Modified Genetic Algorithm to Model Crystal Structures: III. Determination of Crystal Structures Allowing Simultaneous Molecular Geometry Relaxation

VICTOR E. BAZTERRA, MARTA B. FERRARO, JULIO C. FACELLI

Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, (1428) Ciudad Universitaria, Pab. I., Buenos Aires, Argentina, and Center for High Performance Computing, University of Utah, 155 South 1452 East RM 405, Salt Lake City, UT 84112-0190

Received 1 November 2002; accepted 20 May 2003

DOI 10.1002/qua.10726

ABSTRACT: The modified genetic algorithm (MGAC) has been extended to allow simultaneous relaxation of molecular geometry during optimization of the crystalline structure. The method was applied to L-alanine and DL-alanine for two different potential functions. The genetic algorithm was always able to find minima that are likely global minima of the crystalline potential, showing good agreement with the experimental structures. For DL-alanine MGAC located the experimental crystalline structure but also consistently found a different low-energy crystalline structure that it is an excellent candidate for a polymorph. © 2003 Wiley Periodicals, Inc. Int J Quantum Chem 96: 312–320, 2004

Key words: genetic algorithms; crystal structure prediction; alanine; polymorphism

Introduction

The difficulty in predicting crystal structures from first principles arises from the large number of parameters to optimize, multiple local minima, selection of energy potential functions, numerical handling of periodic boundary conditions, etc. Moreover, organic compounds usually crystallize in a number of different polymorphic structures that exhibit different physical and chemical properties. Sometimes polymorphs are difficult to detect because they may crystallize and/or develop under special conditions. Thus, modeling methods that can predict crystal structures and their polymorphs from basic principles are extremely valuable. A recent review article by Beyer et al. [1] listed almost 200 crystals that have been used in published crystal prediction studies.

Correspondence to: J. C. Facelli, e-mail: Julio.Facelli@utah.edu

V. E. Bazterra is the recipient of a Fellowship from CONICET at the Universities of Buenos Aires and Utah.

M. B. Ferraro is a member of Carrera del Investigador del CONICET at the University of Buenos Aires.

The effectiveness of current methods used to predict crystalline structures is still an open problem [2–4]. In the previous articles in this series [5, 6] we presented the modified genetic algorithm to model crystal structures (MGAC) method to search for crystalline structures. The applications presented in the previous work were concerned with the crystal structures of planar aromatic hydrocarbons, for which assuming fixed rigid molecular geometries is a good approximation. In this article we report an extension of the MGAC method to find crystal structures allowing the simultaneous relaxation of the molecular geometry during the genetic algorithm (GA) selection process.

The MGAC method employs a modified GA [7, 8] to generate a population of random structures of the crystals and evolve these structures to find the fittest individuals according to a given "fitness function." Information about different regions of the configurational space is passed between the individual strings by means of the crossover procedure. Operator analogs to crossover, mutation, and natural selection are employed to perform a search able to explore and learn the multidimensional parameter space and determine which regions of that space provide good solutions to the problem. Our GASPlib [5] library has been successfully used, in conjunction with GAlib [9], to predict the crystal structures of benzene, naphthalene, and anthracene using an empirical potential function to calculate the crystal energies that were used as the fitness functions [5]. Also, we used the enthalpy as the fitness function to successfully locate a high-pressure polymorph of benzene [6].

Despite their biologic importance, few amino acids have been used in crystal prediction studies [10, 11]; for this reason, in this work we selected L-alanine and DL-alanine as test cases for this new implementation of MGAC.

Methodology

GENETIC CODING OF THE CRYSTAL STRUCTURE

In an implementation of a GA it is necessary to define a genome such that the information available in it allows the calculation of the fitness function associated with the individual defined by the genome. The coding method used here allows the treatment of crystals with asymmetrical unit cells with any arbitrary number, Z, of molecules per cell. When the molecules in the crystal are treated as flexible bodies without any restriction the resulting genome contains six real numbers giving the crystallographic parameters ($a, b, c, \alpha, \beta, \gamma$) and an array of dimension $3N$ with the Cartesian coordinates of the N atoms belonging to the asymmetrical cell. Therefore, the number of variables to optimize is $3N + 6$. In many cases the search in this large configuration space can be avoided by reducing the dimension of the genome to reflect physical approximations appropriate for molecules with semirigid structures. The extreme case of this approximation was used in our previous work in aromatic hydrocarbons [5] in which the molecular geometry was assumed fix during the optimization of the crystal structure. A less drastic approximation is appropriate for molecules with rigid backbones but mobile side-chains. In this case, which we will label the semirigid approximation, the genome can be expressed by adding to the genome for the rigid approximation the intramolecular dihedral angles for which the torsion energies are comparable to the intermolecular interactions. The genome within the rigid approximation, for an asymmetrical cell with Z molecules, is given by six real numbers corresponding to the crystallographic parameters of the asymmetrical cell ($a, b, c, \alpha, \beta, \gamma$), the position of the center of mass of all molecules in the cell ($\mathbf{r}_1$, $\mathbf{r}_2$, $\mathbf{r}_{3,}$ . . ., $\mathbf{r}_z$), and Z sets of three Euler angles, (Φ_1, Φ_2, $\Phi_{3,}$ . . ., Φ_z), describing the orientation of the molecules with respect to the unit cell. When partial relaxation of the molecular geometry is allowed in the GA, i.e., the semirigid approximation, the genome has to be enlarged with N_{dihed} real numbers corresponding to the values of the dihedral angles that are allowed to vary in the optimization. With these approximations the total number of independent variables in the search space is

$$k = 6 + Z(6 + N_{\text{dihed}}), \qquad (1)$$

where N_{dihed} is the number of varying dihedral angles in each molecule in the asymmetrical cell. A pictorial representation of the genomes for both cases discussed above is given in Figure 1.

Ideally, the prediction of the crystal structures should be done using only the connectivity between the atoms within the molecule. A full optimization without any symmetry constrain has only one parameter unknown, the number of molecules per asymmetrical cell, which has to be assumed previous to the calculations. The crystal parameters including its symmetry can be calculated from the resulting structures. Unfortunately, this strategy has several disadvantages: The number of parame-

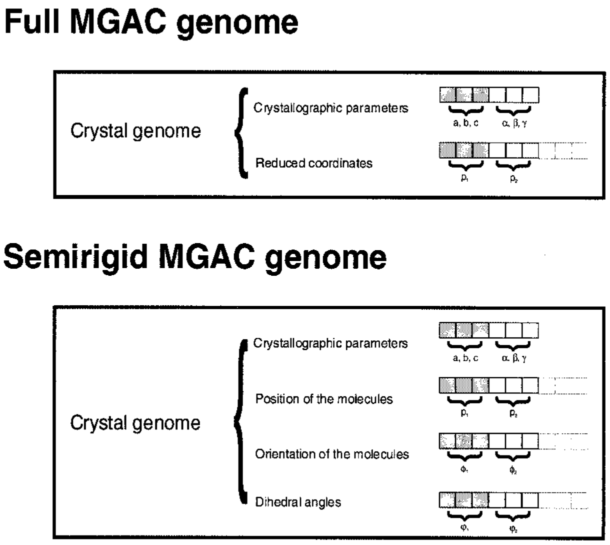


FIGURE 1. Pictorial representation of the genomes used for the unconstrained or full-MGAC and semirigid approximations.

ters to optimize is considerably larger than for the symmetry constrained case, making the problem harder to resolve. Moreover, because there are no symmetry constrains in the optimization, a large number of nonphysical structures are created in the GA process, making it difficult to avoid the contamination of the population from this artifacts. In this work we have chosen to select the symmetry group of the crystal and assume the number of molecules per asymmetrical cell before running the optimization using MGAC. This greatly reduces the number of parameters necessary to define a crystal structure, leading to a smaller configuration space for each choice of number of molecules per cell and space group. When this information is not available multiple runs for different space groups and number of molecules per cell are needed to locate all the possible crystalline structures. While this may be perceived as a serious drawback because there are 235 possible space groups, it has been documented that most of the molecular crystal structures belong to one of the 7 most common space groups [12]. Moreover, for crystals in these space groups the large majority has one or two molecules per asymmetrical cell. Due to these regularities and the fact that the number of parameters in each optimization is greatly reduced by as many times as the number of symmetry operators belonging to a space group, this approach is highly preferred.

For the cases studied here, L-alanine and DL-alanine, the space groups of their crystals as well as the number of molecules per asymmetrical unit cell are known [13, 14]. Therefore, we restricted our search to crystals with four molecules per cell, belonging to the $P2_12_12_1$ and $\mathrm{Pna}2_1$ space groups for L-alanine and DL-alaninne, respectively. Both space groups are orthorhombic and therefore the unit cell is defined by the three cell constants a, b, and c with all the cell angels α, β, $\gamma = 90°$.

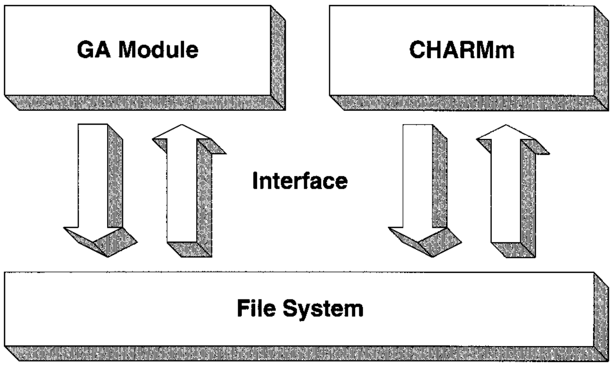


FIGURE 2. Modular structure of the MGAC software used in this work.

POTENTIAL ENERGY FUNCTIONS

Using the information contained in the genomes the energy of each individual was evaluated and its structure relaxed to the local minimum. All calculations were done using the CHARMm code [15] with either the AMBER [16] or CHARMm [17] force fields and considering periodic boundary conditions. To this end, in the new version of MGAC we implemented an interface with the CHARMm code, schematized in Figure 2. Note that while this implementation uses the CHARMm module to calculate the crystal energy and perform its local minimization. The system is modular and the

TABLE I

Comparison of the crystal lattice parameters and R_{wp} [Eq. (2)] factors for the reference and experimental structures of L-alanine using different potentials.

Structure[a]	a (Å)	b (Å)	c (Å)	R_{wp}
Exp.	6.032	12.343	5.784	0.000
Ref_AMBER94	5.758	12.323	5.694	0.416
Ref_CHARMm	5.521	12.483	5.404	1.265

[a] From Ref. [13].

TABLE II

Comparison of the crystal lattice parameters, R_{wp} [Eq. (2)] factors, and relative energies, ΔE,[a] for the predicted structures of L-alanine calculated using the semirigid approximation.

Structure	*a* (Å)	*b* (Å)	*c* (Å)	R_{wp}	ΔE^a (kcal/mol)
Pred_AMBER94	5.774	12.357	5.703	0.418	−0.222
Pred_CHARMm	5.521	12.281	5.404	1.365	0.001

[a] The relative energy is given by the difference between the energy of the predicted structure and the energy of the reference structure for the same force field.

CHARMm module can easily be replaced by any other energy module with similar functionality.

GA OPTIMIZATION

The routine GALib [9] is used to generate the random numbers necessary to construct the initial *genomes* for N_{pop} individuals. The random numbers used to define the length of the crystal axes belong to specified intervals selected according to the expected dimensions of the unit cell; these restrictions have been included to avoid sampling in nonphysical configurations. It is important that the intervals chosen are not too large, to reduce the time required for the local optimizations, but long enough to generate a representative sampling.

The GA operations of mating (crossover), mutation, and selection are used to evolve one generation into the next. The population replacement is made employing the steady-state genetic algorithm (SSGA), which typically replaces only a small proportion of strings in each generation [18–20]. This technique is also known as elitism. Among the population the best individuals, 50% of the population, are copied directly into the next generation. The

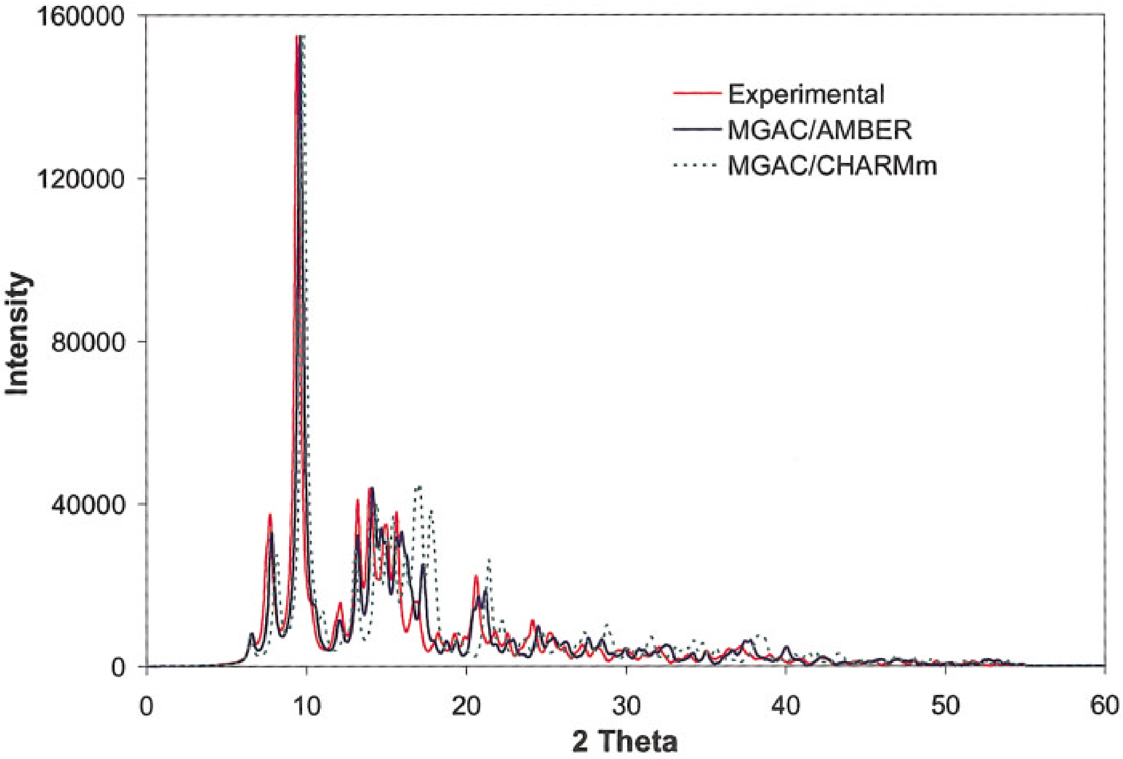
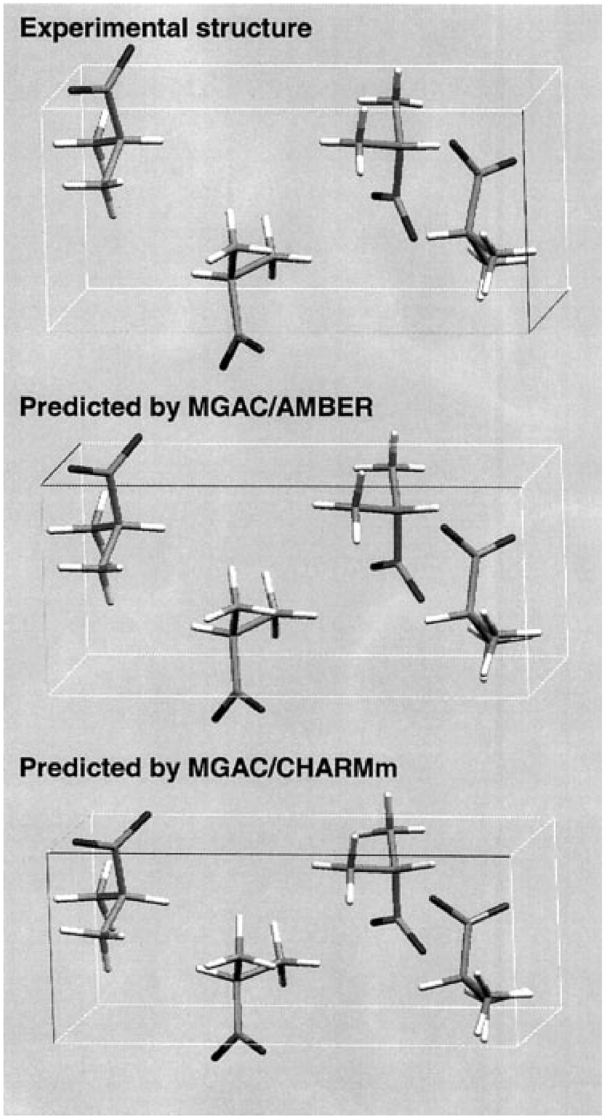


FIGURE 3. Simulated X-ray power diffraction spectra for different structures of L-alanine. (Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.)


Experimental structure

Predicted by MGAC/AMBER

Predicted by MGAC/CHARMm

FIGURE 4. Comparison of the predicted structures of L-alanine with the experimental one.

criteria for fitness probability, selection of the individuals, and mutation used here are those discussed in detail in Ref. [5]. The results reported in this article were obtained using a population formed by $N_{\text{pop}} = 20$ individuals.

Results and Discussion

L-alanine and DL-alanine were chosen as a test case for the new version of MGAC. These molecules are well suited for these tests because they contain a relatively small number of atoms and have a relatively large number of internal degrees of freedom (three dihedral angles) for such small molecules. This allowed extensive testing and comparison of the optimizations procedures described above.

When comparing the predicted (optimized) structures with the experimental ones we use the concept of reference structure that was introduced in our previous work [5]. The reference structures are calculated performing a local minimization of the experimental structures using the different force fields or potentials considered here. By using reference structures when comparing with experimental data it is possible to independently test the deficiencies in the predicted structures that are associated with the GA optimization algorithm from those originating in shortcomings of the force field used in the calculation of the energy.

Several methods have been proposed to compare predicted and experimental crystal structures [4]. Here, we compare the different crystal structures using their lattice parameters and differences between their calculated X-ray powder diffraction spectra. This difference is measured by the weighted profile values, R_{wp} given by [21],

$$R_{wp} = \left\{ \sum_{i}^{N} w_i[\,y_i(\exp) - y_i(\text{pred})]^2 \Big/ \sum_{i}^{N} w_i y_i^2(\exp) \right\}^{1/2}, \qquad (2)$$

TABLE III

Comparison of the crystal lattice parameters, R_{wp} [Eq. (2)] factors, and relative energies, ΔE,[a] for the predicted structures of L-alanine calculated using the unconstraint or full MGAC approximation.

Structure	*a* (Å)	*b* (Å)	*c* (Å)	R_{wp}	ΔE[a] (kcal/mol)
Pred_AMBER94	5.766	12.370	5.701	0.418	−0.024
Pred_CHARMm	5.520	12.483	5.404	1.267	0.001

[a] The relative energy is given by the difference between the energy of the predicted structure and the energy of the reference structure for the same force field.

TABLE IV

Comparison of the crystal lattice parameters and R_{wp} [Eq. (2)] factors for the experimental and reference structures of DL-alanine calculated using different potentials.

Structure[a]	a (Å)	b (Å)	c (Å)	R_{wp}
Exp.	12.026	6.032	5.829	0.000
Ref_AMBER94	12.768	5.667	5.687	0.655
Ref_CHARMm	12.742	5.554	5.402	1.401

[a] From Ref. [14].

where w_i is the statistical weight of the intensity, $w_i = 1/y_i(\text{exp})$, $y_i(\text{exp})$ is the intensity calculated using the experimental structure, and $y_i(\text{pred})$ is the intensity calculated using the predicted structure at channel i.

L-ALANINE

As discussed above we performed a local optimization of the experimental structure using two different force fields to obtain the corresponding reference structures and evaluate their ability to reproduce the structure for this particular molecule. The results of these optimizations are show in Table I. From the comparison of the crystal parameters and the R_{wp} factors it becomes apparent that both the AMBER and CHARMm force fields produce local minima with structures close to the experimental one. The optimization with AMBER gives the closest reference structure to the experimental one.

The first set of optimizations performed using the MGAC were run two or three times for each force field using the semirigid approximation and the symmetry constrains of the $P2_12_12_1$ space group. Under the semirigid approximation all the dihedral angles in L-alanine were allowed to change freely, while the bond distances and bond angles were kept at the values found for the corresponding reference structures. The results of these optimizations are presented in Table II, which shows the crystallographic parameters, R_{wp} factors, and energy differences between the predicted (optimized) and reference structures. It is apparent that the structures predicted by the AMBER and CHARMm force fields are similar to their respective reference structures and consequently to the experimental ones. The X-ray powder diffraction spectra simulations for the experimental and predicted structures are plotted in Figure 3, while their crystal structures are depicted in Figure 4.

The next set of optimizations was run without any constrain in the molecular geometry. The intention here was to verify if the semirigid approximation was valid for L-alanine. Table III shows the results for the full-MGAC algorithm using the AMBER and CHARMm potentials; from the comparison of these results with those in Tables I and II it is evident that the full-MGAC optimization finds the same structures as the semirigid MGAC approximation. While in L-alanine the use of the semirigid approximation may not be necessary, it may be important when MGAC is applied to larger molecules where the number of parameters may be considerably larger for the full MGAC optimizations.

DL-ALANINE

Table IV shows the results of the local optimization of the experimental structures (reference struc-

TABLE V

Comparison of the crystal lattice parameters, R_{wp} [Eq. (2)] factors, and relative energies, ΔE,[a] for the predicted structures of DL-alanine calculated using the semirigid approximation.

Structure	a (Å)	b (Å)	c (Å)	R_{wp}	ΔE[a] (kcal/mol)
Pred_AMBER94					
I	12.785	5.664	5.687	0.622	0.232
II	12.760	5.663	5.701	6.023	0.205
Pred_CHARMm					
I	12.742	5.558	5.403	1.537	0.625
II	12.766	5.471	5.394	9.875	−0.022

[a] The relative energy is given by the difference between the energy of the predicted structure and the energy of the reference structure for the same force field.

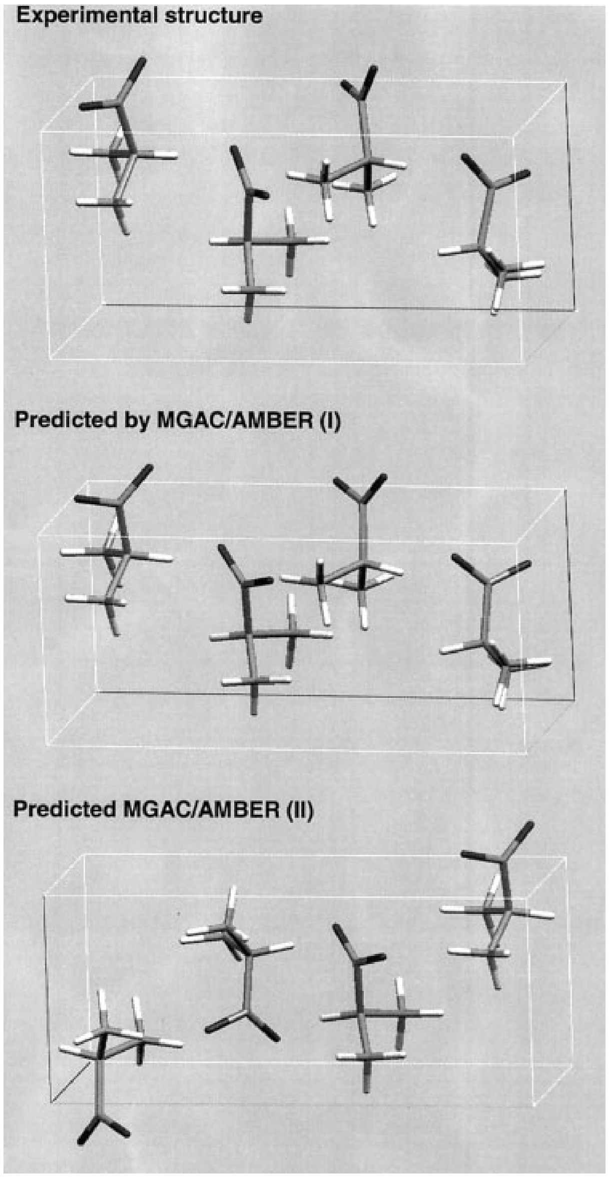


FIGURE 5. Comparison of the two predicted (AMBER) structures of DL-alanine with the experimental one.

tures) for the AMBER and CHARMm potentials. As in L-alanine there is good agreement between the experimental structure and the reference structures. For DL-alanine the semirigid approximation was used in all optimizations; the asymmetrical cell was constructed using the genome parameters to construct the first L-alanine molecule in the cell and the other molecules including the enantiomers were created using the symmetry operations that relate the molecules in the cell for the $Pna2_1$ space group.

The results of the MGAC simulation are given in Table V. For both potentials a number of different simulations consistently predicted two distinctly different structures within 0.7 kcal/mol from the corresponding reference structures. One of these structures, labeled **I** in Table V, closely corresponds to the reference structure and consequently to the experimental one. Again as in L-alanine, for this structure there is good agreement with the experimental structure when the CHARMm or AMBER potentials are used in the calculations. The other structure, labeled **II**, is a completely new structure with energy close, −0.022 kcal/mol for CHARMm and 0.205 kcal/mol for AMBER, to the energy of the corresponding reference structures. While both structures **I** and **II** have similar crystallographic parameters they show large differences in the orientation of the molecules in the unit cell, as depicted in Figure 5 for the AMBER-optimized structures. Figure 6 shows the X-ray powder diffraction spectra simulated for the experimental and predicted structures using the AMBER potential. There is good agreement between those spectra calculated using the experimental and the predicted structure (**I**), but the one corresponding to the second structure (**II**) shows significant differences in the power diffraction spectra. The second structure found by the MGAC method, using both the CHARMm and AMBER potentials, is an excellent candidate as a polymorphic form of DL-alanine.

Conclusions

This article shows how the MGAC method has been extended to allow relaxation of molecular geometry during optimization of the crystalline structure. The method has been implemented allowing either full or partial relaxation of the molecular geometry. The results show that both implementations agree when the semirigid approximation can be justified. The method was applied to L-alanine and DL-alanine for two different potential functions; the genetic algorithm was always able to find minima that are likely global minima of the potential. These global minima of the potential correspond to structures that are close to the experimental structures, showing also good agreement between the calculated X-ray powder diffraction spectra.

For DL-alanine the MGAC method, using both the AMBER and CHARMm potentials, was able to locate the experimental crystalline structure. But, it also consistently found a different crystalline structure at less than 0.2 kcal/mol from the reference structures. This structure shows different X-ray

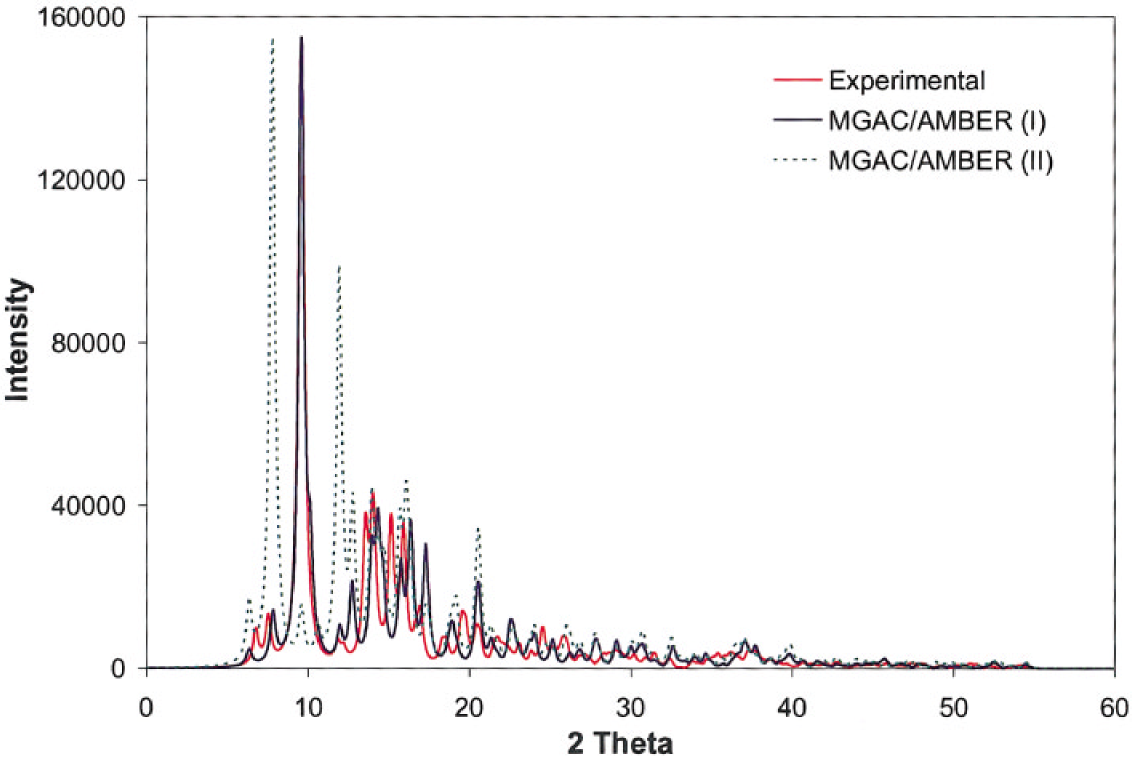


FIGURE 6. Simulated X-ray power diffraction spectra for different structures of DL-alanine. (Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.)

powder diffraction spectrum due to a different orientation of the molecules in the asymmetrical cell and is an excellent candidate for a polymorph of DL-alanine. An extensive literature search did not provide any evidence of polymorphic structures in DL-alanine, which is not surprising as it is well known that polymorphic structures are sometimes difficult to crystallize. Our results do not support the coexistence of these two crystal structures or the contamination of the experimental structure by the second one. We hope the results presented here will encourage experimental work searching for polymorphic structures of DL-alanine.

ACKNOWLEDGMENTS

Financial support for this research by the University of Buenos Aires (UBACYT-X098) and the Argentinean CONICET (PIP 02151/01) is gratefully acknowledged. J. C. F. acknowledges the support from the International and Chemistry Divisions of NSF through Grant INT-0071032. The software for this work used the GALib genetic algorithm package, written by Matthew Wall at the Massachussets Institute of Technology. The Center for High Performance Computing provided computer resources for this project.

References

1. Beyer, T.; Lewis, T.; Price, S. L. Cryst Eng Comm 2001, 44, 1.
2. Gavezzotti, A.; Filippini, G. J Am Chem Soc 1996, 118, 7153.
3. Aakeroy, C. B.; Nieuwenhuyzen, M.; Price, S. L. J Am Chem Soc 1998, 120, 8986.
4. Lommerse, J. P. M.; Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Mooij, W. T. M.; Price, S. L.; Schweizer, B.; Schmidt, M. U.; Van Eijck, B. P.; Verwer, P.; Williams, D. E. Acta Crystallogr B 2000, 56, 697.
5. Bazterra, V. E.; Ferraro, M. B.; Facelli, J. C. J Chem Phys 2002, 116, 5984.
6. Bazterra, V. E.; Ferraro, M. B.; Facelli, J. C. J Chem Phys 2002, 116, 5992.
7. Man, K. M.; Tang, K. S.; Kwong, S. Genetic Algorithms; Springer: Berlin, 1999.
8. Goldberg, D. E. Genetic Algorithms in Search, Optimisation and Machine Learning; Addison-Wesley: New York, 1989.
9. Wall, M. Galib: A C^{++} Library of Genetic Algorithm Components; Mechanical Engineering Department, Massachusets Institute of Technology: Cambridge, MA, 1996.

10. Karfunkel, H. R.; Gdanitz, R. J. J Comput Chem 1992, 13, 1171.
11. Karfunkel, H. R.; Leusen, F. J. J.; Gdanitz, R. J. J Comput-Aided Mater Des 1993, 1, 177.
12. Gavezzotti, A. Acc Chem Res 1994, 27, 309.
13. Simpson, H. J.; Marsh, R. E. Acta Crystallogr 1966, 20, 550.
14. Nandhini, M. S.; Krishnakumar, R. V.; Natarajan, S. Acta Crystallogr C 2001, 57, 614.
15. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J Comput Chem 1983, 4, 187.
16. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. Jr.; Ferguson, D. M.; Spelleyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J Am Chem Soc 1995, 117, 5179.
17. MacKerell, A. D. Jr.; Brooks, B.; Brooks, C. L. III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. In: Schleyer, P. v. R., ed. The Encyclopedia of Computational Chemistry, vol. 1; John Wiley & Sons: Chichester, UK, 1998; p. 271.
18. Syswerda, G. In: Schaeffer, J., ed. Proceedings of the Third International Conference on Genetic Algorithms; Morgan Kaufman: San Mateo, CA, 1989; p. 2.
19. Whitley, D. In: Schaeffer, J., ed. Proceedings of the Third International Conference on Genetic Algorithms; Morgan Kaufman: San Mateo, CA, 1989; p. 116.
20. Whitley, D.; Kauth, J. In: Proceedings of the Rocky Mountain Conference on Artificial Intelligence; Denver, 1988; p. 118.
21. Baerlocher, C. In: Flack, H. D.; Párkányi, L.; Simon, K., eds. Crystallographic Computing 6; International Union of Crystallography, Oxford University Press: Oxford, UK, 1993; p. 89.