JCTC Journal of Chemical Theory and Computation

A Distributed Computing Method for Crystal Structure Prediction of Flexible Molecules: An Application to *N*-(2-Dimethyl-4,5-dinitrophenyl) Acetamide

Victor E. Bazterra,[†,‡] Matthew Thorley,[‡] Marta B. Ferraro,[†] and Julio C. Facelli*[,‡]

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

Received June 23, 2006

Abstract: In this paper, we describe a new distributed computing framework for crystal structure prediction that is capable of performing crystal structure searches for flexible molecules within any space group and with an arbitrary number of molecules in the asymmetric unit. The distributed computing framework includes a series of tightly integrated computer programs for generating the molecule's force field, sampling possible crystal structures using a distributed parallel genetic algorithm, locally minimizing these structures and classifying, sorting, and archiving the most relevant ones. As an example, we report the results of its application to the prediction of the crystal structure of the elusive *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide, a molecule for which its crystal structure proved to be one of the most difficult cases in the last CSP2004 blind test for crystal structure prediction.

Introduction

The crystal structure prediction (CSP) of organic compounds can be described as the process of creating a list of crystal structure candidates, which are likely to be found experimentally, using only the molecular composition and connectivity as input information. The prediction of crystal structures for organic molecules is of great importance in many industries like pharmaceuticals, agrochemicals, pigments, dyes, explosives, and so forth, because many of the macroscopic properties of their products are highly dependent on their crystal structures. The existence of different crystals for the same compound, a phenomenon called polymorphism,[1,2] generally implies a significant variation of the substance's macroscopic properties, such as solubility, bioavailability, vapor pressure, crystal size and color, and so forth, depending on the predominant polymorph present in the material. Therefore, knowledge in advance of a plausible set of possible crystal structures provides important information that can be used for controlling the manufacturing process of solid organic compounds with the desired properties.[3]

Since the 1990s, a diverse group of methods have been developed for CSP.[4] The collective accomplishments of this research community have been summarized in the reports from the three blind tests for CSP conducted under the auspices of the Cambridge Crystallographic Data Center.[4−6] While the results of these workshops show gradual improvement in the predictive capability of the existing methods, there is general agreement that improvements are still necessary in both the force fields needed to better represent the molecular interactions and the optimization techniques used to search the complex energy landscape of molecular crystals.[4] This paper describes our recent developments to improve the latter aspect of CSP when taking advantage of the great progress in parallel computing realized in recent years.

Although the methods used for CSP vary, all of them follow four general procedures:[4] (1) the definition of a molecular model, which may or may not allow the simultaneous change of the molecular conformation during the

* Corresponding author e-mail: julio.facelli@utah.edu.

[†] Universidad de Buenos Aires.

[‡] University of Utah.

10.1021/ct6002115 CCC: $37.00 © 2007 American Chemical Society
Published on Web 12/16/2006

crystal structure searches, (2) an algorithm for searching plausible packing arrangements of the molecules in the crystal structure, (3) a method for ranking the structures according to their relative energies, usually defined by an empirical force field, and (4) a method for the classification and archival of the most relevant structures.

The molecular models can be classified into two groups known as *rigid and flexible*.[6] Within the rigid models, which are the most widely used today, the conformation of the molecule is kept fixed during the crystal structure search process. Usually, the molecular conformations considered are selected a priori by analyzing the molecule's conformational energy profile calculated for the isolated molecule using ab initio methods. This approach is a reasonable approximation when the intermolecular interactions are much smaller than the energy modes associated with the internal degrees of freedom of the molecule. However, this is not the case for flexible molecules for which the intermolecular interaction energies are often of the same order of magnitude than the rotational barriers around single bonds. Methods that allow the concurrent relaxation of the molecular geometry through the global search of the crystal structure are needed for many important applications in which flexible molecules should be considered.

The search for possible packing arrangements of the molecules in a crystal is generally implemented by a global optimization algorithm. Many search techniques have been used for this step, ranging from grid-based searches to stochastic methods like Monte Carlo simulated annealing and genetic algorithms (GAs). Following our previous work, here we continue using our modified GA approach,[4,7,8] which provides a convenient path to parallel implementations that can take advantage of modern computer equipment. It should be noted that our modified GA always includes a local minimization step for all of the structures considered in the GA evolution; that is, only locally minimized structures are included in the GA populations.

In principle, there are two approaches that could be used for CSP; either the search is constrained to run on a relatively small number of space groups one at a time or the search can be run totally unconstrained. The first approach leads to a significant reduction of the search space, and its usefulness lies in the fact that 92.7% of the organic crystals belong to only 18 of 230 space groups[9] and that 91.7% of the known organic crystal structures have only one molecule ($Z' = 1$) in their asymmetric unit. The disadvantage of this approach is that it is not able to find structures that may not be in the most common symmetry groups. Searches without symmetry constraints, based on the $P1$ space group with a variable number of molecules in their unit cell, have been reported, but they are extremely difficult because the significant increase in the search space greatly reduces their ability to produce a representative sampling of the possible packing arrangements.[9] In spite of the drawbacks of the unconstrained searches, in our computer programs, we have implemented both methodologies for completeness. Our preliminary findings of their relative performance of the constrained and unconstrained searches of the crystal structure of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide confirm the lack of predictive powder of unconstrained searches.

The ranking of the structures generated by the search algorithm is generally done according to their calculated energy. This criterion is based on the assumption that the crystallization process is under thermodynamic control. Although there is evidence that this is not always the case,[10,11] it is reasonable to assume that any metastable crystal structure should have a crystal energy close to the global minimum.[12] Alternative criteria for ranking crystal structures are sometimes used, for instance, taking into account the three-dimensional regularities commonly observed in the crystal structures deposited in the Cambridge Structural Database.[13] An extensive comparison of the methods currently available for CSP can be found in the Supporting Information of ref 4.

The comparison and classification of the resulting crystal structures is generally accomplished by sorting them by energy and eliminating from the list those structures that can be superimposed within a given tolerance, this can be done using programs like *CRYCOM*[14] or *COMPACK*.[15] Crystal structures with more than one molecule in the asymmetric unit may exhibit higher symmetry than required for the space group in which their search was performed; the additional symmetry elements in these structures can be found using the *ADDSYM* algorithm in *PLATON*[16] or the symmetry finder in *Cerius2*.[17]

This paper describes a distributed computer framework where several independent computer programs have been integrated to provide a comprehensive environment for CSP. Within this environment it is possible to (1) search for crystal structures within any symmetry group and with an arbitrary number of molecules and molecular types per asymmetric unit; (2) search structures using either the rigid or flexible molecule models; (3) automatically generate the molecule's force field using existing force field libraries; (4) increase the sampling power and the complexity of molecules amenable to CSP studies using the parallel and distributed computing capabilities of the system; and (5) automatically compare, sort, and archive the most relevant structures in a user database.

As an example, we show that this method can correctly predict the structure of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide, a well-recognized "problem" case in the crystal prediction literature.

Computational Methods

Modified Genetic Algorithm for Crystal Structure Prediction (MGAC). GAs are a family of search techniques rooted on the ideas of Darwinian evolution.[18] Operators analogous to crossover, mutation, and natural selection are employed to perform a search able to explore and learn the multidimensional parameter space and to determine which regions of the space provide good solutions to a problem and in which the search should be intensified. To improve their convergence, GAs are commonly coupled with local optimizations at each generation, a practice that has been followed in this work. Therefore, it is important to emphasize that all the crystal structures reported here correspond to local minima of the potential energy surface.

Crystal Structure Model. When using GAs for the prediction of crystal structures, these structures have to be encoded in a *genome* that can be manipulated by the genetic operators as well as used to calculate the energy of the crystal structure they represent. For organic crystals, the molecular geometries are highly constrained by strong covalent bonds, leading to a considerable reduction of the number of parameters that are allowed to change during the global search. This means that it is not necessary to include the molecule's bond lengths and bond angles in the global optimization because their values in the crystal structure are always close to those in the isolated molecule. Therefore, they can be obtained by performing only local optimizations in which these parameters are allowed to change. Because the rotational barriers around single covalent bonds are comparable to the intermolecular interaction energies, their associated dihedral angles can be significantly affected by the intermolecular interactions leading to values in the crystal that are quite different than those in the isolated molecule. Therefore, these dihedral angles must be included in the global optimization to allow the exploration of conformations which may become energetically favorable for the certain packing motifs.

For each molecular species, we define a molecular frame anchored to the rigid structure of the molecule in which the positions of all the atoms can be determined. Using this molecular frame, we define the *genome* for a crystal structure with n molecules in the asymmetric unit, by specifying in the crystal frame the position of the origin of the molecular frames, $\mathbf{R}_1...\mathbf{R}_n$, and the orientation of the molecular axis relative to the crystal frame, $\Theta_1...\Theta_n$. These orientations are given by the corresponding Euler angles. For rigid molecules using these parameters, the symmetry group, and the molecular geometry, it is possible to calculate the position of all the atoms in the unit cell. For flexible molecules, as discussed above, the *genome* needs to be augmented by k scalars, $\Phi_1...\Phi_k$, that give the values of the k single-bond dihedral angles, measured in the molecular frame, that are allowed to change during the global optimization. It is very important to understand that the inclusion of these dihedral angles in the *genome* allows the GA to sample regions of the conformation space with quite different dihedral angles than those used as starting parameters. While these regions may not be energetically favorable for the isolated molecule they may be favorable for certain packing configurations. Note that regardless of which dihedral angles are included in the *genome*, all the intramolecular geometry parameters—bond lengths, bond angles, and dihedral angles—are always locally optimized in every GA generation.

The MGAC program only considers the lattice angles (α, β, γ) as independent parameters in the GA optimization. The lattice lengths (a, b, c) are treated as dependent parameters derived from the position of the molecules in the unit cell. In this respect, MGAC uses an approach similar to the one in *Polymorph Predictor*.[9] The lattice lengths are determined as follows: Given the molecular coordinates and the lattice angles that define the crystal structure, the initial estimates of the values of lattice lengths are chosen such that they define the smaller space that encloses all the molecules in the asymmetric unit. To reduce the chance of producing very short intermolecular distances between molecules in the unit cell and their neighbors, which can lead to spurious results when locally optimized, the asymmetric unit is extended to guarantee a minimal intermolecular distance (by default, 3 Å). Note that this arbitrary determination of the initial values of the lattice parameters does not have any effect on the final crystal structures as they undergo a local minimization in which all the inter- and intramolecular parameters are optimized. This last step allows that the effects of molecular interactions be included in the local refinement of the entire crystal structure.

Modified Genetic Algorithm. Except for the addition of the lattice angles, the crystal structure encoding given in the last section resembles the one used to describe molecular clusters. A very efficient GA scheme for molecular cluster optimization has been previously proposed by Niesse et al.,[19,20] and following this precedence, we have implemented in MGAC the *one-point-crossover*, *two-point-crossover*, *n-point-crossover*, *uniform-crossover*, *arithmetic-crossover*, *inversion-crossover*, *geometric-crossover*, and *gaussian mutation* genetic operators.[20] All of these operators are used in MGAC when acting on the molecular parameters. For the lattice angles, the *inversion-crossover* was removed from the operator list to preserve the crystal system, such that the resulting group of operators always transforms a triclinic structure into another triclinic and so forth.

Starting from a set of crystal structures randomly created (the initial generation or population), one can use these operators to construct a new set of crystal structure candidates for the next generation. The program verifies that the lattice angles define a linear independent set of lattice vectors in three dimensions, eliminating any spurious quasi-planar or linear structures. Any structure that does not meet this test is eliminated and replaced by a new one randomly generated. This procedure is repeated until a complete new population is generated. At each GA evolution, all the new structures (by default, the number of new structures is half of the population size) are relaxed to the corresponding local minima using the local optimization techniques available in CHARMM.[21,22] The local optimization, in which all the inter- and intramolecular degrees of freedom are allowed to change, is performed within the desired space group and produces a new set of candidate solutions that compete with the preexisting solutions for their permanence in the population. This competition is implemented by combining both sets of solutions into a larger population from which the worst candidates are eliminated until the number of structures equals the desired population size. The fitness of the solution is given by the total energy of the crystal structure evaluated after the local optimization. This defines the population from which the next generation can be obtained by repeating the procedure just described above. This evolution is repeated either for a predetermined number of generations or until the diversity of the population reaches such uniformity that the GA procedure becomes a random search.

The algorithm described above can be used to perform constrained searches in any of the 230 space groups with an arbitrary number of molecules and molecule types on the asymmetric unit. Therefore, MGAC can be used by systematically searching solutions in different symmetry groups, normally restricting the search to the most common ones or by unconstrained searches, a methodology that has not yet been proved effective in CSP.

To reach the high sampling power required for these studies, we have used a global parallelization scheme of the genetic algorithm implemented in MGAC. In every generation, the parallelization scheme relies on the simultaneous run of the local optimization of the crystal structures belonging to the same population. The parallelization was done using our adaptive parallel genetic algorithm, *APGA*.[23] *APGA* was designed to perform efficiently on a heterogeneous cluster of computers and to provide a great degree of adaptability and performance in distributed systems. This level of parallelization allows making the execution time of MGAC approximately independent of the population size when a sufficient number of processors is available for the calculations.

The *MGAC* package is written in C++, using *BASH*[24] for scripting, *MPICH*[25] for parallel programming, *GALib*[26] for the GA implementation, and *xerces-c*[27] for parsing input and output *XML*[28] files.

Force Field Generator. An automatic force field generator, *charmmgen*, was implemented on the basis of *antechamber*.[29] Given the molecular composition and connectivity, this program calculates the molecule's force field parameters on the basis of the general amber force field, GAFF,[30] or any other force field library with a similar functional form. In this work, we have used the default parameters of GAFF,[30] which is a general force field with parameters for most organic and pharmaceutical molecules containing H, C, N, O, S, P, and halogens. It uses a simple functional form in which the energy is defined by

$$E = \sum_{\text{bonds}} k_r(r - r_{\text{eq}})^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_{\text{eq}}) + \sum_{\text{dihedrals}} \frac{v_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{\ 12}} - \frac{B_{ij}}{R_{ij}^{\ 6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right]$$

where r_{eq} and θ_{eq} are equilibrium structural parameters; k_r, k_θ, and v_n are the force constants; n is the multiplicity; and γ is the phase angle for the torsional angle parameters. The A, B, and q parameters describe the nonbonded potentials. In this work, we calculated the atomic charges, q_i, using the restrained electrostatic potential approach implemented in the RESP program.[31] This program fits the atomic charges to reproduce the electrostatic potential generated by the molecule's charge distribution, which was calculated using the *Gaussian 03*[32] package at the HF/6-31G* level. The atomic charges used for *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide were calculated at the optimized geometry (HF/6-31G* level) of the isolated molecule, but it was verified that they do not change in any appreciable way for other conformations of the molecule.

All the crystal energy calculations and local optimization are performed using CHARMM[21,22] with the GAFF[30] parameters. At least two unit cells were included in the simulation box in every direction, short range nonbonded interactions were summed up to a cutoff of 14 Å, and the electrostatic interactions were calculated using the Ewald technique.

Analysis of the Resulting Structures. Once a series of MGAC runs has been performed, it is necessary to sort and compare the crystal structures generated in these runs to find the *n* unique ones with the lowest energies. Because of numerical fluctuations, the set of structures generated by the MGAC runs has many similar structures with small energy differences that correspond to the same physical structure. Therefore, it is necessary to detect and remove these duplicate crystal structures from the final list. There are several well-established methodologies for comparing three-dimensional crystal structures, such as a comparison of their computed powder patterns,[33,34] and the *CRYCOM*[14] or *COMPACK*[15] programs, but because the source codes of these programs are not available, they could not be easily integrated into our computer framework. Therefore, we have implemented an alternative method that was easily integrated into our crystal prediction environment.

Our approach is similar to the one implemented in *COMPACK*,[15] based on the comparison of the structures of two finite clusters, but because MGAC preserves the atomic labels of the molecules, we avoid the expensive computational step of building a classification tree to sort the atomic labels of the molecules in the two fragments.

Our method uses two spherical clusters of *n* and *m* molecules, with $m > n$, that represent the three-dimensional structures of the two crystal structures under comparison. We call these two clusters the reference and compared fragments, respectively. The number of molecules in each cluster needs to be sufficiently large to completely characterize the environment surrounding the central molecule of each fragment, which should be one of the molecules in the asymmetric unit. By default, the values of *n* and *m* are set to 16 and 24, respectively, but it should be emphasized that the comparisons are always done using the same number of molecules in each fragment, that is, matching the reference cluster with only a fraction of the molecules in the compared cluster. Our experience shows that very small changes in crystal structures can produce a different set of molecules in the list of the *n* ones closest to the central molecule. This can lead to the classification of two similar crystals as different ones because a dissimilar set of molecules has been included in the two finite clusters. Hence, it is convenient to have a larger compared cluster to ensure that all the molecules in the smaller reference fragment are included in the compared cluster. Because within MGAC we preserve the atomic labeling of the individual molecules, the comparator program can easily search for the best rotation and translation to superimpose the chosen central molecules of each fragment. This is done by minimizing the root-mean-square (RMS) deviations of their atomic coordinates, excluding the hydrogen atoms, using the overlapping points algorithm develop by Kabsch.[35,36] If the resulting RMS is higher than a given threshold (by default 0.5 Å), there is no match between the selected pair of central molecules and a new set may be chosen (see below). If the RMS is lower than the threshold, the comparator searches for the best superposition between the reference fragment and a subset of *n* molecules in the compared fragment. If the resulting RMS is lower than a given value (by default, 1.0 Å), the two crystals are considered similar. If any of these comparisons, that is, between the central molecules and/or between the fragments, do not show a match and both crystal structures have only one molecule in their asymmetric unit, they are considered different. Otherwise, the comparator program repeats the process described above, constructing a new compared cluster for a different central molecule. This process continues until all possible molecules in the asymmetric unit of the compared crystal structure have been used as central molecules of the compared cluster. If no match is found, the entire process is repeated, but applying the inversion operation to the compared fragments. Finally, if no match is found when using the inversion operation, the two crystals are considered different.

The crystal comparator, named *MOLCRY*, has been written in C++ and *Python*[37] as an extension of the Computational Crystallography Toolbox.[38] We have implemented in C++ those sections of the code that are computationally intensive, and Python code is used to interface the different parts of the code into a coherent application. Furthermore, Python was used to develop a set of Web services to automatically process and store the crystal structures produced by MGAC (see below) into a user data base.

When the crystal comparator is used, it is possible to create the list of the *n* unique crystal structures with lowest energy for a given set of MGAC runs. This is accomplished by comparing any candidate to be added to the list against a subset of the structures already in it. This subset of structures is chosen such that their crystal structures do not differ in energy by more than a given amount (by default, 16 kJ/mol) of the new candidate structure. This limits the number of comparisons needed per candidate, assuring that the time required for building the list scales linearly with its size.

Web Services for Crystal Analysis. In order to automate and standardize the process of classification and comparison of the crystal structures, we have created Web services that perform the comparison and classification of the structures. These services provide also an interface to a user database in which the crystal structures studied in our laboratory can be maintained. These services were developed in Python using representational state transfer or REST[39] architecture and the *Django* Web framework.[40] REST was implemented on top of the hypertext transfer protocol or HTTP.[41] This procedure greatly reduces the human time required for the analysis process, opening the possibility of using this framework for high-throughput CSP work. A detailed technical description of the implementation of these Web services is given as part of the Supporting Information.

Results and Discussion

The crystal structure of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide has been recently used as a test case for flexible molecules in the last Cambridge blind test, CSP2004.[4] The crystal structure of this molecule became one of the most difficult to predict by all the crystal prediction methodologies presented in the CSP2004. Actually, none of the 18 participants in the CSP2004 found its experimental structure within the first three best candidates, which was the criterion used in CSP2004 for success.

While the structure of this compound is now known,[4] all the calculations reported here have been done as if performing a blind test; that is, no information of the experimental structure was used a priori in our calculations. A series of 20 MGAC runs for each of the 14 most common space groups in organic molecules—$P1$, $P\bar{1}$, $P2_1$, $C2$, Pc, Cc, $P2_1/c$, $C2/c$, $P2_12_12_1$, $Pca2_1$, $Pna2_1$, $Pbcn$, $Pbca$, and $Pnma$—were executed for structures with one and two molecules per asymmetric unit. Each GA run produced 130 generations with 30 crystal structures each, using a crossover probability of 1.0 and a mutation probability of 0.001. This procedure gave a total of 560 MGAC runs in which approximately 1.1 million crystal structures were evaluated for this molecule. This required an approximated total of 80 000 CPU hours from the computational resources that were available at the CHPC Arches Metacluster[42] and the NCSA Teragrid cluster.[43] The first 200 different structures, sorted by energy, were extracted from the collection of all the structures (including repeated ones) generated by the combined runs. When MOLCRY running on a PC AMD Athlon 64 X2 2.2 GHz was used, this entire selection process was done in 240 min. From this list, 11 planar crystal structures were removed, and the resulting 189 structures were analyzed for missing symmetries using *PLATON*'s *ADDSYM* algorithm.[16]

In Table 1, we present the 20 crystal structures with the lowest energy (the extended list of the 189 crystal structures is provided as Supporting Information). The experimental crystal structure was found ranked third by energy. Note that according to the criterion used in the CSP blind tests this should be considered a successful prediction. The comparison between this structure and the experimental one is given in Figure 1 and Table 2, showing the excellent agreement between them. Figure 2 depicts the first three crystal structures in Table 1 and their corresponding simulated powder diffraction patterns. It is apparent that, in spite of their close energies, they are quite different structures. This makes the agreement between the third structure and the experimental one even more remarkable. Clearly, our results suggest that future experimental work in searching for these polymorphs is highly desired.

It should be noted that the experimental crystal structure was found in a search constrained to the $P2_1$ space group with $Z' = 2$, and not in the $P2_1/c$ with $Z' = 1$ as it is known experimentally. The systematic absence of the experimental structure in the search constrained to the $P2_1/c$ space group suggests that it is harder for the MGAC algorithm to find this crystal structure in its own symmetry group than in a less restricted search at $P2_1$. To try to better understand this behavior, a separated series of 20 constrained GA runs in $P2_1/c$ with $Z' = 1$, $P2_1$ with $Z' = 2$, and $P1$ with $Z' = 4$ was executed using MGAC. Note that in any of these three searches it is possible to find the experimental crystal structure. For each kind of search, we averaged the lowest energy reached by each of them at each generation. These average values and their corresponding standard deviations

Table 1. First 20 Lower-Energy Crystal Structures of *N*-(2-Dimethyl-4,5-dinitrophenyl) Acetamide[a]

ranking	*a* Å	*b* Å	*c* Å	α	*β*	*γ*	volume Å^3	space group	addsym	energy[b] kJ mol
1	4.842	14.142	15.118	90.00	102.53	90.00	1011	$P2_1/c$		0.00
2	14.865	8.097	16.749	90.00	90.00	90.00	2016	*Pbca*		0.39
3	**12.554**	**4.799**	**19.406**	**90.00**	**118.58**	**90.00**	**1027**	$\boldsymbol{P2_1}$	$\boldsymbol{P2_1/c}$	**1.14**
4	4.849	9.659	11.953	109.99	96.84	96.94	514	$P\bar{1}$		1.82
5	8.262	12.971	9.590	90.00	95.65	90.00	1023	$P2_1/c$		2.03
6	10.247	11.064	9.470	90.00	105.71	90.00	1034	$P2_1/c$		2.38
7	5.030	13.115	15.861	90.00	107.79	90.00	996	$P2_1/c$		2.40
8	4.915	14.318	7.319	90.00	102.79	90.00	502	$P2_1$		2.40
9	8.105	16.598	8.179	90.00	112.34	90.00	1018	$P2_1/c$		2.83
10	3.955	16.617	7.809	90.00	98.58	90.00	508	$P2_1$		3.47
11	9.037	15.783	7.481	90.00	112.19	90.00	988	$P2_1/c$		3.59
12	8.422	9.536	15.112	90.00	121.81	90.00	1031	*Pc*	$P2_1/c$	3.73
13	11.821	4.825	19.397	90.00	111.96	90.00	1026	$P2_1$	$P2_1/c$	3.76
14	7.324	14.988	9.524	90.00	99.36	90.00	1032	*Pc*	$P2_1/c$	4.10
15	8.310	8.748	14.823	94.40	95.98	100.99	1047	$P\bar{1}$		4.17
16	8.420	8.814	14.460	86.82	82.59	79.64	1046	$P\bar{1}$		4.23
17	9.408	14.723	14.990	90.00	90.00	90.00	2076	$P2_12_12_1$		4.46
18	4.773	14.640	15.048	90.00	92.83	90.00	1050	$P2_1/c$		4.58
19	9.201	7.995	14.330	90.00	100.15	90.00	1038	$P2_1$		4.59
20	4.828	14.921	14.448	90.00	95.12	90.00	1037	$P2_1$		4.67

[a] The third-ranked structure matches the experimental known structure for this molecule. [b] Relative energies with respect to the lower crystal structure energy of −309.5077 kJ/mol.

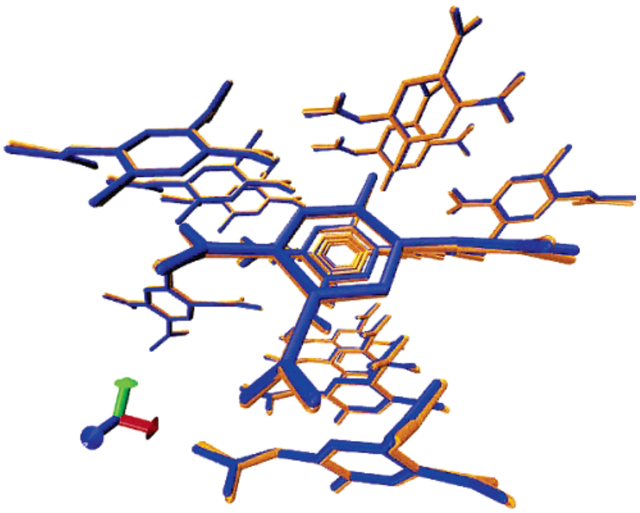

Figure 1. Superposition between experimental (blue) and predicted (orange) crystal fragments (16 molecules) of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide crystal structures. The total RMS for this superposition is 0.158 Å.

Table 2. Comparison between the Experimental and Predicted Crystal Structures of *N*-(2-Dimethyl-4,5-dinitrophenyl) Acetamide[a]

	experimental	predicted	difference
a (Å)	12.569	12.554	−0.33%
b (Å)	4.853	4.799	−1.1%
c (Å)	19.672	19.406	−1.3%
α	90.00	90.00	
β	119.95	118.58	−1.1%
γ	90.00	90.00	
vol (Å^3)	1040	1026	−1.3%
mRMS (Å)		0.076	
RMS (Å)		0.158	

[a] mRMS and RMS represent the best molecular and total root mean square.

are shown in Figure 3. It is interesting to observe that the optimizations with $Z' = 2$, while starting at a higher energy, after 70 generations find better solutions than those with $Z' = 1$. This is a remarkable behavior because it would be expected that the increase in the dimension of the configurational space for $P2_1$ would make it harder to find low-energy structures. Clearly this is what happens in the case of the search in $P1$, where the increased difficulty of finding low-energy structures is clearly noticeable by its slow convergence. MGAC is more efficient, at least for this molecule, when there are two molecules per asymmetric unit, and we speculate that this might indicate that the hypervolume of the attractive basin of the minimum corresponding to the experimental structure is much smaller for the search constrained to $P2_1/c$ and therefore less probable to find in the $P2_1/c$ configurational space than in $P2_1$.

The successful prediction of the experimental crystal structure of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide presented in this paper clearly contrasts with the results presented at the CSP2004. First, none of 18 participants found the experimental structure between the “top three” predictions, and only four participants found it in their extended list (up to 135 structures). Moreover, the best predicted structure exhibits a larger deviation from the experimental structure (RMS of 0.5 Å)[4] than the best structure reported here (RMS of 0.158 Å). This level of agreement shows that both the search methodology and the force field used in this study perform well for this molecule and most likely in similar compounds.

The energy range of the 189 structures collected in the final list is only ∼9.6 kJ/mol (Figure 4). This means that there are ∼20 crystal structures per kJ/mol, but as shown in Figure 4, their distribution as a function of energy is not

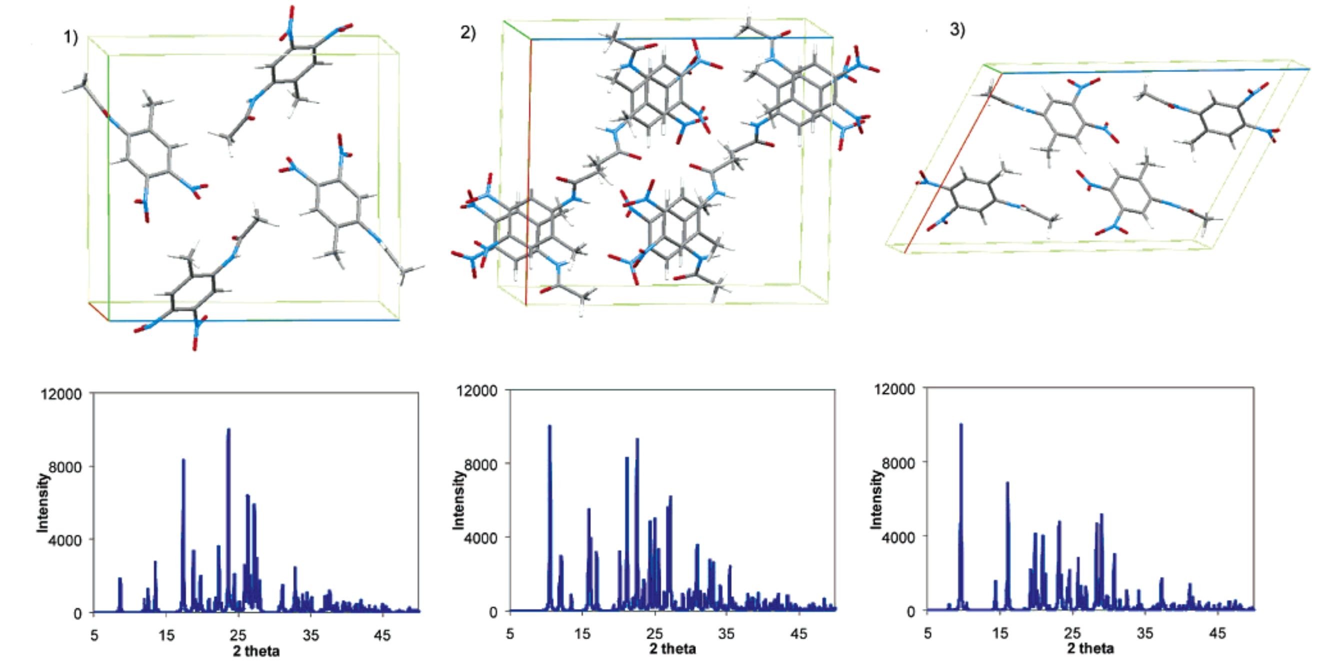

Figure 2. Comparison of the structures and powder diffraction spectra of the three lowest-energy structures of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide found in this study.

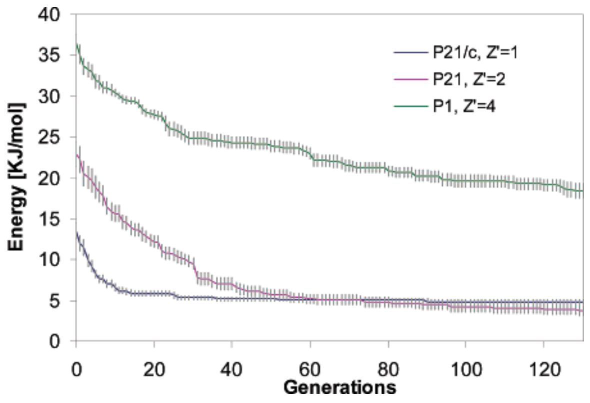

Figure 3. Average of the lowest energy per generation for 20 runs in three settings: $P2_1/c$ ($Z' = 1$), $P2_1$ ($Z' = 2$), and $P1$ ($Z' = 4$).

homogeneous. Even in this narrow range of energies, it is easy to appreciate the fast growth of the number of crystal structures close in energy to the global minimum. This shows that there are a large number of configurations with significant low energies to be considered as acceptable candidates for a minimum. Unfortunately, this can contribute to a rapid stagnation of the GA population and suggest that techniques like multiple independent runs (as used in this work), multiple concurrent populations, or random immigrants are necessary to obtain the desired convergence of the GA.

Conclusions

This paper reports the implementation and testing of a new distributed computing environment for CSP based on our previous work on GA. The environment allows the search of crystal structures for either rigid of flexible molecules without any restriction in the symmetry group or number of

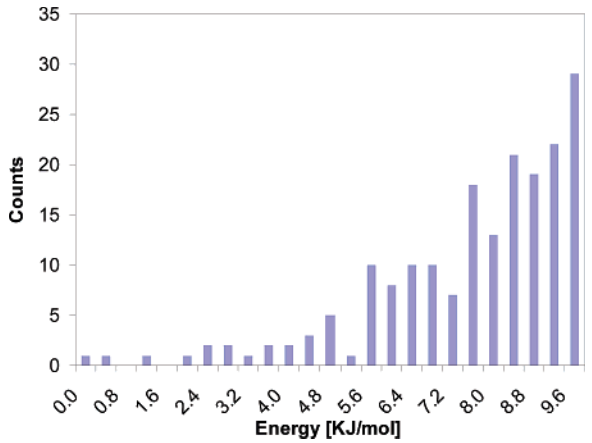

Figure 4. Energy histogram of the first 189 different crystal structures for the *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide molecule.

molecules in the asymmetric unit. The computational environment allows for the automatization of many processes, like the generation of the molecule's force field, execution of multiple GA runs, and the comparison and archival of relevant structures necessary for CSP studies.

As an example, we have shown that the method can predict (using the CSP2004 criterion) the crystal structure of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide, which none of the methods presented at CSP2004 were able to predict correctly. Moreover, our predicted structure agrees much better with the experimental one than any of those found in the extended lists of the CSP2004 blind test.

Another important conclusion that can be drawn from the difficulties encountered in predicting the crystal structure of this molecule is that the prediction of the crystal structures of flexible molecules requires advance search procedures with extensive sampling capabilities as much as it requires greater accuracy in the modeling of the intra- and intermolecular energies.[44,45] Using the computer framework described here, we are performing more extensive tests of the MGAC method including a larger set of molecules and different force fields. The MGAC software will be available, under open-source licensing agreements, later in 2007.

Acknowledgment. The Center for High Performance Computing provided computer resources in the Arches cluster for this project. The Arches cluster was partially funded by a grant from the NIH-National Center for Research Resource (# 1S10RR017214-01). Additional computer resources were provided by NSF Teragrid MCA05T018 grant. The software for this work used the GAlib genetic algorithm package, written by Matthew Wall at the Massachussets Institute of Technology. MBF greatly acknowledge financial support from Universidad de Buenos Aires (UBACYTX035) and form the Argentinean CONICET (PIP5119/05). We thank two anonymous referees for providing critical comments on how to improve the presentation of our results.

Supporting Information Available: We provide a table with the data from the best 189 unique crystal structures of *N*-(2-dimethyl-4,5-dinitrophenyl) acetamide found in this study and the details of the implementation of Web services for crystal structure analysis and archival. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Threlfall, T. L. Analysis of Organic Polymorphs A Review. *Analyst* **1995**, *120*, 2435.

(2) Dunitz, J. D.; Bernstein, J. Disapparing Polymorphs. *Acc. Chem. Res.* **1995**, *28*, 193.

(3) Erk, P.; Hengelsberg, H.; Haddow, M. F.; Gelder, R. v. The Innovative Momentum of Crystal Engineering. *CrystEngComm* **2004**, *6*, 474.

(4) Day, G. M.; Motherwell, W. D. S.; Ammon, H. L.; Boerrigter, S. X. M.; Valle, R. G. D.; Venuti, E.; Dzyabchenko, A.; Dunitz, J. D.; Schweizer, B.; Eijck, B. P. v.; Erk, P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Leusen, F. J. J.; Liang, C.; Pantelides, C. C.; Karamertzanis, P. G.; Price, S. L.; Lewis, T. C.; Nowell, H.; Torrisi, A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; Verwer, P. A Third Blind Test of Crystal Structure Prediction. *Acta Crystallogr., Sect. B* **2005**, *61*, 511.

(5) Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Dzyabchenko, A.; Erk, P.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Lommerse, J. P. M.; Mooij, W. T. M.; Price, S. L.; Scheraga, H.; Schweizer, B.; Schmidt, M. U.; Eijck, B. P. v.; Verwer, P.; Williams, D. E. Crystal Structure Prediction of Small Organic Molecules: A Second Blind Test. *Acta Crystallogr., Sect. B* **2002**, *58*, 647.

(6) 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.; Eijck, B. P. v.; Verwer, P.; Williams, D. E. A Test of Crystal Structure Prediction of Small Organic Molecules. *Acta allogr., Sect. B* **2000**, *56*, 697.

(7) Bazterra, V. E.; Ferraro, M. B.; Facelli, J. C. Modified Genetic Algorithm to Model Crystal Structures. I. Benzene, Naphthalene and Anthracene. *J. Chem. Phys.* **2002**, *116*, 5984.

(8) Bazterra, V. E.; Ferraro, M. B.; Facelli, J. C. Modified Genetic Algorithm to Model Crystal Structures: III. Determination of Crystal Structures Allowing Simultaneous Molecular Geometry Relaxation. *Int. J. Quantum Chem.* **2004**, *96*, 312.

(9) Gdanitz, R. J. Ab Initio Prediction of Possible Molecular Crystal Structures. In *Theoretical Aspects and Computer Modeling of the Molecular Solid State*; Gavezzotti, A., Ed.; John Wiley and Sons: New York, 1997; p 185.

(10) Gavezzotti, A. Ten Years of Experience in Polymorph Prediction: What Next? *CrystEngComm* **2002**, *4*, 343.

(11) Gavezzotti, A. Are Crystals Structures Predictable? *Acc. Chem. Res.* **1994**, *27*, 309.

(12) Day, G. M.; Chisholm, J.; Shan, N.; Motherwell, W. D. S.; Jones, W. An Assessment of Lattice Energy Minimization for the Prediction of Molecular Organic Crystal Structures. *Cryst. Growth Des.* **2004**, *4*, 1327.

(13) Apostolakis, J.; Hofmann, D. W. M.; Lengauer, T. Derivation of a Scoring Function for Crystal Structure Prediction. *Acta Crystallogr., Sect. A* **2001**, *57*, 442.

(14) Dzyabchenko, A. V. Method of Crystal-Structure Similarity Searching. *Acta Crystallogr., Sect. B* **1994**, *50*, 414.

(15) Chisholm, J. A.; Motherwell, W. D. S. COMPACK: A Program for Identifying Crystal Structure Similarity Using Distances. *J. Appl. Crystallogr.* **2005**, *38*, 228.

(16) *PLATON, A Multipurpose Crystallographic Tool*; Utrecht University: Utrecht, The Netherlands, 2005.

(17) *Cerius2*; Accelrys: San Diego, CA, 1997.

(18) Golberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: New York, 1989.

(19) Niesse, J. A.; Mayne, H. R. Global Optimization of Atomic and Molecular Clusters Using the Space-Fixed Modified Genetic Algorithm Method. *J. Comput. Chem.* **1997**, *18*, 1233.

(20) White, R. P.; Niesse, J. A.; Mayne, H. R. A Study of Genetic Algorithm Approaches to Global Geometry Optimization of Aromatic Hydrocarbon Microclusters. *J. Chem. Phys.* **1998**, *108*, 2208.

(21) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187.

(22) MacKerell, A. D.; Brooks, J., B.; Brooks, C. L., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Schreiner, P. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P., Schaefer, H. F., III, Ed.; John Wiley & Sons: Chichester, U. K., 1998; p 271.

(23) Bazterra, V. E.; Cuma, M.; Ferraro, M. B.; Facelli, J. C. A General Framework to Understand Parallel Performance in Heterogeneous Clusters: Analysis of a New Adaptive Parallel Genetic Algorithm. *J. Parallel Distrib. Comput.* **2005**, *65*, 48.

(24) GNU BASH. http://www.gnu.org/software/bash/ (accessed Oct 3, 2006).

(25) MPICH. http://www-unix.mcs.anl.gov/mpi/mpich (accessed Oct 3, 2006).

(26) Wall, M. GAlib. http://lancet.mit.edu/ga/ (accessed Oct 10, 2006).

(27) The Apache Project: Xerces-C++ Parser. http://xml.apache.org/xerces-c/ (accessed Oct 3, 2006).

(28) W3C Extensible Markup Language (XML). http://www.w3.org/XML/ (accessed Oct 3, 2006).

(29) Wang, J. Antechamber. http://amber.scripps.edu/antechamber/antechamber.html (accessed Oct 3, 2006).

(30) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157.

(31) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97*, 10269.

(32) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(33) Gelder, R. d.; Wehrens, R.; Hageman, J. A. A Generalized Expression for the Similarity of Spectra: Application to Powder Diffraction Pattern Classification. *J. Comput. Chem.* **2001**, *22*, 273.

(34) Karfunkel, H. R.; Rohde, B.; Leusen, F. J. J.; Gdanitz, R. J.; Rihs, G. Continuous Similarity Measure between Nonoverlapping X-ray Powder Diagrams of Different Crystal Modifications. *J. Comput. Chem.* **1993**, *14*, 1125.

(35) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr., Sect. A* **1976**, *32*, 922.

(36) Kabsch, W. A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr., Sect. A* **1978**, *34*, 827.

(37) Python Programming Language. http://www.python.org/ (accessed Oct 3, 2006).

(38) Computational Crystallography Toolbox CCTBX. http://cctbx.sourceforge.net/ (accessed Oct 3, 2006).

(39) Fielding, R. T. REST. http://www.ics.uci.edu/∼fielding/pubs/dissertation/top.htm (accessed Oct 3, 2006).

(40) Django. http://www.djangoproject.com/ (accessed Oct 3, 2006).

(41) W3C: HTTP. http://www.w3.org/Protocols/ (accessed Oct 3, 2006).

(42) Arches Metacluster. http://www.chpc.utah.edu/docs/manuals/user_guides/arches/ (accessed Oct 3, 2006).

(43) NCSA TeraGrid IA-64 Linux Cluster. http://www.ncsa.uiuc.edu/UserInfo/Resources/Hardware/TGIA64LinuxCluster/ (accessed Oct 3, 2006).

(44) Eijck, B. P. v.; Mooij, W. T. M.; Kroon, J. Ab Initio Crystal Structure Predictions for Flexible Hydrogen-Bonded Molecules. Part II. Accurate Energy Minimization. *J. Comput. Chem.* **2001**, *22*, 805.

(45) Ouvrard, C.; Price, S. L. Toward Crystal Structure Prediction for Conformationally Flexible Molecules: The Headaches Illustrated by Aspirin. *Cryst. Growth Des.* **2004**, *4*, 1119.

CT6002115