Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Deterministic and stochastic aspects of single-crossover recombination

Ellen Baake

Abstract.
This contribution is concerned with mathematical models for the dynamics of the genetic composition of populations evolving under recombination. Recombination is the genetic mechanism by which two parent individuals create the mixed type of their offspring during sexual reproduction. The corresponding models are large, nonlinear dynamical systems (for the deterministic treatment that applies in the infinite-population limit), or interacting particle systems (for the stochastic treatment required for finite populations). We review recent progress on these difficult problems. In particular, we present a closed solution of the deterministic continuous-time system, for the important special case of single crossovers; we extract an underlying linearity; we analyse how this carries over to the corresponding stochastic setting; and we provide a solution of the analogous deterministic discrete-time dynamics, in terms of its generalised eigenvalues and a simple recursion for the corresponding coefficients.

Mathematics Subject Classification (2000). Primary 92D10, 34L30; Secondary 37N25, 06A07, 60J25.

Keywords. Population genetics, recombination dynamics, Möbius linearisation and diagonalisation, correlation functions, Moran model.

1. Introduction

Biological evolution is a complex phenomenon driven by various processes, such as muation and recombination of genetic material, reproduction of individuals, and selection of favourable types. The area of *population genetics* is concerned with how these processes shape and change the genetic structure of populations. *Mathematical population genetics* was founded in the 1920's by Ronald Fisher, Sewall Wright, and John Haldane, and thus is among the oldest areas of mathematical biology. The reason for its continuing (and actually increasing) attractiveness for both mathematicians and biologists is at least twofold: Firstly, there is a true need for mathematical models and methods, since the outcome of evolution is impossible to predict (and, thus, today's genetic data are impossible to analyse) without their help. Second, the processes of genetics lend themselves most naturally to a mathematical formulation and give rise to a wealth of fascinating new problems, concepts, and methods.

This contribution will focus on the phenomenon of *recombination*, in which two parent individuals are involved in creating the mixed type of their offspring during sexual reproduction. The essence of this process is illustrated in Fig. 1 and may be idealised and summarised as follows.

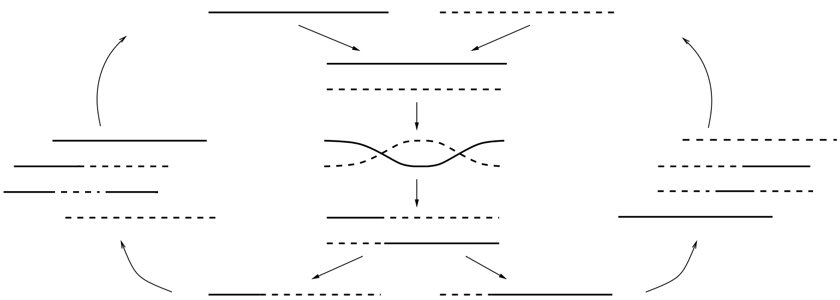

Figure 1. Life cycle of a population under sexual reproduction and recombination. Each line symbolises a sequence of sites that defines a gamete (like the two at the top that start the cycle as 'egg' and 'sperm'). The pool of gametes at the left and the right comes from a large population of recombining individuals. These sequences meet randomly to start the next round of the cycle.

Genetic information is encoded in terms of sequences of finite length. Eggs and sperm (i.e., female and male germ cells or *gametes*) each carry a single copy of such a sequence. They go through the following life cycle: At fertilisation, two gametes meet randomly and unite, thus starting the life of a new individual, which is equipped with both the maternal and the paternal sequence. At maturity, this individual will generate its own germ cells. This process includes recombination, that is, the maternal and paternal sequences perform one or more *crossovers* and are cut and relinked accordingly, so that two 'mixed' sequences emerge. These are the new gametes and start the next round of fertilisation (by random mating within a large population).

Models of this process aim at describing the dynamics of the genetic composition of a population that goes through this life cycle repeatedly. These models come in various flavours: in discrete or continous time; with various assumptions about the crossover pattern; and, most importantly, in a deterministic or a stochastic formulation, depending on whether the population is assumed to be so large that stochastic fluctuations may be neglected. In any case, however, the resulting process appears difficult to treat, due to the large number of possible states and the nonlinearity generated by the random mixture of gametes. Nevertheless, a number of solution procedures have been discovered for the deterministic discrete-time setting [8, 10, 14], and the underlying mathematical structures were investigated within the framework of genetic algebras, see [18, 19, 22]. Quite generally, the solution relies on a certain nonlinear transformation (known as Haldane linearisation) from (gamete or type) frequencies to suitable correlation functions, which decouple from each other and decay geometrically. But if sequences of more than three sites are involved, this transformation must be constructed via recursions that involve the parameters of the recombination process, and is not available explicitly except in the trivial case of independent sites. For a review of the area, see [9, Ch. V.4].

In this contribution, we concentrate on a special case that is both biologically and mathematically relevant, namely, the situation in which at most one crossover happens at any given time. That is, only recombination events may occur that partition the sites of a sequence into two parts that correspond to the sites before and after a given crossover point. We analyse the resulting models in continuous time (both deterministic and stochastic), as well as in discrete time. For the deterministic continuous-time system (Section 2), a simple explicit solution can be given. This simplicity is due to some underlying linearity; actually, the system may even be diagonalised (via a *nonlinear* transformation). In Section 3, we consider the corresponding stochastic process (still in continuous time), namely, the Moran model with recombination. This also takes into account the resampling effect that comes about via random reproduction in a finite population. In particular, we investigate the relationship between the expectation of the Moran model and the solution of the deterministic continuous-time model. We finally tackle deterministic single-crossover dynamics in *discrete* time (Section 4). This setting implies additional dependencies, which become particularly transparent when the so-called *ancestral recombination process* is considered. A solution may still be given, but its coefficients must be determined recursively.

Altogether, it will turn out that the corresponding models, and their analysis, have various mathematical facettes that are intertwined with each other, such as differential equations, probability theory, and combinatorics.

2. Deterministic dynamics, continuous time

2.1. The model. We describe populations at the level of their gametes and thus identify gametes with individuals. Their genetic information is encoded in terms of a linear arrangement of sites, indexed by the set $S := \{0, 1, \dots, n\}$. For each site $i \in S$, there is a set X_i of 'letters' that may possibly occur at that site. To allow for a convenient notation, we restrict ourselves to the simple but important case of *finite* sets X_i; for the full generality of arbitrary locally compact spaces X_i, the reader is referred to [3] and [5].

A *type* is thus defined as a sequence

$$x = (x_0, x_1, \dots, x_n) \in X_0 \times X_1 \times \cdots \times X_n =: X,$$

where X is called the *type space*. By construction, x_i is the i-th coordinate of x, and we define $x_I := (x_i)_{i \in I}$ as the collection of coordinates with indices in I, where I is a subset of S. A *population* is identified with a non-negative measure ω on X. Namely, $\omega(\{x\})$ denotes the frequency of individuals of type $x \in X$ and $\omega(A) := \sum_{x \in A} \omega(\{x\})$ for $A \subseteq X$; we abbreviate $\omega(\{x\})$ as $\omega(x)$. The set of all nonnegative measures on X is denoted by $\mathcal{M}_{\geqslant 0}(X)$. If we define δ_x as the point measure on x (i.e., $\delta_x(y) = \delta_{x,y}$ for $x, y \in X$), we can also write $\omega = \sum_{x \in X} \omega(x) \delta_x$.

We may, alternatively, interpret δ_x as the basis vector of $\mathbb{R}_{\geq 0}^{|X|}$ that corresponds to x (where a suitable ordering of types is implied, and $|X|$ is the number of elements in X); ω is thus identified with a vector in $\mathbb{R}_{\geq 0}^{|X|}$.

At this stage, frequencies need not be normalised; $\omega(x)$ may simply be thought of as the size of the subpopulation of type x, measured in units so large that it may be considered a continuous quantity. The corresponding normalised version $p := \omega/\|\omega\|$ (where $\|\omega\| := \sum_{x\in X} \omega(x) = \omega(X)$ is the total population size) is then a probability distribution on X, and may be identified with a probability vector.

Recombination acts on the links between the sites; the links are collected into the set $L := \left\{\frac{1}{2}, \frac{3}{2}, \dots, \frac{2n-1}{2}\right\}$. We shall use Latin indices for the sites and Greek indices for the links, and the implicit rule will always be that $\alpha = \frac{2i+1}{2}$ is the link between sites i and $i+1$; see Figure 2.

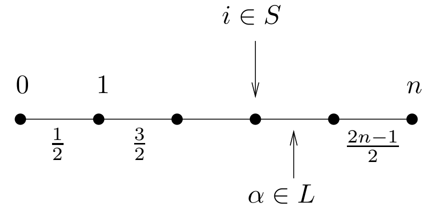

Figure 2. Sites and links.

Let recombination happen in every individual, and at every link $\alpha \in L$, at rate $\varrho_\alpha > 0$. More precisely, for every $\alpha \in L$, every individual exchanges, at rate $\varrho_\alpha/2$, the sites after link α with those of a randomly chosen partner. Explicitly, if the 'active' and the partner individual are of types x and y, then the new pair has types $(x_0, x_1, \dots, x_{\lfloor\alpha\rfloor}, y_{\lceil\alpha\rceil}, \dots, y_n)$ and $(y_0, y_1, \dots, y_{\lfloor\alpha\rfloor}, x_{\lceil\alpha\rceil}, \dots, x_n)$, where $\lfloor\alpha\rfloor(\lceil\alpha\rceil)$ is the largest integer below α (the smallest above α); see Fig. 3. Since every individual can occur as either the 'active' individual or as its randomly chosen partner, we have a total rate of ϱ_α for crossovers at link α. For later use, we also define $\varrho := \sum_{\alpha\in L} \varrho_\alpha$.

In order to formulate the corresponding model, let us introduce the projection operators π_i, $i \in S$, via

$$\begin{array}{rlcl} \pi_i: & X_0 \times X_1 \times \cdots \times X_n & \longrightarrow & X_i \\ & (x_0, x_1, \dots, x_n) & \mapsto & x_i, \end{array} \tag{1}$$

i.e., π_i is the canonical projection to the i-th coordinate. Likewise, for any index set $I \subseteq S$, one defines a projector

$$\begin{array}{rlcl} \pi_I: & X & \longrightarrow & \bigtimes_{i\in I} X_i =: X_I \\ & (x_0, x_1, \dots, x_n) & \mapsto & (x_i)_{i\in I} =: x_I. \end{array}$$

We shall frequently use the abbreviations $\pi_{<\alpha} := \pi_{\{1,\dots,\lfloor\alpha\rfloor\}}$ and $\pi_{>\alpha} := \pi_{\{\lceil\alpha\rceil,\dots,n\}}$, as well as $x_{<\alpha} := \pi_{<\alpha}(x)$, $x_{>\alpha} := \pi_{>\alpha}(x)$. The projectors $\pi_{<\alpha}$ and $\pi_{>\alpha}$ may be

$$\begin{array}{ccc} x_0, \dots, x_n & \overset{\varrho_\alpha}{\rightleftarrows} & x_0, \dots, x_{\lfloor \alpha \rfloor}, y_{\lceil \alpha \rceil}, \dots, y_n \\ y_0, \dots, y_n & & y_0, \dots, y_{\lfloor \alpha \rfloor}, x_{\lceil \alpha \rceil}, \dots, x_n \end{array}$$

$$\begin{array}{ccc} x_0, \dots, x_n & \overset{\varrho_\alpha}{\rightleftarrows} & x_0, \dots, x_{\lfloor \alpha \rfloor}, *\,, \;\dots\;, * \\ *\,, \dots, * & & *\,, \;\dots\;, *\,, x_{\lceil \alpha \rceil}, \dots, x_n \end{array}$$

Figure 3. Upper panel: Recombination between individuals of type x and y. Lower panel: The corresponding 'marginalised' version that summarises all events by which individuals of type x are gained or lost (a '$*$' at site i stands for an arbitrary element of X_i). Note that, in either case, the process can go both ways, as indicated by the arrows.

thought of as *cut and forget* operators because they take the leading or trailing segment of a sequence x, and forget about the rest.

Whereas the π_I act on the types, we also need the induced mapping at the level of the population, namely,

$$\begin{array}{rccl} \pi_I . : & \mathcal{M}_{\geqslant 0} & \longrightarrow & \mathcal{M}_{\geqslant 0} \\ & \omega & \mapsto & \omega \circ \pi_I^{-1} =: \pi_I . \omega, \end{array} \tag{2}$$

where π_I^{-1} denotes the preimage under π_I. The operation . (where the dot is on the line) is the 'pullback' of π_I w.r.t. ω; so, $\pi_I . \omega$ is the marginal distribution of ω with respect to the sites in I. In particular, $(\pi_{<\alpha} . \omega)(x_{<\alpha})$ is the marginal frequency of sequences prescribed at the sites before α, and vice versa for the sites after α.

Now, single-crossover recombination (at the level of the population) means the relinking of a randomly chosen leading segment with a randomly chosen trailing segment. We therefore introduce (elementary) recombination operators (or *recombinators*, for short), $R_\alpha : \mathcal{M}_{\geqslant 0} \to \mathcal{M}_{\geqslant 0}$ for $\alpha \in L$, defined by

$$R_\alpha(\omega) := \frac{1}{\|\omega\|} \big((\pi_{<\alpha} . \omega) \otimes (\pi_{>\alpha} . \omega) \big). \tag{3}$$

Here, the tensor product reflects the independent combination (i.e., the product measure) of the two marginals $\pi_{<\alpha} . \omega$ and $\pi_{>\alpha} . \omega$. R_α is therefore a *cut and relink* operator. $R_\alpha(\omega)$ may be understood as the population that emerges if *all* individuals of the population ω disintegrate into their leading and trailing segments and these are relinked randomly. Note that $\|R_\alpha(\omega)\| = \|\omega\|$.

The recombination dynamics may thus be compactly written as

$$\dot{\omega}_t = \sum_{\alpha \in L} \varrho_\alpha \big(R_\alpha(\omega_t) - \omega_t \big) = \sum_{\alpha \in L} \varrho_\alpha (R_\alpha - \mathbb{1})(\omega_t) =: \varPhi(\omega_t), \tag{4}$$

where $\mathbb{1}$ is the identity operator. Note that (4) is a large, nonlinear system of ordinary differential equations (ODEs).

2.2. Solution of the ODE system. The solution of (4) relies on some elementary properties of our recombinators. Most importantly, they are idempotents and commute with each other, i.e.,

$$\begin{aligned} R_\alpha^2 &= R_\alpha, \qquad \alpha \in L, && (5)\\ R_\alpha R_\beta &= R_\beta R_\alpha, \quad \alpha, \beta \in L. && (6) \end{aligned}$$

These properties are intuitively plausible: if the links before α are already independent of those after α due to a previous recombination event, then further recombination at that link does not change the situation; and if a product measure is formed with respect to two links α and β, the result does not depend on the order in which the links are affected. For the proof, we refer to [5, Prop. 2]; let us only mention here that it relies on the elementary fact that, for $\omega \in \mathcal{M}_{\geqslant 0}$,

$$\begin{aligned} \pi_{<\alpha}.\big(R_\beta(\omega)\big) &= \pi_{<\alpha}.\omega, \qquad \text{for } \beta \geq \alpha, \text{ and}\\ \pi_{>\alpha}.\big(R_\beta(\omega)\big) &= \pi_{>\alpha}.\omega, \qquad \text{for } \beta \leq \alpha; \end{aligned}$$

that is, recombination at or after α does not affect the marginal frequencies at sites before α, and vice versa.

We now define *composite* recombinators as

$$R_G := \prod_{\alpha \in G} R_\alpha \qquad \text{for } G \subseteq L.$$

Here, the product is to be read as composition; it is, indeed, a product if the recombinators are written in terms of their multilinear matrix representations, which is available in the case of finite types considered here (see [2]). By property (6), the order in the composition plays no role. Furthermore, (5) and (6) obviously entail $R_G R_H = R_{G\cup H}$ for $G, H \subseteq L$.

With this in hand, we can now state an explicit solution of our problem, namely,

Theorem 1. *The solution of the single-crossover dynamics* (4) *with initial value* ω_0 *can be given in closed form as*

$$\omega_t = \sum_{G \subseteq L} a_G(t) R_G(\omega_0) =: \varphi_t(\omega_0) \tag{7}$$

with coefficient functions

$$a_G(t) = \prod_{\alpha \in L\setminus G} e^{-\varrho_\alpha t} \prod_{\beta \in G} (1 - e^{-\varrho_\beta t});$$

i.e., φ_t *is the* semigroup *belonging to the recombination equation* (4). □

For the proof, the reader is referred to [5, Thm. 2] or [3, Thm. 3] (the former article contains the original, the latter a shorter and more elegant version of the proof). Let us note that the coefficient functions can be interpreted probabilistically. Given an individual sequence in the population, $a_G(t)$ is the probability that the set of links that have seen at least one crossover event until time t is precisely the set G (obviously, $\sum_{G \subseteq L} a_G(t) = 1$). Note that the product structure of the $a_G(t)$ implies independence of links, a decisive feature of the single-crossover dynamics in continuous time, as we shall see later on. Note also that, as $t \to \infty$, ω_t converges to the stationary state

$$\omega_\infty = \frac{1}{\|\omega_0\|^{n-1}} \bigotimes_{i=1}^{n} (\pi_i . \omega_0), \tag{8}$$

in which all sites are independent.

2.3. Underlying linearity. The simplicity of the solution in Theorem 1 comes as a certain surprise. After all, explicit solutions to large, nonlinear ODE systems are rare – they are usually available for linear systems at best. For this reason, the recombination equation and its solution have already been taken up in the framework of functional analysis, where they have led to an extension of potential theory [21]. We will now show that there is an underlying linear structure that is hidden behind the solution. It can be stated as follows, compare [5, Sec. 3.2] for details.

Theorem 2. *Let* $\bigl\{c_{G'}^{(L')}(t) \mid \varnothing \subseteq G' \subseteq L' \subseteq L\bigr\}$ *be a family of non-negative functions with* $c_G^{(L)}(t) = c_{G_1}^{(L_1)}(t)\, c_{G_2}^{(L_2)}(t)$*, valid for any partition* $L = L_1 \dot\cup L_2$ *of the set* L *and all* $t \geq 0$*, where* $G_i := G \cap L_i$*. Assume further that these functions satisfy* $\sum_{H \subseteq L'} c_H^{(L')}(t) = 1$ *for any* $L' \subseteq L$ *and* $t \geq 0$*. If* $v \in \mathcal{M}_{\geqslant 0}(X)$ *and* $H \subseteq L$*, one has the identity*

$$R_H\Bigl(\sum_{G \subseteq L} c_G^{(L)}(t) R_G(v)\Bigr) = \sum_{G \subseteq L} c_G^{(L)}(t) R_{G \cup H}(v),$$

which is then satisfied for all $t \geq 0$. □

Here, the upper index specifies the respective set of links. Clearly, the coefficient functions $a_G(t)$ of Theorem 1 satisfy the conditions of Theorem 2. The result then means that the recombinators act linearly along the solutions (7) of the recombination equation (4). Theorem 2 thus has the consequence that, on $\mathcal{M}_{\geqslant 0}(X)$, the forward flow of (4) commutes with all recombinators, that is, $R_G \circ \varphi_t = \varphi_t \circ R_G$ for all $t \geq 0$ and all $G \subseteq L$.

But let us go one step further here. The conventional approach to solve the recombination dynamics consists in transforming the type frequencies to certain functions (known as principal components) that diagonalise the dynamics, see [8, 10, 18] and references therein for more. We will now show that, in continuous time, they have a particularly simple structure: they are given by certain correlation functions, known as *linkage disequilibria* (LDE) in biology, which play an important role in applications. They have a counterpart at the level of operators (on $\mathcal{M}_{\geqslant 0}(X)$). Namely, let us define *LDE operators* via

$$T_G := \sum_{H \supseteq G} (-1)^{|H \setminus G|} R_H, \qquad G \subseteq L, \tag{9}$$

where the underdot indicates the summation variable. Note that T_G maps $\mathcal{M}_{\geqslant 0}(X)$ into $\mathcal{M}(X)$, the set of *signed* measures on X. Eq. (9) leads to the inverse $R_G = \sum_{\underset{\cdot}{H} \supseteq G} T_H$ by the combinatorial Möbius inversion formula, see [1, Thm. 4.18]. We then have

Theorem 3. *If ω_t is the solution* (7)*, the transformed quantities $T_G(\omega_t)$ satisfy*

$$\frac{d}{dt} T_G(\omega_t) = -\Big(\sum_{\alpha \in L \setminus G} \varrho_\alpha \Big) T_G(\omega_t), \qquad G \subseteq L. \tag{10}$$

Proof. See [5, Sec. 3.3]. □

Obviously, Eq. (10) is a system of *decoupled, linear, homogeneous* differential equations with the usual exponential solution. Note that this simple form emerged through the *nonlinear* transform (9) as applied to the solution of the *coupled, nonlinear* differential equation (4).

Suitable components of the signed measure $T_G(\omega_t)$ may then be identified to work with in practice (see [5, 6] for details); they correspond to correlation functions of all orders and decouple and decay exponentially. These functions turn out to be particularly well-adapted to the problem since they rely on ordered partitions, in contrast to conventional LDE's used elsewhere in population genetics, which rely on general partitions (see [9, Ch. V.4] for review).

3. Stochastic dynamics, continuous time

3.1. The model. The effect of finite population size in population genetics is, in continuous time, well captured by the *Moran model*. It describes a population of fixed size N and takes into account the stochastic fluctuations due to random reproduction, which manifest themselves via a *resampling effect* (known as genetic drift in biology). More precisely, the finite-population counterpart of our deterministic model is the Moran model with single-crossover recombination. To simplify matters (and in order to clearly dissect the individual effects of recombination and resampling), we shall use the decoupled (or parallel) version of the model, which assumes that resampling and recombination occur independently of each other, as illustrated in Fig. 4. More precisely, in our finite population of fixed size N, every individual experiences, independently of the others,

- resampling at rate $b/2$. The individual reproduces, the offspring inherits the parent's type and replaces a randomly chosen individual (possibly its own parent).

- recombination at (overall) rate ϱ_α at link $\alpha \in L$. Every individual picks a random partner (maybe itself) at rate $\varrho_\alpha/2$, and the pair exchanges the sites after link α. That is, if the recombining individuals have types x and y, they are replaced by the two offspring individuals $(x_{<\alpha}, y_{>\alpha})$ and $(y_{<\alpha}, x_{>\alpha})$, as in the deterministic case, and Fig. 3. As before, the per-capita rate of recombination at link α is then ϱ_α, because both orderings of the individuals lead to the same type count in the population.

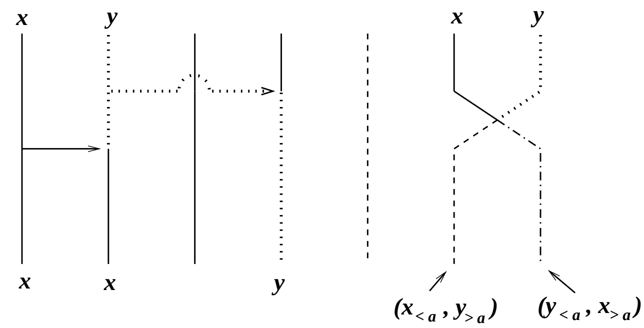

Figure 4. Graphical representation of the Moran model (with parallel resampling and recombination). Every individual is represented by a vertical line; time runs down the page. Resampling is indicated by arrows, with the parent individual at the tail and the offspring at the tip. Recombination is depicted by a crossing between two individuals. Note that the spatial information suggested by the graphical representation does not play a role in the model; one is only interested in the frequencies of the various types.

Note that the randomly chosen second individual (for resampling or recombination) may be the active individual itself; then, effectively, nothing happens. One might, for biological reasons, prefer to exclude these events by sampling from the remaining population only; but this means nothing but a change of time scale of order $1/N$.

To formalise this verbal description of the process, let the state of the population at time t be given by the collection (the random vector)

$$Z_t = \big(Z_t(x)\big)_{x \in X} \in E := \Big\{ z \in \{0,1,...,N\}^{|X|} \,\Big|\, \sum_x z(x) = N \Big\},$$

where $Z_t(x)$ is the number of individuals of type x at time t; clearly, $\sum_{x\in X} Z_t(x) = N$. We also use Z_t in the sense of a (random counting) measure, in analogy with ω_t (but keep in mind that Z_t is integer-valued and counts single individuals, whereas ω_t denotes continuous frequencies in an infinite population). The letter z will be used to denote realisations of Z_t — but note that the symbols x, y, and z are not on equal footing (x and y will continue to be types). The stochastic process $\{Z_t\}_{t\geq 0}$ is the continuous-time Markov chain on E defined as follows. If the current state is $Z_t = z$, two types of transitions may occur:

$$\begin{aligned}
\text{resampling:} \quad & z \to z + s(x,y), \quad s(x,y) := \delta_x - \delta_y, \\
& \text{at rate } \frac{1}{2N} b z(x) z(y) \ \text{ for } (x,y) \in X \times X \qquad (11) \\
\text{recombination:} \quad & z \to z + r(x,y,\alpha), \\
& r(x,y,\alpha) := \delta_{(x_{<\alpha}, y_{>\alpha})} + \delta_{(y_{<\alpha}, x_{>\alpha})} - \delta_x - \delta_y, \\
& \text{at rate } \frac{1}{2N} \varrho_\alpha z(x) z(y) \ \text{ for } (x,y) \in X \times X, \alpha \in L \qquad (12)
\end{aligned}$$

(where δ_x is the point measure on x, as before). Note that, in (11) and (12), transitions that leave E are automatically excluded by the fact that the corresponding rates vanish. On the other hand, 'empty transitions' ($s(x,y) = 0$ or $r(x,y,\alpha) = 0$) are explicitly included (they occur if $x = y$ in resampling or recombination, and if $x_{<\alpha} = y_{<\alpha}$ or $x_{>\alpha} = y_{>\alpha}$ in recombination).

3.2. Connecting stochastic and deterministic models. Let us now explore the connection between the stochastic process $\{Z_t\}_{t\geq 0}$ on E, its normalised version $\{\widehat{Z}_t\}_{t\geq 0} = \{Z_t\}_{t\geq 0}/N$ on E/N, and the solution $\omega_t = \varphi_t(\omega_0)$ (Eq. (7)) of the differential equation. It is easy to see (and no surprise) that

$$\frac{d}{dt}\mathbb{E}(Z_t) = \mathbb{E}\big(\Phi(Z_t)\big), \qquad (13)$$

with Φ of (4). But this does not, *per se*, lead to a 'closed' differential equation for $\mathbb{E}(Z_t)$, because it is not clear whether $\mathbb{E}(\Phi(Z_t))$ can be written as a function of $\mathbb{E}(Z_t)$ alone—after all, Φ is nonlinear. *In the absence of resampling,* however, we have

Theorem 4. *Let $\{Z_t\}_{t\geq 0}$ be the recombination process without resampling (i.e., $b = 0$), and let Z_0 be fixed. Then, $\mathbb{E}(Z_t)$ satisfies the differential equation*

$$\frac{d}{dt}\mathbb{E}(Z_t) = \Phi\big(\mathbb{E}(Z_t)\big)$$

with initial value Z_0, and Φ from (4); therefore,

$$\mathbb{E}(Z_t) = \varphi_t(Z_0), \quad \textit{for all } t \geq 0,$$

with φ_t from (7). Likewise, for all $t \geq 0$,

$$\mathbb{E}(T_G Z_t) = T_G\big(\varphi_t(Z_0)\big).$$

Proof. See [6, Thm. 1 and Cor. 1]. □

The result again points to some underlying linearity, which, in the context of the *stochastic* model, should be connected to some kind of independence. Indeed, the key to the proof of Thm. 4 is a lemma concerning the independence of marginal processes. For $I \subseteq S$, we introduce the 'stretch' of I as

$$J(I) := \{i \in S \mid \min(I) \leq i \leq \max(I)\},$$

and look at the projection of the recombination process on non-overlapping stretches. This is the content of

Lemma 5. *Let $\{Z_t\}_{t\geq 0}$ be the recombination process without resampling (i.e., $b = 0$). Let $A, B \subseteq S$ with $J(A) \cap J(B) = \varnothing$. Then, $\{\pi_A.Z_t\}_{t\geq 0}$ and $\{\pi_B.Z_t\}_{t\geq 0}$ are conditionally (on Z_0) independent Markov chains on E_A and E_B.*

Proof. See [6, Lemma 1]. □

Let us now re-include resampling, at rate $b/2 > 0$, and consider the stochastic process $\{Z_t^{(N)}\}_{t\geq 0}$ defined by both (11) and (12), where we add the upper index here to indicate the dependence on N. Now, Lemma 5 and Thm. 4 are no longer valid. The processes $\{\pi_{<\alpha}.Z_t^{(N)}\}_{t\geq 0}$ and $\{\pi_{>\alpha}.Z_t^{(N)}\}_{t\geq 0}$ are still individually Markov, but their resampling events are coupled (replacement of $y_{<\alpha}$ by $x_{<\alpha}$ is always tied to replacement of $y_{>\alpha}$ by $x_{>\alpha}$). Hence the marginal processes fail to be independent, so that no equivalent of Lemma 5 holds.

Let us, therefore, change focus and consider the normalised version $\{\widehat{Z}_t^{(N)}\}_{t\geq 0} = \{Z_t^{(N)}\}_{t\geq 0}/N$. In line with general folklore in population genetics, in the limit $N \to \infty$, the relative frequencies $\{\widehat{Z}_t^{(N)}\}_{t\geq 0}$ cease to fluctuate and are then given by the solution of the corresponding deterministic equation. More precisely, we have

Proposition 6. *Consider the family of processes $\{\widehat{Z}_t^{(N)}\}_{t\geq 0} = \frac{1}{N}\{Z_t^{(N)}\}_{t\geq 0}$, $N = 1, 2, \ldots$, where $\{Z_t^{(N)}\}_{t\geq 0}$ is defined by (11) and (12). Assume that the initial states are chosen so that $\lim_{N\to\infty} \widehat{Z}_0^{(N)} = p_0$. Then, for every given $t \geq 0$, one has*

$$\lim_{N\to\infty} \sup_{s\leq t} |\widehat{Z}_s^{(N)} - p_s| = 0 \tag{14}$$

with probability 1, where $p_s := \varphi_s(p_0)$ is the solution of the deterministic recombination equation (4). □

The proof is an elementary application of Thm. 11.2.1 of [12]; see Prop. 1 of [6] for the explicit workout.

Note that the convergence in (14) applies for any given t, but need not carry over to $t \to \infty$. Indeed, if resampling is present, the population size required to get close to the deterministic solution is expected to grow over all bounds with increasing t. This is because, for every finite N, the Moran model with resampling and recombination is an *absorbing* Markov chain, which leads to fixation (i.e., to a homogeneous population of uniform type) in finite time with probability one (for the special case of just two types without recombination, the expected time is known to be of order N if the initial frequencies are both 1/2 [13, p. 93]). In sharp contrast, the deterministic system never loses any type, and the stationary state, the complete product measure (8), is, in a sense, even the most variable state accessible to the system. For increasing N, finite populations stay close to the deterministic limit for an increasing length of time.

4. Discrete time

Let us return to the deterministic setting and consider the *discrete-time* version of our single-crossover dynamics (4), that is,

$$\omega_{t+1} = \omega_t + \sum_{\alpha \in L} \widetilde{\varrho}_\alpha \big(R_\alpha - \mathbb{1}\big)(\omega_t) =: \widetilde{\Phi}(\omega_t)\,. \tag{15}$$

Here, the coefficients $\widetilde{\varrho}_\alpha > 0$, $\alpha \in L$, are the *probabilities* for a crossover at link α in every generation (as opposed to the *rates* ϱ_α of the continuous-time setting). Consequently, we must have $0 < \sum_{\alpha \in L} \widetilde{\varrho}_\alpha \leq 1$.

Based on the result for the continuous-time model, the solution is expected to be of the form

$$\omega_t \;=\; \widetilde{\Phi}^t(\omega_0) = \sum_{G \subseteq L} \widetilde{a}_G(t) R_G(\omega_0)\,, \tag{16}$$

with non-negative $\widetilde{a}_G(t)$, $G \subseteq L$, $\sum_{G \subseteq L} \widetilde{a}_G(t) = 1$, describing the (still unknown) coefficient functions arising from the dynamics. This representation of the solution was first stated by Geiringer [14]. The coefficient functions will have the same probabilistic interpretation as the corresponding $a_G(t)$ in the continuous-time model, so that $\widetilde{a}_G(t)$ is the probability that the links that have been involved in recombination until time t are exactly those of the set G.

But there is a crucial difference. Recall that, in continuous time, single crossovers imply *independence* of links, which is expressed in the product structure of the $a_G(t)$ (see Thm. 1). This independence is lost in discrete time, where a crossover event at one link forbids any other cut at other links in the same time step. It is therefore not surprising that a closed solution is not available in this case. It will, however, turn out that a solution can be stated in terms of the (generalised) eigenvalues of the system (which are known explicitly), together with coefficients to be determined via a simple recursion. But it is rewarding to take a closer look at the dynamics first.

Let us introduce the following abbreviations:

$$L_{\leq\alpha} := \{i \in L \mid i \leq \alpha\}\,, \quad L_{\geq\alpha} := \{i \in L \mid i \geq \alpha\}\,,$$

and, for each $G \subseteq L$,

$$G_{<\alpha} := \{i \in G \mid i < \alpha\}\,, \quad G_{>\alpha} := \{i \in G \mid i > \alpha\}\,.$$

Furthermore, we set $\eta := 1 - \sum_{\alpha \in L} \widetilde{\varrho}_\alpha$. The dynamics (15) is then reflected in the following dynamics of the coefficient functions:

Theorem 7. *For all $G \subseteq L$ and $t \in \mathbb{N}_0$, the coefficient functions $\widetilde{a}_G(t)$ evolve according to*

$$\widetilde{a}_G(t+1) = \eta\, \widetilde{a}_G(t) + \sum_{\alpha \in G} \widetilde{\varrho}_\alpha \Big(\sum_{H \subseteq L_{\geq \alpha}} \widetilde{a}_{G_{<\alpha} \cup H}(t) \Big) \Big(\sum_{K \subseteq L_{\leq \alpha}} \widetilde{a}_{K \cup G_{>\alpha}}(t) \Big), \quad (17)$$

with initial condition $\widetilde{a}_G(0) = \delta_{G,\varnothing}$. □

A verbal description of this dynamics was already given by Geiringer [14]; a formal proof may be found in [24, Thm. 3].

The above iteration is easily understood intuitively: A type x resulting from recombination at link α is composed of two segments $x_{<\alpha}$ and $x_{>\alpha}$. These segments themselves may have been pieced together in previous recombination events already, and the iteration explains the possible cuts these segments may carry along. The first term in the product stands for the type delivering the leading segment (which may bring along arbitrary cuts in the trailing segment), the second for the type delivering the trailing one (here any leading segment is allowed). The term $\eta\, \widetilde{a}_G(t)$ covers the case of no recombination.

Let us now have a closer look at the structure of the dependence between links in discrete time. To this end, note first that the set $G = \big\{\alpha_1, \ldots, \alpha_{|G|}\big\} \subseteq L$ with $\alpha_1 < \alpha_2 < \cdots < \alpha_{|G|}$ partitions $L \setminus G$ into $\mathcal{L}_G := \Big\{ I_0^G, I_1^G, \ldots, I_{|G|}^G \Big\}$, where

$$\begin{aligned} I_0^G &= \big\{ \alpha \in L : \tfrac{1}{2} \leq \alpha < \alpha_1 \big\}, \quad I_{|G|}^G = \big\{ \alpha \in L : \alpha_{|G|} < \alpha \leq \tfrac{2n-1}{2} \big\}, \\ \text{and } I_\ell^G &= \big\{ \alpha \in L : \alpha_\ell < \alpha < \alpha_{\ell+1} \big\} \text{ for } 1 \leq \ell \leq |G| - 1. \end{aligned} \quad (18)$$

Cutting all links in G decomposes the original system (of sites and links) into subsystems which are independent of each other from then on. In particular, the links in I_j become independent of those in I_k, for $k \neq j$. The probability that none of these subsystems experiences any further recombination is

$$\lambda_G = \prod_{i=0}^{|G|} \Big(1 - \sum_{\alpha_i \in I_i^G} \widetilde{\varrho}_{\alpha_i} \Big). \quad (19)$$

In particular, $\lambda_\varnothing = \eta = 1 - \sum_{\alpha \in L} \widetilde{\rho}_\alpha \geq 0$. The λ_G are, at the same time, the generalised eigenvalues that appear when the system is diagonalised and have been previously identified by Bennett [8], Lyubich [18] and Dawson [10].

A most instructive way to detail the effect of dependence is the *ancestral recombination process*: start from an individual in the present population, let time run backwards and consider how this individual's type has been pieced together from different fragments in the past. In the four-sites example of Fig. 5, the probability that exactly 1/2 and 3/2 have been cut reads

$$\begin{aligned} \widetilde{a}_{\{\frac{1}{2},\frac{3}{2}\}}(t) = & \widetilde{\varrho}_{\frac{1}{2}} \widetilde{\varrho}_{\frac{3}{2}} \sum_{k=0}^{t-2} \lambda_\varnothing^k \sum_{i=0}^{t-2-k} \lambda_{\frac{1}{2}}^i \lambda_{\{\frac{1}{2},\frac{3}{2}\}}^{t-2-k-i} \\ & + \widetilde{\varrho}_{\frac{1}{2}} \widetilde{\varrho}_{\frac{3}{2}} (1 - \widetilde{\varrho}_{\frac{5}{2}}) \sum_{k=0}^{t-2} \lambda_\varnothing^k \sum_{i=0}^{t-2-k} \lambda_{\frac{3}{2}}^i \lambda_{\{\frac{1}{2},\frac{3}{2}\}}^{t-2-k-i}. \end{aligned} \quad (20)$$

Here, the first (second) term corresponds to the possibility that link 1/2 (3/2) is the first to be cut. Obviously, the two possibilities are not symmetric: If 3/2 is the first to break, an additional factor of $(1 - \widetilde{\varrho}_{5/2})$ is required to guarantee that, at the time of the second recombination event (at 1/2), the trailing segment (sites 2 and 3) remains intact while the leading segment (sites 0 and 1) is cut.

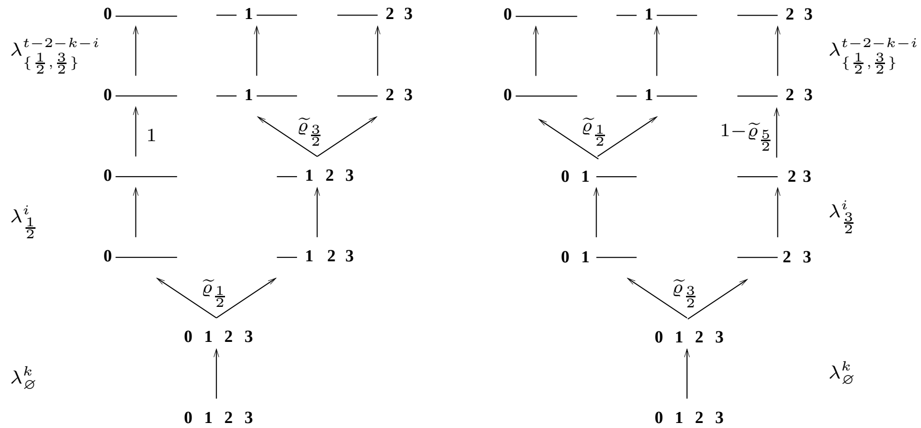

Figure 5. The ancestral recombination process: possible histories of the sequence 0123 (at the bottom). The two panels illustrate the two terms of $\widetilde{a}_{\{1/2,3/2\}}(t)$ in Eq. (20) (left: link 1/2 is cut first; right: link 3/2 is cut first.) Arrows point in the backward direction of time. Blank lines indicate arbitrary leading or trailing segments with which parts of the sequence have joined during recombination (they correspond to the asterisks $(*)$ in Fig. 3). The probability that nothing happens for a while (straight arrows only) is given by (powers of) the generalised eigenvalues (19).

Despite these complications, the discrete-time dynamics can again be solved, even directly at the level of the $\widetilde{a}_G(t)$, albeit slightly less explicitly than in continuous time. Indeed, it may be shown (and will be detailed in a forthcoming paper) that the coefficient functions have the form

$$\widetilde{a}^{(L)}_G(t) = \sum_{H \subseteq G} \gamma^{(L)}_G(H) (\lambda^{(L)}_H)^t,$$

where the upper index has again been added to indicate the dependence on the system. The coefficents $\gamma^{(L)}_G(H)$ $(H \subseteq G)$ are defined recursively as follows. For $G \neq \varnothing$,

$$\gamma^{(L)}_G(H) = \begin{cases} \frac{1}{\lambda^{(L)}_H - \lambda^{(L)}_{\varnothing}} \sum_{\alpha \in H} \widetilde{\varrho}_\alpha \gamma^{(L<\alpha)}_{G_{<\alpha}}(H_{<\alpha}) \gamma^{(L>\alpha)}_{G_{>\alpha}}(H_{>\alpha}), & H \neq \varnothing \\ -\sum_{\varnothing \neq J \subseteq G} \gamma^{(L)}_G(J), & H = \varnothing. \end{cases} \tag{21}$$

Together with the initial value $\gamma^{(L)}_{\varnothing}(\varnothing) = 1$, this may be solved recursively.

A diagonalisation of the system (analogous to that in Thm. 3) may also be achieved via a related, albeit technically more involved recursion [24].

5. Concluding remarks and outlook

The results presented here can naturally only represent a narrow segment from a large area with lively recent and current activities. Let us close this contribution by mentioning some important further directions in the context of recombination.

Our restriction to single crossovers provided a starting point with a uniquely transparent structure (mainly due to the independence of links in continuous time). However, arbitrary recombination patterns (which partition the set of links into two *arbitrary* parts) can also be dealt with, as has been done for the deterministic case in [18, 10]. The underlying mathematical structure will be further investigated in a forthcoming paper, for both the deterministic and the stochastic models.

Above, genetic material was exchanged reciprocally at recombination events, so that the length of each sequence remains constant. But sequences may also shift relative to each other before cutting and relinking (so-called *unequal crossover*), which entails changes in length, see [4] and references therein for more.

The most important aspect of modern population genetics is the backward-in-time point of view. This is natural because evolution is mainly a historical science and today's researchers try to infer the past from samples of individuals taken from present-day populations. We have hinted at this with our version of an ancestral recombination process, but would like to emphasise that this is only a toy version. The full version of this process also takes into account resampling (as in Sec. 3, with $b > 0$) and aims at the law of *genealogies* of *samples* from *finite* populations. This point of view was introduced by Hudson [16]. The fundamental concept here is the *ancestral recombination graph*: a branching-coalescing graph, where branching (backwards in time) comes about by recombination (as in Fig. 5), but lines may also coalesce where two individuals go back to a common ancestor (this corresponds to a resampling event forward in time). For recent introductions into this topic, see [11, Chap. 3], [15, Chap. 5], or [23, Chap. 7]; these texts also contain overviews of how recombination may be inferred from genomic datasets.

Last not least, recombination and resampling are but two of the various processes that act on genes in populations. Further inclusion of mutation and/or selection leads to a wealth of challenging problems, whose investigation has stimulated the exploration of new mathematical structures, concepts, and methods; let us only mention [7], [20], and [17] as recent examples. This development is expected to continue and intensify in the years to come – not least because it concerns the processes that have shaped present-day genomes.

References

[1] M. Aigner, *Combinatorial Theory*, Springer, Berlin, 1979. Reprint 1997.

[2] E. Baake, Mutation and recombination with tight linkage, *J. Math. Biol.* **42** (2001), 455–488.

[3] M. Baake, Recombination semigroups on measure spaces, *Monatsh. Math.* **146** (2005), 267–278 and **150** (2007), 83–84 (Addendum). arXiv:math.CA/0506099.

[4] M. Baake, Repeat distributions from unequal crossovers, *Banach Center Publ.* **80** (2008), 53–70. arXiv:0803.1270.

[5] M. Baake and E. Baake, An exactly solved model for mutation, recombination and selection, *Canad. J. Math.* **55** (2003), 3–41 and **60** (2008), 264–265 (Erratum). arXiv:math.CA/0210422.

[6] E. Baake and I. Herms, Single-crossover dynamics: Finite versus infinite populations, *Bull. Math. Biol.* **70** (2008), 603–624. arXiv:q-bio/0612024.

[7] N. H. Barton, A. M. Etheridge, and A. K. Sturm, Coalescence in a random background, *Ann. Appl. Prob.* **14** (2004), 754-785. arXiv:math/0406174.

[8] J. Bennett, On the theory of random mating, *Ann. Hum. Genet.* **18** (1954), 311–317.

[9] R. Bürger, *The Mathematical Theory of Selection, Recombination, and Mutation*, Wiley, Chichester, 2000.

[10] K. Dawson, The evolution of a population under recombination: How to linearize the dynamics, *Linear Algebra Appl.* **348** (2002), 115–137.

[11] R. Durrett, *Probability Models for DNA Sequence Evolution*, 2nd ed., Springer, New York, 2008.

[12] S. N. Ethier and T. G. Kurtz, *Markov Processes – Characterization and Convergence*, Wiley, New York, 1986. Reprint 2005.

[13] W. Ewens, *Mathematical Population Genetics*, Springer, Berlin, 2nd ed., 2004.

[14] H. Geiringer, On the probability theory of linkage in Mendelian heredity, *Ann. Math. Statist.* **15** (1944), 25–57.

[15] J. Hein, M. H. Schierup, and C. Wiuf, *Gene Genealogies, Variation and Evolution*, Oxford University Press, Oxford, 2005.

[16] R. R. Hudson, Properties of a neutral allele model with intragenic recombination, *Theor. Pop. Biol.* **23** (1983), 183–201.

[17] P.A. Jenkins and Y.S. Song, An asymptotic sampling formula for the coalescent with recombination, *Ann. Appl. Prob.*, in press.

[18] Y. Lyubich, *Mathematical Structures in Population Genetics*, Springer, New York, 1992.

[19] D. McHale and G. Ringwood, Haldane linearization of baric algebras, *J. London Math. Soc.* **28** (1983), 17–26.

[20] P. Pfaffelhuber, B. Haubold, and A. Wakolbinger, Approximate genealogies under genetic hitchhiking, *Genetics* **174** (2006), 1995-2008.

[21] E. Popa, Some remarks on a nonlinear semigroup acting on positive measures, in: O. Carja and I. I. Vrabie, eds., *Applied Analysis and Differential Equations*, pp. 308–319, World Scientific, Singapore, 2007.

[22] G. Ringwood, Hypergeometric algebras and Mendelian genetics, *Nieuw Archief v. Wiskunde* **3** (1985), 69–83.

[23] J. Wakeley, *Coalescent Theory.* Roberts, Greenwood Village, CO, 2009.

[24] U. von Wangenheim, E. Baake, and M. Baake, Single-crossover dynamics in discrete time, *J. Math. Biol.*, in press. arXiv:0906.1678.

Faculty of Technology, Bielefeld University, 33594 Bielefeld, Germany
E-mail: ebaake@techfak.uni-bielefeld.de