# An Ecological Approach to Measuring Locality in Linear Genotype to Phenotype Maps

Tom Seaton, Julian F. Miller, and Tim Clarke

Department of Electronics
University of York, YO10 5DD
{tas507,jfm7,tc2}@ohm.york.ac.uk

**Abstract.** Recent research has considered the role of locality in GP representations. We use a modified statistical technique drawn from numerical ecology, the Mantel test, to measure the locality of integer-encoded GP. Weak locality is identified in a case study on Cartesian Genetic Programming (CGP), a directed acyclic graph representation. A method of varying syntactic program locality continuously through the application of a biased mutation operator is demonstrated. The impact of varying locality under the new measure is assessed over a randomly generated set of polynomial symbolic regression problems. We observe that enforcing higher levels of locality in CGP is associated with poorer performance on the problem set and discuss implications in the context of existing models of GP genotype-phenotype maps.

**Keywords:** Cartesian genetic programming, Locality.

## 1 Introduction

The notion of locality is a well-established property of representations in genetic algorithms, known to have an impact on search performance [1–4]. Locality describes the design heuristic that small changes to a genotype, due to evolutionary operators, should lead to correspondingly small changes in phenotype. The concept can also be related to the assertion that the genotype to phenotype map (GPM) should support a strong causal relationship between the evolving data structure and its decoded expression [5]. Recently, work in genetic programming has focused on extending the original concept of locality from binary strings to standard, GP tree-based representations [6].

This paper considers the design of a novel method of measuring locality, with the aim of assessing indirect, integer-encoded GPM. Our goal is to establish a statistical approach which can then be applied to linear genotypes [7], encompassing methods such as Cartesian Genetic Programming (CGP) [8], Grammatical Evolution [9, 10] or Linear GP [11]. Characteristically, these GP encodings feature an intermediate level of mapping between genotype and fitness evaluation not traditionally incorporated in tree-based GP. The method described here adopts a long standing technique from the field of numerical ecology, the Mantel test [12–18]. The purpose of the test is to provide a means of rigorously determining the significance of measured correlations between distance matrices.

A. Moraglio et al. (Eds.): EuroGP 2012, LNCS 7244, pp. 170–181, 2012.

In Section 2, we briefly review related work on measuring locality in the GA and GP literature and comment on previous definitions. Section 3 provides necessary background on the Mantel statistic and introduces an extension to enable the technique to address correlations in genotype-phenotype maps. Section 4 addresses selection of metrics on the genotype and phenotype space. Section 5 presents a set of preliminary experimental outcomes and analyses in a case study on standard Cartesian GP, over a randomised set of symbolic regression problems. We then discuss these initial results and conclude.

## 2 Related Work

The early work of Rothlauf is generally cited as the seminal work on locality in the study of representations [1]. Rothlauf proposed aggregating the degree of change in phenotype over the local neighbourhood of each genotype, defined with respect to the particular variational operators:

$$L = \sum_{g \in G} \sum_{p' \in adj^*(g)} d_P(p, p') - d_P^- \tag{1}$$

where $L \geq 0.0$ is the level of locality ($L = 0.0$ is maximal). In our notation we use $G$, $d_G$, $P$ and $d_P$ to denote the genotype space, phenotype space and respective distance metrics over each. Equation 1 measures the locality of a map $S : G \to P$ over all neighbouring pairs of genotypes in genotype space, where $adj^*(g)$ denotes the set of phenotypes which correspond to the adjacent neighbours of $g$ in $G$. Distances are summed up under the phenotype metric, relative to the minimum distance in phenotype space, which we denote $d_P^-$. Thus, in a high locality map under Rothlauf's definition, $L$ tends to 0 and genotypes which are neighbours in genotype space also have phenotypes which are similar under the metric applied in that space. The original expression is not normalised with respect to the size of the search space - more recently, extensions to Rothlauf's work on locality in GA for binary strings were proposed by Chiam et. al [19].

Studies of locality in GP [6, 20, 21], by contrast, have considered locality as a direct property of the mapping between genotype and corresponding fitness value. Galvan-Lopez et. al. considered a set of three definitions of locality, derived from Rothlauf's work, and systematically examined each over a set of standard GP benchmarks [6]. This approach would seem appropriate for classical tree GP, where there is no explicit intermediate state between genotype and fitness. However, for indirect GP maps which feature distinct phenotypes, we argue that an understanding of locality should also be sought at the intermediate level. Furthermore, current measures of locality by definition do not consider any relationships at genotype distances beyond the immediate neighbourhood. The method presented in this paper explores an alternative to these aggregative approaches and studies directly the degree of correlation between genotype distances and phenotype distances. There exists some commonality with the method of fitness distance correlation (FDC) extensively addressed in both the GA and GP literature [22]. However, fitness distance correlation develops a measure of problem

difficulty, by considering distances to the generally unknown optimum in fitness space. We are instead preoccupied with understanding locality at the syntactic level - the changes that are introduced into program structure by variations in integer genotype. This enables a problem-independent view, focusing on analysis of the representation and search operators, rather than addressing performance against a particular fitness landscape.

## 3 The Mantel Test

The Mantel Test is a general, non-parametric statistical resampling technique used in the exploration of correlations between two triangular distance matrices [12] Historically, the test was designed to address the analysis of spatial and temporal data from disease clustering. It has seen considerable application in numerical ecology [14–17] and on genetic and linguistic data [18]. In a mathematical sense, the Mantel test provides a permutation-based method of determining the statistical significance of linear or monotonic relationships. The test is applicable in situations when we wish to determine whether a correlation exists in the distances between elements sampled between two metric spaces. Note in particular that it is not appropriate to use standard significance tests because distances derived from the same element cannot generally be considered independent of each other [18]. The technique is applied between two square distance matrices of size $n$, labelled $\mathbf{X}$ and $\mathbf{Y}$. The matrices contain the pair-wise differences calculated between all elements of a sample under two measures of distance $d_X$ and $d_Y$. By way of illustration, in the ecological context $\mathbf{X}$ might represent the geographical distances between samples of a species and $\mathbf{Y}$ corresponding measured genetic distances. Differences are assumed to adhere to the symmetry property of a metric, so both matrices are symmetric with zeros along the diagonal. The original, 'standardised mantel statistic' is then given by the expression [23]

$$r_M = \frac{1}{s-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left( \frac{X_{i,j} - \bar{X}}{\sigma_X} \right) \left( \frac{Y_{i,j} - \bar{Y}}{\sigma_Y} \right) \tag{2}$$

where $r_M$ is the linear correlation coefficient obtained, $\bar{X}$, $\bar{Y}$ and $\sigma_X$, $\sigma_Y$ are the mean and standard deviation calculated for $\mathbf{X}$ and $\mathbf{Y}$ respectively and $s = n(n-1)/2$. This is equivalent to calculating the Pearson-product moment (linear correlation) over the upper-half of the matrix.

### 3.1 Significance Testing on Genotype-Phenotype Maps

For $r_M$ to be a useful statistic under sampling, significance testing should be carried out against the null hypothesis, $H_0$ that the distances in $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated. A key realisation of the Mantel test is that rows and columns of the matrix are exchangeable under the null hypothesis. That is, we expect to be able to freely rearrange the labels of each set of distances. By permuting the rows (and corresponding columns) of $\mathbf{X}$ and recalculating $r_M$, a permutation

distribution can be constructed from which the significance of correlations in the unpermuted data is obtained. Given that the null hypothesis is true, we would expect that the unpermuted data should lie somewhere in the center of this range. The test proceeds by obtaining the original unpermuted coefficient $r_M^0$ and a set of coefficients under permutation of $\mathbf{X}$, denoted $r = \{r_M^1 ... r_M^N\}$, where $N$ is the total number of permutations. Let $\chi \subseteq r$ such that $x \in \chi \geq r_M^0$. The probability of accepting the null hypothesis in the presence of an apparent positive correlation is then given by the one sided test

$$P(H_0|r) \approx \frac{||\chi||}{N} \tag{3}$$

that is the number of instances in which the recalculated coefficient equals or exceeds $r_M^0$, divided by the total number of permutations. A similar test can be carried out for the case of negative correlation.The test does not necessarily have to support a linear model: it may be appropriate to compute $r_M$ using an alternative statistic, such as Spearman rank-based correlation, using the permutation test in exactly the same fashion. The result converges monotonically on the true significance at large $N$. In practice, the number of permutations recommended in the literature varies, but a value in the range of 1000-10000 permutations is typically suggested [23].[1]

For the application of the Mantel statistic to artificial GPM, a method is required to calculate it over particular distance intervals. This is to establish whether a correlation exists only for closer, or more distant, genotypes. A similar situation arises in numerical ecology, where correlations may be limited by time, or by geographic distance. Previous derived techniques of the Mantel statistic have consider correlation as a function of range, such as the 'Mantel correlogram', which applies a model matrix to examine correlations over particular distance classes [14]. We adopt a simplified approach, explicitly sub-dividing the distance matrix. Let $\mathbf{X_U}$ be the upper triangle of $\mathbf{X}$. A set of distance classes are selected such that each distance class $\mathbf{D}_{p,q}$ is a subset of the elements of $\mathbf{X_U}$ where $p \leq X_{i,j} < q$. Hence, a distance class contains the elements over which $r_M$ is computed which fall within the range $(p, q)$. The corresponding set of distances at the same index positions in $\mathbf{Y_U}$ are also found. The coefficient $r_M$ is calculated seperately for each distance class and significance values derived as before, by permuting the original matrix and recomputing $r_M$ over that interval.

## 4 Distance Metrics under the Mantel Statistic

To derive distances between genotypes and phenotypes, an appropriate metric must be selected for each space. Which metrics are suitable is informed by the choice of representation and variation operators. For this initial analysis, we neglect crossover and focus on the mutation operator. Numerous proposals have been put forward for appropriate metric distance measures in GA/GP genotype spaces, see for example the review in [25]. These have included classes such as:

[1] Standard methods for carrying out permutation tests are supported in numerical ecology statistical packages such as `ecodist` and `vegan` [24], in *R*.

M1. *Edit distances* (e.g. Hamming, Levenstein in the case of strings and tree or graph edit distances respectively. [6, 26])
M2. *Subtree distance* (e.g. Tree Alignment [27], Keijzer distance. [28])
M3. *Information compression* (e.g. Normalised Compression Distance. [6])
M4. *Probabilistic measures* (e.g. Subtree Crossover Operator. [29])

Although it is typically practical to define strict metrics for the genotype space under the assumptions made in M1, M2 and M3, those measures based directly on probabilities usually violate one or more of the metric criteria[2]. Calculation of the Mantel statistic only requires that measures adhere to the symmetry criterion [23]. We are therefore in principle free to select from amongst each of the above classes of metric. As noted in [6], there is no priori knowledge of which distance measures are most appropriate to describe differences on program spaces. The approach described here chooses distance measures of types (3) and (4). Our justification for this is one of pragmatism: to avoid being tied to representation specific measures in our analysis and because of the potential complexity of computing edit distances on large graphs (the graph edit distance problem is NP-hard in general). A convenient measure for integer genotypes is the expected number of independent attempts that would be required to generate one genotype from another through a single mutation. Assuming that the two genotypes are mutually reachable (Definition 1), then this is just the inverse of the probability of mutating between both genotypes. We refer to this semimetric as the expected variation distance $\bar{M}$. The measure has the advantage that it defines distance based on the actual transition probability.

**Definition 1 (Mutually reachable genotypes).** *A pair of genotypes (g,g′) are mutually reachable under some variation operator $V$, given that the probability of deriving $g'$ in a single operation $V(g)$ is greater than zero.*

**Definition 2 (Expected variation distance).** *A function $\bar{M} : (g, g', V) \to \mathbb{Z}$ on a pair of mutually reachable genotypes $(g, g')$ where $\bar{M}$ gives the expected number of independent single operations on $g$ such that there is an instance $V(g) = g'$. If $g = g'$, we define $\bar{M}(g, g, V) \equiv 0$.*

**Derivation for CGP with Uniform Mutation Operator.** We can illustrate this approach by calculating the expected variation distance between a pair of standard feed-forward CGP genotypes $g$ and $g'$. Genotypes are assumed to be equal sized integer strings of length $n$, which represent a single row with feedforward connections (see [8] for details), where output is derived from the right-most node. Assume there are $x$ matching integers between $g$ and $g'$ and $n - x = y$ different values. The genotype is split into integers corresponding to connections and functions. From the $y$ different integer values, we have a subset of size $y_F$ different values corresponding to functions and $y_C$ values corresponding

[2] A metric function $d$ on metric space $Q$ satisfies: 1. $d(x, y) \geq 0$; 2. $d(x, y) = 0$ iff $x = y$; 3. $d(x, y) = d(y, x)$ ; 4. $d(x, y) \geq d(x, z) + d(z, y)$ where $x, y, z \in Q$. We adopt the conventional term semimetric when the triangle-inequality is relaxed.

**Table 1.** CGP Search Parameters

| | | | |
|---|---|---|---|
| Representation | CGP | Crossover | No |
| Nodes | 10 | Selection Strategy | $(\mu + \lambda) = (4 + 6)$ |
| Structure | Single row feed-forward | Population Size | 10 |
| Function Set | {+,-,*, ÷} | Fitness Samples | $10 \in \{-2 : 2\}$ |
| Terminal Set | {0,1} | Max Generations | 2000 |
| Mutation Rate | 0.15 | Runs | 500 |

to connections. Assume a mutation operator acts on all values with uniform probability $m$, where a mutation changes the allele to any other feasible integer. Then, the probability of $x$ values remaining the same is $(1-m)^x$. The probability of $y_F$ values from $g$ mutating to the same function as that in $g'$ is $(\frac{m}{F-1})^{y_F}$, where $F$ is the number of possible function choices. Let $\mathbf{y_C}$ be the set of integer values which contribute to $y_C$. Each connection $i \in \mathbf{y_C}$ has $c_i - 1$ possible alternatives (where $c_i$ is in general the total number of inputs, plus all previous nodes). Thus the probability of obtaining the same set of connections is $m^{y_C} \prod_{i \in \mathbf{y_C}} \frac{1}{c_i - 1}$. The total probability $u$ of mutating from one CGP genotype to another is therefore

$$u(g, g') = (1 - m)^x \cdot \left(\frac{m}{F - 1}\right)^{y_F} \cdot m^{y_C} \prod_{i \in \mathbf{y_C}} \frac{1}{c_i - 1} \tag{4}$$

Taking the inverse and collecting terms gives the expected number of independent mutations required, $\bar{M} = \frac{1}{u}$:

$$\bar{M} = \frac{(F - 1)^{y_F}}{m^y \ (1 - m)^x} \prod_{i \in \mathbf{y_C}} c_i - 1 \tag{5}$$

Distances between genotypes under uniform mutation can be computed in other integer representations such as grammar GP in a similar fashion.

## 5 Experiment

To test our approach, 50 biarity tree samples were obtained from a representative CGP genotype space using an arithmetic function set (including the protected division operator). Table 1 summarises the parameters used to initialise each genotype. Sample biarity trees were produced recursively under the uniform mutation operator to a depth of 7 mutations, generating 511 genotypes per sample.[3] The expected variation distance $\bar{M}$ was obtained pair-wise for all members of each sample. The approach provided a set of 50 corresponding matrices each containing ∼150000 genotype distances. CGP phenotypes are directed acyclic graphs. Measurements of the syntactic distance between CGP phenotypes ($d_P$) were derived using the Normalised Compression Distance (NCD) (Equation 6) adapting the procedure of [6] for tree GP, such that

[3] Conceptually, this is a method of biased sampling similar to chain-referral sampling, adopted extensively in sociological research [30]. Sampling via the mutation operator generates trees which partially span the local neighbourhood for each genotype.

$$d_P(p, p') = \frac{C(pp') - min(C(p), C(p'))}{max(C(p), C(p')} \tag{6}$$

where $C$ is a function giving the length in bits of the string representation, under UTF8, of the argument for a particular compressor. Each phenotype was decoded into the prefix string representing the corresponding encoded arithmetic expression. This expression excludes neutral nodes (junk) which do not contribute to the phenotype. Pair-wise application of the NCD gives a measure of similarity between phenotypes in the range of $\{0.0 : 1.0\} + \epsilon$, using the `gzip` algorithm (where $\epsilon \approx 0.1$, an error term induced because the compression is not ideal). To provide a controllable method of exploring the impact of changing locality in a representation, an intermediary bias $u_{\alpha\beta}$ was introduced to the uniform mutation operator:

$$u_{\alpha\beta}(p, p') = \left(1 + e^{-\alpha(d_P(p,p')-\beta)}\right)^{-1} \tag{7}$$

The bias $u_{\alpha\beta}$ was employed to change the expected mutation distance between each pair of genotypes and is a standard sigmoid function, adjusted by a scaling parameter $\alpha$ and translation $\beta$ respectively. For each application of the mutation operator to a genotype, $u_{\alpha\beta}$ defined the probability that a proposed set of mutations will be accepted. The process is repeated until an acceptable mutation is found and returned by the operator. To first order, this gives an adjusted expected variation distance, where

$$\bar{M}_{\alpha\beta}(p, p')^{-1} \approx u_{\alpha\beta}(p, p') \times u(p, p') \tag{8}$$

Hence by scaling $\alpha$, the likelihood that mutations will result in phenotypes which are syntactically similar can be defined. Varying the mutation bias equates to scaling the locality of the mapping. The threshold value assumed in the sigmoid function is set to an intermediate level of similarity, $\beta = 0.2$. An example (for one CGP sample) is given in Figure 1. The graphs are scatterplots, binned into hexagons, illustrating qualitatively the distribution observed between the log-scaled expected variation distance $\bar{M}$ and normalised compression distance $d_P$. The result is shown between two similar maps at low locality ($\alpha = 20, 10$) and at high locality ($\alpha = -10, -20$). This directly compares the change in locality induced by the bias. Inspecting the scatter graphs appears to indicate a weakly positive trend, apparent over short distances. This follows from the decreasing likelihood of making larger syntactic changes to the phenotype, under the uniform mutation operator. Relatively probable mutations, $\bar{M} \sim [0 - 20]$ correspond to smaller changes in compression distance, $d_P \sim [0.1 - 0.3]$. The majority of distances observed in the region $\bar{M} \sim [20 - 40]$ (between genotypes situated on lower branches of the sample) occur with lower probability and correspond to greater variation in syntactic change.

Using the Mantel test, we can validate these qualitative observations. Figure 2 shows the range of corresponding Mantel coefficients $r_M$ calculated over all samples, for linear correlation, as a function of distance. It can be inferred that an overall weak positive correlation exists in the CGP mapping, which falls off as

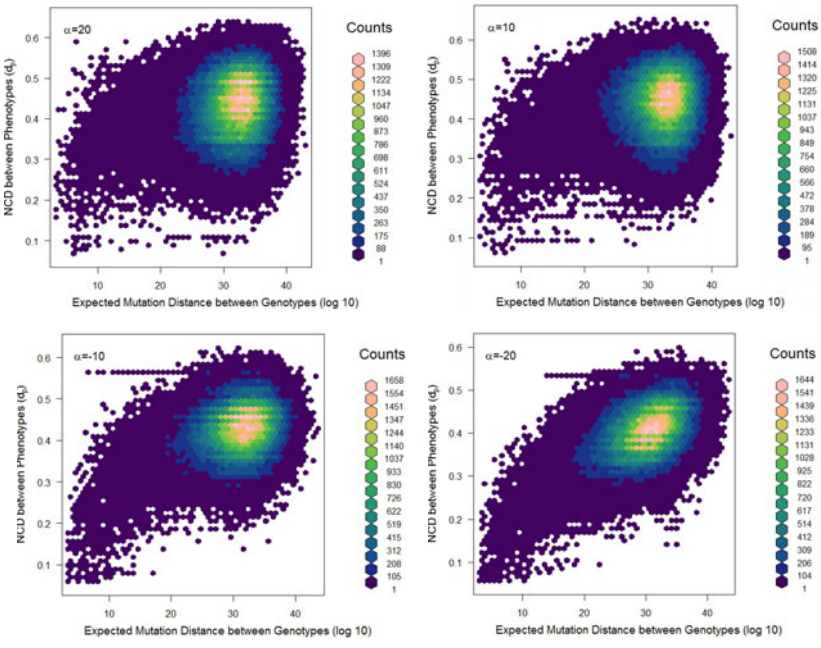


**Fig. 1.** Illustration of the effect of the NCD mutation bias on genotype-phenotype correlation. Top left: Lowest locality. Bottom right: Highest locality.

a function of genotype distance. A set of 1000 permutations was then generated for each distance matrix to test significance at $P(H_0) < 0.005$, for a set of 8 distance classes from $\bar{M} = 0.0 : 40.0$. The correlations found to be significant under permutation are labelled (*). Inclusion of the Mantel test therefore gives a firm basis from which to reject the null hypothesis and accept the correlation. The effects of the mutation bias are also apparent (contrast positive $\alpha$ with negative $\alpha$).

To explore the relationship between syntactic locality and performance, a preliminary experiment was carried out using a randomly generated selection of 38 symbolic regression problems. The problem instances were restricted in complexity to simple 5th order polynomials with integer coefficients in the range of {-2:2}. These are basic problems known to be solvable consistently using only the simple CGP representation analysed, without the requirement for additional features such as modularity. Five instances of each problem were considered, applying the mutation bias with $\alpha = \{-20, -10, 0, 10, 20\}$ and $\beta = 0.2$. Fitness was evaluated by deriving the euclidean distance over the set of uniformly distributed sample points. Other parameters (Table 1) were informed per common previous estimates in CGP [8]. The parameters have not been optimised to account for interaction with the mutation bias. Table 2 shows the corresponding probability of success $\eta$ at each locality level after 2000 generations, estimated in each instance from the fraction of 500 runs which successfully recovered the expression. For the 21 polynomial problems with an average success

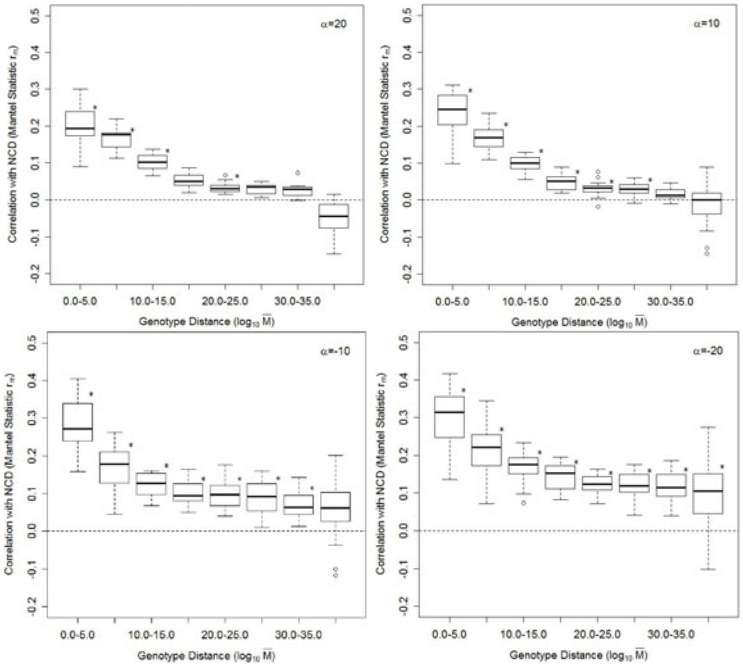


**Fig. 2.** Measured Mantel correlation for CGP with different levels of mutation bias. Top left: Lowest locality. Bottom right: Highest locality.

probability $\bar{\eta}$ greater than 10% (denoted with a †), a general tendancy can be observed towards better performance at lower levels of locality. Of this subset, in 19 of the 20 cases the probability of success was higher for $\alpha = 20$ than $\alpha = -20$. In the remaining cases with success probability below 10%, no measurable trend is observable outside of experimental error. All problems were solved successfully over at least one set of runs.

## 6 Discussion

The weak correlation observed between genotype and phenotype distances in Figures 1 and 2 is consistent with the variation in structure that small mutations can impose in this representation. Altering a single node connection in CGP may cause a large number of functions to be disconnected. Similarly, if the same node is connected to many neighbours, then adjusting it will produce a disproportionate change to the syntax of a program. There is therefore an overall tendancy for parents to produce syntactically similar offspring, but this is offset by the potential for large structural change. The relatively small impact of the mutation bias also suggests this relationship is difficult to suppress, given that it is a direct consequence of utilising a graph-based structure.

**Table 2.** CGP success probability $\eta$ with respect to locality

| Polynomial Expression | α | | | | | $\bar{\eta} \geq 0.1$? |
|---|---|---|---|---|---|---|
| | -20 | -10 | 0 | 10 | 20 | |
| $-1 - 2x^3 + x^4$ | 0.236 | 0.268 | 0.286 | 0.304 | 0.288 | † |
| $-2 - 2x - x^2 - x^4$ | 0.102 | 0.116 | 0.158 | 0.142 | 0.174 | † |
| $-2 - x^3$ | 0.254 | 0.314 | 0.386 | 0.404 | 0.450 | † |
| $-2x^2 + 2x^4 + 2x^5$ | 0.152 | 0.196 | 0.226 | 0.196 | 0.244 | † |
| $-2x^2 - 2x^3$ | 0.556 | 0.570 | 0.594 | 0.616 | 0.676 | † |
| $-2x^3 + x^4$ | 0.646 | 0.758 | 0.792 | 0.814 | 0.790 | † |
| $-2x^3 + x^5$ | 0.196 | 0.278 | 0.350 | 0.360 | 0.394 | † |
| $-x^2 + x^3 + 2x^4$ | 0.684 | 0.722 | 0.790 | 0.840 | 0.866 | † |
| $-x - x^4$ | 0.626 | 0.744 | 0.792 | 0.840 | 0.868 | † |
| $1 - 2x^2 - 2x^3$ | 0.556 | 0.570 | 0.594 | 0.616 | 0.676 | † |
| $1 - x + 2x^3 + x^4$ | 0.290 | 0.324 | 0.370 | 0.322 | 0.326 | † |
| $1 - x + x^2 - x^3$ | 0.766 | 0.840 | 0.878 | 0.894 | 0.924 | † |
| $1 + x^3 + 2x^5$ | 0.170 | 0.214 | 0.198 | 0.198 | 0.216 | † |
| $2 + 2x - x^2 - 2x^3$ | 0.228 | 0.226 | 0.260 | 0.248 | 0.272 | † |
| $2 + x^2 - x^3$ | 0.422 | 0.486 | 0.516 | 0.516 | 0.572 | † |
| $2x^2 - x^3 + 2x^4$ | 0.382 | 0.446 | 0.492 | 0.524 | 0.522 | † |
| $2x^2 - x^3 + x^4 - x^5$ | 0.198 | 0.256 | 0.232 | 0.290 | 0.282 | † |
| $2x^2 + x^3 - 2x^4 - x^5$ | 0.504 | 0.592 | 0.540 | 0.536 | 0.524 | † |
| $2x + x^2 - 2x^3$ | 0.202 | 0.216 | 0.206 | 0.230 | 0.276 | † |
| $x + 2x^2 - 2x^3$ | 0.610 | 0.672 | 0.642 | 0.670 | 0.684 | † |
| $-2 - 2x^2 - 2x^5$ | 0.034 | 0.042 | 0.038 | 0.014 | 0.028 | - |
| $-x + 2x + x^4 - 2x^5$ | 0.122 | 0.122 | 0.112 | 0.110 | 0.116 | † |
| $-2 + 2x - x^2 - 2x^3 + 2x^5$ | 0.002 | 0.006 | 0.000 | 0.000 | 0.000 | - |
| $-2 + x - x^2 - 2x^4 - 2x^5$ | 0.001 | 0.014 | 0.004 | 0.004 | 0.040 | - |
| $-2x^5 - 2x^3 - 2x^2 - 2x - 1$ | 0.016 | 0.016 | 0.004 | 0.040 | 0.008 | - |
| $-2x + 2x^3 + 2x^4 + 2x^5$ | 0.064 | 0.050 | 0.040 | 0.048 | 0.040 | - |
| $-2x + 2x^3 + x^4 - 2x^5$ | 0.006 | 0.020 | 0.006 | 0.006 | 0.006 | - |
| $-2x + 2x - 2x^2 - x^4 - 2x^5$ | 0.002 | 0.002 | 0.000 | 0.000 | 0.002 | - |
| $-2x + x^2 - 2x^3 + 2x^4 - 2x^5$ | 0.006 | 0.010 | 0.004 | 0.006 | 0.008 | - |
| $1 - 2x + x^2 - x^3 - x^5$ | 0.120 | 0.098 | 0.092 | 0.084 | 0.046 | - |
| $1 - x^3 + 2x^4 - 2x^5$ | 0.046 | 0.036 | 0.030 | 0.038 | 0.022 | - |
| $1 + 2x^2 - x^3 + 2x^4 - 2x^5$ | 0.020 | 0.044 | 0.024 | 0.018 | 0.006 | - |
| $1 + 2x - x^2 + 2x^5$ | 0.106 | 0.112 | 0.092 | 0.086 | 0.082 | - |
| $2 - 1x - 2x^2 + 2x^3 - x^4 - x^5$ | 0.008 | 0.002 | 0.002 | 0.002 | 0.002 | - |
| $2 + 2x^2 + x^3 - 2x^4 - x^5$ | 0.070 | 0.068 | 0.050 | 0.048 | 0.042 | - |
| $2 + x - 2x^2 - 2x^3 + x^5$ | 0.038 | 0.034 | 0.030 | 0.030 | 0.042 | - |
| $2 + x - x^2 - x^4 - x^5$ | 0.052 | 0.024 | 0.032 | 0.032 | 0.028 | - |
| $2 + x - x^2 + x^4 - x^5$ | 0.052 | 0.024 | 0.032 | 0.032 | 0.028 | - |

The trend of the symbolic regression results implies, somewhat counter-intuitively, that higher levels of correlation between genotype and phenotype distance tended to produce poorer performance in CGP. We consider three feasible explanations. Firstly, it is likely that the constraints imposed by high locality have restricted the diversity of the search, which may render intermediary schema difficult to reach. Secondly, in Rothlauf's model of locality, poorer performance under higher locality can be associated with fitness landscapes which are misleading [1] or deceptive (for example, GA trap functions). Further investigation of the fitness landscapes for these specific problem instances would be required to determine whether this is the case for this representation. Thirdly, it is unclear how features of the CGP genotype-phenotype map not addressed here, such as high redundancy, or structural bias [26] contribute to the trend.

In practice, using locality as a general performance predictor, or as a method of directly tuning existing genotype-phenotype maps, is clearly a challenging issue in integer encoded GP. Bypassing the intermediary stage and relating genotype and fitness values may lead to better outcomes on individual problems, but provides limited guidance for improving GP representations in general. Despite these outstanding problems, from this initial work we anticipate that the Mantel

test will prove a useful addition to existing approaches in statistical analyses of GP genotype-phenotype maps. It is encouraging that a technique founded in ecology can also contribute to the study of a complex artificial system.

## 7 Conclusions

The Mantel test, a statistical technique from numerical ecology, was adopted to analyse the locality of GP maps. As a case study, we examined an established integer representation, CGP. It was observed that a weakly positive correlation exists for CGP over short genotype distances, when using arithmetic function sets. We introduced a method of scaling the locality of a CGP genotype-phenotype map, by providing a bias into the mutation operator based on the normalised compression distance between phenotypes. To our knowledge, this is the first instance of explicitely controlled locality explored within a graph-based representation. The effect of varying locality on performance was measured for randomly generated polynomial symbolic regression problems. Higher locality was associated with reduced performance over 19 instances. We infer that employing less local maps may be advantageous on these classes of problem. In the future, we intend to test the robustness of this approach by applying it to other non-standard genotype-phenotype maps, such as grammar GP. Direct comparisons with alternative locality measures would also be appropriate.

**Acknowledgements.** Particular thanks are due to Dr. Dan Franks and other members of the York Centre for Complex Systems Analysis for advice concerning the Mantel statistic.

## References

[1] Rothlauf, F.: Representations for Genetic and Evolutionary Algorithms, pp. 33–96. Springer, Heidelberg (2006)
[2] Gottlieb, J.: Empirical Analysis of Locality, Heritability and Heuristic Bias in Evolutionary Algorithms: A Case Study for the Multidimensional Knapsack Problem. Evolutionary Computation 43, 441–475 (2004)
[3] Gen, M., Cheng, R.: Genetic Algorithms and Engineering Optimisation. John Wiley and Sons, Inc. (2000)
[4] Rothlauf, F., Goldberg, D.E.: Pruefer Numbers and Genetic Algorithms: A Lesson on How the Low Locality of an Encoding Can Harm the Performance of GAs. In: Deb, K., Rudolph, G., Lutton, E., Merelo, J.J., Schoenauer, M., Schwefel, H.-P., Yao, X. (eds.) PPSN 2000. LNCS, vol. 1917, pp. 395–404. Springer, Heidelberg (2000)
[5] Droste, S., Wiesmann, D.: On Representation and Genetic Operators in Evolutionary Algorithms. Technical report (SFB) 531: [249], Univ. of Dortmund (1998)
[6] Galván-López, E., McDermott, J., Brabazon, A.: Defining locality as a problem difficulty measure in genetic programming. Genetic Programming and Evolvable Machines, 1–37 (2011)
[7] Oltean, M., Grosnan, C., Diosan, L., Mihaila, C.: Genetic Programming with Linear Representation: A Survey. Int. J. on Artificial Intelligence Tools, 197–238 (2008)
[8] Miller, J.F. (ed.): Cartesian Genetic Programming. Springer, Heidelberg (2011)

[9] Rothlauf, F., Oetzel, M.: On the Locality of Grammatical Evolution. In: Collet, P., Tomassini, M., Ebner, M., Gustafson, S., Ekárt, A. (eds.) EuroGP 2006. LNCS, vol. 3905, pp. 320–330. Springer, Heidelberg (2006)
[10] Fagan, D., O'Neill, M., Galván-López, E., Brabazon, A., McGarraghy, S.: An Analysis of Genotype-Phenotype Maps in Grammatical Evolution. In: Esparcia-Alcázar, A.I., Ekárt, A., Silva, S., Dignum, S., Uyar, A.Ş. (eds.) EuroGP 2010. LNCS, vol. 6021, pp. 62–73. Springer, Heidelberg (2010)
[11] Brameier, M.F., Banzhaf, W.: Linear Genetic Programming. Genetic and Evolutionary Computation. Springer, Heidelberg (2007)
[12] Mantel, N.: The detection of disease clustering and a generalized regression approach. Cancer Research 27, 209–220 (1967)
[13] Dietz, E.J.: Permutation tests for association between two distance matrices. Systematic Zoology 32, 21–26 (1983)
[14] Oden, N.L., Sokal, R.R.: Directional autocorrelation: an extension of spatial correlograms to two dimensions. Systematic Biology 35, 608 (1986)
[15] Legendre, P., Fortin, M.-J.: Spatial pattern and ecological analysis. Vegetatio 80, 107–138 (1989)
[16] Legendre, P., Lapointe, F.J., Cagrain, P.: Modeling brain evolution from behavior: a permutational regression approach. Evolution 48, 1487–1499 (1994)
[17] Lichstein, J.W.: Multiple regression on distance matrices: a multivariate spatial analysis tool. Plant Ecology 188, 117–131 (2006)
[18] Legendre, P., Fortin, M.-J.: Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Molecular Ecology Resources, 831–844 (2010)
[19] Chiam, S.C., Tan, K.C., Goh, C.K., Al Mamun, A.: Improving locality in binary representation via redundancy.. IEEE Trans. on Sys. Man. and Cybernetics (B) 38, 808–825 (2008)
[20] McDermott, J., Galván-Lopéz, E., O'Neill, M.: A Fine-Grained View of GP Locality with Binary Decision Diagrams as Ant Phenotypes. In: Schaefer, R., Cotta, C., Kołodziej, J., Rudolph, G. (eds.) PPSN XI. LNCS, vol. 6238, pp. 164–173. Springer, Heidelberg (2010)
[21] Krawiec, K.: Semantically Embedded Genetic Programming. In: Genetic and Evolutionary Computation Conference, Dublin, Ireland, pp. 1379–1386 (2011)
[22] Jones, T., Forrest, S.: Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In: Proc. of the 6th Int. Conference on Genetic Algorithms, vol. 129, pp. 184–192. Citeseer (1995)
[23] Legendre, P., Legendre, L.: Numerical Ecology, 2nd edn. Developments in Environmental Modelling. Elsevier (1998)
[24] Goslee, S.C., Urban, D.L.: The ecodist Package for Dissimilarity-based Analysis of Ecological Data. Journal Of Statistical Software 22 (2007)
[25] Hien, N.T., Hoai, N.X.: A Brief Overview of Population Diversity Measures in Genetic Programming. In: 3rd Asian-Pacific Workshop on Genetic Programming, pp. 128–139 (2006)
[26] Payne, A.J., Stepney, S.: Representation and Structural biases in CGP. In: IEEE Congress on Evolutionary Computation, vol. 8, pp. 1064–1071. IEEE (2009)
[27] Vanneschi, L.: Theory and Practice for Efficient Genetic Programming. PhD thesis, Univ. of Lausanne (2004)
[28] Keijzer, M.: Efficiently Representing Populations in Genetic Programming. In: Advances in Genetic Programming, vol. 2, pp. 259–278. MIT Press (1996)
[29] Vanneschi, L.: Crossover-Based Tree Distance in Genetic Programming. IEEE Transactions on Evolutionary Computation 12, 506–524 (2008)
[30] Biernacki, P., Waldorf, D.: Snowball Sampling: Problems, Techniques and Chain-Referral Sampling. Socio. Methods And Research 10, 141–163 (1981)