Exact Results from a Coarse Grained Formulation of the Dynamics of Variable-length Genetic Algorithms

Christopher R. Stephens
Instituto de Ciencias Nucleares
UNAM, Circuito Exterior
Mexico D.F. 04510
stephens@nuclecu.unam.mx

Riccardo Poli
Dept. of Computer Science
University of Essex,
Colchester, CO4 3SQ, UK
rpoli@essex.ac.uk

Alden H. Wright
Dept. of Computer Science
University of Montana, USA
wright@cs.umt.edu

Jonathan E. Rowe
School of Computer Science
University of Birmingham, UK
j.e.rowe@cs.bham.ac.uk

Abstract

We consider the dynamics of variable-length Genetic Algorithms (GAs) with strings of length $N \leq N_m$ using a recently developed exact, coarse-grained formulation where the relevant coarse-grained degrees of freedom are "building block" schemata. We derive an exact formal solution of the equations showing how a hierarchical structure in time and degree of coarse-graining emerges, the effect of recombination being to successively form more fine-grained objects from their more coarse-grained building blocks, where in this case the building blocks can come from strings of different lengths. We examine the limit distributions of the dynamics in the case of a flat fitness landscape, one-point homologous crossover and no mutation. By taking advantage of the existence of a set of conserved quantities in the dynamics we provide exact solutions for the cases $N_m = 2,\ 3$ and use these to investigate the phenomenon of inter-length-class allele diffusion. We also study the general case showing what exact results may be easily derived using our particular coarse-grained formulation.

1 Introduction

The dynamics engendered by a "canonical" GA and, indeed, genetic dynamics in general, is exceedingly complicated. This is true even in the case of what one might think of as "toy" fitness landscapes such as counting ones or needle-in-a-haystack. In fact, up until quite recently [1], to our knowledge, no solutions have been found for the dynamics in the presence of recombination for arbitrary string lengths even in the case of a flat fitness landscape, though there has been recent noteworthy progress in the special case of "genepool" recombination [2, 3], where for a given recombination event allele mixing is over the entire population not just between two parents. For binary strings of fixed length, N, the probability distribution that describes the dynamics is obtained by solving 2^N coupled, non-linear difference equations. Important results have been derived about this system of equations by viewing them as a dynamical system [4]. However, these coupled equations, in terms of the underlying string variables, are far removed from traditional elements of GA theory such as the Schema theorem and Building Block Hypothesis (BBH) [5, 6].

The underlying microscopic equations, however, can be rewritten naturally in a basis other than the string basis [7, 8, 9] yielding evolution equations that offer the benefit of a very intuitive interpretation, that illuminate the content of the Schema theorem and the BBH, that naturally coarse grain from string equations to schema equations, that yield an interpolation between the microscopic and the macroscopic and that offer new exact results or simpler proofs of known results. These equations lead to many insights into the dynamics of GAs offering an exact Schema theorem that naturally incorporates a form of the BBH, although it is important to emphasize here that the "building blocks" that naturally emerge in this formulation are dynamic and not necessarily short or even fit! However, creation events due to recombination can be precisely understood in terms of these BBs. Originally applied to a canonical GA (proportional selection, 1-point crossover and mutation) the basic elements have been extended to GAs with arbitrary selection schemes and any homologous crossover [1] and, importantly, have been extended to Genetic Programming (GP) by Poli and coworkers [10, 11].

There has been increasing interest in variable-length representations from different points of view [12, 13, 14] . In this paper we will use a coarse-grained BB formulation to investigate the dynamics of variable length GAs up to a maximum size N_m. We present formal solutions for an homologous crossover operator and arbitrary fitness landscape and mutation showing how the solution naturally admits an interpretation in terms of a hierarchy of BBs. We then consider the asymptotic behaviour of the dynamics for a flat fitness landscape, both at the formal level, discussing generalizations of Geiringer's theorem, and at the explicit level, deriving exact solutions for $N_m = 2$ and 3 and various exact results for arbitrary N_m.

This work is, of course, susceptible to the standard criticisms - what is the relevance of considering a small number of loci and flat fitness landscapes? There are several ways of rebutting such criticism. Firstly, simple models can lead to intuitive insights that would be less transparent in a more complex model. An important example of that is the minimal two-bit deceptive problem [15]. Another example, is the work of Spears [16] where limit distributions for recombination and mutation for fixed length GAs in a flat fitness landscape were investigated in simple two and three-bit problems. Interestingly, even in this case he had to resort to numerical rather than analytical calculations. Additionally, understanding the structure of the dynamics in simple problems can lead to insight about how to construct results or proofs in more general problems and potentially lead to insights which may be of benefit for practitioners.

2 Coarse-Grained Evolution Equations

In this section we introduce the notion of coarse-grained evolution equations in a BB basis, discussing their interpretation and advantages at a formal level. We will not derive the coarse-grained exact evolution equations here but refer the reader to the original literature [7, 9, 10, 11]. Our interest here is variable-length GAs with homologous crossover. As homologous crossover operators conserve length classes [18] we will consider the corresponding evolution equation for strings or schemata within a given length class N, composed of strings of a fixed length, and consider arbitrary string length $N \leq N_m$, where $N_m \in [1, \infty]$. In this case, if one considers the evolution of length N strings then one of the parents in the crossover operation must be a length N string as well while the other parent may be of arbitrary size. The action of the homologous crossover we will use can be simply understood by aligning the two parents at the first loci then implementing a mask defined on the common region of the two strings. For example, with 1111 and 000000 the common region is associated with the first four loci. A one-point crossover between the second and third loci would yield 110000 and 0011 while a crossover between the fifth and sixth loci (of the second string) is not allowed. Hence, the total number of possible masks on the common region is 2^4.

Our primary object of interest will be the proportion of strings of a given type, C_i^N, $P(C_i^N, t)$, or of a given schema, ξ^N, $P(\xi^N, t) = \sum_{C_i^N \in \xi^N} P(C_i^N, t)$, within a length class N. Thus, we define a schema relative to a given length class. However, it is important to note that all proportions will be relative to the total population size summed over all length classes. In the infinite population limit, which we will generally assume throughout, $P(C_i^N, t)$ is simply the probability for finding the string C_i^N. For a string C_i^N we have

$$P(C_i^N, t+1) = \mathcal{P}(C_i^N) P_c(C_i^N, t) + \sum_{C_i \neq C_j} \mathcal{P}(C_j^N \to C_i^N) P_c(C_j^N, t) \quad (1)$$

where the sum is over all length-class N strings that differ by at least one bit from C_i^N. $\mathcal{P}(C_i^N) = (1-p_m)^N$ is the probability that C_i^N remains unmutated and $\mathcal{P}(C_j^N \to C_i^N) = p_m^{d_H(i,j)}(1-p_m)^{N-d_H(i,j)}$ is the probability that the string C_j^N mutate to the string C_i^N, $d_H(i,j)$ being the Hamming distance between the strings C_i^N and C_j^N. Note that mutation preserves the length class of a string or schema. $P_c(C_i^N, t)$ is the probability of finding a string C_i^N after selection and crossover and is given by

$$P_c(C_i^N, t) = (1-p_c) P'(C_i^N, t) + \sum_{j=1}^{N_m} \sum_{m=0, even}^{2^{min(j,N)}-1} (p_c(m) + p_c(\bar{m})) P'(C_i^j(m), t) P'(C_i^N(\bar{m}), t) \quad (2)$$

where $P'(C_i^N, t)$ is the probability for selecting the string C_i^N. $p_c(m)$ is the probability of implementing the mask m on the common region between the two strings and we sum over only even masks as this ensures that the tail comes from the second parent which, without loss of generality we assume to be of length N, and therefore that length is preserved. $\bar{m}$ is the mask conjugate to m. The total number of possible masks on the common region is $2^{min(j,N)}$. $C_i^j(m)$ for a given mask m represents the part of the string C_i^N inherited from the first parent, which we assume to be of length j, and $C_i^N(\bar{m})$ is that part inherited from the second. Both $C_i^j(m)$ and $C_i^N(\bar{m})$ are schemata. (2) has a form similar to that for the fixed length case and can be interpreted similarly, i.e. strings are created by BBs, the difference in this case being that one of the BBs can come from a parent of other than length N. Once again we emphasize that these BBs are dynamical not static schema averages and are neither necessarily small or even fit!

The microscopic equation (1) can be coarse-grained to an arbitrary schema of order $N_2 \leq N$ and defining length $(l-1)$ contained within strings of size N to find

$$P(\xi^N, t+1) = \mathcal{P}(\xi^N) P_c(\xi^N, t) + \sum_{\not\xi_i^N} \mathcal{P}(\not\xi_i^N \to \xi^N) P_c(\not\xi_i^N, t) \quad (3)$$

where the sum is over all schemata, $\not\xi_i^N$, that differ by at least one bit from ξ^N in one of the N_2 defining bits of ξ^N. In other words any schema competing with ξ^N and belonging to the same partition. $\mathcal{P}(\xi^N) = (1-p_m)^{N_2}$ is the probability that ξ^N remains unmutated and $\mathcal{P}(\not\xi_i^N \to \xi^N) = p_m^{d_H(\xi^N, \not\xi_i^N)}(1-p_m)^{N_2 - d_H(\xi^N, \not\xi_i^N)}$ is the probability that the schema $\not\xi_i^N$ mutate to the schema ξ^N with $d_H(\xi^N, \not\xi_i^N)$ being the Hamming distance between the schemata ξ^N and $\not\xi_i^N$. $P_c(\xi^N, t) = \sum_{C_i^N \in \xi^N} P_c(C_i^N, t)$ is the probability of finding a schema ξ^N of length class N after selection and crossover and is given by

$$P_c(\xi^N, t) = (1 - p_c A(\xi, t)) P'(\xi^N, t) + \sum_{j=1}^{N_m} \sum_{m \in \mathcal{M}_r(\xi^N)} (p_c(m) + p_c(\bar{m})) P'(\xi^j(m), t) P'(\xi^N(\bar{m}), t) \quad (4)$$

where $P'(\xi^N,t)$ is the probability for selecting a schema ξ^N from strings of length class N. $\xi^N(m)$ for a given mask m represents the part of the schema ξ^N inherited from the first parent and $\xi^N(\bar{m})$ is that part inherited from the second. Now, $\xi^N(m)$ and $\xi^N(\bar{m})$ are the BBs for the schema ξ^N. Thus, we see that BBs at one level are composed of more primitive (lower order) BBs which in their turn are composed of lower order blocks etc. thus leading to a hierarchical structure. $\mathcal{M}_r$ is the set of crossover masks that end in a 0 that affect ξ^N, i.e. the number of "allele mixing" masks, $N_{\mathcal{M}_r(\xi)}$ is their number. $A(\xi^N,t)$ determines the survival probability of the schema and depends on the properties of the particular schema, such as order and defining length, and, importantly, also depends on the length distribution of the strings and their corresponding fitnesses [18].

As with all coarse grained evolution equations the interpretation of (1) and (2) is very intuitive: (2) tells us how a particular string is selected and survives crossover, or alternatively how it is built up from its BBs. The novel element here compared to standard GAs is that the BBs come from strings of potentially different sizes. (1) then tells us how the string is preserved by mutation or formed by mutation from some other string of the same partition.

We can put the basic equation (1) into a yet more elegant form, the corresponding equation for schemata follows trivially, by introducing a 2^N-dimensional population vector for each length class, $\mathbf{P}^N(t)$, whose elements are $P(C_i^N,t), i=1,...,2^N$. Equation (1) takes the form

$$\mathbf{P}^N(t+1) = \overline{\mathbf{W}}^N \mathbf{P}_c^N(t) \tag{5}$$

where the $N\times N$-dimensional mutation matrix $\overline{\mathbf{W}}^N$ is real, symmetric and time independent and has elements $\overline{\mathbf{W}}_{ij}^N = p_m^{d^H(i,j)}(1-p_m)^{N-d^H(i,j)}$. For selection schemes linear in $P(C_i^N,t)$, $\mathbf{P}_c^N(t)$ can be written as

$$\mathbf{P}_c^N(t) = \overline{\mathbf{F}}^N(t)\mathbf{P}^N(t) + \sum_{j=1}^{N_m}\sum_{m=0,even}^{2^{min(j,N)}-1}(p_c(m)+p_c(\bar{m}))\mathbf{J}^j(m,t) \tag{6}$$

where the "cloning" matrix, $\overline{\mathbf{F}}^N(t)$, is diagonal and describes both selection and survival under crossover. Explicitly, for proportional selection $\overline{\mathbf{F}}^N{}_{ii}(t) = (f(C_i)/\bar{f}(t))(1-p_c)$. Finally, the components of the "source" vector are given by $J^j_{C_i^N}(m,t) = P'(C_i^j(m),t)P'(C_i^N(\bar{m}),t)$ which corresponds to the BB sources, from strings of length j and N respectively, for the string C_i^N. Defining the cloning-mutation matrix $\overline{\mathbf{W}}_s^N(t) = \overline{\mathbf{W}}^N\overline{\mathbf{F}}^N(t)$ we have

$$\mathbf{P}^N(t+1) = \overline{\mathbf{W}}_s^N(t)\mathbf{P}^N(t) + \sum_{j=1}^{N_m}\sum_{m=0,even}^{2^{min(j,N)}-1}(p_c(m)+p_c(\bar{m}))\overline{\mathbf{W}}^N\mathbf{J}^j(m,t) \tag{7}$$

The interpretation of this equation is that $J^j_{C_i^N}(m,t)$ is a source which creates strings C_i^N by bringing BBs from strings of length j and N together. The first term on the right hand side tells us how the strings themselves are propagated, or survive, into the next generation, the destructive effect of crossover renormalizing the fitness of the strings. Note that the equation is linear but for the presence of string creation. It is this division into a linear term and a source that allows for a natural formal solution which leads to further insight into the nature of GA dynamics while at the same time offering the possibility of exact, analytic calculations in certain circumstances.

Needless to say solutions of these dynamical equations are hard to come by. They represent, for binary alleles, $2(2^{N_m}-1)$ coupled non-linear difference equations, or in the continuous time limit - differential equations. Here, we consider the formal solution for the case of homologous crossover and mutation and for any selection scheme linear in $P(C_i^N,t)$. The equation (7) is always of the same form, i.e. a first order, linear, inhomogeneous difference (differential) equation. Its iterated solution is

$$\mathbf{P}^N(t) = \mathbf{D}(t,0)\mathbf{P}^N(0) + \sum_{j=1}^{N_m}\sum_{m=0,even}^{2^{min(j,N)-1}}(p_c(m)+p_c(\bar{m}))\sum_{n=0}^{t-1}\mathbf{D}(t,n)\overline{\mathbf{W}}^N\mathbf{J}^j(m,n) \tag{8}$$

where $\mathbf{D}(t,0)=\prod_{n=0}^{t-1}\overline{\mathbf{W}}_s^N(n)$. The interpretation of (8) follows naturally from that of (7). Considering first the case without mutation, the first term on the right hand side gives us the probability that a string survives from $t=0$ to t without being destroyed by crossover. In other words $\mathbf{D}(t,0)$ is the Greens function or propagator for $\mathbf{P}^N$ [1]. In the case of a flat fitness landscape without mutation for instance $D_{ij}(t,0) = (1-p_c)^t\delta_{ij}$. In the second term, each element, $\mathbf{J}^j(m,n)$, is associated with the creation of a string C_i^N at time n via the juxtaposition of two BBs from strings of length j and N respectively and associated with a mask m. The component corresponding to C_i^N of the matrix $\mathbf{D}(t,n)=\prod_{i=n}^{t-1}\overline{\mathbf{W}}_s(i)$ is the probability that the resultant string survives from its creation at time n to t. The sum over masks, string lengths, j, and n is simply the sum over all possible creation events in the dynamics. In a more explicit notation we will denote the propagator for a string $h_1\cdots h_N$ by $D_{h_1\cdots h_N}(t,t')$.

This formal solution above has a very natural diagramatic interpretation both at the level of fixed length strings which can be extended to the present case.

3 Geiringer's Theorem

For any dynamical system fixed points and their stability are of particular interest. Hence, in this section we will discuss the fixed point distributions for fixed and variable-length GAs. For a fixed-length GA evolving on a flat landscape in the absence of mutation the fixed point $P^*(h_1...h_N)$ of the dynamics for a string $c_i = h_1...h_N$ is

$$P^*(h_1...h_N) = \lim_{t\to\infty} P(c_i, t) = \prod_{i=1}^{N} P(*^{i-1}h_i*^{N-i}, 0) \quad (9)$$

where $*^i$ as a string argument means the symbol $*$ repeated i times. This result is the well known Geiringer's theorem [17] for a general crossover operator. Any population that factorizes in this manner is said to be in linkage equilbrium and the resulting allele frequencies are known as Robbins proportions. This result emerges naturally from equation (8), specialized to the case of a single length class, N, which yields for a flat landscape in the absence of mutation

$$\mathbf{P}^N(t) = (1-p_c)^t\mathbf{P}(0) + \sum_{m=0,even}^{2^N-1} (p_c(m) + p_c(\bar{m})) \sum_{n=0}^{t-1}(1-p_c)^{t-n-1}\mathbf{J}^N(m,n) \quad (10)$$

As $\lim_{t\to\infty}(1-p_c)^t = 0$, hence $\mathbf{P}^N(t) \to 0$ as $t \to \infty$ unless the summation over time leads to a cancellation of this damping factor. Given that the BB constituents of $\mathbf{J}^N(m,n)$ are associated with damping factors $(1 - p_c\frac{N_{\mathcal{M}_r}(C_i^j(m))}{N_{\mathcal{M}}})^t$ and $(1 - p_c\frac{N_{\mathcal{M}_r}(C_i^N(\bar{m}))}{N_{\mathcal{M}}})^t$, where $N_{\mathcal{M}}$ is the total number of non-zero crossover masks, this can only occur if there is no damping of the consituent BBs and this only happens if they are 1-schemata as then $N_{\mathcal{M}_r} = 0$. Thus, the only term that survives in the hierarchical solution of (8) is the product of 1-schemata [9].

The type of recombination employed controls how fast the transient corrections to the limit distribution die out. The damping is controlled by $N_{\mathcal{M}_r}(\xi)$, hence the bigger it is the faster the corresponding transient dies out [1]).

The general approach to equilibrium is characterized by the exponential decay of linkage disequilibrium functions $\Delta_{h_1\cdots h_N} = \langle(h_1 - \langle h_1\rangle)\cdots(h_N - \langle h_N\rangle)\rangle$ where $\langle O\rangle$ denotes the population average of O. Thus, $\langle h_i\rangle = P(*^{i-1}h_i*^{N-i})$. These linkage disequilbrium functions will be seen to be natural variables in which to understand the dynamics and approach to equilbrium. In GAs a set of variables that have also been viewed as natural for considering the dynamics are "building blocks".

The generalization of Geiringer's theorem to the variable length case has recently been derived [18]

$$P^*(h_1\cdots h_N) = P(*^N)\prod_{i=1}^{N}\frac{P(*^{i-1}h_i\#, 0)}{P(*^i\#, 0)}, \quad (11)$$

where

$$P(*^{i-1}h_i\#, 0) = \sum_{N\geq 0} P(*^{i-1}h_i*^N, 0)$$

and

$$P(*^i\#, 0) = \sum_{N\geq 0} P(*^{i+N}, 0).$$

Here, we see a generalization of the concept of Robbins proportions, the corresponding proportions in the variable length case being $\frac{P(*^{i-1}h_i\#,0)}{P(*^i\#,0)}$. We will see in the next section that there are natural analogs of the linkage disequilibrium functions as well.

4 Explicit Solutions - $N_m = 2,\ 3$

In [1] it was shown for fixed length strings in the continuous time limit how an exact explicit solution corresponding to (8) could be found for a flat fitness landscape. Even in this case however, the result is highly non-trivial due to the complicated combinatorics of the various BB creation events. In the case of variable length strings one would expect the combinatorics to be even more complicated. Before considering the general case we will therefore look at some relatively simple cases for $N_m = 2,\ 3$ with no mutation and using one-point crossover where we also include crossover before the first bit and immediately after the last bit of the shortest parent. For $N_m = 2$ we must solve:

$$P(h_1h_2, t+1) = (1-p_c)P(h_1h_2, t) + \sum_{j=1}^{2}\frac{p_c}{min(2,j)+1}\sum_{i=0}^{min(2,j)}P(h_1...h_i*^{j-i}, t)P(*^ih_{i+1}...h_2, t) \quad (12)$$

for strings of length two and

$$P(h_1, t+1) = (1-p_c)P(h_1, t) + \frac{p_c}{2}\sum_{j=1}^{2}\sum_{i=0}^{1}P(h_1...h_i*^{j-i}, t)P(*^ih_{i+1}...h_1, t) \quad (13)$$

for strings of length one. The corresponding "source" terms are respectively

$$J^j_{h_1h_2}(i,t) = P(h_1...h_i*^{j-i}, t)P(*^ih_{i+1}...h_2, t) \quad (14)$$
$$J^j_{h_1}(i,t) = P(h_1...h_i*^{j-i}, t)P(*^ih_{i+1}...h_1, t). \quad (15)$$

The explicit forms of the equations of motion are

$$P(h_1h_2, t+1) = (1-p_cA(h_1h_2))P(h_1h_2, t) + \frac{p_c}{2}P(h_1, t)P(*h_2, t) + \frac{p_c}{3}P(h_1*, t)P(*h_2, t) \quad (16)$$

where $A(h_1h_2) = \left(\frac{1}{2}P(*^1) + \frac{1}{3}P(*^2)\right)$ and

$$P(h_1, t+1) = (1-p_cA(h_1))P(h_1, t) + \frac{p_c}{2}P(*^1)P(h_1*, t) \quad (17)$$

where $A(h_1) = P(*^2)/2$. $P(*^1)$ and $P(*^2)$ are the probabilities to get any string of length one and length two respectively. Note that homologous crossover preserves the length distribution [18].

With this simple $N_m = 2$ problem equations (16) and (17) have an intuitive interpretation that allows us immediately to investigate the phenomenon of allele diffusion between different length classes that is an important characteristic of variable-length genetic dynamics. The factor $P_s(h_1 \cdots h_N) = (1 - p_c A(h_1 \cdots h_N))$ describes the survival probability per generation of a particular length-N string. For length-one strings $P_s(h_1) = (1 - p_c P(*^2)/2)$ so it is only in the presence of length-two strings that that there is a non-zero decay probability. This probability grows as a function of $P(*^2)$ due to the fact that there are more decay channels open to the string. For length-one strings the only creation source is via the 2-schema h_1* which implies a diffusion of alleles of type h_1 from length-two to length-one strings. For length-two strings the two corresponding creation terms are associated with getting the first bit of the string from a parent of length one and the second bit from a 1-schema associated with strings of length two and the first and second bits from 1-schemata associated with strings of length two. This second term is exactly the same as would be found in a fixed-length GA. The novel element is to be able to construct the desired length-two string by interaction between a 1-schema associated with length-two strings and a length-one string. Thus, in order to solve for the dynamics for length-two strings one must first solve for the dynamics of the size one strings. As from (17) one can see that their dynamics depends on the dynamics of the 1-schemata it would seem that the dynamics of the length-one and two strings are inextricably interwined and must be solved for simultaneously. However, this is not so. The reason why not is that there exist constants of the motion that can be exploited. To see this consider $P(h_1\#, t) = P(h_1, t) + P(h_1*, t)$. The 1-schema probability $P(h_1*, t)$ may be determined from (16)

$$\begin{aligned} P(h_1*, t+1) = (1 - p_c A(h_1 h_2)) P(h_1*, t) + \\ \frac{p_c}{2} P(h_1, t) P(*^2) + \frac{p_c}{3} P(h_1*, t) P(*^2, t) \end{aligned} \quad (18)$$

thus adding this to (17) one finds

$$P(h_1\#, t+1) = P(h_1\#, t) \quad (19)$$

and hence $P(h_1\#)$ is an invariant of the motion. It basically expresses the conservation of the allele h_1 associated with the first bit position and in this sense is analogous to the conservation law $P(*^{k-1} h_k *^{N-k}, t) = P(*^{k-1} h_k *^{N-k}, 0)$ for any k associated with fixed length GAs. In the variable-length case however there is no conservation of alleles within a given length class due to the phenomenon of inter-length-class allele diffusion. With this conservation law in hand the equations (17) and (16) can be decoupled. We write (17) as

$$P(h_1, t+1) = D_{h_1} P(h_1, t) + \frac{p_c}{2} P(*^1) P(h_1\#, t) \quad (20)$$

where we now revert to the propagator notation used in section 2, $D_{h_1} = (1 - p_c/2)$ being the survival probability per generation. This equation can be simply solved using equation (8) to yield

$$P(h_1, t) = D_{h_1}^t P(h_1, 0) + (1 - (1 - D_{h_1}^t)) P^*(h_1) \quad (21)$$

where $P^*(h_1) = P(*^1) P(h_1\#)/P(*\#)$ is the fixed point of the dynamics in agreement with the general fixed point of (11). We may expand $P(h_1\#) = P(h_1, 0) + P(h_1*, 0)$ to find

$$\begin{aligned} & P(h_1, t) = \\ & \left((1 - \frac{p_c}{2})^t + (1 - (1 - \frac{p_c}{2})^t) P(*^1) \right) P(h_1, 0) \\ & + (1 - (1 - \frac{p_c}{2})^t) P(*^1) P(h_1*, 0) \end{aligned} \quad (22)$$

Note that even if $P(h_1, t) = 0$ inter-length-class allele diffusion will generate alleles h_1 in length-one strings at some later time. Thus, unlike the fixed length case a particular allele in a given length class may be regenerated without the intervention of mutation. Note that at the fixed point the contributions to h_1 are determined solely by the $t = 0$ proportions of this allele from all possible length classes. Hence, recombination in the variable length case maximally mixes the alleles among all available length classes.

Having found the exact solution for strings of length one we may proceed to strings of length two. As can be seen from equation (16) we need to solve first for the dynamics of the two 1-schemata h_1* and $*h_2$. From (16), one notices that there are no source terms for $*h_2$ from length-one strings. Hence, one finds that

$$P(*h_2, t+1) = P(*h_2, t) \quad (23)$$

and notes that the allele h_2 is conserved in agreement with (11). The 1-schema $P(h_1*, t) = P(h_1\#) - P(h_1, t)$ can be simply solved for to yield

$$P(h_1*, t) = D_{h_1*}^t P(h_1*, 0) + (1 - D_{h_1*}^t) P^*(h_1*) \quad (24)$$

where the survival probability per generation for h_1* is $D_{h_1*} = (1 - \frac{p_c}{2})$ and the fixed point $P^*(h_1*)$ is given by $P^*(h_1*) = P(*^2) P(h_1\#)/P(*\#)$ once again in agreement with equation (11). Note that the exponential approach to this fixed point is the same as for $P(h_1, t)$.

Finally, using the explicit solutions (21), (23) and (24) we may deduce the solution of (16). $P(h_1*, t)$ and $P(h_1, t)$ are a time-dependent source of strings $P(h_1 h_2, t)$. Substituting in (16) the solutions (21), (23) and (24) one finds

$$\begin{aligned} & P(h_1 h_2, t) = \\ & D_{h_1 h_2}^t (P(h_1 h_2, 0) - P(h_1\#) P(*h_2, 0)) + \\ & \frac{P(*h_2, 0)}{P(*^2)} (P(h_1, 0) - P(*^1) P(h_1\#)) (D_{h_1 h_2}^t - D_{h_1}^t) \\ & + P(h_1\#) P(*h_2, 0) \end{aligned} \quad (25)$$

In the limit $t \to \infty$ $D_{C_i^N} \to 0$; thus, we see the fixed point $P^*(h_1h_2) = P(h_1\#)P(*h_2, 0)$ emerging in agreement with equation (11).

The solutions can be put into a more elegant and transparent form by introducing the notion of generalized linkage disequilibrium functions. We define $\Delta_{h_1}(t) = (P(h_1,t) - P(*^1)P(h_1\#))$ and $\Delta_{h_1h_2}(t) = (P(h_1h_2,t) - P(h_1\#)P(*h_2))$. Thus, both these functions characterize deviations from the corresponding fixed points. Immediately we see an important distinction from the fixed length case where a single bit cannot have BBs and linkage occurs between different bits. Here the "building blocks" of h_1 are any length-one string and any string of any length that contains h_1. Due to the phenomenon of inter-length-class allele diffusion there is a concept of linkage disequilibrium for a single bit. This is due to the fact that linkage disequilibrium can be generalized to take into account correlation between corresponding bits in different length classes. Similarly, for h_1h_2 the BBs are the length class two schema $*h_2$ and any string of any length that contains h_1. In both cases we see that one of the BBs is associated with a coarse graining over all possible length classes and hence is not a schema associated with a fixed length class. Explicitly,

$$P(h_1,t) = D^t_{h_1}\Delta_{h_1} + P^*(h_1) \tag{26}$$

and

$$P(h_1h_2,t) = D^t_{h_1h_2}(\Delta_{h_1h_2} + \frac{P(*h_2)}{P(*^2)}\Delta_{h_1}) - D^t_{h_1}\frac{P(*h_2)}{P(*^2)}\Delta_{h_1} + P^*(h_1h_2) \tag{27}$$

We now consider the solution for strings of length $N \leq 3$. For $N_m = 3$ we have

$$P(h_1h_2h_3,t+1) = (1 - p_cA(h_1h_2h_3))P(h_1h_2h_3,t) + \frac{p_c}{2}P(h_1,t)P(*h_2h_3,t) + \frac{p_c}{3}(P(h_1*,t)P(*h_2h_3,t) + P(h_1h_2,t)P(**h_3,t)) + \frac{p_c}{4}(P(h_1**,t)P(*h_2h_3,t)) + P(h_1h_2*,t)P(**h_3,t) \tag{28}$$

where $A(h_1h_2h_3) = (P(*^1)/2 + 2P(*^2)/3 + P(*^3)/2)$. Once again this is a linear equation in $P(h_1h_2h_3,t)$ but with sources for which we have to solve equations for length one and two strings and 1-schemata from two strings and 1- and 2-schemata from length-three strings. Analogously to the case $N_m = 2$ length-one strings satisfy an equation that is coupled to 1-schemata of different length, in this case $P(h_1*,t)$ and $P(h_1**,t)$. However, as in the length-two case using the conservation law $P(h_1\#,t) = P(h_1,t) + P(h_1*,t) + P(h_1h_2*,t) = P(h_1\#,0)$ allows us to write the equation as

$$P(h_1,t+1) = D_{h_1}P(h_1,t) + \frac{p_c}{2}P(*^1)P(h_1\#,t) \tag{29}$$

The solution and associated fixed point are given by (26) as in the case $N_m = 2$ above. Length-two strings satisfy

$$P(h_1h_2,t+1) = (1 - p_cA(h_1h_2))P(h_1h_2,t) + \frac{p_c}{2}P(h_1,t)P(*h_2,t) + \frac{p_c}{3}P(h_1*,t)P(*h_2,t) + \frac{p_c}{3}P(h_1**,t)P(*h_2,t) + \frac{p_c}{3}P(h_1h_2*,t)P(*^2) \tag{30}$$

Thus we see a coupling to length-one and length-three sources. The 1-schemata equations for $P(h_1*,t)$ and $P(*h_2,t)$ however can be solved by eliminating length-three sources using the conservation law $P(*h_2\#,t) = P(*h_2,t) + P(*h_2*,t) = P(*h_2\#,0)$. One obtains

$$P(h_1*,t) = D^t_{h_1*}(\Delta_{h_1*} + \frac{P(*^2)}{P(*\#)}\Delta_{h_1}) - \frac{P(*^2)}{P(*\#)}\Delta_{h_1} + P^*_{h_1*} \tag{31}$$

where Δ_{h_1*}, Δ_{h_1} and $P^*_{h_1*}$ are as above in the $N_m = 2$ case. To solve (28) we still require $P(h_1**,t)$, $P(*h_2*,t)$, $P(**h_3,t)$, $P(*h_2h_3,t)$ and $P(h_1h_2*,t)$. $P(**h_3,t)$ is conserved as the final bit of the longest string cannot mix with anything else and therefore is unaffected by inter-length-class allele diffusion. $P(*h_2*,t)$ can be solved for in terms of the solution of $P(*h_2,t)$. $P(h_1**,t)$ obeys

$$P(h_1**,t+1) = (1 - \frac{p_c}{2}P(*^1) - \frac{2p_c}{3}P(*^2))P(h_1**,t) + \frac{p_c}{2}P(*^3)P(h_1,t) + \frac{2p_c}{3}P(*^3)P(h_1*,t) \tag{32}$$

As we already have the solution for $P(h_1*,t)$ and $P(h_1,t)$ this can simply be solved for. $P(*h_2h_3,t)$ satisfies

$$P(*h_2h_3,t+1) = (1 - \frac{p_c}{3}P(*^2) - \frac{p_c}{4}P(*^3))P(*h_2h_3,t) + \frac{p_c}{3}P(**h_3)P(*h_2,t) + \frac{p_c}{4}P(**h_3)P(*h_2*,t) \tag{33}$$

Once again, given that we have the solutions for $P(*h_2,t)$ and $P(*h_2*,t)$ this can be simply solved. Finally, $P(h_1h_2*,t)$ satisifes

$$P(h_1h_2*,t+1) = (1 - \frac{p_c}{2}P(*^1) - \frac{2p_c}{3}P(*^2) - \frac{p_c}{4}P(*^3))P(h_1h_2*,t) + \frac{p_c}{3}P(*^3)P(h_1h_2,t) + \frac{p_c}{2}P(*h_2*,t)(P(h_1,t) + \frac{2}{3}P(h_1*,t) + \frac{1}{2}P(h_1**,t)) \tag{34}$$

This is the only non-trivial equation left to solve as it is coupled to $P(h_1h_2,t)$. Both equations are first order linear inhomogeneous difference equations and can be decoupled by going to a second order linear inhomogeneous difference equation which can be readily solved. Due to length constraints we will present the results elsewhere. With these solutions in hand $P(h_1 h_2 h_3, t)$ may readily be solved for.

It is worth taking stock of what we have done here. In the case $N_m = 2$, in terms of the underlying string variables, there are six coupled equations to be solved. By going to a coarse-grained schema, or BB basis, one is able to implement the conservation laws most naturally, thereby decoupling the equations and finding an exact, explicit solution. For $N_m = 3$ there are fourteen coupled equations. The only extra complication relative to the $N_m = 2$ case however was the fact that after implementing the conservation laws two equations remained non-trivially coupled and had to be decoupled by going to a higher order difference equation.

5 Explicit Solutions - N_m arbitrary

In this section we wish to make some observations about the general case - N_m arbitrary. An important element, seen in the last section, is the existence of conservation laws which may be used to facilitate the solution of the dynamics. Generally, the conserved quantities are

$$P(*^{i-1} h_i \#, t) = P(*^{i-1} h_i \#, 0) \tag{35}$$

of which there are N_m. Hence, from the dynamical equations one may eliminate N_m variables. As in the above cases of $N_m = 2,\ 3$ one may use this fact to obtain the exact dynamics of certain schemata. These conservation laws are more naturally expressed in terms of schemata rather than strings. For instance, the conservation law $P(1\#, t) = \mathrm{constant}$ in terms of string variables is $P(1,t)+P(11,t)+P(10,t)+P(100,t)+P(101,t)+P(110,t)+P(111,t) = \mathrm{constant}$. This is a difficult constraint to implement at the level of the string equations themselves.

As we have emphasized, with the coarse-grained BB approach advocated here dynamical solutions are built up hierarchically beginning with low order BBs and proceeding to higher ones. As the lowest order ones are 1-schemata it is of interest to investigate the general equation for a 1-schema from length class N. One finds that

$$\begin{aligned} P(*^{i-1} h_i *^{N-i}, t+1) &= A_1 P(*^{i-1} h_i *^{N-i}, t) + \\ A_2 \sum_{j \geq i; j \neq N} A_3(j) P(*^{i-1} h_i *^{j-i}, t) \end{aligned} \tag{36}$$

where

$$A_1 = i \left(\sum_{j>N} \frac{P(*^j)}{N+1} + \sum_{j=i}^{N} \frac{P(*^j)}{j+1} \right) + \sum_{j=1}^{i-1} P(*^j)$$
$$+ P(*^N) \left(\frac{N-i+1}{N+1} \right)$$

$$A_2 = P(*^N)$$
$$A_3 = \left(1 - \delta(j > N) \frac{i}{N+1} - \delta(k \leq j \leq N) \frac{i}{j+1} \right)$$

Note that 1-schemata from other than length-class N strings act as sources for h_i, however, there are no more "primitive", i.e. lower order, sources. Hence, in the sense of section 2 this equation is really homogeneous with no BB sources and hence can be written as

$$\mathbf{P}(t+1) = \mathbf{A}\mathbf{P}(t) \tag{37}$$

where the elements of the matrix $\mathbf{A}$ can be read off from (36) and the values of the coefficients A_1, A_2 and A_3. The diagonalization of this matrix yields the decay rates of the various 1-schemata. With the 1-schemata solution in hand we may start to reconstruct the 2-schemata respecting the hierarchical structure outlined in section 2. We will not pursue this further in this paper restricting attention to some more specific results.

From (35) one immediately sees that the quantity $P(*^{N_m-1} h_{N_m}, t)$ is conserved. Additionally, for the length-one strings all "sources" $P(h_1 *^{j-1})$ for $P(h_1, t)$ appear with the same coefficient, $p_c/2$. Hence, $P(h_1, t)$ satisfies (26) the only difference now being that $P(h_1\#) = \sum_{j=2}^{N_m} P(h_1 *^{j-1}, t)$.

Using the conservation of the last bit of the longest string one may also determine the evolution of the last bit of the next longest string and the last bit of the string of length $N = N_m - 1$ by using the conservation law $P(*^{N_m-2} h_{N_m-1}\#) = \mathrm{constant}$. For the next to last bit of the longest string the solution is

$$\begin{aligned} P(*^{N_m-2} h_{N_m-1} *, t) &= \\ D^t_{*^{N_m-2} h_{N_m-1} *} \Delta_{*^{N_m-2} h_{N_m-1} *} &+ P^*_{*^{N_m-2} h_{N_m-1} *} \end{aligned} \tag{38}$$

where $D_{*^{N_m-2} h_{N_m-1} *} = (1 - (1/N_m)(P(*^{N_m-1}) + P(*^{N_m})))$ and $P^*_{*^{N_m-2} h_{N_m-1} *} = P(*^{N_m}) P(*^{N_m-2} h_{N_m-1}\#)/P(*^{N_m-1}\#)$ which is the expected fixed point from (11).

6 Conclusions

We have investigated the dynamics of variable-length GAs using a coarse-grained BB representation of the dynamical equations. We showed that the formal solution of the equations could be interpreted in an analogous manner to that of the fixed length case, i.e. the hierarchical construction of more fine-grained schemata from their more coarse-grained BBs. The novel element here is that these BBs could come from strings of different lengths. We discussed briefly the fixed point distribution of the equations for a flat fitness landscape using a one-point homologous crossover operator and no mutation showing how a generalization of

Robbins proportions emerged that involved a generalized notion of a BB. We then turned to a more explicit construction of the entire dynamics and quantified the approach to the fixed point. For $N_m = 2,\ 3$ we were able to find explicit solutions utilizing the existence of conservation laws for certain quantities. This in itself shows the utility of the coarse grained BB representation, the $N_m = 3$ problem at the string level corresponding to 14 simultaneous first order difference equations which need to be solved.

From the resultant solutions we were able to investigate the phenomenon of inter-length-class allele diffusion. We saw that the diffusion rates, or mixing times, for different alleles or combination of alleles depended strongly on the length distribution of strings, which in the case of a flat fitness landscape is time independent. For instance, the diffusion rate for the allele h_1 in length-class-three strings is slower than that of the same allele in length-class-two or one strings if $P(*^1) + (4/3)P(*^2) > 1$ which is the case if the proportion of length-three strings is small. We also can see that the closer the string bit to the beginning of the string then typically the faster it mixes, simply because there are more things with which it can mix. In this sense in the variable length case the degree of exploration versus expolitation carried out by recombination is inhomogeneous depending on the bit's position in the string and the distribution of lengths, diversity being encouraged more at the beginning of strings than at the end. Another interesting aspect of inter-length-class allele diffusion is the fact that for a given length class a lost allele from a particular bit position can be recovered if the allele exists in the corresponding bit of another length class string.

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 and the University of Essex for their support. Alden Wright did this work while visiting the University of Birmingham, supported by EPSRC grant GR/R47394.

References

[1] Stephens, C.R. (2001) "Some Exact Results from a Coarse Grained Formulation of Genetic Dynamics". In L. Spector *et al* eds. *Proceedings of GECCO 2001*, 631-638 (Morgan Kaufmann, San Francisco).

[2] Wright, Rowe, J., Poli, R. and Stephens, C.R. (2002) "A Fixed Point Analysis of a Genepool GA with Mutation", accepted for publication (full paper) in *GECCO 2002*.

[3] Mähnig, T. and Mühlenbein, H. (2001) "Optimal Mutation Rate Using Bayesian Priors for Estimation of Dsitribution Algorithms", in *Stochastic Algorithms: Foundations and Applications*, ed. K. Steinhüfel, LNCS Springer-Verlag.

[4] Vose, M.D. (1999) *The Simple Genetic Algorithm: Foundations and Theory*, (MIT Press, Cambridge MA).

[5] Holland, J.H. (1975) *Adaptation in Natural and Artificial Systems* (MIT Press, Cambridge, MA).

[6] Goldberg, D. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison Wesley, MA).

[7] Stephens, C.R. and Waelbroeck, H. (1997), "Effective Degrees of Freedom in Genetic Algorithms and the Block Hypothesis", *Proceedings of the ICGA7*, ed. T. Bäck, 34-41 (Morgan Kaufmann, San Mateo).

[8] Stephens, C.R. and Waelbroeck, H. (1998) "Analysis of the Effective Degrees of Freedom in Genetic Algorithms", *Physical Review* **D57** 3251-3264.

[9] Stephens, C.R. and Waelbroeck, H. (1999) "Schemata Evolution and Building Blocks", *Evol. Comp.* **7(2)** 109-124.

[10] Poli, R. (2000) "Exact Schema theorem and Effective Fitness for GP with one-point crossover", in *Proceedings of GECCO2000*, eds D. Whitley *et al* 469-476 (Morgan Kaufmann).

[11] Poli, R. and McPhee, N.F. (2000) "Exact Schema Theory for GP and Variable-length GAs with Homologous Crossover", *Proceedings of GECCO-2001*, ed. Lee Spector *et al* 104-111 (Morgan Kaufmann, San Mateo).

[12] Nordin, P. (1994) "A Compiling Genetic Programming System that Directly Manipulates the Machine Code", *Advances in Genetic Programming*, ed. K.E. Kinnear Jr., 311-331 (MIT Press).

[13] O' Neil, M. and Ryan, C. (2001) "Grammatical Evolution" *IEEE Transaction on Evolutionary Compuation*, in press.

[14] Wu, A.S. and Banzhaf, W. (1998) "Introduction to the Special Issue: Variable-Length Representation and Noncoding Segments for Evolutionary Algorithms" *Evolutionary Computation* **6(4)**, iii-iv.

[15] Goldberg, D.E. (1987) *"Simple Genetic Algorithms and the Minimal, Deceptive Problem"*, in *Genetic Algorithms and Simulated Annealing*, ed. L. Davis, 74-88 (Pitman, London).

[16] Spears, W.M. (2000) "Limiting distributions for mutation and recombination", in *Proceedings of FOGA 6*, eds. W.M. Spears and W. Martin, (Morgan Kaufmann, San Mateo).

[17] Geiringer, H. (1944) "On the Probability Theory of Linkage in Mendelian Heredity", *Annals of Mathematical Statistics* **15**, 25-27.

[18] Poli, R., Rowe, J., Stephens, C.R. and Wright, A. (2002) "On the Search Biases of Homologous Crossover in Linear Genetic Programming and Variable-length Genetic Algorithms", accepted for publication (full paper) in *GECCO 2002*; University of Essex Computer Science Technical Report TRCSM-352.