Quality Criteria of Genetic Algorithms for Structure Optimization

Ron Wehrens,*[,†] Ernö Pretsch,[‡] and Lutgarde M. C. Buydens[†]

Laboratory of Analytical Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen,
The Netherlands, and Laboratory of Organic Chemistry, Swiss Federal Institute of Technology (ETH),
CH-8092 Zürich, Switzerland

Received August 18, 1997

A set of quality criteria is proposed to evaluate the performance of genetic algorithms for optimization. Instead of concentrating solely on the best solutions proposed by the algorithm, the quality criteria also consider the repeatability of the optimization and the coverage of the search space. They are tested by using various parameter settings of a genetic algorithm providing starting structures for molecular mechanic calculations of organic molecules. These or similar criteria can also be used for other domains of optimization with evolutionary algorithms.

INTRODUCTION

Although stochastic optimization techniques such as evolutionary methods[1] and genetic algorithms (GAs)[2−4] can never guarantee that the best solution is found, in general they are the optimization methods of choice for complex search spaces of high dimension with multiple optima. Their strength of being robust in such situations is coupled with a weakness of being inefficient for fine optimizations. Therefore, ideally hybrid methods are used in which the stochastic procedure provides starting points for a local optimization, such as gradient-descent or Newton−Raphson techniques. In the case of molecular modeling, this means that the stochastic method generates starting structures as input for standard molecular mechanics or molecular dynamics calculations. The design of a GA for a given problem is a complex task because numerous possibilities are available concerning the form of the fitness function, the selection method, the kind of crossover, and the use of sharing operators and generation gaps. In addition, for the chosen procedures, various parameters have to be set, sometimes critically determining the success rate of the method. For this tailoring of the GA to a specific task, objective criteria are needed. Unfortunately, no such measures are available at present. In most cases, solely the fitness of the best individual is monitored over the generations. Clearly, this measure is insensitive and does not reflect several important characteristics of a given GA setup, such as its robustness or its coverage of the search space. In this paper, four criteria are proposed to assess different quality aspects of GAs. They have been designed and tested in with a GA for molecular modeling. The goal of the GA is here to find promising (i.e., low energy) conformers as starting structures for molecular mechanics (MM) structure optimization.[5−7] The same or analogous criteria are also applicable for other problem domains. They can be used to find optimal GA settings and to compare the performance of different optimization methods for a given problem. In essence, the criteria reflect two different characteristics of a stochastic optimization. Two of them describe the coverage of the total search space and the relevant part of it, respectively, and thus are related to the chance to find the global optimum. The other two criteria supply a measure of robustness. Though replicate runs behave differently because of the stochastic nature of the algorithm, they should ideally yield the same final results. To test this characteristic, the ability of the GA to find the same relevant solutions in replicate runs is investigated. Another criterion is related to the reproducibility of not only the final results but of the whole GA run, using principal component analysis (PCA).[8]

The proposed quality criteria can be used to assess the quality of an optimization method and, therefore, provide a means to improve it. Moreover, various optimization strategies can be compared with each other on a more sound basis. The criteria are illustrated using experiments with different settings of a GA for structure optimization, but the principles are generally applicable and can be adapted to other problem domains.

METHODOLOGY

The investigations are made with a hybrid algorithm where the GA provides starting structures for a MM optimization. Ideally, not only the global minimum should be found but all possible states that are accessible at room temperature. The structure that is optimized by the GA is represented in torsion-angle space. For the calculation of the energy, standard values for bond lengths and bond angles are used. The optimization of the bond lengths and angles, as well as the corresponding fine-tuning of the torsion angles, is the task of a subsequent MM2 run.[9] Because the MM optimization may substantially lower the energies, the deepest minimum will not necessarily be obtained with the best individual of the final GA population. Therefore, in such an application the goal of the GA is not to obtain the one best individual but to produce as many of the potentially relevant ones as possible.

Torsion angles are coded as real numbers, and the fitness function is calculated from the MM2 energies as described

* Author to whom correspondence should be addressed.
[†] Laboratory of Analytical Chemistry.
[‡] Laboratory of Organic Chemistry.

S0095-2338(97)00430-7 CCC: $15.00
Published on Web 02/05/1998

Table 1. Settings of the Experiments[a]

parameter	I	II	III	IV	V
mutation rate	0.02	0.02	0.00	0.02	1.00
crossover rate	0.80	0.80	0.80	0.80	0.00
sharing distance	0	10	10	10	0
sharing offset	0	30	180	90	0
tournament size	2	2	2	10	1

[a] Fixed settings: tournament selection, uniform crossover, population size 50, number of generations 100. The two best strings are always copied unchanged to the next generation (elitism).

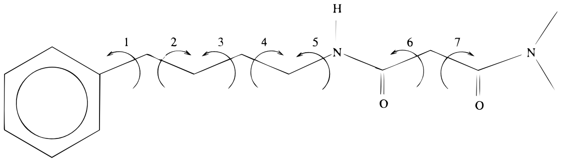

Figure 1. The test molecule: *N*,*N*-dimethyl-*N*′-4-phenylbutylmalonamide. Torsion angles that are optimized are numbered.

previously.[7] Instead of the earlier used roulette wheel method, parents to generate individuals in the next generation are chosen by tournament selection, which takes the best individual of a randomly chosen small subset (the size of which is an input parameter, see Table 1). Uniform crossover was applied with a probability of 0.8; thus, two parent strings have an 80% chance of exchanging the values of randomly selected parameters and a 20% chance of being selected without changes.

The mutation operator was applied at different rates (see Table 1). If it was selected, it replaced the corresponding torsion angle by a randomly chosen value between 0 and 360°. As described previously,[7] for this type of application, the sharing operator is essential to ensure a sufficient spread of the population. In essence, it is an additional mutation if a newly generated individual is too similar to an already existing one of the same generation. The similarity is measured by root mean square difference (RMS) of the torsion angles. If it is smaller than a threshold, a random value between 0 and a predefined maximum is added to a randomly selected torsion angle. The threshold is called the sharing distance, and the maximum size of the mutation is called the sharing offset (cf. Table 1). All test runs were made with *N*,*N*-dimethyl-*N*′-4-phenylbutyl malonamide,[10] depicted in Figure 1. Seven torsion angles are optimized; those around the amide bonds are kept fixed in the planar conformation. The size of the population and the number of generations are 50 and 100, respectively, in all experiments.

Five different parameter combinations, shown in Table 1, are used to illustrate the quality criteria. With a high crossover rate and a low mutation rate, Experiment I uses typical GA parameter settings.[4] In Experiment II the sharing operator is introduced. This operator completely replaces the mutation in Experiment III (i.e., a random change is only applied when necessary to improve the diversity of the population). Here, the maximum size of the forced mutation, the sharing distance, is also increased. In Experiment IV, the tournament size is set to 10, which creates a much higher selection pressure because in each case only the best out of 10 randomly selected individuals contributes to the next generation. To prevent premature convergence, a combination of mutation and a relatively large sharing effect is used, thereby increasing the random component of the search in this experiment. Experiment V serves as a reference and corresponds to a purely random search because there is no selection pressure (tournament size of 1, i.e., random selection), no crossover, and a 100% chance of mutation.

Each experiment consists of five replicate runs using different random seeds and thus different random numbers, both to initiate the population and for the various GA operators. Each of the experiments was repeated three times (denoted by letters A, B, and C) to investigate the reproducibility of the quality criteria. In total, 75 GA runs were performed (five replicate runs per setting, three repeated experiments, five settings), each consisting of 5050 energy evaluations (50 individuals for 101 generations including the 0th starting population). In two of the quality criteria described later, a greedy clustering method is employed on the pooled last population of the five replicate runs of one experiment. The method starts with the best structure and places it in a cluster with all others that are within a certain distance. It then proceeds with the best unclustered structure, and so on, until a predefined energy threshold has been reached. This procedure has the advantage that the clustering concentrates on the low-energy solutions, and yields more meaningful clusters than classical cluster algorithms such as complete linkage or single linkage clustering.[11] The GA software described in reference 7 was used with a few small changes. During the GA runs, all populations are written to files, including the energies of the trial structures. These files are processed by the statistical software package R.[12] The GA runs have been performed on a Silicon Graphics Indy, and the calculation of the quality criteria on a SUN Ultra machine.

RESULTS AND DISCUSSION

For each of the five experimental settings, the mean energy of the best individuals in five replicate runs is plotted as a function of the number of generations in Figure 2. This kind of plot has often been used as quality criterion. On the basis of these plots, Experiment IV seems the most robust one because in every run a low energy individual is found after ~40 generations. Although there is a more diverse behavior at the beginning, Experiment II performs similarly after 40 generations. Experiment III finds solutions of a similar quality, albeit after more generations. Decreasing reproducibility is observed in Experiment I, and the most erratic behavior is found with the pure random search (Experiment V) in which the lowest energies are substantially higher than in the other experiments.

The drawback of such plots is that they only provide information about the fittest individual, or, as in this case, the mean of a set of fittest individuals. No information is available about the rest of population. However, such information would be essential, especially in hybrid applications where the GA only generates starting points for further optimization and where the final global optimum does not necessarily result from the best individual found by the GA. A further disadvantage is that the graphs shown in Figure 2 do not provide any means to judge the coverage of the search space.

Coverage Criteria. *Coverage of the Search Space.* The most important requirement of an optimization is that the

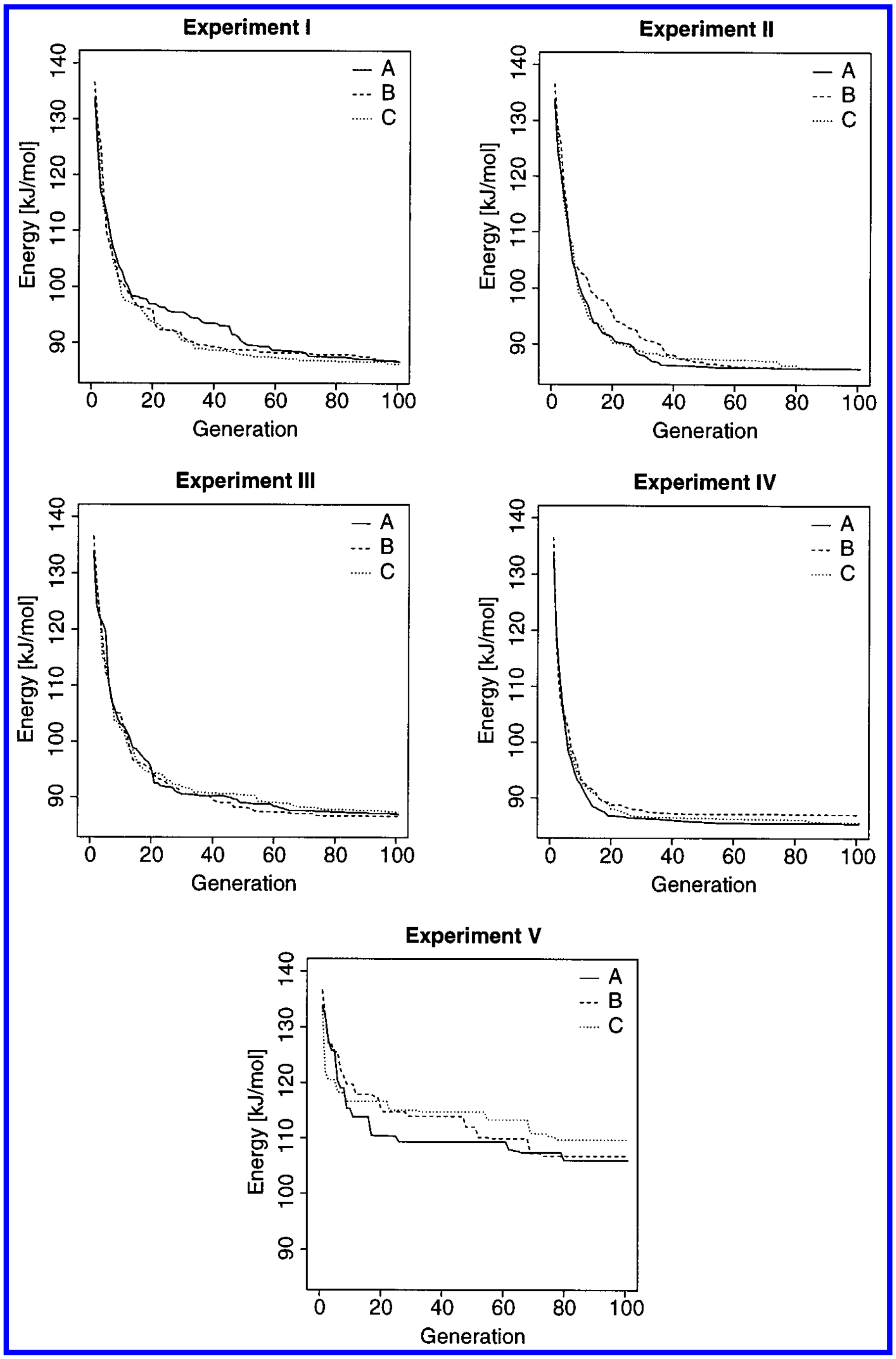

Figure 2. Energies of the best structures during the optimization [kJ/mol]. Plotted are the mean values of five replicate runs for each repeated experiment A, B, and C.

chance to be trapped in a local minimum must be as small as possible. For obvious reasons, information about this cannot be directly assessed by the method itself so that only indirect hints can be provided. These hints are based on the notion that the larger the portion of the search space that is investigated, the more complete the search will be and the lower the probability that the global optimum is missed. A simple measure of coverage has been obtained by dividing the search space into hypercubes of equal sizes and counting the ones that are visited during a run relative to the total number of hypercubes. A similarly defined conformational coverage has been applied for pharmacophore screening and 3D database search.[13]

If the range of each of p parameters is divided into n parts, a total of $k = n^p$ hypercubes are produced. The coverage measure l/k, where l is the number of hypercubes visited, is investigated as a function of the size of the hypercubes in Figure 3. If the total number of energy evaluations is $>k$, then in principle a coverage of 100% is possible. For hypercubes of with edges of 180° (i.e., the search space for each parameter is divided in two parts), a 100% coverage is indeed achieved by all experiments so that no informative measure can be derived. For hypercubes of 120° and smaller, essentially the same relative coverage measures are obtained, showing that this criterion is robust. Because the number of hypercubes increases drastically with a smaller size, a hypercube size of 90° is used for quantitative comparisons (Table 2). The maximal coverage still is 100%, but it is not achieved in any of the experiments. The variation for repeated experiments A−C is small (<10%), a necessary

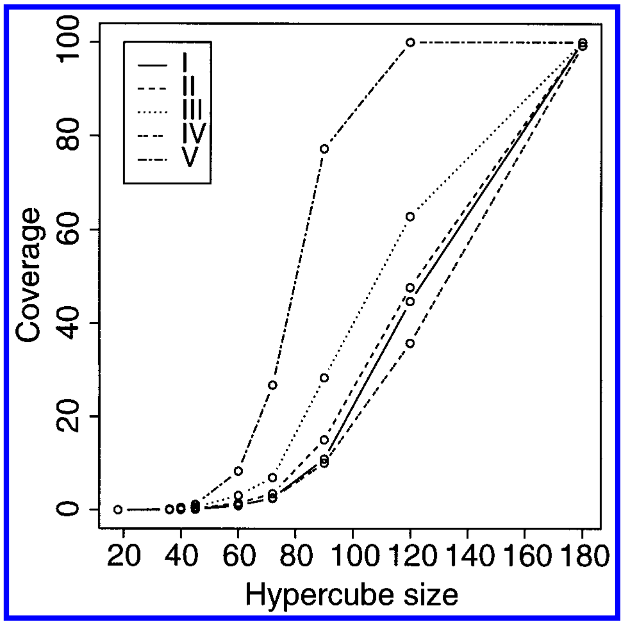

Figure 3. Coverage of the search space, depending on hypercube size. Points are mean values of the A−C experiments.

Table 2. Coverage of Search Space[a]

run	I	II	III	IV	V
A	10.6	13.7	29.1	10.6	77.0
B	11.8	15.5	27.5	9.7	77.2
C	10.3	16.1	28.2	9.7	77.4

[a] Percentage of hypercubes with edges of 90° that is actually visited; maximal coverage, 100%.

requirement for a quality measure. The coverage of the purely random search (Experiment V) is clearly the largest. The coverage is the second largest in Experiment III, in which the mutation operator was switched off but a higher sharing offset was introduced instead. Similarly, comparison of Experiments I and II shows the significant increase of the coverage because of the sharing operator. On the other hand, despite an increased sharing offset, the strong selection pressure in Experiment IV reduces the coverage, even relative to Experiment I. The increased selection pressure clearly biases the optimization toward a more local search. Interestingly, the experiment with the lowest coverage (IV) seemed to have the best performance according to the fitness development of the best individual (Figure 2).

Coverage of the Relevant Search Space: Clustering. According to the previous criterion, the random search (Experiment V) has the best coverage. However, it does not consider the strength of evolutionary optimization methods of being able to concentrate on relevant parts of the search space. An ideal GA should provide as many diverse and relevant final solutions as possible. Those that are too similar to each other would lead to the same structure after fine optimization and are, therefore, as useless as the ones having too high energies so that they would not lead to meaningful structures. Therefore, a measure of the coverage of the relevant search space is calculated by grouping the similar solutions of the final population in clusters and counting the number of low-energy clusters.

For the greedy clustering method described earlier, two parameters must be chosen: the radius of the cluster (i.e., the maximum distance of a member to the center of the cluster) and the energy threshold. Based on the assumption that torsion angle differences of <30° would not lead to an energy barrier, the radius, defined as the maximum absolute deviation in torsion angles, was set to 30°. To estimate the appropriate energy threshold, the relationship between the energies before and after the MM energy minimization was investigated for a large number of solutions (Figure 4). The results show a high degree of correlation. Although the MM2 minimization further lowers the energies by ~120 kJ/mol, the relative energies before and after are roughly equal. Based on these results, the energy threshold for the clustering algorithm was set to 100 kJ/mol, which is ~15 kJ/mol above the energy of the best solution.

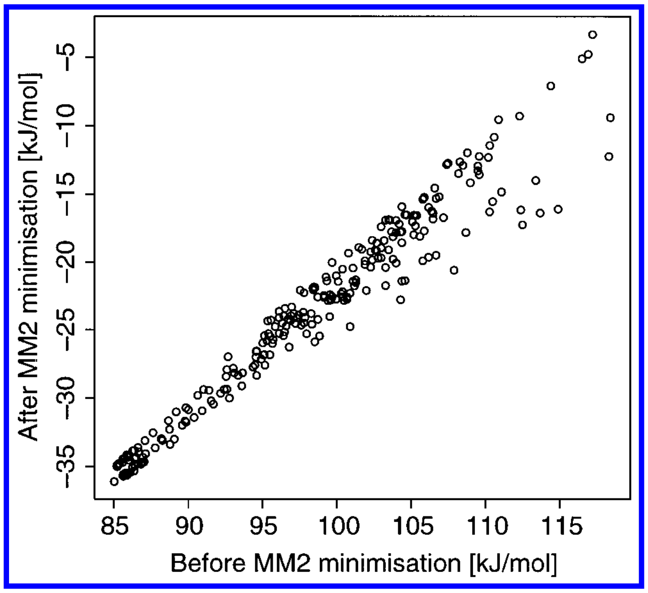

Figure 4. Energies of cluster centers before and after MM2 energy minimization [kJ/mol]. Minimization continued until the energy gradient was <0.01 kJ/mol.

In Figure 5, the size of the clusters obtained with these settings for the pooled final generation of replicate runs is indicated, together with the corresponding energy of the best structure in the cluster. The random Experiment V does not yield any meaningful clusters, only individual solutions are found. Additionally, it can be seen that the energies of these solutions are much higher than those of the other experiments. As Figure 4 shows, none of the solutions of Experiment 5 can give a useful result after MM2 relaxation. Extreme results of another kind are produced by Experiment I. Here, a series of large clusters of 40 individuals or more are found, but only very few small clusters. The size of low-energy clusters decreases in the sequence II, IV, and III.

To improve the reproducibility of this criterion, only clusters with at least three members are considered; this procedure reduces the effect of one individual that is found by chance in one of the last populations. Furthermore, only clusters whose centers have an energy of <100 kJ/mol are considered. The quality criterion then is formed by the number of these clusters. The results displayed for the repeated experiments A−C in Table 3 show that the robustness of the criterion is reasonably good. Besides Experiment V, the quality of Experiment I is also lower than that of the others. Compared with the within-run variation, the differences between the three Experiments II−IV are not

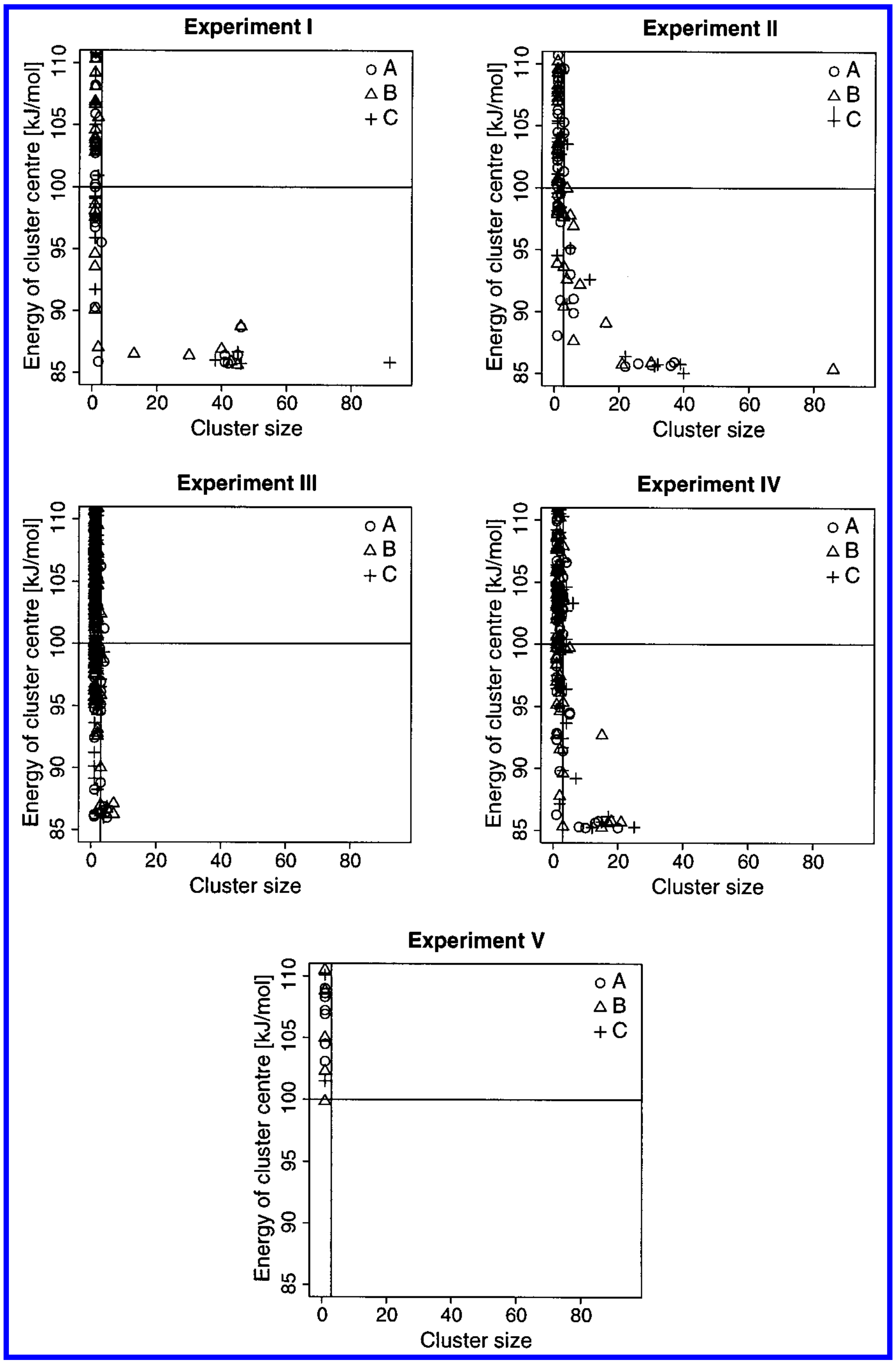

Figure 5. The size of the clusters plotted against the energy of the cluster center (the best structure in the cluster).

Table 3. Number of Clusters Found by Greedy Clustering[a]

run	I	II	III	IV	V
A	6	9	8	8	0
B	6	14	9	10	0
C	4	9	7	11	0

[a] Minimal cluster size, 3; energy cutoff, 100 kJ/mol.

Table 4. Number of Clusters Occurring in More than One Replicate Run[a]

run	I	II	III	IV	V
A	1	3	6	3	0
B	0	6	8	3	0
C	1	2	5	3	0

[a] In all cases the clusters are touched in two or three runs.

big. A sequence based on the product of the mean values of Tables 2 and 3 is: III > II > IV > I > V.

Reproducibility Criteria. The performance of the optimization should be independent of the (usually randomly chosen) starting structures and of the random operators. A robust method should, on repeated application, yield individuals that lead to the same final structures after local optimization; in the present context, after a MM run. In the following, two criteria are proposed to test the robustness of a GA.

Cluster Analysis of Pooled Runs. The same clusters discussed in the previous section can be analyzed in view of the reproducibility by investigating how many of them contain individuals from different runs. In Table 4, the number of clusters is shown containing individuals from at

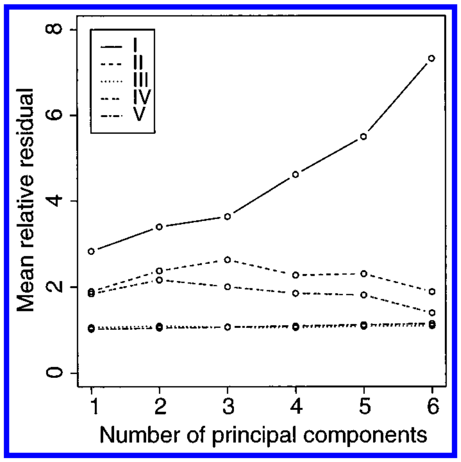

Figure 6. Reproducibility of the search, as measured by the projection on PCA space. Points are mean values of the A−C experiments and depict the ratio of the mean interrun and intrarun residuals after projection.

least two of the five replicate runs. The most striking result is that the large clusters found in Experiment I almost exclusively consist of strings from one single run. In the experiments IA and IC, just one cluster contained members of two replicate runs. Because of the small random component (absence of sharing and low mutation rate), each replicate run is trapped in another local optimum. The result is clearly 0 for Experiment V because no clusters of more than one individual were formed. Experiments II and IV perform quite similarly, with II being of somewhat lower reproducibility. Despite their small size, the clusters formed in Experiment III often consist of strings from different replicate runs, so that the reproducibility of this experiment is the best of all. However, the ideal situation in which each cluster contains individuals from each run was not achieved in any of the experiments. In all cases, the clusters only accommodated individuals from two or three of the five replicate runs. The results for the repetition A, B, and C are quite close to each other, so that the criterion is sufficiently robust.

PCA Residuals. Another measure for the reproducibility of the space coverage applies principal component analysis (PCA). Each individual can be seen as a point in the n -dimensional space, n being the number of parameters (i.e., torsion angles). If the dimensionality of the space is reduced to $k < n$ principal components, the essential data structure is kept because only the least significant $n - k$ dimensions are eliminated. For each replicate run, a PCA basis is calculated and the other $(l - 1)$ runs are projected into that basis. The degree of reproducibility is quantified by the residuals after projection. If the ratio of the mean residual after projection of the other runs on the basis of run i and the residual of projecting run i onto its own basis approximates 1, then the runs cover the same part of the search space. The larger this ratio is, the more different the spaces covered by the runs. This calculation is performed for $k =$ 1...6 principal components and averaged over the PC bases of all runs. To reduce the computational demands, instead of the whole population of 5050 individuals of a run, only every 10th generation was used. This reduction did not significantly influence the results.

The corresponding relative residual plots are shown in Figure 6. Irrespective of the number of principal components, the relative residuals are close to 1 for Experiments III and V; that is, the residuals do not increase if the results of a run are displayed in the space spanned by the principal components of another run. This result shows that because of the high random component, the replicate runs have similar space coverage. Experiment I, in which no sharing operator is used, has the highest relative residuals. Again this result shows that every replicate run goes in a different direction and thus covers another part of the space. From the two intermediate Experiments, II and IV, the values for the latter show a somewhat better reproducibility (see Table 5).

Table 5. Reproducibility Using PCA: Ratio of Between-Run and Within-Run residuals[a]

run	I	II	III	IV	V
A	3.3	2.3	1.1	1.6	1.1
B	3.3	2.7	1.0	2.2	1.1
C	4.2	2.9	1.1	2.2	1.1

[a] Three principal components are used.

Evaluation of the Experiments. Experiments with a high random component, such as the random search of Experiment V, have good coverage and reproducibility properties. However, their coverage of the relevant parts of the search space is bad (V) so that random searches are likely to fail when search spaces are vast. Experiment I, although it represented more or less standard GA settings as advocated in the literature, was the worst of all experiments with the exception of the random experiment. Moderate coverage properties were combined with a bad reproducibility. Clearly, the random component should be enhanced in such a situation. Experiment III performed best on all criteria but one, the coverage of the relevant search space. Experiments II and IV were better for coverage, with the former showing slightly better coverage properties. These results lead to the overall sequence of Experiments III > II > IV > I > V, which is in contrast to what would be concluded from Figure 2.

CONCLUSIONS

The four quality criteria introduced here allow a more fundamental comparison of the performance of various GA settings than the usual monitoring of the best individuals. By counting the number of visited hypercubes, the influence on the coverage of various random operators such as mutation and sharing, can be compared directly. However, because a purely random search would obtain the best rank according to this criterion, an additional measure of coverage is introduced that only focuses on the relevant parts of the search space: the number of clusters of sufficiently good solutions. Two other criteria are suggested to test the reproducibility: they are the number of clusters touched in replicate runs and residuals in spaces spanned by the principal components of other runs. The combined application of the four criteria resulted in a clear sequence of performance of the five test cases that was not possible by simple monitoring of the best individuals.

In each case, repeated experiments were used to show that the performance criteria lead to reproducible results. To achieve a general applicability, the number of parameters needed to calculate the quality measures was kept as small as possible. Two of these, the number of principal components to use in the PCA projections and the size of the hypercubes in the coverage criterion, did not have a large effect on the outcome. The two other parameters are clearly problem dependent; these parameters are the radius of the clusters (i.e., the definition that solutions are too similar to each other) and the quality threshold, above which an individual is of no use. Proper values for these parameters must be estimated from test runs. Further work is in progress to test the usefulness of the quality criteria for other problem domains and their application in fine-tuning GA settings and the comparison between different optimization techniques.

ACKNOWLEDGMENT

R.W. acknowledges a travel grant from The Netherlands Organization for Scientific Research (NWO).

REFERENCES AND NOTES

(1) Rechenberg, I. *Evolutionsstrategie. Optimierungsstrategien technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann−Holzberg, Stuttgart, 1973.

(2) Goldberg, D. E. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley: Readings, MA, 1989.

(3) Lucasius, C. B.; Kateman, G. Understanding and using genetic algorithms. Part 1: Concepts, properties and context. *Chemometr. Intelligent Lab. Syst.* **1993**, *19*, 1−33.

(4) Lucasius, C. B.; Kateman, G. Understanding and using genetic algorithms. Part 2: Representation, configuration and hybridization. *Chemometr. Intelligent Lab. Syst.* **1994**, *25*, 99−145.

(5) Blommers, M. J. J.; Lucasius, C. B.; Kateman, G.; Kaptein, R. Conformational analysis of a dinucleotide photodimer with the aid of the genetic algorithm. *Biopolymers* **1992**, *32*, 45−52.

(6) McGarrah, D. B.; Judson, R. S. Analysis of the genetic algorithm method of molecular conformation determination. *J. Comput. Chem.* **1993**, *14*, 1385−1395.

(7) Brodmeier, T.; Pretsch, E. Application of genetic algorithms in molecular modelling. *J. Comput. Chem.* **1994**, *15*, 588−595.

(8) Jackson, J. E. *A user's guide to principal components*; Wiley: New York, 1991.

(9) Bowen, J. P.; Allinger, N. L. Molecular mechanics: the art and science of parametrization. In *Reviews in Computational Chemistry, vol. 2*; Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH: New York, 1991; pp 81−97.

(10) Brodmeier, T.; Pretsch, E. Genetic algorithms for molecular modelling. In *Software Development in Chemistry 9: Proceedings of the 9th Workshop "Computers in Chemistry"*; Moll, R., Ed.; Gesellschaft Deutsche Chemiker: Frankfurt am Main, 1994; pp 213−224.

(11) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data, An Introduction to Cluster Analysis*; Wiley: New York, 1989.

(12) Ihaka, R.; Gentleman, R. R: A language for data analysis and graphics. *J. Comput. Graphic. Statist.* **1996**, *5*, 299−314.

(13) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of conformational coverage. 2. Applications of conformational models. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 295−304.

CI970430H