Acta Crystallographica Section D
Biological
Crystallography

ISSN 0907-4449

R_{free} and the R_{free} ratio. II. Calculation of the expected values and variances of cross-validation statistics in macromolecular least-squares refinement

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

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

R_{free} and the R_{free} ratio. II. Calculation of the expected values and variances of cross-validation statistics 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

Correspondence e-mail: d.moss@bbk.ac.uk

Received 25 May 1999
Accepted 27 January 2000

The free R factor is used routinely as a cross-validation tool in macromolecular crystallography. However, without any means of deriving quantitative estimates of its expected value and variance, its application has been rather subjective and its usefulness therefore somewhat limited. In the first part of this series, estimates of the expected value of the ratio of the free R factor to the standard R factor at the convergence of the structure refinement were given. Here, estimates of the variance of this ratio are given and are compared with the observed deviations from the expected values for a selection of refined structures. It is discussed how errors in the functional form of the structure-factor model as well as other types of errors might influence this ratio.

1. Introduction

Macromolecular structure refinement from diffraction data presents a particular challenge to the crystallographer when there is a small parameter-to-observation ratio. In such cases, it may be possible to drive down an R factor without improving the model structure (Brändén & Jones, 1990; Kleywegt & Jones, 1995). In order to combat such problems, Brünger introduced the idea of an R_{free} (Brünger, 1992, 1993) based on the standard statistical modelling technique of jack-knifing or cross-validatory statistics (McCullagh & Nelder, 1989). The R_{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.

In an earlier paper (Tickle *et al.*, 1998*b*), we reviewed the development of the use of R_{free} as a cross-validation tool in macromolecular crystallography and noted that the behaviour of this statistic had yet to be put on a firm theoretical footing. The need for more understanding of the behaviour of R_{free} has also been highlighted by Dodson *et al.* (1996). As a first step towards this goal, we derived the expected values of cross-validation statistics arising from a test set of reflections excluded from a converged least-squares refinement of a crystal structure. We assumed that there were only random experimental and model errors and that these errors were reflected in the weighting of the observations and restraints. In particular, we introduced the R_{free} ratio, which is the ratio of R_{free} (based on the test set) to the standard R factor (based on the working set). We derived the expected value of this ratio and noted that the derivation requires the assumption that the variances of all structure amplitudes are equal. The use of the generalized R factors avoids this assumption and is to be

442 Tickle *et al.* • R_{free} and the R_{free} ratio *Acta Cryst.* (2000). D**56**, 442–450

Table 1
Definitions and abbreviations.

Scalars	
a	d/N_a
d	$m - r + D_{\rm rest}$
D	$\sum w_i(\lvert F_{\rm obs}\rvert_i - G\lvert F_{\rm calc}\rvert_i)^2$, the weighted sum of squares of residuals (deviance)
$D_{\rm inc}$	D based on a working set of f reflections
$D_{\rm free}$	D based on a test set of p excluded reflections
$D_{\rm ratio}$	$(fD_{\rm free}/pD_{\rm inc})^{1/2}$
$D_{\rm rest}$	D based on the geometric and other restraints
$D_{\rm total}$	$D_{\rm inc} + D_{\rm rest}$
f	The number of structure amplitude observations included in the refinement in the working set
m	The number of parameters being refined
n	The number of observations, including any restraints, in the refinement
N_a	The number of atoms in the refinement
p	The number of observations excluded from refinement (the test set)
r	The number of restraints included in the refinement ($r = n - f$)
R	$\sum \big\lvert \lvert F_{\rm obs}\rvert_i - G\lvert F_{\rm calc}\rvert_i \big\rvert / \sum \lvert F_{\rm obs}\rvert_i$, the standard R factor
R_G	$[\sum w_i(\lvert F_{\rm obs}\rvert_i - G\lvert F_{\rm calc}\rvert_i)^2 / \sum w_i \lvert F_{\rm obs}\rvert_i^2]^{1/2}$, 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
$R_{G{\rm ratio}}$	$R_{G{\rm free}}/R_{G{\rm inc}}$
w_i	The weight of the ith observation
Σ	$\sum w_i \lvert F_{\rm obs}\rvert_i^2$
$\Sigma_{\rm inc}$	Σ based on a working set of f reflections
$\Sigma_{\rm free}$	Σ 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 target distances)
$\hat{\mathbf{f}}$	The least-squares estimates of $\mathbf{f}$
$\mathbf{g}$	The p excluded observations (structure amplitudes and/or target distances)
$\hat{\mathbf{g}}$	The estimates of $\mathbf{g}$ calculated from $\hat{\mathbf{x}}$
$\hat{\mathbf{x}}$	The least-squares estimates of the m parameters
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 $p \times p$ VCM of the unweighted residuals ($\mathbf{g} - \hat{\mathbf{g}}$) in the test set
$\mathbf{H}$	The $m \times m$ normal matrix given by $\mathbf{A}^T\mathbf{WA}$
$\mathbf{I}_p$	A $p \times p$ unit matrix
$\mathbf{Q}$	The $p \times p$ symmetric matrix given by $\mathbf{W}_{\rm free}^{1/2}\mathbf{BH}^{-1}\mathbf{B}^T\mathbf{W}_{\rm free}^{1/2}$
$\mathbf{W}$	The $n \times n$ symmetric weight matrix of $\mathbf{f}$
$\mathbf{W}_{\rm free}$	The $p \times p$ symmetric weight matrix of $\mathbf{g}$
Σ	The VCM of the experimental and model errors
$\Sigma_{\rm free}$	Σ of the excluded observations only
Abbreviations	
tr	The trace of a square matrix
VCM	The variance-covariance matrix which reflects the random model and experimental errors

preferred. Using these R factors, the $R_{\rm free}$ ratio is written as $R_{G{\rm ratio}}$.

In this paper, we take the work further, exploring the variation of the $R_{G{\rm ratio}}$ about its expected value. This variation is a consequence in part of statistical fluctuations and hence we derive the variances of cross-validation statistics. It is also a consequence of the breakdown of the assumptions behind the statistical model and we discuss the effect on these statistics. We also examine the observed variation from expected values seen both in our own refinements of eye-lens proteins and in crystal structures from the Protein Data Bank (Bernstein *et al.*, 1977). Finally, we examine the variation of cross-validation statistics as a function of the choice and size of test set. Errors in the functional form of the structure-factor model, errors arising from false minima and errors arising from under-refinement may all perturb the $R_{G{\rm ratio}}$ away from its theoretical value; this topic is further explored in §6. Algebraic derivations have been relegated to appendices.

2. Definitions

For convenience, the definitions of symbols used in this paper and its appendices are grouped together in Table 1. Where relevant, all quantities are assumed to be evaluated at the convergence of the refinement.

3. Earlier work

In earlier papers (Tickle *et al.*, 1998*a*,*b*), we have derived the expected values of cross-validation statistics arising from the least-squares refinement of macromolecular structures. We assumed that the reflections had been weighted by the inverse of a VCM which reflected the random experimental and model errors. In this case, we showed that at the convergence of a least-squares refinement the expected value of the sum of squares of the weighted residuals in a working set is given by

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

where the angle brackets denote the statistical expectation. Similarly, the expected value of the sum of squares of the weighted residuals in a test set is given by

$$\langle D_{\rm free}\rangle = p + \sum_{i=1}^{p} w_i \mathbf{b}_i^T \mathbf{H}^{-1}\mathbf{b}_i. \qquad (2)$$

The above summation can be written in terms of the trace of $\mathbf{Q}$ [see *Appendix A*, equation (18)],

$$\langle D_{\rm free}\rangle = p + {\rm tr}(\mathbf{Q}). \qquad (3)$$

The derivation of the above expressions assumes that the weighting of both the working set and the test set is on an absolute scale and that it correctly reflects both the random experimental and model errors.

The right-hand sides of (1) and (2) can be evaluated directly if the least-squares normal matrix $\mathbf{H}$ can be inverted. Although matrix inversion in macromolecular refinement is becoming more feasible with the increasing memory sizes of modern computers, it is still not routinely possible in many laboratories. We therefore developed approximations to (1) and (2) which were expressed in terms of quantities readily available (Tickle *et al.*, 1998*b*). At convergence, the expected value of the sum of squares of the weighted residuals for the free set, $\langle D_{\rm free}\rangle$, can be approximated by

$$\langle D_{\rm free}\rangle \simeq (p/f)(f+d), \qquad (4)$$

Table 2
Basic data for γB- and βB2-crystallin.

	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	174	181
Non-H protein atoms	1478	1472
Ordered solvent molecules	230	92
Number of reflections	26151	18583

while the corresponding expected value for the working set is approximated by

$$\langle D_{\rm inc}\rangle \simeq f - d, \tag{5}$$

where d can be regarded as an effective number of refined parameters. The scaled ratio of the quantities in (4) and (5) gives an estimate of the $R_{G{\rm ratio}}$ as

$$R_{G{\rm ratio}} = \frac{R_{G{\rm free}}}{R_{G{\rm inc}}} \simeq \left(\frac{f+d}{f-d}\right)^{1/2}. \tag{6}$$

For the special case of unrestrained refinement, d reduces to m in the above equations. In this paper, we also use the statistic $D_{\rm ratio}$, whose estimate is

$$D_{\rm ratio} \simeq (f+d)/(f-d). \tag{7}$$

Unlike R-factor ratios, this statistic is less affected by correlations between numerator and denominator and is thus better for comparison with theory than the $R_{G{\rm free}}$ ratio.

It should be noted that the use of the approximate equations (4), (5), (6) and (7) assumes that the restraint weights are on an absolute scale. For geometric restraints, this implies that the reciprocals of the variances of geometric parameters, such as those derived from Engh & Huber (1991), should be used as weights when refining protein structures. The assumption that the reflections are weighted on an absolute scale is not required for the calculation of the estimates of $R_{G{\rm ratio}}$ or $D_{\rm ratio}$, because the right-hand sides of (6) and (7) are independent of the reflection weights.

4. Variation of cross-validation statistics

4.1. Variance of $D_{\rm free}$, $R_{G{\rm free}}$ and $R_{G{\rm ratio}}$

In order to use $D_{\rm free}$ to assess the validity of refinement models, it is necessary know the likely variation of $D_{\rm free}$ about its expected value. Whereas the derivation of the expected value of $D_{\rm free}$ only needs knowledge of the second moments of the error distribution, derivation of expressions for the variance of $D_{\rm free}$ requires assumptions about the fourth moments of the errors. In *Appendix A* it is shown that if these errors are assumed to be normally distributed then the variance of $D_{\rm free}$ can be written in terms of the trace of the matrix $\mathbf{Q}$,

$$\begin{aligned}{\rm var}(D_{\rm free}) &= 2{\rm tr}(\mathbf{I}_p + \mathbf{Q})^2\\ &= 2[p + 2{\rm tr}(\mathbf{Q}) + {\rm tr}(\mathbf{Q}^2)].\end{aligned} \tag{8}$$

Table 3
γB- and βB2-crystallin least-squares structure refinement.

Minimization against working set of reflections plus geometric restraints.

	γB-crystallin	βB2-crystallin
Parameters refined (m)	6844	6266
Reflections in working set (f)	24788	17622
Reflections in test set (p)	1363	961
Geometric restraints (r)	3887	3853
Total observations used (n)	28675	21475
Deviance from geometry ($D_{\rm rest}$)	1137	696
Deviance from reflections ($D_{\rm inc}$)	20701	14522
Deviance from all data ($D_{\rm total}$)	21838	15218
Expected value of deviance ($n - m$)	21831	15209

In contrast to $D_{\rm free}$, $R_{G{\rm free}}$ and $R_{G{\rm ratio}}$ are independent of the absolute scale of the weights; in *Appendix A* we derive the following expression for their fractional standard deviations,

$$\frac{\sigma(R_{G{\rm free}})}{R_{G{\rm free}}} \simeq \frac{\sigma(R_{G{\rm ratio}})}{R_{G{\rm ratio}}} \simeq \frac{[p + 2{\rm tr}(\mathbf{Q}) + {\rm tr}(\mathbf{Q}^2)]^{1/2}}{(2)^{1/2}D_{\rm free}}.$$

These quantities may be used to test whether an observed $R_{G{\rm free}}$ or $R_{G{\rm ratio}}$ deviates significantly from its estimated value as given by Tickle *et al.* (1998*b*). Hypothesis tests will rely on the central limit theorem to ensure that the relevant distributions are approximately normal.

4.2. Calculation of the variance of $D_{\rm free}$, $R_{G{\rm free}}$ and $R_{G{\rm ratio}}$

The evaluation of the expressions for the variances given in the previous section requires the calculation of ${\rm tr}(\mathbf{Q})$ and ${\rm tr}(\mathbf{Q}^2)$. If the weight matrix is diagonal, the trace of $\mathbf{Q}$ can be expressed as

$${\rm tr}(\mathbf{Q}) = \sum_{i=1}^{p} w_i \mathbf{b}_i^T \mathbf{H}^{-1} \mathbf{b}_i$$

and does not require explicit evaluation of $\mathbf{Q}$. However, the trace of $\mathbf{Q}^2$ is the square of the Frobenius norm of $\mathbf{Q}$,

$${\rm tr}(\mathbf{Q}^2) = \|\mathbf{Q}\|^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} \mathbf{Q}_{ij}^2, \tag{9}$$

and each unique element of $\mathbf{Q}$ must be evaluated as

$$\mathbf{Q}_{ij} = (w_i w_j)^{1/2} \mathbf{b}_i^T \mathbf{H}^{-1} \mathbf{b}_j. \tag{10}$$

If the evaluation of ${\rm tr}(\mathbf{Q}^2)$ is computationally too demanding, an approximation may be developed for ${\rm var}(D_{\rm free})$ if it is assumed that the eigenvalues of $\mathbf{Q}$ are equal (*Appendix B*). In this case,

$${\rm var}(D_{\rm free}) \simeq 2p(f+d)^2/f^2. \tag{11}$$

Fractional standard errors are often of more practical value. In *Appendix B*, it is shown that the fractional standard error of $D_{\rm free}$ may be estimated as

$$\sigma(D_{\rm free})/D_{\rm free} \simeq (2/p)^{1/2}. \tag{12}$$

If the smaller errors in $D_{\rm inc}$, $\Sigma_{\rm inc}$ and $\Sigma_{\rm free}$ are ignored, the fractional error in $R_{G{\rm free}}$ and $R_{G{\rm ratio}}$ is given by halving the above estimate,

Table 4
Statistics from γB- and βB2-crystallin least-squares structure refinements.

Statistic	Origin	Reference	γB-crystallin	βB2-crystallin
d	$m - r + D_{\rm rest}$	Definition	4094	3109
$D_{\rm free}$	Observed		1872	1353
$\langle D_{\rm free}\rangle$	$p + {\rm tr}(\mathbf{Q})$	Eq. (3)	1663	1196
	Approximation, $(p/f)(f + d)$	Eq. (4)	1588	1131
$\sigma(D_{\rm free})$	$\{2[p + 2\,{\rm tr}(\mathbf{Q}) + {\rm tr}(\mathbf{Q}^2)]\}^{1/2}$	Eq. (8)	64	55
	Approximation, $(2p)^{1/2}(f + d)/f$	Eq. (11)	61	52
$R_{G{\rm inc}}$	Observed		0.226	0.213
$R_{G{\rm free}}$	Observed		0.291	0.279
$R_{G{\rm ratio}}$	Observed		1.29	1.31
	Estimated $[(f + d)/(f - d)]^{1/2}$	Eq. (6)	1.18	1.20
$\sigma(D_{\rm free})/D_{\rm free}$	$(2/p)^{1/2}$	Eq. (12)	0.038	0.046
$\sigma(R_{G{\rm free}})/R_{G{\rm free}}$	$1/(2p)^{1/2}$	Eq. (13)	0.019	0.023
$s(R_{\rm free})/R_{\rm free}$	$1/p^{1/2}$	Brünger (1997)	0.027	0.032

$$\frac{\sigma(R_{G{\rm free}})}{R_{G{\rm free}}} \simeq \frac{\sigma(R_{G{\rm ratio}})}{R_{G{\rm ratio}}} \simeq 1/(2p)^{1/2}. \qquad (13)$$

5. Observed values of cross-validation statistics

5.1. Calculation of $D_{\rm free}$, $R_{G{\rm free}}$ and $R_{G{\rm ratio}}$ from crystallin refinements

In order to deploy the above theory, we carried out refinements of γB- and βB2-crystallin, which are proteins found in the fibre cells of the eye lens. The crystal data for the proteins are listed in Table 2 and the refinement statistics are given in Table 3. The work was carried out using the least-squares refinement program *RESTRAIN* (Haneef *et al.*, 1985; Driessen *et al.*, 1989), with weighting of observations on an absolute scale as described in Tickle *et al.* (1998*a*). Provided that the number of parameters is small relative to the number of reflections, the estimates of $R_{G{\rm ratio}}$ and $D_{\rm ratio}$ are not particularly sensitive to small errors in the weighting of the structure-amplitude terms relative to the geometrical restraints. This is because errors in $D_{\rm rest}$, and hence d, are small compared with f and so do not significantly affect the value of $R_{G{\rm ratio}}$ (6).

The starting atomic models were those obtained in Tickle *et al.* (1998*a*). For each crystallin, we took one test set comprising about 5% of the data and refined the model against the corresponding working set using the full normal equations matrix, $\mathbf{H}$. All cross terms, including the scale factor G positional parameters and temperature factors, were included in $\mathbf{H}$. The refinements were terminated when the fractional parameter shifts were of the same order as the square root of the machine single precision (Press *et al.*, 1992). At this stage, the refinement was as close as possible to the local function minimum and was at least near the global minimum. The normal matrix was then inverted and used to calculate the matrix $\mathbf{Q}$. The expected values of $D_{\rm free}$ were calculated using (2). The standard deviations of these values were calculated using (8), (9) and (10).

It should be noted that the protocol adopted for deriving these refinement statistics assumes that the initial crystal structure models are free from gross errors. This protocol is quite different from that required for validating a crystal structure. In the latter case, the test set should be separated from the working set at the beginning of structure analysis. In this way, $R_{G{\rm free}}$ is independent of R_G throughout the course of the refinement, allowing the $R_{G{\rm ratio}}$ to assist in the recognition of serious errors in the earlier stages of the analysis. The test set should be reintroduced at the end of refinement, where possible.

The statistics from these refinements are shown in Table 4. The theory presented in this paper and in its predecessor (Tickle *et al.*, 1998*b*) is only applicable at the global least-squares minimum. It refers to the expected values of R-factor statistics at the termination of a refinement but not during the course of a refinement. Hence, the values of statistics shown in the tables are all at the termination of refinement. They were obtained using many more cycles of refinement than had been deemed necessary for publication of the crystallin crystal structures themselves. As the structure refinement converged against the working set, it was interesting to note that $R_{G{\rm ratio}}$ in the βB2-crystallin refinement steadily rose, starting at about

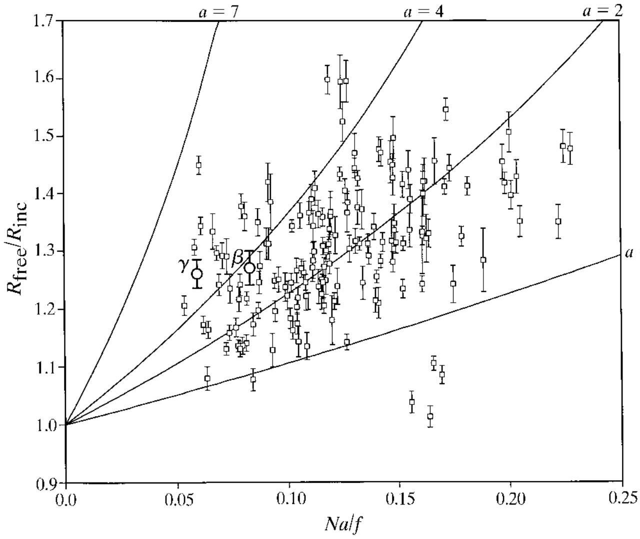

Figure 1
Plot of the $R_{\rm free}/R_{\rm inc}$ ratios and their estimated errors as a function of N_a/f for 157 macromolecular structures in the Protein Data Bank, where N_a is the number of atoms included in the refinement and f is the number of reflections used. Only structures in the resolution range 1.5–2.1 Å are shown. The error bars were calculated as $1/(2p)^{1/2}$ of the $R_{\rm free}$ ratio, where p is the number of reflections in the test set. Also shown are the data points for the γB- and βB2-crystallin structures discussed in this paper, shown as large bold circles and labelled γ and β, respectively. The four curves correspond to different values of the variable a defining the refinement regime used: a = 1 represents 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; a = 7 represents nine parameters per atom (restrained anisotropic refinement).

Table 5
γB- and βB2-crystallin least-squares structure-refinement statistics using 10% test sets.

Mean and s.d. are the column mean and standard deviation, respectively.

	Test set	p	$D_{\rm inc}$	$D_{\rm free}$	$D_{\rm ratio}$
γB-crystallin	0	2633	19468	3876	1.33
	1	2581	19523	3477	1.28
	2	2640	19473	3627	1.31
	3	2640	19477	3776	1.31
	4	2673	19423	3797	1.31
	5	2625	19507	3625	1.29
	6	2567	19573	3738	1.32
	7	2573	19551	3441	1.27
	8	2587	19514	3738	1.32
	9	2616	19508	3860	1.33
Mean		2614	19502	3695	1.31
S.d.			18	133	0.02
S.d./mean			0.001	0.04	0.02
βB2-crystallin	0	1898	13624	2734	1.33
	1	1815	13693	2606	1.33
	2	1868	13615	2607	1.31
	3	1887	13598	2710	1.33
	4	1866	13658	2797	1.35
	5	1855	13648	2714	1.34
	6	1833	13661	2523	1.30
	7	1814	13669	2484	1.30
	8	1853	13614	2831	1.37
	9	1894	13610	2569	1.29
Mean		1858	13639	2657	1.32
S.d.			18	105	0.03
S.d./mean			0.001	0.04	0.02

Table 6
γB- and βB2-crystallin least-squares structure-refinement R factors.

10% test sets were used. Mean and s.d. are the column mean and standard deviation, respectively.

	Test set	$R_{G\rm inc}$	$R_{G\rm free}$	$R_{G\rm free}/R_{G\rm inc}$	$R_{\rm inc}$	$R_{\rm free}$	$R_{\rm free}/R_{\rm inc}$
γB-crystallin	0	0.224	0.299	1.34	0.173	0.226	1.31
	1	0.225	0.288	1.28	0.174	0.221	1.27
	2	0.225	0.292	1.30	0.174	0.221	1.27
	3	0.225	0.295	1.31	0.173	0.222	1.28
	4	0.224	0.295	1.32	0.173	0.222	1.28
	5	0.225	0.290	1.29	0.174	0.219	1.26
	6	0.225	0.297	1.32	0.173	0.223	1.29
	7	0.226	0.287	1.27	0.174	0.219	1.26
	8	0.225	0.294	1.31	0.173	0.228	1.32
	9	0.224	0.298	1.33	0.173	0.222	1.28
Mean		0.225	0.293	1.31	0.173	0.222	1.28
S.d.		0.0005	0.004	0.02	0.0005	0.003	0.02
S.d./mean		0.002	0.01	0.02	0.003	0.01	0.02
βB2-crystallin	0	0.212	0.275	1.30	0.176	0.218	1.24
	1	0.211	0.279	1.32	0.175	0.227	1.30
	2	0.211	0.280	1.33	0.175	0.224	1.28
	3	0.211	0.283	1.34	0.174	0.232	1.33
	4	0.211	0.287	1.36	0.174	0.232	1.33
	5	0.211	0.282	1.34	0.175	0.223	1.27
	6	0.212	0.274	1.29	0.176	0.216	1.23
	7	0.212	0.274	1.29	0.176	0.213	1.21
	8	0.210	0.293	1.39	0.174	0.234	1.34
	9	0.213	0.273	1.28	0.177	0.212	1.20
Mean		0.211	0.280	1.32	0.175	0.223	1.27
S.d.		0.0007	0.006	0.03	0.0010	0.008	0.05
S.d./mean		0.003	0.02	0.03	0.006	0.04	0.04

one standard deviation *less* than the expected value and rising to three standard deviations above that value. This was a consequence more of a rise in $R_{G\rm free}$ than of a fall in R_G.

5.2. Statistical approximations from crystallin refinements

Table 4 also shows approximations to statistics from the refinements of γB- and βB2-crystallin which do not require the calculation of the matrix $\mathbf{Q}$ or any matrix inversion. The expected value and standard deviation of $D_{\rm free}$ calculated from the approximate formulae may be compared with those derived from the more exact calculations. The approximate expressions do not reflect the fact that the expected value and variance of $D_{\rm free}$ depend on the particular test set. The approximate formula for the variance of $D_{\rm free}$ shown in (11) neglects the variation of the eigenvalues of $\mathbf{Q}$. Consideration of (25) shows that this is likely to lead to underestimation of the variance of $D_{\rm free}$. The approximate formulae do indeed underestimate the variance by about 5%.

Finally, Table 4 also shows results from approximate formulae for $R_{G\rm ratio}$ and the fractional standard deviations of $D_{\rm free}$ and $R_{G\rm free}$.

5.3. $R_{\rm inc}$ and $R_{\rm free}$ in the Protein Data Bank

We also examined 157 X-ray structures from the Protein Data Bank (Bernstein *et al.*, 1977) which had been determined to a resolution between 1.5 and 2.1 Å and for which $R_{\rm inc}$ and $R_{\rm free}$ data were available. The statistically more tractable generalized R factors would have been used had they been available from all refinement programs. We plotted the $R_{\rm free}/R_{\rm inc}$ ratios as a function of N_a/f. We also plotted curves (Fig. 1) which represent the estimated $R_{G\rm ratio}$s for different refinement regimes calculated in the same way as described in Tickle *et al.* (1998*b*).

The crystallin $R_{\rm free}/R_{\rm inc}$ ratios lie three to four standard deviations above the expected positions on the $a = 2$ curve, the approximate curve for restrained isotropic refinement.

5.4. Jack-knife tests

The Protein Data Bank results have generally been concerned with single test sets for each protein and, in the case of the crystallins, the variance calculated from (8) measures the variation to be expected in $D_{\rm free}$ owing to model and experimental errors for a *given* test set.

We now consider the effect on the R factors and $R_{G\rm ratio}$ of choosing different test sets. We have examined the effect of different test sets on crystallin refinements, but we have not developed any theory to cover this case, as multiple test sets are not regularly used in practice. However, Brünger (1997) has given an estimate of the variation of $R_{\rm free}$ in such cases and this is included in Table 4.

Structure amplitudes of the two crystallins were used in a jack-knife procedure in which each reflection belonged to a subset containing approximately 10% of the data. Each reflection was assigned to a subset by generating a random

Table 7
γB- and βB2-crystallin least-squares structure-refinement statistics.

Mean and s.d. are the mean and standard deviation for the 20 5% test sets for each structure, respectively.

		p	$D_{\rm inc}$	$D_{\rm free}$	$D_{\rm ratio}$
γB-crystallin	Mean	1307	20770	1809	1.29
	S.d.		17	88	0.03
	S.d./mean		0.001	0.05	0.02
βB2-crystallin	Mean	929	14557	1335	1.32
	S.d.		21	80	0.04
	S.d./mean		0.001	0.06	0.03

number from a uniform distribution between 0 and 1, multiplying this number by ten and taking the integral part. This produced an integer between 0 and 9 which identified the test set to which the reflection was assigned. Each subset was taken in turn and used as a test set while the other nine subsets were used as the working set in refinement. Table 5 shows the statistics resulting from these refinements including $D_{\rm ratio}$. However, R factors and $R_{G\rm ratio}$ are most readily available from refinement outputs and these are shown in Table 6. The jackknife tests were repeated using test sets containing approximately 5% of the data. Summary results are shown in Tables 7 and 8.

The variation of $R_{G\rm ratio}$ between test sets, as measured by the fractional standard deviations (s.d./mean) is shown in Tables 6 and 8. The variation in Table 8 is higher than the statistical variation [$\sigma(R_{G\rm free})/R_{G\rm free}$] shown in Table 4. The use of different test sets produces extra variability. This effect is less pronounced with larger test sets, as shown by the smaller values of s.d./mean in Tables 5 and 6 compared with Tables 7 and 8.

6. Discussion

6.1. Origins of parameter errors and their effect on cross-validation statistics

The origins of parameter errors in a supposedly refined protein structure fall into five categories.

(i) Missing higher resolution X-ray data owing to weak diffraction and also absent data from correlated regions of reciprocal space.

(ii) The choice of functional form of the structure-factor model and associated statistical assumptions. For example, isotropic mean-square displacement parameters are usually used, even though proteins are known to exhibit significant anharmonic disorder. Missing atoms, misidentification of atoms, unmodelled and disordered solvent, lack of absorption corrections, inappropriate rigid-body constraints, erroneous assumptions in likelihood methods, errors in the scaling model and inappropriate weighting of observations may also contribute errors in this category.

(iii) Refinement of an insufficiently accurate atomic model of the protein which results in convergence to a wrong minimum.

(iv) Under-refinement of the model parameters, where convergence is far from being achieved owing to insufficient iterations of the refinement process.

(v) Lack of precision of the observations (X-ray intensities and restraints).

In a well refined structure of a small molecule, errors in category (v) should dominate. However, in a protein structure, where there is limited resolution and the conventional structure factor expression provides a relatively poor model, the dominating errors will usually arise from category (ii). These errors are in the functional form of the structure factor and can be thought of in terms of missing parameters in the structure-factor expression or, equivalently, the imposition of incorrect parameter constraints. For example, the usual isotropic model corresponds to constraining to zero all anharmonic displacement parameters and all off-diagonal parameters in the anisotropic displacement tensors while constraining the diagonal terms to be equal. Errors in category (iii) can also be thought of in terms of positional constraints keeping the refinement away from the correct minimum.

Inappropriate constraints will give rise to higher values of both R_G and $R_{G\rm free}$. The $R_{G\rm ratio}$, however, will increase when the missing parameters are correlated with parameters present in the structure-factor model. For example, at lower resolution the displacement parameters are significantly correlated with the positional parameters owing to unresolved atoms. In such a case, inappropriate displacement-parameter constraints will cause the minimization of R_G to produce erroneous shifts in atomic co-ordinates. These shifts will not produce a downward response in $R_{G\rm free}$, and the $R_{G\rm ratio}$ will therefore increase above that predicted by theory. This probably explains the high $R_{G\rm ratio}$s found at the end of our crystallin refinements. In a similar way, category (iii) errors will also increase the $R_{G\rm ratio}$.

The considerations of the previous paragraph might suggest that protein refinement should produce an $R_{G\rm ratio}$ higher than that predicted by theory because some correlations with missing parameters will always occur. Examination of the figure does not bear this out in all cases. This may be because of category (iv) errors. The initial value of the $R_{G\rm ratio}$, when the working set has just been separated from the test set, will be close to unity. As refinement proceeds, the ratio rises. Some crystallographers may consider that nothing is gained by fully refining a protein structure, while others may only separate their test set from the working set near the end of the refinement. Both these approaches will lead to smaller $R_{G\rm ratio}$s than those predicted by theory and probably account for most of the points below the $a = 2$ curve in Fig. 1. It should be noted that serious use of the $R_{G\rm ratio}$ to detect errors requires separation of the test set before the start of refinement.

6.2. The effect of adding parameters on cross-validation statistics

The above discussion has pointed out that constraining parameters to incorrect values is likely to increase the $R_{G\rm ratio}$ above the value predicted by theory. Conversely, adding parameters which improve a model leads to a reduction in R_G,

Table 8
γB- and βB2-crystallin least-squares structure-refinement R factors.

Mean and s.d. are the mean and standard deviation, respectively, for the 20 5% test sets for each structure.

		$R_{G\rm inc}$	$R_{G\rm free}$	$R_{G\rm free}$/ $R_{G\rm inc}$	$R_{\rm inc}$	$R_{\rm free}$	$R_{\rm free}$/ $R_{\rm inc}$
γB-crystallin	Mean	0.226	0.291	1.29	0.175	0.220	1.26
	S.d.	0.0004	0.007	0.03	0.0003	0.006	0.03
	S.d./mean	0.002	0.02	0.02	0.002	0.03	0.03
βB2-crystallin	Mean	0.213	0.281	1.32	0.176	0.224	1.27
	S.d.	0.0005	0.009	0.04	0.0006	0.009	0.06
	S.d./mean	0.002	0.03	0.03	0.003	0.04	0.04

$R_{G\rm free}$ and the $R_{G\rm ratio}$. However, adding parameters to an already correct model (sometimes known as over-refinement) will tend to increase the $R_{G\rm ratio}$, but in this case the increase is predicted by theory, as can be seen by considering (6). Conversely, reducing the effective number of refined parameters by applying appropriate restraints or contraints (for example, in rigid-body refinement) will reduce the $R_{G\rm ratio}$ and (6) may be used to test whether or not the fall in this statistic confirms the appropriateness of the imposed restraints or constraints.

As the addition of parameters to an already correct protein model is likely to increase the refined $R_{G\rm ratio}$, it might be thought that the parameter set which gives the lowest $R_{G\rm ratio}$ is the most appropriate. However, the goal of a protein refinement is usually to answer a biological question, rather than minimizing any given statistic *per se*. For example, it might be important to know the standard deviations of metal-ligand distances and therefore these distances should be freely refined, even though by restraining them the $R_{G\rm ratio}$ might be reduced.

6.3. Other considerations

Carrying out the work for this paper has forcibly reminded us that there are no criteria for assessing whether a refinement has converged. Parameter shifts which are a fraction of one standard deviation have traditionally been used as a convergence criterion, but this assumes that the shifts are part of a series whose sum will rapidly converge as refinement proceeds. This may not be the case, especially when function minimization does not use a full normal matrix.

The theory developed in this paper has been in the context of the least-squares method. However, many of the statistical properties of least squares hold asymptotically for other estimating regimes such as maximum likelihood or quasi-likelihood (McCullagh & Nelder, 1989). This may also apply to the statistical theory in the present paper.

6.4. Significance of cross-validation statistics

Both our exact and approximate theoretical expressions for $\sigma(D_{\rm free})$ take account of the errors arising from sampling the set of free residual values from the population of sets of free residual values. Also, both take account of the correlations between the free residuals within the test set. The difference is that the exact expression applies to a specific choice of working set, because it takes account of the effect of choosing the working set and hence the model *via* the matrix $\mathbf{Q}$.

The estimates of the standard deviation of $D_{\rm free}$ obtained by choosing different pairs of complementary working and test sets also take account of the sampling errors, even though the sample is from the possible test sets rather than from the population of values for a given test set. However, there are now additional effects owing to the variation of the working set and also correlations between different sets of free residuals. As a result, one would expect the standard deviation of $D_{\rm free}$ between test sets to be greater than that within a test set and this is what is observed. For example, compare the values of $\sigma(D_{\rm free})$ and $\sigma(D_{\rm free})/D_{\rm free}$ in Table 5 with those of s.d.($D_{\rm free}$) and s.d.($D_{\rm free}$)/mean($D_{\rm free}$) in Table 7.

7. Conclusions

Interpretation of all cross-validation statistics requires special care. For a significantly wrong model, we assume (probably correctly) that the $R_{G\rm ratio}$ will significantly exceed its expected value if sufficient refinement is carried out. However, an under-refined incorrect model may yield an $R_{G\rm ratio}$ which is close to the expected value. Thus, while a value of the $R_{G\rm ratio}$ which is close to its expected value is not necessarily a criterion for the correctness of a model structure, an $R_{G\rm ratio}$ significantly larger than the expected value should be a serious warning that the model may need substantial revision.

An important question is how to use the $R_{G\rm ratio}$ to recognize a wrong structure in the early stages of refinement. This requires an investigation of the behaviour of the $R_{G\rm ratio}$ during the refinement of wrong models. This work will now be undertaken.

APPENDIX *A*
Properties of cross-validation statistics

A1. Expected value and variance of $D_{\rm free}$

In earlier papers (Tickle *et al.*, 1998*a*,*b*), we have derived expressions for the expected values and variance–covariance matrices of the residuals at the convergence of a refinement. In order to derive an expression for the variance of the sum of weighted squared residuals, we need to make some assumption about their distribution. Here, we assume that they are normally distributed.

First, we consider a vector of q random variables $\mathbf{x}$ with distribution $N(0, \mathbf{D})$, where $\mathbf{D}$ is the VCM of $\mathbf{x}$. We now transform $\mathbf{x}$ into a vector of independent unit normal deviates $\mathbf{y}$. Defining $\mathbf{y} = \mathbf{D}^{-1/2}\mathbf{x}$, then $\mathbf{y} \simeq N(0, \mathbf{I}_q)$.

The variance of the sum of squared random variables can then be written as

$$\begin{aligned}\mathrm{var}(\mathbf{x}^T\mathbf{x}) &= \mathrm{var}(\mathbf{y}^T\mathbf{D}\mathbf{y})\\ &= \mathrm{var}\left(\sum_i D_{ii}y_i^2 + 2\sum_i\sum_{\substack{j\\ i<j}} D_{ij}y_iy_j\right)\\ &= \sum_i D_{ii}^2\mathrm{var}(y_i^2) + 4\sum_i\sum_{\substack{j\\ i<j}} D_{ij}^2\mathrm{var}(y_iy_j)\\ &\quad + 4\sum_i\sum_{\substack{j\\ i<j}} D_{ii}D_{ij}\mathrm{cov}(y_i^2, y_iy_j)\\ &\quad + 8\sum_i\sum_{\substack{j\\ i<j}}\sum_{\substack{k\\ j<k}} D_{ij}D_{jk}\mathrm{cov}(y_iy_j, y_jy_k). \qquad (14)\end{aligned}$$

The values of the variances in the above equation follow immediately from the properties of independent unit normal deviates, $\mathrm{var}(y_i^2) = 2$ and $\mathrm{var}(y_iy_j) = 1$. The covariances which are between independently distributed random variables are zero [for example, $\mathrm{cov}(y_iy_j, y_ky_l)$ and $\mathrm{cov}(y_i^2, y_j^2)$] and are not shown in (14). The other covariances involve non-independent variables, but these also vanish because the moments of odd order are zero in the following expansions,

$$\begin{aligned}\mathrm{cov}(y_i^2, y_iy_j) &= \langle y_i^3y_j\rangle - \langle y_i^2\rangle\langle y_iy_j\rangle,\\ \mathrm{cov}(y_iy_j, y_jy_k) &= \langle y_iy_j^2y_k\rangle - \langle y_iy_j\rangle\langle y_jy_k\rangle.\end{aligned}$$

(14) therefore reduces to

$$\mathrm{var}(\mathbf{x}^T\mathbf{x}) = 2\sum_i D_{ii}^2 + 4\sum_i\sum_{\substack{j\\ i<j}} D_{ij}^2. \qquad (15)$$

The right-hand side of (15) is twice the square of the Frobenius norm of $\mathbf{D}$, this norm being the square root of the sum of the squares of the matrix elements. The square of the Frobenius norm of a matrix is equal to the trace of its square and therefore we can write

$$\mathrm{var}(\mathbf{x}^T\mathbf{x}) = 2\|\mathbf{D}\|^2 = 2\mathrm{tr}(\mathbf{D}^2). \qquad (16)$$

We now consider the random variables $\mathbf{x}$ to be a vector of weighted residuals of the excluded observations. (16) may then be used to evaluate the variance of the sum of the squares of these weighted residuals, given the matrix of their second moments. An expression for the latter was derived in Tickle *et al.* (1998*b*) equation (19), where we showed that, in the notation of this paper,

$$\mathbf{W}_{\mathrm{free}}^{1/2}\mathbf{D}_{\mathrm{free}}\mathbf{W}_{\mathrm{free}}^{1/2} = \mathbf{I}_p + \mathbf{Q}. \qquad (17)$$

Taking the trace of both sides gives

$$\begin{aligned}\langle D_{\mathrm{free}}\rangle &= \mathrm{tr}(\mathbf{I}_p + \mathbf{Q})\\ &= p + \mathrm{tr}(\mathbf{Q}), \qquad (18)\end{aligned}$$

while by using (17) we can substitute $\mathbf{D} = \mathbf{I}_p + \mathbf{Q}$ in (16) giving

$$\begin{aligned}\mathrm{var}(D_{\mathrm{free}}) &= 2\mathrm{tr}(\mathbf{I}_p + \mathbf{Q})^2\\ &= 2[p + 2\mathrm{tr}(\mathbf{Q}) + \mathrm{tr}(\mathbf{Q}^2)]. \qquad (19)\end{aligned}$$

The traces in the above two equations can be replaced by summations over the eigenvalues of $\mathbf{Q}$. (18) can be written as

$$\langle D_{\mathrm{free}}\rangle = p + \sum_{\mathrm{i}}^{p}\lambda_i, \qquad (20)$$

where λ_i is the ith eigenvalue of $\mathbf{Q}$. Similarly, (19) can be written as

$$\mathrm{var}(D_{\mathrm{free}}) = 2\left(p + 2\sum_{\mathrm{i}}^{p}\lambda_i + \sum_{\mathrm{i}}^{p}\lambda_i^2\right). \qquad (21)$$

These results will be used in *Appendix B*, where approximate expressions will be developed for the standard deviations of D_{free} and R_{free}.

A2. The standard deviations of $R_{G\mathrm{free}}$ and $R_{G\mathrm{ratio}}$

$R_{G\mathrm{free}}$ is defined as

$$R_{G\mathrm{free}} = (D_{\mathrm{free}}/\Sigma_{\mathrm{free}})^{1/2}.$$

A rigorous derivation of the variance of $R_{G\mathrm{free}}$ (unpublished results) follows the same lines as that given above for D_{free} and takes into account the variance of the denominator and the correlation between numerator and denominator. However, a good approximation can be given in cases where $D_{\mathrm{free}} << \Sigma_{\mathrm{free}}$ and consequently the fractional error in the denominator is small. We can then write

$$\sigma^2(R_{G\mathrm{free}}^2) \simeq \sigma^2(D_{\mathrm{free}})/\Sigma_{\mathrm{free}}^2$$

and the fractional variation $R_{G\mathrm{free}}^2$ is then

$$\frac{\sigma(R_{G\mathrm{free}}^2)}{R_{G\mathrm{free}}^2} \simeq \frac{\sigma(D_{\mathrm{free}})}{D_{\mathrm{free}}}.$$

Halving the right-hand side gives us the fractional variation in $R_{G\mathrm{free}}$ as

$$\begin{aligned}\frac{\sigma(R_{G\mathrm{free}})}{R_{G\mathrm{free}}} &\simeq \frac{\sigma(D_{\mathrm{free}})}{2D_{\mathrm{free}}} \qquad (22)\\ &= \frac{[p + 2\mathrm{tr}(\mathbf{Q}) + \mathrm{tr}(\mathbf{Q}^2)]^{1/2}}{(2)^{1/2}D_{\mathrm{free}}}. \qquad (23)\end{aligned}$$

$R_{G\mathrm{ratio}}$ is given by

$$R_{G\mathrm{ratio}} = R_{G\mathrm{free}}(\Sigma_{\mathrm{inc}}/D_{\mathrm{inc}})^{1/2}.$$

If the variation arising from the larger summations over the working set is ignored, the fractional standard deviation of the $R_{G\mathrm{ratio}}$ will be equal to that of $R_{G\mathrm{free}}$,

$$\frac{\sigma(R_{G\mathrm{ratio}})}{R_{G\mathrm{ratio}}} \simeq \frac{[p + 2\mathrm{tr}(\mathbf{Q}) + \mathrm{tr}(\mathbf{Q}^2)]^{1/2}}{(2)^{1/2}D_{\mathrm{free}}}. \qquad (24)$$

APPENDIX *B*
Approximations for the standard deviations of D_{free} and $R_{G\mathrm{free}}$

In *Appendix A*, we have derived expressions for the expected value and variance of D_{free}. Computation of these expressions requires the inversion of the normal matrix. If this is not possible, then we need to find approximations which are more readily calculated. We have already derived approximations

Acta Cryst. (2000). D**56**, 442–450 Tickle *et al.* • R_{free} and the R_{free} ratio **449**

for the expected values of D_{free} and $R_{G\text{ratio}}$ (Tickle *et al.*, 1998*b*) and these are given in (4) and (6), respectively. In this appendix we derive, for both these statistics, approximate expressions for their variances which can be calculated from quantities readily available from a crystal structure refinement.

According to (20) and (21) in *Appendix A*, the mean and variance of D_{free} can be written in terms of the eigenvalues of the symmetric matrix $\mathbf{Q}$ of order p.

(21) can be rewritten in terms of the mean eigenvalue $\langle\lambda\rangle$ and the variation of the eigenvalues about their mean,

$$\operatorname{var}(D_{\text{free}}) = 2\left[p + 2\sum_i^p \lambda_i + p\langle\lambda\rangle^2 + \sum_i^p (\lambda_i - \langle\lambda\rangle)^2\right]. \quad (25)$$

From (4) and (20), we have

$$\sum_i^p \lambda_i = pd/f, \quad (26)$$

$$\langle\lambda\rangle = d/f. \quad (27)$$

If we neglect the last term in equation (25) which expresses the variation of the eigenvalues of $\mathbf{Q}$ about their mean, we obtain from (25), (26) and (27),

$$\begin{aligned}\operatorname{var}(D_{\text{free}}) &\simeq 2[p + 2pd/f + p(d/f)^2] \\ &\simeq 2p(f+d)^2/f^2.\end{aligned}$$

Hence,

$$\sigma(D_{\text{free}}) \simeq (2p)^{1/2}(f+d)/f. \quad (28)$$

Fractional standard errors are often of more practical value. From (4) and (28), the fractional standard error of D_{free} is given by

$$\sigma(D_{\text{free}})/D_{\text{free}} \simeq (2/p)^{1/2}.$$

If $D_{\text{free}} << \Sigma_{\text{free}}$, we can use (22), (23) and (24) in *Appendix A* and the fractional error in $R_{G\text{free}}$ and $R_{G\text{ratio}}$ is given by

$$\frac{\sigma(R_{G\text{free}})}{R_{G\text{free}}} \simeq \frac{\sigma(R_{G\text{ratio}})}{R_{G\text{ratio}}} \simeq 1/(2p)^{1/2}.$$

The authors wish to thank Dr Steve Wood, Department of Biochemistry, University of Southampton for the provision of the computing facilities which made this work possible. They also wish to thank the referees for most useful and detailed comments.

References

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)*, **35**5, 472–475.

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

Brünger, A. T. (1997). *Methods Enzymol.* **277**, 366–396.

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.

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

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

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

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, pp. 391–418. London: Chapman & Hall.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes in C*, p. 398. Cambridge University Press.

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.