Bayesian probability theory applied to the space group problem in powder diffraction

A. J. Markvardsen

ISIS Facility, Rutherford Appleton Laboratory, Chilton, Oxon OX11 0QX, England

Abstract. Crystal structure determination from powder diffraction data has become a viable option for molecules with less than 50 non-hydrogen atoms in the asymmetric unit and this includes the majority of compounds of pharmaceutical interest [1, 2]. The solution of crystal structures, including space group determination, is more challenging from powder diffraction data than from single crystal diffraction data. Here, it will be demonstrated how a Bayesian probability analysis of this problem has helped to provide a new algorithm for the determination of the space group symmetry of a crystal from powder diffraction data. Specifically, the relative probabilities of different extinction symbols are accessed within a particular crystal system. Examples will be presented to illustrate this approach.

INTRODUCTION

X-ray and neutron diffraction techniques are powerful methods for elucidating structural details of molecules. A X-ray (or neutron) beam scattered from a single-crystal produces diffraction spots that are scattered in well-defined directions away from the sample. The intensity detected for each diffraction spot is interpreted to be proportional to the amplitude squared of a Fourier component of the unit-cell electron density. In crystallography, each of these Fourier components is expressed as a structure factor: $F = |F|e^{i\theta}$, where $|F|$ is the amplitude and $e^{i\theta}$ is the phase. The fact that only the structure factor amplitudes are measured in a diffraction experiment leads to the famous 'phase problem', where the electron density must be determined without knowing the structure factor phases.

Not all diffraction experiments, investigating crystalline material, are performed using single-crystals. The main alternative is to use crystalline powder consisting of a large number randomly orientated crystallites. The diffraction pattern from a collection of randomly oriented crystallites no longer provide well-defined spots but rather concentric circles, see Figure 1. Clearly, the information from a powder diffraction experiment is significantly less than that of a single-crystal experiment. Still, powder diffraction experiments are widely used because crystalline powder may be easier to produce than single-crystals suitable for a single-crystal experiment. Therefore, much emphasis has been put into developing the powder diffraction technique. Recent advances in both laboratory and synchrotron powder diffraction equipment and developments in data analysis methods means that it is now possible to solve 50 non-hydrogen atom crystal structure from powder diffraction data, see for instance [1]. Analysis of powder diffraction data has some unique challenges not present in analysis of single-crystal

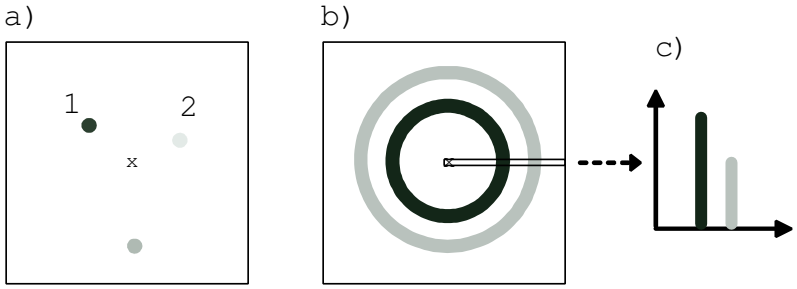

FIGURE 1. (a) Shows a schematic single-crystal diffraction pattern, and (b) the equivalent powder diffraction pattern. A powder diffraction pattern is typically a slice of the 2D plot in (b) as shown in (c).

data. To understand this consider Figure 1. Spots 1 and 2 in Figure 1(a) are completely separated in single crystal data, but in the corresponding powder data these spots become part of the same ring, shown as the inner ring in Figure 1(b). The result of this is what is known as the peak overlap problem in powder diffraction, and essentially means that a given peak in a diffraction pattern cannot be guarantied to originate from just *one* diffraction spot. For instance, from Figure 1(a) both the structure factor intensity 1 and 2 can be reliably estimated, denote these $|F_1|^2$ and $|F_2|^2$, but the powder data (Figure 1(b)) only has information about the sum of these, i.e. $|F_1|^2 + |F_2|^2$. A consequence of the peak overlap problem is that it is often difficult to determine the unit cell and space-group of the crystal; if this cannot be done the analysis process is often not continued. In what follows an intensity or structure factor will also be referred to as a reflection.

The unit cell and space group determination is typically a two stage process. In the work presented here it is assumed that the unit cell has been determined. Space group symmetry affects the diffraction pattern in a number of ways. Most importantly, space group symmetry can cause certain structure factor components to be absent from the diffraction pattern. Such missing components are called systematic absences and identifying absences enable crystallographers to identify possible space groups. A small number of space groups may have the same systematic absences and an extinction symbol notation is used in crystallography to catagorize the space groups according to their systematic absensces.

Traditionally, space group determination, from powder diffraction data is performed manually by inspection of the systematically absent reflections. However, the partial or complete peak overlap can make this manual inspection time consuming and ambiguous. Here, an algorithm is presented that gives a quantitative measure of the relative probabilities of different extinction symbols. The work presented here extends the work in [3] by treating the case where three or more reflections completely overlap and discusses how the relative probability of space groups with the same extinction symbol can be distinguished.

BAYESIAN PROBABILITY ANALYSIS

Let S_{gr} denote a space group and $\mathbf{I^P}$ and $\mathbf{C}$ the data. The probability we would like to evaluate is $p(S_{gr}|\mathbf{I^P},\mathbf{C})$, which is the probability of some space group given the data. If this probability is calculated for all possible S_{gr} it would provide, from a Bayesian viewpoint, the best starting point for selecting a space group.

Let $\mathbf{I^P}$ be a vector of extracted intensity values from a powder diffraction pattern: $\mathbf{I^P} = (I_1^P, I_2^P, \ldots, I_N^P)$, and let the matrix $\mathbf{C}$ hold information about intensity correlations between neighbouring reflections in the powder pattern. For example, if none of the reflections in the diffraction pattern are found to overlap then the correlation matrix is an $N \times N$ diagonal matrix. The data in a powder diffraction pattern may be summarised by a multivariate normal probability distribution

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

where $\mathbf{I^P}$ are the 'observed' intensities and $\mathbf{I} = (I_1, I_2, \ldots, I_N)$ some set of intensity values.

Consider an isolated reflection in the diffraction pattern. A given S_{gr} will either predict this intensity to be present or absent. Imagine that a value for this reflection is measured to $I^P = 10$ with the variance $C = \sigma^2 = 1$, what numbers should be assigned for the present/absent probabilities: $p(present|10,1)$ and $p(absent|10,1)$? Intuitively, the later probability may be expected to be close to zero since the probability of a reflection being absent given that it is measured to 10 ± 1 sounds unlikely. Likewise, what number should be assigned to $p(present|10,1)$? The exercise which follows aims to derive reasonable mathematical expressions for these probabilities and ultimately $p(S_{gr}|\mathbf{I^P},\mathbf{C})$, which takes into account all the available data.

Using Bayes' theorem $p(S_{gr}|\mathbf{I^P},\mathbf{C})$ can be decomposed to

$$p(S_{gr}|\mathbf{I^P},\mathbf{C}) = p(S_{gr})p(\mathbf{I^P},\mathbf{C}|S_{gr})/p(\mathbf{I^P},\mathbf{C}) \quad , \tag{2}$$

where $p(\mathbf{I^P},\mathbf{C})$ is constant since the data are kept constant. Assuming that no space group is favored to any other space group then $p(S_{gr})$ is uniformly distributed in which case it follows from Equation 2 that

$$p(S_{gr}|\mathbf{I^P},\mathbf{C}) \propto p(\mathbf{I^P},\mathbf{C}|S_{gr}) \quad . \tag{3}$$

The likelihood probability density $p(\mathbf{I^P},\mathbf{C}|S_{gr})$ can further be written as the integral

$$p(\mathbf{I^P},\mathbf{C}|S_{gr}) = \int p(\mathbf{I^P},\mathbf{C},\mathbf{I}|S_{gr})d\mathbf{I} \tag{4}$$

where the integrand may be decomposed to

$$p(\mathbf{I^P},\mathbf{C},\mathbf{I}|S_{gr}) = p(\mathbf{I}|S_{gr})p(\mathbf{I^P},\mathbf{C}|\mathbf{I},S_{gr}) \quad . \tag{5}$$

Consider this joint probability density for an isolated reflection in the diffraction pattern, and where the space group S_{gr} predicts this reflection to be absent. Unlike the observed

intensity I^P, the intensity I is assumed to have no uncertainty associated with it, and therefore

$$p(I|absent) = \delta(I) \quad , \tag{6}$$

i.e. an absent reflection cannot have any probability of non-zero intensity. Because of the delta-function in Equation 6, Equation 5 only needs to be considered for $I = 0$. Assuming the data are normally distributed then Equation 6 becomes

$$p(I^P, C = \sigma^2, I|absent) = \delta(I) \exp\left[-(I^P/\sigma)^2/2\right] \tag{7}$$

for an isolated reflection, which is predicted to be absent.

Alternatively, S_{gr} may predict the reflection to be present and an expression for $p(I|present)$ is required. A number of works have been published in the literature of crystallography that discuss this probability density for different scenarios. For instance, different expressions may be assigned to this distribution depending on whether the underlying space group S_{gr} is centric (centrosymmetric) or acentric (non-centrosymmetric); Wilson showed this in 1949 [4], and the following probability expressions are often referred to as the Wilson distributions in crystallography

$$p(I|present, acentric) = \begin{cases} 0 & I < 0 \\ \mu^{-1} \exp(-I/\mu) & I \geq 0 \end{cases} \tag{8}$$

and

$$p(I|present, centric) = \begin{cases} 0 & I < 0 \\ \pi^{-1} (\mu I)^{-1/2} \exp(-I/\mu) & I \geq 0 \end{cases} \quad . \tag{9}$$

From single crystal diffraction data it is often possible to identify whether a dataset exhibit centric/acentric behavior based on an analysis of the measured intensities, although this is not always unambiguous [5]. For powders an equivalent analysis may be expected to give more ambiguous results and here $p(I|present)$ is assigned to the expression in Equation 8 regardless of whether S_{gr} is centric or acentric. In practice, this choice is justified by fitting Equation 8 to the distribution of observed intensities. This is illustrated in Figure 2; this fitting procedure also enables μ to be determined [3]. To summarize we assign

$$p(I^P, C = \sigma^2, I|present) = \mu^{-1} \exp(-I/\mu) \exp\left[-((I^P - I)/\sigma)^2/2\right] \tag{10}$$

for an isolated reflection, which is predicted to be present.

A diffraction pattern consists of many reflections, some of which are isolated, other partially overlap and some are completely overlapping. In assigning the joint probability distribution $p(\mathbf{I}|S_{gr})$, the trivial assumption is made that all reflections are independently distributed, i.e.

$$p(\mathbf{I}|S_{gr}) = \sum_{i=1}^{N} p(I_i|S_{gr}) \quad . \tag{11}$$

Inserting Equation 4 into Equation 3 we have

$$p(S_{gr}|\mathbf{I^P}, \mathbf{C}) \propto p(\mathbf{I^P}, \mathbf{C}|S_{gr}) = \int \sum_{i=1}^{N} p(I_i|S_{gr}) p(\mathbf{I^P}, \mathbf{C}|\mathbf{I}) d\mathbf{I} \tag{12}$$

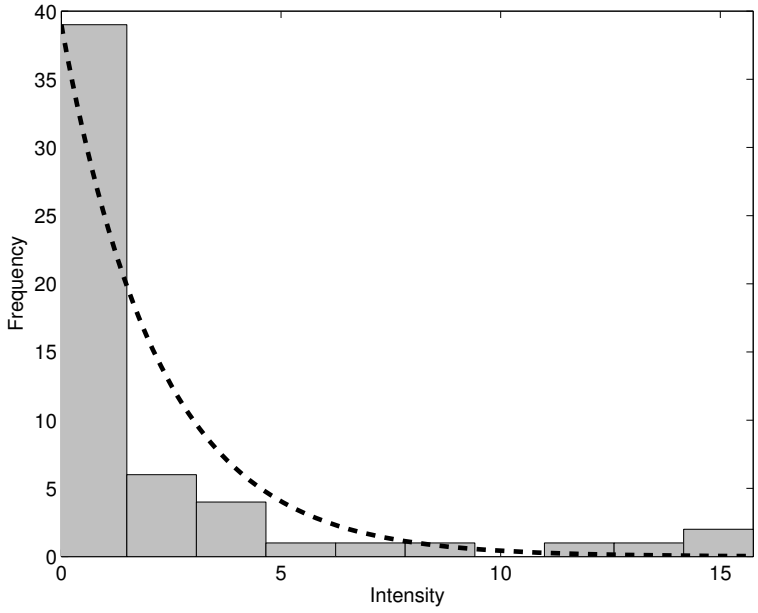

FIGURE 2. Histogram plot of reflection intensity (grouped into a number of intervals) versus the number of isolated intensity values falling into each of these intervals for the compound remacemide nitrate. The Wilson distribution (Equation 8) is plotted as the dashed line.

where $p(I_i|S_{gr})$ is given by either Equation 6 or Equation 8 and the likelihood $p(\mathbf{I^P},\mathbf{C}|\mathbf{I})$ by Equation 1. When peaks in the diffraction pattern partially overlap this integral cannot be evaluated analytically; an efficient and accurate method of dealing with this case is described in [3]. For isolated and completely overlapping reflections the integral in Equation 12 can be evaluated analytically.

Consider an isolated peak in the diffraction pattern that has contributions from N_o completely overlapping reflections; denote these $I_1, I_2, \ldots, I_{N_o}$. The observed intensity for this peak, I^P, measure only a value for the sum of these N_o reflections with some standard deviation σ. The part of the integral in Equation 12 that covers the integration over the variables $I_1, I_2, \ldots, I_{N_o}$ reads

$$\begin{aligned} p(I^P,\sigma|allpresent) &= (2\pi)^{-1/2}\sigma^{-1}\mu^{-N_o}\int_0^\infty\cdots\int_0^\infty dI_1\cdots dI_{N_o} \\ &\quad \exp\left\{-[I^P-(I_1+\cdots+I_{N_o})]^2/(2\sigma^2)-(I_1+\cdots+I_{N_o})/\mu\right\} \quad . \end{aligned} \tag{13}$$

As seen this integral may be solved analytically since the integrand only depends on the sum of the intensities. Make the following change of variables

$$\begin{aligned} I &= I_1+I_2+\cdots+I_{N_o} \\ I_i' &= I_i \text{ for all } i=2,3,\ldots,N_o \quad . \end{aligned} \tag{14}$$

The determinant of the Jacobian for this change of variables is 1 and the integral in Equation 14 becomes

$$p(I^P,\sigma|allpresent) = \frac{1}{(2\pi)^{1/2}\sigma\mu^{N_o}}\int_0^\infty dI \exp\left\{-\frac{(I^P-I)^2}{2\sigma^2}-\frac{I}{\mu}\right\} f(I;N_o) \quad , \qquad (15)$$

where

$$f(I;N_o) = \int_0^I dI'_2 \int_0^{I-I'_2} dI'_3 \cdots \int_0^{I-I'_2-I'_3-\cdots-I'_{N_o}} dI'_{N_o} \quad . \qquad (16)$$

It is found that $f(I;N_o) = I^{N_o-1}/\Gamma(N_o)$ for $N_o \leq 5$ and this expression is assumed valid for all $N_o = 1,2,\ldots$. Equation 15 reads

$$p(I^P,\sigma|allpresent) = \frac{1}{(2\pi)^{1/2}\sigma\mu^{N_o}\Gamma(N_o)}\int_0^\infty dI \exp\left\{-\frac{(I^P-I)^2}{2\sigma^2}-\frac{I}{\mu}\right\} I^{N_o-1} \quad . \qquad (17)$$

This integral can for instance be identified in [6] and the solution to Equation 17 may be written as

$$p(I^P,\sigma|allpresent) = \frac{\sigma^{N_o-1}}{(2\pi)^{1/2}\mu^{N_o}}\exp\left\{\frac{z^2}{2}-\frac{(I^P)^2}{2\sigma^2}\right\} D_{-N_o}(\sqrt{2}z) \quad , \qquad (18)$$

where $z = 2^{-1/2}(\sigma/\mu - I^P/\sigma)$ and D_i a parabolic cylinder function as defined in [6].

IMPLEMENTATION AND DISCUSSION

A program was written (in C/C++) to test the approach outlined above and in connection with the work in [3]. The current implementation of the program accept an input data file that read in values for the $\mathbf{I^P}$ data vector and correlation matrix $\mathbf{C}$, and information about which reflections are treated as completely overlapping. This program is available for free from the author, but is also part of the commercial structure solution package DASH (see `http://www.ccdc.cam.ac.uk` for more information).

To demonstrate the program consider the space group determination of dopamine hydrobromide. The powder diffraction pattern of this compound is shown in Figure 3. The unit cell for this compound has orthorhombic symmetry. The number of orthorhombic extinction symbols is 111 and one of these, denoted $P---$ has no systematic absences. The program outputs a table listing all possible extinction symbols and the relative log probability of each extinction symbol relative to $P---$. The 11 most probable extinction symbols for the dopamine hydrobromide dataset are shown in Table 1. It is found that the majority of orthorhombic extinction symbols are extremely improbable, with only six extinction symbols being more probable than the one corresponding to the extinction symbol having no systematic absences. Of these six possibilities it is clear that $Pbc-$ is much more probable, given the data, than the next choice $Pb--$ and so on. By investigating which reflections are systematically absent for each extinction symbol in Table 1, it is not surprising that the second to sixth ranked choices in that table are more

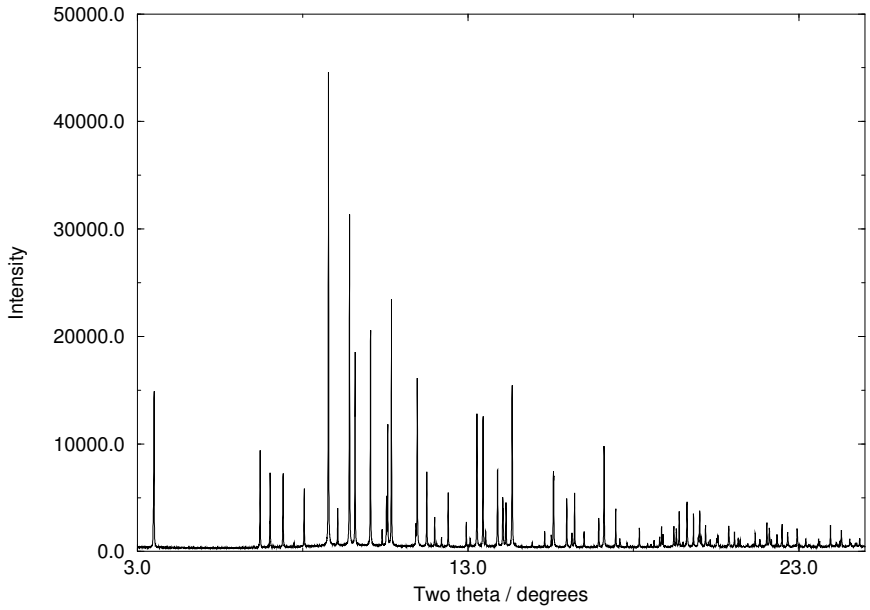

FIGURE 3. Dopamine hydrobromide synchrotron powder diffraction data, measured on BM16 of the ESRF, Grenoble [7].

probable than $P---$ since all these contain subsets of the systematic absent reflections for the most probable choice $Pbc-$.

A powder diffraction dataset is summarized by the multidimensional normal distribution in Equation 1. A representation of the dopamine hydrobromide dataset consisted of 184 observed intensities where only two of these measured the *sum* of 3 or more reflection intensity values, i.e., where 3 or more reflections are considered as completely overlapping and denoted multiplets of order > 2. Numbers in parentheses in Table 1 were obtained using a previous version of the algorithm [3] in which multiplets of order > 2 were not considered. These are similar to the probability ratios calculated with the new algorithm, see Table 1, as this dataset contains only two multiplets of order > 2. Still, the results in Table 1 demonstrate the potential benefit to be gained by including information from higher order multiplets. For instance, the improved probability ratios provide a better discrimination in probability of the $Pbc-$ extinction symbol, and a better overall discrimination of the first three listed extinction symbols from the first 7 listed extinction symbols. The importance of including multiplets of order > 2 in calculating extinction group probabilities scales with the level of peak overlap in the diffraction pattern. A higher peak overlap is, for instance, expected when a laboratory X-ray source is used rather than a synchrotron source.

Table 2 shows the results of running the extinction group program on a diffraction pattern of MgIr [8]. The executation time for generating this table was 4-5 seconds on a laptop 2.5GHz processor. The most probable extinction symbol in Table 2 is the observed one [8].

TABLE 1. Extinction symbols and log probabilities for dopamine hydrobromide, calculated as $\ln[p(E_{gr}|data)/p(P---|data)]$, where E_{gr} is an extinction symbol, for the 9 most probable choices. The numbers in the parentheses show the log probabilities as taken from Table 2 in [3].

Symbol	Prob. ratio	Symbol	Prob. ratio	Symbol	Prob. ratio
$Pbc-$	101.9 (97.9)	$P-2_12_1$	21.1 (21.1)	$P---$	0 (0)
$Pb--$	52.5 (50.3)	$P-2_1$-	15.2 (15.2)	$Pbcb$	-383.4 (-388.5)
$P-c-$	46.1 (43.9)	$P--2_1$	5.9 (5.9)	$P-cb$	-420.3 (-423.6)

TABLE 2. Extinction symbols and log probabilities for MgIr, calculated as in Table 1 for the 9 most probable choices.

Symbol	Prob. ratio	Symbol	Prob. ratio	Symbol	Prob. ratio
$C-c(ab)$	174.4	$C--2_1$	158.3	$Pbca$	72.0
$C--(ab)$	167.7	$C---$	155.7	$Pbaa$	71.3
$C-c-$	162.4	$Pbab$	73.6	$Pbcb$	70.9

CONCLUSIONS

The benefit of the Bayesian approach has been demonstrated for the problem of space group determination from powder diffraction data. The case of completely overlapping reflections was generalized and centric/acentric space group discrimination discussed.

ACKNOWLEDGMENTS

I would like to acknowledge John Johnston, Bill David, Kenneth Shankland and Devinder Sivia for their contributions in developing the space group determination method.

REFERENCES

1. David, W. I. F., Shankland, K., McCusker, L., and Baerlocker, C., editors, *Structure determination from powder diffraction data*, Oxford University Press, Oxford, 2000.
2. David, W. I. F., Shankland, K., and Markvardsen, A. J., *Cryst. Reviews*, **9**, 3–15 (2003).
3. Markvardsen, A. J., David, W. I. F., Johnson, J. C., and Shankland, K., *Acta Cryst.*, **A57**, 47–54 (2001).
4. Wilson, A. J. C., *Acta. Cryst.*, **2**, 318–321 (1949).
5. Giacovazzo, C., *Direct phasing in crystallography*, Oxford University Press, New York, 1998.
6. Gradshteyn, I. S., and Ryzhik, I. M., *Table of integrals, series and products*, Academic Press, San Diego, CA, 2000.
7. Shankland, N., Love, S. W., Watson, D. G., and Shankland, K., *J. Chem. Soc. Faraday Trans.*, **92**, 4555–4559 (1996).
8. Černý, R., Renaudin, G., Favre-Nicolin, V., Hlukhyy, V., and Pöttgen, *Acta Cryst.*, **B60**, 272–281 (2004).