

Initial Population Diversity and Performance Independence

Pedro A. Diaz-Gomez
Computing & Technology Department
Cameron University
Lawton, OK 73505, USA
pdiaz-go@cameron.edu

Dean F. Hougen
School of Computer Science
University of Oklahoma
Norman, OK 73019, USA
hougen@ou.edu

Abstract

It is widely accepted that there is a strong correlation between diversity in the initial population and Genetic Algorithms' performance. One contribution of this paper is to show that there is not such a strong correlation between diversity and Genetic Algorithms' performance, at least with the standard range of diversity measures used in a randomly generated population and with the misuse detection problem.

1. Introduction

It is commonly believed the quality of a possible solution is likely to be better if the initial population is more diverse [21, 1, 2, 13, 22, 19, 23, 28]. This relationship was tested elsewhere with two extreme problems: the one-max and the snake-in-the-box problems [17]. This paper is going to continue an empirical study about the independence of diversity in the initial population and performance of Genetic Algorithms (GAs) [17], using a third problem that could be considered a problem between the previous ones: the misuse detection problem. The one-max problem is considered an easy problem where each bit contributes to the solution [4, 20, 18, 30]. The snake-in-the-box problem is a difficult problem where the GA has to find a longest connected path that obeys certain constraints in a hypercube [3, 26, 27]. The misuse detection problem could be considered a “modest” problem—it depends on the length of the chromosome—where the GA has to find intrusions, knowing the intrusion profiles [10, 8, 9, 11, 24, 25].

1.1 Performance

In evolutionary computation, performance can be considered as (1) the quality of the solution found in a specific number of generations and (2) the number of generations to find the global maximum when resources are sufficient. Correspondingly, performance in this article has two meanings: (1) the best solution found so far during a given

number of generations, and (2) the number of generations needed to reach the global maximum [17]. Because the global maximum may not always be reached, a maximum of 100,000 generations is used.

1.2 Diversity

As with performance, diversity can be looked at in different ways. One approach is to look at differences at the level of the gene. Another is to look at differences at the level of the entire chromosome [14, 17]. This paper is going to show results using a gene-level metric called *entropy* and defined, for the initial population $P(0)$, as

$$H(P(0)) = \frac{1}{l} \sum_{j=1}^l H_j, \quad (1)$$

where l is the length of the chromosome, H_j corresponds to the entropy as in Equation 2 for locus j of the entire population, with the classical measure of uncertainty given by Shannon [29]

$$H_j = - \sum_{i=1}^N p_{ij} * \log_2 p_{ij} \quad (2)$$

where p_{ij} is the probability of occurrence of independent event $a_{i,j}$, and N is the number of trials (that is the population size in this case).

Equation 2—takes into account the probability of occurrence of 1's and 0's for each locus in the entire population. For example, for 3 bits with values 1, 1, 0, according to equation 2, $H = -(\frac{2}{3} \log_2 \frac{2}{3} + \frac{1}{3} \log_2 \frac{1}{3}) = -(-0.39 - 0.5283) = 0.9183$. In this paper, Equation 2 is applied to each locus j of the entire population. The average is taken, as in Equation 1, to obtain in this way a diversity measure of the initial population, in which values range between 0 and 1, where 1 is the highest diversity value.

The result described in this paper applies to gene and chromosome level diversity as defined elsewhere [14].

2. Experimental Setup

The relationship between diversity and performance is analyzed using the misuse detection problem with two difficulty levels: using chromosome lengths 32 and 64. Each case was tested twice in three trials with different sets of intrusions. The initial population is then generated randomly using a different set of seed for each trial. The GA ran with different initial population diversities, and the rest of the GA's parameters (population size, crossover and mutation probabilities, selection pressure and, stop criteria) remained constant. This set up was used to look for a possible linear correlation between initial population diversity and performance.

2.1 Test Problem

The purpose of the misuse detection problem is to find intrusions in log files. The possible solution is encoded in the chromosome as 1 if there is an intrusion, or 0 if there is no intrusion. For example, if the chromosome length is 32, it means that at most there are 32 different types of intrusions. Each position of the chromosome encodes one [8, 7, 9, 11, 12, 16, 15, 24, 25]. The difficulty of the problem depends on the size of search space 2^l , where l is the length of the chromosome, and in the occurrence of a particular user activity that can cause more than one intrusion. The GA receives as input the initial population, a synthetic vector with the activity performed by a user in one session, and the profiles of intrusions to look for. The output is a 32 chromosome which indicates the possible intrusions [11, 10]

2.2 Test Methods

For each misuse detection case with 32 and 64 chromosome length, 90 initial populations are generated using 90 different seeds. This gives a set of initial populations that have a range of diversity values. The population size is maintained constant with 20 individuals for each case. For the case of a chromosome length of 32, the GA is run for each initial population, and the quality of the best solution found thus far is recorded every 5 generations until generation 20, and after that at each 20 generations until generation 60. For the 64 chromosome length, the GA is run for each initial population, and the quality of the best solution found thus far is recorded every 10 generations until generation 80. Besides looking at the best solution thus far, the algorithm continues until a global maximum is found up to 100,000 generations, recording the number of generations needed to reach the global maximum.

This procedure is repeated three times, giving three repetitions, or *trials* of data with different sets of seeds. The data is then analyzed using entropy and two performance

metrics. The procedure is repeated again with a different synthetic vector, for a total of twelve sets of data.

2.3 Parameters and Parameter Justification

The same set of parameters as in [17] are used for comparison purposes: the population size is 20, 2-tournament selection is used, the probability of uniform crossover is 1.0, the probability of mutation is 0.001 per bit, and a stop criterion of a maximum of 100,000 generations is used.

2.4 Data Analysis Methods

Initial testing showed no apparent curve in the diversity vs. performance graphs, so it appears appropriate to look for a linear correlation [17]. The Pearson's correlation coefficient r_{XY} which gives a positive, negative, or no linear relationship between two variables was used [5]. If $r_{XY} \approx 1.0$ then there is a strong to perfect positive linear correlation, and if $r_{XY} \approx -1.0$ there is a strong to perfect negative linear correlation [5]. But what happens with values in between? The Fisher's r to z transform can be used in conjunction with the Z test in order to check whether the correlation between two factors is near zero, i.e., if the Z value is between the critical values -1.96 and 1.96 [5]. Therefore, the Z test and the Fisher's transform are used in this research as a strong score to show the independence between diversity in the initial population and performance.

3. Experimental Results

As described previously, one diversity metric, two performance metrics, two case problems, and three trials give us a total of twelve sets of results: entropy vs. maximum thus far and entropy vs. number of generations for the intrusion detection problem with 32 bits chromosome length, and; entropy vs. maximum thus far and entropy vs. number of generations for the intrusions detection problem with 64 bits chromosome length.

3.1 Performance on Misuse Detection 32-bits

(1) *Entropy vs. maximum thus far*: No statistically significant positive correlation between the entropy metric and the quality of the solution was discovered, except for trial two with one intrusion, as shown by the Pearson's Coefficients and Z values in Tables 1 and 2 respectively. (See Section Analysis for an analysis of results.)

(2) *Entropy vs. number of generations*: No statistically significant correlation between the entropy metric and the number of generations to reach a global maximum was found. The Pearson's coefficients for one intrusion were -0.020 , -0.190 and 0.164 , with the corresponding Z values of -0.19 , -1.80 and 1.54 ; for five intrusions they were

Trial	Generation					
	0	5	10	15	20	40
1	0.236	0.100	0.053	0.078	0.097	0.125
2	-0.003	0.238	0.581	0.371	0.319	0.150
3	0.087	0.151	-0.077	-0.151	-0.148	-0.173
1	0.039	0.080	0.113	0.081	0.085	0.045
2	-0.035	0.108	0.185	0.198	0.179	0.211
3	-0.179	-0.118	-0.125	-0.064	-0.069	-0.059

Table 1: Pearson’s Coefficients for the Misuse Detection 32-bits. Entropy metric. Top one intrusion, bottom five intrusions. No strong correlation found.

Trial	Generation					
	0	5	10	15	20	40
1	2.24	0.94	0.49	0.73	0.91	1.17
2	-0.025	2.26	6.21	3.64	3.08	1.41
3	0.81	1.42	-0.72	-1.42	-1.38	-1.63
1	0.36	0.75	1.06	0.76	0.79	0.42
2	-0.32	1.01	1.75	1.87	1.67	2.00
3	-1.69	-1.10	-1.71	-0.60	-0.64	-0.55

Table 2: Z values for the Misuse Detection 32-bits. Entropy metric. Top one intrusion, bottom five intrusions. No linear correlation found, *except for one intrusion in trial two that is statistically significant.*

0.160, 0.027 and -0.149 with Z values of 1.51, 0.25 and -1.41 . See Figure 7 for the first trial one intrusion case.

3.2 Performance on Misuse Detection 64-bits

(3) *Entropy vs. maximum thus far:* Table 3 shows Pearson’s coefficients for different runs until generation 80. It can be observed that there is no significant linear correlation between the entropy value of the initial population and the quality of the solution for these test sets. As the Pearson’s values are small, in order to see if the two variables (diversity and solution quality) have correlation zero, Z is presented in Table 4, where no values are statistically significant, confirming the independence between diversity in the initial population and performance.

(4) *Entropy vs. number of generations:* No statistically significant correlations were discovered between the entropy metric and the number of generations to reach the goal. The Pearson’s Coefficients, with three intrusions, for the three trials were -0.059 , 0.105 and -0.027 (small values) and the corresponding Z values were -0.55 , 0.98 and -0.25 , which are between the critical values -1.96 and 1.96 ; and the corresponding values for seven intrusions were 0.047 , -0.077 and 0.079 , with Z values of 0.44 , -0.72 and 0.73 .

Trial	Generation					
	0	10	20	30	40	80
1	0.164	0.065	0.097	0.051	0.038	0.047
2	0.185	0.050	0.100	0.090	0.101	0.049
3	0.008	0.028	-0.021	-0.001	-0.062	-0.067
1	0.008	0.072	0.064	0.028	0.101	0.140
2	0.287	0.119	0.081	0.055	0.052	0.002
3	0.040	0.050	0.133	0.121	0.103	0.106

Table 3: Pearson’s Coefficients for Misuse Detection 64-bits. Entropy vs. maximum so far. Top three intrusions, bottom seven intrusions No significant linear correlation found.

Trial	Generation					
	0	10	20	30	40	80
1	1.54	0.61	0.91	0.48	0.35	0.48
2	1.74	0.46	0.94	0.84	0.95	0.46
3	0.07	0.26	-0.19	-0.01	-0.58	-0.62
1	0.07	0.67	0.60	0.26	0.95	1.31
2	2.75	1.17	0.75	0.52	0.48	0.22
3	0.38	0.47	1.25	1.14	0.97	0.99

Table 4: Z values for Misuse Detection 64-bits. Entropy vs. maximum so far. Top three intrusions, bottom seven intrusions. No significant linear correlation found.

4. Analysis

This paper tested independence between diversity in the initial population and performance for GAs using as a test problem the intrusion detection problem with two sizes of search spaces and with different numbers of intrusions. In Figures 1 through 6, each figure shows the quality of the solution for each diversity value, as well as three lines: two dashed lines that give the boundaries for non-statistically significance and the third line is the regression line. It is observed, for instance, in Figure 1, which corresponds to the first generation, that diversity is beneficial (the regression line has a positive slope and it is outside the boundaries of the upper non statistically significant dashed line), however as the algorithm runs (Figures 2 through 6) there is no-statistical difference between diversity and performance. We are sure that no ceiling effect occurs. The maximum is reached at a quality of solution equal to 1. However, the maximum was not reached in any of the snapshots.

One problem, one diversity metric, and two performance metrics were used. A strong correlation between diversity and performance was not found in any of the specific cases tested, using as test statistics the Pearson’s coefficient and the Fisher’s r to z transform.

The Pearson’s coefficient R_{XY} gives a positive, negative, or no linear relationship between two variables [5]. For the particular research, the two variables are diversity

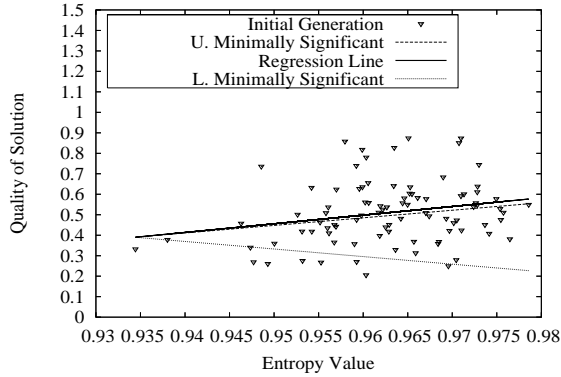


Figure 1: Entropy vs. quality of the solution. Misuse Detection 32-bits. Trial 1. Snapshot at initial generation. 90 Runs.

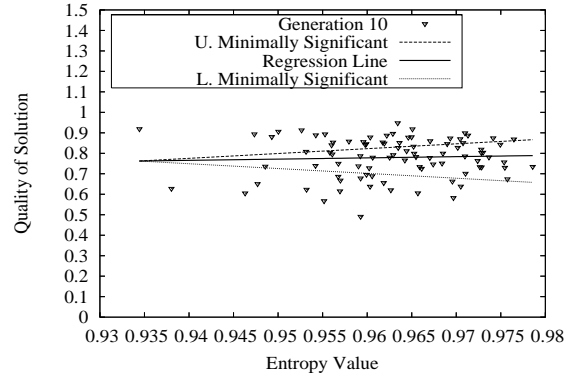


Figure 3: Entropy vs. quality of the solution. Misuse Detection 32-bits. Trial 1. Snapshot at generation 10. 90 Runs.

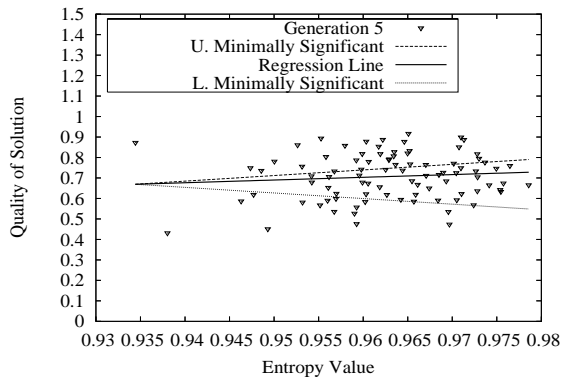


Figure 2: Entropy vs. quality of the solution. Misuse Detection 32-bits. Trial 1. Snapshot at generation 5. 90 Runs.

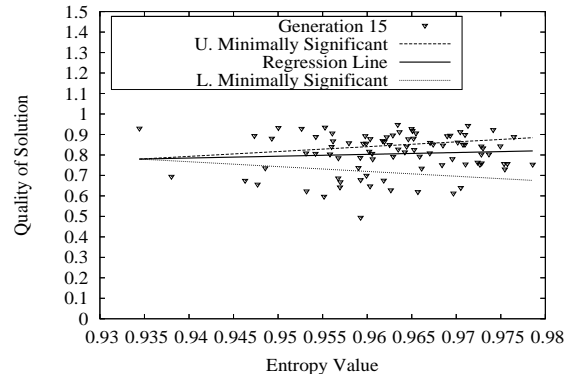


Figure 4: Entropy vs. quality of the solution. Misuse Detection 32-bits. Trial 1. Snapshot at generation 15. 90 Runs.

(X) and performance of a GA (Y). If $R_{XY} \approx 1.0$ then there is a strong-to-perfect positive linear correlation, and if $R_{XY} \approx -1.0$, there is a strong-to-perfect negative linear correlation [5]. None of these values were reached for any of the specific hypotheses, as can be seen in Tables 1 and 3. However, to be more certain about the independence of the two variables X and Y , the Fisher's r to z transform and the Z test were applied, the results of which are in Tables 2 and 4. In all cases, the correlation was almost zero, except for the case with one intrusion in trial two, a situation that could occur by chance.

5. Conclusions & Future Work

Previous study [17] about the independence of diversity in the initial population and performance of GAs was complemented in this paper, with a third problem of moderate difficulty: the misuse detection problem. No strong corre-

lation between initial population diversity and performance was found, corroborating the finding in [17]. But why, in general, is it widely believed that diversity in the initial population is beneficial for Evolutionary Computation? [6] addressed this topic using the Iterated Prisoner's Dilemma (IDP) game as a problem to test the interaction between diversity in the initial population and performance. They showed that apparently a higher diversity was not beneficial for solving the IDP problem and that a correlation between diversity and quality of the solution can be misleading [6].

Population size is usually referred to diversity [13]. The higher the population size, the more diverse a population is thought to be. However this is not always the case. For example, if an initial population is duplicated with the same individuals, its entropy value is going to be the same. This leads us to propose future experiments changing the population size but maintaining the same diversity value and looking for a possible linear correlation between population size and GAs performance.

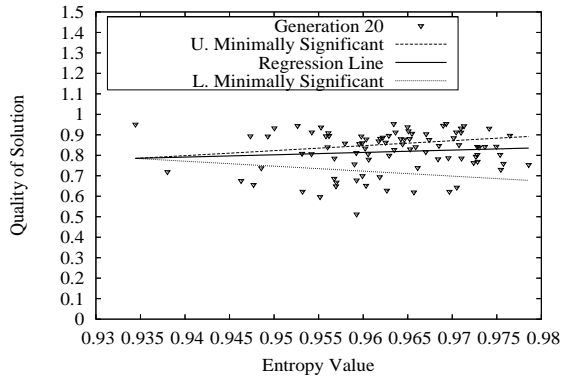


Figure 5: Entropy vs. quality of the solution. Misuse Detection 32-bits. Trial 1. Snapshot at generation 20. 90 Runs.

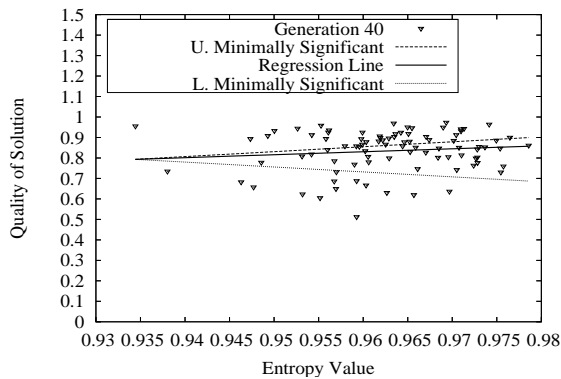


Figure 6: Entropy vs. quality of the solution. Misuse Detection 32-bits. Trial 1. Snapshot at generation 40. 90 Runs.

6 Acknowledgements

Thanks to Dr. Mike Estep and Mr. Mark Polson, from Cameron University, for proof reading this paper.

References

- [1] D. S. Bitterman. New lower bounds for the snake-in-the-box problem: A prolog genetic algorithm and heuristic search approach, 2004. Master Thesis accessed Jun. 2007.
- [2] E. K. Burke, S. Gustafson, and G. Kendall. Diversity in genetic programming: An analysis of measures and correlation with fitness. *IEEE Transactions on Evolutionary Computation*, 8(1):47–62, 2004.
- [3] D. A. Casella and W. D. Potter. New lower bounds for the snake-in-the-box problem: Using evolutionary techniques to hunt for snakes. In *Proceedings of the Florida Artificial Intelligence Research Society Conference*, pages 264–268, 2004.

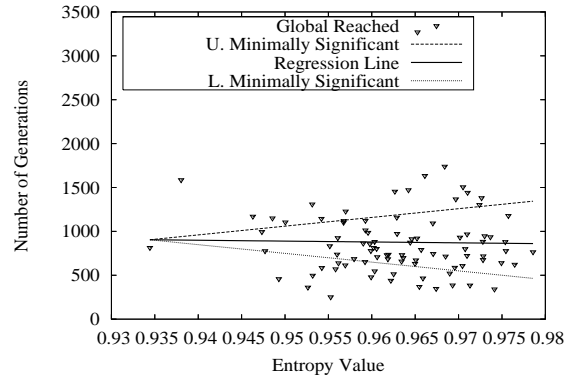


Figure 7: Entropy metric vs. number of generations to reach the global maximum. 100,000 Generations. Misuse Detection 32-bits. 90 Runs.

- [4] C. D. Cheng and A. Kosorukoff. Iterative one-max problem allows to compare the performance of interactive and human-based genetic algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 983–993, 2004.
- [5] P. R. Cohen. *Empirical Methods for Artificial Intelligence*. The MIT Press, 1995.
- [6] P. J. Darwen and X. Yao. Eebic group, 1999.
- [7] P. A. Diaz-Gomez and D. F. Hougen. Analysis and mathematical justification of a fitness function used in an intrusion detection system. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1591–1592, 2005.
- [8] P. A. Diaz-Gomez and D. F. Hougen. Analysis of an off-line intrusion detection system: A case study in multi-objective genetic algorithms. In *Proceedings of the Florida Artificial Intelligence Research Society Conference*, pages 822–823, 2005.
- [9] P. A. Diaz-Gomez and D. F. Hougen. Further analysis of an off-line intrusion detection system: An expanded case study in multi-objective genetic algorithms. In *Proceedings of The South Central Information Security Symposium*, 2005.
- [10] P. A. Diaz-Gomez and D. F. Hougen. Improved off-line intrusion detection using a genetic algorithm. In *Proceedings of the International Conference on Enterprise Information Systems*, pages 66–73, 2005.
- [11] P. A. Diaz-Gomez and D. F. Hougen. A genetic algorithm approach for doing misuse detection in audit trail files. In *Proceedings of the International Conference on Computing*, pages 329–335, 2006.
- [12] P. A. Diaz-Gomez and D. F. Hougen. Three approaches to intrusion detection: Analysis and enhancements. In *Proceedings of the National Computer and Information Security Conference*, 2006.
- [13] P. A. Diaz-Gomez and D. F. Hougen. Empirical study: Initial population diversity and genetic algorithm performance. In *In Proceedings of the Conference on Artificial Intelligence and Pattern recognition*, 2007.
- [14] P. A. Diaz-Gomez and D. F. Hougen. Initial population for genetics algorithms: A metric approach. In *Proceedings of the International Conference on Genetic and Evolutionary Methods*, 2007.

- [15] P. A. Diaz-Gomez and D. F. Hougen. MISUSE DETECTION: A neural network vs. a genetic algorithm approach. In *Proceedings of the International Conference on Enterprise Information Systems*, 2007.
- [16] P. A. Diaz-Gomez and D. F. Hougen. MISUSE DETECTION: An iterative process vs. a genetic algorithm approach. In *Proceedings of the International Conference on Enterprise Information Systems*, 2007.
- [17] P. A. Diaz-Gomez and D. F. Hougen. Initial population diversity does not influence performance. In *Proceedings of the International Conference on Genetic and Evolutionary Methods*, 2008.
- [18] P. Giguere and D. E. Goldberg. Population sizing for optimum sampling with genetic algorithms: A case study of the onemax problem. In *Proceedings of the Third Annual Genetic Programming Conference*, 1998.
- [19] J. J. Grefenstette. Optimization of control parameters for genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-16(1):122–128, 1986.
- [20] G. R. Harik and F. G. Lobo. A parameter-less genetic algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 258–265, 1999.
- [21] Z. M. Jaroslaw Arabas and J. Mulawka. GAVaPS—a genetic algorithm with varying population size. In *Proceedings of the IEEE International Conference on Evolutionary Computation*, pages 73–78, 1995.
- [22] F. G. Lobo and C. F. Lima. A review of adaptive population sizing schemes in genetic algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 228–234, 2005.
- [23] N. E. McPhee and N. Hopper. Analysis of genetic diversity through population history. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1112–1120, 1999.
- [24] L. Mé. Security audit trail analysis using genetic algorithms. In *Proceedings of the International Conference on Computer safety, reliability, and Security*, pages 329–340, 1993.
- [25] L. Mé. GASSATA, a genetic algorithm as an alternative tool for security audit trail analysis. In *Proceedings of the First International Workshop on the Recent Advances in Intrusion Detection*, 1998.
- [26] W. D. Potter, R. W. Robinson, J. A. Miller, K. Kochut, and D. Z. Redys. Using the genetic algorithm to find snake-in-the-box codes. In *Proceedings of the 7th International Conference on Industrial & Engineering Applications of Artificial Intelligence and Expert Systems*, pages 421–426, 1994.
- [27] D. S. Rajan and A. M. Shende. Maximal and reversible snakes in hypercubes. In *Proceedings of the Australasian Conference on Combinatorial Mathematics and Combinatorial Computing*, 1999.
- [28] J. P. Rosca. Entropy-driven adaptive representation. In *Proceedings of the Workshop Genetic Programming: From Theory to Real-World Applications*, pages 23–32, 1995.
- [29] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.
- [30] T.-L. Yu, K. Sastry, D. E. Goldberg, and K. Sastry. Optimal sampling and speed-up for genetic algorithms on the sampled onemax problem. Technical report, Illinois Genetic Algorithms Laboratory, University of Illinois, 2003.