ELSEVIER International Journal of Pharmaceutics 165 (1998) 117–126

international journal of pharmaceutics

Structure solution of Ibuprofen from powder diffraction data by the application of a genetic algorithm combined with prior conformational analysis

K. Shankland [a,*], W.I.F. David [a], T. Csoka [a], L. McBride [b]

[a] *ISIS Facility, Rutherford Appleton Laboratory, Chilton, Didcot, Oxon OX11 0QX, UK*
[b] *Department of Pharmaceutical Sciences, University of Strathclyde, George Street, Glasgow G1 1XW, UK*

Received 10 October 1997; received in revised form 5 December 1997; accepted 11 December 1997

Abstract

The crystal structure of Ibuprofen has been solved from synchrotron X-ray powder diffraction data using a genetic algorithm based model building method. The performance of the algorithm is enhanced if additional prior chemical information is incorporated in the form of hard limits on the values that can be assumed by flexible torsion angles within the molecule. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Ibuprofen; Crystal structure solution; X-ray powder diffraction; Genetic algorithm

1. Introduction

The ability to solve molecular crystal structures directly from powder diffraction data has long been a goal of those who recognise that it is often difficult to grow single crystals of a particular polymorphic form of a molecule under study. This goal is seldom easily achieved, as a series of non-trivial steps must be performed after the initial data collection in order to solve the final crystal structure. Firstly, the diffraction pattern must be indexed to provide information about the size and the shape of the unit cell. From the unit cell volume, it is then possible to determine the number of molecules in the asymmetric unit of the crystal structure. Once the unit cell is known, the individual reflection intensities, peak shape parameters and unit cell dimensions may be treated as variables in a model independent least-squares fit to the diffraction data. This method, originally developed by Pawley (Pawley, 1981), is performed in the Patterson group of the crystal structure. Using the extracted reflection intensities, the space group of the material is then deduced and the data subjected to a further

* Corresponding author.

PII S0378-5173(98)00009-X

refinement based on the method by Pawley. The net result of this procedure is a set of reflection intensities and an associated covariance matrix unique to the material under study. At this point, there are two distinctly different approaches to structure solution. The first approach borrows heavily from the conventional direct methods of structure solution that have been extremely successful when applied to single crystal diffraction data. One limitation of this approach is the inevitable reflection overlap (and consequent loss of intensity information) that occurs when the three dimensions of diffraction space are collapsed onto the one dimension of a powder diffraction pattern as a result of the random orientation of crystallites within the powder sample. In general, conventional direct methods do not deal particularly well with the low resolution, poorly determined reflection intensities recovered from a typical powder diffraction pattern exhibiting a high degree of reflection overlap. However, a variety of strategies have been proposed to counter this (Jansen et al., 1992; Gilmore et al., 1993; Sivia and David, 1994; Altomare et al., 1995) and an increasing number of structures are now being solved in this way.

The second approach to structure solution from powder diffraction data is radically different and can be regarded as a global optimisation procedure. Once the unit cell and space group are known, a crystal structure can be postulated, the corresponding diffraction pattern calculated and that pattern compared with the measured diffraction pattern. The trial structure can then be adjusted in such a way as to minimise the differences between the observed and calculated patterns. Once there is sufficiently good agreement, the structure may be considered solved. Whilst some successes have been reported with this approach in recent years (Tremayne and Harris, 1996), the majority of examples in the literature are structure solutions of rigid or essentially rigid molecules, such as dye substances. The reason for this is clear; highly flexible molecules contain a large number of internal degrees of freedom which need to be varied in addition to the six external degrees of freedom that are varied when solving rigid molecules. This rapidly leads to a combinatorial explosion for all but the simplest of problems, rendering exhaustive search methods impractical and challenging totally random search methods.

We have recently developed a computer program, 'the GAP', that addresses this problem by implementing a genetic algorithm (GA) search method. The method can be applied to flexible molecules and multiple fragments and is therefore well suited to dealing with drug substances, including salt forms. The GA implementation has been described in some detail elsewhere (Shankland et al., 1997). Here, using the previously known crystal structure of Ibuprofen as a test case, we show that the efficiency of the GA method can be improved by using the results of conformational analyses to direct the search to regions of space where the likelihood of finding the correct molecular structure is relatively high.

2. Experimental

2.1. Data collection

X-ray powder diffraction data were collected from a hexane-recrystallised sample of Ibuprofen on station 2.3 of the Daresbury synchrotron radiation source, operating with a flat plate geometry and an incident wavelength of 1.405Å. Data in the range 6°–56° 2θ were collected in 0.01° steps with a count time of 3 s/step, giving a total collection time of just over 4 h. The resultant diffraction pattern is shown in Fig. 1.

2.2. Data fitting

Program SR15LS (David et al., 1992) was used to perform a Pawley refinement of the data in which the reflection intensities, peak shape (Voigt, with an asymmetry correction for axial divergence), cell parameters and instrument zero point were allowed to vary. The initial cell parameters and the correct space group ($P2_1/c$) were taken from a previous single-crystal X-ray diffraction study of Ibuprofen (McConnell, 1974). The final χ^2 for the Pawley fit was ≈ 21, indicating that the diffraction pattern had been reasonably well fitted (Fig. 2). The resultant listing of reflection intensi-

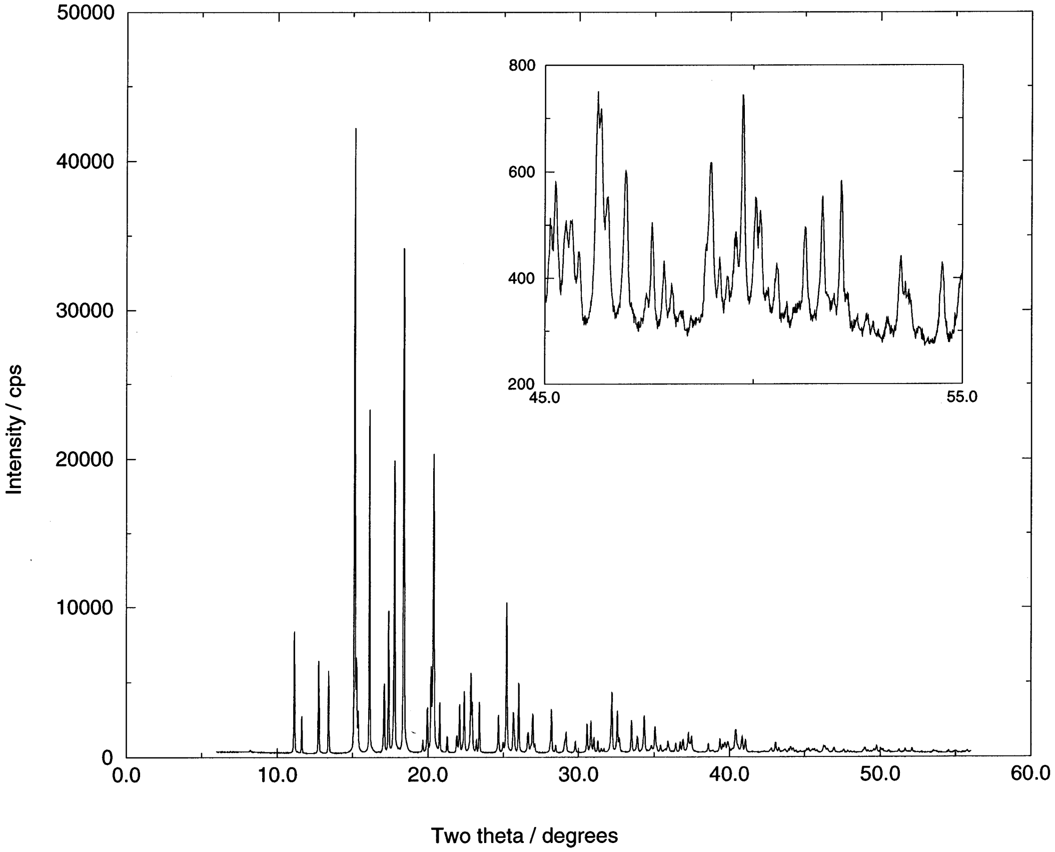


Fig. 1. X-ray powder diffraction data for a sample of hexane-recrystallised Ibuprofen. The inset graph shows that a considerable amount of information is still present in the weaker, high angle region of the diffraction pattern.

ties, together with its associated covariance matrix, was written to file at the end of the fitting procedure.

2.3. The molecular chromosome

A molecular model of an (R)-Ibuprofen molecule was constructed in internal coordinates using standard bond lengths and angles. Those four torsion angles (τ_1, τ_2, τ_3, τ_4), that could not be assigned fixed values were flagged as variables for the GA procedure (Fig. 4). The external degrees of freedom of the single molecule in the asymmetric unit are defined by three coordinates (x, y, z) and three Euler angles (θ, ϕ, ψ) and so the full molecular contents of the asymmetric unit of the crystal structure may then be represented by a sequence of ten real numbers: $x, y, z, \theta, \phi, \psi, \tau_1, \tau_2, \tau_3, \tau_4$.

In GA terminology, this sequence is a chromosome describing the molecular crystal structure and each of the individual variables is a gene, whose value is allowed to vary throughout the lifetime of the GA run (Michalewicz, 1996).

2.4. Bounds on the allowable gene values

Bounds on the positional and orientational parameters were derived from the Euclidian normalisers of the space group (Hahn, 1989) and taken to be: $0 \leq x \leq 0.5$, $0 \leq y \leq 0.25$, $0 \leq z \leq 0.5$, $0° \leq \theta \leq 360°$, $0° \leq \phi \leq 90°$, $0° \leq \psi \leq 360°$.

Two sets of restrictions were placed on the values that could be assumed by the variable

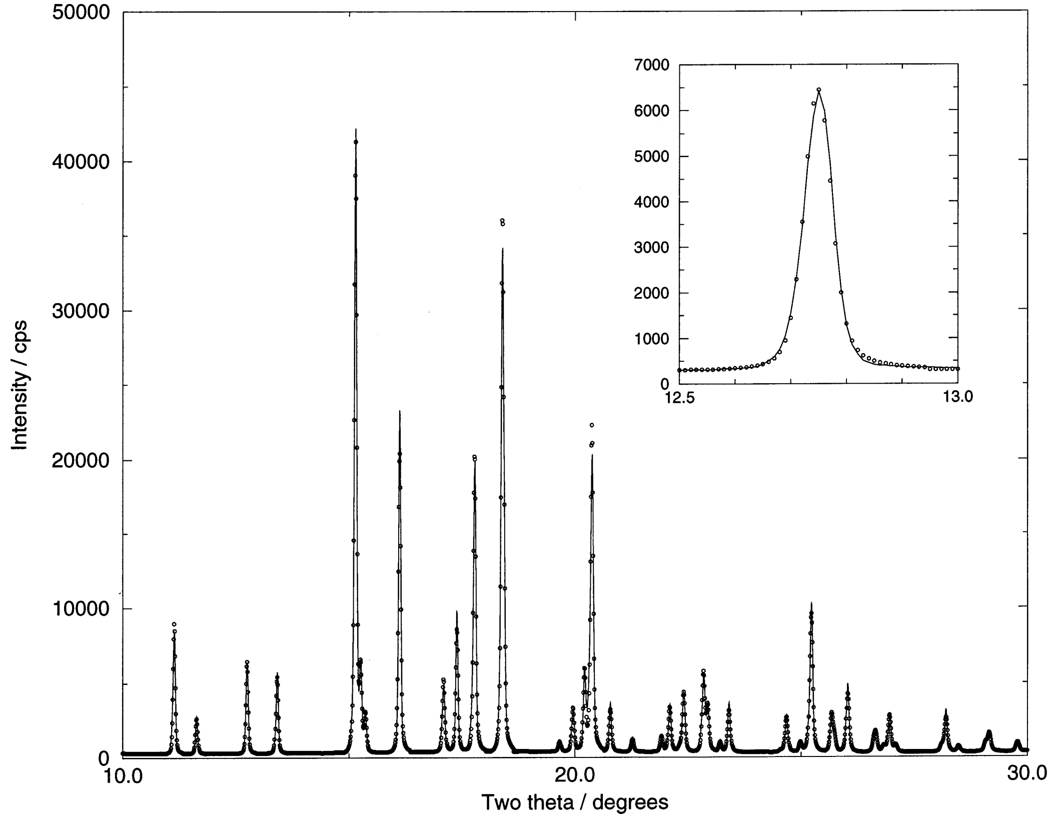


Fig. 2. The result of a model independent fit to the Ibuprofen diffraction data. The calculated profile (——) matches the measured diffraction data (○) well. The inset graph shows that the measured peaks are reasonably well described by the Voigt function used in the fit.

torsion angles within the GA procedure. The first 'general' set was used to ensure that essentially duplicate molecular conformations were not created within the program. For example, given that the X-ray data extend to relatively low resolution, and the fact that hydrogen atoms are very weak X-ray scatterers, the following conformations are essentially indistinguishable (Fig. 3).

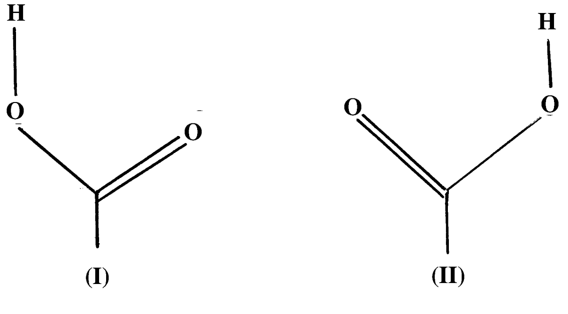


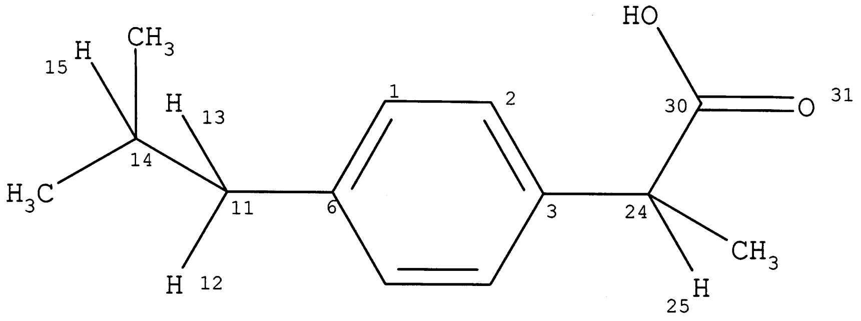
Fig. 3.

Thus, only 180° of the space covered by rotation of the carboxyl group needs be searched.

The second 'specific' set of torsion angle restrictions was used to direct the search to regions of conformational space where the likelihood of finding favourable molecular conformations is greater. These regions of space were identified from a conformational analysis of Ibuprofen (Shankland et al., 1998) and structurally related molecules, and the derived bounds are listed in Fig. 4.

2.5. Molecular evolution

A collection of chromosomes constitutes a population of molecules and the fitness of an individual chromosome is a figure of merit that indicates whether one chromosome is 'better' than another



Variable torsion	General bounds	Specific bounds
τ_1 = 12 - 11 - 6 - 1	-90° to 90°	0° to 60°
τ_2 = 15 - 14 - 11 - 13	0° to 360°	40° to 200°
τ_3 = 25 - 24 - 3 - 2	0° to 360°	-170° to 100°
τ_4 = 31 - 30 - 24 - 25	0° to 180°	120° to 160°

Fig. 4. The Ibuprofen molecule showing the numbering scheme used for those flexible torsion angles flagged in the internal coordinate definition. Note, that this numbering scheme is different from the one used in the conformational analysis of Ibuprofen (Shankland et al., 1998). The bounds on the torsion angles varied within the GA program are listed for the general case and for the case where restrictions have been derived from the conformational analysis of Ibuprofen.

in the context of the given problem. For the powder diffraction problem, we define the fitness function to be:

$$\chi^2 = \sum_{\mathrm{h}}\sum_{\mathrm{k}}[(I_{\mathrm{h}} - c|F_{\mathrm{h}}|^2)(V^{-1})_{\mathrm{hk}}(I_{\mathrm{k}} - c|F_{\mathrm{k}}|^2)]$$

where $I_{\mathrm{h,k}}$ is the extracted intensity from a Pawley refinement of the diffraction pattern; V_{hk}, the covariance matrix from the Pawley refinement; c, scale factor; $F_{\mathrm{h,k}}$, the calculated structure factor from trial structure.

This expression is formally equivalent to fitting a diffraction pattern by traditional Rietveld methods. It has the advantage, however, of being substantially faster to calculate than a standard Rietveld agreement factor (Rietveld, 1969).

Thus, the fitness of an individual chromosome in the powder problem is an indication of how well the trial crystal structure matches the measured diffraction data; the smaller the value of χ^2, the better the match. We therefore start the evolution process by creating a population of molecules with genes initialised at random. We then evaluate the fitness of each chromosome, and thus decide which molecules will survive to form the next population through a 'survival of the fittest' procedure.

Pairs of individual survivors are allowed to breed by uniform and arithmetic crossover mechanisms, used in equal proportions. The single random point uniform crossover operates as:

Parents: $(x_{\mathrm{a}}, y_{\mathrm{a}}, z_{\mathrm{a}}, \theta_{\mathrm{a}}, \phi_{\mathrm{a}}, \psi_{\mathrm{a}}, \tau_{1\mathrm{a}}, \tau_{2\mathrm{a}}, \tau_{3\mathrm{a}}, \tau_{4\mathrm{a}}) + (x_{\mathrm{b}}, y_{\mathrm{b}}, z_{\mathrm{b}}, \theta_{\mathrm{b}}, \phi_{\mathrm{b}}, \psi_{\mathrm{b}}, \tau_{1\mathrm{b}}, \tau_{2\mathrm{b}}, \tau_{3\mathrm{b}}, \tau_{4\mathrm{b}})$

Offspring: $(x_{\mathrm{a}}, y_{\mathrm{a}}, z_{\mathrm{a}}, \theta_{\mathrm{b}}, \phi_{\mathrm{b}}, \psi_{\mathrm{b}}, \tau_{1\mathrm{b}}, \tau_{2\mathrm{b}}, \tau_{3\mathrm{b}}, \tau_{4\mathrm{b}})$ and $(x_{\mathrm{b}}, y_{\mathrm{b}}, z_{\mathrm{b}}, \theta_{\mathrm{a}}, \phi_{\mathrm{a}}, \psi_{\mathrm{a}}, \tau_{1\mathrm{a}}, \tau_{2\mathrm{a}}, \tau_{3\mathrm{a}}, \tau_{4\mathrm{a}})$ where the crossover point is chosen randomly. The arithmetic crossover operates as:

Parents:

$$\vec{A} + \vec{B}$$

Offspring:

$$(\vec{A}r + \vec{B}(1-r)) \text{ and } (\vec{A}(1-r) + \vec{B}r)$$

where $\vec{A}+\vec{B}$ represent parent chromosomes as outlined in the uniform crossover case, and r is a random number between zero and one. Typically, 5–10% of genes in the resultant offspring are mutated, 40% of the time by a uniform mutation and 60% of the time by a non-uniform mutation. The single random point uniform mutation operates as $(x_a, y_a, z_a, \theta_b, \phi_b, \psi_b, \tau_{1b}, \tau_{2b}, \tau_{3b}, \tau_{4b}) \rightarrow (x_a, y_a, z_a, \theta_b^*, \phi_b, \psi_b, \tau_{1b}, \tau_{2b}, \tau_{3b}, \tau_{4b})$ in the case where θ_b has been chosen at random as the gene to be mutated, and has been assigned a new random value θ_b^*, within the allowed bounds. The non-uniform mutation initially operates identically to the uniform mutation, but as the number of generations increases, the bounds on the gene effectively contract around the current value of that gene. Thus, this operator introduces an increasing amount of local searching as the population ages (Michalewicz, 1996). The offspring formed by the crossovers and mutations form the new parent population and the process is repeated. Fig. 5 summarises this cycle and its implementation in 'the GAP' computer program. The net effect is to produce a continually changing population of trial molecular structures in which members iterate towards the global minimum in $\chi^2(x, y, z, \theta, \phi, \psi, \tau_1, \tau_2, \tau_3, \tau_4)$ space. The overall crossover and mutation rates were set such that there was a 95% chance of a parent being selected to participate in a crossover operation, and an 8% chance of an individual gene undergoing a mutation operation. The overall population size was held constant at 224 molecules, with the members of the population distributed in groups of seven over 32 islands, and evolution occurring independently on each island. Migration of members between islands was allowed only every 50th generation, with a 25% chance of an individual member then migrating to another island. A structure was deemed to be solved at a point where a constrained Rietveld refinement of the best trial structure converged to the correct (previously known) crystal structure. Preliminary runs showed that this process could be carried out reliably when χ^2 values of 500 or less were reached. A total of 50 runs were performed with the general torsion angle bounds listed in Fig. 4, and another 50 runs with the specific angle bounds (henceforth referred to as the unconstrained and constrained runs respectively). Each GA run was terminated if a member with a $\chi^2 \leq 500$ was detected or the preset maximum of 10000 generations was reached. All the GA structure solution runs were carried out on two dual 200 MHz Pentium Pro personal computers, oper-

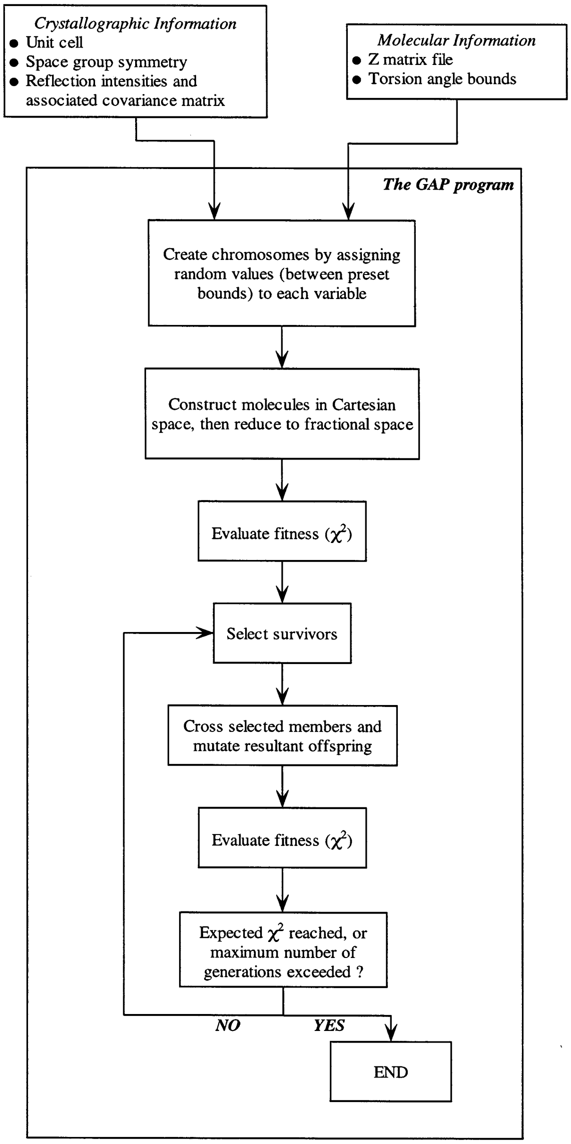

Crystallographic Information
● Unit cell
● Space group symmetry
● Reflection intensities and associated covariance matrix
Molecular Information
● Z matrix file
● Torsion angle bounds
The GAP program
Create chromosomes by assigning random values (between preset bounds) to each variable
Construct molecules in Cartesian space, then reduce to fractional space
Evaluate fitness (χ²)
Select survivors
Cross selected members and mutate resultant offspring
Evaluate fitness (χ²)
Expected χ² reached, or maximum number of generations exceeded ?
NO
YES
END

Fig. 5. A flow chart outlining the operation of 'the GAP' computer program.

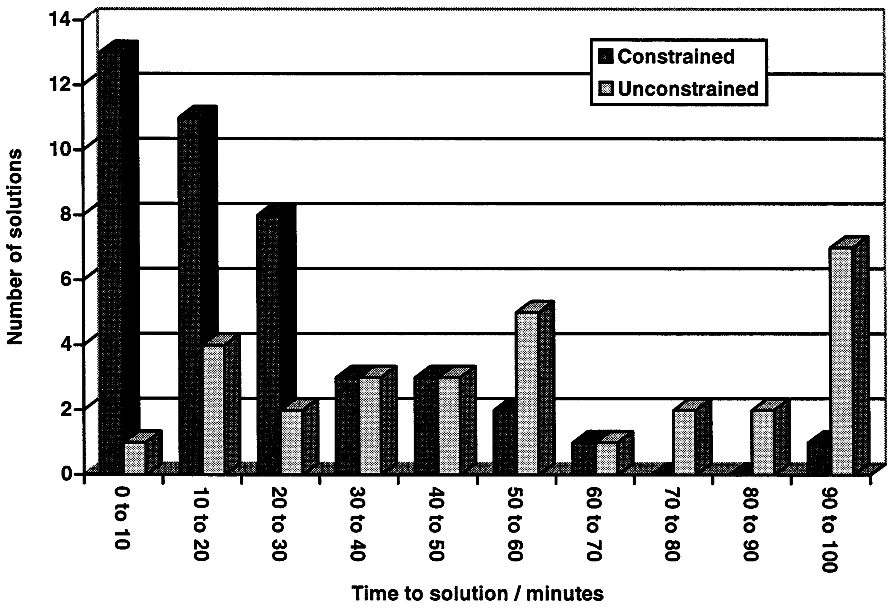


Fig. 6. The performance of the GA in solving the crystal structure of Ibuprofen. Only eight of the constrained runs failed to reach a solution within 10000 generations, compared to 20 of the unconstrained runs.

ating as a four processor parallel computer running Linux 2.0.28. Parallel program execution was accomplished using the MPI subroutine library (Gropp and Lusk, 1996). Structures were displayed and examined using Cerius2 molecular graphics package.

3. Results

The results of the GA runs are summarised in Fig. 6, which shows how quickly refineable structure solutions ($\chi^2 \leq 500$) were obtained in the constrained and unconstrained cases. The use of torsion angle constraints reduced the number of runs that failed to reach a $\chi^2 \leq 500$ in 10000 generations by more than a factor of two. In addition, it dramatically reduced the average time taken to obtain a structure solution. Fig. 7 shows how the crystal structures of five of the constrained GA solutions that terminated with $497 \leq \chi^2 \leq 500$ compare to the known crystal structure of Ibuprofen. The trial structures are all clearly in

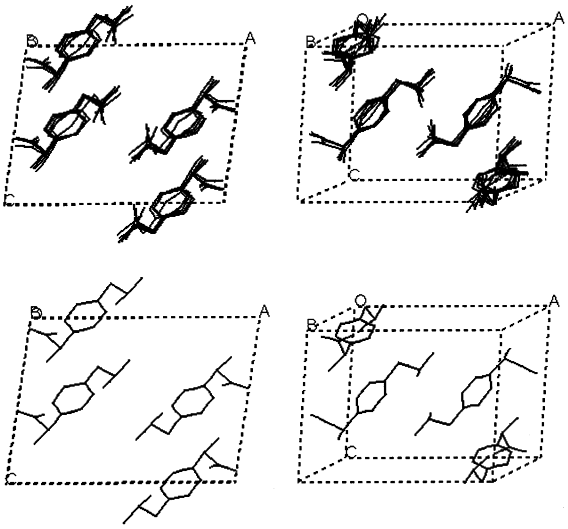


Fig. 7. A comparison of five superimposed trial crystal structures obtained from the constrained GA runs (upper two cell views) with the known single crystal structure of Ibuprofen (lower two cell views). All the trial structures have effectively equal χ^2 values, yet each one occupies a slightly different space. This illustrates the complexity of the χ^2 space that is being searched.

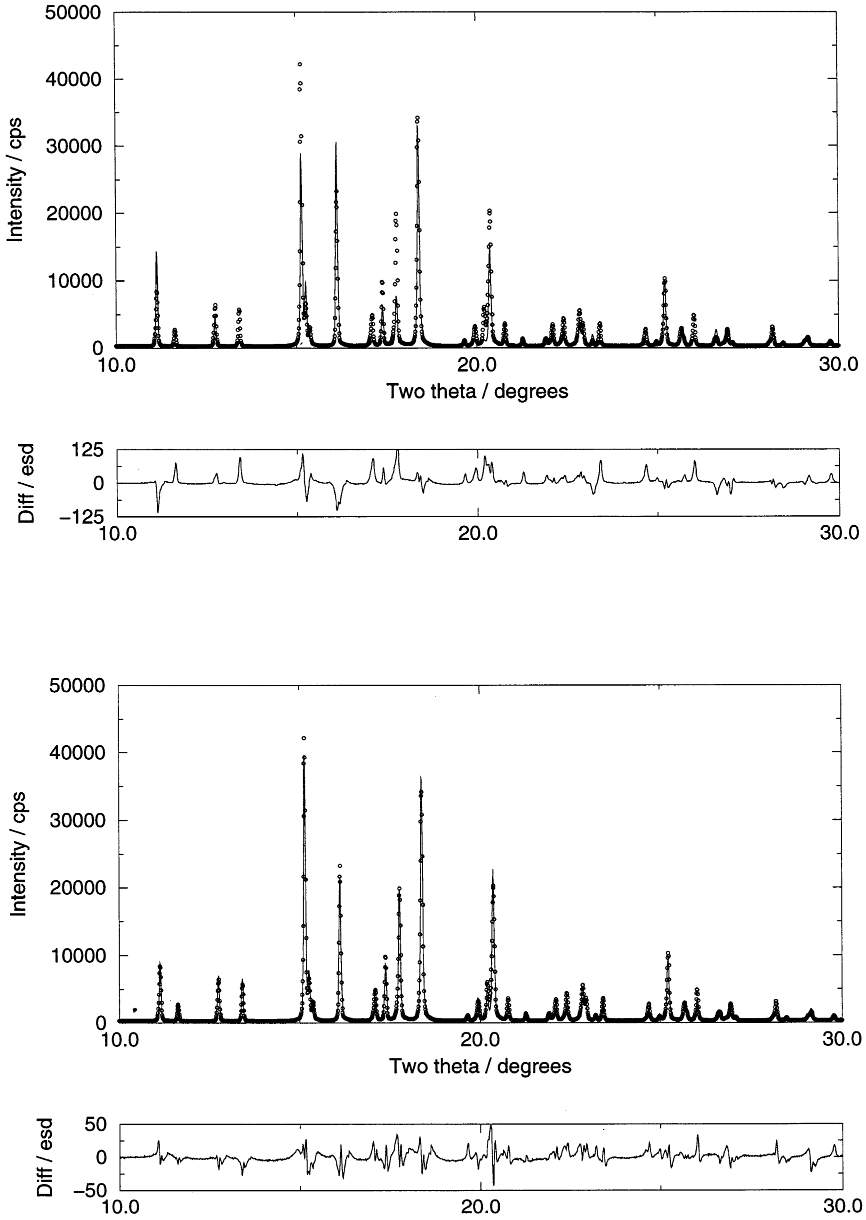


Fig. 8. The result of fitting a successful trial structural model derived from one of the constrained GA runs to the Ibuprofen diffraction data. The upper two plots show that the calculated profile (——) for the model is a good match for the measured diffraction data (○) with only the overall scale factor refined i.e. the model obtained from the GA run has not been adjusted in any way. The 'difference over estimated standard deviation' plot emphasises the deficiencies in the trial model with respect to the data. Once the model is allowed to refine subject to a series of slack geometric constraints that help to hold the atoms in chemically sensible positions during the refinement procedure, the differences between the observed and calculated profiles are greatly reduced (lower two plots).

the correct location within the unit cell. A constrained Rietveld refinement of any one of these trial structures proceeded routinely to give a good fit to the measured diffraction data (Fig. 8), demonstrating that correct solutions had indeed been located.

4. Discussion

The magnitude of the combinatorial explosion that would result were an exhaustive search method to be applied to the Ibuprofen problem is illustrated in Table 1. Such a scenario is, of course, unrealistic in the sense that even a random search would outperform a grid search that takes 31 000 years to complete! Nevertheless, the figures serve as a useful reminder that sheer computing power alone is insufficient to solve such problems on a realistic time scale. This is unlikely to change in the foreseeable future, particularly given that the majority of drug compounds are more complex than Ibuprofen in terms of their conformational flexibility. The GA approach used here has several features that appear to contribute to its successful application in the field of structure solution from powder diffraction data:

- Unlike traditional GAs, which operate on a discrete representation of the problem variables, each variable is represented by a real number that can assume any value between its allowed bounds. This removes one level of abstraction between the problem and its representation within the computer program. It is interesting to note that the discrete grid search outlined in Table 1 is actually extremely coarse compared to the continuous representation used in our GA.
- The fitness function used in our implementation allows the viability of structures to be evaluated much more quickly than functions that rely upon an exact point-by-point comparison of measured and calculated diffraction profiles.
- Elements of 'local fine-tuning' are incorporated into the search procedure via the use of non-uniform mutations. This greatly improves the convergence properties of the algorithm.
- The implicit parallelism of the GA means that computational load can be efficiently distributed over a number of processors, decreasing the already short run times by a factor approximately equal to the number of processors. Structure solutions for Ibuprofen were evaluated at a rate of 300/s on the four processor Linux cluster.

It is also important to realise that the GA solutions are purely structure factor driven and that no intramolecular/intermolecular distance or energy checks are performed at any stage. However, it is clear from the results of the constrained runs that the implicit introduction of intramolecu-

Table 1
The combinatorial explosion for an exhaustive search of possible crystal structures of Ibuprofen, using parameter bounds equivalent to the ones used in the constrained GA structure solution runs

Parameter	Range	Step size	No. of steps	Cumulative combinations	Compute time/years
x	0–0.5	0.01	50	0.5×10^{2}	1.6×10^{-9}
y	0–0.25	0.01	25	12.5×10^{2}	4.0×10^{-8}
z	0–0.5	0.01	50	62.5×10^{3}	2.0×10^{-6}
θ	0–360°	5°	72	4.5×10^{6}	1.4×10^{-4}
ϕ	0–90°	5°	18	8.1×10^{7}	2.6×10^{-3}
ψ	0–360°	5°	72	5.8×10^{9}	1.8×10^{-1}
τ_1	0–60°	5°	12	7.0×10^{10}	2.2
τ_2	40–200°	5°	32	2.2×10^{12}	7.0×10^{1}
τ_3	−170–100°	5°	54	1.2×10^{14}	3.8×10^{3}
τ_4	120–160°	5°	8	9.7×10^{14}	3.1×10^{4}

The compute times are based on an evaluation rate of 1000 structures/s.

lar energy terms, via specific torsion angle bounds, greatly improves the performance of the algorithm. Organic molecules invariably crystallise in relatively low energy conformations, and the improved performance of the constrained runs is realised by reducing the amount of time spent on trial structures in which the intramolecular energy is relatively high, and therefore the probability of the structure being correct is relatively low. The conformational analysis of Ibuprofen shows that favourable conformations tend to appear on the potential energy surface as discrete clusters, and ideally the torsion angle search should reflect this. However, with the implementation of 'the GAP' used here, only one continuous region of space was searched per variable torsion angle, and some unfavourable space was inevitably included within the torsion angle bounds. When this restriction is lifted in future versions of 'the GAP', a further improvement in the performance of the algorithm should be obtained.

Other successful structure solutions achieved with 'the GAP' include cimetidine (X-ray data) and $[^2H]_{10}$-dopamine deuteriobromide (neutron data). These successes are encouraging and suggest that the GA approach is a valuable addition to the methods available for tackling structural problems of the size and conformational complexity typically encountered in drug molecules.

Acknowledgements

We would like to thank Astra Charnwood, and in particular Dr Gerry Steele for financial support for L. McBride.

References

Altomare, A., Burla, M.C., Cascarano, G., Giacovazzo, G., Guagliardi, A., Moliterni, A.G.G., Polidori, G., 1995. Extra—a program for extracting structure factor amplitudes from powder diffraction data. J. Appl. Cryst. 28, 842–846.

David, W.I.F., Ibberson, R.M., Matthewman, J.C., 1992. Rutherford Appleton Laboratory Report RAL-92-032.

Gilmore, C.J., Shankland, K., Bricogne, G., 1993. Applications of the maximum entropy method to powder diffraction and electron crystallography. Proceedings Royal Society Of London—A. Math. Physical Sci. 442, 97–111.

Gropp, W., Lusk, E., 1996. User's guide for mpich, a portable implementation of MPI. Argonne National Laboratory Report MCS-TM-ANL-96/6.

Hahn, T. (Ed.), 1989. International Tables for X-ray Crystallography, vol. A, (Space Group Symmetry). Kluwer Academic Publishers, Dordrecht.

Jansen, J., Peschar, R., Schenk, H., 1992. On the determination of accurate intensities from powder diffraction data. II. Estimation of intensities of overlapping reflections. J. Appl. Cryst. 25, 237–243.

McConnell, J.F., 1974. 2-(4-Isobutylphenyl) Propionic Acid. Cryst. Struct. Commun. 3, 73–75.

Michalewicz, Z., 1996. Genetic Algorithms + Data Structures = Evolution Programs, 3rd ed. Springer-Verlag, Berlin.

Pawley, G.S., 1981. Unit cell refinement from powder diffraction scans. J. Appl. Cryst. 14, 357–361.

Rietveld, H.M., 1969. A profile refinement method for nuclear and magnetic structures. J. Appl. Cryst. 2, 65–71.

Shankland, K., David, W.I.F., Csoka, T., 1997. Crystal structure determination from powder diffraction data by the application of a genetic algorithm. Z. Krist. 212, 550–552.

Shankland, N., Florence, A.J., Cox, P.J., Wilson, C.C., Shankland, K., 1998. Conformational analysis of Ibuprofen by crystallographic database searching and potential energy calculation. Int. J. Pharm. 165, 107–116.

Sivia, D.S., David, W.I.F., 1994. A Bayesian approach to extracting structure-factor amplitudes from powder diffraction data. Acta Cryst. A50, 703–714.

Tremayne, M., Harris, K.D.M., 1996. Crystal structure determination from powder diffraction data. Chem. Mater. 8, 2554–2570.