Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Rearrangement of Cruickshank's formulae for the diffraction-component precision index

D. M. Blow

Biophysics Group, Imperial College of Science, Technology and Medicine, London SW7 2AZ, England

Correspondence e-mail: d.blow@ic.ac.uk

Received 27 November 2001
Accepted 28 February 2002

The formulae for the diffraction-component precision index introduced by Cruickshank [(1999), *Acta Cryst.* D**55**, 583–601] are simplified using two approximations. A rearranged formula for the precision index is presented which can readily be calculated from experimental data. It is shown that the precision varies as $(\text{resolution})^{5/2}$ if R and completeness are maintained. It varies as $(\text{completeness of intensity measurement})^{-5/6}$ and its dependence on the inclusion of solvent atoms is discussed.

1. Introduction

The purpose of crystal structure refinement is to generate coordinates which represent the structure as precisely as possible. In order to evaluate refinement strategy it is necessary to form an estimate of coordinate accuracy. It is not sufficient, for example, to achieve the lowest possible R factor, since at better resolution a more accurate structure may be defined, even though the R factor may be larger.

Cruickshank (1999) introduced the diffraction-component precision index (DPI) to estimate the precision of coordinates obtained by structural refinement of protein diffraction data. He had earlier analysed the precision of refined coordinates from structural analysis of smaller molecules at atomic resolution (Cruickshank, 1949, 1959) and these formulae have been used extensively in small-molecule structure analysis. The present paper develops Cruickshank's formulae to bring out the dependence of coordinate accuracy on parameters which are under the experimenter's control in a macromolecular structure analysis, most importantly the resolution. For proteins, it had become usual to use a method invented by Luzzati (1952) for quite a different purpose, but this method of analysis was cogently criticized by Cruickshank, who demonstrated that his methods can be applied effectively to macromolecular studies. Crucial differences from the analysis of a simple crystal structure arise because the Debye–Waller factor ('B factor') may be too large for individual atoms to be resolved and it may vary substantially between different parts of a macromolecule.

The DPI provides an estimate for the precision of coordinates obtained by structural refinement from diffraction data without including extra precision which may be provided by refinement constraints. The standard deviation of a coordinate x is obtained for an atom whose B factor is an average B for the particular structure which has been refined. This quantity $\sigma(x, B_{\text{avg}})$ is called the 'diffraction-component precision index' by Cruickshank (1999). The DPI has been shown to provide an estimated standard uncertainty within about 15% of that generated by full-matrix inversion of the unrestrained or restrained normal matrix in three cases (Cruickshank, 1999, 2001). He suggests it may be used to give 'a quick and rough guide' to coordinate precision.

The apparent complexity of the formulae presented by Cruickshank seems to have discouraged their use. Using two approximations, the present paper recasts the formulae into a more easily usable form and presents other rearrangements that demonstrate the effects on precision of quantities which the experimenter may be able to control. Expressions for the DPI are generated which show how coordinate accuracy depends on the resolution, the completeness of data collection, the number of fully occupied solvent sites used in refinement, the R factor and the Matthews volume V_M (the volume of crystal per dalton of protein; Matthews, 1968).

2. Analysis

2.1. Cruickshank's formulae

Cruickshank's formulae (Cruickshank, 1999) depend upon the effective number of fully occupied atom sites N_i of each type i and a quantity p, which is the excess of the number of observations, $n_{\rm obs}$, over the number of parameters to be refined, $n_{\rm params}$. The result takes a simpler form if the scattering atoms are all considered to be of the same type, which is a reasonable assumption for a structure which is composed overwhelmingly of C, N and O atoms (with associated H atoms). With this assumption, Cruickshank presents the following formula for the precision of a coordinate x for a C, N or O atom whose B is average,

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

where N_i is the number of C, N and O atoms, $p = n_{\rm obs} - n_{\rm params}$ and $n_{\rm obs} = Cn_{\rm int}$. C is called the *completeness* of the intensity data and $n_{\rm int}$ is the total number of independent intensities obtainable to resolution limit $d_{\rm min}$. In practice, the count N_i includes the O atoms of ordered water molecules whose positions are refined.

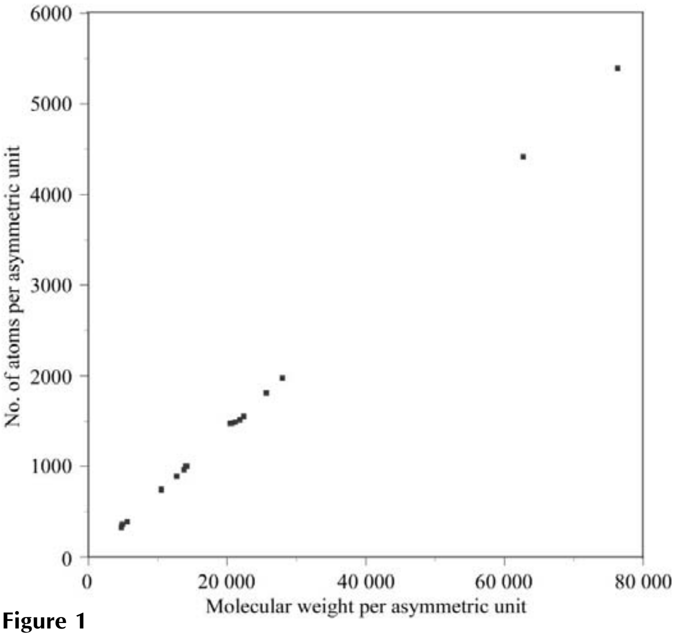

Figure 1
For 18 different proteins used as examples by Cruickshank (1999), the number of non-H atoms in the asymmetric unit (ignoring solvent) is plotted against the molecular weight in the asymmetric unit. Further details are given in §5.

For protein crystallography at low resolution, refinement is carried out with restraints and successful refinement is possible even if p is negative. For this situation, Cruickshank (1999) 'empirically' proposes the use of $R_{\rm free}$ (Brünger, 1992) in place of R and $n_{\rm obs}$ in place of p, to give

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

Both (1) and (2) show the standard deviation of coordinates to be directly proportional to the R factor. However, because of the interactions of $n_{\rm obs}$, C and $d_{\rm min}$, it is difficult to see how the precision depends on the resolution or on the completeness.

As emphasized by Cruickshank (1999), his analysis only applies when the whole structure can be represented by atoms at full occupancy and is not appropriate for structures where a significant fraction of the atoms cannot be assigned unique coordinates.

2.2. The number of independent measurable reflections to a resolution $d_{\rm min}$

If a crystal has a primitive unit-cell volume V, which contains m asymmetric units, the asymmetric unit volume V_a is V/m. To a resolution $d_{\rm min}$, the volume of the accessible sphere of reciprocal space is $(4\pi/3)d_{\rm min}^{-3}$. Assuming Friedel's law to apply, only half of this volume offers independent intensity measurements. For non-centrosymmetric crystals, crystallographic symmetry will cause m intensities in this hemisphere to be identical (except for 'special' reflections with, for example, one index zero). Therefore, the volume of reciprocal space enclosing a full set of independent intensities to resolution $d_{\rm min}$ is $(2\pi/3m)d_{\rm min}^{-3}$.

The volume of the reciprocal unit cell $V^* = 1/V = 1/mV_a$ and there is one intensity associated with each volume V^* of reciprocal space. Thus, the number of independent intensity observations available is

$$n_{\rm int} = \frac{2\pi}{3md_{\rm min}^3}\frac{1}{V*} = \frac{2\pi V_a}{3d_{\rm min}^3}. \qquad (3)$$

A similar formula was given by Blundell & Johnson (1976). Although the formula is not perfectly accurate, because of 'special' reflections and systematic absences arising from screw symmetry and depending upon the particular distribution of reciprocal-lattice points close to the boundary of the sphere of resolution, it is a satisfactory approximation in the context of a 'quick and rough guide'. It may be adjusted to take account of a low-resolution cutoff, where one is applied.

2.3. The number of intensities per dalton depends on the Matthews volume

There is considerable variation in protein crystals with regard to their solvent content. The Matthews volume V_M is the volume of the crystal unit cell divided by the molecular weight of protein which it contains (Matthews, 1968). It is large for protein crystals of high solvent content and usually lies between 1.7 and 3.5 Å^3 Da^{-1}.

Because of the rather uniform composition of protein molecules, predominantly C, N and O atoms, the molecular weight associated with each non-H atom is fairly constant. Even in a metalloprotein or in a protein crystal containing a non-protein ligand or inhibitor, this quantity varies only slightly. For 18 of the 19 proteins and protein complexes used as examples by Cruickshank (1999), the mean molecular weight per non-H atom (ignoring solvent atoms) w is 14.12, with a standard deviation of 2.0% (Fig. 1) (see §5).

The Matthews volume $V_{\rm M}$ can be combined with (3) to estimate the ratio of the available number of independent intensity observations to the number of atoms. The asymmetric unit volume V_a may be written as $MV_{\rm M}$, where M is the molecular weight associated with each crystal asymmetric unit, ignoring solvent. The number of independent intensities per dalton of the asymmetric unit's molecular weight is then

$$\frac{n_{\rm int}}{M} = \frac{2\pi V_{\rm M}}{3d_{\rm min}^3}.$$

Writing $M = wN_{\rm atoms}$, this becomes

$$\frac{n_{\rm int}}{N_{\rm atoms}} = \frac{2\pi w V_M}{3d_{\rm min}^3}. \qquad (4)$$

The existence of a significant amount of phosphorus increases w significantly in nucleic acids. The presence of phosphorus also makes the assumption that all atoms are 'the same type' less valid.

2.4. The resolution necessary to achieve unrestrained refinement depends on the Matthews volume

The DPI takes no account of restraints and deals with direct refinement of parameters from observations. In this case, coordinate refinement is only possible if the number of experimental observations exceeds the number of variables to be refined,

$$n_{\rm obs} > qN_i + v,$$

where q is the number of parameters to be refined for each non-H atom (frequently four, as mentioned by Cruickshank, 1999) and v enumerates other parameters such as overall scale factor which are included in the refinement. In all practical cases v is negligible compared with qN_i and it will be ignored. Using $n_{\rm obs} = Cn_{\rm int}$, this leads to

$$n_{\rm obs} = \frac{2\pi CMV_{\rm M}}{3d_{\rm min}^3} > qM/w.$$

(It is important to emphasize that when part of the structure is disordered and is not approximated by atoms at full occupancy, then qM/w is no longer a good estimate of $N_{\rm i}$. The analysis given here may not apply.)

Hence, for unrestrained refinement to be possible,

$$d_{\rm min} < (2\pi wCV_{\rm M}/3q)^{1/3} \qquad (5)$$

and using the values q = 4, w = 14.1, C = 1, this leads to $d_{\rm min} < 1.94V_{\rm M}^{1/3}$. For a typical $V_{\rm M}$ of 2.4 Å^3 Da^{-1}, this shows that refinement of individual atomic positions and B factors without restraints is impossible for a typical protein if the resolution is worse than about 2.6 Å. Cruickshank (1999) notes cases where p is negative for data at 2.6 and 2.5 Å resolution, respectively. (Standard practice requires a considerable excess of measurements over variables, say three times as many, and unrestrained refinement is inadvisable if $d_{\rm min} < 3^{-1/3} \times 1.94V_{\rm M}^{1/3}$, leading to 1.8 Å for a typical protein.)

2.5. Solvent atoms

Suppose the number of refined solvent atoms $N_{\rm solv}$ to be a fraction s of the atoms in the molecule, so that $s = N_{\rm solv}/N_{\rm atoms}$. The number of parameters $n_{\rm params}$ to be refined is increased by the factor (1 + s). Employing these two expressions, the number of free parameters is given by

$$p = n_{\rm obs} - n_{\rm params} = M\left[\frac{2\pi CV_{\rm M}}{3d_{\rm min}^3} - \frac{q}{w}(1+s)\right]. \qquad (6)$$

3. New forms for the DPI

3.1. A convenient form for calculation

Noting that

$$C = \frac{n_{\rm obs}}{n_{\rm int}} = \frac{3d_{\rm min}^3 n_{\rm obs}}{2\pi V_a},$$

(2) may be rewritten

$$\begin{aligned}\sigma(x, B_{\rm avg}) &= \left(\frac{N_i}{n_{\rm obs}}\right)^{1/2}\left(\frac{3d_{\rm min}^3 n_{\rm obs}}{2\pi V_a}\right)^{-1/3} R_{\rm free} d_{\rm min}\\ &= 1.28N_i^{1/2}V_a^{1/3}n_{\rm obs}^{-5/6}R_{\rm free} \qquad (7)\end{aligned}$$

and in the same way (1) may be written

$$\begin{aligned}\sigma(x, B_{\rm avg}) &= \left(\frac{N_i}{n_{\rm obs} - qN_i}\right)^{1/2}\left(\frac{3d_{\rm min}^3 n_{\rm obs}}{2\pi V_a}\right)^{-1/3} Rd_{\rm min}\\ &= 1.28N_i^{1/2}\left(1 - \frac{qN_i}{n_{\rm obs}}\right)^{-1/2} V_a^{1/3}n_{\rm obs}^{-5/6}R. \qquad (8)\end{aligned}$$

The term in parentheses is close to 1 at high resolution, but may become negative at poor resolution. In the absence of atom sites with partial occupancy, N_i will be the number of atoms in the coordinate file.

3.2. Rearrangement to show the dependence of the DPI on resolution

N_i, the number of ordered scattering atoms in the asymmetric unit, may be represented as $N_{\rm atoms}$(1 + s). Using $N_{\rm atoms} = M/w$ and $n_{\rm obs} = 2\pi CMV_{\rm M}/3d_{\rm min}^3$, (2) may be rewritten

$$\begin{aligned}\sigma(x, B_{\rm avg}) &= [3d_{\rm min}^3(1+s)/2\pi wCV_{\rm M}]^{1/2}C^{-1/3}R_{\rm free}d_{\rm min}\\ &= 0.69[(1+s)/wV_{\rm M}]^{1/2}C^{-5/6}R_{\rm free}d_{\rm min}^{5/2}\\ &= 0.18(1+s)^{1/2}V_{\rm M}^{-1/2}C^{-5/6}R_{\rm free}d_{\rm min}^{5/2}, \qquad (9)\end{aligned}$$

taking the value 14.1 for w. Fig. 2 shows how σ varies with $d_{\rm min}$ for a 'typical' case.

When the conventional R factor is used, the DPI is estimated by (1). Using (6) for p this may be rearranged to

$$\sigma(x, B_{\rm avg}) = 1.0\left[\frac{2\pi w V_{\rm M}}{3(1+s)} - \frac{q d_{\rm min}^3}{C}\right]^{-1/2} C^{-5/6} R d_{\rm min}^{5/2}.$$

If w = 14.1, q = 4, this becomes

$$\sigma(x, B_{\rm avg}) = 0.18\left[\frac{V_{\rm M}}{1+s} - \frac{0.135 d_{\rm min}^3}{C}\right]^{-1/2} C^{-5/6} R d_{\rm min}^{5/2}. \qquad (10)$$

For refinements at good resolution, with many more observations than parameters, the second term in the brackets is a small correction to the first. Apart from this correction, (10) using R is the same as (9) using $R_{\rm free}$. Fig. 3 shows how this correction assumes great importance as the resolution deteriorates.

The difference between (9) and (10) is the correction term in brackets in (10). It may not have any important meaning. Cruickshank uses the word 'empirically' in introducing his use of $R_{\rm free}$. The correction is serious if the number of observations is less than three times the number of variables. The relationship between R and $R_{\rm free}$ is discussed more generally by Tickle *et al.* (1998). In practice, the form using $R_{\rm free}$ appears more meaningful at limited resolution.

4. Discussion

The DPI does not depend on the size of the unit cell or the scattering power of its contents. For larger structures, it may be difficult to achieve such a favourable resolution or such a small R factor, but the DPI has no direct dependence on the size of the structure.

4.1. Equations (9) and (10)

The merit of these two equations is that they plainly show how the coordinate precision depends on quantities which may be within the experimenter's control. The DPI is proportional to the 5/2 power of $d_{\rm min}$. This means, for example, that if the reciprocal resolution $1/d_{\rm min}$ can be extended by a factor 4/3 (say from $d_{\rm min}$ = 2 Å to 1.5 Å), while achieving the same completeness and R factor, the DPI should be halved.

Incomplete data cause coordinate precision to deteriorate. The loss of 20% of the intensity data (C = 0.8) increases the DPI by a factor $C^{-5/6}$ or 1.20. The customary elimination of 5% of intensities from refinement in order to calculate $R_{\rm free}$ increases the DPI only by a factor of 1.04.

(9) shows that the DPI depends on $(1 + s)^{1/2} R_{\rm free}$. This gives a basis for deciding whether the addition of further solvent atoms to represent peaks observed in solvent areas of a difference map can improve the precision of structure analysis. Thus, the addition of extra solvent atoms could reduce the DPI only if they reduced $R_{\rm free}$ by a factor $(1 + s)^{-1/2}$ (Fig. 4). For instance, if $R_{\rm free}$ is currently 0.25, inclusion of solvent atoms equivalent to 20% of the original structure is only justified if $R_{\rm free}$ decreases to less than 0.23 as a result. (Solvent sites ought to be fully occupied for the theory to apply, so this can be no more than a rough guide.)

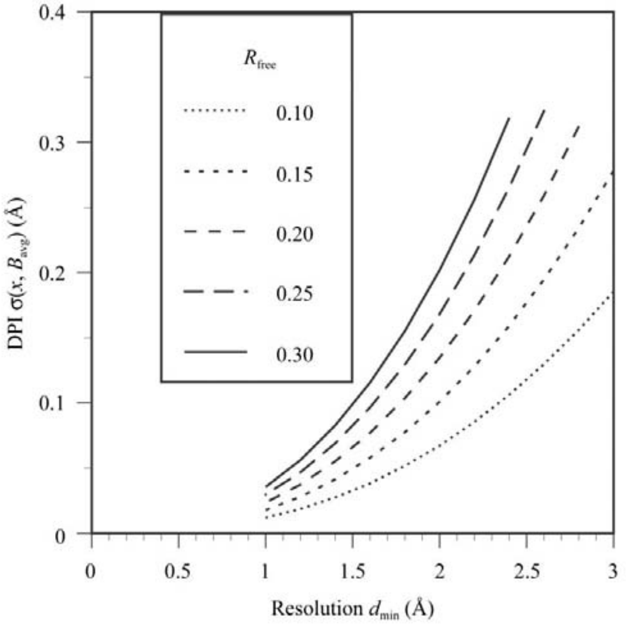

Figure 2
Example of the dependence of the DPI on $R_{\rm free}$ and resolution for a 'typical' protein crystallographic refinement. For this 'typical' case, w = 14.1 and $V_{\rm M}$ = 2.4 Å^3. The completeness C is taken as 0.95, representing complete data with 5% reserved to calculate $R_{\rm free}$. Four parameters are refined per atom and no solvent atoms are included.

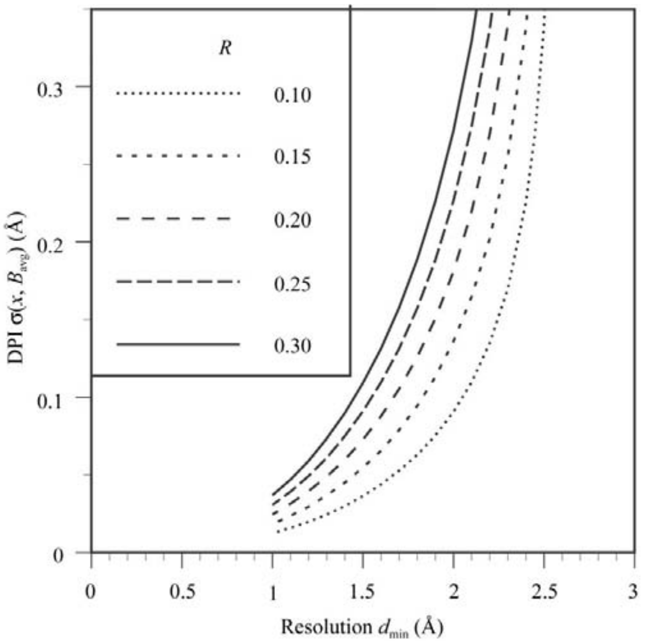

Figure 3
Example of the dependence of the DPI on R and resolution for refinement of the same 'typical' protein structure as in Fig. 2. In this case, C is taken as 1.0 (complete data). It may be observed that the two sets of curves are similar at resolutions superior to 1.8 Å, but the indicated precision deteriorates as the resolution approaches the limit where the number of parameters equals the number of intensity measurements.

Table 1
Comparison of diffraction-component precision index as generated by Cruickshank's formulae and by (7) to (10).

						DPI based on R_{free}			DPI based on R		
Protein	PDB code	s†	V_M† ($\text{Å}^3\,\text{Da}^{-1}$)	R‡	R_{free}‡	Cruickshank‡ (Å)	Eq. (7) (Å)	Eq. (9) (Å)	Cruickshank‡ (Å)	Eq. (8) (Å)	Eq. (10) (Å)
Crambin	1cbn	0.28	1.78	0.090					0.012	0.012	0.011
Rubredoxin	8rxn	0.23	1.68	0.160					0.028	0.028	0.029
Ribonuclease MGMP	1rge	0.29	2.42	0.109					0.027	0.025	0.025
Ribonuclease MSA	1rgh	0.23	2.34	0.106					0.026	0.023	0.025
Plastocyanin 295	1plc	0.15	1.90	0.149					0.061	0.059	0.063
Plastocyanin 173S	1pnc	0.25	1.82	0.132					0.120	0.116	0.132
Plastocyanin 173H	§	0.23	1.82	0.153					0.136	0.132	0.150
TGF-β2 1TGI	2tgi	0.07	3.14	0.173					0.095	0.096	0.094
TGF-β2 1TFG	1tfg	0.09	3.13	0.188					0.136	0.137	0.135
Cd-azurin	1aiz	0.12	2.48	0.168					0.121	0.121	0.120
Lactoferrin	1lfg	0.09	2.79	0.179					0.252	0.245	0.264
Thaumatin C2	1thu	0	2.24	0.184					¶	¶	¶
Concanavalin A	1nls	0.18	2.34	0.128	0.148	0.021	0.020	0.021	0.020	0.018	0.019
HEW lysozyme gr	193l	0.14	2.05	0.184	0.226	0.069	0.067	0.071	0.062	0.061	0.065
HEW lysozyme sp	194l	0.14	2.06	0.183	0.226	0.076	0.074	0.079	0.069	0.067	0.073
γ B crystallin	1gcs	0.16	1.98	0.180	0.204	0.080	0.077	0.084	0.082	0.079	0.088
β B2 crystallin	2bb2	0.06	4.28	0.184	0.200	0.126	0.125	0.127	0.142	0.141	0.144
β purothionin	1bhp	0.21	2.69	0.198	0.281	0.149	0.148	0.148	0.131	0.130	0.129
α1 purothionin	2plh	0.31	2.60	0.155	0.218	0.392	0.375	0.426	¶	¶	¶
EM lysozyme	1jug	0.06	2.12	0.169	0.229	0.162	0.161	0.164	0.172	0.170	0.175
Azurin II (corrected††)	1arn	0.05	2.53	0.188	0.207	0.173††	0.168	0.182	0.237††	0.230	0.270
RNAse A + RI	1dfj	0.01	3.07	0.194	0.286	0.396	0.384	0.418	1.067	¶	¶
FabHyHEL-5 + HEWL	2iff	0.02	2.65	0.196	0.288	0.515	0.532	0.489	¶	¶	¶

† Using data deposited in PDB where not available from original publication. Molecular weight calculated from amino-acid sequence where necessary. ‡ Values given by Cruickshank (1999). § Not in PDB; see Fields *et al.* (1994). ¶ The number of free parameters p is negative. †† Dodd *et al.* (1995) quoted number of reflections 'including anomalous', used as n_{obs} by Cruickshank (1999). The given figures use a revised number of independent reflections calculated from Dodd *et al.* (1995).

4.2. What resolution is achievable?

The formulae give a new slant to the vexed question of deciding the practical limit of resolution. As hinted in the introduction, the resolution which gives the lowest value of R will not usually give the lowest DPI. Cruickshank's formulae indicate that increasing n_{obs} reduces the DPI. However, of course, if this is achieved by 'measuring' reflections which are too weak to provide information, no information is gained and the true precision cannot be improved.

Some authors omit weak 'unobservable' reflections from refinement calculations. They do not always delete these reflections in counting the completeness of data measurement. These matters can be discussed endlessly. It can be argued that the fact that a reflection is weak is a significant observation in itself. This is certainly true at a lower resolution, where most intensities are easily measurable. The other extreme would be to refine far beyond the limit of observable intensities, using large numbers of 'measurements' which are nothing but noise, but recording a larger d_{min}.

For this reason, it is suggested that in deciding the completeness of measurement for use in these equations, 'unobservable' reflections (say less than one or two standard deviations of measurement) should not be included in counting n_{obs}. C in the formulae should represent the fraction of measured and 'observable' reflections. Many of us may argue that unobservable reflections should ideally be included in refinement (with appropriate weight), but this practice will help to prevent unrealistic expectations from extending resolution beyond the practical limit.

5. Validation

5.1. Molecular weight per non-H atom

To estimate w for proteins molecular weights were obtained for 18 of the different proteins used as examples by Cruickshank (1999), omitting one (Fab-HyHEL5) whose molecular weight is not readily available. For proteins whose molecular weights are not stated in the original publication, the molecular weight was generated from the appropriate residues of sequences in the SwissProt database (http://www.expasy.ch) and increased to account for non-protein ligands.

5.2. Test of equations (7) to (10)

These equations were checked by using them to calculate DPI factors in the 23 examples presented by Cruickshank (1999) and the results are summarized in Table 1.[1] There is good agreement except in the case of of azurin II, where confusion arose about the value of n_{obs} (see note to Table 1). The results appear acceptable as a 'quick and rough guide'. The formulae for convenient calculation (7) and (8) give results which are on average 2–3% lower than Cruickshank's, with a standard deviation of 2–3%. Results from (9) and (10) are on average 1–2% higher than Cruickshank's. The standard deviation between (2) and (9) using R_{free} is 4% and is 6% comparing the results of (1) and (10) using R.

[1] Supplementary data have been deposited in the IUCr electronic archive (Reference: li0426). Services for accessing these data are described at the back of the journal.

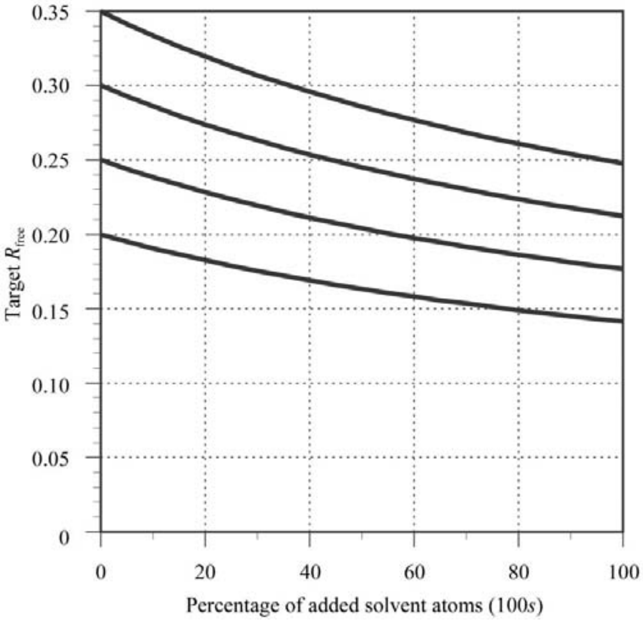

Figure 4
Given $R_{\rm free}$ without added solvent atoms as 0.2, 0.25, 0.3 or 0.35, the curves show the $R_{\rm free}$ to be achieved with added solvent atoms as a given fraction of the structure, if the added solvent is to improve its precision.

I thank Durward Cruickshank for stimulating and constructive comments.

References

Blundell, T. L. & Johnson, L. (1976). *Protein Crystallography*, p. 248. New York: Academic Press.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472-475.
Cruickshank, D. W. J. (1949). *Acta Cryst.* **2**, 65–82.
Cruickshank, D. W. J. (1959). *International Tables for X-ray Crystallography*, Vol. 2, edited by J. S. Kasper & K. Lonsdale, pp. 318–340. Birmingham: Kynoch Press.
Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.
Cruickshank, D. W. J. (2001). *International Tables for Crystallography*, Vol. F, edited by M. G. Rossmann & E. Arnold, pp. 403–414. Dordrecht: Kluwer Academic Publishers.
Dodd, F. E., Hasnain, S. S., Abraham, Z. H. L., Eady, R. R. & Smith, B. E. (1995). *Acta Cryst.* D**51**, 1052–1064.
Fields, B. A., Bartsch, H. H., Bartunik, H. D., Cordes, F., Guss, J. M. & Freeman, H. C. (1994). *Acta Cryst.* D**50**, 709–730.
Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *Acta Cryst.* D**54**, 547–557.