A Genetic Algorithm for Crystal Structure Solution from Powder Diffraction Data†

J. Chem. Research (S), 1998, 390–391†

Kenneth D. M. Harris,*[a] Roy L. Johnston,[a] Benson M. Kariuki[a] and Maryjane Tremayne[b]

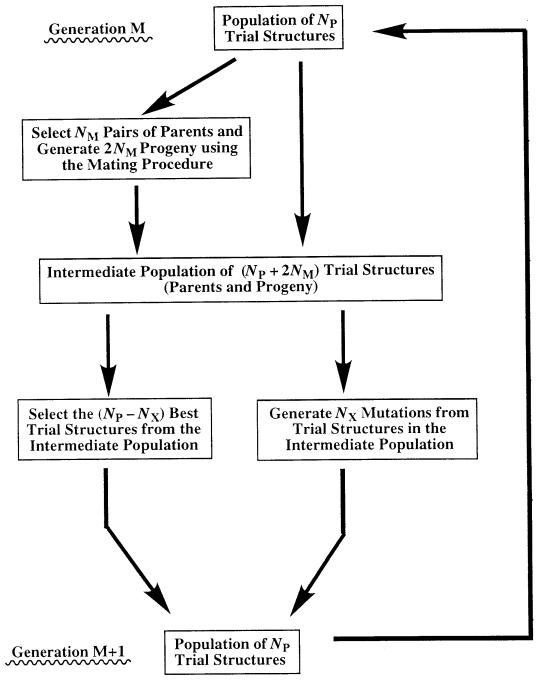
[a]*School of Chemistry, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK*
[b]*Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, UK*

A genetic algorithm is used as the basis of a new technique for crystal structure solution from powder diffraction data; two examples of successful structure solution illustrate the potential of this approach.

Solving crystal structures directly from *powder* diffraction data by traditional structure solution methods is associated with several difficulties, originating primarily from extensive peak overlap in the powder diffractogram. An approach to circumvent this problem[1,2] is to sample trial structures in direct space, and to assess their correctness by comparison (using the profile *R*-factor R_{wp}) between the diffractogram calculated for each trial structure and the experimental diffractogram. In essence, this approach involves searching the $R_{wp}(\mathbf{X})$ hypersurface to find the best structure solution (lowest R_{wp}), where $\mathbf{X}$ represents the set of variables that define the structure. Previously,[1–7] the success of Monte Carlo methods for exploring the $R_{wp}(\mathbf{X})$ hypersurface has been demonstrated. In this paper, we propose a new method[8] for exploring $R_{wp}(\mathbf{X})$, based on a Genetic Algorithm; two examples illustrate the success of this approach.

The Genetic Algorithm (GA) is an optimization technique based on evolutionary principles, in which the fittest members of a population survive and procreate, leading to improved subsequent generations.[9] In the present case, each member of the population is a trial crystal structure, defined by the position, orientation and internal geometry of a 'structural fragment' (representing an appropriate set of atoms within the asymmetric unit). The 'fitness' of each structure (labelled i) in the population depends on R_{wp}, and is quantified using the function $F(i) = 0.5[1 - \tanh(2\pi\{2\rho(i) - 1\})]$, where $\rho(i) = [R_{wp}(i) - R_{min}]/[R_{max} - R_{min}]$. Note that $F(i) = 1$ when $R_{wp} = R_{min}$ (the lowest R_{wp} in the current population) and $F(i) = 0$ when $R_{wp}(i) = R_{max}$ (the highest R_{wp} in the current population). The values of R_{min} and R_{max} are updated for each new generation of the population (see below). The fitness function defined above has been designed from our knowledge of the typical nature of R_{wp} hypersurfaces.

The initial population comprises N_P randomly generated structures, and the set $\mathbf{X}$ of variables that defines each structure can be regarded as its 'genetic code'. Subsequent generations of the population are produced through well-defined evolutionary procedures (Scheme 1). Mating involves selecting pairs of structures (with probability of selection proportional to fitness), and generating progeny by combining genetic information from the two parents. Any progeny that are identical to an existing structure in the population are deleted immediately, preventing premature convergence of the entire population towards a single structure. Diversity of the population is ensured by introducing a few mutant structures within each generation; these are generated by randomly selecting structures from the population and introducing random changes to parts of their genetic codes. Natural selection ensures that only the best structures survive and the overall fitness of the population improves from one generation to the next.



Scheme 1 Procedure for evolution of the population from one generation (M) to the next generation (M + 1) in the GA calculation

The success of our GA approach (using the program GAPSS[10]) is illustrated for structure solution of *p*-methoxybenzoic acid (**1**) and formylurea (**2**). The structures of **1** and **2** were solved previously from powder X-ray diffraction data (Monte Carlo method for **1**;[4] direct methods for **2**[11]), and the same data were used in the present work. Testing the new GA method using previously-known structures in this way allows a definitive assessment of its validity. In both cases, the GA calculation involved the evolution of 100 generations of a population of 100 structures (N_P). In each generation, 100 matings (N_M) and 10 mutations (N_X) were considered.

The GA calculation for **1** used a rigid structural fragment comprising the C and O atoms of the benzoate group (C_6CO_2) and the O atom of the methoxy group. Standard bond lengths and bond angles were used, with the two C—O bond lengths of the carboxylic acid group taken as equal. Thus, each structure was defined by six parameters $\{x_i, y_i, z_i, \theta_i, \phi_i, \psi_i\}$, representing the position of the centre of mass (x_i, y_i, z_i) and the orientation (θ_i, ϕ_i, ψ_i) of the rigid structural fragment. Mating was carried out by a single point cross-over, with the genetic codes of the two parents (i and j) cut between the positional and orientational par-

*To receive any correspondence (*e-mail:* K.D.M.Harris@bham.ac.uk).

†This is a **Short Paper** as defined in the Instructions for Authors, Section 5.0 [see *J. Chem. Research (S)*, 1998, Issue 1]; there is therefore no corresponding material in *J. Chem. Research (M)*.

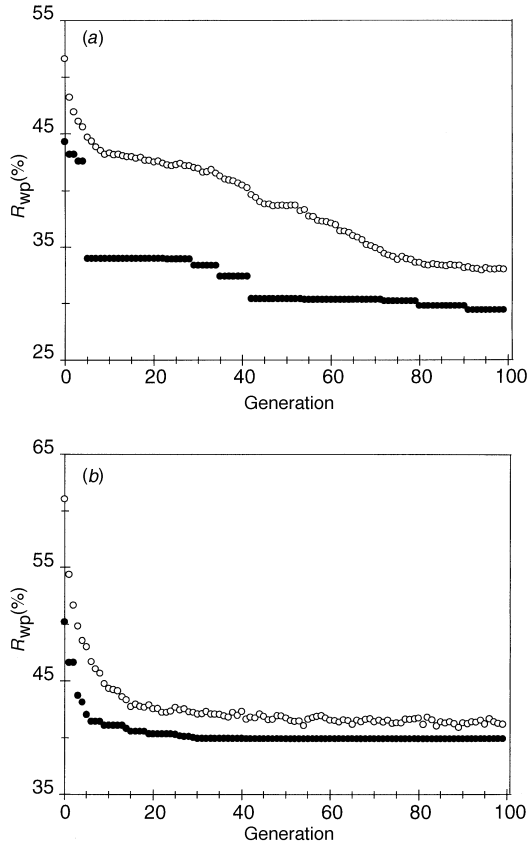


Fig. 1 The evolution of R_{wp} for the best structure in the population (filled circles) and the evolution of the average R_{wp} for the population (open circles) in the GA calculations for: (*a*) **1**; (*b*) **2**

ameters to produce two progeny $\{x_i, y_i, z_i, \theta_j, \phi_j, \psi_j\}$ and $\{x_j, y_j, z_j, \theta_i, \phi_i, \psi_i\}$. Mutations of selected structures were carried out by randomly changing one positional parameter and one orientational parameter.

The GA calculation for **2** used a flexible structural fragment comprising all non-H atoms, with standard bond lengths and bond angles. The internal degrees of freedom were the torsion angles for the two C—N bonds shown: H(CO)—(NH)—(CO)(NH_2). Thus, each structure was defined by eight parameters $\{x_i, y_i, z_i, \theta_i, \phi_i, \psi_i, \tau_i, \chi_i\}$, representing the six parameters described for **1** plus two torsion angles (internal degrees of freedom). For mating and mutation, the eight parameters were considered to comprise four groups $\{(x_i, y_i, z_i); (\theta_i, \phi_i, \psi_i); (\tau_i); (\chi_i)\}$. In mating, two groups were selected at random from one parent and combined with the other two groups taken from the other parent; again, the two progeny generated by cross-over were considered. Mutations involved making a random change to two of the four groups [for the groups (x_i, y_i, z_i) and $(\theta_i, \phi_i, \psi_i)$, only one of the three parameters was changed].

Fig. 1 shows the evolution of R_{wp} in the GA calculations, and Fig. 2 compares the best structure solution with the known structure. Clearly the GA approach has successfully located and discriminated a position for the structural fragment close to its position in the crystal structure. Subsequent Rietveld refinement of these structure solutions (and for **1**, location of the methoxy C atom by difference-Fourier methods) leads straightforwardly to the known crystal structures.

These results demonstrate the potential of the GA approach for structure solution from powder diffraction data, both for rigid and flexible structural fragments. Preliminary comparisons suggest that the GA approach may be significantly more efficient than the Monte Carlo technique (note from Fig. 1 that structures with low R_{wp} are generated very early in the evolutionary process), while we emphasize again the advantages of direct-space methods based on consideration of R_{wp} over the traditional techniques for structure solution from powder diffraction data. A rigorous optimization of our GA approach is currently in progress.

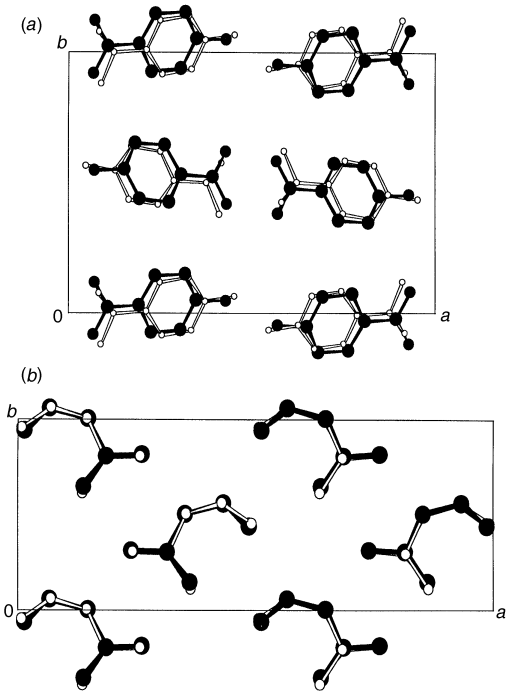


Fig. 2 Comparison between the position of the structural fragment in the best structure solution obtained in the GA calculation (open circles) and the position of the corresponding atoms in the known crystal structure (filled circles) for: (*a*) **1**; (*b*) **2**

We are grateful to EPSRC and Ciba Speciality Chemicals for financial support, and to Dr G Brand for helpful discussions.

Received (by the RSC, Cambridge), 17th June 1997
Transferred to J. Chem. Research, 24th March 1998
Accepted, 9th April 1998
Paper E/8/02319K

References

1 K. D. M. Harris, M. Tremayne, P. Lightfoot and P. G. Bruce, *J. Am. Chem. Soc.*, 1994, **116**, 3543.
2 K. D. M. Harris and M. Tremayne, *Chem. Mater.*, 1996, **8**, 2554.
3 J. M. Newsam, M. W. Deem and C. M. Freeman, *Accuracy in Powder Diffraction II; NIST Special Publ. No. 846*, 1992, p. 80.
4 M. Tremayne, B. M. Kariuki and K. D. M. Harris, *J. Appl. Crystallogr.*, 1996, **29**, 211.
5 M. Tremayne, B. M. Kariuki and K. D. M. Harris, *Angew. Chem., Int. Ed. Engl.*, 1997, **36**, 770 and references therein.
6 D. Ramprasad, G. P. Pez, B. H. Toby, T. J. Markley and R. M. Pearlstein, *J. Am. Chem. Soc.*, 1995, **117**, 10 694.
7 Y. G. Andreev, P. Lightfoot and P. G. Bruce, *Chem. Commun.*, 1996, 2169.
8 K. D. M. Harris, Grant GR/L72268, EPSRC, (submitted on 11th April 1997; awarded on 21st July 1997).
9 D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading (MA), 1989.
10 R. L. Johnston, B. M. Kariuki and K. D. M. Harris, *GAPSS*, 1997, Genetic Algorithm for Powder Structure Solution, University of Birmingham.
11 P. Lightfoot, M. Tremayne, K. D. M. Harris and P. G. Bruce, *J. Chem. Soc., Chem. Commun.*, 1992, 1012.