*Acta Cryst.* (1996). A52, 659–668

# Improved Structure Refinement Through Maximum Likelihood

NAVRAJ S. PANNU[a] AND RANDY J. READ[b]*

[a]*Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta T6G 2H7, Canada, and* [b]*Department of Medical Microbiology and Immunology, University of Alberta, Edmonton, Alberta T6G 2H7, Canada.* *E-mail: rndy@mycroft.mmid.ualberta.ca*

(*Received 4 January 1996; accepted 29 March 1996*)

## Abstract

When crystal structures of proteins or small molecules are used to address questions of scientific relevance, the accuracy and precision of the atomic coordinates are crucial. Accordingly, the atomic model is generally improved by refining it to improve agreement with the observed diffraction data. The refinement of crystal structures is conventionally based on least-squares methods but such procedures are handicapped since conditions necessary for the use of the least-squares target are not satisfied. It is proposed here that refinement should be based on maximum likelihood and two maximum-likelihood targets have been implemented in the program *XPLOR*. Preliminary tests with protein structures give dramatic results. Compared to least-squares refinement, maximum-likelihood refinement can achieve more than twice the improvement in average phase error. The resulting electron-density maps are correspondingly clearer and suffer less from model bias.

## 1. Introduction

To obtain the most accurate possible crystal structure, one typically refines the atomic model to optimize its agreement with the observed diffraction data. However, the quality of the resulting model will depend on the validity of the target function that is optimized. We believe that, since the conventional least-squares target is poorly justified in this case, the refinement procedures are unduly handicapped. A maximum-likelihood target is much better justified and we show that it performs significantly better in macromolecular refinement.

The standard macromolecular refinement programs, *PROLSQ* (Konnert & Hendrickson, 1980), *TNT* (Tronrud, Ten Eyck & Matthews, 1987), *XPLOR* (Brünger, Kuriyan & Karplus, 1987) and *GROMOS* (Fujinaga, Gros & van Gunsteren, 1989), minimize a residual that is the weighted sum of squared deviations between the observed ($|F_o|$) and calculated ($|F_c|$) structure-factor amplitudes, including a relative scale factor $k$:

$$\sum_{hkl} w(|F_o| - k|F_c|)^2. \quad (1)$$

The refinement programs differ primarily in their minimization methods. Even though the atomic model is improved, problems arise because such a least-squares residual is poorly justified, especially early in the refinement. As Silva & Rossmann (1985) have pointed out, what is minimized (ignoring weights) is the r.m.s. deviation between the model electron density and the density computed from Fourier coefficients $|F_o|\exp(i\alpha_c)$. This deviation can be minimized either by improving the model or by introducing systematic errors that obliterate differences from the model in the $|F_o|\exp(i\alpha_c)$ map. Since most macromolecular refinements have an unfavourable parameter-to-observation ratio, the data are typically overfitted, which means that such systematic errors must be introduced.

The least-squares-refinement target could be considered to arise from the principle of maximum likelihood, if the following assumptions hold: the deviation between $|F_o|$ and $k|F_c|$ is a Gaussian, the mean deviation is zero and the standard deviation of the Gaussian is independent of the parameters of the atomic model. This is not true, as shown below, because the errors have a (changing) phase component. For this reason, we should return to first principles and apply a maximum-likelihood analysis to the problem of protein structure refinement, as we (Read, 1990) and Bricogne (1991, 1993) have suggested. At the recent CCP4 Workshop on Macromolecular Refinement (5–6 January 1996, Chester, England), results were also presented from two other implementations of maximum-likelihood refinement, by Garib Murshudov and by Gerard Bricogne & John Irwin. In another crystallographic context, that of multiple isomorphous replacement, a maximum-likelihood treatment has also been applied with good results (Otwinowski, 1991).

## 2. Devising a likelihood function

The principle of maximum likelihood formalizes the idea that the quality of a model is judged by its consistency with the observations. To say that a model is consistent with an observation means that, if the model were correct, there would be a reasonably high probability of making an observation with that value. With the relevant observations taken as a set, then the probability of making the entire set of observations is an excellent measure of the quality of the model. If we assume that the observations are independent, the joint probability

of making the set of observations is the product of the probabilities of making each independent observation. This joint probability is the likelihood function ($L$):

$$L = \prod_{hkl} P(|F_o|; |F_c|). \quad (2)$$

Since it is more convenient to work with sums than products, one typically works with the logarithm of the likelihood function. As well, the maximization problem can be turned into a minimization problem by multiplying by $-1$. Therefore, defining $\mathcal{L} = -\log(L)$ gives the following:

$$\mathcal{L} = -\sum_{hkl} \log[P(|F_o|; |F_c|)]. \quad (3)$$

In the case of crystallographic refinement, it is not strictly true that the diffraction observations are independent; if they were, direct methods and density modification would not work. There is doubtless much useful information to be gained by working with higher-order collections of structure factors (Bricogne, 1993) but, as we will show, useful results are obtained even when independence is assumed.

To apply maximum likelihood, one must start from the probability of making a measurement, given the model, its errors and the measurement errors. We have shown previously that various sources of random error in the model have equivalent effects on the probability distribution for the true structure factor, whether the errors are in atomic positions or temperature factors or whether there are missing or extra atoms; in each case, the distribution of the true structure factor is well approximated by a Gaussian distribution centred on $D\mathbf{F}_c$ (Read, 1990). In the case of acentric structure factors, which make up the bulk of data for macromolecular structures, the distribution $[P_a(\mathbf{F}; \mathbf{F}_c)]$ is a two-dimensional Gaussian in the complex plane, while, for centric structure factors, it is a one-dimensional Gaussian $[P_c(\mathbf{F}; \mathbf{F}_c)]$:

$$P_a(\mathbf{F}; \mathbf{F}_c) = [1/(\pi\varepsilon\sigma_\Delta^2)] \exp[-(\mathbf{F} - D\mathbf{F}_c)^2/\varepsilon\sigma_\Delta^2] \quad (4)$$

$$P_c(\mathbf{F}; \mathbf{F}_c) = [1/(2\pi\varepsilon\sigma_\Delta^2)^{1/2}] \exp[-(\mathbf{F} - D\mathbf{F}_c)^2/2\varepsilon\sigma_\Delta^2]. \quad (5)$$

$\varepsilon$ is the expected intensity factor, $\sigma_\Delta^2 = \Sigma_N - D^2\Sigma_P$, $\Sigma_N$ = distribution parameter of the Wilson intensity distribution for $|F|$ (Wilson, 1949) and $\Sigma_P$ = distribution parameter of the Wilson intensity distribution for $|F_c|$.

The probability of the true structure-factor amplitude ($|F|$), conditional on the calculated amplitude ($|F_c|$), is obtained by integrating over the unknown phase difference to give the following:

$$P_a(|F|; |F_c|) = (2|F|/\varepsilon\sigma_\Delta^2) \times \exp[-(|F|^2 + D^2|F_c|^2)/\varepsilon\sigma_\Delta^2] \times I_0(2|F|D|F_c|/\varepsilon\sigma_\Delta^2) \quad (6)$$

$$P_c(|F|; |F_c|) = (2/\pi\varepsilon\sigma_\Delta^2)^{1/2} \times \exp[-(|F|^2 + D^2|F_c|^2)/2\varepsilon\sigma_\Delta^2] \times \cosh(|F|D|F_c|/\varepsilon\sigma_\Delta^2). \quad (7)$$

The probability distribution required to apply maximum likelihood, however, is the probability of the observed diffraction measurement given the calculated diffraction measurement as the true value is not known. We have used two methods to approximate this distribution, differing in the level of approximation and in the distribution assumed for the observational error. In the first method (MLF1), the measurement error is assumed to be Gaussian in structure-factor amplitudes and a Gaussian approximation is made for the resultant combined distribution, expressed in terms of structure-factor amplitudes. In the second method (MLF2), the measurement error is assumed to be Gaussian in the intensities and a series representation of the resultant combined distribution is expressed in terms of structure-factor amplitudes squared.

### 2.1. *MLF1: an amplitude-based likelihood function*

If the probability of the measurement error $[P(|F_o| - |F|)]$ is assumed to be Gaussian in structure-factor amplitudes with standard deviation $\sigma_F$, then the required probability distribution $P(|F_o|; |F_c|)$ is obtained by convoluting $P(|F|; |F_c|)$ by $P(|F_o| - |F|)$.

$$P(|F_o|; |F_c|) = P(|F|; |F_c|) \otimes P(|F_o| - |F|). \quad (8)$$

As far as we have been able to determine, there is no analytical solution to this convolution for the important acentric case. (A series representation could be derived similarly to MLF2, as discussed below. We believe that it is better to use MLF2 if one goes to the effort of computing the series representation.) However, a good Gaussian approximation can be obtained using the first two central moments of the distribution. The expected value for the acentric case is given by the following:

$$\langle|F_o|\rangle = [(\pi\varepsilon\sigma_\Delta^2)^{1/2}/2]\, \Phi(-1/2, 1, -D^2|F_c|^2/\varepsilon\sigma_\Delta^2). \quad (9)$$

For the centric case, the expected value is

$$\langle|F_o|\rangle = [(2\varepsilon\sigma_\Delta^2/\pi)^{1/2}]\, \Phi(-1/2, 1/2, -D^2|F_c|^2/2\varepsilon\sigma_\Delta^2). \quad (10)$$

In these expressions, $\Phi(a, b, x)$ is Kummer's confluent hypergeometric function, also denoted by ${}_1F_1(a, b, x)$. The variance for both the acentric and centric distributions is given by the following:

$$\sigma_{ML}^2 = \varepsilon\sigma_\Delta^2 + \sigma_F^2 + D^2|F_c|^2 - \langle|F_o|\rangle^2. \quad (11)$$

As $|F_c|$ increases, $\sigma^2_{ML}$ tends towards $\varepsilon\sigma^2_\Delta + \sigma^2_F$ in the centric case and $\frac{1}{2}\varepsilon\sigma^2_\Delta + \sigma^2_F$ in the acentric case because, in the limit, only the component of model error parallel to $\mathbf{F}_c$ contributes to the error in the amplitude. When these moments are used to construct a Gaussian approximation, the negative log-likelihood function ($\mathcal{L}$) is

$$\mathcal{L} = \sum_{hkl} \tfrac{1}{2}\log(2\pi) + \log(\sigma_{ML}) + (1/2\sigma^2_{ML})(|F_o| - \langle|F_o|\rangle)^2 \tag{12}$$

If $\sigma_{ML}$ is assumed to be relatively constant within a cycle of refinement, maximum-likelihood refinement can be approximated as a modified least-squares refinement, in which the following target is minimized:

$$\text{WSSQ} = \sum_{hkl}(1/\sigma^2_{ML})(|F_o| - \langle|F_o|\rangle)^2. \tag{13}$$

This target can readily be implemented in any crystallographic refinement program that uses a least-squares target by weighting each term with $1/\sigma^2_{ML}$, replacing $k|F_c|$ with $\langle|F_o|\rangle$ and replacing $\partial|F_c|/\partial p$ with $(\partial\langle|F_o|\rangle/\partial|F_c|)(\partial|F_c|/\partial p)$, where $p$ is any parameter of the model being refined. The required derivative for the acentric case is given by

$$\partial\langle|F_o|\rangle/\partial|F_c| = (\pi/\varepsilon\sigma^2_\Delta)^{1/2}(D^2|F_c|/2) \times \Phi(1/2, 2, -D^2|F_c|^2/\varepsilon\sigma^2_\Delta); \tag{14}$$

and for the centric case by

$$\partial\langle|F_o|\rangle/\partial|F_c| = (2/\pi\varepsilon\sigma^2_\Delta)^{1/2}D^2|F_c| \times \Phi(1/2, 3/2, -D^2|F_c|^2/2\varepsilon\sigma^2_\Delta). \tag{15}$$

Note that the $|F_c|$ term eliminates the singularity in the derivatives that can arise in least-squares refinement on amplitudes (Schwarzenbach *et al.*, 1989).

Algorithms implementing (9)–(11), (14) and (15) are described in Appendix *A*. The quality of the Gaussian approximation can be judged from a comparison of distributions shown in Fig. 1.

### 2.2. *MLF2: an intensity-based likelihood function*

The second method that we use to derive the required probability distribution works in terms of structure-factor amplitudes squared ($J = |F|^2$). Two advantages are attained by working in $J$ instead of $|F|$. First, measurement errors frequently lead to a negative net intensity, which is reduced to negative $J$; when these legitimate observations are transformed to $|F|$, one has the choice of omitting them, replacing them with zero or replacing them with a non-zero Bayesian posterior value (French & Wilson, 1978). By working in terms of $J$, this problem is avoided. Furthermore, a Gaussian measurement error is better justified in $J$ than in $|F|$. In principle, maximum likelihood is insensitive to variable transformations such as from $|F|$ to $|F|^2$ (Edwards, 1992). If MLF2 did not differ from MLF1 in the distribution assumed for the measurement error, the two likelihood functions would differ only in the precision of the approximation.

The required probability distribution $P(J_o; J_c)$ is derived by multiplying $P(J; J_c)$ with the Gaussian probability of the measurement error $[P(J_o; J)]$ with standard deviation $\sigma_j$ and integrating over the true structure-factor amplitude squared ($J$).

$$P(J_o; J_c) = \int_0^\infty P(J_o; J) \times P(J; J_c)\,\mathrm{d}J. \tag{16}$$

A series representation of $P(J_o; J_c)$ can be computed. For acentric reflections, the distribution is

$$P_a(J_o; J_c) = [1/(2\pi)^{1/2}\varepsilon\sigma^2_\Delta]\exp(-J_o^2/2\sigma_j^2 - D^2J_c/\varepsilon\sigma^2_\Delta) \times \sum_{n=0}^{\infty}(D^2J_c\sigma_j/\varepsilon^2\sigma^4_\Delta)^n(1/n!) \times \exp[(\sigma_j^2 - J_o\varepsilon\sigma^2_\Delta)^2/4\varepsilon^2\sigma^4_\Delta\sigma_j^2] \times \mathsf{D}_{-n-1}([\sigma_j^2 - J_o\varepsilon\sigma^2_\Delta]/\varepsilon\sigma^2_\Delta\sigma_j). \tag{17}$$

$\mathsf{D}_{-n-1}(x)$ is a parabolic cylinder function. For centric reflections,

$$P_c(J_o; J_c) = [1/2(\pi\sigma_j\varepsilon)^{1/2}\sigma_\Delta] \times \exp(-J_o^2/2\sigma_j^2 - D^2J_c/2\varepsilon\sigma^2_\Delta) \times \sum_{n=0}^{\infty}(D^2J_c\sigma_j/2\varepsilon^2\sigma^4_\Delta)^n[1/(2n)!!] \times \exp[(\sigma_j^2 - 2J_o\varepsilon\sigma^2_\Delta)^2/16\varepsilon^2\sigma^4_\Delta\sigma_j^2] \times \mathsf{D}_{-n-1/2}([\sigma_j^2 - 2J_o\varepsilon\sigma^2_\Delta]/2\varepsilon\sigma^2_\Delta\sigma_j). \tag{18}$$

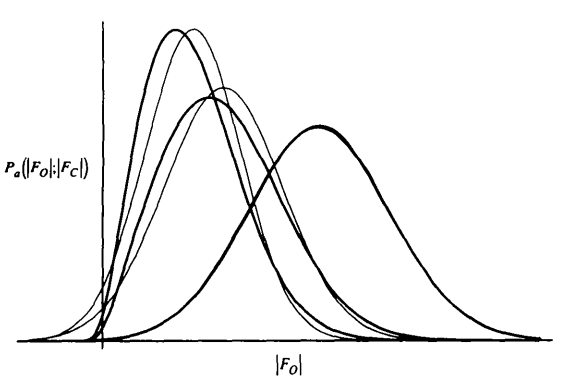


Fig. 1. Comparison of the Gaussian approximation to $P_a(|F_o|; |F_c|)$ (thin lines) with the exact form determined by numerical integration (thick lines). Three pairs of curves are shown, corresponding to weak, average and strong reflections with $D = 0.7$. This figure, Fig. 2 and some of the mathematical derivations were made with the assistance of the program *Mathematica* (Wolfram, 1991).

After eliminating terms that are constant within a cycle of refinement, the negative log likelihood ($\mathcal{L}$) for the acentric case is

$$\mathcal{L} = \sum_{hkl} \log(\varepsilon\sigma_\Delta^2) + D^2 J_c/\varepsilon\sigma_\Delta^2 - \log\left\{\sum_{n=0}^{\infty}(D^2 J_c\sigma_j/\varepsilon^2\sigma_\Delta^4)^n(1/n!) \times \exp[(\sigma_j^2 - J_o\varepsilon\sigma_\Delta^2)^2/4\varepsilon^2\sigma_\Delta^4\sigma_j^2] \times \mathsf{D}_{-n-1}([\sigma_j^2 - J_o\varepsilon\sigma_\Delta^2]/\varepsilon\sigma_\Delta^2\sigma_j)\right\}; \quad (19)$$

and for centric reflections it is

$$\mathcal{L} = \sum_{hkl} \tfrac{1}{2}\log(\varepsilon\sigma_\Delta^2) + D^2 J_c/2\varepsilon\sigma_\Delta^2 - \log\left\{\sum_{n=0}^{\infty}(D^2 J_c\sigma_j/2\varepsilon^2\sigma_\Delta^4)^n[1/(2n)!!] \times \exp[(\sigma_j^2 - 2J_o\varepsilon\sigma_\Delta^2)^2/16\varepsilon^2\sigma_\Delta^4\sigma_j^2] \times \mathsf{D}_{-n-1/2}([\sigma_j^2 - 2J_o\varepsilon\sigma_\Delta^2]/2\varepsilon\sigma_\Delta^2\sigma_j)\right\}. \quad (20)$$

Equations (17)–(20) are derived in Appendix *A*.

Some essential differences between least-squares and maximum-likelihood refinement can be seen in a comparison (Fig. 2) of the derivatives of the target functions, which lead to the atomic shifts in the refinement process.

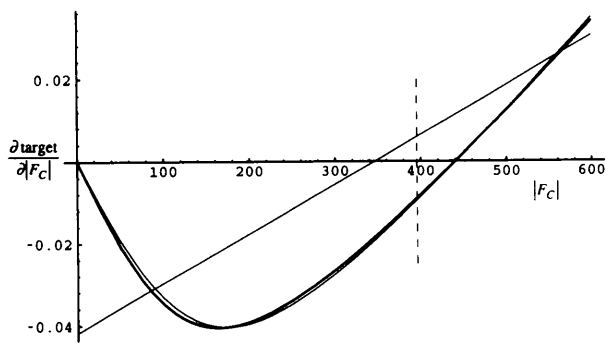


Fig. 2. Comparison of the derivatives, with respect to $|F_c|$ for one reflection, of the refinement targets for least-squares (thin line), MLF1 (thin curve) and MLF2 (thick curve) as a function of $|F_c|$. The example (the 2,12,17 reflection of the gTIM test case, discussed below) is chosen to illustrate the degree to which the least-squares and maximum-likelihood targets can differ. In *XPLOR*, the derivative contributes to a force on each atom to move in a direction that will decrease the refinement target. At the start of refinement, $|F_c|$ is 395.6 (indicated by the dashed vertical line); according to the least-squares target, atoms should move to decrease $|F_c|$ while, according to the maximum-likelihood targets, atoms should move in the opposite direction to increase $|F_c|$. Note that if $|F_c|$ were zero the derivatives for the maximum-likelihood targets would also be zero, reflecting the fact that the true phase would be completely uncertain and that a desired direction of shift could not be inferred.

### 2.3. *Calibration of structure-factor probabilities*

The value of the likelihood function depends on the parameters of the atomic model. It also depends on the resolution-dependent parameters $D$ and $\sigma_\Delta^2$, which characterize the effect of model error on the structure-factor probability distributions. [In fact, $D$ and $\sigma_\Delta^2$ are not independent and can each be computed from the single parameter $\sigma_A$ (Read, 1990).] In principle, it would be best to optimize the likelihood function by adjusting all parameters simultaneously, including coordinates, $B$ factors *and* $\sigma_A$ values. Unfortunately, a problem arises if the $\sigma_A$ values are refined using the same data against which the model is refined: the poor parameter-to-observation ratio allows overfitting of the amplitudes, which results in an overestimation of $\sigma_A$ and hence an underestimation of the errors in the calculated structure factors (Lunin & Urzhumtsev, 1984; Read, 1986). This leads to a positive feedback cycle in which the pressure to overfit becomes stronger. In our first attempt to implement maximum-likelihood refinement, this problem was ignored. As the quality of the likelihood function depends strongly on the accuracy of $\sigma_A$ estimates, the results were unimpressive.

The solution we have adopted is to use cross-validation data (a minority of data omitted from the refinement target) in an active way to provide unbiased estimates of structure-factor accuracy. These data are normally used to compute $R_{\text{free}}$, an unbiased measure of refinement progress (Brünger, 1992). The use of cross-validation data to estimate $\sigma_A$ is complicated, however, by the fact that stable estimates require 500 to 1000 reflections in each resolution shell, especially when the true value is low (Read, 1986). To overcome the problem of instability, we exploit the fact that $\sigma_A$ varies smoothly with resolution. A simple correction, in which a penalty is applied when a $\sigma_A$ value lies far from the line connecting its two neighbours, is sufficient (Read, unpublished).

A better solution would be to refine the $\sigma_A$ values as parameters in the refinement but to make allowance for the fact that they are biased estimates in using them in the likelihood function. Since a theoretical basis for the correction for bias is lacking, however, this solution cannot yet be applied. We are currently studying the effect of refinement bias on the structure-factor distributions, to lay the groundwork for such an improved treatment.

## 3. Test refinements

The two maximum-likelihood targets have been implemented in the program *XPLOR* (Brünger, Kuriyan & Karplus, 1987). Results from runs of the modified *XPLOR* on two test systems will be discussed here. In each test, the suggested weighting factor (WA) for the diffraction terms in the target, obtained by comparing

Table 1. *Refinement statistics for the SGT test case*

The starting model (BT superimposed on SGT) was refined against calculated SGT data in three runs of *XPLOR*, identical except for the target function. In total, 420 cycles of energy-minimization refinement were carried out.

| | Start | Least squares | MLF1 | MLF2 |
|---|---|---|---|---|
| *R* factor | 0.515 | 0.403 | 0.416 | 0.422 |
| $R_{free}$ | 0.542 | 0.511 | 0.525 | 0.528 |
| Mean phase error (°) | 62.2 | 60.0 | 56.7 | 56.5 |
| Mean cos(phase error) | 0.365 | 0.394 | 0.436 | 0.437 |

the gradients from the diffraction and energy terms (Brünger, Karplus & Petsko, 1989), was divided by two.

### 3.1. *Streptomyces griseus trypsin*

The crystal structure of *Streptomyces griseus* trypsin (Read & James, 1988) (SGT) was solved originally by molecular replacement using the structure of bovine trypsin (Chambers & Stroud, 1979) (BT) as a search model. In order to compare the power of the maximum-likelihood and least-squares targets in a case where the phase errors are known exactly, we used data calculated from SGT as error-free amplitudes $|F_o|$ and a superimposed model of BT as a starting structure. Since these two proteins share about 33% sequence identity, BT provides a relatively poor model that will only be capable of refining into a local minimum.

Data from infinity to 2.8 Å resolution (5732 reflections, of which 578 were flagged as cross-validation data) were used for both refinements. (One often omits the low-resolution data for least-squares refinement because of the complications caused by disordered solvent but in this case there is no disordered solvent.) Table 1 shows the results obtained in the different refinements. While none of the refinements could achieve an accurate model, owing to the inadequacies of the starting model, the maximum-likelihood targets gave more than twice as large an improvement in the average phase error. Note that, owing probably to the small number of reflections used in this case, $R_{free}$ provides a weak indication of phase accuracy.

### 3.2. *Trypanosoma brucei glycosomal triosephosphate isomerase*

At an intermediate stage in the refinement of the glycosomal triosephosphate isomerase (gTIM) from *Trypanosoma brucei* (Wierenga, Noble, Vriend, Nauche & Hol, 1991), data to a resolution of 1.83 Å became available to replace the data to 2.4 Å resolution that had been used to that point (Wierenga, Kalk & Hol, 1987). We tested the three refinement targets on this intermediate model, using the observed diffraction data (model and data kindly supplied by Dr R. K. Wierenga). Of 38 812 observed amplitudes, 1014 were flagged randomly as cross-validation data. Because this is a real data set measured from a crystal with disordered solvent, data from infinity to 8.0 Å resolution were omitted in the least-squares refinement while they were used in both maximum-likelihood refinements.

As shown in Fig. 3, both maximum-likelihood target functions achieved a significantly greater improvement in the model, measured by both $R_{free}$ and phase differences with the final model. One example of the increased power of maximum-likelihood refinement is illustrated in Fig. 4; the least-squares refinement has failed to complete a rigid-body shift of a helix that

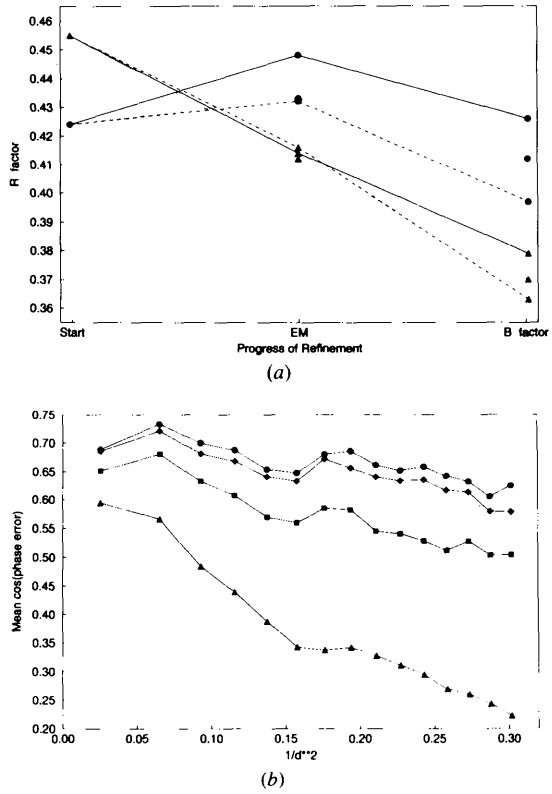


Fig. 3. (*a*) *R* factors through the test refinements of gTIM. The runs were identical except for the target function and the treatment of low-resolution data; for the least-squares refinement, data from infinity to 8 Å were omitted, while they were included for both maximum-likelihood refinements. In each case, 250 cycles of energy minimization (EM) refinement were run, followed by 30 cycles of *B*-factor refinement. The solid lines indicate *R* factors for the least-squares target, the dotted lines indicate *R* factors for the MLF1 target and the dashed lines indicate *R* factors for the MLF2 target. $R_{free}$ values for the three different target functions are represented by circles and *R* values for the three different target functions are represented by triangles. The initial increase in $R_{free}$ probably reflects the fact that all data from 6.0 to 2.4 Å resolution had been used in the previous refinement. (*b*) Phase accuracy after gTIM test refinements. The phase accuracy is computed as the mean cosine of the phase error, which is comparable to the mean figure of merit. Triangles correspond to the starting model, squares to the least-squares model, diamonds to the MLF1 model and circles to the MLF2 model.

is within the convergence radius for the maximum-likelihood refinement. The increased phase accuracy, coupled perhaps with less of a tendency to overfit data, results in an electron-density map that is clearer in regions where the model is still in error (Fig. 5).

As one might expect from the increased precision of the approximation, the MLF2 target gives significantly better results than MLF1 (Fig. 3). This improvement is achieved for a modest computational cost. Compared to an equivalent refinement with the least-squares target, the MLF1 target requires about 1% more computer time, while the MLF2 target requires about 10% more computer time.

## 4. Conclusions

While the current implementations of maximum-likelihood refinement already provide significant benefits, a number of improvements can be envisioned. First, the algorithm for the estimation of $\sigma_A$ does not take into account measurement errors. Either of the likelihood functions derived here, MLF1 or MLF2, can be used to compute $\sigma_A$ values that take into account measurement errors, and these modified likelihood functions will be implemented in the *SIGMAA* algorithm. As is clear from the variance term in the Gaussian approximation MLF1 [equation (11)], observational error has little

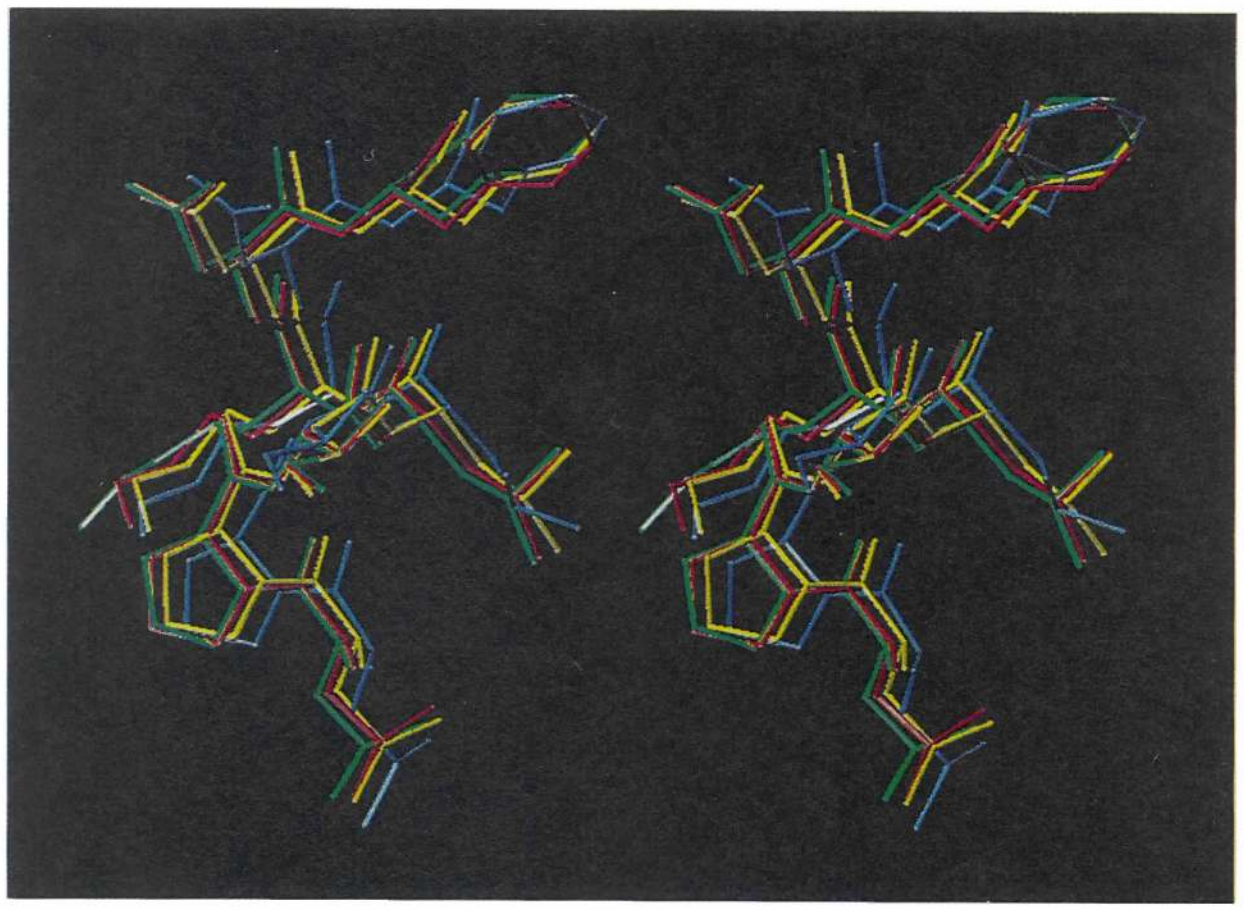

Fig. 4. Rigid-body shift in gTIM test refinements. In this helical region of gTIM, a rigid-body shift can be seen between the starting model (blue) and the final model (green). The least-squares refinement (yellow model) has failed to make the full shift, while the maximum-likelihood target, MLF1 (pink model), has converged to a result close to the final model. This figure and Fig. 5 were drawn using the program *O* (Jones, Zou, Cowan & Kjeldgaard, 1991).

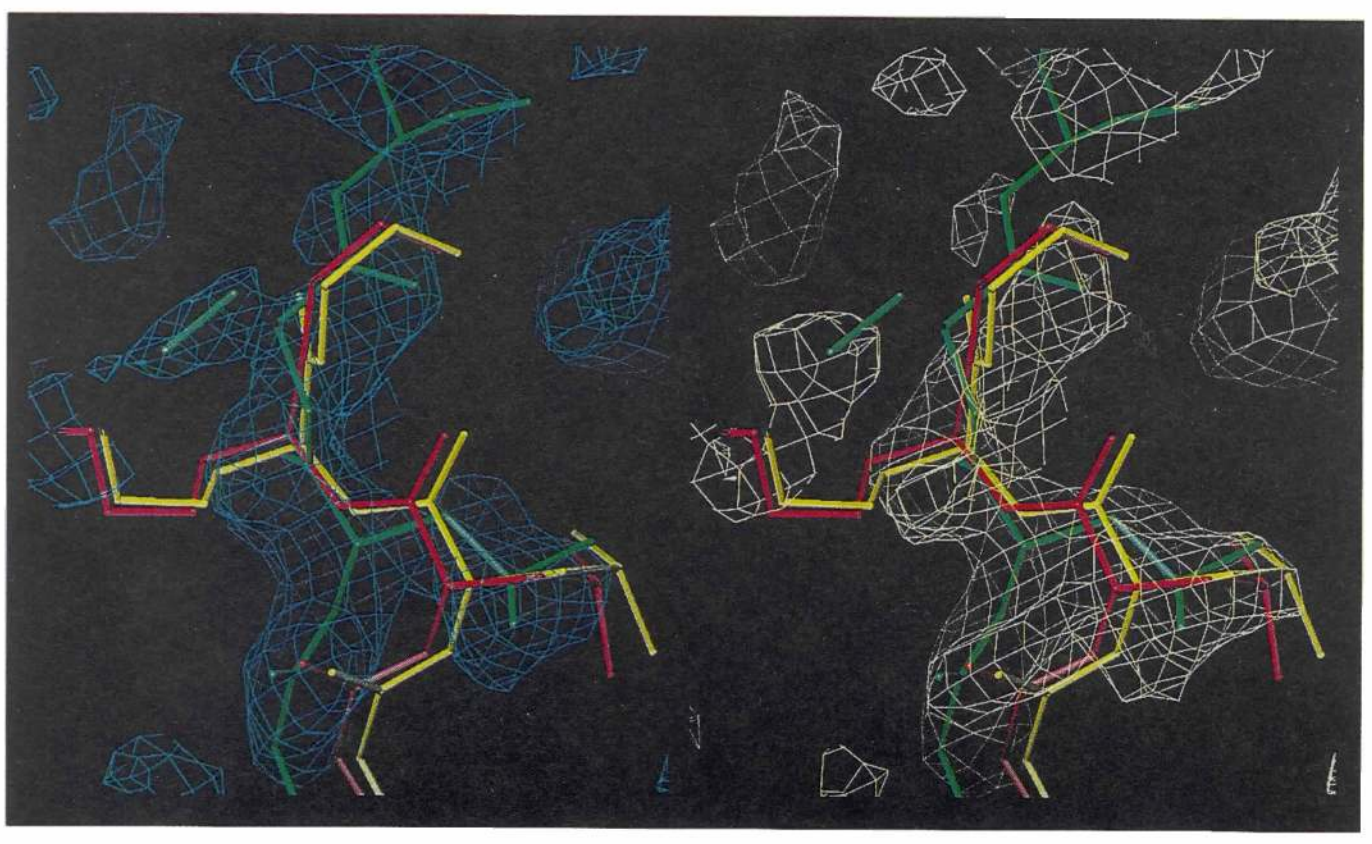

Fig. 5. Electron-density maps from gTIM test refinements. In this region of gTIM, a major conformational change is required to get to the final model (green), but is not within the power of gradient-driven refinement with either the maximum-likelihood target, MLF1 (pink model), or the least-squares target (yellow model). Nonetheless, the general phase improvement through maximum-likelihood refinement makes the change required in the model considerably clearer (left, blue density) than for the least-squares refinement (right, tan density). Each map is contoured at the r.m.s. value of the electron density.

influence on the likelihood function unless the model is quite accurate. Nonetheless, it will become significant at the end of refinement and a proper treatment will be important to obtain an optimal final model.

Arbitrary relative weights between diffraction and geometry terms should not be required, in principle, if each is introduced to maximum likelihood through the appropriate probability distributions. However, we have found that some overweighting of the diffraction terms, relative to the theoretical value, is needed to achieve convergence. This may be necessary in part because the inevitable overfitting of the diffraction amplitudes alters the distribution $P(\mathbf{F}; \mathbf{F}_c)$. In various tests, the comparison of gradients has led to weights that are increased by factors of between 4 and 50, with higher weights being required for less-refined models at lower resolution. Further tests will be required to decide whether these relative weights are optimal.

Finally, the maximum-likelihood approach allows one to include, in a sensible way, any combination of information (Bricogne, 1993). We believe that considerable scope for improvement exists in the simultaneous refinement of structures, for instance, native with liganded, or native with heavy-atom derivatives. In such a refinement, all observations would be fit simultaneously, using models that are restrained to resemble one another to a degree required by the relationships among the measured sets of structure factors.

This work might not have been carried out if not for the opportunity provided by Dr Rik K. Wierenga, who was the host for RJR as a summer visitor to EMBL, Heidelberg, Germany, in May 1993, when the first steps to implementation were taken. Discussions with Bart Hazes and Steven Ness helped greatly in implementing MLF2 into *XPLOR*. Financial support was provided by the Alberta Heritage Foundation for Medical Research, the Medical Research Concil of Canada, the Natural Sciences and Engineering Research Council of Canada and an International Research Scholar award to RJR from the Howard Hughes Medical Institute.

## APPENDIX *A*
## Derivations and implementation

### *A* 1. Implementation of MLF1

The implementation of the MLF1 target requires the computation of a number of confluent hypergeometric functions of the form $\Phi(\alpha, \gamma, x)$, with different arguments $\alpha$ and $\gamma$. When the argument of $x$ is small, a Chebyshev polynomial approximation (Luke, 1977) can be used. For arguments of $x$ larger than ten, it is preferable to use an asymptotic expansion (Slater, 1965). Note that, as $|F_c|$ increases, $\langle|F_o|\rangle$ tends to $D|F_c|$ and $\langle|F_o|\rangle^2$ tends to $D^2|F_c|^2$ (centric case) or $D^2|F_c|^2 + \frac{1}{2}\varepsilon\sigma_\Delta^2$ (acentric case).

### *A* 2. Implementation of MLF2

The distribution $P(J_o; J_c)$ is attained by multiplying $P(J; J_c)$ with a Gaussian probability of measurement errors $[P(J_o; J)]$ with standard deviation $\sigma_j$ and integrating over the true structure-factor amplitude squared, $J$. The distribution $P(J; J_c)$ is obtained *via* a variable transformation of (6) for acentric and (7) for centric structure factors.

$$P_a(J; J_c) = [1/(\varepsilon\sigma_\Delta^2)] \exp[-(J + D^2J_c)/\varepsilon\sigma_\Delta^2] \times I_0(2D[JJ_c]^{1/2}/\varepsilon\sigma_\Delta^2) \quad (21)$$

$$P_c(J; J_c) = [1/(2\pi\varepsilon\sigma_\Delta^2 J)]^{1/2} \exp[-(J + D^2J_c)/2\varepsilon\sigma_\Delta^2] \times \cosh[D(JJ_c)^{1/2}/\varepsilon\sigma_\Delta^2]. \quad (22)$$

The joint probability, $P(J, J_o; J_c)$, is the product of the probability of the observation error and the probability of the true intensity given the calculated intensity. The desired distribution, $P(J_o; J_c)$, is the integral over $J$ of the joint probability.

$$P(J_o; J_c) = \int_0^\infty P(J, J_o; J_c)\, dJ = \int_0^\infty P(J_o; J) \times P(J; J_c)\, dJ. \quad (23)$$

For acentric reflections,

$$P_a(J_o; J_c) = \int_0^\infty [1/(2\pi)^{1/2}\sigma_j\varepsilon\sigma_\Delta^2] \times \exp[-(J - J_o)^2/2\sigma_j^2 - (J + D^2J_c)/\varepsilon\sigma_\Delta^2] \times I_0(2D[JJ_c]^{1/2}/\varepsilon\sigma_\Delta^2)\, dJ \quad (24)$$

$$= [1/(2\pi)^{1/2}\sigma_j\varepsilon\sigma_\Delta^2] \times \exp(-J_o^2/2\sigma_j^2 - D^2J_c/\varepsilon\sigma_\Delta^2) \times \int_0^\infty \exp\{-J^2/2\sigma_j^2 - J[(\sigma_j^2 - J_o\varepsilon\sigma_\Delta^2)/\varepsilon\sigma_\Delta^2\sigma_j^2]\} \times I_0(2D[JJ_c]^{1/2}/\varepsilon\sigma_\Delta^2)\, dJ. \quad (25)$$

Expanding the modified Bessel function into a power series $[I_0(x) = \sum_{n=0}^\infty (x/2)^{2n}/(n!)^2]$ and interchanging integration and summation gives the following:

$$P_a(J_o; J_c) = [1/(2\pi)^{1/2}\sigma_j\varepsilon\sigma_\Delta^2] \times \exp(-J_o^2/2\sigma_j^2 - D^2J_c/\varepsilon\sigma_\Delta^2) \times \sum_{n=0}^\infty (D^2J_c/\varepsilon^2\sigma_\Delta^4)^n[1/(n!)^2] \times \int_0^\infty \exp\{-J^2/2\sigma_j^2 - J[(\sigma_j^2 - J_o\varepsilon\sigma_\Delta^2)/\varepsilon\sigma_\Delta^2\sigma_j^2]\} J^n\, dJ. \quad (26)$$

There exists an antiderivative for this expression (Gradshteyn & Ryzhik, 1980).

$$P_a(J_o;J_c) = [1/(2\pi)^{1/2}\varepsilon\sigma_\Delta^2]\exp(-J_o^2/2\sigma_j^2 - D^2J_c/\varepsilon\sigma_\Delta^2) \times \sum_{n=0}^{\infty}(D^2J_c\sigma_j/\varepsilon^2\sigma_\Delta^4)^n(1/n!) \times \exp[(\sigma_j^2 - J_o\varepsilon\sigma_\Delta^2)^2/4\varepsilon^2\sigma_\Delta^4\sigma_j^2] \times \mathsf{D}_{-n-1}([\sigma_j^2 - J_o\varepsilon\sigma_\Delta^2]/\varepsilon\sigma_\Delta^2\sigma_j). \quad (27)$$

$\mathsf{D}_{-n-1}(x)$ is a parabolic cylinder function. Now, for centric reflections,

$$P_c(J_o;J_c) = (1/2\pi\varepsilon^{1/2}\sigma_\Delta\sigma_j) \times \int_0^{\infty}(1/J^{1/2})\exp[-(J + D^2J_c)/2\varepsilon\sigma_\Delta^2 - (J - J_o)^2/2\sigma_j^2] \times \cosh[D(JJ_c)^{1/2}/\varepsilon\sigma_\Delta^2]\,dJ \quad (28)$$

$$= (1/2\pi\varepsilon^{1/2}\sigma_\Delta\sigma_j) \times \exp(-J_o^2/2\sigma_j^2 - D^2J_c/2\varepsilon\sigma_\Delta^2) \times \int_0^{\infty}(1/J^{1/2})\exp\{-J^2/2\sigma_j^2 - J[(\sigma_j^2 - 2J_o\varepsilon\sigma_\Delta^2)/2\varepsilon\sigma_\Delta^2\sigma_j^2]\} \times \cosh[D(JJ_c)^{1/2}/\varepsilon\sigma_\Delta^2]\,dJ. \quad (29)$$

Expanding the cosine hyperbolic function into a power series [$\cosh(x) = \sum_{n=0}^{\infty}x^{2n}/(2n)!$] and interchanging summation and integration gives

$$P_c(J_o;J_c) = (1/2\pi\varepsilon^{1/2}\sigma_\Delta\sigma_j) \times \exp(-J_o^2/2\sigma_j^2 - D^2J_c/2\varepsilon\sigma_\Delta^2) \times \sum_{n=0}^{\infty}(D^2J_c/\varepsilon^2\sigma_\Delta^4)^n[1/(2n)!] \times \int_0^{\infty}J^{n-1/2}\exp\{-J^2/2\sigma_j^2 - J[(\sigma_j^2 - 2J_o\varepsilon\sigma_\Delta^2)/2\varepsilon\sigma_\Delta^2\sigma_j^2]\}\,dJ. \quad (30)$$

The analytic solution for this expression is

$$P_c(J_o;J_c) = [1/2(\pi\sigma_j\varepsilon)^{1/2}\sigma_\Delta] \times \exp(-J_o^2/2\sigma_j^2 - D^2J_c/2\varepsilon\sigma_\Delta^2) \times \sum_{n=0}^{\infty}(D^2J_c\sigma_j/2\varepsilon^2\sigma_\Delta^4)^n[1/(2n)!!] \times \exp[(\sigma_j^2 - 2J_o\varepsilon\sigma_\Delta^2)^2/16\varepsilon^2\sigma_\Delta^4\sigma_j^2] \times \mathsf{D}_{-n-1/2}([\sigma_j^2 - 2J_o\varepsilon\sigma_\Delta^2]/2\varepsilon\sigma_\Delta^2\sigma_j). \quad (31)$$

The negative logarithm of $P(J_o;J_c)$ for the acentric case gives the following:

$$\tfrac{1}{2}\log(2\pi) + \log(\varepsilon\sigma_\Delta^2) + J_o^2/2\sigma_j^2 + D^2J_c/\varepsilon\sigma_\Delta^2 - \log\left\{\sum_{n=0}^{\infty}(D^2J_c\sigma_j/\varepsilon^2\sigma_\Delta^4)^n(1/n!) \times \exp[(\sigma_j^2 - J_o\varepsilon\sigma_\Delta^2)^2/4\varepsilon^2\sigma_\Delta^4\sigma_j^2] \times \mathsf{D}_{-n-1}([\sigma_j^2 - J_o\varepsilon\sigma_\Delta^2]/\varepsilon\sigma_\Delta^2\sigma_j)\right\}. \quad (32)$$

For centric reflections, the negative logarithm of $P(J_o;J_c)$ is

$$\log[2(\pi\sigma_j\varepsilon)^{1/2}\sigma_\Delta] + J_o^2/2\sigma_j^2 + D^2J_c/2\varepsilon\sigma_\Delta^2 - \log\left\{\sum_{n=0}^{\infty}(D^2J_c\sigma_j/2\varepsilon^2\sigma_\Delta^4)^n[1/(2n)!!] \times \exp[(\sigma_j^2 - 2J_o\varepsilon\sigma_\Delta^2)^2/16\varepsilon^2\sigma_\Delta^4\sigma_j^2] \times \mathsf{D}_{-n-1/2}([\sigma_j^2 - 2J_o\varepsilon\sigma_\Delta^2]/2\varepsilon\sigma_\Delta^2\sigma_j)\right\}. \quad (33)$$

The elimination of constant terms from (32) and (33) leads to the functions implemented in *XPLOR*.

$$\mathcal{L} = \sum_{hkl}\log(\varepsilon\sigma_\Delta^2) + D^2J_c/\varepsilon\sigma_\Delta^2 - \log\left\{\sum_{n=0}^{\infty}(D^2J_c\sigma_j/\varepsilon^2\sigma_\Delta^4)^n(1/n!) \times \exp[(\sigma_j^2 - J_o\varepsilon\sigma_\Delta^2)^2/4\varepsilon^2\sigma_\Delta^4\sigma_j^2] \times \mathsf{D}_{-n-1}([\sigma_j^2 - J_o\varepsilon\sigma_\Delta^2]/\varepsilon\sigma_\Delta^2\sigma_j)\right\} \quad (34)$$

for acentric reflections and

$$\mathcal{L} = \sum_{hkl}\tfrac{1}{2}\log(\varepsilon\sigma_\Delta^2) + D^2J_c/2\varepsilon\sigma_\Delta^2 - \log\left\{\sum_{n=0}^{\infty}(D^2J_c\sigma_j/2\varepsilon^2\sigma_\Delta^4)^n[1/(2n)!!] \times \exp[(\sigma_j^2 - 2J_o\varepsilon\sigma_\Delta^2)^2/16\varepsilon^2\sigma_\Delta^4\sigma_j^2] \times \mathsf{D}_{-n-1/2}([\sigma_j^2 - 2J_o\varepsilon\sigma_\Delta^2]/2\varepsilon\sigma_\Delta^2\sigma_j)\right\} \quad (35)$$

for centric reflections. The numerical algorithm employed to evaluate the parabolic cylinder functions in (34) and (35) can be divided into two different possibilities: one case is when the argument of the parabolic cylinder function is non-positive and the other is when it is positive. In both cases, the algorithm developed relies on evaluating the function $\mathsf{D}_{-\nu}(x)$ for two particular values of $\nu$ and using recursion relations to calculate the special function for the other values of $\nu$ necessary for the series to converge.

In the first case, when the argument of the parabolic cylinder function is non-positive, the special function becomes large as $x \to -\infty$. Thus, to ensure convergence

of the series, $\exp(-x^2/4)D_{-\nu}(x)$ is evaluated and then $x^2/2$ is added to the likelihood function. The algorithm utilizes the relationship with the complement of the error function [erfc(x)] in the acentric case:

$$\mathsf{D}_{-n-1}(x) = (\pi/2)^{1/2}[(-1)^n/n!]\exp(-x^2/4) \times \mathrm{d}^n[\exp(x^2/2)\,\mathrm{erfc}(x/2^{1/2})]/\mathrm{d}x^n, \quad (36)$$

where the complementary error function is defined as

$$\mathrm{erfc}(x) = (2/\pi^{1/2})\int_x^\infty \exp(-\eta^2)\,\mathrm{d}\eta \quad (37)$$

$$= 1 - \mathrm{erf}(x). \quad (38)$$

Thus, for $n = 0, 1$, (36) implies

$$\exp(-x^2/4)\,\mathsf{D}_{-1}(x) = (\pi/2)^{1/2}\,\mathrm{erfc}(x/2^{1/2}) \quad (39)$$

$$\exp(-x^2/4)\,\mathsf{D}_{-2}(x) = \exp(-x^2/2) - x(\pi/2)^{1/2}\,\mathrm{erfc}(x/2^{1/2}). \quad (40)$$

The code for the numerical evaluation of erfc(x) in MLF2 was written by Cody (1969). The following recursion relation is used to calculate higher-order parabolic cylinder functions.

$$\mathsf{D}_{-n-1}(x) = (1/n)[\mathsf{D}_{-n+1}(x) - x\mathsf{D}_{-n}(x)]. \quad (41)$$

Note that both sides of the equation can be multiplied by $\exp(-x^2/4)$ to give a recursion relation involving $\exp(-x^2/4)\,\mathsf{D}_{-n-1}(x)$.

In the centric case, when the argument of the parabolic cylinder function is non-positive, the first two terms ($n = 0, 1$) are evaluated *via* the relationship with the confluent hypergeometric function $\Phi(a, b, x)$:

$$\exp(-x^2/4)\,\mathsf{D}_{-n-1/2}(x) = (\pi^{1/2}/2^{n/2+1/4})\{[1/\Gamma(n/2+3/4)] \times \Phi(1/4 - n/2, 1/2, -x^2/2) - [2^{1/2}x/\Gamma(n/2+1/4)] \times \Phi(3/4 - n/2, 3/2, -x^2/2)\}. \quad (42)$$

The algorithm for the numerical evaluation of $\Phi(a, b, -x)$ was adopted from Slater (1965), Luke (1977) and Baker (1992). A recursion relation similar to (41) can be used to attain higher-order terms in the centric case.

In the case where the argument of the parabolic cylinder function is positive, both acentric and centric likelihood functions can be calculated using the relationship of the parabolic cylinder function with the confluent hypergeometric function $\Psi(a, b, x)$, also denoted by $U(a, b, x)$. Since $\Psi(a, b, x)$ remains bounded as $x$ becomes large, $\exp(x^2/4)\,\mathsf{D}_{-\nu}(x)$ is evaluated.

$$\exp(x^2/4)\,\mathsf{D}_{-\nu}(x) = (1/2^{\nu/2})\,\Psi(\nu/2, 1/2, x^2/2). \quad (43)$$

If the first two terms ($n = 0, 1$) are evaluated using (43) and higher-order terms are evaluated using (41), catastrophic cancellation occurs during the determination of higher-order terms. Therefore, first $\mathsf{D}_{-\nu}(x)$ is evaluated using (43) for $-\nu = \lambda + 1, \lambda$, where $\lambda$ is large enough to ensure convergence. Then the terms $-\nu = \lambda - 1, \lambda - 2, \ldots, 0$ are evaluated using a rearrangement of (41):

$$\mathsf{D}_{-\nu}(x) = \nu\mathsf{D}_{-\nu-1}(x) + x\mathsf{D}_{-\nu-2}(x). \quad (44)$$

The numerical evaluation of $\Psi(a, b, x)$ in MLF2 was adopted from Temme (1983).

Note that as $D^2J_c\sigma_j/\varepsilon^2\sigma_\Delta^4$ increases the infinite summations in (34) and (35) need more terms to converge and it is possible that the numerical values exceed machine precision before convergence occurs. We have recently derived an asymptotic equation valid for large values of $D^2J_c\sigma_j/\varepsilon^2\sigma_\Delta^4$ for acentric reflections. Such asymptotic expressions will compute the likelihood function more efficiently for large parameters and avoid potential overflow. In the two test cases discussed, however, overflow was not a problem. Nonetheless, in order to compute the likelihood function more efficiently for large parameters and avoid potential overflow, the equation derived will be implemented. In the centric case, if overflow occurs, either the MLF1 target for centric reflections can be used or an exact probability density for the observed structure-factor amplitude given the calculated amplitude (assuming a Gaussian observational error in structure-factor amplitudes) that we have derived can be implemented and used.

## References


Baker, L. (1992). *C Mathematical Function Handbook*. New York: McGraw-Hill.

Bricogne, G. (1991). *Acta Cryst.* A**47**, 803–829.

Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–474.

Brünger, A. T., Karplus, M. & Petsko, G. A. (1989). *Acta Cryst.* A**45**, 50–61.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.

Chambers, J. L. & Stroud, R. M. (1979). *Acta Cryst.* B**35**, 1861–1874.

Cody, W. J. (1969). *Math. Comput.* **23**, 631–637.

Edwards, A. W. F. (1992). *Likelihood*. Baltimore: Johns Hopkins University Press.

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

Fujinaga, M., Gros, P. & van Gunsteren, W. F. (1989). *J. Appl. Cryst.* **22**, 1–8.

Gradshteyn, I. S. & Ryzhik, I. M. (1980). *Tables of Integrals, Series, and Products*, Corrected and Enlarged Edition, edited by A. Jeffrey. San Diego: Academic Press.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Konnert, J. H. & Hendrickson, W. A. (1980). *Acta Cryst.* A**36**, 344–350.

Luke, Y. L. (1977). *Algorithms for the Computation of Mathematical Functions*. New York: Academic Press.

Lunin, V. Y. & Urzhumtsev, A. G. (1984). *Acta Cryst.* A**40**, 269–277.

Otwinowski, Z. (1991). *Isomorphous Replacement and Anomalous Scattering*. Proceedings of the CCP4 Study Weekend, 25–26 January 1991, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–86. Daresbury: Science and Engineering Research Council.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.

Read, R. J. & James, M. N. G. (1988). *J. Mol. Biol.* **200**, 523–551.

Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Gonschorek, W., Hahn, T., Huml, K., Marsh, R. E., Prince, E., Robertson, B. E., Rollet, J. S. & Wilson, A. J. C. (1989). *Acta Cryst.* A**45**, 63–75.

Silva, A. M. & Rossmann, M. G. (1985). *Acta Cryst.* B**41**, 147–157.

Slater, L. J. (1965). *Handbook of Mathematical Functions*, edited by M. Abramowitz & I. A. Stegun, pp. 503–535. New York: Dover.

Temme, N. M. (1983). *Numer. Math.* **41**, 63–82.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.

Wierenga, R. K., Kalk, K. H. & Hol, W. G. J. (1987). *J. Mol. Biol.* **198**, 109–121.

Wierenga, R. K., Noble, M. E. M., Vriend, G., Nauche, S. & Hol, W. G. J. (1991). *J. Mol. Biol.* **220**, 995–1015.

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

Wolfram, S. (1991). *Mathematica: a System for Doing Mathematics by Computer*, 2nd ed. Reading, MA: Addison-Wesley.