Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 22 June 2000
Accepted 5 September 2000

A probabilistic approach to space-group determination from powder diffraction data

A. J. Markvardsen,* W. I. F. David, J. C. Johnson and K. Shankland

ISIS Facility, Rutherford Appleton Laboratory, Chilton, Oxon OX11 0QX, England. Correspondence e-mail: a.j.markvardsen@rl.ac.uk

An algorithm for the determination of the space-group symmetry of a crystal from powder diffraction data, based upon probability theory, is described. Specifically, the relative probabilities of different extinction symbols are assessed within a particular crystal system. In general, only a small number of extinction symbols are relatively highly probable and a single extinction symbol is often significantly more probable than any other. Several examples are presented to illustrate this approach.

1. Introduction

Crystal-structure determination from powder diffraction data faces several challenges that do not exist in the determination of crystal structures from single-crystal data. These challenges all principally result from the compression of three dimensions of diffraction data on to the one dimension of a powder diffraction pattern. Although unit-cell determination and structure-factor extraction are the two stages in the structure solution process that suffer most from Bragg peak overlap, space-group determination can also be problematical. Traditionally, space-group determination from powder diffraction data is performed manually by inspection of the systematically absent reflections. However, in a powder diffraction pattern, partial or complete Bragg peak overlap can make this manual inspection time consuming and ambiguous, particularly for orthorhombic and tetragonal symmetries. In this paper, an algorithm is presented that gives a quantitative measure of the relative probabilities of different extinction symbols. Although it is in principle possible to distinguish between different space groups that possess the same extinction symbol, no attempt has been made to do this in this paper. From a practical point of view, this is not a severe problem since there are generally only a small number of space groups that possess the same extinction symbol.

2. Probabilistic approach

In this paper, it is presumed that the unit cell has been previously determined by indexing and that the crystal system is therefore known. Armed with this information, it is possible to perform a profile refinement using either the Le Bail (Le Bail *et al.*, 1988) or Pawley (1981) methods in order to obtain Bragg peak intensities. In this paper, the Pawley method is preferred since it gives access to the full covariance matrix of correlations between Bragg peak intensities. The information content on the Bragg peak intensities in the diffraction pattern is then summarized by a multivariate Gaussian likelihood function,

$$p(\mathbf{I}^{\mathbf{P}}|\mathbf{I}) = (2\pi)^{-N/2}|\mathbf{C}|^{-1/2}\exp[-\tfrac{1}{2}(\mathbf{I}^{\mathbf{P}}-\mathbf{I})^{T}\mathbf{C}^{-1}(\mathbf{I}^{\mathbf{P}}-\mathbf{I})]. \quad (1)$$

The intensities $\mathbf{I}^{\mathbf{P}} = (I_1, I_2, \ldots, I_N)$ are the values determined from the linear-least-squares Pawley refinement with one value for each reflection. The matrix $\mathbf{C}^{-1}$ holds information about intensity correlations between neighbouring reflections in the powder pattern. For example, if none of the reflections in the diffraction pattern are found to overlap then the correlation matrix is an $N \times N$ diagonal matrix with the diagonal elements equal to the variance of each of the N refined intensities. The correlation-matrix element is equal to the expectation value

$$C_{ij} = \langle (I_i^P - I_i)(I_j^P - I_j)\rangle_{p(\mathbf{I}^{\mathbf{P}}|\mathbf{I})}.$$

The off-diagonal elements are thus only of significant value for reflections that are substantially overlapped. It is important to emphasize that, in the Pawley refinement, the intensities are not constrained to be positive. Assuming that the diffraction pattern background has been correctly modelled, the Pawley refined intensity of an isolated peak is generally positive but, for very weak peaks, may be negative by an amount that is similar to the standard uncertainty of the peak intensity. For overlapping peaks, one or more of the refined intensities may be negative. However, the mean of the probability distribution of the sum of the overlapping intensities is a positive value and the statistics of a group of reflections are well behaved.

The first step in the determination of the most probable extinction symbol is therefore a Pawley refinement of the diffraction pattern in the most general extinction group of the crystal system under consideration. For example, if the Bravais lattice is orthorhombic, then the extracted intensities are obtained by performing the Pawley refinement in extinction group *P*––– (corresponding to any of *P*222, *Pmmm*, *Pmm*2, *Pm*2*m*, *P*2*mm*), which has no systematic absences. Denoting the refined Pawley data set by the intensity vector $\mathbf{I}^\mathbf{P}$ and using Bayes's theorem gives the relation

$$p(E_{\rm gr}|\mathbf{I}^\mathbf{P}) = p(E_{\rm gr})p(\mathbf{I}^\mathbf{P}|E_{\rm gr})/p(\mathbf{I}^\mathbf{P}),$$

where $p(E_{\rm gr})$ is here the prior probability distribution. In this paper, it is presumed that all extinction symbols are *a priori* equally probable and thus $p(E_{\rm gr})$ is a constant. Other prior probability options are possible; one obvious choice is to bias the probabilities by the known relative frequencies with which space groups occur in nature for a particular class of compounds. Additionally, as the data do not change from consideration of one extinction symbol to another, $p(\mathbf{I}^\mathbf{P})$ is constant. Together with the prior assignment for $p(E_{\rm gr})$, this means that $p(E_{\rm gr}|\mathbf{I}^\mathbf{P})$ is proportional to $p(\mathbf{I}^\mathbf{P}|E_{\rm gr})$ as a function of $E_{\rm gr}$. In other words,

$$p(E_{\rm gr}|\mathbf{I}^\mathbf{P}) \propto p(\mathbf{I}^\mathbf{P}|E_{\rm gr}). \qquad (2)$$

The quantity $p(\mathbf{I}^\mathbf{P}|E_{\rm gr})$ thus gives the relative probability for each extinction symbol of a crystal system. The different extinction symbols may then be ranked by their relative probabilities. To calculate the probability $p(\mathbf{I}^\mathbf{P}|E_{\rm gr})$, consider first the joint probability density $p(\mathbf{I}^\mathbf{P}, \mathbf{I}|E_{\rm gr})$, which may be written as

$$p(\mathbf{I}^\mathbf{P}, \mathbf{I}|E_{\rm gr}) = p(\mathbf{I}|E_{\rm gr})p(\mathbf{I}^\mathbf{P}|\mathbf{I}, E_{\rm gr}). \qquad (3)$$

Clearly, the last term in (3) does not depend on $E_{\rm gr}$ and, by marginalizing out the components of the intensity vector $\mathbf{I}$ in (3), the equation that will be used to calculate the Bayesian probability table is obtained:

$$p(\mathbf{I}^\mathbf{P}|E_{\rm gr}) = \textstyle\int p(\mathbf{I}|E_{\rm gr})p(\mathbf{I}^\mathbf{P}|\mathbf{I})\,{\rm d}\mathbf{I}. \qquad (4)$$

In this equation, $p(\mathbf{I}^\mathbf{P}|\mathbf{I})$ is the likelihood function from (1) and $p(\mathbf{I}|E_{\rm gr})$ is the probability of observing the intensities $\mathbf{I} = (I_1, I_2, \ldots, I_N)$ given that $E_{\rm gr}$ is the true extinction group. In assigning $p(\mathbf{I}|E_{\rm gr})$, the trivial assumption is made that all of the intensities are independent and identically distributed; in other words, we can write $p(\mathbf{I}|E_{\rm gr})$ as the product

$$p(\mathbf{I}|E_{\rm gr}) = \prod_{i=1}^{N} p(I_i|E_{\rm gr}).$$

This reduces the problem of assigning $p(\mathbf{I}|E_{\rm gr})$ to that of assigning a univariate distribution $p(I_i|E_{\rm gr})$. Depending on the extinction group $E_{\rm gr}$, the intensity I_i will be predicted either to be present or to be absent. When the ith intensity is predicted to be absent by $E_{\rm gr}$, $p(I_i|E_{\rm gr})$ will be a delta function:

$$p(I_i|E_{\rm gr}) = \delta(I_i).$$

If the ith intensity is predicted to be present, then $p(I_i|E_{\rm gr})$ is modelled by an exponential distribution with mean value μ,

$$p(I_i|E_{\rm gr}) = \begin{cases} 0 & \text{for } I_i < 0 \\ \mu^{-1}\exp(-I_i/\mu) & \text{for } I_i \geq 0. \end{cases} \qquad (5)$$

The delta-function assignment for an absence is intuitively straightforward to accept; if an intensity is predicted to be absent, then this is exactly equivalent to assigning zero probability density for any non-zero intensity. The exponential distribution assigned in (5) is the intensity distribution for an acentric reflection (Wilson, 1949) and its general use for all space groups may be questioned. However, in §6, it will be shown that it is an appropriate choice for modelling a presence regardless of whether the underlying space group is centrosymmetric or non-centrosymmetric. As a corollary, however, this simplification does negate the possibility of discriminating between centrosymmetric or non-centrosymmetric space groups. As stated earlier, this is in general not a serious problem, since only a small number of space groups belong to a given extinction symbol.

3. Isolated reflections

In this section, the evaluation of probabilities for the presence or absence of an isolated reflection in the powder diffraction pattern is addressed. Subsequent sections deal with the general case of overlapping reflections.

The following analytical solutions of (4) may be obtained by assuming that the reflection is (a) present and (b) absent:

$$p(I^P|\text{present}) = \frac{1}{2\mu}\exp\left(\frac{\sigma^2}{2\mu^2} - \frac{I^P}{\mu}\right)\text{erfc}\left(\frac{1}{2^{1/2}}\left[\frac{\sigma}{\mu} - \frac{I^P}{\sigma}\right]\right), \qquad (6)$$

where erfc(x) is the complementary error function and

$$p(I^P|\text{absent}) = \frac{1}{(2\pi)^{1/2}\sigma}\exp\left(-\frac{(I^P)^2}{2\sigma^2}\right). \qquad (7)$$

Hence, using (2), the absent/present probability ratio is given by

$$\frac{p(\text{absent}|I^P)}{p(\text{present}|I^P)} = \frac{p(I^P|\text{absent})}{p(I^P|\text{present})} = \left(\frac{\mu 2^{1/2}}{\sigma\pi^{1/2}}\right)\frac{\exp(-z^2)}{\text{erfc}(z)}, \qquad (8)$$

where

$$z = 2^{-1/2}(\sigma/\mu - I^P/\sigma).$$

Figs. 1(a) and 1(b) show the probabilities of a peak being absent and present, respectively, whilst Fig. 1(c) shows the logarithm of the ratio in (8) as a function of I^P and μ for constant $\sigma = 1$. Note that μ is the expected intensity of a reflection that is present; for weak diffraction data, μ will be small, whilst, for strong diffraction data, μ is expected to be large. With the asymptotic approximation $\text{erfc(z)} \sim \pi^{-1/2}z^{-1}\exp(-z^2)$, it may be shown, in the limit of μ tending towards zero, that the ratio in (8) tends towards one. In other words, weak peaks in weak data offer little in the way of discrimination with respect to systematic presences or absences. In the other limit, as μ tends to infinity for a fixed value of I^P, the ratio in (8) becomes proportional to μ. This highlights the need to have a good initial estimate for μ, as it is always possible to make the 'peak absent' case the more probable, irrespective of the data, by choosing μ to be sufficiently large. In §6, it is shown how μ can be automatically calculated from the powder diffraction pattern. A close examination of Fig. 1(c) shows how any suggestion that a peak is present rapidly discriminates against a systematic absence. For example, with $\mu = 20$, a peak with an intensity $I^P = 2$ and

$\sigma = 1$ represents the point where the ratio in (8) is equal to 1. A peak that is only twice as intense ($I^P = 4 \pm 1$) will be ~300 times more likely to be present than absent. Thus strong peaks severely penalize the criterion of systematic absence.

4. Computational considerations

The dimension of the integral in (4) is equal to the number of intensities (typically several hundred) that are evaluated in the Pawley refinement. Integrals of this size are, in general, non-trivial to evaluate numerically and may sometimes be intractable. This section addresses this numerical problem, with the principal savings in computation time coming from the treatment of the correlation matrix in a block-diagonal style suited to its inherent structure. Thus, the integral in (4) is split up into a set of smaller integrals. Integrals with dimension larger than one are solved using the Monte Carlo method. The remaining integrals have analytical solutions.

4.1. The correlation matrix

If all N peaks in a powder diffraction pattern are non-overlapping, then the correlation matrix is diagonal. The likelihood function in (1) becomes a product of N 'one-dimensional' Gaussian distributions and the integral in (4) is reduced to a product of N one-dimensional integrals that can be evaluated analytically.

In general, diffraction patterns exhibit significant peak overlap and thus the correlation matrix takes a more complex form. It will, however, always be approximately block diagonal with blocks of dimensions equal to the number of peaks present in each of the overlap regions. In writing $\mathbf{C}$ in block-diagonal form, it is convenient to ignore small overlaps between peaks. This is accomplished by only keeping the off-diagonal elements of the correlation matrix that satisfy the criterion

$$C_{ij}/(C_{ii}C_{jj})^{-1/2} \geq \eta. \tag{9}$$

Thus, starting from the first reflection in the diffraction pattern, $C_{12}/(C_{11}C_{22})^{1/2} \geq \eta$ is evaluated. The correlation cut-off is typically about 40%. If the condition is not satisfied, C_{12} is set to zero and thus I_1 and C_{11} define the first block. Clearly, blocks having only one member represent an uncorrelated reflection. If reflections 1 and 2 are found to be correlated, then the expressions $C_{13}/(C_{11}C_{33})^{1/2} \geq \eta$ and $C_{23}/(C_{22}C_{33})^{1/2} \geq \eta$ are assessed for correlation between reflection 3 and either of the first two. This process is repeated until a reflection is reached that is not correlated with any previous reflection and thus the end of a block has been reached and a new one begins. At the end of this operation, the correlation matrix $\mathbf{C}$ is transformed into M submatrices $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_M$ of varying dimensions. All of these M submatrices are positive definite because every leading principal submatrix of a positive-definite matrix is also positive definite. In the block-diagonal approximation, the likelihood in (1) may thus be written as the product of M univariate and multivariate Gaussian distributions, *i.e.*

$$\begin{aligned} p(\mathbf{I^P}|\mathbf{I}, \eta) &= \prod_{m=1}^{M} p(\mathbf{I}^{\mathbf{P}}_m|\mathbf{I}_m, \eta) \\ &= [1/(2\pi)^{N/2}] \prod_{m=1}^{M} (1/|\mathbf{C}_m|^{1/2}) \\ &\quad \times \exp[-\tfrac{1}{2}(\mathbf{I}^{\mathbf{P}}_m - \mathbf{I}_m)^T \mathbf{C}_m^{-1} (\mathbf{I}^{\mathbf{P}}_m - \mathbf{I}_m)], \end{aligned}$$

where η is the correlation cut-off criterion in (9) and $\mathbf{I}_m$ and $\mathbf{C}_m$ are the intensities and correlation matrix defining block m. Inserting this simplified likelihood into (4) gives

$$p(\mathbf{I^P}|E_{\rm gr}, \eta) = \prod_{m=1}^{M} \int p(\mathbf{I}_m|E_{\rm gr}) p(\mathbf{I}^{\mathbf{P}}_m|\mathbf{I}_m, \eta)\, {\rm d}\mathbf{I}_m. \tag{10}$$

This may be calculated more easily than (4) when $M > 1$.

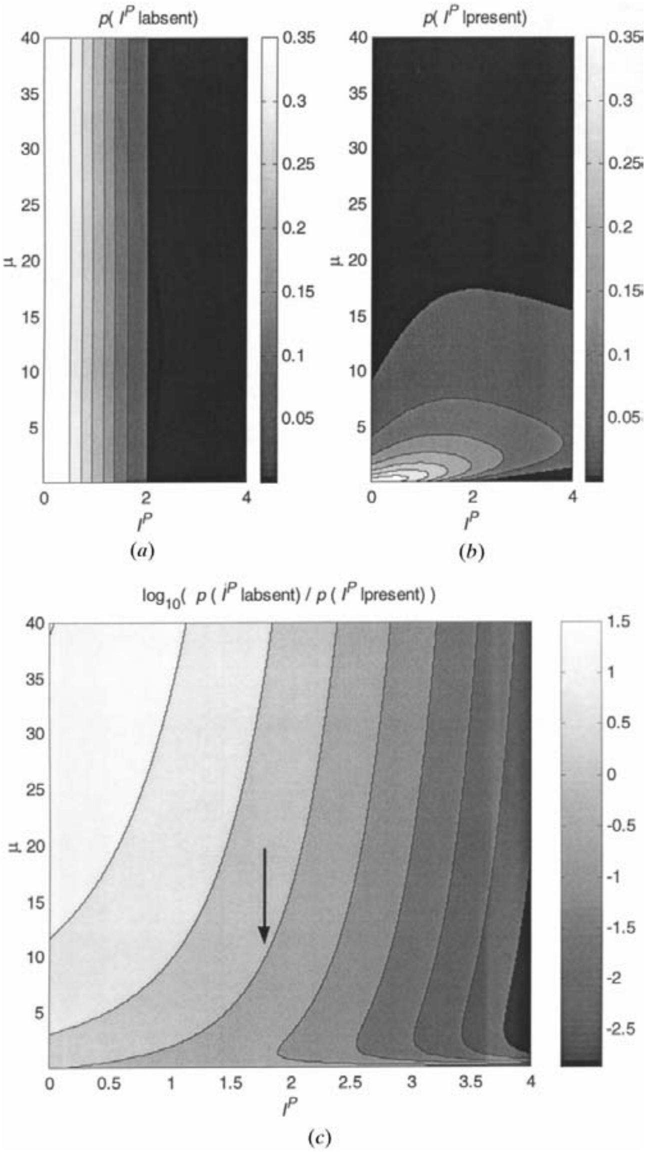


Figure 1
The probabilities of a peak being (a) absent or (b) present, shown as a function of I^P and μ for constant $\sigma = 1$. The $\log_{10}$ of the ratio of the above probabilities is shown as a function of I^P and μ in (c). The arrow points to the contour level corresponding to an equal probability of a reflection being absent or present.

4.2. Calculating a block integral

For block integrals with dimensions higher than one in (10), no analytical solution exists and we use the Monte Carlo technique to find a numerical solution. Our implementation of the Monte Carlo method is described in the subsections below.

4.2.1. Monte Carlo method. Given N samples $\mathbf{x}^{(n)}$ drawn from the probability distribution $p(\mathbf{x})$, then, for a particular function $g(\mathbf{x})$, the expectation value for g is

$$\langle g(\mathbf{x})\rangle_{p(\mathbf{x})} = \textstyle\int g(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} = (1/N)\sum\limits_{n=1}^{N} g(\mathbf{x}^{(n)}) \pm \sigma, \qquad (11)$$

where the error term σ is approximately given by

$$\sigma^2 = (1/N)\sum_{n=1}^{N} g^2(\mathbf{x}^{(n)}) - \left[(1/N)\sum_{n=1}^{N} g(\mathbf{x}^{(n)})\right]^2. \qquad (12)$$

Comparing g in (11) with $p(\mathbf{I}_m|E_{\rm gr})$ in (10), and noting the similarity in form, the mth term of (10) may be written as

$$p(\mathbf{I}^{\mathbf{P}}_m|E_{\rm gr}, \eta) = \langle p(\mathbf{I}_m|E_{\rm gr})\rangle_{p(\mathbf{I}^{\mathbf{P}}_m|\mathbf{I}_m,\eta)}. \qquad (13)$$

4.2.2. Generation of Gaussian samples. Samples from a multivariate Gaussian distribution are required in order to calculate the expectation value in (13). Any such distribution can always be transformed into the standard Gaussian distribution by the change of variables

$$\mathbf{I}^{(n)} = \mathbf{M}\mathbf{y}^{(n)} + \mathbf{I}^{\mathbf{P}}, \qquad (14)$$

where $\mathbf{M}$ is any non-singular factorization of $\mathbf{C}$ such that $\mathbf{C} = \mathbf{M}\mathbf{M}^T$ and $\mathbf{y}$ is distributed by the standard Gaussian, *i.e.* with zero mean values and the identity matrix as covariance matrix. Standard methods for generating samples $\mathbf{y}^{(n)}$ drawn from a standard Gaussian distribution exist (*e.g.* Press *et al.*, 1992) and, using the linear transformation in (14), one can then generate a sample drawn for any multivariate Gaussian distribution. To reduce the computational cost of calculating the matrix–vector multiplication $\mathbf{M}\mathbf{y}^{(n)}$ in (14), the covariance matrix $\mathbf{C}$ is decomposed by a standard Cholesky algorithm to obtain a triangular matrix for $\mathbf{M}$. In addition, by calculating the components of a sample $\mathbf{I}^{(n)}$ one by one, if the ith component is negative, none of the remaining components need be calculated because $p(\mathbf{I}^{(n)}|E_{\rm gr})$ is zero when any of the components of $\mathbf{I}^{(n)}$ are negative [see (5)].

4.2.3. Very improbable integrals. The Monte Carlo integration method works very well for the majority of block integrals in (10) but occasionally it fails. Typically, this happens when a number of the refined intensities in a block possess negative intensity values. In such cases, the probability of generating a Gaussian sample $\mathbf{I}^{(n)}$, which lies within the sample region where $p(\mathbf{I}^{(n)}|E_{\rm gr})$ is non-zero, can be extremely small. Hence, the integration fails as it becomes impractical to produce a non-zero value for the block integral within a reasonable time scale. Rather than discard such blocks, a simple modification of the shape of the integral can render it solvable within an acceptable time scale. For instance, if the problem is to produce Gaussian samples $\mathbf{I}^{(n)}$ where all the components of $\mathbf{I}^{(n)}$ are positive and therefore $p(\mathbf{I}^{(n)}|E_{\rm gr})$ is non-zero, then the Monte Carlo block integral in (13) can be recast into the form

$$\begin{aligned}\langle p(\mathbf{I}_m|E_{\rm gr})\rangle_{p(\mathbf{I}^{\mathbf{P}}_m|\mathbf{I}_m,\eta)} &= \exp\{-(\mathbf{P}\mathbf{I}^{\mathbf{P}}_m)^T\mathbf{C}_m^{-1}\mathbf{P}([\mathbf{1} - \tfrac{1}{2}\mathbf{P}]\mathbf{I}^{\mathbf{P}}_m)\}\\ &\quad\times \langle\exp[(\mathbf{P}\mathbf{I}^{\mathbf{P}}_m)^T\mathbf{C}_m^{-1}\mathbf{I}_m]p(\mathbf{I}_m|E_{\rm gr})\rangle_{p([\mathbf{1}-\mathbf{P}]\mathbf{I}^{\mathbf{P}}_m|\mathbf{I}_m,\eta)}. \qquad (15)\end{aligned}$$

Here the Gaussian sampler takes the form $p([\mathbf{1} - \mathbf{P}]\mathbf{I}^{\mathbf{P}}_m|\mathbf{I}_m, \eta)$, where $\mathbf{1}$ is an $N_m \times N_m$ identity matrix and $\mathbf{P}$ is a diagonal matrix with (a) ones along the diagonal for components of $\mathbf{I}^{\mathbf{P}}_m$, which are negative, and (b) the remaining diagonal elements equal to zero. As none of the components of $[\mathbf{1} - \mathbf{P}]\mathbf{I}^{\mathbf{P}}_m$ of the Gaussian sampler in (15) are negative, we are guaranteed at least a 50% success rate in generating samples $\mathbf{I}^{(n)}$, where the function whose expectation value is to be determined,

$$\exp[(\mathbf{P}\mathbf{I}^{\mathbf{P}}_m)^T\mathbf{C}_m^{-1}\mathbf{I}_m]p(\mathbf{I}_m|E_{\rm gr}),$$

is non-zero. This and other similar manipulations can be used to evaluate numerically even extremely improbable block integrals.

4.3. Repetition of calculations

If the number of intensities in a block is relatively small, then the same integral values may be calculated many times over for different symmetries that possess the same systematic absences within a single block. In order to avoid undue repetition of these computationally expensive calculations, the value of each integral is stored and reused whenever possible.

5. Systematic absences and complete reflection overlap

The effect of systematic absences as well as completely overlapping peaks upon the integrals under consideration must be accounted for before the Monte Carlo method outlined above can be applied.

5.1. Systematic absences

Systematic absences in $E_{\rm gr}$ introduce delta functions into the integral in (4). For example, if $E_{\rm gr}$ dictates that only the first N_R of the N possible intensities are present in the diffraction pattern, then these N_R intensities can be represented by the $N_R \times 1$ vector $\mathbf{I}_{\mathbf{R}} = (I_1, I_2, \ldots, I_{N_R})$ and the integral in (4) reduces to

$$\begin{aligned}p(\mathbf{I}^{\mathbf{P}}|E_{\rm gr}) &= (2\pi)^{-N/2}|\mathbf{C}|^{-1/2}(1/\mu^{N_R})\\ &\quad\times \textstyle\int \exp\left[-\tfrac{1}{2}(\mathbf{I}^{\mathbf{P}} - \tilde{\mathbf{I}})^T\mathbf{C}^{-1}(\mathbf{I}^{\mathbf{P}} - \tilde{\mathbf{I}}) - \mu^{-1}\sum\limits_{i=1}^{N_R} I_i\right]\mathrm{d}\mathbf{I}_{\mathbf{R}},\end{aligned}$$

where $\tilde{\mathbf{I}}$ is an $N \times 1$ vector, the first N_R components of which are the components of $\mathbf{I}_{\mathbf{R}}$ and the remaining elements are zero. The above integral can be rewritten as

$$\begin{aligned} p(\mathbf{I^P}|E_{\rm gr}) &= (2\pi)^{-N/2}|\mathbf{C}|^{-1/2}(1/\mu^{N_R}) \\ &\quad\times \exp[\tfrac{1}{2}(\mathbf{I}_\mathbf{R}^{\rm new})^T\mathbf{C}_\mathbf{RR}^{-1}\mathbf{I}_\mathbf{R}^{\rm new} - \tfrac{1}{2}(\mathbf{I^P})^T\mathbf{C}^{-1}\mathbf{I^P}] \\ &\quad\times \int \exp\Bigg[-\tfrac{1}{2}(\mathbf{I}_\mathbf{R}^{\rm new} - \mathbf{I_R})^T\mathbf{C}_\mathbf{RR}^{-1}(\mathbf{I}_\mathbf{R}^{\rm new} - \mathbf{I_R}) \\ &\quad - \mu^{-1}\sum_{i=1}^{N_R} I_i\Bigg]\,{\rm d}\mathbf{I_R}, \end{aligned}$$

where $\mathbf{C}_\mathbf{RR}^{-1}$ is the first N_R rows and N_R columns of $\mathbf{C}^{-1}$, and $(\mathbf{I}_\mathbf{R}^{\rm new})^T = (\mathbf{I^P})^T\mathbf{C}_\mathbf{NR}^{-1}\mathbf{C}_\mathbf{RR}$ is a new set of Gaussian mean values, where $\mathbf{C}_\mathbf{NR}^{-1}$ is the first N and N_R columns of the inverse correlation matrix $\mathbf{C}^{-1}$.

5.2. Two completely overlapping peaks

In a Pawley refinement, two almost identically overlapping reflections must be treated as a single variable representing the sum of the two different reflection intensities in order that instabilities in the least-squares refinement are avoided. These doublets in the Pawley refinement lead to a likelihood expression of the form

$$p(I^P|I_1, I_2) = (2\pi)^{-1/2}\sigma^{-1}\exp\{-[I^P - (I_1 + I_2)]^2/(2\sigma^2)\},$$

where I_1 and I_2 are two reflection intensities with different Miller indices and I^P is the refined value for the sum of the two intensities. When both I_1 and I_2 are predicted to be present by $E_{\rm gr}$, this results in an entry of the following type in (10):

$$\begin{aligned} &p(I^P|I_1 \text{ and } I_2 \text{ present}) \\ &\quad= (2\pi)^{-1/2}\sigma^{-1}(1/\mu^2)\int_0^\infty\int_0^\infty \exp\{-[I^P - (I_1 + I_2)]^2/(2\sigma^2) \\ &\quad - (I_1 + I_2)/\mu\}\,{\rm d}I_1{\rm d}I_2. \end{aligned} \qquad (16)$$

Clearly, this integral depends only on the sum of the two intensities I_1 and I_2. Therefore, by changing the variables to the sum $I = I_1 + I_2$ and difference $I_D = I_1 - I_2$ intensities, the expression becomes (on integrating out I_D)

$$\begin{aligned} &p(I^P|I_1 \text{ and } I_2 \text{ present}) \\ &\quad= (2\pi)^{-1/2}\sigma^{-1}(1/\mu^2)\int_0^\infty I\exp[-(I^P - I)^2/(2\sigma^2) - I/\mu]\,{\rm d}I. \end{aligned} \qquad (17)$$

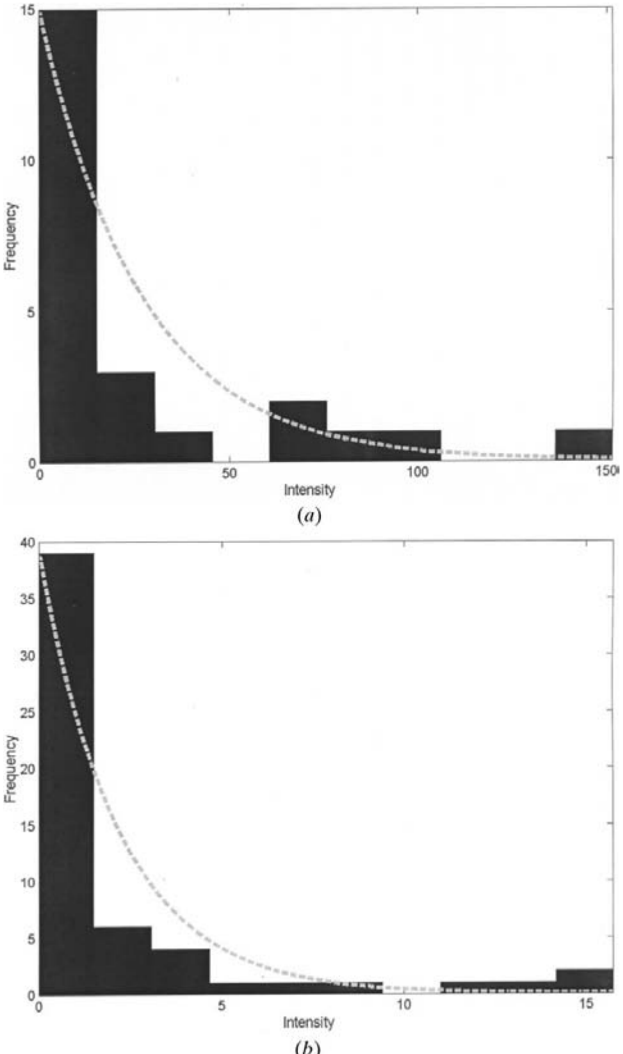
Equation (17) has the analytical solution

$$\begin{aligned} &p(I^P|I_1 \text{ and } I_2 \text{ present}) \\ &\quad= 2^{-1/2}\mu^{-2}\sigma[\pi^{-1/2} - z\exp(z^2)\,{\rm erfc(z)}]\exp[-(I^P)^2/(2\sigma^2)] \end{aligned}$$

with $z = 2^{-1/2}(\sigma/\mu - I^P/\sigma)$. This step from (16) and (17) can easily be extended to a pattern containing any number of correlated singlets, doublets and multiplets.

6. Calculation of μ

As outlined in §3, the choice of an appropriate prior mean intensity value μ is important and it is appropriate to use the diffraction pattern in determining this value. Fortunately, certain Bragg reflections are always present in the diffraction pattern irrespective of the extinction symbol. Furthermore, for the majority of patterns, a number of these reflections will be uncorrelated with others and thus μ can be set equal to the arithmetic mean of such reflections. As a consequence of using all the uncorrelated reflections of the pattern to estimate one value of μ, this value represents μ at any (2θ) location in the diffraction pattern. Fig. 2 shows the intensity and frequency of the reflections used in the calculation of μ for (*a*) dopamine hydrobromide, space group $Pbc2_1$ and (*b*) remacemide nitrate, space group $P2_1/a$. Despite representing non-centrosymmetric and centrosymmetric space groups, respectively, both histograms show reasonable agreement with the exponential distribution $p(I) = \exp(-I/\mu)/\mu$. This ambiguity highlights the difficulties in assigning a centric/acentric distinction solely on the basis of this subset of well determined reflections. In the



(*a*)

(*b*)

Figure 2
Histogram plots of reflection intensity (grouped into a number of intervals) *versus* the number of uncorrelated intensity values falling into each of these intervals for the compounds (*a*) dopamine hydrobromide and (*b*) remacemide nitrate. The arithmetic means of the uncorrelated intensities were found to be 37 and 2.2, respectively. The exponential distributions [equation (5)] are plotted as dashed lines and scaled such that the values at the origin are equal to the frequency numbers for the first intensity intervals. The correlation cut-off, η, is 40% in both cases [see equation (9)].

Table 1
Extinction symbols for orthorhombic space groups, listed in the order in which they appear in *ITCA* (1983) and corresponding probabilities expressed as $\ln[p(E_{\rm gr}\mid\mathbf{I}^{\rm P})/p(E_{---}\mid\mathbf{I}^{\rm P})]$ for dopamine hydrobromide.

The correlated integrated intensities were extracted in space group *Pmmm*.

Symbol	Probability	Symbol	Probability	Symbol	Probability	Symbol	Probability
$P{-}{-}2_1$	5.9	$Pbaa$	−18084.6	$Pn{-}a$	−15534.9	$B{-}(ac)b$	−56981.4
$P{-}2_1{-}$	15.2	$Pbab$	−13107.2	$Pn{-}b$	−5915.3	$Bb{-}{-}$	−56419.8
$P{-}2_12_1$	21.1	$Pban$	−17774.1	$Pn{-}n$	−15224.4	$Bb{-}b$	−56565.9
$P2_1{-}{-}$	−4644.4	$Pbc{-}$	97.9	$Pna{-}$	−19708.7	$Bb(ac){-}$	−56825.6
$P2_1{-}2_1$	−4638.5	$Pbca$	−10015.9	$Pnaa$	−25172.7	$Bb(ac)b$	−56971.7
$P2_12_1{-}$	−4629.2	$Pbcb$	−388.5	$Pnab$	−20195.3	$A{-}{-}{-}$	−64661.5
$P2_12_12_1$	−4623.3	$Pbcn$	−9705.4	$Pnan$	−24862.2	$A2_1{-}{-}$	−69303.2
$P{-}{-}a$	−10108.4	$Pbn{-}$	−14704.3	$Pnc{-}$	−5387.1	$A{-}{-}a$	−74368.6
$P{-}{-}b$	−471.3	$Pbna$	−20168.2	$Pnca$	−15501.0	$A{-}a{-}$	−79919.6
$P{-}{-}n$	−9782.7	$Pbnn$	−19857.7	$Pncb$	−5873.6	$A{-}aa$	−84985.0
$P{-}a{-}$	−12673.3	$Pc{-}{-}$	−5428.3	$Pncn$	−15190.5	$A(bc){-}{-}$	−64654.1
$P{-}aa$	−18137.2	$Pc{-}a$	−15534.5	$Pnn{-}$	−20189.4	$A(bc){-}a$	−74361.2
$P{-}ab$	−13144.6	$Pc{-}b$	−5899.6	$Pnna$	−25653.3	$A(bc)a{-}$	−79912.2
$P{-}an$	−17811.6	$Pc{-}n$	−15208.8	$Pnnb$	−20675.9	$A(bc)aa$	−84977.6
$P{-}c{-}$	43.9	$Pca{-}$	−19708.3	$Pnnn$	−25342.8	$I{-}{-}{-}$	−53517.0
$P{-}ca$	−10066.2	$Pcaa$	−25172.2	$C{-}{-}{-}$	−58406.3	$I{-}{-}(ab)$	−54679.6
$P{-}cb$	−423.6	$Pcab$	−20179.6	$C{-}{-}2_1$	−60009.3	$I{-}(ac){-}$	−53922.8
$P{-}cn$	−9740.5	$Pcan$	−24846.5	$C{-}{-}(ab)$	−59569.0	$I{-}cb$	−55085.4
$P{-}n{-}$	−14754.6	$Pcc{-}$	−5386.6	$C{-}c(ab)$	−61632.1	$I(bc){-}{-}$	−53492.0
$P{-}na$	−20218.6	$Pcca$	−15500.5	$Cc{-}{-}$	−67348.9	$Ic{-}a$	−54654.5
$P{-}nb$	−15225.9	$Pccb$	−5858.0	$Cc{-}(ab)$	−68511.5	$Iba{-}$	−53897.8
$P{-}nn$	−19892.9	$Pccn$	−15174.8	$Ccc{-}$	−67809.0	$Ibca$	−55060.4
$Pb{-}{-}$	50.3	$Pcn{-}$	−20188.9	$Ccc(ab)$	−68971.7	$F{-}{-}{-}$	−94799.8
$Pb{-}a$	−10055.8	$Pcna$	−25652.9	$B{-}{-}{-}$	−56444.7	$F{-}dd$	−109509
$Pb{-}b$	−436.1	$Pcnb$	−20660.3	$B{-}2_1{-}$	−56429.5	$Fd{-}d$	−110222
$Pb{-}n$	−9745.2	$Pccn$	−25327.2	$B{-}{-}b$	−56575.6	$Fdd{-}$	−110833
$Pba{-}$	−12620.7	$Pn{-}{-}$	−5428.8	$B{-}(ac){-}$	−56850.6	$Fddd$	−111193

case of the number of reflections available for estimating μ being too small (*e.g.* less than 5), then an alternative strategy, such as simply taking the arithmetic mean of all the reflection intensities present in the pattern, may be preferred.

7. Performance of the algorithm

Diffraction data (Fig. 3) collected from three crystalline samples are used to illustrate the performance of the algorithm. Synchrotron X-ray data were collected from a 1 mm capillary containing dopamine hydrobromide (Shankland *et al.*, 1996) on Station BM16 of the ESRF, using an incident wavelength of 0.6528 Å. Laboratory X-ray data were collected from a sample of decafluoroquarterphenyl (Smrčok *et al.*, 2000) using Mo $K\alpha$ radiation and time-of-flight neutron diffraction data were collected from a crystalline sample of ZrW_2O_4 (Evans *et al.*, 1999) on the high-resolution powder diffractometer of the ISIS spallation neutron source. In each case, unit cells were determined by conventional indexing procedures and correlated integrated intensities were extracted from the data in appropriate space groups possessing no systematic absences, using Pawley refinement programs based upon the Cambridge Crystallographic Subroutine Library (David *et al.*, 1992). These extracted intensities alone were used as input to a space-group-determination program written in C++ that implements the above algorithm. As outlined in §6, values of μ were determined directly from uncorrelated peaks in the data. Whilst evaluation of the probability tables is computationally non-trivial, the program typically takes only a few seconds to execute on a modern personal computer.

7.1. Results

Table 1 lists each of the extinction symbols for the orthorhombic Laue class and the corresponding probabilities calculated for the dopamine hydrobromide data, with the probabilities expressed as

$$\ln[p(E_{\rm gr}|\mathbf{I}^{\rm P})/p(E_{---}|\mathbf{I}^{\rm P})],$$

where $E_{\rm gr}$ is the extinction group being tested and E_{---} is an appropriate space group that possesses no systematic absences for the orthorhombic Laue class. Table 2 shows the ten most probable extinction-group choices ranked by probability, with the associated reflection conditions for each of these choices. Likewise, Tables 3 and 4 show ranked lists of extinction symbols, corresponding probabilities and associated reflection conditions for the monoclinic decafluoroquarterphenyl and cubic zirconium tungstate data sets, respectively. Table 5 summarizes the most probable extinction symbols (and corresponding space groups) for each of the compounds examined. The correct space groups for the compounds are also shown.

7.2. Discussion

A careful examination of Tables 1 to 4 shows the high degree of discrimination afforded by the algorithm outlined in this paper. Taking the specific example of dopamine hydrobromide, Table 1 shows that the majority of possible extinction

Table 2
Extinction symbols, probabilities expressed as $\ln[p(E_{\rm gr}|\mathbf{I}^{\mathbf{P}})/p(E_{---}|\mathbf{I}^{\mathbf{P}})]$ and reflection conditions for the 10 most probable choices calculated for dopamine hydrobromide.

Symbol	Probability	*hkl*	0*kl*	*h*0*l*	*hk*0	*h*00	0*k*0	00*l*
Pbc–	97.9		*k*	*l*			*k*	*l*
Pb––	50.3		*k*				*k*	
P–*c*–	43.9			*l*				*l*
P–2_12_1	21.1						*k*	*l*
P–2_1–	15.2						*k*	
P––2_1	5.9							*l*
P–––	0							
Pbcb	−388.5		*k*	*l*	*k*		*k*	*l*
P–*cb*	−423.6			*l*	*k*		*k*	
Pb–*b*	−436.1		*k*		*k*		*k*	
P––*b*	−471.3				*k*		*k*	

Table 3
Extinction symbols, probabilities expressed as $\ln[p(E_{\rm gr}|\mathbf{I}^{\mathbf{P}})/p(E_{---}|\mathbf{I}^{\mathbf{P}})]$ and reflection conditions for the monoclinic (*b* unique) choices for the compound decafluoroquarterphenyl.

Symbol	Probability	*hkl* 0*kl* *hk*0	*h*0*l* *h*00 00*l*	0*k*0
*I*1*a*1	153.1	*h*+*k*+*l*	*h*,*l*	*k*
*I*1–1	129.1	*h*+*k*+*l*	*h*+*l*	*k*
*P*1 2_1/*n* 1	60.1		*h*+*l*	*k*
*P*1 2_1/*c* 1	59.6		*l*	*k*
*P*1*n*1	57.1		*h*+*l*	
*P*1*c*1	56.6		*l*	
*P*1 2_1/*a* 1	56.1		*h*	*k*
*P*1*a*1	53.1		*h*	
*P*1 2_1 1	3.0			*k*
*P*1–1	0			
*A*1*n*1	−3619.1	*k*+*l*	*h*,*l*	*k*
*A*1–1	−3640.9	*k*+*l*	*l*	*k*
*C*1*c*1	−4218.7	*h*+*k*	*h*,*l*	*k*
*C*1–1	−4247.9	*h*+*k*	*h*	*k*

symbols for this orthorhombic data set are extremely improbable, with only six extinction symbols being more probable than the one corresponding to the extinction group having no systematic absences. Of these six possibilities (Table 2), it is clear that *Pbc*– is much more probable, given the data, than the next choice, *Pb*––, which is in turn much more probable than *P*–*c*– *etc*. It is not surprising that the second- to sixth-ranked choices are more probable than *P*––– since all contain subsets of the reflection conditions for the most probable choice *Pbc*–. Similarly, those that are less probable than *P*––– all contain additional conditions that are not met by the data. In particular, it is clear from Table 1 that the possibility of the diffraction data corresponding to a face-centred extinction symbol is extremely remote.

Tables 3 and 4 show that this high degree of discrimination also holds for the monoclinic and cubic data sets examined. Indeed, in the cubic case, only one extinction symbol proved to be more likely than the one corresponding to the extinction group having no systematic absences.

Table 5 shows that, in each case, the correct extinction symbol for the data has been determined. It is useful to recall that it is the correct extinction symbol that is determined from the data; if multiple space-group choices exist within the extinction class, then additional information must be brought to bear on the problem of the final space-group choice. Often, this is a trivial matter, as the space-group choice is likely to reflect some intrinsic molecular property. For example, the decafluoroquarterphenyl molecule possesses a centre of symmetry and so a satisfactory structure solution and subsequent Rietveld refinement are obtained in *I*2/*a*. As outlined in

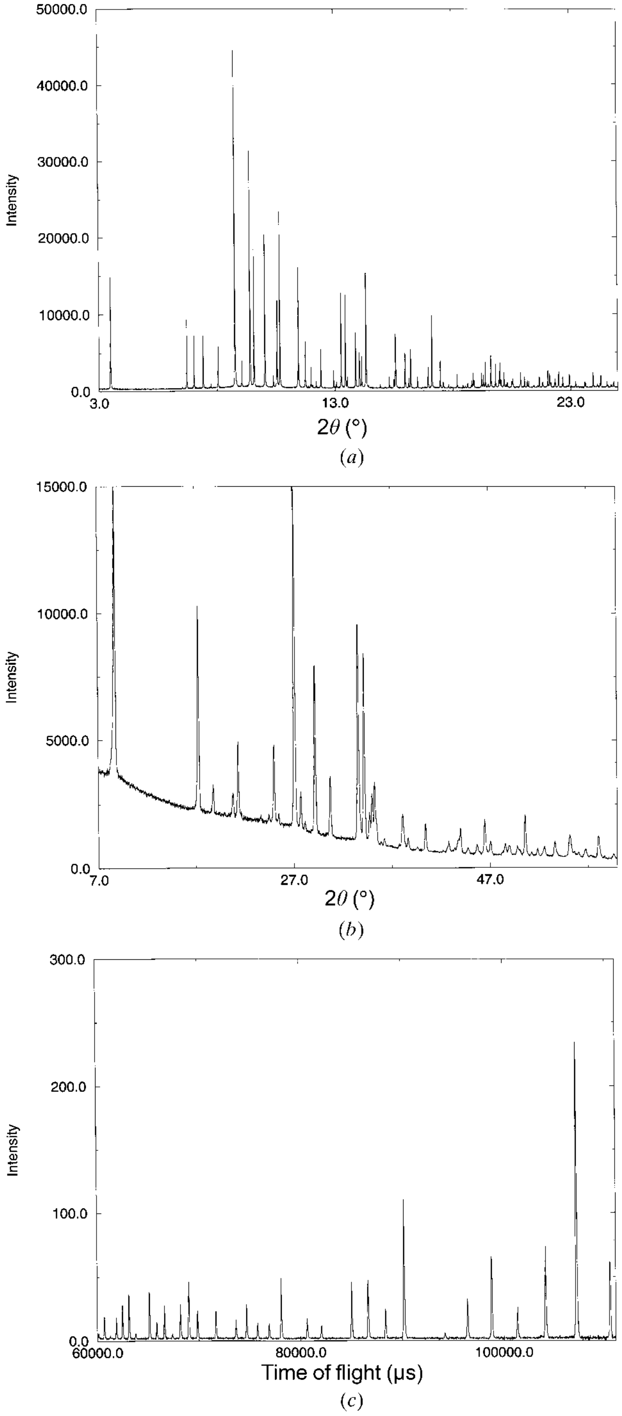


Figure 3
Diffraction data used to test the algorithm: (*a*) dopamine hydrobromide, (*b*) decafluoroquarterphenyl and (*c*) zirconium tungstate.

Table 4
Extinction symbols, probabilities expressed as $\ln[p(E_{\rm gr}|\mathbf{I}^{\mathbf{P}})/p(E_{---}|\mathbf{I}^{\mathbf{P}})]$ and reflection conditions for the cubic choices for the compound zirconium tungstate.

Symbol	Probability	*hkl*	0*kl*	*hhl*	00*l*
$P2_1(4_2)$– –	1.6				*l*
P– – –	0				
$P4_1$– –	−98.6				$l = 4n$
Pn– –	−1110.7		$k+l$		*l*
Pa– –	−1694.4		*k*		*l*
P– –*n*	−8381.1			*l*	*l*
Pn–*n*	−9976.1		$k+l$	*l*	*l*
I– – –	−11321.9	$h+k+l$	$k+l$	*l*	*l*
$I4_1$– –	−11422.2	$h+k+l$	$k+l$	*l*	$l = 4n$
Ia– –	−12625.8	$h+k+l$	k,l	*l*	*l*
I– –*d*	−12942.4	$h+k+l$	$k+l$	$2h+l = 4n,l$	$l = 4n$
Ia–*d*	−13873.3	$h+k+l$	k,l	$2h+l = 4n,l$	$l = 4n$
F– – –	−29211.4	$h+k,h+l,k+l$	k,l	$h+l$	*l*
$F4_1$– –	−29311.7	$h+k,h+l,k+l$	k,l	$h+l$	$l = 4n$
Fd– –	−29568.6	$h+k,h+l,k+l$	$k+l = 4n,k,l$	$h+l$	$l = 4n$
F– –*c*	−30272.4	$h+k,h+l,k+l$	k,l	h,l	*l*
Fd–*c*	−30629.6	$h+k,h+l,k+l$	$k+l = 4n,k,l$	$h\,l$	$l = 4n$

Table 5
The most probable extinction symbols, corresponding space-group choices and true space groups for the three diffraction data sets.

Compound	Extinction symbol	Space-group choices	True space group
Dopamine hydrobromide	*Pbc*–	*Pbca*, $Pbc2_1$	$Pbc2_1$
Decafluoroquarterphenyl	*I*1*a*1	*Ia*, *I*2/*a*	*I*2/*a*
Zirconium tungstate	$P2_1$– –; $P4_2$– –	$P2_13$, $P4_232$	$P2_13$

§6, although discrimination between centrosymmetric and non-centrosymmetric space groups can be achieved in principle through an examination of the distribution of structure-factor magnitudes, we have found that, in practice, the information content of powder diffraction data does not normally allow such a distinction.

The discussion so far has focused on discriminating space groups by examination of reflection intensities. It is worth noting that space groups may also be determined by analysis of Harker lines and sections in a Patterson map. By invoking Parseval's theorem, it is clear that the present analysis is equivalent to such a process. The likelihood function translated from reciprocal space to real space is a weighted difference between observed and model Patterson maps. The probability expressions in (5) enforce upon the model Patterson map the symmetry of a random-atom structure that obeys the space-group symmetry. The model Patterson map is therefore that corresponding to a random-atom structure with the caveat that the Harker lines and sections are consistent with the model space-group symmetry.

8. Conclusions

It has been shown that the problem of space-group determination from powder diffraction data can be placed on a quantitative basis by the application of probability theory to correlated integrated intensities extracted directly from a powder diffraction pattern. The methodology developed removes the need to make subjective judgements about whether or not a peak is 'present' or 'absent' and extends the range of data used in the decision-making process from a few isolated peaks at the start of the pattern to the entire diffraction pattern.

We gratefully acknowledge useful discussions with Dr D. S. Sivia during the early stages of this work.

References

David, W. I. F., Ibberson, R. M. & Matthewman, J. C. (1992). Report RAL-92-032. Rutherford Appleton Laboratory, Chilton, Oxon, England.

Evans, J. S. O., David, W. I. F. & Sleight, A. W. (1999). *Acta Cryst.* B**55**, 333–340.

International Tables for Crystallography (1983). Vol. A, edited by Th. Hahn. Dordrecht: Kluwer Academic Publishers. [Revised editions: 1987, 1989, 1993, 1995, 1996. Abbreviated as *ITCA*(1983).]

Le Bail, A., Duroy, H. & Fourquet, J. L. (1988). *Mater. Res. Bull.* **23**, 447–452.

Pawley, G. S. (1981). *J. Appl. Cryst.* **14**, 357–361.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipies in C.* Cambridge University Press.

Shankland, N., Love, S. W., Watson, D. G., Knight, K. S., Shankland, K. & David, W. I. F. (1996). *J. Chem. Soc. Faraday Trans.* **92**, 4555–4559.

Smrčok, L., Koppelhuber-Bitschnau, B., Shankland, K., David, W. I. F., Tunega, D. & Resel, R. (2000). *Z. Kristallogr.* Submitted.

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