Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 30 November 2001
Accepted 18 March 2002

A maximum-likelihood method for global-optimization-based structure determination from powder diffraction data

Anders J. Markvardsen, William I. F. David* and Kenneth Shankland

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

A maximum-likelihood algorithm has been incorporated into a crystal structure determination from a powder diffraction data framework that uses an integrated-intensity-based global optimization technique. The algorithm is appropriate when the structural model being optimized is not a complete description of the crystal structure under study.

1. Introduction

Recent years have seen significant developments in the methodologies available for tackling the problem of crystal structure determination from powder diffraction data and the range of applicability of the principal techniques (global optimization and direct methods) has been considerably extended (David *et al.*, 2002). In particular, global optimization methods have found great utility in solving molecular organic crystal structures where the known chemical connectivity of the molecule under study can be easily translated into a three-dimensional molecular model. It is the prior chemical information embedded in this model that provides the fundamental strength of the global optimization approach. However, it also introduces a fundamental weakness, in that the success of the structure determination depends crucially on the accuracy of the input molecular model. In many cases, the level of prior chemical knowledge is such that the input model is quite accurate and the correctness of a trial crystal structure produced in the global optimization search can be assessed in a meaningful way. The assessment is normally performed by comparing diffraction data calculated from a trial structure with diffraction data measured from a sample, using a least-squares figure of merit (FOM). That is, either as the weighted sum of squared deviations between the observed ($y_i^{\rm obs}$) and calculated (y_i) diffraction patterns using $\chi^2_{\rm profile} = \sum_i w(y_i^{\rm obs} - y_i)^2$ (Young, 1993), or the weighted sum of squared deviations between observed ($I_i^{\rm data}$) and calculated (I_i) integrated intensities of the diffraction pattern [see equation (7)].

That crystal structures can be solved when the input model is slightly in error has been demonstrated elsewhere. For example, omission of hydrogen atoms in a model of the anti-ulcer drug famotidine had no significant impact on the success rate in obtaining the crystal structure of famotidine form B from powder diffraction data using simulated annealing (McBride, 2000). More significant errors might be introduced owing to a lack of prior knowledge of phenomena such as disorder or as the result of approximations made in order to simplify the input model. For example, there may be too many independent fragments in a crystal structure for the structure to be solved by simultaneously optimizing the positions, orientations and conformations (*i.e.* the structural parameters) of all the fragments. In this case, the structure solution process may be broken down into a series of smaller, seemingly more tractable, problems. Initially, one or more of the fragments in the asymmetric unit cell is simply ignored and the structural parameters of the remaining components are determined. In a subsequent step, the determined components are then kept fixed and the structural parameters of the remaining components are optimized.[1] However, this attractive approach is hampered by the fact that the principle of least-squares refinement is poorly justified when the input model is significantly in error. Put straightforwardly, how can certain components of the crystal structure be optimized against the measured diffraction data when other components that contributed to that measured data are ignored? Problems very similar to those detailed above have been the focus of research in other areas of crystallography, in particular within the field of macromolecular crystallography. For example, the maximum-likelihood approach has recently been applied to improve the threshold for protein structures that might be used in molecular replacement (Read, 2001). In the area of protein structure refinement, maximum likelihood has been demonstrated to be much better than the traditional least-squares approach (Pannu & Read, 1996; Murshudov *et al.*, 1997). The likelihood function has also been utilized in the field of fibre diffraction (Mu & Makowski, 2000). For further references on the likelihood approach see, for instance, Read (1997) and Bricogne (1997*a*) and references therein.

This paper describes how a maximum-likelihood approach can be introduced into a framework for global-optimization-based crystal structure determination from powder diffraction

[1] This is the global optimization equivalent of the standard sequence of direct (or Patterson) methods, partial fragment refinements and difference Fourier calculations that is often employed in structure determination from powder diffraction data.

data. The approach is explained in the context of a real-life problem, that of optimizing the structural parameters of a remacemide ion against diffraction data collected from a powder sample of remacemide nitrate. In the terminology of this paper, the fragment to be optimized against the measured diffraction data is known as *frag*, whilst the fragment that is not being optimized, but whose contribution to the diffraction data is still being considered, is known as *blur*. Thus the positively charged remacemide ion is referred to as *frag*, whilst the negatively charged nitrate group is referred to as *blur*. A second example, remacemide acetate, is also discussed.

2. Maximum likelihood in a powder diffraction context

Firstly, it is assumed that the scattering contribution of the *blur* fragment is randomly distributed throughout the unit cell. The consequence of introducing such a *blur* may be expressed probabilistically as follows. Denote by $\mathbf{I}^{frag} = (I_1^{frag}, \ldots, I_N^{frag})$ the intensities as calculated from the *frag* component, and by $\mathbf{I} = (I_1, \ldots, I_N)$ a set of N intensities. Then, probabilistically, the consequence of introducing a *blur* component may be written as the conditional probability density $p(\mathbf{I}|\mathbf{I}^{frag})$, *i.e.* the probability density of some set of intensities given the position, orientation and conformation of the *frag* component and the fact that the *blur* component is randomly distributed throughout the unit cell. Such a probability density was first studied by Wilson (1949), who considered the statistical consequence of having all the atoms randomly distributed in the unit cell and the application of the central limit theorem to this problem. Subsequently, a number of contributions were made to calculating this type of distribution under different circumstances; see, for instance, Read (1997, and references therein). Of particularly relevance to this study is the work where the *blur* probability density is found for the case of a set of completely overlapping reflections in a powder pattern (Bricogne, 1991). Consider the case where n_a acentric and n_c centric reflections are completely overlapped. The intensity associated with these overlapping reflections is

$$I_i = \sum_{i=1}^{n_a} p_{\mathbf{h}_i}(A_{\mathbf{h}_i}^2 + B_{\mathbf{h}_i}^2) + \sum_{i=1}^{n_c} p_{\mathbf{k}_i} A_{\mathbf{k}_i}^2, \qquad (1)$$

where $p_{\mathbf{h}_i}$ is the multiplicity factor for reflection $\mathbf{h}_i$. By performing the appropriate integral over a surface of a hypersphere of dimension $n-1$ (Bricogne, 1991), the *blur* probability distribution for I_i can be found to be

$$p(I_i|I_i^{frag}) = \frac{1}{2\Sigma_i^{blur}}\left(\frac{I_i}{D_i^2 I_i^{frag}}\right)^{n/4-1/2} \exp\left(-\frac{I_i + D_i^2 I_i^{frag}}{2\Sigma_i^{blur}}\right) \times \mathrm{I}_{n/2-1}\left[\frac{D_i(I_i I_i^{frag})^{1/2}}{\Sigma_i^{blur}}\right], \qquad (2)$$

where $n = 2n_a + n_c$, $\mathrm{I}_n(z)$ is the modified Bessel function of order n. $\Sigma_i^{blur} = |G|(\Sigma_{N,i} - D_i^2\Sigma_{P,i})$ and D_i may be referred to as the Luzzati weighting factor (Read, 2001), which allows for the possibility of incorporating uncertainty into the known part of the structure (*i.e.* the *frag* component in the notation used in §1). With $D_i = 1$, (2) reduces to equation (3.6a) in Bricogne (1991). In this work, we report practical examples (see §4) where substantial benefit is found by employing (2) with $D_i = 1$, *i.e.* where *frag* is treated as being known exactly and therefore the *blur* variance can be written as $\Sigma_i^{blur} = |G|\sum_j f_j(\mathbf{h}_i)^2$, where the index j sums over the atoms of the *blur* component over the whole unit cell and $|G|$ is the number of group elements in the point group of the space group of the crystal. Incorporating additional uncertainty due to errors in the known part can, for example, be achieved straightforwardly by using the approach of Luzzati (1952). Examples of this approach will be reported elsewhere.

The probabilistic model used to derive (2) depends upon the central limit theorem and is thus most accurate when a large number of atoms are included in the blur. Nevertheless, even for a small number of atoms (such as a nitrate or acetate group), the approximation is very good. The dual prior assumptions of random orientation and random positioning of a rigid fragment such as a nitrate group are sufficient to produce a distribution that is close to completely random for all but the lowest $\sin\theta/\lambda$ reflections. Probability distributions associated with 'random-fragment' rather than 'random-atom' behaviour have been discussed theoretically by Bricogne (1997*b*), who expanded the structure-factor distributions in terms of spherical harmonics. The leading term is still the random-atom distribution and for powder diffraction data (where the information content is much lower because of reflection overlap) this term is dominant. The practical validity of the random-atom approximation to the blur probability distribution is to be found in the accuracy with which the location of a 'non-blur' component of a crystal structure can be determined when likelihood optimization (using the random-atom approximation for the blur component) is employed (§4). It is, of course, also worth pointing out that the random-atom approximation has been used with great success in crystallography, particularly in the field of direct methods. For instance, Cochran (1955) uses such an approach to derive the tangent formula and Sim (1960) and Woolfson (1956) 'heavy-atom' weights are derived using the same assumption. Furthermore, single-crystal versions of (2) are extensively used in macromolecular crystallography (see, for instance, Read, 1997, and references therein).

The likelihood function is the probability distribution of the data given the structural parameters (position, orientation and conformation) of the model. These refineable parameters are conveniently summarized by their set of dependent calculated intensities $\mathbf{I}^{frag}$ and the likelihood function may then be written as $L = p(\mathrm{data}|\mathbf{I}^{frag})$. It will be assumed that the data can be summarized by the following multivariable normal distribution:

$$p(\mathbf{I}^{\mathrm{data}}|\mathbf{I}) = (2\pi)^{-N/2}|\mathbf{C}|^{-1/2}\exp[-\tfrac{1}{2}(\mathbf{I}^{\mathrm{data}} - \mathbf{I})^T\mathbf{C}^{-1}(\mathbf{I}^{\mathrm{data}} - \mathbf{I})]. \qquad (3)$$

The intensities $\mathbf{I}^{\mathrm{data}} = (I_1^{\mathrm{data}}, I_2^{\mathrm{data}}, \ldots, I_N^{\mathrm{data}})$ may be determined from a least-squares Pawley (1981) refinement or using the iterative Le Bail method (Le Bail *et al.*, 1988; David *et al.*, 2002). The matrix $\mathbf{C}^{-1}$ holds information about correlations

between the intensities of 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. Equation (3) is an accurate representation of the diffraction pattern if either the Pawley (1981) or LeBail method is used for obtaining $\mathbf{I}^{\rm data}$ and $\mathbf{C}^{-1}$. The likelihood function is expressed in integral form as follows:

$$L = \int p(\mathbf{I}|\mathbf{I}^{frag}) p(\mathbf{I}^{\rm data}|\mathbf{I})\, {\rm d}\mathbf{I}, \tag{4}$$

i.e. the integral of the product of the *blur* and data probability distributions referred to in (2) and (3), respectively. If the correlation matrix $\mathbf{C}$ is diagonal, (4) reduces to a product of N one-dimensional integrals, where, by expanding the modified Bessel function in (2) as a power series, the ith one-dimension integral may be written as the infinite sum

$$L_i = \frac{1}{4\Sigma_i^{blur}} \left[\frac{C_{ii}^{1/2}}{2^{1/2}\Sigma_i^{blur}}\right]^{n/2-1} \exp\left[z^2 - \frac{I_i^{frag}}{2\Sigma_i^{blur}} - \frac{(I_i^{\rm data})^2}{2C_{ii}}\right] \times \sum_{k=0}^{\infty} \left[\frac{I_i^{frag} C_{ii}^{1/2}}{2^{3/2}(\Sigma_i^{blur})^2}\right]^k \frac{1}{k!} {\rm i}^{n/2+k-1}{\rm erfc}(z), \tag{5}$$

where ${\rm i}^n{\rm erfc}(z)$ is the repeated integral of the error function (Abramowitz & Stegun, 1970), $z = 2^{-1/2}[C_{ii}^{1/2}/(2\Sigma_i^{blur}) - I_i^{\rm data}/C_{ii}^{1/2}]$ and C_{ii} is the ith diagonal element of $\mathbf{C}$. Equation (5) is the same probability expression as equation (47) in Mu & Makowski (2000) in the context of fibre diffraction and, for $n = 2$ and $n = 1$, it reduces to the probability expressions of equations (17) and (18) in Pannu & Read (1996). However, the overlap inherent in the diffraction pattern ensures that the correlation matrix contains off-diagonal elements and therefore the integral in (4) cannot be written as a product of one-dimensional integrals. Evaluation of the integral is therefore non-trivial, particularly so in the context of a global optimization search algorithm, where it must be evaluated for each and every trial structure. Accordingly, the likelihood FOM

$$\chi^2_{\rm like} = -\log(L) \tag{6}$$

must be able to be evaluated rapidly if it is to be of use in structure determination. The following section shows how this can be achieved. Finally, notice that if $\mathbf{I}^{frag}$ represents the complete content of the asymmetric unit and there is no *blur* fragment, then the *blur* probability density in (2) becomes the δ function and $\chi^2_{\rm like}$ becomes proportional to the usual least-squares χ-squared FOM, *i.e.*

$$\chi^2_{\rm like} \propto \chi^2 = N^{-1}(\mathbf{I}^{\rm data} - \mathbf{I}^{frag})^T \mathbf{C}^{-1} (\mathbf{I}^{\rm data} - \mathbf{I}^{frag}). \tag{7}$$

Thus, the maximum-likelihood approach yields the same FOM when there is no *blur* component. However, when a *blur* component is included, then the likelihood FOM is expected to improve the success rate of the global optimization method, as this FOM incorporates the contribution of the *blur* fragment in a statistically sound manner.

3. Evaluating the likelihood integral

One option for evaluating the likelihood in (4) is the use of a Monte Carlo integration algorithm described previously (Markvardsen *et al.*, 2001). However, whilst such an algorithm has potential in the context of the refinement of crystal structures against powder diffraction data, it is not fast enough for use with global optimization algorithms. These algorithms require the evaluation of $\chi^2_{\rm like}$ for each trial structure and typically many hundreds of thousands of trial structures are evaluated in a single structure-solution run. Taking the structure solution of hydrochlorothiazide from synchrotron powder diffraction data as an example (23 atoms, 8 degrees of freedom, 204 reflections to 1.5 Å resolution, 9726 points in the profile), the *DASH* program (David *et al.*, 2001) evaluates approximately 3500 trial structures per second using its χ^2 FOM when running on a single processor 800 MHz Intel Pentium III-based computer. Ideally, the evaluation of $\chi^2_{\rm like}$ should take place on a comparable timescale such that its introduction into the global optimization framework does not result in excessively long run times. The following approximation is therefore introduced. If the *blur* probability density in (2) is actually a Gaussian distribution with mean value $\tilde{I}_i^{frag}$ and variance $\tilde{\Sigma}_i^{blur}$, then the likelihood integral of (4) can be evaluated analytically and (6) becomes:

$$\chi^2_{\rm like} = \tfrac{1}{2}\ln[(2\pi)^N|\mathbf{\Sigma}|] + \tfrac{1}{2}(\mathbf{I}^{\rm data} - \tilde{\mathbf{I}}^{frag})^T \mathbf{\Sigma}^{-1} (\mathbf{I}^{\rm data} - \tilde{\mathbf{I}}^{frag}), \tag{8}$$

where $\mathbf{\Sigma} = \mathbf{C} + \tilde{\mathbf{\Sigma}}^{blur}$ and $\tilde{\mathbf{\Sigma}}^{blur}$ is an $N \times N$ diagonal matrix with diagonal elements $\tilde{\Sigma}_i^{blur}$. The $\chi^2_{\rm like}$ obtained from (8) can be rapidly evaluated by first block diagonalizing the inverse correlation matrix in (3) and inverting this matrix to obtain $\mathbf{C}$; this only needs to be done once. $\mathbf{\Sigma}$ may then be calculated and this matrix inverted in order to calculate the second term in (8). It is important to note that, whilst N may be large, the size of the individual blocks produced as a result of the diagonalization procedure does not exceed the size of an overlapping clump of reflections. In a typical high-resolution diffraction pattern collected to modest spatial resolution, overlapping clumps seldom exceed 10 reflections in size. Hence $\mathbf{C}$ and $\mathbf{\Sigma}$ will be block-diagonal with a maximum block size of 10×10; such matrices are trivial to invert.

Given that (8) can be evaluated quickly and efficiently, it appears promising in the context of structure solution. However, for it to be useful, it must approximate (4) as closely as possible and this involves obtaining suitable values of the mean value $\tilde{I}_i^{frag}$ and variance $\tilde{\Sigma}_i^{blur}$. The following method is used. $\partial \ln[p(I_i|I_i^{frag})]/\partial I_i = 0$ implies:

$$(n-2)z_i^{-1} + A_i z^{-1/2} = y_i, \tag{9}$$

where $z_i = I_i I_i^{frag}(\Sigma_i^{blur})^{-2}$, $A_i = {\rm I}_{n/2}(z_i^{1/2})/{\rm I}_{n/2-1}(z_i^{1/2})$ and $y_i = \Sigma_i^{blur}/I_i^{frag}$. For $n \geq 3$, (9) has a unique solution for all values of the ratio y_i (since y_i is always positive). If we denote the values of z_i and I_i which satisfy (9) by $\tilde{z}_i$ and $\tilde{I}_i$, it is seen that for small values of y_i we have $\tilde{z}_i^{-1} \rightarrow y_i^2$ (which corresponds to saying $\tilde{I}_i \cong I_i^{frag}$) and for large values of y_i we have $\tilde{z}_i^{-1} \rightarrow (y_i - 1/n)/(n-2)$. Hence, (9) may be solved by a simple quadratic interpolation. The unique solution $\tilde{I}_i$ from (9)

is used to estimate $\tilde{I}_i^{frag}$ when $n \geq 3$. When $n = 2$ and $n = 1$, some special care is needed in finding an estimate for $\tilde{I}_i^{frag}$. When $n = 2$, $\tilde{I}_i^{frag}$ is taken to be the $\tilde{I}_i$ solution from (9) for $y_i < 1/2$; otherwise, it is set to zero. As a result of the I_i^{-1} singularity at the origin of the *blur* probability distribution in (2) when $n = 1$, it is found that in order to get good estimates for $\tilde{I}_i^{frag}$ for this case, then $\tilde{I}_i^{frag}$ is selected as the $\tilde{I}_i$ solution to the equation $A_i z^{-1/2} = y_i$ when $y_i < 1$; otherwise it is set to zero.

The final step is to find good estimates for $\tilde{\Sigma}_i^{blur}$. For $n \geq 3$, these are simply taken to be $\tilde{\Sigma}_i^{blur} = -\{\partial^2 \ln[p(I_i|I_i^{frag})]/\partial I_i^2|_{I_i=\tilde{I}_i^{frag}}\}^{-1}$ and substituting (9) into this expression, then $\tilde{\Sigma}_i^{blur}$ can be written as:

$$\tilde{\Sigma}_i^{blur} = (\Sigma_i^{blur} y_i)^2 \{[(1 - n/2)y_i + (ny_i - 1)/4]z_i^{-1} + y_i^2/4\}^{-1}|_{I_i=\tilde{I}_i^{frag}}. \quad (10)$$

Notice that $\tilde{\Sigma}_i^{blur}$ has dimension *intensity squared* while Σ_i^{blur} has dimension *intensity*. When $n = 2$, (10) is used for calculating $\tilde{\Sigma}_i^{blur}$ when $y_i < 1/2$; otherwise, it is set to

$$\tilde{\Sigma}_i^{blur} = -\{\partial^2 \ln[p(I_i|I_i^{frag})]/\partial I_i^2|_{I_i=0, I_i^{frag}=2\Sigma_i^{blur}}\}^{-1} = 8(\Sigma_i^{blur})^2.$$

When $n = 1$, to again circumvent the singularity of the *blur* probability at the origin,

$$\tilde{\Sigma}_i^{blur} = 4(\Sigma_i^{blur} y_i)^2 \{(y_i - 1)z_i^{-1} + y_i^2\}^{-1}|_{I_i=\tilde{I}_i^{frag}}, \quad y_i < 1; \quad (11)$$

otherwise,

$$\tilde{\Sigma}_i^{blur} = 4(\Sigma_i^{blur} y_i)^2 \{(y_i - 1)z_i^{-1} + y_i^2\}^{-1}|_{I_i=0, I_i^{frag}=\Sigma_i^{blur}} = 6(\Sigma_i^{blur})^2.$$

The justification for the formulae outlined in this section is based on an analysis in which the likelihood functions in (4) and (8) are compared for various test cases as discussed in Appendix *A*. In the context of an initial structural solution, they are certainly sufficiently accurate, as is demonstrated in the following section.

4. Practical implementation and testing

4.1. Incorporation into a global optimization framework

The algorithms outlined above were incorporated into a modified version of the *DASH* crystal structure determination program. Briefly, *DASH* normally operates as follows. A Pawley (1981) fit to the measured diffraction data is used to extract a set of correlated integrated intensities and an associated covariance matrix. The molecular content of the asymmetric unit is described using internal coordinates (*Z*-matrix format) and in cases where there is more than one independent structural fragment in the asymmetric unit, a separate *Z* matrix is input for each fragment. Simulated annealing is employed to optimize the structural parameters of the input fragments against the extracted intensity data using the χ^2 FOM in (7). Multiple runs are normally performed from different random start points in order to ensure that the global minimum in the χ^2 search space has been located.

In the modified version of *DASH* (henceforth referred to as *ML-DASH*), the FOM employed is $\chi^2_{\rm like}$ as given in (8) but ignoring the slowly varying first term in that equation. Furthermore, in order to put $\chi^2_{\rm like}$ on the same scale as the χ^2 in (7), $\chi^2_{\rm like}$ is taken to be:

$$\chi^2_{\rm like} = N^{-1}(\mathbf{I}^{\rm data} - \tilde{\mathbf{I}}^{frag})^T \boldsymbol{\Sigma}^{-1}(\mathbf{I}^{\rm data} - \tilde{\mathbf{I}}^{frag}). \quad (12)$$

Further, at the fragment input stage, the option of specifying whether or not an input fragment should be treated as a *frag* to be optimized or as a *blur* to be incorporated into the calculations, is provided.

4.2. Reference crystal structure

Remacemide is an anti-convulsant agent synthesized at AstraZeneca, Loughborough, England, whose crystal structure and the crystal structures of many of its salt forms have been investigated by both single-crystal X-ray diffraction and X-ray powder diffraction (McBride, 2000). The compound remacemide nitrate ($C_{17}H_{21}N_2O^+ \cdot NO_3^-$, Fig. 1) was selected as a suitable example for testing the maximum-likelihood method. The acetate salt of remacemide was also examined (§4.6).

The crystal structure of the nitrate salt form (a = 11.7278, b = 8.9339, c = 15.8738 Å, β = 95.95°, $P2_1/a$) was first solved from high-resolution powder diffraction data (Fig. 2) collected at station BM16 of the ESRF in Grenoble, France. The crystal structure solution was obtained using *DASH*, simultaneously optimizing the structural parameters of both the remacemide molecule and the nitrate ion. Full details of the structure are given elsewhere (McBride, 2000) and only relevant points are

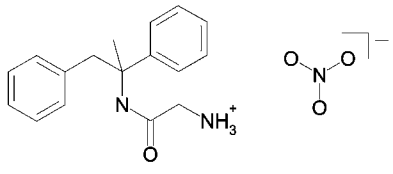

Figure 1
The molecular structure of remacemide nitrate.

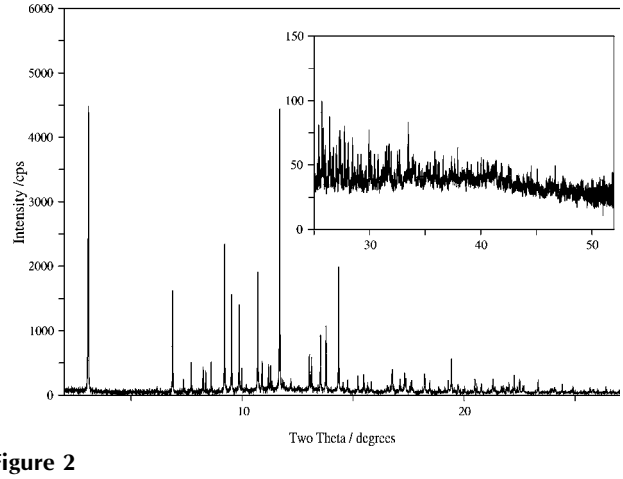

Figure 2
A plot of the diffraction data collected from a sample of remacemide nitrate contained in a 1.0 mm capillary on diffractometer BM16 of the ESRF. The incident wavelength was 0.85070 Å. The inset graph shows an enlarged view of the high-angle region of the pattern.

summarized here. The success rate in finding the correct crystal structure over many repeat runs was ~40%. Subsequent slack-constrained Rietveld refinement of the crystal structure gave an excellent fit to the data with the following profile agreement factors over the data range 2–40°: $\chi^2_{\rm profile}$ = 1.83, R_p = 9.20%, R_{wp} = 9.99% and R_E = 7.38%. The refined structure, shown in projection in Fig. 3, was subsequently verified by laboratory single-crystal X-ray diffraction.

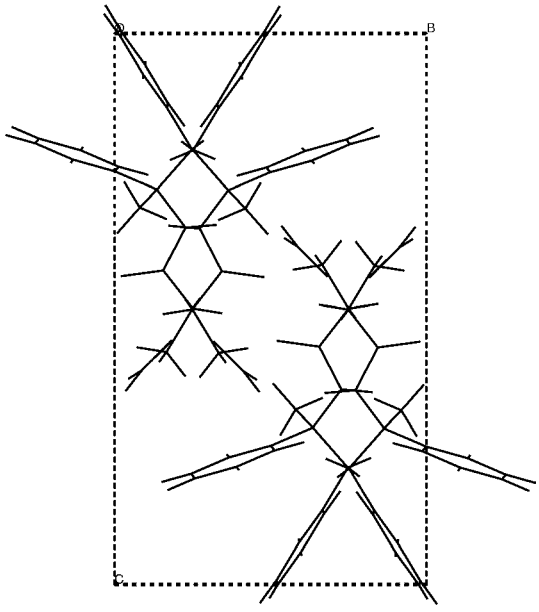

Figure 3
The refined crystal structure of remacemide nitrate, viewed in projection down the *a* axis.

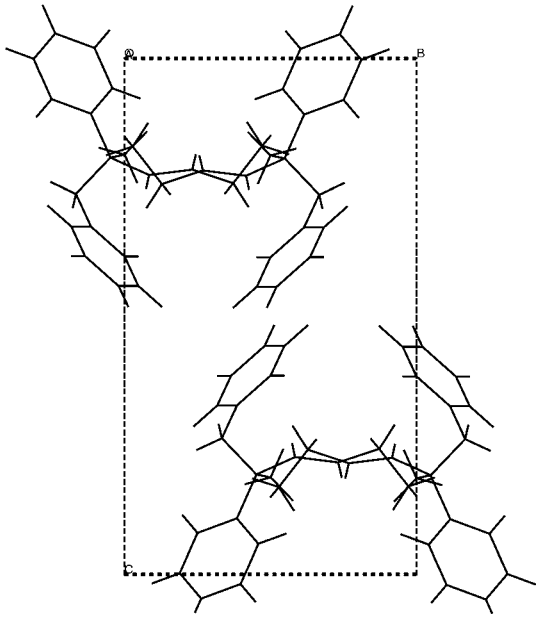

Figure 4
Remacemide nitrate: the best crystal structure obtained by optimizing only the structural parameters of the remacemide ion against the remacemide nitrate diffraction data, using the conventional least-squares version of *DASH*.

Table 1
Remacemide nitrate: the results of 20 *DASH* runs, in which only the structural parameters of the remacemide ion were optimized against the remacemide nitrate diffraction data; the nitrate ion was not included in the calculations.

Run no.	χ^2	$\chi^2_{\rm profile}$
1	702.8604	20.1543
2	703.5668	20.1726
3	704.3631	20.2236
4	712.5563	20.9228
5	712.6614	21.0366
6	712.2556	21.0401
7	724.8441	21.7439
8	720.4251	21.7912
9	720.3148	21.9812
10	739.1662	22.5121
11	703.3167	20.0848
12	702.6350	20.1593
13	703.4621	20.1652
14	704.7296	20.3086
15	713.4292	21.0807
16	720.5380	21.8804
17	721.5416	22.2367
18	722.1905	22.4415
19	739.2380	22.5844
20	758.2836	23.2101

The remacemide nitrate crystal structure has several features that make it attractive for testing the maximum-likelihood approach in cases where the input molecular structure to be optimized is not a complete description of contents of the asymmetric unit. Firstly, the nitrate ion constitutes a significant (~18%) but not excessive percentage of the total scattering power of the crystal. Secondly, as a small, rigid, independent unit, the nitrate ion is representative of many of the salts and solvents that frequently complicate structure solution from powder diffraction data.[2] Thirdly, the fact that the nitrate ion is a discrete entity makes it a straightforward matter to assess the impact of its omission from a standard *DASH* run and its inclusion as a *blur* in an *ML-DASH* run. Finally, the remacemide nitrate crystal structure has a very characteristic 'cartwheel' motif in projection (see Fig. 3) that enables correct solutions to be quickly identified amidst the plethora of trial structures returned from the global optimization runs.

4.3. Benchmark *DASH* trials

A series of 20 *DASH* runs was performed in which the structural parameters of the remacemide and nitrate ions were optimized against the intensities and associated covariance matrix extracted from the remacemide nitrate data. These runs, which essentially repeated the work of McBride, resulted in answers ranging from the correct crystal structure ($\chi^2 = 130$, $\chi^2_{\rm profile} = 3.9$) to false minima ($\chi^2 = 573$, $\chi^2_{\rm profile}$ = 16.0) with the anticipated success rate of ~40%.

[2] They are therefore often omitted from the initial stages of a global optimization structure solution in order to simplify the search space.

Table 2
Remacemide nitrate: the results of 20 *ML-DASH* runs, in which only the structural parameters of the remacemide ion were optimized against the remacemide nitrate diffraction data; the nitrate ion was used in the calculation of the *blur*.

Run no.	χ^2_{like}	$\chi^2_{profile}$
1	72.59	27.48
2	69.67	28.82
3	69.25	27.56
4	70.19	28.69
5	71.54	34.21
6	69.20	28.81
7	53.11	44.61
8	52.95	44.17
9	68.98	27.48
10	74.32	30.40
11	52.90	45.37
12	74.68	27.36
13	53.00	44.84
14	53.30	43.23
15	71.97	32.65
16	70.45	28.58
17	53.34	45.38
18	53.16	44.56
19	70.05	27.09
20	72.66	27.89

4.4. *DASH* trials ignoring the nitrate ion

A series of 20 *DASH* runs was performed in which *only* the structural parameters of the remacemide ion were optimized against the intensities and associated covariance matrix extracted from the remacemide nitrate data. The nitrate ion was *not* included in the calculations at any stage. The χ^2 and $\chi^2_{profile}$ values for these runs are listed in Table 1 and the best solution with regard to the χ^2 FOM (no. 12) is shown in projection in Fig. 4.

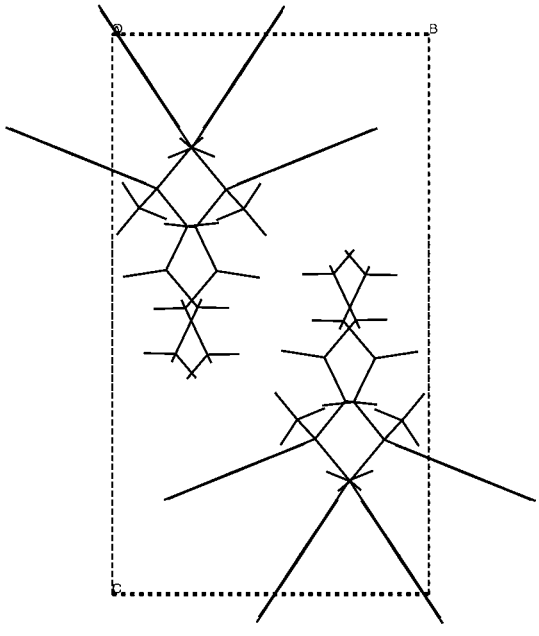

Figure 5
Remacemide nitrate: the best crystal structure obtained by optimizing only the structural parameters of the remacemide ion against the remacemide nitrate diffraction data, using the modified version of *DASH* (*i.e. ML-DASH*) implementing the maximum-likelihood method.

4.5. *ML-DASH* trials using the nitrate ion as a *blur*

A series of 20 *ML-DASH* runs was performed in which *only* the structural parameters of the remacemide ion were optimized against the intensities and associated covariance matrix extracted from the remacemide nitrate data. The nitrate ion was input into *ML-DASH* but was flagged for use as a *blur*. The χ^2_{like} and $\chi^2_{profile}$ values for these runs are listed in Table 2 and the best solution with regard to the χ^2_{like} FOM (no. 11) is shown in projection in Fig. 5 and overlaid upon the refined crystal structure in Fig. 6.

4.6. Example 2: remacemide acetate

The crystal structure of the acetate salt form of remacemide was first solved from high-resolution powder diffraction data, again collected at BM16. The crystal structure (a = 15.4018, b = 6.7683, c = 17.3531 Å, β = 93.13°, $P2_1/c$) solution was obtained

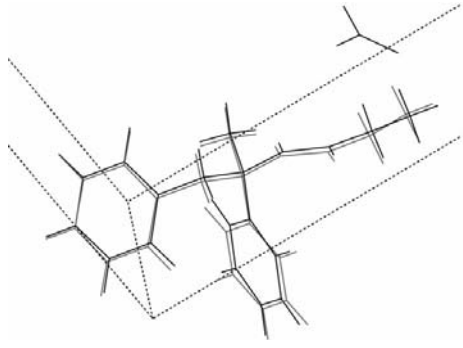

Figure 6
Remacemide nitrate: the asymmetric unit of the best remacemide-only substructure (normal lines) obtained by *ML-DASH*, superimposed upon the asymmetric unit of the refined remacemide nitrate crystal structure (bold lines).

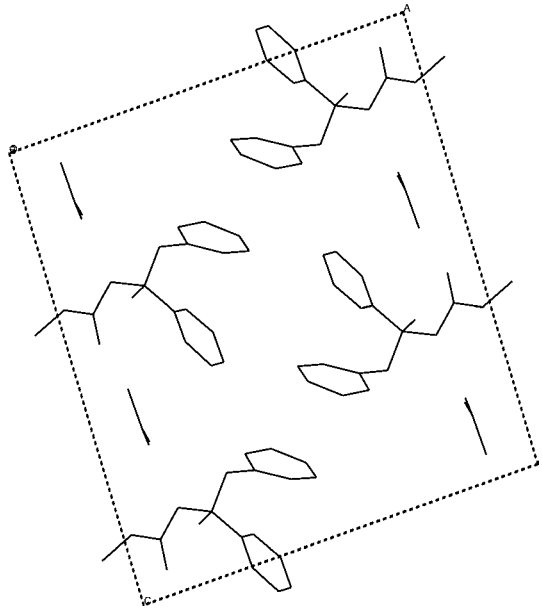

Figure 7
The refined crystal structure of remacemide acetate, viewed in projection down the *b* axis. H atoms are omitted for clarity.

Table 3
Remacemide acetate: the results of ten *DASH* runs, in which only the structural parameters of the remacemide ion were optimized against the remacemide acetate diffraction data; the acetate ion was not included in the calculations.

Run no.	χ^2	$\chi^2_{\rm profile}$
1	260.98	40.41
2	252.45	43.23
3	260.11	42.63
4	252.63	43.87
5	245.49	38.54
6	260.45	43.37
7	252.16	43.23
8	245.86	38.85
9	260.31	43.02
10	245.13	38.52

Table 4
Remacemide acetate: the results of ten *ML-DASH* runs, in which only the structural parameters of the remacemide ion were optimized against the remacemide acetate diffraction data; the acetate ion was used in the calculation of the *blur*.

Run no.	$\chi^2_{\rm like}$	$\chi^2_{\rm profile}$
1	24.68	45.33
2	21.29	54.23
3	21.21	54.23
4	21.66	53.95
5	21.82	54.73
6	21.56	54.37
7	21.02	59.49
8	20.92	55.58
9	21.45	57.53
10	20.78	52.17

using *DASH*, simultaneously optimizing the structural parameters of both the remacemide molecule and the acetate ion. Full details of the structure are given elsewhere (McBride, 2000). Slack-constrained Rietveld refinement of the crystal structure gave an excellent fit to the data with the following profile agreement factors over the data range 2–37°: $\chi^2_{\rm profile}$ = 1.76, R_p = 8.48%, R_{wp} = 9.15% and R_E = 6.89%. The refined structure is shown in projection in Fig. 7. Benchmark *DASH* runs, repeating the 'standard' structure determination process, resulted in a 70% success rate in returning the correct crystal structure of remacemide acetate. A further series of ten standard least-squares *DASH* runs in which *only* the structural parameters of the remacemide ion were optimized against the intensities and associated covariance matrix extracted from the remacemide acetate data was also performed. The acetate ion was *not* included in the calculations at any stage. The χ^2 and $\chi^2_{\rm profile}$ values for these ten runs are listed in Table 3 and the four packing motifs that were observed in the solutions are shown in Fig. 8(*a*), whilst the 'best' structure in terms of its agreement with the refined crystal structure is shown in Fig. 8(*b*). A series of ten *ML-DASH* runs was performed in which *only* the structural parameters of the remacemide ion were optimized against the intensities and associated covariance matrix extracted from the remacemide acetate data. The acetate ion was input into *ML-DASH* but was flagged for use as a *blur*. The $\chi^2_{\rm like}$ and $\chi^2_{\rm profile}$ values for these runs are listed in Table 4 and solutions 2 to 10 are shown overlaid upon the refined crystal structure in Fig. 9.

5. Discussion

It is clear from both the examples discussed in §4 that the least-squares χ^2 FOM is incapable of directing the global optimization of the remacemide fragment when the contribution of the counter ion to the diffraction data is ignored. In

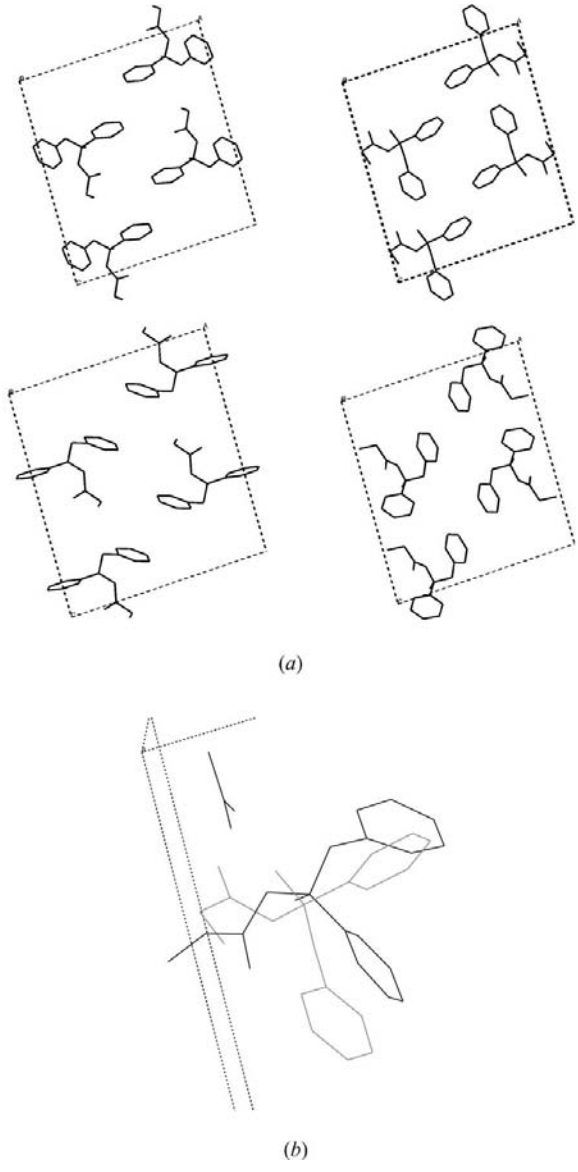

Figure 8
Remacemide acetate. (*a*) The ten solutions returned by *DASH* when optimizing only the structural parameters of the remacemide ion against the remacemide acetate diffraction data can be grouped into four distinct crystal structures. These structures are viewed in projection down the *b* axis and are all distinctly different from the correct structure shown in Fig. 7. (*b*) The solution (normal lines) displaying the best agreement with the correct structure (bold lines). H atoms are omitted for clarity in both figures.

the case of remacemide nitrate, the deliberate omission of 18% of the scattering leads to remacemide-only 'substructures' that are incorrectly placed, oriented and folded as a result of attempting to satisfy the electron density from both the remacemide and the nitrate ions in the crystal. However, when the nitrate ion is incorporated as a *blur*, and the maximum-likelihood FOM is employed, the situation is substantially different.

Examination of both Tables 2 and 4 shows that the $\chi^2_{\rm like}$ FOM does provide the discrimination needed to distinguish between good and bad solutions. In Table 2, seven of the twenty runs (nos. 7, 8, 11, 13, 14, 17 and 18) have values of $\chi^2_{\rm like}$ that are substantially lower than those of the other runs. Examination of the crystal structures associated with these seven runs shows that they all are essentially identical and that they are typified by run no. 11, which is shown in projection in Fig. 5. The 'cartwheel' motif indicates the correctness of the structure in projection whilst Fig. 6 shows that the position, orientation and the conformation of the remacemide ion have been very well determined. The fact that the $\chi^2_{\rm profile}$ values for these seven runs are amongst the highest seen is worthy of comment. The $\chi^2_{\rm profile}$ for any of the solutions obtained where the contribution of the nitrate ion to the diffracted intensity was either ignored or included as a blur is expected to be poor, as the calculation of $\chi^2_{\rm profile}$ is based upon the contribution of the remacemide ion only. However, in the least-squares analysis, $\chi^2_{\rm profile}$ is minimized whilst in the likelihood analysis it is not. Thus it is not surprising that the $\chi^2_{\rm profile}$ associated with the correctly positioned remacemide ion as found from seven of the maximum-likelihood runs is much worse than that associated with the incorrectly positioned remacemide ions returned by conventional least-squares runs.

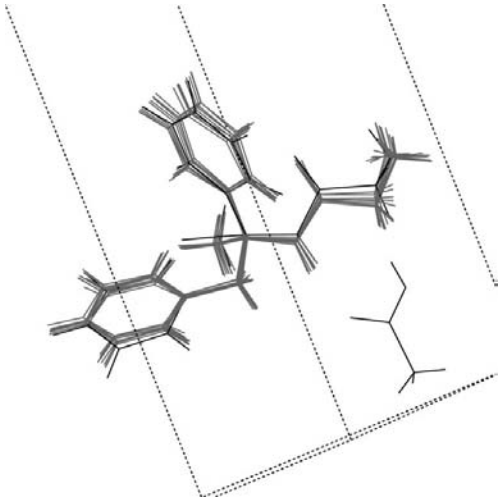

Figure 9
Remacemide acetate: the asymmetric unit of the nine correct remacemide-only substructures (normal lines) obtained by *ML-DASH*, superimposed upon the asymmetric unit of the refined remacemide acetate crystal structure (bold lines).

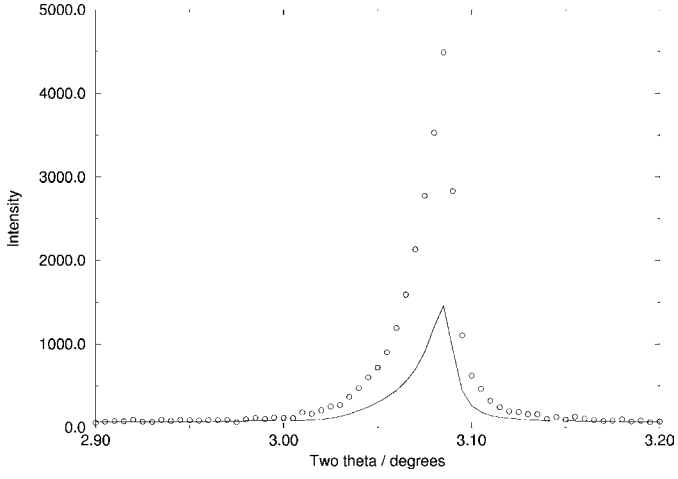

Figure 10
Remacemide nitrate: measured (open circles) and calculated (solid line) diffraction data for the best remacemide-only substructure obtained by *ML-DASH*, *i.e.* the structure shown in Figs. 5 and 6.

The effect of the *blur* component upon the space searched by the global optimization algorithm is dramatically illustrated in Fig. 10, which shows the fit to the first peak in the diffraction profile for the structure shown in Fig. 5. It is clear that the *blur* contribution must confer considerable latitude to the process of fitting the extracted intensity associated with this peak. It is this latitude that ultimately allows the crystal structure to be solved with respect to the remacemide ion component alone.

The results obtained from the remacemide acetate experiments largely echo the results obtained from the remacemide nitrate experiments. The ten *DASH* solutions obtained when the acetate fragment was ignored fell into four distinct groups, as shown in Fig. 8(*a*). It is clear from a comparison with Fig. 7 that in each case the least-squares *DASH* solution has the remacemide ion placed in *approximately* the correct region of space, but that the solutions are sufficiently far away from the correct remacemide-only substructure as to render them useless as starting models for structural refinement; see, for example, Fig. 8(*b*). In stark contrast, any one of the nine *ML-DASH* solutions shown in Fig. 9 would serve as an excellent starting point for structure completion. Solution no. 1, which is actually an incorrect answer, is easily identified by its significantly higher value of $\chi^2_{\rm like}$.

6. Conclusions

The maximum-likelihood method has been introduced to crystal structure determination from powder diffraction data. The results presented in this paper give a strong indication that the likelihood approach has the ability to improve the success rate of global-optimization-based crystal structure determination methods in circumstances where the structural model being optimized is not a complete description of the crystal structure under study. These findings are in broad agreement with other comparisons of least-squares and maximum-likelihood methods in macromolecular crystallography.

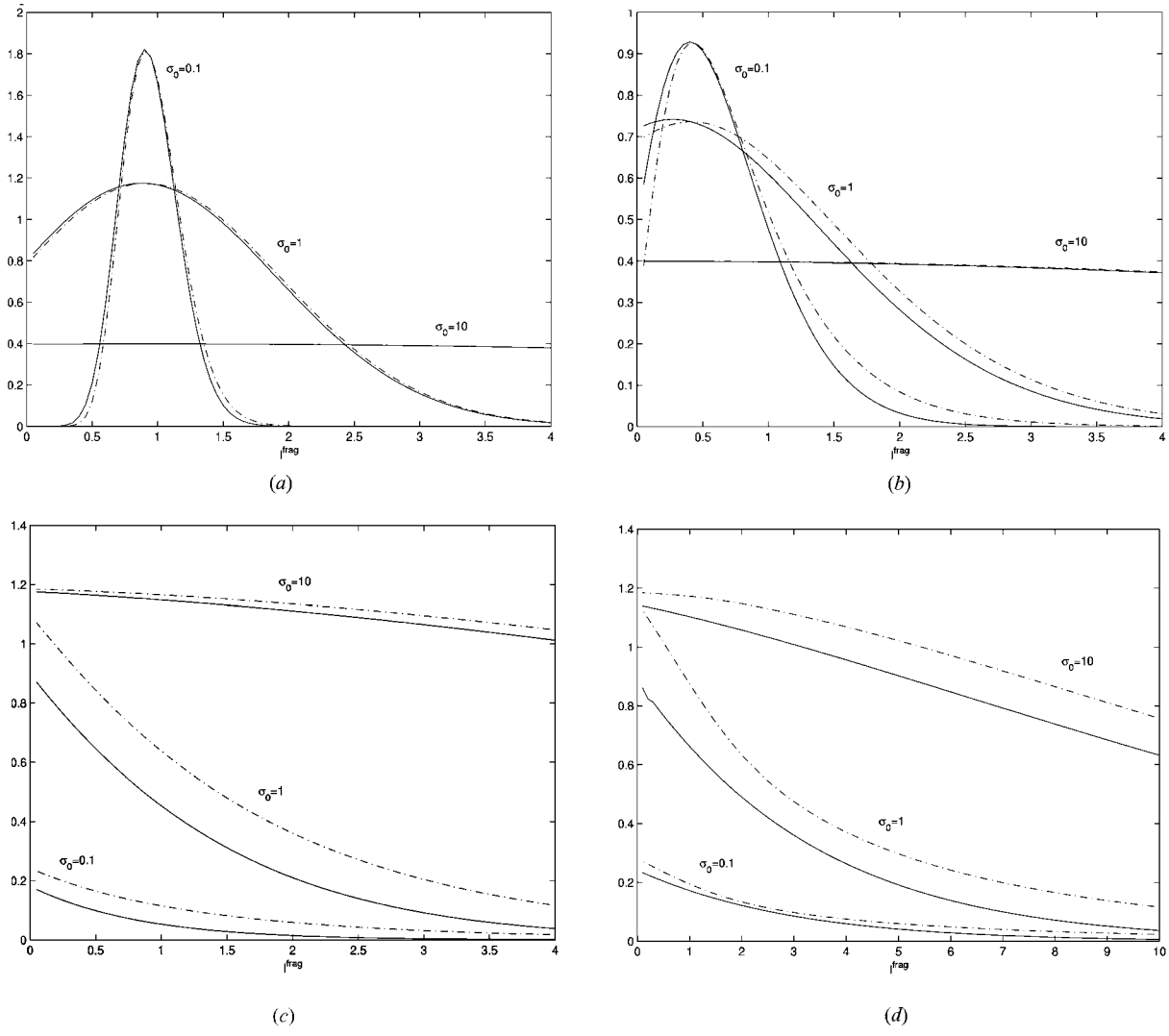

Figure 11
A comparison of the likelihood functions given by (8) (dashed line) and (4) (solid line) with $I^{\rm data} = 1$ and standard deviations $\sigma_0 = 0.1$, $\sigma_0 = 1$ and $\sigma_0 = 10$ as specified in the figures. The individual graphs show the likelihood as a function of I^{frag} for the cases: (*a*) $n = 10$ and $\sigma^{blur} = (\Sigma^{blur})^{1/2} = 0.1$, (*b*) $n = 10$ and $\sigma^{blur} = 0.25$, (*c*) $n = 10$ and $\sigma^{blur} = 0.5$ and (*d*) $n = 3$ and $\sigma^{blur} = 1$. As is expected, when $\sigma^{blur}/I^{\rm data} \leq 0.1$, the likelihood function and its approximation are almost indistinguishable, as illustrated in (*a*). This is a general observation for all n and the remaining graphs only show cases where $\sigma^{blur}/I^{\rm data} > 0.1$. Note that in (*a*) the $\sigma_0 = 1$ line was scaled up by a factor of 3 and the $\sigma_0 = 10$ line scaled up by a factor of 10 for clarity. In the remaining figures, those same lines were scaled up by factors of (*b*) 2 and 10, (*c*) 5 and 30 and (*d*) 5 and 30, respectively.

APPENDIX *A*

A good approximation to the likelihood formula given in (4) must follow the overall behaviour of that formula as a function of the refineable parameters (*i.e.* I_i^{frag}) for fixed values of any data and blur variance, Σ_i^{blur}. However, as the global optimization approach to structure solution involves a search for the global minimum of (4), then the approximation given by (8) may deviate from (4) by some constant value as a function of I_i^{frag} without compromising the validity of the approximation.

Approximating the blur probability in (2) for $n > 3$ by a Gaussian distribution leads to a very good approximation to (8) in any limit. This is not surprising, as when $n > 3$ the blur probability distribution is guaranteed to have a unique maximum away from the origin. As the likelihood is the integral of the product of this blur distribution with a Gaussian distribution [derived from the data, see (3)], the product of these two distributions is expected to be close to another Gaussian with a modified mean value and standard deviation. Figs. 11(*a–d*) shows that this is indeed the case.

Finding a suitable Gaussian approximation to the blur probability distribution is problematic in the following two circumstances: firstly, when $n = 1$ and $\Sigma_i^{blur}/I_i^{frag} \geq 1$ and, secondly, when $n = 2$ and $\Sigma_i^{blur}/I_i^{frag} \geq 1/2$, for reasons explained in §3. However, in both circumstances, these regions of the blur distributions contribute little information in terms of favouring particular I_i values, *i.e.* Σ_i^{blur} is approximately equal to or bigger than I_i^{frag}. Thus, when the value of I_i^{frag} is reached for which $\Sigma_i^{blur}/I_i^{frag} = 1$ ($n = 1$ case) and

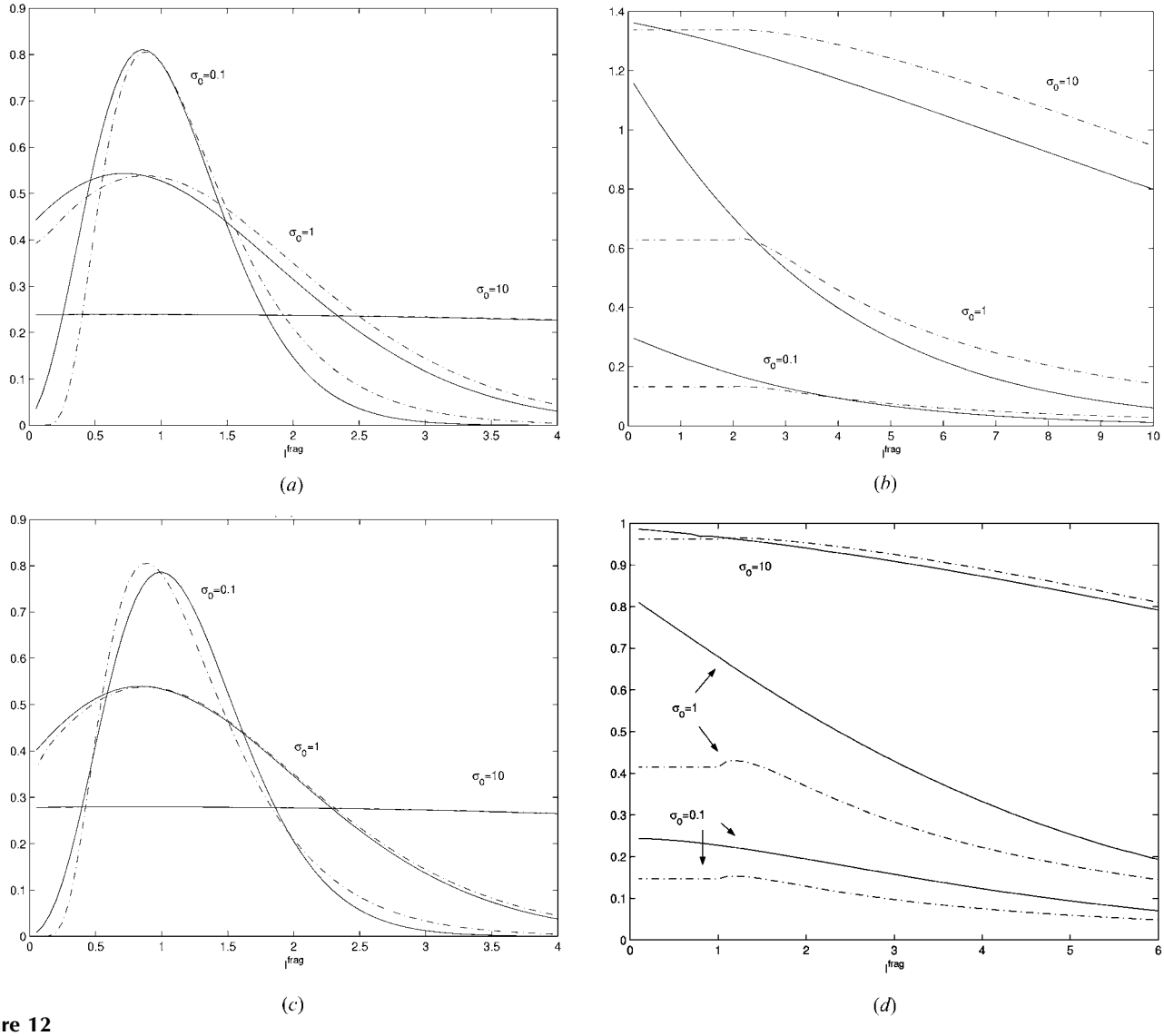

Figure 12
A comparison of the likelihood functions given by (8) (dashed line) and (4) (solid line) with $I^{\rm data} = 1$ and standard deviation $\sigma_0 = 0.1$, $\sigma_0 = 1$ and $\sigma_0 = 10$ as specified in the figures. The figures show the likelihood as a function of I^{frag} for the cases: (*a*) $n = 2$ and $\sigma^{blur} = 0.25$; (*b*) $n = 2$ and $\sigma^{blur} = 1$; (*c*) $n = 1$ and $\sigma^{blur} = 0.25$; (*d*) $n = 1$ and $\sigma^{blur} = 1$. Note that in (*a*) the $\sigma_0 = 1$ line was scaled up by a factor of 1.5 and the $\sigma_0 = 10$ line was scaled up by a factor of 6 for clarity. In the remaining figures, those same lines were scaled up by factors of (*b*) 5 and 35, (*c*) 1.5 and 7 and (*d*) 3 and 25, respectively.

$\Sigma_i^{blur}/I_i^{frag} = 1/2$ ($n = 2$ case) we are *a priori* very uncertain about this calculated intensity and as I_i^{frag} is decreased still further, this uncertainty increases. Hence, this suggests that a suitable approximation is one in which, as I_i^{frag} goes below the threshold values just mentioned, the blur probability distribution is kept constant. This is equivalent to saying that we do not distinguish in terms of likelihood between such I_i^{frag} values. Importantly, this approximation, which is both intuitive and conservative, does not lead to unphysical results in any limits and Figs. 12(*a–d*) shows it to be a reasonable approximation to (4) even in regions where the largest discrepancies are expected, *i.e.* where $\Sigma_i^{blur}/I_i^{\rm data} \leq 0.1$.

It should be clear from §3 that these approximations are applied for block integrals of dimension greater than one in (8). For block integrals of dimension equal to one, the expression in (5) can be evaluated efficiently from a table *via* a simple extrapolation algorithm.

We gratefully acknowledge the cooperation of AstraZeneca Charnwood, and in particular Dr Gerry Steele, in the use of remacemide nitrate and remacemide acetate as examples. We also thank Dr Andy Fitch of the ESRF for his help in collecting the diffraction data and Professor Randy Read of MRC Cambridge for useful discussions on the maximum-likelihood literature in macromolecular crystallography.

References

Abramowitz, M. & Stegun, I. A. (1970). *Handbook of Mathematical Functions.* New York: Dover.
Bricogne, G. (1991). *Acta Cryst.* A**47**, 803–829.

Bricogne, G. (1997*a*). *Methods Enzymol.* **276**, 361–423.
Bricogne, G. (1997*b*). *Methods Enzymol.* **277**, 14–18.
Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.
David, W. I. F., Shankland, K., Cole, J., Maginn, S., Motherwell, W. D. S. & Taylor, R. (2001). *DASH User Manual.* Cambridge Crystallographic Data Centre, Cambridge, England.
David, W. I. F., Shankland, K., McCusker, L. & Baerlocher, C. (2002). Editors. *Structure Determination from Powder Diffraction Data.* Oxford University Press.
Le Bail, A., Duroy, H. & Fourquet, J. L. (1988). *Mater. Res. Bull.* **23** 447–452.
Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.
McBride, L. (2000). PhD thesis, Strathclyde University, Glasgow, Scotland.
Markvardsen, A. J., David, W. I. F., Johnson, J. C. & Shankland, K. (2001). *Acta Cryst.* A**57**, 47–54.
Mu, X. Q. & Makowski, L. (2000). *Acta Cryst.* A**56**, 168–177.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.
Pawley, G. S. (1981). *J. Appl. Cryst.* **14**, 357–361.
Read, R. J. (1997). *Methods Enzymol.* **277**, 110–128.
Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.
Sim, G. A. (1960). *Acta Cryst.* **13**, 511–512.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Woolfson, M. M. (1956). *Acta Cryst.* **9**, 804–810.
Young, R. A. (1993). Editors. *The Rietveld Method.* Oxford University Press.