Genetic Algorithms in Conformational Analysis

Nikhil Nair and Jonathan M. Goodman*

Department of Chemistry, Lensfield Road, Cambridge, CB2 1EW

Received September 22, 1997

A genetic algorithm-based method has been designed and shown to be effective for the conformation searching of unbranched alkanes. A measure of diversity is defined and used to investigate the best parameter settings. Except for very short alkanes, genetic algorithms are very much more effective than Monte Carlo searches. The procedure can be used for molecules other than unbranched alkanes, and PM-toxin A is used as an example for which the method works well.

INTRODUCTION

Exploration of the conformational space of molecules is an important and intrinsically difficult process.[1−3] For flexible molecules, the process may be prohibitively difficult, even for quite small systems. The folding of peptides and other biological polymers can only be understood by studying fairly large molecules, because short oligopeptides do not show secondary or tertiary structure. Unbranched alkanes alter their ground-state properties from being extended to being folded once they pass a certain size,[4] and so they are the simplest examples of this class of molecules. Accurate force field parameters for alkanes are available, and so the difficulty of this sort of exploration lies only in the conformation search and not in the force field. For this study we have used the MM2* force field[5] as implemented in MacroModel.[6]

Genetic algorithms have shown great promise in conformation searching.[7] They have already been used to study alkane conformations,[8] but this earlier work only investigated short alkanes for which the global minimum will be the extended conformation.[4] Longer alkane chains have folded global minima, which are harder to find. Will genetic algorithms be effective in investigating the conformation space of such systems?

METHODS

The genetic algorithm program follows standard procedures.[9] Conformations of the alkanes differ principally in their dihedral angles, and so the "chromosome" for the genetic algorithm consists of a string of real numbers corresponding to the dihedral angles of the alkane in sequence. The program is limited to unbranched alkanes, so the mapping of dihedral angles to numbers in the string is straightforward. The program starts by randomly generating a pool of conformations. By default, this contains 10 different geometries. The "fitness" of each conformation is then calculated by minimization of each using MacroModel and the MM2* force field and using the energy of the minimized structure as the measure of fitness. This evaluation of fitness alters the chromosome for each structure, as the dihedral angles may change during the process. This evaluation may be regarded as corresponding to a Lamarckian view of evolution,[10] but the changes introduced tend not to be very large and the structures of interest are the minima.

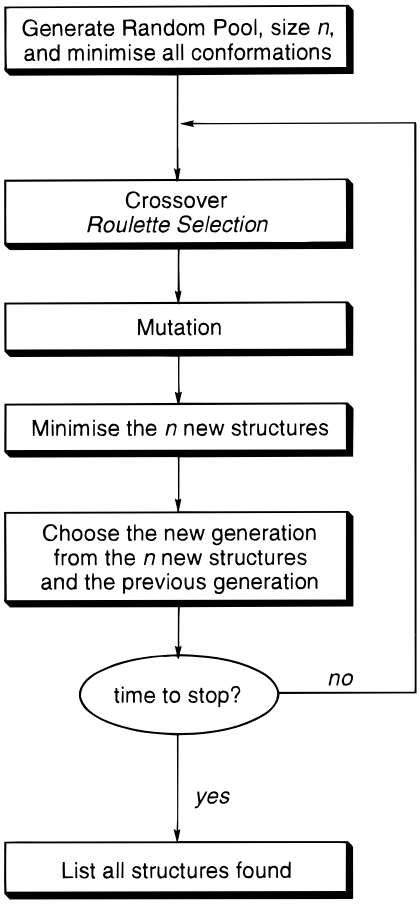

Figure 1. The genetic algorithm procedure.

The algorithm is illustrated as a flowchart in Figure 1. Each new generation is produced by combining and mutating the chromosomes of the present pool of conformations. First, the chromosomes are altered using a crossover operation: two chromosomes are selected, the position at which to break the chromosomes is chosen, then each is split into two at this position. The pieces are recombined to form two new chromosomes. The choice of chromosomes for crossover is decided by a roulette-based method: the chance of a

S0095-2338(97)00433-2 CCC: $15.00
Published on Web 02/06/1998

particular chromosome being selected is proportional to the Boltzmann factor of the energy of the corresponding conformation. Thus, low-energy conformations are more likely to be selected than high-energy conformations. The "temperature" of the crossover roulette selection, *s*, has a default setting of 10 000 K, which means that the bias toward low-energy conformations is small. The two chromosomes needed for each crossover are chosen separately, so it is possible for the same chromosome to be chosen twice. In such a case, the crossover would have no effect. The crossover rate can be adjusted from zero (no crossover) to one (as many crossover steps as there are members of each generation).

After crossover has been completed, mutations occur by altering the dihedral angles of each structure. Randomly altering all the angles would lose the information present in the parent conformation, so only a few of the angles are altered. This alteration is controlled by a parameter *m*. The chance of a particular torsion angle being mutated is *m* divided by the number of flexible torsion angles in the alkane. Using the default setting ($m = 0.4$), about two torsion angles in every five conformations are altered, so that more than half of the conformations are likely to remain unaffected by that mutation step. Because the dihedral angles prefer to be 60°, 180° or −60°, the change in angle is biased toward these figures by choosing random numbers from three superimposed normal distributions.

These two procedures create a new generation of conformations, which are minimized by MacroModel. However, there is no reason to suppose that they will be better than their parent conformations. The new pool, therefore, is chosen from the parents and the children, again using a roulette-based method, with the chance of a particular conformation being chosen increasing with its Boltzmann factor. This time, the selection temperature, *r*, has a default value of 1000 K, so the choice is biased toward low-energy conformations. An energy penalty, *d*, with a default value of 10 kJ mol^{-1}, is introduced to reduce the chance of the same structure being chosen more than once, because this would reduce the diversity of the pool. As soon as the correct number of chromosomes has been selected, the new pool is subjected to selection, crossover, mutation, and replacement, as before, to form the next generation.

When the required number of generations has been completed, the program lists all the conformations that have been encountered, sorted by energy. The program treats enantiomeric conformations as being identical, and so does not produce two different entries for such species. Energies may be repeated, however, because dihedral angles are always measured from one end of the chain. Thus, a conformation with a twist at the beginning is considered to be different from a conformation with the same twist at the end of the chain.

It is important that the pool contains a diverse set of structures. Diversity is hard to define, however, particularly for an angular measure for which dihedral angles of +179° and −179° are almost equivalent, despite being numerically very different. This problem can be overcome by expressing each angle as a complex number with modulus 1 [*i.e.*, replacing each angle θ_k with $z_k = exp(i\theta_k)$]. The modulus of the mean of all the complex numbers z_k will be a measure of diversity. Graphically, every angle θ_k corresponds to a point on a circle with unit radius. The mean position of these points will be somewhere inside the circle, and its distance from the center of the circle will be a measure of the diversity. If the points are evenly distributed around the circle, the mean position will be the center of the circle. If the points are clustered around a particular part of the circle, the mean position will be away from the center of the circle and the diversity of the points will be lower. If all the points are in the same place, the mean position will be on the circle, and the diversity will be minimal. We can, therefore, define the diversity of a set of torsion angles as $1 - |\bar{z}|$, which is convenient to express as a percentage. The diversity of a population of molecules is the average of the diversities of the individual angles. After minimization, a randomly initialized group of infinite size should have a diversity of <100% because there is a preference for 60° and 180° torsion angles. The diversity of the initial generations increases slowly toward ~80%, which is the same as would be obtained if the torsion angles were divided equally between extended and staggered conformations.

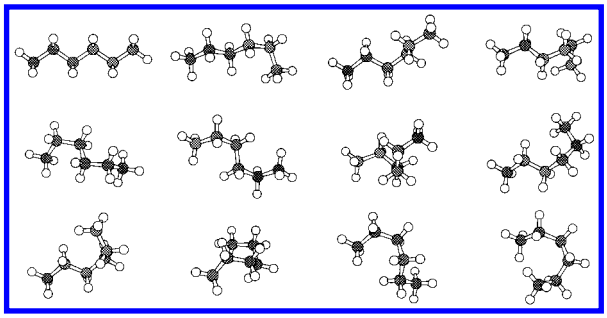

Figure 2. All the minima of hexane.

The executable for the program, compiled for Silicon Graphics Workstations, is available on http://www.ch.cam.ac.uk/MMRG/software/. Note that MacroModel, which is available from Professor Clark Still at Columbia University, is required to run the program.

RESULTS

The method was first applied to hexane, a molecule that is sufficiently simple for us to be confident that all conformations have been found. MacroModel was used to perform a systematic search, and this generated the 12 conformations shown in Figure 2. MacroModel automatically discards mirror image conformations conformations that differ only in the direction in which the chain is numbered. The genetic algorithm code does not have this second refinement, so it should be able to find 20 distinct conformations of hexane, eight of which are pairs of the same energy.

The results are given in Table 1. Every run involves random changes, so the results each represent the average of eight runs. The program only provides information between generations, so the number of minimizations required to find the global minimum must be a multiple of the pool size.

Table 1 results show that a pool size of two is too small for this conformational search, but that the search is not very sensitive to pool size for pools with 4−20 members. The runs with a pool size of 20 appear to take a long time to find the global minimum, but this is because the number of steps must be a multiple of the pool size, so the individual

Table 1. Results of Genetic Algorithm Calculations on Hexane[a]

pool size	options[b]	steps[c] needed to find global minimum	initial diversity	diversity after 200 steps[c](%)	minima after 200 steps[c]	minima after 500 steps[c]
2	default	64	0.43	16	12.7	15.5
4	default	35	0.54	33	12.9	15.4
6	default	33	0.73	37	15.3	15.7
8	default	24	0.72	47	14.9	16.5
10	default	21	0.70	58	15.0	16.4
20	default	33	0.77	69	17.3	18.1
10	*n = 0*	95	0.72	55	14.6	16.7
10	*m = 0*	20	0.70	48	14.7	15.6
10	*m = 0.1*	24	0.74	52	15.3	16.2

[a] Every entry is the average of eight separate runs. [b] The probability of a crossover step is *c*. The probability of a particular torsion angle being mutated is *m/3*, because there are three torsion angles being rotated. Default values: *c = 1.0*; *m = 0.4*. [c] Each minimization is a step. The number of generations is the number of steps divided by the pool size.

runs gave results of either 20 or 40 steps. The initial diversity increases slightly with pool size, as does the total number of structures found. The diversity after 200 steps increases with pool size, and higher diversities at this stage seem to be related to finding more minima after 500 steps. However, the search is nearing completion after 200 steps. The importance of crossover is demonstrated by the poor results of the runs that did not use it and relied entirely on mutation. After 500 steps, however, the no-crossover runs had caught up with the crossover runs, and found very slightly more minima, although the small difference is unlikely to be significant. The effect of reducing or removing mutation steps is less dramatic than removing crossover. A systematic search or Monte Carlo search would outperform the genetic algorithm approach for this small molecule, but the table demonstrates that the genetic algorithm can generate most low-energy conformations with a wide range of settings for its parameters. This tolerance to initial settings is important because there are no obvious criteria for choosing good parameters for a particular system except by trial and error.

The smallest unbranched alkane with a nonlinear global minimum structure is $C_{18}H_{38}$, according to MM2*.[4] The global minimum has an energy of 46.09 kJ mol^{-1}, and the extended form has an energy of 47.13 kJ mol^{-1}. Table 2 shows the results of a series of genetic algorithm calculations on this molecule. The default settings were used, except for those described in the table.

Table 2 results demonstrate that a wide range of settings lead to good results for this system. The only option that was very unsuccessful was *B*, in which the replacement temperature, *r*, was set very high, with the result that the algorithm was hardly biased toward low-energy structures for each new generation. Although all the other choices of options were successful at finding low-energy structures, small pool sizes and lower values for the duplication penalty, *d*, tended to reduce the total number of structures found. The mean and the median values for the lowest energies found are rather similar, and the standard deviations of the lowest energies for each group of eight runs are usually <1, showing that the runs did not vary widely in their results. The diversities of the final populations show some correlation with the total number of structures found.

Table 2. Results of Genetic Algorithm Calculations on $C_{18}H_{38}$[a]

		lowest energy structure			structures found		diversity (%)	
pool size	options[b]	mean	median	standard deviation	mean	standard deviation	initial	final
10	default	46.6	46.6	0.52	431	36	70.7	20.0
10	*A*	47.2	47.1	0.53	513	24	69.4	20.0
20	default	46.8	46.1	1.24	522	31	75.4	22.3
20	*A*	47.6	47.1	0.74	571	48	75.0	23.1
20	*B*	54.5	54.5	1.41	778	43	76.2	29.9
20	*C*	47.0	47.1	0.34	237	20	75.2	17.1
20	*D*	47.5	47.1	1.54	461	29	74.5	22.1
20	*E*	47.6	47.5	1.19	258	43	75.4	18.5
40	default	46.8	47.1	0.63	637	35	78.4	24.7
40	*A*	46.7	46.6	0.66	671	41	79.3	26.3

[a] Every entry is the average of eight separate runs, each of which had 2000 steps. [b] Default settings: $d = 10$ kJ mol^{-1}; $r = 1000$ K; $s = 10000$ K. Default settings were used in every case, except *A*: $d = 50$ kJ mol^{-1}; *B*: $r = 10\,000$ K; *C*: $r = 100$ K; *D*: $s = 1000$ K; *E*; $s = 100$ K.

Table 3. Results of Genetic Algorithm Calculations on $C_{39}H_{80}$[a]

		lowest energy structure			structures found		diversity (%)	
pool size	*d*	mean	median	standard deviation	mean	standard deviation	initial	final
10	10	80.9	79.45	6.2	445	72	68.24	14.95
10	50	80.0	81.76	7.4	472	40	66.84	17.79
20	10	79.0	79.37	3.0	529	27	73.50	16.25
20	50	75.9	74.08	6.0	596	27	72.64	20.00
40	50	75.8	75.63	3.3	672	27	76.27	25.00
80	50	81.5	80.84	3.2	796	15	78.96	39.70

[a] Every entry is the average of eight separate runs, each of which had 2000 steps.

A far more demanding test of a conformation search routine is $C_{39}H_{80}$, the results for which are given in Table 3. This molecule was chosen because it has the shortest chain for which the lowest energy structure we have found contains two different twists in the chain, according to the MM2* force field. It represents, therefore, a much more demanding conformation searching problem than chain lengths up to $C_{17}H_{36}$, which prefer to be linear, or alkanes with between 18 and 38 carbon atoms, which we believe prefer to have a single twist.[4] A Monte Carlo search with 10000 steps did not find a lower energy structure than the extended one, which has an energy of 102.55 kJ mol^{-1}. The genetic algorithm was only allowed to run for a fifth of this time, but every run found lower energy structures than this. A purely random search, generating and minimizing 6000 structures with randomly generated torsion angles, found a conformation with an energy of 100.6 kJ mol^{-1}.

Table 3 results show again that the genetic algorithm is not very sensitive to the settings. Increasing the size of the pool increases the number of structures that were found, but at the expense of finding fewer low-energy structures. A pool size of 10 or 20, with a small duplication penalty, gives a low diversity after 2000 minimizations, and the results are less good. Pool sizes of 20 or 40, with a duplication penalty of 50 kJ mol^{-1}, seem to give the best compromise between finding low-energy structures and finding many structures.

Alkanes are not usually the subjects of synthetic or biological studies. This genetic algorithm can, however, be applied to systems of more general interest. For example,

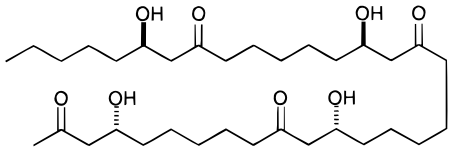

Figure 3. PM-toxin A.

PM-toxin A (Figure 3), a corn host-specific pathotoxin, has recently been synthesised.[11] Is it possible to perform a conformation analysis of this compound in a reasonable amount of time?

A 10000 step Monte Carlo search found two structures with an energy of <-70 kJ mol^{-1}, the lowest being -78.35 kJ mol^{-1}. An alternative strategy began by using the genetic algorithm to do a conformation search on the 33-carbon-atom backbone of the molecule (pool size of 20; 200 generations). This search generated 1011 conformations. The program *Acca*[12] was used to convert the alkyl chain into PM-toxin A. The hydroxyl groups were each included with two different orientations, 180° apart, so this generated 16 176 conformations that were ordered by the energy of the alkyl chain. We hoped, therefore, that the lower-energy conformations would be toward the beginning of this file. Minimization of the first 1000 structures from this file produced 48 structures with energies of <-70 kJ mol^{-1}. The lowest energy structure had an energy of <-100 kJ mol^{-1}. This structure is >20 kJ mol^{-1} lower in energy than the lowest energy structure found by the Monte Carlo search, which required twice the computer time to generate because twice as many structures were minimized. Continuing the reminimization of the converted structures, so that the total computer time was roughly equivalent to that for the Monte Carlo search (4000 minimizations for the genetic algorithm, 6000 for the reminimization of the converted structures), generated 168 structures with energies below -70 kJ mol^{-1}, but did not find another structure lower in energy than -100 kJ mol^{-1}. It is clear that the genetic algorithm followed by mutation strategy is much better than a Monte Carlo conformation search for this particular structure.

CONCLUSION

A genetic algorithm-based conformation searching program has been developed for linear alkanes. The program may be downloaded from the Cambridge Chemistry Department WWW server on URL: http://www.ch.cam.ac.uk/MMRG/software/. This program outperforms Monte Carlo conformation searches and appears not to be very sensitive to pool size and the other adjustable parameters. Combined with *Acca*,[12] the genetic algorithm was used to perform a conformation search on PM-toxin A. This method appears to represent a useful strategy for the conformation analysis of such systems.

ACKNOWLEDGMENT

The Royal Society, the EPSRC and the Cambridge Centre for Molecular Recognition are thanked for their support.

REFERENCES AND NOTES

(1) Saunders: M.; Houk, K. N.; Wu, Y. D.; Still, W. C.; Lipton, M.; Chang, G.; Guida, W. C. Conformations of cycloheptadecane − a comparison of methods for conformational searching. *J. Am. Chem. Soc.* **1990**, *112*, 1419−1427.

(2) Ngo, J. T.; Karplus, M. Pseudosystematic conformational search. Application to cycloheptadecane *J. Am. Chem. Soc.* **1997**, *119*, 5657−5667.

(3) Goodman, J. M.; Still, W. C. An unbounded systematic search of conformational space. *J. Comput. Chem.* **1991**, *12*, 1110−1117.

(4) Goodman, J. M. What is the longest unbranched alkane with a linear global minimum conformation? *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 876−878.

(5) Allinger, N. L. Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V_1 and V_2 torsional terms. *J. Am. Chem. Soc.* **1977**, *99*, 8127−8134.

(6) Mohamedi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. Macro-Model- An integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comp. Chem.* **1990**, *11*, 440−467.

(7) (a) McGarrah, D. B.; Judson, R. S. Analysis of the genetic algorithm method of molecular conformation determination. *J. Comput. Chem.* **1993**, 14, 1385−1395. (b) Herrmann, F.; Suhai, S. Energy minimisation of peptide analogues using genetic algorithms. *J. Comput. Chem.* **1995**, *16*, 1434−1444. (c) Meza, J. C.; Judson, R. S.; Faulkner, T. R.; Treasurywala, A. M. A comparison of a direct search method and a genetic algorithm for conformational searching. *J. Comput. Chem.* **1996**, *17*, 1142−1151.

(8) Brodmeier, T.; Pretsch, E. Application of genetic algorithms in molecular modeling. *J. Comp. Chem.* **1994**, *15*, 588−595.

(9) (a) Holland, J. H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*; University of Michigan: Ann Arbor, MI, 1975. (b) Davis, L. *Handbook of Genetic Algorithms*; Van Nostrand Reinhold: New York, 1991.

(10) Lamarck, J. B. *Philosophie Zoologique*; 1809.

(11) Hayakawa, H.; Ohmori, M.; Takamichi, K.; Matsuda, F.; Miyashita, M. Asymmetric total synthesis of PM-toxin A, a corn host-specific pathotoxin *Chem. Commun.* **1997**, 1219−1220.

(12) *Acca* is partially described in Goodman, J. M.; Leach, A. G. Rapid Conformation Searching. Part 2. Similar Compounds. *J. Chem. Soc., Perkin Trans. 2* **1997**, 1205−1208. It may be downloaded from the URL: http://www.ch.cam.ac.uk/SGTL/accadoc/.

CI970433U