Acta Cryst. (1996). D**52**, 257–266

Statistical Expectation Value of the Debye–Waller Factor and $E(hkl)$ Values for Macromolecular Crystals

ROBERT H. BLESSING, D. Y. GUO AND DAVID A. LANGS

Hauptman-Woodward Research Institute (formerly the Medical Foundation of Buffalo), 73 High Street, Buffalo, New York 14203, USA

(*Received 22 May 1995; accepted 10 October 1995*)

Abstract

If the unit-cell distribution of atomic mean-square displacement parameters $B = 8\pi^2\langle u^2\rangle$ is assumed to be normal, with mean $\mu = \langle B\rangle$ and variance $\sigma^2 = \langle(B - \langle B\rangle)^2\rangle$, the statistical expectation value of the Debye–Waller factor $W^2 = \exp(-2Bs^2)$, where $s = (\sin\theta)/\lambda$, is $\langle W^2\rangle = \exp[-2(\mu - \sigma^2 s^2)s^2]$. This result has been incorporated into procedures for scaling and normalizing measured Bragg intensities to their Wilson expectation values. The procedures can determine both isotropic $\mu(B)$ and $\sigma(B)$ and anisotropic $\mu(U^{ij})$ and $\sigma(U^{ij})$ distribution parameters. Tests with experimental data and refined structural models for several protein crystals show that the procedures yield reliable normalized structure-factor amplitudes for direct-methods applications, with values of $R = \sum_h \big||E_o| - |E_c|\big| / \sum_h |E_o|$ averaging ~5%.

1. Introduction

In general, the different atoms of a crystal structure have different values for their mean-square displacement parameters,

$$B = 8\pi^2\langle u^2\rangle. \tag{1}$$

For example, B values due to thermal vibration are usually larger for atoms at the periphery of a molecule than for atoms near the molecular center of mass, or larger for side-chain atoms than for main-chain atoms. We have found that, at room temperature, root-mean-square deviations from mean B values are typically about 25% in organic molecular crystals and 50% or more in crystals of biological macromolecules. Atomic B values tend to be larger and more broadly distributed in biomolecular crystals because, in general, the crystals are highly hydrated, the molecules are loosely packed, and atomic displacements due to disorder are commonly as large as or larger than the displacements due to thermal vibration.

2. Expectation value for the function $W^2 = \exp[-2B(\sin\theta)^2/\lambda^2]$

We assume that the distribution of atomic B values over the crystal chemical unit can be fairly approximated by a normal distribution,

$$p(B) = [1/\sigma(2\pi)^{1/2}]\exp -[(B - \mu)^2/(2\sigma^2)], \tag{2}$$

with mean,

$$\mu = \langle B\rangle, \tag{3}$$

and variance,

$$\sigma^2 = \langle(B - \langle B\rangle)^2\rangle. \tag{4}$$

Then, the statistical expectation value for the Debye–Waller factor,

$$W^2 = \exp(-2Bs^2), \tag{5}$$

where $s = (\sin\theta)/\lambda$, can be obtained according to,

$$\langle f(x)\rangle = \int_{-\infty}^{+\infty} p(x) f(x)\, dx,$$

$$\langle\exp(-2Bs^2)\rangle = 1/[\sigma(2\pi)^{1/2}] \int_{-\infty}^{+\infty} \exp -[(B - \mu)^2/(2\sigma^2)] \times \exp(-2Bs^2)\, dB. \tag{6}$$

As shown in *Appendix A*, completing the square in the integrand simplifies (6) to,

$$\langle\exp(-2Bs^2)\rangle = \exp[-2(\mu - \sigma^2 s^2)s^2]. \tag{7}$$

Thus, due to the spread of the distribution of atomic B values, the expectation value for the Debye–Waller factor corresponds to an effective overall B value,

$$B_{\text{eff}} = \langle B\rangle - \langle(B - \langle B\rangle)^2\rangle s^2, \tag{8}$$

that is smaller than the mean B, and that decreases with increasing $(\sin\theta)/\lambda$.

3. Application to Wilson scaling

Wilson (1949) showed that for structures of N atoms per unit cell uniformly and randomly distributed in space group $P1$ or $P\bar{1}$ the squared structure-factor amplitudes have statistical expectation values,

$$\langle|F(\mathbf{h})|^2\rangle = \sum_{a=1}^{N} [f_a(s)\, W_a(s)]^2, \tag{9}$$

where $|\mathbf{h}| = 2s$, with expected variances,

$$\langle(|F|^2 - \langle|F|^2\rangle)^2\rangle = \begin{cases} \langle|F|^2\rangle & \text{in } P1, \text{ or} \\ 2\langle|F|^2\rangle & \text{in } P\bar{1}. \end{cases} \tag{10}$$

In higher symmetry space groups the expectation values become,

$$\langle|F(\mathbf{h})|^2\rangle = \varepsilon(\mathbf{h}) \sum_a [f_a(s)\, W_a(s)]^2, \tag{11}$$

where $\varepsilon(\mathbf{h}) \geq 1$ is the degeneracy of the reciprocal lattice point $\mathbf{h}$. The degeneracy factors are given by,

$$\varepsilon(\mathbf{h}) = m_L\, \varepsilon'(\mathbf{h}), \tag{12}$$

where m_L is the lattice multiplicity and ε' is a point-group dependent projection symmetry multiplier. The factor $m_L = 1$, 2, 4, or 3 allows for the systematic extinction of a fraction $(m_L - 1)/m_L$ of the reflections due to lattice centering so that the total scattering is concentrated in the allowed fraction $1/m_L$ of the reflections; the factor $\varepsilon' = 1$, 2, 4, 8, 3, 6, or 12 allows for the multiple enhancement of the average intensities for certain classes of zonal or axial reflections due to superposition of symmetrically equivalent atoms in projection onto mirror planes or rotation axes (Rogers, 1965, 1980; Iwasaki & Ito, 1977).

When the Wilson expectation values of $|F|^2$ are used to estimate a measurement scaling factor $k = |F|(\text{absolute})/|F|(\text{relative})$, it is usually assumed that a single isotropic $\langle B\rangle$ value, the same for all atoms, is an adequate approximation. Then (11) gives,

$$|F(\mathbf{h})|^2_{\text{meas}} \Big/ \left[\varepsilon(\mathbf{h}) \sum_a f_a^2(s)\right] \simeq k^{-2} \exp(-2\langle B\rangle s^2). \tag{13}$$

The statistical expectation value $\langle\exp(-2Bs^2)\rangle$ given by (7) should, however, be a better approximation. It gives,

$$|F(\mathbf{h})|^2_{\text{meas}} \Big/ \left[\varepsilon(\mathbf{h}) \sum_a f_a^2(s)\right] \simeq k^{-2} \exp[-2(\mu - \sigma^2 s^2)s^2], \tag{14}$$

and k, μ, and σ can be estimated from a logarithmically linearized least-squares fit,

$$\ln\left\{|F(\mathbf{h})|^2_{\text{meas}} \Big/ \left[\varepsilon(\mathbf{h}) \sum_a f_a^2(s)\right]\right\} \simeq -2\ln k - 2\mu s^2 + 2\sigma^2 s^4, \tag{15}$$

analogous to the commonly used Wilson plot (Wilson, 1942). Equation (15) is quadratic in s^2 and, since necessarily $\sigma > 0$, the curvature must be concave. Thus (15) will give a smaller k and larger μ than would a conventional Wilson plot which, since it tacitly assumes $\sigma = 0$, is linear in s^2.

In practice, we initially assume $\sigma = 0$ and roughly approximate k and μ using (15). Then, by non-linear least-squares iterations based on (14), we first refine k and μ with $\sigma = 0$, and then fit k, μ, and $\sigma > 0$. In order to escape the $\sigma = 0$ least-squares minimum, the iterations to fit $\sigma > 0$ are started from $\mu = 2\mu'$ and $\sigma = \mu'/2$, where the primed value of μ is that fitted with $\sigma = 0$. After fitting for isotropic B values, we fit for overall anisotropic b^{ij} values by iterations based on,

$$|F(\mathbf{h})|^2_{\text{meas}} \Big/ \left[\varepsilon(\mathbf{h}) \sum_a f_a^2(s)\right] \simeq k^{-2} \exp\{-2[\mathbf{h}^T \boldsymbol{\mu} \mathbf{h} - (\mathbf{h}^T \boldsymbol{\sigma} \mathbf{h})^2]\}, \tag{16}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are symmetric matrices which must be positive definite. The starting values for the anisotropic tensor parameters μ^{ij} and σ^{ij} $(i \leq j = 1, 2, 3)$ are obtained by expansions of the isotropic scalar parameters μ and σ analogous to the expansion (Johnson & Levy, 1974) of B_{iso} to b^{ij} values,

$$b^{ij} = 2\pi^2 a^{*i} a^{*j} U^{ij} = \tfrac{1}{4} a^{*i} a^{*j} \cos\alpha^{*ij} B_{\text{iso}}. \tag{17}$$

Details of the least-squares fitting based on (14) through (17) are given in *Appendix B*. We have found that even with $\sigma = 0$ the iterative least-squares procedure is better than the widely used Wilson-plot (Wilson, 1942) or K-curve (Karle & Hauptman, 1953) procedures or methods based on analysis of the Patterson origin peak (Rogers, 1965; Nielsen, 1975; Blessing & Langs, 1988). The iterative fit gives a direct and reliable evaluation of the overall anisotropy of mean-square atomic displacements (Levy, Thiessen & Brown, 1970) and, since the iterative fit does not require that the intensities be averaged in shells of scattering angle or summed to construct the Patterson origin peak, the many weak high-resolution data help offset bias from non-random distribution statistics in the few strong low-resolution data. Most importantly, as shown below, the scaling procedures based on (14) through (17) lead to meaningful experimental normalized structure-factor amplitudes for direct-methods applications with protein crystals.

4. Normalized structure-factor amplitudes

Experimental normalized structure-factor amplitudes are calculated from the fitted scale factor and the expectation

Table 1. *Crystal data for test structures*

Position and mean-square displacement parameters for H atoms were not available so they were omitted from all calculations except the experimental $|F_o|$ normalization denominators in (18).

	2-Zinc pig insulin	Rubredoxin	Crambin ($T \simeq 150$ K)
Space group	$R3$	$P2_1$	$P2_1$
Unit-cell dimensions (Å,°)	$a = b = 82.5, c = 34.0$	$a = 19.97, b = 41.45,$ $c = 24.41\ \beta = 108.39$	$a = 40.763, b = 18.492,$ $c = 22.333\ \beta = 90.61$
V (Å^3)	200409	19185	16833
Z	9	2	2
Formula	$C_{512}H_{821}N_{129}O_{152}S_{12}Zn_{2/3}.281H_2O$	$C_{243}H_{300}N_{57}O_{85}S_6Fe.102H_2O$	$C_{202}H_{313}N_{55}O_{64}S_6.C_2H_5OH.84H_2O$
M_r	16713.89	7465.23	6287.82
ρ_{calc}	1.246	1.292	1.241
ρ_{meas}	1.245 (by flotation in toluene/bromobenzne)		
d (Å)	$\infty < d \leq 1.5$	$12.8 \leq d \leq 1.0$	≤ 0.83
No. of data	13414 (98% complete)	18529 (91% complete)	28727 (95% complete)
Reference	(*a*), (*b*)	(*c*)	(*d*), (*e*)

References: (*a*) Baker *et al.* (1988). (*b*) Blundell & Johnson (1976) show a conventional Wilson plot. (*c*) Dauter, Sieker & Wilson (1992). (*d*) Hope (1988). (*e*) Teeter, Roe & Heo (1993).

value for the Debye–Waller factor as,

$$|E_o(\mathbf{h})| = k|F_o(\mathbf{h})| \Big/ \left\{ \varepsilon(\mathbf{h}) \left[\sum_a f_a^2(\mathbf{h}) \right] \times \exp[-2\mathbf{h}^T \boldsymbol{\mu} \mathbf{h} + 2(\mathbf{h}^T \boldsymbol{\sigma} \mathbf{h})^2] \right\}^{1/2} \tag{18}$$

These are to be compared with the amplitudes of model-calculated crystal structure factors normalized to their Wilson expectation values, for which four cases need be considered. In the general case of [1] unequal atoms with unequal mean-square atomic displacements,

$$E_c(\mathbf{h}) = F_c(\mathbf{h}) \Big/ \left[\varepsilon(\mathbf{h}) \sum_a p_a^2 f_a^2(\mathbf{h}) \exp(-2\mathbf{h}^T \mathbf{b}_a \mathbf{h}) \right]^{1/2}, \tag{19}$$

where,

$$F_c(\mathbf{h}) = \sum_{a=1}^{N} p_a f_a(\mathbf{h}) \exp(2\pi i\, \mathbf{h}^T \mathbf{r}_a - \mathbf{h}^T \mathbf{b}_a \mathbf{h}), \tag{20}$$

and $0 < p_a \leq 1$ allows for partial atomic site occupation in disordered structures. In the approximation of [2] atoms at rest (or with equal mean-square displacements),

$$E_c^0(\mathbf{h}) = \left[\sum_a p_a f_a(\mathbf{h}) \exp(2\pi i\, \mathbf{h}^T \mathbf{r}_a) \right] \Big/ \left[\varepsilon(\mathbf{h}) \sum_a p_a^2 f_a^2(\mathbf{h}) \right]^{1/2}. \tag{21}$$

In the further approximation of [3] point-atoms at rest,

$$E_c^z(\mathbf{h}) = \left[\sum_a p_a Z_a \exp(2\pi i\, \mathbf{h}^T \mathbf{r}_a) \right] \Big/ \left[\varepsilon(\mathbf{h}) \sum_a p_a^2 Z_a^2 \right]^{1/2} \tag{22}$$

where $Z_a = f_a(\mathbf{h})|_{\mathbf{h}=0}$ replaces $f_a(\mathbf{h})$ in (21). And in the still further approximation of [4] equal point-atoms at rest,

$$E_c^1(\mathbf{h}) = \left[\sum_a p_a \exp(2\pi i\, \mathbf{h}^T \mathbf{r}_a) \right] \Big/ \left[\varepsilon(\mathbf{h}) \sum_a p_a^2 \right]^{1/2}. \tag{23}$$

5. Test results with protein data

We have compared results from (18)–(23) using experimental diffraction data and refined structural models for a number of protein crystal structures, three of which – insulin and rubredoxin at room temperature and crambin at ~150 K – are listed in Table 1. These three were chosen as illustrations because, as their calculated mass densities indicate, they are among the few protein crystal structures for which an essentially complete structural model including solvent structure has been refined. Distribution statistics for the refined atomic B values of the non-H atoms are given in Table 2, which shows that the B distributions are skewed toward larger than average B values (for solvent atoms and peripheral protein atoms) and are more sharply peaked (for core protein atoms) about the average B value than normal distributions. The distribution abnormalities follow the sequence: insulin < rubredoxin < crambin.

Table 3 summarizes tests of the atoms-at-rest approximations (21)–(23) and the normalization procedure (14)–(18) applied to 'error-free' model-calculated data. The $E_c - E_c^0$ and $E_c - E_c^z$ columns of Table 3 (and similar $E_c - E_c^1$ results not reprinted in Table 3) show that the atoms-at-rest hypotheses introduce substantial errors in both amplitudes and phases of the high-resolution data. In contrast, the $|E_o| - |E_c|$ column, second from the right in Table 3, shows that taking the atoms-in-motion $|F_c|$ values from (20) to be synthetic $|F_{meas}|$ or $|F_o|$ data and applying the normalization procedure (14)–(18) yields synthetic $|E_o|$ that agree very well with the calculated $|E_c|$ from (19), with values of

Table 2. *Distribution statistics for the refined B values ($Å^2$) of the non-H atoms of the structures listed in Table 1*

	Mean $\langle B\rangle$	Standard deviation $\langle(B-\langle B\rangle)^2\rangle^{1/2}$	Moment coefficients of skewness kurtosis* c_3	c_4
Insulin	29.57	19.26	1.09	0.53
Rubredoxin	15.82	15.06	1.85	2.48
Crambin	6.45	6.72	3.85	18.60

* The moment coefficients of skewness and kurtosis are, respectively, $c_3 = \langle(B-\langle B\rangle)^3\rangle / \langle(B-\langle B\rangle)^2\rangle^{3/2}$ and $c_4 = \langle(B-\langle B\rangle)^4\rangle / \langle(B-\langle B\rangle)^2\rangle^2 - 3$. For normal distributions $c_3 = c_4 = 0$. Values of $c_3 > 0$ indicate positively skewed distributions with an abnormally large population or long tail with $B > \langle B\rangle$; and values of $c_4 > 0$ indicate distributions abnormally sharply peaked about $\langle B\rangle$.

$R = \sum_{\mathbf{h}} \big||E_o| - |E_c|\big| / \sum_{\mathbf{h}} |E_o|$ averaging ~5%. This constitutes an essential validation of the normalization procedure.

The scale factors and B distribution parameters from the normalization procedure applied to the experimental $|F_o|$ as well as to the model-calculated $|F_c|$ data sets are given in Table 4. The two sets of fitted parameters k, $\mu(B)$, and $\sigma(B)$ are highly consistent with one another, but comparison with the distribution statistics in Table 2 shows that the fitting procedure systematically underestimated both the means and standard deviations of the atomic B distributions. This is because the actual distributions are strongly skewed and sharply peaked, and cannot be closely approximated by normal distributions. Table 4 also shows that along with the underestimation of the mean-square displacement parameters $\mu(B)$ and $\sigma(B)$ there is a correlated overestimation of the scale factor k. Ideally we would expect $k = 1$ from $|F_c|$ data, but the $|F_c|$ fitted k values in Table 4 are some 15 to 30% too large. The overestimated k values were, however, largely compensated by the underestimated $\mu(B)$ and $\sigma(B)$ values, and the net effect of the systematic errors of fit was an only slight deterioration of the agreement between the synthetic $|E_o|$ and model $|E_c|$ at low resolution, as shown in the second column from the right in Table 3.

Application of the normalization procedure to the experimental $|F_o|$ data is summarized in Table 5, which shows that the agreement between the experimental $|E_o|$ and model $|E_c|$ is as good as the $|F_o|$ *versus* $|F_c|$ agreement. Predictably, the atoms-at-rest $|E_c^0|$ values do not agree as well as the atoms-in-motion $|E_c|$ values do with the experimental $|E_o|$ data, especially at high resolution. The low-resolution $R(|F|)$ values in Table 5 also show that there remain some small problems with the $|F_o|$ data and/or with the scale factor, mean-square atomic displacements, and/or solvent structure in the $|F_c|$ models of the crambin and rubredoxin crystals. We think that the uniformity of the $R(|F|)$ and $\langle|F_o|\rangle/\langle|F_c|\rangle$ statistics over resolution subsets for the 2-zinc pig insulin structure testifies to the great care the Dorothy Hodgkin and her co-workers devoted to determining the interstitial water structure in the crystals.

Table 3. *Agreement statistics in cumulative resolution subsets for calculated data for the (non-H) structures listed in Table 1*

E_c from (19) and (20) for unequal atoms with unequal B values, E_c^0 from (21) for unequal atoms at rest (or with equal B values), E_c^z from (22) for unequal point-atoms at rest, and *synthetic* $|E_o|$ from (14)–(18) applied to $|F_c|$ from (20). The statistics tabulated are the normalized mean absolute amplitude differences, $R = \sum_{\mathbf{h}} \big||E_1| - |E_2|\big| / \sum_{\mathbf{h}} 0.5(|E_1| + |E_2|)$, and the $|E|$-weighted mean absolute phase differences, $\langle|\Delta\varphi|\rangle = \sum_{\mathbf{h}} |E_1E_2|\,|\varphi_1 - \varphi_2| / \sum_{\mathbf{h}} |E_1E_2|$, compiled cumulatively for resolution subsets of the n reflections with $\infty > d > d_{\min}$.

$d_{\min}$ (Å)	n	$E_c - E_c^0$ R	$\langle\lvert\Delta\varphi\rvert\rangle$*	$E_c - E_c^z$ R	$\langle\lvert\Delta\varphi\rvert\rangle$*	$\lvert E_o\rvert - \lvert E_c\rvert$ R	$\lvert E_o\rvert - \lvert E_c^0\rvert$ R
Insulin							
1.5	13744	0.364	22.7	0.357	22.0	0.050	0.359
2	5836	0.290	16.2	0.280	15.3	0.027	0.290
2.5	2997	0.228	11.2	0.217	10.6	0.019	0.230
3	1737	0.171	7.8	0.161	7.4	0.023	0.176
3.5	1093	0.136	6.3	0.125	5.8	0.031	0.146
4	738	0.122	4.9	0.108	4.6	0.039	0.139
5	379	0.098	3.9	0.088	3.7	0.056	0.132
6	216	0.083	2.4	0.077	2.3	0.069	0.130
8	93	0.067	1.4	0.063	1.4	0.079	0.132
10	47	0.062	0.9	0.058	0.8	0.086	0.134
Rubredoxin							
1	20437	0.347	20.0	0.356	20.8	0.049	0.351
1.5	6122	0.298	15.0	0.293	14.7	0.048	0.298
2	2588	0.235	10.7	0.223	10.1	0.065	0.244
2.5	1336	0.174	7.5	0.165	7.1	0.084	0.196
3	786	0.132	5.4	0.124	5.0	0.100	0.171
3.5	499	0.104	4.2	0.093	4.0	0.113	0.170
4	335	0.094	3.5	0.080	3.2	0.125	0.176
5	171	0.089	2.5	0.076	2.2	0.144	0.191
6	103	0.068	1.8	0.060	1.7	0.153	0.186
8	44	0.050	1.1	0.046	1.0	0.163	0.195
10	23	0.032	0.9	0.028	0.8	0.167	0.188
Crambin							
0.83	30258	0.242	11.6	0.259	12.7	0.016	0.243
1	18353	0.212	9.7	0.219	10.1	0.018	0.214
1.5	5542	0.154	6.4	0.153	6.3	0.021	0.158
2	2390	0.114	4.3	0.107	4.0	0.036	0.124
2.5	1237	0.083	2.9	0.077	2.7	0.054	0.101
3	733	0.059	2.1	0.056	2.0	0.064	0.090
3.5	472	0.044	1.4	0.042	1.4	0.072	0.090
4	322	0.036	1.1	0.034	1.1	0.078	0.094
5	168	0.028	0.8	0.025	0.7	0.086	0.102
6	98	0.022	0.4	0.019	0.4	0.091	0.104
8	45	0.016	0.1	0.012	0.2	0.094	0.104
10	22	0.013	0.05	0.010	0.05	0.096	0.107

6. Conclusions

Notwithstanding its shortcomings for describing strongly skewed and sharply peaked B distributions, the normalization procedure based on (14)–(18) can produce entirely reliable experimental $|E(hkl)|$ amplitudes for protein crystals if data extending to ~2.5 Å resolution or better are measured (see *Appendix C*). Common notions that the nature of protein crystals or difficulties inherent in measuring their Bragg diffraction data present insuperable obstacles to reliable normalization are

Table 4. *B distribution parameters fitted by (14)–(16) to experimental* $|F_o|^2$ *and model-calculated* $|F_c|^2$ *data for the structures listed in Table 1*

	Fitted to $\|F_o\|^2$			Fitted to $\|F_c\|^2$		
	k	$\mu(B)$ (Å^2)	$\sigma(B)$ (Å^2)	k	$\mu(B)$ (Å^2)	$\sigma(B)$ (Å^2)
Insulin	3.865 (15)	15.30 (5)	0	1.327 (6)	15.63 (6)	0
	3.441 (22)	20.40 (23)	6.45 (14)	1.145 (8)	22.0 (3)	7.18 (15)
Rubredoxin	0.02456 (12)	6.37 (3)	0	1.435 (6)	5.86 (3)	0
	0.01998 (16)	10.66 (13)	4.05 (6)	1.213 (10)	9.12 (12)	3.43 (6)
Crambin	0.4968 (19)	3.69 (2)	0	1.192 (5)	3.70 (2)	0
	0.539 (4)	3.34 (8)	0.04 (4)	1.321 (8)	3.23 (7)	0.01 (3)

	i,j	1,1	2,2	3,3	1,2	1,3	2,3
	k	3.434 (22)					
Insulin	$\mu(U^{ij})$ (Å^2)	0.251 (5)	0.251 (5)	0.226 (6)	0.126 (4)	0	0
	$\sigma(U^{ij})$ (Å^2)	0.072 (5)	0.072 (5)	0.055 (6)	0.036 (4)	0	0
	k	0.02012 (15)					
Rubredoxin	$\mu(U^{ij})$ (Å^2)	0.152 (2)	0.109 (3)	0.124 (3)	0	0.046 (2)	0
	$\sigma(U^{ij})$ (Å^2)	0.050 (1)	0.048 (2)	0.037 (2)	0	0.015 (1)	0
	k	0.4895 (18)					
Crambin	$\mu(U^{ij})$ (Å^2)	0.0410 (3)	0.0551 (5)	0.0461 (4)	0	0.0066 (2)	0
	$\sigma(U^{ij})$ (Å^2)	0	0	0	0	0	0

Notes: For insulin, Blundell & Johnson (1976) report $\mu(B) = 13.4$ Å^2 from a conventional Wilson plot. For crambin, with both the observed and the calculated data, the attempt to fit $\sigma(B) > 0$ failed to improve on the statistics of fit obtained with $\sigma(B) = 0$. Atypically and unrealistically, the fits with $\sigma(B) > 0$ produced increases in k and decreases in $\mu(B)$. These effects are presumably due to the very sharply peaked distribution of atomic B values (Table 2) in crambin crystals at ~150 K. For all three structures, the anisotropic fit gave no improvement with the calculated $|F_c|^2$ data because they were based on isotropic refinement models. The experimental $|F_o|^2$ data, however, manifested significant anisotropy when fitted by (16). The $\sigma(U^{ij})$ values are not uncertainties in the U^{ij} values, which are given as e.s.d.'s of fit in parentheses; the $\sigma(U^{ij})$ values are breadth parameters of the anisotropic unit-cell distributions of the anisotropic U^{ij} values. The overall anisotropies expressed as $\max(U^{ii})/\min(U^{ii})$ are 1.11, 1.39, and 1.34 for the insulin, rubredoxin, and crambin structures, respectively. The overall U^{ij} values could be used to improve the structure refinements (Sheriff & Hendrickson, 1987).

unduly pessimistic. There are, however, several noteworthy characteristics that distinguish protein data sets qualitatively from small-molecule data sets.

First, the large average values of the mean-square atomic displacements in protein crystals, and the broad distributions about the average values, render useless the hypothesis of atoms at rest – or even atoms with equal mean-square displacements – and the approximations (21)–(23) for calculating 'error-free' $E(hkl)$ data for protein crystals.

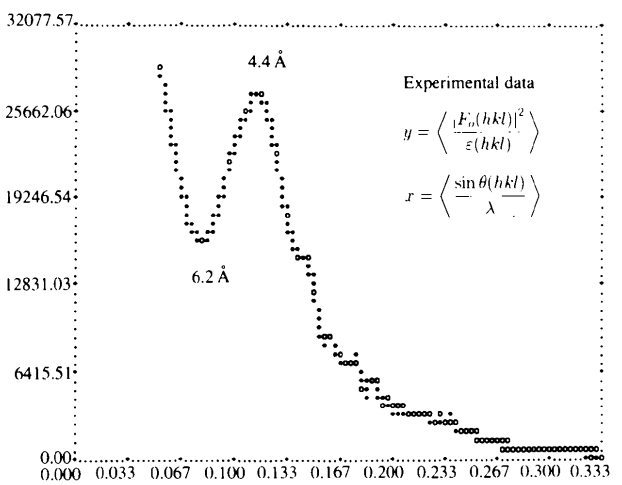


Fig. 1. Multiplicity-weighted local averages $\langle |F_{\mathbf{h}}|^2_{\text{meas}}/\varepsilon_{\mathbf{h}} \rangle$ *versus* $\langle (\sin\theta_{\mathbf{h}})/\lambda \rangle$ for the experimental 2-zinc pig insulin data of Baker *et al.* (1988). Symbols are o for the local averages of 135 data each, and * for cubic-spline interpolated values. The local minimum at $d = \lambda/(2\sin\theta) = 6.2$ Å and maximum at 4.4 Å are typical; most protein data sets exhibit a similar pair of local extrema due to ubiquitous non-random structural motifs (see Table 6 and text).

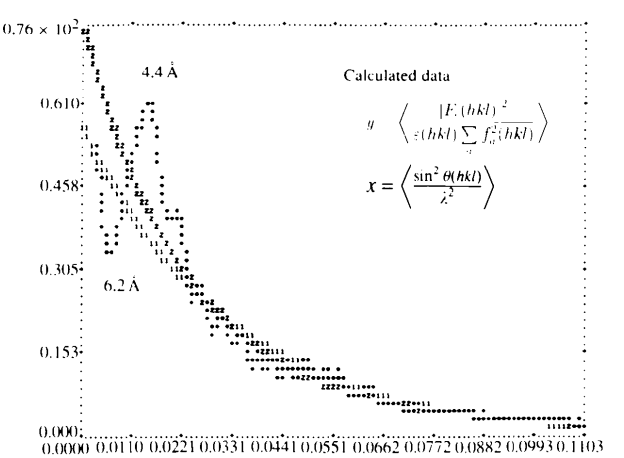


Fig. 2. Multiplicity-weighted local averages $\langle |F_c|^2/(\varepsilon \sum_a f_a^2) \rangle$ *versus* $\langle (\sin\theta)^2/\lambda^2 \rangle$ from calculated data for the 2-zinc insulin structure (Baker *et al.*, 1988). Symbols are * for the cubic-spline interpolated data curve, 1 for equation (14) fitted with $\sigma(B) = 0$, and 2 for (14) fitted with $\sigma(B) > 0$. The fitted parameters are given in Table 4. The local minimum and maximum in the calculated data occur at 6.2 and 4.4 Å resolution, respectively, as in the experimental data shown in Fig. 1.

Table 5. *Agreement statistics in cumulative resolution subsets for observed versus calculated $|E|$'s for the structures listed in Table 1*

$|E_o|$ from the experimental $|F_o|$ *via* (14)–(18) *versus* $|E_c|$ from (19) and $|E_c^0|$ from (21). H atoms were omitted from all calculations except the experimental $|F_o|$ normalization denominators in (18). The statistics tabulated are the normalized mean absolute amplitude differences, $R_F = \sum_h ||F_o| - k|F_c||/\sum_h |F_o|$, and $R_E = \sum_h ||E_o| - |E_c||/\sum_h |E_o|$, compiled cumulatively for resolution subsets of the n reflections with $\infty > d > d_{min}$. The scaling factor in R_F, $k = \sum_h |F_o|/\sum_h |F_c|$, was computed for the full data set and not re-computed for the lower resolution subsets. The values were $k = 1.214$, 1.135 and 1.209 for the insulin, rubredoxin and crambin data sets, respectively.

d_{min}	n	$\langle\|F_o\|\rangle/\langle k\|F_c\|\rangle$	R_F	$\langle\|E_o\|^2\rangle$	$\langle\|E_c\|^2\rangle$	R_E	$\langle\|E_c^0\|^2\rangle$	R_E
Insulin								
1.5	13414	1.00	0.164	1.00	0.92	0.202	0.95	0.356
2	5591	0.98	0.132	0.98	0.98	0.140	0.99	0.304
2.5	2936	0.97	0.123	0.95	0.96	0.124	0.95	0.270
3	1737	0.96	0.118	1.01	1.02	0.115	0.96	0.227
3.5	1093	0.96	0.110	1.04	1.02	0.102	0.91	0.192
4	738	0.97	0.113	0.99	0.94	0.105	0.82	0.186
5	379	0.96	0.140	0.73	0.67	0.128	0.59	0.189
6	216	0.96	0.157	0.66	0.59	0.145	0.52	0.196
8	93	0.97	0.175	0.86	0.74	0.165	0.66	0.204
10	47	0.98	0.213	0.91	0.77	0.202	0.70	0.236
Rubredoxin								
1	18529	1.00	0.161	1.00	1.02	0.172	1.02	0.354
1.5	6111	1.02	0.125	0.97	0.91	0.126	0.93	0.314
2	2577	1.02	0.118	0.99	0.92	0.116	0.94	0.262
2.5	1326	1.00	0.112	0.98	0.94	0.108	0.94	0.216
3	776	0.98	0.110	1.04	1.00	0.103	0.99	0.189
3.5	489	0.98	0.120	1.04	1.00	0.110	0.93	0.180
4	325	0.96	0.141	0.89	0.87	0.129	0.79	0.198
5	161	0.93	0.219	0.55	0.56	0.197	0.52	0.256
6	93	0.87	0.269	0.55	0.62	0.236	0.58	0.278
8	43	0.84	0.578	0.64	0.77	0.340	0.71	0.366
10	13	0.68	0.696	0.51	0.93	0.625	0.90	0.621
Crambin								
0.83	28727	1.00	0.156	1.00	1.01	0.171	1.01	0.261
1	18277	1.00	0.135	1.05	1.04	0.137	1.03	0.232
1.5	5542	1.00	0.115	0.95	0.91	0.115	0.92	0.188
2	2390	1.00	0.112	1.08	0.98	0.115	0.99	0.159
2.5	1237	0.98	0.114	1.10	1.00	0.115	0.99	0.140
3	733	0.97	0.117	1.18	1.07	0.114	1.07	0.127
3.5	472	0.94	0.126	1.09	1.02	0.115	1.00	0.126
4	322	0.92	0.146	0.95	0.92	0.127	0.89	0.141
5	168	0.88	0.202	0.68	0.72	0.165	0.70	0.175
6	98	0.84	0.248	0.68	0.78	0.202	0.76	0.207
8	45	0.78	0.353	0.82	1.06	0.288	1.03	0.285
10	22	0.69	0.556	0.84	1.35	0.445	1.32	0.438

Second, as illustrated in Figs. 1 and 2, protein crystals exhibit characteristically structured distributions of local average intensity against scattering angle, with typically a local minimum $\langle|F|^2\rangle$ at $d \simeq 6$ Å resolution and a local maximum at $d \simeq 4$ Å. Others (*e.g.*, French & Wilson, 1978) have noted these features before; we think they are due to the ubiquitous protein structural motifs listed in Table 6. The twisting course of a protein's main-chain 1,2 $C^\alpha - C^\alpha$ repeat units places many side-chain atoms near crystal planes with ~4 Å spacings in many directions, and these densely populated families of planes have higher than average Bragg reflectivity. Similarly, the 1,3 $C^\alpha - C^\alpha$ repeat units place many atoms near planes with ~6 Å spacings, but these densely populated families of planes interleave one another, reflect beams that interfere destructively, and have lower than average reflectivity.

Third, additional local minima and maxima are sometimes discernible in very low resolution data from protein crystals with unit cells larger than those of our insulin, rubredoxin or crambin examples. These average intensity oscillations are attributable to the non-uniformity of the average electron-density distributions over the unit cells of protein crystals due to their being partitioned into more-or-less well defined protein and solvent regions. In liquid water with mass density $1.00\,mg\,mm^{-3}$ the volume per water molecule is $v(H_2O) = 29.92\,Å^3$, and the average electron density is,

$$\langle\rho_s\rangle = 10\,e/29.92\,Å^3 = 0.334\,e\,Å^{-3}.$$

In a protein crystal, the average electron density in the protein regions can be approximated as,

Table 6. *Fundamental repeat distances in protein crystals from standard average bond lengths, valence angles, conformation angles, and water···water hydrogen bonds*

Repeat unit	Repeat distance (Å, °)
$C(\alpha_i)—C(\alpha_{i+1})$	3.82
$C(\alpha_{i-1})—C(\alpha_{i+1})$	5.42 in α-helices
	6.92 in β-sheets
$H_2O\cdots H_2O$	2.75 O—O in ice
H_2O	109.5 O—O—O
/ \	
$H_2O\cdots H_2O$	4.49 O···O

$$\langle \rho_p \rangle = \sum_{p=1}^{N_p} Z_p / [V_{\text{cell}} - N_s\, v(\text{H}_2\text{O})],$$

where N_p is the number of protein atoms (H atoms included), and N_s is the number of water molecules, per unit cell. Assuming the molecular volume in liquid water to be fair approximation in the solvent regions, we get,

$$\langle \rho_p \rangle = 0.450, 0.453 \text{ and } 0.431 \text{ e Å}^{-3},$$

for our insulin, rubredoxin, and crambin examples, respectively, and we take,

$$\langle \rho_p \rangle / \langle \rho_s \rangle \simeq 4/3,$$

as a rule-of thumb.

Finally, the solute–solvent partitioning in protein crystals profoundly affects the scattering-angle distribution of their normalized structure factors. This is shown in Table 7 and Fig. 3, which summarize calculations for the protein non-H atoms and for the water O atoms in insulin crystals. Three sets of E's, one for the complete structure and one for each of the two substructures, were calculated separately, with each set normalized for its own chemical composition. The protein and solvent substructures each have $\langle |E|^2 \rangle > 1$ at low resolution, with remarkably large $|E|^2$ values for some very low-angle reflections, but the complete protein-plus-solvent structure has $\langle |E|^2 \rangle < 1$ for its low-resolution subsets. This implies that the protein scattering is mainly out of phase with the solvent scattering, which, due to the large mean-square displacements of the water molecules, is mainly confined to the low-angle reflections. The effect on low-resolution structure factors is so large that calculations of 'error-free' data for protein crystals are utterly unrealistic below ~6 Å resolution if the solvent substructure is not included.

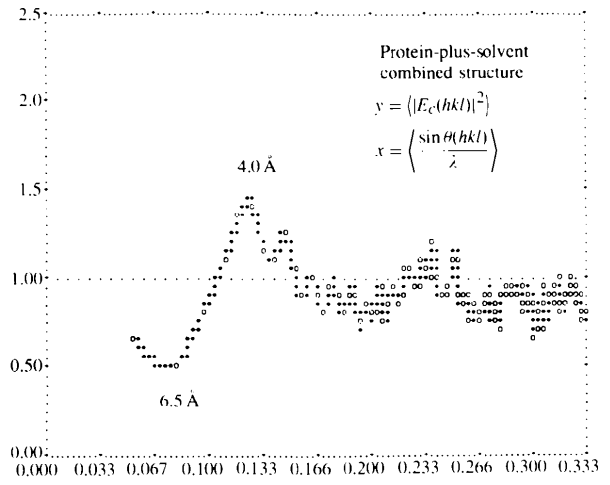
An anti-phase relationship between beams scattered by the protein and solvent substructures is to be expected since, for the most part, the solvent molecules occupy the interstitial space midway between neighboring protein molecules, and beams scattered by the two interleaved substructures interfere destructively. The $\langle |E|^2 \rangle$ *versus* $\langle (\sin\theta)/\lambda \rangle$ plots in Fig. 3 show that the combined structure and the protein-only substructure each exhibit minima at ~6 Å and maxima at ~4 Å, presumably due to the 1,3 and 1,2 $C^\alpha—C^\alpha$ repeats (Table 6), respectively. The water-only substructure exhibits a clear minimum at ~5 Å, possibly due to interleaved 1,3 $H_2O\cdots H_2O$ repeats, but an anticipated maximum at ~3 Å, due to the 1,2 $H_2O\cdots H_2O$ repeats, is flattened on account of the large mean-square displacements of the water molecules. By equation (7), with $\mu = 53.7$ and $\sigma = 16.1$ Å from Table 7, the Debye–Waller factors for the water substructure at $s = 1/(2d)$ are $\langle W^2 \rangle = 0.36$ and 0.07 at $d = 5$ and 3 Å. respectively, showing that for $d < 3$ Å the water scattering is practically nullified by thermal vibration and disorder.

Copies of the FORTRAN source codes and ASCII files of users' instructions for our normalization programs *SORTAV*, *BAYES*, *LEVY* and *EVAL* are available on request from their author, RHB.

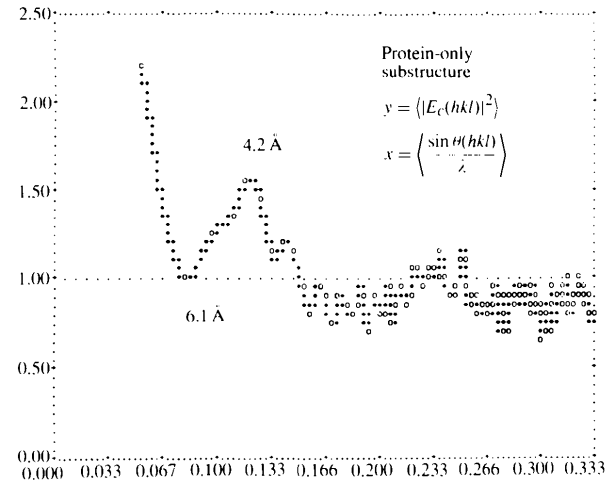
Table 7. *Substructure data on the non-H atomic B distributions and atoms-in-motion $|E_c|^2$ values from (19) for 2-zinc pig insulin (Baker et al., 1988)*

The three columns on the right give results for the protein-plus-solvent combined structure, the protein-only substructure, and the solvent-only substructure, respectively. The tabulated *hkl* are those for the first 15 reflections in order of increasing scattering angle.

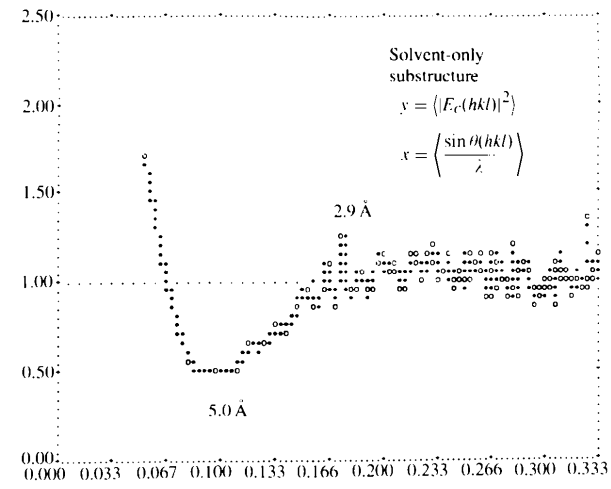
		Protein plus solvent	Protein only	Solvent only
ρ_m (mg mm^{-3})		1.143	0.807	0.336
$\langle B \rangle$		29.6	21.1	53.7
$\langle (B - \langle B \rangle)^2 \rangle^{1/2}$		19.3	12.0	16.1
c_3		1.09	1.8	0.41
c_4		0.43	4.2	−0.74
$d_{\min}$	n		$\langle \lvert E_c \rvert^2 \rangle$	
1.5	13414	0.92	0.92	1.00
2	5591	0.98	1.00	0.98
2.5	2936	0.96	1.05	0.90
3	1737	1.02	1.22	0.82
3.5	1093	1.02	1.38	0.75
4	738	0.94	1.46	0.77
5	379	0.67	1.51	0.95
6	216	0.59	1.71	1.30
8	93	0.74	2.65	2.13
10	47	0.77	3.93	3.52
h k l			$\lvert E_c(hkl) \rvert^2$	
−2 1 0		0.79	3.93	2.03
−1 1 1		0.74	18.24	25.08
−2 0 1		0.56	9.57	12.16
−3 0 0		0.77	34.56	56.39
−2 3 1		0.05	0.51	0.50
−3 2 1		0.56	0.86	1.25
−4 2 0		0.02	1.09	3.14
−4 1 1		0.44	0.54	1.71
−1 4 1		3.61	14.33	5.85
−1 0 2		0.04	0.13	0.04
−4 4 1		1.58	7.01	3.42
−5 4 0		1.40	5.37	2.31
−5 1 0		0.22	2.06	2.21
−2 2 2		0.99	2.92	0.78
−5 3 1		0.86	7.34	7.59



(*a*)



(*b*)



(*c*)

Fig. 3. Multiplicity-weighted local averages $\langle |E_c|^2 \rangle$ *versus* $\langle (\sin\theta)/\lambda \rangle$ for 2-zinc pig insulin corresponding to Table 7. Symbols are o for the local averages of 137 data each, and * for cubic-spline interpolated values. The local extrema are discussed in the text.

APPENDIX *A*. Simplification of equation (6)

The integrand in (6),

$$f(B) = \exp -[(B-\mu)^2/(2\sigma^2)]\exp(-2Bs^2),$$

is simplified by first expanding and collecting terms to obtain,

$$f(B) = \exp\{-1/(2\sigma^2)[B^2 - 2(\mu - 2\sigma^2 s^2)B + \mu^2]\},$$

then completing the binomial square by adding and subtracting the square of half the first-degree term in B. After some simplification we obtain,

$$f(B) = \exp\{-1/(2\sigma^2)[(B - \mu + 2\sigma^2 s^2)^2 + 4\sigma^2 s^2(\mu - \sigma^2 s^2)]\},$$

which we rewrite as a product of two exponentials,

$$f(B) = \exp[-2(\mu - \sigma^2 s^2)s^2] \times \exp -[(B - \mu + 2\sigma^2 s^2)^2/(2\sigma^2)].$$

The first exponential factors out of the integrand, and, since $\int_{-\infty}^{+\infty} \exp[-(x+a)^2]\mathrm{d}x = \int_{-\infty}^{+\infty} \exp(-u^2)\mathrm{d}u$, the second exponential integrates to the reciprocal normalization constant $\sigma(2\pi)^{1/2}$ so that (6) simplifies to (7).

APPENDIX *B*. Iterative least-squares fitting based on equations (14), (15) and (16)

The data fitted are the individual unique reflection data. It is neither necessary nor advantageous to average the data in shells of scattering angle, and since the data-to-parameter ratio is large the refinements have large radii of convergence and converge quickly. In a step prior to the least-squares fitting, the data are processed by a Bayesian procedure (French & Wilson, 1978) that improves the experimental values $|F|^2_{\mathrm{meas}}$, $\sigma(|F|^2_{\mathrm{meas}})$, $|F|_{\mathrm{meas}}$ and $\sigma(|F|_{\mathrm{meas}})$, especially for the weak reflections with $-3\sigma(I) < I < +3\sigma(I)$.

The residual minimized in the least-squares fitting is,

$$\begin{aligned}\chi^2 &= \sum_{\mathbf{h}} mw\Delta^2 = \sum_{\mathbf{h}} m\Delta^2/\sigma^2(\Delta) \\ &= \sum_{\mathbf{h}} m(y_o - y_c)^2/[\sigma^2(y_o) + \sigma^2(y_c)],\end{aligned} \tag{24}$$

where m is the point-group multiplicity of reflection **h** and where it is essential that the weights be based on variance estimates for both the observed *and* calculated ordinates.

For the iterative refinements based on (14) and (16) the observed and calculated ordinates are,

$$y_o = |F|^2_{\mathrm{meas}} \Big/ \left(\varepsilon \sum_a f_a^2\right)$$

$$y_c = \begin{cases} k^{-2}\exp(-2\mu s^2), \\ k^{-2}\exp(-2\mu s^2 + 2\sigma^2 s^4), \\ k^{-2}\exp(-2\mathbf{h}^T\boldsymbol{\mu}\mathbf{h}), \text{ or} \\ k^{-2}\exp[-2\mathbf{h}^T\boldsymbol{\mu}\mathbf{h} + 2(\mathbf{h}^T\boldsymbol{\sigma}\mathbf{h})^2]. \end{cases} \quad (25)$$

The y_o are measured relative intensities normalized to the absolute scale of scattering by a structure of uniformly randomly distributed atoms at rest, and the y_c are normalized Wilson expectation values on the relative experimental scale for vibrating or disordered atoms. The corresponding variance estimates are,

$$\begin{aligned} \sigma^2(y_o) &= \sigma^2(|F|^2_{\text{meas}}) \Big/ \left(\varepsilon \sum_a f_a^2\right)^2 \\ &= y_o{}^2[\sigma(|F|^2_{\text{meas}})/|F|^2_{\text{meas}}]^2 \\ \sigma^2(y_c) &= y_c^2\sigma_w^2, \end{aligned} \quad (26)$$

where σ_w^2 is the normalized variance (10) of the appropriate Wilson distribution, namely,

$$\sigma_w^2 = \begin{cases} 1 & \text{for acentric reflections, or} \\ 2 & \text{for centric reflections.} \end{cases} \quad (27)$$

Similarly, for the initial approximation by logarithmically linearized fit based on (15) with $\sigma = 0$,

$$\begin{aligned} y_o &= \ln\left[|F|^2_{\text{meas}} \Big/ \left(\varepsilon \sum_a f_a^2\right)\right] \\ y_c &= -2\ln k - 2\mu s^2, \end{aligned} \quad (28)$$

and,

$$\begin{aligned} \sigma^2(y_o) &= [\sigma(|F|^2_{\text{meas}})/|F|^2_{\text{meas}}]^2 \\ \sigma^2(y_c) &= \sigma_w^2. \end{aligned} \quad (29)$$

Unlike the many least-squares problems in which $\sigma^2(y_c)$ is negligible compared with $\sigma^2(y_o)$, the present problem presents a reverse situation since, typically, $\langle\sigma(|F|^2_{\text{meas}})\rangle/\langle|F|^2_{\text{meas}}\rangle \leq 0.05$ but $\sigma_w = 1$ or $(2)^{1/2}$. Furthermore, for the iterative least-squares fitting, the weighting factors must be re-evaluated in each cycle since their values depend on the value of the fitted parameters. This compounds the non-linearity of the problem.

To deal with the non-linearity and the strong numerical correlations among the fitted parameters, which often have correlation coefficients $\rho(k, \mu)$, $\rho(k, \sigma) < -0.9$ and $\rho(\mu, \sigma) > 0.9$, we have programmed the iterative fitting to estimate optimum shift-damping factors $0 \leq f \leq 1$ in each cycle (Hamilton, 1964). The fitting typically converges to a standardized root-mean-square error of fit,

$$Z = [\chi^2/(n-m)]^{1/2} \simeq 1, \quad (30)$$

for n data and m fitted parameters, but the normalized root-mean-square error of fit is always very large,

$$R = \left(\chi^2/\sum wy_o^2\right)^{1/2} > 0.5. \quad (31)$$

The large R values are not surprising since the iterative fitting corresponds to a structure refinement for which the structural model is a single super-pseudoatom, with the scattering power of whole unit cell, positioned at the unit-cell origin (Sheriff & Hendrickson, 1987). The optimizations of the shift-damping factors, significance tests for an improved fit as more parameters are included, and the criterion for convergence at cycle n,

$$R_n < R_{n-1} \quad \text{and} \quad f_n|\delta/\sigma|_{\max} < 10^{-6}, \quad (32)$$

are based on R rather than Z because the parameter-dependent weights appear in both the numerator and denominator of R but in only the numerator of Z.

APPENDIX C. Variability of fitted k and B parameters

One of the referees of this paper asked how the values of the scale and mean-square displacement parameters obtained using the statistical expectation value of the Debye–Waller factor compare with values obtained by other methods, and how the values vary with data resolution. Data pertaining to these questions are given in Table 8 for the insulin structure; corresponding data from the rubredoxin and crambin structures show the same trends.

Table 8 shows that the methods that tacitly assume $\sigma(B) = 0$ tend to overestimate k and underestimate $\mu(B)$ more than the fit based on (14) with $\sigma(B) > 0$ does. In general, we have found that the best agreement with 'true' refined-structure k and B values is obtained using (14); relatively good agreement results from analysis of the Patterson origin peak; and relatively poor agreement results from a conventional Wilson plot. Table 8 also shows that, while cutting back the data resolution from 1.5 to 2 Å actually improved the fitted k and B parameters, further cutting back from 2 to 4 Å gave worse and worse parameter estimates. We think that the improvement at $d_{\min} = 2$ Å occurred because, with the data set truncated to $s^2_{\max} = 1/(2d_{\min})^2 = 0.0625$ Å^2, the fit of (14) could better accommodate the concave curvature of the data shown in Fig. 2. This behaviour suggests that more accurate k and B parameters might be obtained if an analog of (14) were derived for a Gramm–Charlier expansion about a normal distribution (Johnson & Levy, 1974) to include skewness and kurtosis terms to fit data with strongly concave curvature.

The data presented in Table 8 illustrate our general experience that reliable estimation of k and B parameters requires data to ~2.5 Å or better resolution. Absent reliable k and B parameters, lower resolution data sets

Table 8. *Effects of different fitting methods and data-resolution limits on the scale and mean-square atomic displacement parameters fitted to the model-calculated $|F_c|^2$ data for the insulin structure*

d_{min} (Å)	k	$\mu(B)$ (Å^2)	$\sigma(B)$ (Å^2)	Fitting method	Reference
0	1	29.6	19.3	Refined structure averages	(*a*)
1.5	1.24	18.1	0	Rogers $P(000)$ analysis	(*b*)
1.5	1.74	15.4	0	Wilson plot, (15) with $\sigma = 0$	(*c*)
1.5	1.33	15.6	0	(14) with $\sigma = 0$	(*d*)
1.5	1.14	22.0	7.2	(14)	(*e*)
2	1.09	25.9	10.7	(14)	(*e*)
2.5	1.19	27.2	0.006	(14)	(*e*)
3	1.25	26.8	0.02	(14)	(*e*)
3.5	1.42	2.4	0.6	(14)	(*e*)
4	2.27	0.0	0	(14)	(*e*)

References: (*a*) Baker *et al.*, 1988; (*b*) Rogers, 1965; Blessing & Langs, 1988; (*c*) Wilson, 1949; (*d*) Levy *et al.*, 1970; (*e*) this work.

cannot be normalized using (18), and one must resort to normalization to the empirical local average intensity,

$$|E_o(\mathbf{h})| = |F_o(\mathbf{h})|/[\varepsilon(\mathbf{h})\langle|F_o|^2/\varepsilon\rangle]^{1/2}, \qquad (33)$$

as suggested by Main (1985). Experimental $|E|$ magnitudes estimated by (33) are generated by our program *BAYES*.

We are grateful for help from several colleagues. Grant Moss pointed out the simplification by completing the square in equation (6); Eleanor Dodson supplied the diffraction data and refined coordinates for 2-zinc pig insulin; Zbyszek Dauter supplied the diffraction data and refined coordinates for rubredoxin; Håkon Hope supplied the diffraction data, and Martha Teeter supplied the refined coordinates, for crambin; Dave Smith provided helpful advice. We also gratefully acknowledge support from USDHHS PHS NIH grants GM46733, GM34073, DK19856 and HL32303.

References

Baker, E. N., Blundell, T. L., Cutfield, J. F., Cutfield, S. M., Dodson, E. J., Dodson, G. G., Hodgkin, D. M. C., Hubbard, R. E., Isaacs, N. W., Reynolds, C. D., Sakabe, K., Sakabe, N. & Vijayan, N. M. (1988). *Philos. Trans. R. Soc. London*, **319**, 369–456.

Blessing, R. H. & Langs, D. A. (1988). *Acta Cryst.* A**44**, 729–735.

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*, p. 334. New York: Academic Press.

Dauter, Z., Sieker, L. C. & Wilson, K. S. (1992). *Acta Cryst.* A**48**, 42–59.

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

Hamilton, W. C. (1964). *Statistics in Physical Science*, pp. 155–156. New York: The Roland Press Co.

Hope, H. (1988). *Acta Cryst.* B**44**, 22–26.

Iwasaki, H. & Ito, T. (1977). *Acta Cryst.* A**33**, 227–229.

Johnson, C. K. & Levy, H. A. (1974). In *International Tables for X-ray Crystallogrpahy*, Vol. IV, edited by W. C. Hamilton & J. A. Ibers, pp. 311. Birmingham, England: Kynoch Press. (Present distributor Kluwer Academic Publishers, Dordrecht.)

Karle, J. & Hauptman, H. (1953). *Acta Cryst.* **6**, 473–476.

Levy, H. A., Thiessen, W. E. & Brown, G. M. (1970). Am. Crystallogr. Assoc. Meeting, New Orleans, Louisiana, March 1970. Abstract No. B6.

Main, P. (1985). In *Crystallographic Computing 3: Data Collection, Structure Detemination, Proteins, and Databases*, edited by G. M. Sheldrick, C. Krüger & R. Goddard, p. 208. Oxford University Press.

Nielsen, K. (1975). *Acta Cryst.* A**31**, 762–763.

Rogers, D. (1965). *Computing Methods in Crystallography*, edited by J. S. Rollet, pp. 117–148. Oxford: Pergamon Press.

Rogers, D. (1980). *Theory and Practice of Direct Methods in Crystallography*, edited by M. F. C. Ladd & R. A. Palmer, pp. 82–92. New York: Plenum Press.

Sheriff, S. & Hendrickson, W. A. (1987). *Acta Cryst.* A**43**, 118–121.

Teeter, M. M., Roe, S. M. & Heo, N. H. (1993). *J. Mol. Biol.* **230**, 292–311.

Wilson, A. J. C. (1942). *Nature (London)*, **150**, 151, 152.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.