Curso 2006/07
CIENCIAS Y TECNOLOGÍAS/2
I.S.B.N.: 978-84-7756-736-3

JESÚS IVÁN DA SILVA GONZÁLEZ

Últimos avances en la resolución de estructuras cristalinas a partir de datos de difracción en polvo

Directores
JAVIER GONZÁLEZ PLATAS
CARMELO GIACOVAZZO

SERVICIO DE PUBLICACIONES
UNIVERSIDAD DE LA LAGUNA

SOPORTES AUDIOVISUALES E INFORMÁTICOS
Serie Tesis Doctorales

Esta Tesis se ha realizado gracias a la financiación por parte del proyecto Marie-Curie de la Comunidad Europea, a través del cual obtuve la beca para mi estancia en Bari, Italia (Ref. HPMT-CT-2001-00425), y por parte de la Universidad de La Laguna y de Cajacanarias, a través de la beca de colaboración en el Servicio Integrado de Difracción de Rayos-X (Ref. SIDIX-1).

AGRADECIMIENTOS

En primer lugar, quiero expresar mi más sincero agradecimiento a la Dra. Cristina González Silgo y al Dr. Javier González Platas, quienes me han ayudado tanto a lo largo de estos años y me han hecho descubrir y disfrutar un tema de investigación realmente bonito.

Al Profesor Carmelo Giacovazzo, por brindarme la oportunidad de trabajar, durante algo más de un año, en un grupo de investigación tan importante en su campo, como es el *Istituto di Cristallografia*.

Por otra parte, no me olvido de agradecer a mi familia y a todos mis amigos quienes, en algún momento, me han tenido que soportar (...tranquilos, que todavía les queda mucho por aguantar!!).

Respecto al grupo de Bari, gracias a todos por haberme hecho sentir como en casa durante todo ese tiempo que compartimos y, sobre todo, por esos cafés y bocadillos!! Debo dar un agradecimiento especial a Marcello, quien se portó realmente de maravilla conmigo. Gracias, amigo.

A mis padres, por invitarme a este maravilloso viaje

A Ana, por acompañarme durante todos esos años

Índice

Index

Introduction

1.1 State of the Art

Diffraction by polycrystalline solids offers a wide variety of applications to material scientists and, today, it is one of the most important experimental techniques that are available to solid-state research. From the beginning, the powder diffraction method has been very popular, mainly used for various procedures such as phase identification and quantitative phase analysis, the precise determination of unit cell parameters or line-broadening analysis, aimed to the study of structure imperfections.

Half a century after the introduction of the powder method, developed independently by Debye and Scherrer in 1916 (Germany) and by Hull in 1917 (United States), the diffraction theory was well established for those techniques just mentioned above and several applications were reported in diverse material science fields such as metallurgy or clay mineralogy. During the last decades, these techniques have become more accurate.

So, from a powder diffraction pattern we can obtain information about the microstructure of the solid under study; but, moreover, it also contains information about its own crystal structure. Because of this fact, the effort to extract this three-dimensional information was very important although,

directly available from the pattern, there is only one-dimensional data. It would be the introduction of the powerful Rietveld method (Rietveld, 1969) for refining crystal structures from powder diffraction data what made the extraction of that structural information a realistic process.

Diverse and very important developments in the instrumentation (in conventional laboratory X-rays as well as in synchrotron X-rays and neutron radiation) and, like in most research fields, the increasing power, storage and availability of computers have completely transformed the powder method. Thus, these advances have greatly benefited the modern powder diffraction applications (*e.g.* phase, microstructural and structural analysis).

Although since its introduction the powder method has been widely used, it has been in the recent years when its significance has been emphasized because of its importance in the characterization of high-T_c superconductors, zeolites, nanocrystallite solids and, principally, the continuous and considerable development of the structure determination for those materials that can be only obtained in polycrystalline form. At the present time, the method is very used in several fields of material science or chemistry and, more recently, is being applied to pharmaceutical compounds. Furthermore, the method does not restrict us either to analyse only one diffraction pattern or to measure it at room temperature; in fact, the versatility of the modern applications allows us to deal with data collected under non-ambient conditions (pressure, temperature) or even in a dynamic mode. The expansion of this last area of research would not be possible without the availability of X-ray synchrotron and neutron radiation sources (which become each time more intense) and, besides it, the fast development of new detector technology which, for example, allows us to complete the measurements in a shorter time range.

Unfortunately, even though with the help provided by those improvements, the structure determination using powder diffraction data continues to be much more difficult than using single crystal data. This fact is owed almost entirely to the collapse of the three-dimensional crystallographic information of the sample onto the one-dimensional powder diffraction pattern. This collapse introduces an uncertainty in the data extracted from the pattern, creating important problems (often hard or even not possible to overcome) in the determination of the unit cell and in the process of applying Direct or Patterson Methods, in order to achieve the crystal structure determination. Actually, all the steps of the pathway to structure solution, when working with single crystal data, are less complicated to get through than their powder diffraction equivalents.

So, if the single crystal approach is much easier, why then powder diffraction should be used to solve the structure of crystalline materials? The answer to this question is very simple: because there are many compounds that cannot be obtained in single crystal form (either because the dimension of the crystal is not big enough or because it has been obtained a crystal of poor quality) but, on the contrary, can be crystallized in powder form. Of course if, for a certain material we want to work with, a single crystal is available, then single crystal diffraction data will be much better for our purpose of determining its structure.

1.1.1 Using powder diffraction data to solve crystalline structures

On the *ab initio* structure determination pathway by using powder diffraction data, a series of particular procedure steps must be carried out. They range

from the initial collection of data to the final refinement stage and can be summarized as follows:

i) it is required that the sample must be carefully prepared, in order to reduce some problems (*e.g.* preferred orientation effects) that can arise from the measurement, which should preferably be done with a high resolution diffraction instrument. The fact of spending a little time at this step could benefit on saving a lot of time and effort to overcome those possible problems on the subsequent steps. So, contrary to what it can seem, this step of the process is not a trivial task.

ii) the resulting diffraction peaks observed in the measured powder pattern are singled out and submitted to an indexing procedure, obtaining the cell dimensions and the corresponding crystal system.

iii) the space group is determined by inspection of the systematic absences.

iv) to every indexed reflection an integrated intensity is assigned by pattern fitting procedure (see §1.1.6) and, from them, structure factor amplitudes $|F_{hkl}|$ are derived.

v) solution of the phase problem: the obtained set of structure factors is used as input for Direct or Patterson Methods, in order to achieve structure solution.

vi) the resulting atomic coordinates are refined by using the Rietveld method.

The phase assignment problem, present in the fifth step, shows harder difficulties for the powder diffraction case than for the single-crystal diffraction case, which comes out from the smaller available set of unequivocal Bragg intensities. This is one of the biggest problems that must be overcome and is, at present, a subject intensively studied.

1.1.2 Powder diffraction data measurement

Depending on what information we want to obtain from the compound under study, the experiment must be conceived differently.

Before we can proceed with the diffraction measurement, there are three general issues we have to take into account:

a) what radiation is most suitable for our purposes? The answer to this question (*i.e.* laboratory X-ray, synchrotron or neutron radiation diffractometer) depends, in fact, on what is our purpose. For example, neutron radiation is most appropriated for compounds with high absorption or containing light atoms (*i.e.* deuterated samples). On the other hand, synchrotron radiation offers by far the best resolution (*i.e.* narrowest diffraction peaks); but it is clear that with a conventional laboratory X-ray diffractometer there is much more measurement-time available, offering a resolution between those of neutron and synchrotron radiations. As a general rule, a first attempt of the study is made using a laboratory diffractometer and, if unsuccessful results are obtained, then a new data set is obtained, by using a more suitable radiation.

b) the second aspect is related with the diffractometer geometry and configuration in which the measurement is carried out. Reflection geometry is

suitable for high absorption materials (and when there is enough amount of material available), but it can easily suffer from aberrations such as preferred orientation or sample transparency, which will cause unbiased reflection intensities and peak shifts. Transmission geometry is almost always used for data collection with synchrotron and neutron radiation, becoming more popular on laboratory diffractometers; the main advantages of this geometry are the almost complete elimination of the preferred orientation effects and that a small amount of sample is needed for the experiment. On the contrary, lower peak intensities will be obtained (compared with the reflection geometry) and sample absorption will vary as a function of the diffraction angle, so a correction of the calculated intensities has to be made.

c) finally, once we have chosen the diffraction equipment and geometry, we have to decide the most suitable angular range in which measure the diffraction profile. It will depend on the study that will be carried out: if we want to index the pattern, just the low angle part of the pattern will be enough; if we focus on structure determination, a bigger range will be needed and the measuring time should be as long as possible, in order to reach a good counting statistics.

1.1.3 Powder pattern indexing

The development of the modern powder diffraction has greatly benefited from the advances achieved in the powder-pattern indexing procedure. One reason for this contribution is the fact that indexing the powder pattern is an initial requirement for the most part of applications.

The aim of pattern indexing is the determination of the crystal symmetry and the unit-cell dimensions, together with the *hkl* indices assignment for each reflection. The only initial information used is the distribution of $d=2\cdot\sin\theta/\lambda$ spacings, which is obtained from the powder diffraction pattern.

This fundamental problem of powder crystallography has been studied with much effort for over more than half a century, being the first important approach to the problem reported by Runge (1917), just after the powder method was introduced.

Diverse approaches for the pattern indexing process have been presented and included in computer programs: an exhaustive successive dichotomy algorithm is used in DICVOL91 (Boultif & Louër, 1991); the Runge-Ito-de Wolff method has been used for the program written by Visser (1969); a trial-and-error method based on the Miller indices permutation for low-angle lines is used in TREOR90 (Werner *et al.*, 1985); more recently, the Monte Carlo and grid search methods have been included in the McMaille program (Le Bail, 2004). The success rate achieved by these algorithms is high, often provided that the quality of the diffraction pattern is good enough; that is, the absolute error on the peak positions extracted from the pattern should be lower than about 0.03º 2θ.

Actually, the only big difficulty that we can find when applying the indexing procedure is the presence of spurious peaks in the diffraction pattern (*i.e.* the presence of another crystalline phase mixed with that one we are interested in).

1.1.4 The reflection overlap problem

As it has been said before, instead of the three-dimensionally well located reflections that result when measuring single-crystal diffraction data, in the case of powder data the principal problem is the fact that, as a consequence of the random orientation of the crystallites that constitute the sample which is being measured, the whole (three dimensional) set of reflections is rotationally projected onto the one-dimensional diffraction pattern that we obtain on a powder diffraction experiment. Then, the only initial information that is shown in the measured pattern is the intensity distribution of the diffraction peaks, as a function of the reciprocal plane distances *d*. Thus, this collapse of the reciprocal space brings on an overlap of the existing diffraction peaks, which derives in a loss of information.

The consequences of this fact can be summarized as follows:

i) the resulting reflection overlap can be either accidentally or exact. The later happens when two or more reciprocal nodes, with different Miller indices between them, are all located at the same distance from the origin of the reciprocal lattice. In this case, the profile that will be obtained in the diffraction pattern will be the summation of the individual intensities associated to each node. Particularly for compounds that show high crystal symmetry, several non-equivalent diffraction peaks can exactly overlap, which implies a direct consequence for the analysis of the structure and/or microstructure: under these circumstances, it is not clear which is the individual contribution of each set of diffraction planes to the whole observed intensity peak. It is clear that the reflection overlap can easily be a very serious problem, which

restricts the complexity of the crystal structure that can successfully be treated.

ii) as the interplanar distance d increases, the total number of reflections that are projected onto the diffraction pattern also increases (with d^3); this fact results on a problematic increase of the reflection overlap degree at high diffraction angles. However, at low angles the density of reflection is lower (the peaks are more separated) and this is an advantage when we are at the stage of indexing the powder diffraction pattern, as we can easily single out the reflection positions.

iii) to the fact of resolution losing, also diffraction line broadening has its own contribution. This broadening comes out from the convolution of two contributions to the reflection profile: the instrumental $g(x)$ profile and the intrinsic $f(x)$ profile. The first one includes instrumental aspects such as spectral radiation dispersion or the configuration in which the measurement has been made (divergence/receiving slits of the diffractometer, for example), while the second is due to specimen defects such as microstrain or small crystallite size. These contributions to the peak broadening become even more problematic to treat when they present anisotropic behaviour.

1.1.5 Space group definition

Once the pattern has been indexed and the unit cell has been determined, the associated space group must be found. This is also a critical step of the structure solving process and is often difficult and ambiguous. It is normally carried out by manually inspection of the systematic absences on the diffraction pattern, which are related with the symmetry operators belonging to the actual space group.

In fact, from this procedure we only obtain the extinction symbol to which the space group of the structure belongs. Although for some cases the extinction symbol univocally indicates the space group, it often corresponds to different space groups, so it is needed some extra information in order to correctly choose the right one.

1.1.6 Modelling of powder diffraction data

In modern powder diffraction, the pattern modelling techniques are a basic tool. There are two different methods currently applied: the pattern decomposition, which does not requires any structural information, and the Rietveld method, for which the knowledge of prior structural information is needed.

In the pattern modelling technique, the whole measured diffraction pattern is fitted, normally by means of a least-squares refinement, with a calculated model. Thus, the calculated intensity $y(x_i)$ at a point x_i is calculated as a function of the integrated intensity $I_{\mathbf{k}}$ of the reflections **k** within the diffraction

pattern and a normalized analytical function Φ is used to simulate the individual peak profiles:

$$y_{cal}(x_i) = \sum_{\mathbf{k}} I_{\mathbf{k}} \Phi(x_i - x_{\mathbf{k}}) + b(x_i) \qquad (1.1)$$

where $b(x_i)$ is the background intensity and the sum runs over those reflections that contribute to the intensity at point x_i, being the intensity proportional to the squared structure factor $|F_{\mathbf{k}}|^2$ (see §1.1.7); actually, all the reflections existing in the pattern contribute to that intensity, but it is considered as a good approximation that the contribution comes only from those reflections lying on a small angular range around the point x_i (in fact, it is the angular range of reflection profile calculation which is limited). Then, the values of the adjustable parameters in the model are estimated using the least squares technique. Normally, the peak shape functions used to model the pattern are derived from the Gaussian (G) and the Lorentzian (L) profiles, *e.g.* the pseudo-Voigt function is a sum of both G and L components, with a mixing factor η (η=0 for G profile and η=1 for L profile); the Pearson VII function is $(L)^m$, with the line-shape parameter m=1 for L peak-shape and m=∞ for the G shape (see Appendix B for a description of these profile functions).

More parameters can be introduced in equation (1.1) in order to take into account some other contributions, like preferred orientation, peak asymmetry, absorption correction, corresponding weight percentage on a phase mixture...

- Pattern decomposition method: the fundamental aim of this modelling technique is to extract, from the whole diffraction pattern, an integrated intensity corresponding to each *hkl* reflection; being the position of each

diffraction peak constrained by the known lattice parameters (which are adjustable in the fitting procedure). Of course, when the space group is known, only allowed reflections by the extinction conditions are used in the method. Then, what we obtain by using this technique is a set of reflection intensities, whatever the degree of overlap between them. Generally, an equipartition principle of the overall integrated intensity is applied for the case of exactly overlapping reflections. This set is then converted to structure factor amplitudes $|F_{obs}|$ and used as input data for structure solution methods.

The most used approaches for the decomposition procedure are the Pawley technique (Pawley, 1981), which is based on a least-squares fitting of the diffraction pattern, and the Le Bail technique, derived from the procedure used by Rietveld for the partition of calculated intensities in the last stage of the refinement (Le Bail *et al.*, 1988). This last method is known as being very stable and robust, while the first one needs some additional treatment to eliminate some instability (Jansen et al. 1992; Sivia & David, 1994).

It is very important to make clear the critical role that plays the decomposition of the diffraction pattern and the precision of the structure factor amplitude in the structure solution step. Unfortunately, the existing overlap between non-equivalent reflections introduces some bias on the set of extracted structure factors. In order to estimate the total amount of reliable information extracted from the powder diffraction pattern, Altomare *et al.* (1995a) and David (1999) have proposed different algorithms, which are based on a systematic study of reflection overlap, to evaluate the total number of statistically independent observations as a function of the angular range.

- The Rietveld method: the principal difference with the pattern decomposition method is that, in the Rietveld approach, the integral intensity $I_{\mathbf{k}}$ of each reflection **k** is calculated using available structure information, which is included in the model. Then, the equation (1.1) becomes

$$y_{cal}(x_i) = S\sum_{\mathbf{k}} m_{\mathbf{k}} (Lp)_{\mathbf{k}} \left|F_{\mathbf{k}}^2\right| P_{\mathbf{k}} \Phi(x_i - x_{\mathbf{k}}) + b(x_i) \qquad (1.2)$$

where S is the scale factor, $m_{\mathbf{k}}$ is the reflection multiplicity, $(Lp)_{\mathbf{k}}$ is the Lorentz-polarization factor and $P_{\mathbf{k}}$ is the preferred-orientation correction function. To create the whole powder diffraction pattern, diverse parameters used to describe the reflection profile (width and shape) are adjusted, together with the structural parameters, until a best fit of the modelled pattern to the observed pattern is obtained.

1.1.7 Methodology to solve unknown structures from powder data

The advances reached on the techniques commented in the previous points (*e.g.* methods to determine the unit cell parameters, to extract the integrated intensities from the pattern and to refine the crystal structure from powder data) has brought the possibility to solve crystal structures from powder diffraction data. Thus, in the recent years, the field of structure solution has become a very active area of research, in constant development. At this time, new complex structures are solved and reported regularly, as the procedures used to treat powder data are well established. The obstacles that can be found in solving structures can be of diverse nature (*e.g.* limited resolution data or complex chemical formula) and frequently the number of atoms present in the unit cell is used as a complexity indicator.

Crystal structure analysis is usually based on diffraction phenomena caused by the interaction of matter with some type of radiation, in general X-rays, electrons, or neutrons. The diffracted intensity is simply related to the squared structure factor $|F_{\mathbf{h}}|^2$, where

$$F_{\mathbf{h}} = \sum_{j=1}^{N} f_j \exp(-2\pi \mathrm{i}\mathbf{h}\cdot\mathbf{r}_j) \tag{1.3}$$

where $\mathbf{h} \equiv h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$ is a vector in the reciprocal space and $\mathbf{r}_j \equiv x_j\mathbf{a} + y_j\mathbf{b} + z_j\mathbf{c}$ the positional vector of the j-th atom in the unit cell.

The crystal structure is defined if the electron density distribution ρ is known. ρ may be written in the form:

$$\rho(\mathbf{r}) = \mathrm{T}[F_{\mathbf{h}}] = \frac{1}{V}\sum_{\mathbf{h}} F_{\mathbf{h}} \exp(-2\pi \mathrm{i}\mathbf{h}\cdot\mathbf{r}) \tag{1.4}$$

where T is the Fourier transform function.

Since the diffraction data provide the magnitudes of the structure factors but not their phases, (1.4) cannot be used to directly obtain the electron density distribution from the experimental data. This is the so called *phase problem*.

The problem must in principle have a solution (even if not necessarily unique), but in this case the unknown quantities (the atomic positions $\mathbf{r}_j$) appear as argument of trigonometric functions:

$$|F_{\mathbf{h}}|^2 = \sum_{j=1}^{N} f_j^2 + 2\sum_{j>k=1}^{N} f_j f_k \cos 2\pi\mathbf{h}\cdot(\mathbf{r}_j - \mathbf{r}_k) \tag{1.5}$$

(1.5) is a system of non-linear equations, and its solutions cannot be obtained in any analytical way, even though the number of relationships greatly exceeds the number of unknowns.

1.1.8 Direct Methods

Methods which try to derive the structure factor phases directly from the observed amplitudes through mathematical relationships are called *Direct Methods*. They are responsible of the great deal of crystal structures solved in the last years, but a considerable effort has must been done to adapt these approaches from the single-crystal to the powder diffraction case, introduce new methodology and develop the corresponding software. These methods follow the strategy stated in §1.1.1 (the chances of success rely on the accuracy of the decomposition step) and are the most generally used for structure analysis when high-quality powder diffraction data is available.

Historically, the first mathematical relationships capable of giving phase information were obtained by Harker and Kasper (1948) in the form of inequalities. In 1953 Hauptman and Karle established the basic concepts and the probabilistic foundations of Direct Methods. Also in 1952 Sayre was able to derive a very important relationship

$$F_{\mathbf{h}} = \vartheta_{\mathbf{h}} \sum_{\mathbf{k}} F_{\mathbf{k}} \, F_{\mathbf{h-k}} \tag{1.6}$$

where $\vartheta_{\mathbf{h}}$ is a known value.

A crucial role in Direct Methods is played by the *structure invariants*. They are linear combinations of phases which are independent of the choice of origin.

Therefore, their value depends only on the crystal structure, and consequently may be estimated (in principle) from the observed diffraction magnitudes.

The most general structure invariants is represented by the product

$$F_{\mathbf{h}_1} F_{\mathbf{h}_2} \ldots F_{\mathbf{h}_n} = \left| F_{\mathbf{h}_1} F_{\mathbf{h}_2} \ldots F_{\mathbf{h}_n} \right| \exp\left[\mathrm{i}(\phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \ldots + \phi_{\mathbf{h}_n}) \right] \qquad (1.7)$$

when

$$\mathbf{h}_1 + \mathbf{h}_2 + \ldots + \mathbf{h}_n = 0 \qquad (1.8)$$

The simplest structure invariant is F_{000}, its phase is always zero. The structure invariant $F_{\mathbf{h}} F_{-\mathbf{h}} = \left| F_{\mathbf{h}} \right|^2$ does not contain any phase information, while the triplet invariant $F_{\mathbf{h}} F_{-\mathbf{k}} F_{-\mathbf{h}+\mathbf{k}}$, with phase $\phi_{\mathbf{h}} - \phi_{\mathbf{k}} - \phi_{\mathbf{h}-\mathbf{k}}$, plays a primary role in the probabilistic procedures for phase determination. With a simple extension it is possible to define the quartet invariant $F_{\mathbf{h}} F_{-\mathbf{k}} F_{-\mathbf{l}} F_{-\mathbf{h}+\mathbf{k}+\mathbf{l}}$ with phase $\phi_{\mathbf{h}} - \phi_{\mathbf{k}} - \phi_{\mathbf{l}} - \phi_{\mathbf{h}-\mathbf{k}-\mathbf{l}}$. In the same form we can define quintets, sextets, etc.

Structure *semi-invariants* are single phases or linear combinations of phases which are invariant with respect to a shift of permissible origin. A basic property of a structure semi-invariant (Giacovazzo, 1980) is its capability of being transformed into a structure invariant by adding one or more pairs of symmetry-equivalent phases. For instance, in a given space group possessing the symmetry operator $\mathbf{C} \equiv (\mathbf{R}, \mathbf{T})$, the phase $\phi_{\mathbf{H}}$ is a semi-invariant if it possible to find a reflection **h** such that

$$\Psi = \phi_{\mathbf{H}} - \phi_{\mathbf{h}} + \phi_{\mathbf{hR}} \qquad (1.9)$$

is an invariant, this is $\mathbf{H} - \mathbf{h} + \mathbf{hR} = 0$.

The normalized structure factors

A central role in Direct Methods is played by the normalized structure factors, defined as

$$\left|E_{\mathbf{h}}\right| = \left|F_{\mathbf{h}}\right| \Big/ \sqrt{\varepsilon_{\mathbf{h}} \sum} \tag{1.10}$$

where $\varepsilon_{\mathbf{h}}$ is the Wilson coefficient depending on the specific indices (*hkl*) (for the tabulation of $\varepsilon_{\mathbf{h}}$ see Giacovazzo, 1980). From (1.10) one can immediately obtain (under the hypothesis that the atomic positions are random variables with uniform distribution throughout the unit cell)

$$< \left|E_{\mathbf{h}}\right|^2 >= 1 \tag{1.11}$$

So far we have implicitly assumed that the observed structure factor moduli are on the absolute scale, but in general the values of $\left|F_{\mathbf{h}}\right|^2_{\text{obs}}$ obtained from the intensities are on a relative scale. In the case that we assume an overall isotropic thermal motion equal for all the atoms, we may write

$$\left|F_{\mathbf{h}}\right|^2_{\text{obs}} = K \left|{}^{o}F_{\mathbf{h}}\right|^2 \exp(-2Bs^2) \tag{1.12}$$

where K is a scale factor, $\left|{}^{o}F_{\mathbf{h}}\right|$ is the structure amplitude in absolute scale for atoms at rest, B is the overall isotropic temperature factor, and $s = \sin\theta / \lambda$.

Wilson (1942) proposed a method to derive the values K and B. From (1.12) one obtains

$$\text{Ln}\left(\frac{< \left|F_{\text{obs}}\right|^2 >}{\sum_s}\right) = \text{Ln}\, K - 2B < s^2 > \tag{1.13}$$

where

$$\Sigma_s = \sum_{j=1}^{N} {}^{o} f_j^2 \tag{1.14}$$

where ${}^{\circ} f_j$ denote the atomic scattering factor for *j*th atom at rest.

If $\mathrm{Ln}\left(\dfrac{< |F_{\mathrm{obs}}|^2 >}{\Sigma_s} \right)$ is plotted against $< s^2 >$ and then the best straight line is derived, the intercept of the line on the vertical axis will give us Ln *K* and its slope the value of 2*B*. This is called the *Wilson plot*.

Probabilistic methods

The basic probability formula for triplet invariants was derived by Cochran (1955):

$$\mathrm{P}(\boldsymbol{\Phi}) = \frac{1}{2\pi\, I_o(G)} \exp(G \cos \boldsymbol{\Phi}) \tag{1.15}$$

where

$$G = 2\left[\sigma_3 / \sigma_2^{-3/2} \right] \left| E_{\mathbf{h}}\, E_{\mathbf{k}}\, E_{\mathbf{h-k}} \right| \tag{1.16}$$

and

$$\sigma_i = \sum_{j=1}^{N} Z_j^i \tag{1.17}$$

where Z_j is the atomic number of the *j*th atom. $\mathrm{P}(\boldsymbol{\Phi})$ is a so-called von Mises distribution and *G* is its concentration parameter.

If more than one pair of phases $\phi_{\mathbf{k}_j}$, $\phi_{\mathbf{h}-\mathbf{k}_j}$, with j=1, 2,..., r, are known, all defining the same $\phi_{\mathbf{h}}$ through triplet relations such us $\phi_{\mathbf{h}} - \phi_{\mathbf{k}} - \phi_{\mathbf{h}-\mathbf{k}}$, then (Karle & Hauptman , 1956)

$$\tan\theta_{\mathbf{h}} = \frac{\sum_{j=1}^{r} G_j \sin(\phi_{\mathbf{k}_j} + \phi_{\mathbf{h}-\mathbf{k}_j})}{\sum_{j=1}^{r} G_j \cos(\phi_{\mathbf{k}_j} + \phi_{\mathbf{h}-\mathbf{k}_j})} \tag{1.18}$$

where

$$G_j = 2\left[\sigma_3 / \sigma_2^{3/2}\right] \left| E_{\mathbf{h}} E_{\mathbf{k}_j} E_{\mathbf{h}-\mathbf{k}_j} \right| \tag{1.19}$$

gives the most probable value of $\phi_{\mathbf{h}}$. The relation (1.18) is known as the *tangent formula* and it plays a central role in the phase determination process.

Solving crystal structures with Direct Methods

The following general scheme for a crystal structure solution may be used:

1. *Normalization*: The values of the normalized structure factors are calculated as we have described before.
2. *Setting up phase relationships*: Triplet and quartet relationships are calculated among the reflections with large |*E*| values in order to estimate those with the highest reliability. The triplet and quartet search must take into account the space group symmetry.
3. *Definition of an optimum starting set of phases*: The choice of the reflections in the starting set is very important because the success of the process will depend on them. In general the starting set is formed by the

reflections which fix the origin (up to three reflections), one reflection if it is necessary to fix the enantiomorph, and a limited number of reflections (usually five or more) which are assigned by different techniques (symbolic addition, magic integers permutation, etc). Given the phases in the starting set, all the other phases can be determined one after the other in a chain process.

4. *Figures of merit*: They are functions which allow an a priori estimate of the goodness of each phase set as representative of the correct solution.

5. *Electron density maps*: The phase set with the highest CFOM (combined figure of merit) is used for obtain the electron density map. The most modern programs find the peaks and supply a list of maxima sorted in decreasing order of height. This list may then be analysed in terms of distances and angles.

6. *Completing and refining the structure*: If the model obtained according to the above procedures is not complete, there are different methods which recovery the complete structure. The most widely used methods of structure refinement are the least-squares Fourier-methods.

1.1.9 Direct-space Methods

An alternative strategy that can be adopted, with the benefit of avoiding the critical step of decomposing the diffraction pattern into a set of integrated intensities by decomposition methods, is to generate several structure models using Direct-space Methods and independently of the powder diffraction data.

These methods have been most successfully developed for molecular compounds, for which the atoms connectivity is known, or for organic compounds, where the resolution of the powder data is often low, and hence we could only work with a small and poorly determined set of structure factor magnitudes. The molecule can be easily described in terms of internal coordinates (bond distances, bond angles and torsion angles), given the connectivity and typical bond lengths and angles. At this point, the variables are some torsion angles and the position and orientation of the whole molecule. Then, a lot (sometimes millions) of possible trial structures are generated by varying those variables and their corresponding diffraction patterns are calculated. The last step is to find the correct structure between all the trials used, which is done by direct comparison of the observed and calculated patterns; so the trial that better fits the observed diffraction data should be the correct one. The power of these methods can be found on the fact that they use prior chemical knowledge but the weakness relies, fundamentally, on the high computing times needed.

A quite large number of algorithms using this global optimisation techniques have been developed: Grid Search (Reck *et al*., 1998; Hammond *et al*., 1997), Monte Carlo (Harris *et al*., 1994, Tremayne *et al*., 1997), Simulated Annealing (Andreev *et al*., 1998; David *et al*., 1998; David & Sivia, 2002), Genetic Algorithm (Shankland *et al*., 1997; Kariuki *et al*., 1997; Harris *et al*., 1998) or Parallel Tempering (Favre-Nicolin & Cerný, 2004).

1.1.10 Hybrid methods

There are also other developed approaches that use both reciprocal and direct space to solve the crystal structure. For example, the program FOCUS uses a method created for the study of zeolite structures, which is based on the generation of a large number of electron-density maps using chemical information from zeolite framework structures in real space and combining with the extracted structure factor magnitudes and random starting phases in reciprocal space (McCusker *et al.* 1996). Another example is the program EXPO which, starting only with the well defined heavy-atoms positions provided by Direct Methods, is able to locate the light-atoms forming the polyhedra around the heavy-atoms by generating several trials that take into account the allowed polyhedral coordination geometry (Altomare *et al.* 2000b).

1.2 Scope of the present Thesis

This thesis has been focused, fundamentally, in various steps that precede the application of Direct Methods (*e.g.* the measurement, the decomposition process and the space group determination). It will be structured as follows:

- In chapter 2, it will be discussed the role that plays the extraction of reflection intensities from the powder diffraction pattern in the *ab initio* structure determination. Thus, the efficiency of the phasing process will be related with various parameters such as the peak-shape function used, the number of cycles carried out in the decomposition or the degree of reflection overlapping.

- Chapter 3 will be focused in an analogous discussion; but this time, the experimental measurement of the powder diffraction pattern will be the process to relate with the phasing procedure. It will be seen how the use of different strategies or time of measurement can directly affect the efficiency of Direct Methods.
- Finally, in chapter 4, a new algorithm for the determination of the extinction group will be developed and tested with several structures; this algorithm will be based on a statistical analysis of the normalized intensities extracted from the powder diffraction pattern.

Chapter 2.

Powder pattern decomposition versus direct phasing

2.1 Introduction

Two main approaches can be followed for solving crystal structures by powder diffraction data: the global optimisation method and the two stage method. As we have seen before, the first one (Direct Space approach) is particularly useful for the solution of organic crystal structures and, generally, in case of structures with known connectivity. The second method (reciprocal space approach) involves the extraction of the integrated intensities *via* pattern decomposition, followed by the application of Direct Methods for phasing reflections and calculating the electron density map. It is the obliged choice when the structural information is minimal.

For this study, we consider the two stage method and, in particular, we focus our attention on the decomposition process, by analysing its role into the

phasing procedure. Two algorithms are widely used to carry out the decomposition process: the Pawley (Pawley, 1981) and Le Bail (Le Bail *et al*., 1988) methods. Owing to problems like peak overlapping, background noise and preferred orientation, the accuracy of the extracted integrated intensities may be low and Direct Methods inefficient. Great efforts have been done for improving the extraction process: the Pawley method has been modified with a Bayesian approach to reduce the intensity correlation and provide positive integrated intensity values (Jansen *et al*., 1992, Sivia & David, 1994), whereas for the Le Bail algorithm the practice of randomly partitioning the intensities of clusters of severely overlapping reflections (Altomare *et al*., 2001, 2003, 2004a) seems more rewarding.

This work aims at clarifying the role of profile fitting, peak-shape function and other critical parameters into the classical Le Bail procedure for full pattern decomposition and the effects of the decomposition in the subsequent phasing process. Our analysis was helped by the use of EXPO2004 (Altomare *et al*., 2004c).

2.2 About the decomposition process criteria

In any decomposition process the observed profile intensity y_i at the i-th step is modelled by summing the contributions from neighbouring reflections to the background *via* the formula (Rietveld, 1969):

$$y_{c,i} = \sum_{\mathbf{k}} I_{\mathbf{k}} \Omega(i,\mathbf{k}) + y_{b,i}$$

where $I_{\mathbf{k}}$ is the integrated intensity of reflection **k**, Ω(i,**k**) is the reflection profile function associated with reflection **k** and calculated at step *i* and $y_{b,i}$ is the background value at step *i*. To estimate behaviour and efficiency of the decomposition process we will use two criteria: the profile factor

$$R_p = 100(\sum_i |y_i - y_{c,i}|) / \sum_i y_i ,$$

where the summations go over the total number of profile points used in the procedure, and the crystallographic factor

$$R_F = 100(\sum_{\mathbf{k}} \left| |F_{\mathbf{k}}^{est}| - |F_{\mathbf{k}}^{true}| \right|) / \sum_{\mathbf{k}} |F_{\mathbf{k}}^{true}| ,$$

where $F_{\mathbf{k}}^{true}$ is the structure factor of reflection ***k*** calculated from the published model of the crystal structure under study, and $F_{\mathbf{k}}^{est}$ is that obtained from the extracted intensities through the formula:

$$F_{\mathbf{k}}^{est} = (I_{\mathbf{k}} / m_{\mathbf{k}} L_{\mathbf{k}})^{1/2} .$$

$m_{\mathbf{k}}$ is the multiplicity of the reflection **k** and $L_{\mathbf{k}}$ is its Lorentz-polarization factor. The first agreement factor assesses the quality of the fit between observed and calculated profiles, given the functions chosen to describe peak and background shapes, whereas the second one indicates the accuracy of the structure factor extraction from the powder pattern and it is more directly related to the efficiency of Direct Methods.

The decomposition process will be analysed by using 31 test structures. The reader can find their code names, chemical information and relative references in Table 2.1.

Table 2.1 - Code name, crystal chemical data and experimental details for the test structures (X: conventional X-ray data; S: synchrotron data; N: neutron data). The numbers on the first column corresponds to the structure reference, which can be found on Appendix A.

Structure Code	Space group	Cell content	Number of reflections	Resolution (Å)
AGPZ (X)[1]	P bca	$Ag_8N_{16}C_{24}H_{24}$	258	1.20
ALDX (X)[2]	P $2_12_12_1$	$C_{88}N_8O_{32}Cl_4H_{100}$	582	1.34
ALPHA (X)[3]	P $2_12_12_1$	$O_{16}C_{20}N_4$	183	1.34
AND1N (N)[4]	P 2_1/n	$C_{28}N_{20}O_8H_{44}$	293	1.46
AND2N (N)[4]	P 2_1/n	$C_{24}N_{24}O_{20}H_{32}$	380	1.35
BACO (N)[5]	C 2/m	$Ba_4C_8O_{20}D_8$	272	0.99
BAMO (X)[6]	P 2_1	$Ba_4Mo_{12}O_{40}$	1220	0.89
CAINE (X)[7]	P $\bar{1}$	$C_{30}N_4O_4Cl_2$	546	1.46
CIME (S)[8]	P 2_1/n	$S_4C_{40}N_{24}H_{64}$	924	1.13
COPPER (X)[9]	P $\bar{1}$	$Cu_2O_8C_{60}N_4H_{48}$	908	1.42
CROX (X)[10]	P $\bar{1}$	Cr_8O_{21}	657	1.08
DADA (S)[11]	P $2_12_12_1$	$Ti_8K_4Si_{12}O_{40}$	518	1.05
DFQP (X)[12]	I 2/a	$F_{40}C_{96}$	348	1.39
EMT (S)[13]	P 6_3/mmc	$Si_{96}O_{192}Na_8$	670	1.33
GAPO (S)[14]	P bca	$Ga_{32}P_{32}O_{128}F_8C_{56}$	1235	1.17
IBUPS (S)[15]	P 2_1/c	$O_8C_{52}H_{72}$	380	1.50
LAMO (X)[16]	P 2_1/a	$LaMo_5O_8$	271	1.36
LASI (N)[17]	P 2_1/c	$La_8Si_8O_{28}$	253	1.38
LEV (S)[18]	R $\bar{3}$m	$Si_{54}O_{108}C_{48}N_6$	323	1.13
MCM (S)[19]	P 6/mmm	$Si_{72}O_{144}$	480	1.30

Table 2.1 (Continuation).

Structure Code	Space group	Cell content	Number of reflections	Resolution (Å)
NBPO (S)[20]	C 2/c	$Nb_{20}O_{120}P_{28}$	1201	1.0
NIZR (S)[21]	P 2_1/n	$Ni_4Zr_8P_4O_{16}$	628	1.17
OTHYM (X)[22]	P 2_1/n	$O_{12}C_{44}$	453	1.34
SAPO-40 (X)[23]	P mmn	$Si_{32}O_{64}N_2C_{48}$	717	1.20
SBPO (S)[24]	P 2_1/n	$Sb_8P_{14}O_{48}$	1071	0.95
SGT (S)[25]	I 4_1/amd	$Si_{64}O_{128}C_{104}$	454	1.07
SULPH (N)[26]	P bcm	S_8D_{16}	220	0.99
UTM1 (S)[27]	C 2/m	$Si_{44}O_{88}$	1133	1.01
VFI (S)[28]	P 6_3	$(H_2O)[Al_{18}P_{18}O_{72}]$	787	1.07
YONO (S)[29]	P 2_1	$Y_8O_{26}N_2H_{18}$	680	0.96
YURI (X)[30]	P 2_1/c	$Na_4S_4O_{16}C_{12}F_{12}$	243	1.45

2.2.1 The role of the peak-shape function

The peak-shape function Ω is determined by both instrumental (wavelength distribution in the primary beam, beam divergence, detector system, etc.) and sample effects (finite crystallite size, transparency, imperfections and inhomogeneities of the crystalline material, etc.). EXPO2004 provides some choices of the analytical peak-shape function (see Appendix B for a description of the most used functions): their parameters are refined during the extraction process to bring the calculated profile as close as possible to the experimental one. We used in our tests two functions: the Pearson VII (PVII) (Brown & Edmonds, 1980) and the pseudo-Voigt (PV) defined by Young and Wiles (1982). We applied EXPO2004 to all the test structures: the experimental

patterns were decomposed in single diffraction intensities by using PVII and PV as profile functions (the background has been treated in the same way for all the two cases). The R_p and R_F values obtained at the end of the decomposition process are reported in Figure 2.1. It can be seen that:

a) the R_p values corresponding to the PV are systematically higher than those obtained by the PVII function (the averaged R_p values are $<R_p$-PVII$>$=5.2 and $<R_p$-PV$>$=7.2);

b) the average R_F values are $<R_F$-PVII$>$=43.1 and $<R_F$-PV$>$=44.0: the discrepancy between the two profile functions is negligible and for most structures the two functions lead to comparable results;

c) it frequently happens that lower R_p values do not correspond to lower R_F values.

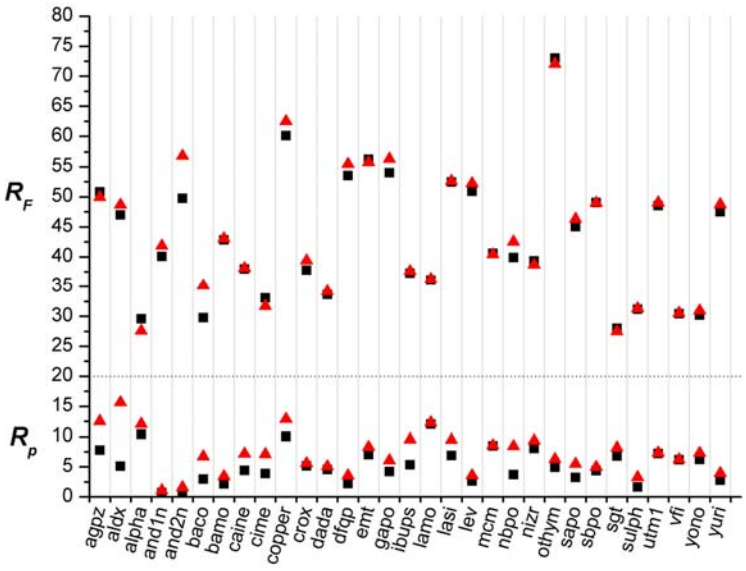


Figure 2.1 EXPO2004 decomposition process: R_F values (top part) and R_p values (bottom part) obtained by using the Pearson VII (squares) and the pseudo-Voigt (triangles) peak-shape functions.

It may be concluded that the choice of the peak-shape function has a limited influence on the accuracy of the intensities extracted from a powder pattern, though it may influence the agreement between the experimental and calculated profiles (very low R_p values do not necessarily correspond to good estimates of the |F|'s). As a rule of thumb, the value of R_p mostly depends on the fit between the chosen profile function and the experimental one, while the R_F value also depends on the peak overlapping.

As the choice of the peak-shape function is not crucial (in the section 2.2.3 we will see that R_p depends also on the number of refinement cycles) we will use, for the next calculations, only the Pearson VII function (the traditional default profile of EXPO2004).

2.2.2 About the profile fitting

The test powder patterns were decomposed by EXPO2004 by using the peak-shape function PVII. The background has been described by a polynomial function whose order was automatically chosen by the program. The values of R_p and their corresponding R_F values for each test structures are shown in Fig. 2.2.

We calculated, for the 31 test structures, the correlation factor between R_p and R_F and we obtained the value of -0.04206 which indicates that R_p and R_F are loosely correlated.

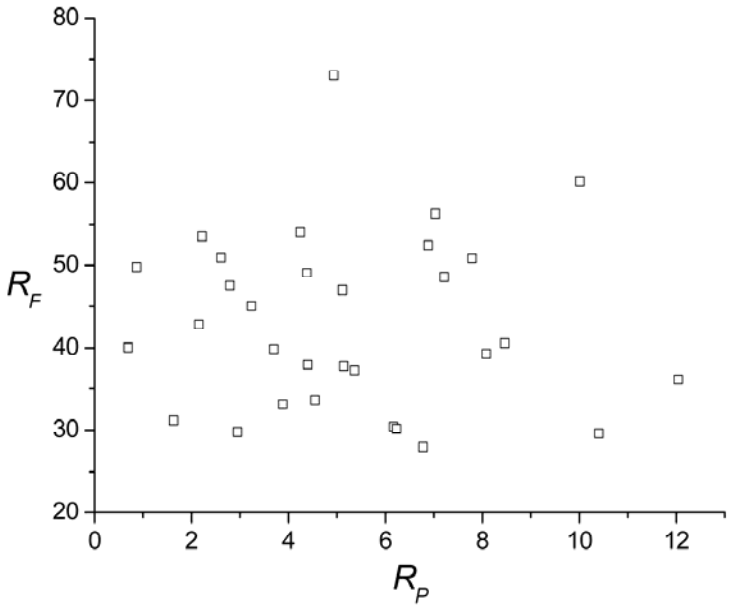


Figure 2.2 R_F and R_p values obtained by the EXPO2004 decomposition process using the Pearson VII profile function for the 31 test structures.

2.2.3 About the number of profile refinement cycles

In general, the Le Bail algorithm converges rapidly to better intensities (typically within 10 refinement cycles). To verify what happens if the user increases the number of refinement cycles we increased them up to 100 and we reprocessed all test structures. The results may be so described:

a) for some structures both R_p and R_F decrease. This is the case shown in Fig. 2.3a for CIME;

b) for some structures R_p is stable or decreases and R_F increases. This is the case shown in Fig. 2.3b for MCM;

c) for other structures R_p and R_F are stable during the refinement.

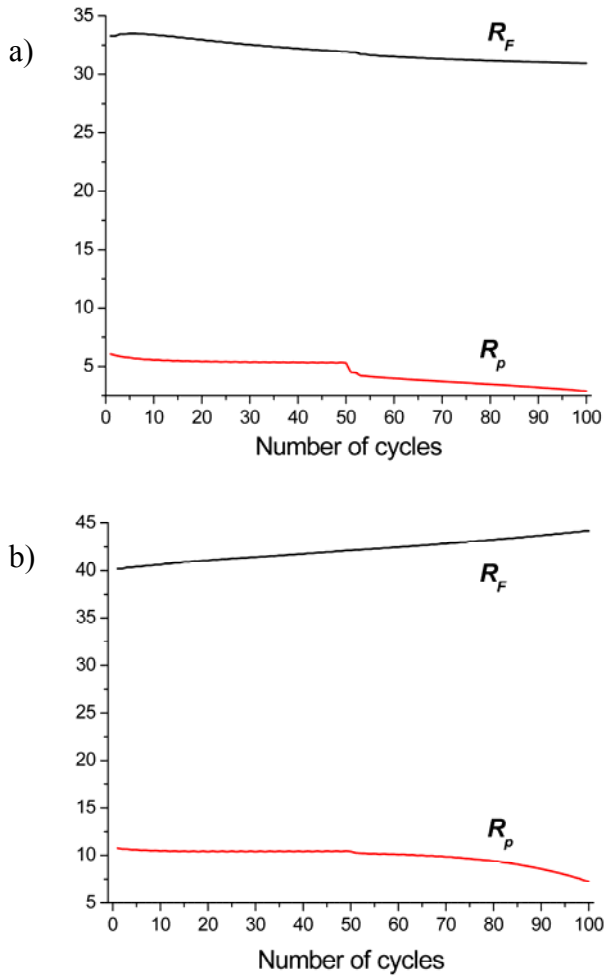


Figure 2.3 R_F values (top part) and R_p values (bottom part) obtained by EXPO2004 versus the number of refinement cycles for a) CIME and b) MCM.

The average (for all the test structures) values of R_p and R_F after 100 profile refinement cycles are: $\langle R_p \rangle$=4.2, $\langle R_F \rangle$=43.4 respectively. If compared with the corresponding averages reported in §2.2.1, obtained after at most 10

refinement cycles, they suggest that a) R_p usually decreases by increasing the number of cycles, while R_F remains practically constant; b) the convergence rate of a typical decomposition process is program dependent, and critically depends on the presence and on the values of the dumping parameters used in the least squares procedures; c) often it isn't worth trying to improve the decomposition process by increasing the number of cycles: this can introduce divergences and in general this does not lead to better integrated intensities; d) there is not a global minimum for the fitting process: as discussed in the section 2.4, this is mostly due to the limited information contained in a powder pattern.

2.2.4 About the overlapping

In order to understand how the experimental peak overlap invalidates the decomposition process we recall the algorithm proposed by Altomare *et al*. (1995a) (see also David, 1999), which is able to estimate the amount of information contained in an experimental powder pattern in terms of independent reflection intensities. *E.g*., two completely overlapping reflections are equivalent to a single observation; two partially overlapping reflections are statistically equivalent to a number of independent reflections lying between 1 and 2. Let NREF be the number of reflections in the observed 2θ range. The algorithm is able to establish that the information provided by the full experimental pattern containing NREF overlapping reflections is statistically equivalent to that provided by $NREF_{eq}$ non overlapping reflections. Let PERC = $NREF_{eq}$ /NREF: in Fig. 2.4 we show, for each test structure, the value of PERC versus the R_F value. The calculated correlation coefficient between R_F

and PERC is -0.69: such a value suggests that PERC may be used as a statistical criterion to assess the expected efficiency of the decomposition process: the smaller PERC, the lower the efficiency of the decomposition. It may also be useful to note that the quality of the experimental data largely varies with the structure.

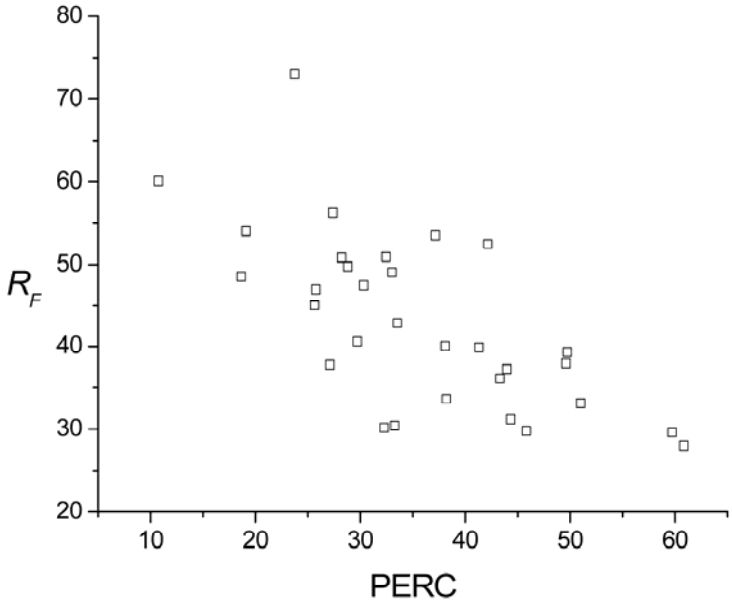


Figure 2.4 R_F versus PERC for the 31 test structures.

2.3 About the phasing process

We focus our attention on how the application of Direct Methods in EXPO2004 is influenced by the previous steps of the phasing procedure. E. g.,

we will deal with the unit cell determination and with the definition of the space group.

About the unit cell determination. The modulus of EXPO2004 devoted to the pattern indexing is constituted by the program N-TREOR (Altomare *et al*., 2000a). The peak search procedure is performed by combining the peak profile fit (Huang & Parrish, 1975) and the derivative approach (Savitzky & Golay, 1964; Sonneveld & Visser, 1975; Huang,1988). The profile fitting includes the decomposition of overlapping peaks. We have verified if the use of different profile functions (*e.g*. the Pearson VII and the Pseudo-Voigt) affects the efficiency of the process: no fundamental differences were obtained, only small changes in the values of the M_{20} figure of merit (de Wolff, 1968).

About the definition of the space group. The space group determination in EXPO2004 is made via a probabilistic approach which is able to associate a probability value to each extinction symbol compatible with the previously established lattice symmetry (Altomare *et al*., 2004b; for a different approach see also Markvardsen *et al*., 2001). The identification of the extinction group depends on the statistics of the weak reflection intensities. We have checked the relative efficiency of the profile functions PVII and PV: we have not found any relevant difference. On the contrary, a relevant role may be played by the presence of impurity peaks: they may overlap with reflections which are expected to be systematically absent, so modifying the correct intensity statistics.

The Direct Methods applications. We applied the EXPO2004 Direct Methods section in its default configuration to the integrated intensities produced by the standard Le Bail decomposition process. The structure factor moduli extracted from the diffraction pattern were scaled by the Wilson method to obtain normalized amplitudes, the triplet invariants were estimated and then the

tangent formula was applied by using a multisolution approach. For each structure 100 phasing trials were generated: the best solution (practically that with the lowest average phase error) was singled out. The corresponding phase errors (say $< \Delta\phi_{100} >$) are reported in Table 2.2. The table suggests that a remarkable percentage of test structures cannot be solved in default mode owing to the large average phase errors produced by the tangent procedure.

Table 2.2 – Average phase error $< \Delta\phi_{100} >$ of the best solution found by processing with Direct Methods the set of integrated intensities extracted by EXPO2004.

Structure Code	$< \Delta\phi_{100} >$	**Structure Code**	$< \Delta\phi_{100} >$
AGPZ	0.00	LAMO	9.47
ALDX	54.35	LASI	22.98
ALPHA	44.11	LEV	23.29
AND1N	22.50	MCM	45.88
AND2N	38.68	NBPO	0.86
BACO	2.05	NIZR	9.77
BAMO	11.39	OTHYM	63.45
CAINE	65.96	SAPO-40	47.44
CIME	5.19	SBPO	0.00
COPPER	80.41	SGT	2.90
CROX	8.00	SULPH	0.00
DADA	28.61	UTM1	16.81
DFQP	35.04	VFI	46.01
EMT	41.92	YONO	10.00
GAPO	36.48	YURI	16.67
IBUPS	30.68		

This unsatisfactory result may be explained as follows. The efficiency of Direct Methods relies on the reliability of the triplet invariants, which depends on the Cochran (1955) parameter

$$G = \frac{2\,|E_{\mathbf{h}} E_{\mathbf{k}} E_{-\mathbf{h}-\mathbf{k}}|}{\sqrt{N_{eq}}},$$

where E is the normalised structure factor, $N_{eq} = \sigma_2^3 / \sigma_3^2$ and $\sigma_p = \sum_{j=1}^{N} Z_j^p$. N is the number of atoms in the (primitive) unit cell, Z is the atomic number. When some heavy atoms are present, $N_{eq} < N$: in this case the crystal structure solution is easier. For single crystal diffraction data (small molecules) the efficiency of Direct Methods is usually related to $NAT_{eq} = N_{eq} / m$, where m is the number of symmetry operators of the point group of the crystal structure. If powder diffraction data are available, at least two supplementary parameters should be considered to predict how much efficient Direct Methods are: the value of PERC and data resolution (RES). This last parameter is suggested by the fact that powder data resolution is usually lower than for single crystal: *e.g.*, for organic molecules it is not infrequent that the powder pattern is practically flat for RES < 1.3Å. In Fig. 2.5 we plot $\sqrt{NAT}_{eq}$(DRES/PERC) (PERC is divided by 100) *versus* $< \Delta\phi_{100} >$: $DRES = RES_p / RES_s$ is the ratio between the experimental resolution for the powder under study and the typical resolution for single crystal data (*e.g.*, atomic resolution). We choose RES_s=1.0 Å and we obtained a correlation coefficient of 0.75 for the pair $[\sqrt{NAT}_{eq}(\text{DRES/PERC}), < \Delta\phi_{100} >]$. In conclusion, when powder data are

used, Direct Methods *see* a structure more complex than the real one, with $NAT_{eq}(\mathrm{DRES/PERC})^2$ atoms in the asymmetric unit. In some cases the simulated structural complexity corresponds to a small protein, owing to the large value of $(\mathrm{DRES/PERC})^2$. The reader can usefully compare some of the results quoted in Table 2.2 with the average phase errors obtained by the tangent procedure for some small proteins (Burla *et al.*, 1999).

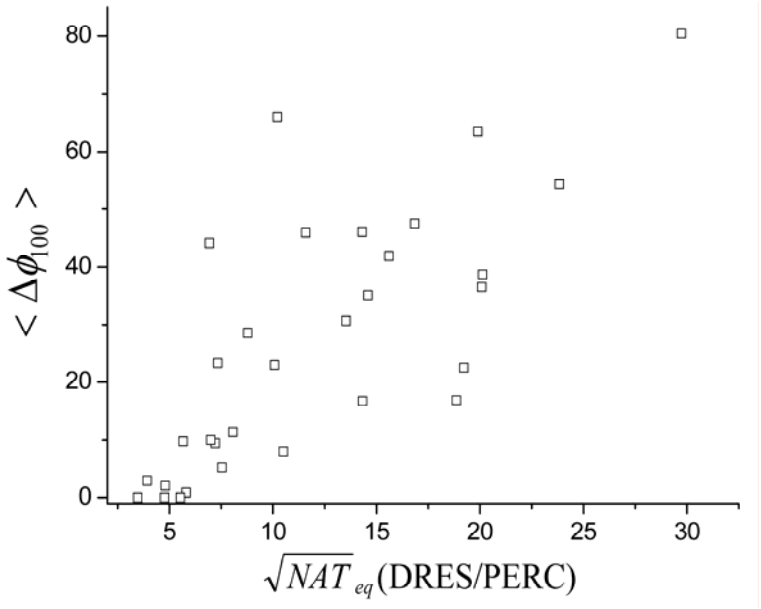


Figure 2.5 $< \Delta\phi_{100} >$ versus $\sqrt{NAT}_{eq}(\mathrm{DRES/PERC})$ for the 31 test structures.

2.4 Conclusions

The role of the full pattern decomposition process in the powder *ab-initio* solution has been investigated by using the extraction algorithms implemented in EXPO2004 package and by studying how the choice of the profile function, the accuracy of the profile refinement, the degree of reflection overlapping and the determination of the peak-shape parameters influences the accuracy of the extracted intensities and the subsequent Direct Methods procedure. The results suggest that:

a) systematic attempts to improve the profile fitting do not necessarily lead to more accurate extracted intensities;

b) a suitable choice of the peak-shape function may improve the profile fitting but not necessarily the decomposition efficiency;

The rationale of the above conclusions is the following. In the Rietveld refinement step, the fit between calculated and experimental profile is a strong criterion owing to the fact that a strong restraint is available: the knowledge of a structural model. Such prior information does make the correct choice of the profile function be the necessary key for the success of the structural refinement. When no structural information is available, as in the case here described, the full pattern decomposition may converge to structure factor moduli far away from the true ones: this effect is not due to inadequacy of the decomposition methods, but to the insufficient information contained in the powder pattern. This lack of information is the strongest obstacle to the improvement of the decomposition process and identifies in the development of the experimental techniques the most useful tool for increasing the quality of the extracted intensities.

On the other side, our tests on direct phasing suggest that amount of information provided by the experimental powder pattern depends on several factors like: the nature of the radiation (X-ray, neutron), the quality of the experiment (*e.g.*, synchrotron data are usually more informative), the structure complexity, the chemical composition of the compound, etc. Some loss of information is unavoidable in powder diffraction. Therefore, in difficult cases, the powder pattern decomposition has to be improved by special techniques. An example is that recently proposed by Altomare *et al*. (2001, 2003, 2004a), based on a Monte Carlo approach: several sets of extracted intensities, called decomposition trials, are submitted one after the other to Direct Methods. Each decomposition trial is obtained by partitioning in a specific way the overall intensity of one or more clusters of overlapping reflections under the constraint that the intensity of each cluster is constant. When the correct intensity partitions are approximated the gain of information increases the value of PERC (and therefore lowers the value $NAT_{eq}(\mathrm{DRES/PERC})^2$). To give an example, two perfectly overlapping reflections correspond to a single observation in absence of any Monte Carlo application. If, among the different partitions of the doublet one is nearly correct, we have two correct intensities available (corresponding to two "observations"). Of course, the application of such methods requires some brute computer force, but it improves the efficiency of Direct Methods.

Chapter 3.

About the variable counting time techniques in powder diffraction data

3.1 Introduction

High quality data are essential for crystal structure solution, particularly when Direct Methods instead than Global Optimization Techniques are used; the variable-counting-time technique (VCT) was introduced by Madsen & Hill (1994) in Rietveld analysis of X-ray powder diffraction data. Basically, the counting time per each measuring step increases as a function of the scattering angle 2θ, in order to counteract the average peak intensity fall-off. This arises from the natural scattering-factor decrease, the thermal vibration and the Lorentz-polarisation factor. In their work, Madsen & Hill compared the VCT measurement strategy with the fixed-counting-time (FCT) scheme, showing that the first one brings on some benefits in the Rietveld analysis: *e.g.*, better

goodness-of-fit and profile-factors, more accurate determination of the atomic positions and more reliable quantitative phase analysis.

Shankland *et al*. (1997) first applied the VCT strategy in order to improve the quality of the experimental data for the structure determination process. By now, there is yet a small number of structures which have been solved using VCT: we found only one paper in which Direct Methods and VCT techniques were jointly applied (Shankland *et al*., 1997). In all the other cases VCT techniques were applied in combination with Direct-Space Methods.

The present work aims at completing the study of the potentiality of the VCT protocol, with special attention to its influence on the full pattern decomposition and on the direct phasing process. This analysis is carried out by using 6 test structures: four of them (i.e., AMMO, BA-TART, MN-TART and LOPEZITE) contain some heavy atoms, two (FR3S and ASPAR) are light atom structures. Their code names, chemical information and corresponding references are summarized in Table 3.1, where RES (the experimental data resolution) is also given.

Table 3.1 Code names and crystal chemical data for the test structures. NAT is the number of non-hydrogen atoms in the asymmetric unit; RES is the experimental data resolution (in Å). The numbers on the first column corresponds to the structure reference, which can be found on Appendix A.

Structure Code	Space Group	Cell Content	NAT	RES
AMMO [31]	P 2_1/c	$Mo_{28}N_{24}O_{112}H_{128}$	41	1.02
FR3S [32]	P $2_12_12_1$	$C_{36}O_{32}H_{72}$	17	0.86
ASPAR [33]	P 2_1	$C_8N_2O_8H_{14}$	9	0.98
BA-TART [34]	P $2_12_12_1$	$Ba_4C_{16}O_{24}H_{16}$	11	0.86
MN-TART [35]	P 2_1	$Mn_4C_{16}O_{32}H_8$	26	0.86
LOPEZITE [36]	P -1	$Cr_8K_8O_{28}$	22	0.98

3.2 Experimental

3.2.1 Instrumental conditions

The diffractometer used for our study was a Panalytical X'Pert Pro. It was used in Debye-Scherrer geometry with a hybrid monochromator for parallel-beam geometry. Although this geometry is suitable for weakly aborting specimens we were forced to use it due to the small quantity of compound available for each test structure. All the samples were introduced in a 0.3mm diameter capillary. The detector used was an X'Celerator (a fast X-ray detector based on a real-time multiple-strip detection technology), working in continuous step mode with 2.149º active length. In all cases, Soller slits of 0.02º were placed into the incident and diffracted beam to reduce the asymmetry of peaks due the axial divergence.

3.2.2 Data Collection

In accordance with the algorithm developed and coded in the *XRDTIME* program by Madsen & Hill (1994), for each test structure we applied the VCT strategy by using blocks of 5º in 2θ. Since the official software controlling the X'Pert Pro Diffractometer works only with measurement programs using a fixed counting time, we wrote a measurement program per each block, by fixing the corresponding counting time (let it be T_i for the i-th block). Accordingly, the full pattern arises by merging several diffrattograms (a diffractogram per each block). An example of a typical time/step curve is shown in Fig. 3.1.

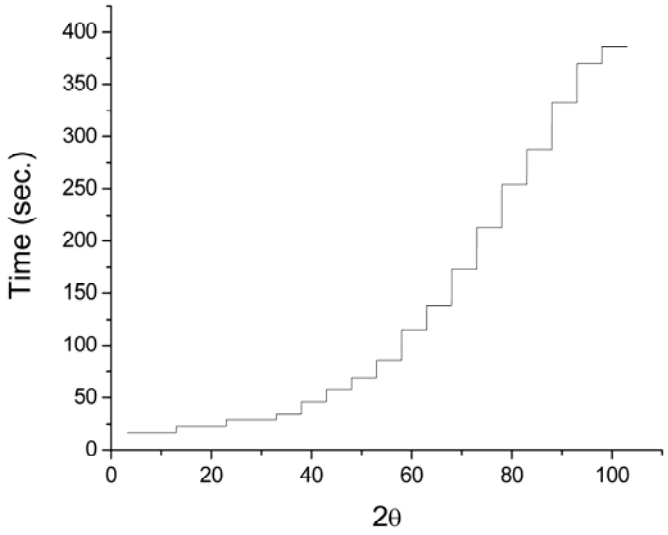


Figure 3.1 A typical time/ step curve

Let us denote by the word *cycle* a full pattern measuring time of about two hours, no matter if the FCT or the VCT strategy is used. The two strategies were applied by using the same 2θ range and the same number of cycles. The different experimental patterns (in VCT or FCT mode) used for our tests were obtained by a simple and progressive summing of the individual cycles (the maximum number of summed cycles was varied structure by structure). This allowed us to compare the effects of the two counting time schemes on the full pattern decomposition and on the structure solution process, as a function of the total measuring time.

3.3 Structure determination

3.3.1 Intensity extraction

The crystal structure solution of the test structures was attempted by applying the two-step approach of EXPO2004 (Altomare *et al*., 2004c).

The experimental FCT pattern was (in default mode) decomposed *via* the Le Bail method (Le Bail *et al*., 1988): the Pearson-VII function was used as the peak-shape function, and a polynomial was introduced for describing the background, whose order was automatically chosen by the program. The extracted integrated intensities were then processed by Direct Methods.

The experimental VCT pattern is usually affected by a background difficult to be modelled by a polynomial. This is due to the "stair-steps" counting time scheme. In Figure 3.2 we show the case of MN-TART corresponding to 18 cycles: it can be noticed the dependence of the background on the counting time scheme.

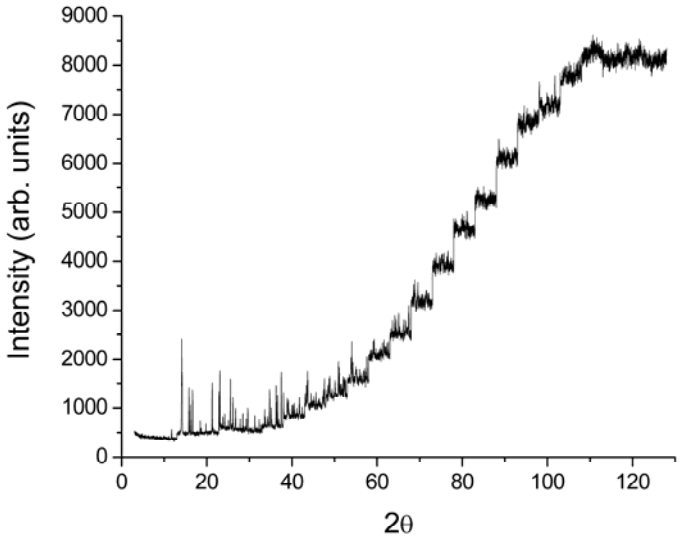


Figure 3.2 MN-TART: experimental pattern (counting intensities versus 2θ) corresponding to 18 cycles

For the VCT case, two different procedures were devised:

a) the intensities of each 2θ block were immediately *normalized* with respect to the corresponding counting time. The combination of the various normalised blocks led to a new full pattern diagram which was then submitted to the same intensity extraction and Direct Methods procedures as the FCT diagrams. This approach was applied to all the test structures and provided to be less useful than the approach described in b). We will not mention it anymore in this work.

b) the intensity extraction was carried out per block, by using directly the experimental VCT pattern: the normalisation with respect to the block counting time was then applied to the F^2's. We have implemented in EXPO2004 an automatic normalizing procedure, based on the assumption that

the intensity jumps present in the experimental diffractogram are directly related with the measuring-time jumps. It consists of the following steps:

i) the minimum experimental intensity (say Y_{min}) in the block with the highest measuring time is singled out. Let *m* be the order number of this block and T_m its measuring time.

ii) the whole time/step curve is multiplied by the constant C defined by the relation $CT_m = Y_{min}$: then the counting time value of the block corresponding to the highest measuring time is made equal to Y_{min} (C has the dimension of counts/ unit of time).

iii) in any i^{th} block the quantity $C \cdot T_i$ is subtracted from any observed intensity (i.e., $Y_{obsnew} = Y_{obs} - C \cdot T_i$). The whole modified pattern is eventually shifted up in order to make positive all the intensity values.

iv) the intensity extraction procedure is applied and the intensities are time–normalised.

In Fig. 3.3a, for ASPAR, we show the observed diffractogram (black line) and the time/step curve (red line) scaled as described at the point ii). In Fig. 3.3b we plot the pattern modified in accordance with the point iii).

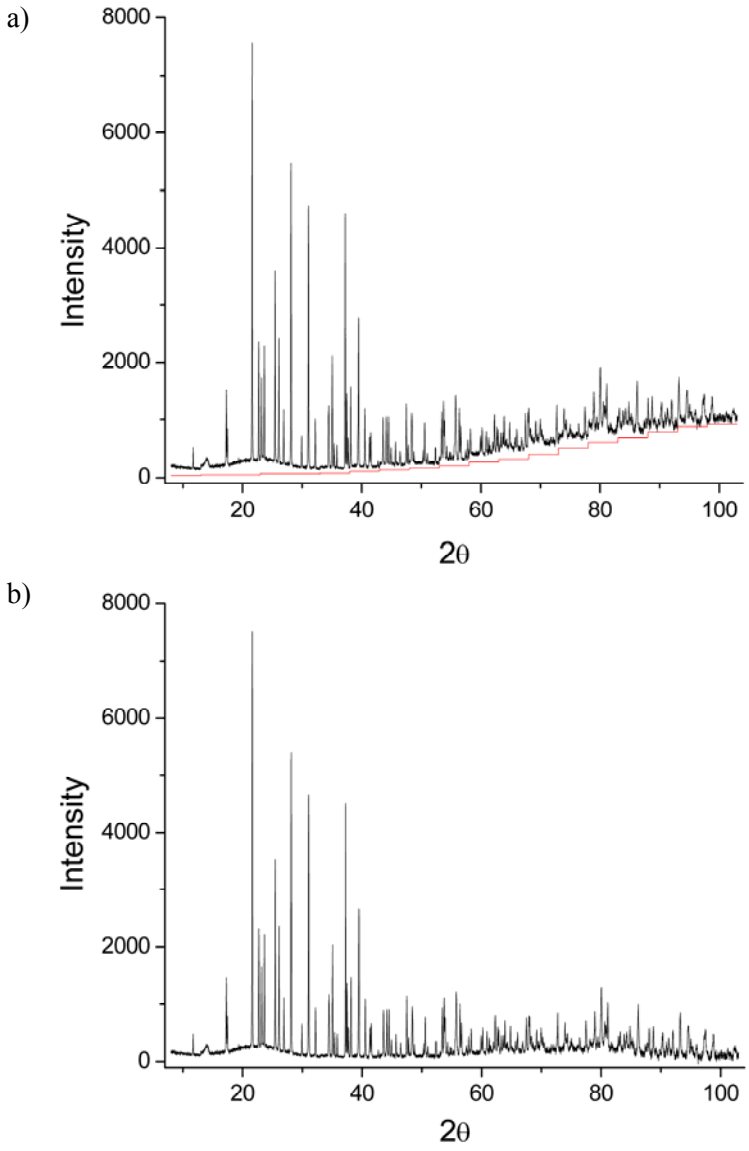


Figure 3.3 ASPAR: a) the experimental pattern (black line) and the scaled time/step curve (red line); b) the pattern available at the end of the step iii).

In order to assess the relative efficiency of the intensity extraction process when applied to FCT and VCT data, the crystallographic residual

$$R_F = 100\left(\sum_{\mathbf{k}}\left|\left|F_{\mathbf{k}}^{est}\right| - \left|F_{\mathbf{k}}^{true}\right|\right|\right)\Big/\sum_{\mathbf{k}}\left|F_{\mathbf{k}}^{true}\right|$$

may be used, where $\left|F_{\mathbf{k}}^{\text{true}}\right|$ represents the structure factor modulus of reflection **k** calculated from the published structural model, and $\left|F_{\mathbf{k}}^{\text{est}}\right|$ is the structure factor modulus obtained from the pattern decomposition through the formula:

$$\left|F_{\mathrm{k}}^{est}\right| = (I_{\mathrm{k}} / m_{\mathrm{k}} L_{\mathrm{k}})^{1/2}.$$

$I_{\mathbf{k}}$ and $m_{\mathbf{k}}$ are the extracted integrated intensity and the multiplicity of the reflection **k**, respectively, and $L_{\mathbf{k}}$ is its Lorentz-polarization factor. Since the phases of the structure factors are *hidden* in the moduli, smaller R_F values suggest more straightforward crystal structure solutions (Altomare *et al.*, 2005), Fig. 3.4 shows, for each test structure, R_{F} *versus* the number of cycles for FCT and VCT data. It can be seen that:

a) the R_F values decrease (on the average) with increasing values of the cycle number for both FCT and VCT data. Indeed, when the number of cycles is low, the bad counting statistics can provoke unreliable intensity extractions.

b) the decreasing rate of R_F tends to vanish at large measuring times. Indeed the better counting statistics cannot overcome the lack of information fixed by the peak overlapping.

c) the R_F factor is systematically lower for the VCT case (the improvement is not very meaningful for AMMO).

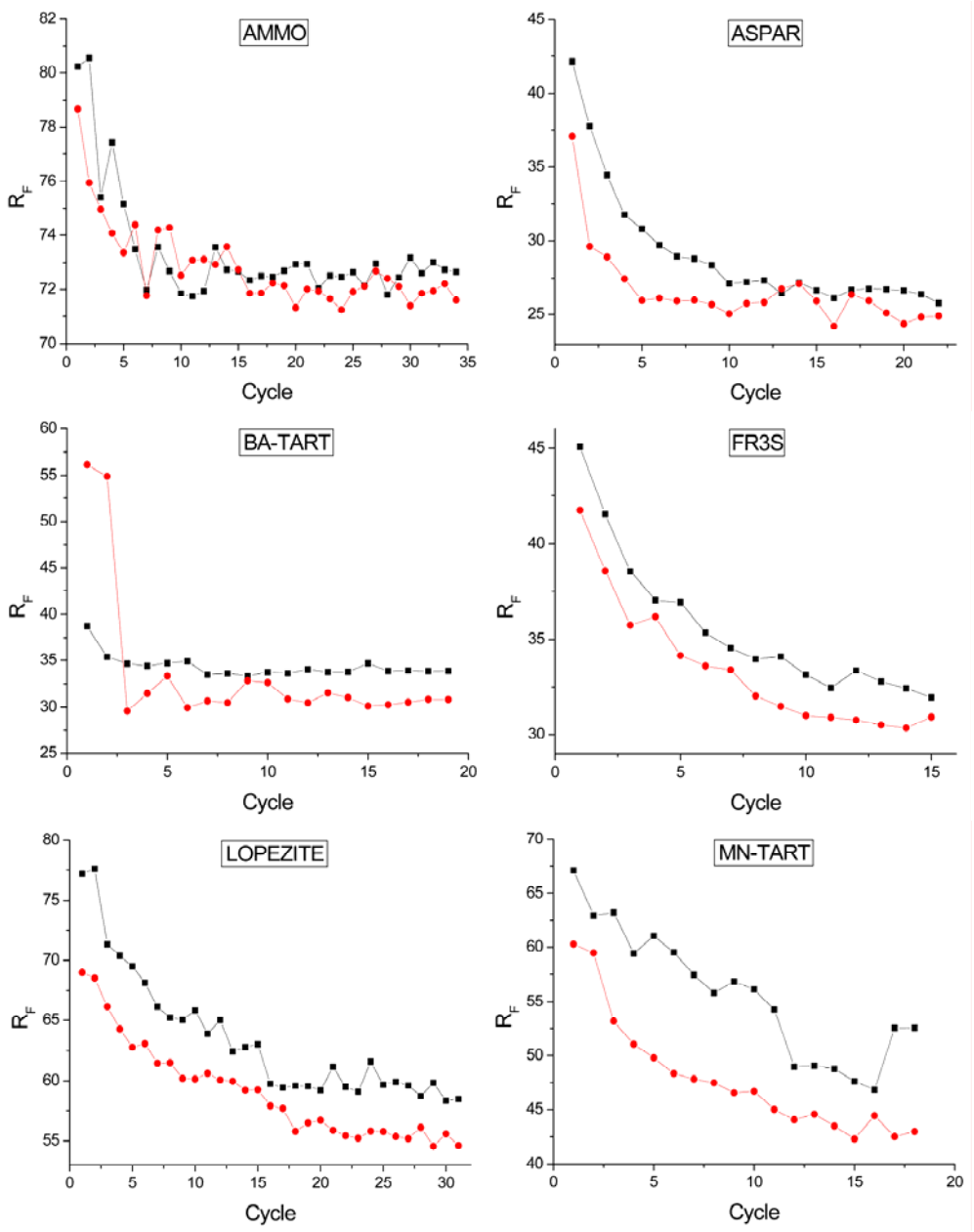


Figure 3.4 R_F versus the cycle number for all the test structures, for FCT (black squares) and VCT (red circles) data.

It may be concluded that, as expected, the time spent for the data collection contributes to define the quality of the powder pattern, no matter whether FCT or VCT scheme is chosen. The decomposition process however is in general more efficient when the VCT scheme is applied: indeed it is able to improve the counting statistics at higher angles, just where a bigger density of reflections is found. The number of cycles behind which the quality of the decomposition process is stationary is structure dependent (and, possibly, instrumentation dependent).

The R_F criterion is a quite useful tool for evaluating, *post-mortem*, the efficiency of the decomposition process (that is, when the structure has been solved and refined). An *ante mortem* criterion was suggested by Altomare *et al.* (1995a) [see also David, 1999.]: it estimates the amount of information contained in an experimental powder pattern in terms of independent reflection intensities. The algorithm establishes that the information provided in a full experimental pattern containing NREF overlapping reflections is statistically equivalent to that provided by $NREF_{eq}$ non overlapping reflections. The factor PERC (PERC= $NREF_{eq}$/NREF) was defined as a statistical criterion related to the efficiency of the decomposition process: a small value of PERC suggests low accuracy in the intensity extraction and consequently a more difficult solution process. In Figure 3.5 we show, for MN-TART (the same behaviour is obtained from the other test structures), the value of PERC calculated by EXPO2004 *versus* the number of cycles, using VCT and FCT strategies. At low cycle numbers PERC is smaller for VCT data: this is mainly due to the fact that it is dramatically worse at low angle values, owing to the unsatisfactory counting statistics. At large cycle numbers PERC is larger for VCT data: this is due to the fact the VCT strategy substantially improves the counting statistics at high angle values, where the largest part of the reflections

lie. The above results suggest that the VCT strategy can exploit a larger amount of information during the intensity extraction step.

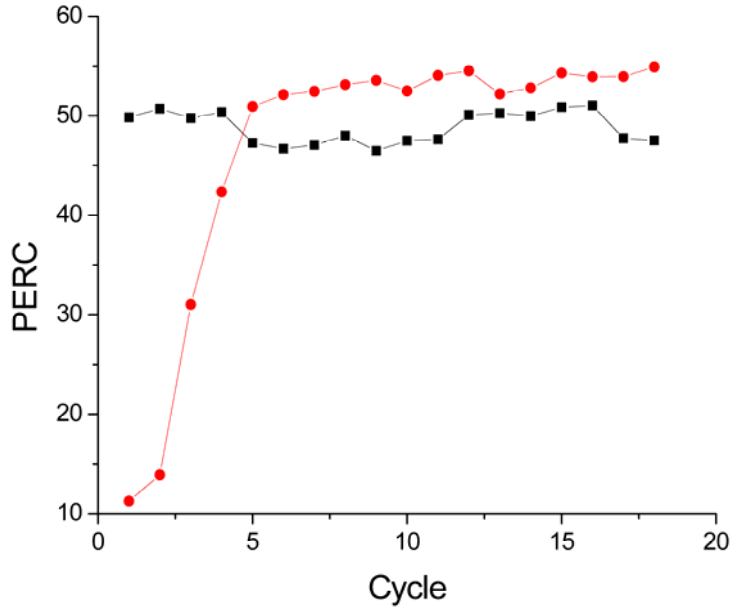


Figure 3.5 - MN-TART: the value of PERC calculated by EXPO2004 using VCT (red circles) and FCT (black squares) strategies, *versus* the number of cycles.

It may be useful to characterize the experimental FCT diagrams with respect to the VCT normalised diagrams obtained at the end of the step iv) of our automatic procedure. In Figure 3.6 we show, for each test structure, the correlation curves (CORR) of the two diagrams calculated at the 20th cycle. Each point of the curves measures the correlation in the corresponding 2θ block. We observe: a) for two structures (AMMO and BA-TART) CORR is close to unity in all the blocks. That suggests that the FCT strategy provides sufficiently good counting statistics even at high 2θ values, and that no

substantial gain of information is obtained in VCT mode; b) for two structures (ASPAR and FR3S) CORR substantially differs from the unity only in the high resolution blocks, where the VCT strategy may give rise to a better count statistics; c) for LOPEZITE and MN-TART CORR significantly differs from the unity in most of the blocks, so suggesting better count statistics for the VCT strategy. However CORR is too small for the high resolution blocks of MN-TART for which some additional investigation is needed (see §3.3.3).

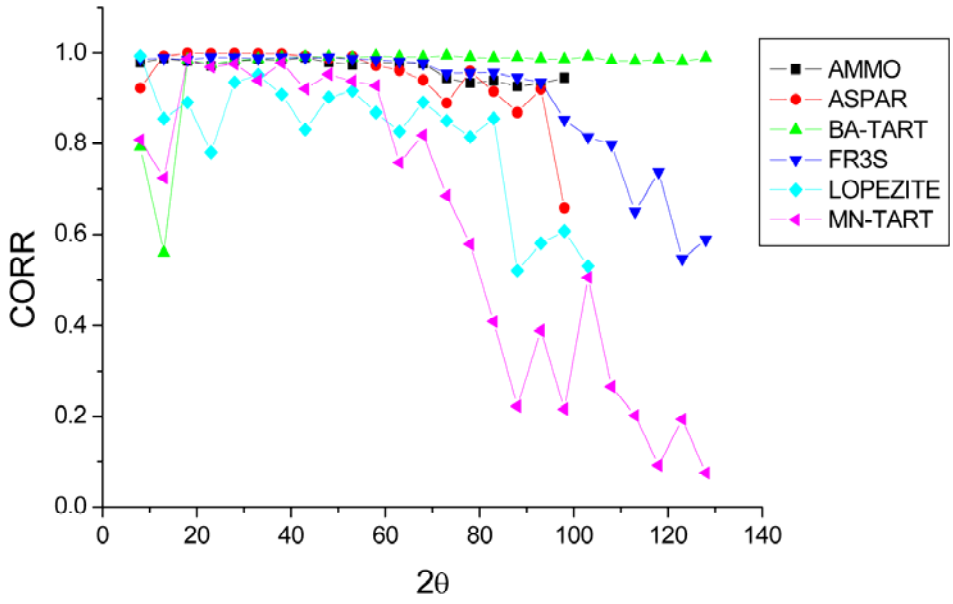


Figure 3.6 - For each test structure, the correlation (CORR) between the experimental FCT diagrams and the corresponding VCT normalised diagrams obtained at the end of the step iii) are given, for diffraction data collected at the 20th cycle.

3.3.2 Structure Solution

The phasing process has been carried out in the default way *via* the Direct Methods section of EXPO2004. The crystal structure solution was attempted by using the data corresponding at each cycle: for a given set of data EXPO2004 selects 10 trial solutions, among which we identified that with the minimum average phase error (say $<\Delta\varphi>_{min}$; the phase error is calculated with respect to the published structural model). In Figure 3.7 $<\Delta\varphi>_{min}$ is plotted *versus* the number of cycles. We observe: a) the trend of the error is oscillating. That is due to the stochastic nature of Direct Methods (an occasionally wrong intensity associated to a single reflection actively used in the phasing process may heavily influence the pathway); b) as a general trend, the phase error is smaller for VCT data, particularly at large cycle number: it is dramatically better for FR3S, ASPAR, MN-TART and LOPEZITE; c) for MN-TART in FCT mode the absence of the phase error in the initial cycles indicate that Direct Methods were not able to phase any reflection; d) for AMMO and BA-TART the FCT and VCT strategies produce nearly equivalent results, in accordance with the conclusions suggested by Fig. 3.6.

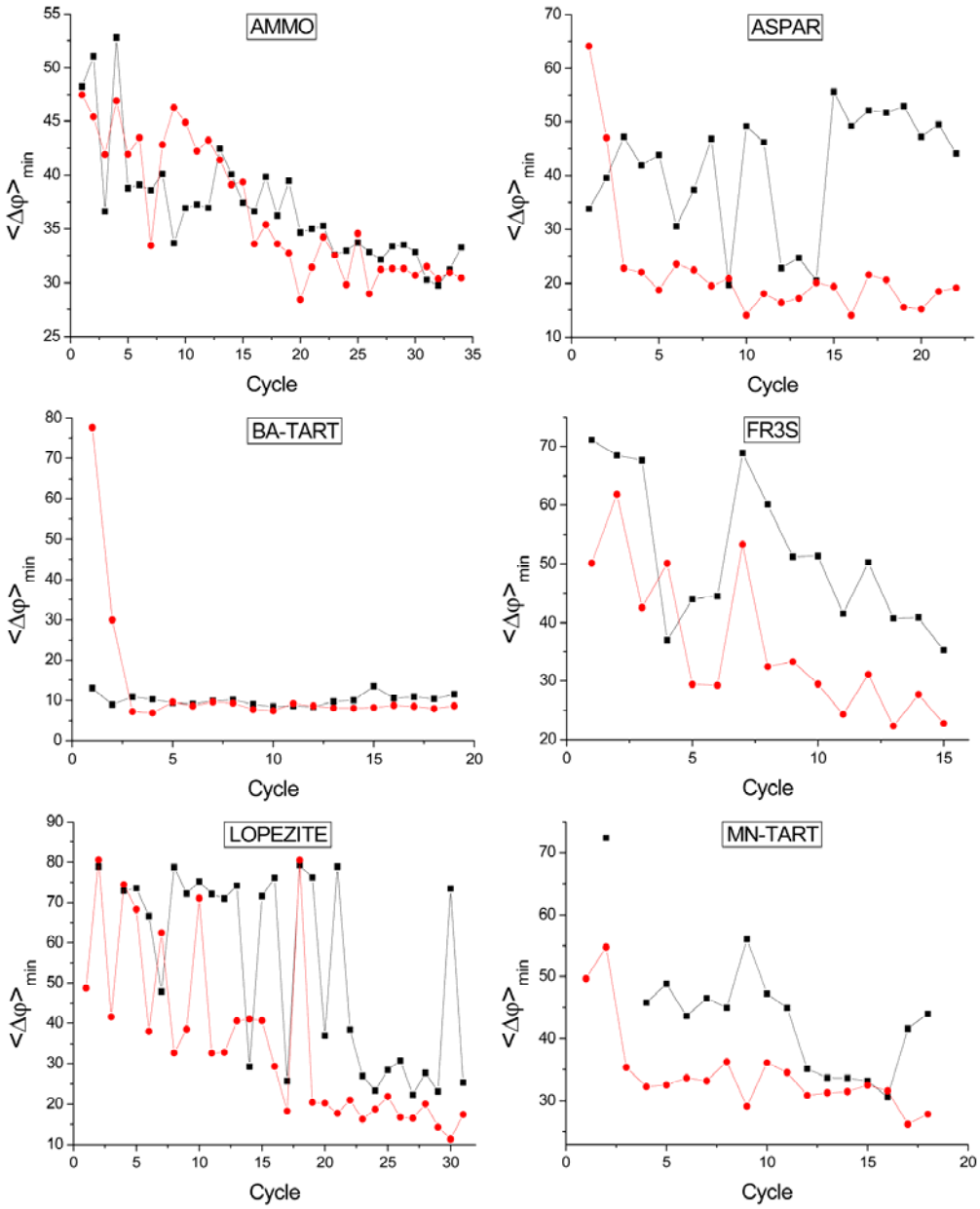
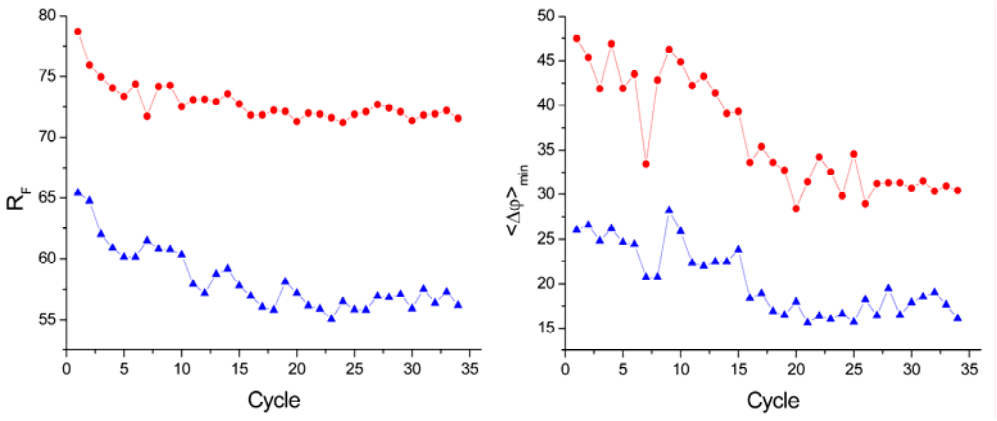


Figure 3.7 $<\Delta\varphi>_{min}$ *versus* the number of cycles for both FCT (black squares) and VCT (red circles) strategies.

In order to further asses the relative performances of the VCT and FCT strategies for crystal structure solution we analysed the electron density maps calculated at the end of the default EXPO2004 direct phasing, by Fourier transforming the trial solution having the minimum average phase error. In Table 3.2 we give, for each test structure and for the last cycle, the number of Fourier peak positions (NP) close (within a distance less than 0.6Å) to the true ones and their average distance (<d>). We note: a) for MN-TART and LOPEZITE $NP_{VCT}>NP_{FCT}$ and $<d>_{VCT} < <d>_{FCT}$. In accordance with the conclusions suggested by Fig. 3.6 the VCT mode provides better results; b) for FR3S and ASPAR $NP_{VCT}=NP_{FCT}$ and $<d>_{VCT} < <d>_{FCT}$. Again the VCT strategy is more useful. c) for AMMO $NP_{VCT}=NP_{FCT}$ but $<d>_{VCT} < <d>_{FCT}$, while for BA-TART $NP_{VCT}<NP_{FCT}$ but $<d>_{VCT} < <d>_{FCT}$. For this case the two strategies provide nearly equivalent results, in accordance with Fig. 3.6.

3.4 The best 2θ interval

Let us consider two test structures, AMMO and MN-TART. For AMMO the VCT strategy does not provide substantial additional information. Data were collected up to RES=1.02; however, owing to the (relatively) large unit cell, a great number of reflections at high 2θ angles severely overlap. We have therefore processed again the normalised VCT pattern of AMMO (20^{th} cycle) by using only data up to 1.4Å. In Fig. 3.8 we compare the R_F and the $<\Delta\varphi>_{min}$ values obtained at 1.4Å with those obtained at 1.02 Å; limiting the experimental 2θ interval makes the crystal structure solution more straightforward: for the 20^{th} cycle there are 28 found atoms, with <d>=0.3212.



a) b)

Figure 3.8 - AMMO: R_F (Fig. 3.8a) and $<\Delta\varphi>_{min}$ (Fig. 3.8b) obtained at RES = 1.4Å (blue triangles) and 1.02 Å (red circles), respectively.

For MN-TART Fig. 3.6 shows too low values of CORR for the highest resolution blocks. This suggests that the diffraction signal is practically absent: in a situation where using the VCT mode amplifies the noise rather than improves the count statistics. In Fig. 3.9 the VCT pattern of MN-TART at the end of the step iii) is shown: its inspection does not indicate in which blocks the noise is dominant. We cut the upper 2θ limit to 80° and we ran again EXPO2004 Direct Methods, showing a slight improvement on the results: 17 found atoms, with <d>=0.3178.

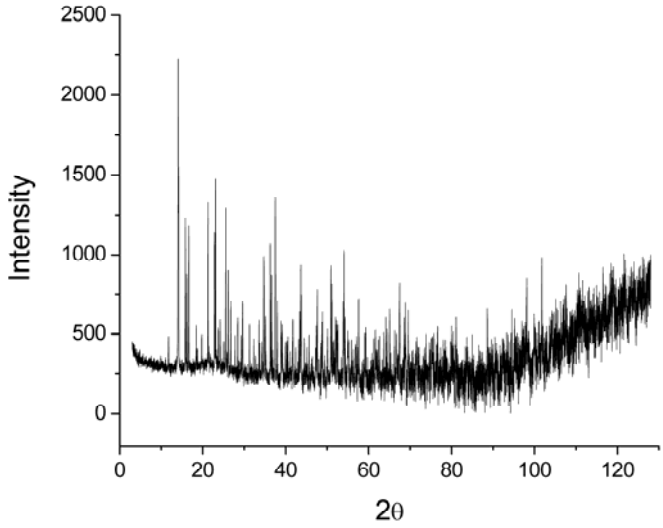


Figure 3.9 - MN-TART: the pattern available at the end of the step iii).

The above results suggest that, when the VCT mode is used, a supplementary data collection limited at the highest resolution blocks in FCT mode can provide useful information (through the calculation of CORR) about the quality of the high resolution data.

Table 3.2 NP is the number of electron density peaks whose positions are found close (within a distance less than 0.6Å) to the published ones, and <d> is the corresponding average distance.

Structure Code	FCT		VCT	
	NP	<d>	NP	<d>
AMMO	23	0.2944	23	0.2748
FR3S	15	0.1415	16	0.1416
ASPAR	9	0.1169	9	0.0881
BA-TART	7	0.2496	6	0.2072
MN-TART	10	0.3539	12	0.3324
LOPEZITE	13	0.2466	16	0.1904

3.5 Conclusions

In the present work we have evaluated the influence of the data collection strategy on the crystal structure solution process. Experimental data have been collected by using FCT and VCT schemes, and the phasing process has been performed *via* the Direct Methods section of EXPO2004. We have used six test structures having different complexity: four of them with heavy atoms and two of organic nature. The performances of the two strategies have been checked *via* different parameters such as the crystallographic residual R_F, the average phase error, the number of correctly located atoms at the end of the phasing process. The VCT strategy often (but not always) provides better results than the FCT scheme, essentially when it improves the data quality at medium and high 2θ angles, just where the reflection density becomes higher. As a rule of thumb, the more time we spend for data collection the more straightforward is the crystal structure solution: the improvement rate declines when the count statistics reaches a sufficient quality. The VCT strategy is particularly effective for organic structures (owing to their low scattering power at high angles) provided terminal 2θ blocks devoid of structural information are excluded from the calculations.

Chapter 4.

Space group determination from powder data

4.1 Introduction

When working with powder diffraction data, after the pattern has been successfully indexed, the following important step is the determination of the space group in which the compound has crystallized. Traditionally, it has been carried out by directly observation of the experimental profile by the researcher, in order to check what systematic extinctions are present and so deduct the corresponding extinction group. But, as we are dealing with powder diffraction data, the hard problem of reflection overlapping, due to the collapse of the three-dimensional reciprocal space, is still present (as well as the uncertainty about the background definition). In the space group determination procedure, this problem implies that saying that a certain reflection is 'present' or 'absent' is not anymore an easy task.

The direct visual inspection of the diffraction pattern was avoided in a recent approach proposed by Markvardsen *et al.* (2001), who elaborated an algorithm, based on a Bayesian probability approach, for calculating the relative probabilities of the different extinction symbols compatible with the actual crystal system of the sample. The intensities used by the procedure to estimate those probabilities were extracted from the pattern by means of the Pawley (1981) method. Although their test showed that the correct extinction symbol often coincides with the correct one, the algorithm also showed how difficult is to significantly discriminate between the correct and incorrect extinction symbols, in terms of their assigned probabilities. Also, the tests were carried out with a small number of test structures.

In this chapter a probabilistic algorithm, implemented in EXPO2004 program (Altomare *et al*., 2004c), will be developed. It is based on a statistical analysis of the normalized z intensities ($z = |E|^2$, where $E_{\mathbf{h}}$ are the normalized structure factors $F_{\mathbf{h}}$), aimed to determine the extinction symbol of the sample under study.

4.2 The algorithm

The algorithm proceeds as follow:

1. The unit cell must be already determined by indexing the experimental powder pattern; the crystal system is therefore assumed to be known.
2. For each crystal system, the relative extinction symbols are stored (see Table 3.2 of *International Tables for Crystallography*, Vol. A, 1992). They are 14 for the monoclinic system, unique axis *b*; 111 for the orthorhombic;

31 for the tetragonal; 12 for the trigonal–hexagonal systems (only hexagonal axes considered); 18 for the cubic.

3. The full pattern decomposition is performed, according to the Le Bail algorithm, in the space group having the largest Laue symmetry and no extinction conditions (*e.g. P*12/*m*1 in the monoclinic system, *P*2/*m*2/*m*2/*m* for the orthorhombic, *P*4/*mmm* for the tetragonal, *P*6/*mmm* for the trigonal–hexagonal systems, and *Pm*–3*m* for the cubic).

4. The extracted intensities are then normalized, according to the classical Wilson method, using as space groups the Laue groups mentioned in point 3. This is an arbitrary choice, because the statistical Wilson parameter ε varies with the space group, and this is unknown at this stage. The choice, however, proved not to be critical for the success of the algorithm.

5. The *z* statistics are used to define the probability of the extinction symbol.

   To give a representative example, let us consider the orthorhombic system, where any space group may be represented by the general string

$$M\ r_1/s_1\ r_2/s_2\ r_3/s_3 \tag{4.1}$$

   M denotes the unit-cell type; r_j, j = 1–3, are the symmetry elements along the three axes; s_j, j = 1–3, are the symmetry elements perpendicular to the axes. We assume that the occurrence probabilities for the 2_1 screw axes are

$$p(2_{1[100]}) = 1 - \langle z_{h00} \rangle_{h=2n+1}\ , \tag{4.2a}$$

$$p(2_{1[010]}) = 1 - \langle z_{0k0} \rangle_{k=2n+1}\ , \tag{4.2b}$$

$$p(2_{1[001]}) = 1 - \langle z_{00l} \rangle_{l=2n+1}\ . \tag{4.2c}$$

The probabilities are equal to unity if the z averages are close to zero, and vanish if the averages are equal to (or larger than) unity. The same equations define the probabilities for the twofold axes:

$$p(2_{[100]}) = 1 - p(2_{1[100]}) ;$$

$$p(2_{[010]}) = 1 - p(2_{1[010]}) ;$$

$$p(2_{[001]}) = 1 - p(2_{1[001]}) .$$

Let us now consider s_1 (the same criterion will hold for s_2 and s_3). It may coincide with b, c, n and d, for which the following probabilities arise:

$$p(b) = 1 - \langle z_{0kl} \rangle_{k=2n+1} ,$$

$$p(c) = 1 - \langle z_{0kl} \rangle_{l=2n+1} ,$$

$$p(n) = 1 - \langle z_{0kl} \rangle_{k+l=2n+1} ,$$

$$p(d) = 1 - \langle z_{0kl} \rangle_{k+l \neq 4n} .$$

The probability for the mirror plane is assumed to be

$$p(m) = 1 - \max [p(b), p(c), p(n), p(d)] .$$

The probabilities for the allowed unit-cell types may be derived as follows. First we calculate

$$p'(A) = 1 - \langle z_{hkl} \rangle_{k+l=2n+1} ,$$

$$p'(B) = 1 - \langle z_{hkl} \rangle_{h+l=2n+1} ,$$

$$p'(C) = 1 - \langle z_{hkl} \rangle_{h+k=2n+1} ,$$

$$p'(I) = 1 - \langle z_{hkl} \rangle_{h+k+l=2n+1} .$$

from which we obtain the probabilities

$$p(A) = p'(A)\,[1 - p'(B)]\,[1 - p'(C)]\,, \tag{4.3a}$$

$$p(B) = p'(B)\,[1 - p'(A)]\,[1 - p'(C)]\,, \tag{4.3b}$$

$$p(C) = p'(C)\,[1 - p'(A)]\,[1 - p'(B)]\,, \tag{4.3c}$$

$$p(I) = p'(I)\,. \tag{4.3d}$$

Equations (4.3*a*)–(4.3*c*) express the probabilities that the cell is *A*, *B* or *C* without being *F*. The probabilities for the *F* and the *P* cells are, respectively,

$$p(F) = 1 - \langle z_{hkl} \rangle_{[g]}$$

where [*g*] is the subset of reflections for which not all the indices are odd or even, and

$$p(P) = 1 - \max\,[p(A), p(B), p(C), p(I), p(F)]\,.$$

Each extinction symbol of the orthorhombic system is compared with the general string given in (4.1); the probability is calculated as the product of the probabilities of the symmetry elements involved in the extinction symbol. For example, for the symbol *P* - - - we can write the following expression:

$$p(P - - -) = p(P)\; p(2_{[100]})\; p(m \perp \mathbf{a})\; p(2_{[010]})\; p(m \perp \mathbf{b})\; p(2_{[001]})\; p(m \perp \mathbf{c})$$

Let us give the probability expression for an extinction symbol with a centred cell, *e.g. Bb*−*b*. The string *b*−*b* implies the following symmetry elements:

$$b \perp \mathbf{a},\; b \perp \mathbf{c},\; 2_{1[010]}\,.$$

The presence of *B* additionally implies

$$c \perp \mathbf{a},\; n \perp \mathbf{b},\; a \perp \mathbf{c},\; 2_{1[100]},\; 2_{1[001]}\,.$$

Accordingly (see Table 3.2 of *ITCr*), the probability for the extinction symbol *Bb-b* has to be calculated as

$$p(Bb\text{-}b) = p(B)\, p(2_{1[100]})\, p(b,c\perp\mathbf{a})\, p(2_{1[010]})\, p(n\perp\mathbf{b})\, p(2_{1[001]})\, p(a,b\perp\mathbf{c}) .$$

The scheme is easily extended to other crystal systems; we will examine some distinctive examples to clarify possible obscurities.

(a) The occurrence probability of the unit-cell type may be fixed as follows.

- Monoclinic system: the 14 extinction symbols include *A*, *C*, *I* and *P* cells; since the extinction conditions are mutually exclusive, we have $p(A) = p'(A)$, $p(C) = p'(C)$, $p(I) = p'(I)$, $p(P) = 1 - \max[p(A), p(C), p(I)]$.
- Tetragonal system: $p(P) = 1 - p(I)$.
- Trigonal–hexagonal systems: $p(P) = 1 - p(R)$.
- Cubic system. $p(I) = p'(I)$, $p(F) = p'(F)$, $p(P) = 1 - \max[p(I), p(F)]$.

(b) In the trigonal–hexagonal systems, *P* - - - may represent both trigonal (*e.g. P*3, *P*-3, *P*321 *etc*.) and hexagonal (*P*6, *P*-6, *P*6*mm*, *P*-62*m etc*.) space groups. Then:

- $p(3 \text{ or } 6) = 1 - \max[p(3_1), p(6_1), p(6_3)]$.
- $p(m \perp \mathbf{c}) = 1$ (no glide plane perpendicular to **c** in these systems).
- $p(2_{[100]})$ or $p(2_{[1\text{-}10]}) = 1$ (no extinction condition for screws in the plane normal to **c**).
- $p(m \perp \mathbf{a}) = 1 - p(c \perp \mathbf{a})$ (only *c* glide planes can occur).

In conclusion, *P* - - - is equivalent to the string *P*(3 or 6)/*m* 2/*m* 2/*m* and therefore $p(P\text{ - - -}) = p(P)\, p(3 \text{ or } 6)\, p(m \perp \mathbf{a})\, p[m \perp (\mathbf{a} - \mathbf{b})]$.

6. For each crystal system, the probabilities of the various extinction symbols are calculated and presented to the user in order of probability. Clicking on an extinction symbol in the list opens the sublist of the compatible space groups, from which the user can choose the preferred one.

4.2.1 About peak overlapping

As has been said, in powder diffraction the collapse of the three-dimensional lattice onto the 2θ axis hinders the unambiguous estimation of the diffracted intensities; it may frequently occur that for most of the reflections the assigned z value is uncertain, and that uncertainty can critically weaken any criterion based on the standard z statistics. It seems more useful to associate each reflection with a proper weight, depending on the accuracy of the intensity estimate.

Let us suppose that two symmetry-independent reflections perfectly overlap. For non-centrosymmetric space groups, the joint probability distribution $P(z_1, z_2)$ is given by

$$P(z_1, z_2) = \exp(-z_1 - z_2)\,.$$

The probability that the sum of z_1 and z_2 is equal to S is given by

$$P(S) = \int_0^S P(z_1, S - z_1)\mathrm{d}z_1$$

$$= \int_0^S \exp(-z_1)\exp(-S+z_1)\mathrm{d}z_1 = \int_0^S \exp(-S)\mathrm{d}z_1 = S\exp(-S) \qquad (4.4)$$

For a cluster of n perfectly overlapping reflections, with normalized diffraction moduli $z_1, z_2, \ldots, z_n$, the probability $P(S)$ is given by

$$P(S) = \int_0^S \cdots \int_0^S \exp[(-z_1 - z_2 - \cdots - (S - z_1 - z_2 - \cdots - z_n)]\mathrm{d}z_1 \ldots \mathrm{d}z_{n-1} =$$

$$= S^{n-1}\exp(-S) \qquad (4.5)$$

where $S = z_1 + z_2 + \cdots + z_n$. Equation (4.5) suggests that the overall intensity of a cluster of n overlapping reflections is distributed differently from the single reflection intensity (see Fig. 4.1). In accordance with the central limit theorem, for increasing values of n the distribution $P(S)$ becomes more similar to a Gaussian, with mean value $\langle S \rangle$ equal to the sum of the mean values ($\langle S \rangle = n\langle z \rangle = n$) and variance equal to the sum of the z variances. This trend is clearly visible in Fig. 4.1.

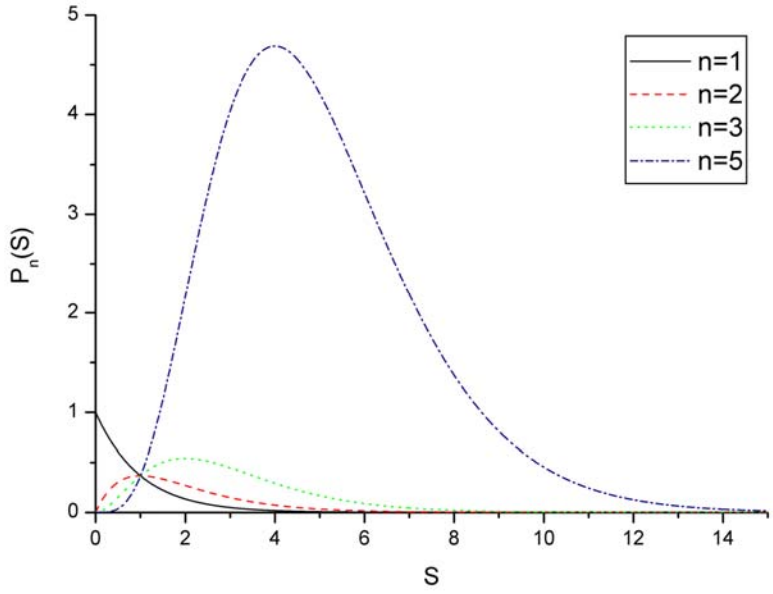


Figure 4.1 Equation 4.5 plotted for different values of *n*.

Let us now calculate, for the case when *S* is the measured intensity of a cluster of *n* perfectly overlapping reflections, the variance σ_z^2 associated with the generic intensity component *z* (note that the Le Bail algorithm, in the absence of additional prior information, would assign the same intensity *S*/*n* to all the reflections of the cluster). We have

$$\sigma_z^2 = \int_0^S z^2 \exp(-z)\mathrm{d}z - \left[\int_0^S z \exp(-z)\mathrm{d}z\right]^2 \tag{4.6}$$

Since

$$\int_0^S z^n \exp(-z)\mathrm{d}z = n! - \exp(-S)\sum_{k=0}^{n}\frac{n!}{k!}S^k$$

we have

$$\int_0^S z^2 \exp(-z)\mathrm{d}z = 2 - \exp(-S)(2 + 2S + S^2)\,,$$

$$\int_0^S z\ \exp(-z)dz = 1 - \exp(-S)(1 + S)$$

from which we obtain

$$\sigma_z^2 = 1 - S^2 \exp(-S) - (1 + S)^2 \exp(-2S) \qquad (4.7)$$

It is easily seen that:

a) $\sigma_z^2 = 0$ when $S = 0$. In this case the error on the intensity estimate due to overlapping is vanishing, as should be expected.

b) If S tends to infinity then σ_z^2 tends to unity (the value of the variance for the acentric Wilson distribution). This is the maximal uncertainty caused by the perfect overlapping condition.

c) Equation (4.7) always provides σ_z^2 values lying in the interval (0, 1) (see Fig. 4.2).

To treat clusters not constituted by perfectly overlapping reflections, we first specify when a reflection is defined as single, namely when no other reflection is closer than

$$\mathrm{tr}\theta = s \cdot FWHM$$

where *FWHM* is the full width at half-maximum calculated at the cluster 2θ angle and s is a parameter fixed by the user (in our tests $s = 0.7$). Accordingly, two reflections for which

$$t = (2\theta_1 - 2\theta_2)/\mathrm{tr}\theta < 1$$

belong to the same cluster.

Let us now relax the condition of the perfect overlapping. For a general cluster the following variance may be associated to the single intensity components:

$$\sigma_z^2 = (1-t)^q [1 - S^2 \exp(-S) - (1+S)^2 \exp(-2S)]. \qquad (4.8)$$

Equation. (4.8) agrees well with the common sense. Indeed:

a) when t tends to unity σ_z^2 tends to zero, as for single reflections (for these reflections there is no error due to the overlapping).

b) if $t = 0$ we have a case of perfectly overlapping reflections, and (4.8) coincides with (4.7). Larger values of t diminish the variance value.

c) q may be a positive, integral or fractional number - it damps or strengthens the effect of the t variable. In our tests we used $q = 2$.

The above results suggest that the presence of a symmetry element with a non-vanishing translation component may be better assessed by the average

$$\langle z_w \rangle = \left(\sum w_j z_j\right) \Big/ \sum w_j \ , \qquad (4.9)$$

than by $\langle z \rangle$, where $w = 1/\sigma_z^2$. According to (4.9) $w = 1$ is the minimum weight to assign in the case of strong overlapping with a very intense reflection, in all other cases $w > 1$. Since $\sigma_z^2 = 0$ for single reflections and for reflections in a cluster with overall vanishing intensity, the criterion $w = 1/\sigma_z^2$ would cause a singularity. We prefer then to use the following weight:

$$w = \exp\left(-2\sigma_z^2\right) ,$$

giving w = 1 for single reflections and for reflections belonging to a cluster with vanishing intensity, and w < 1 for overlapping reflections.

Until now, this mathematical treatment ignores the multiplicity m_i, i=1,...,n of the reflections belonging to the cluster. So, we assume $S = m_1 z_1 + m_2 z_2 + \cdots + m_n z_n$ and we replace (4.6) by

$$\sigma^2(z_i) = \int_0^{S'_i} z_i^{\,2} P(z_i) dz_i - \left[\int_0^{S'_i} z_i \; P(z_i) dz_i \right]^2 ,$$

where $S_i' = S / m_i$. Then, the equation (4.8) is replaced by

$$\sigma^2(z_i) = (1 - t^q)[1 - S_i'^2 \; \exp(-S'_i) - (1 + S'_i)^2 \exp(-2S'_i)] .$$

Similar calculations may be performed for the centrosymmetric case. For example,

$$P(S) = (2\pi)^{-n/2} \exp(-S/2) \cdot$$
$$\cdot \int_0^S \cdots \int_0^S [(z_1 z_2 \ldots z_{n-1})(S - z_1 - z_2 - \cdots - z_{n-1})]^{-1/2} \, \mathrm{d}z_1 \ldots \mathrm{d}z_{n-1}$$

The exact value of the right-hand side of the above equation is difficult to calculate for n larger than two. More practical is the use of the central limit theorem, which may be applied by calculating the following expression:

$$\left(\sigma_z^2\right)_1 = (2\pi)^{-1/2} \int_0^S z^{3/2} \exp(-z/2) \mathrm{d}z - (2\pi)^{-1} \left[\int_0^S z^{1/2} \exp(-z/2) \right]^2 \mathrm{d}z \ .$$

Since

$$\int_0^u z^{\nu-1}\exp(-\mu z)\mathrm{d}z = \mu^{-\nu}\gamma(\nu,\mu u)$$

where

$$\gamma(\alpha,x)=\int_0^x \exp(-t)t^{\alpha-1}\mathrm{d}t$$

is the incomplete gamma function, we obtain

$$\left(\sigma_z^2\right)_1 = \left[4/(\pi)^{1/2}\right]\cdot\left[\gamma(5/2,S/2)-(\pi)^{1/2}\gamma^2(3/2,S/2)\right].$$

The values of $\left(\sigma_z^2\right)_1$ (shown in Fig. 4.2) are always in the interval (0, 2). It may be noted that, for $k \to \infty$, $\left(\sigma_z^2\right)_1$ coincides with the variance of the Wilson distribution for the centrosymmetric case.

Since the space group is unknown at this stage, we decided to simplify the procedure by applying the results obtained for the non-centrosymmetric case to all the extinction symbols. This choice proved not to be critical.

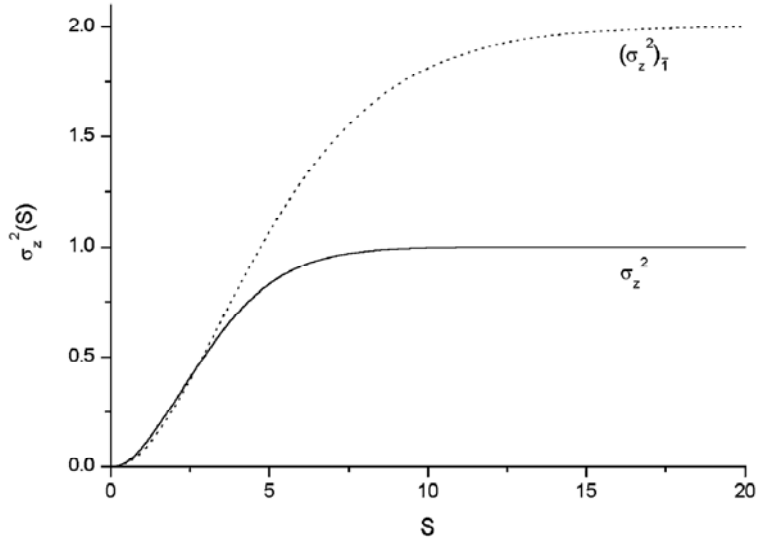


Figure 4.2 σ_z^2 and $\left(\sigma_z^2\right)_{\bar{1}}$ *versus* the overall cluster intensity *S*.

4.2.2 About the scarcity of the experimental information

Quite often the number, *n*, of reflections obeying the extinction rule of a screw axis is quite small, thus weakening the predictive power of our probabilistic criteria. Assuming that $\langle z_w \rangle$ is sufficiently indicative only if it contains at least four reflections, we replace *p*(screw) as follows:

$$\text{if } p(\text{screw}) < 0.5 \text{ then } p(\text{screw}) = p(\text{screw}) \exp[(4 - n)/8] ,$$

$$\text{if } p(\text{screw}) > 0.5 \text{ then } p(\text{screw}) = p(\text{screw}) \exp[-(4 - n)/8] .$$

4.2.3 A more stable statistical criterion

Let us denote by $\langle z_w \rangle_{[g]}$ and $\langle z_w \rangle_{[ng]}$ two weighted z averages, the first with respect to [g] (the set of reflections obeying the extinction rule) and the second with respect to [ng] (the complementary set). For example, for $c \perp \mathbf{a}$, [g] = {0kl} with l = 2n + 1, and [ng] = {0kl} with l = 2n.

It may occur that both the conditions $\langle z_w \rangle_{[g]} << 1$ and $\langle z_w \rangle_{[ng]} << 1$ hold; the first inequality suggests the presence of an extinction rule, the second reduces this expectation. It may also occur that $\langle z_w \rangle_{[g]} \approx 1$ and $\langle z_w \rangle_{[ng]} >> 1$. The first condition excludes the presence of an extinction rule, the second suggests more care. We then decided to derive the elementary probabilities of glide planes and centred unit cells by comparing the average $\langle z_w \rangle_{[g]}$ with $\langle z_w \rangle_{[ng]}$:

$$P(\text{glide}) \text{ and } p'(\text{centred cell}) = 1 - \langle z_w \rangle_{[g]} / \langle z_w \rangle_{[ng]} \,. \qquad (4.10)$$

The criterion (4.10) proved to be sufficiently robust for practical applications; it does have a statistical basis, which may be described by illustrating a simple example. Let us suppose that the true space group is $P2_12_12_1$; if the space group was known a priori, we would prefer to decompose the diffraction pattern in that space group and normalize the intensities according to the Wilson rule:

$$z = |F|^2 \Big/ \left(\varepsilon \sum_{\mathrm{N}} f_j^2 \right) ,$$

where ε = 1 always, except for the reflections (h00), (0k0) and (00l), for which ε = 2.

In the absence of any prior information and in accordance with §4.3, the pattern decomposition and normalization are both made in the space group $P2/m2/m2/m$, where

$\varepsilon = 1$ for (hkl); $\varepsilon = 2$ for (0kl); (h0l); (hk0) ;

$\varepsilon = 4$ for (h00); (0k0); (00l) .

As a practical result, the reflections (0kl), (h0l) and (hk0) will have $\langle z \rangle$ values smaller than their true values, so simulating the presence of glide planes. Since the same ε is applied to [g] and to [ng], the ratio $\langle z_w \rangle_{[ng]} / \langle z_w \rangle_{[g]}$ is insensitive to the error introduced by the normalization process and may constitute a more efficient criterion.

4.2.4 A limit for the elementary probabilities

The occurrence probability for an extinction symbol is the product of several elementary probabilities. If one of them is zero the other probability can no longer influence the final value. We have therefore imposed the following limitation on the elementary probabilities:

$$\text{if } p < 0.1 \text{ then } p = 0.1 \ .$$

Before presenting the results obtained with this algorithm, some additional developments (regarding the intensities extraction from the powder diffraction pattern and the visual inspection of it) must be introduced because, as we will see later, their contribution to the success of the procedure of finding the space group is very important.

4.3 Impurity peak removal

Impurity peaks, when not eliminated from the diffraction pattern, can affect the intensity extraction process, with bad effects on the efficiency of the procedure for space group determination previously explained (as well as for crystal structure solution). A new algorithm has been developed to identify and delete impurity peaks eventually present in the powder diffraction pattern.

Let us suppose that a peak search procedure has been used, that the diffraction pattern has been indexed, and that full pattern decomposition has been accomplished. The simplest condition for recognizing an impurity peak is the following: the 2θ-distance of the peak from the closest reflection is larger than the full width at half maximum (*FWHM*).

Impurity peaks are also recognised as follows:

1) The pairs of contiguous reflections ***h*** and ***k*** for which

$$2\theta_{\mathbf{k}} - 2\theta_{\mathbf{h}} \geq \frac{1}{2}\left[FWHM(\mathbf{h}) + FWHM(\mathbf{k})\right]$$

   are considered.

2) The intensity *y(i)* measured at each *i*-th count step between $2\theta_{\mathbf{h}}$ and $2\theta_{\mathbf{k}}$ is checked: the presence of an impurity peak is assessed if the next conditions are verified for at least one *i*:

$$2\theta_i - 2\theta_{\mathbf{h}} > \frac{1}{2}FWHM(\mathbf{h})$$

$$2\theta_{\mathbf{k}} - 2\theta_i > \frac{1}{2}FWHM(\mathbf{k})$$

$$y(i) > y(\mathbf{h}), y(\mathbf{k})$$

The peaks considered as impurities are deleted by using a filter technique as follows:

a) a Gaussian profile $y^c(i)$ centred at $2\theta_{max}$ is calculated: $2\theta_{max}$ is the 2θ position corresponding to the maximum intensity [say y_{max}] of the points satisfying the conditions stated at the step 2. We have:

$$y^c(i) = \begin{cases} y_{max} G(\Delta_i) & i \in \chi_{peak} \\ 0.0 & otherwise \end{cases}$$

$G(\Delta_i)$ is the Gaussian profile function, and χ_{peak} is the peak existence range;

b) the observed profile count $y(i)$ is replaced by:

$$y^n(i) = y(i) - y^c(i).$$

4.4 Background Improvement

The background estimate plays a central role in the determination of the integrated intensities extracted from a powder diffraction pattern, particularly for small intensities, which are the most sensitive to the background modification. As the procedure for the space group determination here developed uses the statistical analysis of the normalized structure factors extracted from the experimental profile, it strongly depends on the peak intensities and requires accurate background evaluation. The algorithm uses the automatic background determination procedure implemented in EXPO (Altomare *et al.*, 1995b, 1999), which is based on the assumption that the

observed diffraction profile counts outside the reflection peak areas are background counts. The procedure is efficient, but it may fail because of two main reasons:

1) Impurity peaks are present in the pattern. In this case, since no reflection corresponds to the impurity peak, any automatic procedure will assume that the impurity peak counts are background points. Consequently, the background curve fitting those points will be overestimated (sometimes severely).

2) At very large 2θ values, where peak overlap is strong, it is difficult to discriminate background from Bragg signal.

Owing to the above two considerations, the strategy of EXPO allowed the user to easily modify, *via* a graphic interface, the background points fixed by the automatic procedure, by adding or deleting background points. This option is maintained, but it has been implemented in a new automatic procedure which should avoid the two above mentioned sources of errors. It is described below.

The EXPO strategy of dividing the observed pattern into intervals (Altomare *et al.*, 1995b) is maintained. In each interval a peak search procedure is used for locating the maxima of the experimental profile after having multiplied the count values by -1. The 2θ positions corresponding to such maxima are considered as background points. Since they may occasionally be located in peak areas the following steps are executed:

a) Let $y(i)$ be the observed intensity corresponding to the i-th background point. The polynomial curve best fitting the background points is calculated by a least squares procedure which associates to each point a weight $w(i) = 0.5$ if

$$y(i) \geq 1.1 \cdot y(i-1) \text{ or } y(i) \geq 1.1 \cdot y(i+1),$$

otherwise $w(i)$ is 1.0.

b) Let $bg(i)$ be the polynomial value at the i-th point. We consider the ratio

$$rbg(i) = [y(i)-bg(i)]/bg(i),$$

its mean value

$$rbg_{med} = \frac{\sum_{i=1}^{N} rbg(i)}{N}$$

and its standard deviation

$$rbg_{stdv} = \sqrt{\frac{\sum_{i=1}^{N} (rbg(i) - rbg_{med})^2}{N\text{-}1}}.$$

The summation goes over the N points satisfying the condition $y(i) > bg(i)$. If $rbg_{stdv} < 0.01$ the process passes to the step d); if $rbg(i) > sogh \cdot rbg_{stdv}$ the i-th point is removed from the set of background points (*sogh* is a suitable threshold whose value is clarified in step c). The aim of step b) is to remove background points having large deviation from the best fitted background polynomial (*e.g.* points corresponding to impurity contributions, points located in the peak areas). If no point is eliminated the procedure jumps to step d).

c) The selected background points are used for a new polynomial calculation. The process comes back to a) and is repeated up to ten times. The *sogh* value depends on the cycle number *nc*:

$$sogh = 0.5 + 0.05 \cdot (nc\text{-}1).$$

d) If the 2θ distance between two adjacent background points is larger than 0.5° then some background points are added (at intervals of 0.5°). Their intensity values are obtained by interpolation. This step d) aims at avoiding large 2θ intervals without background points, which may ill condition the least squares background polynomial calculation.

e) The background points are used for a new polynomial calculation.

f) The percentage *pg* of the pattern points whose current background values $bg(i)$ are larger than the observed values $y(i)$ is calculated with respect to the full number of observations in the profile. If *pg* is not in the interval (0.05, 0.08), the background value in each point of the pattern is increased (if $pg < 0.05$) or decreased (if $pg > 0.08$) by constant factors (0.98 for decreasing and 1.003 for increasing). The step f) is repeated until the condition on *pg* is satisfied. This procedure is particularly effective when the background is underestimated (*e.g.* when the pattern is noisy).

4.5 More selective criteria for identifying the correct extinction group

If we are examining single crystal data under the hypothesis of the space group SG, and we find a few non-zero intensities corresponding to reflections which should be systematically absent, we should reject (by excluding Renninger effects) the assumed SG, even in the case in which $\langle z \rangle$ is close to zero (the average is over the systematically absent reflections). This means that the statistical analysis adopted by the actual algorithm for selecting the most

probable extinction group from powder data, based only on the *z*-statistics, may fail in unfortunate cases. The risk is higher for powder data because reflections overlap in clusters and each *z*-value may be wrongly evaluated by the decomposition algorithm. In order to improve the efficiency of the score based on the *z*-statistics, we have introduced in the automatic procedure the following additional criteria:

a) if a cell is supposed to be centred on the basis of the *z*-statistics, but four or more single (*i.e.*, non overlapping with any other) reflections, expected to be systematically absent, are found with $d \geq 2.25$ Å and $z_w \geq 0.45$, then the hypothesis is rejected;

b) if a glide plane is supposed to be present on the basis of the *z*-statistics, but two or more single reflections, expected to be systematically absent, are found with $d \geq 2.25$ Å and $z_w \geq 0.45$, then the hypothesis is rejected;

c) the same is made for a supposed screw axis when two single reflections are found with $d \geq 2$ Å and $z_w \geq 0.35$.

If, for a given extinction group, at least one of the previous conditions is verified, the probability value of the extinction group provided by the *z*-statistics [say $P_{old}(ext.\ group)$] is replaced by:

$$P_{new}(ext.\ group) = 0.01 \cdot P_{old}(ext.\ group) \ .$$

To gain experimental information, also clusters constituted by reflections with reduced multiplicity when this is fixed by the same symmetry operator (*e.g.*, in the orthorhombic system when both the reflections are of type (*hk*0) with $h+k =$ odd) are tested by the above criteria.

4.6 Applications

The statistical criteria described in §4.3, together with the additional procedures of §4.4, §4.5 and §4.6, have been applied to 59 test structures; their code names are quoted in Table 4.1. The structures cover all (but the triclinic) crystal systems; for the sake of brevity, for each structure we provide not the main crystallochemical data but only the true space group and references to where the reader can find supplementary information. We decided to verify the robustness of our algorithm by using a large set of test structures for several reasons:

i. Each pattern shows features that are experiment dependent, *e.g.* a different degree of overlapping, a preferred orientation or a background of variable complexity. In particular, the set of test structures contains traditional X-ray generator data as well as synchrotron and neutron data.

ii. Each pattern reflects eventual pseudosymmetries (*e.g.* of pseudotranslational nature) present in the crystal structure, thus making the identification of the correct extinction symbol difficult.

iii. The definition of the correct extinction symbol is more difficult in some crystal systems.

Table 4.1 - Code names of the test structures and their space groups. The numbers appearing on the fist column correspond to the main references, available on Appendix A.

Code name	Space group
ADAM [31]	*F m–3m*
AGPZ [1]	*Pbca*
ALDS, ALDX [2]	$P2_12_12_1$
ALPHA [3]	$P2_12_12_1$
AND1, AND2 AND1N, AND2N [4]	$P2_1/n$
BACO [5]	*C2/m*
BAMO [6]	$P2_1$
BENZ [38]	$P2_1/a$
CeO_2 [39]	*Fm-3m*
CF3BR [40]	$P2_1/a$
CFCL [41]	*Fdd2*
CFI [42]	*Cmca*
CIME [8]	$P2_1/n$
Cs_2BeF_4 [43]	*Pnma*
CUPZ [44]	*Pbca*
DADA [11]	$P2_12_12_1$
DFQP [12]	*I2/a*
ECH1 [45]	$I4_1/a$
EMT [13]	$P6_3/mmc$
F1A [46]	*R-3*
FAMO [47]	$P2_1/c$
FORMYLUREA [48]	$Pn2_1a$
FR3S [32]	$P2_12_12_1$
GAPO [14]	*Pbca*
IBUPS [15]	$P2_1/c$
K_2BeF_4 [43]	*Pnma*
KUOS [49]	*P4/ncc*

Code name	Space group
LaB_6 [50]	*Pm-3m*
LAMO [16]	$P2_1/a$
LASI [17]	$P2_1/c$
LASO [51]	*C2/c*
LASOKA [52]	*Pnma*
LEV [18]	*R-3m*
MCM-22 [19]	*P6/mmm*
MES [53]	$P2_1/c$
METYL [54]	*I222*
NAC [55]	$I2_13$
NBPO [20]	*C2/c*
NIZR [21]	$P2_1/n$
NORB2 [56]	$P2_1/c$
OTHYM [22]	$P2_1/n$
PBS [57]	*Pbca*
Rb_2BeF_4 [43]	*Pnma*
RUCO [58]	*Ibam*
SAPO-40 [23]	*Pmmn*
SBPO [24]	$P2_1/n$
SCAMPHOR1 [59]	$P2_12_12_1$
SGT [25]	$I4_1/amd$
SULPH [26]	*Pbcm*
UTM1 [27]	*C2/m*
VFI [28]	$P6_3$
VNI [60]	$P4_12_12$
Y_2O_3 [61]	*Ia-3*
YONO [29]	$P2_1$
YURI [30]	$P2_1/c$

All the above-mentioned features can influence the robustness of any probabilistic approach aimed at determining space-group symmetry; the method may be successful in some cases and inadequate in others. The use of a large set of test structures allowed us to verify the general applicability of our algorithm, and suggests to the reader the nature of the obstacles that they have to overcome in defining the correct space group.

We have applied the following procedure:

a) The experimental patterns are decomposed into single diffraction intensities by the modulus EXTRA (Altomare *et al.*, 1995b) of EXPO2004 and the resulting intensities are normalized.

b) The probabilistic algorithm illustrated above is applied. The results of our tests are shown in Table 4.2, where it is provided, for each structure, the correct extinction symbol (cexts) and the order number at which it is ranked by the probabilistic criteria (orn; orn = 1 means that cexts has the highest probability value).

Non-crystallographic symmetry can hinder the identification of the correct extinction group (*e.g.* non-crystallographic symmetry can simulate centred cells). To illustrate this statement, for each crystal system we regrouped the test structures into two sets; set *A* contains the structures for which EXPO2004 did not find (via a statistical analysis of the diffraction intensities) any important pseudotranslational symmetry, and set *B* contains the structures with relevant pseudotranslational symmetry effects. For set *B*, perc > 0.16, where perc is the percentage of electrons satisfying the pseudosymmetry [see Altomare *et al.* (1996)]. The set *B* structures are the best candidates to show a centred cell or to simulate it.

Table 4.2 – For each test structure it is given the correct extinction symbol (*cexts*), the order number *orn* at which it is ranked by the automatic procedure for the identification of the extinction symbol. The order number *orn2* refer to the case in which the pseudotranslation correction is applied.

MONOCLINIC			
Codename	**cexts**	**orn**	**orn2**
AND1	P 1 2_1/n 1	1	
AND1N	P 1 2_1/n 1	1	
AND2	P 1 2_1/n 1	1	
AND2N	P 1 2_1/n 1	1	
BENZ	P 1 2_1/a 1	1	
CF3BR	P 1 2_1/a 1	1	
CIME	P 1 2_1/n 1	1	
FAMO	P 1 2_1/c 1	1	
IBUPS	P 1 2_1/c 1	1	
LAMO	P 1 2_1/a 1	1	
LASI	P 1 2_1/c 1	1	
MES	P 1 2_1/c 1	2	
NORB2	P 1 2_1/c 1	1	
OTHYM	P 1 2_1/n 1	1	
YONO	P 1 2_1 1	1	
YURI	P 1 2_1/c 1	1	
BACO	C 1 - 1	1	10
BAMO	P 1 2_1 1	2	1
DFQP	I 1 a 1	1	5
LASO	C 1 c 1	1	3
NBPO	C 1 c 1	1	2
NIZR	P 1 2_1/n 1	1	1
SBPO	P 1 2_1/n 1	1	1
UTM1	C 1 - 1	1	3

ORTHORHOMBIC			
Codename	**cexts**	**orn**	**orn2**
ALDS	P 2_1 2_1 2_1	1	
ALDX	P 2_1 2_1 2_1	1	
Cs_2BeF_4	P n - a	1	
DADA	P 2_1 2_1 2_1	3	
FR3S	P 2_1 2_1 2_1	1	
GAPO	P b c a	1	
K_2BeF_4	P n - a	1	
LASOKA	P n - a	1	
Rb_2BeF_4	P n - a	1	
SAPO	P - - n	1	
SCAMPHOR1	P 2_1 2_1 2_1	2	
AGPZ	P b c a	29	2
ALPHA	P 2_1 2_1 2_1	1	1
CFCL	F d d -	1	1
CFI	C - c (ab)	1	16
CUPZ	P b c a	10	1
FORMYLUREA	P n - a	1	1
METYL	I - - -	2	44
PBS	P b c a	1	2
RUCO	I b a -	3	15
SULPH	P b c -	1	1

Table 4.2 (Continuation)

TETRAGONAL			
Codename	**cexts**	**orn**	**orn2**
ECH1	I 4_1/a - -	5	27
KUOS	P n c c	1	9
SGT	I a - d	1	9
VNI	P 4_1 2_1 -	3	3

TRIGONAL / HEXAGONAL			
Codename	**cexts**	**orn**	**orn2**
EMT	P - - c	1	
MCM	P - - -	2	
F1A	R - -	7	7
LEV	R - -	1	5
VFI	P 6_3 - -	2	2

CUBIC			
Codename	**cexts**	**orn**	**orn2**
ADAM	F - - -	1	7
CeO_2	F - - -	3	16
LaB_6	P - - -	4	4
NAC	I - - -	1	18
Y_2O_3	I a - -	1	12

From Table 4.2, the following can be observed:

i. The extinction symbol is correctly identified (orn and/or orn2 = 1) for 46 structures; orn/orn2 = 2 for other six (the difference between the meaning of orn and orn2 will be explained in the next point).

ii. Identifying the extinction symbol is easier for the structures belonging to set *A*. There are a few exceptions, which will be commented on shortly. For SCAMPHOR1 orn = 2. Fig. 4.3 shows that a non-negligible percentage of atoms have their *x* coordinates close to zero or their *z* coordinates close to 0.25. In such conditions the space group $P2_12_12_1$ may simulate the glide planes $n \perp \mathbf{b}$, $a \perp \mathbf{c}$ and $b \perp \mathbf{c}$, which explains why the first ranked extinction symbol is *P-nb* rather than the correct one $P2_12_12_1$.

The value of orn for VNI is 3; the extinction symbol of this structure (for which c = 73.65 Å) cannot be found without the crystallochemical information employed by McCusker et al. (1996).

iii. The recognition of the correct extinction symbol for the test structures belonging to the set *B* is more difficult. The algorithm works well in some cases but fails badly in others (see, for example, AGPZ and CUPZ, where orn is 29 and 10, respectively). The failure is often due to the erroneous identification of the unit-cell type (it may be difficult to distinguish centred from pseudocentred cells), but also to erroneous identification of symmetry elements. We will take two examples to illustrate how the algorithm may be modified to face such problems.

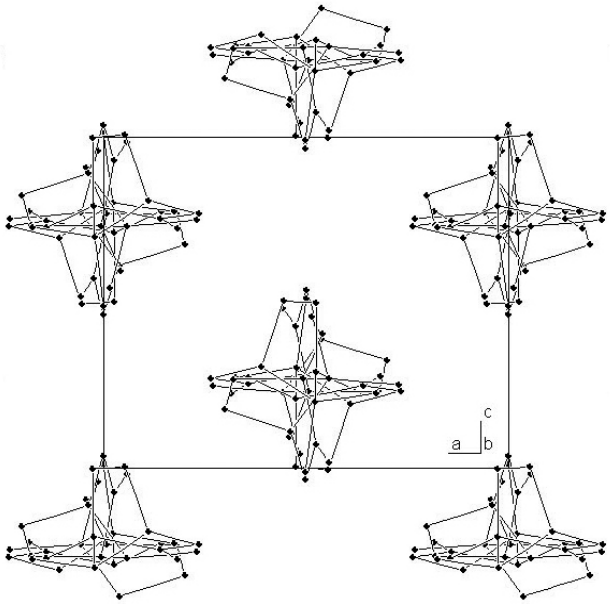


Figure 4.3 The crystal structure of SCAMPHOR1, viewed along the **b** axis.

For BAMO, EXPO2004 found a pseudotranslational vector **u** = **a**/2 + **b**/4 + **c**/2, with perc = 0.20; the corresponding condition for the substructure reflections is $2h + k + 2l = 4n$. The pseudotranslational symmetry simulates a glide plane, $n \perp \mathbf{b}$, which contributes to the large probability value of the selected symbol $P\text{-}2_1/n\text{-}$ rather than to $P\text{-}2_1\text{-}$. In order to simplify the recognition of the correct extinction symbol we have modified the algorithm as follows. Suppose that a pseudotranslational symmetry and the corresponding perc value have been found by EXPO2004, and that our probabilistic algorithm ranks the extinction symbols in order of probability (list 1). If a *P* cell is chosen as the most probable one, a new run is automatically launched in which glide planes and screw axes simulated by the pseudotranslational symmetry are identified, and their established elementary probabilities are multiplied by $[1 - \text{perc}(\mathbf{u})]^2$ (actually, we do not modify the probabilities of the screw axes). A supplementary list of the most probable extinction symbols (list 2) is then added to the original one, which can help the user to identify the correct extinction symbol. In Table 4.2 we show the orn value (orn2) corresponding to list 2. Remaining with the BAMO example, the probability of $P\text{-}2_1\text{-}$ and $P\text{-}2_1/n\text{-}$ are evaluated as follows: in list 1,

$$p(P\text{-}2_1/n\text{-}) = p(P)\, p(2_{1[010]})\, p(n \perp \mathbf{b}) ;$$

$$p(P\text{-}2_1\text{-}) = p(P)\, p(2_{1[010]})\, p(m \perp \mathbf{b}) ;$$

where $p(m \perp \mathbf{b}) = 1 - p(n \perp \mathbf{b})$.

In list 2

$$p(P\text{-}2_1/n\text{-}) = p(P)\, p(2_{1[010]})\, p(n \perp \mathbf{b})[1 - \mathrm{perc}(\mathbf{u})]^2 ;$$

$$p(P\text{-}2_1\text{-}) = p(P)\, p(2_{1[010]})\, \{1 - [p(n \perp \mathbf{b})(1 - \mathrm{perc}(\mathbf{u})]^2\} .$$

The above modifications lead to orn2 = 1 for BAMO (see Table 4.2). Generally speaking, list 2 provides the user with a different rank order alternative from that suggested by list 1.

Conversely, let us assume that a centred unit cell has been selected by our algorithm. If we suspect that the cell is actually pseudocentred, we have to diminish the elementary probability of screw axes and glide planes simulated by the centring. For example, for CUPZ, EXPO2004 found the pseudotranslation vector $\mathbf{u} = \mathbf{a}/2 + \mathbf{b}/2 + \mathbf{c}/2$ with perc = 0.47; the corresponding condition for the substructure reflections is $h + k + l = 2n$, corresponding to an *I* cell, and the selected extinction group is *Ibca*. We can rerun the program to reduce the probabilities of the centred cell, and of screw axes and glide planes simulated by the *I* centring (*e.g.* $n \perp \mathbf{a}$, $n \perp \mathbf{b}$, $n \perp \mathbf{c}$, $2_{1[100]}$, $2_{1[010]}$ and $2_{1[001]}$); the corresponding elementary probabilities (actually, only of the centred cell and glide planes) are multiplied by $[1 - \mathrm{perc}(\mathbf{u})]^2$, and a supplementary list of the most probable extinction symbols (list 2) is again added to the first list. CUPZ has orn2 = 1 in the new list.

The program automatically manages a limited list of pseudotranslational vectors; these are listed in Table 4.3, together with the corresponding reflection conditions and simulated symmetry elements.

It may be worth noting that the two lists offer two different ranking criteria: the first is the pure application of the statistical criteria defined

above, while the second aims to reduce the effects of the pseudotranslational symmetry. Usually, if an extinction symbol is highly ranked in one list, it is badly ranked in the second. Examination of the two lists can make it easier the identification of the correct extinction symbol.

iv. Finally, we observe that in several cases (about ten) the discrimination in terms of probabilities between the correctly identified extinction group and the other groups is significant. For example, for ADAM $p = 0.547$ for the correct symbol and $p = 0.07$ for the next one.

Table 4.3 List of the pseudotranslational vectors (**u**) managed by the algorithm. For each **u**, the reflection condition and the simulated symmetry elements are given.

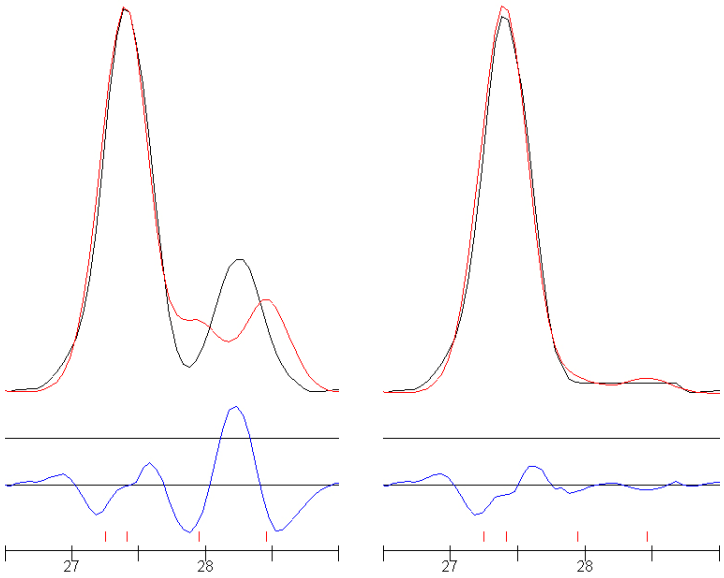
u		**Reflection condition**		**Simulated symmetry elements**
a/2	↔	$h = 2n$	→	$a \perp \mathbf{b}$; $a \perp \mathbf{c}$; 21[100]
b/2	↔	$k = 2n$	→	$b \perp \mathbf{a}$; $b \perp \mathbf{c}$; 21[010]
c/2	↔	$l = 2n$	→	$c \perp \mathbf{a}$; $c \perp \mathbf{b}$; 21[001]
a/2 + **b**/2	↔	$h + k = 2n$	→	C; $b \perp \mathbf{a}$; $a \perp \mathbf{b}$; $n \perp \mathbf{c}$; $2_{1[100]}$; $2_{1[010]}$
a/2 + **c**/2	↔	$h + l = 2n$	→	B; $a \perp \mathbf{c}$; $c \perp \mathbf{a}$; $n \perp \mathbf{b}$; $2_{1[100]}$; $2_{1[001]}$
b/2 + **c**/2	↔	$k + l = 2n$	→	A; $b \perp \mathbf{c}$; $c \perp \mathbf{b}$; $n \perp \mathbf{a}$; $2_{1[010]}$; $2_{1[001]}$
a/2 + **b**/2 + **c**/2	↔	$h + k + l = 2n$	→	I; $n \perp \mathbf{a}$; $n \perp \mathbf{b}$; $n \perp \mathbf{c}$; $2_{1[100]}$; $2_{1[010]}$; $2_{1[001]}$
-**a**/3 + **b**/3 + **c**/3	↔	$-h + k + l = 3n$	→	R; $3_{1[001]}$ or $6_{2[001]}$
a/2 + **b**/4	↔	$2h + k = 4n$	→	$a \perp \mathbf{b}$; $2_{1[100]}$
a/4 + **b**/2	↔	$h + 2k = 4n$	→	$b \perp \mathbf{a}$; $2_{1[010]}$

Table 4.3 (Continuation)

u		Reflection condition		Simulated symmetry elements
a/2 + **c**/4	↔	$2h + l = 4n$	→	$a \perp \mathbf{c}$; $2_{1[100]}$; $4_{1[001]}$
a/4 + **c**/2	↔	$h + 2l = 4n$	→	$c \perp \mathbf{a}$; $2_{1[001]}$
b/2 + **c**/4	↔	$2k + l = 4n$	→	$b \perp \mathbf{c}$; $2_{1[010]}$; $4_{1[001]}$
b/4 + **c**/2	↔	$k + 2l = 4n$	→	$c \perp \mathbf{b}$; $2_{1[001]}$
a/4 + **b**/2 + **c**/2	↔	$h + 2k + 2l = 4n$	→	$n \perp \mathbf{a}$; $2_{1[010]}$; $2_{1[001]}$
a/2 + **b**/4 + **c**/2	↔	$2h + k + 2l = 4n$	→	$n \perp \mathbf{b}$; $2_{1[100]}$; $2_{1[001]}$
a/2 + **b**/2 + **c**/4	↔	$2h + 2k + l = 4n$	→	$n \perp \mathbf{c}$; $2_{1[100]}$; $2_{1[010]}$

4.6.1 Application of the impurity peak removal procedure

In Figure 4.4 it can be seen how this algorithm works for CFCL (see Table 4.1) into the 2θ interval (26.5°-29º). The intensity extraction process (the first step of the procedure for the determination of the extinction group), performed in the space group *Pmmm*, is unable to fit the experimental profile in the range (26.5°-29º) by means of the two reflections lying at 27.94º and 28.46º (see Fig.4.4a). Our algorithm corrects the experimental profile as in Fig.4.4b, providing nearly vanishing intensities to the two reflections. Actually, these two reflections are extinct for the right extinction group *Fdd2*: accordingly, the application of our algorithm makes the determination of the correct space group easier.



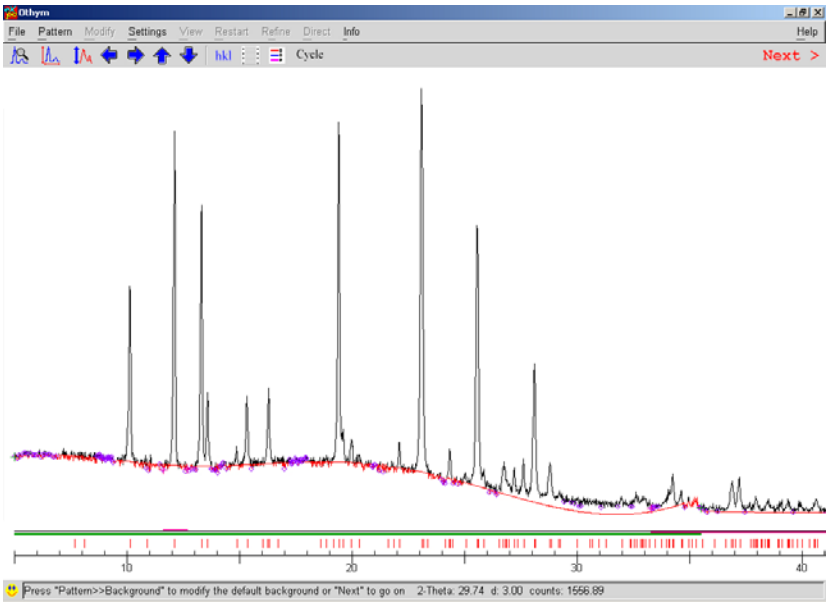
a) b)

Figure 4.4 – The interval 25.5-29.5º of the diffraction pattern, after the intensity extraction procedure, is shown for the CFCL structure: a) on a normal run and b) applying the impurity removal algorithm.

4.6.2 Application of the improved automatic background procedure

This procedure provides a more reasonable background evaluation and, combined with the other new features described in this work, contributes to improve the identification of the extinction group. An example is shown in Figure 4.5 where for OTHYM (see Table 4.1) the background evaluations by the old (Fig. 4.5a) and the new (Fig. 4.5b) procedures are respectively given. It may be useful to note how better the new polynomial in the range 26°-35° is.

a)



b)

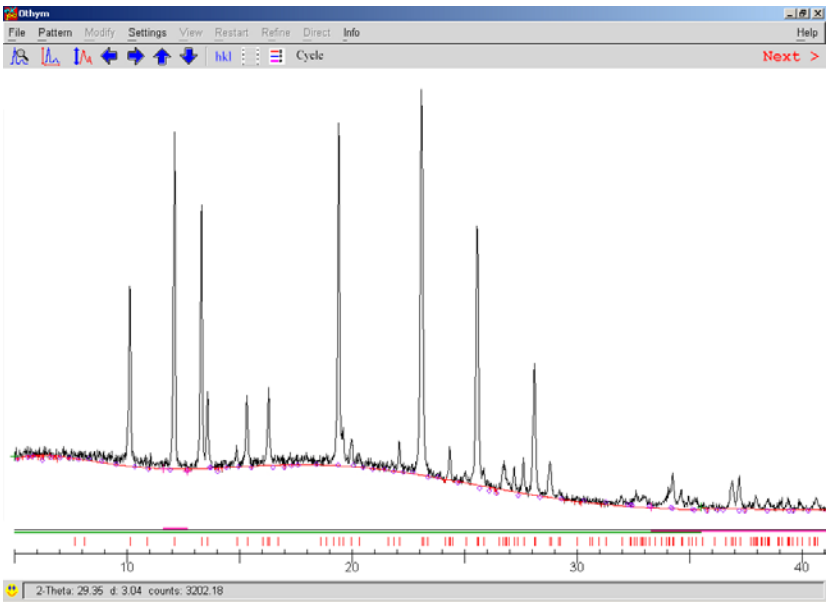


Figure 4.5 – The background (red line) calculated by the old (a) and the new (b) procedure is shown for the OTHYM structure.

4.7 The graphical interface

Our last observation concerns the usefulness of direct visual inspection of the diffraction pattern. As with any other statistical method, this approach provides criteria based on the averages of useful variables, rather than on the study of the profiles of specific peaks. Thus, visual inspection of the pattern is a useful complement to the technique. We will see two examples:

The automatic procedure is able to recognise the correct extinction symbol for 46 of the 59 test structures. For them *orn* (and/or *orn2*) is 1. One can wonder if the user inspection of the diffraction diagram can discard the wrong extinction symbols having probability larger than that of the true one. To facilitate the user control, the standard graphical interface of EXPO2004 has been modified. Such interface provides a window showing the extinction symbols ranked according to their probability: clicking on one of them, the list of the compatible space groups is supplied. In addition, the new graphical interface allows the user to gain the following information:

a) the list of the reflections which, according to the selected extinction symbol, are expected to be systematically absent. For each of them the following specifications are given: if the reflection is single or in overlapping, the corresponding z_w value and the symmetry operator(s) responsible for the extinction.

b) an histogram for each symmetry operator, giving the number of systematically absent reflections (*nsar*) *versus* the z_w values. The histogram provides an immediate evaluation of the distribution of the intensities associated to potentially systematically absent reflections. As an example, we consider the test structure METYL: the highest score extinction group is *I - (ac)* - (P=0.214); the second, in the order,

is *I* - - - (*P*=0.213) which is the correct one. For *I* - *(ac)* - we show in Figures 4.6a, 4.6b, 4.6c three histograms corresponding to the three symmetry operators *I*, *a*, *c* respectively. Figure 4.6c shows, in the ellipse, a single reflection (blue colour) which is expected to be systematically absent. In Figure 4.7 the corresponding peak is shown, which clearly eliminates the extinction symbol *I* - *(ac)* - from the list of candidates.

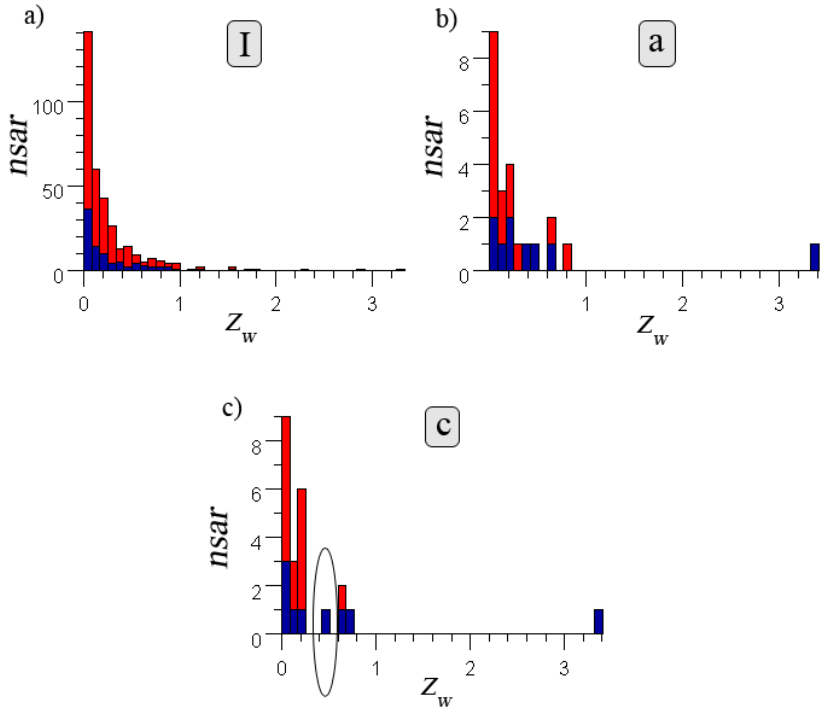


Figure 4.6 – Histogram plots of METYL corresponding to the symmetry operators *I*, *a*, *c*.

c) the reflections which are expected to be absent for the selected extinction symbol are marked by a vertical bar [see Figure 4.7, where the diffraction pattern of METYL in the range (9º- 14.5º) is shown according to the extinction group *I - (ac) -*]. These bars are situated below the observed pattern, just over the set of bars defining the peak positions, and have different colours depending on whether the reflection is single (or belongs to a whole extinct cluster) or in overlapping. In default, black vertical bar will mark single systematically absent reflections.

This graphical interface allows to easily correcting most of the errors made by the automatic procedure. For example:

METYL- The top ranked extinction symbol is *I - (ac) -*; *I - - -*, the true one, is ranked second. The experimental diagram shows, at $2\theta = 13.26º$, an evident peak (see Fig. 4.7) with z_w=0.180, marked by a black vertical bar (the reflection is considered as single by the program). The user can then safely discard *I - (ac) -* in favour of *I - - -*.

Working with this visual help, for the DADA, METYL, ECH1, MCM, CeO_2 and LaB_6 cases, the extinction groups with higher probability than the true one can be discarded with just a quick inspection of the pattern by the user.

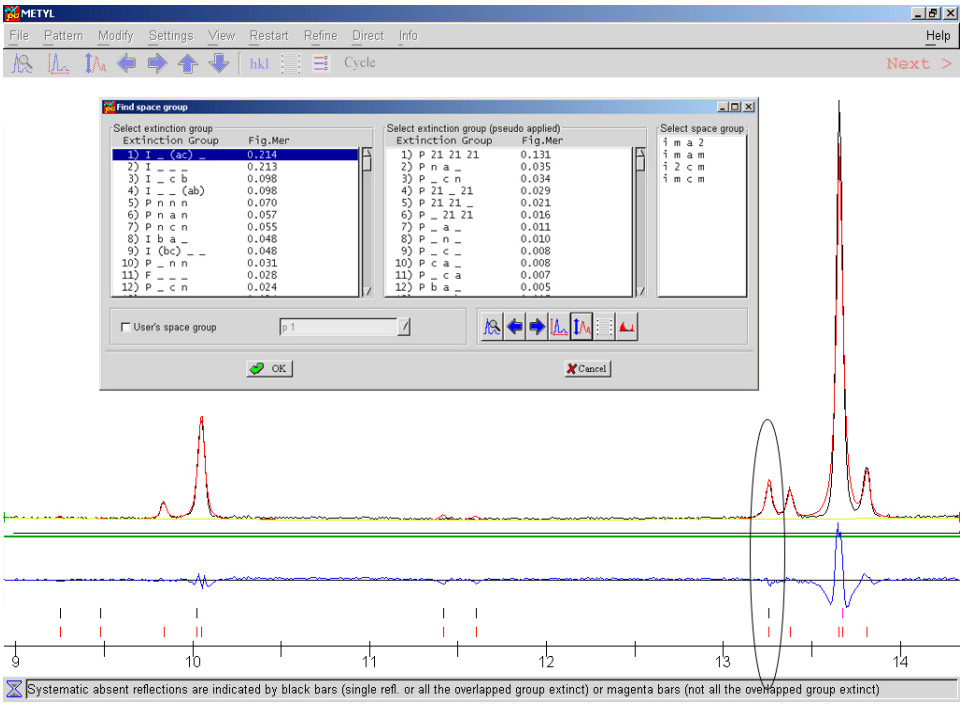


Figure 4.7 – Details of the diffraction pattern for METYL, at the final step of the Space-Group Determination Procedure. Tick marks at bottom show reflections for the *P 2/m 2/m 2/m* space group, while the upper tick marks show extinct reflections for the *I - (ac)* - extinction group (black tick marks correspond to single reflections or those belonging to a whole extinct cluster).

4.8 Conclusions

We have described a statistically based algorithm that is able to provide a probability value for each diffraction symbol compatible with the established lattice symmetry. The algorithm, which has been implemented in the program EXPO2004 and is fully automatic, employs statistical criteria, reduces the errors in the background definition, eliminates impurity peaks and offers to the

user a powerful graphical interface. In most cases the correct diffraction symbol has the largest probability value; in several cases (when the probability value associated with the correct diffraction symbol is much larger than the others) the diffraction symbol is unambiguously defined. For some test structures the correct diffraction symbol is not associated with the largest probability value; these cases have been analysed and it has been shown that the failures frequently result from the effects of pseudosymmetry. However, the graphical support helps the user to discard unreliable diffraction symbols even if estimated with large probability.

Conclusions

- A systematic study with a set of test structures has been carried out, in order to investigate the influence of the full pattern decomposition procedure on the *ab initio* structure solution. From this study we can conclude that choosing a certain peak-shape function (as well as increasing the number of refinement cycles in the intensity extraction process) can sometimes improve the fit of the observed powder diffraction pattern (which implies lower values of R_P) but, contrary to what it could seem, this improvement does not necessarily implies an enhancement on the decomposition efficiency and, subsequently, on the Direct methods procedure, which was confirmed by calculating the crystallographic residue R_F (i. e. not always a lower R_P value corresponds to a lower R_F value). This contradiction comes from the fact that we are treating the data without any structural information and, therefore, we have no restraint available which could tell us how to share the integrated intensity extracted from a cluster between the various reflections that form it. At this point, we have seen that the study of the reflection overlapping degree present on the powder data becomes a better consideration to assess, *a priori*, the efficiency of the decomposition process.

- The role played by the powder data collection step in the crystal structure solution pathway has been evaluated by using different sets of data obtained with both Fixed-Counting-Time (FCT) and Variable-Counting-Time (VCT) measurement schemes. The results coming out from processing the data by means of Direct Methods showed that the VCT scheme offers better efficiency for the phasing process (mainly for organic structures) in terms of average phase error and the correct location of the atoms at the end of the process. This arises from a better accuracy on the extraction of the intensities from the powder pattern (i.e. lower R_F values), as the VCT strategy improves the quality of the data just in the range of highest reflection density. For both FCT and VCT strategies, the fact of increasing the total measurement time brings on better results for the structure solution; but this improvement can not be maintained infinitely, as the information available from the diffractogram is restricted by the information lack that comes from the reflection overlapping. In certain cases this limitation can be overcome, although only in part, by excluding high-2θ data from the calculations (data which, instead of offering more information, increases the error on the intensities estimates).

- A fully automatic algorithm for the determination of the extinction group has been developed. It is based on a statistical study of the normalized integrated intensities that have been extracted from the powder diffractogram and it assigns a probability value to every

extinction symbol which is compatible with the actual lattice symmetry. After testing the algorithm with several test structures, the results have showed that in most of the cases the right extinction symbol is correctly defined as the most probable. The algorithm also provides analysis and correction of the probabilities due to the effects of pseudosymmetry. Apart from the statistical analysis, the algorithm is able to reduce the errors in the background modelling, to eliminate spurious peaks from the pattern and provides the user with a powerful graphical interface, which helps to easily discard wrong extinction groups that may be assigned a high probability.

References

Altomare, A., Cascarano, G., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G., Burla, M. C., Polidori, G. (1995a). *J. Appl. Cryst.*, **28**, 738-744.

Altomare, A., Burla, M. C., Cascarano, G., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G., Polidori, G. (1995b). *J. Appl. Cryst*. **28**, 842-846.

Altomare, A., Foadi, J., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G. (1996). *J. Appl. Cryst.* **29**, 674-681.

Altomare, A., Burla, M. C., Camalli, M., Carrozzini, B., Cascarano, G., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G., Polidori, G., Rizzi, R. (1999). *J. Appl. Cryst*. **32**, 339-340.

Altomare A., Giacovazzo C., Guagliardi A., Moliterni A. G. G., Rizzi R., Werner, P.-E. (2000a). *J. Appl. Cryst.* **33**, 1180-1186.

Altomare, A., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G., Rizzi, R. (2000b). *J. Appl. Cryst.* **33**, 1305-1310.

Altomare A., Giacovazzo C., Guagliardi A., Moliterni A. G. G., Rizzi R. (2001). *J. Appl. Cryst.* **34**, 704-709.

Altomare A., Caliandro R., Cuocci C., Giacovazzo C., Moliterni A.G.G., Rizzi R. (2003). *J. Appl. Cryst.* **36**, 906-913.

Altomare A., Caliandro R., Cuocci C., da Silva I., Giacovazzo C., Moliterni A. G. G., Rizzi R. (2004a). *J. Appl. Cryst*. **37**, 204-209.

Altomare A., Caliandro R., Camalli, M., Cuocci C., da Silva I., Giacovazzo C., Moliterni A. G. G., Spagna, R. (2004b). *J. Appl. Cryst*. **37**, 957-966.

Altomare, A., Caliandro, R., Camalli, M., Cuocci, C., Giacovazzo, C., Moliterni, A. G. G., Rizzi, R. (2004c). *J. Appl. Cryst.* **37** 1025-1028.

Altomare, A., Caliandro, R., Camalli, M., Cuocci, C., Giacovazzo, C., Moliterni, A. G. G., Rizzi, R., Spagna, R., Gonzalez-Platas, J. (2004d). *Z. Kristallogr.* **219**, 833-837.

Altomare, A., Cuocci, C., da Silva, I., Giacovazzo, C., Moiliterni, A. G. G, Rizzi, R. (2005). *Powder Diffraction* **20**, 327-333.

Amoureux, J. P., & Bee, M. (1980) *J. Phys. C: Solid State Phys.* **13**, 3577-3583.

Andreev, Y. G., MacGlashan, G. S., Bruce, P. G. (1997). *Phys. Rev*. B**55**, 12011-12017.

Andreev Y.G. & Bruce P.G. (1998) *J. Chem Soc. Dalton Trans*., 4071-4080.

Antic, B., Mitric, M., Rodic, D. (1995) *J. Magn. Magn. Mater.* **145**, 349-356.

Arumugam, S., Glidewell, C., Harris, K. D. M. (1992). *J. Chem. Soc., Chem. Commun.*, 724-726.

Baerlocher C., McCusker L. B., Chiappetta R. (1994). *Microporous Mater.* **2,** 269-280.

Banerjee S., Mukherjee A. K., Neumann M. A., Louer D. (2002) *Acta Cryst*. A**58** (Supplement), C264.

Bataille, T., Auffrédic, J.-P., Louër, D. (1999). *Chem. Mater*. **11**(6), 1559-1567.

Benet-Buchholz, J., Haumann, T., Boese, R. (1998) *J. Chem. Soc., Chem. Commun*., 2003-2004.

Boultif, A. & Louër, D. (1991). *J. Appl. Cryst.* **24**, 987-993.

Brauer, G. & Gradinger, H. (1954) *Z. Anorg. Allg. Chem.* **277**, 89-95.

Brown A. & Edmonds J. W. (1980) *Adv. X-ray Anal*. **23**, 361.

Brunelli, M., Fitch, A. N., Mora, A. J. (2002). *J. Solid State Chem.* **163**, 253-258.

Brunton, G. (1973). *Materials Research Bulletin* **8**, 271-274.

Burla, M. C., Camalli, M., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G., Spagna, R. (1999) *Acta Cryst*. A**55**, 991-999.

Camblor, M. A., Corma, A., Díaz-Cabañas, M.-J., Baerlocher, C. (1998). *J. Phys. Chem.* B**102**, 44-51.

Cernik, R. J., Cheetham, A. K., Prout, C. K., Watkin, D. J., Wilkinson, A. P., Willis, B. T. M. (1991). *J. Appl. Cryst.* **24**, 222-226.

Chernyshev, V. V. & Schenk, H. (1998). *Z. Kristallogr.* **213**, 1-3.

Chernyshev, V. V., Fitch, A. N., Sonneveld, E. J., Kurbakov, A. I., Makarov, V. A., Tafenko, V. A. (1999). *Acta Cryst*. B**55**, 554-562.

Christelle, S. C., Michaud, F., Léger, J. M., Deslandes, E. (2003). *Carbohydrate Research* **338**, 2413-2416.

Christensen, A. N. (1992). *Acta Chem. Scandin.* **46**, 240-243.

Christensen, A. N. (1994). *Z. Kristallog.* **209**, 7-13.

Christensen, A. N., Hazell, R. G., Hewat, A. W., O'Reilly, P. J. (1991). *Acta Chem. Scandin.* **45**, 469-473.

Christensen, A. N., Nielsen, M., O'Reilly, K. P. J., Wroblewski, T. (1992). *Acta Chem. Scandin.* **46**, 224-230.

Christensen, A. N., Hazell, R. G., Lehmann, M. S. Nielsen, M. (1993). *Acta Chem. Scand.* **47**, 753-756.

Clarke, S., Cockcroft, J. K., Fitch, A. N. (1993) *Z. Kristallog.* **206**, 87-95.

Cochran, W. (1955). *Acta Cryst*. **8**, 473-478.

Cockcroft, J. K. & Fitch, A. N. (1990). *Z. Kristallog.* **193**, 1-19.

Cockcroft, J. K. & Fitch, A. N. (1991). *Z. Kristallog.* **197**, 121-130.

Courbion, G. & Ferey, G. (1988) *J. Solid State Chem.* **76**, 426-431.

da Silva, I., González-Silgo, C., González-Platas, J., Rodríguez-Carvajal, J., Martínez-Sarrión, M. L., Mestres, L. (2005). *J. Solid State Chem.* **178**, 1601-1608.

Dadachov, M. S. & Le Bail, A. (1997). *Eur. J. Solid State Inorg. Chem*. **34**, 381-390.

David W. I. F., Shankland K., Shankland N. (1998). *Chem. Commun.* 931-932.

David W. I. F (1999). *J. Appl. Cryst.* **32**, 654-663.

David W. I. F. & Sivia, D. S. (2002). *In Structure Determination from Powder Diffraction Data*, Edited by David, W. I. F., Shankland, K., McCusker, L. B. & Baerlocher, Ch. Oxford Univ. Press, pp.136-161.

de Wolff, P. M. (1968). *J Appl. Cryst*. **1**, 108-113.

Derissen, J. L., Endeman, H. J., Peerdeman, A. F. (1968). *Acta Cryst.* B**24**, 1349-1354.

Eliseev, A. A., Efremmov, V. A., Kuzmicheva, G. M., Konovalova, E. S., Lazorenko, V. I., Paterno, Y. B., Khlyustova, S. Y. (1986). *Kristallografiya* **31**, 803-805.

Estermann, M. A., McCusker, L. B., Baerlocher, C. (1992). *J. Appl. Cryst.* **25**, 539-543.

Evans, H. T. jr., Gatehouse, B. M., Leverett, P. (1975). *Journal of the Chemical Society*, Dalton Transactions, **6**, 505-514.

Favre-Nicolin, V. & Cerný, R. (2004). *Z. Kristallogr*. **219**, 847-856.

Fitch, A. N. & Cole, M. (1991). *Mat. Res. Bull*. **26**, 407-414.

Florence, A. J., Shankland, N., Shankland, K., David, W. I. F., Pidcock, E., Xu, X., Johnston, A., Kennedy, A. R., Cox, P. J., Evans, J. S. O., Steele, G., Cosgrove, S. D., Frampton, C. S. (2005). *J. Appl. Cryst.* **38**, 249-259.

Giacovazzo, C. (1980). *Direct Methods in Crystallography*. Academic London.

González-Silgo, C., González-Platas, J., Ruiz-Pérez, C., López, T., Torres, M. (1999). *Acta Cryst.* C**55**, 740-742.

Hammond R. B., Roberts K. J., Docherty R., Edmondson M. (1997) *J. Phys. Chem.* B**101**, 6532.

Harker, D. & Kasper, J. S. (1948). *Acta Cryst*. **1**, 70.

Harris K. D. M., Tremayne M., Lightfoot P., Bruce P. G. (1994) *J. Am. Chem. Soc*. **116**, 3543.

Harris K. D. M., Johnston R. L., Kariuki B. M. (1998). *Acta Cryst*. A**54**, 632-645.

Harris K. D. M., Treymayne M., Kariuki B. M. (2001). *Angew. Chem. Int. Ed*. **40**, 1626-1651.

Harris, K. D., Habershon, S., Cheung, E. Y., Johnston, R. L., (2004). *Z. Kristallogr*. **219**, 838-846.

Hauptman, H. & Karle, J. (1953). *The solution of the phase problem. I. The centrosymmetric crystal*, ACA Monograph, No. 3 Polycrystal Book Service, New York.

Hibble S. J., Cheetham A. K., Bogle A. R. L., Wakerley H. R., Cox D. E. (1988) *J. Am. Chem. Soc.* **110**, 3295-3296.

Huang, T. C. (1988). *Aust. J. Phys*. **41**, 201-212.

Huang, T. C. & Parrish, W. (1975). *Appl. Phys. Lett*. **27**, 123-124.

Jansen, J., Peschar, R., Schenk, H. (1992). *J. Appl. Cryst.* **25**, 231-236.

Jouanneaux, A., Verbaere, A., Piffard, Y., Fitch, A. N., Kinoshita, M. (1991a). *Eur. J. Solid State Inorg. Chem.* **28**, 683-699.

Jouanneaux, A., Verbaere, A., Guyomard, D., Piffard, Y., Oyetola, S., Fitch, A. N. (1991b). *Eur. J. Solid State Inorg. Chem.* **28**, 755-765.

Jouanneaux, A., Fitch, A. N., Cockcroft, J. K. (1992). *Mol. Phys.* **71**, 45-50.

Kariuki, B. M., Serrano Gonzalez, H., Johnston, R. L., Harris, K. D. M. (1997). *Chem. Phys. Lett.* **280**, 189-195.

Karle, J. & Hauptman, H. (1956). *Acta Cryst*. **9**, 635-651.

Langford, J. I. & Louër, D. (1996). *Rep. Prog. Phys.* **59**, 131-234.

Lasocha, W. & Schenk, H. (1997). *J. Appl. Cryst.* **30**, 909-914.

Lasocha, W., Milart, P., Rafalska-Lasocha, A., Schenk, H. (2001). *Z. Kristallogr*. **216**, 117-121.

Le Bail (2004). *Powder Diffraction* **19**, 249-254.

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

Louër, D. (1998). *Acta Cryst.* A**54**, 922-933.

Madsen, I. C. & Hill, R. J. (1994). *J. Appl. Cryst.* **27**, 385-392.

Markvardsen, A. J., David, W. I. F., Johnson, J. C., Shankland, K. (2001). *Acta Cryst*. A**57,** 47-54.

Masciocchi N., Moret M., Cairati P., Ragaini, F., Sironi, A., (1993) *J. Chem. Soc., Dalton Trans.* 471-475.

Masciocchi N., Moret M., Cairati P., Sironi A., Ardizzoia G. A., La Monica G., (1994) *J. Am. Chem. Soc.* **116**, 7668-7676.

McCusker, L. (1988). *J. Appl. Cryst.* **21**, 305-310.

McCusker, L. B. (1993). *Mat. Sci. Forum* **133-136**, 423-434.

McCusker, L. B., Baerlocher Ch., Jahn E., Bülow M. (1991). *Zeolites* 308-313.

McCusker, L. B., Grosse-Kunstleve, R. W., Baerlocher, C., Yoshikawa, M., Davis, M. E. (1996). *Microporous Mater.* **6**, 295-309.

Meden, A., Grosse-Kunstleve, R. W., Baerlocher, C., McCusker, L. B. (1997). *Z. Kristallog.* **212**, 801-807.

Norby, P., Christensen, A. N., Fjellvåg, H., Lehmann, M. S., Nielsen, M. (1991). *J. Solid State Chem.* **94**, 281-293.

Nowell H., Attfield J. P., Cole J. C., Cox P. J., Shankland K., Maginn S. J., Motherwell W. D. S. (2002). *New J. Chem*. **26**, 469-472.

Pawley, G. A. (1981). *J. Appl. Cryst.* **14**, 357-361.

Perez, C., Espinola, C. G., Foces-Foces, C., Nunez-Coello, P., Carrasco, H., Martin, J. D. (2000) *Org. Lett.* **2(**9), 1185-1188.

Plévert J., Yamamoto K., Chiari G., Tatsumi T. (1999) *J. Phys. Chem*. B**103**, 8647-8649.

Reck G., Kretshmer R. G., Kutschabsky L., Pritzkow W. (1998). *Acta Cryst*. A**44**, 417-421.

Rietveld, H. M. (1969). *J. Appl. Cryst.* **2**, 65-71.

Ruiz-Pérez, C., Hernández-Molina, M., González-Silgo, C., López, T., Yanes, C., Solans, X. (1996). *Acta Cryst.* C**52**, 2473-2475.

Runge, C. (1917). *Phys. Z.* **18**, 509-515.

Savitzky, A. & Golay, M. J. E. (1964). *Anal. Chem.* **36**, No. 8, 1627-1639.

Sayre, D. (1952). *Acta Cryst*. **5**, 60-65.

Shankland, K., David, W. I. F., Csoka, T. (1997). *Z. Kristallogr.* **212**, 550-552.

Shankland, K., David W. I. F., Csoka, T., McBride L. (1998) *Int J. Pharm*. **165**, 117-126.

Shankland, K., McBride, L., David, W. I. F., Shankland, N., Steele, G. (2002). *J. Appl. Cryst.* **35**, 443-454.

Simon-Colin, C., Michaud, F., Léger, J.-M., Deslandes, E. (2003). *Carbohydrate Research* **338,** 2413-2416.

Sivia, D. S. & David, W. I. F. (1994). *Acta Cryst.* A**50**, 703-714.

Smrčok L'., Kopelhuber B., Shankland K., David W. I. F., Tunega D., Resel R. (2001). *Z. Kristall.* **216**, 63-66.

Sonneveld, E. J. & Visser, J. W. (1975). *J. Appl. Cryst*. **8**, 1-7.

Structure Determination by Powder Diffractometry Round Robin (SDPDRR) (1998) http://www.cristal.org/SDPDRR/index.html.

Structure Determination from Powder Diffraction Data. Edited by David, W. I. F., Shankland, K., McCusker, L. B., Baerlocher, Ch. (2002). Oxford University Press.

Tremayne, M., Kariuki, B. M., Harris, K. D. M. (1997). *Angew. Chem.* **109**, 788.

Turner, G. W., Tedesco, E., Harris, K. D. M., Johnston, R. L., Kariuki, B. M. (2000). *Chem. Phys. Lett.* **321**, 183-190.

Visser, J. W. (1969). *J. Appl. Cryt.* **2**, 89-95.

Werner, P.-E., Eriksson, L., Westdahl, M. (1985). *J. Appl. Cryst.* **18**, 367-370.

Werner, P.-E., Moustiakimov, M., Marinder, B.-O., Knight, K. S. (1997). *Z. Kristallog.* **212**, 665-670.

Weiss, E., Corbelin, S., Cockcroft, J. K., Fitch, A. N. (1990). *Chem. Ber.* **123**, 1629-1634.

Williams, J. H., Cockcroft, J. K., Fitch, A. N. (1992). *Angew. Chem. Int. Ed. Engl.* **31**, No.12, 1655-1657.

Wilson, A. J. C. (1942). *Nature* **150**, 151.

Young, R. A. & Wiles D. B. (1982). *J. Appl. Cryst*. **15,** 430-438.

Zah-Letho, J. J., Jouanneaux, A., Fitch, A. N., Verbaere, A., Tournoux, M. (1992). *Eur. J. Solid State Inorg. Chem.* **29**, 1309-1320.

Appendix A.

Test structures references

	Code name	Reference
1	AGPZ	Masciocchi *et al*., (1994)
2	ALDS, ALDX	SDPDRR (1998)
3	ALPHA	Harris *et al*., (2001)
4	AND1, AND1N, AND2, AND2N	Chernyshev *et al*. (1999)
5	BACO	Christensen (1992)
6	BAMO	Werner *et al.* (1997)
7	CAINE	Nowell *et al*. (2002)
8	CIME	Cernik *et al.* (1991)
9	COPPER	Banerjee *et al*. (2002)
10	CROX	Norby *et al.* (1991)
11	DADA	Dadachov & Le Bail (1997)
12	DFQP	Smrčok *et al*. (2001)
13	EMT	Baerlocher *et al*. (1994)
14	[$GaPO_4$]-ZON	Meden *et al*. (1997)
15	IBUPS	Shankland *et al*. (1998)
16	LAMO	Hibble *et al.* (1988)
17	LASI	Christensen, (1994)
18	LEV	McCusker (1993)
19	MCM-22	Camblor *et al*. (1998)

20	NBPO	Zah-Letho *et al.* (1992)
21	NIZR	Jouanneaux *et al.* (1991a)
22	OTHYM	Kariuki *et al*. (1997)
23	SAPO-40	Estermann *et al*. (1992)
24	SBPO	Jouanneaux *et al.* (1991b)
25	SGT	McCusker (1988)
26	SULPH	Cockcroft & Fitch (1990)
27	UTM1	Plévert *et al*. (1999)
28	VFI	McCusker *et al*., (1991)
29	YONO	Christensen *et al*., (1992)
30	YURI	Andreev *et al*. (1997)
31	AMMO	Evans *et al*., (1972)
32	FR3S	Christelle *et al*. (2003)
33	ASPAR	Derissen *et al.* (1968)
34	BA-TART	González-Silgo *et al*. (1999)
35	MN-TART	Ruiz-Pérez *et al*. (1996)
36	LOPEZITE	Brunton, G. (1973)
37	ADAM	Amoureux & Bee (1980)
38	BENZ	Williams *et al*. (1992)
39	CeO_2	Brauer & Gradinger (1954)
40	CF3BR	Jouanneaux *et al*. (1992)
41	CFCL	Cockcroft & Fitch (1991)
42	CFI	Clarke *et al.* (1993)
43	Cs_2BeF_4, K_2BeF_4, Rb_2BeF_4	da Silva *et al*., (2005)
44	CUPZ	Masciocchi *et al*., (1994)
45	ECH1	Bataille *et al.*, (1999)
46	F1A	Perez *et al.* (2000)
47	FAMO	Shankland *et al*. (2002)
48	FORMYLUREA	Arumugam *et al*. (1992)

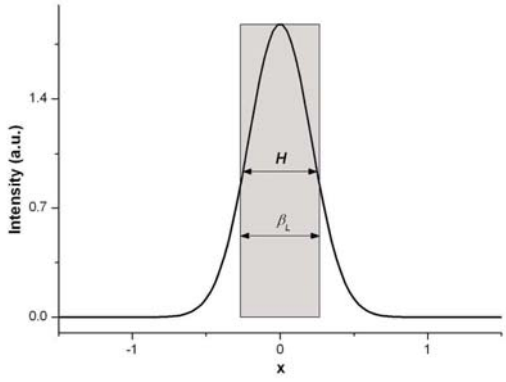
49	KUOS	Fitch & Cole (1991)
50	LaB_6	Eliseev *et al.* (1986)
51	LASO	Lasocha & Schenk (1997)
52	LASOKA	Lasocha *et al*. (2001)
53	MES	Christensen *et al*. (1993)
54	METYL	Weiss *et al*. (1990)
55	NAC	Courbion & Ferey (1988)
56	NORB2	Benet-Buchholz *et al*. (1998)
57	PBS	Christensen *et al*. (1991)
58	RUCO	Masciocchi *et al.*, (1993)
59	SCAMPHOR1	Brunelli *et al*., (2002)
60	VNI	McCusker *et al*. (1996)
61	Y_2O_3	Antic *et al.* (1995)

Appendix B.

Peak-shape functions

Here we describe the peak-shape functions most widely used for the modelling of powder diffraction data, in both pattern decomposition and Rietveld methods.

- **Gaussian**



The Gaussian peak-shape function is defined as

$$\Omega_0(x) = G(x) = a_G \exp\left(-b_G x^2\right)$$

where

$$a_G = \frac{2}{H}\sqrt{\frac{\ln 2}{\pi}} \qquad \text{and} \qquad b_G = \frac{4\ln 2}{H^2}$$

The integral breadth of this function is

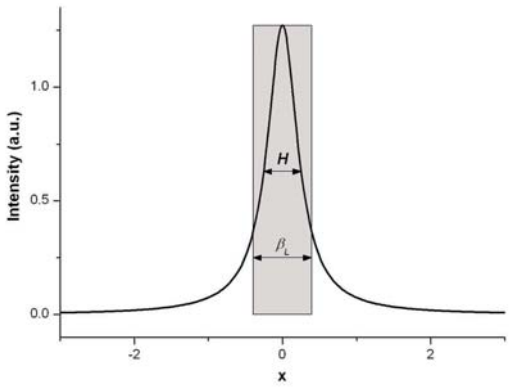
$$\beta_G = \frac{1}{a_G} = \frac{H}{2}\sqrt{\frac{\pi}{\ln 2}}$$

And the Full-Width-at-Half-Maximum (H) is

$$H^2 = \left(U + D_{ST}^2\right)\tan^2\theta + V\tan\theta + W + \frac{I_G}{\cos^2\theta}$$

where U, V, W, D_{ST}^2 and I_G are adjustable parameters on the least squares procedure.

- **<u>Lorentzian</u>**



The Gaussian peak-shape function is defined as

$$\Omega_1(x) = L(x) = \frac{a_L}{1 + b_L x^2}$$

where

$$a_L = \frac{2}{\pi H} \qquad \text{and} \qquad b_L = \frac{4}{H^2}$$

The integral breadth of this function is

$$\beta_L = \frac{1}{a_L} = \frac{\pi H}{2}$$

The Full-Width-at-Half-Maximum is the same as for the Gaussian case.

- **<u>Pearson VII</u>**

The expression for the Pearson-VII (PVII) function is

$$\Omega_6(x) = PVII(x) = \frac{a_{VII}}{\left(1 + b_{VII} x^2\right)^m}$$

where

$$a_{VII} = \frac{\Gamma(m)}{\Gamma(m) - 1/2} \frac{2\sqrt{2^{1/m} - 1}}{\sqrt{\pi} H} \ , \qquad b_{VII} = \frac{4\left(2^{1/m} - 1\right)}{H^2}$$

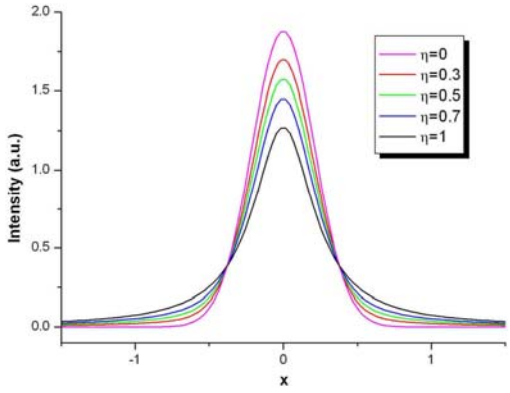
and

$$m = m_0 + 100 \frac{X}{2\theta} + 10000 \frac{Y}{(2\theta)^2} \ .$$

The Full-Width-at-Half-Maximum expression is the same as for the Gaussian and the Lorentzian and *m*, *X* and *Y* are also adjustable parameters.

When $m = 1$, the PVII is equivalent to a Lorentzian but as *m* increases the peak becomes more Gaussian in character.

- **Pseudo-Voigt**



The pseudo-Voigt (pV) peak-shape function is defined as

$$\Omega_5(x) = pV(x) = \eta L(x) + (1-\eta)G(x)$$

with

$$\eta = \eta_0 + X \cdot 2\theta$$

where

$$0 \le \eta \le 1$$

The $pV(x)$ function is a linear combination of a Lorentzian (L) and a Gaussian (G) with the same H, so there are two parameters characterising the peak shape: $pV(x)=pV(x,\eta,H)$. If $L(x)$ and $G(x)$ are normalised, $pV(x)$ is also normalised. The integral breadth of a normalised pseudo-Voigt function is just the inverse of the maximum value. The Full-Width-at-Half-Maximum is the same for $L(x)$, $G(x)$ and $pV(x)$.

Resúmenes en español

Introducción

Hoy en día, una de las técnicas experimentales más importantes disponibles para el estudio del estado sólido es la difracción de polvo policristalino. Desde sus comienzos, el método de difracción en polvo ha sido usado, principalmente, para análisis y estudio cuantitativo de fases, la determinación precisa de parámetros de celda y el estudio de imperfecciones estructurales mediante el análisis del ensanchamiento de picos de difracción.

Durante las últimas décadas, los importantes avances en la instrumentación usada (tanto para fuentes de rayos-X en laboratorios convencionales, como para radiación de neutrones o sincrotrón) y el gran aumento en la potencia de cálculo de los ordenadores han permitido la aplicación de los diversos avances teóricos que se han desarrollado, enfocados a obtener la información estructural a partir del patrón de difracción. Desafortunadamente, aún con la ayuda de dichos avances, la resolución estructural mediante difracción en polvo continúa siendo mucho más difícil que mediante difracción en monocristal, debido fundamentalmente al hecho de que la información cristalográfica tridimensional del compuesto en estudio se encuentra proyectada en un patrón de difracción monodimensional. Este hecho introduce una incertidumbre en los datos extraídos a partir del difractograma, creando

problemas en todos los pasos envueltos en el proceso de determinación de la estructura cristalina.

El proceso de resolución estructural *ab initio*, a partir de datos de difracción en polvo, se puede resumir en los siguientes pasos:

- Medida de datos de difracción:

El modo en el que se llevará a cabo el experimento dependerá del tipo de información que queramos obtener. Por ejemplo, si queremos obtener sólo los parámetros de celda nos bastará con medir la zona de bajos ángulos y obtener las primeras reflexiones; si queremos datos de alta resolución puede que necesitemos usar radiación sincrotrón; si trabajamos con poca cantidad de muestra o ésta presenta efectos de transparencia, lo más adecuado será usar una geometría de transmisión.

- Indexación:

Las posiciones angulares de los picos de difracción obtenidos en el difractograma son usadas en el proceso de indexación, cuyo objetivo es el de determinar la simetría del cristal y los parámetros de la celda unidad, así como asignar los índices *hkl* a cada reflexión. Actualmente, el principal problema que podemos encontrar en este proceso es el de la presencia de picos impureza, que no correspondan al compuesto cuya estructura queremos determinar; si ello no sucede, el éxito del proceso está prácticamente garantizado, siempre que el error en las posiciones de los picos extraídas sea menor de 0.03º en 2θ, aproximadamente.

- Determinación del grupo espacial:

Una vez indexado el difractograma se debe determinar el grupo espacial, proceso que resulta muchas veces difícil y ambiguo, ya que normalmente se lleva a cabo mediante una inspección visual del patrón de difracción, para comprobar qué condiciones de extinción de reflexiones se cumplen y así deducir el grupo de extinción. Aunque para algunos grupos de extinción sólo corresponde un grupo espacial, normalmente necesitaremos alguna información extra (volumen aproximado de la molécula, por ejemplo) para encontrar el grupo espacial correcto entre todos aquellos que compartan las mismas extinciones sistemáticas.

- Descomposición del patrón de difracción:

El objetivo en este paso es el de asignar una intensidad integrada a cada reflexión presente en el difractograma (aquellas que sean permitidas por las condiciones de extinción del grupo espacial). Este proceso se realiza mediante el ajuste, normalmente mediante un refinamiento de mínimos cuadrados, del difractograma experimental con un modelo teórico. Básicamente, en dicho modelo, la intensidad observada en un punto cualquiera del difractograma es calculada en función de las intensidades integradas de las reflexiones del patrón de difracción (en realidad, sólo de aquellas reflexiones dentro de un cierto rango angular, en torno al punto) y una contribución del fondo. El principal problema que se encuentra en este proceso es el del solapamiento de los picos de difracción, el cual provoca que la asignación de las intensidades no sea una tarea trivial, ya que diversos esquemas de repartición de intensidades entre las reflexiones que forman un grupo solapado pueden conducir a ajustes de perfil análogos. Ello, unido al hecho de que el solapamiento hace más difícil

la estimación del fondo (no existen zonas de fondo entre los picos), conduce a estimaciones erróneas de las intensidades integradas en un proceso que resulta de gran importancia en la resolución estructural, y más aún si se usan los Métodos Directos.

- Solución al problema de las fases:

La distribución de la densidad electrónica (y, por tanto, la estructura cristalina) está directamente relacionada con los factores de estructura de las reflexiones, que son expresiones complejas. Por otra parte, la intensidad difractada es proporcional al cuadrado de los factores de estructura; así, el proceso de extracción ofrece una estimación de los módulos de dichos factores de estructura, pero no de sus fases. Esta estimación se lleva a cabo, de forma probabilística, mediante los Métodos Directos. El proceso se compone de los siguientes pasos básicos:

1. Modificación de los factores de estructura derivados de la descomposición del difractograma, usando la información de la composición química del compuesto y la simetría del grupo espacial, para la obtención de los llamados factores de estructura normalizados.
2. Cálculo de determinadas combinaciones lineales entre las fases de las reflexiones, que resultan ser independientes de la elección del origen (tripletes y cuartetes).
3. Definición de un conjunto inicial de fases, del cual se derivarán todas las restantes mediante un proceso en cadena usando las relaciones calculadas en el paso anterior. Este paso se repite varias veces, modificando las fases del conjunto inicial, con lo cual al final se obtienen varias soluciones distintas.

4. Cálculo de las figuras de mérito correspondientes a cada solución, para estimar *a priori* la bondad de cada conjunto de fases y determinar la posible solución.
5. A partir del conjunto con mejor figura de mérito se calcula el mapa de densidad electrónica. Los picos presentes en este mapa son analizados en función de sus intensidades y en términos de distancias y ángulos, para así asignar un determinado átomo a su posición.
6. En caso de que el modelo estructural no esté completo, se debe completar y refinar. Esto se lleva a cabo, normalmente, mediante métodos de mínimos cuadrados.

Objetivos de esta Tesis

El trabajo desarrollado en esta Tesis está enfocado, principalmente, en el estudio de los procesos que preceden a la aplicación de los Métodos Directos y está estructurada de la siguiente manera:

En el capítulo 2 se discute el papel que desarrolla el proceso de extracción de intensidades en la resolución estructural, así como la influencia que diversos parámetros como la función de pico usada, el número de ciclos llevados a cabo en la descomposición o el grado de solapamiento de picos, tienen sobre la eficiencia del proceso de asignación de fases a las reflexiones.

El capítulo 3 está enfocado en estudiar la influencia que tiene el modo de realizar la toma de datos de difracción (mediante diferentes estrategias o tiempos de medida) sobre el posterior proceso de asignación de fases.

Finalmente, en el capítulo 4 se expone el desarrollo de un nuevo algoritmo para la determinación del grupo de extinción. Dicho algoritmo, probado con un gran número de estructuras test, se basa en un análisis estadístico de las intensidades extraídas del patrón de difracción.

Capítulo 2.

Descomposición del difractograma de polvo para la asignación de fases de las reflexiones por Métodos Directos.

Introducción

Para este estudio hemos usado el método *clásico* de resolución estructural de dos etapas (extracción de intensidades, seguido de aplicación de los Métodos Directos) para analizar la influencia que tiene la descomposición del difractograma de polvo en el proceso de asignación de fases.

El objetivo es estudiar el papel desarrollado por el ajuste del perfil, las distintas funciones de forma de pico y otros parámetros en el procedimiento Le Bail para la extracción de intensidades y sus efectos en dicho proceso de asignación de fases.

Metodología y resultados

El procedimiento de descomposición del difractograma fue analizado mediante el uso de 31 estructuras test (ver Tabla 2.1, pág. 28), utilizando el programa EXPO2004.

Los datos de todas las estructuras fueron tratados con EXTRA, un algoritmo para la descomposición del patrón de difracción incluido en el programa EXPO2004, usando dos funciones de perfil diferentes (pseudo-Voigt y Pearson VII) para la extracción de intensidades (ver el Apéndice B para una descripción de las funciones de perfil). Los correspondientes valores obtenidos para los parámetros R_P y R_F (ver fig. 2.1, pág. 30) muestran que con la función Pearson VII se alcanzan, en general, mejores resultados para el ajuste de perfil (R_P menor), aunque la diferencia media con los resultados correspondientes a la función pseudo-Voigt es pequeña. En cuanto al parámetro R_F, la diferencia media entre las dos funciones es insignificante y para la mayoría de las estructuras se obtienen resultados comparables. Además, para algunas estructuras, la obtención de valores de R_P más bajos con una función de perfil no corresponden a valores más bajos de R_F. Visto que la elección de la función de forma de pico no resulta un paso crucial, en los cálculos siguientes se usó solamente la función Pearson VII. En las figuras 2.2 y 2.3 (pág. 32 y 33, respectivamente) se muestran los valores obtenidos para los dos parámetros R_P y R_F, tanto para el conjunto total de estructuras al final del refinamiento del difractograma, como para dos estructuras en particular en función del ciclo de refinamiento. Como se puede observar, no existe una clara correlación entre dichos factores de acuerdo.

Otro parámetro que resulta de interés, calculado en los test, ha sido el porcentaje de observaciones independientes presente en el difractograma

(PERC). En la figura 2.4 (pág. 35) se puede observar cómo en este caso el parámetro PERC presenta una alta correlación con el parámetro R_F, lo cual sugiere que se puede usar como un indicador de la eficiencia del proceso de descomposición.

Los factores de estructura previamente obtenidos con el método de extracción de Le Bail fueron usados para la aplicación de los Métodos Directos, los cuales, para cada una de las estructuras test, generaron un conjunto de 10 soluciones distintas, de entre las cuales la mejor (la de error medio de fase más bajo) fue elegida para cada conjunto; cabe señalar que estos errores medios de las fases los podemos determinar puesto que las estructuras son conocidas y, por lo tanto, podemos calcular las fases correctas para compararlas con las obtenidas por los Métodos Directos. El correspondiente valor del error de fase cometido para cada estructura se muestra en la tabla 2.2 (pág. 37); éstos sugieren que para un considerable número de estructuras no podrían ser resueltas de modo automático, debido a los grandes valores de error de fase producidos por la aplicación de los Métodos Directos.

Dado que el parámetro PERC presenta una alta correlación con la eficiencia en la extracción de intensidades (es decir, con R_F), es de esperar que también la presente con la eficiencia de los Métodos Directos (es decir, con el error de fase medio), lo cual, efectivamente, se puede comprobar en la figura 2.5 (pág. 39). Esta figura muestra que el error de fase medio puede ser estimado *a priori* gracias al valor obtenido en el cálculo de la expresión $\sqrt{\mathrm{NAT}}_{\mathrm{eq}}(DRES/PERC)$, la cual contiene información sobre los tipos de átomos presentes en la estructura, la resolución hasta la cual se ha medido los datos de difracción y el porcentaje de observaciones independientes, previamente calculado.

Conclusiones

La influencia que, en la resolución de estructuras *ab initio*, tiene el procedimiento de extracción de intensidades ha sido estudiado mediante el uso de un conjunto de estructuras test. A partir de este estudio podemos concluir que la elección de una determinada función de perfil para modelar la forma de los picos (así como el hecho de incrementar el número total de ciclos de refinamiento llevados a cabo en el proceso de extracción) a veces puede mejorar el ajuste del patrón de difracción observado (lo cual implica una obtención de valores más bajos para el parámetro de acuerdo R_P) pero, al contrario de lo que puede parecer, esta mejora no implica necesariamente un incremento en la eficiencia de la descomposición y, consecuentemente, de los Métodos Directos, lo cual fue confirmado con el cálculo del índice de acuerdo cristalográfico R_F (es decir, no siempre un valor más bajo de R_P corresponde a un valor más bajo de R_F). Esta contradicción proviene del hecho que estamos tratando los datos sin usar ninguna información estructural y, por lo tanto, no tenemos a disposición ninguna condición de contorno que nos pueda indicar de qué forma repartir la intensidad integrada que ha sido extraída de un conjunto de reflexiones en solapamiento entre dichas reflexiones que forman el grupo, lo que provoca cometer errores (a menudo muy significativos) en la estimación de este conjunto de intensidades. Estos errores se propagarán a los siguientes cálculos llevados a cabo por los Métodos Directos, teniendo graves consecuencias (a veces, incluso ligeros errores en un pequeño conjunto de factores de estructura puede marcar la diferencia entre resolver o no la estructura del cristal). Así, llegados a este punto, hemos comprobado que el estudio del grado de solapamiento de picos presente en el difractograma resulta un mejor parámetro para evaluar, *a priori*, la eficiencia del proceso de descomposición de intensidades.

Capítulo 3.

Medida de datos de difracción en polvo mediante la técnica de tiempo variable.

Introducción

Para la resolución de estructuras cristalinas en polvo, particularmente cuando trabajamos con Métodos Directos, necesitamos datos de alta calidad. La técnica de medida en tiempo variable (*variable-counting-time*, VCT) fue introducida por Madsen y Hill en 1994 para el análisis Rietveld, mostrando que ésta ofrece diversas ventajas frente a la técnica clásica de medida en tiempo fijo (*fixed-counting-time*, FCT), como por ejemplo un mejor ajuste del perfil, la determinación de las posiciones atómicas con mayor precisión o un análisis cuantitativo más exacto.

El trabajo llevado a cabo en este capítulo pretende completar el estudio de las ventajas de la técnica VCT, esta vez para la extracción de intensidades a partir

del difractograma y en el proceso de asignación de fases a cada una de las reflexiones. Para ello hemos usado seis estructuras test (ver tabla 3.1, pág. 44).

Metodología y resultados

Todas las muestras fueron medidas en un difractómetro convencional (Panalytical X'Pert Pro), usando tanto el protocolo FCT como el VCT; para éste último se usaron curvas de tiempos análogas a la de la figura 3.1 (pág. 46), calculadas con el programa *XRDTIME*, de Madsen y Hill. Diferentes ciclos de medida se obtuvieron tras sumar, progresivamente, los difractogramas obtenidos para diferentes repeticiones de medidas de una duración de dos horas, aproximadamente, para ambos protocolos.

Los datos experimentales fueron introducidos en el programa EXPO2004 para la extracción de las intensidades integradas. En el caso FCT el proceso fue totalmente automático, mientras que para el caso VCT se necesitó de un tratamiento previo de los datos para solventar el problema de los "escalones" presentes en los difractogramas VCT , lo cual se realizó con la ayuda de la curva de tiempos (fig. 3.3, pág. 50). Una vez modificado el difractograma VCT, se efectúa el proceso de extracción de manera automática, de igual manera que en el caso FCT, con la única salvedad de que el conjunto de intensidades obtenido finalmente es normalizado con respecto al tiempo de medida, en función de la posición angular de las reflexiones.

El cálculo del parámetro R_F muestra que la eficiencia del procedimiento de descomposición del difractograma, para ambos protocolos de medida, aumenta en función del ciclo de medida, además de ser sistemáticamente mayor para el caso VCT (fig. 3.4, pág.52), debido a que con este esquema se consigue una

mejora de la estadística de cuentas a alto ángulo, que es justo la zona donde hay una mayor concentración de reflexiones.

Por otra parte, el cálculo del parámetro PERC muestra que, efectivamente, en las medidas VCT la cantidad de información ofrecida por el difractograma es mayor que para el caso FCT, a excepción de los primeros ciclos, para algunas estructuras (fig.3.5, pág. 54).

El cálculo de la correlación entre los difractogramas FCT y VCT para el último ciclo de medida, en las diferentes zonas angulares fijadas por la curva de tiempos (fig. 3.6, pág. 55), resulta ser de utilidad para estimar la mejoría en la estadística de los datos mediante el uso del esquema VCT frente al FCT y, consecuentemente, en la eficacia de los Métodos Directos. Esto último se comprobó mediante el cálculo del error de fase medio para la mejor solución ofrecida por el programa EXPO2004 (aquella con error más bajo) para todas las estructuras, usando ambos protocolos de medida, y para todos los ciclos. Los resultados (fig. 3.7, pág. 57) muestran que efectivamente, como norma general, el error de fase es menor para el caso VCT; como excepción están los casos de AMMO y BA-TART, para las que se obtienen valores comparables, como se podía deducir del anterior cálculo de la correlación entre los difractogramas.

Además, para cada una de las soluciones escogidas se calculó el correspondiente mapa de densidad electrónica, comparando seguidamente los picos de densidad encontrados con las posiciones atómicas correctas: en la tabla 3.2 (pág. 60) se muestra la cantidad de átomos correctamente localizados (con una distancia menor de 0.6Å), para el último ciclo correspondiente a cada estructura. Se puede comprobar que, a excepción del caso de BA-TART, el protocolo VCT ofrece la localización correcta de un mayor número de átomos y con una precisión también mayor que usando el protocolo FCT.

Por último, otro parámetro a tener en cuenta es la resolución hasta la cual se ha llegado en la realización de la medida, ya que el hecho de alcanzar una mayor resolución no implica obtener mejores resultados. Para comprobar su influencia hemos repetido los cálculos para dos estructuras, pero con la diferencia de que hemos usado una resolución menor. Los resultados muestran una mejora de la eficiencia de los Métodos Directos, tanto para la estructura AMMO (en este caso debido a que la zona excluida de los cálculos presentaba un fuerte solapamiento de picos) como para la MN-TART (la zona excluida no presentaba señal sino ruido, prácticamente).

Conclusiones

Se ha evaluado el papel que lleva a cabo el paso de medida de datos de difracción en polvo en el proceso de resolución estructural, usando para ello diferentes conjuntos de datos obtenidos tanto con el esquema de medida en tiempo fijo (FCT) como con el de tiempo variable (VCT). Los resultados obtenidos al procesar los datos mediante Métodos Directos han mostrado que el esquema VCT ofrece una mejor eficiencia en el proceso de fasado (principalmente para estructuras orgánicas), en términos de error medio de fase y localización correcta de los átomos al final del proceso. Este hecho es consecuencia directa de una extracción de intensidades más precisa del patrón de difracción (valores más bajos de R_F), ya que la estrategia VCT mejora la calidad de los datos que se usarán posteriormente en los cálculos, justamente en la zona de mayor densidad de reflexiones y, por lo tanto, donde existe mayor posibilidad de cometer errores en la estimación de las intensidades debido al solapamiento de picos. Tanto para el esquema FCT como para el VCT, el hecho de incrementar el tiempo total de medida conlleva la obtención

de mejores resultados en la resolución estructural; pero esta mejora no se puede mantener indefinidamente, ya que la cantidad de información disponible en el difractograma tiene una limitación, la cual viene impuesta principalmente por el solapamiento de las reflexiones. En ciertos casos esta limitación puede ser superada, aunque solamente en parte, excluyendo de los cálculos la zona de datos correspondiente a los mayores ángulos 2θ (datos que, en vez de ofrecer más información, hacen incrementar el error en las estimaciones de intensidad).

Capítulo 4.

Determinación del grupo espacial para datos de difracción en polvo.

Introducción

En el proceso de resolución estructural, una vez que hemos indexado el patrón de difracción, el siguiente paso que debemos dar es el de determinar el grupo espacial en el cual ha cristalizado el compuesto. Esto normalmente se lleva a cabo mediante una inspección visual por parte del investigador, para comprobar cuáles son las extinciones sistemáticas presentes y así deducir el grupo espacial. Pero al trabajar con datos de difracción en polvo nos encontramos con la gran dificultad del solapamiento de los picos, lo cual implica que el hecho de poder asegurar si una reflexión está o no presente en el difractograma no sea una tarea trivial.

En este apartado se ha desarrollado un algoritmo basado en un estudio estadístico de las intensidades normalizadas *z*, el cual está enfocado a la determinación del grupo de extinción del compuesto bajo estudio.

Metodología y resultados

El proceso seguido para el funcionamiento del algoritmo se puede resumir en los siguientes pasos:

1. La celda unidad debe haber sido ya determinada. Por lo tanto, sabemos cual es el sistema cristalino.
2. Los símbolos de extinción para cada sistema cristalino, junto con sus condiciones de extinción, han sido previamente tabulados y almacenados.
3. Se realiza la extracción de intensidades a partir del patrón de difracción, mediante el método de Le Bail, usando para ello el grupo espacial con mayor simetría de Laue y sin condiciones de extinción. Es decir: *P*12/*m*1 para el sistema monoclínico; *P*2/*m*2/*m*2/*m* para el ortorrómbico; *P*4/*mmm* para el tetragonal; *P*6/*mmm* para los sistemas trigonal y hexagonal; *Pm*-3*m* para el cúbico.
4. Se normalizan las intensidades extraídas, de acuerdo al método clásico de Wilson.
5. Por último, el algoritmo define las probabilidades de cada grupo de extinción usando la estadística de *z*.

Para ver de una forma básica cómo funciona el algoritmo, consideremos por ejemplo el sistema ortorrómbico, en el cual cualquier grupo espacial se puede escribir de la siguiente manera:

$$M\ r_1/s_1\ r_2/s_2\ r_3/s_3$$

donde M es el tipo de celda y r_j, j=1−3 y s_j, j=1−3 son los elementos de simetría a lo largo y perpendicularmente a los tres ejes, respectivamente.

El algoritmo asignara así una determinada probabilidad de existencia a cada uno de los elementos de simetría posibles; por ejemplo, en este caso, la probabilidad de existencia para los ejes 2_1 a lo largo de **a**, **b** y **c**, respectivamente, se define como

$$p(2_{1[100]}) = 1 - \langle z_{h00} \rangle_{h=2n+1} ,$$

$$p(2_{1[010]}) = 1 - \langle z_{0k0} \rangle_{k=2n+1} ,$$

$$p(2_{1[001]}) = 1 - \langle z_{00l} \rangle_{l=2n+1} .$$

Así, la probabilidad es nula si el valor medio de z es igual o mayor que la unidad (lo cual significa que existen reflexiones con intensidad no nula para esa condición de extinción y, por lo tanto, no se cumple) y máxima cuando está cercana a cero. Un cálculo análogo correspondería a los planos de simetría y al tipo de celda. Para aquellos elementos que no estén asociados a una condición de extinción, las probabilidades correspondientes son calculadas en base a las probabilidades asignadas a los elementos que sí tienen condición de extinción. Por ejemplo:

$$p(2_{[100]}) = 1 - p(2_{1[100]})$$

o

$$p(m) = 1 - \max\ [p(b), p(c), p(n), p(d)] .$$

Finalmente, la probabilidad de cada grupo de extinción es calculada como el producto de las probabilidades de los elementos de simetría involucrados en él. Por ejemplo:

$$p(P\text{ - - -}) = p(P)\; p(2_{[100]})\; p(m\perp\mathbf{a})\; p(2_{[010]})\; p(m\perp\mathbf{b})\; p(2_{[001]})\; p(m\perp\mathbf{c})\;.$$

Si los valores estimados para las intensidades integradas fueran lo bastante precisos, el algoritmo debería funcionar correctamente sin más consideraciones. Pero el problema habitual del solapamiento de picos hace que los valores de *z* asignados no sean los correctos para un conjunto de las reflexiones presentes (en ocasiones, para muchas de ellas), lo cual reduce la efectividad del algoritmo. Por ello, para evitar todo lo posible los efectos negativos que implica el solapamiento, resulta más útil el asignar a cada reflexión un determinado peso en función de su varianza, la cual depende directamente tanto de su grado de solapamiento con las reflexiones adyacentes como de la intensidad total del grupo al cual pertenece. Así, en los cálculos de probabilidad de los elementos de simetría, en vez de usar $\langle z \rangle$ resulta más ventajoso usar los valores de las intensidades pesadas

$$\langle z_w \rangle = \left(\sum w_j z_j\right) \Big/ \sum w_j\;,$$

donde el peso viene calculado como

$$w = \exp\left(-2\sigma_z^2\right),$$

siendo w=1 para aquellas reflexiones que sean individuales y w<1 para aquellas que estén en solapamiento.

El algoritmo, brevemente descrito, fue completado y mejorado con otros criterios secundarios, como la limitación de las probabilidades mínimas a 0.1 en vez de a cero, el uso de las *z* tanto de las reflexiones que cumplen una

determinada condición de extinción como del conjunto complementario, o la modificación de la probabilidad de ciertos elementos de simetría en función del número de reflexiones existentes que cumplen su condición de extinción, por ejemplo. Todo ello, junto a un mejorado tratamiento del ajuste del background en el programa de extracción de intensidades y otro algoritmo desarrollado para eliminar posibles picos impureza presentes en el difractograma, fue aplicado a los datos de 59 estructuras test (ver tabla 4.1, pág. 86), abarcando todos los sistemas cristalinos (a excepción del sistema triclínico). Los resultados obtenidos se muestran en la tabla 4.2 (pág. 88); para cada estructura se muestra el símbolo de extinción correcto (cexts) y el puesto en el que ha sido propuesto dicho símbolo, en orden de probabilidad, por el algoritmo (orn): orn=1 quiere decir que al símbolo correcto se le ha asignado la máxima probabilidad. Por otra parte, el parámetro orn2 ofrece la posición asignada después de realizar diversas correcciones en los valores de las probabilidades de los elementos de simetría, los cuales son llevados a cabo cuando se encuentra efectos importantes de pseudosimetría; en la tabla 4.3 (pág. 93) se muestran los vectores de pseudotraslación que el algoritmo está en grado de manejar (básicamente, las probabilidades de los elementos de simetría simulados por los vectores pseudotraslacionales existentes vienen reducidas en función de su efecto).

Los resultados muestran que para un gran número de estructuras (46) el símbolo de extinción es correctamente identificado (orn y/o orn2 = 1), mientras que para otras seis la posición en la lista fue la segunda, lo cual es una muestra de la gran efectividad del algoritmo. Además, el identificar el símbolo correcto resulta más fácil cuando no existen efectos de pseudosimetría y más difícil para los sistemas cristalinos de mayor simetría. En este último caso, la ayuda de la interfaz gráfica implementada resulta muy útil para descartar símbolos de

extinción incorrectos a los que pueda haber sido asignada una alta probabilidad de existencia. Por otra parte, para varias estructuras (sobre diez de ellas) la discriminación alcanzada en términos de probabilidad entre el símbolo de extinción correctamente identificado y el segundo de la lista es muy significativo; por ejemplo, para el caso de la estructura ADAM se obtiene p=0.547 y p=0.07 para el primero y segundo símbolo, respectivamente.

Conclusiones

Se ha desarrollado un nuevo algoritmo para la determinación del grupo de extinción. Éste está basado en un estudio estadístico de las intensidades integradas normalizadas que han sido previamente extraídas del patrón de difracción, asignando una determinada probabilidad a cada símbolo de extinción que sea compatible con el tipo de red del compuesto. Los resultados de los tests realizados con un gran número de estructuras han mostrado que, en la mayoría de los casos, el símbolo de extinción correcto ha sido apropiadamente definido como el más probable. El algoritmo también incluye un análisis (junto con la correspondiente corrección de los valores de las probabilidades) de los posibles efectos debidos a la presencia de pseudosimetrías. Aparte del análisis estadístico, el algoritmo es capaz de reducir los errores en el ajuste del background, de eliminar posibles picos de impurezas existentes en el difractograma y ofrece al usuario una útil y sencilla interfaz gráfica, la cual resulta de mucha ayuda para descartar fácilmente grupos de extinción incorrectos a los que se les haya asignado un alto valor de probabilidad.