

A new approach for indexing powder diffraction data based on whole-profile fitting and global optimization using a genetic algorithm

Benson M. Kariuki,^a Scott A. Belmonte,^b Malcolm I. McMahon,^b Roy L. Johnston,^a Kenneth D. M. Harris^{a*} and Richard J. Nelmes^{b*}

^a*School of Chemistry, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK, and*

^b*Department of Physics and Astronomy, University of Edinburgh, Edinburgh EH9 3JZ, UK.*

E-mail: k.d.m.harris@bham.ac.uk

(Received 19 October 1998; accepted 18 December 1998)

This paper describes a new technique for indexing powder diffraction data. The lattice parameters (unit-cell dimensions) $\{a, b, c, \alpha, \beta, \gamma\}$ define the parameter space of the problem, and the aim is to find the optimal lattice parameters for a given experimental powder diffraction pattern. Conventional methods for indexing consider the measured positions of a limited number of peak maxima, whereas this new approach considers the whole powder diffraction profile. This new strategy offers several advantages, which are discussed fully. In this approach, the quality of a given set of lattice parameters is determined from the profile R -factor, R_{wp} , obtained following a Le Bail-type fit of the intensity profile. To find the correct lattice parameters (*i.e.* the global minimum in R_{wp}), a genetic algorithm has been used to explore the $R_{wp}(a, b, c, \alpha, \beta, \gamma)$ hypersurface. (Other methods for global minimization, such as Monte Carlo and simulated annealing, may also be effective.) Initially, a number of trial sets of lattice parameters are generated at random, and this 'population' is then allowed to evolve subject to well defined evolutionary procedures for mating, mutation and natural selection (the fitness of each member of the population is determined from its value of R_{wp}). Examples are presented to demonstrate the success and underline the potential of this new approach for indexing powder diffraction data.

Keywords: powder diffraction; genetic algorithms; indexing; whole-profile fitting.

1. Introduction

In general, crystal structure determination from diffraction data involves the following stages: (a) determination of the unit cell (*i.e.* indexing the diffraction pattern), (b) space group assignment, (c) structure solution, and (d) structure refinement. For powder diffraction data, the structure solution stage is generally regarded as the most challenging, and significant advances have been made recently in the scope and power of methods for structure solution [for reviews see Cheetham & Wilkinson (1992), Harris & Tremayne (1996), Poojary & Clearfield (1997) and subsequent work described by Harris *et al.* (1994), Tremayne *et al.* (1997), Andreev *et al.* (1997), Kariuki *et al.* (1997), Harris *et al.* (1998), Shankland *et al.* (1997), David *et al.* (1998)]. However, it is not possible to proceed at all with structure solution unless the correct unit cell is found in stage (a). Indeed, we have found that reliable indexing of powder diffraction data using existing techniques is often the limiting step in the structure determination process. In contrast to the recent advances in techniques for structure solution, there has been relatively little fundamental development of indexing methods since the pioneering

work over 20 years ago [see Langford & Louër (1996) for a recent review]. For these reasons, and in the light of the more demanding structural problems now being addressed and the availability of faster computers, there is now a pressing need to re-assess the strategies for indexing powder diffraction patterns.

In this paper, we describe a new indexing technique, based on whole-profile fitting and global optimization using a genetic algorithm. Examples are given to illustrate the successful application of our method. Although these examples are for synchrotron X-ray powder diffraction data, our method is general, and may also be applied to conventional laboratory X-ray powder diffraction data or neutron powder diffraction data.

2. Background to indexing powder diffraction patterns

The unit-cell dimensions (lattice parameters) $\{a, b, c, \alpha, \beta, \gamma\}$ determine the positions of peaks in a powder diffraction pattern. The aim of indexing is to determine the correct lattice parameters by using information on the peak posi-

tions in the experimental pattern. To our knowledge, all indexing strategies reported to date are based on using peak positions obtained directly from the experimental powder diffraction pattern, although different strategies are adopted for extracting this information [general reviews are given by Shirley (1980), Louër (1992), Langford & Louër (1996)]. Currently, the most widely used indexing programs are *ITO* (Visser, 1969), *TREOR* (Werner *et al.*, 1985) and *DICVOL* (Boultif & Louër, 1991). All these programs require the measured positions of peak maxima for a number of peaks. However, experimental patterns typically have extensive peak overlap, and sometimes peak displacements, which can lead to several problems, now discussed.

(i) Only a subset of the complete diffraction data (the selected peaks) is actually used in the indexing procedure. As overlap is more severe at high values of $\sin(\theta)/\lambda$, the selected peaks tend to be confined to the low-angle region. For materials with low symmetry and large unit cells, the overlap of peaks is particularly severe, making extraction of peak positions difficult even at low values of $\sin(\theta)/\lambda$, and reducing the limiting value of $\sin(\theta)/\lambda$ above which reliable peak positions cannot be extracted. In many cases, certain peaks which are important in the indexing process may be obscured or completely unresolved due to peak overlap.

(ii) In many cases, the limitations described above are exacerbated by factors that give rise to a reduction in the crystallinity of the sample; for example, small particle size, stacking faults and other defects that cause peak broadening. These factors can lead directly or indirectly (e.g. through peak overlap) to serious errors (e.g. greater than $\pm 0.02^\circ$ in 2θ) in measured peak positions. For large unit cells, even a precision of $\pm 0.02^\circ$ in measured 2θ values may be insufficient to distinguish one solution from another.

(iii) Other difficulties (often more serious than those described above) can arise from defects that displace the peaks such that they do not actually fit any lattice exactly. This problem, which is encountered particularly often in powder diffraction data recorded at high pressure, seriously limits the ability of the existing indexing programs to find any solution.

(iv) The presence of impurity phases in samples studied by powder diffraction is often unavoidable, and often constitutes an intractable problem in carrying out indexing by existing methods [unless the existence and identity of the impurity phase(s) are known *a priori*]. Thus, if an impurity peak is inadvertently selected for the indexing procedure, it is generally impossible to find the correct unit cell [although some cases of successful indexing under these circumstances have been reported (Dinnebier *et al.*, 1997; Williams *et al.*, 1992)].

For reliable indexing by the existing methods, it is recommended that at least 20 measured peak positions should be used, especially when the symmetry is expected to be low. These indexing procedures are particularly vulnerable to experimental errors in the measured peak

positions. Random errors must be small [preferably less than $\delta(2\theta) \simeq 0.02^\circ$], and systematic errors (such as a significant zero offset or the inclusion of peaks due to impurity phases) will often lead to failure.

There is clearly a pressing need for new indexing strategies that do not require extracted peak positions, but instead use all the available d -spacing information in a powder pattern (including strongly overlapped regions). In this paper, we report a new approach of this kind. It requires two new aspects: (i) a method to assess the quality of fit of the whole powder diffraction profile for a given set of trial lattice parameters $\{a, b, c, \alpha, \beta, \gamma\}$, and (ii) an efficient method for searching parameter space to find the best set of lattice parameters (*i.e.* the correct unit cell). In our method, we assess the correctness of each set of trial lattice parameters by using a profile-fitting procedure based on a slightly modified version of the conventional weighted profile R -factor (R_{wp}), and we explore parameter space to find the global minimum in R_{wp} using a genetic algorithm.

This approach is fundamentally different from any of the indexing methods available hitherto, and is intrinsically more robust for dealing with the problems discussed above. Peak overlap is implicitly taken care of, and the powder diffraction data are used directly as measured. Furthermore, other information that may affect peak positions in the powder diffraction pattern [such as (hkl)-dependent effects or zero point error] may be incorporated directly into the calculation.

3. Method for whole-profile fitting

First, we describe our method for constructing a powder diffraction profile for a set of trial lattice parameters $\{a, b, c, \alpha, \beta, \gamma\}$, thereby allowing $R_{wp}(a, b, c, \alpha, \beta, \gamma)$ to be determined. The given set of lattice parameters determines the peak positions, and parameters describing the peak shape and peak width are used in the Le Bail profile-fitting procedure (Le Bail *et al.*, 1988; Toraya, 1993) to partition the observed intensities, producing a calculated pattern from which R_{wp} is calculated. From our experience, we have found that this method for obtaining a calculated pattern is orders of magnitude faster than least-squares fitting of individual intensities and does not suffer from the potential stability problems of the least-squares method. It should be noted that the Le Bail technique is not based on the minimization of R_{wp} . However, R_{wp} will be smallest when all the parameters of the model are correct. Tests carried out on data recorded for materials of known structure confirm that this method for calculating R_{wp} can successfully identify and discriminate the correct lattice parameters (see §5).

The correct partitioning of the observed data does require a reasonable initial estimate of peak shapes and widths. Such an estimate is obtained either from experience with similar patterns or by fitting representative peaks individually across the pattern. For this purpose, two or

three sharp peaks that may be assigned confidently as non-overlapped with any other peak should be chosen.

The presence and position of peaks at low $\sin(\theta)/\lambda$ can be crucial for the correct indexing of a pattern. However, these peaks are often weak, and using the conventional R_{wp} to assess the agreement between the experimental and calculated data may fail to recognize the significance of these peaks. The conventional R_{wp} assesses the agreement between observed and calculated *intensities*, and in our indexing method (as presently applied) it would be most strongly affected by discrepancies in the positions of the strongest peaks. Discrepancies in the positions of weak peaks will have little effect on the conventional R_{wp} , even though such peaks may be crucial for successful indexing. To overcome this limitation, we use a modified version of R_{wp} . The pattern is split into different regions (defined by the user), and the weighted profile R -factor is then calculated separately for each region. These R -factors are summed to obtain the overall R_{wp} .

$$R_{wp} = \sum_{\text{region}} \left[\sum_i w_i (y_i - y_{ci})^2 / \sum_i w_i y_i^2 \right]^{1/2}, \quad (1)$$

where i runs over the points in each region. With this procedure, the residuals for a given region are scaled according to the total intensity of the region (the lower the total intensity, the higher the scaling factor). Thus, a region containing only weak peaks can make as significant a contribution to the modified R_{wp} as regions that contain peaks of much greater absolute intensity. The importance of this effect is well illustrated in §6. There is also scope for overcoming the problems discussed above by using other definitions of R_{wp} with weighting schemes (w_i) based on d -spacing rather than intensity.

By the nature of the whole-profile method, peak overlap is handled directly and the complete experimental d -spacing information is used. Fitting the whole profile also gives a natural measure of the errors in the d -spacings (e.g. broad peaks are associated with larger error than narrow peaks). An important advantage of whole-profile fitting is that the presence of a minority impurity phase does not limit the success of the procedure. Peaks due to impurity phases simply contribute a constant amount to the R -factor, and the global minimum in R_{wp} will arise when the majority phase is indexed correctly; an example of this is given in §6.

Furthermore, the lattice parameters need not be the only parameters optimized in our approach. The parameters defining the peak shapes and peak widths could be subjected to restrained refinement, and parameters describing the zero offset and/or (hkl)-dependent sample effects could also be included in the optimization. In principle, such features could also be considered in the existing indexing methods (some of which already allow the zero offset to be varied). However, as these methods use only a subset of the available d -spacing information, a significant increase in the number of variable parameters is

not necessarily justified (and could invalidate these methods completely, on the grounds that the number of variable parameters is too large in comparison with the number of observables).

4. Search methods

The aim of our indexing strategy is to search parameter space $\{a, b, c, \alpha, \beta, \gamma\}$ in order to find the parameter set in best agreement (lowest R_{wp}) with the experimental powder diffraction pattern. In effect, we require to search the hypersurface $R_{wp}(a, b, c, \alpha, \beta, \gamma)$ to find the global minimum. We note that some of these parameters may be fixed by the metric symmetry of the lattice, and it is only in the triclinic case that all six parameters are treated as variables.

In our search procedure, the cell volume is restricted to sensible ranges consistent with our knowledge of the system under study (recalling that any powder diffraction pattern can be indexed by a unit cell of sufficiently large volume). Even with a restricted volume range, however, a systematic grid search through parameter space would require an unfeasible length of time, and more efficient search methods are required. A number of different approaches for global optimization may be used, including Monte Carlo methods, simulated annealing methods and genetic algorithms. In the present work, we have used a genetic algorithm, based on the recognition that certain features of typical $R_{wp}(a, b, c, \alpha, \beta, \gamma)$ hypersurfaces make them particularly suited for exploration using genetic algorithms (see §5). In this context, we note that a method for indexing powder diffraction data by using a genetic algorithm within the framework of existing indexing strategies (i.e. requiring a set of extracted peak positions) has also been proposed recently (Paszkowicz, 1996). Genetic algorithms have also been used recently in fitting SAXS/SANS data (Chacon *et al.*, 1998).

5. Genetic algorithms

The genetic algorithm (GA) is an optimization technique based on evolutionary principles, in which the fittest members of a population survive and procreate, leading to improved subsequent generations (Goldberg, 1989; Keane, 1996; Harris *et al.*, 1998). In the present case, the population comprises a number of sets of trial lattice parameters $\{a, b, c, \alpha, \beta, \gamma\}$ (or a subset of these parameters, depending on the crystal system), and the population is allowed to evolve through several generations (see Fig. 1) using the evolutionary operations of mating, mutation and natural selection. The quality ('fitness') of each set of trial lattice parameters depends on its value of R_{wp} (low R_{wp} corresponds to high fitness). In the examples reported in §6, the fitness (F) of a given member of the population is defined as $F = 1 - (R_{wp} - R_{\min}) / (R_{\max} - R_{\min})$, where $R_{\min}$ and $R_{\max}$ denote the lowest and highest values of R_{wp} in the current population ($R_{\min}$ and $R_{\max}$ are updated for each new generation of the population). Specific details of our

strategy for carrying out the GA calculation will be given in future publications, although we note that it is similar to the strategy implemented in our GA technique for structure solution (Harris *et al.*, 1998).

For GA methods to lead to efficient global optimization, an important requirement is that the evolutionary procedure is able to recognize certain combinations of parameters (known as schemata) that are associated with high fitness. Thus, if a subset of the parameters defining a member of the population is close to optimal but the other parameters are not optimal, it is important that the GA process can identify and propagate this optimal subset of parameters. Importantly, we believe that such schemata are expressed strongly in the case of indexing powder diffraction data. For example, in the $R_{wp}(a,c)$ surface shown in Fig. 2, there are several strong local minima connected by deep 'valleys' to the global minimum, which corresponds to the correct (known) unit cell. These valleys provide an expression of the schemata in this surface, and they help the GA to locate the global minimum successfully despite the presence of the local minima. Clearly the existence and proper handling of such schemata will assist the GA calculation to reach the global minimum rapidly. The

existence of strong schemata in indexing powder diffraction data is readily understood by recognizing that certain parts of the experimental data depend specifically on only one parameter or a well defined group of parameters. For example, the positions of $\{h00\}$ peaks depend only on the parameter a , the positions of $\{hk0\}$ peaks depend only on the subset of parameters $\{a,b,\gamma\}$, and so on.

6. Examples of the application of our methodology

Our new indexing method has been applied successfully in a number of cases, and is illustrated here for unit-cell determination of orthorhombic GaAs-II (space group *Cmcm*). Importantly, the experimental data contain an impurity phase, thus representing a particularly challenging problem for indexing (see §1). The unit cell was determined previously from powder X-ray diffraction data as $a = 4.9703$ (1) Å, $b = 4.7801$ (3) Å, $c = 5.2717$ (4) Å. The use of known cells in this preliminary work allows a definitive test of the validity of our method.

Fig. 3(a) shows the raw data with the impurity peaks highlighted by arrows. The successful indexing of this pattern relies on correctly fitting the position of the two

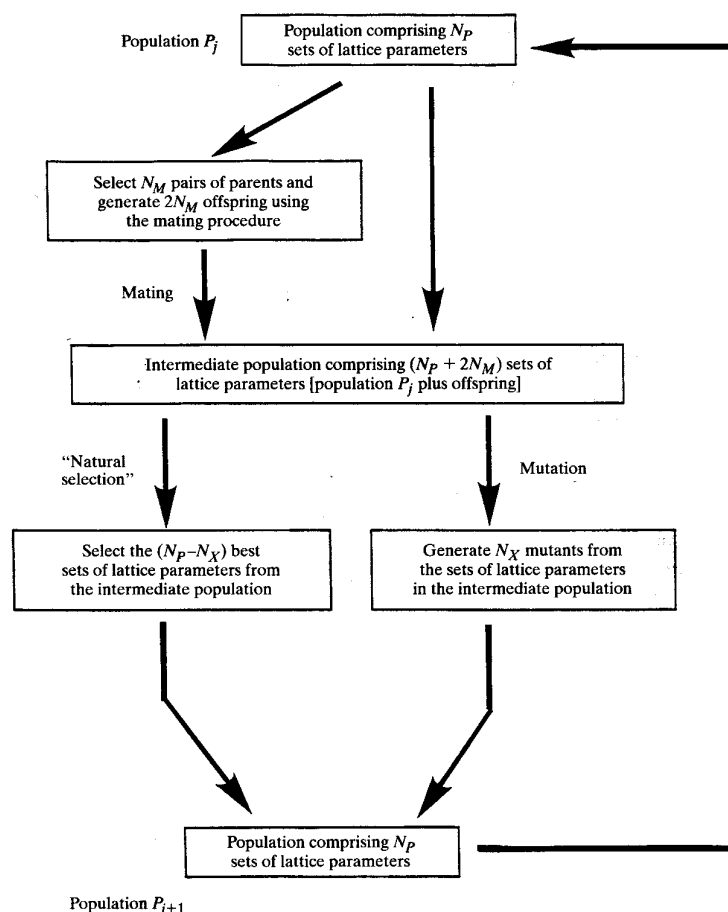


Figure 1

Procedure for evolution of the population from one generation (P_j) to the next generation (P_{j+1}) in the GA calculation.

very weak peaks at low angle (marked by asterisks and shown enlarged in the inset), but these two peaks would have negligible effect on the conventional R_{wp} . However, by considering two regions: (1) $6^\circ \leq 2\theta < 9.5^\circ$, (2) $9.5^\circ \leq 2\theta < 30^\circ$, and using our modified R_{wp} (defined in §3), both regions influence the fit significantly. As illustrated in Fig. 3(b), this procedure effectively scales the intensity in region (1) relative to region (2) during calculation of R_{wp} , although it is important to note that the raw data are not actually modified. There is no limit to the number of such regions that can be defined, and experience will lead to an understanding of the optimum strategies to apply in this regard.

Fig. 3(b) also shows the fit to the observed data obtained from the best (and correct) indexing of the unit cell given by our GA calculation. The inclusion of just the strongest impurity peak causes all of the commonly available indexing programs to fail completely, but by using our whole-profile fitting method it has been possible to index the majority orthorhombic phase. [The presence of two calculated peaks either side of the strongest impurity peak in Fig. 3(b) is a consequence of the Le Bail method of fitting whereby the intensity of the impurity peak is partitioned almost equally into the calculated peak positions given by the correct unit cell on either side of the impurity peak.]

The GA calculation involved the evolution of 500 generations of a population of 100 (N_P) sets of trial lattice parameters $\{a, b, c\}$. In each generation, 25 mating operations (N_M) (leading to 50 offspring) and 20 mutations (N_X) were considered. Each of the lattice parameters $\{a, b, c\}$ was constrained to be less than 8 Å during the calculation, and the cell volume was constrained to be less than 200 Å³. Fig. 4 shows the evolution of the value of R_{wp} for the best

member of the population and the average value of R_{wp} for the population as a function of generation number during the GA calculation. Clearly, the quality of the population improves during the GA calculation; the average R_{wp} decreases quickly in the early stages of the evolution, with the rate of improvement decreasing in the later stages. The set of lattice parameters corresponding to the lowest R_{wp} at the end of the GA calculation are $a = 4.968$ Å, $b = 4.773$ Å, $c = 5.285$ Å, in good agreement with the values found previously, $a = 4.9703$ Å, $b = 4.7801$ Å, $c = 5.2717$ Å.

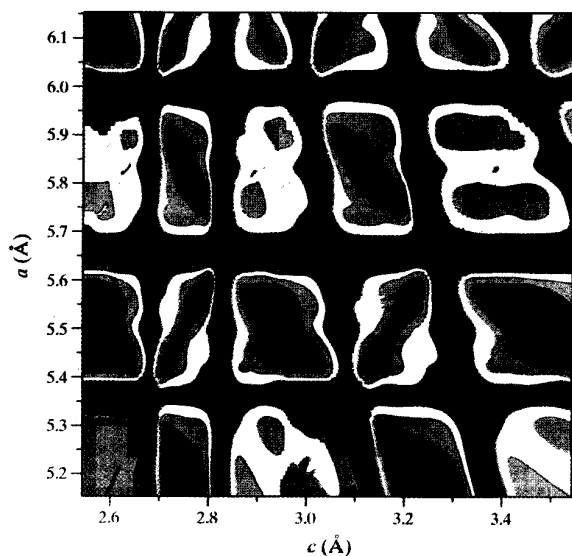


Figure 2

Contour plot of an R -factor surface around the global minimum of tetragonal Sr ($a = 5.653$ Å, $c = 3.043$ Å). The 'valleys' leading to the global minimum show directly the existence of schemata in this particular R -factor space.

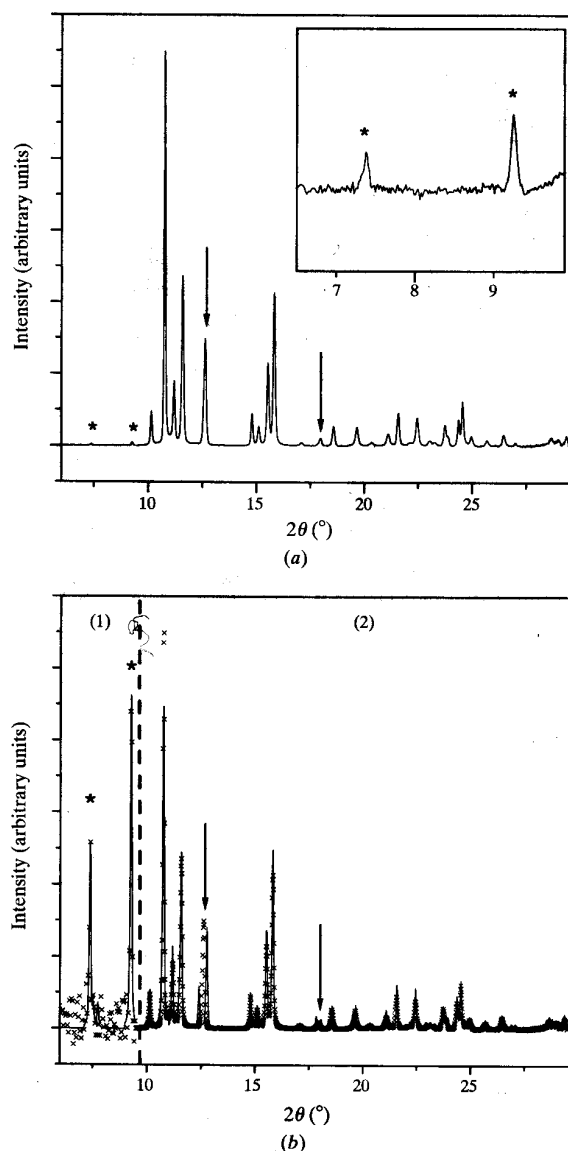


Figure 3

(a) Raw data collected from orthorhombic phase-II of GaAs. The arrows highlight the impurity peaks and the asterisks mark the positions of very weak reflections, the existence of which is crucial for correct indexing of the pattern (an enlargement of the region around the weak peaks is shown in the inset). (b) The Le Bail fit of the calculated data (solid line) to the observed data (crosses) when the pattern is correctly indexed. Regions (1) and (2) are discussed in the text.

7. Conclusions

The results reported here demonstrate clearly the utility of our new approach for indexing powder diffraction data. A rigorous optimization of our strategy is currently in progress, and further aspects of its implementation are being developed. In addition to the opportunities for further development discussed above [such as inclusion, within the optimization, of parameters describing the zero offset and (*hkl*)-dependent effects], other issues that we are currently investigating include: careful exploration of the definition of the profile *R*-factor used to assess the quality of each set of trial lattice parameters [for example, weighting schemes based on $\sin(\theta)/\lambda$ rather than intensity may be advantageous]; the exploration of other strategies for carrying out the GA calculation (including different definitions of fitness functions, different procedures for the mating and mutation operations *etc.*); the use of other search techniques, such as Monte Carlo and simulated annealing methods; the inclusion, within our whole-profile indexing strategy, of information on the positions of selected low-angle peaks, which may be known accurately; applications to the case of indexing multiphase powder-diffraction patterns (requiring two independent sets of lattice parameters to be determined); and the exploration of potential opportunities to merge aspects of indexing and structure solution, using our new indexing method based on consideration of the whole profile and direct space structure solution strategies employing similar figures of merit. [We note that another method for merging indexing and structure solution, also based on using the whole profile, has been discussed (Bartell *et al.*, 1987; Bartell & Caillat, 1987)].

Given the advantages of carrying out indexing based on whole-profile fitting of the powder diffraction pattern, and

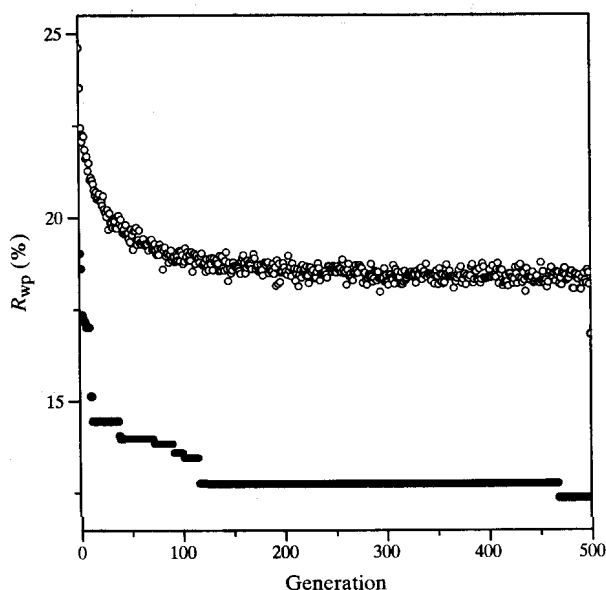


Figure 4
Evolution of R_{wp} for the best unit cell (closed circles) and the average R_{wp} for the population (open circles) as a function of generation number in the GA calculation.

the opportunities for further development and optimization of the new strategy reported here, we are confident that the future application of this method will open up important new opportunities in the analysis, understanding and exploitation of powder diffraction data in many areas of solid state sciences.

We are grateful to EPSRC for research grant support (KDMH, RJN), and CCLRC for support and assistance with facility development (RJN). SAB would like to thank Lachlan M. D. Cranswick for some assistance on software-related matters.

References

- Andreev, Y. G., MacGlashan, G. S. & Bruce, P. G. (1997). *Phys. Rev. B*, **55**, 12011–12017.
- Bartell, L. S. & Caillat, J. C. (1987). *J. Appl. Cryst.* **20**, 461–466.
- Bartell, L. S., Caillat, J. C. & Powell, B. M. (1987). *Science*, **236**, 1463–1465.
- Boultif, A. & Louër, D. (1991). *J. Appl. Cryst.* **24**, 987–993.
- Chacon, P., Moran, F., Diaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Biophys. J.* **74**, 2760–2775.
- Cheetham, A. K. & Wilkinson, A. P. (1992). *Angew. Chem. Int. Ed. Engl.* **31**, 1557–1570.
- David, W. I. F., Shankland, K. & Shankland, N. (1998). *Chem. Commun.* pp. 931–932.
- Dinnebier, R. E., Olbrich, F., van Smaalen, S. & Stephens, P. W. (1997). *Acta Cryst.* **B53**, 153–158.
- Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison-Wesley.
- Harris, K. D. M., Johnston, R. L. & Kariuki, B. M. (1998). *Acta Cryst.* **A54**, 632–645.
- Harris, K. D. M. & Tremayne, M. (1996). *Chem. Mater.* **8**, 2554–2570.
- Harris, K. D. M., Tremayne, M., Lightfoot, P. & Bruce, P. G. (1994). *J. Am. Chem. Soc.* **116**, 3543–3547.
- Kariuki, B. M., Serrano-González, H., Johnston, R. L. & Harris, K. D. M. (1997). *Chem. Phys. Lett.* **280**, 189–195.
- Keane, A. J. (1996). *Modern Heuristic Search Methods*, edited by V. Rayward-Smith, I. Osman, C. Reeves & G. D. Smith. New York: Wiley.
- Langford, J. I. & Louër, D. (1996). *Rep. Prog. Phys.* **59**, 131–234.
- Le Bail, A., Duroy, H. & Fourquet, J. L. (1988). *Mater. Res. Bull.* **23**, 447–452.
- Louër, D. (1992). *Accuracy in Powder Diffraction II: NIST Special Publ. No. 846*, edited by E. Prince & J. K. Stalick, pp. 92–104. Gaithersburg, MA: US Department of Commerce.
- Paszkowicz, W. (1996). *Mater. Sci. Forum*, **228–231**, 19–24.
- Poojary, D. M. & Clearfield, A. (1997). *Acc. Chem. Res.* **30**, 414–422.
- Shankland, K., David, W. I. F. & Csoka, T. (1997). *Z. Kristallogr.* **212**, 550–552.
- Shirley, R. (1980). *Data Accuracy for Powder Indexing*. Natl. Bur. Stand. (US) Spec. Publ. No. 567361.
- Toraya, H. (1993). *The Rietveld Method*, edited by R. A. Young pp. 254–275. IUCr/Oxford University Press.
- Tremayne, M., Kariuki, B. M. & Harris, K. D. M. (1997). *Angew. Chem. Int. Ed. Engl.* **36**, 770–772.
- Visser, J. W. (1969). *J. Appl. Cryst.* **2**, 89–95.
- Werner, P.-E., Eriksson, L. & Westdahl, M. (1985). *J. Appl. Cryst.* **18**, 367–470.
- Williams, J. H., Cockcroft, J. K. & Fitch, A. N. (1992). *Angew. Chem. Int. Ed. Engl.* **31**, 1655–1657.