In David Fogel, editor, Proceedings of the Second IEEE Conference on Evolutionary Computation,
Piscataway New Jersey: IEEE Service Center, 1995.

Simple Genetic Algorithm Parameter Selection for Protein Structure Prediction

George H. Gates, Jr., Laurence D. Merkle, Gary B. Lamont
Department of Electrical and Computer Engineering
Graduate School of Engineering
Air Force Institute of Technology
Wright-Patterson AFB, OH 45433
{ggates, lmerkle, lamont}@afit.af.mil

Ruth Pachter
Wright Laboratory
3005 P St., Ste. 1
Wright-Patterson AFB, OH 45433-7702
pachterr@ml.wpafb.af.mil

ABSTRACT

Selection of run-time parameters is a critical step in the application of genetic algorithms. Numerous investigations have discussed parameter set selection, both theoretically and empirically. Theoretical work has focused on the choice of population size [7, 8, 9, 13, 16], while empirical studies cover a wide range of GA parameters [3, 4, 10, 15]. Theory suggests population sizes which increase exponentially with string length. The available experimental data suggests small populations perform consistently well, but the test problems are limited to small string lengths. Thus, we still do not have a complete understanding of how parameters should be chosen, especially for problems with large string lengths.

This study extends Schaffer's results by performing a similar empirical analysis of GA parameters on a real-world application, with longer string lengths and a very large number of local optima. Relationships between population size, mutation rates, and crossover rates similar to those reported by Schaffer are shown.

1. Introduction

Selecting run-time parameters is notably the most difficult part of successfully applying genetic algorithms to search and optimization problems. Several investigations have discussed parameter set selection both theoretically and through experimental analysis. Theoretical work is chiefly aimed at the choice of population size [7, 8, 9, 13, 16], and provides conflicting guidelines. Goldberg has argued that extremely small populations may be warranted for serial SGAs [8], but according to more recent theoretical results, the population size necessary to statistically guarantee correct decisions increases with string length [9].

Empirical studies covering a wide range of GA parameters and combinations thereof (e.g. [3, 4, 10, 15]) suggest that small populations perform consistently well across a range of problems. One of the most comprehensive empirical analyses of parameter settings is that of Schaffer *et. al.* [15]. Schaffer uses a set of ten test functions, including De Jong's five-function test suite. The study identifies ranges of population size, crossover rates, and mutation rates that exhibit good online performance over the range of test functions. It also evaluates the effects of one- and two-point crossover and finds that the latter is always at least as good as the former when considering online performance. Table 1 lists the parameters suggested by Schaffer and those proposed earlier by De Jong [3] and Grefenstette [10].

Table: 1: Comparison of Empirically Determined GA Parameter Settings

Author	Population Size	Crossover Rate	Mutation Rate
Schaffer	20 – 30	0.75 – 0.95	0.005 – 0.01
De Jong	50 – 100	0.60	0.001
Grefenstette	30	0.95	0.01

This study extends Schaffer's work by performing a similar empirical analysis of GA parameters. His study used pedagogical test problems with relatively short string lengths compared to those required by our problem domain. In light of theory indicating population size should grow exponentially with string length, the applicability of Schaffer's results to our real-world application, with significantly longer string lengths, is unclear. Additionally, the number of local optima in our application far exceeds the number found in any of the test problems.

2. Background

We are interested in the performance of GAs applied to the minimization of polypeptide (protein) energies. The energy function is characteristically non-linear and contains many local optima [2]. The primary determinants of a protein's three-dimensional structure, and thus the energetics of the system, are its independent dihedral angles [17]. Our GA operates on individuals which encode these dihedral angles [6], which necessitates string lengths consderably larger than those used by Schaffer (e.g. 240 for the relatively small protein [Met]-Enkephalin).

This application illustrates a difficulty with existing theoretical population sizing guidelines. For example, Goldberg [9] shows that a population size of

$$n = 2c\left(\frac{\sigma^2_{rms}(m-1)}{d^2}\right)\chi^k \qquad (1)$$

is sufficient to ensure a specified level of confidence in building block decision making, where we have omitted additional terms which account for noisy operators,

c is a function of the confidence level α,
$\sigma^2_{rms}(m-1)$ is the fitness variance ($m = l/k$),
l is the string length,
k is the estimated order of deception,
d is the signal difference we wish to detect, and
χ is the cardinality of the encoding alphabet.

Population sizes suggested by Equation 1, together with both the following conservative assumptions:

1% sampling error is allowed ($\alpha = 0.01, c \approx 6$),
the signal difference is $d = 0.1$,
the estimated order of deception is $k = 5$,
the maximum energy is $f_{max} \approx 75^2 \cdot 10^9$,
the minimum energy is $f_{min} \approx 0$,

and each of three assumptions regarding the fitness variance:

- Maximum: $\sigma^2_{rms} = \frac{(f_{max}-f_{min})^2}{4}$
- Estimate based on a sample of 40,000 random conformations: $\sigma^2_{rms} \approx 10^{19}$
- Minimum: $\sigma^2_{rms} = \frac{(f_{max}-f_{min})^2}{2\chi^l}$

are summarized in Table 2. The best case (minimum variance) calculation indicates no population is required. This is clearly not a useful result. The other cases also do not provide useful results because the computational cost associated with the resulting population sizes is beyond our capability to implement.

Table: 2: Theoretical Population Size Required for Optimal Solution Convergence of [Met]-enkephalin

Variance Assumption	Calculation Parameters			Population Size
	l	χ	σ^2_{rms}	
Maximum	240	2	1.65×10^{23}	2.98×10^{29}
Estimated	240	2	9.20×10^{18}	1.66×10^{25}
Minimum	240	2	8.95×10^{-48}	1.62×10^{-41}

Alternatively, the empirically determined parameters suggested by De Jong, Goldberg, or Schaffer could be used (Table 1). However, each of these recommendations are based on experiments using pedagogical functions with comparatively short encodings and relatively few local optima. The large string lengths required for this application, as well as the large number of local optima present in the fitness landscape, bring into question the applicability of previous experimental findings.

In the next section, we describe experiments similar to Schaffer's to determine appropriate parameter settings for the [Met]-Enkephalin energy minimization problem. In Section 4 we present the results of these experiments. Section 5 summarizes the paper and presents conclusions.

3. Experiment Design

The following experiments are modeled after those described by Schaffer [15]. They are designed to identify the ranges of parameter settings conducive to good GA performance for an application with large string lengths and many local optima such as the energy minimization of [Met]-enkephalin. We consider population sizes $n \in \{10, 20, 30, 50, 100, 200\}$, mutation rates $p_m \in \{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1\}$, and crossover rates $p_c \in \{0.05, 0.15, 0.25, \ldots, 0.95\}$. With the exception of using two-point crossover exclusively and regular binary encoding instead of gray code, all controls are set exactly as in Schaffer's study. A complete factorial design for the 420 remaining parameter combinations is performed, and ten repetitions with different random number seeds are run for each combination.

Our objective function, which we seek to minimize, is the CHARMM [1] energy function

$$\begin{aligned} E = & \sum_{(i,j)\in\mathcal{B}} K_{r_{ij}}(r_{ij} - r_{eq})^2 \\ & + \sum_{(i,j,k)\in\mathcal{A}} K_{\Theta_{ijk}}(\Theta_{ijk} - \Theta_{eq})^2 \end{aligned}$$

$$+ \sum_{(i,j,k,l)\in\mathcal{D}} K_{\Phi_{ijkl}}[1+\cos(n_{ijkl}\Phi_{ijkl}-\gamma_{ijkl})]$$

$$+ \sum_{(i,j)\in\mathcal{N}} \left[\left(\frac{A_{ij}}{r_{ij}}\right)^{12} - \left(\frac{B_{ij}}{r_{ij}}\right)^{6} + \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \right] \tag{2}$$

where the four terms represent the energy due to bond stretching, bond angle deformation, dihedral angle deformation, and non-bonded interactions respectively.

The particular biomolecule investigated here is the pentapeptide [Met]-enkephalin. This molecule is chosen because its native conformation is known and it has been used as a test problem for many other energy minimization investigations (e.g. [11, 12]).

The encoding used is an affine mapping of each dihedral angle (ranging from $-180°$ to $180°$) into 10 consecutive bits. This encoding yields approximately a third of one degree precision. Twenty-four dihedral angles determine [Met]-enkephalin's structure, hence the string length is 240.

Because the simple GA does not reliably find the accepted global optimum using the CHARMM energy function, our definition of "doing well" differs from Schaffer's. We choose the following definition to closely approximate his: at least 10% (42) of the cells in the design locate a value within 10 kcal/mol of the best known solution (-35.1155 kcal/mol) at least 50% of the time (5 out of 10 repetitions) [15]. The raw data is analyzed using the Kruskal-Wallis test to identify the members of the *best online pool.* Test cells (parameter set combinations) belong to the best online pool if their performance cannot be statistically distinguished from the best performing cell.

The experiments are performed using the 1990 version of GENESIS running on a SPARC10 workstation. The generation gap and scaling window parameters are set to 1 and the elitist strategy is used. To guarantee that our online performance criteria is met, the maximum total trials is set to 10,000.

4. Results

Table 3 lists the 23 members of the best online pool starting with the settings that exhibit the best average online performance. Cases with normalized mean online performance (NMOP) near 1 correspond to prematurely converged GAs, while cases with NMOP near 0 correspond to effective search. The Kruskal-Wallis statistic, calculated from the raw sample data, is $h = 31.691$, which is less than the critical value of the Chi-square distribution at the $\alpha < .05$ significance level. Thus we accept the null hypothesis and conclude that these 23 parameter settings exhibit the same online performance.

Table: 3: Best Online Pool

Population Size	Crossover Rate	Mutation Rate	Mean Online Performance
10	0.75	0.005	0.014678
10	0.85	0.005	0.014678
10	0.95	0.005	0.014920
20	0.95	0.002	0.015697
20	0.65	0.002	0.015831
20	0.75	0.002	0.017045
10	0.55	0.005	0.017276
10	0.65	0.005	0.017276
20	0.85	0.002	0.017567
20	0.35	0.002	0.017919
30	0.15	0.002	0.018720
20	0.25	0.005	0.019704
20	0.45	0.002	0.020177
10	0.15	0.005	0.021270
10	0.25	0.005	0.021270
20	0.55	0.002	0.021707
30	0.05	0.002	0.022144
20	0.55	0.005	0.022557
10	0.05	0.005	0.022617
20	0.15	0.002	0.022739
10	0.35	0.005	0.023407
10	0.45	0.005	0.023407
20	0.25	0.002	0.023686

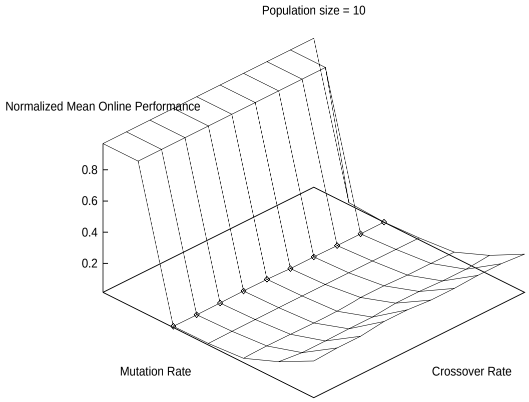

Fig. 1: NMOP (Population Size = 10)

Figures 1 through 6 show the average online performance for all 420 parameter set combinations. The diamonds mark the locations of the members of the best online pool. The results show relationshiops between population size, mutation rates, crossover rates, and NMOP similar to those reported by Schaffer [15]. Most evident is a very strong inverse relationship between population size and mutation rate to achieve good NMOP. In contrast, there is no evidence of any strong relationship between population size and crossover probability to achieve good NMOP.

Figures 1 thru 3 show premature convergence is highly likely when the mutation rate is too low with respect to population size. Optimal GA perfor-

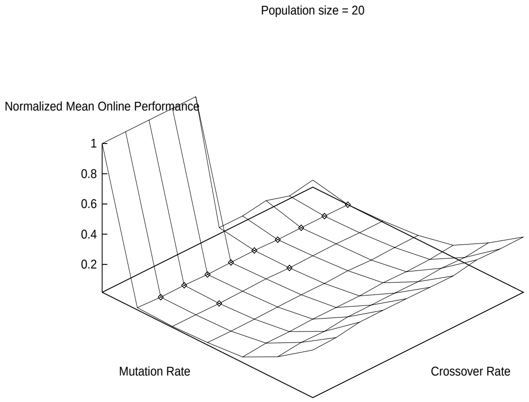

Fig. 2: NMOP (Population Size = 20)

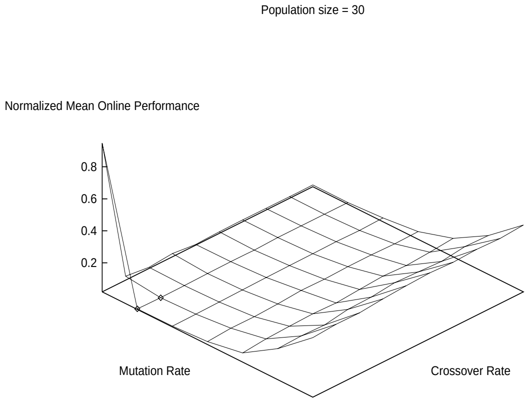

Fig. 3: NMOP (Population Size = 30)

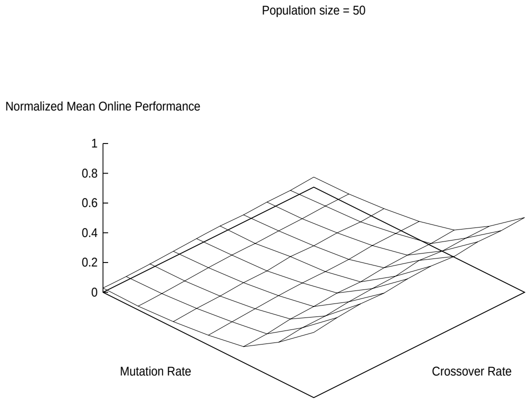

Fig. 4: NMOP (Population Size = 50)

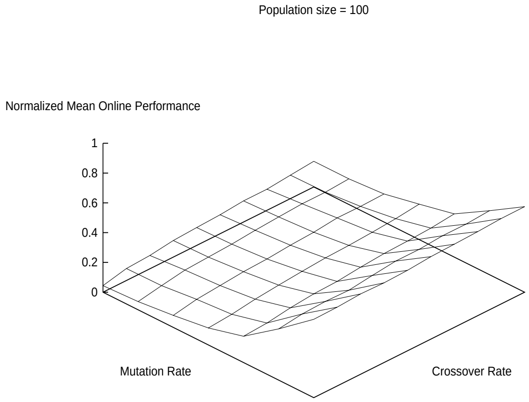

Fig. 5: NMOP (Population Size = 100)

mance (NMOP near 0) occurs when the mutation rate is matched appropriately with population size. For mutation rates greater than the optimal value, the GA performs successively more random search.

5. Conclusions

The results of Section 4 suggest the following meta-level algorithm for choosing optimal population size/mutation rate combinations and obtaining good GA performance:

1. Choose a relatively small population size (10-30).
2. Start with a very small mutation rate (one that will make the GA converge prematurely.
3. Run the GA.
4. When convergence is detected, stop.
5. Increase the mutation rate slightly.
6. Repeat from step 3 until you have exhausted the preset number of function evaluations.

The parameter settings that enable the SGA to perform the best optimization of [Met]-enkephalin's energy potential fall very close the ranges observed by Schaffer for general function optimization [15]. From these results we conclude that we've observed "good" SGA performance on this optimization problem. The results also confirm that choosing a population size, mutation rate, and crossover rate within the ranges specified by Schaffer (10–30, 0.005–0.1, 0.65–0.95) is a good starting point, even for an application with a large string length and many local optima. SGA online performance has been shown to be extremely sensitive to the combined choice of population size and mutation rate. Finally, an algorithm has been proposed that should find good mutation rates with relatively little computational cost when small population sizes are used. The performance of this algorithm will be the subject for future investigation.

References

[1] Brooks, Bernard R. and others. "CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamic Calculations," *Jounal*

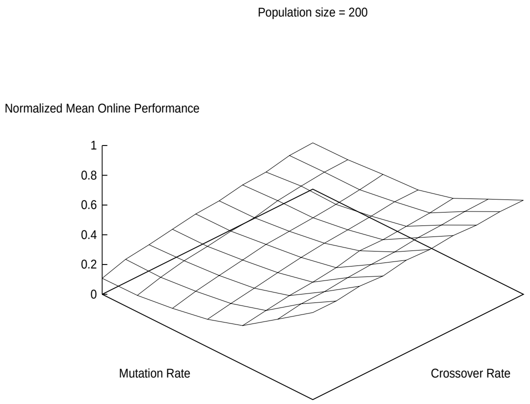

Fig. 6: NMOP (Population Size = 200)

of Computational Chemistry, *4*(2):187–217 (1983).

[2] Chan, Hue Sun and Ken A. Dill. "The Protein Folding Problem," *Physics Today*, 24–32 (February 1993).

[3] DeJong, Kenneth A. *An Analysis of the Behavior of a Class of Genetic Adaptive Systems.*. PhD dissertation, The University of Michigan, Ann Arbor MI, 1975.

[4] Fogarty, Terence C. "Varying the Probability of Mutation in the Genetic Algorithm." In Schaffer [14], 104–109.

[5] Forrest, Stephanie, editor. *Proceedings of the Fifth International Conference on Genetic Algorithms*, San Mateo CA: Morgan Kaufmann Publishers, Inc., July 1993.

[6] Gates, Jr., George H. *Predicting Protein Structure Using Parallel Genetic Algorithms*. MS thesis, AFIT/GCS/ENG/94D-03, Graduate School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH 45433, December 1994.

[7] Goldberg, David E. *Optimal Initial Population Size for Binary-Coded Genetic Algorithms*. Technical Report TCGA Report Number 85001, Department of Engineering Mechanics, University of Alabama,Tuscaloosa AL 35486: University of Alabama, The Clearing House for Genetic Algorithms, November 1985.

[8] Goldberg, David E. "Sizing Populations for Serial and Parallel Genetic Algorithms." In Schaffer [14], 398–405.

[9] Goldberg, David E., et al. "Genetic Algorithms, Noise, and the Sizing of Populations," *Complex Systems*, *6*:333–362 (1992).

[10] Grefenstette, John J. "Optimization of Control Parameters for Genetic Algorithms," *IEEE Transactions on Systems, Man, & Cybernetics*, 122–128 (1986).

[11] LeGrand, Scott M. and Kenneth M. Merz Jr. "The Application of the Genetic Algorithm to the Minimization of Potential Energy Functions," *Journal of Global Optimization*, *3*:49–66 (1991).

[12] Nayeem, Akbar, et al. "A Comparative Study of the Simulated-Annealing and Monte Carlo-with-Minimization Approaches to the Minimum-Energy Structures of Polypeptides: [Met]-Enkephalin," *Journal of Computational Chemistry*, *12*(5):594–605 (1991).

[13] Reeves, Colin R. "Using Genetic Algorithms with Small Populations." In Forrest [5], 92–99.

[14] Schaffer, J. David, editor. *Proceedings of the Third International Conference on Genetic Algorithms*, San Mateo, California: Morgan-Kaufmann Publishers, Inc., June 1989.

[15] Schaffer, J. David. "A Study of Control Parameters Affecting Online Performance of Genetic Algorithms for Function Optimization." In Schaffer [14], 51–60.

[16] Thierens, Dirk and David E. Goldberg. "Mixing in Genetic Algorithms." In Forrest [5], 38–45.

[17] Vásquez, Maximiliano, et al. "Conformational Energy Calculations on Polypeptides and Proteins," *Chemical Reviews*, *94*:2183–2239 (1994).