

Coarse-Grained Dynamics for Generalized Recombination

Christopher R. Stephens and Riccardo Poli

Abstract—An exact microscopic model for the dynamics of a genetic algorithm with generalized recombination is presented. Generalized recombination is a new model for the exchange of genetic material from parents to offspring that generalizes and subsumes standard operators, such as homologous crossover, inversion and duplication, and in which a particular gene in the offspring may originate from *any* parental gene. It is shown that the dynamics naturally coarse grains, the appropriate effective degrees of freedom being schemata that act as building blocks. It is shown that the Schema dynamics has the same functional form as that of strings and we derive a corresponding exact Schema theorem. To exhibit the qualitatively new phenomena that can occur in the presence of generalized recombination, and to understand the biases of the operator, we derive a complete, exact solution for a two-locus model without selection, showing how the dynamical behavior is radically different to that of homologous crossover. Inversion is shown to potentially introduce oscillations in the dynamics, while gene duplication leads to an asymmetry between homogeneous and heterogeneous strings. All nonhomologous operators lead to allele “diffusion” along the chromosome. We discuss how inferences from the two-locus results extend to the case of a recombinative genetic algorithm with selection and more than two loci providing evidence from an integration of the exact dynamical equations for more than two loci.

Index Terms—Adaptive systems, genetic algorithms (GAs), search methods.

I. INTRODUCTION

COARSE-GRAINED formulations of the dynamics of evolutionary algorithms (EAs) offer many advantages relative to “microscopic” ones based purely on the string/chromosome degrees of freedom. These benefits have been exhibited in both the standard genetic algorithm (GA) [31], and in variable-length GAs/linear genetic programming (GP) and GP itself [19]. The main advantage is that they provide a simpler but deeper understanding of the role of homologous recombination, [19], [31], wherein the most appropriate effective degrees of freedom for describing recombination are not strings/chromosomes, but coarse-grained “building blocks,” with which the

EA builds strings. The form that these building blocks take depends on the representation used. For instance, in GAs, they are a particular subset of schemata that form an alternative and more appropriate basis—the building block basis (BBB) [3]. In the case of variable-length strings and trees, they are generalizations of those found in fixed length GAs—building block hyperschemata [19]. Additionally, coarse-grained formulations have shown a theoretical commonality between GAs and GP that was not previously apparent, thus leading to a unification of the theoretical underpinnings of both areas. They have also led to a deeper understanding of approximate Schema theorems, such as Holland’s [12], and the Building Block Hypothesis [9]. As they are exact, they have also served as a bridge between this latter work and dynamical systems models [34] (see, for example, [20]).

Up to now, coarse-grained formulations have been studied for GAs and GP with both homologous and subtree crossover and, for GAs, in the presence of standard point mutation. In nature, however, there are many more ways of combining parental genetic material into an offspring than homologous crossover, many of which have been used in EAs and have been found to be useful by practitioners. Gene duplication, for example, has been studied in biology [4], as well as in the context of GAs [25] and GP [14], while inversion was one of the operators used by Holland [12] in the original formulation of the GA. Additionally, there is little to no theoretical analysis in the evolutionary computation (EC) literature concerning inversion and duplication, at least not based on an underlying exact dynamical model.

In this paper,¹ the notion of generalized recombination is introduced, which can account for *any* redistribution of parental genes into the offspring. This requires the generalization of the concept of a crossover mask to that of a generalized crossover mask (GCM), with an associated generalized recombination distribution (GRD). Generalized recombination generalizes and subsumes many other common genetic operators, including homologous crossover and inversion, as well as gene duplication and deletion. Thus, by studying this more general operator theoretically, we are simultaneously developing a framework within which several different familiar and used operators can be studied. Generalized recombination can also lead to qualitatively new phenomena that are not present in the case of homologous crossover, such as periodic oscillations in the dynamics of strings and schemata in the presence of inversion, and a preference for homogeneous strings and schemata over heterogeneous ones in the presence of duplication, e.g., for

Manuscript received June 23, 2005; revised November 22, 2005 and July 3, 2006. This work was supported in part by ESPRC under Grant GR/T24616/01, in part by the Conacyt Project 30422-E and the Universidad Nacional Autonoma de México (UNAM) Macroproyecto—“Tecnologías para la Universidad de la Información y la Computación” and in part by Leverhulme Trust. The work of C. R. Stephens was supported by a Sabbatical Fellowship from DGAPA of the UNAM.

C. R. Stephens is with the Instituto de Ciencias Nucleares, Universidad Nacional Autonoma de México, A. Postal 70-543, México, D.F. 04510 (e-mail: stephens@nucleares.unam.mx).

R. Poli is with the Department of Computer Science, University of Essex, Wivenhoe CO4 3SQ, U.K. (e-mail: rpoli@essex.ac.uk).

Digital Object Identifier 10.1109/TEVC.2006.884043

¹This paper is an extension of earlier, preliminary work [23], [33] presented at CEC 2005.

two loci a potential preference for 11 over 10 even if they are equally fit.

We initially study generalized recombination in the context of a variable-length representation with mutation, deriving an exact dynamical systems type model for describing it. More importantly, however, we present a detailed analysis of how the dynamics of generalized recombination, as in the case of homologous crossover, is much more naturally represented in terms of building block schemata, as opposed to strings. The difference in this case is that due to the presence of operators other than homologous recombination, a richer diversity of building blocks enters into the dynamics. We show that the dynamics is form invariant under a coarse graining when passing from strings to schemata, and hence derive an exact Schema theorem for a fixed-length GA evolving in the presence of mutation and generalized recombination. Interpreted in terms of the concept of *effective* fitness [27]–[30], this exact Schema theorem states that, as in the case of homologous crossover, those schemata which are more effectively fit propagate preferentially. This coarse-grained formulation of generalized recombination, as in previous analyses, offers a further theoretical unification for GAs, showing that building blocks also appear naturally in the context of a GA where a gene of the offspring is derived from any of the parental genes.

After introducing the exact Schema theorem for generalized recombination and mutation, we focus our attention on a detailed analysis of a two-locus model without selection in the infinite population limit. Of course, one might question to what extent a two-locus model can illuminate the more complicated multilocus case. It is wise to remember, however, that in population biology, such models have played a crucial role, permitting the qualitative, and sometimes quantitative, analysis of a host of important phenomena (see, for instance, [1] and references therein). Even in EC, such models have made important appearances, such as in the deceptive two-bit problem [6], and in previous analyses of the effects of recombination and mutation [26]. The model we will present has the advantage of being exactly soluble, while at the same time being quite transparent. Additionally, all the interesting phenomena observed in this two-locus model have also been shown to be present in the case of multilocus models with selection, as we explicitly demonstrate by considering some three-locus results with different fitness landscapes. One may also question the relevance of an infinite population model. The relevance of the infinite population model for finite population dynamics has been discussed extensively in the context of the canonical GA [34]. The same arguments apply here. In particular, for a population of size N , one expects finite population effects to be $O(N^{-1/2})$ and, hence, small for large populations. Additionally, many of the elements we discuss here also enter into a formulation of the dynamics in terms of a Markov chain, which is the appropriate general framework for the finite population model.

II. GENERALIZED RECOMBINATION

In nature, there are a multitude of ways that genetic material can be distributed from parents to offspring. Some involve two

parental chromosomes, such as “homologous” recombination and translocation—the latter being the breakage and removal of a large segment of DNA from one chromosome, followed by the segment’s attachment to a different chromosome. Others involve only one parental chromosome, such as inversion, duplication, and deletion. “Homologous” in biology can have different meanings. In the context of meiosis [15] in diploids, it refers to the recombination that takes place between “homologous” pairs of chromosomes, a homologous pair being such that the i th locus in each member of the pair codes for the same gene, even though the particular allele might be different, e.g., green eyes versus blue eyes as opposed to green eyes and brown hair. It can also refer, however, to the fact that a subset of genes or nucleotides in a pair have the same structure, and hence might serve as a preferred point at which exchange of genetic material can take place. Translocation is of this type, and in EC is often termed “unequal” crossing over in order to distinguish it from the more familiar homologous crossover, where parental chromosomes are first “aligned” so that homologous genes are in corresponding positions.

The most well-known operator for transferring genetic material in GAs is homologous recombination where alleles at a particular locus have their origin in the corresponding genetic loci of the parental chromosomes. Such recombination can be succinctly modeled using the concept of a recombination mask, \mathbf{m} , which, for strings of length ℓ , can be represented by an ℓ -dimensional vector $\mathbf{m} = (m_1, m_2, \dots, m_\ell)$, where $m_i = 0, 1$ indicates from which parent the i th allele is taken—0 meaning take it from the i th locus of the first parent, and 1, from the i th locus of the second parent. The total number of possible masks is 2^ℓ . Associated with them is a recombination distribution, denoted by $p_c(\mathbf{m})$. If the probability to perform crossover is p_{xo} , then $p_c(\mathbf{m})$ is the conditional probability for choosing the mask \mathbf{m} given that crossover was implemented. Hence, $\sum_{\mathbf{m}} p_c(\mathbf{m}) = 1$ and $p_{xo} \times p_c(\mathbf{m})$ is the probability to crossover using the mask \mathbf{m} . It is the choice of $p_c(\mathbf{m})$ that specifies if one is considering one-point, two-point, uniform crossover, etc.

Although binary masks are sufficient to model homologous genetic operators in a fixed length setting, to describe more general ones where an allele in a particular locus of the offspring comes from a different locus in either one of the parents, a new level of generality is required. We will term such a generalization—a GCM. The associated probability distribution over the GCMs generalizes the concept of a recombination distribution and will be termed a GRD. A GCM can be specified mathematically using several equivalent representations. If we consider the more general case of variable-length strings, in order to be able to account for unequal crossing over as well as homologous crossover, a GCM has to specify how ℓ alleles, for an offspring of length ℓ , are obtained from two parents of lengths ℓ_1 and ℓ_2 , respectively.

We call the representation of a GCM that is closest to that of a standard crossover mask a *recombination vector*, \mathbf{v} . As the i th allele in the offspring could come from any locus in the parents, the components of this vector can take values from the set

$\mathcal{N}_{\ell_1+\ell_2} = \{1, \dots, \ell_1+\ell_2\}$, values from 1 to ℓ_1 denoting that the allele originated in the first parent, and values between ℓ_1+1 and $\ell_1+\ell_2$ signifying that it came from the second. Thus, in this representation, we denote a GCM by $\mathbf{v}(\ell, \ell_1, \ell_2) = (v_1, \dots, v_\ell)$, with $v_i \in \mathcal{N}_{\ell_1+\ell_2}$. For example, for $\ell = 3$, $\mathbf{v}(3, 2, 4) = (1, 5, 3)$ represents a GCM that takes genetic material from parents of lengths 2 and 4, respectively, and recombines it into an offspring of length 3. The notation $(1, 5, 3)$ for the components of $\mathbf{v}(3, 2, 4)$ signifies that the first gene of the offspring came from the first gene of the first parent, the second gene from the third gene of the second parent, and the last from the first of the second parent. The total number of GCMs is $(\ell_1+\ell_2)^\ell$. The associated distribution of probabilities, $p_c(\mathbf{v}(\ell, \ell_1, \ell_2))$, then determines the GRD. For a fixed-length representation, we will drop the string length arguments. In this case, the total number of recombination vectors is $(2\ell)^\ell$.

Another equivalent representation for GCMs uses arrays (crossover matrices) instead of bit strings (vectors) to represent recombination events. An appropriate crossover matrix for representing the formation of a length ℓ offspring from length ℓ_1 and ℓ_2 parents will have ℓ rows and $(\ell_1 + \ell_2)$ columns. The first ℓ_1 columns indicate which alleles are copied from the first parent, while columns $\ell_1 + 1$ through $\ell_1 + \ell_2$ indicate what is provided by the second. The matrix elements are either 0 or 1, where a 1 in row r and column c means that locus r in the offspring is filled with the allele from locus c in the first parent if $c \leq \ell_1$. If $c > \ell_1$, it is filled with the allele from locus $c - \ell_1$ of the second parent. As an offspring would not be fully specified if some of its alleles were undefined, or would be overly specified if we tried to place more than one allele in a locus, in each row of a crossover matrix there must be exactly one 1 (with all other elements in the row being 0). For a fixed length representation, the matrices are of constant size $\ell \times 2\ell$.

Finally, another useful representation is that of a *recombination pair*, $\mathbf{r}(\ell, \ell_1, \ell_2) \equiv (\mathbf{m}, \mathbf{v})$, which is a hybrid between the notion of a standard crossover mask and a recombination vector with a reduced cardinality alphabet. Here, $\mathbf{m} = (m_1 \dots m_\ell)$ is an ℓ -dimensional vector (i.e., $m \in \{0, 1\}^\ell$) whose components specify which parent contributes the alleles to fill each locus in the offspring. So, for example, $m_i = 0$ means locus i will be filled with an allele from parent 1, while $m_i = 1$ means parent 2 will contribute the allele instead. Having specified from which parent a particular allele originates, one must specify from which locus. This is achieved by specifying $v(\ell, \ell_1, \ell_2) = (v_1, \dots, v_\ell)$, an ℓ -component vector of integers with components $v_i \in \{1, \dots, \ell_1\}$ if $m_i = 0$ (i.e., $v \in \mathcal{N}_{\ell_1}$), and $v_i \in \{1, \dots, \ell_2\}$ if $m_i = 1$. The elements of $\mathbf{v}(\ell, \ell_1, \ell_2)$ specify which alleles from a particular parent will be transferred to the offspring. In the case of a recombination pair, for variable-length strings, the cardinality of the alphabet from which the v_i are taken depends on the corresponding m_i . Hence, this representation is less convenient than that of a recombination vector for variable-length strings. For fixed-length strings, however, it is perfectly natural, and actually presents several advantages relative to the recombination vector notation.

As an example of how the different representations of a GCM work, consider standard one-point crossover for $\ell = 3$. The associated crossover matrices are

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

each invoked with probability 1/2. These are equivalent to the recombination vectors $v_1 = (1, 5, 6)$ and $v_2 = (1, 2, 6)$, or to the recombination pairs $r_1 = (011, (1, 2, 3))$ and $r_2 = (001, (1, 2, 3))$, or to the more traditional crossover masks 011 and 001.

Although, for formal manipulations we will use the more powerful recombination pair notation as we will treat the two-locus case extensively in this paper, we will represent the GRD there using the recombination vector notation. The GRD then is a vector with components p_{ab} , where the indices a and b take values from one to four, one and two corresponding to the first and second loci of the first parent and three and four the corresponding loci of the second parent. Thus, for example, p_{13} gives the probability for applying the GCM associated with finding the first locus of the offspring from the first locus of the first parent and the second locus from the first locus of the second parent.

III. EVOLUTION EQUATION IN THE STRING BASIS

Let us start by writing down the exact equations for the expected population at time $t + 1$ for a finite population of fixed-length strings of length ℓ and cardinality \mathcal{A} evolving in the presence of selection, mutation, and generalized recombination

$$E(P_I(t + 1)) = \sum_J M_I^J \left((1 - p_{xo}) P_J^I(t) + p_{xo} \sum_{\mathbf{r}} p_c(\mathbf{r}) \sum_K \sum_L \lambda_J^{KL}(\mathbf{r}) P_K^I(t) P_L^I(t) \right) \quad (1)$$

where $P_I^I(t)$ is the probability to select string $I \in \Omega$ at time t , where Ω is the set of such strings, $P_I(t)$ being the proportion of string type I at time t . M_I^J is the probability that string J mutates to string I . p_{xo} is the probability that recombination occurs and $p_c(\mathbf{r})$ is the conditional probability to apply the recombination pair \mathbf{r} . $\lambda_J^{KL}(\mathbf{r})$ is an indicator function that takes value 1 when the result of applying the GCM to the parents K and L is J and is zero, otherwise.² The first term in (1) arises from the cloning of J and its subsequent mutation into I , while the second term represents all the ways in which J may be constructed from other strings via generalized recombination and subsequently mutated.

²Note that the arrangement of indices I, J, K , and L leave (1) in ‘‘covariant’’ form [3]. Thus, by using the rules of tensor calculus each element in the equation may be transformed to a different coordinate basis, such as the Walsh or BBB, using the appropriate coordinate transformation matrix.

Equation (1) represents the relation between the actual values of $P_I(t)$ at generation t , and the *expected* value $E(P_I)$ at generation $t + 1$, where the expectation value is over an ensemble of realizations of the stochastic process that takes the system from t to $t + 1$. Generally, $E(P_I(t)) \neq P_I(t)$ except in the infinite population limit. As the expected string frequencies in generation $t + 1$ are expressed in terms of the selection probabilities of *strings* at generation t , we will say that this evolution equation is expressed in the *string basis*.

We can now directly pass to the more general case of variable-length strings. Using the notation $P_{I^\ell}(t)$ for the proportion of genotypes of type $I^\ell \in \Omega_\ell$ in strings of length ℓ , we have

$$\begin{aligned} E(P_{I^\ell}(t+1)) &= \sum_{J^\ell} M_{I^\ell}^{J^\ell} \\ &\times \left((1 - p_{xo}) P'_{J^\ell}(t) + p_{xo} \sum_{\ell_1, \ell_2} \sum_{\mathbf{r}(\ell, \ell_1, \ell_2)} p_c(\mathbf{r}(\ell, \ell_1, \ell_2)) \right. \\ &\times \left. \sum_{K^{\ell_1} \in \Omega_{\ell_1}} \sum_{L^{\ell_2} \in \Omega_{\ell_2}} \lambda_{J^\ell}^{K^{\ell_1} L^{\ell_2}}(\mathbf{r}(\ell, \ell_1, \ell_2)) P'_{K^{\ell_1}}(t) P'_{L^{\ell_2}}(t) \right) \end{aligned} \quad (2)$$

where the notation is an obvious generalization of that above for the fixed-length case. $\lambda_{J^\ell}^{K^{\ell_1} L^{\ell_2}}(\mathbf{r}(\ell, \ell_1, \ell_2))$ is once again an indicator function with value 1 if the offspring J^ℓ is formed given parents K^{ℓ_1} and L^{ℓ_2} and a GCM $\mathbf{r}(\ell, \ell_1, \ell_2)$, and zero, otherwise. Although P'_{I^ℓ} accounts for any selection (with replacement) mechanism, the relation between P'_{I^ℓ} and P_{I^ℓ} depends on the specifics of the selection operator. For instance, for proportional selection $P'_{I^\ell}(t) = (f_I/\bar{f}(t)) P_{I^\ell}(t)$, where $\bar{f}(t)$ is the average population fitness. Although (1) and (2) are valid for arbitrary alphabets, the specific values that $M_{I^\ell}^{J^\ell}$ takes depend on the cardinality of the alphabet. For binary alleles, for instance, $M_{I^\ell}^J = p_m^{d(I, J)} (1 - p_m)^{\ell - d(I, J)}$, where $d(I, J)$ is the Hamming distance between I and J and p_m is the mutation rate. Equations (1) and (2) describe the dynamics of genetic systems that recombine genetic material from two parents in a general way, and extend known results for exact evolution equations for the canonical GA [34], and for variable-length GAs with either homologous [19], [22] or subtree-type [21] crossover, to the more general class of variable or fixed-length strings and generalized recombination.

Inheritance via recombination can be understood by considering it gene by gene. This is reflected in the nature of $\lambda_{J^\ell}^{K^{\ell_1} L^{\ell_2}}(\mathbf{r}(\ell, \ell_1, \ell_2))$ as an indicator function, whereby the truth of the assertion that J^ℓ is the offspring of K^{ℓ_1} and L^{ℓ_2} with respect to the GCM $\mathbf{r}(\ell, \ell_1, \ell_2)$ can be seen as the logic conjunction (for all i) of the assertions that allele J_i^ℓ is a result of the crossover between alleles $K_{v_i}^{\ell_1}$ and $L_{v_i}^{\ell_2}$ using the GCM component $r_i(\ell, \ell_1, \ell_2)$ which determines m_i and v_i . Thus, one may write

$$\lambda_{J^\ell}^{K^{\ell_1} L^{\ell_2}}(\mathbf{r}(\ell, \ell_1, \ell_2)) = \prod_{i=1}^{\ell} \lambda_{J_i^\ell}^{K_{v_i}^{\ell_1} L_{v_i}^{\ell_2}}(r_i(\ell, \ell_1, \ell_2)) \quad (3)$$

where we represent *false* with 0 and *true* with 1, and so the product represents a logical and. Note that allele J_i^ℓ is a result of the crossover between alleles $K_{v_i}^{\ell_1}$ and $L_{v_i}^{\ell_2}$ in two and only two mutually exclusive cases: 1) $m_i = 0$ and $J_i^\ell = K_{v_i}^{\ell_1}$ or 2) $m_i = 1$ and $J_i^\ell = L_{v_i}^{\ell_2}$. Therefore

$$\lambda_{J_i^\ell}^{K_{v_i}^{\ell_1} L_{v_i}^{\ell_2}}(r_i(\ell, \ell_1, \ell_2)) = \left((1 - m_i) \delta_{J_i^\ell}^{K_{v_i}^{\ell_1}} + m_i \delta_{J_i^\ell}^{L_{v_i}^{\ell_2}} \right) \quad (4)$$

where $+$ acts as a logic or for mutually exclusive events, $1 - x$ represents the negation of x and δ_i^j is the Kronecker delta ($\delta_i^j = 1$ if $j = i$ and zero, otherwise) which we use as an equality predicate.

As an example, consider for fixed length strings with $\ell = 3$ the recombination pair $\mathbf{r} = (001, (3, 1, 2))$. In this case

$$\begin{aligned} \lambda_{J^{KL}}^{KL}((001, (3, 1, 2))) &= \left(1\delta_{J_1}^{K_3} + 0\delta_{J_1}^{L_3} \right) \left(1\delta_{J_2}^{K_1} + 0\delta_{J_2}^{L_1} \right) \left(0\delta_{J_2}^{K_1} + 1\delta_{J_2}^{L_2} \right) \\ &= \delta_{J_1}^{K_3} \delta_{J_2}^{K_1} \delta_{J_3}^{L_2}. \end{aligned} \quad (5)$$

Naturally, the action of the projection operators is the same, no matter how the GCM is represented. So, if we use the recombination vector $\mathbf{v} = (3, 1, 5)$ to represent the recombination event \mathbf{r} , we have $\lambda_{J^{KL}}^{KL}((3, 1, 5)) = \delta_{J_1}^{K_3} \delta_{J_2}^{K_1} \delta_{J_3}^{L_2}$.

Note that (1) is functionally identical to that for the case of standard mask-based crossover [3], the only difference being the different recombination distribution, and hence the different set of $\lambda_{J^{KL}}^{KL}(\mathbf{r})$ that are nonzero. As in the standard homologous crossover case, for an alphabet of cardinality \mathcal{A} , we have \mathcal{A}^ℓ coupled, first-order difference equations to solve. The chief problem, however, is the fact that, on the right-hand side with the term $\sum_{\mathbf{r}} p_c(\mathbf{r}) \sum_K \sum_L \lambda_{J^{KL}}^{KL}(\mathbf{r}) P'_K(t) P'_L(t)$, we have $\mathcal{A}^\ell \times \mathcal{A}^\ell \times (\mathcal{A}^\ell)^\ell = \mathcal{A}^{3\ell} \times \ell^\ell$ possible contributing terms, corresponding to the \mathcal{A}^ℓ possible configurations of the parental strings and the $(\mathcal{A}^\ell)^\ell$ GCMs.

A. String Example for $\ell = 2$

For example, for $\ell = 2$, there are 16 GCMs denoted in recombination vector notation by $\{(v_1, v_2)\}$, where v_1 and v_2 run over the values 1, 2, 3, and 4. The sums over the strings J and K run over the values 1 to \mathcal{A}^ℓ . Thus, for an arbitrary GRD, even at the two bit level there are $16 \times 4 \times 4 = 256$ terms $\lambda_{J^{KL}}^{KL}(\mathbf{r})$ to compute for a given string I . Symbolically, however, the terms are quite simple

$$\textit{cloning} \quad \lambda_{J_1 J_2}^{K_1 K_2 L_1 L_2}(1, 2) = \delta_{J_1}^{K_1} \delta_{J_2}^{K_2} \quad (6)$$

$$\textit{inversion} \quad \lambda_{J_1 J_2}^{K_1 K_2 L_1 L_2}(2, 1) = \delta_{J_1}^{K_2} \delta_{J_2}^{K_1} \quad (7)$$

$$\textit{crossover} \quad \lambda_{J_1 J_2}^{K_1 K_2 L_1 L_2}(1, 4) = \delta_{J_1}^{K_1} \delta_{J_2}^{L_2} \quad (8)$$

$$\textit{cross+inv} \quad \lambda_{J_1 J_2}^{K_1 K_2 L_1 L_2}(4, 1) = \delta_{J_1}^{L_2} \delta_{J_2}^{K_1} \quad (9)$$

$$\textit{dup 1} \quad \lambda_{J_1 J_2}^{K_1 K_2 L_1 L_2}(1, 1) = \delta_{J_1}^{K_1} \delta_{J_2}^{K_1} \quad (10)$$

$$\textit{dup 1} \quad \lambda_{J_1 J_2}^{K_1 K_2 L_1 L_2}(2, 2) = \delta_{J_1}^{L_1} \delta_{J_2}^{L_1} \quad (11)$$

$$\textit{dup 2} \quad \lambda_{J_1 J_2}^{K_1 K_2 L_1 L_2}(1, 3) = \delta_{J_1}^{K_1} \delta_{J_2}^{L_1} \quad (12)$$

$$\textit{dup 2} \quad \lambda_{J_1 J_2}^{K_1 K_2 L_1 L_2}(2, 4) = \delta_{J_1}^{K_2} \delta_{J_2}^{L_2} \quad (13)$$

where we use the recombination vector notation and show only those GCMs that correspond to creation of genotype J using K

as the first parent. The corresponding second parent terms can be found by interchanging K and L on the right-hand side of (6)–(13) and letting $(v_1, v_2) \rightarrow (v'_1, v'_2)$, where $v'_i = (v_i + 1) \bmod 4 + 1$. The meaning of these terms, as alluded to in equations (6)–(13), is the following: the terms represented by GCMs (1,2) and (3,4) are *cloning* terms due to the application of a trivial standard crossover mask, where both offspring alleles come from the corresponding loci of only one of the parents. The *inversion* term is represented by GCMs (2,1) (inversion of first parent) and (4,3) (inversion of second parent). The GCMs (1,4) and (3,2) represent the results of standard *one-point crossover*, while the GCMs (4,1) and (2,3) represent the results of standard *one-point crossover* followed by an *inversion* (or vice versa). The terms denoted by *duplication 1*—one-parent duplication—are associated with the GCMs (1,1), (2,2), (3,3), and (4,4) and represent duplication of an allele from a single locus of a single parent. Finally, the *duplication 2*—two-parent duplications—GCMs (1,3), (2,4), (3,1), and (4,2) represent gene duplication as well, but where the two genes of the offspring come from the same locus but in different parents.

Substituting these expressions in (1), computing all terms, expanding the sums \sum_K and \sum_L , setting $I_1 = i'$, $I_2 = j'$, $J_1 = i$, and $J_2 = j$ and omitting the argument t for conciseness, one finds (14) shown at the bottom of the page, where for simplicity, we are restricting attention to a binary alphabet and \bar{i} signifies the bit complement of i . The first term on the right-hand side is a cloning-mutation term due to the fact that with probability $(1 - p_{xo})$ strings are copied without recombination, P'_{ij} being the probability to select the genotype ij , and $M'_{ij'}$ being the probability to mutate it to the genotype $i'j'$. The meaning of the different terms in (14) is inherited from the meaning of the corresponding terms of the GRD, the p_{ab} being the notation for the GCM probability associated with the GCM (a, b) . The Kronecker delta, ensures that the contribution from gene duplication from a single parent is only present for homogeneous offspring, i.e., those with both allele values the same. Note that of the 256 possibilities there are only 44 nonzero terms in (14). However, in order to compute which terms are nonzero, all have to be computed. By way of comparison, the canonical GA with one-point crossover, where $p_{14} = p_{32} = 1/2$ with all other GCMs zero, has only eight nonzero terms out of the $2 \times 4 \times 4$ possible ones.

IV. COARSE-GRAINED EVOLUTION EQUATIONS

A. Strings

For both homologous and generalized recombination, it is clear that there is a great deal of redundancy in the string representation. In the case of homologous recombination, it has been found that a coarse-grained representation in terms of schemata makes the dynamics much more transparent; partly due to the fact that the number of terms on the right-hand side of (1) reduces to 2^ℓ for a binary alphabet in the case of general homologous crossover which, when compared with 8^ℓ in the string basis, is a substantial reduction in complexity. For a particular type of recombination distribution, such as one-point crossover, where there are only $(\ell - 1)$ nonzero masks, the simplification is even greater from $(\ell - 1)4^\ell$ to $(\ell - 1)$. One is naturally inclined to ask whether an appropriate simplification can be effected in this more general case. The answer is in the affirmative.

However, before proceeding further, we wish to clarify what we mean by *coarse graining*: The underlying microscopic degrees of freedom of the dynamical system are strings, a string specifying a state in the configuration space of the system. A schema, however, is a coarse-grained object, in a similar sense to that of a “block-spin” variable in the physics of phase transitions [10], [11], [13]. Moreover, as in these other areas, the coarse-grained variables are obtained by summing over the possible states of the degrees of freedom one is coarse graining. For instance, to get the schema $1*$, one sums over the possible states for the second bit. Thus, we speak of a coarse-grained formulation as we are using coarse-grained variables, understood in the canonical accepted sense, to describe the dynamics.³

One can simply illustrate how coarse-grained variables, i.e., schemata in this case, naturally arise by considering one partic-

³There is a second notion of coarse-graining which refers to whether the dynamics being considered is the complete dynamics or a projection onto a configuration space of reduced dimensionality. To give an example: with the schemata $1*$ and $0*$, for a binary system, one may completely determine the dynamics in the one-dimensional subspace associated with the first locus. Alternatively, with $1*$ and $**$ one may reconstruct the dynamics of $0*$. However, one may not generically reconstruct the dynamics of any string. Thus, $1*$ and $0*$ give a projected dynamics not the complete one. However, the set 11 , $1*$, $*1$, and $**$ do determine the complete dynamics, even though one is dealing with coarse-grained variables. Below, we will deal almost exclusively with the notion of coarse-graining as expressed in the use of coarse-grained variables as opposed to its use in the determination of a projected dynamics.

$$\begin{aligned}
 E(P_{i'j'}(t+1)) = \sum_{i,j=0,1} M'_{ij'} \left\{ (1 - p_{xo}) P'_{ij} + p_{xo} \left[(p_{11} + p_{33} + p_{22} + p_{44}) P'_{ij} \delta_i^j \right. \right. \\
 + \left((p_{11} + p_{33}) P'_{i\bar{j}} + (p_{22} + p_{44}) P'_{\bar{i}j} \right) \delta_i^j + (p_{12} + p_{34}) P'_{ij} + (p_{21} + p_{43}) P'_{ji}(t) \\
 + (p_{13} + p_{31} + p_{14} + p_{32} + p_{23} + p_{41} + p_{42} + p_{24}) P'_{ii} P'_{jj} \\
 + (p_{23} + p_{41} + p_{13} + p_{31}) P'_{ii} P'_{\bar{j}\bar{j}} + (p_{13} + p_{31} + p_{14} + p_{32}) P'_{\bar{i}\bar{i}} P'_{jj} \\
 + (p_{24} + p_{42} + p_{14} + p_{32}) P'_{ii} P'_{j\bar{j}} + (p_{14} + p_{32}) P'_{\bar{i}\bar{i}} P'_{j\bar{j}} + (p_{13} + p_{31}) P'_{\bar{i}\bar{i}} P'_{j\bar{j}} \\
 \left. \left. + (p_{24} + p_{42} + p_{23} + p_{41}) P'_{ii} P'_{j\bar{j}} + (p_{41} + p_{23}) P'_{\bar{i}\bar{i}} P'_{j\bar{j}} + (p_{42} + p_{24}) P'_{\bar{i}\bar{i}} P'_{j\bar{j}} \right] \right\} \quad (14)
 \end{aligned}$$

ular locus, J_i , of the offspring. Assuming that $m_i = 0$ so that J_i comes from the first parent and $J_i = K_{v_i}$ from the action of $\delta_{I_i}^{K_{v_i}}$, then the contribution from parents K and L after generalized recombination and before mutation is

$$\begin{aligned} & \sum_{K_{v_i}} \sum_{L_{v_i}} \left((1 - m_i) \delta_{J_i}^{K_{v_i}} + m_i \delta_{J_i}^{L_{v_i}} \right) \\ & \quad \times P'_{K_1 \dots K_{v_i} \dots K_\ell} P'_{L_1 \dots L_{v_i} \dots L_\ell} \\ & = \sum_{K_{v_i}} \delta_{J_i}^{K_{v_i}} P'_{K_1 \dots K_{v_i}} \sum_{L_{v_i}} P'_{L_1 \dots L_{v_i} \dots L_\ell} \\ & = P'_{K_1 \dots J_{i v_i} \dots K_\ell} P'_{L_1 \dots *_{v_i} \dots L_\ell} \end{aligned} \quad (15)$$

where $J_{i v_i}$ denotes that the allele corresponding to J_i in the offspring is the same as that in the v_i th locus of the parent and $*_{v_i}$ is the standard wildcard symbol signifying that the allele values at the locus L_{v_i} have been summed over, thus leaving the marginal probability $P'_{L_1 \dots *_{v_i} \dots L_\ell}$. For example, $\sum_{j \in \mathcal{A}} P'_{ij} = P'_{i0} + P'_{i1} + \dots + P'_{i(\mathcal{A}-1)} = P'_{i*}$. One can also think of $\sum_{K_{v_i}} \delta_{J_i}^{K_{v_i}}$ as restricting the sum over K_{v_i} to only one value. As $L_1 \dots *_{v_i} \dots L_\ell$ is clearly a schema we see that, as in the case of homologous recombination, the notion of schemata, and therefore coarse-graining, naturally emerges. In distinction to the homologous case, however, the locus v_i that gives rise to the i th locus of the offspring is not necessarily the i th locus of the first parent.

With this simple example in mind, we can now proceed to the following.

Theorem 1: The expected frequency of a string I at the next generation in a generational GA with arbitrary selection with replacement, mutation, and generalized recombination is

$$\begin{aligned} E(P_I(t+1)) &= \sum_J M_I^J \left((1 - p_{xo}) P'_J(t) \right. \\ & \quad \left. + p_{xo} \sum_{\mathbf{r}} p_c(\mathbf{r}) P'_{J^r}(t) P'_{J^{\bar{r}}}(t) \right) \end{aligned} \quad (16)$$

where $P'_{J^r} = \sum_{K_1} \dots \sum_{K_\ell} \prod_{i=1}^{\ell} \left((1 - m_i) \delta_{J_i}^{K_{v_i}} + m_i \right) P'_{K_1 \dots K_\ell}$ and $P'_{J^{\bar{r}}} = \sum_{L_1} \dots \sum_{L_\ell} \prod_{i=1}^{\ell} \left((1 - m_i) + m_i \delta_{J_i}^{L_{v_i}} \right) P'_{L_1 \dots L_\ell}$ are the selection probabilities for the building block schemata J^r and $J^{\bar{r}}$.

Proof: The only part of the right-hand side that differs from (1) is the string construction term which can be written as

$$\begin{aligned} & \sum_K \sum_L \lambda_J^{KL}(\mathbf{r}) P'_K(t) P'_L(t) \\ & = \sum_{K_1 \dots K_\ell} \sum_{L_1 \dots L_\ell} \prod_{i=1}^{\ell} \left((1 - m_i) \delta_{I_i}^{K_{v_i}} + m_i \delta_{I_i}^{L_{v_i}} \right) \\ & \quad \times P'_{K_1 \dots K_\ell} P'_{L_1 \dots L_\ell}. \end{aligned} \quad (17)$$

As m_i and $(1 - m_i)$ satisfy

$$(1 - m_i)^2 = (1 - m_i) \quad m_i^2 = m_i \quad m_i(1 - m_i) = 0 \quad (18)$$

we may write

$$\begin{aligned} & \prod_{i=1}^{\ell} \left((1 - m_i) \delta_{J_i}^{K_{v_i}} + m_i \delta_{J_i}^{L_{v_i}} \right) \\ & = \prod_{i=1}^{\ell} \left((1 - m_i) \delta_{J_i}^{K_{v_i}} + m_i \right) \\ & \quad \times \prod_{j=1}^{\ell} \left((1 - m_j) + m_j \delta_{J_j}^{L_{v_j}} \right). \end{aligned} \quad (19)$$

Thus, (17) can be written as

$$\begin{aligned} & \sum_{K_1 \dots K_\ell} \prod_{i=1}^{\ell} \left((1 - m_i) \delta_{J_i}^{K_{v_i}} + m_i \right) P'_{K_1 \dots K_\ell} \\ & \quad \times \sum_{L_1 \dots L_\ell} \prod_{j=1}^{\ell} \left((1 - m_j) + m_j \delta_{J_j}^{L_{v_j}} \right) P'_{L_1 \dots L_\ell} \end{aligned} \quad (20)$$

the two terms of which define the building block schemata J^r and $J^{\bar{r}}$ with selection probabilities

$$\begin{aligned} P'_{J^r} &= \sum_{K_1} \dots \sum_{K_\ell} \prod_{i=1}^{\ell} \left((1 - m_i) \delta_{J_i}^{K_{v_i}} + m_i \right) P'_{K_1 \dots K_\ell} \\ P'_{J^{\bar{r}}} &= \sum_{L_1} \dots \sum_{L_\ell} \prod_{i=1}^{\ell} \left((1 - m_i) + m_i \delta_{J_i}^{L_{v_i}} \right) P'_{L_1 \dots L_\ell}. \end{aligned} \quad (21)$$

Hence, substituting into (17), we have

$$\sum_K \sum_L \lambda_J^{KL}(\mathbf{r}) P'_K(t) P'_L(t) = P'_{J^r}(t) P'_{J^{\bar{r}}}(t). \quad (22)$$

Corollary 1: Let N be the population size. For fixed t , in the limit $N \rightarrow \infty$, $E(P_I(t+1)) \rightarrow P_I(t+1)$, the probability to find a genotype I at time $t+1$, and so one obtains a deterministic equation for the evolution of the $P_I(t)$.

Theorem 1 extends previous results obtained for homologous crossover to the case of generalized recombination. It is striking that the functional form of the dynamical equation is identical to that found for homologous crossover [30], [31], even though the range of genetic operators covered by the equation goes far beyond simple homologous crossover. Just as in the homologous case, for each mask the corresponding string sums have disappeared. This means that instead of having to consider $\mathcal{A}^{2\ell}$ terms for every GCM, one only has to consider one, associated with the schemata J^r and $J^{\bar{r}}$, which are just the building blocks for the string J . The key difference between generalized and homologous recombination then does not lie in whether or not building blocks are used, but in the nature of those building blocks, and generalized recombination leads to a much richer set of them.

The differences can be simply illustrated considering $\ell = 2$. To see how one is led to different building blocks, consider the

GRD where only p_{41} and p_{23} are nonzero (for simplicity, we put $p_{x0} = 1$ and $p_m = 0$). From (1)

$$\begin{aligned} E(P_{J_1 J_2}(t+1)) &= p_{41} \sum_{K_1 K_2} \sum_{L_1 L_2} \delta_{J_1}^{L_2} \delta_{J_2}^{K_1} P'_{K_1 K_2}(t) P'_{L_1 L_2}(t) \\ &+ p_{23} \sum_{K_1 K_2} \sum_{L_1 L_2} \delta_{J_1}^{K_2} \delta_{J_2}^{L_1} P'_{K_1 K_2}(t) P'_{L_1 L_2}(t) \\ &= p_{41} P'_{J_2^*}(t) P'_{*J_1}(t) + p_{23} P'_{*J_1}(t) P'_{J_2^*}(t) \end{aligned} \quad (23)$$

where we have taken the opportunity in the first line to also illustrate how the coarse graining that corresponds to this GRD appears, by starting off with the string representation of the equation, rather than the coarse-grained one. The building blocks in this case are J_2^* and $*J_1$, whereas for homologous one-point crossover, where the corresponding GRD is p_{14} and p_{32} and the rest zero, the building blocks would be J_1^* and $*J_2$. Thus, we see how generalized recombination may use building blocks that are not available in homologous crossover.

From (16), as introduced in [27]–[30],⁴ one defines the *effective* fitness, $f_I^{\text{eff}}(t)$, of the string I to be, in the case of proportional selection

$$E(P_I(t+1)) = \frac{f_I^{\text{eff}}(t)}{\bar{f}(t)} P_I(t) \quad (24)$$

where $\bar{f}(t)$ is the average population fitness as defined previously. Thus, in the case at hand, we have

$$\begin{aligned} f_I^{\text{eff}} = \frac{\bar{f}(t)}{P_I(t)} \sum_J M_I^J \left[(1 - p_{x0}) \frac{f_J}{\bar{f}(t)} P_J(t) \right. \\ \left. + p_{x0} \sum_{\mathbf{r}} p_c(\mathbf{r}) \frac{f_{J^{\mathbf{r}}}(t)}{\bar{f}(t)} P_{J^{\mathbf{r}}}(t) \frac{f_{J^{\mathbf{r}}}(t)}{\bar{f}(t)} P_{J^{\mathbf{r}}}(t) \right] \end{aligned} \quad (25)$$

where $f_{J^{\mathbf{r}}}(t) = \sum_{J \in J^{\mathbf{r}}} f_J P_J(t) / \sum_{J \in J^{\mathbf{r}}} P_J(t)$ is the fitness of the schema $J^{\mathbf{r}}$. Strings with $f_I^{\text{eff}}(t) > \bar{f}(t)$ increase in number, while those with $f_I^{\text{eff}}(t) < \bar{f}(t)$ decrease. The effective fitness depends on all the genetic operators not just selection, hence a string I , may increase in frequency relative to another J , even if $f_I < f_J$ if it is more favored by mutation or crossover so that $f_I^{\text{eff}}(t) > f_J^{\text{eff}}(t)$. The latter is governed by the selection weighted linkage disequilibrium coefficient

$$\Delta_J^{\mathbf{r}}(t) = P'_J(t) - P'_{J^{\mathbf{r}}}(t) P'_{J^{\mathbf{r}}}(t). \quad (26)$$

If $\Delta_J^{\mathbf{r}}(t) < 0$ for a given GCM, then generalized recombination increases the proportion of genotype J in the next generation, relative to the proportion one would have with $p_{x0} = 0$; whereas, if $\Delta_J^{\mathbf{r}}(t) > 0$ the proportion is decreased relative to the no recombination limit.

B. Schemata

One lesson that previous work on coarse-grained formulations has taught us is that schemata naturally emerge in any study of homologous recombination. We see from (16) that

⁴This is a generalization of that of [7], [8], [17], and [18], where only the destructive effect of crossover is considered.

this is also true for generalized recombination. The fact that schemata have emerged in the dynamics of strings evolving under the action of generalized recombination is manifest at the level of the string proportions and selection probabilities in (21), which are probabilities associated with hyperplanes of Ω . The form of (16) also tells us that in order to describe the dynamics of strings in this coarse-grained formulation, one must simultaneously understand the dynamics of schemata and, in particular, building block schemata, as the string frequencies at $t+1$ depend on the building block frequencies at t .

For homologous crossover, one of the most remarkable features of the analog of (16) is its form invariance under a further coarse graining [30], [31], i.e., that the functional form of the equations for a schema is identical to that of the equations for the strings themselves. This means that building blocks for a string are composed, in their turn, by other more coarse-grained (lower order—less defining bits/more wildcards) building blocks, which in their turn, etc., the whole hierarchy terminating at 1-schemata, i.e., schemata with only one defining bit and $(\ell-1)$ wildcards, as the latter cannot be composed of more elementary objects. It is precisely the existence of this form of invariance and the hierarchical nature of the relationship between the different building blocks that has led to so many new results using the coarse-grained formulation. We are thus led to consider whether for generalized recombination the same features appear, which can then be further exploited to gain a better theoretical understanding and derive new practical results.

The goal is to coarse grain (16) by passing from an arbitrary string to an arbitrary schema. This coarse graining can be effected in terms of a standard mask, $n = (n_1, \dots, n_\ell)$, as introduced in Section II, and an associated coarse-graining operator, C_I^n , of the form

$$C_I^n = \sum_{I'_1 \dots I'_\ell} \prod_{i=1}^{\ell} \left((1 - n_i) \delta_{I'_i}^{I_i} + n_i \right) \quad (27)$$

which will project out from any vector defined on the configuration space Ω , a coarse-grained vector that naturally lives on a subspace Ω^n of Ω . Further, the operator C_I^n can be written as a tensor product

$$C_I^n = C_{I_1}^{n_1} \otimes C_{I_2}^{n_2} \otimes \dots \otimes C_{I_\ell}^{n_\ell}. \quad (28)$$

As an example, consider

$$\begin{aligned} C_I^{010} P_I &= C_{I_1}^0 C_{I_2}^1 C_{I_3}^0 P_{I'_1 I'_2 I'_3} \\ &= \sum_{I'_1=0,1} \sum_{I'_2=0,1} \sum_{I'_3=0,1} \left(1 \delta_{I_1}^{I'_1} + 0 \right) \\ &\quad \times \left(0 \delta_{I_2}^{I'_2} + 1 \right) \left(1 \delta_{I_3}^{I'_3} + 0 \right) P_{I'_1 I'_2 I'_3} \\ &= P_{I_1^* I_3}. \end{aligned} \quad (29)$$

We denote the bits of the mask n that correspond to zeros, \mathcal{N}_0 , and those that correspond to ones, \mathcal{N}_1 . With this representation in hand, we may state the following theorem, obtained by coarse graining (16) with the operator (27).

Theorem 2: Exact Schema theorem for generalized recombination: The expected frequency of a schema I^n at the next generation in a generational GA of cardinality \mathcal{A} with arbitrary selection with replacement, mutation, and generalized recombination is

$$E(P_{I^n}(t+1)) = \sum_{J^n} M_{I^n}^{J^n} \left((1-p_{xo})P'_{J^n}(t) + p_{xo} \sum_{\mathbf{r}} p_c(\mathbf{r})P'_{J^{rn}}(t)P'_{J^{rn}}(t) \right) \quad (30)$$

where

$P'_{J^{nr}} = \sum_{J_1' \dots J_\ell'} \prod_{i=1}^{\ell} \left((1-m_i) \left((1-n_i)\delta_{J_i'}^{J_i} + n_i \right) + m_i \right) P'_{J_i'}$ and $P'_{J^{nr}} = \sum_{J_1' \dots J_\ell'} \prod_{i=1}^{\ell} \left(m_i \left((1-n_i)\delta_{J_i'}^{J_i} + n_i \right) + (1-m_i) \right) P'_{J_i'}$ are the selection probabilities for the building blocks J^{nr} and J^{nr} of the schema J^n with respect to the GCM $\mathbf{r} = (m, v)$ and $M_{I^n}^{J^n}$ is the mutation probability from the schema J^n to the schema I^n .

Proof: Acting with C_I^n on the left-hand side of (16) one immediately obtains

$$\begin{aligned} C_I^n E(P_{I^n}(t+1)) &= E \left(\sum_{I_1'} \dots \sum_{I_\ell'} \prod_{i=1}^{\ell} \left((1-n_i)\delta_{I_i'}^{I_i} + n_i \right) \right. \\ &\quad \left. \times P_{I_1' \dots I_\ell'}(t+1) \right) \\ &= E(P_{I^n}(t+1)). \end{aligned} \quad (31)$$

We now consider the action of C_I^n on the right-hand side: the mutation operator can be written as a tensor product $\mathbf{M}(\ell) = \bigotimes_{i=1}^{\ell} \mathbf{M}(1)$, where $\mathbf{M}(1)$ is the mutation matrix for the one-allele case. The indicator function $\lambda_J^{KL}(\mathbf{r})$ can also be written

as a tensor product $\lambda_J^{KL}(\mathbf{r}) = \bigotimes_{i=1}^{\ell} \lambda_{J_i}^{K_{v_i} L_{v_i}}(\mathbf{r}_i)$. Hence, for the cloning term on the right-hand side

$$\begin{aligned} C_I^n \mathbf{M}(\ell) &= \bigotimes_{i=1}^{\ell} C_{I_i}^{n_i} \bigotimes_{i=1}^{\ell} \mathbf{M}(1) = \bigotimes_{i=1}^{\ell} (C_{I_i}^{n_i} \mathbf{M}(1)) \\ &= \bigotimes_{i=1}^{\ell} \sum_{J_i} \left((1-n_i)M_{I_i}^{J_i} + n_i \right) \end{aligned} \quad (32)$$

where to arrive at (32) we have used the column stochasticity of the mutation matrix. The expression $\left((1-n_i)M_{I_i}^{J_i} + n_i \right)$ shows clearly that the coarse-grained mutation operator acts *only* on those alleles that correspond to $n_i = 0$. Thus

$$\begin{aligned} C_I^n \sum_J M_I^J P_J'(t) &= \sum_{J_i: i \in \mathcal{N}_0} \bigotimes_{i \in \mathcal{N}_0} M_{I_i}^{J_i} \sum_{J_i: i \in \mathcal{N}_1} P_{J_1 \dots J_\ell}(t) \\ &= \sum_{J^n} M_{I^n}^{J^n} P_{J^n}'(t) \end{aligned} \quad (33)$$

where $P_{J^n}'(t) = \sum_{J_i: i \in \mathcal{N}_1} P_{J_1 \dots J_\ell}(t)$ is the probability to select the schema J^n , the defining bits of which are determined by the set \mathcal{N}_0 , and $M_{I^n}^{J^n} = \bigotimes_{i \in \mathcal{N}_0} M_{I_i}^{J_i}$ is the mutation matrix which acts on the defining characters of the schema J^n . Thus, $M_{I^n}^{J^n}$ is the mutation matrix projected down onto the subspace Ω^n defined by C_I^n .

We now only need consider the final term $B_I = C_I^n \sum_J \sum_K M_I^J \lambda_J^{KL}(m, v) P_K P_L$. In analogy to the mask n , we divide the mask m up into the set where $m_i = 0$, \mathcal{M}_0 , and the set where $m_i = 1$, \mathcal{M}_1 and write

$$\begin{aligned} &\prod_{i=1}^{\ell} \left((1-n_i)\delta_{J_i}^{J_i} + n_i \right) \\ &= \prod_{i \in \mathcal{M}_0} \left((1-n_i)\delta_{J_i}^{J_i} + n_i \right) \\ &\quad \times \prod_{i \in \mathcal{M}_1} \left((1-n_i)\delta_{J_i}^{J_i} + n_i \right). \end{aligned} \quad (34)$$

$$\begin{aligned} B_I &= \sum_{J_i: i \in \mathcal{M}_0} \sum_{K_1 \dots K_\ell} \prod_{i \in \mathcal{M}_0} \left((1-n_i)M_{I_i}^{J_i} + n_i \right) \prod_{i=1}^{\ell} \left((1-m_i)\delta_{J_i}^{K_{v_i}} + m_i \right) P'_{K_1 \dots K_\ell} \\ &\quad \times \sum_{J_i: i \in \mathcal{M}_1} \sum_{L_1 \dots L_\ell} \prod_{i \in \mathcal{M}_1} \left((1-n_i)M_{I_i}^{J_i} + n_i \right) \prod_{i=1}^{\ell} \left((1-m_i) + m_i \delta_{J_i}^{L_{v_i}} \right) P'_{L_1 \dots L_\ell} \end{aligned} \quad (35)$$

$$\begin{aligned} &= \sum_{J_1 \dots J_\ell} \sum_{K_1 \dots K_\ell} \prod_{i=1}^{\ell} \left((1-m_i) \left((1-n_i)M_{I_i}^{J_i} + n_i \right) + m_i \right) \left((1-m_i)\delta_{J_i}^{K_{v_i}} + m_i \right) P'_{K_1 \dots K_\ell} \\ &\quad \times \sum_{J_1 \dots J_\ell} \sum_{L_1 \dots L_\ell} \prod_{i=1}^{\ell} \left(m_i \left((1-n_i)M_{I_i}^{J_i} + n_i \right) + (1-m_i) \right) \left((1-m_i) + m_i \delta_{J_i}^{L_{v_i}} \right) P'_{L_1 \dots L_\ell} \end{aligned} \quad (36)$$

$$= \sum_{J_1 \dots J_\ell} \prod_{i=1}^{\ell} \left((1-m_i) \left((1-n_i)M_{I_i}^{J_i} + n_i \right) + m_i \right) P'_{J_r} \sum_{J_1 \dots J_\ell} \prod_{i=1}^{\ell} \left(m_i \left((1-n_i)M_{I_i}^{J_i} + n_i \right) + (1-m_i) \right) P'_{J_r} \quad (37)$$

$$= \sum_{J^n} M_{I^n}^{J^n} P'_{J^{nr}} P'_{J^{nr}} \quad (38)$$

Then, using (19) we can write B_I as shown in (35)–(38) at the bottom of the previous page, where the selection probabilities of the two building block schemata J^{nr} and $J^{n\bar{r}}$ of the schema J^r , which can mutate into the schema I^n are given by

$$P'_{J^{nr}} = \sum_{J'_1 \dots J'_\ell} \prod_{i=1}^{\ell} \left((1 - m_i) \left((1 - n_i) \delta_{J'_i}^{J_i} + n_i \right) + m_i \right) P'_{J^{nr}} \quad (39)$$

$$P'_{J^{n\bar{r}}} = \sum_{J'_1 \dots J'_\ell} \prod_{i=1}^{\ell} \left(m_i \left((1 - n_i) \delta_{J'_i}^{J_i} + n_i \right) + (1 - m_i) \right) P'_{J^{n\bar{r}}}. \quad (40)$$

■

Corollary 1: Let N be the population size. For fixed t , in the limit $N \rightarrow \infty$, $E(P_{I^n}(t+1)) \rightarrow P_{I^n}(t+1)$ the probability to find the schema I^n at time $t+1$ and so one obtains a deterministic equation for the evolution of the schema frequencies $P_{I^n}(t)$.

To give a bit more insight into the proof: Ω^n is defined, for a schema of order q , to be the \mathcal{A}^q -dimensional subspace of Ω projected out by the action of \mathcal{C}_I^n . For instance, for $\ell = 3$, $\mathcal{A} = 2$, $n = 010$, and $q = 2$, Ω consists of the eight strings $111, \dots, 000$, while Ω^n consists of the four schemata $1*1, 1*0, 0*1$, and $0*0$.

Note that in passing from (36) to (37), the ranges on the products were extended to ℓ by introducing the projection operators $(1 - m_i)$ and m_i . To intuitively see that (37) just gives the building block selection probabilities $P'_{J^{nr}}$ and $P'_{J^{n\bar{r}}}$, consider the case of one locus in particular, i . There are four possibilities for the pair (n_i, m_i) —(0,0), (0,1), (1,0), and (1,1). Then, considering $\sum_{J_1 \dots J_\ell} \prod_{i=1}^{\ell} \left((1 - m_i) \left((1 - n_i) M_{I_i}^{J_i} + n_i \right) + m_i \right) P'_{J^r}$: if $(n_i, m_i) = (0,0)$, then in this case the i th offspring locus came from a mutation of the v_i th locus of the first parent. Similarly, for $(n_i, m_i) = (0,1)$, the i th offspring locus came from the second parent, the v_i th locus of the first parent then being a wildcard symbol $*$. For $(n_i, m_i) = (1,0)$, the i th offspring locus is a $*$ due to the action of n_i on the v_i th transmitted locus from the first parent. Finally, for $(n_i, m_i) = (1,1)$, the v_i th locus of the first parent was already a $*$ and cannot be coarse grained further. Thus, only when $(n_i, m_i) = (0,0)$ does the v_i th locus pass from the first parent via a mutation into the offspring. Similarly, only for $(n_i, m_i) = (0,1)$ does the v_i th locus pass from the second parent via a mutation into the offspring.

Comparing (16) and (30), we see that the functional form of the dynamics for strings and schemata is the same, and hence the dynamics is covariant under a coarse graining. This covariance, or functional “self-similarity,” of the equations can be observed nicely by noting that

$$\begin{aligned} & \sum_{J'_1 \dots J'_\ell} \prod_{i=1}^{\ell} \left((1 - m_i) \left((1 - n_i) \delta_{J'_i}^{J_i} + n_i \right) + m_i \right) \\ & \times \sum_{K_1 \dots K_\ell} \prod_{i=1}^{\ell} \left((1 - m_i) \delta_{J'_i}^{K_{v_i}} + m_i \right) P'_{K_1 \dots K_\ell} \end{aligned}$$

$$= \sum_{K_1 \dots K_\ell} \prod_{i=1}^{\ell} \left((1 - m'_i) \delta_{J'_i}^{K_{v_i}} + m'_i \right) P'_{K_1 \dots K_\ell}$$

where $(1 - m'_i) = (1 - n_i)(1 - m_i)$ and $m'_i = n_i(1 - m_i) + m_i$, and we have once again used the projection operator properties of (18). Note that $(1 - m'_i) = 1$ if and only if $m_i = 0$ and $n_i = 0$, while $m'_i = 1$ if and only if $m_i = 1$ or $n_i = 1$ and $m_i = 0$ which manifests the constraint that one cannot coarse grain a locus that has already been coarse grained. The content of this equation is that the coarse graining onto the schema building block J^{rn} can be achieved in two equivalent ways: one involves a composition of coarse grainings, wherein strings K can be coarse-grained via a recombination pair \mathbf{r} , that involves a mask m , to yield a building block J^r of a string J , which can be further coarse-grained using a mask n , to give a building block J^{nr} of the schema J^n . The second, equivalent, way of achieving the coarse graining is to implement it using a composite mask, m' that involves both m and n and whose elements are as above.

The coarse-graining operators (27) form a semi-group, and are actually just an explicit representation of the renormalization group [2], [32]. To see this: to form a semi-group, they must obey a composition law of the form

$$\mathcal{C}_I^{nm} = \mathcal{C}_I^n \mathcal{C}_I^m \quad (41)$$

and also there must exist a unit element. That the composition law (41) exists follows from the relation:

$$\begin{aligned} & \left(\sum_{I_1 \dots I_\ell} \prod_{i=1}^{\ell} \left((1 - m_i) \delta_{I_i}^{J_i} + m_i \right) \right) \\ & = \left(\sum_{I'_1 \dots I'_\ell} \prod_{i=1}^{\ell} \left((1 - n_i) \delta_{I'_i}^{J_i} + n_i \right) \right) \\ & \times \left(\sum_{I''_1 \dots I''_\ell} \prod_{i=1}^{\ell} \left((1 - m_i) \delta_{I''_i}^{J_i} + m_i \right) \right) \quad (42) \end{aligned}$$

for the projection operators, where, once again, $(1 - m'_i) = (1 - n_i)(1 - m_i)$ and $m'_i = n_i(1 - m_i) + m_i$. The unit element is given by the mask $n = (0, 0, \dots, 0)$. This semi-group at the level of the probabilities P_I could be interpreted as a simple manifestation of the rules of marginalizing probabilities. This is indeed also the case with coarse graining in statistical physics when talking directly about probabilities. However, in the equations that govern the dynamics of these probabilities one must also understand how mutation, recombination and selection transform under a coarse graining. In arriving at theorem (1), we have shown exactly how they transform, that under the coarse graining they are form invariant.

In analogy with (24), one may define the effective fitness of the schema I^n via

$$E(P_{I^n}(t+1)) = \frac{f_{I^n}^{\text{eff}}(t)}{f(t)} P_{I^n}(t) \quad (43)$$

where, again, $\bar{f}(t)$ is the average population fitness. Thus

$$f_{I^n}^{\text{eff}} = \frac{\bar{f}}{P_{I^n}} \sum_{J^n} M_{I^n}^{J^n} \left[(1 - p_{xo}) \frac{f_{J^n}}{\bar{f}} P_{J^n} + p_{xo} \sum_{\mathbf{r}} p_c(\mathbf{r}) \frac{f_{J^n \mathbf{r}}}{\bar{f}} P_{J^n \mathbf{r}} \frac{f_{J^n \mathbf{r}}}{\bar{f}} P_{J^n \mathbf{r}} \right]. \quad (44)$$

Equation (30), formulated in terms of effective fitness, states that: a GA evolving under the action of the operators selection, mutation, and generalized recombination gives an increasing number of trials to *effectively* fit schemata. Equation (30), and indeed, its equivalent (16), both clearly show that the population evolves in an analogous fashion to that with purely homologous crossover—by combining lower order building blocks into higher order ones, which in their turn form yet higher order blocks, etc. The end points of this process are the strings themselves, the highest order objects possible, and one-schemata, the lowest order objects possible.

V. EXAMPLES

To illustrate some of the general features of the coarse-grained equations, we will first consider some general properties of the equations for $\ell = 2$ and $\ell = 3$ for an arbitrary fitness landscape. We will subsequently give a solution to the two-locus case for no selection and an infinite population. As our main interest in this paper is evolution under generalized recombination, we will set the mutation rate $p_m = 0$ in the rest of this paper. For simplicity, we will also in this section set $p_{xo} = 1$.

A. $\ell = 2$

The evolution equations for a generic string of length $\ell = 2$ from (16) are

$$\begin{aligned} E(P_{ij}(t+1)) &= p_{11} P'_{i*} \delta_i^j + p_{12} P'_{ij} + p_{13} P'_{i*} P'_{j*} \\ &+ p_{14} P'_{i*} P'_{*j} + p_{21} P'_{ji} + p_{22} P'_{*i} \delta_i^j \\ &+ p_{23} P'_{*i} P'_{j*} + p_{24} P'_{*i} P'_{*j} + p_{31} P'_{j*} P'_{i*} \\ &+ p_{32} P'_{*j} P'_{i*} + p_{33} P'_{i*} \delta_i^j + p_{34} P'_{ij} \\ &+ p_{41} P'_{j*} P'_{*i} + p_{42} P'_{*j} P'_{*i} \\ &+ p_{43} P'_{ji} + p_{44} P'_{*i} \delta_i^j. \end{aligned} \quad (45)$$

If one replaces the generic, arbitrary alleles i and j with particular values from Ω , the Kronecker deltas are zero or one according to whether one is considering the heterogeneous case $i \neq j$, or the homogeneous one $i = j$. Then, the equations simplify further. For example, if $i = j = 1$ and all GCMs have equal probability ($p_{ij} = 1/16$), we obtain

$$\begin{aligned} E(P_{11}(t+1)) &= \frac{1}{8} \\ &\times \left(P'_{1*}{}^2 + P'_{1*} + 2P'_{1*} P'_{*1} + P'_{*1}{}^2 + 2P'_{11} + P'_{*1} \right). \end{aligned} \quad (46)$$

Notice that in order to solve for the dynamics of the strings here, we need to solve first for the dynamics of the building blocks i^* , $*i$, j^* , and $*j$. Of course, only two of these one-schemata are independent.

As an example, the evolution equation for the schema i^* (a building block for ij) can be simply obtained from (46) by combining the equations for ij and $\bar{i}j$ to find

$$E(P_{i^*}(t+1)) = (p_{1*} + p_{3*}) P'_{i^*} + (p_{2*} + p_{4*}) P'_{*i} \quad (47)$$

where $p_{a*} = \sum_b p_{ab}$ represents a coarse graining of the GCMs themselves. We immediately notice an important difference with the case of homologous crossover, where in (47) $p_{2*} = p_{4*} = 0$, and hence there is no contribution from the schema $*i$, and therefore $E(P_{i^*}(t+1)) = P'_{i^*}(t)$. Thus, we see that the dynamics of the most elementary and fundamental objects under recombination—one-schemata—are quite different in the presence of generalized recombination.

B. $\ell = 3$

Turning now to the case of $\ell = 3$: The general form of the evolution equations for the generic string ijk in terms of building blocks contains 216 terms—a number that, although quite big, is only a tiny fraction of the 13,824 terms one would get in the absence of coarse graining!

Note that the expected frequency of ijk at generation $t+1$ depends not only on its frequency at generation t but is a linear function of the selection probabilities of not only that string but all its permutations, as well as a (generally) quadratic function of the selection probabilities of the lower order schemata (building blocks) that compose it, i.e.,

$$\begin{aligned} E(P_{ijk}(t+1)) &= p_{123} P'_{ijk} + p_{132} P'_{ikj} \\ &+ p_{213} P'_{jik} + p_{231} P'_{kij} + p_{312} P'_{jki} + p_{321} P'_{kji} + b_{ijk}(t). \end{aligned} \quad (48)$$

Again, in order to solve for the string dynamics we need to have the dynamics of the building blocks that determine the driving term $b_{ijk}(t)$. One of the building blocks, for example, is ij^* , the evolution equation of which is found from (48) by considering $P_{ij^*}(t) = \sum_{k=1}^A P_{ijk}(t)$ to find

$$\begin{aligned} E(P_{ij^*}(t+1)) &= p_{11*} P'_{i**} \delta_{ij} + p_{12*} P'_{aj*} + p_{13*} P'_{i*j} \\ &+ p_{14*} P'_{i**} P'_{j**} + p_{15*} P'_{i**} P'_{*j*} \\ &+ p_{16*} P'_{i**} P'_{**j} \\ &+ p_{21*} P'_{j*i} + p_{22*} P'_{*i*} \delta_{ij} \\ &+ p_{23*} P'_{*ij} + p_{24*} P'_{**i} P'_{j**} \\ &+ p_{25*} P'_{**i} P'_{*j*} + p_{26*} P'_{**i} P'_{**j} \\ &+ p_{31*} P'_{j*i} + p_{32*} P'_{*ji} \\ &+ p_{33*} P'_{**i} \delta_{ij} + p_{34*} P'_{**i} P'_{j**} \\ &+ p_{35*} P'_{**i} P'_{*j*} + p_{36*} P'_{**i} P'_{**j} \\ &+ p_{41*} P'_{j**} P'_{i**} + p_{42*} P'_{*j*} P'_{i**} \\ &+ p_{43*} P'_{**j} P'_{i**} + p_{44*} P'_{i**} \delta_{ij} \\ &+ p_{45*} P'_{ij*} + p_{46*} P'_{i*j} \\ &+ p_{51*} P'_{j**} P'_{*i*} + p_{52*} P'_{*j*} P'_{*i*} \\ &+ p_{53*} P'_{**j} P'_{*i*} + p_{54*} P'_{j*i} \\ &+ p_{55*} P'_{**i} \delta_{ij} + p_{56*} P'_{*ij} \\ &+ p_{61*} P'_{j**} P'_{**i} + p_{62*} P'_{*j*} P'_{**i} \\ &+ p_{63*} P'_{**j} P'_{**i} + p_{64*} P'_{*ji} \\ &+ p_{65*} P'_{*ji} + p_{66*} P'_{**i} \delta_{ij} \end{aligned} \quad (49)$$

where we have collected terms involving the same schema and where $p_{ab*} = \sum_c p_{abc}$. Notice that the building block ij^* , in its turn, will depend on the dynamics of its own building blocks, such as i^{**} , the equation for which is found from (49) by considering $P_{i^{**}}(t) = \sum_{j=1}^{\mathcal{A}} P_{ij^*}(t)$

$$E(P_{i^{**}}(t+1)) = (p_{1^{**}} + p_{4^{**}})P'_{i^{**}} + (p_{2^{**}} + p_{5^{**}})P'_{*i^*} + (p_{3^{**}} + p_{6^{**}})P'_{**i}. \quad (50)$$

This hierarchical structure is studied at length in [24] in order to investigate the asymptotic behavior of the dynamics and especially its fixed points. For now though, we turn to the explicit, exact solution of the case $\ell = 2$.

VI. TWO-LOCUS SOLUTION

In both population genetics ([1] and references therein) and EC (see, for example, [6] and [26]) two-locus models have played an important role, leading to improved understanding in the context of potentially analytically solvable models. In this section, we study generalized recombination in the context of a simple two-locus model with no mutation and no selection. Thus, we choose to initially study only the intrinsic biases of the generalized recombination operator \mathcal{G} . As has been seen in previous work, this can lead to practical recipes for practitioners [16].

In this context, one would like to see if and how the dynamics of GAs based on generalized recombination differ from their homologous counterparts. This is a naturally complicated question, given that generalized recombination really captures several different basic operators, including homologous recombination, inversion (and more generally—permutations), different types of gene duplication and combinations of these different operators. Here, we investigate the solutions to (45) in the infinite population limit in the absence of selection, i.e., $P'_{ij}(t) = P_{ij}(t)$, where $P_{ij}(t)$ is a probability, and where, for simplicity, we set $p_{x0} = 1$.

In the infinite population limit $E(P_I(t+1)) \rightarrow P_I(t+1)$ and then (1) and the equations derived from it become deterministic equations for the string and/or schema proportions and describe the corresponding dynamical system. The results derived from the infinite population model neglect the variance inherent in the dynamics due to limited sampling, an effect which is expected to vary as $n^{-1/2}$, where n is the population size. Hence, for large population sizes or short runs, one would expect the analysis below to be an accurate representation of what actually occurs. For small populations, one would have to consider directly the Markov chain for this model. The transition matrix elements that enter in this case would be obtained by using a multinomial distribution with success probabilities given by the right-hand side of (1). Similarly, one would expect our subsequent analysis to give a good qualitative description of the biases engendered by generalized recombination even if selection is included as long as the selection is weak, as will be illustrated later.

A. Building Block Dynamics

In order to solve (45), just as in the case of standard homologous recombination [31], we need to hierarchically solve for the dynamics of the building blocks of the genotype ij . These are: i^* and $*i$. From (45), we can determine the equations for the building block schemata by considering $P_{i^*}(t) = \sum_j P_{ij}(t)$ and $P_{*j}(t) = \sum_i P_{ij}(t)$. The equations for the one-schemata are⁵

$$P_{i^*}(t+1) = (p_{1^*} + p_{3^*})P_{i^*}(t) + (p_{2^*} + p_{4^*})P_{*i}(t) \quad (51)$$

$$P_{*i}(t+1) = (p_{*1} + p_{*3})P_{i^*}(t) + (p_{*2} + p_{*4})P_{*i}(t). \quad (52)$$

Equations (51) and (52) form a coupled linear system, similar to that of mutation in a one-locus system. To solve these equations, we need to determine the eigenvalues and eigenvectors of the matrix

$$\mathbf{W} \equiv \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} (p_{1^*} + p_{3^*}) & (p_{2^*} + p_{4^*}) \\ (p_{*1} + p_{*3}) & (p_{*2} + p_{*4}) \end{pmatrix}. \quad (53)$$

In the case of pure mask-based recombination, only p_{12} , p_{34} , p_{14} , and p_{32} are nonzero. Hence, the evolution of the schema i^* is independent of the schema $*i$, i.e., the two equations decouple, giving as solution $P_{i^*}(t) = P_{i^*}(0)$ and $P_{*i}(t) = P_{*i}(0)$. More generally, noting that, as $\sum_{i=1}^{\mathcal{A}} p_{i^*} = \sum_{i=1}^{\mathcal{A}} p_{*i} = 1$, then $a + b = c + d = 1$, the eigenvalues of matrix (53) are

$$\begin{aligned} \lambda_+ &= 1 \\ \lambda_- &= \frac{1}{2}(a + d - b - c) \\ &= \frac{1}{2}[(p_{14} - p_{41}) + (p_{32} - p_{23}) + (p_{34} - p_{43})] \end{aligned} \quad (54)$$

with corresponding eigenvectors

$$\mathbf{e}_+ = 2^{-1/2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \mathbf{e}_- = (b^2 + c^2)^{-1/2} \begin{pmatrix} b \\ -c \end{pmatrix}. \quad (55)$$

The transformation matrix $\mathbf{\Lambda} \equiv (\mathbf{e}_+, \mathbf{e}_-)^{-1}$, formed from the eigenvectors and (55), diagonalizes \mathbf{W} and rotates the vector $\mathbf{P}(t) = (P_{1^*}, P_{*1})^T \rightarrow (\tilde{P}_+(t), \tilde{P}_-(t))^T$ such that the diagonalized equations can be immediately integrated, then rotated back to the original schema basis to find

$$P_{i^*}(t) = \frac{(cP_{i^*}(0) + bP_{*i}(0))}{(b+c)} + \frac{b(a-c)^t}{b+c} (P_{i^*}(0) - P_{*i}(0)) \quad (56)$$

$$P_{*i}(t) = \frac{(cP_{i^*}(0) + bP_{*i}(0))}{(b+c)} - \frac{c(a-c)^t}{b+c} (P_{i^*}(0) - P_{*i}(0)). \quad (57)$$

⁵Note that i and j are purely symbolic values so that P_{i^*} and P_{*i} are sufficient to cover the $2\mathcal{A}$ possibilities $P_{0^*}, \dots, P_{(\mathcal{A}-1)^*}, P_{*0}, \dots, P_{*(\mathcal{A}-1)}$.

From this solution, we can examine the fixed point. As $|a - c| \leq 1$, except for $a = 1, c = 0$ or $a = 0$ and $c = 1$, the time dependent term vanishes asymptotically, giving as fixed point

$$P_{i^*}^* = \lim_{t \rightarrow \infty} P_{i^*}(t) = \frac{(cP_{i^*}(0) + bP_{*i}(0))}{(b+c)} \quad (58)$$

$$P_{*i}^* = \lim_{t \rightarrow \infty} P_{*i}(t) = \frac{(cP_{i^*}(0) + bP_{*i}(0))}{(b+c)}. \quad (59)$$

Interestingly, in this case, the fixed point is the same for the schemata i^* or $*i$, though this proportion is approached from opposite directions. The behavior of the transients on approaching the fixed point also depends sensitively on the value of $(a - c)$.⁶ If $(a - c) > 0$, then the fixed point is approached monotonically. However, for $(a - c) < 0$, the factor $(-1)^t$ implies the presence of oscillations. However, as $|a - c| \leq 1$ these oscillations are damped and vanish asymptotically. We will now consider some particular cases of interest.

- 1) $b = c = 0$ —in this case, $a = d = 1$, $\lambda_{\pm} = 1$, and $P_{i^*}^* = P_{i^*}(0)$, $P_{*i}^* = P_{*i}(0)$; thus the initial proportions are preserved. This type of recombination is homologous and preserves gene frequencies at a given locus.
- 2) $a = d = 0$ —here, $b = c = 1$, $\lambda_{\pm} = 1, -1$ and the associated eigenvectors are $\mathbf{e}_+ = (1/\sqrt{2})(1, 1)^T$ and $\mathbf{e}_- = (1/\sqrt{2})(1, -1)^T$. There is now no fixed point, but rather a cycle of period two, where $P_{i^*}(t) = P_{i^*}(0)$ for t even and $P_{*i}(0)$ for t odd. Similarly, $P_{*i}(t) = P_{*i}(0)$ for t even and $P_{i^*}(0)$ for t odd. This leads to lateral diffusion of alleles along the string from one genetic locus to another.
- 3) $a = b = c = d = 1/2$ —in this case, $\lambda_{\pm} = 1, 0$ with the same eigenvectors as for case 2). Now, there are no oscillations ($(a - c) = 0$) and the fixed point $P_{i^*}^* = P_{*i}^* = (P_{i^*}(0) + P_{*i}(0))/2$ is reached after only one generation.
- 4) $P_{i^*}(0) = P_{*i}(0)$ —when this condition holds, irrespective of the generalized recombination probabilities, this remains a fixed point. This condition is satisfied both at the center of the simplex [34], as well as at its vertices. It is equivalent to having equal proportions for heterogeneous genotypes.

We have discussed the asymptotic behavior in terms of the four parameters a, b, c , and d . However, we wish to understand the dynamics in terms of the generalized recombination probabilities p_{ab} . For case 1) above, $(p_{1^*} + p_{3^*}) = (p_{*2} + p_{*4}) = 1$ and $(p_{*1} + p_{*3}) = (p_{2^*} + p_{4^*}) = 0$, the latter being equivalent to there being no duplication or inversion or any combination that includes them. This means that there are no genetic operators that lead to lateral diffusion of alleles along the string from one genetic locus to another. The resultant fixed point for the one schemata is on the Robbins/Geiringer manifold [5]. Similarly, for case 2), we have $(p_{1^*} + p_{3^*}) = (p_{*2} + p_{*4}) = 0$ and $(p_{*1} + p_{*3}) = (p_{2^*} + p_{4^*}) = 1$. Under these conditions, the only nonzero terms are those associated with inversion. There is no homologous recombination or duplication. Thus, pure inversion without duplication or homologous crossover leads to periodic behavior.

⁶Note that due to the identities $a + b = 1$ and $c + d = 1$ this is equivalent to $(b - d)$.

We may also investigate the biases of a particular genetic operator, investigating the solutions in the absence of the other operators. Thus, for instance, for duplication from one parent, then $p_{ii} \neq 0$, while all other generalized recombination probabilities are zero. In this case, $a = c = (p_{11} + p_{33})$ and $b = d = (p_{22} + p_{44}) = (1 - (p_{11} + p_{33}))$. Additionally, we have $\sum_{i=1}^4 p_{ii} = 1$, i.e., $b + c = a + d = 1$. Hence, there is no transient term and the fixed point is

$$P_{i^*}^* = P_{*i}^* = P_{i^*}(0) + (p_{22} + p_{44})(P_{*i}(0) - P_{i^*}(0)) \quad (60)$$

which is reached after one generation. For cloning, p_{12} and p_{34} are the only nonzero GCMs, hence, $a = d = 1$ and $b = c = 0$. In this case, the fixed point is trivially

$$P_{i^*}^* = P_{i^*}(0) \quad P_{*i}^* = P_{*i}(0). \quad (61)$$

For inversion, the only nonzero probabilities are p_{21} and p_{43} . In this case, $a = d = 0$ and $b = c = 1$, and the asymptotic behavior is governed by the two cycle of 2) above with

$$P_{i^*}^* = \frac{1}{2}((1 + (-1)^t)P_{i^*}(0) + (1 - (-1)^t)P_{*i}(0)) \quad (62)$$

$$P_{*i}^* = \frac{1}{2}((1 - (-1)^t)P_{i^*}(0) + (1 + (-1)^t)P_{*i}(0)). \quad (63)$$

For two-parent duplication, the appropriate nonzero recombination probabilities are p_{13}, p_{24}, p_{31} , and p_{42} . Hence, $a = c = (p_{13} + p_{31})$ and $b = d = (p_{24} + p_{42})$ with $b + c = a + d = 1$. As $a = c$ there are no transients and the fixed point

$$P_{i^*}^* = P_{*i}^* = P_{i^*}(0) + (p_{24} + p_{42})(P_{*i}(0) - P_{i^*}(0)) \quad (64)$$

is reached after one generation just as in the case of one-parent duplication. Note that this fixed point is of the same form as that found for one parent duplication. For homologous crossover, the nonzero entries in the recombination distribution are p_{14} and p_{32} . In this case, $a = d = 1$, $b = c = 0$ and the fixed point is the same as that for cloning. Finally, for crossover and inversion the corresponding recombination distribution entries are p_{41} and p_{23} which implies $a = d = 0$ and $b = c = 1$. In this case, the fixed point is the same as that for inversion above.

In Fig. 1, using (56), we see a graph of the evolution of the one-schema 1^* for different GRDs. The direct integration of (51) and (52) yields exactly the same curves as expected. The initial condition used is an asymmetric one, where $P_{11}(0) = P_{00}(0) = 0.1$, $P_{01}(0) = 0.6$, and $P_{10}(0) = 0.2$; hence, $P_{1^*}(0) = 0.3$. The fixed point behavior described in points 1)–4) and equations (60)–(64) is clearly visible. For one-parent duplication, the fixed point is reached after one generation at a value $P_{1^*}^* = P_{1^*}(0) + P_{*1}(0)/2 = 0.3 + 0.7 = 0.5$. For inversion, one sees the characteristic oscillations between the values 0.3 and 0.7 associated with $P_{1^*}(0)$ and $P_{*1}(0)$. For homologous crossover, the fixed point is the initial proportion $P_{1^*}(0) = 0.3$, i.e., the allele frequency at a given locus is preserved. Finally, considering all GCMs with equal probability—the All curve in Fig. 1—one sees that the system reaches a fixed point in one generation.

The general features we have just delineated for $\ell = 2$, are, as we will see in Section VII, also present for $\ell > 2$, and represent

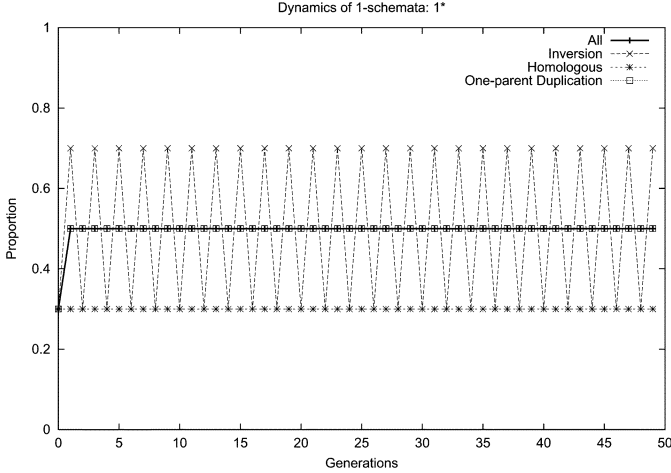


Fig. 1. Dynamical evolution of the one-schema 1* for different GRDs and no selection.

qualitatively new phenomena when compared with the normal homologous forms of crossover with which we are familiar. The lateral diffusion of alleles, relative to the homologous case, leads to a fixed point where for a given offspring locus, the allele frequency at that locus depends not only on the allele frequency at the same parental loci but also on the allele frequencies at other genetic loci. For $\ell > 2$, instead of a pair of linear coupled equations for the one-schemata, one has ℓ coupled equations whose solution can be found by solving the corresponding eigensystem.

B. Solution for Strings

With the solutions for the one-schemata in hand, we can now proceed to determine the solutions for the strings themselves from (45) by substituting in the solutions (56) and (57), which are the contributions from the building blocks to yield

$$P_{ij}(t+1) = (1 - p_{xo})P_{ij}(t) + p_{xo}(p_{12} + p_{34})P_{ij}(t) + p_{xo}(p_{21} + p_{43})P_{ji}(t) + p_{xo}F_{ij}(t) \quad (65)$$

where

$$F_{ij}(t) = (C_{ij} + D_{ij}(a - c)^t + E_{ij}(a - c)^{2t}) \quad (66)$$

and the matrices C_{ij} , D_{ij} , and E_{ij} are given by

$$C_{ij} = (p_{14} + p_{41} + p_{32} + p_{23} + p_{13} + p_{31} + p_{24} + p_{42})A_iA_j + (p_{11} + p_{22} + p_{33} + p_{44})A_i\delta_i^j \quad (67)$$

$$D_{ij} = \left[\begin{aligned} &((p_{14} + p_{32})(bA_jB_i - cB_jA_i) \\ &+ (p_{23} + p_{41})(bA_iB_j - cB_iA_j) \\ &+ (p_{13} + p_{31})b(A_jB_i - B_jA_i) \\ &- (p_{24} + p_{42})c(A_jB_i - B_jA_i) \\ &+ ((p_{11} + p_{33})b - (p_{22} + p_{44})c)B_i\delta_i^j \end{aligned} \right] \quad (68)$$

$$E_{ij} = ((p_{13} + p_{31})b^2 + (p_{24} + p_{42})c^2 - (p_{14} + p_{41} + p_{32} + p_{23})bc)B_iB_j \quad (69)$$

where

$$A_i = \frac{(cP_{i*}(0) + bP_{*i}(0))}{(b + c)} \quad B_i = \frac{(P_{i*}(0) - P_{*i}(0))}{(b + c)}. \quad (70)$$

To solve (65), we need to also have the equation for P_{ji} , which is just (65) with $i \leftrightarrow j$. The matrices C_{ij} and E_{ij} are both symmetric matrices, D_{ij} however is not. Hence, $P_{ji}(t)$ satisfies

$$P_{ji}(t+1) = (1 - p_{xo})P_{ji} + p_{xo}(p_{12} + p_{34})P_{ji}(t) + p_{xo}(p_{21} + p_{43})P_{ij}(t) + p_{xo}F_{ji}(t) \quad (71)$$

where

$$F_{ji}(t) = (C_{ij} + D_{ji}(a - c)^t + E_{ij}(a - c)^{2t}). \quad (72)$$

Equations (65) and (71) are linear coupled inhomogeneous first-order difference equations and can be solved as with (51) and (52), by determining the corresponding eigensystem. Putting $p_{xo} = 1$, the relevant matrix is

$$\mathbf{W}' = \begin{pmatrix} (p_{12} + p_{34}) & (p_{21} + p_{43}) \\ (p_{21} + p_{43}) & (p_{12} + p_{34}) \end{pmatrix} \quad (73)$$

whose eigenvalues and eigenvectors are given by $\lambda_{\pm} = (p_{12} + p_{34}) \pm (p_{21} + p_{43})$; $\mathbf{e}_+ = 2^{-1/2}(1 \ 1)^T$ and $\mathbf{e}_- = 2^{-1/2}(1 \ -1)^T$. In the eigenvector basis, $\mathbf{P}(t) = (P_{ij}, P_{ji})^T \rightarrow (\tilde{P}_+(t), \tilde{P}_-(t))^T$ such that

$$\tilde{P}_+(t+1) = \lambda_+ \tilde{P}_+(t) + \tilde{F}_+(t) \quad (74)$$

$$\tilde{P}_-(t+1) = \lambda_- \tilde{P}_-(t) + \tilde{F}_-(t) \quad (75)$$

where

$$\tilde{F}_+(t) = \frac{1}{2^{1/2}}(F_{ij}(t) + F_{ji}(t))$$

$$\tilde{F}_-(t) = \frac{1}{2^{1/2}}(F_{ij}(t) - F_{ji}(t)) \quad (76)$$

which can be immediately integrated to yield

$$\tilde{P}_{\pm}(t) = \lambda_{\pm}^t \tilde{P}_{\pm}(0) + \sum_{n=0}^{t-1} \lambda_{\pm}^{t-n-1} \tilde{F}_{\pm}(n). \quad (77)$$

Rotating back to the original basis one finds

$$P_{ij}(t) = \frac{1}{2}(\lambda_+^t + \lambda_-^t)P_{ij}(0) + \frac{1}{2}(\lambda_+^t - \lambda_-^t)P_{ji}(0) + \frac{1}{2} \sum_{n=0}^{t-1} (\lambda_+^{t-n-1}(F_{ij}(n) + F_{ji}(n)) + \lambda_-^{t-n-1}(F_{ij}(n) - F_{ji}(n))). \quad (78)$$

There now only remains to do the summations to obtain the final answer

$$\begin{aligned}
P_{ij}(t) = & \frac{1}{2}(\lambda_+^t + \lambda_-^t) P_{ij}(0) + \frac{1}{2}(\lambda_+^t - \lambda_-^t) P_{ji}(0) \\
& + C_{ij} \left(\frac{1 - \lambda_+^t}{1 - \lambda_+} \right) + E_{ij} \left(\frac{(a-c)^{2t} - \lambda_+^t}{(a-c)^2 - \lambda_+} \right) \\
& + \frac{1}{2}(D_{ij} + D_{ji}) \left(\frac{(a-c)^t - \lambda_+^t}{a-c - \lambda_+} \right) \\
& + \frac{1}{2}(D_{ij} - D_{ji}) \left(\frac{(a-c)^t - \lambda_-^t}{a-c - \lambda_-} \right). \quad (79)
\end{aligned}$$

Note how this solution has been created—hierarchically, as in the case of homologous crossover [31]. One can solve first for the order one building blocks, which then serve as a “source” for construction of order 2 building blocks, which serve as a source for the order 3, etc., until one arrives at the strings themselves. The difference here is that inversion can couple different building blocks of the same order, unlike the homologous case where they are decoupled.

In the limit $t \rightarrow \infty$, in the case where the cloning or inversion probabilities are less than one, the fixed point of (79) is

$$P_{ij}^* = \lim_{t \rightarrow \infty} P_{ij}(t) = \frac{C_{ij}}{(1 - \lambda_+)}. \quad (80)$$

Explicitly, in terms of the GRD

$$\begin{aligned}
P_{ij}^* = & \left[\frac{(p_{14} + p_{41} + p_{32} + p_{23} + p_{13} + p_{31} + p_{24} + p_{42})}{(1 - p_{12} - p_{34} - p_{21} - p_{43})} \right. \\
& \times \frac{((p_{*1} + p_{*3})P_{i*}(0) + (p_{2*} + p_{4*})P_{*i}(0))}{((p_{*1} + p_{*3}) + (p_{2*} + p_{4*}))} \\
& \times \left. \frac{((p_{*1} + p_{*3})P_{*j}(0) + (p_{2*} + p_{4*})P_{j*}(0))}{((p_{*1} + p_{*3}) + (p_{2*} + p_{4*}))} \right] \\
& + \left[\frac{(p_{11} + p_{22} + p_{33} + p_{44})}{(1 - p_{12} - p_{34} - p_{21} - p_{43})} \right. \\
& \times \left. \frac{((p_{*1} + p_{*3})P_{i*}(0) + (p_{2*} + p_{4*})P_{*i}(0))}{((p_{*1} + p_{*3}) + (p_{2*} + p_{4*}))} \delta_i^j \right].
\end{aligned}$$

To get some intuition about the nature of this fixed point, we will consider some interesting limits associated with different initial populations and different recombination probability distributions. Beginning with a random initial population, where $P_{ij}(0) = 1/4$, the fixed point becomes

$$\begin{aligned}
P_{ij}^* = & \frac{(p_{14} + p_{41} + p_{32} + p_{23} + p_{13} + p_{31} + p_{24} + p_{42})}{4(1 - p_{12} - p_{34} - p_{21} - p_{43})} \\
& + \frac{(p_{11} + p_{22} + p_{33} + p_{44})}{2(1 - p_{12} - p_{34} - p_{21} - p_{43})} \delta_i^j. \quad (81)
\end{aligned}$$

Only in the case where there is no one-parent gene duplication, i.e., $p_{ii} = 0$, is the center of the simplex a fixed point. In the presence of one-parent gene duplication, homogeneous strings

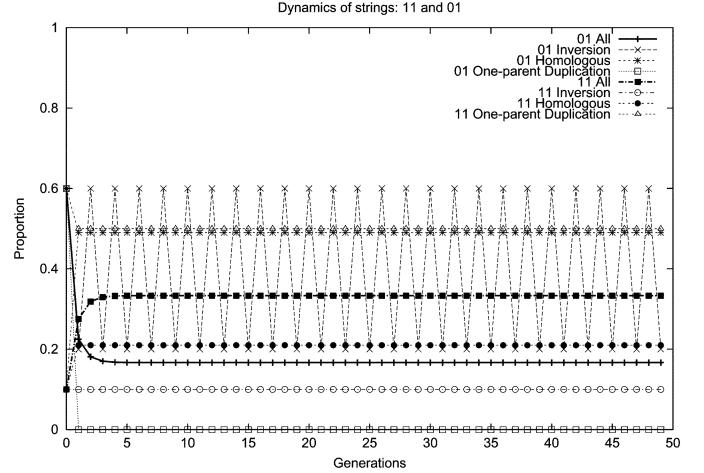


Fig. 2. Dynamical evolution of the strings 11 and 01 for different GRDs and no selection.

are favored over their heterogeneous counterparts. For instance, for GCMs with equal probabilities of 1/16, the asymptotic proportions of homogeneous and heterogeneous strings are 1/3 and 1/6, respectively. Thus, homogeneous strings have higher *effective* fitness [31] than heterogeneous ones.

In Fig. 2, we see a graph of the solution (79) for the strings 11 and 01 for the same asymmetric initial conditions used for Fig. 1, and for the same GCMs. Notice the presence of four different fixed points (two-cycle in the case of inversion) for each string type. This is a much richer behavior than in the case of simple homologous crossover, where the unique Geiringer limit $P_{ij}^* = P_{i*}(0)P_{*j}(0)$ holds. The Geiringer limits for 11 and 10, with the previously stated initial conditions are $P_{11}^* = 0.3 \times 0.7 = 0.21$ and $P_{01}^* = 0.7 \times 0.7 = 0.49$, both of which agree with the asymptotic limits seen in Fig. 2. For one-parent duplication, the expected fixed points for 11 and 10 are from (81) $(0.7 + 0.3)/2$ and 0, respectively, once again in agreement with the graph and showing the higher effective fitness associated with homogeneous strings. Note also the oscillations present in the heterogeneous string 01 and their corresponding absence in the homogeneous string 11.

VII. EFFECTS OF SELECTION AND $\ell > 2$

One is prompted to wonder whether the phenomena seen in the previous section are peculiarities of the two-locus case or are more robust, with analogues for $\ell > 2$ and when selection is present. We begin with selection for $\ell = 2$. Of course, it is not feasible to determine what happens for an arbitrary landscape and so we restrict ourselves to some generic observations. We first consider in Fig. 3 results found using a “Schemulator,”⁷ a Java-based simulator that integrates the exact coarse-grained equations for any ℓ and any fitness landscape. We first consider the effects for a nonepistatic unication landscape with fitnesses $f_{11} = 12$, $f_{10} = f_{01} = 11$, and $f_{00} = 10$. What is clearly seen is a similar bias as found from the no selection case of Fig. 2, but superimposed on a selection dominated “trend.” The oscillations for inversion only are clearly visible. However, here, as only heterogeneous strings can oscillate, and as the optimal

⁷The program is available from the authors.

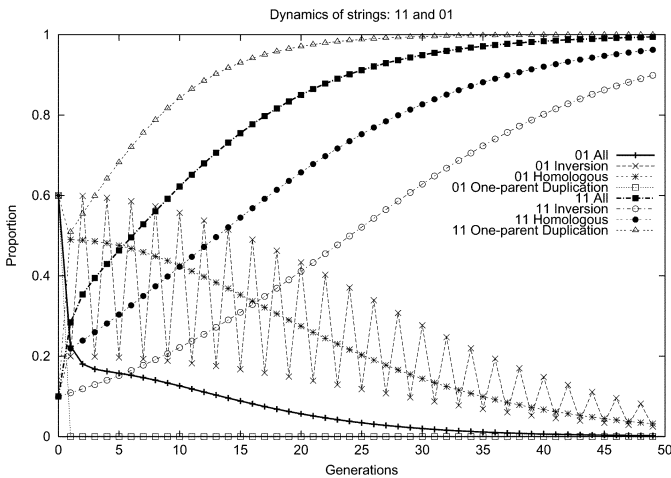


Fig. 3. Dynamical evolution of the strings 11 and 01 for different GRDs in a one max landscape.

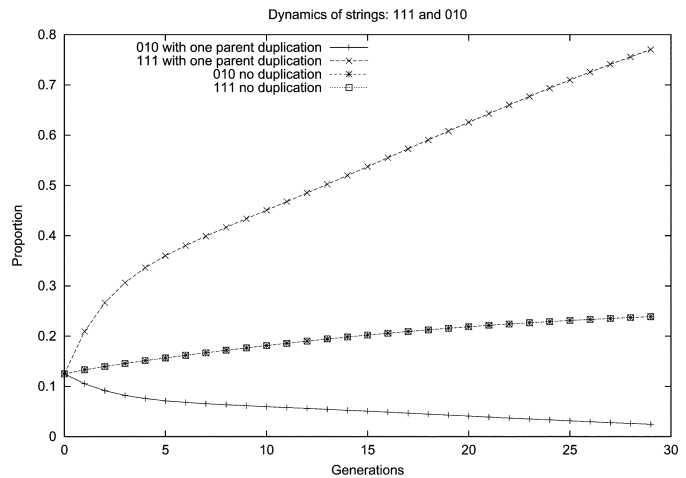


Fig. 5. Dynamical evolution of the strings 111 and 010 with uniform recombination and comparing one-parent duplication and no duplication on the problem with degenerate genotype-phenotype map.

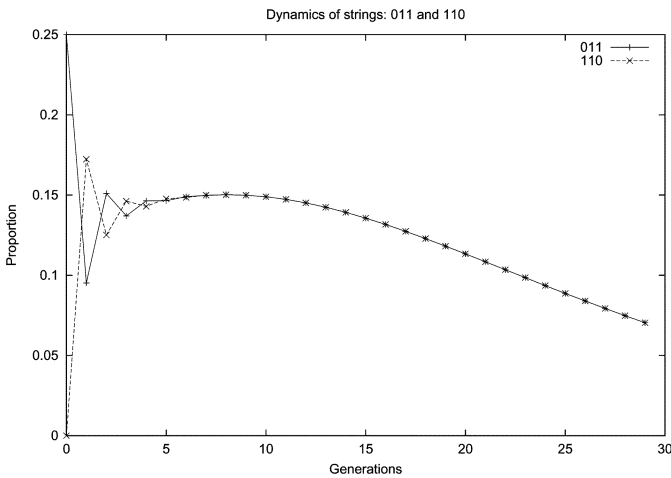


Fig. 4. Dynamical evolution of the strings 011 and 110 with inversion and uniform recombination.

string is homogeneous, the oscillations diminish in amplitude. For one-parent duplication, the proportion of 01 strings vanishes after one generation, as in the no selection case. For 01 with all GCMs present, also as in the no selection case, there is a sharp initial decrease. In distinction to that case though, in the presence of selection, the proportion continues to diminish, but at a much reduced rate. As the selection pressure diminishes, the curves of Fig. 3 will imitate those of Fig. 2 ever more closely, while for increasing selection pressure the phenomena due to inversion and duplication will be less and less noticeable. We see then that, although we have only exactly solved the no selection case, observed phenomena such as lateral allele diffusion, oscillations, preference for homogenous strings, etc., are also present with selection.

Turning our attention now to the case $\ell = 3$, the results are shown in Fig. 4, where the fitness landscape here is now such that $f_{111} = 13$, while the fitness of the Hamming distance one neighbors 110, 101, and 011 is 12, that of the Hamming distance two neighbors 100, 010, and 001 is 11 and $f_{000} = 10$. We restrict ourselves to uniform homologous recombination with the corresponding $p_c(m, (1, 2, 3)) = 0.05$ and with full inversion

implemented with $p_c(000, (3, 2, 1)) = p_c(111, (3, 2, 1)) = 0.3$. We also start with the initial condition $P_{110}(0) = 0, P_{011}(0) = 0.25$ with all other frequencies being 0.125. Note the characteristic oscillations that inversion can induce superimposed with a selective trend just as in the case of $\ell = 2$. The extra feature we wish to point out here, is the fact that the addition of inversion has added a qualitatively new feature to the search properties of the algorithm. With the chosen initial conditions and only homologous crossover, it would be impossible to generate the string 110. This will also be true of building blocks as well as strings, as is evident in the observation that the one-schemata equations are coupled for any ℓ . Thus, there exists the possibility of generating important building blocks that are not currently present in the search by lateral allele diffusion, induced, for example, by inversion. For example, the building block $1****$ could be generated via a GCM $[11111, (6, 5, 4, 3, 2, 1)]$ acting on the building block $*****1$. As a final illustration, we consider in Fig. 5 the effect of one-parent duplication for $\ell = 3$. The landscape we consider here is associated with a degenerate genotype-phenotype map wherein there are two optimal genotypes—111 and 010 with fitness 13. The Hamming distance one neighbors of *either* of these strings have fitness 12 and the Hamming distance two neighbors 001 and 100 have fitness 11. The chosen initial population is random. The most notable feature here is the “symmetry breaking” of the genotype-phenotype map due to the presence of duplication, the homogeneous optimum being more effectively fit than its heterogeneous counterpart. In this sense, even after optima are found, the genetic operators are continuing the evolutionary search seeking those optima that are the most evolutionarily robust—in this case, the homogeneous optimum.

So, although for $\ell > 2$ and/or including selection, exact solutions are not available, we can see that the principal derived predictions from the two-locus no-selection case—oscillations, homogeneous/heterogeneous asymmetry, lateral allele diffusion, etc.—can all be observed in the more general case. This has been explicitly checked by integrating the equations for both $\ell = 3$ and $\ell = 4$ using the Schemulator. In general, all these phenomena occur simultaneously, for instance, for $\ell = 4$, one

might have inversion restricted to the first two loci, leading to oscillations there, while restricting duplication to the last two loci, and having a preference for homogeneous alleles there. Due to the richness of the potential behaviors that emerge from generalized recombination, to make the dynamics more transparent we have chosen to illustrate them individually.

VIII. CONCLUSION

The main results of this paper are twofold: first, the introduction and theoretical analysis of a generalized notion of exchange of genetic material—generalized recombination—which extends and subsumes many currently used genetic operators, such as homologous recombination, inversion, and duplication; and second, the demonstration that this more general EA is most naturally treated in a coarse-grained formulation, wherein the natural dynamical effective degrees of freedom are building block schemata not strings.

We showed that generalized recombination requires an extension of the notion of crossover mask and recombination distribution to that of a TCM and GRD. GCMs could be explicitly represented in different ways—through a recombination vector, a crossover matrix, or a recombination pair. With these representations in hand, an exact string evolution equation was derived for both variable-length and fixed-length representations, including mutation. It was then shown that the dynamics was written much more naturally in terms of building block schemata, that emerge by the actions of projection operators that are the natural representation of the generalized recombination operator and which implement coarse grainings. It was then shown that the resulting string equation, written in terms of building block schemata, under a coarse graining, yielded a functionally identical equation for schemata, thus leading to a new exact Schema theorem for generalized recombination. The coarse graining projection operators were shown to exhibit a semi-group structure, thus giving an explicit realization of the renormalization group.

Given that homologous recombination, inversion, and duplication have all been found to be useful by practitioners, we do not need to justify the utility of generalized recombination, though it does remain to be seen to what extent the extra diversity, above and beyond the standard operators when considered individually, can help in evolutionary search. Having an exact theoretical framework also allows for a better understanding of how the different genetic operators work, as has been exhibited not only in the context of the exact two-locus solution, but also in the results of the Schemulator, where several interesting phenomena were observed. Among these were the appearance of oscillations in the frequencies of strings and schemata in the presence of permutations, higher effective fitness for homogeneous versus heterogeneous strings and schemata in the presence of duplication, and the lateral diffusion of alleles for any nonhomologous operator, the latter allowing for a more “mutation”-like effect, where an allele that did not originally exist at a particular locus can be generated by transferring it laterally from some other locus. It should be emphasized that, although we have studied these phenomena in the context of an infinite population model, they are in fact robust—appearing also in the finite size context, though naturally, the intrinsic extra “noise” due to

finite population effects can make their identification more difficult. In terms of analytical results for $\ell > 2$, although exact solutions are, of course, very difficult to obtain, much can be said about the asymptotic behavior. This and other analytical results are discussed in [24].

ACKNOWLEDGMENT

The authors thank the referees for comments on the manuscript.

REFERENCES

- [1] R. Bürger, *The Mathematical Theory of Selection, Recombination, and Mutation*. Chichester, U.K.: Wiley, 2000.
- [2] C. R. Stephens, C. Chryssomalakos, and A. Zamora, “Coarse graining in genetic dynamics: A renormalization group analysis of a simple genetic system,” *Rev. Mex. Fis.*, vol. 50, pp. 388–396, 2004.
- [3] C. Chryssomalakos and C. R. Stephens, “What basis for genetic dynamics?,” in *Proc. GECCO*, K. Deb, Ed., 2004, pp. 1018–1029.
- [4] A. G. Clark, “Invasion and maintenance of a gene duplication,” *Proc. Nat. Acad. Sci.*, vol. 91, pp. 2950–2954, 1994.
- [5] H. Geiringer, “On the probability theory of linkage in Mendelian heredity,” *Ann. Math. Stat.*, vol. 15, no. 1, pp. 25–57, Mar. 1944.
- [6] D. E. Goldberg, “Simple genetic algorithms and the minimal deceptive problem,” in *Genetic Algorithms and Simulated Annealing*, L. Davis, Ed. London, U.K.: Pitman, 1987, pp. 74–88.
- [7] —, “Genetic algorithms and Walsh functions: Part I. A gentle introduction,” *Complex Syst.*, vol. 3, pp. 123–152, 1989.
- [8] —, “Genetic algorithms and Walsh functions: Part II. Deception and its analysis,” *Complex Syst.*, vol. 3, pp. 153–171, 1989.
- [9] —, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley, 1989.
- [10] N. Goldenfeld, *Phase Transitions and the Renormalization Group*. Reading, MA: Addison-Wesley, 1992.
- [11] N. Goldenfeld, A. McKane, and Q. Hou, “Block spins for partial differential equations,” *J. Stat. Phys.*, vol. 93, pp. 699–714, 1998.
- [12] J. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: Univ. Michigan Press, 1975.
- [13] L. Kadanoff, *Statistical Physics*. Singapore: World Scientific, 2000.
- [14] J. R. Koza, “Gene duplication to enable genetic programming to concurrently evolve both the architecture and work-performing steps of a computer program,” in *Proc. IJCAI*, Montreal, Canada, 1995, vol. 1, pp. 734–740.
- [15] B. Lewin, *Genes VIII*. Englewood Cliffs, NJ: Prentice-Hall, 2003.
- [16] N. F. McPhee and R. Poli, “A schema theory analysis of the evolution of size in genetic programming with linear representations,” in *Proc. Euro Genetic Program.*, Milan, Italy, Apr. 2001, LNCS, pp. 18–20.
- [17] P. Nordin and W. Banzhaf, L. Eshelman, Ed., “Complexity compression and evolution,” in *Proc. 6th Int. Conf. Genetic Algorithms*, Pittsburgh, PA, 1995, pp. 310–317.
- [18] P. Nordin, F. Francone, and W. Banzhaf, “Explicitly defined introns and destructive crossover in genetic programming,” in *Proc. Workshop Genetic Programming: From Theory to Real World Applications*, J. P. Rosca, Ed., Tahoe City, CA, 1995, pp. 6–22.
- [19] R. Poli, “Exact schema theory for genetic programming and variable-length genetic algorithms with one-point crossover,” *Genetic Program. Evol. Mach.*, vol. 2, no. 2, pp. 123–163, 2001.
- [20] R. Poli, N. F. McPhee, and J. E. Rowe, “Exact schema theory and Markov chain models for genetic programming and variable-length genetic algorithms with homologous crossover,” *GPEH*, vol. 5, pp. 31–70, 2004.
- [21] R. Poli and N. F. McPhee, “General schema theory for genetic programming with subtree-swapping crossover: Part I,” *Evol. Comput.*, vol. 11, no. 1, pp. 53–66, 2003.
- [22] R. Poli, J. E. Rowe, and N. F. McPhee, “Markov chain models for GP and variable-length GAs with homologous crossover,” in *Proc. Genetic Evol. Comput. Conf.*, San Francisco, CA, Jul. 2001, pp. 7–11.
- [23] R. Poli and C. R. Stephens, “Theoretical analysis of generalized recombination,” in *Proc. CEC*, B. McKay, Ed., 2005, pp. 411–418.
- [24] —, “Theoretical analysis of generalized recombination,” Dept. Comput. Sci., Univ. Essex, Essex, U.K., Tech. Rep. CSM-426, 2005.
- [25] H. Sawai and S. Adachi, “A comparative study of gene-duplicated GAs based on pfGA and SSGA,” in *Proc. GECCO*, 2000, pp. 74–81.

- [26] W. M. Spears, "The equilibrium and transient behavior of mutation and recombination," in *Proc. FOGA 6*, W. Spears and W. Martin, Eds., San Francisco, CA, 2001, pp. 74–88.
- [27] P. F. Stadler and C. R. Stephens, "Landscapes and effective fitness," *Commun. Theor. Biol.*, vol. 8, pp. 389–431, 2003.
- [28] C. R. Stephens, "Effective fitness landscapes for evolutionary systems," in *Proc. Congr. Evol. Comput.*, P. J. Angeline, Z. Michalewicz, M. Schoenauer, X. Yao, and A. Zalzala, Eds., Washington, D.C., Jul. 6–9, 1999, vol. 1, pp. 703–714.
- [29] C. R. Stephens and J. M. Vargas, "Effective fitness as an alternative paradigm for evolutionary computation I: General formalism," *Genetic Program. Evol. Mach.*, vol. 1, no. 4, pp. 363–378, Oct. 2000.
- [30] C. R. Stephens and H. Waelbroeck, "Effective degrees of freedom in genetic algorithms and the block hypothesis," in *Proc. 7th Int. Conf. Genetic Algorithms*, T. Bäck, Ed., East Lansing, 1997, pp. 34–40.
- [31] —, "Schemata evolution and building blocks," *Evol. Comput.*, vol. 7, no. 2, pp. 109–124, 1999.
- [32] C. R. Stephens, "The renormalization group and the dynamics of genetic systems," *Acta Phys. Slov.*, vol. 52, pp. 515–524, 2002.
- [33] C. R. Stephens and R. Poli, "Coarse graining in an evolutionary algorithm with recombination, duplication and inversion," in *Proc. CEC*, B. McKay, Ed., 2005, pp. 1683–1691.
- [34] M. D. Vose, *The Simple Genetic Algorithm: Foundations and Theory*. Cambridge, MA: MIT Press, 1999.
- [35] W. B. Langdon and R. Poli, *Foundations of Genetic Programming*. Berlin, Germany: Springer-Verlag, 2002.



Chris Stephens received the undergraduate degree at Queen's College, Oxford, U.K., and the Ph.D. degree in quantum and statistical field theory from the University of Maryland, College Park.

He is a Professor at the Institute for Nuclear Sciences of the Universidad Nacional Autónoma de México (UNAM)—the oldest university in the Americas. He then had several postdoctoral positions, including Imperial College, London, and the University of Utrecht, where he worked with Gerard 't Hooft, the 1999 Nobel Laureate in Physics, before moving to Mexico City. He has had visiting positions at various leading academic institutions, including the Weizmann Institute, the Joint Institute for Nuclear Research, Dubna, Dublin Institute for Advanced Studies, the University of Birmingham, and the University of Essex. He is also a founding partner of Adaptive Technologies, Inc., and Adaptive Technologies SA de CV—research companies dedicated to the production of agent-based technolo-

gies for dynamical optimization in finance and industry. His research interests are very broad, having published over 80 refereed research articles with over 1000 citations in a wide array of international journals—ranging from *Classical and Quantum Gravity* to the *Journal of Molecular Evolution*. An overriding theme in the vast majority of the work, however, has been the Renormalization Group—a general methodology for solving complex, nonlinear problems with many degrees of freedom via coarse graining—and, more recently, applying it to the area of genetic dynamics.

Dr. Stephens received the Jorge Lomnitz Prize of the Mexican Academy of Sciences and a Leverhulme Professorship from the Leverhulme Trust, U.K. He is a member of the Editorial Board of Genetic Programming and Evolvable Hardware and has been invited to present plenary talks or tutorials at major international conferences such as EUROGP and GECCO.



Riccardo Poli is a Professor in the Department of Computer Science at the University of Essex. He has coauthored *Foundations of Genetic Programming* (Springer-Verlag, 2002) with W. B. Langdon. He has published over 180 refereed papers on evolutionary algorithms (particularly genetic programming), neural networks, and image/signal processing. His main research interests include genetic programming (GP) and the theory of evolutionary algorithms.

Prof. Poli was elected a Fellow of the International Society for Genetic and Evolutionary Computation (ISGEC), in recognition of sustained and significant contributions to the field and the community, in July 2003. He has been cofounder and Co-Chair of EuroGP, the European Conference on Genetic Programming for 1998, 1999, 2000, and 2003. He was the Chair of the GP theme at the Genetic and Evolutionary Computation Conference (GECCO) 2002 (the largest conference in the field) and was Co-Chair of the prestigious Foundations of Genetic Algorithms (FOGA) Workshop in 2002. He has been (the first non-U.S.) General Chair of GECCO in 2004, and served as a member of the business committee for GECCO 2005. He is Technical Chair of the International Workshop on Ant Colony Optimization and Swarm Intelligence (ANTS 2006) and Competition Chair for GECCO 2006. He is an Associate Editor of *Evolutionary Computation* (MIT Press), *Genetic Programming and Evolvable Machines* (Springer), and the *International Journal of Computational Intelligence Research* (IJCIR). He has been program committee member of over 50 international events. He has presented invited tutorials on GP at ten international conferences. He is a member of the EPSRC Peer Review College and has attracted, as Principal Investigator or Co-Investigator, funding for over \$1.8M from EPSRC, DERA, Leverhulme Trust, Royal Society, and others.