equation (II.11*ab*) coincide. Then, in the matrices **H** and $\mathbf{M}^-(rs)$ also the great terms

$$(x^+_{pq}-y^+_{pq})^{-1},\ (x^+_{pq}-y^-_{mn})^{-1},\quad (x^-_{mn}-y^-_{mn})^{-1},\ (x^-_{mn}-y^+_{pq})^{-1} \qquad (A.2)$$

are to be taken into consideration. Since these additional large terms are outside the first row of the matrix $\mathbf{M}^-(rs)$, a simple consideration shows that the coincidence of the poles in (*A*.1) does not change the value of the quotient $\det \mathbf{M}^-(rs)/\det \mathbf{H}$, (10). Another situation appears if (6) is satisfied for two (or more) different pairs of indices (rs), *i.e.* when as well as (6)

$$\theta^+_{00}(\mathbf{k}) = \theta^-_{ab}(\mathbf{k}) + 2\pi j + \eta,\quad (ab) \neq (rs), \qquad (A.3)$$

also holds. In this case, additional large terms appear in the first rows of the matrices $\mathbf{M}^-(rs)$ and **H** and they change quotient (10). The conditions (6) and (*A*3) mean physically that either the Bragg reflection condition is satisfied for two wave vectors $\mathbf{K}^-_{rs}$ and $\mathbf{K}^-_{ab}$ or the incident wave is near the grazing reflection angle [see § 3(iii) of Litzman & Dub (1990)]. These cases were not considered in the present paper.

References

BRÜMMER, O., HÖCHE, H. R. & NIEBER, J. (1979). *Phys. Status Solidi A*, **53**, 565–570.

CATICHA, A. & CATICHA-ELLIS, S. (1982). *Phys. Rev. B*, **25**, 971–982.

KOHRA, K. & MATSUSHITA, T. (1972). *Z. Naturforsch. Teil A*, **27**, 484–487.

LITZMAN, O. (1986). *Acta Cryst.* A**42**, 552–559.

LITZMAN, O. & DUB, P. (1990). *Acta Cryst.* A**46**, 247–254.

LITZMAN, O. & RÓZSA, P. (1980). *Czech. J. Phys. (Engl. Transl.)*, B**30**, 816–826.

PINSKER, Z. G. (1978). *Dynamical Scattering of X-rays in Crystals*. New York: Springer.

ZACHARIASEN, W. H. (1946). *Theory of X-ray Diffraction in Crystals*. New York: Wiley.

Acta Cryst. (1990). A**46**, 900–912

Structure-Factor Probabilities for Related Structures

BY RANDY J. READ

Department of Medical Microbiology and Infectious Diseases, University of Alberta, Edmonton, Alberta, Canada T6G 2H7

(*Received* 17 *January* 1990; *accepted* 9 *May* 1990)

Abstract

Probability relationships between structure factors from related structures have allowed previously only for either differences in atomic scattering factors (isomorphous replacement case) or differences in atomic positions (coordinate error case). In the coordinate error case, only errors drawn from a single probability distribution have been considered, in spite of the fact that errors vary widely through models of macromolecular structures. It is shown that the probability relationships can be extended to cover more general cases. Either the atomic parameters or the reciprocal-space vectors may be chosen as the random variables to derive probability relationships. However, the relationships turn out to be very similar for either choice. The most intuitive is the expected electron-density formalism, which arises from considering the atomic parameters as random variables. In this case, the centroid of the structure-factor distribution is the Fourier transform of the expected electron-density function, which is obtained by smearing each atom over its possible positions. The centroid estimate has a phase different from, and more accurate than, that obtained from the unweighted atoms. The assumption that there is a sufficient number of independent errors allows the application of the central limit theorem. This gives a one- (centric case) or two-dimensional (non-centric) Gaussian distribution about the centroid estimate. The general probability expression reduces to those derived previously when the appropriate simplifying assumptions are made. The revised theory has implications for calculating more accurate phases and maps, optimizing molecular replacement models, refining structures, estimating coordinate errors and interpreting refined *B* factors.

1. Introduction

A model of a crystal structure always has errors in any parameters used to describe the structure: atomic coordinates, atomic scattering factors, thermal motion parameters, or even cell dimensions. In addition, the approximations of spherically symmetric atoms and of harmonic (or even isotropic) thermal motion will lead to small errors. In refining a structure, we attempt to minimize these errors as far as possible, but it is best to keep their existence in mind and to be aware of their effects.

0108-7673/90/110900-13$03.00

Errors in the model will lead to errors in the calculated structure factors. In principle, if we know the probability of various errors in the model we can deduce the probability of various errors in the calculated structure factor. Having an estimate of the probability distribution of the true structure factor can be useful in a number of circumstances. The centroid estimate (probability-weighted average, or expected value) of the structure factor will minimize the root-mean-square (r.m.s.) error in both the structure factor and in electron-density maps computed from it (Blow & Crick, 1959). To combine phase information from several sources (Rossmann & Blow, 1961; Hendrickson & Lattman, 1970), we need the probability distribution of the true phase. From the opposite point of view, we might hope to use the disagreement between the calculated and observed structure-factor amplitudes to draw inferences about the probability of errors in the model.

In deriving probability relationships, I will be considering the combination of differences in both atomic coordinates and scattering factors (§ 2). This is fairly general: differences in cell dimensions will show up as differences in fractional coordinates, and differences in B factor (including errors from the assumption of harmonic isotropic thermal motion) or atom type will show up as differences in scattering factor. Only the effect of errors in the measurements of the structure-factor amplitudes will be ignored. In this work, only the conditional distributions involving pairs of structure factors with the same *hkl* will be derived, not those involving higher orders such as triplets. The probability distributions will be shown to reduce to those derived previously, when the appropriate simplifying assumptions are made.

The revised theory has a number of implications (explored in § 3) for the practice of macromolecular crystallography. Numerical tests supporting the general probability distributions and their physical interpretation will be presented in § 4. Finally, a brief overview of the most important results will be given in § 5.

2. Theory

A model of the electron density in a crystal, often expressed as a set of atomic positions and thermal motion parameters, can be considered as one member of a related pair of structures. The model crystal has cell dimensions similar (within experimental error) to those of the true crystal, it belongs to the same space group, and it contains electron density that is, preferably, non-randomly related to the true electron density. Accordingly, probability relationships derived for atomic models also apply to isomorphous derivatives.

The strategy for obtaining structure-factor probability distributions will be the following. I will assume that the conditions of the central limit theorem apply, in other words, that there is a sufficient number of independent finite contributions to the difference between two structure factors, none of them dominating. The overall distribution of the difference, then, tends toward a Gaussian. The center of the Gaussian distribution is displaced by the sum of the expected values of the contributions to the difference. We need not be concerned with the form of the distribution for the individual contributions, only with their variances. The variance of the Gaussian is the sum of the individual variances. This strategy will give a conditional probability distribution, but other distributions involving the two structure factors can be obtained from this.

Table 1. *Definitions of terms and notation*

$\bar{x}$ = mean value of x

$\langle x \rangle$ = expected value, or probability-weighted average, of x

$\mathbf{F} = \sum_{j=1}^{N} \mathbf{f}_j \exp[2\pi i \mathbf{s} \cdot (\mathbf{r}_j + \Delta\mathbf{r}_j)]$,
where $\mathbf{s}$ is the reciprocal-space vector [$|\mathbf{s}| = 2(\sin\theta)/\lambda$], the $\mathbf{r}_j$ are atomic coordinates (in Å), and the $\Delta\mathbf{r}_j$ are positional difference vectors

$\Sigma_N = \sum_{j=1}^{P} |\mathbf{f}_j|^2 + \sum_{j=P+1}^{N} |\mathbf{f}_j|^2$
$= \Sigma_P + \Sigma_Q$, where the P atoms constitute the partial structure and the Q atoms the missing structure
$= \langle |\mathbf{F}|^2/\varepsilon \rangle$, where ε is a correction factor for the expected intensity in a reciprocal-lattice zone

$\mathbf{F}_P = \sum_{j=1}^{P} \mathbf{f}_j \exp(2\pi i \mathbf{s} \cdot \mathbf{r}_j)$

$\mathbf{G} = \sum_{j=1}^{N} \mathbf{g}_j \exp(2\pi i \mathbf{s} \cdot \mathbf{r}_j)$
$= |\mathbf{G}| \exp(i\alpha_G)$

$m = \langle \cos(\alpha_F - \alpha_G) \rangle$
= figure of merit

Consider a related pair of crystals $\mathfrak{F}$ (with Fourier transform mate **F**) and $\mathfrak{G}$ (Fourier transform mate **G**). Without loss of generality, the crystals can be considered to have the same number of atoms, since extra atoms in one crystal can be considered to have a zero scattering factor in the other. The matched atoms in the two structures have, in general, different scattering factors and positions that differ by a shift vector. For reasons that will become clear below, it is necessary to consider scattering-factor and coordinate differences simultaneously.

It will be most convenient to consider the Fourier transforms of the electron-density distributions for the two crystals, given in (1). (Unless otherwise specified, sums throughout are taken over all atoms in the unit cell. Some terms and notation are defined in Table 1.)

$$\mathbf{G} = \sum_j \mathbf{g}_j \exp(2\pi i \mathbf{s} \cdot \mathbf{r}_j), \qquad (1a)$$

$$\mathbf{F} = \sum_j \mathbf{f}_j \exp[2\pi i \mathbf{s} \cdot (\mathbf{r}_j + \Delta\mathbf{r}_j)]. \qquad (1b)$$

For generality, complex scattering factors **f** and **g** are assumed. **F**, **G**, **f** and **g** are all functions of the reciprocal-space vector **s**, but for clarity this

dependence is left implicit. The variables $\mathbf{s}$ and $\mathbf{r}_j$ are used instead of the more familiar $\mathbf{h}$ and $\mathbf{x}_j$ for two reasons: first, the coordinate errors $\Delta\mathbf{r}_j$ will then be expressed in ångström units; second, we will be interested in the continuous Fourier transform of the probability distribution of $\Delta\mathbf{r}_j$, more appropriately expressed in terms of $\mathbf{s}$ than $\mathbf{h}$.

The centroid of the distribution of $\mathbf{F}$ given $\mathbf{G}$ will be defined in terms of a complex multiplier, $\mathbf{D}$,

$$\langle \mathbf{F} \rangle = \mathbf{DG}, \tag{2}$$

$$\mathbf{D}(\mathbf{s}) = \langle \mathbf{F}/\mathbf{G} \rangle = \left\langle \left[\sum_j \mathbf{f}_j \exp(2\pi i \mathbf{s} \cdot \Delta\mathbf{r}_j) \times \exp(2\pi i \mathbf{s} \cdot \mathbf{r}_j) \right] \Big/ \mathbf{G} \right\rangle, \tag{3a}$$

or

$$\mathbf{D}(\mathbf{s}) = \langle \mathbf{FG}^*/\mathbf{GG}^* \rangle = \left\langle \left\{ \sum_j \sum_k \mathbf{f}_j \exp(2\pi i \mathbf{s} \cdot \Delta\mathbf{r}_j) \mathbf{g}_k^* \times \exp[2\pi i \mathbf{s} \cdot (\mathbf{r}_j - \mathbf{r}_k)] \right\} \Big/ |\mathbf{G}|^2 \right\rangle. \tag{3b}$$

Equation (3b) is more convenient under some circumstances, such as when there is a dependence between the parameters associated with two atoms. As shown explicitly here, $\mathbf{D}$ is in general a complex function of $\mathbf{s}$. The dependence is made explicit because, under circumstances to be discussed below, $\mathbf{D}$ can be either real or complex, and a function of either $\mathbf{s}$ or just resolution.

The variance of $\mathbf{DG}$ as an estimate of $\mathbf{F}$ is given in general form as

$$\langle |\mathbf{F} - \mathbf{DG}|^2 \rangle = \langle (\mathbf{F} - \mathbf{DG})[\mathbf{F}^* - (\mathbf{DG})^*] \rangle. \tag{4}$$

Further development of (3) and (4) requires the assignment of the random variables and some specification of their underlying probability distributions. Either reciprocal-space or real-space variables can be the random variables, depending on the circumstances.

(*a*) *Differences between atoms as random variables*

We can consider that $\mathbf{f}_j$ and $\Delta\mathbf{r}_j$ are the random variables. Such will be the case, for instance, when we have some *a priori* estimates of the probable errors in a molecular replacement model. The general expression for the variance is developed as follows:

$$\langle |\mathbf{F} - \mathbf{DG}|^2 \rangle = \left\langle \sum_j \sum_k \{\mathbf{f}_j \exp[2\pi i \mathbf{s} \cdot (\mathbf{r}_j + \Delta\mathbf{r}_j)] - \mathbf{Dg}_j \exp(2\pi i \mathbf{s} \cdot \mathbf{r}_j)\} \times \{\mathbf{f}_k^* \exp[-2\pi i \mathbf{s} \cdot (\mathbf{r}_k + \Delta\mathbf{r}_k)] - \mathbf{D}^* \mathbf{g}_k^* \exp(-2\pi i \mathbf{s} \cdot \mathbf{r}_k)\} \right\rangle. \tag{5}$$

The cross terms will tend to cancel for atoms that give independent contributions to the difference. However, the contributions of symmetry-related atoms are not independent, in fact are identical, for certain classes of reflections. The number of identical contributions arising from symmetry can be denoted by the expected intensity factor ε (see, for example, Stewart & Karle, 1976). If the remaining cross terms arise from atoms giving independent contributions, (5) will simplify as follows:

$$\langle |\mathbf{F} - \mathbf{DG}|^2 \rangle = \varepsilon \sigma_\Delta^2, \tag{6a}$$

where

$$\sigma_\Delta^2 = \sum_j \langle |\mathbf{f}_j \exp[2\pi i \mathbf{s} \cdot (\mathbf{r}_j + \Delta\mathbf{r}_j)] - \mathbf{Dg}_j \exp(2\pi i \mathbf{s} \cdot \mathbf{r}_j)|^2 \rangle = \sum_j \langle |\mathbf{f}_j \exp(2\pi i \mathbf{s} \cdot \Delta\mathbf{r}_j) - \mathbf{Dg}_j|^2 \rangle. \tag{6b}$$

Note that σ_Δ^2 is in general a function of $\mathbf{s}$. We will see below that, under certain conditions, it is a function only of resolution.

If there is a sufficient number of independent differences between $\mathfrak{F}$ and $\mathfrak{G}$, the distribution of $\mathbf{F}$ will be a Gaussian with variance $\varepsilon\sigma_\Delta^2$ about $\mathbf{DG}$. In the non-centric case, the variance is distributed in the complex plane, giving rise to the following conditional probability distribution:

$$p_N[\mathbf{F}; \mathbf{G}] = \frac{1}{\pi\varepsilon\sigma_\Delta^2} \exp\left(-\frac{|\mathbf{F} - \mathbf{DG}|^2}{\varepsilon\sigma_\Delta^2} \right). \tag{7}$$

Various other probability distributions can be obtained from this by standard manipulations. Examples of such manipulations can be found in Srinivasan & Parthasarathy (1976). One example is the conditional distribution of the phase difference, which can be obtained by changing variables, fixing $|\mathbf{F}|$ and renormalizing

$$p_N[\Delta\alpha; |\mathbf{F}|, |\mathbf{G}|] = \exp[X \cos(\Delta\alpha)]/2\pi I_0(X), \tag{8a}$$

where

$$X = 2|\mathbf{F}||\mathbf{DG}|/\varepsilon\sigma_\Delta^2. \tag{8b}$$

For centric reflections, the variance is distributed only in the magnitude, and the following conditional probabilities are obtained:

$$p_C[\mathbf{F}; \mathbf{G}] = (2\pi\varepsilon\sigma_\Delta^2)^{-1/2} \times \exp(-|\mathbf{F} - \mathbf{DG}|^2/2\varepsilon\sigma_\Delta^2), \tag{9}$$

$$p_C[\Delta\alpha; |\mathbf{F}|, |\mathbf{G}|] = \exp[(X/2) \cos(\Delta\alpha)]/2 \cosh(X/2), \tag{10}$$

where X is as defined in (8b).

These expressions have the same mathematical form as those derived previously for structure-factor probabilities. With the appropriate approximations,

outlined below, these equations reduce to the earlier probability distributions.

(i) *Luzzati's distribution.* Luzzati (1952) considered the case in which $\mathbf{f}_j = \mathbf{g}_j$ for all atoms, and the $\Delta\mathbf{r}_j$ are drawn independently from a single probability distribution, $p(\Delta\mathbf{r})$. Under these circumstances, $p(\Delta\mathbf{r})$ is independent of all other parameters and $(3a)$ can be rearranged to give

$$\begin{aligned}\mathbf{D}(\mathbf{s}) &= \langle \exp(2\pi i \mathbf{s}\,.\,\Delta\mathbf{r})\rangle \\ &\quad \times \sum_j \mathbf{g}_j \exp(2\pi i \mathbf{s}\,.\,\mathbf{r}_j)/\mathbf{G} \\ &= \langle \exp(2\pi i \mathbf{s}\,.\,\Delta\mathbf{r})\rangle \\ &= \int_{\substack{\text{all}\\ \text{space}}} p(\Delta\mathbf{r}) \exp(2\pi i \mathbf{s}\,.\,\Delta\mathbf{r})\, \mathrm{d}\Delta\mathbf{r}. \qquad (11)\end{aligned}$$

$\mathbf{D}$ is the Fourier transform of the probability distribution of $\Delta\mathbf{r}$, as noted by Luzzati. In the formulation presented here, $p(\Delta\mathbf{r})$ is not assumed to have any special symmetry so that $\mathbf{D}$ is in general complex.

The variance is obtained from $(6b)$ using the specified assumptions

$$\begin{aligned}\sigma_\Delta^2 &= \sum_j \langle |\mathbf{g}_j \exp(2\pi i \mathbf{s}\,.\,\Delta\mathbf{r}_j) - \mathbf{D}\mathbf{g}_j|^2\rangle \\ &= \langle |\exp(2\pi i \mathbf{s}\,.\,\Delta\mathbf{r}_j) - \mathbf{D}|^2\rangle \sum_j |\mathbf{g}_j|^2 \\ &= (1 - |\mathbf{D}|^2)\, \Sigma_N. \qquad (12)\end{aligned}$$

Luzzati considered a particular case in which $p(\Delta\mathbf{r})$ is a three-dimensional isotropic Gaussian. The Fourier transform of such a distribution is the real function of resolution given in (13):

$$D(|\mathbf{s}|) = \exp[-(2\pi^2/3)\langle|\Delta\mathbf{r}|^2\rangle|\mathbf{s}|^2]; \qquad (13a)$$

$$D(|\mathbf{s}|) = \exp[-(8\pi^2/3)\langle|\Delta\mathbf{r}|^2\rangle(\sin\theta/\lambda)^2]. \qquad (13b)$$

Note the correspondence to overall isotropic thermal motion, seen in most familiar form in $(13b)$. The analogy to thermal motion will be explored below in the section on the expected electron-density function.

In his Appendix 3, Luzzati (1952) expressed D in terms of σ^2, the mean-square displacement of the atom in the direction of the diffraction vector (or, since the distribution is spherically symmetric, in any direction). Luzzati's expression can be obtained from $(13a)$ by the substitution $\langle|\Delta\mathbf{r}|^2\rangle = 3\sigma^2$.

The correspondence to isotropic thermal motion has been obscured because D is usually expressed in terms of the mean absolute value of $\Delta\mathbf{r}$ rather than its mean-square value. For a spherically symmetric Gaussian distribution, $\langle|\Delta\mathbf{r}|\rangle^2 = (8/3\pi)\langle|\Delta\mathbf{r}|^2\rangle$. With the substitution in $(13a)$, Luzzati's (1952) equation (51) is obtained for D:

$$D(|\mathbf{s}|) = \exp[-(\pi^3/4)\langle|\Delta\mathbf{r}|\rangle^2|\mathbf{s}|^2]. \qquad (14)$$

(ii) *Individual distributions for each atom.* Often we know more about the distributions of the differences between two structures. For instance, molecular replacement models will be very close to the true structure in highly conserved regions such as the active site of an enzyme, but will differ much more in surface loops with low amino acid sequence homology.

The names of the atoms involved are irrelevant to the accuracy of the electron-density model. What matters is how close atom k in $\mathfrak{G}$ is likely to be to the nearest atom j in $\mathfrak{F}$. If this did not give a one-to-one matching (for instance, if one atom in the molecular replacement model were the nearest to several atoms in the true structure), we would have to consider $(3b)$. However, we will assume for now that a one-for-one matching is possible. Then in principle we could reorder the atoms and set $\mathbf{g}_j = \langle \mathbf{f}_j \exp(2\pi i \mathbf{s}\,.\,\Delta\mathbf{r}_j)\rangle$ ($\Delta\mathbf{r}_j$ being the distance to the nearest atom) in $(3a)$, so that $\mathbf{D} = 1$,

$$\langle\mathbf{F}\rangle = \mathbf{G} = \sum_j \langle \mathbf{f}_j \exp(2\pi i \mathbf{s}\,.\,\Delta\mathbf{r}_j)\rangle \exp(2\pi i \mathbf{s}\,.\,\mathbf{r}_j). \qquad (15)$$

If the only uncertainties are assumed to be in the positions of the atoms,

$$\langle\mathbf{F}\rangle = \sum_j \mathbf{d}_j \mathbf{f}_j \exp(2\pi i \mathbf{s}\,.\,\mathbf{r}_j), \qquad (16a)$$

where

$$\mathbf{d}_j(\mathbf{s}) = \int_{\substack{\text{all}\\ \text{space}}} p(\Delta\mathbf{r}_j) \exp(2\pi i \mathbf{s}\,.\,\Delta\mathbf{r}_j)\, \mathrm{d}\Delta\mathbf{r}_j. \qquad (16b)$$

If the errors are independent, $(6b)$ becomes

$$\begin{aligned}\sigma_\Delta^2 &= \sum_j \langle |\mathbf{f}_j \exp(2\pi i \mathbf{s}\,.\,\Delta\mathbf{r}_j) \\ &\quad - \langle \mathbf{f}_j \exp(2\pi i \mathbf{s}\,.\,\Delta\mathbf{r}_j)\rangle|^2\rangle \\ &= \sum_j |\mathbf{f}_j|^2 - |\langle \mathbf{f}_j \exp(2\pi i \mathbf{s}\,.\,\Delta\mathbf{r}_j)\rangle|^2, \qquad (17)\end{aligned}$$

and if the errors are only in the coordinates,

$$\sigma_\Delta^2 = \sum_j |\mathbf{f}_j|^2 (1 - |\mathbf{d}_j|^2). \qquad (18)$$

At this point it should be clear why it is necessary to consider the differences in scattering factor and in coordinates simultaneously. The difference between $\mathbf{f}_j$ and $\mathbf{g}_j$ could be considered an error, but, as just shown, the appropriate correlation between this 'error' and the effect of coordinate error in fact reduces the deviation between $\mathbf{F}$ and $\mathbf{G}$, improving the model of the electron density. One reason for allowing complex scattering factors is now also clear, as complex $\mathbf{g}_j$ is necessary to allow for coordinate-error distributions that lack a center of symmetry.

Complex scattering factors also allow scattering groups larger than single atoms. We have assumed that differences between the atomic coordinates are independent, but constraints of bonding and packing will lead to dependence among the parameters of neighboring atoms. However, the structure could then

be considered to be made up of an effectively smaller number of independent fragments. As long as this number is sufficiently large, the central limit theorem can still be invoked.

(iii) *The expected electron-density function.* The results obtained so far in reciprocal-space terms have a more intuitive interpretation in real space. From Parseval's theorem, it follows straightforwardly that the expected value of a Fourier transform is the Fourier transform of the expected value, since the expected value of a quantity minimizes its r.m.s. error. This is just a generalization of the result of Blow & Crick (1959). They showed that the r.m.s. error in an electron-density map is minimized by using, in the Fourier transform, structure factors that minimize the r.m.s. error in the complex plane. Given the more general statement, we can easily reverse the question addressed by Blow & Crick to obtain the following: the expected value of a structure factor is the Fourier transform of the expected value of the electron density.

Since the electron density is the sum of atomic densities, the expected density is the sum of the expected densities for each atom. For each atom, the expected density is the probability-weighted sum of its densities in all possible positions. In simple terms, the electron density for each atom is smeared out to represent the uncertainty in its position. More formally, the expected electron density of atom j is the convolution of its electron-density distribution with the probability distribution $p(\Delta\mathbf{r}_j)$. From the convolution theorem, this corresponds to multiplying the scattering factor by the Fourier transform of $p(\Delta\mathbf{r}_j)$, $\mathbf{d}_j$ [defined in (16b)].

From (16a) we see that the probability distribution of shift vectors plays the same role in the calculation of the expected structure factor as thermal motion plays in the calculation of conventional structure factors. In the expected electron density, atoms are smeared out over a range of possible positions; in real electron density, atoms are smeared out over a range of alternative positions. Because of this correspondence and the equivalence of the mathematics, we can take advantage of the extensive work that has been done on models of thermal motion.

An intuitive picture of the form of (16) and (17) can be given. When the atoms are smeared over their distribution of possible positions in forming the expected electron density, their effective scattering power is reduced and some of the scattering power is thus missing from the model. This missing scattering power can be treated as being distributed randomly, so that $\mathbf{F}$ has a Wilson (1949) distribution centered on $\langle\mathbf{F}\rangle$, the Fourier transform of the expected density. As with a standard Wilson distribution, a random distribution of scattering power through the entire unit cell is more restrictive than necessary. All that is required is that the fractional part of the dot product $\mathbf{s}.\mathbf{r}$ be distributed randomly over the range 0 to 1.

A number of special cases of the expected density formalism are of interest in certain circumstances. We have already seen how Luzzati's (1952) distribution arises, and a number of other cases will now be summarized briefly.

Case of a perfect but incomplete model. This case can be treated by assigning the atoms to two classes. For the included atoms, or the 'P' atoms, $p(\Delta\mathbf{r})$ is a δ function at the origin ($d_j = 1$, the Fourier transform of a δ function). For the missing atoms, or the 'Q' atoms, $p(\Delta\mathbf{r})$ is a uniform distribution ($d_j = 0$ in general, 1 for the reciprocal-lattice-origin term). Equations (16a) and (18) become

$$\langle\mathbf{F}\rangle = \mathbf{F}_P. \qquad (19a)$$

$$\sigma_\Delta^2 = \Sigma_Q. \qquad (19b)$$

In this case, (7) reduces to the conditional probability derived by Sim (1959) and (9) to that derived by Woolfson (1956). Note that if the position of none of the atoms is known, $P = 0$, so that $\langle\mathbf{F}\rangle = 0$ and $\sigma_\Delta^2 = \Sigma_N$. Then (7) and (9) reduce to Wilson's distributions (Wilson, 1949).

Case of uniform Gaussian errors, incomplete model. This has been treated by Srinivasan & Ramachandran (1965). For each of the P atoms, $p(\Delta\mathbf{r})$ is the same finite Gaussian distribution, and D is given by (13a). The Q atoms, again, have uniform distributions, for which $d_j = 0$ in general. Equations (16a) and (18) become

$$\langle\mathbf{F}\rangle = D\mathbf{F}_P. \qquad (20a)$$

$$\sigma_\Delta^2 = (1 - D^2)\Sigma_P + \Sigma_Q. \qquad (20b)$$

Case of individual Gaussian errors. Equations (16a) and (18) apply, the only modification being that d_j is the following real function of resolution:

$$d_j(|\mathbf{s}|) = \exp\left[-(2\pi^2/3)\langle|\Delta\mathbf{r}_j|^2\rangle|\mathbf{s}|^2\right]. \qquad (21)$$

In practice, $\langle\mathbf{F}\rangle$ is most easily computed by adding $(8\pi^2/3)\langle|\Delta\mathbf{r}_j|^2\rangle$ to the B factors of the atoms in the model, $\mathfrak{G}$.

In the previous two cases the centroid structure-factor estimate differed only by a scale factor from the structure factor calculated, without considering errors, from the atoms in the model. Now the atoms most likely to be in error contribute with the lowest weight to determining the centroid estimate and, hence, the phase. The relative weights vary with the resolution. Note that missing atoms can be handled by assigning infinitely broad Gaussian error distributions.

More complicated distributions. Just as with complicated thermal motion models, complicated error models could easily have too many parameters to be relevant, given the resolution of the available data. One can imagine anisotropic Gaussian distributions analogous to anisotropic thermal vibrations, but these might be useful only in special circumstances. For instance, models of proteins in crystals that exhibit anisotropic diffraction (Sheriff & Hendrickson, 1987) probably have anisotropic error distributions.

Distance geometry methods (*e.g.* Havel & Wüthrich, 1985) generate an ensemble of structures that satisfy distance restraints, for instance from two-dimensional NMR data. In this case, the expected electron density would be the ensemble average.

Of more general interest is the question of the probability distribution for missing atoms. These atoms will be excluded from the volume occupied by the included atoms. Their expected density should therefore be distributed only through the unoccupied volume, not through the entire unit cell. In effect, this is already being done when models for the disordered solvent are varied to optimize the agreement with observed structure factors (Phillips, 1980). For loops that have not yet been fitted, bonding constraints could reduce the accessible volume further. The expected density could be the mean from a number of possible conformations generated, for instance, by systematic search (Moult & James, 1986).

(*b*) *Reciprocal-space vector as random variable*

When, for example, a molecular replacement model is being constructed, there is conceptually an ensemble of possibilities for the true structure. However, the true structure is a particular choice from that ensemble and it becomes less appropriate to treat atomic parameters as random variables. But instead we can treat the reciprocal-space vector $\mathbf{s}$ for a set of structure factors as the random variable, averaging the effect of differences in atomic parameters over reciprocal space.

I will only deal here with scattering factors for spherical atoms (real functions of resolution) and coordinate shifts that are equally frequent in all directions. Then it is appropriate to average over a spherical shell in reciprocal space, starting with the definition of $\mathbf{D}$ in (3*b*). (An overall anisotropic distribution of coordinate differences could be dealt with by averaging over planes in reciprocal space orthogonal to the principal axes of the distribution.) D has the same value at $\mathbf{s}$ and $-\mathbf{s}$, so it is real valued. Finally, since the same value of D is assumed to apply for each $\mathbf{s}$ in the shell, independent of $|\mathbf{G}|$, the expected value of the ratio is the ratio of expected values. [The result in (22) can also be obtained by finding the value of D that minimizes $\langle|\mathbf{F}-D\mathbf{G}|^2\rangle$ over a shell of reciprocal space.]

$$\begin{aligned} D(|\mathbf{s}|) &= \langle \mathbf{F}\mathbf{G}^*/|\mathbf{G}|^2\rangle \\ &= \langle \mathbf{F}\mathbf{G}^*\rangle/\langle|\mathbf{G}|^2\rangle \\ &= \frac{1}{\langle|\mathbf{G}|^2\rangle}\sum_j\sum_k \int_{\text{shell}} p(\mathbf{s})\, f_j g_k \\ &\quad \times \exp\left[2\pi i \mathbf{s}\,.\,(\mathbf{r}_j+\Delta\mathbf{r}_j-\mathbf{r}_k)\right] \mathrm{d}\mathbf{s}. \end{aligned} \tag{22}$$

Since all reciprocal-space vectors of magnitude $|\mathbf{s}|$ are equally probable, $p(\mathbf{s})$ is spherically symmetric. The numerator, then, is a double sum of Fourier transforms of spherically symmetric functions. For such a function, the angular variables can be integrated out and the Fourier transform can be expressed in terms of the radial distribution function (James, 1948), in this case a one-dimensional δ function at $|\mathbf{s}|$,

$$\begin{aligned} D(|\mathbf{s}|) &= \frac{1}{\langle|\mathbf{G}|^2\rangle}\sum_j\sum_k \int_0^\infty \delta(|\mathbf{s}|) f_j g_k \\ &\quad \times \frac{\sin(2\pi s|\mathbf{r}_j+\Delta\mathbf{r}_j-\mathbf{r}_k|)}{2\pi s|\mathbf{r}_j+\Delta\mathbf{r}_j-\mathbf{r}_k|}\,\mathrm{d}s \\ &= \frac{1}{\langle|\mathbf{G}|^2\rangle}\sum_j\sum_k f_j g_k \\ &\quad \times \frac{\sin(2\pi|\mathbf{s}||\mathbf{r}_j+\Delta\mathbf{r}_j-\mathbf{r}_k|)}{2\pi|\mathbf{s}||\mathbf{r}_j+\Delta\mathbf{r}_j-\mathbf{r}_k|}. \end{aligned} \tag{23}$$

When the argument of the $\sin(x)/x$ function in (23) is greater than π (in other words, when the distance between atom j in $\mathfrak{F}$ and atom k in $\mathfrak{G}$ is greater than half the resolution), the contributions to the double sum will oscillate from negative to positive. These contributions will tend to cancel, so we could neglect them. We can approximate further by considering only terms $j=k$ in the numerator and denominator. (If the vectors $\mathbf{r}_j+\Delta\mathbf{r}_j-\mathbf{r}_k$ and $\mathbf{r}_j-\mathbf{r}_k$ have similar radial distributions, errors from the approximation will tend to cancel.)

$$D(|\mathbf{s}|) = \sum_j f_j g_j \frac{\sin(2\pi|\mathbf{s}||\Delta\mathbf{r}_j|)}{2\pi|\mathbf{s}||\Delta\mathbf{r}_j|} \Big/ \sum_j g_j^2. \tag{24}$$

Equation (24) can be used to demonstrate a result obtained by Luzzati (1952). With his assumptions, that $f_j = g_j$ and that the shift vectors are drawn from a single spherically symmetric distribution,

$$D(|\mathbf{s}|) = \frac{1}{N}\sum_j \frac{\sin(2\pi|\mathbf{s}||\Delta\mathbf{r}_j|)}{2\pi|\mathbf{s}||\Delta\mathbf{r}_j|}. \tag{25}$$

Given that the frequency of occurrence of values $|\Delta\mathbf{r}_j|$ will reflect the radial probability distribution, D approximates to the Fourier transform of $p(\Delta\mathbf{r})$. However, in most realistic cases, the size of the shifts will be correlated to the strength of scattering, so the assumptions leading to (25) will not be justified.

If the average is over a sufficient number of independent reflections, the central limit theorem can be invoked and the distribution of **F** about $D\mathbf{G}$ will tend toward a Gaussian. However, we must remember the expected intensity factor in (6*a*),

$$\sigma_\Delta^2 = \langle|\mathbf{F} - D\mathbf{G}|^2/\varepsilon\rangle. \tag{26}$$

Since $D = \langle\mathbf{FG}^*/|\mathbf{G}|^2\rangle$ [(3*b*)],

$$\begin{aligned}\sigma_\Delta^2 &= \langle|\mathbf{F}|^2/\varepsilon\rangle - D^2\langle|\mathbf{G}|^2/\varepsilon\rangle \\ &= \sum_j f_j^2 - D^2\sum_j g_j^2.\end{aligned} \tag{27}$$

Because of the symmetry between the reciprocal-space and real-space variables, the probability relationships are rather similar with either choice of random variable.

(*c*) *Relationships in terms of normalized structure factors*

Srinivasan & Ramachandran (1965) showed that the effects of missing atoms and coordinate errors are equivalent in terms of normalized structure factors. Since the probability relationships derived here have the same mathematical form, the effects of all the differences between two related structures will be equivalent. Another reason to consider normalized structure factors is that, even though **D** reflects the effects of coordinate errors, it also includes corrections for overall differences in scale factor and thermal motion. These will disappear with normalized structure factors. Srinivasan & Ramachandran (1965) define a single parameter σ_A that characterizes the probability distributions of normalized structure factors. The parameter that plays the same role in the more general distributions will be termed σ_E. Because **s** will be used as the random variable, I will only consider the case developed in the previous section. Therefore D is a real-valued function of $|\mathbf{s}|$.

First we define the normalized variables:

$$\mathbf{E}_F = \mathbf{F}/\langle|\mathbf{F}|^2\rangle^{1/2}, \tag{28a}$$

$$\mathbf{E}_G = \mathbf{G}/\langle|\mathbf{G}|^2\rangle^{1/2}. \tag{28b}$$

Then, incorporating the results of (22),

$$\begin{aligned}\langle\mathbf{E}_F\rangle &= D\mathbf{G}/\langle|\mathbf{F}|^2\rangle^{1/2} \\ &= \frac{\langle\mathbf{FG}^*\rangle\mathbf{G}}{\langle|\mathbf{F}|^2\rangle^{1/2}\langle|\mathbf{G}|^2\rangle} \\ &= \sigma_E\mathbf{E}_G,\end{aligned} \tag{29a}$$

where

$$\begin{aligned}\sigma_E &= \frac{\langle\mathbf{FG}^*\rangle}{(\langle|\mathbf{F}|^2\rangle\langle|\mathbf{G}|^2\rangle)^{1/2}} \\ &= \frac{\langle\mathbf{E}_F\mathbf{E}_G^*\rangle}{(\langle|\mathbf{E}_F|^2\rangle\langle|\mathbf{E}_G|^2\rangle)^{1/2}} \\ &\cong \langle\mathbf{E}_F\mathbf{E}_G^*\rangle,\end{aligned} \tag{29b}$$

$$\begin{aligned}\langle|\mathbf{E}_F - \sigma_E\mathbf{E}_G|^2\rangle &= \langle|\mathbf{E}_F|^2\rangle - \sigma_E^2\langle|\mathbf{E}_G|^2\rangle \\ &\cong 1 - \sigma_E^2.\end{aligned} \tag{30}$$

The approximate equalities of (29*b*) and (30) are exact if the normalization is carried out such that $\langle|\mathbf{E}|^2\rangle = 1$.

3. Practical implications

(*a*) *Refinement*

In the ideal case, refinement will minimize the differences between the observed and calculated structure factors in the complex plane. For various reasons (non-linearity of the refinement problem, geometrical restraints), atoms are not necessarily shifted to their correct positions by refinement. It is commonly seen that atomic B factors become inflated for atoms in the incorrect positions. The inflation of the B factors can be related numerically to the size of the coordinate error. This will be shown by determining the scattering factor for each atom that would minimize the error, averaged over shells in reciprocal space. (Because of the averaging over shells, the imaginary terms disappear.)

$$\begin{aligned}\Psi &= \langle|\mathbf{F} - \mathbf{G}|^2/\varepsilon\rangle \\ &= \Big\langle\sum_k\sum_l \{f_k f_l \cos[2\pi\mathbf{s}\,.\,(\mathbf{r}_k + \Delta\mathbf{r}_k - \mathbf{r}_l - \Delta\mathbf{r}_l)] \\ &\quad - 2f_k g_l \cos[2\pi\mathbf{s}\,.\,(\mathbf{r}_k + \Delta\mathbf{r}_k - \mathbf{r}_l)] \\ &\quad + g_k g_l \cos[2\pi\mathbf{s}\,.\,(\mathbf{r}_k - \mathbf{r}_l)]\}\Big\rangle,\end{aligned} \tag{31}$$

$$\begin{aligned}\partial\Psi/\partial g_j &= \Big\langle\sum_k 2g_k \cos[2\pi\mathbf{s}\,.\,(\mathbf{r}_k - \mathbf{r}_j)] \\ &\quad - 2f_k \cos[2\pi\mathbf{s}\,.\,(\mathbf{r}_k + \Delta\mathbf{r}_k - \mathbf{r}_j)]\Big\rangle = 0,\end{aligned}$$

so

$$\begin{aligned}g_j &= \langle f_j \cos(2\pi\mathbf{s}\,.\,\Delta\mathbf{r}_j)\rangle \\ &\quad + \Big\langle\sum_{k\neq j} f_k \cos[2\pi\mathbf{s}\,.\,(\mathbf{r}_k + \Delta\mathbf{r}_k - \mathbf{r}_j)] \\ &\quad - g_k \cos[2\pi\mathbf{s}\,.\,(\mathbf{r}_k - \mathbf{r}_j)]\Big\rangle \\ &= f_j\frac{\sin(2\pi|\mathbf{s}||\Delta\mathbf{r}_j|)}{2\pi|\mathbf{s}||\Delta\mathbf{r}_j|} \\ &\quad + \sum_{k\neq j} f_k\frac{\sin(2\pi|\mathbf{s}||\mathbf{r}_k + \Delta\mathbf{r}_k - \mathbf{r}_j|)}{2\pi|\mathbf{s}||\mathbf{r}_k + \Delta\mathbf{r}_k - \mathbf{r}_j|} \\ &\quad - g_k\frac{\sin(2\pi|\mathbf{s}||\mathbf{r}_k - \mathbf{r}_j|)}{2\pi|\mathbf{s}||\mathbf{r}_k - \mathbf{r}_j|}.\end{aligned} \tag{32}$$

The expected values in (32) were evaluated as in (23) above. The physical interpretation of (32) is that the

optimal scattering factor for an atom will allow it to account both for the density of the nearest atom and for the unaccounted density of neighboring atoms. Ignoring the terms $k \neq j$, (32) indicates that, averaging over a sphere in reciprocal space, an error $\Delta\mathbf{r}_j$ is best accounted for by smearing the atom over the surface of a sphere with radius $|\Delta\mathbf{r}_j|$. In terms of the expected density formalism, this corresponds to knowing the size, but not the direction, of the coordinate shift.

The $\sin(x)/x$ function in the leading term of (32) can be modeled fairly well as isotropic thermal motion, as long as $|\Delta\mathbf{r}_j|$ is small compared to the resolution (up to about $|\Delta\mathbf{r}_j||\mathbf{s}| = 0{\cdot}33$),

$$\frac{\sin(2\pi|\mathbf{s}||\Delta\mathbf{r}_j|)}{2\pi|\mathbf{s}||\Delta\mathbf{r}_j|} \cong \exp\left(-\frac{2\pi^2}{3}|\Delta\mathbf{r}_j|^2|\mathbf{s}|^2\right). \quad (33)$$

This means, for example, that the resolution-dependent effect of a 1 Å error can be modeled as an increase of $(8\pi^2/3)$ Å^2 in the B factor, up to about 3 Å resolution. The functions in (33) are compared in Fig. 1.

Both the 'real' B factor and the 'error' B factor minimize the structure-factor error. They are intimately intertwined and one cannot hope to separate them entirely. Some workers, recognizing that B factors are error sinks, avoid refining them until late in refinement. Because the expected electron-density model will give more accurate phases, this might not be the best approach. The potential for improvement in an electron-density map that comes from the increased phase accuracy will be demonstrated below.

It is desirable to obtain an expected density model, since this gives the best estimate of the phases. One might question to what extent this occurs in least-squares refinement. The least-squares refinement of protein structures commonly minimizes the residual $[w(|\mathbf{F}| - |\mathbf{G}|)^2]$. The assumption is that refinement of the difference between the amplitudes should minimize the difference in the complex plane. As a number of authors (Wilson, 1976; Silva & Rossmann, 1985) have discussed, this corresponds to the assumption that there is no phase error. As phase errors increase, or as the model errors become larger relative to the resolution, this assumption becomes less tenable. A feature of the expected density model is that the scattering power is reduced, so that the mean value of the amplitudes is reduced. However, least-squares refinement is based on the assumption that $|\mathbf{G}|$ has a Gaussian distribution centered on $|\mathbf{F}|$, so least squares will tend to inhibit the reduction of scattering power. In the most extreme case, a completely random model, the best estimate of $\mathbf{F}$ is 0; this clearly would not minimize the refinement residual.

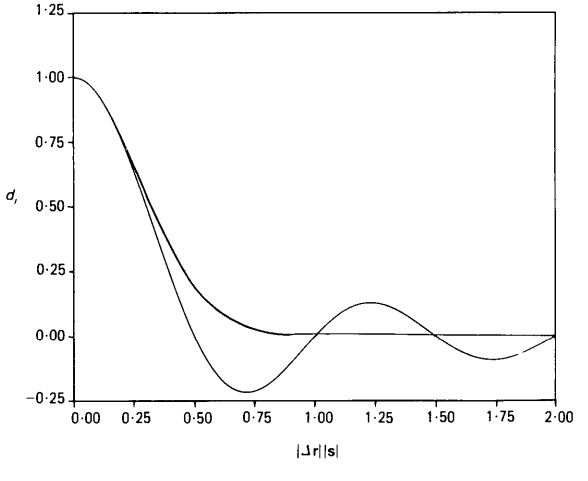

Fig. 1. Comparison of the functions $\sin(2\pi|\mathbf{s}||\Delta\mathbf{r}|)/2\pi|\mathbf{s}||\Delta\mathbf{r}|$ (thick line) and $\exp[-(2\pi^2/3)|\Delta\mathbf{r}|^2|\mathbf{s}|^2)$ (thin line).

When the assumption of Gaussian errors breaks down in a minimization problem, it is advisable to go back to first principles and apply maximum-likelihood methods, using more-accurate probability distributions, such as those derived here. [See, for example, the discussion of maximum likelihood by Mitra, Ahmed & Das Gupta (1985).] Such methods will be investigated in future work.

(*b*) *Molecular replacement*

In molecular replacement methods (Rossmann, 1972), the accuracy of the model is of great importance to success. As the expected electron density is more accurate than the unweighted density, it should be valuable in making molecular replacement structure solutions more straightforward, especially in marginal cases.

Molecular replacement models are commonly improved by editing out the portions expected to deviate most widely. This is a rather extreme and arbitrary action, which corresponds to assuming that the deleted atoms of the model bear no resemblance to any part of the new protein structure. Clearly, if reasonable estimates could be made for the probable errors, B-factor weighting would provide a much better model. Currently much effort is being expended in knowledge-based modeling of protein structures (*e.g.* Sutcliffe, Haneef, Carney & Blundell, 1987). Estimates of the reliability of the different parts of the resulting models would be extremely useful, and could be developed as a side product of the database investigations. Some information along these lines has already been obtained; the r.m.s. deviation of backbone atoms is a function of the degree of sequence homology (Chothia & Lesk, 1986; Hubbard & Blundell, 1987).

Some proteins, such as immunoglobulin Fab fragments, have domains with variable hinge angles. In constructing an expected electron-density model, the uncertainty could be modeled by anisotropic distributions, or by using the ensemble average from a number of possible hinge angles. Similar ideas have been

shown useful in practice. S. J. Remington (personal communication) has used the average density from several possible models for molecular replacement, obtaining better results than with any single model. Otwinowski *et al.* (1988) used isotropic B factors to model uncertainty for atoms in one domain of a molecular-replacement model.

(c) *Estimation of coordinate error*

What is estimated as σ_A by the method of Read (1986) is more correctly the parameter σ_E defined in this paper. It plays the same mathematical role as σ_A, but has a different physical interpretation. Under certain assumptions, $\sigma_E = \sigma_A$, and a plot of $\ln(\sigma_A)$ *vs* $|\mathbf{s}|^2$ has a slope proportional to the mean-square coordinate error of the atoms included in the model (Read, 1986). These assumptions are rarely satisfied in real cases. Each assumption will be stated, and the effect of violating it will be examined.

(1) *In the expansion of* (29*b*) *to double sums over atoms, only the terms* $j = k$ *are significant.* Then,

$$\sigma_E = \frac{\sum_j \langle f_j g_j \exp(2\pi i \mathbf{s} \cdot \Delta\mathbf{r}_j)\rangle}{\left[\sum_j f_j^2 \sum_j g_j^2\right]^{1/2}}, \qquad (34a)$$

where

$$\langle f_j g_j \exp(2\pi i \mathbf{s} \cdot \Delta\mathbf{r}_j)\rangle = f_j g_j \frac{\sin(2\pi|\mathbf{s}||\Delta\mathbf{r}_j|)}{2\pi|\mathbf{s}||\Delta\mathbf{r}_j|}. \qquad (34b)$$

For errors that are small compared to the average distance between neighboring atoms, this is probably reasonable. But, for example, if an atom in the model were halfway between two atoms in the true structure, at least two terms of the double sum would be significant. One conclusion is that, for atoms in the tightly packed interior of a protein molecule, errors greater than the van der Waals radius do not have much meaning.

(2) *The scattering factor* g_j *is either zero (missing atom) or differs from* f_j *by an overall resolution-dependent scale factor.* Consider a complete model with no coordinate errors but with errors in the individual B factors. σ_E would reduce to the parameter α defined by Hauptman (1982) for the isomorphous replacement case. At higher resolution, the errors in scattering factor would increase and σ_E would decrease, mimicking the effects of coordinate error.

(3) *The missing atoms are selected randomly, independent of* f_j. Consider a perfect partial model. If the missing atoms tend to have higher B factors than the included atoms, as would often be the case, the relative scattering power of the missing structure will decrease with resolution. Then σ_E will increase with resolution, giving an indicated negative mean-square error. By the same token, coordinate errors would be partially masked in a partial structure with errors. However, if only disordered solvent were missing, it might be sufficient to ignore the low-resolution data to which the disordered solvent atoms contribute. This is relevant to the end of refinement, when only the least-well-ordered solvent is still missing from the model.

(4) *The coordinate errors are drawn independently from a single isotropic Gaussian distribution.* The assumption that the distribution is Gaussian is less important than that it is isotropic. If all the other assumptions are satisfied, σ_E is the Fourier transform of $p(\Delta\mathbf{r})$. In principle, then, $p(\Delta\mathbf{r})$ can be obtained by taking the Fourier transform of σ_E. I have already argued that it is not necessary that the errors for the atoms be independent, as long as the structure can be considered to be made up of a sufficiently large number of fragments with independent errors.

(5) *For the atoms included in the model, there is no correlation between* $f_j g_j$ *and* $\Delta\mathbf{r}_j$. This is one of the most questionable assumptions. To begin with, there is usually a correlation between B factors and errors. In a molecular replacement model, the largest errors are usually in surface regions, which have the highest B factors. Atoms with high thermal motion have poorly defined density and are difficult to fit or to refine accurately.

Even if there were no correlation of $f_j g_j$ with coordinate error, a correlation would develop from refinement of the atomic B factors. To see the relevance of B-factor refinement, consider a structure that satisfies Luzzati's assumptions. Before any estimates of the individual coordinate errors are available, σ_E reduces to the expression for D in (25), which approximates the Fourier transform of the error distribution. After B-factor refinement, the individual B factors will reflect the size of the coordinate errors, the increase in B (in the ideal case) being approximately equivalent to the $\sin(x)/x$ term in (34). In effect, refinement takes atoms from a single error class and separates them into a large number of classes. Substituting $g_j \cong f_j \sin(2\pi|\mathbf{s}||\Delta\mathbf{r}_j|)/(2\pi|\mathbf{s}||\Delta\mathbf{r}_j|)$ into (34) gives the following:

$$\sigma_E \cong \left\{\frac{1}{\sum_j f_j^2}\sum_j\left[\frac{f_j \sin(2\pi|\mathbf{s}||\Delta\mathbf{r}_j|)}{2\pi|\mathbf{s}||\Delta\mathbf{r}_j|}\right]^2\right\}^{1/2}$$

$$= \left\{\frac{1}{N}\sum_j\left[\frac{\sin(2\pi|\mathbf{s}||\Delta\mathbf{r}_j|)}{2\pi|\mathbf{s}||\Delta\mathbf{r}_j|}\right]^2\right\}^{1/2}. \qquad (35)$$

Before B-factor refinement, σ_E is the average effect of all the coordinate errors. In contrast, after refinement σ_E is more like a r.m.s., not an average. It is no longer the Fourier transform of the coordinate-error distribution. Therefore, the Luzzati plot and the σ_A plot will not necessarily give meaningful answers once the contributions of the atoms to the calculated structure factors have been weighted according to expected positional error.

The effect of inflated B factors will, however, be much less significant at the end of a refinement. For example, an error of 0·2 Å corresponds to an increase of only about 1 Å^2 in the B factor. Another way to look at this is to consider that the $\sin(x)/x$ terms in (35) will all be close to 1. For numbers close to 1, the mean and the r.m.s. are nearly equal.

In principle, new methods to estimate coordinate error could be developed, starting from expressions such as (34). One difficulty would be that the combinations of error and scattering factor that give the variation of σ_E with resolution will not be unique. Some mathematical form would have to be assumed for the joint distribution $p(f_j g_j, \Delta \mathbf{r}_j)$. It could be assumed, for instance, that the dependence of error on scattering power is of the form derived by Cruickshank (1949), which would only be reasonable near the end of refinement. Furthermore, it would have to be assumed that the individual B factors are increased by an amount related to the coordinate error.

Further work will be required to resolve these questions. In the meantime, one should be aware that tools such as the σ_A or Luzzati plots suffer from a number of systematic errors. They should be used for comparative, rather than absolute, measures of coordinate error.

4. Numerical tests

(*a*) *Structure-factor and phase accuracy*

The structure factor computed from the expected electron density should be a more-accurate estimate of the true structure factor than one computed without a consideration of errors. The test case will be a molecular replacement model. Such a case has been chosen both for its relevance and because it violates a number of the assumptions of previous work in this area.

Bovine trypsin (BT; Chambers & Stroud, 1979) is about 33% identical in sequence to *Streptomyces griseus* trypsin (SGT), and was used as a molecular replacement model for the solution of that structure (Read & James, 1988). The two molecules are about the same size, but where the sequences are different there is no one-to-one correspondence between pairs of atoms. In determining the size of coordinate shifts (or 'errors') that relate atoms of the two structures, then, it is necessary first to decide with which SGT atom each BT atom should be paired. For the purpose of this test, each BT atom is paired with the SGT atom nearest to it in the overlapped structures. The distance is assigned as the error for that atom of BT as a model for SGT. This rule can result in several atoms of BT being paired with a single atom of SGT, but that occurs rarely and only in regions that are very different in the two structures. The individual error estimates obtained in this way are unrealistically precise, but serve to test the theory. A second, less precise, set of estimates is obtained by the following averaging procedure. For each residue, the error assigned for the main-chain atoms is the r.m.s. error in a five-residue window centered on that residue. Similarly, the error for the side-chain atoms is the r.m.s. value for the side chains in a five-residue window.

Four sets of structure factors (**G**) were computed from the BT molecular replacement model for SGT: (1) unweighted; (2) weighted individually using (21) for d_j (B-factor weighting); (3) weighted individually using $\sin(x)/x$ from the leading term of (32) for d_j [$\sin(x)/x$ weighting]; (4) weighted by the r.m.s. value in the five-residue window, using (21) for d_j (smoothed B-factor weighting). Apart from the error

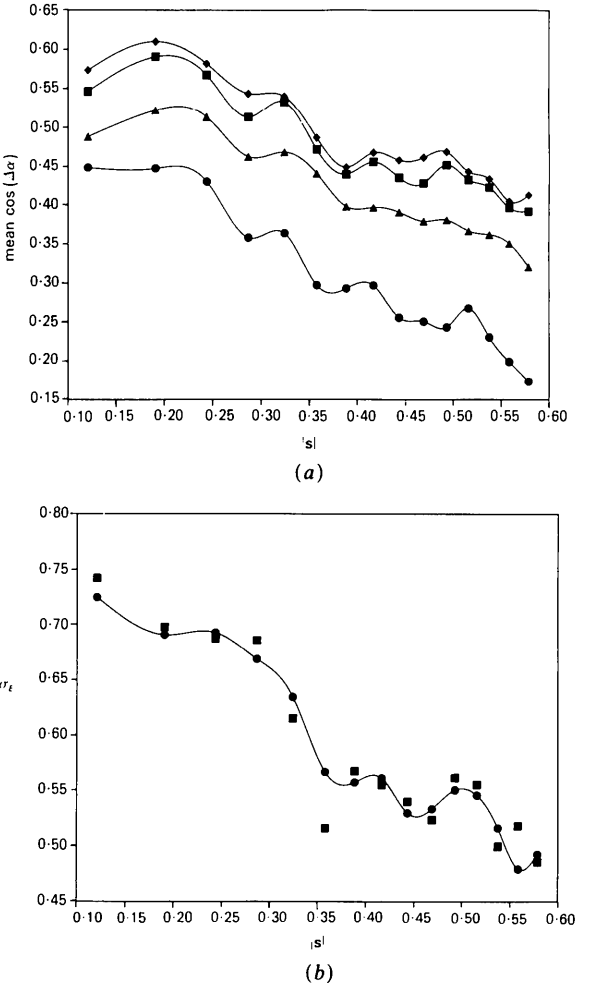

Fig. 2. (*a*) Improvement of phase accuracy of the BT model for SGT with weighting of the atomic contributions. The mean cosine of the phase difference is shown as a function of resolution. Circles correspond to the unweighted model, squares to the model with B-factor weighting, triangles to the model with smoothed B-factor weighting and diamonds to the model with $\sin(x)/x$ weighting. (*b*) Comparison of σ_E values estimated as a function of resolution by the method of Read (1986), indicated by circles, or by (36), indicated by squares.

Table 2. *Agreement measures for test data sets*

Weighting	R factor*	R.m.s. ($\|\mathbf{F}\|-\|\mathbf{G}\|$)	Intensity correlation†	Mean $\cos(\Delta\alpha)$
Unweighted	0·553	249·8	0·608	0·276
Individual B	0·540	229·1	0·666	0·453
Smoothed B	0·612	254·0	0·632	0·397
$\sin(x)/x$	0·539	227·5	0·672	0·469

* $R = \sum \|\|\mathbf{F}\|-\|\mathbf{G}\|\| / \sum |\mathbf{F}|$.

† $r = \dfrac{\sum (|\mathbf{F}|^2 - \overline{|\mathbf{F}|^2})(|\mathbf{G}|^2 - \overline{|\mathbf{G}|^2})}{[\sum (|\mathbf{F}|^2 - \overline{|\mathbf{F}|^2})^2 \sum (|\mathbf{G}|^2 - \overline{|\mathbf{G}|^2})^2]^{1/2}}$.

weighting, only an overall B factor of 17·3 Å^2 was applied to the BT model. Refined B factors were used to compute **F** from the final SGT model.

Fig. 2(a) shows that the phases obtained from any of the weighted models are much more accurate than those obtained from the unweighted model. As one might expect from Fig. 1, there is only a small improvement in going from B-factor weighting to $\sin(x)/x$ weighting. The results obtained using smoothed B-factor weighting show that an improvement in phase accuracy can be expected even when the information about the size of coordinate errors is imprecise. Table 2 gives some measures of overall structure-factor agreement for the four models. Note that the coefficient of correlation between $|\mathbf{F}|^2$ and $|\mathbf{G}|^2$ is a better indicator of phase accuracy than the more customary measures of agreement.

In Fig. 2(b), one set of values of σ_E is estimated by the method of Read (1986), a method based on the probability distributions of (7) and (9). These are compared to the 'true' values, obtained by using means over resolution shells for the expected values in (29b).

$$\sigma_E = \sum_{\text{shell}} \mathbf{FG}^* \Big/ \left(\sum_{\text{shell}} |\mathbf{F}|^2 \sum_{\text{shell}} |\mathbf{G}|^2 \right)^{1/2}. \qquad (36)$$

For this test, **G** is computed with B-factor weighting. The good agreement between the two sets of σ_E values suggests that the distribution of **F** about $\langle \mathbf{F} \rangle$ is consistent with (7) and (9). In addition, the mean figures of merit predicted for the four sets of structure factors agree equally well with the mean cosines of the phase differences in Fig. 2(b) (results not shown).

(b) *Electron-density maps*

From the point of view of the practising crystallographer, what is most important is whether the improvement in phase accuracy leads to a noticeable improvement in the interpretability of electron density maps. Fig. 3 shows a comparison between two SGT maps, one phased by the unweighted BT model and the other phased by the BT model weighted with individual B factors. There is a marked improvement in the connectivity of the model, as well as a marked reduction in model bias.

(c) *Estimation of coordinate error*

The data in Fig. 2 demonstrate that the estimation of σ_E is reliable, and that the mathematical form of the probability expressions is reasonable. What this tells us is that the various sources of error in the calculated structure factor combine to give a Gaussian distribution. However, to hope to use the variation of σ_E with resolution for estimating coordinate error, we must understand the physical significance of σ_E. This requires determining which approximations in the theory can be justified.

Fig. 4 shows two plots of $\ln(\sigma_E)$ *vs* $|\mathbf{s}|^2$, for the unweighted and B-factor-weighted cases. If σ_E is interpreted as σ_A, unreasonable estimates of r.m.s. coordinate error and model completeness (Σ_P/Σ_N) are obtained. The BT model has about the same number of atoms as SGT, and the r.m.s. distance to the nearest atom is 1·49 Å. In both cases, the line that is predicted for an overall Gaussian error of 1·49 Å falls off much too steeply. The least-squares lines indicate, before B weighting, that the r.m.s. error is 0·61 Å and $\Sigma_P/\Sigma_N = 0{\cdot}34$, and, after B weighting, that the r.m.s. error is 0·42 Å and $\Sigma_P/\Sigma_N = 0{\cdot}50$. (This illustrates the potential change in indicated coordinate error from the refinement of B factors.) For the unweighted data, the problem is mostly with the assumption of a Gaussian error distribution; the curve derived from the Fourier transform of the actual error distribution [$\sigma_E = D$ in (25)] agrees fairly well with the higher-resolution data. However, only the curves derived from (34) agree well both before and after B weighting. In this test case, therefore, the effect of differences in scattering factor is small compared to the effect of coordinate error before, but not after, B weighting.

The curves derived from (25) and (34) overestimate σ_E significantly at resolutions below 6 Å in this test. The assumptions must therefore be less valid at low resolution. There are a number of possibilities. More of the cross terms discarded from (29b) are significant at low resolution. The atoms with the largest errors, which contribute to σ_E primarily at low resolution, are on the surface of the protein; the distribution of atoms around them is far from spherically symmetric. (However, in a real crystal there would be at least poorly ordered solvent, not a vacuum.) Also, at lower

resolution it is more difficult to satisfy the requirement that the fractional part of $\mathbf{s} \cdot \mathbf{r}$ be randomly distributed over the range 0 to 1. Nonetheless, the agreement is reasonable for most of the data, even in this rather extreme test case. This shows that the assumptions leading to (34) are fairly robust.

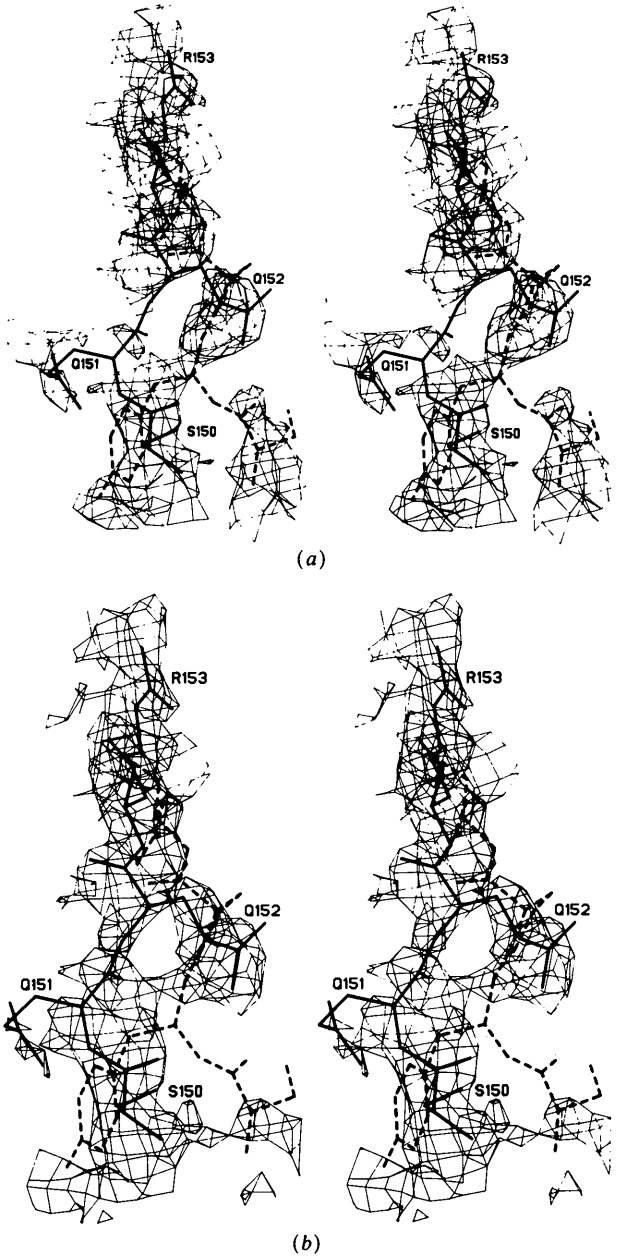

Fig. 3. Comparison of two electron-density maps computed with coefficients $(2m|\mathbf{F}| - D|\mathbf{G}|)\exp(i\alpha_G)$ (Read, 1986). In both maps, $|\mathbf{F}|$ is calculated from the final refined model of SGT, shown in solid lines. $\mathbf{G}$ is calculated from the BT molecular replacement model, shown in dashed lines, either with no weighting (*a*) or with individual *B*-factor weights (*b*). The contours are at 1·2 times the r.m.s. value of the map, 0·40 e Å^{-3} for (*a*) and 0·46 e Å^{-3} for (*b*). For clarity, only contours within 1·7 Å of an atom in the figure are shown.

5. Concluding remarks

Most differences between two related crystals can be considered to reside in either the coordinates of the atoms or in their scattering factors. The effect of these differences on the probability distributions of the structure factors has been derived using two choices of random variable.

It is appropriate to consider the atomic parameters to be random variables when there is *a priori* knowledge of the probability of differences between the structures. For instance, we know that molecular replacement models will be more reliable in some

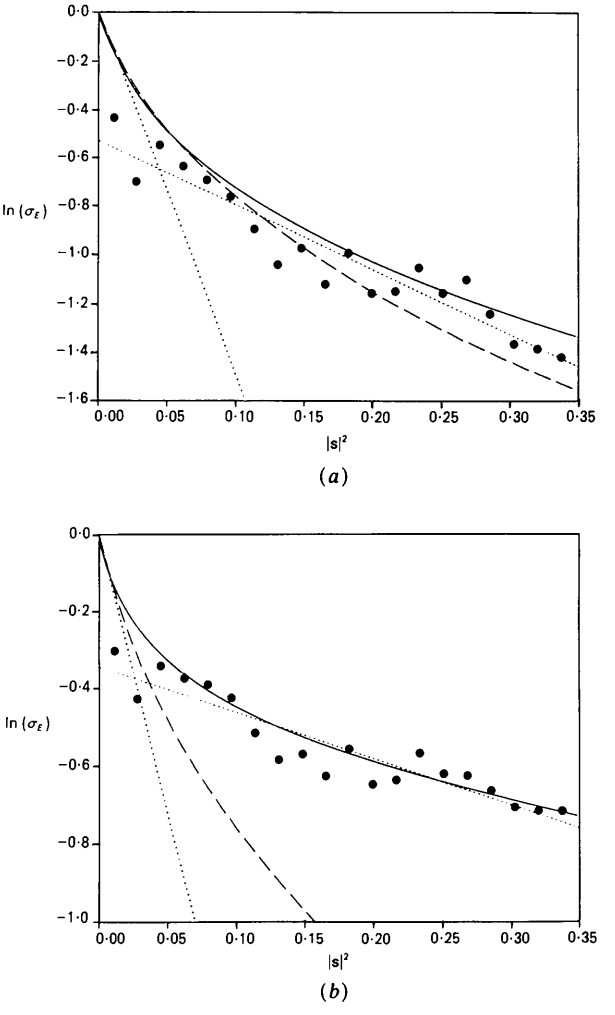

Fig. 4. Comparison of σ_E values [computed with (36)] with various theoretical curves, plotted as $\ln(\sigma_E)$ *vs* $|\mathbf{s}|^2$. The dotted line is predicted by the theory of Luzzati (1952) for an isotropic Gaussian error of 1·49 Å r.m.s., while the chain-dotted line is obtained by taking σ_E to be the Fourier transform of the actual distribution of coordinate differences (Luzzati, 1952). The dashed line gives the least-squares fit to the points. Finally, the solid curve is computed using (34) for σ_E. The $\mathbf{F}$ and f_j values are from the final refined model of SGT. The $\mathbf{G}$ and g_j values are from the BT models, either with no weighting (*a*) or with individual *B*-factor weights (*b*).

regions than others. The centroid of the structure-factor probability distribution is obtained, in this case, by taking the Fourier transform of the expected electron-density function. In other terms, each atom of a molecular replacement model would be smeared over its distribution of possible positions. It is assumed that there is a sufficient number of independent contributions to the difference in the structure factors, so that the central limit theorem applies and the probability distribution is a Gaussian about the centroid estimate.

For a model of a crystal structure, it is preferable to consider the average effect of a specific set of errors on a set of structure factors, in other words to consider the reciprocal-space vector as the random variable. The probability distributions underlying the differences between the model and the true structure enter through the frequencies of the errors over all the atoms. Essentially the same probability distributions of structure factors arise as in the previous case, because of the symmetry between real and reciprocal space in the Fourier transform.

Considered in terms of normalized structure factors, all sources of error have the same effect, which can be summarized in a single parameter, σ_E. This parameter plays the same role in the probability distributions as σ_A in the distributions of Srinivasan & Ramachandran (1965). Therefore, the methods suggested previously to estimate phase probabilities and to calculate electron-density maps (Read, 1986) are still valid. However, the interpretation of the parameter σ_E is different. In particular, the variation of σ_E with resolution cannot be attributed entirely to coordinate error. Methods such as the Luzzati (1952) plot and the σ_A plot (Read, 1986) to estimate coordinate error will therefore suffer from a number of sources of systematic error.

It is a pleasure to acknowledge helpful discussions with Marie E. Fraser and Trevor N. Hart. The author is an Alberta Heritage Foundation for Medical Research Scholar.

References

BLOW, D. M. & CRICK, F. H. C. (1959). *Acta Cryst.* **12**, 794-802.
CHAMBERS, J. L. & STROUD, R. M. (1979). *Acta Cryst.* B**35**, 1861-1874.
CHOTHIA, C. & LESK, A. M. (1986). *EMBO J.* **5**, 823-826.
CRUICKSHANK, D. W. J. (1949). *Acta Cryst.* **2**, 65-82.
HAUPTMAN, H. (1982). *Acta Cryst.* A**38**, 289-294.
HAVEL, T. F. & WÜTHRICH, K. (1985). *J. Mol. Biol.* **182**, 281-294.
HENDRICKSON, W. A. & LATTMAN, E. E. (1970). *Acta Cryst.* B**26**, 136-143.
HUBBARD, T. J. P. & BLUNDELL, T. L. (1987). *Protein Eng.* **1**, 159-171.
JAMES, R. W. (1948). *The Optical Principles of the Diffraction of X-rays.* Ithaca, NY: Cornell Univ. Press.
LUZZATI, V. (1952). *Acta Cryst.* **5**, 802-810.
MITRA, G. B., AHMED, R. & DAS GUPTA, P. (1985). *Structure and Statistics in Crystallography*, edited by A. J. C. WILSON, pp. 151-181. New York: Adenine Press.
MOULT, J. & JAMES, M. N. G. (1986). *Proteins*, **1**, 146-163.
OTWINOWSKI, Z., SCHEVITZ, R. W., ZHANG, R.-G., LAWSON, C. L., JOACHIMIAK, A., MARMORSTEIN, R. Q., LUISI, B. F. & SIGLER, P. B. (1988). *Nature (London)*, **335**, 321-329.
PHILLIPS, S. E. V. (1980). *J. Mol. Biol.* **142**, 531-554.
READ, R. J. (1986). *Acta Cryst.* A**42**, 140-149.
READ, R. J. & JAMES, M. N. G. (1988). *J. Mol. Biol.* **200**, 523-551.
ROSSMANN, M. G. (1972). *The Molecular Replacement Method.* New York: Gordon and Breach.
ROSSMANN, M. G. & BLOW, D. M. (1961). *Acta Cryst.* **14**, 641-647.
SHERIFF, S. & HENDRICKSON, W. A. (1987). *Acta Cryst.* A**43**, 118-121.
SILVA, A. M. & ROSSMANN, M. G. (1985). *Acta Cryst.* B**41**, 147-157.
SIM, G. A. (1959). *Acta Cryst.* **12**, 813-815.
SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-ray Crystallography.* Oxford: Pergamon Press.
SRINIVASAN, R. & RAMACHANDRAN, G. N. (1965). *Acta Cryst.* **19**, 1008-1014.
STEWART, J. M. & KARLE, J. (1976). *Acta Cryst.* A**32**, 1005-1007.
SUTCLIFFE, M. J., HANEEF, I., CARNEY, D. & BLUNDELL, T. L. (1987). *Protein Eng.* **1**, 377-384.
WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318-321.
WILSON, A. J. C. (1976). *Acta Cryst.* A**32**, 781-783.
WOOLFSON, M. M. (1956). *Acta Cryst.* **9**, 804-810.

Acta Cryst. (1990). A**46**, 912-915

On Inclined Cubic Sublattices of Cubic Lattices

BY V. FREI

Department of Semiconductor Physics, Faculty of Mathematics and Physics, Charles University, Ke Karlovu 5, 121 16 *Praha* 2, *Czechoslovakia*

(*Received* 1 *January* 1989; *accepted* 19 *June* 1990)

Abstract

Cubic sublattices of cubic lattices are described which share only some of the point-symmetry operations with the original lattices; the common operations form the point groups $\bar{3}m$, $\bar{3}$, $4/m$, $2/m$ or $\bar{1}$. Some properties of these sublattices, including the centred ones, are shown and tentative terminology, notation and classification are introduced. All the different types of inclined primitive cubic sublattices L_n for

0108-7673/90/110912-04$03.00