communications in mathematical
match
and in computer chemistry

no. 38, October 1998

ISSN 0340-6253

MATCDY (38) 123-135 (1998)

Application of a Genetic Algorithm in structure determination from powder diffraction data

Benson M. Kariuki, Roy L. Johnston, Kenneth D. M. Harris,*
Katerina Psallidas, Shinbyoung Ahn and Heliodoro Serrano-González

September 16, 1998

School of Chemistry, University of Birmingham, Edgbaston,
Birmingham, B15 2TT, United Kingdom

Abstract

This paper describes the development of a Genetic Algorithm for solving crystal structures directly from powder diffraction data. A number of examples are given to illustrate the application of this approach for the structure determination of molecular crystals. Our approach adopts the direct-space philosophy for structure solution from powder diffraction data, in which trial structures are generated independently of the experimental diffraction data and the quality of each trial structure is assessed by comparing the calculated and experimental powder diffraction patterns using the weighted profile R-factor, R_{wp}. The use of the profile R-factor in this approach implicitly takes care of the overlap of peaks in the powder diffraction pattern, and avoids the need to extract the intensities of individual reflections from the experimental powder diffraction data. In our Genetic Algorithm, a population of trial crystal structures is allowed to evolve subject to well-defined rules governing mating, mutation and natural selection, and the fitness of each structure is defined as a function of R_{wp}. The aim is to find the minimum value of R_{wp}, which corresponds to the best structure solution.

*Author for correspondence.

1 Introduction

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 for the material of interest. However, many crystalline solids (including several of academic and industrial importance) cannot be prepared as single crystals, and, in such cases structure determination is generally attempted using powder diffraction data.

It is relevant here to emphasize that crystal structure determination from diffraction data (either single crystal or powder) can be divided into three stages: (a) unit cell determination and space group assignment; (b) structure solution; and (c) structure refinement. The aim of the structure solution stage is to derive a good approximation to the crystal structure by direct consideration of the experimental diffraction data, but starting from no knowledge of the arrangement of atoms within the unit cell. If this approximate structural model is a sufficiently good representation of the true structure, a good quality crystal structure may then be obtained by refinement of this model against the experimental diffraction data in the structure refinement stage. For powder diffraction data, refinement of crystal structures can now be carried out fairly routinely (usually using the Rietveld profile refinement technique), whereas the solution of crystal structures directly from powder diffraction data is a significantly greater challenge. Here we focus on this structure solution stage of the structure determination process.

The traditional approach 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 using the types of structure solution calculation used for single crystal diffraction data (*e.g.* direct methods or Patterson methods). However, as there is usually extensive overlap of peaks in the powder diffraction pattern, it is often impossible to extract unambiguous values of the intensities of the individual diffraction maxima. To circumvent the problems associated with extracting the intensities of individual reflections directly from the powder diffraction pattern, progress has been made in recent years in the development of an alternative strategy for structure solution, the so-called 'direct-space' approach [1, 2], in which the powder diffraction data is used directly in its 'raw' digitized form.

In the direct-space strategy, trial crystal structures are generated in direct-space, with the suitability of each trial structure assessed by direct comparison between the powder diffraction pattern calculated for the trial structure and the experimental powder diffraction pattern. This comparison is quantified using the weighted profile R-factor (R_{wp}), which considers the whole digitized intensity profile rather than the integrated intensities of individual diffraction maxima, and thus implicitly takes care of peak overlap. Clearly, the aim is to find the trial crystal structure corresponding to the lowest value of R_{wp} and the direct-space strategy is therefore equivalent to exploring a hypersurface $R_{wp}(X)$ to find the best structure solution (lowest R_{wp}). Here $\{X\}$ represents the set ('string') of variables ('parameters') that define the structure. This paper describes the application of a Genetic Algorithm to accomplish global minimization with respect to the $R_{wp}(X)$

hypersurface.

The Genetic Algorithm (GA) is an optimization technique, based on the principles of evolution, and involves familiar evolutionary operations such as mating (crossover), mutation and natural selection. Through natural selection, the fittest members of a population survive and procreate, passing their genetic information into subsequent generations. GAs have found many applications in science, engineering and business [3]; chemical applications range from studies of protein folding to conformational optimization of long chain molecules [4] and geometry optimization of atomic clusters [5]. A crucial feature of the GA approach is that it operates essentially in a parallel manner, with many different regions of parameter space 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.

The possibility of using GA techniques in structure solution from powder diffraction data has been realized independently by two research groups [6, 7, 8, 9, 10]. Our approach described here and the approach of Shankland *et al.* [8] differ in the definition and handling of the fitness function as well as other aspects concerning the way in which the genetic Algorithm is implemented.

2 Methodology

Our GA approach for structure solution from powder diffraction data has been implemented in the program GAPSS [11]. A schematic flow chart describing the operation of this program is shown in Figure 1. The method and the program have been discussed in detail elsewhere (see in particular ref. [9]).

Before running GAPSS, it is necessary to know the lattice parameters and space group (obtained by indexing the diffraction pattern) and it is necessary to define a structural fragment [1, 2]. Ideally, the structural fragment should include all atoms with significant scattering power (*i.e.* all non-hydrogen atoms in the case of powder X-ray diffraction) within the asymmetric unit , but in many cases it may be desirable to omit certain atoms from the structural fragment in order to restrict the number of variables to be optimized (the omitted atoms may be found later by difference Fourier techniques).

Each member of the population is a trial crystal structure, defined by a set of variables {X}, representing the position (centre of mass or a pre-defined pivot atom) {x, y, z}, orientation $\{\theta, \phi, \psi\}$ and flexible torsion angles $\{\tau_1, \tau_2 \ ...\}$ of the structural fragment. The initial population P_0 comprises N_p randomly generated structures. In general, the population P_{j+1} (*i.e.* generation j+1) is produced from the previous population P_j by the operations of mating, mutation and natural selection. 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 performed during the evolution from population P_j to population P_{j+1}.

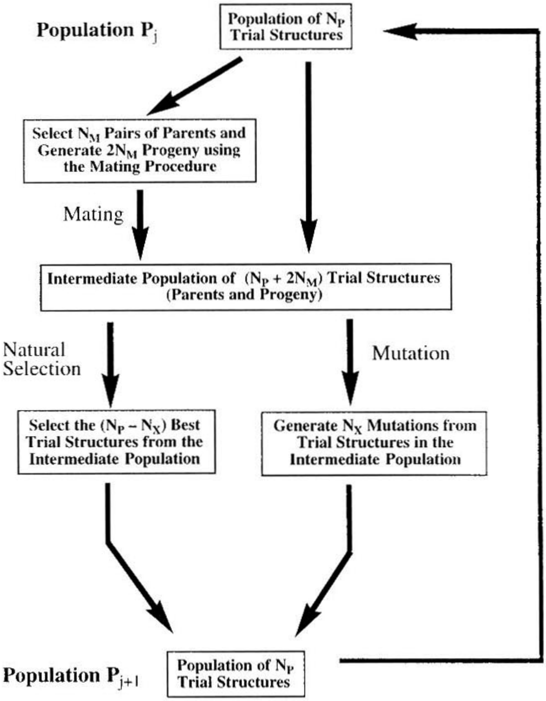

Population P_j
Population of N_P Trial Structures
Select N_M Pairs of Parents and Generate $2N_M$ Progeny using the Mating Procedure
Mating
Intermediate Population of ($N_P + 2N_M$) Trial Structures (Parents and Progeny)
Natural Selection
Mutation
Select the ($N_P - N_X$) Best Trial Structures from the Intermediate Population
Generate N_X Mutations from Trial Structures in the Intermediate Population
Population P_{j+1}
Population of N_P Trial Structures

Figure 1: Flow chart representing the evolution of the population from one generation (population P_j) to the next generation (population P_{j+1}) in the program GAPSS.

The probability of a given structure surviving into subsequent generations and taking part in mating depends on its fitness. In our GA, the fitness of a given trial crystal structure is defined as a function of its weighted profile R-factor, R_{wp}, which describes how well it fits the experimental powder diffraction data. To date, most of our work has used the following *tanh* fitness function (although exponential and linear functions have also been investigated [9]):

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

where

$$\rho = \frac{R_{wp} - R_{min}}{R_{max} - R_{min}} \tag{2}$$

and R_{min} and R_{max} are the lowest and highest values of R_{wp} in the current population [6, 7, 9, 10]. The appearance of R_{max}–R_{min} (*i.e.* the range of R_{wp} values in the current population) as the denominator of ρ ensures that the fitness function is dynamically scaled. The *tanh* fitness function was adopted because it discriminates strongly between good structures (low R_{wp} values) and poor structures (high R_{wp}), but does not discriminate strongly among different good structures or among different poor structures. We are currently investigating the potential advantages of altering the nature of the fitness function at different stages during the evolution of the GA calculation.

Prior to mating, a structure (with fitness F) is chosen from the population at random and is allowed to take part in mating if $F > R$, where R is a random number in the range 0 – 1. This selection procedure is repeated to find a second structure that is allowed to mate with the first, and so on, until N_m pairs of parents have been selected. This approach is a variant of the roulette wheel selection procedure [3]. Mating (crossover) is accomplished by combining parameters defining the genetic information of the two selected parents. The type of crossover adopted depends on the complexity of the structural fragment. Thus, for a rigid structural fragment, we have used single point crossover, with positional {x, y, z} and orientational $\{\theta, \phi, \psi\}$ parameters exchanged between the two parents. For a structural fragment with two torsional degrees of freedom, the eight parameters in each string are partitioned into four groups $\{x, y, z \mid \theta, \phi, \psi \mid \tau_1 \mid \tau_2\}$. Two offspring are generated from each mating operation; each offspring possesses two complementary groups from each parent, and three possible pairs of offspring may be generated in this way from a given pair of parents. Similarly, for a structural fragment with four torsional degrees of freedom, the ten parameters in each string are partitioned into six groups $\{x, y, z \mid \theta, \phi, \psi \mid \tau_1 \mid \tau_2 \mid \tau_3 \mid \tau_4\}$. Two offspring are generated from each mating operation, and each offspring now takes three complementary groups from each parent.

Since each mating operation leads to two offspring, a total of $2N_m$ offspring are produced in each GA cycle. This creates an intermediate population of (N_p+2N_m) structures (see Figure 1). The R_{wp} values are then calculated for all offspring, new values of R_{min} and R_{max} are determined for the intermediate population, and the new fitness values are calculated for all members of the intermediate population. If two or more structures are close to identical, all but one of these structures is eliminated from the intermediate population. The structures in the intermediate population are then ranked according to their fitness, in preparation for the natural selection process (see below).

In each generation, mutant structures are also generated in order to maintain diversity within the population. To do this, a specified number (N_x) of structures are selected at random from the intermediate population, and mutant structures are generated by making random changes to their genetic information. For structural fragments with two torsional degrees of freedom, mutation is carried out by randomly selecting two of the four groups of parameters $\{x, y, z \mid \theta, \phi, \psi \mid \tau_1 \mid \tau_2\}$ and assigning a new random value to one parameter within each of the selected groups. Similarly, for structural fragments with four torsional degrees of freedom, mutation is carried out by assigning new random values to one parameter in each of two groups selected at random from the six groups of parameters $\{x, y, z \mid \theta, \phi, \psi \mid \tau_1 \mid \tau_2 \mid \tau_3 \mid \tau_4\}$.

The structures within the population in the next generation are obtained by selecting the (N_p− N_x) best (highest fitness) members of the intermediate population together with the N_x mutant structures. This selection procedure (the analogue of 'natural selection' in biological evolution) is 'elitist', since the best structures survive into successive generations, irrespective of whether they are parents or offspring. Thus, R_{min} cannot increase from one generation to the next. The overall quality of the population, assessed by the average value of R_{wp} (denoted R_{ave}) for the population, generally improves from one generation to the next. The complete cycle (involving mating, mutation and natural selection) is repeated for a specified number (N_g) of generations, or until convergence is reached. For the examples reported here, the GA parameters had the following values: N_p = 100; N_m = 50 and 100 (for ortho-thymotic acid and L-glutamic acid respectively), N_x = 10; N_g = 100 (although far fewer than 100 generations are generally required to find the correct structure solution [6, 7, 9]).

3 Examples

Initially, to demonstrate the application of our GA method for structure solution from powder diffraction data, two previously known structures, para-methoxybenzoic acid and formylurea, were studied. These examples have been discussed in detail elsewhere [7, 9]. In both cases, our GA calculation found the correct structure solution within a very low number of generations (5 and 6 respectively). In this paper, we discuss the application of our GA to two known crystal structures containing a highly flexible molecule (the α and β phases of L-glutamic acid) as well as to a previously unknown crystal structure (ortho-thymotic acid).

3.1 The α and β phases of L-glutamic acid

The L enantiomer of glutamic acid $HO_2C(CH_2)_2CH(NH_2)CO_2H$ crystallizes as two polymorphs, known as the α and β phases. In both of these molecular solids the L-glutamic acid molecules exist as zwitterions, as shown in Figure 2a. Both structures are orthorhombic and have space group $P2_12_12_1$ [12, 13]. The unit cell parameters are: [α phase] a = 10.282 Å, b = 8.779 Å, c = 7.068 Å; [β phase] a = 5.159 Å, b = 17.300 Å, c = 6.948 Å. For both polymorphs, there is one molecule in the asymmetric unit.

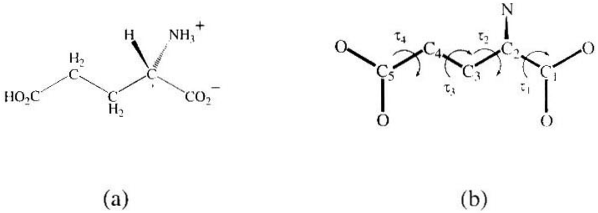


Figure 2: (a) Molecular structure of the L-glutamic acid zwitterion. The chiral centre is indicated by an asterisk. (b) Structural fragment used in the GA structure solution calculation for L-glutamic acid.

The structural fragment used in our GA stucture solution calculations for both polymorphs of L-glutamic acid comprised all of the non-hydrogen atoms of the molecule (Figure 2b). Standard geometries (bond lengths and angles) were used, with all C–O bond lengths taken to be equal. The four torsion angles (τ_1–τ_4) describing the flexibility of the L-glutamic acid molecule are indicated in Figure 2b. The position of the structural fragment was defined by the {x, y, z} coordinates of a pivot atom, taken as the central carbon atom (C_3) of the L-glutamic acid molecule (see Figure 2b). For these calculations, the positional, orientational and torsional parameters were all discretized, with grid sizes of 0.01 for all fractional coordinates and 10^o for all angles.

3.2 Ortho-thymotic acid

The structure solution of ortho-thymotic acid (Figure 3a) represents the first application of our GA method to solve a previously unknown structure [6]. Measurement of the powder X-ray diffractogram for ortho-thymotic acid, unit cell determination (a = 11.08 Å, b = 8.15 Å, c = 11.78 Å, β = 100.2°) and space group assignment ($P2_1/n$) have been discussed elsewhere [6]. There is one molecule in the asymmetric unit.

In our GA structure solution calculation for ortho-thymotic acid, the structural fragment

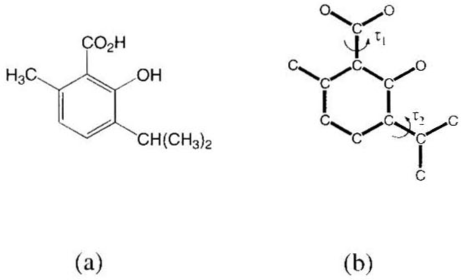


Figure 3: (a) Molecular structure of ortho-thymotic acid. (b) Structural fragment used in the GA structure solution calculation for ortho-thymotic acid.

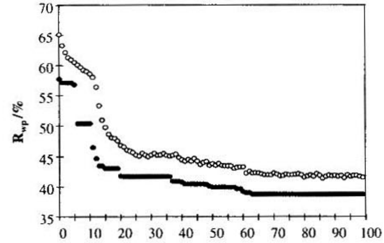
(Figure 3b) comprised all non-hydrogen atoms in the molecule. Standard geometries (bond lengths and bond angles) were used, with the lengths of the two C–O bonds in the carboxylic acid group taken to be equal. The structural fragment was allowed a certain degree of flexibility, defined by the two torsion angles τ_1 (describing rotation about the C–C bond between the carboxylic acid group and the benzene ring) and τ_2 (describing rotation about the C–C bond between the iso-propyl group and the benzene ring), as indicated in Figure 3b. The position of the structural fragment was defined by the {x, y, z} coordinates of the centre of mass of the molecule, and no discretization of the parameters was adopted.

4 Results

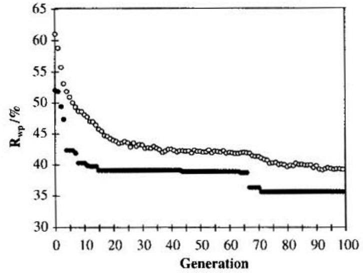
4.1 The α and β phases of L-glutamic acid

The progress of the GA structure solution calculation can be monitored by constructing an 'Evolutionary Progress Plot' (EPP), which shows the best (R_{min}) and average (R_{ave}) values of R_{wp} for the population as a function of the generation number. Figure 4 shows EPPs for the α (Figure 4a) and β (Figure 4b) phases of L-glutamic acid. For the β phase, there is a rapid initial drop in R_{min}, whereas for the α phase the EPP is more step-like. There is clearly rapid convergence of the GA structure solution calculation for both phases.

The best structure solutions (*i.e.* those with the lowest R_{wp} values after 100 generations) for the α and β phases are shown in Figure 5, where they are compared with the known crystal structures [12, 13] of the α and β phases. In both cases, the structure solution

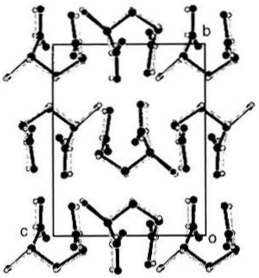


(a)

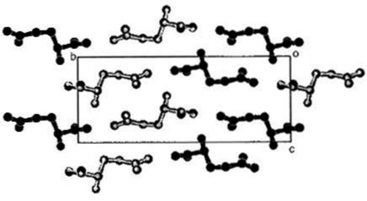


(b)

Figure 4: Evolutionary Progress Plot showing the evolution of R_{min} (filled circles) and R_{ave} (open circles), as a function of generation number, in the GA structure solution calculation for: (a) the α phase and (b) the β phase of L-glutamic acid.



(a)



(b)

Figure 5: Comparison between the position of the structural fragment in the best structure solution obtained in the GA structure solution calculation (open circles) and the positions of the corresponding atoms in the known crystal structure (filled circles) for: (a) the α phase and (b) the β phase of L-glutamic acid.

generated by the GA calculation is in excellent agreement with the known structure and, in each case, the structure solution refines easily (using the Rietveld refinement technique) to the known crystal structure. The maximum distance between an atom in the GA structure solution and the corresponding atom in the known crystal structure is less than 0.5 Å for both the α and β phases. From Figure 5, it is apparent that the L-glutamic acid molecules have significantly different conformations in the two phases [12, 13].

4.2 Ortho-thymotic acid

In the EPP for ortho-thymotic acid (Figure 6), R_{min} and R_{ave} both decrease rapidly in the early generations. R_{min} flattens out after the 39^{th} generation, with a further drop at the 80^{th} generation. For ortho-thymotic acid, the crystal structure was not known prior to the GA calculation, so the quality of the structure solution must be judged by comparing the calculated and experimental powder diffraction profiles following full Rietveld refinement, as well as by assessing the chemical and structural plausibility of the final refined structure. Rietveld refinement from the best structure solution obtained in the GA calculation gave R_{wp} = 3.2 %.

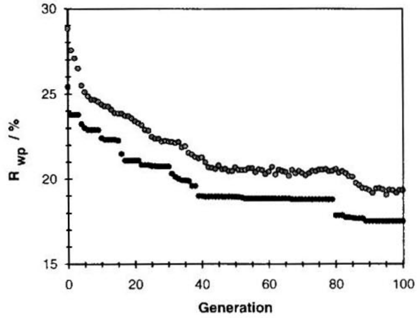


Figure 6: Evolutionary Progress Plot showing the evolution of R_{min} (dark circles) and R_{ave} (light circles), as a function of generation number, in the GA structure solution calculation for ortho-thymotic acid.

The final refined crystal structure, shown in Figure 7, which has been discussed elsewhere [6], is completely reasonable on structural and chemical grounds. For example, the structure is found to exhibit the familiar carboxylic acid dimer motif, without this (or any other type of) intermolecular contact being imposed during the GA structure solution calculation. The best structure solution in the plateau region (see Figure 6) extending from the 39^{th} to the 79^{th} generation (with $R_{min} \approx 19$ %) also refines to the same structure,

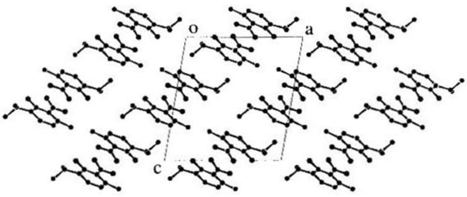


Figure 7: Final refined crystal structure of ortho-thymotic acid (hydrogen atoms not shown) viewed along the **b** axis.

indicating that the correct structure solution has been found relatively early in the GA calculation.

5 Concluding Remarks

We have demonstrated the successful application of a Genetic Algorithm within the framework of the direct-space approach for crystal structure solution from powder diffraction data. For all examples of structure solution considered, the correct structure solution was found 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 other direct-space techniques, although several aspects of our approach are currently undergoing rigorous optimization. With regard to future development and optimization of the GA approach, we are currently exploring two strategies [9]: (i) experimenting with fundamental aspects of the GA technique, leading to new and optimized procedures for exploring the $R_{wp}(X)$ hypersurface; and (ii) developing new ways of defining the hypersurface, so that global optimization may be achieved more efficiently.

Acknowledgments

The authors wish to thank the EPSRC, the University of Birmingham and Glaxo Wellcome plc for financial support.

References

[1] K. D. M. Harris, M. Tremayne, P. Lightfoot and P. G. Bruce, *J. Amer. Chem. Soc.* **116**, 3543 (1994).

[2] K. D. M. Harris and M. Tremayne, *Chem. Mater.* **8**, 2554 (1996).

[3] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley, Reading, MA, 1989).

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

[5] D. M. Deaven, N. Tit, J. R. Morris, and K. M. Ho, *Chem. Phys. Lett.* **256**, 195 (1996).

[6] B. M. Kariuki, H. Serrano-González, R. L. Johnston, and K. D. M. Harris, *Chem. Phys. Lett.* **280**, 189 (1997).

[7] K. D. M. Harris, R. L. Johnston, B. M. Kariuki, and M. Tremayne, *J. Chem. Res. (S)* 390 (1998).

[8] K. Shankland, W. I. F. David, and T. Csoka, *Z. Kristall.* **212**, 550 (1997).

[9] K. D. M. Harris, R. L. Johnston, and B. M. Kariuki, *Acta Cryst. A* **54**, 632 (1998).

[10] K. D. M. Harris, B. M. Kariuki, M. Tremayne, and R. L. Johnston, *Mol. Cryst. Liq. Cryst.* **313**, **1** (1998).

[11] R. L. Johnston, B. M. Kariuki, and K. D. M. Harris, "GAPSS: Genetic Algorithm for Powder Structure Solution", University of Birmingham (1997).

[12] M. S. Lehman, T. F. Koetzle and W. C. Hamilton, *J. Cryst. Mol. Struct.* **2**, 225 (1972).

[13] M. S. Lehman and A. C. Nunes, *Acta Cryst. B* **36**, 1621 (1980).