Characterizing a Tunably Difficult Problem in Genetic Programming

Omer A. Chaudhri, Jason M. Daida, Jonathan C. Khoo, Wendell S. Richardson, Rachel B. Harrison, and William J. Sloat

The University of Michigan, Artificial Intelligence Laboratory and Space Physics Research Laboratory
2455 Hayward Avenue, Ann Arbor, Michigan 48109-2143, daida@eecs.umich.edu, (734) 647-4581, (734) 764-5137 FAX

Abstract

This paper examines the behavioral phenomena that occur with the tuning of the binomial-3 problem. Our analysis identifies a distinct set of phenomena that may be generalizable to other problems. These phenomena also bring into question whether GA theory has any bearing on GP theory.

1 INTRODUCTION

A common assumption in genetic programming theory is that phenomena that occur under genetic algorithms (GA) have roughly analogous counterparts in genetic programming (GP). After all, genetic programming is a derivative from genetic algorithms. Many of their procedural steps are similar—indeed, GP is more alike GA than is different. Where the two differ lies in what is seemingly a secondary issue: that GP (often) uses a tree representation to structure information rather than, say, a linear representation. Given the various implementations of GP and GA, this "difference" is even less clear, since some GP systems use a linear representation and some GA systems use a tree-like representation. Consequently, given this degree of similarity between GA and GP, the assumption of analogous phenomena is understandable.

Is the assumption correct? We would claim perhaps, under some circumstances. If we took the position, however, that phenomena that occur under GP have *no* counterpart in GA, we would need to re-examine just what phenomena does occur under GP. If the phenomena were found to be substantially different, it would beg the issue of whether theory in GA has *any* bearing on GP theory.

This paper re-examines phenomena that occur under GP as a problem is tuned for difficulty. (For this paper, we define *phenomena* as behavioral occurances, patterns, that are observed in the output from a GP system.) Although the output we obtain is strictly static (i.e., non-time varying and represents but a single snapshot of performance for a particular trial), by tuning and observing the output, we can infer something about GP's internal dynamics. (The process we use is akin to finding an unknown transfer function to a black box.)

For this paper then, we examine in detail the behavioral phenomena that occurs with the tuning of a *single* problem. Our investigation represents a longitudinal study with the purpose of understanding GP *dynamics*. This is in contrast to examining the behavioral phenomenon that occurs under *multiple* problems. Multiple problems represent a latitudinal approach to understanding *gross characteristics* of GP. For example, latitudinal studies, like [Luke and Spector 1998], have been used to discuss efficacy of methods (e.g., crossover v. mutation). Longitudinal studies complement latitudinal ones. This paper represents one of the few longitudinal studies in our field. (For another, see [Langdon, 1998]).

The paper is organized in the following manner. Section 2 provides a description of the experiment as well as the binomial-3 problem. Section 3 presents the results collected from our experiments. Section 4 discusses the several different phenomena apparent in GP. Section 5 concludes the paper.

2 EXPERIMENT DESCRIPTION

2.1 BINOMIAL-3 PROBLEM DESCRIPTION

The binomial-3 problem is an instance taken from symbolic regression and involves solving for the function $f(x) = (1 + x)^3 = 1 + 3x + 3x^2 + x^3$. The term *binomial* refers to the sequence of coefficients in this polynomial; the "3" refers to the order of this polynomial.

We define the binomial-3 problem as follows. Fitness cases are 50 equidistant points generated from the equation $f(x) = (1 + x)^3$ over the interval [-1, 0). Raw fitness score is the sum of absolute error. A hit is defined as being within 0.01 in ordinate of a fitness case for a total of 50 hits. The stop criterion is when an individual in a population first scores 50 hits. Adjusted fitness is the reciprocal of the quantity one plus raw fitness score.

A function set is a subset of {+, –, ×, ÷}, which corresponds to arithmetic operators of addition, subtraction, multiplication and protected division. We define protected division as the operator that returns one if the denominator is exactly zero. Typical function sets include {+, –, ×, ÷}, which we presume for this paper. Other sets may include other permutations such as {+, ×} or {–, ×}.

A terminal set is a subset of {X, R}, where X is the symbolic variable and R is a set of ephemeral random constants (ERCs). We presume that the ephemeral random constants are uniformly distributed over a specified interval of the form $[-a_R, a_R]$, where a_R is a real number that specifies the range for ERCs. We require that each ERC is generated but once at population initialization and *is not changed in value during the course* *of a GP run.* Typical terminal sets include either {X} (a binomial-3 problem without ERCs) or {X, R} (a binomial-3 problem with ERCs).Tuning is achieved by varying the value associated with a_R.

2.2 BINOMIAL-3 PROBLEM BACKGROUND

The binomial-3 problem shares many properties that are common to other problems in GP. It requires symbol manipulation. It also allows for *nocs* (i.e. non-coding segments, also known as *introns* or unexpressed code). The problem affords GP to choose from multiple approaches to solve for the same problem. For example, equivalent solutions include $(1 + x)^3$, $(1 + x)(1 + 2x + x^2)$, $(x - -1)^3$ and $(x + 1) \div (1 \div (1 + (x \div 0.5) + (x \div (1 \div x)))$. In addition to these equivalent approaches, there exists a number of approximate approaches (e.g. rational polynomials that fit all 50 points, but not necessarily anywhere else). Furthermore, there are several ways to generate numerical coefficients. For example, the coefficient 2 can be generated by using an ERC that (approximately) equals this value. It can be generated with the value 0.5 and taking the reciprocal of that value. It can also be generated through distribution, e.g., $(x + x)$. We surmise that the total number of ways to solve the binomial-3 problem to be on the order of a few thousand (i.e., see [Daida, et al. 1999]).

The choice of coefficients, form, and order of the target function $f(x)$ for the binomial-3 problem was purposeful and deliberate. The use of $f(x) = (1 + x)^3$ has allowed for an extended mathematical treatment [Daida, Bertram et al. 1999].

The binomial-3 problem does not share an antecedent with a related test problem in GA research, but its domain has an extended history in GP. One of the earliest, intuitive applications of GP has involved data modeling under the moniker of symbolic regression. See [Koza 1992] and also [Banzhaf et al. 1998].

In spite of these works, we recognize that from a purely practical standpoint, there exist modifications to standard GP that may be better suited for data modeling. This seems to have been particularly true in the generation of parameter constants, which standard GP does awkwardly with ERCs. Recent developments in GP indicate methods that appear to generate constants with greater efficacy than as with using ERCs (e.g., [Angeline 1996; Raidl 1998]).

Our interest in using the binomial-3 problem has been to illustrate behavioral phenomena that may be reflective of the class of problems subsumed under data modeling. ERCs can also be used to address building block issues, as well as fitness landscapes. See [Daida, Polito et al., 1999; Daida, Bertram et al. 1999].

For examples of other tunably difficult problems, please see [Soule et al. 1996; Punch et al. 1996].

2.3 EXPERIMENT PROCEDURE

We used a patched version of lilgp [Zongker and Punch 1995] to generate our data as discussed in[Daida, Bertram, et al. 1999].

The GP parameters were similar to those mentioned in Chapter 7 [Koza 1992]: population size = 500; crossover rate = 0.9; replication rate = 0.1; population initialization with ramped half-and-half; initialization depth of 2-6 levels; and fitness-proportionate selection. Other parameter values were maximum generations = 200 and maximum tree depth = 26.

We ran two sets of experiments; the first involved varying the tuning parameter a_R, which indicates the range of R in the form of $[-a_R, a_R]$. We used 15 such values of a_R, where a_R = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 10, 100, 1000. Fifteen data sets were collected in all. Each data set consisted of 600 trials for a total of 9,000 runs for the first set. (Note: results for a_R = 0.1, 1, 2, 3, 10, 100, 1000 were reported in [Daida, Polito et al., 1999, 2000].)

The second set of experiments included optimized sets of R. In these experiments, we narrowed the number of ERCs to one or two specified values. In particular, we ran experiments for R = {-1}, {1}, {0}, {-1, 1}. For R = {-1, 1}, we set the probability for choosing either -1and 1 as equally likely. In this set of experiments, too, we also ran one control with no ERCs. Five data sets were collected in all. Each data set also consisted of 600 trials for a total of 3,000 runs.

Taken with the other experiments, there were 12,000 runs total. All trials were run on Sun Ultra workstations.

3 RESULTS

Figure 1 illustrates how problem difficulty changes over the range $0 < a_R \leq 1000$. The *x*-axis denotes the value of the tuning parameter a_R. The *y*-axis denotes the percentage of the total number of trials (i.e., 600) with a score that fell in either perfect, the upper decile, or the upper quartile of the total number of hits (i.e., at least 50, 45, 38 hits or better, respectively). For example, upper-decile coordinates of (100, 5) means that 5% out of 600 trials scored at least 45 hits for the tuning parameter a_R =100. Note that there is a performance peak at a_R =1 (also referred to as the Unity data set, or simply Unity) and as a_R diverges from Unity, the performance drops rapidly.

Figures 2 and 3 summarize the results from the first set of experiments (i.e., values of a_R that belong to the set {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 10, 100, 1000}). Figure 4 summarizes the results from the second set of experiments (i.e., Control and ERC-specific experiments: R = {-1}, {1}, {0}, {-1, 1}). Each plot shows 600 points, with each point corresponding to a best-of-trial individual. Rows are arranged by data set.

For the second and third columns in Figures 2, 3 and 4, we added a small amount of uniform random noise to both (*x*, *y*) coordinates of each point. We did this for visualization only. The quantities corresponding to node count, depth, and generation are integer values—because of this, a single dot could correspond to many data points. The noise was added to displace points visually away from each other. That technique was not repeated for the first column, only because adjusted fitness is a real-valued quantity.

The first column in Figures 2, 3 and 4 shows the effect of tuning concerning node versus adjusted fitness. When a_R pro-

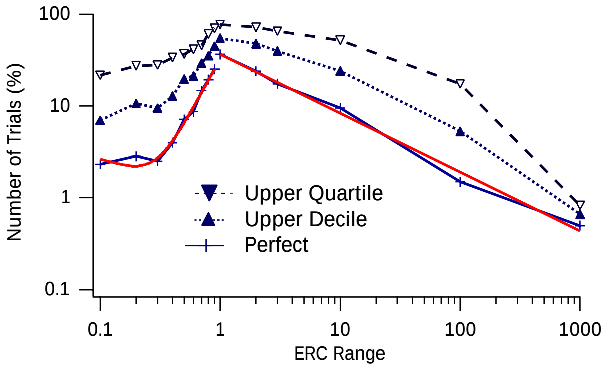


Figure 1 Hit Score v. a_R. This log-log plot shows the relationship between the tuning parameter a_R and the hit score.

gressed from 0.1 to 1, a horizontal cloud of points moved to a vertical cluster, i.e., from lower to higher fitness. When a_R progressed from 1 to 1000, the cluster of points shifted from right to left, i.e. , from higher to lower fitness. In Figure 4, the large solid vertical line indicates those best-of-trial individuals that had perfect adjusted fitness score.

The second column in Figures 2, 3 and 4 shows the effect of tuning concerning node count versus the generation in which a best-of-trial individual was found. Generally, the earlier an individual was identified in a trial, the smaller that individual was and the less computational effort it took to evaluate that individual. Note that for $a_R \neq 1$ in Figures 2 and 3, an increased density of points can be found to the right (i.e., at higher generation numbers). The exception was at a_R = 1,000 which shows none of these patterns. Figure 4 shows that most of the best-of-trial individuals were discovered in earlier generations.

The last column in Figures 2, 3, and 4 shows the effect of ERC range concerning node count versus the depth of the best-of-trial individuals. The lines indicate the upper and lower limits for the numbers of nodes that can be present in a tree for a certain depth. In general, the pattern of clusters appears to move toward the right (i.e., greater depths), with the exception of a_R = 1000, R={0}, and Control, which appears to have their best-of-trial individuals uniformly distributed over that region. Remaining ERC-specific experiments indicate their best-of-trial individuals are typically less deep.

Figure 5 has three separate plots that correlate the frequency of trials greater than a given adjusted fitness score. Figure 5a plots the results for $a_R \geq 1$, and Figure 5b does the same for values of $a_R \leq 1$. Figure 5c plots the results for Control and ERC-specific results.

Figure 6 shows a histogram of all 26 million ERC values in Unity (i.e., 30,000 individuals at 500 individuals per population snapshot per trial for 600 trials). Each population snapshot was taken when a best-of-trial individual was identified.

4 DISCUSSION

The following are seven phenomena that occurred in the course of adjusting problem difficulty for the binomial-3 problem. We describe some of the phenomena in terms of content and context. For this paper, the term *content* refers to the information contained in or pointed to by a node (i.e., a node points to a particular element in either function or terminal sets). The term *context* refers to an ordered and labelled subtree in which a node is a member.

- Problem difficulty was driven by varying content and not increased combinatorial search space
- Cumulative measures for adjusted fitness were bounded by a_R = 1 (for the most part)
- Various attractors appeared during tuning (content- and context-dependent behavior)
- Difficulty influenced the size of individuals, as well as the generation in which a best-of-trial individual is identified (for the most part)
- Specific shape trends persisted in spite of tuning: content- and context-free behavior
- Optimal ERC value was found in tuning, not in ERC distributions
- Domains of distinct behavior occurred

Difficulty was Driven by Varying Content

Problem difficulty was driven by content and not by changes in combinatorial search space. This was originally discussed in [Daida, Polito 1999, 2000], which noted that the combinatorial search space for the binomial-3 problem is statistically invariant despite changing a_R. This observation continued to hold true for the extended range of tuning described in this paper.

This type of tuning is particular to GP, inasmuch as the binomial-3 problem does not share an antecedent in GA research.

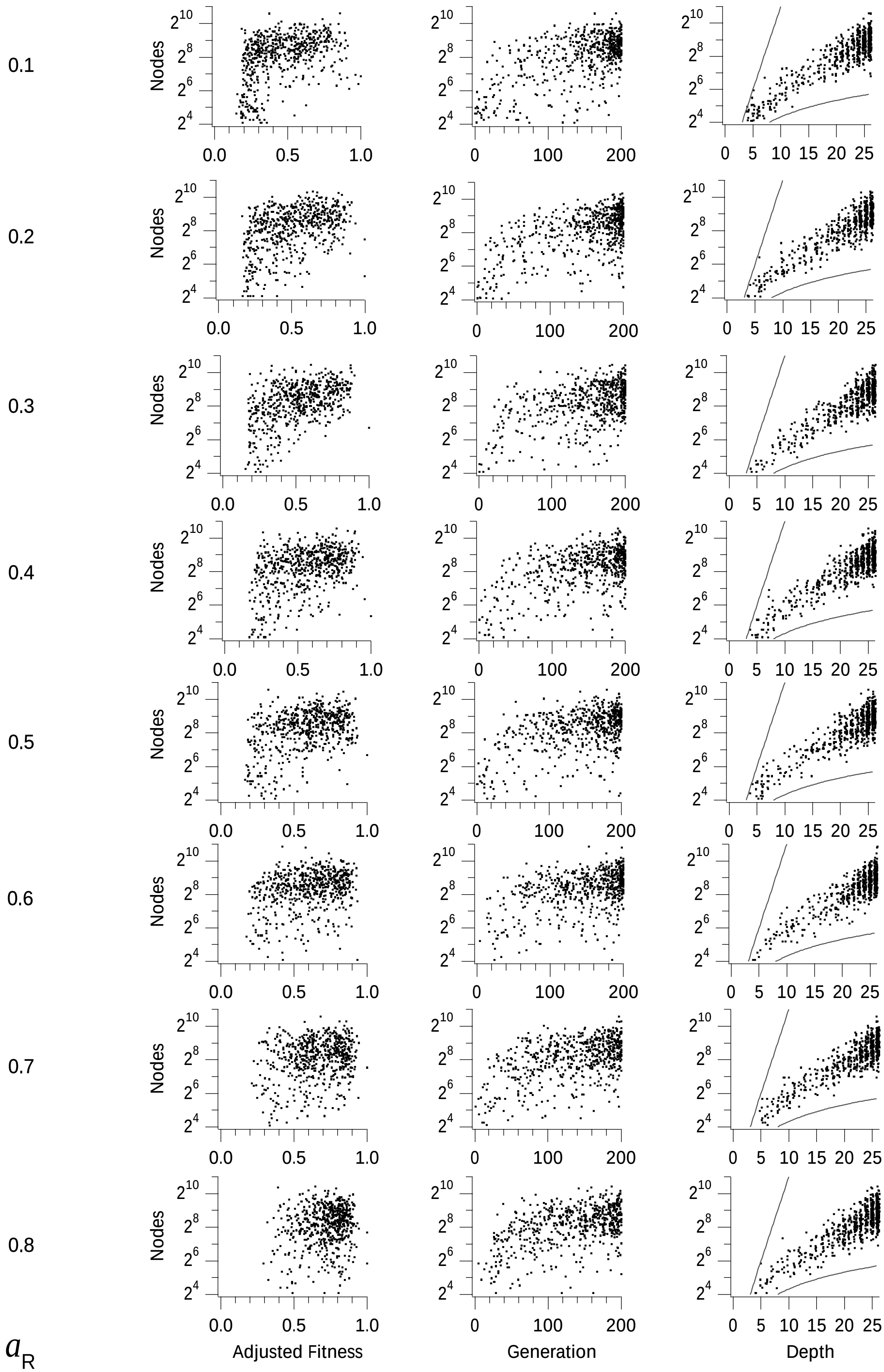
Cumulative Measures of Adjusted Fitness were Bounded

Note that for the first part of the experiment, that the cumulative adjusted fitness measures for $a_R < 1$ and $a_R > 1$ were bounded by the curve for a_R = 1 (as shown in Figures 5a and 5b). Even curves with small deviations from a_R lay completely within bounds described by the curve for a_R = 1. Generally, as a_R diverges from 1, the culmulative distribution curves became more distant from the curve for a_R = 1, further supporting a_R = 1 as the optimal parameter for the binomial-3 problem in the given tuning range. Note that there was a class of exceptions (which is described later in this section).

Various Attractors Appeared During Tuning

Intuitively, we would expect the cluster of points to approach the adjusted fitness score of 1, as GP applies selection pressure for individuals to correctly solve the binomial-3 problem . (See Column 1 of Figures 2, 3, and 4.) Surprisingly, other distinct attractors emerged in our results, which was apparent among the symmetric ERC ranges. For values of a_R that were close to 1, a cluster of points localized around an adjusted score of 0.8 and not 1.0. This localization was also evident in ERC-specific results. For each of them, a significant density of points surrounded this apparent attractor. In some experiments, yet another attractor was apparent at 0.2 for data sets whose val-

Figure 2. Best-of-Trial Results. For $a_R = 0.1, 0.2, \ldots, 0.8$



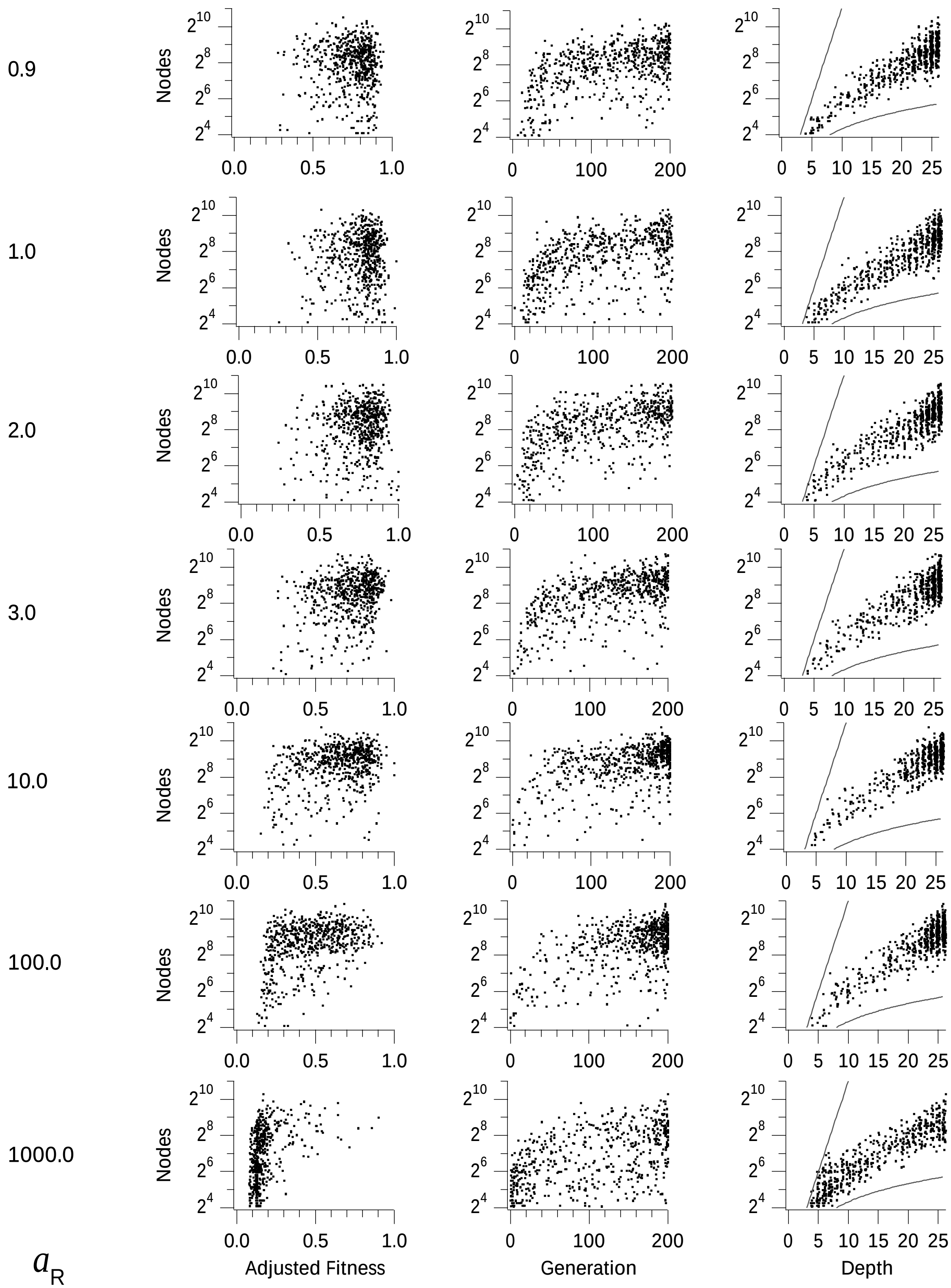


Figure 3. Best-of-Trial Results. For a_R = 0.1, 1, 2, 3, 10, 100, 1000

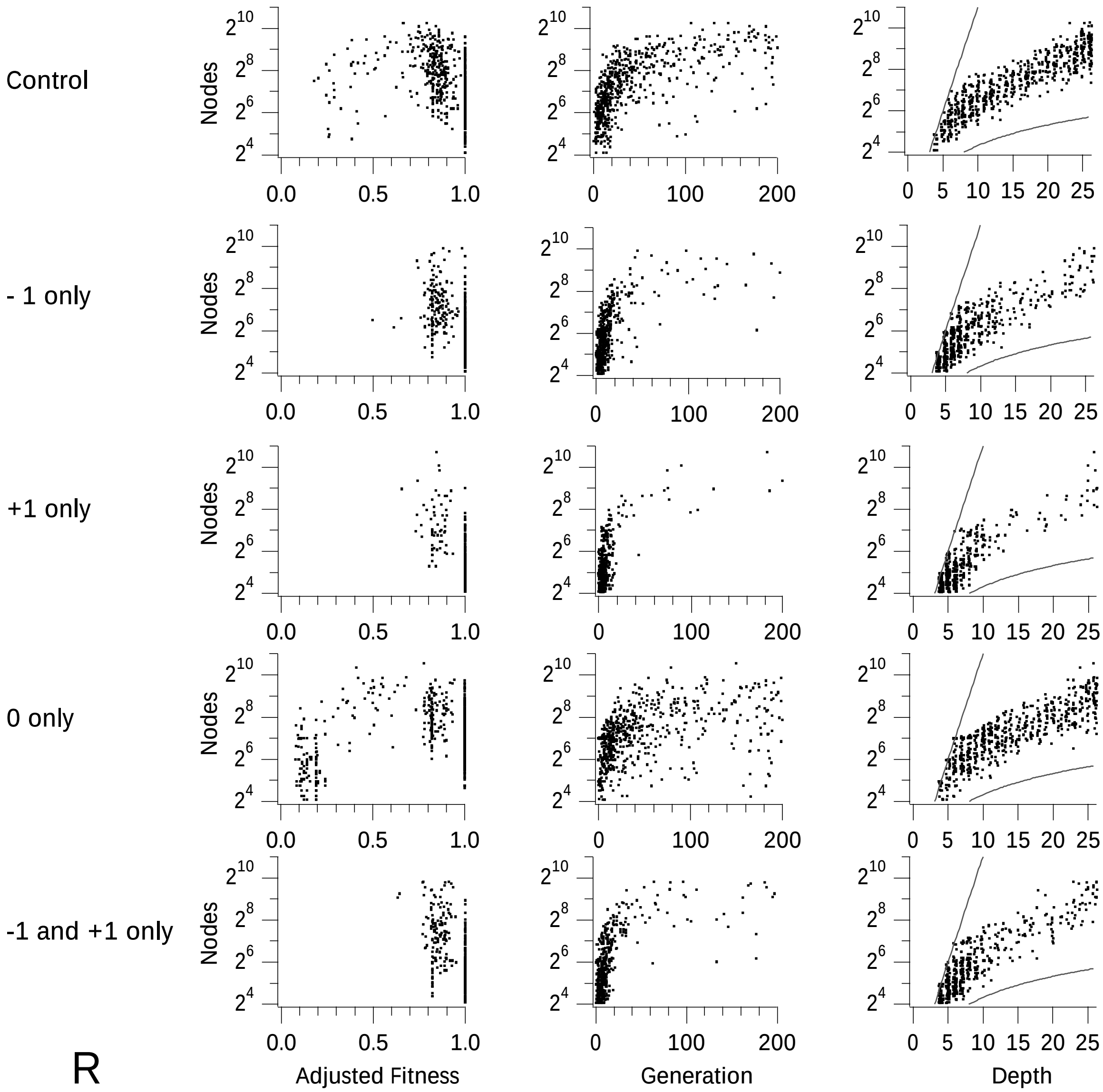


Figure 4. Best-of-Trial Results. For Control, R = {-1}, {1}, {0}, {-1, 1}

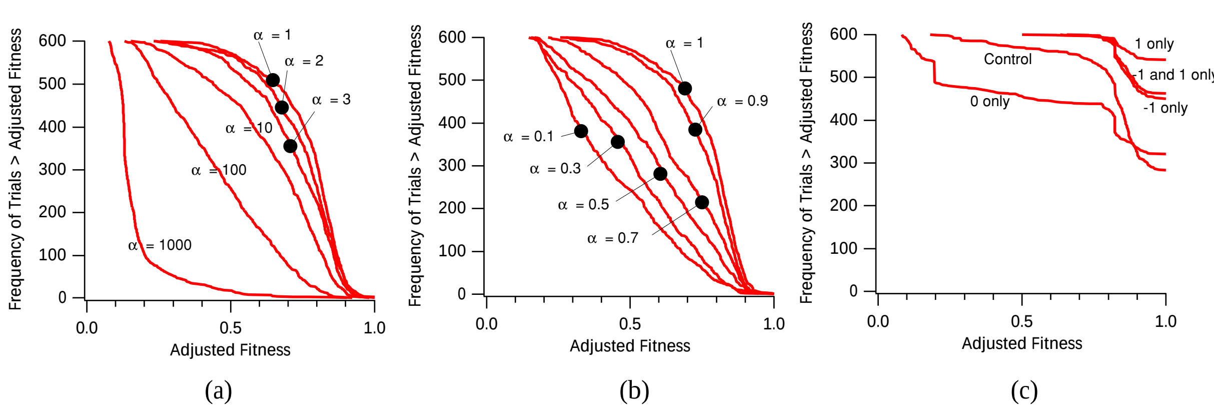


Figure 5. Culmulative Distributions for Adjusted Fitness

ues of a_R are not proximal to 1 (i.e. a_R = 0.1, 0.2, 100, and 1000). This evidence would suggest that such behavior is indicative of content acting negatively on GP performance.

We note that context- and content-dependent behavior was anticipated in [O'Reilly and Oppacher 1995]. Context- and content-dependent behavior are somewhat addressed in theory concerning GAs and epistasis, but not to the degree that O'Reilly and Oppacher suggested.

Difficulty Influenced Individual Size (for the most part)

Greater computational effort often indicates increased problem difficulty, which was evident as the size of a solution generally increased as the difficulty of the binomial-3 problem increased. This behavior occurred for many results shown in Figures 2 and 3 (Column 2), with the exception of a_R =1000. In Figure 4, there was some consistency with higher generations producing larger individuals, but primarily we noted that the opposite situation occurred. These experiments corresponding to Figure 4 demonstrated convincingly that GP can be treated as a simple problem with minimal computing time. For R = {-1}, {+1}, and {-1, 1}, nearly all individuals were produced in the first several generations at relatively small sizes. For Control and R={0}, the trends were not as well defined, and so it can be argued that Control and R={0} increased problem difficulty when compared to other ERC-specific experiments. However, they decreased problem difficulty when compared to the lot of symmetric ERC ranges.

Specific Shape Trends Persisted in spite of Tuning

In the third column of Figures 2, 3, and 4, we note that GP consistently produced a pattern that suggests upper and lower bounds for plots of nodes v. depth. Although the distribution of points within these bounds changed with respect to tuning, the bounds themselves did not seem to change much. This strongly suggests that these bounds were invariant to tuning, which further implies that there were context- and content-free mechanisms that limit individual size and shape. What has been surprising has been the robustness of this shape with respect to tuning.

Langdon, Soule et al. (1999) argued that simple random drift accounts for behaviors like these. Their claim has merit, inasmuch as Flajolet and Oldyzko's work (1982) on binary trees (in general) appears to apply to their GP data.

If Langdon, Soule et al.'s (1999) assertion proves true, it would mean that in spite of an expansive search space of combinatorial possibilities, there are theoretical limits of just how much of that space can be explored for *any* problem using standard GP. Given the relative positions of upper and lower boundaries, their assertion also carries the implication that just a fraction of this space can be searched.

This type of growth is different from that assumed in GA theory, since it would imply that random drift substantially limits search to a fraction of the total number of possibly attainable solutions.

Optimal Value Identified in Tuning, not in ERC Distribution

We note that the optimal tuning parameter (a_R = 1, and consequently, the best ERC building block) was not apparent in the ERC distributions (Figure 6, from [Daida, Polito 2000]),

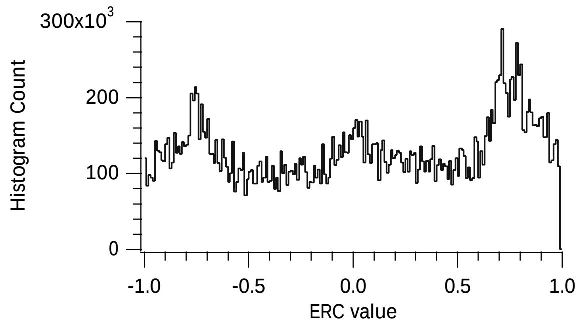


Figure 6. Distributions of ERCs for a_R=1. From [Daida, Polito 2000]

but in the tuning curves for binomial-3 difficulty (Figure 1). The purpose of visualizing the ERC distributions was to observe what frequency of ERC values existed among individuals at the end of a GP run. The local maximums at ±0.75 ran counter to initial expectations of greater densities around the optimal parameter values of ±1.0. It appeared that selection for those ERC values was not as strong as anticipated; perhaps other internal factors contributed to the distribution. However, when we observed how symmetric ERC ranges affected GP performance, we note the following: clearly, as a_R diverged from 1, the percent of individuals with perfect adjusted fitness scores decreased dramatically. There was no evidence to suggest any increase in performance for values of a_R >1000. Therefore, a_R = 1 was likely the global maximum among symmetric ERC ranges, suggesting an optimal value for GP in this tuning range.

We believe that for other symbolic regression problems, there could exist other optimal tuning ranges or optimal sets of parameter values. We do not suggest that for all regression problems that the particulars of the phenomena that we have identified here would hold. For example, the ERC range of [-1, 1] is suitable for the binomial-3 problem; it would probably be a different range for a different problem. However, we do suggest there is a correlation between optimal parameter values and the roots of a regression problem, further indicating that for distinct multiple roots, there may very well be multiple maximums.

This is an unexpected result with respect to GA theory. GA theory, particularly those concerned with schema theorems, infers that selection would sift for best values and that these best values would be retained via crossover. Consequently, ERC frequency distribution curves should show the greatest frequency counts for optimal ERC values. Instead they did not. Rather, *the optimal value of 1 showed up in the difficulty curves as being the easiest* a_R.

Domains of Distinct Behavior Occurred

The symmetric ERC ranges and ERC-specific cases demonstrated very distinct behaviors, suggesting two different strategies that were being employed to solve for the binomial-3 problem (see also [Daida, et al. 2000]). The ERC-specific experiments yielded smaller individuals at earlier generations, whereas symmetric ERC ranges typically yielded larger individuals at later generations. We also observe separate behaviors in Figure 5, as there were potentially different limiting factors. If we were to transpose the results in Figure 5c on to both 5a and 5c, we would find that the ERC-specific experiments were clearly not bounded by the culmulative distribution curve corresponding to $a_R = 1$.

It was possible to further tune the difficulty of the binomial-3 problem by judiciously selecting the components of which GP could use. The results shown in this paper suggest that just the content of the terminal sets (which are particular to GP) are likely of particular significance.

5 CONCLUSION

Is the assumption that GP phenomena is similar to that found in GA a correct assumption? The phenomena discussed in this paper suggested that although it is possible that this assumption is correct, there were several significant behaviors that do not have a clear antecedent in GA theory. For example, there was evidence that suggests performance characteristics were largely dependent on context and content. However, there was equally substantial evidence that related behavioral patterns to context- and content-free factors. Although we clearly demonstrated how significant alterations in content can dramatically impact GP performance, there were consistent and specific shape trends that constrained just how much of the search space was considered. Apparently, at least two mechanisms in GP exist in tension—a context and content-free mechanism, as well as a context- and content-dependent mechanism.

We have noted that a practical implication of our work is that it pays to optimize for the terminal (and function) set. The gains were substantial, which suggests that optimizing the components that GP would use is a crucial preparatory step.

Finally, we suggest that although GA and GP share many similar processes, the theory behind each may be quite different.

ACKNOWLEDGMENTS

We gratefully acknowledge the other group members who have contributed to the binomial-3 study: S. Stanhope, J. Polito 2, R. Bertram, S. Chaudhary, S. Yalcin, P. Litvak, D. Ampy, M. Ratanasavetavadhana, and H. Li.

REFERENCES

Angeline, P. J. (1996). "An Investigation into the Sensitivity of Genetic Programming to the Frequency of Leaf Selection During Subtree Crossover." In J. R. Koza, D. E. Goldberg, D. B. Fogel, and R. L. Riolo (Eds.), *Genetic Programming 1996: Proceedings of the First Annual Conference: July 28–31, 1996*, Stanford University. Cambridge: The MIT Press. pp. 21–29.

Banzhaf, W., P. Nordin, et al. (1998). *Genetic Programming: An Introduction: On the Automatic Evolution of Computer Programs and Its Applications.* San Francisco: Morgan Kaufmann Publishers, Inc.

Daida, J. M., R. B. Bertram, et al. (1999). "Analysis of Single-Node (Building) Blocks in Genetic Programming." In L. Spector, W. B. Langdon, U.-M. O'Reilly, and P. J. Angeline (Eds.), *Advances in Genetic Programming 3*. Cambridge: The MIT Press, pp. 217–241.

Daida, J. M., J.A. Polito 2, et al. (1999). "What Makes a Problem GP-Hard? Analysis of a Tunably-Difficult Problem in Genetic Programming." In W. Banzhaf, J. Daida, A. Eiben, M. Garzon, et al. (Eds.), *Proceedings of the Genetic and Evolutionary Computation Conference, July 13–17, 1999, Orlando, Florida*. San Francisco: Morgan Kaufmann Publishers, Inc. pp. 982–989.

Daida, J. M., J.A. Polito 2, et al. (2000). "What Makes a Problem GP-Hard? Analysis of a Tunably-Difficult Problem in Genetic Programming." *Journal of Genetic Programming and Evolutionary Machines.* Accepted.

Flajolet, P. and A. Oldyzko (1982). "The Average Height of Binary Trees and Other Simple Trees," *Journal of Computer and System Sciences*, 25, pp. 171–213.

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge: The MIT Press.

Langdon, W.B., T. Soule, et al. (1999). "The Evolution of Size and Shape." In L. Spector, W. B. Langdon, U.-M. O'Reilly, and P. J. Angeline (Eds.), *Advances in Genetic Programming 3*. Cambridge: The MIT Press, pp. 163–190.

Langdon, W. B. and R. Poli (1998). "Why Ants Are Hard." In J. R. Koza, W. Banzhaf, K. Chellapilla, et al (Eds.), *Genetic Programming 1998: Proceedings of the Third Annual Conference, July 22–25, 1998*, University of Wisconsin, Madison. San Francisco: Morgan Kaufmann Publishers. pp. 193–201.

Luke, S. and L. Spector (1998). "A Revised Comparison of Crossover and Mutation in Genetic Programming." In J. R. Koza, W. Banzhaf, K. Chellapilla, et al (Eds.), *Genetic Programming 1998: Proceedings of the Third Annual Conference, July 22–25, 1998*, University of Wisconsin, Madison. San Francisco: Morgan Kaufmann Publishers. pp. 208–213.

O'Reilly, U.-M. and F. Oppacher (1995). "The Troubling Aspects of a Building Block Hypothesis for Genetic Programming." In L. D. Whitley and M. D. Vose (Eds.), *Foundations of Genetic Algorithms 3*. San Francisco: Morgan Kaufmann Publishers. pp. 73–88.

Punch, W. F., D. Zongker, et al. (1996). "The Royal Tree Problem, a Benchmark for Single and Multiple Population Genetic Programming." In P. J. Angeline and K. E. Kinnear, Jr. (Eds.), *Advances in Genetic Programming.* Cambridge: The MIT Press. pp. 299–316.

Raidl, G. R. (1998). "A Hybrid GP Approach for Numerically Robust Symbolic Regression." In J. R. Koza, W. Banzhaf, K. Chellapilla, et al (Eds.), *Genetic Programming 1998: Proceedings of the Third Annual Conference, July 22–25, 1998*, University of Wisconsin, Madison. San Francisco: Morgan Kaufmann Publishers. pp. 323–328.

Soule, T. and J. A. Foster (1997). "Code Size and Depth Flows in Genetic Programming." In J. R. Koza, K. Deb, M. Dorigo, et al (Eds.), *Genetic Programming 1997: Proceedings of the Second Annual Conference, July 13-16, 1997*, Stanford University. San Francisco: Morgan Kaufmann Publishers. pp. 313–320.

Soule, T, J.A. Foster, and J. Dickinson (1996). "Using Genetic Programming to Approximate Maximum Cliques." In J. R. Koza, D. E. Goldberg, D. B. Fogel, and R. L. Riolo (eds.) *Genetic Programming 1996: Proceedings.* Cambridge: The MIT Press, pp. 400–405.

Zongker, D. and W. Punch (1995). *lilgp*. Lansing, Michigan State University Genetic Algorithms Research and Applications Group. http://garage.cps.msu.edu/software/software-index.html.