Allele Diffusion in Linear Genetic Programming and Variable-Length Genetic Algorithms with Subtree Crossover

Riccardo Poli[1], Jonathan E. Rowe[2], Christopher R. Stephens[3], and Alden H. Wright[4]

[1] Department of Computer Science, University of Essex, UK
rpoli@essex.ac.uk
[2] School of Computer Science, The University of Birmingham, UK
j.e.rowe@cs.bham.ac.uk
[3] Instituto de Ciencias Nucleares, UNAM, Mexico
stephens@nuclecu.unam.mx
[4] Computer Science Department, University of Montana, USA
wright@cs.umt.edu

Abstract. In this paper we study, theoretically, the search biases produced by GP subtree crossover when applied to linear representations, such as those used in linear GP or in variable length GAs. The study naturally leads to generalisations of Geiringer's theorem and of the notion of linkage equilibrium, which, until now, were applicable only to fixed-length representations. This indicates the presence of a diffusion process by which, even in the absence of selective pressure and mutation, the alleles in a particular individual tend not just to be swapped with those of other individuals in the population, but also to diffuse *within* the representation of each individual. More precisely, crossover attempts to push the population towards distributions of primitives where each primitive is equally likely to be found in any position in any individual.

1 Introduction

Schemata are sets of points in a search space sharing some syntactic feature. For example, in the context of GAs operating on binary strings, the syntactic representation of a schema is usually a string of symbols from the alphabet {0,1,*}, where the character * is interpreted as a "don't care" symbol. Typically schema theorems are descriptions of how the number of members of the population belonging to a schema vary over time. If $\alpha(H,t)$ denotes the probability that at time t a newly created individual samples (or matches) the schema H, which we term the *total transmission probability* of H, then an exact schema theorem for a generational system is simply

$$E[m(H,t+1)] = M\alpha(H,t), \tag{1}$$

where M is the population size, $m(H,t+1)$ is the number of individuals sampling H at generation $t+1$ and $E[\cdot]$ is the expectation operator. Holland's [4] and other (e.g. [15]) worst-case-scenario schema theories normally provide a lower bound for $\alpha(H,t)$ or, equivalently, for $E[m(H,t+1)]$. Only recently schema theorems which provide the exact value for $\alpha(H,t)$ have become available for fixed-length GAs with one-point crossover and mutation [24, 25] and other homologous crossovers [23]. Even more recent is the development of exact schema theorems for variable-length GAs, linear GP and tree-based GP. These now cover a variety of crossover and mutation operators including one-point crossover [11, 10, 12], standard and other subtree-swapping crossovers [13, 17, 7], different types of subtree mutation and headless chicken crossover [16, 8], and, finally, the class of homologous crossovers [18].

Exact schema theorems provide probabilistic models of the expected behaviour of a GA or a GP system which can be used to understand the system and study its behaviour. This can be done either through simulation (i.e., by "running" the equations) or through mathematical analysis. Although exact GP schema equations have become available only very recently, early studies indicate their utility. For example, simulations and analyses of exact GP schema equations for the case of linear, variable-length representations under different crossover and mutation operators [17, 7, 9, 20] indicate that they can provide a deeper understanding of emergent phenomena such as bloat [6, 21, 5].

In general, the availability of exact models for different operators facilitates the formal study of the biases of those operators. Knowledge of these biases is very important because these biases can interfere (but not necessarily in a negative way) with the intended bias of selection. So, this knowledge allows for a better informed choice of operators, parameter settings and even initialisation strategies for particular problems. For example, the knowledge of the biases of the operators obtained from exact schema theories allows one to initialise the population so as to minimise the biases of the operators in the early generations — a particularly important stage in a run. Steps in this direction have recently been made in [17, 7, 9], where a particular type of Gamma program-length distribution has been shown to present minimum length biases for variable-length linear systems under GP subtree crossover (we will review this particular result later).

In this paper we continue the study of the biases of the standard subtree-swapping GP crossover operator for the case where it is applied to linear structures, such as the ones used in linear GP and in variable-length GAs. In the case of linear structures, standard GP crossover involves randomly selecting two crossover points, one in each parent, and producing the offspring by swapping the substrings to either the left or the right hand side of the crossover points.

Our study is based on the use of exact schema evolution equations and on the analysis of their fixed points and naturally leads to a generalisation of Geiringer's theorem, and of the notion of linkage equilibrium. Both of these concepts, until now, were applicable only to fixed-length representations. This characterises a diffusion process by which, even in the absence of selective pressure and mutation, alleles drift not just between individuals, but also between positions within individuals. This means that uniform populations are not necessarily fixed-points for GP, unlike the fixed-length GA case.

The paper is organised as follows. We provide some background information on the GP schema theory for subtree crossover acting on linear structures in Section 2. We describe Geiringer's theorem for fixed-length GAs in Section 3 and introduce our extension of Geiringer's theorem in Section 4. Experimental results backing up the theory are presented in Section 5. We discuss our results in Section 6 and, finally, we draw some conclusions in Section 7.

2 GP Schema Theory Background

Syntactically a GP schema is a tree composed of functions from the set $\mathcal{F} \cup \{=\}$ and terminals from the set $\mathcal{T} \cup \{=\}$, where $\mathcal{F}$ and $\mathcal{T}$ are the function and terminal sets used in a GP run. The primitive $=$ is a "don't care" symbol which stands for a *single* terminal or function. A schema H represents the set of all programs having the same shape as H and the same non-$=$ nodes as H.

As discussed in [17], when only unary functions are used in GP, schemata (and programs) can only take the form $(h_1(h_2(h_3....(h_{N-1}h_N)....)))$ where $N > 0$, $h_i \in \mathcal{F} \cup \{=\}$ for $1 \leq i < N$, and $h_N \in \mathcal{T} \cup \{=\}$. Therefore, they can be written unambiguously as strings of symbols of the form $h_1h_2h_3....h_{N-1}h_N$. In order to make the notation more compact, in the following we will represent repeated symbols in a string using the power notation where x^y means x repeated y times. Particularly important for the GP schema theory are schemata containing "don't care" symbols only, since they represent all the programs of a particular shape. Using the power notation they can be represented as $(=)^N$ for any $N > 0$.

In [17] we proved that the total transmission probability for a linear GP schema of the form $h_1...h_N$ under standard crossover with uniform selection of the crossover points and no mutation can be written in the following form

$$\alpha(h_1...h_N,t) = (1-p_{xo})p(h_1...h_N,t) + \\ p_{xo}\sum_{k>0}\frac{1}{k}\sum_{i=0}^{\min(N,k)-1} p(h_1...h_i(=)^{k-i},t)\sum_{n=N-i}^{\infty}\frac{p((=)^{n-N+i}h_{i+1}...h_N,t)}{n}, \tag{2}$$

where p_{xo} is the crossover probability and $p(H,t)$ is the selection probability of the schema H. In fitness proportionate selection $p(H,t) = m(H,t)f(H,t)/(M\bar{f}(t))$, where $m(H,t)$ is the number of individuals matching the schema H at time t, $f(H,t)$ is their mean fitness, $\bar{f}(t)$ is the mean fitness of the individuals in the population and M is the population size. In the equation each k represents the length of a first parent, each n represents the length of a second parent, and each i is a valid crossover point.

This equation can be used to study, among other things, the evolution of size in linear GP/GA systems. This is because it can be specialised to describe the transmission probability of schemata of the form $(=)^N$. The quantity $\alpha(H,t)$ represents a probability. However, for an infinite population $\alpha(H,t)$ can also be interpreted as the proportion of the population matching schema H at generation $t+1$, a quantity that we will denote with $\Phi(H,t+1)$. Also, we should note that if the fitness landscape is flat, then $p(H,t) = \Phi(H,t)$. So, under the assumptions of infinite population and flat landscape the specialisation of Equation 2 to schemata of the form $(=)^N$ leads to the following length-evolution equation:

$$\Phi((=)^N,t) = (1-p_{xo})\Phi((=)^N,t) \\ + p_{xo}\sum_{k}\sum_{i=0}^{\min(N,k)-1}\frac{\Phi((=)^k,t)}{k}\sum_{n=N-i}^{\infty}\frac{\Phi((=)^n,t)}{n}. \tag{3}$$

In [17] we showed both empirically and mathematically that, in these conditions, a family of fixed-point distributions of lengths exists and that this family is the following family of discretised Gamma distributions

$$\Phi((=)^N, t) = Nr^{N-1}(r-1)^2, \tag{4}$$

where $r = (\mu - 1)/(\mu + 1)$ and μ is the mean length of the individuals in the population. We also proved that the mean size of the programs at generation $t+1$, $\mu(t+1)$, in a linear GP system with standard crossover, uniform selection of the crossover points, no mutation and an infinite population is

$$\mu(t+1) = \sum_N Np((=)^N, t) \tag{5}$$

and, therefore, that on a flat landscape,

$$\mu(t+1) = \mu(t). \tag{6}$$

For alternative proofs of some of these results and other related results, such as a time evolution equation and a fixed point for the variance of the length distribution, see [20].

3 Geiringer's theorem

In this section we briefly introduce Geiringer's theorem [3], an important result with implications both for natural population genetics and evolutionary algorithms [1, 2, 22]. Geiringer's theorem indicates that, in a population of fixed-length chromosomes repeatedly undergoing crossover (in the absence of mutation and selective pressure), the probability of finding a generic string $h_1 h_2 \cdots h_N$ approaches a limit distribution which is only dependent on the distribution of the alleles h_1, h_2, etc. in the initial generation. More precisely, if $\Phi(h_1 h_2 \cdots h_N, t)$ is the proportion of individuals of type $h_1 h_2 \cdots h_N$ at generation t (i.e. $\Phi(h_1 h_2 \cdots h_N, t) = m(h_1 h_2 \cdots h_N, t)/M$) and $\Phi(h_i, t)$ is the proportion of individuals carrying allele h_i then

$$\lim_{t \to \infty} \Phi(h_1 h_2 \cdots h_N, t) = \prod_{i=1}^{N} \Phi(h_i, 0). \tag{7}$$

This result is valid for all homologous crossover operators which allow any two loci to be separated by recombination. Strictly speaking the result is valid only for infinite populations.

If one interprets $\Phi(h_1 h_2 \cdots h_N, t)$ as a probability distribution of the possible strings in the population, we can interpret Equation 7 as saying that such a distribution is converging towards independence. When at a particular generation t the frequency of any string in a population $\Phi(h_1 h_2 \cdots h_N, t)$ equals $\prod_{i=1}^{N} \Phi(h_i, t)$, the population is said to be in *linkage equilibrium* or *Robbins' proportions*.

It is trivial to generalise Geiringer's theorem to obtain the expected fixed-point proportion of a generic linear fixed-length GA schema H for a population undergoing crossover:

$$\lim_{t \to \infty} \Phi(H, t) = \prod_{i \in \Delta(H)} \Phi(*^{i-1} h_i *^{N-i}, 0), \tag{8}$$

where $\Delta(H)$ is the set of indices of the defining symbols in H, h_i is one such defining symbols and we used the power notation x^y to mean x repeated y times. (Note that $\Phi(*^{i-1} h_i *^{N-i}, t)$ coincides with the frequency of allele h_i, $\Phi(h_i, t)$.)

4 A Geiringer-theorem-type result for linear GP representations and subtree crossover

A full extension of Geiringer's theorem to linear, variable-length structures and standard GP crossover would require two steps: (a) proving that, in the absence of mutation and of selective pressure and for an infinite population, a distribution $\Phi(h_1 h_2 \cdots h_N, t)$, where the alleles can be considered independent stochastic variables, is a fixed point, and (b) showing that the system indeed moves towards that fixed point. In this paper we prove (a) mathematically and provide experimental evidence for (b).

Our objective is to identify the fixed point to which an infinite linear GP population converges under the effect of crossover only (i.e. on a flat landscape). Instead of just providing the equation for the fixed point and proving that it is indeed a fixed point, we prefer to describe the reasoning that led us to guess the form of the fixed point since this better illustrates its meaning.

Imagine a population of strings of different lengths and focus attention on a particular non-terminal allele a at a particular locus l of a particular string s (we will consider the case of terminal alleles later). Subtree crossover allows for the migration of allele a to different strings, for example to strings of length different from the length of s. So, subtree crossover promotes a process of "diffusion" of alleles between different length classes. Unlike the case of homologous crossover in fixed length strings, in general, this process does not keep the alleles in their original position, i.e. allele a might migrate to loci different from l. Because of this, in repeated applications of crossover, a copy of the allele can be placed back into the original string s (which may now have a different length and allele composition) but at a different locus, effectively creating a sort of gene duplication (indeed unequal crossing over seems to be the mechanism of gene duplication in nature [19]). So, subtree crossover also promotes another type of allele diffusion: diffusion within length classes. Put another way, crossover is trying to spread each non-terminal allele as thinly as possible over every non-terminal locus available in the population.

Let us calculate the total number, $n(a,t)$, of non-terminal alleles of type a in a population of size M at generation t. With the "don't care" symbol $\bar{a}$ we denote any non-terminal allele different from a, while with $\boxed{\bar{a}}$ we denote any sequence of $\bar{a}$'s, including the empty sequence. The "don't care" symbol "=" will be used to represent any terminal symbol. Then we can write

$$\begin{aligned}
n(a,t) &= M\big(1 \times \Phi(a =, t) + 2 \times \Phi(aa =, t) + 1 \times \Phi(\bar{a}a =, t) + 1 \times \Phi(a\bar{a} =, t) \\
&\qquad +3 \times \Phi(aaa =, t) + 2 \times \Phi(\bar{a}aa =, t) + 2 \times \Phi(a\bar{a}a =, t) + \ldots\big) \\
&= M\big(\Phi(\boxed{\bar{a}}a\boxed{\bar{a}} =, t) + \Phi(\boxed{\bar{a}}a\boxed{\bar{a}}a\boxed{\bar{a}} =, t) + \Phi(\boxed{\bar{a}}a\boxed{\bar{a}}a\boxed{\bar{a}}a\boxed{\bar{a}} =, t) + \ldots\big) \\
&= M \sum_{n \geq 1} \Phi((\boxed{\bar{a}}a)^n \ \boxed{\bar{a}} =, t).
\end{aligned}$$

So, whether the population is finite or infinite the expected number of non-terminal alleles of type a per individual is $\sum_{n \geq 1} \Phi((\boxed{\bar{a}}a)^n \ \boxed{\bar{a}} \ =, t)$, which we expect not to vary with t. However, as discussed in Section 2 the mean program length, $\mu(t)$, is also expected to be time independent. So, the average number of non-terminal alleles of type

a per non-terminal locus is:

$$c(a) = \frac{\sum_{n\geq 1} \mathit{\Phi}((\boxed{a}a)^n \ \boxed{a} =, 0)}{\mu(0) - 1}. \tag{9}$$

If we assume that there is no length bias in the choice of the crossover points and that the two crossover points are chosen independently, then after some time crossover will have mixed the alleles sufficiently so that the presence of a given non-terminal allele in a given locus is independent of any other non-terminal allele and locus. We also expect that the probability of finding an allele of type a at a generic locus within a string of any length will be constant and equal to $c(a)$.

Let us now consider the effects of crossover on terminal alleles and loci. Again, let us focus our attention on terminal allele a in a string s of length l. Clearly allele a occupies locus l in s, but we expect that crossover will sooner or later move a to strings of length different from l. So, also for terminal alleles there is a diffusion effect which promotes their migration to strings of different length. However, it is impossible to obtain more than one copy of a terminal in a particular string because terminal alleles can only occupy the terminal locus. So, there cannot be a diffusion process of terminal alleles within a length class.

Let us calculate the total number, $n(a, t)$, of terminal alleles of type a in a population of size M at generation t:

$$\begin{aligned} n(a,t) &= M\big(1 \times \mathit{\Phi}(a,t) + 1 \times \mathit{\Phi}(= a,t) + 1 \times \mathit{\Phi}(== a,t) + \ldots\big) \\ &= M \sum_{n\geq 0} \mathit{\Phi}((=)^n a, t). \end{aligned}$$

So, whether the population is finite or infinite the expected number of terminal alleles of type a per individual is $\sum_{n\geq 0} \mathit{\Phi}((=)^n a, t)$, which, again, we expect not to vary with t. Because there is only one terminal locus per individual, the average number of terminal alleles of type a per terminal locus is:

$$c(a) = \sum_{n\geq 0} \mathit{\Phi}((=)^n a, 0). \tag{10}$$

We can assume that after some time, crossover will have mixed the alleles sufficiently so that the probability of finding a given terminal allele in a string is independent of the string length and is equal to $c(a)$.

The independence arguments above mean that the conditional probability of finding a specific string within the class of strings of length N will be described by the following probability distribution

$$\Pr\{h_1 h_2 \ldots h_N | \text{length} = N\} = \prod_{i=1}^{N} c(h_i). \tag{11}$$

This result allows us to calculate the fixed point proportion of strings of type $h_1 h_2 \ldots h_N$

$$\mathit{\Phi}(h_1 h_2 \ldots h_N, \infty) = \mathit{\Phi}((=)^N, \infty) \times \prod_{i=1}^{N} c(h_i), \tag{12}$$

where $\Phi((=)^N, \infty)$ is the discrete gamma distribution given in Equation 4, which represents the fixed point for length evolution. If we extend the definition of $c(a)$ to accept the argument "=" (by setting $c(=) = 1$), it is easy to see that Equation 12 is also valid for schemata.

The arguments reported above led us to the following

Theorem 1. *A fixed point distribution for the proportion of a linear, variable-length schema $h_1 h_2 \cdots h_N$ under subtree crossover for an infinite population initialised at the fixed point length distribution operating on a flat fitness landscape is given in Equation 12.*

Proof. Since the fitness landscape is flat, $p(H,t) = \Phi(H,t)$ for any schema. Also, because the population is infinite, $\alpha(H,t) = \Phi(H,t+1)$. So, Equation 2 becomes

$$\begin{aligned}\Phi(h_1...h_N, t+1) &= (1-p_{xo})\Phi(h_1...h_N, t) + \\ &p_{xo} \sum_k \frac{1}{k} \sum_{i=0}^{\min(N,k)-1} \Phi(h_1...h_i(=)^{k-i}, t) \sum_{n=N-i}^{\infty} \frac{\Phi((=)^{n-N+i} h_{i+1}...h_N, t)}{n}.\end{aligned} \tag{13}$$

We can prove that Equation 12 is a fixed point for this equation, by substituting the right-hand side of Equation 12 into the right-hand side of this equation and then showing that the resulting expression for $\Phi(h_1...h_N, t+1)$ has exactly the same form as the right-hand side of Equation 12.

From the substitution we obtain:

$$\begin{aligned}\Phi(h_1...h_N, t+1) &= (1-p_{xo})\Phi((=)^N, \infty) \prod_{i=1}^{N} c(h_i) + \\ &p_{xo} \sum_k \frac{1}{k} \sum_{i=0}^{\min(N,k)-1} \Phi((=)^k, \infty) \prod_{\iota=1}^{i} c(h_\iota) \sum_{n=N-i}^{\infty} \frac{1}{n} \Phi((=)^n, \infty) \prod_{\iota=i+1}^{N} c(h_\iota) \\ &= \prod_{i=1}^{N} c(h_i) \Bigg((1-p_{xo})\Phi((=)^N, \infty) + \\ &\qquad p_{xo} \sum_k \frac{1}{k} \sum_{i=0}^{\min(N,k)-1} \Phi((=)^k, \infty) \sum_{n=N-i}^{\infty} \frac{1}{n} \Phi((=)^n, \infty) \Bigg).\end{aligned}$$

Note that the large factor in parentheses is entirely equivalent to the right-hand side of Equation 3. Because by hypothesis $\Phi((=)^N, \infty)$ is a fixed point for the length distribution, then the factor must be equivalent to $\Phi((=)^N, \infty)$ and so

$$\Phi(h_1...h_N, t+1) = \Phi((=)^N, \infty) \prod_{i=1}^{N} c(h_i), \tag{14}$$

which proves that Equation 12 is a fixed point for the distribution of strings.

It is interesting to rewrite Equation 12 is a slightly different form. If $\nu(h_1 h_2 \ldots, a)$ represents the number of times symbol a appears in the string or schema $h_1 h_2 \ldots$, and

$\mathcal{C}$ represents our primitive set then

$$\Phi(h_1 h_2 \dots h_N, \infty) = \Phi((=)^N, \infty) \prod_{a \in \mathcal{C}} (c(a))^{\nu(h_1 h_2 \dots, a)}. \tag{15}$$

So, for example if $\mathcal{C} = \{\sqrt{}, \sin, x, y, z\}$ and the population is initialised so that $c(\sqrt{}) = c(\sin) = 1/2$ and $c(x) = 1/3$, then $\Phi(\sqrt{\sqrt{\sin x}}, \infty) = (1/2)^2 \times (1/2) \times (1/3) = 1/24$. Interestingly, in the case of a binary alphabet, the probability of sampling a given string is only a function of the unitation value (the number of ones) of the string.

As we have not provided a formal proof that the length distribution converges towards a discrete-Gamma fixed point[1] we cannot prove that our theorem holds if the length distribution is other than at its fixed point.

It is also important to note that, in the absence of a proof of the stability of the family of fixed points in Equation 12, we cannot rigorously claim that any population will always converge to an independent allele distribution. However, the arguments preceeding the theorem can be considered as an informal proof of convergence since they consider the mixing/diffusion effects of crossover over a number of generations. Additionally, the experimental results described in the following section strongly corroborate these conjectures.

5 Experimental results

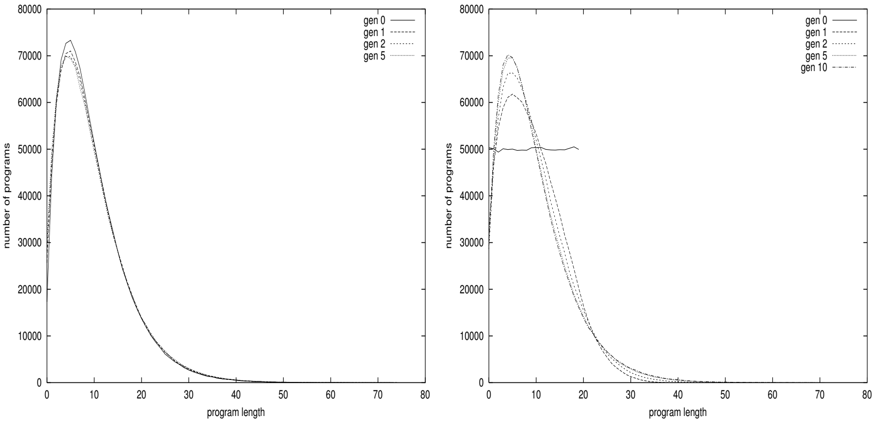
In order to check the theoretical results in this paper we set up a population of variable length strings consisting of 1,000,000 individuals. All individuals had the same terminal allele, 0, while two types of non-terminal alleles were used: alleles of type 0 and alleles of type 1. The majority of alleles were of type 0 and represented a “background” against which alleles of type 1 could be more easily traced. Alleles of type 1 are a “contrast medium” inoculated in the representation for the purpose of studying the diffusion of non-terminal alleles. Initially, alleles of type 1 were restricted to appear at only one specific non-terminal locus (which was varied between experiments). All strings which included that locus had nodes of type 1 at that locus. All other loci were occupied by alleles of type 0.

In our experiments we used two different initial length distributions: a distribution closely resembling a Gamma distribution with mean 10.5, and a uniform distribution with the same mean. Each population was run for 100 generations. The system was a generational GP/GA system with subtree crossover applied with 100% probability and a flat fitness landscape. Multiple independent runs were not required since the population size was set to be sufficiently large so as to remove any significant statistical variability and therefore to approximate the infinite-population behaviour (for each program length we had tens of thousands of individuals on average).

We start by checking what happens to the length distribution over time. Figure 1(a) shows that the distribution of program length is indeed at a fixed point when the population is initialised using a discrete Gamma distribution. In the figure the small variations in the plots for the first and second generations are due to the slight inaccuracy

[1] However, in [17] experimental evidence was provided that corroborates the hypothesis that a discrete gamma length distribution is asymptotically approached when the landscape is flat and the population is large.

of our Gamma-deviate generation algorithm. Only the first few generations are shown because the plots for later generations simply coincide with the plot for generation 5. Figure 1(b) illustrates how quickly the system converges to a discrete Gamma distribution even when initialised using a radically different distribution (in this case a uniform length distribution).

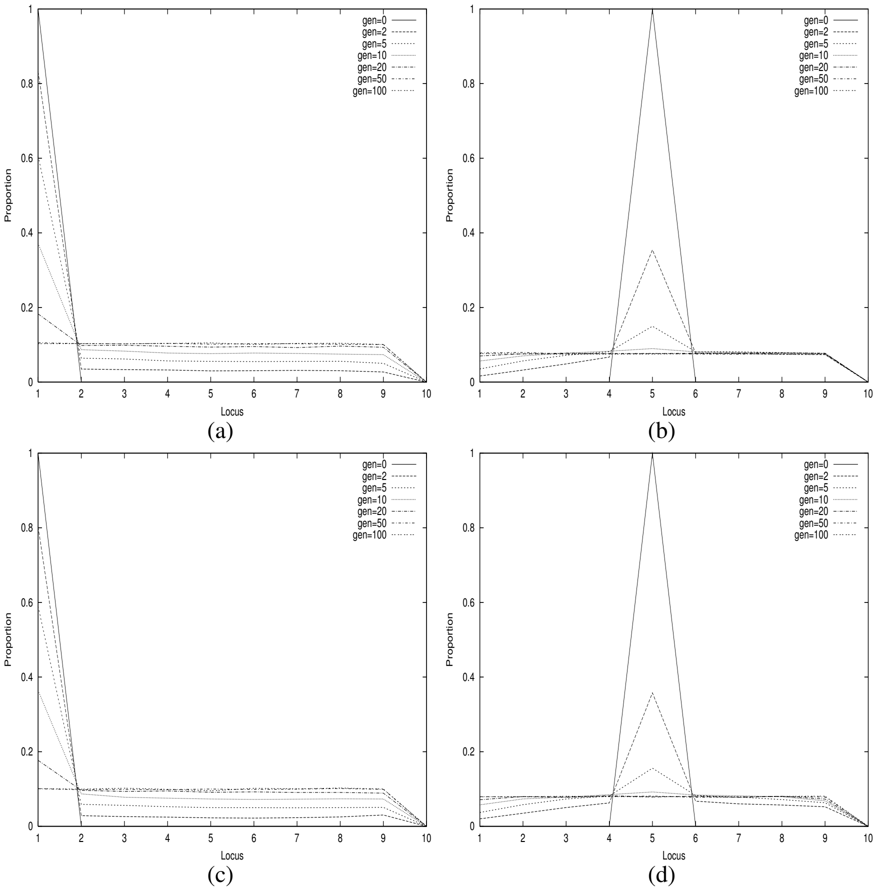


(a) (b)

Fig. 1. Plots of the number of programs vs. program length for different generations for a population of 1,000,000 individuals initialised with an approximately-Gamma (a) and an approximately uniform (b) length distribution.

Considering now the allele dynamics: Figure 2(a) shows how the distribution of alleles of type 1 varies within programs of length 10 over a number of generations in a population initialised at the Gamma-distribution length fixed-point. In the initial generation all non-terminal loci at position 1 were occupied by alleles of type 1 (i.e. only programs of length 1 did not include any 1's). So, our "contrast medium" was maximally isolated within the representation. Nonetheless, the "lateral" diffusion process very quickly spreads the "dye" and, within 20 generations or so, the distribution of alleles becomes uniform (since no terminal 1's were allowed, the proportion of 1's in locus 10 was always 0). At generation 100 the value of the proportion of alleles of type 1 averaged over the non-terminal loci was 0.103046, which, as shown in the first row of Table 1, matches very closely the value predicted by the theory on the basis of the frequency of non-terminal alleles of type 1 at generation 0.

Figure 2(b) shows what happens if we initialise the population so that all non-terminal loci at position 5 are occupied by alleles of type 1. Again, the initial length distribution is approximately a discrete Gamma. In this case the diffusion of alleles of type 1 is even faster. At generation 100 the proportion of alleles of type 1 averaged over the non-terminal loci is 0.075467. This is lower than the value in the previous paragraph because the average frequency of non-terminal 1's at generation 0 was lower (programs of length 1, 2, 3, 4 and 5 did not include any 1's). As before (see the second row of Table 1) this matches very closely our theoretical prediction based on the allele frequencies at generation 0.



(a) (b)

(c) (d)

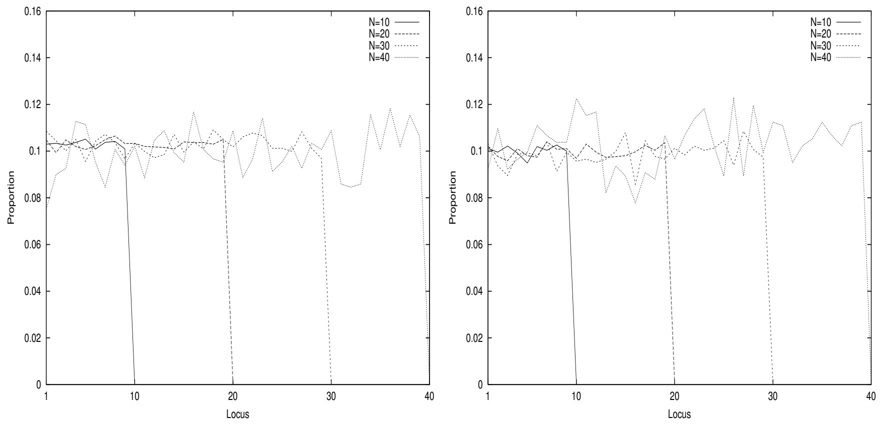
Fig. 2. Plots of the relative proportion of non-terminal alleles of type 1 vs. locus position within programs of length 10 for different generations. The population was initialised with an approximately Gamma length distribution in (a) and (b), and a uniform length distribution in (c) and (d). In the initial generation the alleles of type 1 were confined to non-terminal locus 1 in (a) and (c), and to non-teminal locus 5 in (b) and (d).

Table 1. Comparison between empirically measured and theoretical value of the fixed point frequency of non-terminal 1's, $c(1)$.

Initial distribution	Locus of initial 1's	Measured Frequency	Theoretical Frequency
gamma	1	0.103046	0.103444
gamma	5	0.075467	0.077442
gamma	10	0.042532	0.042695
gamma	15	0.020574	0.020963
uniform	1	0.100201	0.099964
uniform	5	0.079757	0.078857
uniform	10	0.052157	0.052604

The situation is no different for populations initialised with a uniform length distribution, as indicated in Figures 2(c) and 2(d). In this case the values to which the frequency of 1's is approaching are: 0.100201 for a population initialised with ones in non-terminal locus 1, and 0.079757 for a population initialised with ones in non-terminal locus 5. Again, these values are very close to the theoretical predictions based on generation 0 information, as indicated in the fifth and sixth rows of Table 1. The table reports also other limit values, both measured and theoretically predicted, which further corroborate the theory.

The picture is not very different for classes of program of length other than 10, as illustrated in Figure 3. The figures show plots of the relative frequencies of non-terminal alleles of type 1 measured at generation 100 as a function of the locus position, for programs of lengths 10, 20, 30 and 40. In the initial population, all non-terminal loci at position 1 were occupied by alleles of type 1. Lengths were Gamma distributed in Figure 3(a) and uniformly distributed in Figure 3(b). Despite minor statistical oscillations, the proportions of alleles of type 1 approach a length- and locus-independent value, as expected from the theory.



(a) (b)

Fig. 3. Plots of the relative proportion of non-terminal alleles of type 1 vs. locus position within programs of length 10, 20, 30 and 40 at generation 100. The population was initialised with an approximately-Gamma length distribution in (a) and with a uniform length distribution in (b). In the initial generation the alleles of type 1 were confined to non-terminal locus 1.

To further verify that indeed under subtree crossover the population tends towards an independent allele distribution, we performed an experiment with exactly the same set up as in Figure 2(a) but this time we kept track of the co-occurrence of pairs of non-terminal alleles within the class of programs of length 10. So, for each generation we obtained a set of four 9×9 co-occurrence frequency matrices, one for each possible choice of a pair of the non-terminal alleles 0 and 1. An element at position (r, c) of the co-occurrence matrix for non-terminal alleles a and b, represented the average number of times allele a was present in locus r while at the same time allele b was present in locus c in strings of length 10. Once normalised by the total number of strings of length 10, the diagonal elements of the 0/0 and 1/1 matrices represent the proportions of alleles of type 0 and 1, respectively, present at each locus. So, the diagonals represent the same information as in Figure 2(a). The off-diagonal elements, however, may reveal correlations between pairs of alleles and loci.

Figure 4 shows plots of the average and standard deviation of the off-diagonal elements of the co-occurrence matrices for different pairs of non-terminal alleles and different generations. Initially the correlation between pairs of alleles is high. However, the off-diagonal elements converge rather quickly towards three different constant values and after 30 or 40 generations all the off-diagonal elements of each co-occurrence matrix are approximately identical. This indicates that, for strings of length 10, there is no remaining pairwise correlation in the population. The asymptotic value to which the frequencies of the off-diagonal elements converge coincides almost perfectly with those predicted by the theory on the basis of generation 0 information. Since in these circumstances $c(1) \approx 0.103444$, the theoretical values are: $(1 - c(1))^2 \approx 0.8038$ for the 0/0 allele pair, $c(1) \times (1 - c(1)) \approx 0.0927$ for the 0/1 and 1/0 allele pairs, and $c(1)^2 \approx 0.0107$ for the 1/1 allele pair. The picture is exactly the same for strings of other lengths.

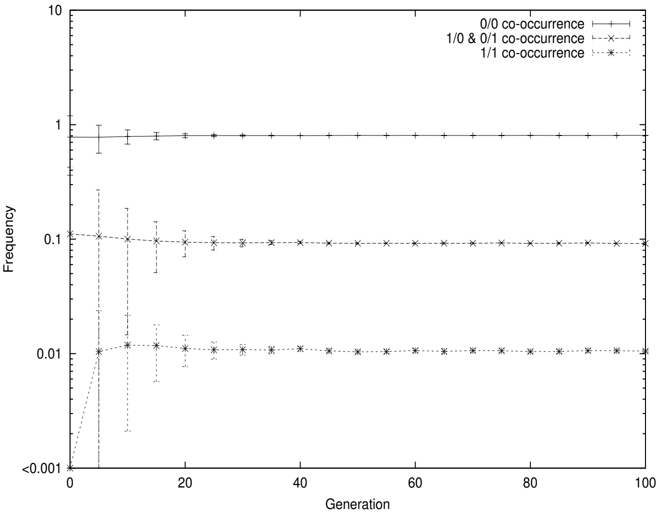


Fig. 4. Plots of the mean relative frequency of co-occurrence of pairs of non-terminal alleles vs. generation within programs of length 10. The population was initialised with an approximately Gamma length distribution. In the initial generation the alleles of type 1 were confined to non-terminal locus 1.

6 Discussion

In previous research [17, 7] the length biases of standard GP crossover when applied to linear representations were studied. In that work all the complexities involved in following the propagation of alleles (functions and terminals) were removed with only schemata of the form $(=)^N$, in the absence of selective pressure, being studied using exact schema evolution equations [13]. In those studies it was discovered that the program-length distribution tends towards a discrete Gamma distribution. This has important implications, such as that GP tends to sample exponentially more often shorter programs than longer ones. It also implies that, even in the presence of selective pressure, GP will be unable to fully converge to solutions of a specific length.

These results were the starting point for the work presented in this paper. Here, we focused our attention on the effects of subtree crossover on the primitive distribution, and therefore, ultimately, on the precise way in which this type of crossover explores the search space. We think that our empirical and theoretical results can be interpreted in a relatively simple way: subtree crossover generates both a "vertical" (between strings) primitive-mixing process and a "lateral" (within string) primitive diffusion process.

Let us first consider the effects of vertical mixing. A vertical mixing behaviour is present in most crossover operators described in the literature on fixed length GAs. It is well known that this destroys "linkage", i.e. correlations, between different allele positions in the population. In the fixed length case the asymptotic convergence towards independence described by Geiringer's theorem is the result of the decay of correlations due to the mixing effect of crossover. Because vertical mixing is performed also by subtree crossover, it is not surprising to see that GP is also moving towards an independent fixed-point string distribution. In other words, vertical primitive mixing is the reason why the right hand side of Equation 12 is a product, like the right hand side of Equation 7. Note that in fixed-length GAs the decay of correlations is exponential in time and in the case of a continuous time evolution can be solved for exactly [23] showing that higher order correlations decay faster (exponentially) than lower order ones. It is likely that a similar behaviour characterises also subtree crossover.

Let us now consider the effects of lateral diffusion. To the best of our knowledge a phenomenon of this type has not been reported for any other operator, and so it seems a very special feature of GP with subtree crossover. The effect of the diffusion process is that, unlike the case of fixed length representations and homologous crossover, the population moves towards a state where the probability of finding a particular allele at a given locus is locus-independent. This is why Equation 12 can be rewritten in the form in Equation 15, which emphasises the multinomial nature of the process (the probability of any specific string being in the population is only a function of the number of alleles of each type in the string, not where such alleles are located).

There are different ways in which we can interpret the effects of lateral diffusion. In one way, this is somewhat analogous to mutation. One of the features of mutation is that it allows for the reintroduction of a lost primitive at a particular position. With homologous crossover in fixed length strings, as is well known, once an allele has been lost at a particular bit position it cannot be recovered by selection and crossover. However, thanks to the lateral diffusion effect this is not the case for subtree crossover as our experiments plainly show. Imagine that a fit string requires a 1 at a particular bit position and that in the current population it does not exist then lateral diffusion can provide the missing primitive in a similar manner to that of mutation. The key difference is that with mutation there is no conservation law at work, with crossover on fixed length strings the frequency of a given primitive at a given bit position is preserved while in the case of subtree crossover on variable length strings only the frequency of an primitive in the entire population is preserved. This mutation-like behaviour of subtree crossover, which is unavoidably also present on non-flat fitness landscapes, is likely to be a reason for the convergence (or, rather, lack of convergence) behaviour shown by standard GP.

Another way of interpreting the effects of lateral diffusion is that, in some circumstances, they can provide an automatic restart mechanism. If at some point in a run selection becomes absent or very weak, crossover will try to push the population towards a Gamma length distribution and will mix and diffuse its primitives. So, if the selective pressure remains low for long enough, the population will move to a state which has some of the characteristics of a typical initial population: no correlation between primitives and locus-independent primitive frequencies. This is not exactly equivalent to reinitialising the population (a restart), because the mean length of the population will be the mean length reached when the selection pressure dropped, and the frequency of each primitive will be the mean frequency present when the selection pressure dropped. Also, if the initial population included some correlations, e.g. if it was the result of inoculation (the initialisation of the population with programs believed to be good candidate solutions), then lateral diffusion could not be seen as a restarting mechanism. In any case, if the designer of the GP system wanted the system to work by discovering progressive improvements of the best program found so far, the behaviour just described might appear undesirable.[2]

A consequence of the diffusive bias of subtree crossover is that the way the search space is sampled in the initial generations depends perhaps more heavily on the initial primitive frequencies than on the actual structure of the programs in the initial generation. This suggests that clever initialisation procedures such as inoculation might not necessarily produce the desired effects (after inoculation, crossover will tend to destroy any specific arrangement of primitives as crossover moves the population toward an independent node distribution).

7 Conclusions

In this paper we have presented theoretical results describing the asymptotic behaviour of a linear GP system, or a variable length GA, evolving in a flat fitness landscape with no mutation and using subtree crossover. We provided experimental evidence that firmly corroborates the theory, showing an almost perfect match between the predictions of the theory based on generation 0 data and the observed primitive frequencies at later generations.

In part, the behaviour we have observed and characterised is what one would expect: a) crossover shuffles the primitives present in different individuals and b) primitives which left an individual due to an earlier crossover event can come back at a different position in a later crossover event, resulting in a sort of gene duplication. What is perhaps surprising is that this second effect is a real diffusion process which attempts to push the population towards a locus- and length-independent primitive distribution where each primitive is equally likely to be found in any position of any individual.

[2] In [14], on the basis of a theoretical analysis of the amount of genetic material exchanged by the parents to form the offspring, we conjectured that in some cases GP with subtree crossover might be more like a set of stochastic hill-climbers working in parallel than like a genetic algorithm. The restarting behaviour described above seems to further corroborate the hill-climbing conjecture and to refine it by indicating that, in some circumstances, a GP system might behave a bit like a hill-climber with restarts.

Knowing this bias and the length biases [17] of standard crossover is important because it allows the users of GP systems to evaluate whether this type of crossover provides the desired search behaviour for the system. If this is not the case, then the knowledge of the search biases of other operators, which has recently started emerging both from empirical studies and schema-theoretic analyses, allows an informed choice for an alternative. In addition, as discussed in the paper, the knowledge of these biases can explain emergent GP phenomena such as the inability of GP to converge.

Acknowledgements

CRS would like to thank the University of Birmingham for a visiting Professorship and DGAPA-PAPIIT grant IN100201. RP and CRS would like to thank the Royal Society for their support. JER, AHW and RP would like to thank EPSRC for their support.

References

[1] L. B. Booker. Recombination distributions for genetic algorithms. In *FOGA-92, Foundations of Genetic Algorithms*, Vail, Colorado, 24–29 July 1992. Email: booker@mitre.org.

[2] L. B. Booker, D. B. Fogel, D. Whitley, P. J. Angeline, and A. E. Eiben. Recombination. In T. Bäck, D. B. Fogel, and T. Michalewicz, editors, *Evolutionary Computation 1: Basic Algorithms and Operators*, chapter 33. Institute of Physics Publishing, 2000.

[3] H. Geiringer. On the probability theory of linkage in Mendelian heredity. *Annals of Mathematical Statistics*, 15(1):25–57, March 1944.

[4] J. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, USA, 1975.

[5] W. B. Langdon, T. Soule, R. Poli, and J. A. Foster. The evolution of size and shape. In L. Spector, W. B. Langdon, U.-M. O'Reilly, and P. J. Angeline, editors, *Advances in Genetic Programming 3*, chapter 8, pages 163–190. MIT Press, Cambridge, MA, USA, June 1999.

[6] N. F. McPhee and J. D. Miller. Accurate replication in genetic programming. In L. Eshelman, editor, *Genetic Algorithms: Proceedings of the Sixth International Conference (ICGA95)*, pages 303–309, Pittsburgh, PA, USA, 15-19 July 1995. Morgan Kaufmann.

[7] N. F. McPhee and R. Poli. A schema theory analysis of the evolution of size in genetic programming with linear representations. In *Genetic Programming, Proceedings of EuroGP 2001*, LNCS, Milan, 18-20 Apr. 2001. Springer-Verlag.

[8] N. F. McPhee, R. Poli, and J. E. Rowe. A schema theory analysis of mutation size biases in genetic programming with linear representations. In *Proceedings of the 2001 Congress on Evolutionary Computation CEC 2001*, Seoul, Korea, May 2001.

[9] N. F. McPhee, R. Poli, and J. E. Rowe. A schema theory analysis of mutation size biases in genetic programming with linear representations. In *Proceedings of the 2001 Congress on Evolutionary Computation CEC2001*, pages 1078–1085, COEX, World Trade Center, 159 Samseong-dong, Gangnam-gu, Seoul, Korea, 27-30 May 2001. IEEE Press.

[10] R. Poli. Exact schema theorem and effective fitness for GP with one-point crossover. In D. Whitley, D. Goldberg, E. Cantu-Paz, L. Spector, I. Parmee, and H.-G. Beyer, editors, *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 469–476, Las Vegas, July 2000. Morgan Kaufmann.

[11] R. Poli. Hyperschema theory for GP with one-point crossover, building blocks, and some new results in GA theory. In R. Poli, W. Banzhaf, and *et al.*, editors, *Genetic Programming, Proceedings of EuroGP 2000*. Springer-Verlag, 15-16 Apr. 2000.

[12] R. Poli. Exact schema theory for genetic programming and variable-length genetic algorithms with one-point crossover. *Genetic Programming and Evolvable Machines*, 2(2), 2001. Forthcoming.

[13] R. Poli. General schema theory for genetic programming with subtree-swapping crossover. In *Genetic Programming, Proceedings of EuroGP 2001*, LNCS, Milan, 18-20 Apr. 2001. Springer-Verlag.

[14] R. Poli and W. B. Langdon. On the search properties of different crossover operators in genetic programming. In J. R. Koza, W. Banzhaf, K. Chellapilla, K. Deb, M. Dorigo, D. B. Fogel, M. H. Garzon, D. E. Goldberg, H. Iba, and R. Riolo, editors, *Genetic Programming 1998: Proceedings of the Third Annual Conference*, pages 293–301, University of Wisconsin, Madison, Wisconsin, USA, 22-25 July 1998. Morgan Kaufmann.

[15] R. Poli and W. B. Langdon. Schema theory for genetic programming with one-point crossover and point mutation. *Evolutionary Computation*, 6(3):231–252, 1998.

[16] R. Poli and N. F. McPhee. Exact GP schema theory for headless chicken crossover and subtree mutation. In *Proceedings of the 2001 Congress on Evolutionary Computation CEC 2001*, Seoul, Korea, May 2001.

[17] R. Poli and N. F. McPhee. Exact schema theorems for GP with one-point and standard crossover operating on linear structures and their application to the study of the evolution of size. In *Genetic Programming, Proceedings of EuroGP 2001*, LNCS, Milan, 18-20 Apr. 2001. Springer-Verlag.

[18] R. Poli and N. F. McPhee. Exact schema theory for GP and variable-length GAs with homologous crossover. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, San Francisco, California, USA, 7-11 July 2001. Morgan Kaufmann.

[19] M. Ridley. *Evolution*. Blackwell Scientific Publications, Boston, 1993.

[20] J. E. Rowe and N. F. McPhee. The effects of crossover and mutation operators on variable length linear structures. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, San Francisco, California, USA, 7-11 July 2001. Morgan Kaufmann.

[21] T. Soule, J. A. Foster, and J. Dickinson. Code growth in genetic programming. In J. R. Koza, D. E. Goldberg, D. B. Fogel, and R. L. Riolo, editors, *Genetic Programming 1996: Proceedings of the First Annual Conference*, pages 215–223, Stanford University, CA, USA, 28–31 July 1996. MIT Press.

[22] W. M. Spears. Limiting distributions for mutation and recombination. In W. M. Spears and W. Martin, editors, *Proceedings of the Foundations of Genetic Algorithms Workshop (FOGA 6)*, Charlottesville, VA, USA, July 2000. In press.

[23] C. R. Stephens. Some exact results from a coarse grained formulation of genetic dynamics. In L. Spector, E. D. Goodman, A. Wu, W. B. Langdon, H.-M. Voigt, M. Gen, S. Sen, M. Dorigo, S. Pezeshk, M. H. Garzon, and E. Burke, editors, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, pages 631–638, San Francisco, California, USA, 7-11 July 2001. Morgan Kaufmann.

[24] C. R. Stephens and H. Waelbroeck. Effective degrees of freedom in genetic algorithms and the block hypothesis. In T. Bäck, editor, *Proceedings of the Seventh International Conference on Genetic Algorithms (ICGA97)*, pages 34–40, East Lansing, 1997. Morgan Kaufmann.

[25] C. R. Stephens and H. Waelbroeck. Schemata evolution and building blocks. *Evolutionary Computation*, 7(2):109–124, 1999.