Z. Kristallogr. Suppl. **26** (2007) 27-32
© by Oldenbourg Wissenschaftsverlag, München

Likelihood methods with protein powder diffraction data

J. P. Wright[1,*], A. J. Markvardsen[2] and I. Margiolaki[1]

[1]ESRF, 6 Rue Jules Horowitz, BP-220, F-38043, Grenoble, France.
[2]ISIS Facility, Rutherford Appleton Laboratory, Chilton, OX11 0QX, UK
[*] Contact author; e-mail: wright@esrf.fr

Keywords: powder diffraction, rotation function, likelihood, protein crystallography

Abstract. Likelihood methods for powder diffraction data are developed in the Gaussian approximation for the probability distribution of the intensities arising from a model. These methods are applied to improve the intensity extraction from severely overlapped profiles for a protein sample and also for the computation of the rotation function arising in molecular replacement problems.

Introduction

There is increasing anecdotal evidence that single crystal models for protein structures do not give a very good fit to powder data, particularly the low angle reflections. This is not surprising given the widespread acceptance of crystallographic models which are built into phased electron density maps but only refine to give R-factors of the order 20%. The remaining differences between the model and data cannot be accounted for by the statistical uncertainties on the data alone and the concept of an "error bar" (or probability distribution) for the intensities computed from a model can be introduced to give more rational figures of merit.
For single crystal data there is already an extensive literature [1] dealing with the application of likelihood methods where sometimes the errors on the data can be neglected entirely and probability distributions for the structure factor are derived. Unfortunately for powder diffraction data the peak overlap problem means that the data errors are rarely negligible and the data are linear in terms of intensity so that distributions in terms of intensities are more convenient than structure factors.
Bricogne [2] has previously described the mathematical background and application to powder data to the point where diffractometer counts can be modelled. For practical implementation purposes we have chosen to break up the problem by using instead the correlated integrated intensities derived from a Pawley refinement [3,4]. This allows software which is complex in terms of crystallographic computations to be developed independently of the sophisticated methods for modelling diffractometer counts.
We shall describe an implementation of the computation of a likelihood based figure of merit for powder diffraction data where the distributions of intensities arising from the model are assumed to be Gaussian. Two examples where the model uncertainties are particularly large

are used to illustrate the application of this approach for protein powder diffraction data; the intensity extraction problem and the computation of a rotation function for molecular replacement.

Computation of the likelihood integral in the Gaussian approximation

The likelihood of the combined system of model and data can be expressed by the integral:

$$L = \int P(model)P(data)d\mathbf{I} \,. \tag{1}$$

where P(model) and P(data) are probability distributions for the model and data as a function of the intensities respectively. The usual least squares figures of merit which are commonly in use in Rietveld software are simply the limiting case of having zero uncertainty on the model and a delta function for *P*(*model*). The goodness of fit as a function of the integrated intensities, **I**, can be described by the quadratic form:

$$\chi^2 = \left(\mathbf{I}^o - \mathbf{I}\right)^T \mathbf{W}^o \left(\mathbf{I}^o - \mathbf{I}\right), \tag{2}$$

and the associated data probability distribution is given by $P(data) \propto \exp(-\chi^2/2)$. $\mathbf{I}^o$ is the vector of observed intensities and $\mathbf{W}^o$ the weight matrix derived from a Pawley refinement [4]. Notice $\mathbf{W}^o$ may be singular due to exact peak overlaps in the data. David [5] has shown that the figure of merit in (2) is proportional to the usual Rietveld χ^2. The off diagonal elements in $\mathbf{W}^o$ take account of the peak overlaps and this matrix is sometimes written as $\mathbf{C}^{-1}$, the inverse of the covariance matrix from the Pawley refinement, e.g., the least squares matrix. An analogous quadratic form is assumed for the model and the model weight matrix is taken to be non-singular. With these assumptions the likelihood integral is given by:

$$L \propto \int \exp\{-(\mathbf{I}^o - \mathbf{I})^T \mathbf{W}^o (\mathbf{I}^o - \mathbf{I})\} \frac{\exp\{-(\mathbf{I}^m - \mathbf{I})^T \mathbf{W}^m (\mathbf{I}^m - \mathbf{I})\}}{\det(\mathbf{W}^m)} d\mathbf{I} = \int \frac{\exp(-f(\mathbf{I})/2)}{\det(\mathbf{W}^m)} d\mathbf{I} \,. \tag{3}$$

The terms in the exponentials are both quadratic and can be rewritten as follows

$$2f(\mathbf{I}) = \left(\mathbf{I}^o - \mathbf{I}\right)^T \mathbf{W}^o \left(\mathbf{I}^o - \mathbf{I}\right) + \left(\mathbf{I}^m - \mathbf{I}\right)^T \mathbf{W}^m \left(\mathbf{I}^m - \mathbf{I}\right) = f(\mathbf{I}^s) + (\mathbf{I}^s - \mathbf{I})^T \mathbf{S}(\mathbf{I}^s - \mathbf{I}) \,. \tag{5}$$

where $\mathbf{S} = \mathbf{W}^o + \mathbf{W}^d$ and $\mathbf{I}^s$ is the set of intensities which maximise $f(\mathbf{I})$, that is

$$\left.\frac{\partial f(\mathbf{I})}{\partial \mathbf{I}}\right|_{\mathbf{I}=\mathbf{I}^s} = \mathbf{W}^o \left(\mathbf{I}^o - \mathbf{I}^s\right) + \mathbf{W}^m \left(\mathbf{I}^m - \mathbf{I}^s\right) = 0 \tag{7}$$

which implies

$$\left(\mathbf{W}^o + \mathbf{W}^m\right)\mathbf{I}^s = \mathbf{S}\mathbf{I}^s = \mathbf{W}^o\mathbf{I}^o + \mathbf{W}^m\mathbf{I}^m \tag{8}$$

The vector of intensities $\mathbf{I}^s$ are computed using (8) and using a sparse Cholesky decomposition procedure on the matrix **S** as described in [4]. Inserting (5) into (3) the likelihood integral evaluates analytically to

$$L \propto \det(\mathbf{S}) \exp(-f(\mathbf{I}^s)/2) / \det(\mathbf{W}^m) \,. \tag{9}$$

In obtaining (9) it is assumed that **S** is non-singular. The Gaussian approximation used to obtain (9) refers to the assumption that the errors on the model intensities are assigned to be Gaussian. The justification for this will be based on how well the likelihood expression in (9) works in practical examples. This likelihood expression offers a simple platform for taking

into account uncertainty in both the model and data. More accurate model probability distributions may be proposed, see e.g. [2,6], which results in likelihood functions that are more difficult to evaluate than (9). The likelihood is computed as the logarithm of the likelihood in order to avoid numerical overflows and also to allow constant factors to be more conveniently neglected.

$$\log(L) = \log(\det(\mathbf{S})) - \log(\det(\mathbf{W}^m)) - f(\mathbf{I}^s)/2 + const \qquad (10)$$

The term $f(\mathbf{I}^s)$ can be calculated using Eq. (5). When the model weight matrix is diagonal then $\log(\det(\mathbf{W}^m))$ is simply the sum of the logarithms of the diagonal elements (since log(ab)=log(a)+log(b)). The $\log(\det(\mathbf{S}))$ contribution appears non-trivial at first sight but can be reached provided that a Cholesky decomposition of **S** can be formed. Since $\det(\mathbf{AB})=\det(\mathbf{A})\det(\mathbf{B})$ can be applied to the Cholesky factors and the determinant of a triangular matrix is the product of the diagonal elements we compute
$\log(\det(\mathbf{S})) = \log(\det(\mathbf{L}\mathbf{L}^T)) = 2\sum_i \log(\mathbf{L}_{ii})$.

Application to Intensity Extraction

Due to peak overlaps, a powder diagram does not define a unique set of intensities, different Pawley and Le Bail fitting programs may partition overlapped peaks differently. As a first example we will investigate the set of intensities $\mathbf{I}^s$ which result from the compromise between model and data in Eq (8) in the hope that they might be better than those arising from our Pawley fits [4]. In order to avoid the introduction of any structural model bias a simplified form of Wilson statistics is used. Calculated intensities are found by normalising the data in resolution shells and the observed variance of the extracted intensities is used for the weighting ($\mathbf{W}^m_{ij}=\delta_{ij}/\sigma^m_i{}^2$) as follows

$$\mathbf{I}^m_i = \varepsilon_i \langle I/\varepsilon \rangle_{shell} = \frac{\varepsilon_i}{N_{shell}} \sum\nolimits_{k\in shell} I^o_k/\varepsilon_k \qquad (11)$$

$$(\boldsymbol{\sigma}^m_i)^2 = s\sqrt{2}\varepsilon_i^2 \left(\langle (I/\varepsilon)^2 \rangle_{shell} - \langle I/\varepsilon \rangle^2_{shell} \right) \quad \text{for centric reflections} \qquad (12)$$

$$(\boldsymbol{\sigma}^m_i)^2 = s\varepsilon_i^2 \left(\langle (I/\varepsilon)^2 \rangle_{shell} - \langle I/\varepsilon \rangle^2_{shell} \right)/\sqrt{2} \quad \text{for acentric reflections} \qquad (13)$$

where ε is the systematic enhancement factor, the number of times (hkl) transforms into itself under the symmetry operations of the space group and s is a scale factor for the weights. The value of s is varied to maximise the value of the likelihood and is typically close to one. The factors of $\sqrt{2}$ are included to give a broader distribution of intensities for centric reflections compared to acentric reflections. This simplified model avoids the difficulty of modelling the protein and solvent contributions to the structure factor by using the mean and variance of the initially extracted intensities as model values.

This model was applied to powder profiles for the protein trypsin collected at the high resolution powder diffraction beamline (ID31, λ=1.25 Å) and also to area detector data collected at the materials science beamline (ID11, λ=0.484 Å), both at the ESRF. The intensities, $\mathbf{I}^s$, were compared to values found for a single crystal taken from the same sample (also measured at beamline ID11) by computing a correlation coefficient between the structure factors. The $\mathbf{I}^s$ intensities are a significant improvement on the values extracted during the Pawley fit for both the datasets. For the area detector data the overall values are: $CC(\mathbf{I}^s)=0.308$ versus $CC(\mathbf{I}^o)=0.232$ and for the high resolution data $CC(\mathbf{I}^s)=0.729$ is also better than $CC(\mathbf{I}^o)=0.612$.

Figure 1 (left) shows the correlation coefficients as a function of resolution and Fig. 1 (right) shows a small region of both the ID31 and ID11 data. The highly resolved (ID31) data show most improvement at higher angles where there is more peak overlap while the area detector data (ID11) are improved more at low resolution, where the peak overlap is already signifi-

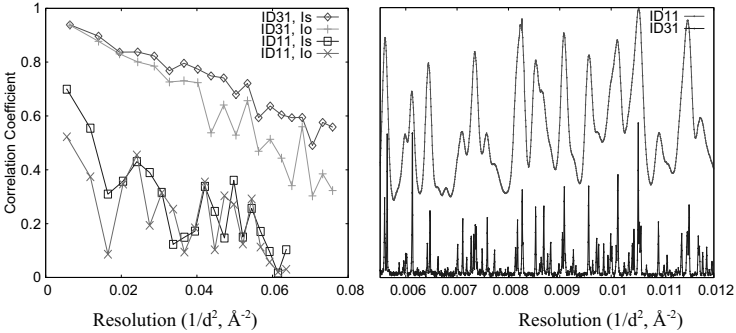

Figure 1: Left: Correlation coefficient between extracted structure factors and single crystal data versus resolution for high-resolution (ID31) and area detector (ID11) data. Right: Comparison of the powder profiles in the low angle region.

cant. For the high angle part of the area detector data the improvements are only marginal, which is perhaps due to the very severe peak overlap presenting an unrecoverable loss of information.

Computation of rotation functions

One of the major motivations for this work has been to find a way to compute a rotation function which takes into account the peak overlaps in powder diffraction data. In single crystal molecular replacement problems the orientation of a protein molecule can often be found independently of the position in the unit cell by looking at the correlation between the observed Patterson function and the Patterson for a single molecule. This is sometimes rationalised in terms of intra-molecular vectors generally lying closer to the Patterson origin than inter-molecular vectors. For powder data no unique Patterson function exists and so a likelihood based approach working entirely in reciprocal space has been developed.

Here a point by point calculation is carried out where the molecule is oriented in the unit cell in spacegroup P1 and a set of structure factors are computed. The structure factors for symmetry related molecules are found by applying the symmetry operators to the hkl indices. For reflections which transform into themselves the relative phase term is fixed by symmetry and the structure factors are added directly. Otherwise, the symmetry generated structure factors are then treated as statistically *independent* contributions to the total structure factor. This gives the set of contributing structure factors which will add together in a way that depends on the position of the molecule in the unit cell. The computed intensity is given by $I = |F|^2 = F^*F$ and can be written in terms of the relative phase terms as:

$$I = \left(\sum F_i\right)\left(\sum F_j\right)^* = \sum_i |F_i|^2 + 2\sum_{i<j} |F_i||F_j| \cos\phi_{ij} \,. \qquad (14)$$

The intra-molecular terms in the first summation in (14) do not depend on the relative positions of the molecules in the unit cell. The inter-molecular terms vary with the position of the molecule in the unit cell and represent the interference function between the molecules. We may treat the second term as a one-dimensional random walk with $N(N-1)/2$ steps each having an average step length $2<|F_i||F_j|>$. The mean contribution to the intensity from this term is zero, assuming the relative phases, ϕ_{ij}, are randomly distributed and the standard deviation is:

$$\sigma_r = \frac{1}{2}\left\langle |F_i||F_j| \right\rangle \sqrt{N(N-1)}\,, \qquad (15)$$

where N is the number of independent contributions to the intensity in (14). Figure 2 illustrates these concepts graphically to show how reflections can be predicted as being weak with small uncertainties or strong with either small or large error *without* positioning the molecule in the unit cell.

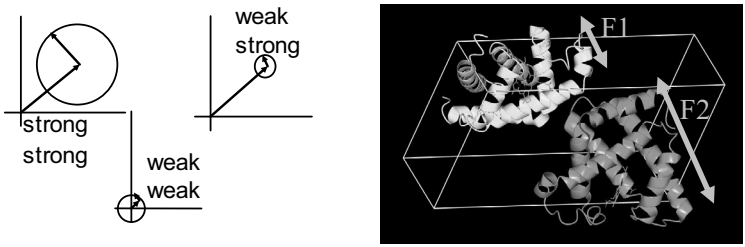

Figure 2: Left: Illustration of the way the intensity is partially determined by the orientation of the molecule in the unit cell for different reflections. Right: Illustration of the differing contributions to the structure factor for the two symmetry related molecules of myoglobin in space group $P2_1$.

Where there is only a single independent contribution to the structure factor (e.g. 0k0 reflections in $P2_1$ and one molecule in the asymmetric unit) the computed intensity is known and no uncertainty is introduced by the lack of position for the protein molecule in the unit cell. To avoid numerical difficulties in this case and also to more realistically reflect reality we introduce an additional uncertainty arising from missing atoms in the model (a Wilson type contribution) and also a term which is proportional to the magnitude of the bulk solvent correction, which has been computed from an exponential scaling model. Using the following equations for $\mathbf{I}^m$ and $\mathbf{W}^m$ the likelihood can be computed as a function of the orientation of the molecule in the unit cell.

$$\mathbf{I}^m = \sum |F_i|^2 \qquad (16)$$

$$\mathbf{W}^m = 1/(\sigma_r^2 + \sigma_s^2 + \sigma_w^2) \qquad (17)$$

Figure 3 shows a one dimensional cut through the likelihood based powder rotation function for the same trypsin data used above for intensity extraction and model 1S81 from the pdb database. The likelihood figure of merit is compared to the scenario where the data are taken to be the likelihood enhanced intensity extraction obtained in the previous section and assumed to be error free. This represents the case where single crystal molecular replacement software is used, which does not have access to overlap information and the model errors are much greater than the uncertainties on the data. In both test cases there is a maximum at the true orientation of the molecule in the unit cell. The signal to noise is much better for the high resolution data (right), and the improvement from using the likelihood based figure of

merit is much greater for the more severely overlapped data (left). Compared to using only the extracted intensities and a standard single-crystal chi-squared figure of merit; additional noise is introduced into the rotation function due to incorrect partitioning of the overlapped peaks and this problem is alleviated by the likelihood based figure of merit.

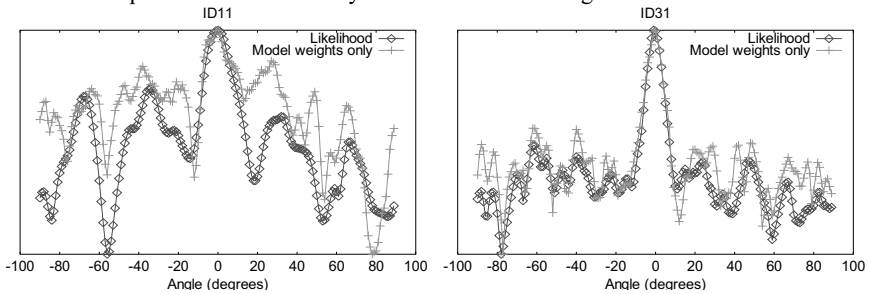

Figure 3: Rotation functions using the likelihood figure of merit and model weights only for trypsin and model pdb code 1S81 (space group $P2_12_12_1$). The 1D section is a rotation of the molecule around crystallographic b-axis. Left: Severely overlapped area detector data from ID11. Right: High resolution ID31 powder profile

Concluding remarks

We have shown that likelihood integrals can be performed in a computationally efficient manner in the Gaussian approximation for overlapped powder diffraction data. This methodology sacrifices strict correctness for computational viability by insisting on a Gaussian distribution for the model probability distribution. Our results show that derived intensities from a Pawley fit [4] modified using the likelihood expression yields new intensities that give an improved match to single-crystal intensities. The degree of improvement naturally depends on the precise details of the initial intensity extraction, but we hope that this will be of wide application as centric and acentric reflections are being matched to different distributions. The computation of a rotation function for molecular replacement from powder diffraction data is shown to be viable and the quality of the resulting likelihood based function appears to be significantly better than the function resulting from using extracted intensities.

References

1. McCoy, A. J., 2004, *Acta Cryst.* **D60**, 2169-2183.
2. Bricogne, G., 1991, *Acta Cryst.*, **A47**, 803.
3. Pawley, G. S., 1981, *J. Appl. Cryst.*, **14**, 357-361.
4. Wright, J.P., 2004, *Z. Krist*, **219,** 791-802.
5. David, W.I.F, 2004, *J. Appl. Cryst.* **37,** 621-628.
6. Markvardsen, A.J., David, W.I.F. and Shankland, K., 2002, *Acta Cryst*., **A58**, 316.

Acknowledgements. We thank the ESRF for provision of synchrotron beamtime.