phasing procedure. The uncertainty in individual invariant relationships and the relative instability of existing phasing algorithms makes it possible for less-reliable invariants to succeed occasionally where more precise sets have failed. Nevertheless, it must remain true that the most precise invariants have a statistically better chance of providing a solution independent of the methods used to apply these invariants.

In summary, this study has shown that, for the eleven structures examined, normalized structure factors, estimated from a Wilson plot using an exponential scaling function, the overall rescale and the random-atom expectation value are best suited for use in direct methods.

The authors wish to acknowledge the assistance of the Australian Research Grants Committee (Grant: C7915302) during the tenure of this study.

References

DEBYE, P. (1915). *Ann. Phys.* (*Leipzig*), **46**, 809.
DECLERCQ, J. P., GERMAIN, G. & VAN MEERSSCHE, M. (1972). *Cryst. Struct. Commun.* **1**, 13–15.
DEWAR, R. B. K. (1970). In *Crystallographic Computing Techniques*, edited by F. R. AHMED, pp. 63–65. Copenhagen: Munksgaard.
HALL, S. R. (1978). *Acta Cryst.* A**34**, S348.
HALL, S. R., RASTON, C. L. & WHITE, A. H. (1978). *Aust. J. Chem.* **31**, 685–688.
HALL, S. R. & SUBRAMANIAN, V. (1980). *ESCAN*. Program for the comparison of E values. XRAY76 System. Univ. of Western Australia.
HALL, S. R. & SUBRAMANIAN, V. (1982*a*). *Acta Cryst.* A**38**, 590–598.
HALL, S. R. & SUBRAMANIAN, V. (1982*b*). *Acta Cryst.* A**38**, 598–608.
HAUPTMAN, H. (1975). *Acta Cryst.* A**31**, 680–687.
KARLE, I. L. (1976). In *Crystallographic Computing Techniques*, edited by F. R. AHMED, pp. 27–70. Copenhagen: Munksgaard.
KARLE, J. & HAUPTMAN, H. (1953). *Acta Cryst.* **6**, 473–476.
LADD, M. F. C. (1978). *Z. Kristallogr.* **147**, 279–296.
MAIN, P. (1976). In *Crystallographic Computing Techniques*, edited by F. R. AHMED, pp. 97–105. Copenhagen: Munksgaard.
MAIN, P., FISKE, S. J., HULL, S. E., LESSINGER, L., GERMAIN, G., DECLERCQ, J. P. & WOOLFSON, M. M. (1980). *MULTAN*80. Univ. of York.
SKELTON, B. W. & WHITE, A. H. (1981). Personal Communication.
WILSON, A. J. C. (1942). *Nature* (*London*), **150**, 151.

Acta Cryst. (1982). A**38**, 590–598

Normalized Structure Factors. II. Estimating a Reliable Value of B

BY S. R. HALL AND V. SUBRAMANIAN†

Crystallography Centre, University of Western Australia, Nedlands 6009, *Australia*

(*Received* 6 *May* 1981; *accepted* 20 *January* 1982)

Abstract

A reliable estimate of the overall temperature factor B is shown to be important to the calculation of normalized structure factors, and to the application of structure-invariant phasing methods. Methods for obtaining improved estimates of B from the Wilson plot procedure are examined. The use of Bayesian statistics, the inclusion of missing data, the application of least-squares weights and the compensation for Debye scattering effects in the Wilson plot are considered. Estimates of B are compared for fourteen refined structures, including three proteins.

† Deceased 27 December 1981.

0567-7394/82/050590-09$01.00

Introduction

The standard method for estimating the overall temperature factor B and the structure-factor scale **k** from measured intensity data is by a linear least-squares fit to data in a Wilson plot (Wilson, 1942). In this plot of $\ln[|F_h^2|/\langle|F_h^2|\rangle]$ *versus* s^2 the slope of the fitted line is $-2B$ and the intercept at $s^2 = 0$ is $-2\ln(\mathbf{k})$. Because the Wilson-plot method is simple and computationally convenient, it is widely used in many crystallographic laboratories for scaling data. It is therefore surprising that the computer programs applying this technique often produce quite different estimates of B and **k** from the same data. In fact, it is not uncommon for estimates to differ by as much as a

factor of two from program to program (see Tables 1 and 2). Estimates of B also depart significantly from the 'true' (refined) values of B.

Moreover, it has been demonstrated that the exponential scale $\mathbf{k}\exp(Bs^2)$ provides consistently better estimates of normalized structure factors than the other commonly used alternatives (Subramanian & Hall, 1982). This emphasizes the need for procedures that would reliably estimate B and $\mathbf{k}$ values for routine structure analysis. In this study the factors which are most important to this objective are identified. They include the treatment of weak intensity data; the compensation for missing data; the averaging and weighting procedures; and the use of a selective least-squares process to account for the effects of Debye scattering. Fourteen refined structures, including those of the three proteins rubredoxin, insulin, and crambin are used to examine the effect of these factors on the Wilson plot process (see Table 1). A general procedure incorporating these features has been applied in the program *GENEV* (Hall, 1981) for the *XTAL* system (Hall, Stewart & Munn, 1980).

Table 1. *Test structures*

$$R = \sum||F_o| - |F_c||/\sum|F_o|.$$

	Formula	Space group	R value	Reference
HCPP	$C_{15}H_{16}O_{10}$	$P\bar{1}$	0·037	(*a*)
CLEPX	$C_{30}H_{18}Cl_4$	$P\bar{1}$	0·037	(*a*)
BEKA4	$C_{58}H_{90}N_2O_6$	$P\bar{1}$	0·055	(*a*)
STIK4	$C_{14}H_{18}O_6$	$P2_1$	0·050	(*a*)
PDCPS	$C_{42}H_{37}Cl_2F_6PPdSb_2$	$P2_1/c$	0·051	(*a*)
CANON2	$C_{18}H_{18}O_5$	$P2_1/n$	0·058	(*b*)
ANTH1	$C_{34}H_{26}O_4$	$P2_1/c$	0·034	(*a*)
TEMPL	$C_{21}H_{34}O_4N_3Cl$	$P2_12_12_1$	0·056	(*a*)
CORT	$C_{21}H_{28}O_5$	$P2_12_12_1$	0·058	(*a*)
K22BR	$C_{35}H_{48}O_6$	*Iba*2	0·049	(*c*)
KCPP	$C_{16}H_{19}KO_{11}$	*Pcab*	0·042	(*a*)
RUBRDN	rubredoxin	*R*3	0·126	(*d*)
INSULN	2-Zn pig insulin	*R*3	0·113	(*e*)
CRAMBN	crambin	$P2_1$	—	(*f*)

References: (*a*) Skelton & White (1981); (*b*) Hall, Raston & White (1978); (*c*) Declercq, Germain & Van Meerssche (1972); (*d*) Watenpaugh, Sieker, Herriott & Jensen (1973); (*e*) Isaacs & Agarwal (1978); (*f*) Teeter & Hendrickson (1979).

The importance of a reliable B estimate

The 'true' B value for each test structure was obtained from a least-squares fit to the Wilson-plot ratios evaluated with an expectation value $\langle|F_h^2|\rangle$ calculated from the refined atomic coordinates [see expression (6) in Subramanian & Hall, 1982]. Typical Wilson plots for four of the test structures (BEKA4, CANON2, CORT and KCPP) are shown in Fig. 1 with the data points (shown as *) based on the refined structure-factor expectation value. The Wilson-plot ratios, based on the random-atom expectation value, are shown as O

Table 2. *Estimated overall temperature factor and scale*

Estimated B (in Å^2) and k values obtained using random-atom expectation values. All data are rescaled so that $|\overline{E}|^2 = 1{\cdot}0$.

		Temperature factors					Structure factor scales				
Test	s^2 (max)	Refined (1)	*GENEV* (2)	*EVAL* (3)	*NORMSF* (4)	*NORMAL* (5)	Refined (6)	*GENEV* (2)	*EVAL* (3)	*NORMSF* (4)	*NORMAL* (5)
HCPP	0·29	4·6	4·5	4·0	3·8	3·9	0·140	0·147	0·160	0·165	0·170
CLEPX	0·25	4·0	3·8	3·3	3·1	3·1	0·269	0·270	0·297	0·303	0·354
BEKA4	0·24	4·5	5·0	3·8	3·1	3·2	0·285	0·279	0·349	0·372	0·385
STIK4	0·36	4·8	4·5	3·8	4·1	4·0	0·157	0·159	0·183	0·171	0·182
PDCPS	0·42	3·6	3·5	3·4	3·5	3·4	1·200	1·268	1·274	1·264	1·300
CANON2	0·24	3·8	4·0	2·7	2·0	2·1	0·079	0·079	0·096	0·106	0·112
ANTH1	0·36	4·5	4·4	4·4	4·6	4·4	0·140	0·148	0·152	0·142	0·154
TEMPL	0·29	4·4	4·6	3·8	3·9	3·7	0·462	0·479	0·556	0·549	0·583
CORT	0·32	3·3	3·3	2·8	3·3	3·2	0·978	0·952	1·081	0·975	1·042
K22BR	0·22	4·8	5·4	3·8	3·3	3·2	0·917	0·894	1·140	1·219	1·315
KCPP	0·36	3·2	3·0	2·5	2·9	2·8	0·840	0·922	1·039	0·908	0·992
RUBRDN	0·10	12·9	13·5	9·5	9·9	10·3	0·362	0·356	0·431	0·425	0·455
INSULN	0·11	15·4	15·4	15·4	15·0	15·4	1·240	1·242	1·242	1·290	1·275
CRAMBN	0·11	7·7*	7·8	7·6	8·1	8·3	1·069*	1·076	1·093	1·054	1·069

Column 1 B values obtained from Wilson plot using expectation value calculated from refined atomic positions.
Column 2 Program *GENEV* (Hall, 1981) including Bayesian, range-fill and inflection-point least squares.
Column 3 Program *EVAL* (Hall, 1978).
Column 4 Program *NORMSF* (Hall, (1972).
Column 5 Program *NORMAL* (Main *et al.*, 1980).
Column 6 k values obtained from last refined cycle of structure-factor least squares, except in the case of RUBRDN and INSULN where k's were obtained in the renormalization procedure to set $|\mathscr{E}(\mathbf{h})|^2 = 1.0$.
* Values supplied by Hendrickson (1980).

and are connected by a solid line. The near-linear relationship between the 'refined' ratios and s^2 provides a relatively precise estimate of the true overall B for each structure.

There are a variety of factors responsible for the large differences in B estimates given in Table 2. Before analysing these factors it is important to establish that errors in B and $\mathbf{k}$ do have a significant effect on the reliability of the structure-invariant relationships.

The four structures CLEPX, BEKA4, CANON2 and K22BR were selected to test the effect of the large discrepancies in estimated B values. The methods employed in this study are similar to those used in the analysis of scaling functions by Subramanian & Hall

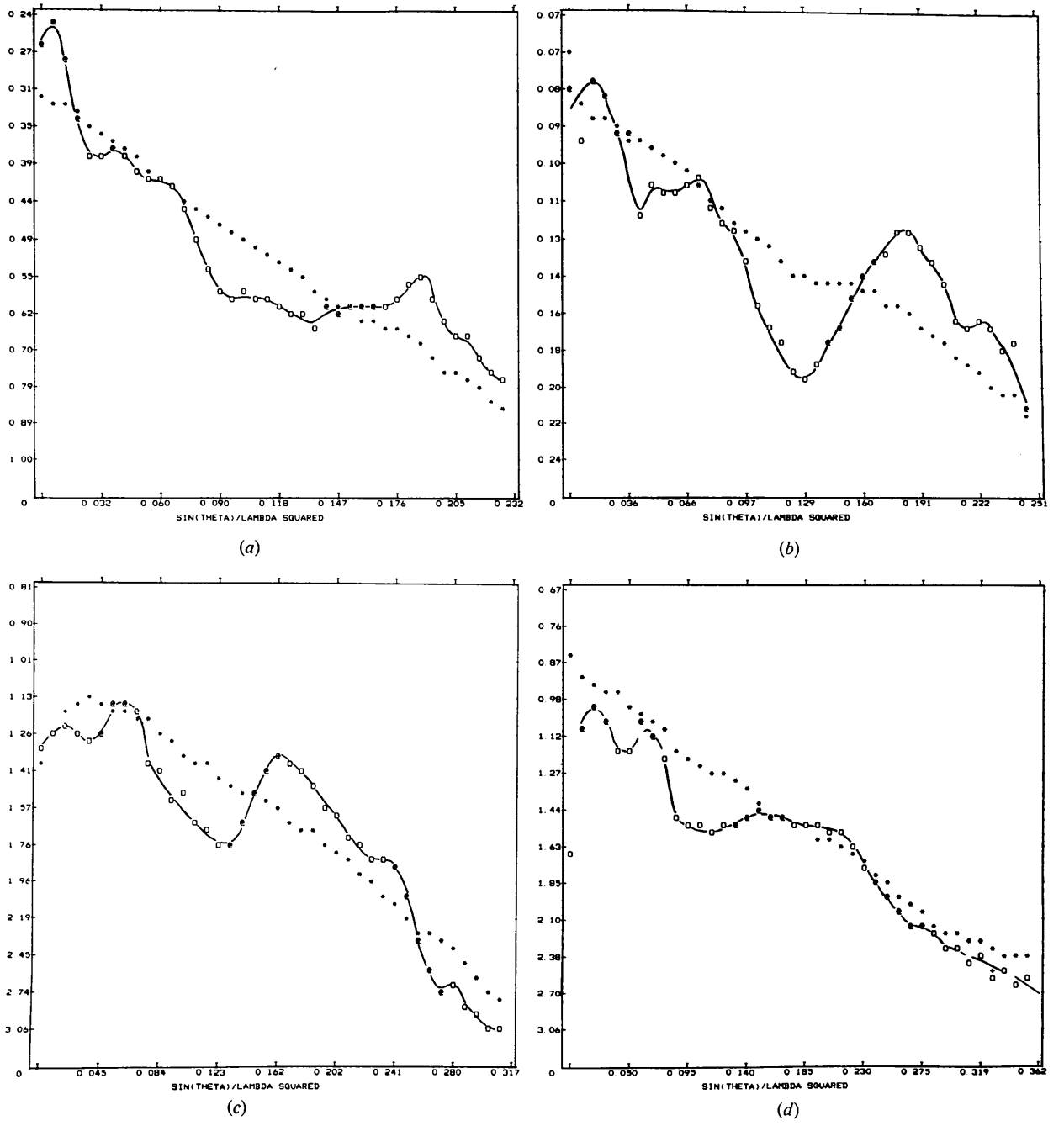


Fig. 1. Wilson plots for the test data of (*a*) BEKA4, (*b*) CANON2, (*c*) CORT and (*d*) KCPP. The abscissa axis is s^2 and the ordinate axis is $\ln[|F^2|/\langle|F^2|\rangle]$; the scale is in terms of the $|F|$-scale $\mathbf{k}$. Data points denoted by O are for the random-atom expectation value $\langle|F^2|\rangle = \varepsilon\sum f^2$. Data points denoted by * are for $\langle|F^2|\rangle$ calculated from the refined atomic coordinates [see expression (6) of Subramanian & Hall (1982)].

(1982). Two sets of E values were calculated: one based on the refined B, and the other on the lowest of the B values estimated by Wilson-plot software. The results of mean $\Delta E^2 (=||E^2| - |\mathscr{E}^2||)$ and $\Delta E^2/|\mathscr{E}_h^2|$, where $|\mathscr{E}_h|$ is the 'true' normalized structure factor, are tabulated in Table 3. The correspondence between calculated $|\mathscr{E}_h|$ and the estimated $|E_h|$ values is poorest for those based on the low B values.

The application of refined phases to sets of triplet and quartet structure-invariant relationships generated for the different $|E_h|$ values provides information on their reliability in phasing procedures. One test, the percentage violations for triplets and quartets, is shown in Table 4. A much more sensitive criterion of invariant reliability, the weighted root-mean-square differences for ψ_3 and ψ_4, is tabulated in Table 5.

The results of these tests show clearly that inaccurate B values give rise to less accurate structure invariants and therefore a potential increase in phasing errors. The magnitudes of these errors are not large; but nevertheless their effect on the phasing process can be significant. It is well recognized in some phasing procedures that small changes in phase reliability often have a profound effect on the structure solution process. It is the sensitivity of structure-invariant procedures to these small changes, particularly in the initial stages, that makes the need for the best possible B and **k** estimates so crucial.

Table 3. *Comparison of* $|\mathscr{E}_h|$ *and* $|E_h|$ *for different B values*

ΔE^2 and $\Delta E^2/\mathscr{E}^2$ are the mean difference and the mean fractional difference between the $|E_h|$'s estimated using the listed B value and the calculated quasi-normalized structure factors $|\mathscr{E}_h|$. All differences are for $|\mathscr{E}_h| > 1{\cdot}0$.

	B (Å^2)	ΔE^2	$\Delta E^2/\mathscr{E}^2$	B (Å^2)	ΔE^2	$\Delta E^2/\mathscr{E}^2$
CLEPX	4·0	0·49	0·22	3·1	0·53	0·23
BEKA4	4·5	0·80	0·32	3·1	0·88	0·35
CANON2	3·8	0·87	0·31	2·0	1·02	0·36
K22BR	4·8	0·76	0·36	3·2	0·81	0·37

Table 4. *Percentage of structure invariants violated*

T, PQ and NQ are the percentages of triplet, positive quartet and negative quartet with estimated ψ values that differ from $\langle\psi\rangle$ by more than $\pi/2$. These are compared for the refined B and the lowest estimated B (see Table 2). See Subramanian & Hall (1982) for definitions. B (Å^2) is the B value used in the estimation of E_h.

Test	B (Å^2)	T	PQ	NQ	B (Å^2)	T	PQ	NQ
CLEPX	4·0	2·4	0·6	12·7	3·1	2·1	0.7	11·9
BEKA4	4·5	4·8	0·6	35·3	3·1	7·9	2·3	39·7
CANON2	3·8	9·8	0·4	2·6	2·0	9·6	13·9	5·1
K22BR	4·8	7·4	—	—	3·2	13·3	—	—

Table 5. *Weighted phase discrepancies*

R.m.s.d. ψ_3 and ψ_4 are the weighted root-mean-square differences between the estimated and expected values (in degrees). These are compared for the refined B and the lowest estimated B (see Table 2). See Subramanian & Hall (1982) for definitions. B (Å^2) is the B value used in the estimation of $|E_h|$.

	B (Å^2)	R.m.s.d. ψ_3	R.m.s.d. ψ_4	B (Å^2)	R.m.s.d. ψ_3	R.m.s.d. ψ_4
CLEPX	4·0	5·	6·	3·1	11·	10·
BEKA4	4·5	19·	48·	3·1	26·	70·
CANON2	3·8	37·	3·	2·0	37·	8·
K22BR	4·8	44·	—	3·2	52·	—

Treatment of weak data

It is a common practice in the reduction of raw intensity data to classify reflections with intensities below a certain threshold (say $3\sigma I$) as unobserved and to consider them as an unreliable source of structural information. For this reason, unobserved reflections are frequently excluded entirely from certain calculations. There are practical as well as historical reasons for this. In the past all visually measured film intensities below the calibration-strip limit were, indeed, unobserved, and their exclusion meant substantial savings in computing time with little apparent effect on the results. Mainly through the growth of accurate electron density studies, a better appreciation of the importance of low-intensity data has developed in recent years; nevertheless, there still remains in many laboratories considerable inertia to use weak data in every calculation.

The correct treatment of weak data is particularly important to Wilson-plot estimates of B and **k**. Thermal motion and X-ray scattering effects cause the average intensity to decrease rapidly with s^2. As a result, weak data tend to predominate in the high-angle regions of a Wilson plot and give rise to significant systematic errors in the B and **k** estimates.

Until recently there was little agreement on how negative or weak intensities should be included in crystallographic calculations. This has been convincingly resolved by French & Wilson (1978) who emphasize the Bayesian nature of intensity statistics. The procedure proposed by French & Wilson provides a statistically correct method for estimating the structure factors and their standard deviations for net intensities below 3σ.

Limited Bayesian treatment of weak data

The correct application of the procedure of French & Wilson (1978) requires a careful analysis of the raw intensity data and some independent measurements of equivalent reflections. The complete Bayesian treatment they propose is highly desirable for accurate electron density studies. However, because of the

additional data measurement and computational requirements it is unlikely that data used in routine structure analyses will be treated in this way. Yet in such cases it is possible to improve the overall precision of the intensities by a partial application of Bayesian statistics to data with $I < 3\sigma I$. This is of particular benefit to data sets with a relatively large proportion of negative intensities, which would otherwise be used in the Wilson-plot process as zero.

A simple one-pass approach to Bayesian statistics is possible provided that the average intensity for all shells of reciprocal space is assumed to be constant. Whereas this is a poor approximation to the full Bayesian treatment, it still is able to provide estimates for the measured structure factor $|F_{\mathbf{h}}^m|$ that are better than the arbitrary value of zero. In this study, this limited Bayesian approach was adopted, assuming a uniform average intensity of $20\sigma I$. Thereafter, the following steps were applied:

(1) For reflections with $I < 3\sigma I$ the ratio of the *net* intensity I and its standard deviation σI is calculated as

$$J = I/\sigma I \quad \text{for} \quad J = -3 \text{ to } +3.$$

(2) From Table 6 [an abbreviation of Table 1 from French & Wilson (1978)] the appropriate values of $R_F(J)$ and $R_S(J)$ are selected.

(3) The new values of $|F_{\mathbf{h}}^m|$ and $\sigma|F_{\mathbf{h}}|$ are estimated as

$$|F_{\mathbf{h}}^m|^2 = R_F^2(J)\,\sigma I\,\mathrm{Lp} \qquad (1)$$

$$\sigma^2(F_{\mathbf{h}}) = R_S^2(J)\,\sigma I\,\mathrm{Lp}, \qquad (2)$$

where Lp is the Lorentz–polarization factor ($|F_{\mathbf{h}}^2| = \mathrm{Lp}I$).

The effect of this simple correction was tested with the weak data sets of the structures HCPP, ANTH1, and KCPP. In each case the B estimate improved by about 0·1 Å^2.

Compensation for missing weak data

The effect of completely omitting weak data is predictable. It causes the average intensity for a given s^2 range to increase. Since the proportion of weak data increases with s^2, the B value estimated from a Wilson plot tends to be too low. This in turn results in fewer than expected large $|E_{\mathbf{h}}|$ values at high angles. Therefore, if reasonable B estimates are to be obtained, it is necessary to have some sort of compensation procedure for missing data in the Wilson-plot algorithm.

Table 6. *Bayesian posterior moments*

Abbreviated table of French & Wilson (1978). Assumes that $I/\sigma I$ for all shells is 20.

	Centrosymmetric distribution		Noncentrosymmetric distribution	
J	R_F	R_S	R_F	R_S
−3	0·30	0·23	0·47	0·24
−2	0·36	0·26	0·54	0·27
−1	0·44	0·31	0·65	0·31
0	0·57	0·37	0·81	0·38
1	0·82	0·44	1·06	0·36
2	1·22	0·45	1·37	0·31
3	1·62	0·35	1·69	0·25

All Wilson-plot procedures compute for each of the n ranges sums of the type

$$\sum m, \sum m|F_{\mathbf{h}}^2|, \sum m\varepsilon \sum^{N} f^2 \text{ and } \sum ms^2,$$

where m is the reflection multiplicity. The expected reflection population of the ith shell of reciprocal space is simply

$$n_i = \frac{32}{3V^*}(s_i^3 - s_{i-1}^3), \qquad (3)$$

where V^* is the volume of the reciprocal cell. The number of missing reflections in the ith shell is therefore

$$\Delta_i = n_i - (\sum m)_i. \qquad (4)$$

If the mean $|F_{\mathbf{h}}^2|$ of the missing reflections is known, then the sums for the ith range can be suitably adjusted.

An estimate of the mean $|F_{\mathbf{h}}^2|$ for missing data can be made in several ways. One way is to calculate the average $\sigma|F_{\mathbf{h}}^2|$ for the weaker intensities in the ith range, and adjust each Wilson-plot sum with an $|F_{\mathbf{h}}^2|$ value estimated from $\sigma|F_{\mathbf{h}}^2|$ using Bayesian statistics. A simpler variation to this approach is to set the missing $|F_{\mathbf{h}}^2|$ values equal to $q\sigma|F_{\mathbf{h}}^2|/2$, where $q\sigma|F_{\mathbf{h}}^2|$ is the known data cut-off value. Another way, useful when $\sigma|F_{\mathbf{h}}^2|$ estimates are not available, is to find the $(|F_{\mathbf{h}}^2|)_{\min}$ value in each range and add the missing reflections with $|F_{\mathbf{h}}^2| = |F_{\mathbf{h}}^2|_{\min}/2$.

This last approach was used to compensate for the 1850 reflections omitted from the data set of the protein rubredoxin, RUBRDN (Watenpaugh, Sieker, Herriott & Jensen, 1973). Fig. 2(*a*) shows the Wilson plot with the data points before and after compensation denoted by * and ○, respectively. The B values estimated from these plots were 9·5 and 13·5 Å^2, respectively. The latter value is within 5% of the refined value given in Table 1.

Wilson-plot least-squares weights

The application of least squares to the Wilson-plot process requires the correct evaluation of the number and the variance of the contributing reflections. In general experience has shown that the application of least squares without pre-averaging into reciprocal-space shells tends to be unreliable because the estimates of $\sigma|F_{\mathbf{h}}^2|$ have an unacceptably high correlation with

$|F_h^2|$. The Wilson plot may be considered as a set of **r** points $p_i(R, s^2)$, **i** = 1 to **r**, where $\overline{s^2}$ is the mean s^2 for the range, and R the natural logarithm of the ratio of the mean $|F_h^m|^2$ and the mean expectation value $\langle|F_h^2|\rangle$. The only parameter in the ratio term for the ith shell, R_i, to contain significant experimental errors, is the value $|F_h^2|$.

$$R_i = \ln \frac{(\overline{|F_h^2|})_i}{(\langle|F_h^2|\rangle)_i}. \tag{5}$$

The mean value of $|F_h^2|$ for the ith shell is

$$(\overline{|F_h^2|})_i = \frac{\sum^{n_i} w_h |F_h^2|}{\sum^{n_i} w_h}, \tag{6}$$

where the n_i is the number of reflections contributing to the ith shell, and w_h is the individual weight associated with each value of $|F_h^2|$. By definition,

$$w_h = (\sigma^2 |F_h^2|)^{-1}. \tag{7}$$

This leads directly to the variance of the mean $|F_h^2|$ for the ith shell as

$$\sigma^2(\overline{F^2})_i = \left(\sum^{n_i} w_h\right)^{-1}. \tag{8}$$

In the Wilson-plot least-squares process each point should be weighted with the reciprocal variance

$$\omega_i = (\sigma^2 R_i)^{-1}. \tag{9}$$

The variance of R_i can be obtained in terms of its components by differentiating (5) with respect to $|F_h^2|$,

$$\sigma^2 R_i = \frac{\sigma^2(\overline{F_h^2})_i}{(F_h^2)_i^2}. \tag{10}$$

Expanding (10) from (8)

$$\sigma^2 R_i = \left[(\overline{F^2})_i^2 \left(\sum^{n_i} w_h\right)\right]^{-1} \tag{11}$$

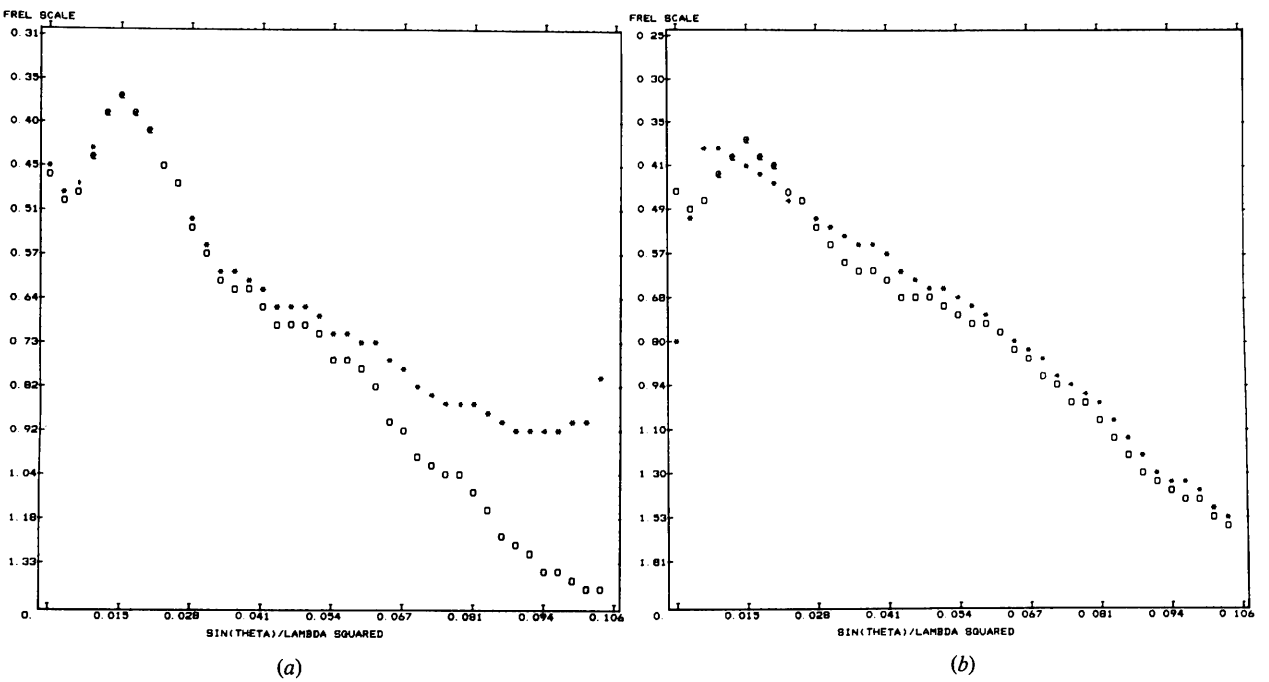
and the least-squares weight for the ith shell from (9) is

$$\omega_i = (\overline{F^2})_i^2 \sum^{n_i} w_h. \tag{12}$$

The individual reflection weight w_h can be expressed in terms of the measured structure factor $|F_h|$ and its standard deviation $\sigma|F_h|$

$$w_h = 2|F_h|\sigma|F_h|. \tag{13}$$

The application of the least-squares weights ω_i to the Wilson plot using values of w_h calculated from (13) is usually unreliable. Its unreliability arises principally because the reflection weight is dependent on the terms $|F_h^2|$ and $\sigma|F_h^2|$ and these are generally highly correlated. In other words, the weight expression contains a systematic bias due to the estimate of $\sigma|F_h^2|$ being dependent on the value of $|F_h^2|$. This situation is analogous to several other crystallographic cal-



(*a*) (*b*)

Fig. 2. Wilson plots for the test data from the protein rubredoxin, RUBRDN. Abscissa and ordinate scales are as defined for Fig. 1. (*a*) Wilson plots obtained with random-atom expectation value: the * data points are *before* compensation for missing reflections; the O points correspond to data *after* compensation. (*b*) Wilson plot using the expectation value given by refined atomic coordinates is denoted as *. The O data points are defined as in (*a*).

culations involving individual weights, such as weighted Fouriers (Davis, Maslen & Varghese, 1978).

For the test structures used in this study all values of $\sigma|F_{\mathbf{h}}^2|$ were derived from intensity statistics alone and this precluded the application of $w_{\mathbf{h}}$ as in (13). For this reason $w_{\mathbf{h}}$ were set to unity. The average weight ω_i for each value of R_i then simplifies to

$$\omega_i = n_i(\overline{F^2})_i^2. \qquad (14)$$

The application of ω_i from (14) also proved unreliable because it gives low weight to the high-angle points that are so critical to the estimation of the B value.

In summary, the application of least-squares weights based on $\sigma|F_{\mathbf{h}}^2|$ proved to be impractical because of its strong correlation with $|F_{\mathbf{h}}^2|$. The most reliable weight, ω_i, for each point $p_i(R, \overline{s^2})$ of a Wilson plot was found to be unity.

There is also another reason why the application of the above weights to the Wilson-plot least squares was not effective. As a general rule statistical weights are only applicable in a linear least-squares process in the absence of significant systematic errors. It is a common occurrence in a Wilson plot that data points depart from the fitted least-squares line by over three standard deviations. These deviations are mainly due to the presence of short-range translational symmetry in the structure (Debye, 1915) and this largely invalidates the normal use of ω_i in the Wilson plot.

Allowance for Debye scattering

In practice the reliability of a B estimated from a Wilson plot is much more dependent on allowing for Debye scattering effects than on the correct least-squares weights. This is particularly true if reflection data are truncated at an s^2 value where the Debye effects are large. In this study the test structures with the lowest $s_{\max}^2$ had consistently worse estimates of B from conventional Wilson-plot least-squares procedures (see Table 2).

Two methods of allowing for Debye scattering effects were studied using four of the test structures, CLEPX, BEKA4, CANON2 and K22BR. In one method, known conformational information was incorporated in the expectation value used in the Wilson-plot ratio. In another, points in the Wilson plot which were less dependent on Debye scattering effects were used in the linear least-squares process.

Debye expectation value

A commonly used approach to compensate for Debye scattering effects incorporates known conformational information in a Debye form of the expectation value $\langle|F_{\mathbf{h}}^2|\rangle$ used in the Wilson-plot procedure (Main, 1976; Main *et al.*, 1980; Hall, 1978). The Debye expectation value has the form

$$\langle|F_{\mathbf{h}}^2|\rangle = \sum_j^N \sum_k^N f_j f_k \frac{\sin 4\pi s d_{jk}}{4\pi s d_{jk}}, \qquad (15)$$

where d_{jk} is the distance in ångströms between the jth and kth atoms. Because the Debye expectation value is usually a closer approximation to the mean value of $|F_{\mathbf{h}}^2|$ than the random atom for each range, the resulting Wilson-plot ratio [see (5)] will be a smoother function of s^2 [see Fig. 7 of Subramanian & Hall (1982)]. It follows that provided sufficient conformational information is available for inclusion in (15), this might be expected to improve the Wilson-plot estimate of B and **k**.

Experience has shown, however, that the results of this approach are often disappointing. The application of the Debye expectation value to CANON2 and CORT using the rigid fragment of these molecules did not provide significant improvements in the estimate of B. The Debye estimate of B for CANON2 was marginally better than the random-atom value by 0·1 Å^2 but the value for CORT was worse by the same amount. These and other tests suggest that the use of the Debye expectation value results only in marginal, if any, improvements in the estimate of B.

There are also several drawbacks to this approach. The first is the need to know *a priori* conformational information of the structure. The second is the time and effort required to prepare the coordinate information used in the calculation of the Debye expectation value. For particularly difficult analyses the additional effort would be worthwhile even for marginal improvements to the estimates of B, but for routine analyses application of the Debye expectation value appears unwarranted. It should also be noted that the use of Debye expectation value in the calculation of normalized structure factors is not recommended for routine phasing procedures (Ladd, 1978; Subramanian & Hall, 1982).

Inflexion-point least squares

A more effective approach to account for Debye scattering in Wilson-plot least squares is proposed here. It is based on the close similarity of Debye scattering curves for a broad range of structure types. This similarity is expected because the structural components that give rise to the dominant features of a Debye curve are generally present in most structures. To illustrate this, a simple six-membered ring system was selected as typical of interatomic distances found in a large number of structures.

A series of scattering curves based on the ratio of the Debye and random-atom expectation values for this six-atom ring structure was calculated for a range of s^2.

These curves are shown in Fig. 3 for the nearest-neighbour C–C distances of 1·5, 1·4 and 1·3 Å. They are characterized by two dominant features; a large peak at about $s^2 = 0{\cdot}2$, and a large trough at about $s^2 = 0{\cdot}1$. Less conspicuous but of equal importance is the low-angle portion of the curve where the ratio approaches the value of 1·0. These are familiar features in the Wilson plots for a wide range of molecular structures, and particularly those with the hexagonal motif. The peak close to $s^2 = 0{\cdot}2$ is frequently responsible for poor estimates of B, especially if s^2_{max} is close to 0·2. This is the case for the data sets of CLEPX, BEKA4, CANON2 and K22BR (see Table 2).

It is also evident from Fig. 3 that the Debye curve 'inflexion points' on either side of $s^2 = 0{\cdot}2$ are close to the median line of the Wilson plot. In theory it should be possible to calculate the position of the Debye inflexion points from $d^2R'/d^2(s^2) = 0$, where R' $(=R + 2Bs^2)$ is the Wilson ratio R [see (5)]. Although an initial estimate of B may be obtained by conventional methods it is difficult to estimate reliably the change of slope of R'. Numerical interpolation methods are cumbersome even with elaborate smoothing techniques. Fitting a polynomial function to R' is potentially a more reliable approach but it is still likely to require iterative procedures.

There is a less complicated method of utilizing inflexion points. Fig. 3 shows that the inflexion points on either side of the 0·2 peak are within quite predictable zones of s^2. This observation is well supported by the Wilson plots for a wide range of structures. In addition, the precision of the s^2 value used to locate the inflexion point of the Debye curve does not seem critical to the Wilson-plot process. To illustrate this, Wilson-plot inflexion points for all test structures were assumed to occur at fixed s^2 values: five data points closest to $s^2 = 0{\cdot}15$ and five closest to $s^2 = 0{\cdot}26$ were selected for a least-squares calculation. Also included in this calculation were five points with the highest values of R. These latter values provide low-angle information which is relatively insensitive to the Debye fluctuation (see Fig. 2). Each point is weighted according to the number of contributors to that shell.

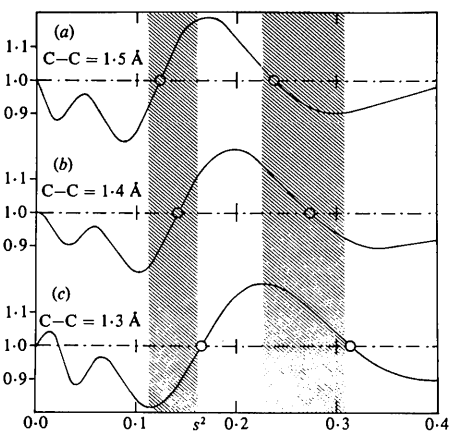


Fig. 3. Plots of the ratio $\sum_j^6 \sum_k^6 f_j f_k (\sin 4\pi s d_{jk}/4\pi s d_{jk}) / \sum^6 f_j^2$ *versus* s^2 for the interatomic distances d_{jk} of a six-membered carbon ring. Plots (*a*), (*b*) and (*c*) are for nearest-neighbour C–C distances of 1·5, 1·4 and 1·3 Å, respectively. The shaded bands indicate zones of s^2 where these curves deviate least from the median line.

For all test structures except ANTH1 the inflexion-point least-squares procedure gave B estimates closer to the refined values than those obtained by conventional Wilson-plot programs (see Table 2). For ANTH1, the B estimate of 4·2 Å^2 was still acceptably close. No attempt was made in these tests to optimize the values of s^2 used to locate the inflexion points, though this would have further improved the B estimates. The approximation of inflexion points at fixed values of s^2, in fact, emphasizes the robustness of this approach and the improvements that can result. Estimates of B from *GENEV* (see Table 2, column 2) agree on the average to within 5% of the refined values; the B values from programs employing conventional Wilson-plot methods have average differences of about 16%. These averages exclude the B values for the three proteins, where inflexion-point least squares could not be applied because the s^2_{max} is below 0·15.

Table 2 also lists the refined and estimated structure-factor scales for each of the test structures. The **k** values were obtained by rescaling the $|E_h|$ values calculated from the estimated B value so that the mean $|E_h^2| = 1{\cdot}0$. The refined **k** values are from the final structure-factor least-squares cycle. The importance of the correct B values to the estimate of **k** is evident from Table 2.

Conclusions

The deleterious effect of incorrect B and **k** values on the normalized structure factors and hence on the structure-invariant phase relationships is demonstrated. A number of approaches to improve the Wilson-plot estimate of the overall temperature factor B has been examined. The use of Bayesian statistics is suggested in the reduction of intensity data, or, if this is not possible, the application of a more limited Bayesian treatment during the Wilson-plot calculation is recommended. A simple method to compensate for missing weak reflections is proposed.

The use of least-squares weights in the Wilson plot was considered and found unreliable because of the high correlation between $|F_h^2|$ and $\sigma|F_h^2|$, and the systematic effects of Debye scattering.

The dominant influence of Debye scattering effects on the Wilson-plot process was studied separately. The application of the Debye expectation value to R [see

(5)] was found to provide only marginal, if any, improvement in the B estimate, and to have a number of drawbacks. The concept of Debye-curve inflexion points is introduced and a straightforward and relatively robust method for improving the least-squares process, based on predictable features of a Debye curve, is described. Values of B estimated by the inflexion-point method are, on average, 10% better than those calculated by conventional methods.

The authors wish to acknowledge the assistance of the Australian Research Grants Committee (Grant: C7915302) during the tenure of this work.

References

DAVIS, C. L., MASLEN, E. N. & VARGHESE, J. N. (1978). *Acta Cryst.* A**34**, 371–377.

DEBYE, P. (1915). *Ann. Phys. (Leipzig)*, **46**, 809.

DECLERCQ, J. P., GERMAIN, G. & VAN MEERSSCHE, M. (1972). *Cryst. Struct. Commun.* **1**, 13–15.

FRENCH, S. & WILSON, K. (1978). *Acta Cryst.* A**34**, 517–525.

HALL, S. R. (1972). XRAY72 system of crystallographic programs : program *NORMSF* edited by J. M. STEWART *et al.* Tech. Rep. TR192, Computer Science Center, Univ. of Maryland, College Park, Maryland.

HALL, S. R. (1978). *Acta Cryst.* A**34**, S348.

HALL, S. R. (1981). *GENEV:* Generation of E values. *The XTAL system of crystallographic programs.* Tech. Rep. TR873. Computer Science Center, Univ. of Maryland, College Park, Maryland.

HALL, S. R., RASTON, C. L. & WHITE, A. H. (1978). *Aust. J. Chem.* **31**, 685–688.

HALL, S. R., STEWART, M. J. & MUNN, R. J. (1980). *Acta Cryst.* A**36**, 979–989.

HENDRICKSON, W. A. (1980). Personal communication.

ISAACS, N. W. & AGARWAL, R. C. (1978). *Acta Cryst.* A**34**, 782–791.

LADD, M. F. C. (1978). *Z. Kristallogr.* **147**, 279–296.

MAIN, P. (1976). *Crystallographic Computing Techniques*, edited by F. R. AHMED, pp. 97–105. Copenhagen; Munksgaard.

MAIN, P., FISKE, S. J., HULL, S. E., LESSINGER, L., GERMAIN, G., DECLERCQ, J. P. & WOOLFSON, M. M. (1980). *MULTAN*80. Univ. of York.

SKELTON, B. W. & WHITE, A. H. (1981). Personal communication.

SUBRAMANIAN, V. & HALL, S. R. (1982). *Acta Cryst.* A**38**, 577–590.

TEETER, M. M. & HENDRICKSON, W. A. (1979). *J. Mol. Biol.* **127**, 219–223.

WATENPAUGH, K. D., SIEKER, L. C., HERRIOTT, J. R. & JENSEN, L. H. (1973). *Acta Cryst.* B**29**, 943–956.

WILSON, A. J. C. (1942). *Nature (London)*, **150**, 151.

Acta Cryst. (1982). A**38**, 598–608

Normalized Structure Factors. III. Estimation of Errors

BY S. R. HALL AND V. SUBRAMANIAN†

Crystallography Centre, University of Western Australia, Nedlands 6009, *Australia*

(*Received* 6 *May* 1981; *accepted* 20 *January* 1982)

Abstract

A method for calculating the expected errors in $|E_h|$ values is outlined. It is based on the precision of the measured data and the Wilson-plot parameters; and allows for errors arising from the use of the profile scaling function and/or the index rescaling procedure in the normalization scheme. Six refined structures are used to test the estimated errors in $|E_h|$ against values deduced from a comparison with the 'true' normalized structure factor $|\mathscr{E}_h|$.

Introduction

One of the most serious obstacles to structure solution by statistical invariant methods is the sensitivity of all phasing procedures to errors in the initial phase relationships. The generation of a single incorrect phase in the early stages of a phasing procedure can often result in the failure of the entire process. For this reason computer programs place a strong emphasis on the choice of initial starting phases and on the order in which the invariants are processed.

There are a number of different approaches to the selection of starting phases but all of them depend on one fundamental quantity, namely, the magnitude of

† Deceased 27 December 1981.

0567-7394/82/050598-11$01.00