RAPID COMMUNICATIONS

PHYSICAL REVIEW E VOLUME 51, NUMBER 4 APRIL 1995

Prediction of the lowest energy structure of clusters using a genetic algorithm

Yehuda Zeiri
Department of Physics, Nuclear Research Center–Negev, P.O. Box 9001, Beer-Sheva, Israel
(Received 31 October 1994)

An optimization approach using a genetic algorithm in the search for a global minimum is described. The method is based on the use of control variables to form the genotypes in each generation. This procedure allows an accurate representation of the control variables, and consequently, a high resolution determination of the optimum solution. A set of genetic operators, appropriate for the operation on genes represented by real numbers, is introduced. The method is used to predict the lowest energy structures of Ar_nH_2 microclusters with $n=4$, 5, 6, 7, and 12. Comparison between the performance of this optimization approach and the well established simulated annealing method clearly demonstrates the superiority of the genetic algorithm based search.

PACS number(s): 02.70.−c, 07.05.−t, 36.40.+d

The considerable recent interest in the physics and chemistry of atomic and molecular clusters motivated many theoretical studies devoted to the search for the lowest energy structure of such systems [1,2]. The interest in impure microclusters which include an infrared active molecule increased due to the development of a spectroscopic technique [3] which allows the determination of structural properties of such systems. Hence, recent theoretical investigations were aimed at obtaining detailed understanding of such systems. For example, molecular dynamics (MD) investigations were performed to study the structural, dynamical [4], and spectral [5] properties of Lennard-Jones clusters. The MD method was also utilized in the investigation of phase changes of Ni clusters [6] and the use of clusters in the catalysis of hydrogen chemisorption onto the Si(111) surface [7].

There is a large literature describing various approaches used for the energy minimization of molecular systems [8]. The main difficulty in such optimizations is the high dimensionality of the energy hypersurfaces governing the structure of the systems which gives rise to many local minima. Hence the main challenge of any search method is to converge to the global minimum without being trapped in a local minimum. In the following we describe a method, based on genetic algorithms, which allows an efficient search of the global minimum in multidimensional energy hypersurfaces of molecular systems.

Genetic algorithms (GA's) are global optimization methods based on several metaphors from biological evolution. The name is derived form the ability of the algorithm to simulate selection in an evolving population of living creatures attempting to adapt to their environment. Genetic algorithms have been applied successfully in a wide variety of fields [9–11]. The conventional GA's differ from traditional optimization methods in four important respects. (a) They employ an encoding of the control variables (usually as binary bit strings termed "genes") rather than the variables themselves. (b) GA's search from one population of solutions to another population, rather than from individual to individual. (c) The GA uses only objective function information, not derivatives. (d) GA's use probabilistic, not deterministic, transition rules.

In the following, a brief outline of an approach based on GA's will be described. A detailed description of the method will be given elsewhere [12].

The main deviation of the present approach form the conventional GA's is in the use of real numbers to construct the individuals in a generation. Since we shall be dealing with optimization of molecular structures, the elements of a gene correspond to the coordinates of the various atoms in the system, the control variables. The main advantage of this representation is its capability to deal with problems in which the control variables are continuous. Hence the encoding and decoding of the genes can be avoided and the optimum solution can be found with any required accuracy.

The first generation, containing $N_{\text{pop}}=100$ individuals, is formed by randomly placing the Ar atoms and the H_2 molecule (fixed at its equilibrium distance) in a "box" whose dimensions are large compared with the expected size of the cluster. Once the initial generation is formed, the potential energy E_i associated with each structure (gene) is calculated. The fitness of the ith gene in the pth generation, f_i^p, is calculated by scaling the E_i's in the range span by the best and worst genes in the population and normalizing the scaled value by the value of the best solution. These values serve as the probability according to which a gene is selected to be used as a parent in the preparation of the next generation.

The following procedure was used to form the next generation. First, the highest fitness k_{best} genes ($k_{\text{best}}/N_{\text{pop}}\simeq 0.1$) are transferred to the new generation. The second step is to form k_{rand} ($k_{\text{rand}}/n_{\text{pop}}\simeq 0.05$) random genes which are included in the new generation to allow the flow of new structures into the simulation. The remaining number of genes needed to complete the new generation are constructed by the application of six GA operators to individuals in the present generation. The following operators were used. (1) n-mutation (O_{mute}): one of the k_{best} genes is chosen randomly and some of its elements (randomly chosen) were modified by the addition of a random number evaluated from a normal Gaussian distribution. (2) Inversion (O_{inv}): the elements in a segment, randomly chosen, of the parent gene were inverted to generate a son. (3) Two-point cross link (O_{cros2}): two parent genes were connected head-to-tail to from a cyclic object which was then disconnected

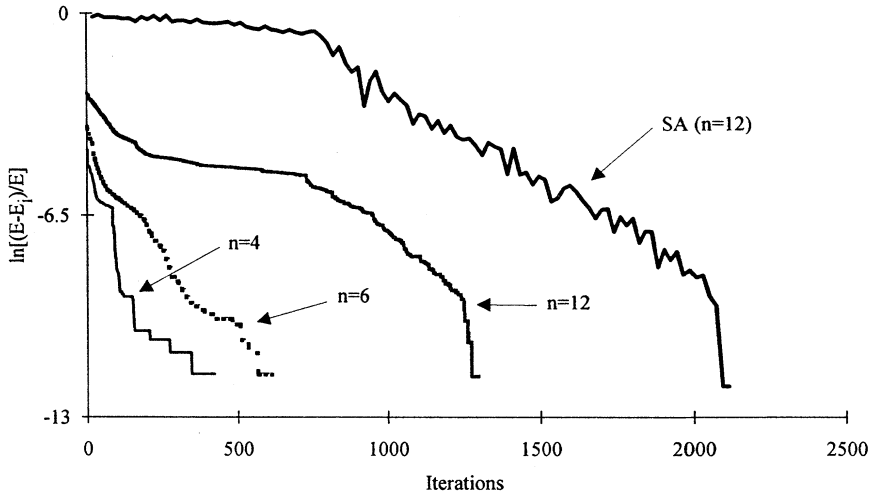

FIG. 1. Convergence rate of the GA method as obtained in typical calculations of the lowest energy structures of three Ar_nH_2 ($n=4$, 6, and 12) clusters. Also shown is the convergence rate as obtained in an SA calculation for $Ar_{12}H_2$, marked as "SA ($n=12$)" (note that the iteration number in this case was divided by 1000).

randomly to form two new genes. (4) n-point cross link ($O_{\text{cros}n}$): the elements of two parent genes were copied to form two "sons." The rearrangement of the parent's elements was accomplished by the following steps: a random number ζ was chosen from a uniform distribution in the range 0−1. If $\zeta \geq 0.5$ the element of parent 1 was copied to son 1 and the corresponding element from parent 2 to son 2, while, if $\zeta < 0.5$, the element of parent 2 was copied to son 1 and the corresponding element from parent 1 to son 2. (5) Arithmetic average (O_{av}): the arithmetic average of the elements of two parent genes were used to form a son. (6) Geometric average (O_{geom}): the geometric average of the elements of two parent genes were used to form a son.

These operators were applied to genes in the pth generation to form sons which were transferred to the $p+1$ generation. In all cases, the probability to choose the ith gene as a parent was proportional to its fitness, f_i^p.

Once the new generation was completed, the energies and corresponding fitness values of the individuals were computed. A uniform initial probability for the application of each one of these operators, P_{O_α}, was used in the first N_p iterations. These probabilities were modified every $N_p=20$ generations according to the number of times each operation was successful in generating a son whose fitness was better than $f_{k_{\text{best}}}^p$. To ensure that the search did not converge to a local minimum a reshuffling procedure was introduced. Namely, if the best gene in the population did not alter for $N_s=8$ generations, the k_{best} genes of generation p were copied to generation $p+1$ in addition to $(N_{\text{pop}}-k_{\text{best}})/2$ randomly created genes and $(N_{\text{pop}}-k_{\text{best}})/2$ genes which were generated by the application of O_{mute} to the k_{best} genes of generation p. To complete the description of the method we have to specify the conditions which were used to terminate a calculation. Two termination conditions were used: first, if the number of iterations exceeded 5000 the calculation was terminated. This condition did not occur in any of the calculations reported below. The second condition is related to the variation of the highest fitness structure in the population. If this best gene did not alter for $N_{\text{iter}}=50$ generations it was assumed the convergence was reached.

To ensure that the global minimum was found, for each microcluster $N_{\text{run}}=5$ different simulations were performed. At the end of each simulation the highest fitness k_{best} genes were stored. Finally, an additional calculation was performed where the initial population contained the $N_{\text{run}}k_{\text{best}}$ highest fitness genes obtained together with $(N_{\text{pop}}-N_{\text{run}}k_{\text{best}})$ random structures. This additional simulation resulted usually in a 0–10 % improvement in the lowest energy structure.

The optimization method described above was applied to the search for the lowest energy structures of five Ar_nH_2 ($n=4$, 5, 6, 7, and 12) clusters using the interaction potentials of Ref. [7]. The results of the present approach will be compared with those obtained by the simulated annealing (SA) approach [7].

The convergence rates of our GA based method to the lowest energy structure in some typical calculation for three cluster sizes is shown in Fig. 1 (the curves marked according to the cluster size as $n=4$, $n=6$, and $n=12$). The measure of convergence rate is defined as the natural logarithm of the normalized energy difference between the lowest energy in generation i and the global minimum found in the search. In all the GA based calculations the initial convergence is exponential. During the later stages the convergence rate changes in steps. This stepwise change is associated with the formation of a subgroup of "good" genes which were then improved by the action of the GA operators. It should be noted that this subgroup of good genes is usually formed shortly after a reshuffling process. These results clearly indicate that increase in the size of the cluster results in a corresponding increase in the number of generations needed to reach the lowest energy structure. However, it is clear that the number of generations needed to converge depends on the accuracy requirement in the search. In the present case the required accuracy was set to 1×10^{-4} kcal/mole.

Let us turn now to a comparison of the results obtained using the GA approach with those of the SA [7]. The energies associated with the most stable geometry of the various microclusters investigated are summarized in Table I. Inspection of these results shows that the GA method yields results which agree extremely well with those obtained using the SA approach. Moreover, for all the microclusters considered, the energy of the most stable geometry calculated by

TABLE I. Energies of the most stable Ar_nH_2 clusters as calculated using the GA method described here and by the SA approach [7]. All energies are in units of eV. The values in parentheses represent energies of geometries other than the most stable.

n	SA	GA
4	−0.09533	−0.09623
5	−0.13352	−0.13495
6	−0.17861	−0.18253 (−0.17879)
7	−0.21678	−0.21855
12	−0.50085	−0.49946

GA is slightly lower than the corresponding value obtained by SA (except for $n=12$). These negligible differences are due to a slight variation in the orientation of the H_2 relative to the structure of the Ar atoms. We have plotted the structure of the various Ar_nH_2 clusters and found that they agree very well with those presented in Ref. [7]. The only case where a slightly larger difference between the results of the two methods was detected is for Ar_6H_2. For this cluster an additional calculation was performed. In this additional calculation we discarded in each generation all genes whose potential energy was lower than that obtained by the SA approach. The energy associated with the most stable geometry obtained in this calculation is added in parentheses in Table I. Figure 2 describes the most stable structures of Ar_6H_2 as obtained in the energy unrestricted [Fig. 2(a)] and the energy restricted [Fig. 2(b)] calculations. Comparison of Fig. 2(b) with Fig. 1(c) in Ref. [7] indicates that these two geometries are practically identical. However, the geometry in Fig. 2(a) is a slightly different one.

To assess the potential value of the GA based method as an efficient optimization approach, the rate at which it converges towards the global minimum should be compared to that of the widely used SA method. Unfortunately, it is not possible to base such a comparison on the results of Ref. [7] since no details of the SA calculations were given. Hence we recalculated the lowest energy structures of these microclusters using the SA approach [13]. In these SA calculations a geometric cooling schedule [13] (i.e., $T_{K+1}=0.9T_K$) was employed where the magnitude of the initial temperature was chosen to yield an average increase acceptance probability of

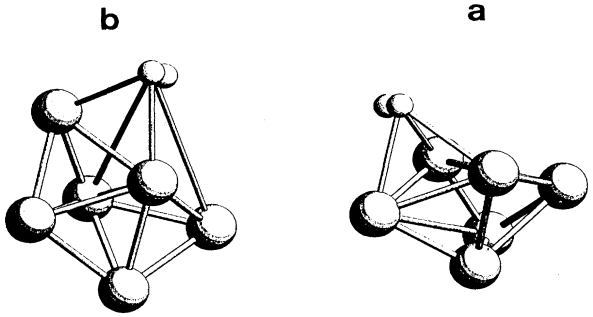

FIG. 2. Two geometries for the Ar_6H_2 cluster as obtain in energy unrestricted (a) and energy restricted (b) calculations. The large balls represent Ar atoms and the small balls H atoms.

TABLE II. Average number of cost function evaluations $\langle N_{\text{func}}\rangle$, for each cluster using simulated annealing (SA) and the genetic algorithm based (GA) optimization approach.

	$\langle N_{\text{func}}\rangle$	
n	SA (units of 10^{-3})	GA (units of 10^{-3})
4	1832	42
5	1975	45
6	1621	57
7	2080	100
12	2137	225

80%. To improve performance, cluster structures which included an Ar-Ar or Ar-H distance smaller than 3.024 Å were rejected without further examination. As in the GA simulations, for each cluster five independent calculations were performed. The energies obtained for the various clusters by these SA calculations were identical (to at least three significant digits) to those obtained by the GA approach.

To assess the relative efficiency of the two optimization methods, we shall compare the average number of cost function (cluster potential energy) evaluations required to reach convergence, $\langle N_{\text{func}}\rangle$. Here, $\langle\ \rangle$ denotes the average over the five independent runs performed by each method. The results are summarized in Table II. These results clearly demonstrate the superior overall convergence of the GA based approach as compared to the SA method. We turn now to a comparison between the convergence rates obtained by the GA and the SA approaches. The curve marked as "SA ($n=12$)" in Fig. 1 represents the convergence rate of a typical calculation for $Ar_{12}H_2$ using the SA approach. To fit this curve into the range of iterations used in Fig. 1, the number of iterations in this case was scaled by a factor of 1000. Comparison between the convergence rates of the GA and SA results for $Ar_{12}H_2$ shows a number of differences. However, before the discussion of these differences, the meaning of the convergence rates shown in Fig. 1 should be examined. The computational effort involved in an iteration of each approach is determined by the number of cost function evaluations. In the case of the GA based method, each iteration involves N_{pop} evaluations of the cluster energy, while in the SA approach each iteration involves a single evaluation of the energy function. Hence a meaningful comparison of the convergence rates of the two methods should examine the variation of the E_i as a function of the number of energy evaluations. Since we used $N_{\text{pop}}=100$ in the GA calculations, each iteration in Fig. 1 corresponds to 100 evaluations of the cluster energy. Thus in terms of cost function evaluations, the scaling factor of the SA results shown in Fig. 1 is 10 (and not 1000, see above) as compared to the GA results.

The SA results (Fig. 1) show practically two stages in the convergence rate, a very slow initial convergence followed by a much faster one. On the other hand, the GA results exhibit a larger number of steplike distinct convergence stages, where in all cases the first stage is the fastest. Moreover, the convergence rates of all the distinct stages in the GA calculation are more rapid than those of the SA calculation (note that the SA results are scaled by a factor of 10).

These differences, together with the results shown in Table II, demonstrate the superiority of the GA based method over the SA one. A more elaborate comparison between these two optimization methods can be found in Ref. [12].

To summarize, an optimization method based on a genetic algorithm in which the control variables are used to construct the genes has been developed. The main advantage of the approach is its capability to deal with problems in which the control variables are continuous. As a result, the encoding and decoding of the decision variables can be avoided. Hence the search for the optimum solution using this method is much more accurate than in conventional GA procedures. The method was applied to the search for the lowest energy structures of Ar_nH_2 clusters. The optimized cluster structures obtained compare extremely well with the results obtained using the simulated annealing approach. Comparison between the convergence rates of these two methods clearly indicates that the GA based approach is much more efficient.

It is believed that the high efficiency and accuracy of the GA based method presented here makes it a very useful approach for structural studies of a large variety of atomic and molecular systems [12].

[1] M. R. Hoare and P. Pal, Adv. Phys. **20**, 161 (1971).

[2] See, for example, *Elemental and Molecular Clusters*, edited by G. Benedek, T. P. Martin, and G. Paccioni, Springer Series in Materials Science Vol. 6 (Springer, Berlin, 1988); *Proceedings of the International Symposium on the Physics and Chemistry of Small Clusters*, edited by P. Jena, B. K. Rao, and S. N. Khana, *NATO Advanced Study Institute Series* (Plenum, New York, 1987), p. 193.

[3] T. E. Gough, D. G. Knight, and G. Scoles, Chem. Phys. Lett. **97**, 155 (1983); T. E. Gough, M. Mengel, P. A. Rowntree, and G. Scoles, J. Chem. Phys. **83**, 4958 (1985).

[4] I. L. Garzon, X. P. Long, R. Kawai, and J. H. Weare, Chem. Phys. Lett. **158**, 525 (1989); I. L. Garzon, X. P. Long, R. Kawai, and J. H. Weare, Z. Phys. D **12**, 81 (1989).

[5] L. Perera and F. G. Amar, J. Chem. Phys. **93**, 4884 (1990).

[6] Z. B. Guvenc, J. Jellinek, and A. F. Voter, in *Physics and Chemistry of Finite Systems: From Cluster to Crystals*, Vol. 374 of *NATO Advanced Study Institute, Series C: Mathematical and Physical Sciences*, edited by P. Jena *et al.* (Kluwer Academic, Dordrecht, 1992), Vol. I, p. 411.

[7] J. N. Beauregard and H. R. Mayne, Surf. Sci. Lett. **280**, L253 (1993).

[8] See, for example, W. J. Hehre, L. Random, P. von R. Schleyer, and J. A. Pople, *Ab Initio Molecular Orbital Theory* (Wiley, New York, 1986).

[9] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, Reading, MA, 1989).

[10] L. Davis, *Handbook of Genetic Algorithms* (Van Nostrand Reinhold, New York, 1991).

[11] *Proceedings of the 3rd International Conference on Genetic Algorithms*, edited by J. D. Schaffer (Morgan Kaufman, San Mateo, 1989).

[12] Y. Zeiri (unpublished).

[13] See, for example, P. J. M. van Laarkoven and E. H. L. Aarts, *Simulated Annealing: Theory and Applications* (Reidel, Dordrecht, 1987); G. T. Parks, in *Advances in Nuclear Science and Technology*, edited by J. Lewins and M. Becker (Plenum, New York, 1990), Vol. 21, p. 195.