Molecular Crystals and Liquid Crystals Science and Technology. Section A. Molecular Crystals and Liquid Crystals

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gmcl19

Solving Crystal Structures from Powder Diffraction Data using Genetic Algorithms

Benson M. Kariuki [a], Kenneth D. M. Harris [a] & Roy L. Johnston [a]

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

Version of record first published: 24 Sep 2006

To cite this article: Benson M. Kariuki, Kenneth D. M. Harris & Roy L. Johnston (2001): Solving Crystal Structures from Powder Diffraction Data using Genetic Algorithms, Molecular Crystals and Liquid Crystals Science and Technology. Section A. Molecular Crystals and Liquid Crystals, 356:1, 469-481

To link to this article: http://dx.doi.org/10.1080/10587250108023725

Mol. Cryst. and Liq. Cryst., 2001, Vol. 356, pp. 469-481
Reprints available directly from the publisher
Photocopying permitted by license only

Solving Crystal Structures from Powder Diffraction Data using Genetic Algorithms

BENSON M. KARIUKI, KENNETH D.M. HARRIS
and ROY L. JOHNSTON

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

(Received September 9, 1999)

Many solids can be prepared only as microcrystalline powders and are not suitable for investigation by single crystal diffraction methods. To establish the structural properties of such materials, it is necessary to solve the crystal structure directly from powder diffraction data. Here we highlight fundamentals and applications of a method for structure solution from powder diffraction data in which a hypersurface based on the profile R-factor is searched using Genetic Algorithm techniques.

Keywords: powder diffraction; genetic algorithm; crystal structure solution

1 INTRODUCTION

The ability to solve crystal structures directly from powder diffraction data promises to open up many new avenues of structural science [1–3]. Many materials cannot be prepared as single crystals appropriate for conventional single crystal diffraction studies, and structure determination from powder diffraction may represent the only viable approach for

establishing the structural properties of such materials. However, significant challenges must be overcome in developing and applying methods for structure solution from powder diffraction data. This paper gives a brief overview of the problems and challenges encountered in this field, and focuses on our recent work involving the application of Genetic Algorithms.

Crystal structure determination from diffraction data (single crystal or powder) comprises the following stages: (1) unit cell determination and space group assignment, (2) structure solution, and (3) structure refinement. Structure solution starts from no knowledge of the arrangement of atoms or molecules in the unit cell, and aims to derive a good approximation to the crystal structure from direct consideration of the experimental diffraction data. If a sufficiently good structure solution is obtained, a high quality structure may then be derived by structure refinement. For powder diffraction data, refinement is usually carried out using the Rietveld profile refinement technique [4].

Techniques for powder structure solution can be divided into "traditional" and "direct-space" approaches, and here we focus only on the latter. In direct-space approaches, trial structures are generated (in direct space) independently of the experimental powder diffraction data, and the suitability of each trial structure is assessed by comparing the powder diffraction pattern calculated for the trial structure and the experimental powder diffraction pattern. This comparison is quantified using an appropriate R-factor. The aim of structure solution is to find the trial structure with lowest R-factor, and the process is equivalent to exploring a hypersurface R(**X**) to find the global minimum (structure solution), where {**X**} represents the set of variables that define the structure (see below). In principle, any technique for global optimization may be used, and there has been much success in using Monte Carlo [5–8] and Simulated Annealing [9–13] search algorithms in this field. Recently, Genetic Algorithms have also been applied [14–21] for direct-space structure solution from powder diffraction data. Here we describe fundamental and applied aspects of our implementations of Genetic Algorithms in this field.

In direct-space structure solution, the structure is defined by a "structural fragment", which represents an appropriate collection of atoms within the asymmetric unit. The variables in the calculation represent the position, orientation and intramolecular geometry of the structural fragment. The position is specified by the coordinates {x y z} of the centre of mass or a pre-defined pivot atom, the orientation is specified by

rotation angles $\{\theta\ \phi\ \psi\}$ around a set of orthogonal axes and the intramolecular geometry is specified by a set of variable torsion angles $\{\tau_1\ \tau_2 \ldots \tau_n\}$. In general, bond lengths, bond angles and any known torsion angles are fixed in the calculation, and are taken either from standard values for the type of molecule under study or from the known geometry of a similar molecule. 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) in the asymmetric unit, although it may be advantageous to omit certain atoms to reduce the number of variables (the omitted atoms may be found later by difference Fourier methods).

Most direct-space approaches, including the Genetic Algorithm method described in this paper, have used the weighted profile R-factor R_{wp} (the R-factor normally used in Rietveld refinement) to assess the agreement between calculated and experimental powder diffraction patterns. R_{wp} considers the whole digitized intensity profile directly "as measured", rather than the integrated intensities of individual diffraction maxima, and thus takes implicit consideration of peak overlap. Clearly the use of R_{wp} to assess the correctness of trial structures requires that the peak shape and peak width parameters used to construct the calculated powder diffraction pattern are consistent with those defining the experimental powder diffraction pattern (this can be ensured by prior analysis of the peak shapes and peak widths in the experimental powder diffraction pattern).

2 FOUNDATIONS OF GENETIC ALGORITHMS

The Genetic Algorithm (GA) is an optimization technique based on the principles of natural evolution, and involves the familiar evolutionary operations of mating, mutation and natural selection. By natural selection, the fittest members of a population survive and procreate, leading to improved individuals in subsequent generations, and leading ultimately to the optimal individual. The possibility of using GA techniques in structure solution from powder diffraction data was realized independently by two research groups. Our approach [14–18] and the approach of Shankland and co-workers [19–21] differ in the definition and handling of the fitness function as well as other aspects of the way in which the GA is implemented. Details can be found in the papers cited. Our GA approach is implemented in the program GAPSS [22], a schematic flow chart for which is shown in Figure 1.

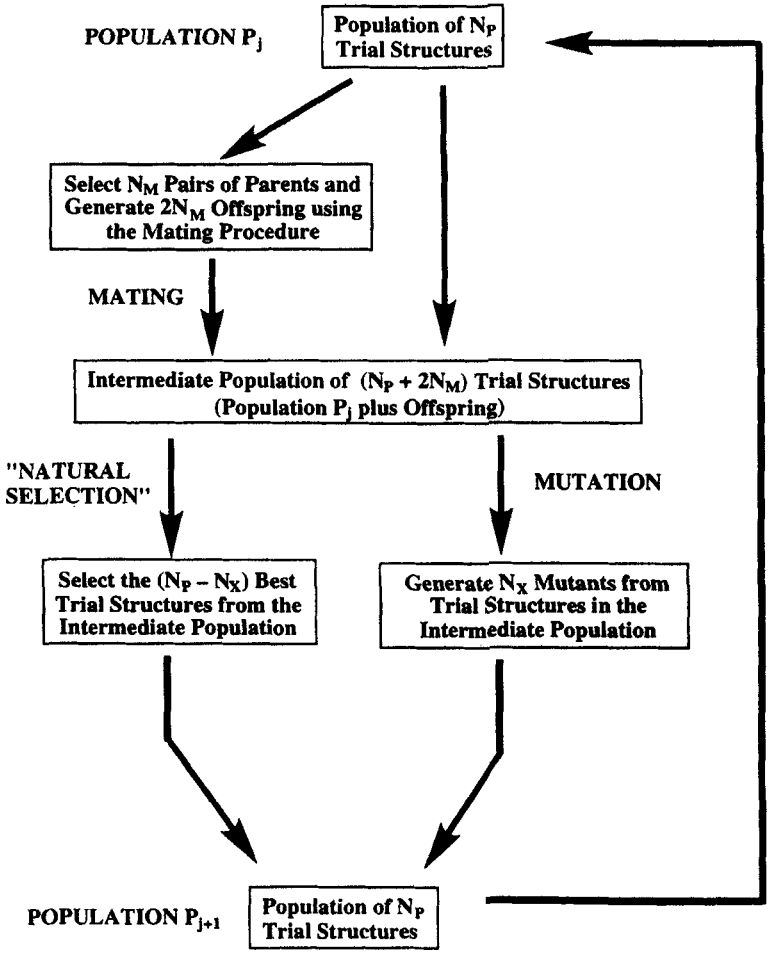

FIGURE 1 Flow chart showing the evolution of the population from one generation (population P_j) to the next generation (population P_{j+1}) in our GA strategy for powder structure solution.

In GA structure solution, the evolution of a population of trial crystal structures is investigated. Each member of the population is defined by a set $\{\mathbf{X}\}$ of variables representing its "genetic code". The initial population contains N_p randomly generated structures, and the population then evolves through subsequent generations by applying the operations of mating, mutation and natural selection. The number N_p of structures in the population is constant for all generations, and N_m mating operations and N_x mutation operations are performed in the evolution of the population from one generation to the next.

The probability that a given structure survives into subsequent generations and the probability that it takes part in mating depend on its "fitness". In our GA approach, the fitness (F) of a structure is a function of R_{wp}, and the following fitness functions have been considered:

$$F(\rho) = [1 - \tanh\{2\pi (2\rho - 1)\}]/2$$
$$F(\rho) = \exp(-S\rho)$$
$$F(\rho) = 1 - \rho^n \quad \text{(including the linear fitness function with } n = 1)$$

where $\rho = [R_{wp}-R_{min}]/[R_{max}-R_{min}]$, and R_{min} and R_{max} are the lowest and highest values of R_{wp} in the current population (note that $0 \leq \rho \leq 1$). The use of ρ in the definition of fitness function introduces the advantages of dynamic scaling.

Before each mating operation, a structure (with fitness F) is chosen from the population at random and is allowed to participate in mating if $F > \mathcal{R}$, where $\mathcal{R}$ is a random number generated between 0 and 1. A second structure is then found using the same selection procedure, and the two structures are allowed to mate with each other. This procedure is repeated until N_m pairs of parents have been selected. Mating (crossover) between two selected parents is carried out by combining the genetic codes of the two parents, although the actual strategy used depends on the complexity of the structural fragment.

As an example, for a structural fragment with two torsional degrees of freedom, one method for mating involves partitioning the eight variables that define each structure into four groups $\{x\ y\ z \mid \theta\ \phi\ \psi \mid \tau_1 \mid \tau_2\}$. Two offspring are then generated, with each offspring taking two complementary groups from each parent. This operation may be carried out in three different (but equi-probable) ways from a given pair of parents, and thus mating the parents $\{x_a\ y_a\ z_a \mid \theta_a\ \phi_a\ \psi_a \mid \tau_{1a} \mid \tau_{2a}\}$ and

$\{x_b\ y_b\ z_b \mid \theta_b\ \phi_b\ \psi_b \mid \tau_{1b} \mid \tau_{2b}\}$ leads with equal probability to one of the following pairs of offspring:

- $\{x_a\ y_a\ z_a \mid \theta_a\ \phi_a\ \psi_a \mid \tau_{1b} \mid \tau_{2b}\}$ and $\{x_b\ y_b\ z_b \mid \theta_b\ \phi_b\ \psi_b \mid \tau_{1a} \mid \tau_{2a}\}$
- $\{x_a\ y_a\ z_a \mid \theta_b\ \phi_b\ \psi_b \mid \tau_{1a} \mid \tau_{2b}\}$ and $\{x_b\ y_b\ z_b \mid \theta_a\ \phi_a\ \psi_a \mid \tau_{1b} \mid \tau_{2a}\}$
- $\{x_a\ y_a\ z_a \mid \theta_b\ \phi_b\ \psi_b \mid \tau_{1b} \mid \tau_{2a}\}$ and $\{x_b\ y_b\ z_b \mid \theta_a\ \phi_a\ \psi_a \mid \tau_{1a} \mid \tau_{2b}\}$

Many other methods for mating the two parents may be implemented, each with specific advantages under particular circumstances.

Each mating operation leads to two offspring, and a total of $2N_m$ offspring are therefore produced in each generation. These offspring together with all N_p structures from the previous generation give rise to an intermediate population of (N_p+2N_m) structures. The structures in the intermediate population are ranked according to fitness, in preparation for the natural selection process (described below). If two or more structures are identical within pre-defined tolerance limits, all but one of these structures is eliminated from the intermediate population.

In each generation, some mutant structures are also generated in order to maintain diversity within the population (we note that mutations create new genetic material within the population, whereas mating redistributes the existing genetic material). In our mutation procedure, N_x structures are selected at random from the intermediate population, and mutants are generated from these structures by making random changes to one or more of the variables in the set $\{\mathbf{X}\}$. The random changes may represent new random values (static mutation) or small random displacements from the existing value (dynamic mutation).

As shown in Figure 1, the next generation is constructed by taking the (N_p-N_x) members of highest fitness from the intermediate population (analogous to "natural selection" in biological evolution) together with the N_x mutant structures generated from the intermediate population. On passing from one generation to the next, the best structure in the population (corresponding to R_{min}) must either improve or remain the same, and the overall quality of the population (assessed from the average value of R_{wp}) usually improves. The complete evolutionary cycle involving mating, mutation and natural selection is repeated for a specified number (N_g) of generations, or until convergence is reached.

An important requirement for GA methods to lead to efficient global optimization is that the evolutionary procedure must be able to recognize certain combinations of variables (known as schemata) that are associated

with high fitness. Thus, if a subset of the variables defining a member of a population is close to optimal, but the other variables are not optimal, it is important that the GA calculation can identify the specific subset of variables that is close to optimal and can retain and propagate this subset of variables in the evolutionary process. In developing efficient implementations of GA methods, it is important to understand the nature of schemata in the powder structure solution problem. The variables that define the structure [i.e. $\{x\ y\ z\ \theta\ \phi\ \psi\ \tau_1\ \tau_2\ ...\ \tau_n\}$ in the implementation described above] could actually be defined in different ways, and schemata may be expressed more strongly for one definition of variables than another (even though the different choices of variables define the same structure). It is also important that the mating procedure is defined such that schemata are preserved and propagated, as an inappropriate crossover method may actually serve to destroy schemata that exist in the population. Clearly the existence and proper handling of schemata facilitates the GA calculation to reach the global minimum in R-factor rapidly.

In comparison with other approaches for global optimization, it is relevant to note that GA methods operate in a parallel manner, with many different regions of parameter space investigated simultaneously. Furthermore, a feature of GA methods is that information from different regions of parameter space is passed actively between different members of the population by the mating operation.

In the implementation of GA techniques, there is considerable scope for diversity in the methods and rules for carrying out the different evolutionary operations and in the definition of fitness function. Furthermore, details of the flow-chart shown in Figure 1 may differ from one implementation to another. In order to fully understand how to optimize the GA approach for particular types of structural problem or particular sets of diffraction data, it is necessary to consider and assess a wide range of strategies for carrying out the different evolutionary operations [16].

3 EXAMPLES OF APPLICATIONS

Several structures of varying degrees of complexity have been solved from powder diffraction data using our GA method, and two case studies are highlighted here: (i) structure solution of two previously-known structures (polymorphs) of L-glutamic acid, for which the R_{wp} hypersurface is

defined by 10 variables; (ii) structure solution of the previously-unknown structure of heptamethylene-bis(diphenylphosphane oxide), for which the R_{wp} hypersurface is defined by 18 variables.

3.1 CASE STUDY 1: POLYMORPHS OF L-GLUTAMIC ACID

L-glutamic acid $HO_2C(CH_2)_2CH(NH_2)CO_2H$ crystallizes in two different polymorphic forms, denoted the α and β phases. In both crystal structures [23,24], the molecules exist as zwitterions. Both structures have the orthorhombic space group $P2_12_12_1$ and the unit cells are: α phase, a = 10.28 Å, b = 8.78 Å, c = 7.07 Å; β phase, a = 5.16 Å, b = 17.30 Å, c = 6.95 Å. Both structures have one molecule in the asymmetric unit.

Powder X-ray diffraction patterns were recorded at ambient temperature in transmission mode on a Siemens D5000 diffractometer, using Ge-monochromated $CuK_{\alpha 1}$ radiation and a linear position-sensitive detector covering 8° in 2θ. The total 2θ range was 3° to 50° (step size 0.02°), recorded over 2 hours.

For both polymorphs, the structural fragment used in the GA structure solution calculations comprised all non-hydrogen atoms of the molecule (Figure 2). Standard bond lengths and angles were used, with all C–O bond lengths taken to be equal (the C–O single and C=O double bonds may be assigned later during Rietveld refinement). Each structure was defined by 10 variables $\{x\ y\ z\ \theta\ \phi\ \psi\ \tau_1\ \tau_2\ \tau_3\ \tau_4\}$ (the torsion angles defining the molecular conformation are shown in Figure 2). The position of the structural fragment was defined by the coordinates of the central carbon atom (C_3). The positional, orientational and torsional variables were all discretized, with grid sizes of 0.01 for all fractional coordinates and 10° for all angles (note that most of our other applications of the GA method have not discretized the variables in this way).

The GA calculation involved 100 generations of a population of 100 structures. In each generation, 200 offspring (100 pairs of parents) and 10 mutations were generated, and the tanh fitness function was used. For mating and mutation, the 10 variables were considered in six groups $\{x\ y\ z \mid \theta\ \phi\ \psi \mid \tau_1 \mid \tau_2 \mid \tau_3 \mid \tau_4\}$. In the mating operation between two selected parents, the six groups from each parent were distributed between the two offspring with no restriction on which combination of groups may come from each parent (this was determined on a random basis for each mating operation). For mutations, two groups were selected at random and a random change was made to one variable in each of the selected groups.

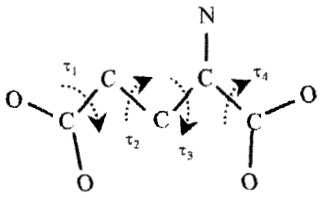

FIGURE 2 The structural fragment used in the GA structure solution calculations for the α and β phases of L-glutamic acid.

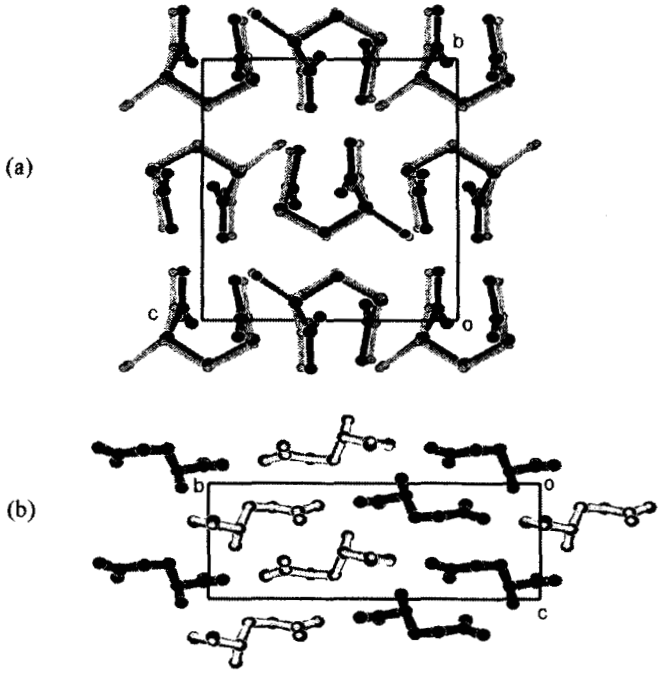

FIGURE 3 The position of the structural fragment in the best structure solution obtained in the GA structure solution calculation (light shading) and the positions of the corresponding atoms in the known crystal structure (dark shading) for (a) the α phase and (b) the β phase of L-glutamic acid.

The best structure solutions (i.e. with lowest R_{wp} in the final generation) for the α and β phases are shown in Figure 3, with the known crystal structures [23,24] also shown for comparison. The maximum distance between an atom in the structure solution and the corresponding atom in the known crystal structure is less than 0.5 Å in each case. For each phase, the structure solution is in excellent agreement with the known structure, and in each case the structure solution refines readily (using Rietveld refinement) to the known structure. Note that the molecular conformation is significantly different in the α and β phases, and the correct conformation has been found successfully by the GA calculation in each case.

3.2 CASE STUDY 2: $Ph_2P(O).(CH_2)_7.P(O)Ph_2$

Our structure determination of $Ph_2P(O).(CH_2)$ $.P(O)Ph_2$ [17] is one of the most complex molecular crystal structure that has been reported to be solved from powder diffraction data, involving a total of 18 degrees of freedom (12 variable torsion angles).

The powder X-ray diffraction pattern was recorded at ambient temperature in transmission mode on a Siemens D5000 diffractometer, using Ge-monochromated $CuK_{\alpha 1}$ radiation and a linear position-sensitive detector covering 8° in 2θ. The total 2θ range was 5° to 60° (step size 0.02°), recorded over 6 hours. The powder diffraction pattern was indexed by the program ITO [25], giving the monoclinic unit cell: a = 12.59 Å, b = 10.20 Å, c = 22.89 Å, β = 105.5°. From systematic absences, the space group was assigned as $P2_1/n$, and density considerations suggest that there is one molecule in the asymmetric unit.

In the GA structure solution calculation, the structural fragment comprised all non-hydrogen atoms of the molecule. Standard bond lengths and bond angles were used and the phenyl rings were constrained to be planar. The molecule was subjected to translation and reorientation within the unit cell, together with variation of all 12 torsion angles (Figure 4) defining the molecular conformation. Thus, each structure was defined by 18 variables $\{x\ y\ z\ \theta\ \phi\ \psi\ \tau_1\ \tau_2\ \tau_3\ ...\ \tau_{12}\}$.

The GA calculation involved 500 generations of a population of 100 structures. In each generation, 100 offspring (50 pairs of parents) and 20 mutations were generated and the linear fitness function was used. For mating and mutation, the 18 variables were subdivided into 14 groups $\{x\ y\ z \mid \theta\ \phi\ \psi \mid \tau_1 \mid \tau_2 \mid \tau_3 \mid ... \mid \tau_{12}\}$. In the mating operation between two

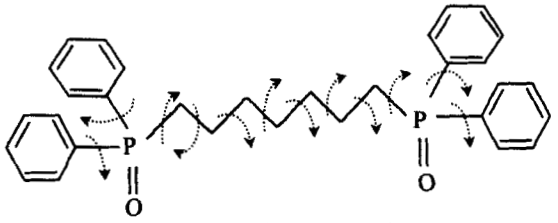

FIGURE 4 Molecular structure of $Ph_2P(O).(CH_2)_7.P(O)Ph_2$ showing the variable torsion angles in the GA structure solution calculation.

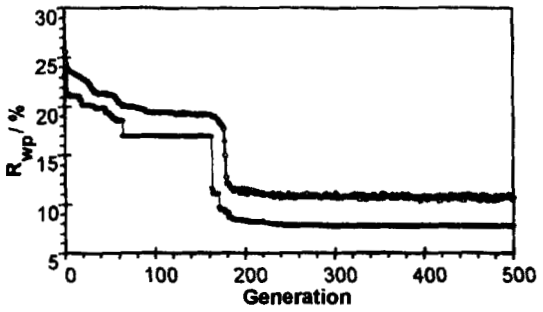

FIGURE 5 Evolutionary Progress Plot showing the evolution of the best (R_{min}, lower plot) and average (R_{ave}, upper plot) values of R_{wp} for the population as a function of generation number in the GA structure solution calculation for $Ph_2P(O).(CH_2)_7.P(O)Ph_2$.

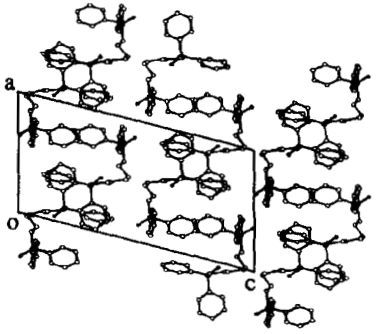

FIGURE 6 Final refined crystal structure of $Ph_2P(O).(CH_2)_7.P(O)Ph_2$ (hydrogen atoms not shown) viewed along the b-axis.

selected parents, the 14 groups from each parent were distributed between the two offspring, with no restriction on which combination of groups may come from each parent (this was determined on a random basis for each mating operation). For mutation, seven groups were selected at random and a random change was made to one variable in each of the selected groups.

Figure 5 shows the evolution of R_{wp} during the GA calculation, and demonstrates a systematic improvement in the overall quality of the population. The lowest value of R_{wp} in the population decreases progressively, with a particularly significant event at generation 163.

The structure with lowest R_{wp} in the final generation was taken as the starting model for Rietveld refinement using the GSAS program [26]. The positions of all non-hydrogen atoms were refined, with standard geometric restraints applied to bond lengths and angles. The final Rietveld refinement gave R_{wp} = 5.0 % and R_p = 3.8 %. In the crystal structure (Figure 6), the molecular conformation has one *gauche* bond in the $(CH_2)_7$ chain and the other parts of the chain are close to all-*trans* conformations. The crystal structure of $Ph_2P(O).(CH_2)_7.P(O)Ph_2$ is described in detail elsewhere [27].

4 CONCLUDING REMARKS

Although the successful application of GA techniques for powder structure solution has been demonstrated, considerable scope remains for further development and optimization of the strategies for implementing GA techniques in this field. In this regard, we are exploring fundamentals of the GA technique, to optimize the procedures for searching R(**X**) hypersurfaces, and developing new definitions of the hypersurface, to allow global optimization to be achieved more efficiently.

Solid state scientists now have access to many different techniques (both traditional and direct-space approaches) for structure solution from powder diffraction data. The future application of these techniques promises to reveal new and important insights into structural properties of solids across the full range of solid state and materials sciences.

Acknowledgements

We thank EPSRC, the University of Birmingham, the Nuffield Foundation, Glaxo Wellcome plc, Ciba Specialty Chemicals and MSI for support of our research. Collaborations with Dr S. Ahn, Miss P. Calcagno and Dr D. Philp are gratefully acknowledged.

References

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

[2] J.I. Langford, D. Louër, *Rep. Progr. Phys.*, **59,** 131 (1996).

[3] D.M. Poojary, A. Clearfield, *Acc. Chem. Res.*, **30,** 414 (1997).

[4] H.M. Rietveld, J. *Appl. Crystallogr.*, **2,** 65 (1969).

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

[6] D. Ramprasad, G.B. Pez, B.H. Toby, T.J. Markley, R.M. Pearlstein, *J. Amer. Chem. Soc.*, **117,** 10694 (1995).

[7] M. Tremayne, B.M. Kariuki, K.D.M. Harris, *Angew. Chemie, Int. Ed. Engl.*, **36,** 770 (1997).

[8] K.D.M. Harris, B.M. Kariuki, M. Tremayne, *Mat. Sci. Forum*, **278–291,** 32 (1998).

[9] J.M. Newsam, M.W. Deem, C.M. Freeman, *Accuracy in Powder Diffraction II: NIST Special Publ. No. 846,* pp. 80–91 (1992).

[10] Y.G. Andreev, P. Lightfoot, P.G. Bruce, *Chem. Commun.*, 2169 (1996).

[11] Y.G. Andreev, G.S. MacGlashan, P.G. Bruce, *Phys. Rev. B,* **55,**12011 (1997).

[12] W.I.F. David, K. Shankland, N. Shankland, *Chem. Commun.*, 931 (1998).

[13] G.E. Engel, S. Wilke, O. König, K.D.M. Harris, F.J.J. Leusen, *J. Appl. Crystallogr.,* in press.

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

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

[16] K.D.M. Harris, R.L. Johnston, B.M. Kariuki, *Acta Crystallogr.,* **A54,** 632 (1998).

[17] B.M. Kariuki, P. Calcagno, K.D.M. Harris, D. Philp, R.L. Johnston, *Angew. Chemie, Int. Ed. Engl.*, **38,** 831 (1999).

[18] B.M. Kariuki, K. Psallidas, K.D.M. Harris, R.L. Johnston, R.W. Lancaster, S.E. Staniforth, S.M. Cooper, *Chem. Commun.*, 1677 (1999).

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

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

[21] T. Csoka, W.I.F. David, K. Shankland, *Mat. Sci. Forum*, **278–281**, 294 (1998).

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

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

[24] M.S. Lehman, A.C. Nunes, *Acta Crystallogr.*, **B36,** 1621 (1980).

[25] J.W. Visser, *J. Appl. Crystallogr.,* **2,** 89 (1969).

[26] A.C. Larson, R.B. Von Dreele, Los Alamos Lab. Report No. LA-UR-86–748 (1987).

[27] P. Calcagno, B.M. Kariuki, S.J. Kitchin, J.M.A. Robinson, D. Philp, K.D.M. Harris, paper submitted.