Rapid Prediction of Optimum Population Size in Genetic Programming using a Novel Genotype – Fitness Correlation

David C. Wedge
Manchester Interdisciplinary Biocentre
School of Chemistry
The University of Manchester
131 Princess Street
Manchester, M1 7DN, United Kingdom
+44 (0) 161 3061545
david.wedge@manchester.ac.uk

Douglas B. Kell
Manchester Interdisciplinary Biocentre
School of Chemistry
The University of Manchester
131 Princess Street
Manchester, M1 7DN, United Kingdom
+44 (0) 161 3064492
dbk@manchester.ac.uk

ABSTRACT

The main aim of landscape analysis has been to quantify the ‘hardness’ of problems. Early steps have been made towards extending this into Genetic Programming. However, few attempts have been made to extend the use of landscape analysis into the prediction of ways to make a problem *easy*, through the optimal setting of control parameters. This paper introduces a new class of landscape metrics, which we call ‘Genotype-Fitness Correlations’. An example of this family of metrics is applied to six real-world regression problems. It is demonstrated that genotype-fitness correlations may be used to estimate optimum population sizes for the six problems. We believe that this application of a landscape metric as guidance in the setting of control parameters is an important step towards the development of an adaptive algorithm that can respond to the perceived landscape in ‘real-time’, i.e. during the evolutionary search process itself.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning – *Parameter learning.*

General Terms: Algorithms, Performance, Design, Theory

Keywords: Landscape, real-world, genotype-fitness correlation, control parameters

1. INTRODUCTION

’ theorem asserts that, when averaged across all possible problems, all search algorithms have the same performance [32]. One implication of this assertion is that matching an algorithm to a problem is necessary in order to give good performance. Much research into the nature of individual problems has focused on ‘adaptive landscapes’ [16, 33]. In particular, the ‘ruggedness’ or smoothness of a landscape is known to affect algorithm performance. Smooth landscapes are generally amenable to simple hill-climbing algorithms while rugged landscapes generally require the injection of a stochastic or more complex search element into the search procedure.

This paper moves towards filling three gaps in problem landscape research. The first is the lack of a satisfactory measure of landscape ruggedness for Genetic Programming (GP) problems. The second is the scarcity of research into ‘real world’ problems. The third is the lack of a link between theory and practice [25]. We attempt to fill these gaps by showing that measures of landscape smoothness may be used directly to predict the optimal value of a control parameter, population size, for a variety of ‘real-world’ problems.

The landscape metric that we introduce measures the extent to which changes in the genotype lead to changes in the fitness function. If small changes in genotype correspond to small changes in fitness while large genotype changes correspond to large fitness changes the landscape may be regarded as smooth. On the other hand if the two quantities are uncorrelated the landscape is rugged. We use a particular measure of inter-tree distance that is closely related to the genetic operators in which we are interested. Similarly, we use a measure of fitness change that scales well for the types of problem (regression problems) in which we are interested. However, the aim is to introduce a *class* of metrics which share the property that they are calculated as correlations between genotypic distance and fitness changes during a genetic operation.

In this paper we demonstrate that genotype-fitness correlations may be calculated ‘offline’, i.e. independently of an evolutionary run. However, in principle these correlations may be calculated in real-time, i.e. during the evolutionary process. This raises the prospect of an *adaptive* evolutionary algorithm that alters its control parameters in response to the prevailing landscape. If an indicator of problem hardness, i.e. landscape ruggedness, can be calculated during an evolutionary run, it should be possible to adapt the evolutionary algorithm to suit the landscape. Calculating a single (non-adaptive) parameter to describe the problem hardness assumes an isotropic landscape. In contrast, an adaptive algorithm has the obvious advantage that the landscape experienced can vary during the course of a run and control parameters will be adjusted accordingly.

As genotype-fitness correlations are calculated for a particular genetic operator, different values may be obtained for each operator. This allows a comparison of different operators and an investigation into the relationship between different operators’ correlations and the values of various control parameters. Specifically, we demonstrate that genotype-fitness correlations are a good predictor of optimum population size during a GP run. This finding is an important step in moving landscape metrics from predicting problem hardness into predicting optimum control parameters.

GECCO’08, July 12–16, 2008, Atlanta, Georgia, USA.
Copyright 2008 ACM 978-1-60558-130-9/08/07...$5.00.

This section has provided an overview of the paper. Section 2 provides a review of the background to our research, covering the two strands that we aim to bring together: theoretical research into problem landscapes (Section 2.1) and practical investigations aimed at optimizing control parameters (Section 2.2). Some approaches that combine theoretical and practical aspects are also covered (Section 2.3). Section 3 describes how we calculate the genotype-fitness correlation we have used for our research. Section 4 defines the problem tasks that we have used to test our metric and Section 5 reports the results of these tests. Section 6 is a discussion of the consequences of our findings. Section 7 is a conclusion and suggested further research. Section 8 contains acknowledgements and Section 9 is a list of references.

2. BACKGROUND

2.1 Landscape Metrics

Developments within the theoretical field may be traced back at least to Kauffman's treatment of 'correlated' landscapes [16]. Problems with a totally smooth landscape may be solved trivially by a mutation-only single-step hill-climber. On the other hand rugged landscapes contain many local optima and a simple hill-climber will not improve after reaching a local optimum. Such landscapes require operators that can 'jump beyond the correlation lengths in the underlying landscape' (Kauffman, [16]). This is equivalent to jumping from a hill into some other part of the search space. This could be another hill or a valley, possibly with lower fitness. Such jumps are valuable in avoiding becoming trapped in local optima. In this analogy, the correlation length is equivalent to the width of the hills.

Manderick *et al.* [20] assessed the relationship between correlation length and GA performance using the auto-correlation function and genetic operator correlation as tools. The auto-correlation is the correlation between pairs of points separated by a specified distance [31]. Thus different auto-correlations may be obtained for points separated by one mutation, two mutations, and so on. Genetic operator correlations, on the other hand, are calculated for specific genetic operators. Different correlations may therefore be obtained for individual genetic operators. Manderick *et al* found that the most effective operators took steps that remained within the correlation length of the landscape.

The conclusions of Kauffman and Manderick are somewhat contradictory. Kauffman argues that one should jump beyond the correlation length whereas Manderick's results suggest that one should stay within it. This conflict may be related to the well-known need for both exploration and exploitation [12]. Exploration may be visualised as a jump to another hill (or valley) whereas exploitation is the process of climbing up the current hill. The need to balance the two suggests that an intermediate jump-size should be optimal, allowing both exploration and exploitation.

Jones and Forrest [15] introduced a measure that enabled comparisons between the difficulties of different problems. They calculated a 'fitness distance correlation' (FDC) as the correlation between the fitnesses of sampled individuals with the distances of these individuals from the global optimum. Assuming that the objective should be maximised they observed that easy problems generally had FDC values below -0.15, hard problems had FDC values between -0.15 and +0.15 and deceptive problems had FDC values above 0.15. The main drawback of FDC is that it requires the location and fitness of the global optimum to be known. For real-world problems this is almost never the case.

Altenberg [2, 3] pointed out the need to use measures that are related to the genetic operators. Manderick's auto-correlation and Jones' distance measure do not meet this criterion for operators other than single-step mutation. Altenberg argues that the evolvability of a population, that is the proportion of offspring that are fitter than their parents, is a better guide to GA performance. However, he suggests that the amount of improvement in fitness, as well as the frequency of improvement, should be taken into account when predicting performance [3]. As with correlation lengths there seems to be an optimum evolvability. This is because more cautious operators may give higher evolvability statistics. Thus, as the step size tends towards zero the evolvability is expected to approach 0.5. This is because offspring will have very similar fitnesses to parents, with the same number of offspring better and worse than their parents, at least in the early stages of an evolutionary run. Larger step sizes are likely to have lower evolvability. However they have a better chance of producing *large* improvements in fitness. Rechenberg [26] and Back *et al* [4] found that an evolvability of 0.2 gave the most rapid improvement on a variety of tasks.

Predictive statistics have rarely been applied to genetic programming (GP). This is perhaps due to the difficulty in defining problems with tunably defined landscapes, in contrast to the GA field where NK landscapes [16] have been extensively studied. Altenberg [1] has suggested that GPs may evolve evolvability through the construction of structures that facilitate further improvement. The best way to measure evolvability was considered to be the correlation of parent and offspring fitnesses: in order for progress to be made, the action of genetic operators on 'good' parents must have a high probability of producing offspring that are also good.

A concept related to correlation length, the 'error threshold', has been applied to GP by Ochoa *et al.* [23]. The error threshold is the mutation rate beyond which 'structures obtained by the evolutionary process are destroyed more frequently than selection can reproduce them'. Ochoa *et al* postulate that the optimum mutation rate would be one close to the error threshold and would therefore balance the demands for exploration and exploitation.

The FDC has been extended into GP through the use of tree-based edit distance measures [8, 10, 21]. The edit distance between a pair of individuals is calculated by overlaying the individuals and counting the minimum number of node changes that would have to be made to transform one individual into the other.

Vanneschi and co-workers [28, 29] have used the negative slope coefficient (NSC) to predict problem hardness. They plot offspring fitness against parent fitness using data generated with the Hastings-Metropolis algorithm. They split the data into bins and plot the median point within each bin. The points are then connected and the slope of each line segment calculated. Negative gradients are considered to be an indication of problem hardness since they are a sign of deception: fitter parents are likely to give less fit offspring within a region with a negative slope. This approach may be seen as an extension of Altenberg's evolvability measure. Unfortunately the results have been found to be highly dependent on the method used to 'bin' data. Some success has been achieved with 'size-driven bisection' [29] but this approach requires the arbitrary setting of parameters to guide partitioning.

2.2 Control Parameter Prediction

The practical task of predicting optimal control parameters for real world problems has generally been pursued separately from the theoretical description of predictive parameters. While a fixed mutation rate of 1/ string_length is believed to give reasonable results with a variety of datasets it has been found that the optimum mutation and crossover rates decrease with population size [13, 14]. More complex interactions between control parameters have been observed by some researchers [22, 24].

Banzhaf *et al* [5] applied four different mutation rates between 5% and 80%, with crossover occurring the rest of the time. They used a fixed population size (3000 individuals) and trees of fixed size. The performance on two classification tasks was assessed using the classification accuracy and the proportion of good runs (those within the top 5% of all runs) as performance indicators. It was found that a 50-50 balance of crossover and mutation gave the best results on these tests.

Luke and Spector [18, 19] similarly varied the crossover-mutation rates and also varied the population size. They used four different tasks (Boolean 6-multiplexer, lawnmower, artificial ant and symbolic regression) to assess the various runs. They found that a crossover rate of 0.6-0.7 gave the best results although the change in performance in response to changes in the crossover-mutation balance was small.

2.3 Combined Approaches

A small number of recent studies have combined the theoretical and practical strands of research by relating a chosen landscape metric to algorithm performance. Corne and co-workers have used finite state machines to simulate the progress of search algorithms on problems including MAX-ONES and NKp landscapes [9, 27]. This enables them to predict the effect of altering the values of control parameters such as mutation rate, population size and tournament size on the performance of a GA.

Burke *et al* [7] consider a number of population descriptors and their use as predictors of GP performance. Their primary aim is to look at the efficacy of diversity measures in predicting performance during the course of a GP run. They sidestep the requirement in FDC to know the global optimum by calculating edit distances to the best individual found *so far*. They find that the average edit distance between the population of individuals and this individual correlates well with the best fitness achieved. Burke *et al* make the observation that regression problems have weaker correlations between diversity measures and fitnesses than other types of problem suggesting that the diversity measures that they used did not capture some features of these problems.

3. GENOTYPE – FITNESS CORRELATIONS

We introduce a new type of measure that we believe is a good indicator of GP landscapes; our measure is the correlation between step size in genotypic space and the change in phenotypic fitness.

The distance measure that we have used is related to those used by de Jong [10] and others [8, 21, 30]. First we overlay a pair of trees. Starting at the root node we then count the number of identical nodes between the trees. If and when we reach a node that differs between the trees we ignore the subtree below that node and backtrack to the next uncounted node. The distance between the trees is then calculated as

$$d = (n_{max} - n_{same}) / n_{max} \quad (1)$$

where n_{max} is the number of nodes in the larger of the 2 trees and n_{same} is the number of nodes which the 2 trees have in common.

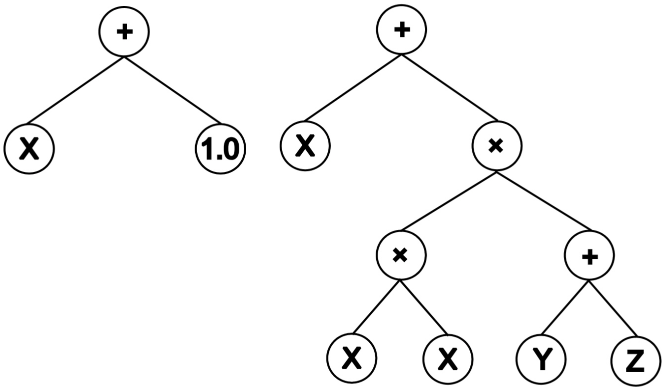

Figure 1. These trees have a distance of 1/3 if normalised w.r.t. the smaller tree, 7/9 if normalised w.r.t. the larger tree.

Our distance metric differs from earlier measures in two ways. First it is normalised with respect to the *larger* of the pair of trees, i.e. the denominator in equation (1) is n_{max} rather than n_{min}. The reason for this is that we wanted trees of very different sizes to have a large inter-tree distance. This can only be achieved by normalising with respect to the larger tree. For example, the trees in Figure 1 have an inter-tree distance of 7/9 using our method and 1/3 if normalised with respect to the smaller tree.

The second way in which our distance metric is novel is that subtrees below differing nodes are ignored. We have chosen to do this for a number of reasons. There is a practical reason, which is that the distance may be calculated more quickly. This is important because one aim is to design a distance metric that is easy and quick to calculate: if the metric is too cumbersome to calculate the improvement in performance may be negated by the computational demands of the metric. Secondly, it gives more emphasis to root nodes. This is illustrated by Figure 2. The 2 trees differ only by the root node and therefore have a distance of 1/3 if all nodes are considered. However, our method gives a distance of 1.0. All nodes are considered to be different since the trees differ at the root node. It has been shown that nodes closer to the root of a tree generally have more effect on phenotypic behaviour and that an emphasis on these nodes is desirable [11]. The final reason for ignoring subtrees below differing nodes is that we wanted a distance metric that would relate closely to the genetic operators being investigated. In this study we have used subtree mutation and subtree crossover as the only genetic operators. For this reason it is difficult to interconvert 2 individuals that differ only at points near the roots of the trees. Our distance metric has worked well with the genetic operators used within this study. However, if alternative operators, such as point mutation, were under investigation it is likely that an alternative distance metric would be a superior indicator of phenotypic variation.

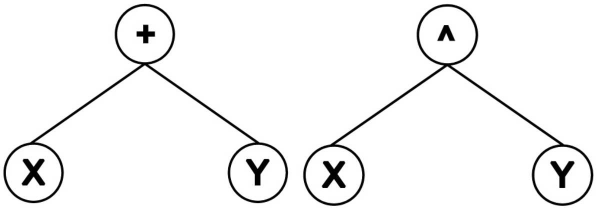

Figure 2. These trees have a distance of 1/3 if all nodes are compared, 1 if sub-nodes below a differing node are ignored.

For every mutation or crossover operation the parent-offspring distance, d_{PO}, was measured. In the case of crossover, two different parent offspring distances were calculated. We distinguished between parent 1, which was defined as the parent that provided the root node, and parent 2, which provided the replacement subtree.

In addition to calculating the parent-offspring genotypic distance, the ratio of parent to offspring fitness was calculated for each new offspring. This ratio, R_{PO}, is more precisely defined in (2).

$$R_{PO} = \log(\max(f_{parent},f_{offspring})/\min(f_{parent},f_{offspring})) \qquad (2)$$

The use of a logarithmic scale was used because the fitness function, f, was the mean square error of each predictor. This value varied across several orders of magnitude so logarithmic scaling was used. Again, parent 1 and parent 2 were distinguished for crossover giving rise to two different R_{PO} values, $R_{PO,1}$ and $R_{PO,2}$.

Our fitness-distance measure differs in several ways from those used by previous authors. It is the 'distance' (R measure) between parent and offspring, not the distance to the global optimum used by Jones and Forrest [15] nor even that to the best-solution-so-far used by Burke [7]. Further, our fitness-distance measure is not a measure of evolvability because it does not matter whether the offspring is more or less fit than its parent(s): it is the log of the ratio of the larger to the smaller fitness in either case. It should also be noted that we consider the whole landscape reachable by the operator, rather than calculating median or mean values as in some alternative approaches [29].

An important property of R_{PO} and d_{PO} is that they are both distance metrics, and therefore have the following properties-

- When parent and offspring are identical they have the value 0. In all other circumstances R_{PO} and d_{PO} are positive.
- Both measures are symmetric, i.e. $R_{PO} = R_{OP}$ and $d_{PO} = d_{OP}$.
- Both measures meet the triangle inequality, i.e. $R_{XY} + R_{YZ} \geq R_{XZ}$ and $d_{XY} + d_{YZ} \geq d_{XZ}$

This is commensurate with the view of the quantities as step sizes i.e. estimates of the distances traveled during a genetic operation in fitness space and genotype space, respectively.

The correlation between R_{PO} and d_{PO}, which we call the genotype-fitness correlation (GFC), is defined by (3), in which *n* is the number of samples used in the calculation and s_R and s_d are the standard deviations of R_{PO} and d_{PO} respectively.

$$GFC = \frac{\sum_{i=1}^{n}\left(R_i - \overline{R}\right)\left(d_i - \overline{d}\right)}{(n-1)s_R s_d} \qquad (3)$$

We believe that the GFC is an indicator of the smoothness of the landscape. If it has a very high value the landscape experienced is highly correlated. In this situation, most individuals are located in regions where the operator in question moves around on the side of a hillside but does not jump beyond the correlation length of the landscape. On the other hand, if this value is very low, most individuals are located in uncorrelated regions. These could be either very rugged landscapes or neutral landscapes [6].

4. PROBLEM DEFINITION

A series of tests was performed in order to find out whether the GFC could be used to predict optimum control parameters, i.e. those parameters that give the lowest mean square error (MSE) within a fixed number of fitness evaluations. Six tasks were identified and many GP runs were carried out on each task using a range of control parameters, with MSEs calculated for every run. In parallel to these runs, calculations of GFC were made for each task. The relationship between GFC values and final fitnesses (MSEs) was then investigated.

The function set comprised four arithmetic functions: addition, subtraction, multiplication and protected division. Generated trees had a maximum depth of 17 and length of 250. A tournament of size 2 was used for selection and the population was evolved in steady-state mode.

All of the problems studied were regression tasks using real-world or pseudo real-world data and they are readily available on the Internet. Two (Boston and kin32nh) were obtained from the Delve repository at Toronto University[1]. The remaining datasets were obtained from Luis Torgo's repository at the University of Porto[2]. The data size and dimensionality (number of predictor variables) are given in Table 1.

Table 1. Dataset properties

dataset	size	dimensionality
auto mpg	392	7
Boston	506	13
cart	40768	10
delta ailerons	7129	5
kin32nh	8192	32
machine	209	6

Two control parameters were varied during the investigation: the crossover/mutation balance and the population size. All offspring were either mutated or recombined from their parent(s), i.e. no cloning occurred. The crossover rate was varied in steps of 0.1 between 0.2 and 0.9. Mutation took place in all other cases, giving corresponding mutation rates that varied between 0.8 and 0.1. The population size comprised the set of values {5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400}. These parameter sweeps give rise to 144 different training regimes. 100 GP runs were carried out using each regime.

[1] *http://www.cs.toronto.edu/~delve/data/datasets.html*

[2] *http://www.liaad.up.pt/~ltorgo/Regression/DataSets.html*

Each run was conducted for 20,000 evaluations, equivalent to between 50 and 4000 generations depending on the population size. Only two operators were used: single-point subtree mutation and single-point subtree crossover. Mutation and crossover points were selected randomly and only one individual was created in each operation. In the case of crossover the parent that provided the root node was identified as ‘parent 1’ and the second parent as ‘parent 2’[3], with the two parents selected through separate tournaments. It is expected that parent 1 will have a larger influence on the fitness of the offspring. We used Langdon’s ‘length neutral’ mutation operator [17]. It fixes the length of the replacement subtree to be between 50% and 150% of the length of the removed subtree, thereby limiting the appearance of program bloat. Correlation calculations were carried out separately for each operator.

The measures of fitness step size and genotype step size were those defined in section 3. The performance of the algorithms was assessed as the mean square error (MSE), averaged over 100 runs of a particular training regime.

GFC calculations were carried out separately. These involved the creation of a random population of 100 solutions followed by the repeated application of a single genetic operator to randomly selected solutions, i.e. a random walk was performed across the whole population. Each walk was continued for 1000 evaluations and separate walks were conducted for the mutation and crossover operators. 100 runs were performed and quoted values have been averaged over all runs.

The time taken to calculate a parent-offspring distance was of the order of 10 µs on a 2.4GHz PC. This is negligible compared to the time required for fitness evaluation: even for the smallest dataset (machine) this was of the order of 10 ms on the same computer.

All code has been produced in-house in the Java programming language. Programs were executed via Condor, a workload management system that distributes processing across a pool of computers.

5. RESULTS

5.1 Control Parameters

Figures 3 and 4 show the variation in mean MSE with, respectively, crossover rate and population size. In the crossover rate plane (Figure 3) there is a bowl shape with the lowest values of mean MSE occurring for medium values of crossover rate and high values occurring for both high and low rates of crossover. All problems show a preference for high crossover rates (above 0.5) and perform particularly badly for low crossover rates. A crossover rate of 0.6 is within the optimum range for all datasets except Cart (for which it is the second best rate). This suggests that there is an optimal crossover rate that, for the domains studied, is largely independent of problem landscape.

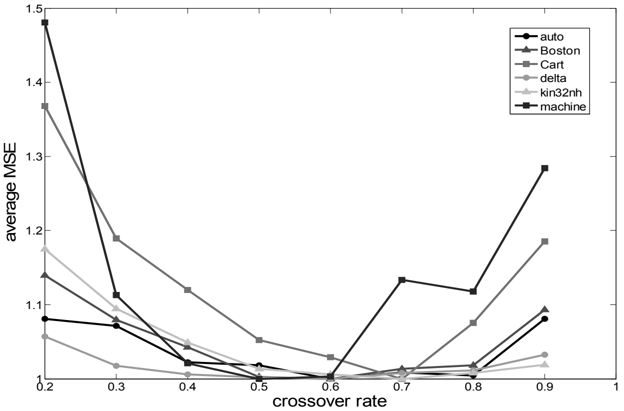

Figure 3. Variation in MSE with crossover rate, normalized to the minimum MSE achieved.

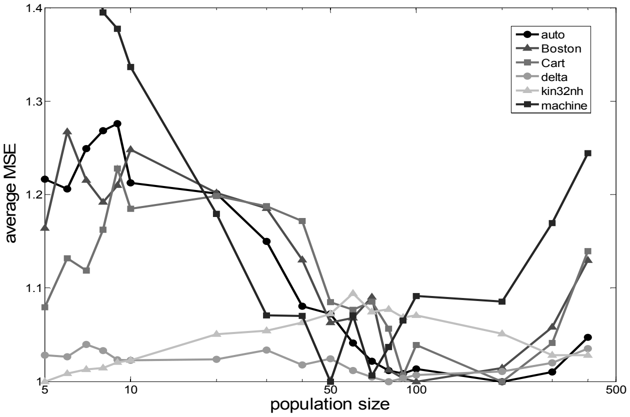

Figure 4. Variation in MSE with population size, normalized to the minimum MSE achieved.

Table 2. Optimum control parameters

dataset	optimum crossover rate	optimum population size
auto mpg	0.6-0.8	90-200
Boston	0.5-0.6	90-100
cart	0.7	90-200
delta ailerons	0.3-0.8	70-100
kin32nh	0.6-0.8	3-6
machine	0.5-0.6	50-70

The MSE varies less smoothly in response to population size (see Figure 4), although a bowl shape is again seen for all datasets except kin32nh. However, the minimum occurs at very different values for different datasets.

Table 2 shows the optimum population size and crossover rate for each dataset. When comparing the values for different population sizes, the values have been averaged across all crossover rates. Similarly, when comparing different crossover rates, values for the

[3] This procedure differs slightly from standard GP crossover, in which 2 offspring are produced and both parents will provide the root node of one offspring.

different population sizes have been averaged.[4] In cases where there are several parameters that give results which are almost as good as each other (average MSE within 1% of each other) a range of optimum values has been supplied.

5.2 Correlation Statistics

Table 3. Mean GFC values, with S.D. Calculation method used a population of 100 and random selection for 1000 evaluations.

dataset	crossover GFC (parent 1)	crossover GFC (parent 2)	mutation GFC
auto mpg	0.134±0.055	0.124±0.037	0.216±0.067
Boston	0.175±0.065	0.120±0.033	0.196±0.069
cart	0.052±0.046	0.084±0.035	0.120±0.060
delta ailerons	0.157±0.063	0.125±0.036	0.206±0.064
kin32nh	0.308±0.066	0.181±0.046	0.304±0.079
machine	0.182±0.064	0.135±0.039	0.218±0.076

Table 3 shows the calculated GFC values. For crossover, GFC values were calculated separately for parent 1 and parent 2 (see Section 3). Figures 5 and 6 plot the midpoint of the range of optimal population sizes against these values. Figure 7 is the corresponding plot using mutation GFCs. The Pearson coefficients for the plots are 0.92, 0.88 and 0.83 for, respectively, crossover (parent1), crossover (parent 2) and mutation. The high degree of linearity of these plots indicates that the optimum population size may be estimated from a single GFC by assuming a simple linear model.

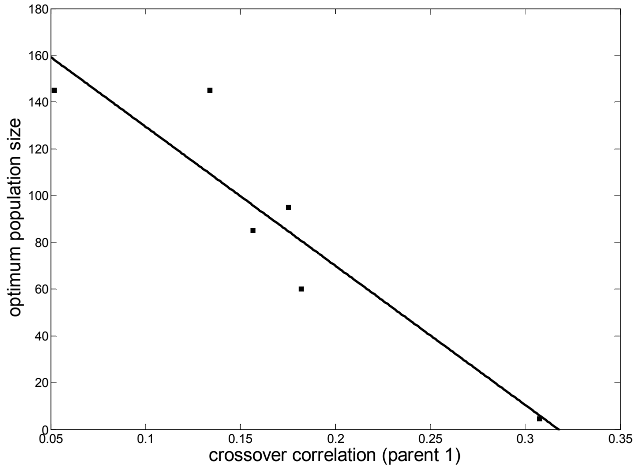

Figure 5. The midpoint of the optimal population range plotted against the crossover GFC (parent 1) for 6 different datasets.

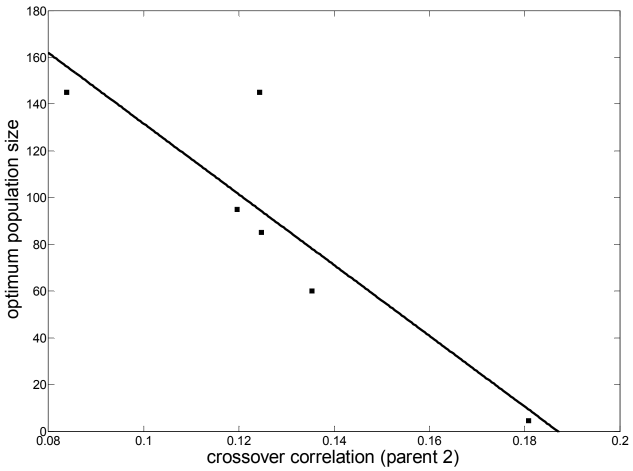

Figure 6. The midpoint of the optimal population range plotted against the crossover GFC (parent 2) for 6 different datasets.

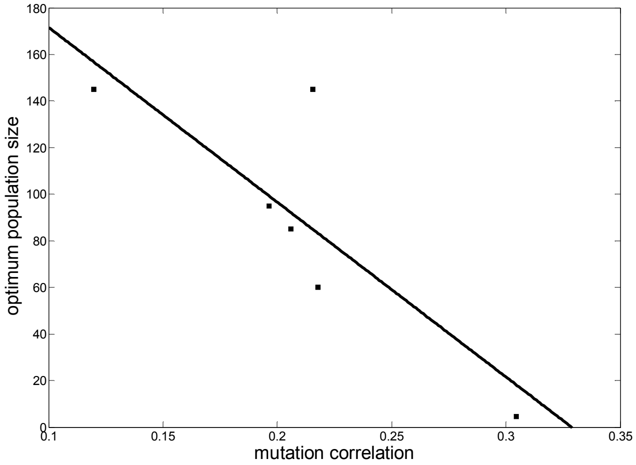

Figure 7. The midpoint of the optimal population range plotted against the mutation GFC for 6 different datasets.

This finding is very pleasing but also slightly surprising. It is intended that further work will investigate the extent to which the linear relationship holds. For example, as the GFC increases we might expect the optimum population size to approach 1 asymptotically (implying a hill-climber), while as the GFC approaches 0 we would expect the optimum population size to approach the total number of evaluations (random search). It is possible that the 'true' relationship between GFC and optimum population size is exponential but that a straight line gives a reasonable approximation for most real-world problems. The investigation of additional datasets with GFCs outside the range included here is required to test the range of applicability of the linear approximation.

6. DISCUSSION

The results obtained indicate that the GFC is a very useful predictive tool. It has been calculated using just 1000 evaluations, averaged over 100 runs. However, for the datasets studied, a linear model is able to predict reliably which population size will give the best performance at the end of a run lasting 20,000 evaluations. Overall, the number of evaluations carried out on each problem was

[4] This procedure assumes that the two parameters have orthogonal effects. In order to test that this was the case, two-way analysis of variance (ANOVA) tests were performed. They indicated that interactions between crossover rate and population size could be ignored for all datasets except machine and kin. For these 2 datasets the effects due to either crossover rate or population by themselves were very much higher than their combined effect.

nearly 300 million (100 runs of 20,000 evaluations each on 144 different control parameter combinations). In contrast, each GFC has been obtained from just 100,000 evaluations.

The power of GFCs as a predictive tool is dependent upon an assumption that population size is linearly related to GFC. Additional datasets are required to further test this hypothesis. However, we believe that the GFC as it stands gives useful guidance concerning optimum population sizes.

In future work we hope to show that the calculation of GFCs 'online', i.e. during an evolutionary run, may have a further benefit. If carried out online they require negligible additional computation: fitness values would be calculated already for the purposes of selection and the genotype distance calculation is very fast. They could therefore be incorporated into an adaptive GP that altered its control parameters in response to the landscape experienced. The results reported in Section 5 imply that an adaptive population size would be more profitable than an adaptive crossover/mutation rate.

The negative correlation between GFC and population size has been demonstrated but the reasons for the relationship are not yet certain. We suspect that it is related to the need to maintain diversity and the associated need for a balance between exploration and exploitation. If diversity is lost too early in an evolutionary run, the population is likely to settle in a local optimum, i.e. the exploration phase is too short. On the other hand, too much diversity may be linked to a failure to converge within the time allowed, i.e. the exploration phase is too long, allowing little time for exploitation. With all datasets the lowest diversities are observed for small populations (which are evolved for a large number of generations). It has been shown that these small populations are optimal for problems with a high GFC, i.e. those with smooth landscapes, whereas problems with rugged landscapes are solved more effectively by evolving a large population for a small number of generations. This implies, as one might have expected, that it is best for problems with smooth landscapes to use exploitation whereas those with rugged landscapes are solved more effectively using a high level of exploration.

7. CONCLUSIONS AND FURTHER WORK

This study has described a novel family of metrics, known as genotype-fitness correlations (GFCs), that provide a useful measure of GP landscape. A particular metric, suitable for the investigation of regression problems using subtree mutation and crossover, has been used to measure the ruggedness of the landscape for six different regression tasks. It has been shown that the GFC is a good guide to the optimal population size, with (for a given number of evaluations) rugged landscapes showing a preference for large populations while smoother landscapes are more effectively conquered using smaller populations.

Obtaining the GFC 'offline', as we have done here, is useful for giving an indication of optimal population size before carrying out an evolutionary run. In future work we intend to show that GFCs can also be obtained in 'online' mode, using information collected during an evolutionary run. This raises the prospect of an adaptive evolutionary algorithm based on these ideas, capable of adjusting its control parameters as it experiences different landscapes.

We plan to extend GFCs to other genetic operators such as tree-based point mutation or single-point binary string crossover. This is likely to involve the definition of new types of GFCs, suited to the particular operators under investigation.

While a crossover-mutation ratio of 0.6 was found to work well within this study, this might not be the case for other mutation operators, such as swap or point mutations. In cases where the optimal crossover-mutation ratio varies, it is hoped that it will become possible to predict the best balance between different operators using GFCs.

8. ACKNOWLEDGEMENTS

We thank the UK Home Office for funding, Joshua Knowles for helpful comments on an earlier draft of this paper and Ian Cottam for setting up and administering the Condor system through which the GPs were run.

9. REFERENCES

[1] Altenberg, L. The Evolution of Evolvability in Genetic Programming. In *Advances in Genetic Programming* (Ed. K. E. Kinnear), MIT Press, 1994, 47-74.

[2] Altenberg, L. The Schema Theorem and Price's Theorem. In *Foundations of Genetic Algorithms*, Estes Park, Colorado, USA, July 31 - August 2 1995. Morgan Kaufmann, 23-49.

[3] Altenberg, L. Fitness Distance Correlation Analysis: An Instructive Counterexample. In *Proceedings of the 7th International Conference on Genetic Algorithms*, East Lansing, MI, USA, July 19-23 1997. Morgan Kauffman, 57-64.

[4] Bäck, T., Hoffmeister, F. and Schwefel, H.-P. A Survey of Evolution Strategies. In *Proceedings of the International Conference on Genetic Algorithms*, San Diego, CA, USA, 1991. Morgan Kaufmann, 2-9.

[5] Banzhaf, W., Francone, F. and Nordin, P. The Effect of Extensive Use of the Mutation Operator on Generalization in Genetic Programming using Sparse Data Sets. In *Proceedings of the 4th Conference on Parallel Problem Solving From Nature*, Berlin, Germany, 1996. Springer-Verlag, 300-310.

[6] Borenstein, Y. and Poli, R. Fitness distributions and GA hardness. In *Proceedings of the 8th Conference on Problem Solving in Nature*, Birmingham, UK, Sep 18-22 2004. Springer, 11-20.

[7] Burke, E. K., Gustafson, S. and Kendall, G. Diversity in Genetic Programming: An Analysis of Measures and Correlation with Fitness. *IEEE Transactions on Evolutionary Computation*, 8, 1, Feb 2004, 47-62.

[8] Clergue, M., Collard, P., Tomassini, M. and Vanneschi, L. Fitness Distance Correlation And Problem Difficulty For Genetic Programming. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, New York, USA, 2002. Morgan Kaufmann, 724-732.

[9] Corne, D. W., Oates, M. J. and Kell, D. B. Landscape State Machines: Tools for Evolutionary Algorithm Performance Analyses and Landscape/Algorithm Mapping. In *Evoworkshops*, Essex, UK, 14-16 April 2003. Springer, 187-198.

[10] de Jong, E. D., Watson, R. A. and Pollack, J. B. Reducing Bloat and Promoting Diversity using Multi-Objective Methods. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, San Francisco, CA, USA, 7-11 July 2001. Morgan Kaufmann, 11-18.

[11] Ekárt, A. and Németh, S. A metric for genetic programs and fitness sharing. In *Proceedings of the European Conference on Genetic Programming*, Edinburgh, Scotland, UK, 2000. Springer, 259-270.

[12] Eshelman, L. J., Caruana, R. and Schaffer, J. D. Biases in the Crossover Landscape. In *Proceedings of the 3rd International Conference on Genetic Algorithms*, George Mason University, Fairfax, Virginia, USA, 1989. Morgan Kaufmann, 10-19.

[13] Grefenstette, J. J. Optimisation of control parameters for genetic algorithms. *IEEE Transactions on Systems, Man and Cybernetics*, 16, 1, Jan-Feb 1986, 122-128.

[14] Hesser, J. and Männer, R. Towards an Optimal Mutation Probability for Genetic Algorithms. In *Proceedings of the 1st Conference on Parallel Problem Solving from Nature*, Dortmund, Germany, Oct. 1-3 1990. Springer-Verlag, 23-32.

[15] Jones, T. and Forrest, S. Fitness Distance Correlation as a Measure of Problem Difficulty for Genetic Algorithms. In *Proceedings of the 6th International Conference on Genetic Algorithms*, Pittsburgh, PA, USA, Jul 15-19 1995. Morgan Kaufmann, 184-192.

[16] Kauffman, S. and Levin, S. Towards a General Theory of Adaptive Walks on Rugged Landscapes. *Journal of Theoretical Biology*, 128, 1, Sep 1987, 11-45.

[17] Langdon, W. B. The Evolution of Size in Variable Length Representations. In *Proceedings of the IEEE International Conference on Evolutionary Computation*, Anchorage, Alaska, USA, May 4-9 1998. IEEE Press, 633-638.

[18] Luke, S. and Spector, L. A Comparison of Crossover and Mutation in Genetic Programming. In *Proceedings of the 2nd Conference on Genetic Programming*, Stanford University, CA, USA, Jul 13-16 1997. Morgan Kaufmann, 240-248.

[19] Luke, S. and Spector, L. A Revised Comparison of Crossover and Mutation in Genetic Programming. In *Proceedings of the 3rd Conference on Genetic Programming*, University of Wisconsin, Madison, Wisconsin, USA, Jul 22 - 25 1998. Morgan Kaufmann, 208-213.

[20] Manderick, B., de Weger, M. K. and Spiessens, P. The Genetic Algorithm and the Structure of the Fitness Landscape. In *Proceedings of the 4th International Conference on Genetic Algorithms*, San Diego, CA, USA, 1991. Morgan Kaufmann, 143-150.

[21] O'Reilly, U.-M. Using a Distance Metric on Genetic Programs to Understand Genetic Operators. In *Late-breaking papers at the IEEE International Conference on Systems, Man, and Cybernetics*, Orlando, FL, USA, Oct 12-15 1997. IEEE, 4092-4097.

[22] Oates, M., Corne, D. and Loader, R. Investigation of a characteristic bimodal convergence-time/mutation-rate feature in evolutionary search. In *Proceedings of the 1999 Congress on Evolutionary Computation*, Washington, DC, USA, Jul 6-9 1999. IEEE, 2182-2188.

[23] Ochoa, G., Harvey, I. and Buxton, H. Error Thresholds and their Relation to Optimal Mutation Rates. In *Proceedings of the 5th European Conference on Artificial Life*, Lausanne, Switzerland, Sep 13-17 1999. Springer, 54-63.

[24] Ochoa, G., Harvey, I. and Buxton, H. On Recombination and Optimal Mutation Rates In *Proceedings of the 8th Genetic and Evolutionary Computation Conference (GECCO)*, Orlando, Florida, USA, Jul 13-17 1999. Morgan Kaufmann, 488-495.

[25] Radcliffe, N. J. and Surry, P. D. Fundamental Limitations on Search Algorithms: Evolutionary Computing in Perspective. In *Computer Science Today* (Ed. J. van Leeuwen), Springer, 1995, 275-291.

[26] Rechenberg, I. *Evolutionsstrategie - Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Fromman-Holzboog, Stuttgart, 1973.

[27] Rowe, W., Corne, D. W. and Knowles, J. Developing Landscape State Machines for Improved Algorithm Performance Prediction. In *IEEE Congress on Evolutionary Computation*, Vancouver, Canada, July 16-21 2006. IEEE Press, 2944 - 2951.

[28] Vanneschi, L., Clergue, M., Collard, P., Tomassini, M. and Verel, S. Fitness Clouds and Problem Hardness in Genetic Programming. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, Seattle, WA, USA, Jun 26-30 2004. Springer, 690-701.

[29] Vanneschi, L., Tomassini, M., Collard, P. and Verel, S. Negative Slope Coefficient: A Measure to Characterize Genetic Programming Fitness Landscapes. In *Proceedings of the 9th European Conference on Genetic Programming*, Budapest, Hungary, Apr 10-12 2006. Springer, 178-189.

[30] Wedge, D. C., Gaskell, S. J., S.J., H., Kell, D. B., Lau, K. W. and Eyers, C. Peptide Detectability Following ESI Mass Spectrometry: Prediction using Genetic Programming. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, London, UK, 7-11 July 2007.

[31] Weinberger, E. D. Correlated and Uncorrelated Fitness Landscapes and How to Tell the Difference. *Biological Cybernetics*, 63, 5, Sep 1990, 325-336.

[32] Wolpert, D. H. and Macready, W. G. No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation*, 1, 1, Apr 1997, 67-82.

[33] Wright, S. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proceedings of the 6th International Congress on Genetics*, Ithaca, NY, USA, 1932. 356-366.