

# INTENSITY STATISTICS AND NORMALIZATION

ROBERT H. BLESSING, D.Y. GUO, and DAVID A. LANGS  
*Hauptman-Woodward Institute*  
73 High Street  
Buffalo, New York 14203-1196, USA  
E-mail: blessing@hwi.buffalo.edu

## 1. Introduction

The "fundamental theorem" of X-ray crystallography is (see, e.g., Coppens, 1996; Bricogne, 1996) the dual-space Fourier transform relationship

$$F(\mathbf{h}) \stackrel{\mathcal{F}}{=} \rho(\mathbf{r}) \stackrel{\mathcal{F}^{-1}}{=} \quad (1)$$

between crystal structure factors in reciprocal space,

$$F(\mathbf{h}) = \mathcal{F}[\rho(\mathbf{r})] = \int_V d^3\mathbf{r} \rho(\mathbf{r}) \exp(+2\pi i\mathbf{h}\cdot\mathbf{r}), \quad (2)$$

and the unit-cell electron density distribution in crystal space,

$$\rho(\mathbf{r}) = \mathcal{F}^{-1}[F(\mathbf{h})] = V^{-1} \sum_{\mathbf{h}} F(\mathbf{h}) \exp(-2\pi i\mathbf{h}\cdot\mathbf{r}). \quad (3)$$

In crystal space  $\mathbf{r} = x\mathbf{a} + y\mathbf{b} + z\mathbf{c} = \sum_i r^i \mathbf{a}_i$ , and in reciprocal space  $\mathbf{h} = h\mathbf{a}^* + k\mathbf{b}^* + \ell\mathbf{c}^* = \sum_i h_i \mathbf{a}^{*i}$ , where  $\mathbf{a}^{*i} = \mathbf{a}_i \times \mathbf{a}_j / V$ ,  $V = (\mathbf{a}_i \times \mathbf{a}_j) \cdot \mathbf{a}_k$ ,  $\mathbf{a}^{*i} \cdot \mathbf{a}_j = \delta_j^i$ ,  $\delta_j^i = 1$  if  $i = j$ ,  $\delta_j^i = 0$  if  $i \neq j$ , and  $i, j, k = 1, 2, 3$  (see, e.g., Shmueli, 1996; Sands, 1996). Due to the dual-space reciprocity,  $\mathbf{h}\cdot\mathbf{r}$  reduces to  $\mathbf{h}\cdot\mathbf{r} = hx + ky + \ell z$ .

The structure factor is given by an atomic summation,

$$F(\mathbf{h}) = \sum_{a=1}^N f_a(\mathbf{h}) W_a(\mathbf{h}) \exp(2\pi i\mathbf{h}\cdot\mathbf{r}_a), \quad (4)$$

where, for atom  $a$  among the  $N$  atoms of the unit cell,  $\mathbf{r}_a$  is the atomic position vector,  $W_a(\mathbf{h})$  is the atomic Debye-Waller disorder and/or thermal vibration factor, and  $f_a(\mathbf{h})$  is the atomic X-ray scattering factor. The Debye-Waller factor is the Fourier transform of the probability density of atomic displacements from the mean atomic position,

$W_a(\mathbf{h}) = \mathcal{F}[p_a(\mathbf{r}_a - \langle \mathbf{r}_a \rangle)]$ , and, at X-ray wavelengths sufficiently far removed from the crystal's resonant electronic absorption edges, the atomic scattering factor is the Fourier transform of the atomic electron density,  $f_a(\mathbf{h}) = \mathcal{F}[\rho_a(\mathbf{r} - \mathbf{r}_a)] = \mathcal{F}[|\psi_a(\mathbf{r} - \mathbf{r}_a)|^2]$ . For spherically symmetric or spherically averaged atomic densities,  $f_a(\mathbf{h}) = f_a(|\mathbf{h}|) = f_a(h) = \mathcal{F}[4\pi r^2 |R_a(r)|^2]$ , where  $R_a(r)$  is the radial part of the atomic wavefunction  $\psi_a(\mathbf{r})$ . The Fourier transform product ( $f_a W_a$ ) in reciprocal space is the Fourier transform of the convolution product ( $\rho_a * p_a$ ) in crystal space, i.e.,  $f_a(\mathbf{h}) W_a(\mathbf{h}) = \mathcal{F}[\rho_a(\mathbf{r} - \mathbf{r}_a) * p_a(\mathbf{r}_a - \langle \mathbf{r}_a \rangle)]$ . As the reciprocal space radius  $|\mathbf{h}|$  increases through  $0 \leq |\mathbf{h}| < \infty$ , the Debye-Waller factor decreases through  $1 \geq W_a > 0$ , and the atomic scattering factor decreases through  $Z_a \geq f_a > 0$ .

The electron density (3) is a real-valued, non-negative function, but the structure factor (2) or (4) is in general a complex-valued function,

$$\begin{aligned} F &= \sum_n f_n W_n \exp(i\phi_n) = \sum_n f_n W_n (\cos \phi_n + i \sin \phi_n) \\ &= \sum_n f_n W_n \cos \phi_n + i \sum_n f_n W_n \sin \phi_n = \sum_n A_n + i \sum_n B_n = A + iB, \\ F(\mathbf{h}) &= A(\mathbf{h}) + iB(\mathbf{h}) = |F(\mathbf{h})| [\cos \phi(\mathbf{h}) + i \sin \phi(\mathbf{h})] = |F(\mathbf{h})| \exp[i\phi(\mathbf{h})]. \quad (5) \end{aligned}$$

Structure factor magnitudes  $|F(\mathbf{h})|$  can be obtained from kinematical diffraction measurements of Bragg reflection intensities,

$$I(\mathbf{h}) \propto |F(\mathbf{h})|^2 = F^*(\mathbf{h}) F(\mathbf{h}), \quad (6)$$

where  $F^* = |F|e^{-i\phi} = A - iB$  is the complex conjugate of  $F$ , with the imaginary unit  $i$  everywhere replaced by  $-i$ . The corresponding structure factor phases  $\phi(\mathbf{h})$ , however, cannot be measured experimentally [although, in specialized experiments designed to measure dynamical diffraction effects (see, e.g., Weckert, Schwegle, and Hümmer, 1993; Weckert and Hümmer, 1997) certain three-phase sums,  $\phi(\mathbf{h}, \mathbf{k}) = \phi(\mathbf{h}) + \phi(\mathbf{k}) + \phi(-\mathbf{h}-\mathbf{k})$ , can be measured]. Possible values for the unmeasurable phases are constrained by the requirement that, in conjunction with a sufficient subset that contains the largest elements of the in-principle infinite set of measurable magnitudes, the phases should produce via (3) an electron density distribution that is real, non-negative, and atomic, i.e., a unit-cell density of the form  $\rho(\mathbf{r}) = \sum_n \rho_n(\mathbf{r} - \mathbf{r}_n)$ , which exhibits distinct local maxima corresponding to a superposition of resolved atomic densities.

For hard, small-unit-cell crystals containing  $Z$  asymmetric crystal chemical units per unit cell and  $N/Z \leq 100$  independent non-hydrogen atoms per crystal chemical unit, with the atoms packed more-or-less tightly in metallic, covalent, ionic, or molecular structures, the usual experimental situation is that the number  $n$  of Bragg reflection intensities measurable as significant above background greatly exceeds the number  $3N/Z$  of unknown atomic coordinates; commonly,  $n \geq 100N/Z$ . The unmeasurable phases are then, in principle, over-determined by systems of simultaneous equations based on (4). The equations are, however, transcendental, so

straightforward analytical solution is precluded, and we are left with the crystallographic phase problem solvable in principle, but in practice difficult to solve.

For soft, large-unit-cell crystals of biological macromolecules containing  $N/Z \geq 1000$  independent non-hydrogen atoms in loosely packed biomolecules and solvate water molecules, the usual experimental situation is much less favorable, because the atomic Debye-Waller factors in large-molecule crystals are generally smaller than those in small-molecule crystals by factors of about  $e^{-2} = 0.1$  to  $e^{-10} = 5 \times 10^{-5}$ . This results in much steeper fall-off of the atomic  $[f_a(\mathbf{h}) W_a(\mathbf{h})]$  values with increasing  $|\mathbf{h}|$ , so that  $n \leq 10 N/Z$  and experimental resolution limits  $d_{\min} = 1/|\mathbf{h}|_{\max} \geq 1.5 \text{ \AA}$  are commonplace. In practice,  $|\mathbf{h}|_{\max}$  is defined operationally as the  $|\mathbf{h}|$  at which the local average measurement precision,  $\langle \sigma(|F(\mathbf{h})|^2) \rangle / \langle |F(\mathbf{h})|^2 \rangle$ , falls to  $\sim 50\%$  due to the fall-off of  $[f_a(\mathbf{h}) W_a(\mathbf{h})]$  values with increasing  $|\mathbf{h}|$ . In consequence, biomolecular crystal structure analyses are usually carried out, not at atomic resolution, but at resolutions corresponding to small groups of atoms. Atomic structures within the groups are assigned from a database of known structures of chemical functional groups determined in atomic-resolution analyses of small-unit-cell crystals. Biochemical functional groups typically contain four to ten non-hydrogen atoms and have group diameters of  $\sim 3.5 \text{ \AA}$ . The functional groups in proteins include, of course,  $C^{\alpha}_{1/2}-CO-NH-C^{\alpha}_{1/2}$  main-chain peptide groups and side-chain groups of the 20 common amino acids. For macromolecular crystals that do not diffract to atomic resolution, the "atomicity" constraint translates to a requirement for distinct high-density volumes enveloped by isodensity surfaces at  $\rho \propto \sigma(\rho) = \langle (\rho - \langle \rho \rangle)^2 \rangle^{1/2}$  surrounding resolved groups or chains of atoms separated by distances that exceed the experimental resolution limit.

With due allowance for experimental resolution limits, the over-determinacy, non-negativity, and atomicity constraints on the unknown phases imply magnitude-conditioned relationships among them. Based on the existence of such relationships, and using hypothetical random-atom structures as starting models, the probabilistic phase estimation methods that are the subject of this book have been developed. The probabilistic treatments of the crystallographic phase problem may be broadly classified as either frequentist or Bayesian. Frequentist approaches are formulated in terms of normalized structure factors,

$$E(\mathbf{h}) = F(\mathbf{h}) / \left\{ \sum_{a=1}^N [f_a(\mathbf{h}) W_a(\mathbf{h})]^2 \right\}^{1/2}, \quad (7)$$

and Bayesian approaches are formulated in terms of unitary structure factors,

$$U(\mathbf{h}) = F(\mathbf{h}) / \sum_{a=1}^N f_a(\mathbf{h}) W_a(\mathbf{h}). \quad (8)$$

As will be shown below, for structures in space group  $P1$  or  $P\bar{1}$ , the square of the denominator on the right-hand side of (7) is equal to the probabilistic or statistical

expectation value of the squared structure factor magnitude,  $\langle |F(\mathbf{h})|^2 \rangle = \sum_a [f_a(\mathbf{h}) W_a(\mathbf{h})]^2$ , so the  $E$  normalization (Hauptman and Karle, 1953) is such that  $\langle |E|^2 \rangle = 1$ . From (4) and (5), the denominator on the right-hand side of (8) represents the maximum possible structure factor magnitude, with all atoms scattering in phase with  $\phi_a = 0$ , so the  $U$  normalization (Harker and Kasper, 1948) is such that  $|U| \leq 1$ .

For structures that can be fairly approximated as being composed of equal atoms with equal mean-square atomic displacements, substituting (4) into (7) and (8) reduces them to

$$E(\mathbf{h}) \approx N^{-1/2} \sum_{a=1}^N \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_a), \quad U(\mathbf{h}) = N^{-1} \sum_{a=1}^N \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_a), \quad \text{and} \quad U(\mathbf{h}) \approx E(\mathbf{h})/\sqrt{N}.$$

Thus, in terms of normalized or unitary structure factors, structures of equal atoms with equal mean-square displacements are equivalent to structures of equal point-atoms at rest. Structure factor normalization thus represents a simplifying idealization of crystal structure that has been widely employed in developing probabilistic theory for the crystallographic phase problem. The crystallographic literature on intensity statistics and normalization is sizeable, and excellent summary reviews of it (Shmueli and Wilson, 1996; Giacovazzo, 1996) are presented in the new, generally excellent, Volume B of the *International Tables for Crystallography*.

## 2. Debye-Waller Factors

The atomic Debye-Waller factor appearing in equations (4) through (8) is (see, e.g., Johnson and Levy, 1974) given by

$$W_a(\mathbf{h}) = \exp(-2\pi^2 \langle u_a^2 \rangle / d_h^2), \quad (9)$$

where  $\langle u_a^2 \rangle$  is the mean-square value of the displacement of atom  $a$  from its mean position,  $u_a = \mathbf{r}_a - \langle \mathbf{r}_a \rangle$ , due to disorder and/or thermal vibration perpendicular to the Bragg reflecting crystal planes with Miller indices  $(h, k, \ell) = (\mathbf{h} \cdot \mathbf{a}, \mathbf{h} \cdot \mathbf{b}, \mathbf{h} \cdot \mathbf{c})$  and interplanar spacing  $d_h = 1/|\mathbf{h}|$ , where  $|\mathbf{h}| = (\sum_i \sum_j h_i h_j \mathbf{a}^i \cdot \mathbf{a}^j)^{1/2} = 2(\sin \theta_h)/\lambda$ . In general, atomic displacements are anisotropic and, for trivariate Gaussian distributions of rectilinear displacements about the mean atomic position, (9) becomes

$$W_a(\mathbf{h}) = \exp(-\mathbf{h}^T \mathbf{b}_a \mathbf{h}) = \exp(-2\pi^2 \mathbf{H}^T \mathbf{U}_a \mathbf{H}), \quad (10)$$

where we employ matrix instead of vector notation in the exponential arguments in which  $\mathbf{h}^T = [h \ k \ \ell]$  and  $\mathbf{H}^T = [h^* \ k^* \ \ell^*]$  are row matrices,  $\mathbf{h}$  and  $\mathbf{H}$  are the corresponding column matrices, and  $\mathbf{b}$  and  $\mathbf{U}$  are square matrices that represent positive-definite, symmetric, second-rank tensors with six ( $i \leq j = 1, 2, 3$ ) independent components for an atom with site symmetry 1. The  $\mathbf{b}$  and  $\mathbf{U}$  matrix elements are

related by

$$b^{ij} = 2\pi^2 a^{*i} a^{*j} U^{ij}, \quad U^{ij} = U^{ji}, \quad U^{ii} = \langle (u^i)^2 \rangle, \quad (11)$$

and necessary and sufficient conditions for positive definiteness of the mean-square displacement tensors are given by the determinantal inequalities

$$U^{11} > 0, \quad \begin{vmatrix} U^{11} & U^{12} \\ U^{21} & U^{22} \end{vmatrix} > 0, \quad \text{and} \quad \begin{vmatrix} U^{11} & U^{12} & U^{13} \\ U^{21} & U^{22} & U^{23} \\ U^{31} & U^{32} & U^{33} \end{vmatrix} > 0. \quad (12)$$

In an isotropic approximation (9) reduces to

$$W_s(\mathbf{h}) = W_s(|\mathbf{h}|) = \exp[-B_{iso,s}(\sin \theta_n)^2/\lambda^2], \quad (13)$$

where  $B_{iso} = 8\pi^2 \langle u^2 \rangle$ . Equivalent isotropic (scalar) values can be obtained by contraction of anisotropic (tensor) values,

$$B_{iso,eq} = (8\pi^2/3) \sum_i \sum_j (\mathbf{a}_i \cdot \mathbf{a}_j) a^{*i} a^{*j} U^{ij} = (4/3) \sum_i \sum_j (\mathbf{a}_i \cdot \mathbf{a}_j) b^{ij}, \quad (14)$$

and, conversely, equivalent anisotropic values can be generated by expansion of isotropic values,

$$U_{eq}^{ij} = [B_{iso}/(8\pi^2)] (\mathbf{a}^{*i} \cdot \mathbf{a}^{*j}) / (a^{*i} a^{*j}). \quad (15)$$

### 3. Statistical Expectation Values and Probability Distributions

In (7), the normalized structure factor  $E(\mathbf{h})$  is defined with respect to the statistical expectation value of the squared structure factor magnitude  $\langle |F(\mathbf{h})|^2 \rangle$ . For any function  $y$  of a random variable  $x$  distributed according to a distribution density function  $p(x)$ , the statistical expectation value for  $y$  is given by

$$\langle y \rangle = \int_{-\infty}^{+\infty} y(x) p(x) dx / \int_{-\infty}^{+\infty} p(x) dx. \quad (16)$$

If  $p(x)$  is a normalized probability density function, the denominator in (16) is unity, and

$$P = \int_a^b p(x) dx \quad (17)$$

gives the probability that  $a \leq x \leq b$ . For a representative finite sample of  $n$  values of the quantity  $y$ , (16) can be approximated by a weighted average,

$$\langle y \rangle = \sum_{i=1}^n w_i y_i / \sum_{i=1}^n w_i , \quad (18)$$

where the weight  $w_i$  for each  $y_i = y(x_i)$  is proportional to the relative distribution density  $p(x_i)$ . For distributions  $p(x)$  with finite variance,

$$\text{var}(x) = \sigma^2(x) = \langle (x - \langle x \rangle)^2 \rangle , \quad (19)$$

appropriate weights are the reciprocal variances,

$$w_i = 1/\sigma^2(y_i) , \quad (20)$$

where, to approximation by a first-order Taylor expansion,

$$\text{var}(y) = \sigma^2(y) = (dy/dx)^2 \sigma^2(x) . \quad (21)$$

If the  $\sigma(y_i)$  values are constant, unit weights  $w_i = 1$  are appropriate, and (18) reduces to the arithmetic average  $\langle y \rangle = (\sum_i y_i)/n$ .

If  $y = y(\mathbf{x})$  is a multivariate function of  $n$  random variables,  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ , the joint probability density function of  $\mathbf{x}$  is the function  $p_j(\mathbf{x})$  such that

$$P = \int_{a_1}^{b_1} \dots \int_{a_n}^{b_n} p_j(\mathbf{x}) d^n \mathbf{x} . \quad (22)$$

gives the probability that  $a_1 \leq x_1 \leq b_1, \dots$ , and  $a_n \leq x_n \leq b_n$ . The marginal probability density function  $p_M(x_i)$  of an element  $x_i$  of  $\mathbf{x}$ , irrespective of the values of the other elements, is

$$p_M(x_i) = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} p_j(\mathbf{x}) d^{n-1} \mathbf{x} , \quad (23)$$

where the  $(n-1)$ -fold integration is performed over all the elements of  $\mathbf{x}$  except  $x_i$ . If  $p_j(\mathbf{x}, \mathbf{y})$  is the combined joint probability density function for two sets of random variables  $\mathbf{x}$  and  $\mathbf{y}$ , then the conditional probability density function for  $\mathbf{x}$  given fixed, particular values for the elements of  $\mathbf{y}$  is

$$p_C(\mathbf{x}|\mathbf{y}) = p_j(\mathbf{x}, \mathbf{y})/p_M(\mathbf{y}) . \quad (24)$$

Therefore,

$$p_j(\mathbf{x}, \mathbf{y}) = p_C(\mathbf{x}|\mathbf{y}) p_M(\mathbf{y}) = p_C(\mathbf{y}|\mathbf{x}) p_M(\mathbf{x}) \quad (25)$$

from which follows Bayes's Theorem,

$$p_C(\mathbf{x}|\mathbf{y}) = p_C(\mathbf{y}|\mathbf{x}) p_M(\mathbf{x})/p_M(\mathbf{y}) . \quad (26)$$

The multivariate analog of the univariate relationship (21) is

$$\sigma^2(\mathbf{y}) = \sum_{i=1}^n \sum_{j=1}^n (\partial y/\partial x_i)(\partial y/\partial x_j) \rho_{ij} \sigma(x_i) \sigma(x_j) , \quad (27)$$

where  $\rho_{ij}$  is the correlation coefficient,

$$\rho_{ij} = \frac{\text{cov}(x_i, x_j)}{[\text{var}(x_i) \text{var}(x_j)]^{1/2}} = \frac{\langle (x_i - \langle x_i \rangle) (x_j - \langle x_j \rangle) \rangle}{[\langle (x_i - \langle x_i \rangle)^2 \rangle \langle (x_j - \langle x_j \rangle)^2 \rangle]^{1/2}} , \quad (28)$$

which has the property  $-1 \leq \rho_{ij} \leq +1$ . For the joint variation of two or more functions of the same set of random variables,

$$\text{cov}(y_i, y_j) = \sum_{k=1}^n \sum_{\ell=1}^n (\partial y_i/\partial x_k)(\partial y_j/\partial x_\ell) \text{cov}(x_k, x_\ell) . \quad (29)$$

If  $\mathbf{x}$  represents a set of  $n$  parameters fitted by least-squares minimization to a set of  $m > n$  data  $\mathbf{y}$ , then

$$\text{var}(x_i) = \text{cov}(x_i, x_i) \quad \text{and} \quad \text{cov}(x_i, x_j) = a^{ij} \chi^2/(m - n) , \quad (30)$$

where  $a^{ij}$  is an element of the matrix  $[a^{ij}]$  inverse to the matrix  $[a_{ij}]$  of coefficients of the least-squares normal equations, and  $\chi^2 = \sum_i w_i [y_i - y(\mathbf{x})]^2$  is the minimized least-squares residual. Equations (21) and (27) through (30) form the basis of propagation-of-error calculations of the effects of experimental measurement uncertainties on the uncertainties of measurement-derived quantities.

#### 4. Wilson Expectation Values $\langle |F(\mathbf{h})|^2 \rangle$ and the Wilson Plot

Historically, the literature on intensity statistics and normalization dates back to an exchange of letters to *Nature* between S.H. Yü (1942) and A.J.C. Wilson (1942). Wilson considered the squared structure factor magnitude from (4) for a structure in space group  $P1$ ,

$$\begin{aligned} |F(\mathbf{h})|^2 &= F(\mathbf{h}) F^*(\mathbf{h}) \\ &= \sum_a \sum_b f_a(\mathbf{h}) f_b(\mathbf{h}) W_a(\mathbf{h}) W_b(\mathbf{h}) \exp [2\pi i \mathbf{h} \cdot (\mathbf{r}_a - \mathbf{r}_b)] \\ &= \sum_a [f_a(\mathbf{h}) W_a(\mathbf{h})]^2 + \sum_a \sum_{b \neq a} f_a(\mathbf{h}) f_b(\mathbf{h}) W_a(\mathbf{h}) W_b(\mathbf{h}) \exp [2\pi i \mathbf{h} \cdot (\mathbf{r}_a - \mathbf{r}_b)] , \end{aligned} \quad (31)$$

and pointed out that if the  $|F(\mathbf{h})|^2$  are locally averaged in spherical shells of  $|\mathbf{h}| = 1/d_h = 2(\sin \theta_h)/\lambda$  then, in shells in which  $d_h$  does not greatly exceed the near-neighbor values of the interatomic distances  $|\mathbf{r}_a - \mathbf{r}_b|$ , the arguments  $\phi = 2\pi\mathbf{h}\cdot(\mathbf{r}_a - \mathbf{r}_b)$  of the functions  $\exp(i\phi) = \cos \phi + i \sin \phi$  will sample the range  $0 \leq \phi \pmod{2\pi} < 2\pi$  more-or-less uniformly, so that  $\cos \phi$  and  $\sin \phi$  will oscillate between positive and negative values and average to practically zero. The statistical expectation value for (31) will then be

$$\langle |F(\mathbf{h})|^2 \rangle = \sum_a [f_a(\mathbf{h}) W_a(\mathbf{h})]^2. \quad (32)$$

Usually Bragg intensities (6) are measured, not on the absolute scale, but on a relative experimental scale, so that

$$|F|_{\text{abs}} = k |F|_{\text{rel}}, \quad (33)$$

and, under the approximation of isotropic mean-square atomic displacements (13) that are approximately the same for all atoms of the unit cell, (32) and (33) yield

$$\langle |F(\mathbf{h})|_{\text{rel}}^2 \rangle = k^{-2} \exp[-2B_{\text{iso}}(\sin \theta_h)^2/\lambda^2] \sum_a f_a^2(\mathbf{h}), \quad (34)$$

from which the absolute scaling factor  $k$  and the overall mean-square atomic displacement parameter  $B_{\text{iso}}$  can be estimated by means of a least-squares straight line fitted to a plot of  $\ln \langle |F(\mathbf{h})|_{\text{meas}}^2 / \sum_a f_a^2(\mathbf{h}) \rangle_{|\mathbf{h}|}$  vs.  $\langle (\sin \theta_h)^2 / \lambda^2 \rangle_{|\mathbf{h}|}$ , where the notation  $\langle x \rangle_{|\mathbf{h}|}$  denotes a local spherical  $|\mathbf{h}|$ -shell average. From empirical estimates for  $k$  and  $B_{\text{iso}}$ , experimental estimates of normalized (7) or unitary (8) structure factor magnitudes can be obtained as

$$|E(\mathbf{h})|_{\text{meas}} = |F(\mathbf{h})|_{\text{meas}} k \exp[+B_{\text{iso}}(\sin \theta_h)^2/\lambda^2] [\sum_a f_a^2(\mathbf{h})]^{-1/2}, \quad (35)$$

$$|U(\mathbf{h})|_{\text{meas}} = |F(\mathbf{h})|_{\text{meas}} k \exp[+B_{\text{iso}}(\sin \theta_h)^2/\lambda^2] [\sum_a f_a(\mathbf{h})]^{-1}. \quad (36)$$

## 5. The Wilson Distributions

Considering (31) and (32) further, Wilson (1949) derived the marginal probability density functions for structure factor magnitudes and intensities to be expected from uniform random unit-cell distributions of atomic positions in the space groups  $P1$  and  $P\bar{1}$ . Assuming a uniform random distribution of the atomic phase components  $0 \leq \phi_a = 2\pi\mathbf{h}\cdot\mathbf{r}_a \pmod{2\pi} < 2\pi$ , the  $P1$  derivation applies the central limit theorem separately to the real part,  $A = \sum_a f_a W_a \cos \phi_a = |F| \cos \phi$ , and the imaginary part,  $B = \sum_a f_a W_a \sin \phi_a = |F| \sin \phi$ , of the structure factor (4) and (5). The  $P\bar{1}$  derivation follows the real part of the  $P1$  derivation, allowing for halving of the number of independent atoms since the atoms occur in pairs at positions  $+\mathbf{r}_a$  and  $-\mathbf{r}_a$ , which restricts the structure factor phases to  $\phi(\mathbf{h}) = 0$  or  $\pi$ .



The resulting Wilson probability density functions for crystal structure factor magnitudes  $|F| = (A^2 + B^2)^{1/2}$  are then the Gaussian forms:

Acentric  $P1$  distribution

$$p_M(|F|) = (2|F|/\langle|F|^2\rangle) \exp(-|F|^2/\langle|F|^2\rangle), \quad \text{var}(|F|) = \langle|F|^2\rangle \quad (37)$$

Centric  $P\bar{1}$  distribution

$$p_M(|F|) = [2/(\pi\langle|F|^2\rangle)]^{1/2} \exp[-|F|^2/(2\langle|F|^2\rangle)], \quad \text{var}(|F|) = \langle|F|^2\rangle \quad (38)$$

where  $\langle|F|^2\rangle$  is the Wilson expectation value given by (32). For normalized structure factor magnitudes,  $|E| = |F|/\langle|F|^2\rangle^{1/2}$ , the probability density functions are:

$$\text{Acentric} \quad p_M(|E|) = 2|E| \exp(-|E|^2), \quad (39)$$

$$\lim_{|E| \rightarrow 0} p_M(|E|) = 0, \quad \text{var}(|E|) = \langle|E|^2\rangle = 1;$$

$$\text{Centric} \quad p_M(|E|) = (2/\pi)^{1/2} \exp(-|E|^2/2), \quad (40)$$

$$\lim_{|E| \rightarrow 0} p_M(|E|) = (2/\pi)^{1/2}, \quad \text{var}(|E|) = 2\langle|E|^2\rangle = 2.$$

And for intensities,  $I = |F|^2$ , the probability density functions are:

$$\text{Acentric} \quad p_M(I) = \langle I \rangle^{-1} \exp(-I/\langle I \rangle), \quad (41)$$

$$\lim_{I \rightarrow 0} p_M(I) = \langle I \rangle^{-1}, \quad \text{var}(I) = \langle I \rangle;$$

$$\text{Centric} \quad p_M(I) = (2\pi I \langle I \rangle)^{-1/2} \exp[-I/(2\langle I \rangle)], \quad (42)$$

$$\lim_{I \rightarrow 0} p_M(I) = \infty, \quad \text{var}(I) = 2\langle I \rangle.$$

The probability density functions (39) through (42), and the corresponding cumulative distribution functions

$$N(x) = \int_0^x p_M(x) dx, \quad (43)$$

are illustrated in Figure 1. The acentric  $P1$  distributions are narrower than the centric  $P\bar{1}$  distributions, with lower probabilities of very small or very large values of  $|E|$ ,  $|F|$ , or  $I$ .

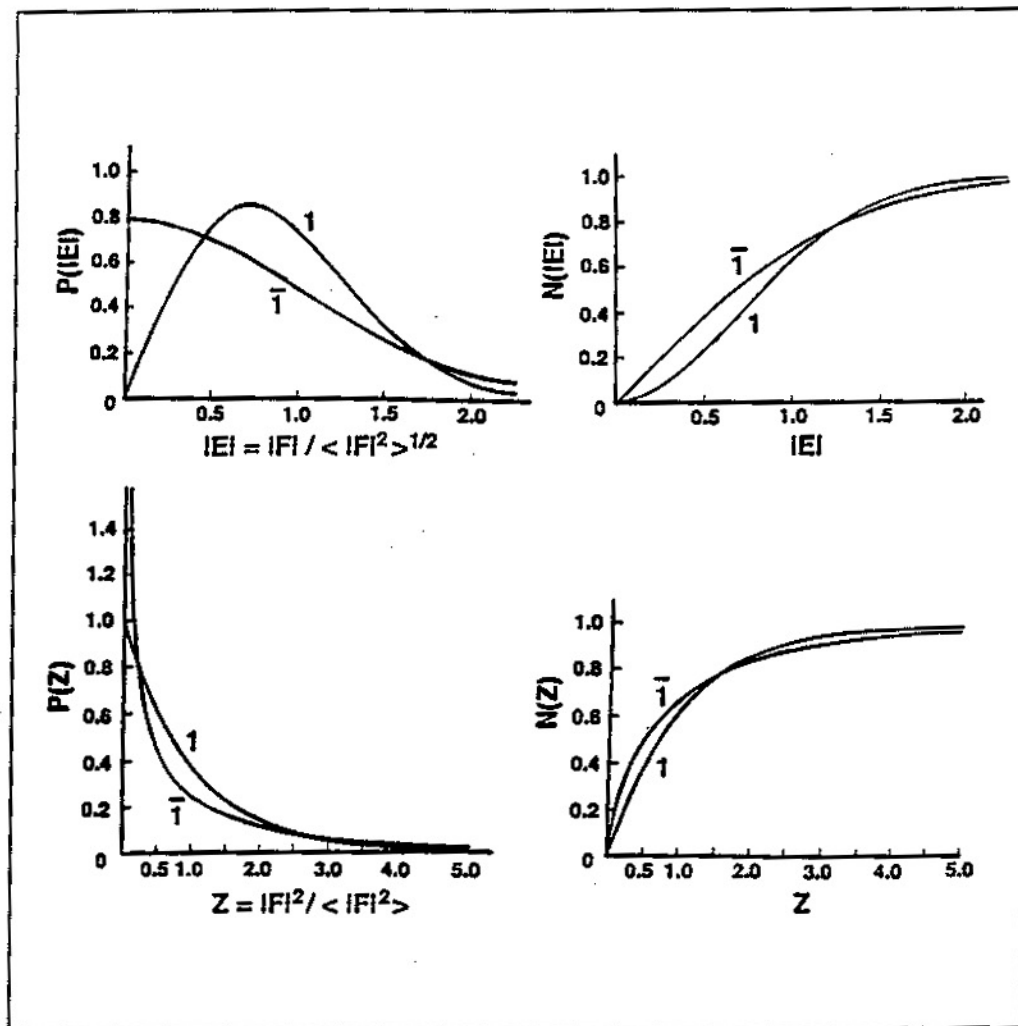


Figure 1. The Wilson distribution probability density functions,  $p(|E|)$  and  $p(|F|^2)$ , and cumulative distribution functions,  $N(|E|) = \int_0^{|E|} p(|E|) d|E|$  and  $N(Z) = \int_0^Z p(Z) dZ$ , where  $Z = |F|^2 / \langle |F|^2 \rangle$ .

Acentric P1

$$p(|E|) = 2|E| \exp(-|E|^2)$$

$$p(|F|^2) = \langle |F|^2 \rangle^{-1} \exp(-|F|^2 / \langle |F|^2 \rangle)$$

$$N(|E|) = 2 \int_0^{|E|} x \exp(-x^2) dx = 1 - \exp(-|E|^2)$$

$$N(Z) = \int_0^Z \exp(-x) dx = 1 - \exp(-Z)$$

Centric P1

$$p(|E|) = (2/\pi)^{1/2} \exp(-|E|^2/2)$$

$$p(|F|^2) = (2\pi |F| \langle |F|^2 \rangle)^{-1/2} \exp[-|F|^2 / (2\langle |F|^2 \rangle)]$$

$$N(|E|) = (2/\pi)^{1/2} \int_0^{|E|} \exp(-x^2/2) dx = \text{erf}(|E|/\sqrt{2})$$

$$N(Z) = (2\pi)^{-1/2} \int_0^Z x^{-1/2} \exp(-x/2) dx = \text{erf}[(Z/2)^{1/2}]$$

The  $P1$  and  $P\bar{1}$  Wilson distributions are archetypal. Their forms hold even for space groups of higher symmetry, if the asymmetric crystal chemical unit of the unit cell locally approximates a random-atom  $P1$  or  $P\bar{1}$  structure. The adaptation to higher symmetry is an integral factor  $\epsilon \geq 1$  multiplying the distribution parameter  $\langle |F|^2 \rangle$  so that (32), (34) and (35) become

$$\langle |F(\mathbf{h})|^2 \rangle = \epsilon(\mathbf{h}) \sum_n [f_n(\mathbf{h}) W_n(\mathbf{h})]^2, \quad (44)$$

$$\langle |F(\mathbf{h})|_{\text{ref}}^2 \rangle = k^{-2} \exp[-2B_{\text{iso}}(\sin \theta_h)^2/\lambda^2] \epsilon(\mathbf{h}) \sum_n f_n^2(\mathbf{h}), \quad (45)$$

$$|E(\mathbf{h})|_{\text{meas}} = |F(\mathbf{h})|_{\text{meas}} k \exp[+B_{\text{iso}}(\sin \theta_h)^2/\lambda^2] [\epsilon(\mathbf{h}) \sum_n f_n^2(\mathbf{h})]^{-1/2}. \quad (46)$$

We call the factor  $\epsilon(\mathbf{h})$  the *degeneracy* of the reciprocal lattice point  $\mathbf{h}$  because it accounts for symmetry-dependent multiple enhancements of  $P1$  or  $P\bar{1}$   $|F(\mathbf{h})|^2$  expectation values. The degeneracy factor is given by

$$\epsilon(\mathbf{h}) = m_L \epsilon'(\mathbf{h}), \quad (47)$$

where  $m_L$  is the lattice multiplicity,

$$\begin{aligned} m_L = & 1 \text{ for primitive P-lattices,} \\ & 2 \text{ for C-, B-, A-, or I-centered lattices,} \\ & 4 \text{ for F-centered lattices, or} \\ & 3 \text{ for R-centered lattices on hexagonal axes,} \end{aligned} \quad (48)$$

and

$$\epsilon'(\mathbf{h}) = 1, 2, 4, 8, 3, 6, \text{ or } 12 \quad (49)$$

is a projection symmetry multiplier for certain classes of zonal or axial reflections in particular reciprocal lattice point groups (Rogers, 1965, 1980). The  $m_L$  enhancements arise from the systematic extinction of a fraction  $[1 - (1/m_L)]$  of the reflections due to lattice centering and the consequent concentration of the total scattering in the allowed fraction  $1/m_L$  of the reflections. The  $\epsilon'$  enhancements arise from superposition of symmetrically equivalent atoms in projection onto mirror planes or rotation axes. In the triclinic point groups  $1$  and  $\bar{1}$ ,  $\epsilon' = 1$  for all reflections. In all point groups,  $\epsilon'(hkl) = 1$  for all non-axial, non-zonal, general reflections; but in the monoclinic point group  $2/m$  ( $b$ -axis unique), for example, the zonal  $h0l$  and axial  $0k0$  reflections are special, and have  $\epsilon'(h0l) = 2$  and  $\epsilon'(0k0) = 2$ , due to superposition in projection of mirror-equivalent and rotation-equivalent atoms, respectively. A useful complete table of  $\epsilon'$  values has been given by Iwasaki and Ito (1977).

## 6. Wilson Normalization with a Statistical Expectation Value of the Debye-Waller Factor

Underlying the normalization equations (44) through (46) and (36) is the assumption that the atomic Debye-Waller factors vary little from atom to atom in the unit cell. For many crystals this is hardly the case. In small-unit-cell molecular crystals, mean-square atomic displacements due to thermal vibration are usually larger for atoms at the periphery of a molecule than for atoms near the molecular center of mass, and larger for conformationally flexible than for conformationally rigid functional groups. In crystals of biological macromolecules, displacements due to disorder and/or thermal vibration are generally larger for atoms at the biomolecular surface than for atoms in the biomolecular core, larger for side-chain than for main-chain atoms, and larger for solvate water molecules than for atoms of the biomolecule.

The crude approximation of constant atomic Debye-Waller factors can be replaced by a less crude approximation if, instead of factoring a squared-Debye-Waller factor  $W^2$  out of the atomic sum (32), we factor-out the statistical expectation value  $\langle W^2 \rangle = \langle \exp(-2Bs^2) \rangle$ , where  $s = (\sin \theta)/\lambda$ . Assuming that the unit-cell distribution of atomic B values can be fairly approximated by a normal distribution

$$p(B) = [(2\pi)^{1/2} \sigma_B]^{-1} \exp [-(B - \mu_B)^2 / (2\sigma_B^2)] , \quad (50)$$

with mean  $\mu_B = \langle B \rangle$  and variance  $\sigma_B^2 = \langle (B - \langle B \rangle)^2 \rangle$ , it has been shown (Blessing, Guo, and Langs, 1996) that (16) yields the expectation value

$$\langle W^2 \rangle = \langle \exp(-2Bs^2) \rangle = \exp [-2(\mu_B - \sigma_B^2 s^2)s^2] . \quad (51)$$

This indicates that, due to the spread of the unit-cell distribution of atomic mean-square displacements, the expectation value of the Debye-Waller factor corresponds to an effective overall B value,  $B_{\text{eff}} = \langle B \rangle - \langle (B - \langle B \rangle)^2 \rangle s^2$ , that is smaller than the mean B and that decreases with increasing  $(\sin \theta)/\lambda$ .

Normalization effects of the spread of the distribution of mean-square atomic displacements can be sizeable because, especially in macromolecular crystals,  $\langle (B - \langle B \rangle)^2 \rangle^{1/2} - \langle B \rangle$  is not uncommon for averages of structure-refined B values. The latter generally exhibit distributions that are positively skewed (since, of physical necessity,  $B_{\text{min}} > 0$ ) and more sharply peaked than normal distributions. This suggests that (51) might be improved by employing an expansion about (50) to derive an expression for  $\langle W^2 \rangle$  that includes, in addition to the dispersion term in  $(s^2)^2$ , a skewness term in  $(s^2)^3$  and a kurtosis term in  $(s^2)^4$ . It has, however, been shown that in practice, for data sets that extend to  $d_{\text{min}} \lesssim 2.5 \text{ \AA}$  resolution, such refinements are not necessary. Normalization via (51) of data from several well-determined protein crystal structures was shown to produce  $|E_{\text{obs}}|$  values that agree with  $|E_{\text{calc}}|$  values calculated from the refined  $r_s$  and  $B_s$  parameters as well as the un-normalized  $|F_{\text{obs}}|$  values agree with the corresponding  $|F_{\text{calc}}|$  values (Blessing, Guo, and Langs, 1996).

In terms of (51), with  $s_h = (\sin \theta_h)/\lambda$ , the empirical normalization equations (45), (46), and (36) become

$$\langle |F(\mathbf{h})|_{\text{rel}}^2 \rangle = k^{-2} \exp [-2(\mu_B - \sigma_B^2 s_h^2) s_h^2] \epsilon(\mathbf{h}) \sum_a f_a^2(\mathbf{h}), \quad (52)$$

$$|E(\mathbf{h})|_{\text{meas}} = |F(\mathbf{h})|_{\text{meas}} k \exp [(\mu_B - \sigma_B^2 s_h^2) s_h^2] [\epsilon(\mathbf{h}) \sum_a f_a^2(\mathbf{h})]^{-1/2}, \quad (53)$$

$$|U(\mathbf{h})|_{\text{meas}} = |F(\mathbf{h})|_{\text{meas}} k \exp [(\mu_B - \sigma_B^2 s_h^2) s_h^2] [\sum_a f_a(\mathbf{h})]^{-1}. \quad (54)$$

Via the Debye-Waller-factor relationships (10) and (13) through (15), these are readily recast in terms of overall anisotropic mean-square displacement parameters as

$$\langle |F(\mathbf{h})|_{\text{rel}}^2 \rangle = k^{-2} \exp \{-2[\mathbf{h}^T \boldsymbol{\mu}_b \mathbf{h} - (\mathbf{h}^T \boldsymbol{\sigma}_b \mathbf{h})^2]\} \epsilon(\mathbf{h}) \sum_a f_a^2(\mathbf{h}), \quad (55)$$

$$|E(\mathbf{h})|_{\text{meas}} = |F(\mathbf{h})|_{\text{meas}} k \exp [\mathbf{h}^T \boldsymbol{\mu}_b \mathbf{h} - (\mathbf{h}^T \boldsymbol{\sigma}_b \mathbf{h})^2] [\epsilon(\mathbf{h}) \sum_a f_a^2(\mathbf{h})]^{-1/2}, \quad (56)$$

$$|U(\mathbf{h})|_{\text{meas}} = |F(\mathbf{h})|_{\text{meas}} k \exp [\mathbf{h}^T \boldsymbol{\mu}_b \mathbf{h} - (\mathbf{h}^T \boldsymbol{\sigma}_b \mathbf{h})^2] [\sum_a f_a(\mathbf{h})]^{-1}, \quad (57)$$

where here, as in (10), we employ matrix instead of vector notation in the exponential arguments in which  $\mathbf{h}^T = [h \ k \ \ell]$  is a row matrix,  $\mathbf{h}$  is the corresponding column matrix, and  $\boldsymbol{\mu}_b$  and  $\boldsymbol{\sigma}_b$  are symmetric matrices in which, to a first approximation from fitted isotropic, scalar values  $\mu_B$  and  $\sigma_B$ ,

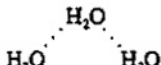
$$\mu_{b, \text{eq}}^{ij} = \mu_B a^{*i} a^{*j/4} \quad \text{and} \quad \sigma_{b, \text{eq}}^{ij} = \sigma_B a^{*i} a^{*j/4}. \quad (58)$$

## 7. Non-Wilson Characteristics of Intensity Distributions from Protein Crystals

The Wilson distributions have very wide ranges of effective applicability, but significant departures from the Wilson distributions do occur when the asymmetric crystal chemical unit does not locally approximate a  $P1$  or  $P\bar{1}$  structure of uniformly randomly distributed equal atoms. Deviant cases include: small, highly symmetric, highly heteroatomic structures; pseudosymmetric structures in which the crystallographic asymmetric unit has noncrystallographic symmetry or quasi-symmetry; and heavy-atom structures in which a small subset of the atoms of the asymmetric unit scatters much more strongly than the other atoms. Structure factor probability density functions for such cases have been derived as Edgeworth, Gram-Charlier, or Fourier-Bessel series expansions about the  $P1$  or  $P\bar{1}$  Wilson probability density functions, but discussion of these analyses would be beyond the scope of this chapter (see, e.g., Shmueli and Wilson, 1996; Castleden and Fortier, 1994).

Intensity and structure factor distributions from protein crystals depart from the Wilson distributions in several characteristic ways, because unit-cell distributions of atomic positions in protein crystals are characteristically nonuniform. Some 25 to 65% of the unit cell volume in protein crystals is occupied by solvent, mainly liquid-like water, filling the space between the large protein molecules (Matthews, 1968). Since  $\text{H}_2\text{O}\cdots\text{H}_2\text{O}$  hydrogen bond distances  $\text{O}\cdots\text{O}$  are  $\sim 2.8$  Å, while protein C-O, C-N, C-C covalent bond lengths are  $\sim 1.2$  to  $\sim 1.5$  Å, average electron densities are lower in the solvent regions in protein crystals than in the protein regions. Simple empirical calculations (Blessing, Guo, and Langs, 1996) show that  $\langle \rho_{\text{protein}} \rangle / \langle \rho_{\text{solvent}} \rangle \approx 4/3$ . In addition, protein molecules have intricately folded polymeric  $-\text{C}^\alpha-\text{CO}-\text{NH}-\text{C}^\alpha-$  structures with the fundamental repeat distances summarized in Table 1. These

Table 1. Fundamental repeat distances in protein crystals from standard bond lengths, valence angles, and conformation angles in peptides and water-water hydrogen bond geometry.

Repeat Unit	Repeat Distance
$\text{C}(\alpha_i)\cdots\text{C}(\alpha_{i+1})$	3.82 Å
$\text{C}(\alpha_{i-1})\cdots\text{C}(\alpha_{i+1})$	5.42 Å in $\alpha$ -helices 6.92 Å in $\beta$ -sheets
$\text{H}_2\text{O}\cdots\text{H}_2\text{O}$	2.75 Å O $\cdots$ O in ice
	109.5° O $\cdots$ O $\cdots$ O
$\text{H}_2\text{O}\cdots\text{H}_2\text{O}$	4.49 Å O $\cdots\cdots\cdots$ O

ubiquitous molecular repeat distances in the range  $\sim 6$  Å  $> |r_a - r_b| > \sim 3$  Å, along with  $\cdots\text{protein}\cdots(\text{H}_2\text{O})_x\cdots\text{protein}\cdots(\text{H}_2\text{O})_x\cdots$  lattice or sub-lattice repeat distances in the range  $|r_a - r_b| > \sim 30$  Å, cause reflections with  $d_n \geq 3$  Å to violate the condition  $d_n \leq |r_a - r_b|$  that underlies the deduction of the Wilson expectation values (32) and (44) (see also Harker, 1953). As a result, plots of  $\ln \langle |F(\mathbf{h})|_{\text{meas}}^2 / [\langle \epsilon(\mathbf{h}) \sum_a f_a^2(\mathbf{h}) \rangle]_{|\mathbf{h}|} \rangle$  vs.  $\langle (\sin \theta_n)^2 / \lambda^2 \rangle_{|\mathbf{h}|}$  characteristically show pronounced nonlinear oscillations for  $d = \lambda / (2 \sin \theta) \geq 3$  Å, the most prominent deviations being a local minimum at  $d \approx 6$  Å attributable to destructive interference of beams Bragg reflected from interleaved crystal planes corresponding to  $\text{C}(\alpha_{i-1})\cdots\text{C}(\alpha_i)\cdots\text{C}(\alpha_{i+1})$  repeats, and a local maximum at  $d \approx 4$  Å attributable to constructive interference of beams reflected by adjacent planes corresponding to  $\text{C}(\alpha_i)\cdots\text{C}(\alpha_{i+1})$  repeats. Typical examples of these effects are discussed in more detail and illustrated in Figures 1, 2, and 3 of Blessing, Guo, and Langs (1996); Figure 5 of Bricogne (1984); Figure 2 of French and Wilson (1978); and Figures 4 and 6 of Luzatti (1955).

## 8. Data Reduction and Error Analysis Procedures

Probabilistic phasing methods depend critically on normalized structure factor data sets that are as accurate and complete as possible. Foremost considerations include specimen crystal quality, instrument performance and calibration, and measurement strategy and technique. No less important are data processing procedures to reduce the in general multiply redundant set of raw intensity measurements to a unique set of structure factor magnitudes and, in the process, assess and preserve experimental accuracy and precision.

### 8.1. BACKGROUND SUBTRACTION, PEAK INTEGRATION, AND NET INTENSITY ESTIMATION

The procedures we employ for diffraction data from small-unit-cell crystals measured with four-circle diffractometers and single-reflection or point detectors have been described in some detail elsewhere (Blessing, 1987, 1989). The scheme of the procedures we employ for diffraction data from protein crystals measured using the oscillation method and area detectors is diagramed in Figure 2. Typically we process the oscillation frame images using the *Denzo* program (Otwinowski, 1993; Gewirth, Otwinowski, and Minor, 1995) to determine the crystal orientation and reflection indexing, the fully or partially recorded status of each reflection spot image, and the Lorentz- and polarization-corrected full- and partial-reflection net intensities  $|F|^2 = (Lp)^{-1} (I_{\text{peak}} - I_{\text{background}})$  and their statistical experimental uncertainties  $\sigma(|F|^2)$ . In place of the *Scalepack* program that is part of the *Denzo* program package, we employ our programs *denzox* and *sortav* to evaluate interframe scale factors (Hamilton, Rollett and Sparks, 1968), scale the full and partial reflections, and sum the scaled partial reflections.

### 8.2. DATA MERGING AND EXPERIMENTAL UNCERTAINTY ESTIMATION

We also use the *sortav* program to evaluate and apply, when necessary, an empirical correction (Blessing, 1995) for residual anisotropic absorption-like errors not corrected by the interframe scaling; to average equivalent measurements using robust/resistant averaging weights to down-weight measurements that are outliers from multiple-measurement sample *medians* (Blessing, 1997a) [a procedure we have found to be superior to our earlier practice of normal-probability down-weighting of outliers from unit-weighted sample *means* (Blessing and Langs, 1987)]; and to perform a bivariate analysis of variance against  $|F|^2$  and  $(\sin \theta)/\lambda$  in order to improve the experimental uncertainty estimates obtained by propagation-of-error calculations applying (21) and (27) through (30) at each stage of the data processing.

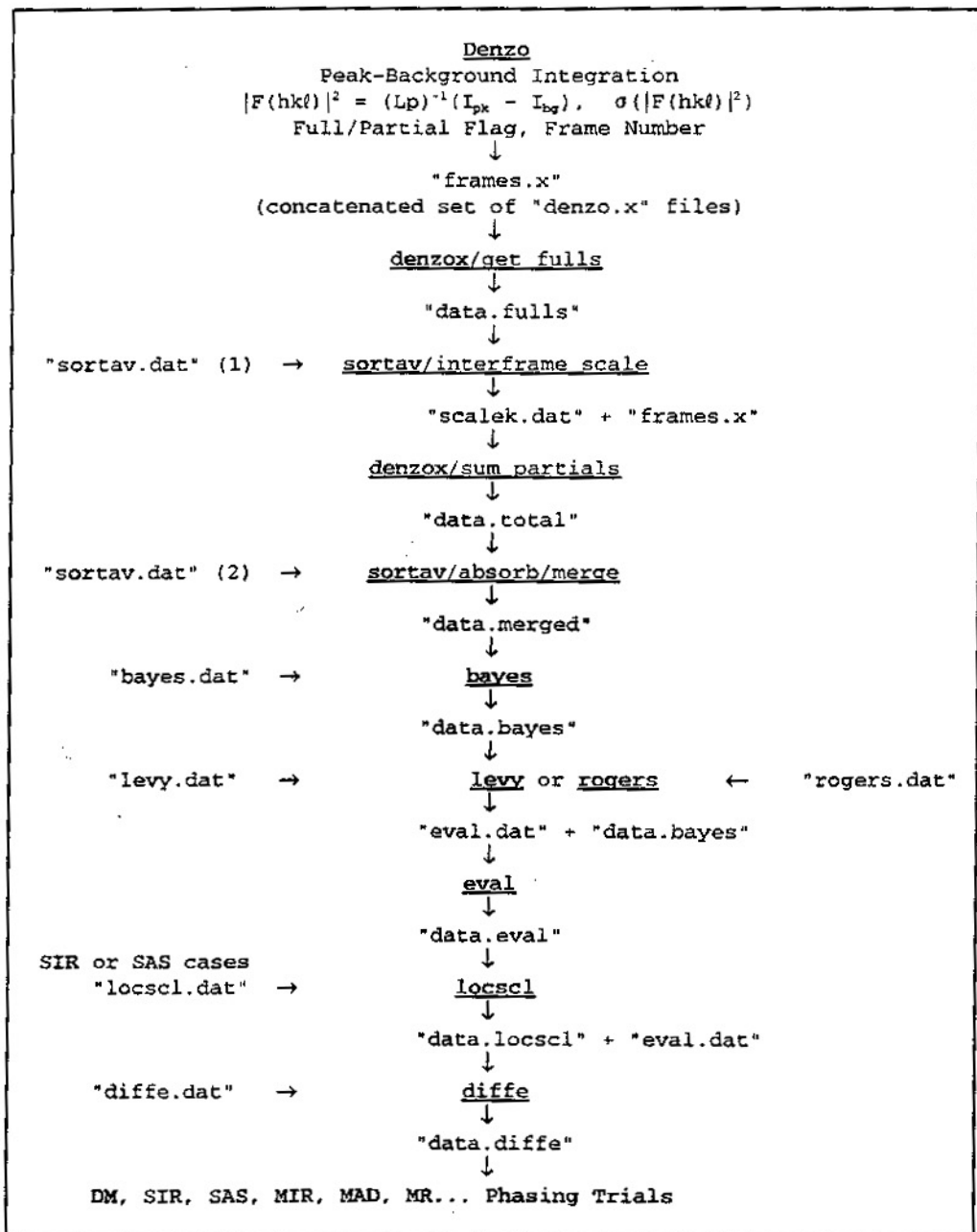


Figure 2. Flow chart for processing diffraction data from protein crystals. Program names are indicated as program, and file names are indicated as "file". Files named "program.dat" are control data files, and files named "data.program" are reflection data files.



### 8.3. BAYESIAN POST-PROCESSING

The Laue-group or point-group unique data set resulting from the *sortav* processing is then post-processed with our program *bayes*, which applies a Bayesian statistical analysis (French and Wilson, 1978) to improve weak-reflection data with  $|F|^2 \leq 3\sigma(|F|^2)$  and to derive appropriate values for the corresponding  $|F|$  and  $\sigma(|F|)$  data. Especially for protein crystals, the Bayesian post-processing can be important, because improving weak-reflection data is tantamount to increasing experimental resolution.

The post-processing applies Bayes's theorem (26) in the form

$$p_C(J|I) \propto p_C(I|J) p_M(J), \quad (59)$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior},$$

$$\text{Bayesian} \propto \text{Normal} \times \text{Wilson},$$

where  $J = |F_o(\mathbf{h})|^2$  represents the "true" intensity, and  $I = |F(\mathbf{h})|^2_{\text{meas}}$  represents the measured intensity, for a given reflection  $\mathbf{h}$ . The *a priori* expectation  $p_M(J)$  is the Wilson distribution (41) if  $\mathbf{h}$  is an acentric reflection, or (42) if  $\mathbf{h}$  is a centric reflection. From the central limit theorem, the likelihood is a normal distribution of measurement errors,

$$p_C(I|J) = [(2\pi)^{-1/2}\sigma]^{-1} \exp [-(I - J)^2/(2\sigma^2)], \quad (60)$$

which, due to the statistical experimental uncertainty  $\sigma = \sigma(I) = \sigma(|F(\mathbf{h})|^2_{\text{meas}})$ , can yield  $I < 0$  when  $J = 0$  even though  $J \geq 0$  is a physical necessity. The *a posteriori* Bayesian distributions are then

$$\text{Acentric} \quad p_C(J|I) \propto \exp [-(I - J)^2/(2\sigma^2)] \exp (-J/\langle J \rangle), \quad (61)$$

$$\text{Centric} \quad p_C(J|I) \propto J^{-1/2} \exp [-(I - J)^2/(2\sigma^2)] \exp [-J/(2\langle J \rangle)], \quad (62)$$

which, after "completing the square" and rearranging and collecting terms in the exponential arguments, become

$$\begin{cases} \text{Acentric} & p_C(J|I) \propto \exp \left\{ -\left[ J - \left[ I - \left( \frac{\sigma^2}{\langle J \rangle} \right) \right] \right]^2 / (2\sigma^2) \right\}, \\ & \left\{ \begin{array}{l} p_C(J|I) \propto \exp \left\{ -\left[ (J/\sigma) - \left[ (I/\sigma) - \left( \frac{\sigma}{\langle J \rangle} \right) \right] \right]^2 / 2 \right\} \\ p_C(J|I) = 0 \end{array} \right. \quad \begin{array}{l} \text{for } J \geq 0, \\ \text{for } J < 0; \end{array} \end{cases} \quad (63)$$

$$\begin{cases} \text{Centric} & p_C(J|I) \propto J^{-1/2} \exp \left\{ -\left[ J - \left\{ I - \left[ \frac{\sigma^2}{2\langle J \rangle} \right] \right\} \right]^2 / (2\sigma^2) \right\}, \\ & \left\{ \begin{array}{l} p_C(J|I) \propto J^{-1/2} \exp \left\{ -\left[ (J/\sigma) - \left\{ (I/\sigma) - \left[ \frac{\sigma}{2\langle J \rangle} \right] \right\} \right]^2 / 2 \right\} \\ p_C(J|I) = 0 \end{array} \right. \quad \begin{array}{l} \text{for } J \geq 0, \\ \text{for } J < 0; \end{array} \end{cases} \quad (64)$$

where the physical requirement  $J \geq 0$ , even if  $I < 0$  due to peak-minus-background statistical fluctuations when  $J = 0$ , is noted explicitly.

The probability density functions (63) and (64) are then used in (16) to improve the measured intensity data to Wilson-conditional Bayesian expectation values,

$$|F_o|^2 = \int_0^{\infty} J p_C(J|I) dJ \quad \text{and} \quad \sigma^2(|F_o|^2) = \int_0^{\infty} (J - |F_o|^2)^2 p_C(J|I) dJ,$$

and

$$|F_o| = \int_0^{\infty} J^{1/2} p_C(J|I) dJ \quad \text{and} \quad \sigma^2(|F_o|) = \int_0^{\infty} (J^{1/2} - |F_o|)^2 p_C(J|I) dJ, \quad (65)$$

all of which, including those for weak  $J = 0$  reflections that yield  $I < 0$ , are non-negative.

The experimental variables in (63) and (64) all appear in the dimensionless arguments  $\{(I/\sigma) - [\sigma/(q\langle J \rangle)]\}$ , where  $I = |F(\mathbf{h})|_{\text{meas}}^2$ ,  $\sigma = \sigma(|F(\mathbf{h})|_{\text{meas}}^2)$ ,  $q = 1$  for acentric  $\mathbf{h}$ ,  $q = 2$  for centric  $\mathbf{h}$ , and the Wilson distribution parameter  $\langle J \rangle = \langle J(\mathbf{h}) \rangle$  is estimated empirically from the local spherical  $|\mathbf{h}|$ -shell average of the measured intensities as  $\langle J(\mathbf{h}) \rangle = \epsilon(\mathbf{h}) \langle |F|_{\text{meas}}^2 / \epsilon \rangle_{|\mathbf{h}|}$ . The integrals (65) have been evaluated numerically and tabulated against arguments,  $-4 \leq (I/\sigma) - [\sigma/(q\langle J \rangle)] \leq +4$ , which cover the range in which effects of the Bayesian treatment are significant (French and Wilson, 1978). The magnitudes of the changes  $||F_o|^2 - |F|_{\text{meas}}^2|$  and  $|\sigma(|F_o|^2) - \sigma(|F|_{\text{meas}}^2)|$  depend primarily on  $I/\sigma$  and secondarily on  $\sigma/(q\langle J \rangle)$ ; the magnitudes increase with decreasing  $I/\sigma$  and increasing  $\sigma/(q\langle J \rangle)$ . In general, the effects of the Bayesian treatment are: all  $|F|_{\text{meas}}^2 < 0$  are replaced by  $|F_o|^2 \geq 0$ , and most  $|F|_{\text{meas}}^2$  with  $0 < |F|_{\text{meas}}^2 \leq 3\sigma(|F|_{\text{meas}}^2)$  are replaced by  $|F_o|^2 \leq |F|_{\text{meas}}^2$ ;  $\sigma(|F_o|^2) \leq \sigma(|F|_{\text{meas}}^2)$  since the Bayesian treatment tends to reduce measurement uncertainties by imposing the Wilson distribution requirements; and  $|F_o| \leq (|F_o|^2)^{1/2}$  since, even if negative and positive errors in  $|F|_{\text{meas}}^2$  are equally likely, negative errors are less likely than positive errors in the necessarily non-negative  $|F_o|$  and  $|F_o|^2$ . If  $I/\sigma \gg \sigma/(q\langle J \rangle)$ , the distributions (63) and (64) reduce to zero-mean, unit-variance normal distributions of  $(I - J)/\sigma$ . In practice, if  $I \geq 3\sigma(I)$  the commonly used relationships  $|F_o| = (|F_o|^2)^{1/2} = I^{1/2}$  and  $\sigma(|F_o|) = \sigma(|F_o|^2)/(2|F_o|) = \sigma(I)/(2I^{1/2})$  are valid, and the Bayesian modifications are negligible.

#### 8.4. STRUCTURE FACTOR NORMALIZATION

Since the Bayesian post-processing requires the evaluation of the local spherical  $|\mathbf{h}|$ -shell averaged intensities, our *bayes* program also produces a set of locally normalized data,

$$|E_o(\mathbf{h})| = |F_o(\mathbf{h})| / [\epsilon(\mathbf{h}) \langle |F_o|^2 / \epsilon \rangle_{|\mathbf{h}|}]^{1/2}. \quad (66)$$

To derive globally normalized data (53) or (56), the unique  $|F_o|$  data set from the *bayes* program is analyzed with our program *levy* (Blessing, Guo, and Langs, 1996) to evaluate by least-squares fit the parameters  $k$ ,  $\mu_b$ , and  $\sigma_b$  of (52), or  $k$ ,  $\mu_b^{ij}$  and  $\sigma_b^{ij}$  of (55). In turn, the fitted parameters are used in our program *eval* to obtain the normalized structure factor magnitudes (53) or (56). For both the locally and globally normalized  $|E|$  values,  $\sigma(|E|)$  values are evaluated by propagation-of-error calculations based on (21) and (27) through (30) to include the effects of error-of-fit uncertainties of the normalization parameters. An important feature of the *levy* program is that it uses a logarithmically linearized least-squares fit based on (45) only to obtain a first approximation to the scale and mean-square displacement parameters, which the program then refines by properly weighted, iterative non-linear least-squares fit to the individual-reflection data,  $|F_o(\mathbf{h})|^2/[\epsilon(\mathbf{h}) \sum_s f_s^2(\mathbf{h})]$ , rather than logarithms of local spherical  $|\mathbf{h}|$ -shell data averages (Levy, Thiessen, and Brown, 1970). The individual-reflection fitting provides a direct evaluation of anisotropy of the mean-square-displacements distribution parameters  $\mu_b^{ij}$  and  $\sigma_b^{ij}$ , and it allows the relatively many Wilson-distributed high-resolution data to overcome, or at least counteract, effects of non-Wilson distributions of the relatively few low-resolution data. [We have also found that the procedure in our program *levy* is often superior to a corresponding procedure in our earlier program *rogers* (Blessing and Langs, 1988), which estimates the parameters  $k$  and  $\mu_b^{ij}$  through an analysis of the Patterson origin peak (Rogers, 1965; Nielsen, 1975).] For low-resolution data sets with  $d_{\min} > 2.5 \text{ \AA}$ , global normalization via the *levy-eval* or *rogers-eval* programs might be unreliable, and it might be better to resort to the local normalization (66) provided by the *bayes* program.

## 9. Treatment of SIR and SAS Data

The cases of single isomorphous replacement (SIR) and single-wavelength anomalous scattering (SAS), and their extensions to the multiple isomorphous replacement (MIR) and multi-wavelength anomalous dispersion (MAD) cases, coupled with Patterson and molecular replacement (MR) analyses provide the classical tools of protein crystallography for dealing with the phase problem (see, e.g., Rossman and Arnold, 1996; Vijayan and Ramaseshan, 1996). Much of the current research on so-called direct methods (DM) probabilistic phasing is directed toward integrating DM with SIR, MIR, SAS, MAD, and MR techniques.

### 9.1. LOCAL SCALING

Given an SIR pair of data sets,  $|F_{\text{Nat}}(\mathbf{h})|$  from a native protein crystal and  $|F_{\text{Der}}(\mathbf{h})|$  from an isomorphous heavy-atom derivative crystal, or an SAS data set of Bijvoet or Friedel pairs,  $|F(+\mathbf{h})|$  and  $|F(-\mathbf{h})|$  from a crystal measured at an X-ray wavelength at which the crystal exhibits significant anomalous dispersion due to damped resonant

scattering, classical SIR or SAS methods seek to determine the substructure of heavy atoms, or of atoms that dominate the anomalous scattering, from Patterson syntheses computed with squared difference coefficients  $(|F_{\text{Der}}| - |F_{\text{Nat}}|)^2$  or  $(|F_{+\text{h}}| - |F_{-\text{h}}|)^2$ . These, like any analyses based on difference data, are highly susceptible to effects of experimental errors, since difference data  $\Delta x = x_2 - x_1$  have uncertainties  $\sigma(\Delta x) = [\sigma^2(x_1) + \sigma^2(x_2)]^{1/2}$  that are necessarily larger than either of the individual  $x_1$  and  $x_2$  data uncertainties.

To treat such cases, we employ our program *locscl* (Blessing, 1997b) to apply the local scaling procedure introduced by Matthews and Czerwinski (1975). The procedure assumes that errors that obscure the real differences between  $|F_1|$  and  $|F_2|$  pairs of data sets can be in large part empirically corrected by locally variable scale factors  $q = q(\mathbf{h})$  defined by

$$\Delta|F| = |F_1| - q|F_2| \quad (67)$$

and estimated by least-squares fit minimizing

$$\chi^2 = \sum_{\mathbf{h}=\Delta\mathbf{h}}^{\mathbf{h}+\Delta\mathbf{h}} w[ (|F_1|/|F_2|) - q ]^2, \quad (68)$$

where  $w = w(\mathbf{h}) = \sigma^{-2}(|F_1|/|F_2|)$ , and the notation

$$\sum_{\mathbf{h}=\Delta\mathbf{h}}^{\mathbf{h}+\Delta\mathbf{h}} x_{\mathbf{h}} = \sum_{\eta=\mathbf{h}-\Delta\mathbf{h}}^{\mathbf{h}+\Delta\mathbf{h}} \sum_{\kappa=\mathbf{k}-\Delta\mathbf{k}}^{\mathbf{k}+\Delta\mathbf{k}} \sum_{\lambda=\ell-\Delta\ell}^{\ell+\Delta\ell} x_{\eta\kappa\lambda} \quad (69)$$

denotes summation over a local block of reciprocal lattice points surrounding, but not including, the point of interest,  $\mathbf{h}\ell$ . For example,  $\Delta\mathbf{h} = \Delta\mathbf{k} = \Delta\ell = 1$  defines a raster or three-dimensional moving window of  $(3 \times 3 \times 3) - 1$  points for the local scale factor fit. The raster semidimensions need not, however, be equal. For a crystal with unit cell dimensions  $c \gg a > b$ , and therefore reciprocal cell dimensions  $c^* \ll a^* < b^*$ , one might choose semidimensions  $\Delta\mathbf{h} = 2$ ,  $\Delta\mathbf{k} = 1$ ,  $\Delta\ell = 4$  and use a raster of  $(5 \times 3 \times 9) - 1$  points to sample local blocks of the reciprocal lattice that have edges of roughly equal length along  $a^*$ ,  $b^*$ , and  $c^*$ . Our program *locscl* chooses statistically optimum raster semidimensions by analyzing the global variation of the locally fitted scale factors and their error-of-fit uncertainties as the semidimensions  $\Delta\mathbf{h}$ ,  $\Delta\mathbf{k}$ , and  $\Delta\ell$  are iteratively varied in proportion to the unit cell dimensions  $a$ ,  $b$ , and  $c$ , so that the raster retains the shape of an approximately rhombic parallelepiped with edges parallel to the  $a^*$ ,  $b^*$ , and  $c^*$  axes of the reciprocal lattice. The *locscl* program also applies (21) and (27) through (30) to propagate the error-of-fit uncertainties of the local scale factors into the  $\sigma(|F|)$  and  $\sigma(|E|)$  values corresponding to the locally scaled  $|F|$  and  $|E|$  values.

## 9.2. DIFFERENCE STRUCTURE FACTOR NORMALIZATION

Among early efforts to exploit probabilistic phasing methods in protein crystallography were applications of the MULTAN program (see, e.g., Main, 1976, 1985) employing SIR (Wilson, 1978) or SAS (Mukherjee, Helliwell, and Main, 1989) difference-magnitude data. In connection with recent further work to develop stronger probabilistic methods for phasing difference magnitudes (Langs, Guo, and Hauptman, 1995; Smith, Nagar, Rini, Hauptman, and Blessing, 1997), we have developed a program *diffe* (Blessing, 1997c) that implements the following difference normalization procedures.

### 9.2.1. SIR Differences

In the SIR case the magnitude differences of interest are

$$\Delta = |F_{\text{Der}}| - |F_{\text{Nat}}|, \quad (70)$$

which, given the corresponding locally scaled  $|E|$  magnitudes, can be calculated as

$$\Delta = (\epsilon_h \sum_{a,\text{Der}} |f_a|^2)^{1/2} |E_{\text{Der}}| - (\epsilon_h \sum_{a,\text{Nat}} |f_a|^2)^{1/2} |E_{\text{Nat}}|. \quad (71)$$

We recall that for structure factors  $F = |F| \exp(i\phi)$  with  $|F| \leq \sum_a f_a$ , the Wilson distributions give the intensity expectation value  $\langle |F|^2 \rangle = \epsilon_h \sum_a f_a^2$ . Therefore, for difference structure factors  $F_\Delta = |\Delta| \exp(i\phi_{\text{Heavy}})$  with  $|\Delta| \leq |F_{\text{Heavy}}| \leq \sum_{a,\text{Heavy}} |f_a|$ , we expect squared SIR difference magnitudes with

$$\langle |\Delta|^2 \rangle \leq \langle |F_{\text{Heavy}}|^2 \rangle = \epsilon_h \sum_{a,\text{Heavy}} |f_a|^2 = \epsilon_h [(\sum_{a,\text{Der}} |f_a|^2) - (\sum_{a,\text{Nat}} |f_a|^2)]. \quad (72)$$

Thus, greatest-lower-bound estimates for SIR difference-E magnitudes can be calculated as

$$|E_\Delta| = \frac{|\sum_{a,\text{Der}} |f_a|^2|^{1/2} |E_{\text{Der}}| - |\sum_{a,\text{Nat}} |f_a|^2|^{1/2} |E_{\text{Nat}}|}{q[(\sum_{a,\text{Der}} |f_a|^2) - (\sum_{a,\text{Nat}} |f_a|^2)]^{1/2}}, \quad (73)$$

where, following (51),

$$q = q_0 \exp(q_1 s^2 + q_2 s^4), \quad \text{in which } s = (\sin \theta)/\lambda, \quad (74)$$

is a least-squares-fitted normalization scaling function that imposes the requirement  $\langle |E_\Delta|^2 \rangle = 1$  and is intended to empirically adjust for effects of: imperfect isomorphism of the derivative and native crystals; inaccurately known heavy-atom content due to multiple derivative sites and/or disordered partial site occupancies; and differences between the unit-cell distributions of mean-square atomic displacements in the heavy-atom substructure, the derivative crystal, and the native protein crystal.

The *diffe* program accepts user-supplied input cutoff values  $x_{\min}$ ,  $y_{\min}$ , and  $s_{\max}$  to limit the processing to data pairs with  $\min[|E_{\text{Nat}}|/\sigma(|E_{\text{Nat}}|), |E_{\text{Der}}|/\sigma(|E_{\text{Der}}|)] \geq x_{\min}$ ,  $[|E_{\text{Der}}| - |E_{\text{Nat}}|]/[\sigma^2(|E_{\text{Der}}|) + \sigma^2(|E_{\text{Nat}}|)]^{1/2} \geq y_{\min}$ , and  $s_h = (\sin \theta_h)/\lambda \leq s_{\max}$ , where typically  $x_{\min} = 3$ ,  $y_{\min} = 1$ , and  $s_{\max}$  is determined by inspection of a plot of the spherical shell averages  $\langle |E_{\Delta}|^2 \rangle$ , vs.  $\langle s \rangle$  from a preliminary run of the *diffe* program with unlimited  $s_{\max}$ . The purpose of the  $s_{\max}$  input cutoff is to prevent generating spuriously large  $|E_{\Delta}|$  values for high-resolution data pairs measured with large uncertainties due to the general fall-off of scattering intensity with increasing scattering angle. The program propagates the error-of-fit uncertainties of the difference normalization scaling parameters into  $\sigma(|E_{\Delta}|)$  values, and accepts a user-supplied output cutoff value  $z_{\min}$  so that  $|E_{\Delta}|$  values for which  $|E_{\Delta}|/\sigma(|E_{\Delta}|) < z_{\min}$ , where typically  $z_{\min} = 3$ , are rejected as too unreliable to be used in subsequent phasing calculations.

### 9.2.2. SAS differences

In the SAS case the magnitude differences of interest are

$$\Delta = |F_{+h}| - |F_{-h}|, \quad (75)$$

which, given the corresponding locally scaled  $|E|$  magnitudes, can be calculated as

$$\begin{aligned} \Delta &= (\epsilon_h \sum_a |f_a|^2)^{1/2} (|E_{+h}| - |E_{-h}|), \\ \Delta &= [\epsilon_h \sum_a (f_a^0 + f_a'')^2 + (f_a''')^2]^{1/2} (|E_{+h}| - |E_{-h}|). \end{aligned} \quad (76)$$

Then, for structure factors  $F_{\Delta} = |\Delta| \exp\{i[\Phi_{\Delta} + (\pi/2)]\}$  with  $|\Delta| \leq 2|F''| \leq 2\sum_a f_a''$  and with  $\Phi_{\Delta}$  representing the phase of the  $F_{\Delta}^0 + F''$  component of the total structure factor,  $F = F^0 + F' + F'' = F_{\text{non-}\Delta} + F_{\Delta}^0 + F'' + F''$ , we expect squared SAS difference magnitudes with

$$\langle |\Delta|^2 \rangle \leq 2\langle |F''|^2 \rangle = 2\epsilon_h \sum_a (f_a''')^2. \quad (77)$$

Thus, greatest-lower-bound estimates for SAS difference-E magnitudes can be calculated as

$$|E_{\Delta}| = \frac{(\sum_a |f_a|^2)^{1/2} ||E_{+h}| - |E_{-h}||}{2q[\sum_a (f_a''')^2]^{1/2}}, \quad (78)$$

where, again,  $q = q_0 \exp(q_1 s^2 + q_2 s^4)$  with  $s = (\sin \theta)/\lambda$  is a least-squares-fitted normalization scaling function that imposes the requirement  $\langle |E_{\Delta}|^2 \rangle = 1$  and, in the SAS case, is intended to empirically adjust for effects of: inaccurately known chemical composition of the unit cell; multiple sites and/or disordered partial site occupancies in the anomalously scattering substructure; inaccuracies in the values of

the anomalous scattering corrections  $f'$  and  $f''$  and in the assumption that they are independent of both the magnitude and direction of  $\mathbf{h}$ ; and differences between the unit-cell distributions of mean-square atomic displacements in the strongly anomalously scattering substructure and the structure overall.

The *diffe* program again accepts user-supplied input cutoffs  $x_{\min}$ ,  $y_{\min}$ , and  $s_{\max}$ , propagates the error-of-fit uncertainties of the normalization parameters, and accepts an output cutoff  $z_{\min}$ . The use of the data selection variables in the SAS case is directly analogous to their use in the SIR case described above.

## 10. Acknowledgements

Our own research related to the subjects of this chapter has been supported by grants from the US National Institutes of Health: at present grant no. GM46733, and in the past grants no. GM34073, HL32303, and DK19856. We are grateful for this support, and for helpful research discussions with our colleagues G. David Smith, Grant R. Moss, and Herbert A. Hauptman.

## 11. References

- Blessing, R.H. (1987). Data Reduction and Error Analysis for Accurate Single-Crystal Diffraction Intensities. *Crystallogr. Rev.* 1, 3-58
- Blessing, R.H. (1989). DREADD - data reduction and error analysis for single crystal diffractometer data. *J. Appl. Cryst.*, 22, 396-397.
- Blessing, R.H. (1995). An Empirical Correction for Absorption Anisotropy. *Acta Cryst.*, A51, 33-38.
- Blessing, R.H. (1997a). Outlier Treatment in Data Merging. *J. Appl. Cryst.* 30, in press.
- Blessing, R.H. (1997b). LOCSCL: a program to statistically optimize local scaling of SIR and SAS data. *J. Appl. Cryst.* 30, 176-177.
- Blessing, R.H. (1997c). Normalization of SIR and SAS Difference-Magnitudes for Phasing Heavy-Atom Substructures. *J. Appl. Cryst.* 30, submitted.
- Blessing, R.H., Guo, D.Y., and Langa, D.A. (1996). Statistical Expectation Value of the Debye-Waller Factor and  $|E(hk\ell)|$  Values for Macromolecular Crystals. *Acta Cryst.*, D52, 257-266.
- Blessing, R.H., and Langa, D.A. (1987). Data averaging with normal down-weighting of outliers. *J. Appl. Cryst.* 20, 427-428.
- Blessing, R.H., and Langa, D.A. (1988). *A Priori* Estimation of Scale and Overall Anisotropic Temperature Factors From the Patterson Origin Peak. *Acta Cryst.*, A44, 729-735.
- Bricogne, G. (1984). Maximum Entropy and the Foundations of Direct Methods. *Acta Cryst.* A40, 410-445.
- Bricogne, G. (1996). Fourier transforms in crystallography: theory, algorithms, and applications. In *International Tables for Crystallography*, Vol. B, edited by U. Shmueli, pp. 23-106. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Castleden, I.R., and Fortier, S. (1994). Intensity Statistics. I. *Acta Cryst.* A50, 9-17.
- Coppins, P. (1996). The Structure Factor. In *International Tables for Crystallography*, Vol. B, edited by U. Shmueli, pp. 10-22. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- French, S., and Wilson, K.S. (1978). On the Treatment of Negative Intensity Observations. *Acta Cryst.* A34, 517-525.

- Gewirth, D., Otwinowski, Z., and Minor, W. (1995). *The HKL Manual*, 4<sup>th</sup> ed. New Haven, Connecticut: Yale University, Department of Biophysics and Biochemistry.
- Giacovazzo, C. (1996). Direct Methods. In *International Tables for Crystallography*, Vol. B, edited by U. Shmueli, pp. 201-229. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Hamilton, W.C., Rollett, J.S., and Sparks, R.A. (1965). On the relative scaling of X-ray photographs. *Acta Cryst.* 18, 129-130.
- Harker, D. (1953). The Meaning of the Average of  $|F|^2$  for Large Values of the Interplanar Spacing. *Acta Cryst.* 6, 731-736.
- Harker, D., and Kasper, J.S. (1948). Phases of Fourier Coefficients Directly from Crystal Diffraction Data. *Acta Cryst.* 1, 70-75.
- Hauptman, H., and Karle, J. (1953). *Solution of the Phase Problem. I. The Centrosymmetric Crystal*. Am. Crystallogr. Assoc. Monograph No. 3. Dayton, Ohio: Polycrystal Book Service.
- Iwasaki, H., and Ito, T. (1977). Values of  $\epsilon$  for obtaining normalized structure factors. *Acta Cryst.* A33, 227-229.
- Johnson, C.K., and Levy, H.A. (1974). Thermal Motion Analysis Using Bragg Diffraction Data. In *International Tables for X-Ray Crystallography*, Vol. IV, edited by J.A. Ibers and W.C. Hamilton, pp. 311-336. Birmingham, England: The Kynoch Press.
- Langs, D.A., Guo, D.Y., and Hauptman, H.A. (1995). Use of 'Random-Atom' Phasing Methods to Determine Macromolecular Heavy-Atom Replacement Positions. *Acta Cryst.* D51, 1020-1024.
- Levy, H.A., Thiessen, W.E., and (in part) Brown, G.M. (1970). A Least-Squares Method for Converting Observed Intensities into Normalized Structure Factors. Am. Crystallogr. Assoc. Meeting, Tulane Univ., New Orleans, Louisiana, March 1970. Abstract No. B6.
- Luzatti, V. (1955). Sur l'Emploi des Méthodes Statistiques dans l'Étude de la Structure Cristalline des Protéines. *Acta Cryst.* 8, 795-806.
- Main, P. (1976). Recent Developments in the MULTAN System. The Use of Molecular Structure. In *Crystallographic Computing Techniques*, edited by F.R. Ahmed with K. Huml and B. Sedláček, pp. 97-105. Copenhagen: Munksgaard Publishers.
- Main, P. (1985). MULTAN - a program for the determination of crystal structures. In *Crystallographic Computing 3: Data Collection, Structure Determination, Proteins, and Databases*, edited by G.M. Sheldrick, C. Krüger, and R. Goddard, pp. 206-215. Oxford, England: Oxford University Press.
- Matthews, B.W. (1968). Solvent Content of Protein Crystals. *J. Mol. Biol.* 33, 491-497.
- Matthews, B.W., and Czerwinski, E.W. (1975). Local Scaling: A Method to Reduce Systematic Errors in Isomorphous Replacement and Anomalous Scattering Measurements. *Acta Cryst.* A31, 480-497.
- Mukherjee, A.K., Helliwell, J.R., and Main, P. (1989). The Use of MULTAN to Locate Positions of Anomalous Scatterers. *Acta Cryst.* A45, 715-718.
- Nielsen, K. (1975). Calculation of  $E$  Values by Means of the Origin Peak in the Patterson Function. *Acta Cryst.* A31, 762-763.
- Otwinowski, Z. (1993). Oscillation Data Reduction Program. In *Proceedings of the CCP4 Study Weekend: Data Collection and Processing*, 29-30 January 1993, edited by L. Sawyer, N. Isaacs, and S. Bailey, pp. 56-62. Daresbury, England: SERC Daresbury Laboratory.
- Rogers, D. (1965). Statistical Properties of Reciprocal Space. The Scaling of Intensities. In *Computing Methods in Crystallography*, edited by J.S. Rollett, pp. 117-132 and 133-148. Oxford, England: Pergamon Press.
- Rogers, D. (1980). Calculation of  $|E|$  Values. In *Theory and Practice of Direct Methods in Crystallography*, edited by M.F.C. Ladd and R.A. Palmer, pp. 82-92. New York: Plenum Press.
- Rossmann, M.G., and Arnold, E. (1996). Patterson and Molecular Replacement Techniques. In *International Tables for Crystallography*, Vol. B, edited by U. Shmueli, pp. 230-263. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Sands, D.E. (1996). Dual Bases in Crystallographic Computing. In *International Tables for Crystallography*, Vol. B, edited by U. Shmueli, pp. 331-344. Dordrecht, The Netherlands: Kluwer Academic Publishers.



- Shmueli, U. (1996). Reciprocal Space in Crystallography. In *International Tables for Crystallography*, Vol. B, edited by U. Shmueli, pp. 2-9. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Shmueli, U., and Wilson, A.J.C. (1996). Statistical Properties of the Weighted Reciprocal Lattice. In *International Tables for Crystallography*, Vol. B, edited by U. Shmueli, pp. 184-200. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Smith, G.D., Nagar, B., Rini, J.M., Hauptman, H.A., and Blessing, R.H. (1997). The Use of SnB to Determine an Anomalous Scattering Substructure. *Acta Cryst.* D53, submitted.
- Vijayan, M., and Ramaseshan, S. (1996). Isomorphous Replacement and Anomalous Scattering. In *International Tables for Crystallography*, Vol. B, edited by U. Shmueli, pp. 264-279. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Weckert, E., and Hümmer, K. (1997). Multiple-Beam X-ray Diffraction for Physical Determination of Reflection Phases and Its Applications. *Acta Cryst.* A53, 108-143.
- Weckert, E., Schwegle, W., and Hümmer, K. (1993). Direct phasing of macromolecular structures by three-beam diffraction. *Proc. Roy. Soc. London*, A442, 33-46.
- Wilson, A.J.C. (1942). Determination of Absolute from Relative X-Ray Intensity Data. *Nature*, 150, 152.
- Wilson, A.J.C. (1949). The Probability Distribution of X-Ray Intensities, *Acta Cryst.* 2, 318-320. See also the discussion in Shmueli and Wilson (1996).
- Wilson, K.S. (1978). The Application of MULTAN to the Analysis of Isomorphous Derivatives in Protein Crystallography. *Acta Cryst.* B34, 1599-1608.
- Yü, S.H. (1942). A New Synthesis of X-Ray Data for Crystal Analysis. *Nature*, 149, 638-639, 729. Determination of Absolute from Relative X-Ray Intensity Data. *Nature*, 150, 151-152.