Parametric Sensitivity and Search-Space Characterization Studies of Genetic Algorithms for Computer-Aided Polymer Design

Anantha Sundaram[†] and Venkat Venkatasubramanian*[,‡]

Laboratory for Intelligent Process Systems, School of Chemical Engineering, Purdue University,
West Lafayette, Indiana 47907

Received May 27, 1998

Genetic Algorithms (GAs) and evolutionary methods have been demonstrated to be flexible and efficient optimization techniques with potential for locating global optima under general conditions for computer-aided molecular design (CAMD). However, they often need customization requiring detailed study of parametric sensitivity and search-space character before optimal internal parameters are determined. This paper describes such a parametric sensitivity study for GA-based polymer design. The objective of the study was to study the influence of the internal parameters of the GA on its performance on a large scale polymer design problem. The study yielded qualitative trends that clearly identified the structure of the target sought or the nature of the search-space to be the key factor determining the GA's performance. The study was then focused toward studying the structure to fitness correlation in the underlying search-space. The results of this study indicate that the performance of the GA could be enhanced by diversified sampling schemes, adaptive parameter tuning, and interactive inclusion of additional design knowledge.

1. INTRODUCTION

Computer-aided molecular design (CAMD) involves the design of molecular structures that satisfy a set of desired properties or performance measures. The task of proposing viable candidates for this purpose is a difficult albeit significant one. The problem has been addressed using a variety of techniques in literature. The design problem of concern here is the proposition of two- or three-dimensional molecular structures that based on an available prediction method are expected to achieve the desired performance. Typically, the molecular structure is assumed to be composed of different functional units and may include additional three-dimensional structural information. The performance criteria can include physical and chemical properties as well as biological activity or performance measures relevant to the domain of application of the material. The predictor method is usually a group contribution, QSAR-based technique that models the structure as a collection of molecular subunits or groups that make a fixed contribution to the property of the molecule. Depending on the domain, the prediction method could be more general to include highly nonlinear neural networks or other black-box models also.

Different techniques have been explored for the solution to this problem over the last several years. Knowledge-base systems,[1,2] machine-learning techniques,[3] and enumeration-based algorithms[4,5] have been quite successful for a variety of problems. Rigorous mathematical formulations have also been proposed and solved for a variety of design problems including solvents,[6,7] refrigerants,[8,9] and polymers.[10−13] The desirable features of any method to this problem are generality of application, ability to handle nonlinear objectives and local optima that arise as a result, ease of implementation and adaptability, computational ease in handling large search-spaces, and robustness to approximations/uncertainties in the property predictors. In spite of their advantages in some domains, all the methods described typically lack one or more of these features.

Genetic Algorithms (GAs)[14,15] have been adapted[16] and successfully applied as a solution framework for the molecular design problem as defined above.[17−20] They have been demonstrated to handle large-scale design problems. In addition, they offer ease and flexibility of implementation across domains and can be adapted to address the different aspects of the design problem.[21] The underling rich chemistry of the design problem is exploited in a straightforward manner leading to a transparent solution framework. In addition, GAs offer complete decoupling of the inverse design strategy from the nature and functionality involved in the property predictor. However, they have the following drawbacks.

(1) Parametric sensitivity: The performance of GAs or GA-based strategies is intimately tied to the values of the different parameters that control various aspects of the algorithm. The discovery of an optimal setting for the parameters or even the existence of one can be determined only with a detailed trial and error approach.

(2) Convergence: GAs are stochastic in nature and several aspects of their implementation are customized to the domain to improve performance. So, in general, no convergence guarantees can be offered, and no definite conclusions made about the expected quality of solutions.

In their work, Venkatasubramanian et al.[22] evaluated the performance of a GA-based search strategy in discovering different molecules with known structures. The algorithm was set the task of identifying that exact structure given only their properties as the desired targets. The results, though

* Author to whom all correspondence should be addressed.
[†] E-mail: anantha@lips.ecn.purdue.edu.
[‡] E-mail: venkat@lips.ecn.purdue.edu.

10.1021/ci980209k CCC: $15.00 © 1998 American Chemical Society
Published on Web 09/25/1998

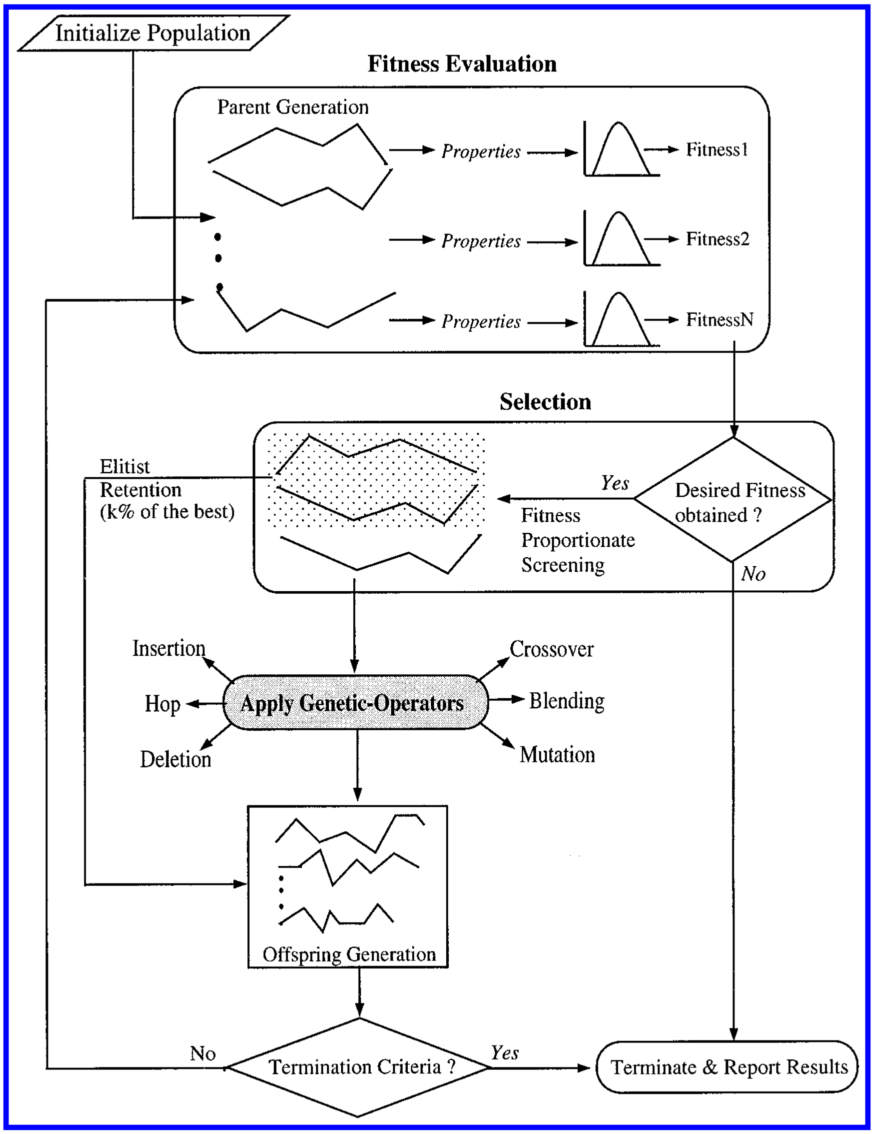


Figure 1. Overall framework of GA-based CAMD.

encouraging, varied widely in terms of the success rate (number of successful target identifications out of several runs of the algorithm) as well as the quality of the final solutions obtained. This indicated that some of the factors mentioned above may be at play, and to obtain any improvement in performance improvement, a detailed parametric sensitivity study needed to be performed. Such a study could determine the influences of the structure of the search-space, the underlying parameter space and their interaction, on the performance of the algorithm. This would help to establish whether an optimal setting could be obtained, independent of the nature of the target structure or design problem. In this work, we used the GA-based CAMD algorithm of Venkatasubramanian et al.[22] to perform a parametric sensitivity study in an effort to systematically determine optimal parameter settings. In addition, we tried to characterize the search-space and identify strategies that would allow the GA to recognize the underlying structure of the search-space and exploit it to improve convergence. In the next section, we briefly outline the GA-based CAMD algorithm and the parameters involved. In section 4.3 the parametric sensitivity study is introduced. The results of the study and the major conclusions are discussed in section 4.4. The search-space characterization effort is detailed in section 5, and conclusions are drawn based on them. Some avenues for performance enhancements are outlined in section 6 including some current efforts.

2. GA BASED CAMD

The GA-based CAMD methodology of Venkatasubramanian et al.[17,22] is outlined in Figure 1. The problem of concern was the determination of the structure of the repeat unit of the polymer, given a set of desired property values. The properties under consideration were density, glass transition temperature, thermal conductivity, heat capacity, and bulk modulus. The algorithm was initialized with a population of randomly created molecules of varying sizes (number of functional groups), by combining functional groups from a predefined list called the base-group set. The population size or the number of molecules in the population was fixed a priori. The properties of each polymer in the population were evaluated using the group contribution techniques of van Krevelen.[23] The property values were compared to the desired values, and the difference represented a fitness measure between 0 and 1. A fitness function transforms the shortcoming in the property values of a molecule compared to the desired property values, into a real value between 0 and 1. Once the fitness value for each of the molecules was determined, molecules were drawn randomly but in proportion to their fitness values from the population. This allowed the fitter candidates in the population to be chosen more often than others and gave them a greater chance for propagating their desirable characteristics. The chosen molecules (called parents) underwent the genetic operations indicated in Figure 1 to create an offspring population. A small fraction of the parent population was retained without any changes in the offspring population to avoid loss of good solutions. This policy is called elitism. The use of the genetic operators was stochastic, and the frequency of their use was determined by preset probabilities. The parameters associated with the fitness function, the elitist policy as well as the operator probabilities, play a crucial role in determining the performance of the GA-based CAMD. These were examined in the parametric sensitivity studies. An optimal setting for these parameters would enhance the general performance of the GA across different target property sets.

The objective of this study was to determine if such a setting existed and to establish rules for obtaining them in such a case. In the absence of an optimal setting we wished to determine the factors affecting the performance and explain the results. The roles played by the different parameters are discussed below.

Fitness Function. The fitness function is a transformation that groups together the different levels of desirability in the five different properties into a single fitness score. The following fitness function was used by Venkatasubramanian et al.[22]

$$F = \exp\left(-\alpha\left[\sum_{i=1}^{5}\left\{\frac{P_i - P_{i,\text{desired}}}{P_{i,\max} - P_{i,\min}}\right\}^2\right]\right) \qquad (1)$$

where P_i is the value of property "i" of the molecule and $P_{i,\text{desired}}$ is the desired value of that property. $P_{i,\max}$, and $P_{i,\min}$ are the maximum and minimum bounds on the property values. The quantity $P_{i,\max} - P_{i,\min}$ is usually referred to as the *allowed tolerance* in the property value. The parameter α is the decay factor. The magnitude of the decay factor determines how strictly a candidate is penalized for not meeting the desired property values. A large value of the decay factor would maintain strict standards for meeting the desired property values, and small shortcomings will be heavily penalized leading to small fitness values. On the other hand, a small decay factor would be more lenient, and even large deviations from the desired property values would lead to moderate fitness scores. The effect of the decay factor is very significant in the mechanics of the genetic search especially in influencing the so-called selection pressure of the algorithm. For instance, in the case of a large decay factor, small differences in the deviations from desirability between two molecules tend to get greatly amplified leading to widely different fitness scores. In combination with a fitness proportionate selection, this would lead to greater focus of the search on only the highly promising regions. But in the extreme they may lead to premature convergence of the GA to a population of similar and suboptimal solutions. When a small decay factor is used, the fitness function is more forgiving. In this case the GA may accept solutions with large deviations from desirability thus increasing the diversity of the sampled solutions. The disadvantage in this case would be the oscillatory nature of the algorithm as it superficially samples vastly different areas of the search-space.

Elitism. The fitness proportionate selection scheme followed by the application of genetic operators often leads to manipulation of the different components of the parents as well as recombination of these characteristics to create the next generation. While this is a desirable feature of the GA in that it allows a wider or more global sampling of the search-space, it might lead to a regular loss of good characteristics from the population. In a situation like that of a GA-based CAMD that allows the use of several different genetic operators, it often leads to poor convergence properties. To ensure that good solutions are not lost regularly, a fraction of the top solutions from the parent population is placed in the next generation without any changes. This strategy is called elitism. The exact fraction of retention can influence the selection pressure. When a large fraction of the good solutions are retained from one generation to another, the GA can maintain only narrow population diversity. On the other hand, a small fraction might not retain enough of the desirable characteristics to overcome convergence problems.

Genetic Operators. Genetic operators represent the different moves that parts of a population make in moving from one region in the search-space to another. Two important operators that are used in GA-based CAMD as in many other GA-based applications are crossover and mutation. Crossover brings two different flavors to the search process. In manipulating the structural properties, i.e., components of molecules within a population, crossover reinforces certain groups and combinations while eliminating others. In examining newer combinations of the genetic/base-group pool in the population, crossover examines the effect of large leaps in the structure space on the quality of solutions. While this is necessary for any search possessing a global sampling character, it is deficient in the sense that nothing is added to the gene/base-group pool of the initial population. One needs an additional operator namely mutation to perform this task. Mutation introduces groups into the population on a regular basis thus providing an opportunity for novelty to occur in a functional group sense. The relative probabilities of crossover and mutation determine the balance and trade-off between exploitation and exploration.

The performance of the GA involves two important mechanisms. They are the introduction of random change and the survival of the fittest. Of the parameters described above, the genetic operators are the primary mechanisms that introduce random change. Among the operators used in the CAMD procedure, crossover and mutation by their very nature are the most significant. The decay factor and the elitist policy are parameters that enforce the survival of the fittest paradigm. Together, all these parameters are instrumental in the performance of the GA, and tuning them can significantly alter it. Hence, they were chosen to be the parameters of choice for the sensitivity study.

The following sections describe the different studies and present the results and conclusions based on them. Parametric sensitivity refers to the variation in performance of the GA-based CAMD as a function of the internal parameters

Table 1. Parameter Settings Examined

crossover/mutation	0.50,[a] 1, 2, 3, 5, 10, 30, 300
elitism	2, 5, 10,[a] 15, 20
decay factor (α)	0.0001, 0.0005, 0.001,[a] 0.002, 0.01

[a] Indicates setting used in the original polymer design case-study.[22]

of the GA such as the ones described above. In this regard, a parameter-space is the space of all the internal parameters of the GA spanning the entire region of the values they can assume. In contrast, the search-space refers to the combinatorial space of molecular structures from which the GA samples its candidates and the fitness values (determined from their properties) associated with them. In this sense, it is sometimes referred to as the structure-fitness space. A fitness function presupposes an evaluation procedure that determines the properties of interest corresponding to any candidate examined as well as a set of target or desired property values. However, except under the controlled conditions of the sensitivity and search-space studies, the target property set may or may not correspond to a target structure. Enumeration studies explore the complete search-space while sampling studies look at large and diverse aggregates of the search-space, in order to obtain clues about their global character.

3. PARAMETRIC SENSITIVITY STUDIES

The sensitivity of the performance of the GA-based CAMD to three different parameters was examined in this study. They were the ratio of crossover to mutation probability, the decay-factor, and the percentage of elitist retention. The general aspects of their influence on genetic search were discussed in the previous section. Each of the parameters described in the previous section was stepped through a series of settings, and these are indicated in Table 1. While studying a particular setting for one of the three parameters, the other parameters were fixed at their base-case (a) values. The effect of interactions between these parameters was not studied. The GA-based algorithm for CAMD is a stochastic one, and conclusions can only be based on the average performance over several runs. In addition, we proposed to examine the influence of interaction between structural characteristics of the target and the parameter settings, on performance. Toward these goals, four different target structures were identified from the original case-study.[22] They are shown in Figure 2. The choice of these targets was based on their varied structure as well differences in the algorithm's success rate in identifying them, in the original case-study. From a property predictor perspective, the calculation of properties for these four structures varied from simple additive group-contributions (targets 1 and 2) to more complex corrections (targets 3 and 4), that tend to introduce additional nonlinearity into the problem. These corrections actually reflect the effect of the so-called epistasis or group-ordering dependency of the property evaluation from molecular structure. The choice of the four targets was also motivated by a desire to study this effect in the light of parametric sensitivity. For a given setting of the parameters, the algorithm was set the task of identifying the target structure given the properties of the structure evaluated using the prediction model. The GA-based CAMD started with an initial population of size 100 and successively iterated over a maximum of 1000 generations. This run was repeated 50 times starting with a different initial population each time. The set of 50 runs was repeated for the same parameter setting using the properties of each of the other three targets as desirables.

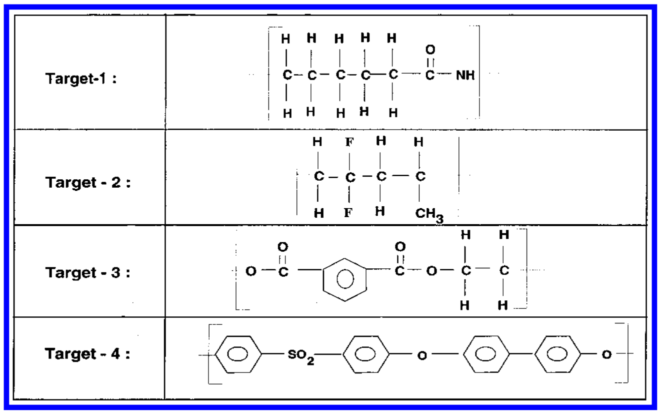


Figure 2. Target structures for sensitivity studies.

The success rate and the convergence rate over 50 runs of the algorithm for each of the given targets was recorded. The success rate is defined as the percentage of runs out of the 50 in which convergence to the target structure was obtained. The convergence rate is the average generation (over the successful runs) in which the target was identified. Then, the difference in the success rate and/or convergence rate over the base-case performance was determined, and significance tests were performed to evaluate the statistical significance of this difference. This difference was used as the basis of comparison between the different parameter settings. The significance tests are performed to test the hypothesis that two different settings of a particular parameter actually give rise to significantly different results in the success rate and/or the convergence rates. In order to do this one needs to take into account the combined standard error in the case of the success rates and the standard deviation in the case of the convergence rates. The test statistic used for the test of significance in the success rate difference is given by[24]

$$z_{\text{success}} = \frac{\left(\dfrac{s_1 - s_2}{N}\right)}{\sqrt{\dfrac{2p_s(1-p_s)}{N}}} \tag{2}$$

where s_1 and s_2 are the number of successful runs for the two different parameter settings to be compared respectively, N is the number of runs in each case (= 50), and p_s is the pooled success estimate which is the average success rate obtained by combining the two sets. For N as large as 50, this statistic can be assumed to follow a normal distribution[24] i.e., $N(0,1)$ and hypothesis tests for the significance of the difference can be performed using the confidence levels from this distribution. In all the success rate plots presented in the following sections, the z_{success} is shown on the Y-axis instead of the actual difference $s_1 - s_2$ where s_2 is the success rate for a defined base-set of parameter values. Similarly when comparing the convergence rates for different parametric scenarios, the scaling factor used to scale the actual difference is the combined standard deviation of the

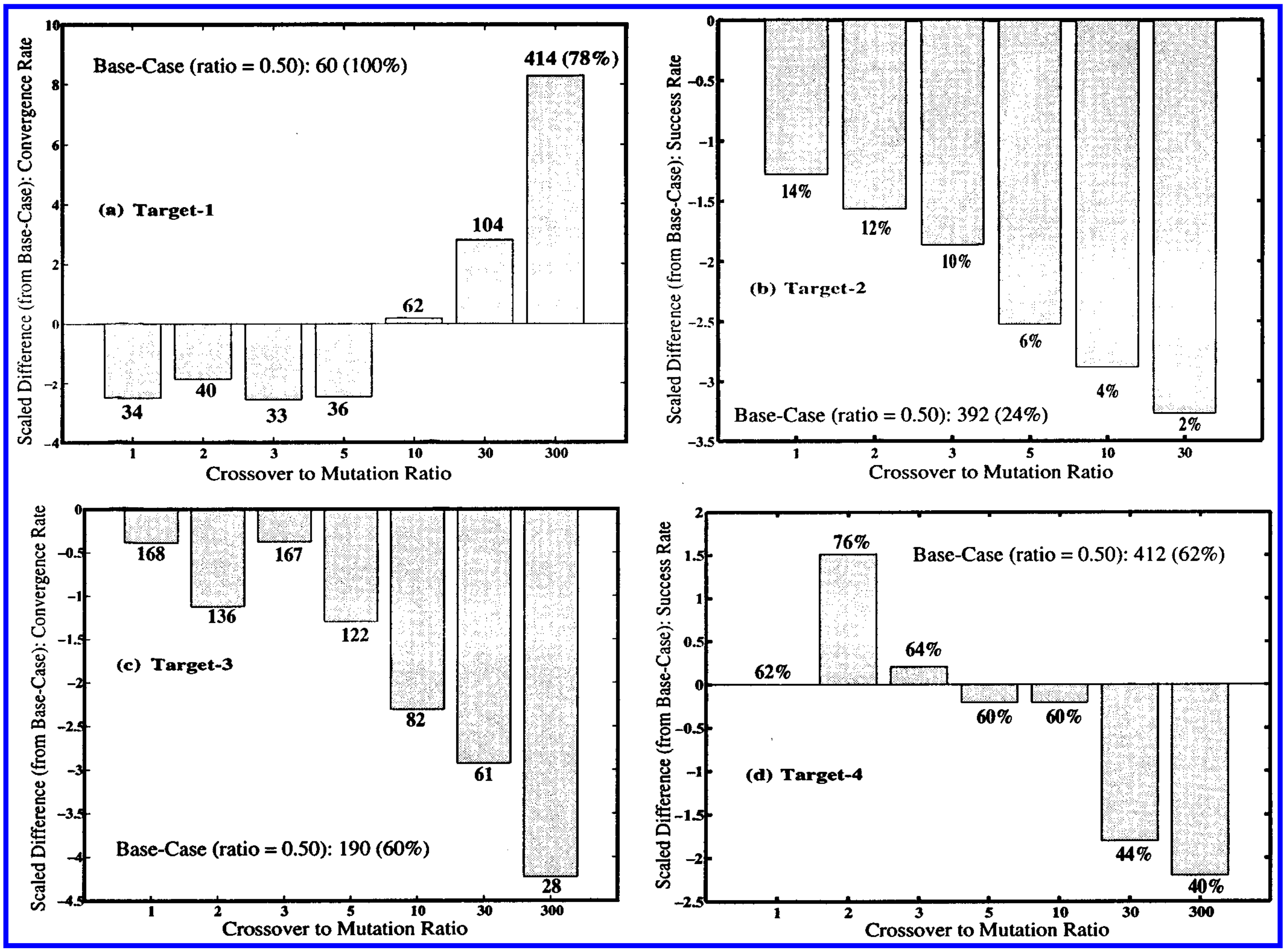


Figure 3. Variation of performance of GA-based CAMD on crossover to mutation probability ratio.

convergence rates across the two sets of runs. In this case the test-statistic used for the significance tests is[25]

$$Z_{\text{convergenc}} = \frac{c_1 - c_2}{\sqrt{\dfrac{v_1}{N_1} + \dfrac{v_2}{N_2}}} \tag{3}$$

where c_1 and c_2 are the convergence rates, v_1 and v_2 are the convergence rate variances estimated from the successful runs among the 50 runs in each case, and N_1 and N_2 are the successful runs in each case. For large N_1 and N_2, the above statistic follows a t-distribution with ν degrees of freedom where ν is given by[25]

$$v = \frac{\left(\dfrac{v_1}{N_1} + \dfrac{v_2}{N_2}\right)^2}{\dfrac{\left(\dfrac{v_1}{N_1}\right)^2}{N_1 + 1} + \dfrac{\left(\dfrac{v_2}{N_2}\right)^2}{N_2 + 1}} - 2 \tag{4}$$

The scaling factor used in the convergence rate plots in the following sections is the denominator of the test-statistic in (3), which is also the combined standard deviation of the two cases being compared. Other statistics such as the average and maximum fitness values of each generation in every run were also recorded for all the different cases examined.

4. RESULTS AND DISCUSSION

4.1. Influence of Crossover to Mutation Probability Ratio. The results of the parametric sensitivity study for the crossover to mutation frequency ratio are shown in Figure 3. The X-axis of all the charts in Figure 3 are the crossover to mutation probability ratios examined in the study. The Y-axis of the charts indicate the scaled difference between the performance of the GA with a particular probability ratio setting and that of the GA with the base-case setting of the ratio. For targets 2 and 4 (Figure 3b,d), the success rate differences were significant enough across the different ratio settings to use as the distinguishing feature. For these figures, the difference in the success rate is plotted along the Y-axis. For targets 1 and 2 (Figure 3a,c) the success rates did not vary significantly across the different settings, but the convergence rate did. In such cases, the convergence rate differences were used to distinguish performance. They are plotted on the Y-axis of Figure 3a,c. In each case, the performance differences were scaled using the combined standard deviations of the runs being compared. Scaling the results in this fashion allowed statistically significant differences to be amplified against statistically negligible differences on the bar graphs. The numbers on the bars in Figure 3b,d indicate the actual success rates (not differences

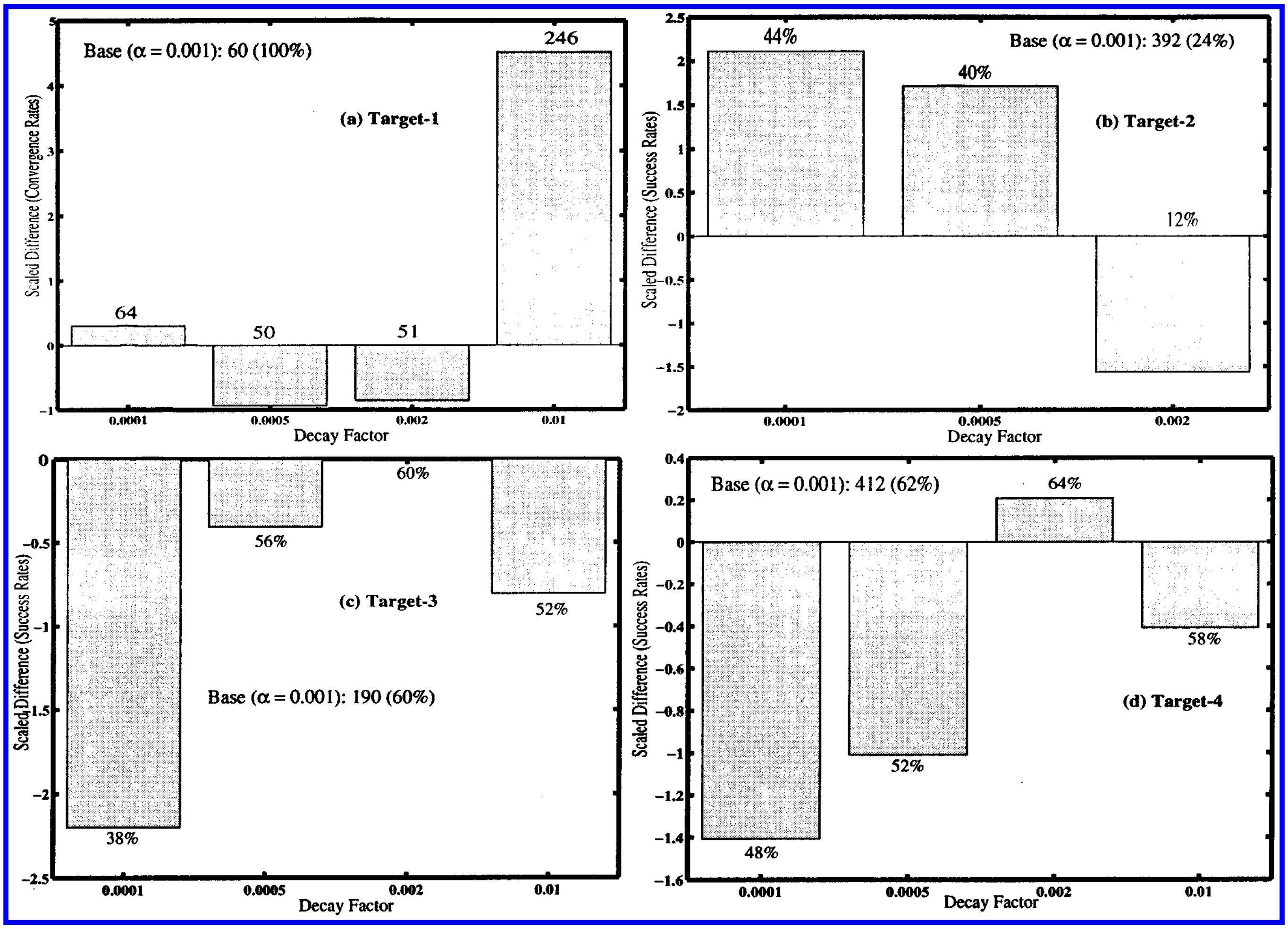


Figure 4. Variation of performance of GA-based CAMD with decay factor.

or scaled values) for that setting and that particular target. Similarly, the numbers on the bars in Figure 3a,c indicate the actual convergence rates for that particular setting and target combination. The base-case convergence rates for a given target are also indicated on each chart and the base-case success rates in brackets.

From an inspection of Figure 3a, it is clear that a nominal increase in the ratio improved performance (i.e., success rate), while a very large increase actually degraded the performance. This trend can be explained on the basis of the structure of target 1. Target 1 contains only two different groups in its structure of which the "CH_2" group also occurred in a number of high fitness molecules in the population and remained preserved from one generation to the next. This was also helped by a special crossover operator that occasionally converts all carbon groups with different side chains to methyl groups. However, the CONH group is rather selectively preserved in the sense that it is not always present in the high fitness members of the population. A large increase in the crossover probability at the expense of mutation probability suppresses the regular reintroduction of this group. In fact, in a number of runs that failed when the ratio was increased to a large value of 300, the final population did not contain a single CONH group. For this setting, the algorithm was able to locate the target only 78% of the time, whereas the success rate was always 100% for lower settings.

In locating target 2, any increase in the crossover probability diminished the performance. The structure of target 2, has three different side-chain groups, the identification of which greatly benefited from an increased frequency of side-chain mutation. All crossover operators except for the special operator discussed above affect only combinations along the main-chain. Side-chain mutation was the only operator that substituted groups other than hydrogen groups on the side chains. Consequently, increasing the frequency of this operator improved the performance, or conversely, decreasing that frequency led to performance loss. The performance of the GA-based CAMD while locating targets 3 and 4 showed a different trend altogether (see Figure 3c,d). The properties of these two molecules are dependent upon the ordering of the different functional groups that compose them. Target 4 (see Figure 2) especially, contains three subsequences of groups that have to be identified and combined in the right order to obtain the exact set of target properties. Large crossover rates were detrimental as the critical subsequences were broken by randomly chosen crossover cut-points.

In the case of target 3 (see Figure 2) the sensitivity to ordering was relatively less. This is because there was only one subsequence that needed to be captured in the exact order for an exact property match, once all the other groups had been identified. A nominal increase in crossover helped the GA to examine different combinations leading to consistent identification of the critical subsequence. Once identified, this sequence was not broken frequently. This is because the subsequence was short, compared to the design length and the probability of a crossover cut-point disturbing the

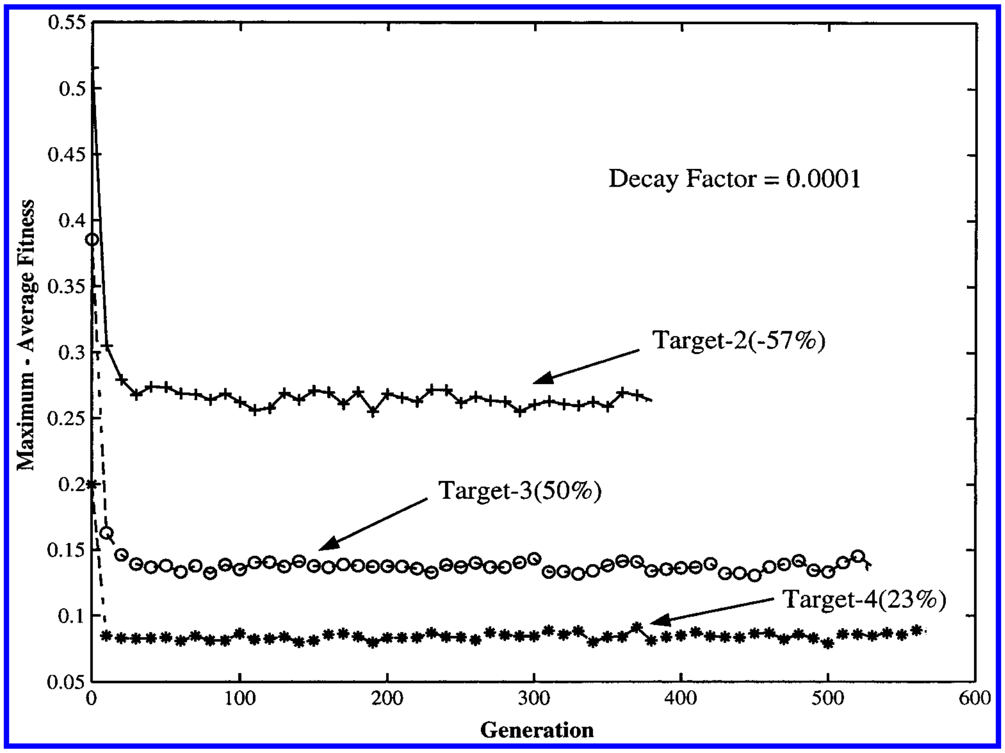


Figure 5. Fitness evolution with generation (average over 50 runs).

sequence was low. Thus, the trend went through a maximum in success rate for target 4 as the crossover to mutation probability ratio was increased.

The results of the sensitive studies for the crossover to mutation probability ratio indicate a strong dependence of the performance of the GA on the structure of the target sought. In particular the desirability of a particular ratio setting in terms of overall performance improvement depends upon the components of the target structure as well as the nature of group-contributions (linear or nonlinear) that determine the properties for that structure. This underlines the importance of the structure of the search-space and its interaction with the internal parameters of the GA in determining its performance.

4.2. Effect of Decay Factor. Figure 4 shows the effect of changing the decay factor on the performance of the GA-based CAMD in locating the four different targets. The trends varied widely across the targets as indicated. Due to the lack of differentiation between the different settings in terms of the success rate, for target 1 alone, convergence rate differences were used as the basis of comparison. As in the previous study, the performance difference was determined over the base-case performance corresponding to a setting of $\alpha = 0.001$. The *X*-axis on all the charts are the decay factor levels. While the effect on performance was not significant for target 1 except at a large α value, in other cases sharply distinct patterns were found. Success rates improved with lower values of α for target 2. For a large $\alpha = 0.01$, the algorithm failed to converge in all the runs for this target and is not shown on the chart. The performance of the GA in locating targets 3 and 4 worsened with lower values of the decay factor, but higher values did not bring significant improvement in performance. Some insight into the reason behind these differences was obtained by looking at a different set of observations of the GA's performance. The decay factor affects the selective pressure within the population. One measure of the selective pressure within the population is the difference between the maximum and average fitness of its members. Figure 5 shows a plot of this difference as a function of generation (averaged over 50 runs) for targets 2−4. The plots were obtained for a decay factor of 0.0001, which was the lowest value examined. This corresponded to a scenario where the objective function was lenient in screening the candidates. A large difference between maximum and average fitness in the population indicates that the population was well differentiated in terms of fitness. Hence, increasing the decay factor or tightening the penalty did not lead to a large performance gain. On the other hand, a small difference between maximum and average fitness values indicates that several solutions were being lumped together in high fitness brackets. This afforded a better chance for average fitness solutions to propagate into the subsequent generations which would lead to a "dilution" of the future populations with less fit members. Certainly, in this case, one would expect that increasing the penalty by increasing the decay factor would eliminate weaker solutions better and improve performance. As shown in Figure 5, the ordering of performance enhancement in going to a larger decay factor (keeping all other parameters at their base values) was target 2(−57%) < target 3(+50%), target 4(+23%) in keeping with the relative ordering in the magnitude of the difference between the maximum and average fitness of the population. These values are the average of the actual (not scaled) percentage success rate differences obtained in switching from a low α of 0.0001 to higher values of $\alpha = 0.0005, 0.001, 0.002$. The difference in the performance improvement between target 3 and target 4 was not statistically significant even though the actual values do not follow the relative ordering expected.

The primary result of the decay factor analysis was whether a setting is optimal or not depends upon the target

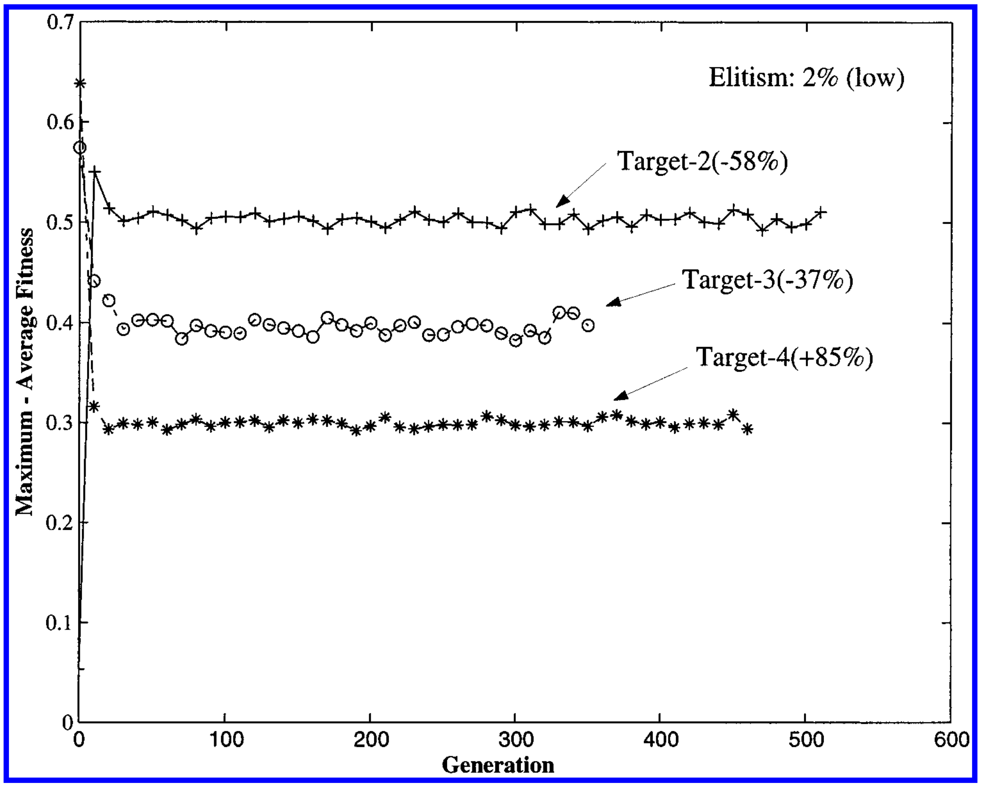


Figure 6. Fitness variation with generations (averaged over 50 runs).

properties and how well the GA is performing in achieving those objectives. The relative magnitude of the difference between the maximum and average fitness in the population between different targets was found to be a qualitative indicator of the desired decay factor setting. A large difference suggested opting for less discrimination between solutions in a population to gain larger diversity by decreasing the decay value. In contrast, a small fitness difference between maximum and average suggested use of greater discrimination within the population by increasing the decay value.

4.3. Effect of Elitism. Elitism like decay factor affects the selection pressure in the population but in a different manner. As in the decay factor study, the difference between the maximum and average fitness within a population was observed. Figure 6 shows a plot similar to Figure 5, that of the difference between maximum and average fitness in a population as a function of generation.The elitism value corresponding to the runs shown on the plot is 2%. This was the lowest value studied in the elitism sensitivity study. As in the previous case, a large value of the maximum to average difference was symptomatic of a large fraction of members with fitnesses between the maximum and the average fitness of the population. Therefore, a higher value of retention allowed several solutions close to the average fitness to be passed on to the next generation. This reduced the focus of the GA, leading to poorer performance. On the other hand, a small difference between average and maximum fitness indicates presence of a number of high fitness members.

In this case, a larger value of elitism allowed the GA to retain more of the high fitness solutions and preserving the characteristics of the promising locations in the structure space. As indicated in Figure 6, the magnitude of the difference between maximum and average fitness in a generation follows target 2 > target 3 > target 4. The average percentage improvement in performance obtained by increasing the elitist retention fraction from 2% to higher values (5%, 10%, 15%, and 20%), followed the ordering target 4(+85%) > target 3(−37%) > target 2(−58%).

There was a large jump in the performance improvement for the GA in locating target 4(+85%) as opposed to the performance degradation obtained in the case of target 3(−37%) though the magnitudes of the maximum to average fitness difference in the population were fairly close to one another. We suspect that this may be due to the difference in the structures of target 3 and target 4 and the nature of subsequence contribution. Retaining several molecules in a population that contain the critical subsequences helped to sustain them against constant depletion due to crossover. The structure corresponding to target 4, as described earlier, contains three subsequences that need to be identified by the GA as opposed a single compact subsequence in target 3. Between the two targets, the effect of crossover is definitely more damaging to the GA's performance in the case of target 4. This was also seen previously while analyzing the effect of crossover to mutation ratio. The performance of the algorithm gained much more due to an increase in elitist retention while locating target 4 (relative to the search for target 3), because a larger fraction of the high-fitness solutions were shielded from the deleterious effects of crossover.

The results of the elitism study again revealed a strong interaction between the nature of the parameters and the search-space in determining the sensitivity of the GA. The magnitude of the difference between the maximum and average fitness within a population roughly determines what a desirable setting would be in this case also. An increase in the elitist retention percentage was found to improve the GA's performance when the maximum to average fitness difference was small and when this difference was large, decreasing the elitist percentage was favorable. As in the decay factor study, the trade-off involved was between discrimination and diversity within the population. Additional discrepancies in the quantitative differences in the GA's performance across the different targets are suspected to be due to the differences in the specific molecular structures of the targets.

The parametric sensitivity studies indicated a large dependence of the performance of the GA-based CAMD on the values of the parameters examined. In spite of the absence of an optimal setting, very strong qualitative correlation between the search-space and the parameters was the repeating theme in all the studies performed. This result though interesting by itself, definitely warrants a closer look at the search-space to characterize it. The final objective of the analysis was to achieve performance improvement. In this regard, given the results of the parametric sensitivity study two issues remain to be addressed. Firstly, what was the nature of the underlying search-space? How were the solutions distributed? A different analysis of the search-space independent of the GA's mechanics was likely to reveal a clearer picture of the search-space. This would yield insight into the character of the molecular search-space relevant to this case-study. Nevertheless, given the common dynamics of different molecular design problems the study would yield widely applicable generalizations. Secondly, we wanted to determine the extent of the correlation between fitness and structure in the search-space and whether the fitness based objective was sufficient to guide a search to globally optimal structures. This would expose any inherent deficiencies in the knowledge used by the search procedure to perform the design.

5. SEARCH-SPACE CHARACTERIZATION

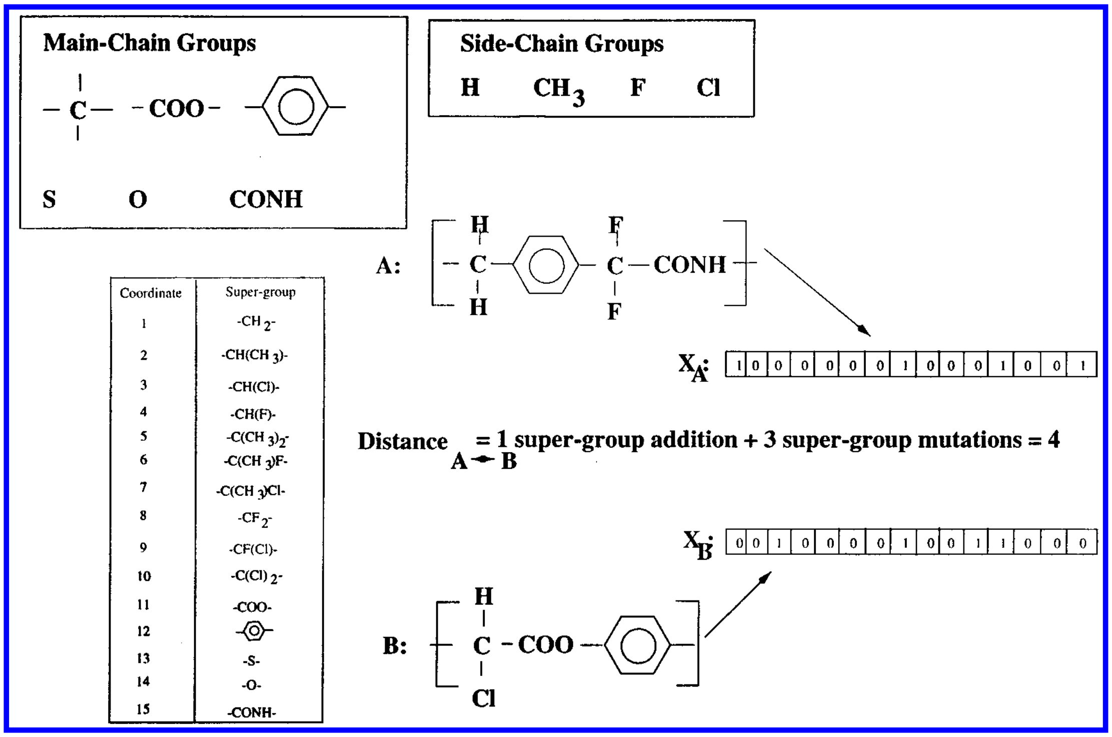
In this section, we describe the search-space characterization study for the polymer design problem. The analysis was structured to be independent of the mechanics of the GA-based CAMD. In other words, the goal was to obtain an idea of the underlying distribution of solutions in the search-space. Every structure in the search-space has a fitness value associated with it. This gives a measure of the desirability of the candidate. The local neighborhood of a structure in the search-space contains very similar structures that differ slightly in one or more components. They are called 1-mutant, 2-mutant, or *n*-mutant neighbors of the structure under consideration, depending upon the number of functional unit changes between them. The nature of the local search-space could be established by tracing the local variation of fitness across these neighbors. Given a target structure (and therefore a target set of properties), one can structurally map the entire neighborhood in relation to the target as 1-mutant, 2-mutant, and so on until the *n*-mutant neighbors. Here, "*n*" is the maximum number of changes allowed, given the constraint on the length of the designs.

In this process, one would also obtain a fitness neighborhood corresponding to the above structural neighborhood. Given this information, one can then try to establish a correlation between the two neighborhoods. The first step in mapping the structure of the search-space was to establish an alphabet for the representation. The logical alphabet set was the base-group list as it was the smallest level of detail at which the designs were manipulated by the search procedure. Since the alphabet set included two different types of groups i.e., main-chain and side-chain groups, they were grouped together to create a set of super-groups. Super-groups are functional units that can occur independently on the main-chain and have a valency of two. They are unique combinations of main-chain and side-chain groups such that the resulting functional units have a valency of two. We considered a small case-study involving six main-chain groups and four side-chain groups in the base-group list as shown in Figure 7. The super-group list obtained from this base-list is also shown in the figure.

Any structure created as a combination of the main-chain and side-chain groups can also be represented as a combination of the super-groups. Alternatively, the same structure could be mapped into an integer coordinate system where each dimension represents the number of occurrences of a particular super-group assigned to that dimension. Such a coordinate system would have as many dimensions as the number of super-groups involved. This alternative representation is shown for an example structure in Figure 7. Now, given two molecules A and B one can obtain two vectors X_A and X_B which are the super-group based coordinates of the two molecules. Using X_A and X_B, one can determine a pseudo-Hamming distance between molecules A and B in the structural sense. This distance is the number of super-group mutations, additions, and deletions required to completely convert molecule A to molecule B. For binary representations, the number of positional differences between two strings is the Hamming distance.[15] In that sense, the structural distance obtained between two structures in this case is a "pseudo-Hamming" distance involving nonbinary representations.[21] One can then map the entire search-space based on pseudo-Hamming distances from a predefined structure.

A large-scale enumeration case-study of the order of the problem examined by the GA-based CAMD is impossible to perform. Instead, we used random walks starting from either a randomly chosen molecule or the target molecule to study the structure of the search-space. By changing the target structure, one can get additional information about the differences or similarities in the landscape. A walk comprises of an initially chosen structure and all structures that are obtained by making successive single changes to the structure. Hence a structure at any step in the walk is a 1-mutant neighbor (neighbor that is one change away) of the structure at the previous step and is therefore a 1-mutant move away. A 1-mutant move is one of mutation, addition, deletion, or position-change operations involving a super-group in the chosen molecule, and the actual move made is randomly selected from among those operations at each step. The length of the walk is the total number of 1-mutant moves made starting from the initial structure.

In particular, three different types of walks were used. A target structure was first identified and its properties determined using the group contribution predictor. In the first type of walk, one started at a randomly chosen structure and stepped through a series of 1-mutant moves. The second type of walk was also a random walk, but the starting point was always at the target under consideration. Finally, the third type of walk was a hill-climbing walk that started at a



Coordinate	Super-group
1	$-CH_2-$
2	$-CH(CH_3)-$
3	-CH(Cl)-
4	-CH(F)-
5	$-C(CH_3)_2-$
6	$-C(CH_3)F-$
7	$-C(CH_3)Cl-$
8	$-CF_2-$
9	-CF(Cl)-
10	$-C(Cl)_2-$
11	-COO-
12	-⌬-
13	-S-
14	-O-
15	-CONH-

Figure 7. Super-group determination and distance calculation.

randomly chosen initial structure. In this case, a 1-mutant operation was accepted only if it led to an improvement in fitness. If the fitness did not improve, the move was discarded, and another 1-mutant move was made. The walk was terminated if after 10 000 such attempts no fitness improvement could be produced. The maximum length of the walk was 1000 steps for the random walks and 100 steps for the hill climbing walks. The results were based on 500 different random or hill-climbing walks. The fitness of the solution obtained the distance from the target structure as well as the distance from the starting point in the walk were calculated for each step of every walk. In addition, for the hill-climbing case, the number of trials attempted before obtaining a fitness improvement was recorded for each step of every walk. The target structure was then changed, and the whole sequence of walks was repeated.

Some of the analysis is similar to the ones done by Weinberger,[26] Manderick et al.,[27] Kauffman,[28] and Jones and Forrest[29] for different problems. However, there are some important differences. These studies involved problems with fixed length strings and binary alleles. In this study however, the length of the strings (molecules) was variable between a minimum of two and a maximum of 15 super-groups. The maximum size considered was two units more than the maximum size of designs explored by the GA-based CAMD. Secondly, the number of alleles allowed at each position of the string was 252 (the number of super-groups) instead of two as in the binary case. However, this problem can be cast in an alternative representation where each molecule is represented in the super-group coordinate space of 252 coordinates. Each coordinate represents the number of times the super-group corresponding to that coordinate position occurs in the molecule. As each coordinate could contain integers from zero to 12, this would give rise to a total binary string length of 252 × 4 = 1008 binary coordinates. But the number of nonzeros is very small in this representation, due to restrictions on the total number imposed by the maximum length constraint implicit in the search. Besides, a 1-mutant move in the sense of a single super-group change does not correlate exactly to a 1-mutant move in the binary sense for mutation. In the binary equivalent of a single super-group mutation, a 1-to-0 change, at a position corresponding to the super-group being replaced, will always be accompanied by a 0-to-1 change, at the position corresponding to the super-group chosen as replacement. Hence, a 1-mutant move in the super-group space corresponds to a 2-mutant move in the binary space.

The results of this study are summarized in Figures 9−12. For the purely random-walk case (first kind) the autocorrelation between fitness values of structures that were separated by a certain number of 1-mutant moves was calculated as a function of the separation and averaged over 500 different walks. The autocorrelation function used here is the same one used in time series analysis and in this case quantifies the correlation between the fitness values that are separated by a finite number steps along a random walk. Along a random walk, we record the sequence of fitness values obtained as F_1, F_2, F_3, ..., F_M, where M is the length of the random walk. The autocorrelation[27] function ρ_t between fitnesses "t" moves apart is given by

$$\rho_t = \frac{V(t)}{\sigma^2} \tag{5}$$

$$V(t) = \frac{1}{M} \sum_{j=0}^{M=t} (F_j - F_{\mathrm{av}})(F_{j+t} - F_{\mathrm{av}}) \tag{6}$$

where $V(t)$ is the estimated autocovariance of fitnesses "t" moves apart, F_{av} is the mean of the fitness sequence, and σ^2 is the variance of the stochastic process.

The autocorrelation function for the first type of random walk as a function of the move separation is plotted in Figure 9. It clearly shows that the correlation decreases exponen-

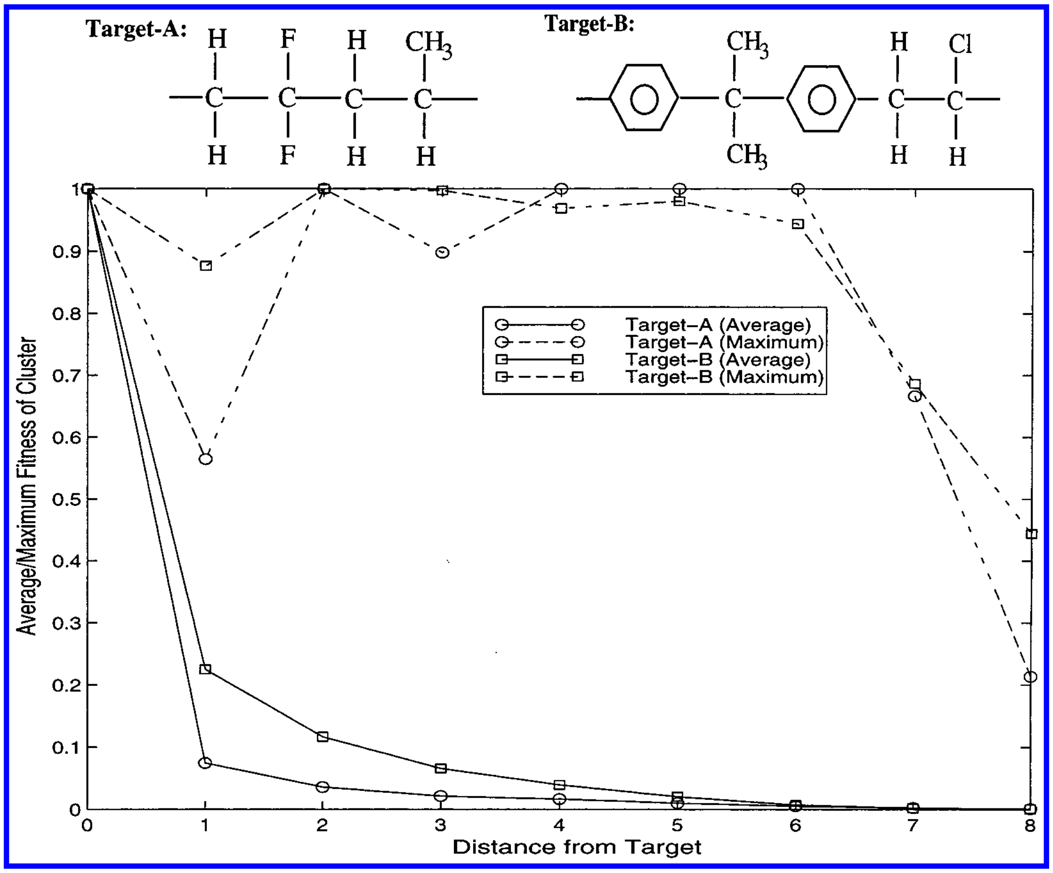


Figure 8. Fitness variation with distance from target, based on complete enumeration.

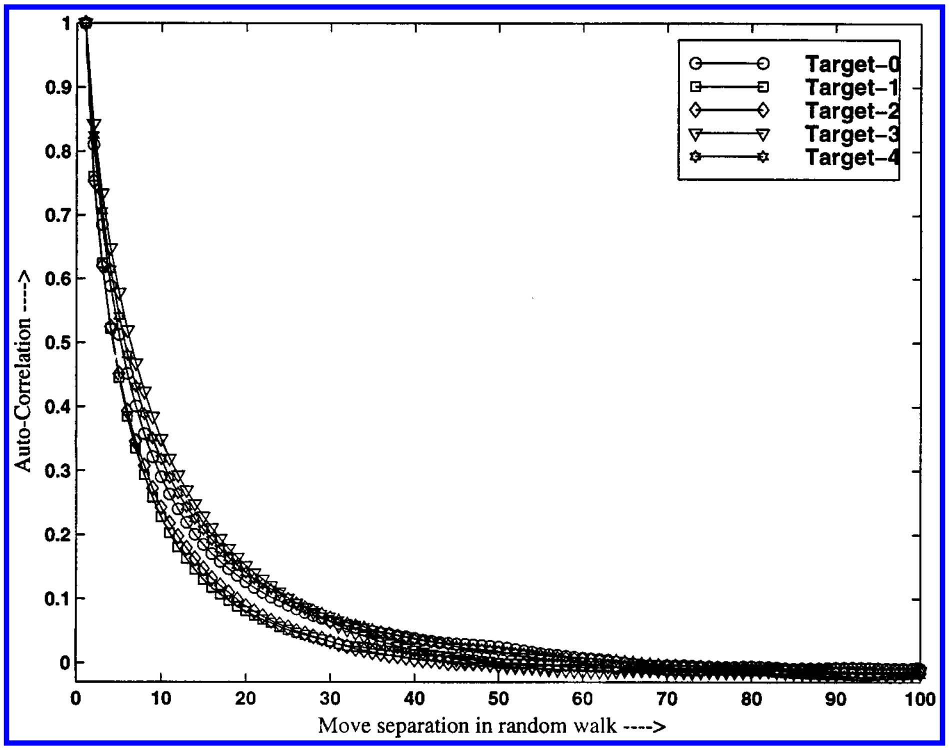


Figure 9. Autocorrelation function for fitness for different targets.

tially with increasing structural separation between solutions. This correlation trend also seems rather independent of the target sought, i.e., does not change significantly with the evaluation function except at large distances. The essential character of the search-space pointed by the analysis is that there is a gradual decay of the fitness correlation with increasing structural dissimilarity. This implies that small changes to molecules would in general lead to smaller fitness changes, and large changes may or may not lead to large changes in fitness due to a diminished correlation to the parent's fitness. The success of any informed search procedure depends on the ability of the procedure to locate

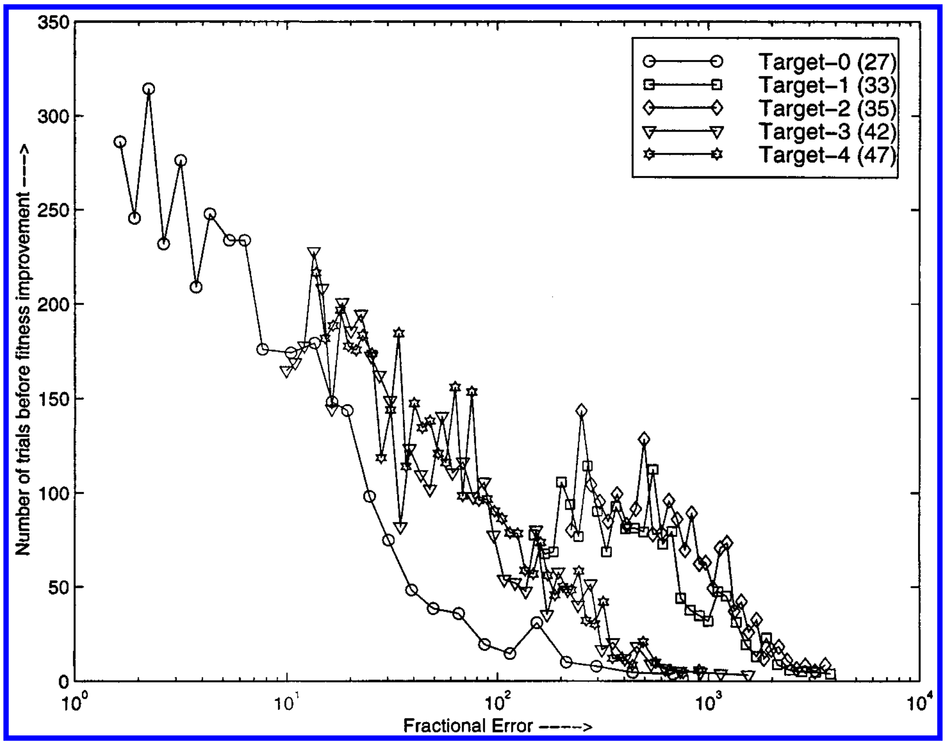


Figure 10. Variation of accessability to local optima with error improvement.

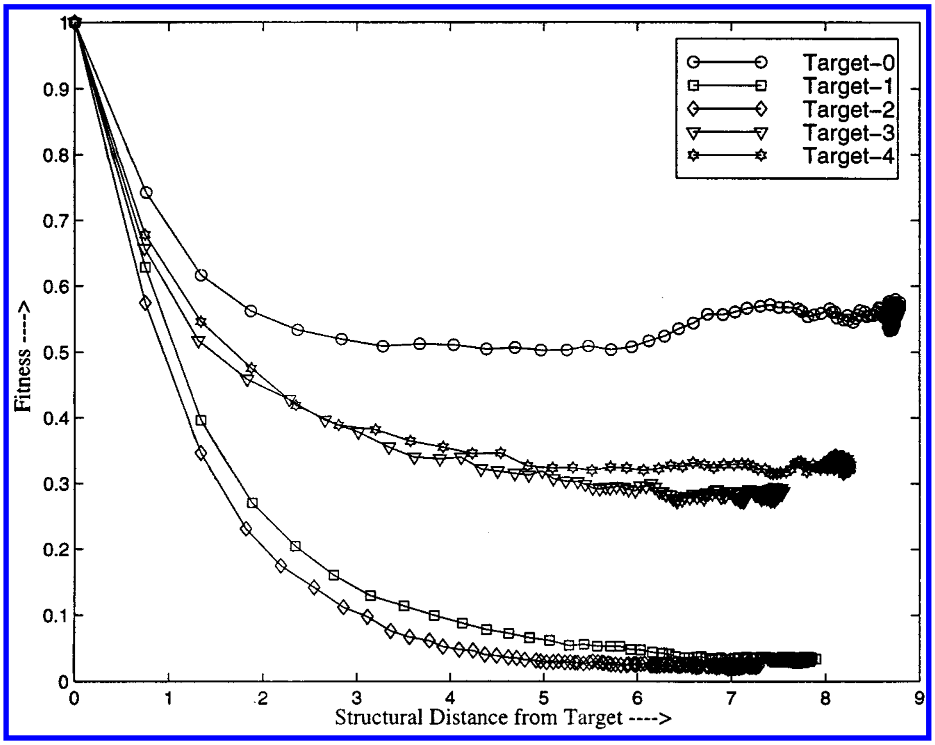


Figure 11. Fitness variation with distance for walks starting at the target (average over 500 walks).

high-fitness local-optima and to manipulate it incrementally.

Figure 10 shows the results obtained from the hill-climbing walks starting from a random initial structure for each of the five targets. The plot shows the number of trials made before obtaining a decrease in error as a function of the error. The error shown here is the fractional error (FE), which is the sum of squares of the errors in the five properties, scaled by the allowed tolerance. The tolerance allowed in these cases was ±0.5% of the property value sought, for each of the properties. This is typically a design parameter chosen by the chemist/designer. A low fractional error corresponds to a high fitness. The average length of a hill-climbing walk is shown in brackets against each target in the legend of Figure 10. The plots trace the number trials along a hill-climbing walk from left to right with respect to the figure. In every case, it is clear that as one moves farther and farther along the path of a hill-climbing walk, i.e., from left to right, the number of trials required to obtain an improvement in fractional error increases. This means that in all the cases examined, it became more difficult to access local optima as the hill-climbing walk progressed. In the case of targets 3 and 4 the hill-climbing walks were truncated much later,

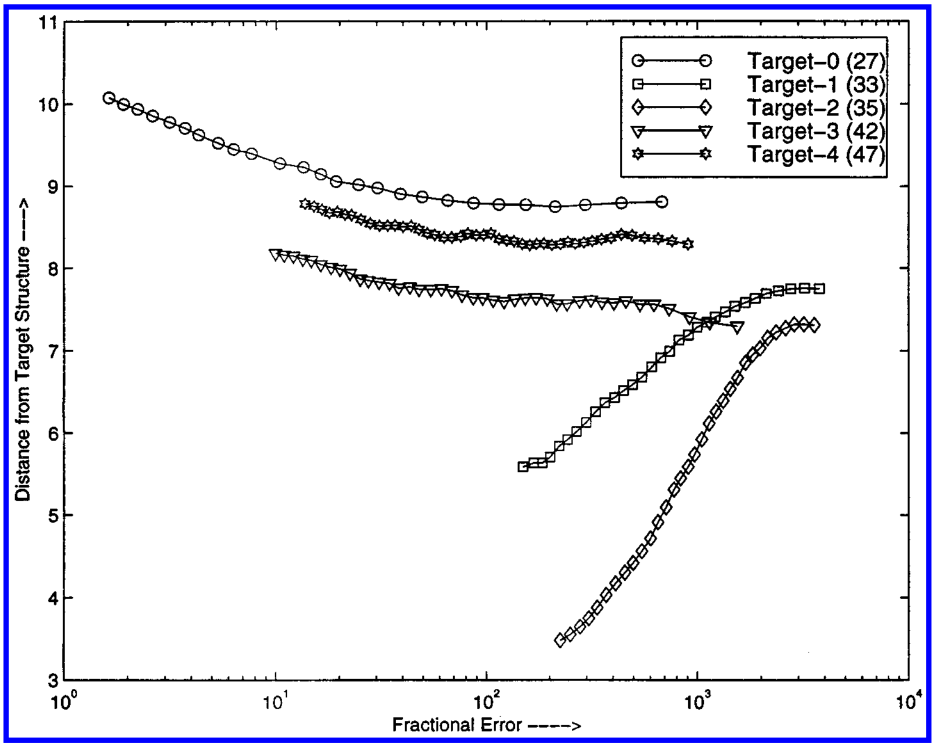


Figure 12. Fractional error variation with distance from target for hill-climbing walks.

on average, compared to targets 0, 1, and 2. The ultimate fractional error at the end of the walk was also higher and the fitness correspondingly lower for searches involving targets 1 and 2, compared to the others.

The hill-climbing search for target 0 contained the shortest walks while also yielding the lowest ultimate fractional error. The accessibility of local-optima at any step in the walk (toward a given target) not only depended on the fitness of the solution at that step but also sharply varied depending on the target property sought. For this case-study, the property sets such as those for targets 0, 3, and 4 seemed to be located in the high probability density regions of the multivariate distribution of the five properties. On the other hand, the property values for targets 1 and 2 were located in a lower probability density region. This reflected on the accessibility to local optima differently for each target. Hence, local-search procedures would be more susceptible to entrenchment in local optima when looking for targets 0, 3, and 4 as opposed to the search for target 1 or 2. As seen before, the local fitness correlation was fairly similar across different targets. However, the accessibility to local fitness optima varied sharply across the different targets making searches across the similarly correlated spaces vastly different. More importantly, the fitness to structural-similarity correlation possessed drastically different trends across targets, as shown by the hill-climbing analysis.

This is further in evidence in Figures 11 and 12. Figure 11 shows the plot of the fitness variation with distance from the target along a random-walk starting at the target based on 500 such walks. It is apparent from the figure that a random-walk starting from the target leads to lesser and lesser fit solutions in all cases examined. However, the fitness settles to a large value even after several steps, in the case of a walk starting at target 0 as compared to the walks starting at target 1 or target 2, where the fitness drop with distance is quite precipitous. This demonstrates a greater accessibility to high-fitness solutions while looking for target 0 compared to the rest even far away from the target.

Irrespective of the abundance of local optima in terms of the fitness values, there was always a single known global optimum structure in each of the five cases examined. It was not obvious whether fitness based local optima corresponded to structures similar to the known global optimal structure. Some insight into this could be obtained from Figure 12, where the fractional error in property values is plotted against distance to the global-optimal target structure during a hill-climbing walk. The plots are based on the average of 500 hill-climbing walks simulated as described earlier. For targets 1 and 2, as the walk progressed successively through solutions with lower fractional error (and higher fitness) than the previous one, the distance to the target sought also increased. This implies that using fitness improvement alone as the guiding principle, an algorithm is likely to identify optima that are structurally close to the global optimum. On the other hand, for targets 0, 3, and 4 the distance or structural dissimilarity to the global optimum was either constant or increased with fitness improvement. Hence, a hill-climbing algorithm looking for one of these targets (especially target 4) will be able to gain fitness improvement without actually sampling structures similar to the target structure.

In summary, the first major result of the search-space studies was that there was reasonable correlation between the fitness of neighboring solutions and this correlation decreased exponentially with distance of separation between structures, irrespective of the nature of the target. Secondly, the difficulty in accessing local-optima from a given structure varied from one target to another. Thirdly, fitness landscape around the target for a fixed decay factor varied dramatically from one target to another. Finally, accessing higher and higher fitness optima did not always guarantee sampling of structures closer to the known global optimum. Given the correlated nature of the landscape a search procedure can keep sampling high fit solutions that are "far" away from the global optimum. From a purely optimization point of view, it is not important whether the solutions (molecules) follow a desirable structural pattern, as long as the resulting property values are close to those desired, i.e., not the global optimum but very close. However, practical aspects of molecular design enforce that structures be realistic or conform to some rules dictated by other considerations than just property achievement. In all the cases we studied the targets were always structurally "simpler" than the high fitness local-optima that the GA or hill-climbing search located. In a practical situation where the designer does not know all the characteristics of the structure sought, this becomes a crucial issue. In such cases, additional knowledge that outlines the simplicity or complexity of the structures in addition to their property based fitness needs to be incorporated into the solution algorithm. This knowledge could be in the form of objectives weighted for complexity[22] and/or by the use of knowledge augmented operators for solution manipulation. Some avenues toward realistic design are outlined in the next section.

6. CONCLUSIONS

Evolutionary design of molecules possessing desired properties offers great scope for large-scale global optimization in general nonlinear landscapes. However, due to their stochastic nature, they cannot offer any convergence guarantees. Moreover, the efficiency and convergence properties of the search strongly depend on the underlying parameter space, the nature of the search-space and the interaction between the two. In this paper the sensitivity of a GA-based CAMD was studied using a large-scale case-study, and the trends observed were explained on the basis of the interaction between the parameters and the nature of the search. The study also clearly highlighted the absence of a single optimal setting for the parameters examined. Moreover, a parameter setting found to be beneficial for performance in searching for a particular target was found nonoptimal for a different target. The results implied that an optimal or desirable setting could be defined only on a run to run basis. Any method to locate better performing settings should be adaptive and should dynamically guide the parameter values based on the performance of the genetic algorithm in seeking that particular target. Such methods are currently being examined toward enhancing the algorithm's performance.

The target-specific nature of the settings exposed a critical influence of the nature of the search-space on the mechanics of the genetic algorithm. This prompted a structure-space characterization study that was performed using enumeration for small-scale problems and using random and hill-climbing walks for larger cases. The major outcome of the case-study was the fact that the structure of the fitness landscape was drastically altered by the target property settings. While in some cases, the landscape was amenable to search using convexity based algorithms, in other cases it remained rather flat but reasonably correlated for small changes. The performance of the GA in locating the different molecules could not be correlated consistently to the results of the search-space characterization except for target 0.

This target was located 0% of the time by the GA for different parameter settings studied. It was pointed out in the hill-climbing analysis that this landscape was the most likely to provide highly fit but at the same time structures highly dissimilar to the global optimum. However, the performance in locating the other targets could not be compared on a common footing as their performance varied widely across different parameter settings, and the complete space of these settings was not examined to use an average. The effectiveness of crossover and crossover-like operators in the search was not quantified in the search-space characterization study. These would drastically alter the way the GA manipulates the solutions as opposed to a simple hill-climber that only makes local moves. These are some issues that need to be addressed in the future. Nevertheless, the study yielded some clues into the nature of the search-space that were fairly generic and independent of the type of search algorithm employed.

The first one, pointed by the enumeration study, was the breadth as well as the depth of the sampling is crucial to performance. In other words, diversity should not only include variety in terms of the distances between the solutions examined but also enough samples at a given distance of separation. This becomes more profound for nonbinary representations. Secondly, searches in highly correlated landscapes need not yield solutions close to the global optimum in a structural sense, and the location of the target property set in the property space is pivotal in determining how drastically the fitness changes around the target. Finally, the accessibility to local optima may play an important role in determining the adequacy of a sampling scheme in locating globally optimal solutions and as such varies with the target properties sought. Some possible areas for future work based on these results are discussed below.

The fitness based objective pools together errors in different properties. This leads to loss in resolution between solutions when searching the space. By considering each of the property objectives separately, and using ideas from multiobjective optimization, significant gains have been obtained in other GA-based algorithms. Such algorithms are currently being adapted to this problem to examine improvements in sampling and consequently in performance. The local high fitness optima that were located in the search-space characterization study did not always possess structural features similar to the target. This suggests including additional design knowledge that brings to bear some formulation-based decision making that can isolate these desirable structural characteristics.

While this knowledge is domain specific, it is anticipated that significant gains could be obtained by making the search procedure interactive. Such a procedure provides a framework for easy and transparent inclusion of such design decisions supplied by the user. This would allow the designer to dynamically change the focus of the evolutionary search by altering the structure of the solutions as well as the internal parameters of the algorithm itself. Furthermore, in common industrial practice, there are rules that govern the design of formulations. Such rules capture experiential design know-how and are often valuable in suggesting synthetically feasible candidates for closer examination. In an interactive evolutionary framework, one can envision such rules occurring as constraints at different levels in the construction of solutions. Efficient incorporation of formulation rules into the genetic operators to ensure creation of feasible designs is an important area that is being pursued. With the use of an interactive framework incorporating constrained genetic-operators that obey formulation rules, diverse sampling strategies, and adaptive parameter tuning methods, it is envisioned that the GA-based design system would be a flexible and powerful tool for highly efficient, realistic as well as novel product design.

REFERENCES AND NOTES

(1) Gani, R.; Brignole, E. A. Molecular Design of Solvents for Liquid Extraction using UNIFAC. *Fluid Phase Equilibria* **1983**, *13*, 331−340.

(2) Derringer, G. C.; Markham, R. L. A Computer-based Methodology for Matching Polymer Structures with Required Properties. *J. Appl. Polym. Sci.* **1985**, *30*, 4609−4617.

(3) Bolis, G.; Pace, L. D.; Fabrocini, F. A machine learning approach to computer-aided molecular design. *J. Comput. Aided. Molecular Design.* **1991**, *5*, 617−628.

(4) Joback, K. G.; Stephanopoulos, G. Designing Molecules Possessing Desired Physical Property Values In Proceedings of the *FOCAPD*; Snowmass, CO, 1989; pp 363.

(5) Skvortsova, M. I.; Baskin, I. I.; Slovokhotova, O. L.; Palulin, V. A.; Zefirov, N. S. Inverse Problem in QSAR/QSPR Studies for the Case of Topological Indices Characterizing Molecular Shape (Kier Indices). *J. Chem. Inf Comput. Sci.* **1993**, *33*, 630−634.

(6) Macchietto, S.; Odele, O.; Omatsone, O. Design of Optimal Solvents for Liquid-Liquid Extraction and Gas Absorption Processes. *Trans I Chem E.* **1990**, *68A*.

(7) Odele, S.; Macchietto, S. Computer Aided Molecular Design: A Novel Method for Optimal Solvent Selection. *Fluid Phase Equlibria* **1993**, *82*, 47−54.

(8) Duvedi, A.; Achenie, L. E. K. Designing Environmentally Safe Refrigerants Using Mathematical Programming. *Chem. Eng. Sci.* **1996**, *51*, 3727−3739.

(9) Churi, N.; Achenie, L. E. K. Novel Mathematical Programming Model for Computer Aided Molecular Design. *Ind. Eng. Chem. Res.* **1996**, *35*, 3788−3794.

(10) Vaidyanathan, R.; El-Halwagi, M. Computer-Aided Design of High Performance Polymers. *J. Elast. Plast.* **1994**, *26*.

(11) Vaidyanathan, R.; El-Halwagi, M. Computer-Aided Synthesis of Polymers and Blends with Target Properties. *Ind. Eng. Chem. Res.* **1996**, *35*, 627−634.

(12) Maranas, C. D. Optimal Computer-Aided Molecular Design: A Polymer Design Case Study. *Ind. Eng. Chem. Res.* **1996**, *35*, 3403−3414.

(13) Maranas, C. D. Optimal Molecular Design under Property Prediction Uncertainty. *AICHE J.* **1997**, *43*.

(14) Holland, J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, MI, 1975.

(15) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: Reading, MA, 1989; p 412.

(16) Venkatasubramanian, V.; Sundaram, A. *Genetic Algorithms: Introduction and Applications*; Encyclopedia of Computational Chemistry, 1998.

(17) Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. Computer-Aided Molecular Design Using Genetic Algorithms. *Computers Chem. Eng.* **1994**, *18*, 833−844.

(18) Glen, R. C.; Payne, A. W. R. A genetic algorithm for the automated generation of molecules with constraints. *J. Comp.-Aid. Mol. Design.* **1995**, *9*, 181−202.

(19) Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. Genetic Algorithmic Approach for Computer-Aided Molecular Design. In *Computer-Aided Molecular Design*; 1995; pp 396−414.

(20) Devillers, J. Designing Molecules with Specific Properties from Intercommunicating Hybrid Systems. *J. Chem. Inf. Conmput. Sci.* **1996**, *36*, 1061−1066.

(21) Venkatasubramanian, V.; Sundaram, A.; Chan, K.; Caruthers, J. M. Computer-aided Molecular Design using Neural-Networks and Genetic Algorithms. In *Genetic Algorithms in Molecular Modeling*; Devillers, J., Ed.; Academic Press: London, 1996; pp 271−302.

(22) Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 188−195.

(23) Krevelen, D. W. v.; Hoftyzer, P. J. *Properties of Polymers, their Estimation and Correlation with Chemical Structure*; 1990.

(24) Collett, D. *Modelling Binary Data*; 1991.

(25) Montgomery, D. C.; Runger, G. C. *Applied Statistics and Probability for Engineers*; John Wiley & Sons: New York, 1994.

(26) Weinberger, E. Correlated and Uncorrelated Fitness Landscapes and How to tell the Difference? *Biological Cybernetics* **1990**, *63*, 325−336.

(27) Manderick, B.; Weger, M. d.; Speissens, P. The Genetic Algorithm and the Structure of the Fitness Landscape. In *Proceedings of the Fourth International Conference on Genetic Algorithms*; Belew, R. K., Booker, L. B., Ed.; Morgan Kaufmann: San Diego, CA, 1991; pp 143−150.

(28) Kauffman, S. *Origins of Order: Self Organization and Selection in Evolution;* Oxford University Press: New York, 1993; p 709.

(29) Jones, T.; Forrest, S. Fitness Distance Correlation as a Measure of Problem Difficulty for Genetic Algorithms. In *Proceedings of the Sixth International Conference on Genetic Algorithms*; Morgan Kaufmann: Pittsburgh, PA, 1995; pp 184−192.

CI980209K