Acta Cryst. (1998). A**54**, 632–645

The Genetic Algorithm: Foundations and Applications in Structure Solution from Powder Diffraction Data

Kenneth D. M. Harris,* Roy L. Johnston and Benson M. Kariuki

School of Chemistry, University of Birmingham, Edgbaston, Birmingham B15 2TT, England.
E-mail: k.d.m.harris@bham.ac.uk

(*Received 1 December 1997; accepted 25 February 1998*)

Abstract

Recently, new methods based on the use of genetic algorithms have been explored and developed for solving crystal structures directly from powder diffraction data. In implementing genetic algorithms in such applications, several different aspects of the technique and strategy are open to optimization, leading to a versatile and powerful approach. In this paper, the fundamental concepts underlying genetic algorithms are discussed and the implementation of the genetic algorithm for structure solution from powder diffraction data is described. The opportunities, scope and potential for future developments in the foundations and applications of genetic algorithms in this field are highlighted. The genetic algorithm approach adopts the 'direct-space' philosophy for structure solution, with trial structures generated independently of the experimental diffraction data and the quality of each structure assessed by comparing the calculated and experimental powder diffraction patterns; in this work, this comparison is made using the profile R factor R_{wp}. In the genetic algorithm, a population of trial structures is allowed to evolve subject to well defined rules governing mating, mutation and 'natural selection'. The 'fitness' of each structure in the population is a function of its profile R factor. The successful application of the genetic algorithm approach for structure solution of molecular crystals from powder diffraction data is demonstrated with examples of previously known and previously unknown structures.

1. Introduction to structure solution from powder diffraction data

1.1. *The traditional approach*

The determination of crystal structures from single-crystal X-ray diffraction data can generally be carried out straightforwardly, provided single crystals of appropriate size and quality are available. However, many crystalline solids cannot be prepared in the form of appropriate single crystals and are therefore not amenable to structural characterization by conventional single-crystal X-ray diffraction techniques. In such cases, progress relies on the availability of techniques for crystal-structure determination using powder diffraction data (Christensen *et al.*, 1985; Cheetham & Wilkinson, 1991, 1992; McCusker, 1991; Rudolf, 1993; Harris & Tremayne, 1996; Langford & Louër, 1996; Poojary & Clearfield, 1997) or microcrystal diffraction data (recorded using synchrotron X-radiation) (Harding, 1996; Harding *et al.*, 1994; Gray *et al.*, 1997; Noble *et al.*, 1997).

The traditional approach (Christensen *et al.*, 1985; McCusker, 1991; Cheetham & Wilkinson, 1991, 1992; Rudolf, 1993; Harris & Tremayne, 1996; Langford & Louër, 1996; Poojary & Clearfield, 1997) for solving crystal structures directly (*ab initio*) from powder diffraction data has been to extract the intensities $I(hkl)$ of individual reflections directly from the powder diffraction pattern and then to solve the structure by using these intensities $I(hkl)$ in the types of calculation (*e.g.* direct methods and the Patterson method) adopted for single-crystal diffraction data. However, this approach is associated with intrinsic difficulties, originating primarily from peak overlap in the powder diffraction pattern. Essentially, single-crystal and powder diffraction patterns contain the same information; however, in the case of single-crystal diffraction data, this information is distributed in three-dimensional space, whereas in the case of powder diffraction data the three-dimensional information is compressed into one dimension. As a consequence, there is usually considerable overlap of peaks in a powder diffraction pattern, leading to difficulties and ambiguities in extracting values of the relative intensities $I(hkl)$ of individual diffraction maxima that are sufficiently reliable to lead to successful structure solution.

Thus, as peak overlap in the powder diffraction pattern limits the potential for successful structure solution by the traditional approach, much attention has been devoted to the development of improved techniques for extracting accurate relative intensities for overlapping peaks. Recent progress in this regard includes the application of relations between the structure factors derived from direct methods and the Patterson function (Jansen *et al.*, 1992), the use of an iterative procedure involving calculation of a squared

ISSN 0108-7673

Patterson map and subsequent back-transformation to give a new set of structure factors for overlapped reflections (Estermann *et al.*, 1992; Estermann & Gramlich, 1993), the development of a method based on entropy maximization of a Patterson function (David, 1987, 1990) and the use of Bayesian fitting procedures (Sivia & David, 1994). We also note that the technique of entropy maximization and likelihood ranking has been implemented successfully in strategies for structure solution from powder diffraction data (Gilmore *et al.*, 1993; Gilmore, 1996). In this approach, groups of overlapping peaks are handled in a rational manner allowing intensity information for these peaks to be used productively, together with the intensities of the non-overlapping peaks, in the structure solution process.

In view of the intrinsic problems associated with extracting accurately the intensities $I(hkl)$ of individual reflections directly from the powder diffraction pattern, progress has also been made in recent years in the development of an alternative strategy for structure solution – the so-called 'direct-space' approach – in which the need to extract such intensity information from the powder diffraction pattern is avoided and the powder diffraction data is used directly in its 'raw' digitized form.

1.2. *The 'direct-space' strategy*

In the 'direct-space' strategy (Harris *et al.*, 1994; Harris, Kariuki & Tremayne, 1998; Harris & Tremayne, 1996) for crystal-structure solution from powder diffraction data, trial structures are generated in direct space, independently of the experimental data, with the suitability of each trial structure assessed by directly comparing the powder diffraction pattern calculated for the trial structure and the experimental powder diffraction pattern. This comparison can be quantified using the weighted profile R factor (R_{wp}, see §4.3), as used in Rietveld refinement. Importantly, R_{wp} considers the whole digitized intensity profile rather than the integrated intensities $I(hkl)$ of individual diffraction maxima. Thus, R_{wp} implicitly takes care of peak overlap (provided an appropriate peak-shape function is known). The 'direct-space' strategy does not require $I(hkl)$ values to be extracted from the experimental powder diffraction pattern and thus overcomes the major problem associated with the traditional approaches for structure solution.

In essence, the 'direct-space' strategy involves exploring a hypersurface, $R_{wp}(\mathbf{X})$, to find the best structure solution (lowest R_{wp}), where $\{\mathbf{X}\}$ represents the set of variables that define the structure. Expressed in this way, structure solution becomes equivalent to global optimization and therefore any technique for global optimization can, in principle, be used as a method for 'direct-space' structure solution.

In applying the 'direct-space' strategy, the structure is generally defined by a 'structural fragment', which represents an appropriately chosen collection of atoms within the asymmetric unit. The variables in $\{\mathbf{X}\}$ describe features such as the position, orientation and intramolecular geometry of the structural fragment. For example, for a structure comprising one molecule of well defined (rigid) geometry in the asymmetric unit, the set $\{\mathbf{X}\}$ may comprise six variables $\{x, y, z, \theta, \varphi, \psi\}$ defining the position and orientation of the structural fragment. Fewer variables may be required if elements of molecular symmetry and crystal symmetry coincide – for example, for a planar molecule lying in a mirror plane with fixed z coordinate, three variables $\{x, y, \psi\}$ are sufficient (where ψ represents rotation about the z axis). If the molecular conformation is not known with certainty beforehand, it is necessary to include a number of torsion angles $(\tau, \chi, \ldots)$ as variables within the structure solution calculation; thus more than six variables $\{x, y, z, \theta, \varphi, \psi, \tau, \chi, \ldots\}$ are required to define the structural fragment.

Initial work on direct-space structure solution focused on the development of Monte Carlo (Harris *et al.*, 1994; Harris, Kariuki & Tremayne, 1998; Kariuki *et al.*, 1996; Tremayne *et al.*, 1996*a*,*b*; Ramprasad *et al.*, 1995; Tremayne, Kariuki & Harris, 1997; Elizabé *et al.*, 1997; Tremayne, Kariuki, Harris, Shankland & Knight, 1997) and simulated annealing (Newsam *et al.*, 1992; Andreev *et al.*, 1996, 1997) methods for exploring the $R_{wp}(\mathbf{X})$ hypersurface. In the present paper, we describe a method for 'direct-space' structure solution based on the application of a genetic algorithm to explore the $R_{wp}(\mathbf{X})$ hypersurface.

In §2, we consider the general concepts underlying the application of genetic algorithms in global optimization, and then consider (§§3–5) details of our implementation of the genetic algorithm technique for structure solution from powder diffraction data. Opportunities for future extension and optimization of this technique are discussed in §§4 and 6.

2. Genetic algorithms

2.1. *Fundamentals*

The genetic algorithm (abbreviated as GA) is an optimization technique (Goldberg, 1989; Cartwright, 1993; Keane, 1996) based on the principles of evolution. The technique involves familiar evolutionary operations such as mating, mutation and 'natural selection', through which the fittest members of a population survive and procreate, passing their genetic information onto subsequent generations to produce descendants of improved quality. The use of GAs for global optimization relies on the fact that the GA should ultimately produce offspring that are optimal (or near to optimal) with respect to the criterion used to define the quality

('fitness') of the individual members of the population. In principle, the GA can be applied to any problem in which the quantity (G) to be optimized [such as potential energy or crystallographic R factor (see below)] can be written as a function of a string (set) of variables $\Gamma = \{\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_n\}$.

By analogy with biological evolution, genetic nomenclature is often employed in discussing the GA method. Thus, the strings Γ are equivalent to chromosomes and the individual variables γ_i within a string correspond to genes. The specific values taken by these variables for a particular member of the population correspond, in genetic terminology, to alleles.

At the start of the GA calculation, a specified number of strings (comprising the initial population) are generated at random in order to provide a diverse starting point for the calculation. The number of strings in the population is typically of the order of tens or hundreds.

The fitness (F) of a string is a measure of its quality with respect to the property G being optimized and fitness can therefore be defined as an appropriate function of G. For example, in our application of GAs in structure solution from powder diffraction data, structures giving rise to good agreement between experimental and calculated diffraction data have high fitness. The evolution of the GA is such that the strings of highest fitness have the best chance of passing on their characteristics (*i.e.* genetic information) to the next generation. The fitness function should be able to provide good discrimination between good and bad strings within the population at all stages of the evolution of the population.

As in biological evolution, the population is allowed to evolve through subsequent generations *via* the procedures of mating, mutation and 'natural selection'. In the mating procedure (also known as crossover), pairs of strings (parents) are selected from the population on the basis of their fitness. The mating procedure is accomplished by cutting and splicing the strings representing the two parents; thus, new individuals (offspring) are produced by mixing genetic information (exchanging genes) from the two parents. The simplest mating procedure corresponds to single-point crossover, in which the parent strings are cut at a single position, and the cut segments are swapped. Thus, if mating between parent 1 $\{a_1, b_1, c_1, d_1, e_1\}$ and parent 2 $\{a_2, b_2, c_2, d_2, e_2\}$ is carried out by single-point crossover between the third (c_i) and fourth (d_i) genes, the two offspring $\{a_1, b_1, c_1, d_2, e_2\}$ and $\{a_2, b_2, c_2, d_1, e_1\}$ are generated. More complicated mating procedures, such as double-point crossover and extensions thereof, can also be carried out and should be advantageous when the strings comprise a large number of variables. Some of these alternative mating procedures are discussed in §4.5.2.

After generating the offspring, only the strings of highest fitness from the set of offspring and parents are chosen to form the next generation, analogous to the process of 'natural selection'. This step is crucial in leading to optimization within the GA method and ensures that the overall quality of the population increases from one generation to the next.

In the mutation procedure, strings are selected randomly from the population, and random changes are made to parts of their genetic information to generate new strings (mutants). The introduction of mutants within the population is necessary to maintain genetic diversity (*i.e.* to prevent in-breeding within the population) and to prevent the GA calculation converging on a non-optimal string (stagnation). In general, the introduction of mutants within the population allows new regions of parameter space to be explored. For a given population, certain regions of parameter space might not be accessible through the mating procedure alone, as mating does not introduce new genes into the population (instead, mating mixes the existing genetic information in different ways).

Advantages of the GA method for optimization (in comparison with alternative approaches based on minimization) include the fact that the members of the initial population are not required to be close to the global minimum, and the fact that the GA calculation does not stop when a local minimum on the hypersurface is reached. In general, the GA calculation is carried out for a specified number of generations or until some pre-selected criterion (based, for example, on the value of G for the best member of the population) is achieved.

An important consideration in applying the GA method is the size of the population. Populations that are too small tend to become dominated by a single string and its offspring, leading to loss of diversity and stagnation of the population. If the population is too large, on the other hand, the best strings may become overwhelmed by a large number of marginally poorer strings, leading to inefficient convergence towards the optimal string.

A necessary criterion for the applicability of the GA method in optimization is that certain combinations of variables in the string can be recognized to be associated with high (or low) fitness. These groups of genes are known as schemata, and their existence lies at the heart of the ability of the GA to improve the fitness of a population by mating. Clearly this is an important requirement governing the potential for success of the GA technique. Thus, if a subset of the genes within a given string is close to optimal but the other genes are not optimal, it is important that the GA calculation can recognize (*e.g.* on the basis of fitness) the existence of the subset of genes that are close to optimal and can retain and propagate this subset of genes in the subsequent evolutionary process. If such schemata do not exist, the GA method becomes little more than a random search procedure and will not be an efficient approach for global optimization.

Another crucial feature of the GA approach is that it operates effectively in a parallel manner; many different strings, and hence different regions of parameter space, are investigated simultaneously. Furthermore, information concerning different regions of parameter space is passed actively between the individual strings by the mating procedure, disseminating genetic information throughout the population. This exchange of information between strings (*i.e.* combining information from different regions of the hypersurface) is a significant feature that lends efficiency to the GA approach (it is noteworthy that this feature is absent in most alternative approaches for optimization, such as running a large number of Monte Carlo calculations in parallel over different regions of the hypersurface). The implicit parallel nature of the GA approach makes it an efficient and robust method for optimization and makes it particularly advantageous for the optimization of functions of many variables.

2.2. *Applications of genetic algorithms*

GAs have found many applications in science, engineering and business (Goldberg, 1989; Keane, 1996) and a number of applications in chemistry have been reported (Cartwright, 1993), ranging from studies of protein folding to conformational optimization of long-chain molecules. In molecular modelling applications, in which the aim is to locate the global minimum on a potential energy hypersurface, it has been found (Brodmeir & Pretsch, 1994) that the GA is generally superior to alternative approaches based on molecular dynamics calculations or Monte Carlo sampling. GAs are also finding increasing use in optimization of the structures of atomic clusters (Hartke, 1995; Deaven & Ho, 1995; Deaven *et al.*, 1996).

Recently, the opportunity of using GA techniques in structure solution from powder diffraction data was recognized independently and at the same time by two research groups (Shankland *et al.*, 1997; Kariuki *et al.*, 1997; Harris, Johnston, Kariuki & Tremayne, 1998; Harris, Kariuki, Tremayne & Johnston, 1998). As discussed above, in developing strategies for applying genetic algorithms, several aspects of the method may be implemented in different ways. In this regard, the approaches of the two groups differ in the details of the strategies for carrying out mating, mutation and natural selection in evolving the population from one generation to the next and differ in the choice of fitness function used to assess the quality of each structure within the population. In our approach, the fitness of each member of the population is determined using an appropriate function of R_{wp} (see §4.3); this choice gives emphasis to the issues discussed in §1.2 and the philosophy of using directly the digitized experimental powder diffraction data 'as measured'. Furthermore, as discussed in §4.3, the use of appropriate functions of R_{wp} (as opposed to R_{wp} itself) provides considerable scope for optimization of the GA strategy and allows the advantages of dynamic scaling. In the approach of Shankland *et al.* (1997), the fitness of members of the population is given by the figure of merit χ^2, which is based on the intensities of individual reflections $I(hkl)$ extracted from the powder diffraction pattern using the Pawley refinement procedure (Pawley, 1981; Toraya, 1993). It is important to emphasize that the definition of χ^2 incorporates the covariance matrix (as derived from the Pawley refinement) and the use of the covariance matrix in this way serves to overcome problems that may otherwise arise (see §1.1) when considering the intensities of individual reflections extracted from the powder diffraction pattern. Clearly an approach based on using χ^2 rather than R_{wp} to compare the experimental and calculated powder diffraction data allows a faster calculation of the fitness function. It is clear that, in implementing genetic algorithms in this field, several different aspects of the technique and strategy are open to optimization and different strategies may each be advantageous to a greater or lesser extent depending on the particular problem under investigation. In this paper, we discuss fundamental aspects of our implementation of the GA method for structure solution from powder diffraction data and highlight several opportunities for future developments and improvements. Three examples, encompassing both previously known and previously unknown crystal structures, are used to illustrate the strategies for applying this method.

3. Structures studied

To demonstrate the application of our GA method for structure solution from powder diffraction data, three examples are discussed. Two previously known structures (*para*-methoxybenzoic acid and formylurea) are considered as test cases.

The structure of *para*-methoxybenzoic acid ($P2_1/a$; $a = 16.97$, $b = 10.96$, $c = 3.97$ Å, $\beta = 98.1^\circ$) was solved previously (Tremayne *et al.*, 1996*a*) from powder X-ray diffraction data using the Monte Carlo method and was also determined previously (Colapietro & Domenicano, 1978) from single-crystal X-ray diffraction data. The structure of formylurea ($Pn2_1a$; $a = 16.82$, $b = 6.06$, $c = 3.67$ Å) was solved previously (Lightfoot *et al.*, 1992) from powder X-ray diffraction data using direct methods. The experimental powder X-ray diffraction patterns recorded for *para*-methoxybenzoic acid and formylurea in our previous studies of these materials (Lightfoot *et al.*, 1992; Tremayne *et al.*, 1996*a*) were also used for the GA structure-solution calculations reported here.

The structure solution of *ortho*-thymotic acid represented the first application of our GA method to solve a previously unknown structure. Measurement of the powder X-ray diffractogram for *ortho*-thymotic acid,

unit-cell determination ($a = 11.08, b = 8.15, c = 11.78$ Å, $\beta = 100.2°$) and space-group assignment ($P2_1/n$) have been discussed elsewhere (Kariuki *et al.*, 1997). There is one molecule of *ortho*-thymotic acid in the asymmetric unit.

4. The GA method for structure solution from powder diffraction data

4.1. *Preliminaries*

We now describe specific aspects of our GA method for structure solution from powder diffraction data, as embodied within the program *GAPSS* (Johnston *et al.*, 1997; Kariuki *et al.*, 1997; Harris, Johnston, Kariuki & Tremayne, 1998; Harris, Kariuki, Tremayne & Johnston, 1998). A schematic illustration of the method governing the evolution of the population from one generation to the next in this program is shown in Fig. 1 and the methodology is discussed in more detail in §§4.2–4.9. Before running the GA program, we require to know the lattice parameters and space group (determined directly from the powder diffraction pattern) and it is necessary to make an appropriate choice of the structural fragment (see §§1.2 and 4.2). As discussed in §§1.2 and 4.2, each structure in the population is characterized by a string of parameters $\{\mathbf{X}\}$. The values of these parameters are real numbers (note that most applications of GAs in other fields use strings of binary numbers).

4.2. *The structural fragment*

In the GA approach for structure solution, each member of the population is a trial crystal structure, defined by a set of variables $\{\mathbf{X}\}$, representing the position, orientation and internal geometry of the structural fragment. The choice of structural fragment for any particular problem is not necessarily unique. Ideally, the structural fragment should include all significantly scattering atoms in the asymmetric unit (*i.e.* all non-H atoms in the case of powder X-ray diffraction) but in many cases it may be desirable to omit certain atoms (to be found later by difference Fourier techniques) from the structural fragment in order to restrict the number of parameters in $\{\mathbf{X}\}$ (*i.e.* by limiting the number of internal degrees of freedom). The factors governing the choice of structural fragment are analogous to those discussed previously (Kariuki *et al.*, 1996; Harris & Tremayne, 1996; Elizabé *et al.*, 1997) in the context of the Monte Carlo technique for structure solution.

In the GA calculation for *para*-methoxybenzoic acid, the structural fragment (Fig. 2*a*) was a rigid unit comprising the C and O atoms of the benzoate group (C_6CO_2) and the O atom of the methoxy group. Standard geometries (bond lengths and bond angles) were used, with the carboxylic acid group simplified by setting the two C—O bond lengths to be equal and with all atoms constrained to lie in the same plane. With one molecule of *para*-methoxybenzoic acid in the asymmetric unit, the crystal structure is defined by six degrees of freedom, representing the position (x, y, z) of the centre of mass of the structural fragment and the orientation (θ, φ, ψ) of the structural fragment relative to a space-fixed axis system. Thus, each member of the population in the GA calculation is defined by a string of six variables: $\{\mathbf{X}\} = \{x, y, z, \theta, \varphi, \psi\}$.

For formylurea (Fig. 2*b*) and *ortho*-thymotic acid (Fig. 2*c*), the structural fragments comprised all non-H atoms of the molecule. Standard geometries (bond lengths and bond angles) were used, with the exception that for *ortho*-thymotic acid the lengths of the two C—O bonds in the carboxylic acid group were taken to be equal. For both formylurea (Fig. 2*b*) and *ortho*-thymotic acid (Fig. 2*c*), the structural fragment was allowed some degree of flexibility, with two internal degrees of freedom (internal rotations about specified bonds) considered in each case. The structural fragment in each case is defined by a string of eight variables: $\{\mathbf{X}\} = \{x, y, z, \theta, \varphi, \psi, \tau, \chi\}$. Three variables define the position (x, y, z), three variables define the orientation (θ, φ, ψ), and two torsion angles (τ, χ) define the intramolecular geometry of the structural fragment. For formylurea, τ and χ are torsion

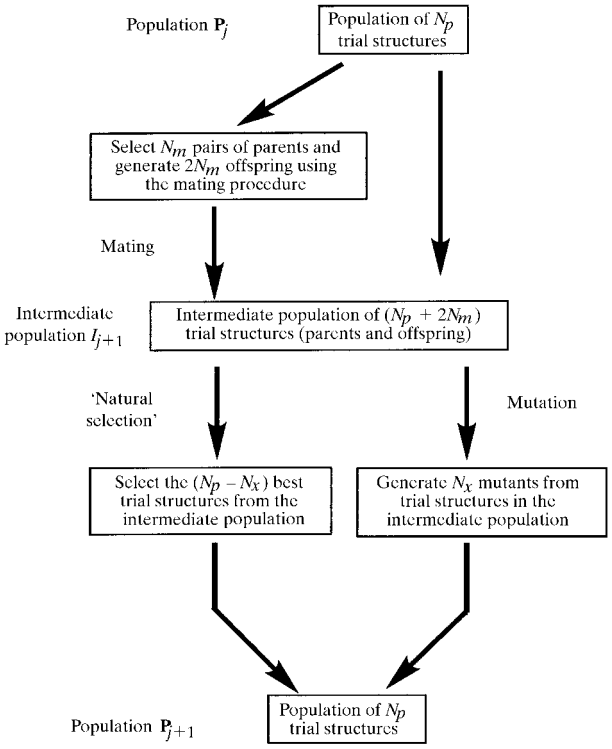

Fig. 1. Procedure for evolution of the population from one generation (population $\mathbf{P}_j$) to the next generation (population $\mathbf{P}_{j+1}$) in our GA method for structure solution from powder diffraction data.

angles for rotation about the two C—N bonds. For *ortho*-thymotic acid, τ is the torsion angle for rotation about the C—C bond between the carboxylic acid group and the benzene ring and χ is the torsion angle for rotation about the C—C bond between the isopropyl group and the benzene ring. All other torsion angles for *ortho*-thymotic acid were fixed such that the benzene ring and all atoms directly bonded to it were coplanar.

4.3. *The fitness function*

In our GA approach for structure solution, the probability of a given structure surviving into subsequent generations (through 'natural selection') and the probability of a given structure taking part in mating depend on its fitness. The fitness of a given structure depends on the weighted profile R factor [*i.e.* $F = f(R_{wp})$], which quantifies the level of disagreement between the calculated and experimental powder diffraction patterns. The profile R factor takes into consideration the entire powder diffraction profile (not the integrated intensities of diffraction maxima) and therefore implicitly takes care of the occurrence of peak overlap. Importantly, R_{wp} considers the digitized experimental diffraction data directly, with no manipulation of these data (as would be performed, for example, in extracting integrated peak intensities); every point in the digitized powder diffraction profile is considered as an individual intensity measurement. The weighted profile R factor compares the calculated powder diffraction pattern point by point against the experimental powder diffraction pattern, as follows:

$$R_{wp} = 100\left\{\frac{\sum_{i=1}^{N} w_i[y_i(\text{exp.}) - y_i(\text{calc.})]^2}{\sum_{i=1}^{N} w_i[y_i(\text{exp.})]^2}\right\}^{1/2},$$

where y_i(exp.) is the intensity of the ith data point in the experimental powder diffraction profile, y_i(calc.) is the intensity of the ith data point in the calculated powder diffraction profile and w_i is the weighting factor for the ith data point. For a given trial structure, the powder diffraction profile is calculated using the following information: (*a*) lattice parameters (to determine peak positions); (*b*) atomic positions and atomic displacement parameters (to determine peak intensities); (*c*) 2θ-dependent analytical functions to describe the peak shapes and peak widths; and (*d*) a description of the background intensity. The shape of a peak in a powder diffractogram depends on features of both the instrument and the sample and different types of peak-shape function are appropriate under different circumstances. The most widely used peak-shape function for powder X-ray diffraction data is the pseudo-Voigt function, which allows flexible variation of the Gaussian and Lorentzian character of the peak shape. Analytical functions are also used to describe the 2θ dependence of the peak width.

In our work so far using the GA method, three types of fitness function have been considered:

(i) exponential: $F(\rho) = \exp(-S\rho)$;

(ii) tanh: $F(\rho) = \frac{1}{2}\{1 - \tanh[2\pi(2\rho - 1)]\}$;

(iii) power: $F(\rho) = 1 - \rho^n$;

where

$$\rho = (R_{wp} - R_{\min})/\Delta R, \quad \Delta R = R_{\max} - R_{\min}$$

and $R_{\min}$ and $R_{\max}$ are the lowest and highest values of R_{wp} in the current population, respectively. In each case, $F(\rho)$ takes its highest value [$F(\rho) = 1$] when $\rho = 0$ [*i.e.* when $R_{wp} = R_{\min}$] and takes its lowest value when $\rho = 1$ [*i.e.* when $R_{wp} = R_{\max}$]. The values of $R_{\min}$ and $R_{\max}$ are continually updated as the population evolves during the GA calculation and the fitness function is said to be 'dynamically scaled'. For a given fitness function, the ability to discriminate between different structures in the population depends on the value of the denominator ΔR in ρ, which can be regarded as a scaling factor. In general, as ΔR becomes smaller, there is a greater level of discrimination in fitness between a given pair of structures. The question of whether ΔR tends to increase or decrease as the GA evolves depends to a large extent on whether mutant structures (see §4.7) are included in determining $R_{\max}$. If mutant structures are omitted from the calculation of $R_{\max}$, ΔR will generally decrease as the GA proceeds, which is desirable for efficient dynamic scaling.

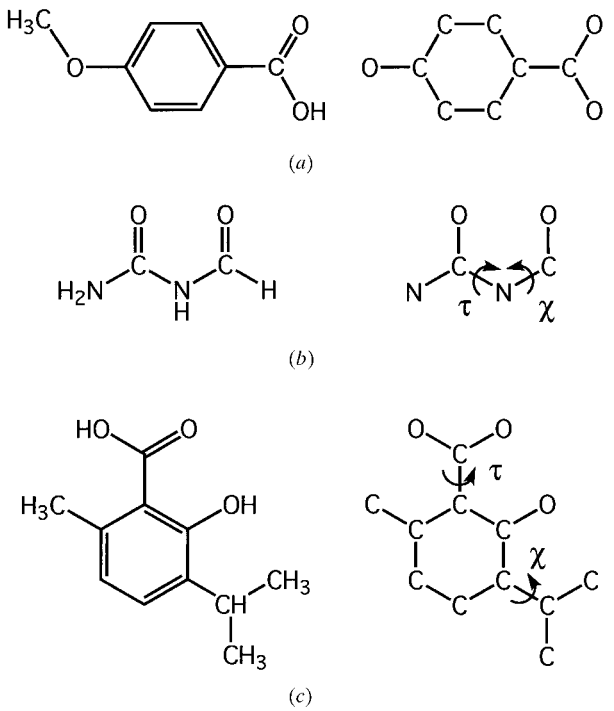

Fig. 2. Molecular structures and definitions of the structural fragments used in the GA structure solution calculations for: (*a*) *para*-methoxybenzoic acid; (*b*) formylurea; (*c*) *ortho*-thymotic acid.

All three fitness functions have maximum fitness $F(0) = 1$ when $R_{wp} = R_{\min}$ (*i.e.* for the best member of the population). For the power and tanh functions, the value of minimum fitness (*i.e.* for $R_{wp} = R_{\max}$) is $F(1) = 0$. For the exponential function, on the other hand, the minimum fitness (for $\rho = 1$) is $F(1) = \exp(-S)$ [thus, for $S \gtrsim 5$, the minimum fitness is $F(1) \lesssim 0.01$]. The major difference between the three types of fitness function considered concerns their behaviour for values of R_{wp} *between* $R_{\min}$ and $R_{\max}$, as shown in Fig. 3.

The exponential function is concave and becomes more concave as S increases. This function discriminates well between different good structures, as a wide range of values of $F(\rho)$ is covered by structures with low R_{wp} and hence low ρ (the curve is steepest at $\rho = 0$). However, the shallow nature of the curve around $\rho = 1$ means that a wide range of poor structures (*i.e.* with high values of ρ) all have very low fitness. For example, for the exponential function with $S = 5$ (see Fig. 3), all structures with $\rho \gtrsim 0.5$ have $F(\rho) \lesssim 0.1$.

The power function is convex for $n > 1$ (and becomes increasingly convex as n increases), linear for $n = 1$ and concave for $n < 1$. In general, the power function is used with $n > 1$. The power function with $n > 1$ gives good discrimination between poor structures (for which the curve is steepest) but little discrimination between good structures. For example, for the power function with $n = 3$ (see Fig. 3), all structures with $\rho \lesssim 0.5$ have $F(\rho) \gtrsim 0.9$. In this regard, the behaviour of the power function with $n > 1$ is directly opposite to the behaviour of the exponential function.

The tanh function presents different behaviour in that it does not discriminate significantly among good structures [*e.g.* all structures with $\rho \lesssim 0.3$ have $F(\rho)$ close to 1] and does not discriminate significantly among poor structures [*e.g.* all structures with $\rho \gtrsim 0.7$ have $F(\rho)$ close to 0]. However, the tanh function does give good discrimination between different structures in the intermediate region $0.3 \lesssim \rho \lesssim 0.7$. In this regard, we note the approximate step-like character of the tanh function. Another fitness function that has similar properties to the tanh function but is less steep in the intermediate region, is the cosine function $F(\rho) = \frac{1}{2}[1 + \cos(\pi\rho/2)]$.

All results presented in §5 were obtained using the tanh function. Our preliminary studies using the power, exponential and tanh functions suggest that the tanh function tends to produce structures with low R factor slightly more quickly than the power and exponential functions. The cosine function has not yet been assessed in our work.

Further studies are currently in progress to explore the optimum choices of fitness function for structural fragments of differing complexity and to assess the use of different types of fitness function at different stages during the GA calculation. For example, it may be advantageous to use the tanh function (which discriminates mainly between good structures and poor structures but provides little discrimination among different good structures) in the early stages of the GA calculation and to use the exponential function (which provides better discrimination between different good structures) in the later stages of the GA calculation.

4.4. *The overall GA strategy*

The initial population $\mathbf{P}_o$ ('zeroth' generation) for the GA calculation comprises N_p randomly generated structures. During the GA calculation, the population evolves through a sequence of generations, with a given population $\mathbf{P}_{j+1}$ (generation $j + 1$) generated from the previous population $\mathbf{P}_j$ (generation j) by the operations of mating, mutation and 'natural selection'. The overall scheme for generating population $\mathbf{P}_{j+1}$ from population $\mathbf{P}_j$ in our implementation of the GA method is summarized in Fig. 1. The number (N_p) of structures in the population remains constant for all generations, and N_m mating operations and N_x mutation operations are considered in the evolution of a given population $\mathbf{P}_j$ to generate the next population $\mathbf{P}_{j+1}$. Details of the different components of the scheme shown in Fig. 1 are discussed in §§4.5–4.9.

4.5. *The mating procedure*

4.5.1. *Selection of parents*. The probability of selecting a given structure to take part in mating (as a 'parent') is related to its fitness, such that structures with high fitness are more likely to be selected. In our procedure for selecting parents, a structure (with fitness F_s) is chosen

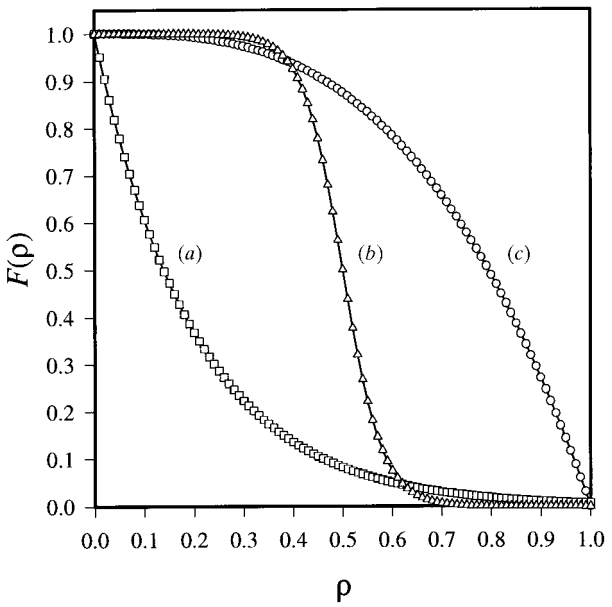

Fig. 3. Graphs showing the fitness functions discussed in the text: (*a*) exponential (for $S = 5$); (*b*) tanh; (*c*) power (for $n = 3$).

from the population at random and a random number R (with $0 \leq R \leq 1$) is generated. The randomly selected structure is then allowed to take part in mating if $F_s > R$. This selection procedure is continued to find a second structure that is allowed to mate with the first. Pairs of structures selected consecutively in this way are allowed to mate with each other, until the required number N_m of mating operations has been carried out. Note that a given structure could be selected several times as a parent for mating operations.

Our approach is a variant of the so-called 'roulette wheel' selection procedure (Goldberg, 1989), in which strings in a given generation ($\mathbf{P}_j$) are selected with probability proportional to their fitness and copied into a temporary population (a given string may be copied more than once). Mating and mutation operations are then performed by selecting strings from this temporary population. The main difference between our approach and this 'roulette wheel' approach is that our mating operation is carried out directly on consecutive pairs of strings in the order that they are selected from the population $\mathbf{P}_j$, whereas the 'roulette wheel' approach involves two stages: (i) selection of strings to form the temporary population, and then (ii) selection of pairs of strings from the temporary population as parents for mating. The order in which the strings are selected from population $\mathbf{P}_j$ in stage (i) of the 'roulette wheel' approach is unimportant, whereas the order of selecting strings from population $\mathbf{P}_j$ in our mating procedure is crucial. The relative merits of these two approaches have not yet been assessed.

An alternative approach (called 'rank selection') that may be used for selecting parents is to choose the N members of the population with highest fitness and to perform a specified number of mating operations by choosing parents at random from this subpopulation. This approach may avoid problems that can arise when the population contains a small number of structures of very high fitness; with our present method, such structures would tend to be selected very often as parents and would therefore dominate the mating procedure such that the offspring exhibited little diversity. The suitability of using the rank selection procedure within our GA approach for structure solution from powder diffraction data is currently being explored.

4.5.2. *The mating operation*. We now consider the actual methods that we have used to generate offspring by combining the parameters in the strings $\{\mathbf{X}\}$ that define the two selected parents.

For *para*-methoxybenzoic acid, with a rigid structural fragment defined by six parameters, mating was carried out by single-point crossover, with the strings for the two selected parents cut and spliced between the positional and orientational parameters to produce two offspring. Thus, the parents $\{x_a, y_a, z_a|\theta_a, \varphi_a, \psi_a\}$ and $\{x_b, y_b, z_b|\theta_b, \varphi_b, \psi_b\}$ lead to the two offspring $\{x_a, y_a, z_a|\theta_b, \varphi_b, \psi_b\}$ and $\{x_b, y_b, z_b|\theta_a, \varphi_a, \psi_a\}$. We note that, for a given pair of parents, this single-point crossover procedure will always generate the *same* pair of offspring. In using this mating procedure, it is clearly important to exclude the possibility that a given pair of parents is chosen more than once in a given generation.

We now assess some general points concerning the handling of a rigid structural fragment, as for *para*-methoxybenzoic acid. Although the use of single-point crossover between the positional and orientational parameters is attractive in view of the physical significance associated with separating the positional and orientational information, there is no guarantee that this is actually the most efficient approach for finding the optimal structure solution. We are currently carrying out a systematic investigation of this issue, considering single-point crossover at randomly selected positions within the string and multiple-point crossover. With these alternative mating procedures, a given pair of parents could produce several different pairs of offspring, overcoming the problem encountered (as discussed above) with single-point crossover at a fixed position within the string.

In the mating procedures for formylurea and *ortho*-thymotic acid, the eight parameters in each string were considered to comprise four groups $\{x, y, z|\theta, \varphi, \psi|\tau|\chi\}$. To carry out the mating operation between two selected parents, the four groups were divided into two sets of two groups. This can be performed in three different ways:

(i) $\{x, y, z|\theta, \varphi, \psi\}$ and $\{\tau|\chi\}$
(ii) $\{x, y, z|\tau\}$ and $\{\theta, \varphi, \psi|\chi\}$
(iii) $\{x, y, z|\chi\}$ and $\{\theta, \varphi, \psi|\tau\}$.

In a given mating operation, one of these ways of dividing the four groups was chosen (with equal probability) and two offspring were generated by taking the first set of two groups from the first parent and the second set of two groups from the second parent, and *vice versa*. Thus, mating the parents $\{x_a, y_a, z_a|\theta_a, \varphi_a, \psi_a|\tau_a|\chi_a\}$ and $\{x_b, y_b, z_b|\theta_b, \varphi_b, \psi_b|\tau_b|\chi_b\}$ will lead, with equal probability, to one of the following three pairs of offspring:

(i) $\{x_a, y_a, z_a|\theta_a, \varphi_a, \psi_a|\tau_b|\chi_b\}$
and $\{x_b, y_b, z_b|\theta_b, \varphi_b, \psi_b|\tau_a|\chi_a\}$
(ii) $\{x_a, y_a, z_a|\theta_b, \varphi_b, \psi_b|\tau_a|\chi_b\}$
and $\{x_b, y_b, z_b|\theta_a, \varphi_a, \psi_a|\tau_b|\chi_a\}$
(iii) $\{x_a, y_a, z_a|\theta_b, \varphi_b, \psi_b|\tau_b|\chi_a\}$
and $\{x_b, y_b, z_b|\theta_a, \varphi_a, \psi_a|\tau_a|\chi_b\}$.

Many other options exist for mating and each of these may be more or less advantageous in different circumstances. One promising mating procedure is to consider an appropriate weighted average (interpolation) of the corresponding parameters from the two parents,

leading from two parents $\{x_a, y_a, z_a|\theta_a, \varphi_a, \psi_a|\tau_a|\chi_a\}$ and $\{x_b, y_b, z_b|\theta_b, \varphi_b, \psi_b|\tau_b|\chi_b\}$ to one offspring $\{x_o, y_o, z_o|\theta_o, \varphi_o, \psi_o|\tau_o|\chi_o\}$ with $\xi_o = (1 - \lambda)\xi_a + \lambda\xi_b$, where ξ represents each of the parameters x, y, z, θ, φ, ψ, τ and χ. The parameter λ is in the range $0 < \lambda < 1$ and may in general depend on the relative values of fitness for the two parents.

4.6. *The intermediate population*

The number of mating operations in each generation is denoted N_m and, since each mating leads to two offspring, the number of offspring produced is $2N_m$. This creates an intermediate population ($\mathbf{I}_{j+1}$) comprising $N_p + 2N_m$ structures – *i.e.* N_p structures from the previous generation ($\mathbf{P}_j$) and $2N_m$ offspring generated by the mating procedure (taking parents from generation $\mathbf{P}_j$). After the N_m mating operations have been completed, the R_{wp} values for all offspring are calculated, new values of $R_{\min}$ and $R_{\max}$ are determined for the intermediate population, and the values of fitness for all members of the intermediate population are calculated. If two or more structures are identical, all but one of these structures is eliminated from the intermediate population. Note that although each of the N_p structures carried through to the intermediate population from the previous generation will have the same value of R_{wp} that it had in the previous generation, its value of fitness may change, as fitness depends on the current values of $R_{\min}$ and $R_{\max}$. The structures in the intermediate population are then ranked according to their fitness, in preparation for the 'natural selection' process (§4.8).

4.7. *The mutation procedure*

In each generation, a certain number of mutant structures are generated in order to maintain diversity within the population. In our mutation procedure, a specified number N_x of 'parent' structures are selected at random from the intermediate population, and a new mutant structure is generated from each selected 'parent' by introducing random changes to some aspects of its genetic information. It is important to note that the 'parent' structures used to create the mutants are not replaced by the mutants but remain within the intermediate population.

In principle, the mutation procedure could be introduced in several different ways within the overall scheme (Fig. 1) for converting population $\mathbf{P}_j$ to population $\mathbf{P}_{j+1}$. However, it is important (as in the scheme shown in Fig. 1) that the mutant structures are allowed the opportunity to take part in mating operations *before* the process of 'natural selection' is carried out. Thus, while several of the mutants will themselves not represent good structures (and will be rejected from the population at the first 'natural selection' step), they may nevertheless be able to pass useful genetic information into the population through the mating procedure.

For *para*-methoxybenzoic acid, mutation was carried out by assigning new random values to one (randomly selected) positional parameter (*i.e.* x, y or z) and one (randomly selected) orientational parameter (*i.e.* θ, φ or ψ) in each of the selected ('parent') structures. For formylurea and *ortho*-thymotic acid, mutation was carried out by randomly selecting two of the four groups of parameters $\{x, y, z|\theta, \varphi, \psi|\tau|\chi\}$ [this can be performed in six different ways (see §4.5.2)] and giving a new random value to one parameter within each of the selected groups. Thus, when selecting the group (x, y, z) or the group (θ, φ, ψ), only one (randomly selected) parameter of the three within the group was given a new value. This technique is known as static mutation, in that the mutations are generated by assigning completely new random values to one or more parameters for a selected structure.

An alternative technique is dynamic mutation, in which selected parameters are subjected to random *displacements* from their values in the 'parent' structure. Thus, for a particular parameter η in the set $\{\mathbf{X}\}$, the new (mutated) value η_m is given by

$$\eta_m = \eta_p + (\mathscr{R} \times \Delta\eta_{\max}),$$

where η_p is the value of η in the 'parent' structure, $\mathscr{R}$ is a random number between -1 and $+1$ and $\Delta\eta_{\max}$ is a maximum allowed displacement. Dynamic mutation is particularly useful for fine-tuning the population in the later stages of the GA calculation, when static mutation may cause too great a perturbation and may lead to structures with very low fitness (and a high probability of being rejected in the next generation). We are currently optimizing the strategy of starting with static mutation and then introducing dynamic mutation in the later stages of the GA calculation.

4.8. *'Natural selection'*

The population in the $(j + 1)$th generation ($\mathbf{P}_{j+1}$) is produced by taking the $N_p - N_x$ best (highest fitness) members of the intermediate population $\mathbf{I}_{j+1}$ together with the N_x mutant structures. The values of R_{wp} for the mutants are calculated and the new values of $R_{\min}$ and $R_{\max}$ for population $\mathbf{P}_{j+1}$ are evaluated, allowing the fitness of each structure in the new population $\mathbf{P}_{j+1}$ to be determined. The complete cycle involving mating, mutation and 'natural selection' is then repeated for a specified number (N_g) of generations or until some predetermined termination criterion (*e.g.* based on reaching a sufficiently low value of $R_{\min}$) is satisfied.

Since the intermediate population $\mathbf{I}_{j+1}$ includes all N_p structures from the previous generation $\mathbf{P}_j$ together with the $2N_m$ new offspring, it is guaranteed that the value of R_{wp} for the best structure in population $\mathbf{P}_{j+1}$ must be less than or equal to the value of R_{wp} for the best structure in population $\mathbf{P}_j$. Thus, $R_{\min}$ cannot increase from one generation to the next. The population size (N_p) remains

constant from one generation to the next and the best structures in a given generation are almost certain to be carried forward into the next generation (*i.e.* it is unlikely that all offspring generated by mating will have higher fitness than the fittest members of the population in the previous generation).

Such approaches in which 'natural selection' is carried out on an intermediate population comprising all offspring and all structures from the previous generation are described as 'elitist'. The elitist approach has a number of advantages over possible alternatives, such as carrying forward only the best offspring to construct the next generation. An alternative elitist strategy is to maintain a 'pool' of structures from previous generations for occasional (random) reintroduction into the population. This approach may be particularly advantageous when the population improves from a good structure but starts to stagnate (*i.e.* to show no significant improvement) after a few generations.

Another method that may be used to ensure diversity of the population and avoid inbreeding is to split the population into a number of subpopulations which evolve quasi-independently. A small amount of mating is allowed between the subpopulations, allowing new genes occasionally to enter the 'gene pool' for a given subpopulation. We have not yet assessed this approach with regard to structure solution from powder diffraction data.

In summary, the process of 'natural selection' ensures that the best structures survive into successive generations. The overall quality of the population – assessed by the average (mean) value of R_{wp} (denoted $R_{\rm ave}$) for the population – generally improves from one generation to the next. However, if (as in the present case) mutants are included in the calculation of $R_{\rm ave}$, the value of $R_{\rm ave}$ may sometimes increase slightly on passing from one generation to the next.

4.9. *The choice of parameters for the GA calculation*

We have carried out some preliminary work to optimize the parameters N_p, N_m and N_x which are input into the GA calculation, although a more detailed optimization is currently in progress. For all three structures described here, the GA calculation involved a population size of 100 structures ($N_p = 100$). In each generation, 50 mating operations ($N_m = 50$) were carried out, giving rise to 100 offspring. The size of the intermediate population was 200 structures, of which the best 90 structures (*i.e.* $N_p - N_x$) were passed forward (through the process of 'natural selection') to the next generation. In each generation, 10 mutant structures ($N_x = 10$) were generated. Although all the GA calculations described here were run for $N_g = 100$ generations, far fewer than 100 generations were actually required for the correct structure solution to be obtained (see §5).

5. Examples of structure solution using the GA method

The progress of the GA structure solution calculation can be monitored by plotting the evolution of the best ($R_{\rm min}$) and average ($R_{\rm ave}$) values of R_{wp} as a function of the generation number (n_g); we refer to this as the evolutionary progress plot (EPP). Such plots are shown in Figs. 4, 5 and 6 for *para*-methoxybenzoic acid, formylurea and *ortho*-thymotic acid, respectively. In all cases, $R_{\rm min}$ and $R_{\rm ave}$ decrease rapidly in the early generations ($n_g \lesssim 10$) and the plots of $R_{\rm min}$ *versus* n_g for *para*-methoxybenzoic acid and formylurea show approximate convergence at $n_g = 43$ and $n_g = 19$, respectively. For *ortho*-thymotic acid, $R_{\rm min}$ appears to converge at $n_g = 39$ (corresponding to the beginning of a long plateau in the $R_{\rm min}$ *versus* n_g plot), although there is a further significant drop in $R_{\rm min}$ at $n_g = 80$ (see below). The general behaviour of the plots of $R_{\rm min}$ *versus* n_g

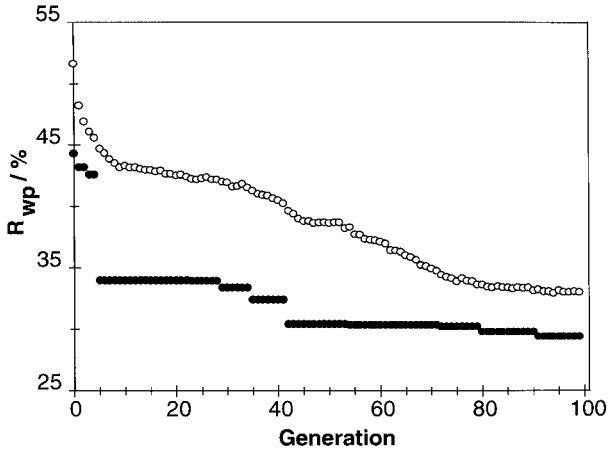

Fig. 4. Evolutionary progress plot for *para*-methoxybenzoic acid, showing the evolution of $R_{\rm min}$ (filled circles) and $R_{\rm ave}$ (open circles) as a function of generation number (n_g).

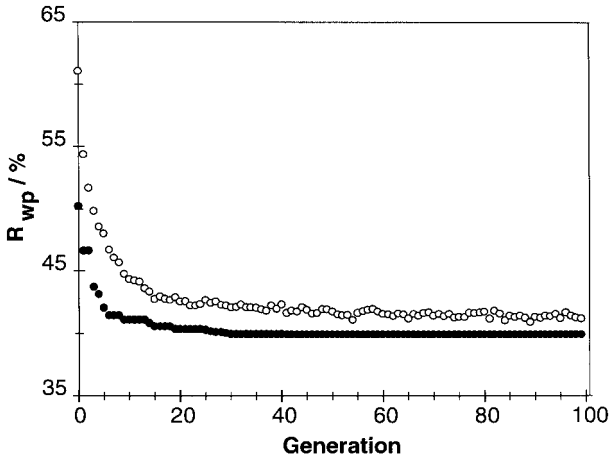

Fig. 5. Evolutionary progress plot for formylurea, showing the evolution of $R_{\rm min}$ (filled circles) and $R_{\rm ave}$ (open circles) as a function of generation number (n_g).

suggests that the most significant improvements in the quality of the best structure solution occur in the early stages of the GA calculation.

The crystal structures of *para*-methoxybenzoic acid and formylurea were known prior to the work described here and the success of the GA method is demonstrated by comparing (Figs. 7 and 8) the best structure found from the GA calculation [*i.e.* the structure corresponding to R_{min} in the final generation ($n_g = 100$)] and the position of the structural fragment in the known structure. In both cases, the best structure solution generated by the GA calculation is very close to the known structure – the maximum distance (d_{max}) and average distance (d_{ave}) between corresponding atom positions in the structure solution and the known structure are $d_{max} = 0.66$ Å and $d_{ave} = 0.47$ Å for *para*-methoxybenzoic acid (Fig. 7) and $d_{max} = 0.93$ Å and $d_{ave} = 0.44$ Å for formylurea (Fig. 8). In both cases, the structure solution found from the GA calculation refines readily to give the known structure on Rietveld refinement [in the present work, all Rietveld refinement calculations were carried out using the *GSAS* program (Larson & Von Dreele, 1987)].

The above discussion considered the best structures obtained at the end of the GA calculations (*i.e.* after 100 generations) and we now assess how early the correct structure solution has actually been obtained in these GA calculations. For *para*-methoxybenzoic acid (Fig. 4), it is found that the best structure after only five generations (at which R_{min} drops from 43 to 34%) refines to the known structure, whereas, for formylurea (Fig. 5), it is found that the best structure after six generations refines to the known structure. Thus, after a very small number of generations in both cases, the GA calculation has successfully located and discriminated a position, orientation and intramolecular geometry for the structural fragment refinably close to its true position, orientation and intramolecular geometry in the crystal structure. In this regard, it is important to emphasize that none of the (randomly chosen) starting structures was close to the correct structure (it is clear from Figs. 4 and 5 that the value of R_{min} for the zeroth generation is large in both cases).

For *ortho*-thymotic acid, the crystal structure was not known prior to the GA calculation and the quality of the structure solution can only be assessed by comparing the calculated and experimental powder diffraction profiles following full Rietveld refinement, as well as assessing the chemical and structural plausibility of the final refined structure. Rietveld refinement (Fig. 9) from the best structure solution obtained in the GA calculation gave $R_{wp} \approx 3.2\%$, and the final refined crystal structure is shown in Fig. 10. This structure, which has been discussed elsewhere (Kariuki *et al.*, 1997), is completely reasonable on structural and chemical grounds. For example, the structure is found to exhibit the familiar carboxylic acid dimer recognition motif, without this (or any other type of) intermolecular contact being imposed

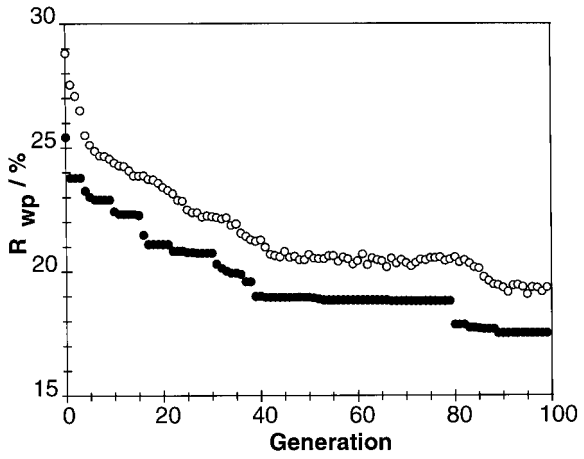

Fig. 6. Evolutionary progress plot for *ortho*-thymotic acid, showing the evolution of R_{min} (filled circles) and R_{ave} (open circles) as a function of generation number (n_g).

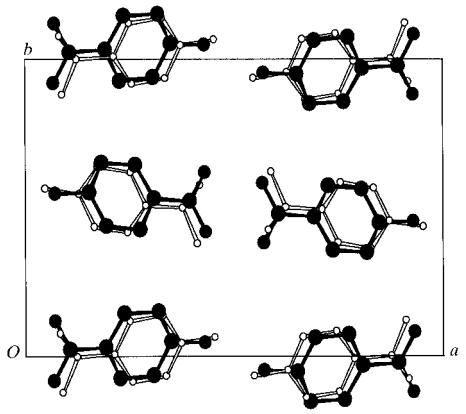

Fig. 7. Comparison between the position of the structural fragment in the best structure solution obtained in the GA calculation for *para*-methoxybenzoic acid (open circles) and the position of the corresponding atoms in the known crystal structure (filled circles).

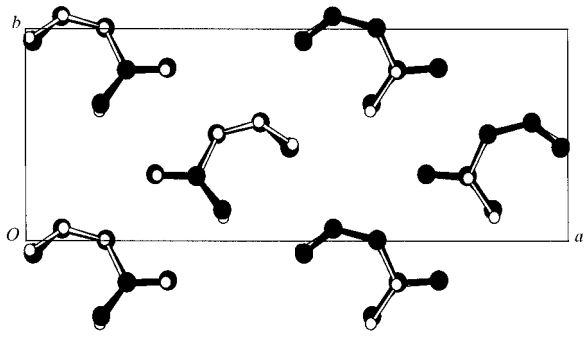

Fig. 8. Comparison between the position of the structural fragment in the best structure solution obtained in the GA calculation for formylurea (open circles) and the position of the corresponding atoms in the known crystal structure (filled circles).

during the GA calculation. The best structure solution in the plateau region (see Fig. 6) extending from $n_g = 39$ to $n_g = 79$ (with $R_{\min} \approx 19\%$) also refines to the same structure, emphasizing again that the correct structure solution has been found comparatively early in the GA calculation.

6. Concluding remarks

The results presented in this paper demonstrate the success of the GA method for crystal structure solution from powder diffraction data, particularly in the case of molecular crystals. In all cases, the correct structure solution was found readily and after a relatively small number of generations in the GA calculation. Indeed, preliminary comparisons suggest that the GA method may be a faster approach for finding the correct structure solution than the Monte Carlo technique. At the heart of this is the implicit parallel nature of the GA technique, which simultaneously considers a large number of structures (of the order of $N_p + 2N_m + N_x$) in each generation, spanning a wide range of parameter space, and passing information (through the mating operation) between different regions of parameter space. The Monte Carlo method, on the other hand, follows a single structure sequentially as it moves across the $R_{wp}(\mathbf{X})$ hypersurface. Systematic studies to assess the relative merits of the Monte Carlo and GA approaches are in progress.

With regard to future development and optimization of the GA approach, there are two basic strategies, both of which we are currently exploring: (i) to develop fundamental aspects of the GA technique, leading to new and optimized strategies and procedures for applying it to explore the $R_{wp}(\mathbf{X})$ hypersurface; and (ii) to consider new ways of defining the hypersurface such that global optimization may be achieved more efficiently using the existing GA methodology. We now consider each of these aspects briefly.

First, several plans for fundamental developments in the GA methodology (*e.g.* in relation to new definitions of fitness functions and new approaches for mating and mutation *etc.*) have been discussed in §4. Another approach that we are currently considering involves a rough local minimization (*i.e.* a few cycles of refinement) of R_{wp} with respect to the parameters in the set $\{\mathbf{X}\}$ for all structures generated during the GA calculation. With this approach, all structures in the population represent local minima (or close to local minima) on the $R_{wp}(\mathbf{X})$ hypersurface, facilitating identification of the global minimum and facilitating the identification (and removal) of members of the population that are in the same local minimum as each other. From applications of the GA method for global optimization in other fields (Brodmeir & Pretsch, 1994; Ho, 1997), it has been found

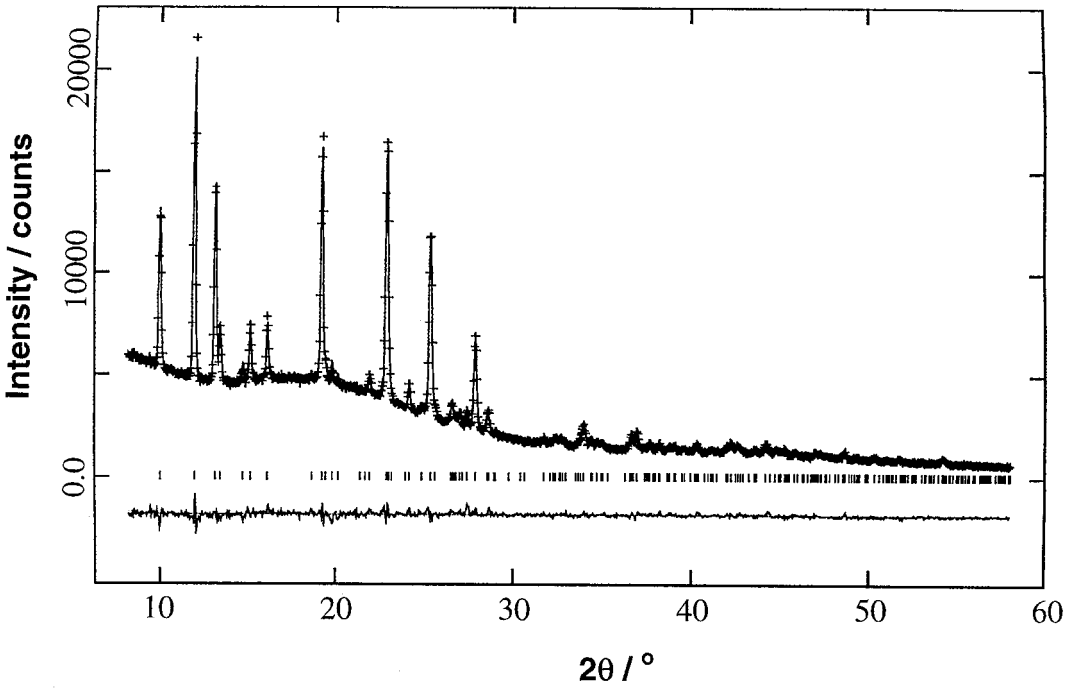

Fig. 9. Experimental (+ marks), calculated (solid line) and difference (lower line) powder X-ray diffraction profiles for the Rietveld refinement of *ortho*-thymotic acid. Reflection positions are marked. The calculated powder diffraction profile is for the final refined crystal structure, details of which are given in Kariuki *et al.* (1997).

that the GA works best when the members of the population are near local minima on the hypersurface.

Second, we consider some opportunities for redefinition of the hypersurface explored in the GA structure solution calculation. The applications of the GA method described in this paper considered the conventional weighted profile R factor R_{wp}, as used routinely in Rietveld refinement calculations. There is considerable scope for modifying the definition of R_{wp}, producing a hypersurface that may be more readily and more efficiently explored using our existing GA method. An important opportunity in this regard is to consider strategies in which the definition of R factor and/or the range of experimental data may be altered at different stages during the GA calculation. Thus, for example, it may be advantageous to focus on the low-angle (low-resolution) data at the start of the calculation, allowing the approximate position of the structural fragment within the unit cell to be established, and to introduce the higher-angle (higher-resolution) data progressively as more knowledge on the structural fragment emerges in the later stages of the calculation.

A significant advantage of considering $R_{wp}(\mathbf{X})$ (or redefinitions thereof) in the structure-solution calculation is that it is based purely on experimental data and is not biased by the introduction of any arbitrary parameters or assumptions. An alternative opportunity for redefining the hypersurface, however, is to combine the powder diffraction data with other 'direct-space' information, such as the computed potential energy $E(\mathbf{X})$. Thus, an alternative strategy for direct-space structure solution is to consider a new hypersurface $S(\mathbf{X})$, defined as an appropriate function of $E(\mathbf{X})$ and $R_{wp}(\mathbf{X})$: *i.e.* $S = \mathscr{F}(E, R_{wp})$. Provided a reliable potential energy parameterization is available for the system of interest (and is known *a priori* to be reliable), this type of combined approach may have significant advantages over the consideration of $R_{wp}(\mathbf{X})$ alone. The key to this approach lies in appropriate definition of the function $\mathscr{F}$, and in this regard we are currently exploring the optimization of this function for use in direct-space structure solution. We note, however, that it is only valid to consider $E(\mathbf{X})$ in structure solution calculations when the structural fragment represents a 'chemically sensible' unit, such as a complete molecule [this, of course, is not a limitation with regard to exploration of the $R_{wp}(\mathbf{X})$ hypersurface]. Finally, we note that GA methods using only the computed potential energy have been applied previously (Bush *et al.*, 1995) for crystal-structure prediction, and other alternative structure solution stategies based on consideration of computed energies have been reported (Hammond *et al.*, 1997).

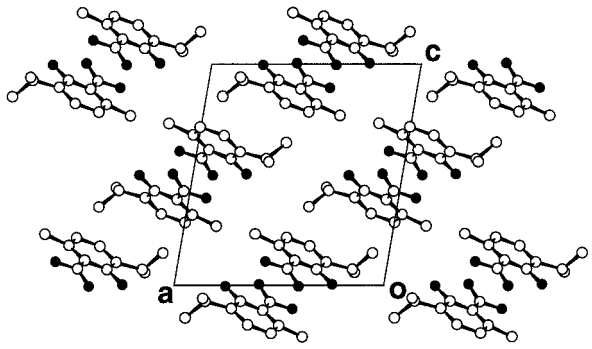

Fig. 10. Final refined crystal structure of *ortho*-thymotic acid (H atoms not shown) viewed along the *b* axis.

These future developments of the GA methodology, together with rigorous optimization of the strategies for its application, will significantly widen the scope and increase the efficiency of the GA approach for structure solution from powder diffraction data. Such developments will lead to faster and more reliable structure solution and will open up opportunities for exploring hypersurfaces of greater complexity.

We are grateful to EPSRC, Ciba Specialty Chemicals, the Royal Society and the University of Birmingham for financial support, and to Heliodoro Serrano-González, Katerina Psallidas, Dr Guy Brand, Dr Riccardo Poli and Dr Bill Langdon for discussions in connection with this work. Dr Maryjane Tremayne is thanked for recording some of the powder diffraction data discussed here.

References

Andreev, Y. G., Lightfoot, P. & Bruce, P. G. (1996). *Chem. Commun.* pp. 2169–2170.

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

Brodmeir, T. & Pretsch, E. (1994). *J. Comput. Chem.* **15**, 588–595.

Bush, T. S., Catlow, C. R. A. & Battle, P. D. (1995). *J. Mater. Chem.* **5**, 1269–1272.

Cartwright, H. M. (1993). *Applications of Artificial Intelligence in Chemistry.* Oxford University Press.

Cheetham, A. K. & Wilkinson, A. P. (1991). *J. Phys. Chem. Solids*, **52**, 1199–1208.

Cheetham, A. K. & Wilkinson, A. P. (1992). *Angew. Chem. Int. Ed. Engl.* **31**, 1557–1570.

Christensen, A. N., Lehmann, M. S. & Nielsen, M. (1985). *Aust. J. Phys.* **38**, 497–505.

Colapietro, M. & Domenicano, A. (1978). *Acta Cryst.* B**34**, 3277–3280.

David, W. I. F. (1987). *J. Appl. Cryst.* **20**, 316–319.

David, W. I. F. (1990). *Nature (London)*, **346**, 731–734.

Deaven, D. M. & Ho, K. M. (1995). *Phys. Rev. Lett.* **75**, 288–291.

Deaven, D. M., Tit, N., Morris, J. R. & Ho, K. M. (1996). *Chem. Phys. Lett.* **256**, 195–200.

Elizabé, L., Kariuki, B. M., Harris, K. D. M., Tremayne, M., Epple, M. & Thomas, J. M. (1997). *J. Phys. Chem. B*, **101**, 8827–8831.

Estermann, M. A. & Gramlich, V. (1993). *J. Appl. Cryst.* **26**, 396–404.

Estermann, M. A., McCusker, L. B. & Baerlocher, C. (1992). *J. Appl. Cryst.* **25**, 539–543.
Gilmore, C. J., (1996). *Acta Cryst.* A**52**, 561–589.
Gilmore, C. J., Shankland, K. & Bricogne, G. (1993). *Proc. R. Soc. London Ser. A*, **442**, 97–111.
Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* Reading, MA: Addison Wesley.
Gray, M. J., Jasper, J. D., Wilkinson, A. P. & Hanson, J. C. (1997). *Chem. Mater.* **9**, 976–980.
Hammond, R. B., Roberts, K. J., Docherty, R. & Edmondson, M. (1997). *J. Phys. Chem. B*, **101**, 6532–6536.
Harding, M. M. (1996). *J. Synchrotron Rad.* **3**, 250–259.
Harding, M. M., Kariuki, B. M., Cernik, R. & Cressey, G. (1994). *Acta Cryst.* B**50**, 673–676.
Harris, K. D. M., Johnston, R. L., Kariuki, B. M. & Tremayne, M. (1998). *J. Chem. Res.* pp. 390–391.
Harris, K. D. M., Kariuki, B. M. & Tremayne, M. (1998). *Mater. Sci. Forum*, **278–281**, 32–37.
Harris, K. D. M., Kariuki, B. M., Tremayne, M. & Johnston, R. L. (1998). *Mol. Cryst. Liq. Cryst.* **313**, 1–14.
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.
Hartke, B. (1995). *Chem. Phys. Lett.* **240**, 560–565.
Ho, K. M. (1997). Personal communication.
Jansen, J., Peschar, R. & Schenk, H. (1992). *J. Appl. Cryst.* **25**, 237–243.
Johnston, R. L., Kariuki, B. M. & Harris, K. D. M. (1997). *GAPSS: Genetic Algorithm for Powder Structure Solution*, University of Birmingham, England.
Kariuki, B. M., Serrano-González, H., Johnston, R. L. & Harris, K. D. M. (1997). *Chem. Phys. Lett.* **280**, 189–195.
Kariuki, B. M., Zin, D. M. S., Tremayne, M. & Harris, K. D. M. (1996). *Chem. Mater.* **8**, 565–569.
Keane, A. J. (1996). *Modern Heuristic Search Methods*, edited by V. Rayward-Smith, I. Osman, C. Reeves & G. D. Smith, pp. 255–272. New York: Wiley.
Langford, J. I. & Louër, D. (1996). *Rep. Prog. Phys.* **59**, 131–234.
Larson, A. C. & Von Dreele, R. B. (1987). Los Alamos Laboratory Report No. LA-UR-86-748. Los Alamos, NM, USA.
Lightfoot, P., Tremayne, M., Harris, K. D. M. & Bruce, P. G. (1992). *J. Chem. Soc. Chem. Commun.* pp. 1012–1013.
McCusker, L. B. (1991). *Acta Cryst.* A**47**, 297–313.
Newsam, J. M., Deem, M. W. & Freeman, C. M. (1992). *Accuracy in Powder Diffraction II. NIST Special Publ.* No. 846, pp. 80–91. NIST, Gaithersburg, MD, USA.
Noble, G. W., Wright, P. A., Lightfoot, P., Morris, R. E., Hudson, K. J., Kvick, Å. & Graafsma, H. (1997). *Angew. Chem. Int. Ed. Engl.* **36**, 81–83.
Pawley, G. S. (1981). *J. Appl. Cryst.* **14**, 357–361.
Poojary, D. M. & Clearfield, A. (1997). *Acc. Chem. Res.* **30**, 414–422.
Ramprasad, D., Pez, G. B., Toby, B. H., Markley, T. J. & Pearlstein, R. M. (1995). *J. Am. Chem. Soc.* **117**, 10694–10701.
Rudolf, P. R. (1993). *Mater. Chem. Phys.* **35**, 267–272.
Shankland, K., David, W. I. F. & Csoka, T. (1997). *Z. Kristallogr.* **212**, 550–552.
Sivia, D. S. & David, W. I. F. (1994). *Acta Cryst.* A**50**, 703–714.
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. (1996*a*). *J. Appl. Cryst.* **29**, 211–214.
Tremayne, M., Kariuki, B. M. & Harris, K. D. M. (1996*b*). *J. Mater. Chem.* **6**, 1601–1604.
Tremayne, M., Kariuki, B. M. & Harris, K. D. M. (1997). *Angew. Chem. Int. Ed. Engl.* **36**, 770–772.
Tremayne, M., Kariuki, B. M., Harris, K. D. M., Shankland, K. & Knight, K. S. (1997). *J. Appl. Cryst.* **30**, 968–974.