Acta Cryst. (1998). D**54**, 547–557

$R_{\rm free}$ and the $R_{\rm free}$ Ratio. I. Derivation of Expected Values of Cross-Validation Residuals Used in Macromolecular Least-Squares Refinement

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 21 July 1997; accepted 16 October 1997)

Abstract

The last five years have seen a large increase in the use of cross validation in the refinement of macromolecular structures using X-ray data. In this technique a test set of reflections is set aside from the working set and the progress of the refinement is monitored by the calculation of a free *R* factor which is based only on the excluded reflections. This paper gives estimates for the ratio of the free *R* factor to the *R* factor calculated from the working set for both unrestrained and restrained refinement. It is assumed that both the X-ray and restraint observations have been weighted correctly and that there is no correlation of errors between the test and working sets. It is also shown that the least-squares weights that minimize the variances of the refined parameters, also approximately minimize the free *R* factor. The estimated free *R*-factor ratios are compared with those reported for structures in the Protein Data Bank.

1. Introduction

One of the problems in macromolecular crystallography is that the crystallographer cannot always be sure that an apparently fully refined structure is free from large systematic errors. The agreement between the model of the molecular structure and the X-ray diffraction data from which it has been derived is measured by the crystallographic *R* factor, but it is well known that structures with acceptable values of this parameter can have significant errors (Brändén & Jones, 1990; Kleywegt & Jones, 1995*a*). The *R* factor is susceptible to manipulation by leaving out weak data or by overfitting the data with too many parameters and so is not a completely reliable guide to accuracy. In small-molecule crystallography, where the number of X-ray intensity observations usually exceeds the number of parameters in the model by at least an order of magnitude, the *R* factor is a more sure guide to both accuracy and precision.

In 1992 Brünger introduced the idea of an $R_{\rm free}$ (Brünger, 1992, 1993), based on the standard statistical modelling technique of jack-knifing or cross-validatory residuals (McCullagh & Nelder, 1983). The $R_{\rm free}$ is the same as the conventional *R* factor, but based on a test set consisting of a small percentage (usually ~5–10%) of reflections excluded from a structure refinement. The remaining reflections included in the refinement are known as the working set. The $R_{\rm free}$ value, unlike the *R* factor, cannot be driven down by refining a false model because the reflections on which it is based are excluded from this process. $R_{\rm free}$ is only expected to decrease during the course of a successful refinement. Consequently, a high value of this statistic and a concomitant low value of *R* may indicate an inaccurate model. The procedure assumes that the reflections removed for the cross-validation test have been randomly selected and have errors uncorrelated with those that remain in the set used in the refinement. This assumption may be partly invalidated by the presence of non-crystallographic symmetry. Ideally, the refinement should be repeated several times, removing non-overlapping sets of reflections each time.

The $R_{\rm free}$ is highly correlated with the phase accuracy of the atomic model (Brünger, 1992, 1993) and can detect various types of errors in the structure including phase errors and partial mistracing of the structure. It has also been used in evaluating different refinement protocols, such as the optimization of the weights used during refinement. It is particularly useful in preventing the overfitting of data (Kleywegt & Brünger, 1996).

Kleywegt & Jones (1995*a*,*b*) have shown that with low-resolution data it is possible to completely mistrace a structure, deliberately tracing it backwards through the density, and still achieve an acceptable *R* factor. The $R_{\rm free}$, on the other hand, could not be duped so easily, and remained at a high value, close to that expected for a random set of scatterers, throughout the refinement.

The use of $R_{\rm free}$ is thus a valuable guide to the process of refinement, particularly for low-resolution data, and its use and publication are widely encouraged. A recent review (Kleywegt & Brünger, 1996) indicated that the use of the measure is becoming more widespread with it being reported in 44% of articles describing macromolecular X-ray structures.

However, the usefulness of $R_{\rm free}$ is limited by the fact that what is an 'acceptable' value is often not evident.

Table 1. *Definitions of symbols*

A superscript T denotes a matrix transpose.

Scalars	
f	The number of structure amplitude observations included in the refinement (the working set)
m	The number of parameters being refined
n	The number of observations, including any restraints, in the refinement
p	The number of observations excluded from refinement (the test set)
r	The number of restraints included in the refinement ($r = n - f$)
w_i	The weight of the ith observation
σ_i	The standard deviation of the ith observation
$\Delta_i = \vert F_{\rm obs}\vert_i - G\vert F_{\rm calc}\vert_i$	An X-ray residual
D	The weighted refinement residual
$D_{\rm inc}$	The contribution to D from the working set of f reflections
$D_{\rm free}$	D based on a test set of p excluded reflections
$D_{\rm rest}$	The contribution to D from the r restraints ($D_{\rm rest} = D - D_{\rm inc}$)
$\vert F\vert$	The structure amplitude
G	The least-squares X-ray scale factor
N_a	The number of atoms being refined
$R = \frac{\sum \vert \vert F_{\rm obs}\vert_i - G\vert F_{\rm calc}\vert_i \vert}{\sum \vert F_{\rm obs}\vert_i}$	The standard R factor
$R_G = \left[\frac{\sum w_i(\vert F_{\rm obs}\vert_i - G\vert F_{\rm calc}\vert_i)^2}{\sum w_i \vert F_{\rm obs}\vert_i^2}\right]^{1/2}$	The generalized R factor
$R_{\rm inc}$ & $R_{G\rm inc}$	R factors based on all reflections in the working set
$R_{\rm free}$ & $R_{G\rm free}$	R factors based on a test set of p excluded reflections
Column matrices	
$\mathbf{a}_i$	The ith row of $\mathbf{A}$
$\mathbf{b}_i$	The ith row of $\mathbf{B}$
$\mathbf{f}$	The n observations employed in the refinement (structure amplitudes and restraints)
$\hat{\mathbf{f}}$	The least-squares estimate of $\mathbf{f}$ calculated at the convergence of the refinement
$\mathbf{g}$	The p excluded observations
$\hat{\mathbf{g}}$	The least-squares estimate of $\mathbf{g}$ calculated from $\hat{\mathbf{x}}$ at the convergence of the refinement
$\hat{\mathbf{x}}$	The least-squares estimate of the m parameters
$\mathbf{\Delta}_{\rm free} = \mathbf{g} - \hat{\mathbf{g}}$	The least-squares residual associated with the excluded observations
Rectangular matrices	
$\mathbf{A}$	The least-squares design matrix of derivatives of order $n \times m$
$\mathbf{B}$	The $p \times m$ matrix analogous to $\mathbf{A}$ but involving the excluded observations
$\mathbf{D}_{\rm free}$	The $p \times p$ variance-covariance matrix of the excluded residuals ($\mathbf{D}_{\rm free} = \langle \mathbf{\Delta}_{\rm free}\mathbf{\Delta}^T_{\rm free}\rangle$)
$\mathbf{H}$	The $m \times m$ normal matrix given by $\mathbf{A}^T\mathbf{W}\mathbf{A}$
$\mathbf{S}$	The $p \times n$ matrix given by $\mathbf{B}\mathbf{H}^{-1}\mathbf{A}^T\mathbf{W}$
$\mathbf{W}$	The $n \times n$ symmetric weight matrix and $\mathbf{W}^{-1}$ is the VCM (variance-covariance matrix) of the included observations. This matrix reflects the random experimental and model errors
$\mathbf{W}_{\rm free}$	The $p \times p$ symmetric weight matrix of $\mathbf{g}$ and $\mathbf{W}^{-1}_{\rm free}$ is the VCM of the excluded observations

One would expect $R_{\rm free}$ to always be higher than R even when there are no systematic errors in the model structure, but is not clear how much higher if should be. At present we merely have a number of rules of thumb (Kleywegt & Brünger, 1996).

Cruickshank has estimated that the expected value of the free R factor (EFRF) is given by

$$\mathrm{EFRF} = [N_{\rm obs}/(N_{\rm obs} - N_{\rm par})]^{1/2}R,$$

where $N_{\rm obs}$ is the number of observations, $N_{\rm par}$ is the number of parameters, and R is the conventional R factor (Dodson *et al.*, 1996). Bacchi *et al.* (1996) use this expression in an extension of the self-validation Hamilton test to assess the significance of any observed drop in $R_{\rm free}$ during refinement.

The need for more understanding of the behaviour of $R_{\rm free}$ was highlighted by Dodson *et al.* (1996). In spite of the enthusiasm for its use, actual applications of $R_{\rm free}$ have remained somewhat subjective without an understanding of its statistical basis. For example, if non-crystallographic symmetry (NCS) constraints are relaxed during a structure refinement, how much should $R_{\rm free}$ rise during subsequent refinement if the restrained model is correct? Without understanding how $R_{\rm free}$ varies as a function of the number of restraints and/or number of parameters it is only possible to make rather subjective judgements.

This paper begins to answer these questions by deriving the expected value of the free residual from which estimates of both $R_{\rm free}$ and the ratio of $R_{\rm free}$ to R are calculated.

2. Theory

2.1. *Error assumptions*

The expected or estimated values of residuals and R factors will be derived on the assumption that the weights used in the structure refinement correctly reflect the errors which include not only experimental errors in measuring X-ray intensities but also errors in the functional form of the structure-factor model which produce random and uncorrelated perturbation in the residuals. These model errors, which may arise from complicated atomic disorder, are an important source of random error in protein structures and are the reason why R factors of refined macromolecular structures are usually higher than their small-molecule counterparts. It is assumed in the derivation of the statistics in this paper that these random errors have been correctly accounted for in the weighting of the X-ray data and of any restraints in the refinement process.

Some model errors, such as the absence of a bulk solvent correction, lead to correlated model errors in reciprocal space. In the theory to be presented in this paper, such correlation could be accommodated if refinement took place with a weight matrix with off-diagonal terms but in practice computational difficulties preclude the use of such matrices in macromolecular refinement. All expressions derived in this paper that use a diagonal weight matrix assume that correlated errors are absent. Model errors such as missing or misplaced atoms are similarly assumed to be absent. On the other hand, no assumptions are made about the completeness or otherwise of the reflection data set.

In order that the model errors in the free reflections are uncorrelated with those reflections used in refinement, the reflections in one set must not be related to those in the other set by crystallographic and non-crystallographic symmetry. No reflection in the free set must be related to one in the main set by pseudosymmetry. Care is needed when selecting reflections from data sets where Bijvoet pairs have been kept separate. Another case arises when there are domains or molecules in the asymmetric unit related by a non-crystallographic axis which is along a rational direction in the crystal lattice (*e.g.* the pseudo-dyads in rhombohedral insulin).

2.2. *Definitions*

For convenience the definitions of symbols commonly used in this paper and its appendices are given in Table 1.

2.3. *The expected value of the free residual*

The algebra for the derivation of the expected value of the free residual is presented in *Appendix A*. There it is first shown that the second moment matrix of residuals corresponding to a test set of observations excluded from least-squares refinement is equal to the sum of the variance–covariance matrix (VCM) of the omitted observations and the VCM of the corresponding quantities calculated from the parameter estimates at the convergence of a refinement. The expected value of the sum of squared residuals associated with the excluded observations is then obtained by taking matrix traces.

The theory then draws on results from an earlier paper (Tickle *et al.*, 1998) where it is shown that when the weighting is on an absolute scale, the expected value of the sum of a subset of s weighted residuals at the convergence of a least-squares refinement is,

$$\left\langle \sum_{i=1}^{s} w_i \Delta_i^2 \right\rangle = s - \sum_{i=1}^{s} w_i \mathbf{a}_i^T \mathbf{H}^{-1} \mathbf{a}_i, \tag{1}$$

where the angled brackets denote statistical expectation.

When the above summation is over all n observations (reflections and restraints),

$$\sum_{i=1}^{n} w_i \mathbf{a}_i^T \mathbf{H}^{-1} \mathbf{a}_i = m, \tag{2}$$

and hence

$$\left\langle \sum_{i=1}^{n} w_i \Delta_i^2 \right\rangle = n - m.$$

In *Appendix A* equation (21) shows that the expected value of the residual associated with p excluded observations in the test set, is given by,

$$\begin{aligned} \langle D_{\rm free} \rangle &= \left\langle \sum_{i=1}^{p} w_i (|F_{\rm obs}|_i - G|F_{\rm calc}|_i)^2 \right\rangle \\ &= p + \sum_{i=1}^{p} w_i \mathbf{b}_i^T \mathbf{H}^{-1} \mathbf{b}_i. \end{aligned} \tag{3}$$

It should be noted that the derivation of this equation does not assume that the test set has been randomly selected from reciprocal space.

We now consider the f structure-amplitude observations *included* in the refinement (the working set). From equation (1) the expected value of the residual associated with these observations at convergence is given by

$$\begin{aligned} \langle D_{\rm inc} \rangle &= \left\langle \sum_{i=1}^{f} w_i (|F_{\rm obs}|_i - G|F_{\rm calc}|_i)^2 \right\rangle \\ &= f - \sum_{i=1}^{f} w_i \mathbf{a}_i^T \mathbf{H}^{-1} \mathbf{a}_i. \end{aligned} \tag{4}$$

The similarities between equations (3) and (4) will be noted. In the linear approximation, both expressions for the expected value are independent of the observations (structure amplitudes and restraints).

Using these results it is possible to obtain estimates of the ratio of $R_{\rm free}$ to R for models with only random uncorrelated errors. This $R_{\rm free}$ ratio is estimated first for unrestrained refinement and then for refinements with geometrical restraints. These $R_{\rm free}$ ratios are the starting point for understanding $R_{\rm free}$ ratios where systematic errors are present.

2.4. *Random exclusion of observations in unrestrained refinement*

The calculation of the expected values of the residuals using equations (3) and (4) is a computationally intensive task in macromolecular refinement, involving the inversion of the normal matrix **H**. In general we need statistics which are more readily available during the refinement process.

We first consider the case of unrestrained refinement. In this case $f = n$ and from equations (2) and (4),

$$\sum_{i=1}^{f} w_i \mathbf{a}_i^T \mathbf{H}^{-1} \mathbf{a}_i = m, \qquad (5)$$

and,

$$\langle D_{\rm inc} \rangle = f - m. \qquad (6)$$

Equation (3) gives the expected value of the free residual based on any given subset of p observations. In order to derive simpler expressions, we must now assume that the test reflections have been randomly chosen from reciprocal space. In this case when p is large we assume that the sums of the quadratic forms are approximately proportional to the number of observations involved. Thus,

$$\sum_{i=1}^{p} w_i \mathbf{b}_i^T \mathbf{H}^{-1} \mathbf{b}_i \simeq (p/f) \sum_{i=1}^{f} w_i \mathbf{a}_i^T \mathbf{H}^{-1} \mathbf{a}_i. \qquad (7)$$

Hence, from equations (3), (5) and (7),

$$\langle D_{\rm free} \rangle \simeq (p/f)(f + m). \qquad (8)$$

From equations (6) and (8) the estimated ratio of the residuals for the case of random uncorrelated errors is

$$\frac{D_{\rm free}}{D_{\rm inc}} \simeq \frac{p(f+m)}{f(f-m)}. \qquad (9)$$

Unlike equations (3) and (4), the above equation is independent of the scale of the weights.

2.5. *The expected value of R_{Gfree} in unrestrained refinement*

The expected values of the residuals derived in the previous section may be used to give estimated values of the generalized R factor, R_G, defined by

$$R_G^2 = \frac{\sum w_i (|F_{\rm obs}|_i - G|F_{\rm calc}|_i)^2}{\sum w_i |F_{\rm obs}|_i^2}.$$

We define $R_{\rm Gfree}^2$ as R_G^2 based on p excluded reflections. If there are only random and uncorrelated errors then the numerator of $R_{\rm Gfree}^2$ may be approximated by $\langle D_{\rm free} \rangle$ and using equation (8) we can write

$$R_{\rm Gfree}^2 \simeq \frac{p(f+m)}{f \sum_{i=1}^{p} w_i |F_{\rm obs}|_i^2}. \qquad (10)$$

We define $R_{\rm Ginc}^2$ as R_G^2 based on all f included reflections and if the weights have been scaled so that the residual in the numerator is equal to its expected value then

$$R_{\rm Ginc}^2 = \frac{f - m}{\sum_{i=1}^{f} w_i |F_{\rm obs}|_i^2}. \qquad (11)$$

2.6. *The ratio of R_{Gfree} to R_G in unrestrained refinement*

From equations (9), (10) and (11) we have

$$\frac{R_{\rm Gfree}^2}{R_{\rm Ginc}^2} \simeq \frac{f D_{\rm free}}{p D_{\rm inc}} \simeq \frac{f+m}{f-m}.$$

Thus, we may write the ratio of the generalized R factors for the case of random uncorrelated errors as

$$\frac{R_{\rm Gfree}}{R_{\rm Ginc}} \simeq \left(\frac{f+m}{f-m}\right)^{1/2}. \qquad (12)$$

These results give the expected ratio of the generalized R factor of the test set to that of the working set at the convergence of an unrestrained refinement in the case where there are only random uncorrelated errors which are correctly reflected in the weights employed. The results depend only on the number of reflections and the number of parameters.

2.7. *R_{Gfree} and the R_{Gfree} ratio in restrained refinement*

In a restrained refinement, such as is typical in macromolecular crystallography, the ratios derived in the previous two sections would only be applicable if $R_{\rm Gfree}$ were calculated from a random selection of residuals including both structure-amplitude observations and restraints. Since R factors are traditionally only based on structure amplitudes, the estimation of $R_{\rm Gfree}$ ratios for restrained refinement requires further analysis.

The number of observations this time is n and f of these are structure amplitudes, the balance consisting of r geometrical, thermal or other restraints which make a contribution $D_{\rm rest}$ to the minimized residual at convergence. From equation (1) we have

$$\langle D_{\rm rest} \rangle = r - \sum_{i=1}^{r} w_i \mathbf{a}_i^T \mathbf{H}^{-1} \mathbf{a}_i, \qquad (13)$$

where the summation is taken over the restraint obser-

Table 2. *Estimates of R_{Gfree} and R_{Gfree} ratios for converged least-squares structure refinements conducted with various constraint regimes*

It is assumed in each case that experimental and model errors are random and uncorrelated and that the weights correctly reflect the model and experimental errors. The last column gives expressions for a refinement where there are r hard constraints, reducing the number of parameters to $m - r$. The constrained expressions are only valid if imposition of the constraints does not invalidate the error assumptions given above.

	No restraints	Restraints	Constraints
$R_{G\text{free}}$	$\left\{\dfrac{p(f+m)}{f\sum_{i=1}^{p} w_i\lvert F_{\text{obs}}\rvert_i^2}\right\}^{1/2}$	$\left\{\dfrac{p[f+(m-r+D_{\text{rest}})]}{f\sum_{i=1}^{p} w_i\lvert F_{\text{obs}}\rvert_i^2}\right\}^{1/2}$	$\left\{\dfrac{p(f+m-r)}{f\sum_{i=1}^{p} w_i\lvert F_{\text{obs}}\rvert_i^2}\right\}^{1/2}$
$\dfrac{R_{G\text{free}}}{R_G}$	$\left\{\dfrac{f+m}{f-m}\right\}^{1/2}$	$\left\{\dfrac{f+(m-r+D_{\text{rest}})}{f-(m-r+D_{\text{rest}})}\right\}^{1/2}$	$\left\{\dfrac{f+(m-r)}{f-(m-r)}\right\}^{1/2}$

vations. Using this equation and equation (2) the summation over the structure-amplitude observations can be written

$$\begin{aligned}\sum_{i=1}^{f} w_i\mathbf{a}_i^T\mathbf{H}^{-1}\mathbf{a}_i &= m - \sum_{i=1}^{r} w_i\mathbf{a}_i^T\mathbf{H}^{-1}\mathbf{a}_i\\ &= m - r + \langle D_{\text{rest}}\rangle. \qquad (14)\end{aligned}$$

This result and equations (3) and (7) give the following approximation for $\langle D_{\text{free}}\rangle$

$$\langle D_{\text{free}}\rangle \simeq (p/f)(f + m - r + D_{\text{rest}}).$$

Similarly an approximation to $\langle D_{\text{inc}}\rangle$ can be derived from equations (4), (7) and (14),

$$\begin{aligned}\langle D_{\text{inc}}\rangle &= f - \sum_{i=1}^{f} w_i\mathbf{a}_i^T\mathbf{H}^{-1}\mathbf{a}_i\\ &\simeq f - (m - r + D_{\text{rest}})\end{aligned}$$

Hence, an estimate of $R_{G\text{free}}$ at the convergence of a correctly weighted restrained refinement with only random uncorrelated errors is

$$R_{G\text{free}} \simeq \left\{\frac{p[f + (m - r + D_{\text{rest}})]}{f\sum_{i=1}^{p} w_i\lvert F_{\text{obs}}\rvert_i^2}\right\}^{1/2}.$$

The estimated ratio of the free residual to the included residual is given by,

$$\frac{D_{\text{free}}}{D_{\text{inc}}} \simeq (p/f)\left[\frac{f + (m - r + D_{\text{rest}})}{f - (m - r + D_{\text{rest}})}\right],$$

and hence,

$$\frac{R_{G\text{free}}}{R_{G\text{inc}}} \simeq \left[\frac{f + (m - r + D_{\text{rest}})}{f - (m - r + D_{\text{rest}})}\right]^{1/2}. \qquad (15)$$

2.8. *Minimum variance weights minimize R_{free}*

The expected values derived above are only applicable for correctly weighted least-squares refinements. However, R_{free} has also been used to optimize the weighting of geometrical or temperature-factor terms in refinement by adjusting the weights so as to minimize R_{free} (Brünger, 1992, 1993). It is, therefore, of interest to enquire how R_{free} responds to variations in weighting.

Appendix C shows that the weights $\mathbf{W}$ which correctly reflect experimental and model errors, minimize the variance of both the refined parameters $\hat{\mathbf{x}}$, and also the expected value of the sum of the squares of the unweighted residuals in the test set. Hence, the choice of these weights approximately minimizes R_{free}. One method of estimating such weights has been described by Tickle *et al.* (1998).

2.9. *The use of standard R factors*

The R-factor expressions used so far in this paper have been based on the generalized R factor, R_G. However, identical expressions can be derived for the estimated values of the standard free R factor, R_{free}, and the standard R_{free} ratio. For example, in *Appendix B* the following estimate of the R_{free} ratio in unrestrained refinement is derived which is based on standard R factors,

$$\frac{R_{\text{free}}}{R_{\text{inc}}} \simeq \frac{f + m/2}{f - m/2}.$$

The numerator and denominator in the above equation are first-order binomial approximations to square roots and, therefore, another estimate of the standard R_{free} ratio is

$$\frac{R_{\text{free}}}{R_{\text{inc}}} \simeq \left(\frac{f + m}{f - m}\right)^{1/2}.$$

The right-hand side is now the same as for the $R_{G\text{free}}$ ratio [equation (12)]. These expressions may be compared with the estimate used by Bacchi *et al.* (1996) which in our notation is

$$\frac{R_{\text{free}}}{R_{\text{inc}}} \simeq \left(\frac{f}{f - m}\right)^{1/2}.$$

The derivation of simple estimates for the standard

R_{free} and its ratio requires the assumption that the variances of all structure amplitudes are equal. The use of generalized R factors, whose estimates do not require this assumption, is to be preferred.

3. Discussion

3.1. *Restrained and unrestrained R_{Gfree} ratios*

R_{free} has been employed (Kleywegt & Brünger, 1996) to detect situations where addition of extra parameters to a refinement brings no significant improvement in the model. For example the extra parameters may concern the relaxation of non-crystallographic symmetry or the employment of individual atomic temperature factors. Relevant estimates of $R_{G\text{free}}$ are summarized in Table 2 and help to make these discussions more quantitative. This table gives the more important expressions derived earlier for a refinement at convergence with only random uncorrelated errors and weighting which correctly reflects the model and experimental errors.

The constrained and unrestrained statistics in Table 2 show what happens when constraints are imposed, assuming that the reduction in parameters does not introduce model error. It is seen that a reduction in the expected values of $R_{G\text{free}}$ and the $R_{G\text{free}}$ ratio takes place. A reduction also takes place when restraints are imposed, assuming again that the above conditions are not violated.

3.2. *The meaning of the R_{free} ratio*

The estimated R_{free} ratios derived in this paper are the values that should be achievable at the end of a structure refinement when only random uncorrelated errors exist in data and model provided that the observations have been properly weighted (see below).

A larger $R_{G\text{free}}$ ratio than that predicted by these formulae may indicate that parameter shifts have taken

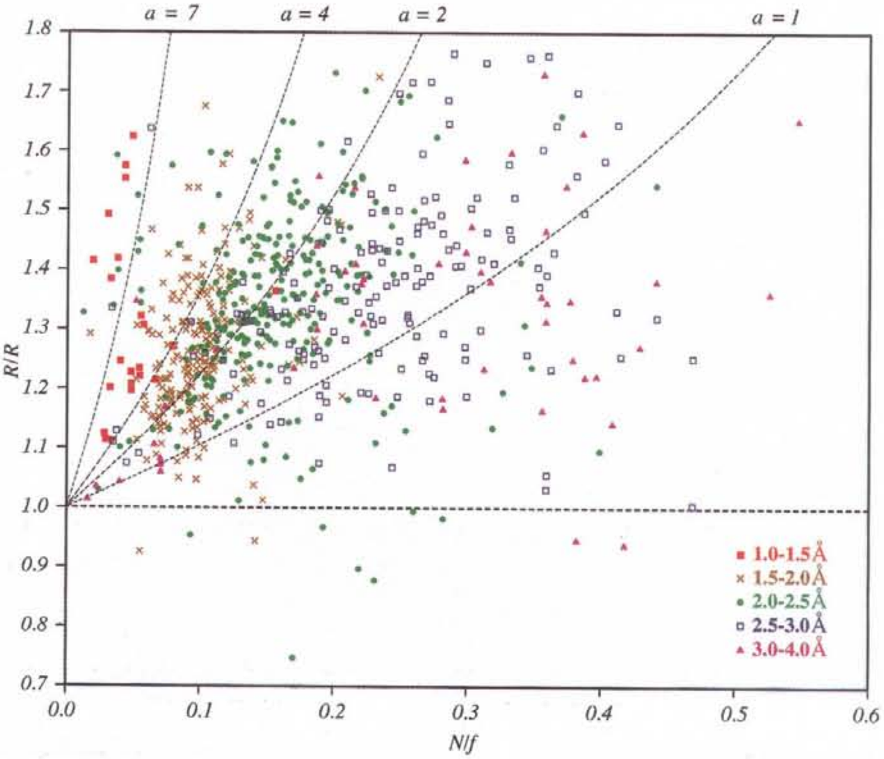

Fig. 1. Plot of the R_{free}/R ratio as a function of the ratio N_a/f for 725 macromolecular structures in the Protein Data Bank, where N_a is the number of atoms included in the refinement and f the number of reflections used. The data points are colour coded according to their resolution range, as shown in the key on the bottom right of the graph. The high-resolution data points tend to be close to the vertical axis with the other points tending to spread further to the right the lower their resolution, as might be expected. Also shown are four dotted curves corresponding to different values of the variable a. The curves shown are for: $a = 1$ which corresponds to three parameters per atom (*i.e.* restrained refinement of atomic coordinates only, plus an overall temperature factor); $a = 2$ is for four parameters per atom (restrained refinement of coordinates plus individual isotropic temperature factors); $a = 4$ represents unrestrained refinement with four parameters per atom; and $a = 7$ represents nine parameters per atom (restrained anisotropic refinement).

place which have minimized the residual without significantly improving the model. This may arise when errors in the model are sufficiently large for the refinement to descend into a false minimum.

A smaller $R_{G\mathrm{free}}$ ratio than that predicted by these formulae may indicate that the refinement has not reached convergence since the initial value of the ratio immediately after the division of the data into a working set and a test set will be approximately unity. Interpretation requires care since a wrong model which has not been fully minimized against the data may produce the same ratio as a fully minimized correct model.

In macromolecular refinement, model error is usually the major contributor to R and R_{free} at the end of the refinement. Paucity of diffraction data means that thermal and static disorder in the more mobile parts of the molecule cannot be accurately modelled. These model errors may cause random perturbations in the structure-amplitude residuals which are indistinguishable from random experimental errors. The $R_{G\mathrm{free}}$ ratio will not be affected by the magnitude of these errors provided that the latter have independent effects on the included and excluded residuals.

3.3. *Observed R_{free} ratios from the Protein Data Bank*

We have examined R_{free} ratios for crystal structures in the Protein Data Bank (Bernstein *et al.*, 1977) as on 1 June 1997. Fig. 1 shows a plot of the R_{free} ratio as a function of N_a/f, where N_a is the number of atoms included in the refinement and f is the number of reflections used, for 725 macromolecular structures for which all these values are reported. The points are colour coded according to resolution range.

We define the R_{free} ratio as $y = R_{\mathrm{free}}/R$ and $x = N_a/f$. Values of y range from about 0.8 to 1.8. By substituting $a = (m - r + D_{\mathrm{rest}})/N_a$ into equation (15), we have

$$y = \left(\frac{1 + ax}{1 - ax}\right)^{1/2}. \qquad (16)$$

Fig. 1 shows the curves corresponding to equation (16) for different values of a.

In order to make the comparison between experimental and theoretical values easier, a function of y was sought which is a linear function of x. By squaring and

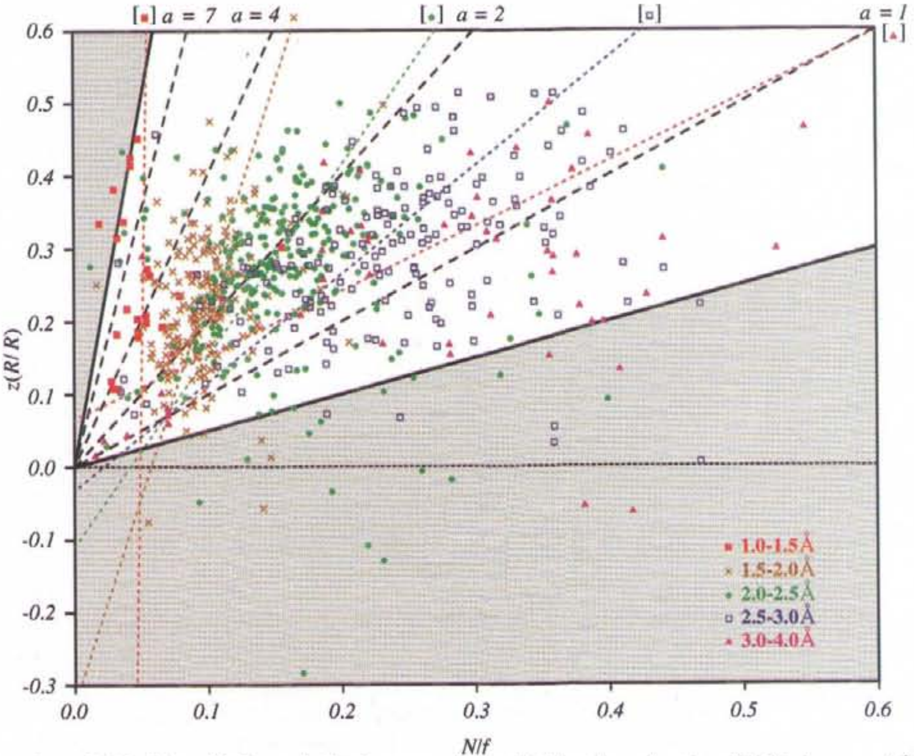

Fig. 2. Plot of the same data as in Fig. 1 but with the vertical axis representing z which is a linear function of N_a/f, where $z = (y^2 - 1)/(y^2 + 1)$ and $y = R_{\mathrm{free}}/R$. Here the curves from Fig. 1, which corresponded to different values of a, are now straight lines. Again, the data points are colour coded according to resolution range, as shown in the key on the right. Also shown are five coloured lines representing least-squares lines fitted to the data points in each of these five different resolution ranges. The lines are identified by the appropriate point markers in square brackets outside the graph border. The grey regions show the regions from which the data points have been excluded in the least-squares lines calculations, namely points outside the sector bounded by the lines corresponding to $a = 10$ and $a = 0.5$.

rearranging the terms in equation (16) we arrive at

$$z = ax,$$

where

$$z = (y^2 - 1)/(y^2 + 1).$$

Fig. 2 shows a plot of z against x where the points are colour coded as in Fig. 1. The coloured straight lines in Fig. 2 are least-squares lines fitted to the data points in the particular resolution range represented by points of the same colour. For example, the pink triangular points represent data between 3 and 4 Å resolution and the pink line is the least-squares line through the pink triangular points. The pecked black lines emanating from the origin in Fig. 2 are plots of $z = ax$ for the same values of a as shown in Fig. 1

We requested information from some of the authors whose structures were outliers in Fig. 2. It became apparent that very unusual $R_{\rm free}$ ratios are normally not the result of careful refinement protocols. The coloured lines were, therefore, plotted ignoring the points in the darker regions outside the sector bounded by the black lines of slopes 0.5 and 10. The choice of these slopes as cutoffs was somewhat arbitrary but the removal of these outliers caused the coloured lines to pass nearer to the origin.

The plots of $z = ax$ represent refinement regimes with different numbers of parameters per atom. The gradients of the lines (a) increase with the number of parameters per atom. In the absence of relevant information in the Data Bank, it was assumed that in restrained refinements, $r = 2.5N_a$ and $D_{\rm rest} = r/5$. These estimates ignore temperature-factor restraints (if any) because our survey of the latter revealed widely different restraint protocols. Using these values, for restrained refinements

$$a = (m/N_a) - 2.$$

It can be seen that the $z = 2x$ line (isotropic temperature factors) passes through the constellations of orange crosses and pink triangular points, representing structures between 2.5 and 1.5 Å resolution and is close to the green pecked line (2.5–2.0 Å data). Similarly the $z = x$ line (overall temperature factor) lies close to the pink line which is fitted to the 4–3 Å data. Even in the absence of details of restraint procedures, the $z = ax$ lines can be seen to pass through areas of the plot where the particular refinement regime is most relevant. The large spread of values about the straight lines is unlikely to be solely a statistical effect and may well say something about the quality of the refinements.

Comparison of the lines $z = 2x$ and $z = 4x$, which differ only in respect of restraints, shows how restraints lower the $R_{\rm free}$ ratio. Non-crystallographic symmetry (NCS, see *Introduction*) might give rise to lower than predicted $R_{\rm free}$ ratios. However, a check on structures in our plots which exhibit NCS did not reveal any obvious systematic effects.

4. Conclusions

Values of $R_{\rm free}$ are affected by all types of error in the model and the data. The $R_{\rm free}$ ratio, however, is independent of random errors and provides a statistic which can be compared with its theoretically estimated value and used to detect systematic model or weighting errors at the convergence of least-squares refinement. However, achievement of a theoretical value of the ratio is not by itself proof of the correctness of the model or of the quality of the refinement. Nevertheless is would still be helpful if refinement programs printed out the calculated and estimated value of the $R_{\rm free}$ ratio using the expressions shown in Table 2. This would encourage a better understanding of $R_{\rm free}$ than exists at present. Calculation of the observed and theoretical values of these ratios has already been implemented in the refinement program *RESTRAIN* (Driessen *et al.*, 1989).

At low resolution the number of data excluded for cross validation may be small and in these circumstances the precision of free residuals is important. This will be the subject of part II of this work.

APPENDIX A

A1. *Statistical properties of free residuals. Derivation of the expected value of the free residual*

Here we derive the statistical expectation of the free residual $\langle D_{\rm free}\rangle$, at the convergence of a least-squares refinement. The normal equations of least-squares refinement at convergence can be written

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

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

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

The structure amplitudes and target distances corresponding to the excluded observations in a test set can be expressed in terms of the parameter estimates by the truncated Taylor expansion,

$$\begin{aligned}\hat{\mathbf{g}} - \langle\mathbf{g}\rangle &= \mathbf{B}(\hat{\mathbf{x}} - \langle\hat{\mathbf{x}}\rangle)\\ &= \mathbf{B}\mathbf{H}^{-1}\mathbf{A}^T\mathbf{W}(\mathbf{f} - \langle\mathbf{f}\rangle)\\ &= \mathbf{S}(\mathbf{f} - \langle\mathbf{f}\rangle). \qquad (17)\end{aligned}$$

Thus, the column of residuals of the excluded observations is given by

$$\begin{aligned}\Delta_{\rm free} &= \mathbf{g} - \hat{\mathbf{g}}\\ &= \mathbf{g} - \langle\mathbf{g}\rangle - \mathbf{S}(\mathbf{f} - \langle\mathbf{f}\rangle).\end{aligned}$$

Assuming that the errors in **f** and **g** are uncorrelated, the

VCM of the residuals associated with the excluded observations $\mathbf{D}_{\rm free} = \langle \boldsymbol{\Delta}_{\rm free}\boldsymbol{\Delta}_{\rm free}^{T}\rangle$ is given by

$$\begin{aligned}\mathbf{D}_{\rm free} &= \left\langle(\mathbf{g}-\langle\mathbf{g}\rangle)(\mathbf{g}-\langle\mathbf{g}\rangle)^{T}\right\rangle \\ &\quad + \left\langle\mathbf{S}(\mathbf{f}-\langle\mathbf{f}\rangle)(\mathbf{f}-\langle\mathbf{f}\rangle)^{T}\mathbf{S}^{T}\right\rangle \\ &= \mathbf{W}_{\rm free}^{-1} + \mathbf{S}\mathbf{W}^{-1}\mathbf{S}^{T}. \qquad (18)\end{aligned}$$

Thus, $\mathbf{D}_{\rm free}$ is equal to the sum of the VCM of the excluded observations and the VCM of the corresponding quantities calculated from the refined parameters. From equation (18) it follows that,

$$\mathbf{D}_{\rm free}\mathbf{W}_{\rm free} = \mathbf{I}_{p} + \mathbf{S}\mathbf{W}^{-1}\mathbf{S}^{T}\mathbf{W}_{\rm free}, \qquad (19)$$

where $\mathbf{I}_p$ is a unit matrix of order p. Using the fact that

$$\mathbf{H} = \mathbf{A}^{T}\mathbf{W}\mathbf{A},$$

and noting that the trace of a product of matrices is invariant under a cyclic permutation of the order of matrix multiplication, we take the trace of both sides of equation (19) to give

$$\begin{aligned}{\rm tr}(\mathbf{D}_{\rm free}\mathbf{W}_{\rm free}) &= p + {\rm tr}(\mathbf{S}\mathbf{W}^{-1}\mathbf{S}^{T}\mathbf{W}_{\rm free}) \\ &= p + {\rm tr}(\mathbf{B}\mathbf{H}^{-1}\mathbf{A}^{T}\mathbf{W}\mathbf{A}\mathbf{H}^{-1}\mathbf{B}^{T}\mathbf{W}_{\rm free}) \\ &= p + {\rm tr}(\mathbf{W}_{\rm free}\mathbf{B}\mathbf{H}^{-1}\mathbf{B}^{T}). \qquad (20)\end{aligned}$$

If the p excluded observations are structure amplitudes and we assume that the weight matrix is diagonal, then equation (20) can be written as

$$\begin{aligned}\langle D_{\rm free}\rangle &= \left\langle\sum_{i=1}^{p} w_i(|F_{\rm obs}|_i - G|F_{\rm calc}|_i)^2\right\rangle \\ &= p + \sum_{i=1}^{p} w_i\mathbf{b}_i^{T}\mathbf{H}^{-1}\mathbf{b}_i \qquad (21)\end{aligned}$$

where the angle brackets denote statistical expectation, $\langle D_{\rm free}\rangle$ is the expected value of the residual associated with the given p excluded observations and $\mathbf{b}_i$ is the ith row of $\mathbf{B}$. The expected value of a single weighted excluded residual in the summation is obtained by taking a single diagonal term from equation (19) which gives

$$\langle w_i(|F_{\rm obs}|_i - G|F_{\rm calc}|_i)^2\rangle = 1 + w_i\mathbf{b}_i^{T}\mathbf{H}^{-1}\mathbf{b}_i.$$

Two notable assumptions have been made in the above analysis. First, it has been assumed that the test-set residuals are uncorrelated with those in the working set. Equation (18) is invalid if this assumption is not true.

Second, it is assumed that the refinement has used a weight matrix $\mathbf{W}$ which correctly reflects the experimental and model errors. Equation (21) is invalid if this assumption is not true. When a diagonal weight matrix is used, as is almost always the case in practice, correlated errors in reciprocal space will not be correctly represented. In *Appendix C* it is shown that the use of weight matrices which do not correctly account for the errors will lead to test-set residuals with larger variance–covariance matrices.

APPENDIX *B*

*B*1. *The ratio of R_{free} to R in unrestrained refinement*

The derivations in the body of this paper have been given in terms of the generalized R factor. In this appendix we derive the expected values of residuals expressed as the sum of unweighted absolute differences. Hence, we obtain an estimate of the $R_{\rm free}$ ratio expressed in terms of the standard R factor which is defined as

$$R = \frac{\sum \big||F_{\rm obs}|_i - G|F_{\rm calc}|_i\big|}{\sum |F_{\rm obs}|_i}.$$

The variance of an included residual $\big||F_{\rm obs}|_i - G|F_{\rm calc}|_i\big|$ at the convergence of a refinement (see *Appendix B* in Tickle *et al.*, 1998) can be written as

$$\langle(|F_{\rm obs}|_i - G|F_{\rm calc}|_i)^2\rangle = \sigma_i^2 - \mathbf{a}_i^{T}\mathbf{H}^{-1}\mathbf{a}_i.$$

To make further progress we need to assume a distribution for each residual $|F_{\rm obs}|_i - G|F_{\rm calc}|_i$. If a normal distribution is assumed, we can write

$$\left\langle\big||F_{\rm obs}|_i - G|F_{\rm calc}|_i\big|\right\rangle = (2/\pi)\left\langle(|F_{\rm obs}|_i - G|F_{\rm calc}|_i)^2\right\rangle^{1/2}, \qquad (22)$$

and hence,

$$\left\langle\sum_{i=1}^{f}\big||F_{\rm obs}|_i - G|F_{\rm calc}|_i\big|\right\rangle = (2/\pi)\sum_{i=1}^{f}(\sigma_i^2 - \mathbf{a}_i^{T}\mathbf{H}^{-1}\mathbf{a}_i).$$

The square root in the above summation may be expanded binomially to the first order giving

$$\begin{aligned}&\left\langle\sum_{i=1}^{f}\big||F_{\rm obs}|_i - G|F_{\rm calc}|_i\big|\right\rangle \\ &\qquad \simeq (2/\pi)\sum_{i=1}^{f}\sigma_i(1 - w_i\mathbf{a}_i^{T}\mathbf{H}^{-1}\mathbf{a}_i/2).\end{aligned}$$

We now assume that σ_i is the same for all observations so that we can write

$$\left\langle \sum_{i=1}^{f} \big| |F_{\rm obs}|_i - G|F_{\rm calc}|_i \big| \right\rangle$$
$$\simeq (2\sigma/\pi) \sum_{i=1}^{f} (1 - w\mathbf{a}_i^T \mathbf{H}^{-1} \mathbf{a}_i/2).$$

Using equation (5) to simplify the right-hand side, we can write

$$\left\langle \sum_{i=1}^{f} \big| |F_{\rm obs}|_i - G|F_{\rm calc}|_i \big| \right\rangle \simeq \frac{2\sigma(f - m/2)}{\pi}. \qquad (23)$$

By analogous reasoning the sum over the p excluded reflections can be approximated as

$$\left\langle \sum_{i=1}^{p} \big| |F_{\rm obs}|_i - G|F_{\rm calc}|_i \big| \right\rangle \simeq \frac{2\sigma p(f + m/2)}{\pi f}. \qquad (24)$$

We have now derived approximations for the numerators of $R_{\rm inc}$ and $R_{\rm free}$. Their denominators are, on average, proportional to the number of reflections in the respective summations. Hence, dividing equation (24) by equation (23) and multiplying by f/p we have

$$\frac{R_{\rm free}}{R_{\rm inc}} \simeq \frac{f + m/2}{f - m/2}.$$

APPENDIX *C*

*C*1. *Minimum variance weights minimize the expected value of R_{free}*

First we show that the least-squares refinement using a weight matrix $\mathbf{W}$ which is the inverse of the VCM of the observations, minimizes the VCM's of both the refined parameters $\hat{\mathbf{x}}$ and the free residuals $\mathbf{g} - \hat{\mathbf{g}}$.

Consider a function $\mathbf{t}$ which is a function of the refined least-squares parameters which is linear to within a first-order Taylor approximation,

$$\delta\mathbf{t} = \mathbf{L}\delta\mathbf{x}.$$

Now consider two column matrices $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$, defined below, which are both unbiased estimates of $\mathbf{t}$.

$$\hat{\mathbf{u}} - \langle \hat{\mathbf{u}} \rangle = \mathbf{P}(\mathbf{f} - \langle \mathbf{f} \rangle) \qquad (25)$$

$$\hat{\mathbf{v}} - \langle \hat{\mathbf{v}} \rangle = \mathbf{Q}(\mathbf{f} - \langle \mathbf{f} \rangle), \qquad (26)$$

where,

$$\mathbf{P} = \mathbf{L}\mathbf{H}^{-1}\mathbf{A}^T\mathbf{W}, \qquad (27)$$

$$\mathbf{Q} = \mathbf{L}\mathbf{H}_{\rm u}^{-1}\mathbf{A}^T\mathbf{U}. \qquad (28)$$

$\mathbf{U}$ is any weight matrix and $\mathbf{H}_{\rm u} = \mathbf{A}^T\mathbf{U}\mathbf{A}$. From equations (27) and (28) and from the definitions of $\mathbf{H}$ and $\mathbf{H}_{\rm u}$,

$$\mathbf{P}\mathbf{A} = \mathbf{Q}\mathbf{A}. \qquad (29)$$

We wish to show that the VCM of $\hat{\mathbf{u}}$ is smaller than that of $\hat{\mathbf{v}}$. Because $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ are unbiased estimators, $\langle \hat{\mathbf{u}} \rangle = \langle \hat{\mathbf{v}} \rangle$ and thus the VCM of $\hat{\mathbf{v}}$ can be expressed as

$$\begin{aligned}
&\langle (\hat{\mathbf{v}} - \langle \hat{\mathbf{v}} \rangle)(\hat{\mathbf{v}} - \langle \hat{\mathbf{v}} \rangle)^T \rangle \\
&= \langle (\hat{\mathbf{u}} - \langle \hat{\mathbf{u}} \rangle + \hat{\mathbf{v}} - \hat{\mathbf{u}})(\hat{\mathbf{u}} - \langle \hat{\mathbf{u}} \rangle + \hat{\mathbf{v}} - \hat{\mathbf{u}})^T \rangle \\
&= \langle (\hat{\mathbf{u}} - \langle \hat{\mathbf{u}} \rangle)(\hat{\mathbf{u}} - \langle \hat{\mathbf{u}} \rangle)^T \rangle + \langle (\hat{\mathbf{v}} - \hat{\mathbf{u}})(\hat{\mathbf{v}} - \hat{\mathbf{u}})^T \rangle \\
&\quad + \langle (\hat{\mathbf{u}} - \langle \hat{\mathbf{u}} \rangle)(\hat{\mathbf{v}} - \hat{\mathbf{u}})^T \rangle + \langle (\hat{\mathbf{v}} - \hat{\mathbf{u}})(\hat{\mathbf{u}} - \langle \hat{\mathbf{u}} \rangle)^T \rangle. \qquad (30)
\end{aligned}$$

The last two terms of the above equation are the transpose of each other and are each zero matrices as shown by the following analysis which uses equations (25), (26), (27) and (29).

$$\begin{aligned}
\langle (\hat{\mathbf{v}} - \hat{\mathbf{u}})(\hat{\mathbf{u}} - \langle \hat{\mathbf{u}} \rangle)^T \rangle &= \langle (\mathbf{Q} - \mathbf{P})(\mathbf{f} - \langle \mathbf{f} \rangle)(\mathbf{f} - \langle \mathbf{f} \rangle)^T \mathbf{P}^T \rangle \\
&= (\mathbf{Q} - \mathbf{P})\mathbf{W}^{-1}\mathbf{W}\mathbf{A}\mathbf{H}^{-1}\mathbf{L}^T \\
&= (\mathbf{Q} - \mathbf{P})\mathbf{A}\mathbf{H}^{-1}\mathbf{L}^T \\
&= 0.
\end{aligned}$$

Hence from equation (30),

$$\langle (\hat{\mathbf{v}} - \langle \hat{\mathbf{v}} \rangle)(\hat{\mathbf{v}} - \langle \hat{\mathbf{v}} \rangle)^T \rangle = \langle (\hat{\mathbf{u}} - \langle \hat{\mathbf{u}} \rangle)(\hat{\mathbf{u}} - \langle \hat{\mathbf{u}} \rangle)^T \rangle + \langle (\hat{\mathbf{v}} - \hat{\mathbf{u}})(\hat{\mathbf{v}} - \hat{\mathbf{u}})^T \rangle.$$

Since the VCM's are positive definite

$$\langle (\hat{\mathbf{u}} - \langle \hat{\mathbf{u}} \rangle)(\hat{\mathbf{u}} - \langle \hat{\mathbf{u}} \rangle)^T \rangle < \langle (\hat{\mathbf{v}} - \langle \hat{\mathbf{v}} \rangle)(\hat{\mathbf{v}} - \langle \hat{\mathbf{v}} \rangle)^T \rangle.$$

Thus, the VCM of $\hat{\mathbf{u}}$ which is calculated with $\mathbf{W}$ is less than the VCM of $\hat{\mathbf{u}}$ which is calculated with another weight matrix $\mathbf{U}$. Making the substitution $\hat{\mathbf{u}} = \hat{\mathbf{x}}$ and setting $\mathbf{L}$ to a unit matrix, this analysis shows that by using the weight matrix $\mathbf{W}$, we minimize the variance of $\hat{\mathbf{x}}$. Substituting $\hat{\mathbf{u}} = \hat{\mathbf{g}}$ and $\mathbf{L} = \mathbf{B}$, the same analysis shows that $\mathbf{W}$ also minimizes the VCM of $\hat{\mathbf{g}}$. From equation (18) the VCM of the residuals associated with the excluded observations $\mathbf{D}_{\rm free}$ is the sum of the constant matrix $\mathbf{W}_{\rm free}^{-1}$ and the VCM of $\hat{\mathbf{g}}$. Hence, $\mathbf{D}_{\rm free}$ and its trace are also minimized by choosing $\mathbf{W}$ as the weight matrix. The trace of $\mathbf{D}_{\rm free}$ is the expected value of the unweighted sum of squared residuals,

$$\left\langle \sum_{i=1}^{p} (|F_{\rm obs}|_i - G|F_{\rm calc}|_i)^2 \right\rangle,$$

where the summation is taken over the p reflections in the test set. By using the normal approximation in equation (22) we can say that the sum of absolute differences

$$\left\langle \sum_{i=1}^{p} \big| |F_{\rm obs}|_i - G|F_{\rm calc}|_i \big| \right\rangle,$$

and hence $R_{\rm free}$ are approximately minimized by choosing $\mathbf{W}$ as the least-squares weight matrix.

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

References

Bacchi, A., Lamzin, V. S. & Wilson, K. S. (1996). *Acta Cryst.* D**52**, 641–646.

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.

Brändén, C.-I. & Jones, T. A. (1990). *Nature (London)*, **343**, 687–689.

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

Brünger, A. T. (1993). *Acta Cryst.* D**49**, 24–36

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

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

Kleywegt, G. J. & Brünger, A. T. (1996). *Structure*, **4**, 897–904.

Kleywegt, G. J. & Jones, T. A. (1995*a*). *Structure*, **3**, 535–540.

Kleywegt, G. J. & Jones, T. A. (1995*b*). *Braille for pugilists*. In *Making the Most of Your Model*, edited by W. N. Hunter, J. M. Thornton & S. Bailey, pp. 11–24. Warrington: Daresbury Laboratory.

McCullagh, P. & Nelder, J. A. (1983). *Generalised Linear Models*, pp. 209–221. London: Chapman & Hall.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *Acta Cryst.* D**54**, 243–252.