STOCHASTIC MODELS IN BIOLOGICAL SCIENCES
BANACH CENTER PUBLICATIONS, VOLUME 80
INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
WARSZAWA 2008

REPEAT DISTRIBUTIONS FROM UNEQUAL CROSSOVERS

MICHAEL BAAKE

Fakultät für Mathematik, Universität Bielefeld
Postfach 100131, 33501 Bielefeld, Germany
E-mail: mbaake@math.uni-bielefeld.de

Abstract. It is a well-known fact that genetic sequences may contain sections with repeated units, called repeats, that differ in length over a population, with a length distribution of geometric type. A simple class of recombination models with single crossovers is analysed that result in equilibrium distributions of this type. Due to the nonlinear and infinite-dimensional nature of these models, their analysis requires some nontrivial tools from measure theory and functional analysis, which makes them interesting also from a mathematical point of view. In particular, they can be viewed as quadratic, hence nonlinear, analogues of Markov chains.

1. Introduction. Recombination is a by-product of (sexual) reproduction, which leads to the mixing of parental genes by exchanging genes (or sequence parts) between homologous chromosomes (or DNA strands). This is achieved through an alignment of the corresponding sequences, along with crossover events that lead to a reciprocal exchange of the induced segments. In this process, an imperfect alignment may result in sequences that differ in length from the parental ones; this is known as *unequal crossover* (UC). Imperfect alignment is facilitated by the presence of repeated elements (as is observed within some rDNA sequences, compare [10]), and is believed to be an important driving mechanism for the evolution of the corresponding copy number distribution. The perhaps best studied case of repeated elements concerns microsatellites, see [9] and references given there for a summary. An important observation is that, within a population, the copy numbers vary, and often (at least approximately) follow a distribution of geometric type (meaning a geometric distribution or a finite convolution product thereof), see [9, 13, 3] and references therein for some experimental examples and findings.

2000 *Mathematics Subject Classification*: Primary 92D25, 92D10; Secondary 60B10, 46B50.

Key words and phrases: recombination, crossover dynamics, quadratic operators, stationary distributions, generating functions.

It is a pleasure to thank E. Baake and O. Redner for their cooperation, and C. Bank for carefully reading the manuscript.

The paper is in final form and no version of it will be published elsewhere.

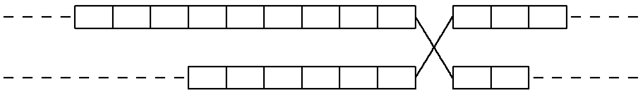

Fig. 1. Snapshot after an unequal crossover event as described in the text. Rectangles denote the relevant blocks, while the dashed lines indicate possible extensions with other elements that are disregarded here.

The microsatellites themselves may follow an evolutionary course independent of each other and thus give rise to evolutionary innovation. For a detailed discussion of these topics, see [9, 23] and references therein; for a brief introduction to molecular evolution, see also [8], or [7, 25] for a thorough overview. In this paper, which is mainly based on previous work by Redner [19, 18], we shall focus on the distribution of the copy numbers only, and disregard further aspects of the possible evolution of the repeated units themselves. We rather aim at analysing some simple models in order to understand the observed copy number or repeat distributions. Moreover, we are primarily interested in models that preserve the mean copy number, though our setting will be adequate to accommodate also more general models. In view of possible applications to systems where the copy number (slowly) changes with time, it seems natural to set up a frame that can cope with such a situation as well.

In the entire model class to be described below, one considers individuals whose genetic sequences contain a section with repeated units. These may vary in number, $i \in \mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$, where $i = 0$ is explicitly allowed and corresponds to no unit being present (yet). The composition of these sections (with respect to mutations that might have occurred) and the rest of the sequence are ignored here, as are details of the actual alignment process (e.g., whether partial loops of longer pieces are formed in order not to disturb the alignment outside the repeat region), see also [4] for a first discussion of possible models in this direction.

In the course of time, recombination events take place in which a random pair of individuals is formed and their respective sections are randomly aligned, possibly imperfectly with 'overhangs'. Then, both sequences are cut at a common position between two building blocks and their right (or left) fragments are interchanged. This so-called *unequal crossover* is schematically depicted in Figure 1. Obviously, the total number of relevant units is conserved in each event. While this is clearly a stochastic process, it is nevertheless interesting to investigate its deterministic limit, at least as a first step towards a better general understanding of this model class. To contribute to this first step, and to summarise what has been done in this direction so far, is the main aim of this contribution.

2. Description of the deterministic limit. As a first step for the analysis of crossover dynamics, we assume the population size to be (effectively) infinite, i.e., large enough so that random fluctuations may be neglected (finite populations will briefly be mentioned later on). We write $\mathcal{M}(X)$ for the (finite) measures on a space X, denote the restriction to positive measures by a superscript $+$, and indicate a restriction to measures of total variation r by a corresponding subscript (see [20] or [27] for a short summary of the measure theory needed here). Then, the distribution of the copy numbers over our population is described by a probability measure (or vector) $\boldsymbol{p} \in \mathcal{M}_1^+(\mathbb{N}_0)$, which we identify with an element $\boldsymbol{p} = (p_k)_{k\in\mathbb{N}_0}$ in the appropriate subset of $\ell^1(\mathbb{N}_0)$. Since we do not consider any genotype space other than $\mathbb{N}_0$ in this article, reference to it will be omitted in the sequel, so we write ℓ^1 instead of $\ell^1(\mathbb{N}_0)$ from now on. These spaces are complete in the metric derived from the usual ℓ^1 norm, which is the same as the total variation norm here. The metric is denoted by

$$d(\boldsymbol{p},\boldsymbol{q}) = \|\boldsymbol{p}-\boldsymbol{q}\|_1 = \sum_{k\geq 0} |p_k - q_k|. \tag{1}$$

Let us consider the above process (as well as various more general ones) on the level of the induced dynamics on the probability measures (i.e., in the infinite population limit mentioned above). With the notation just introduced, the dynamics can be described by means of the *recombinator*

$$\mathcal{R}(\boldsymbol{p})_i := \frac{1}{\|\boldsymbol{p}\|_1} \sum_{j,k,\ell\geq 0} T_{ij,k\ell}\, p_k\, p_\ell. \tag{2}$$

Here, $T_{ij,k\ell} \geq 0$ denotes the probability that a pair (k,ℓ) turns into (i,j), so, for normalisation, we require

$$\sum_{i,j\geq 0} T_{ij,k\ell} = 1, \qquad \text{for all } k,\ell \in \mathbb{N}_0. \tag{3}$$

The factor $p_k\, p_\ell$ in (2) describes the probability that a pair (k,ℓ) is formed, i.e., we assume that two individuals are chosen independently from the population. We assume further that, for all i,j,k,ℓ,

$$T_{ij,k\ell} = T_{ji,k\ell} = T_{ij,\ell k}, \tag{4}$$

i.e., that $T_{ij,k\ell}$ is symmetric with respect to both index pairs, which is reasonable and follows from the corresponding symmetry of the underlying process, compare Figure 1. Then, the summation over j in (2) represents the breaking-up of the pairs after the recombination event. These two ingredients (symmetry and summation) lead to the quadratic nature of the iteration process, see below for more and [14, 15] for the appearance of similar types of equations in a different class of biological models.

Condition (3) and the presence of the prefactor $1/\|\boldsymbol{p}\|_1$ in the defining Eq. (2) make $\mathcal{R}$ norm non-increasing, i.e., $\|\mathcal{R}(\boldsymbol{x})\|_1 \leq \|\boldsymbol{x}\|_1$, and positive homogeneous of degree 1, i.e., $\mathcal{R}(a\boldsymbol{x}) = |a|\mathcal{R}(\boldsymbol{x})$, for $\boldsymbol{x} \in \ell^1$ and $a \in \mathbb{R}$. Furthermore, $\mathcal{R}$ is a positive operator with $\|\mathcal{R}(\boldsymbol{x})\|_1 = \|\boldsymbol{x}\|_1$ for all positive elements $\boldsymbol{x} \in \ell^1$. Thus, it is guaranteed that $\mathcal{R}$ maps $\mathcal{M}_r^+$, the space of positive measures of total variation r, into itself. This subspace is complete in the topology induced by the norm $\|.\|_1$, i.e., by the metric d from (1). (For $r = 1$, the prefactor on the right hand side of (2) is redundant, but improves numerical stability of an iteration with the nonlinear mapping $\mathcal{R}$.)

Given an initial configuration $\boldsymbol{p}_0 = \boldsymbol{p}(0)$, the dynamics may be taken in discrete time steps, with subsequent generations,

$$\boldsymbol{p}(t+1) = \mathcal{R}(\boldsymbol{p}(t)), \qquad t \in \mathbb{N}_0. \tag{5}$$

This iteration reflects the following: due to random mating, it is sufficient to consider the dynamics at the level of the single strands, which will be combined into pairs again randomly in the next generation, according to the Hardy-Weinberg equilibrium [7].

Our treatment of this case will be set up in a way that also allows for a generalisation of the results to the analogous process in continuous time, where generations are overlapping,

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\,\boldsymbol{p}(t) = \varrho\,(\mathcal{R} - \mathbb{1})(\boldsymbol{p}(t)), \qquad t \geq 0. \tag{6}$$

This reflects what is called *instant mixing*, i.e., the instantaneous formation of pairs, their recombination and separation. In other words, the actual duration of the diplophase (or "paired phase") is neglected, which is an approximation that is justified as long as recombination is rare on the time scale of an individual life span.

Obviously, the (positive) parameter ϱ in (6) only leads to a rescaling of the time t. We therefore choose $\varrho = 1$ without loss of generality. Furthermore, it is easily verified that the fixed points of (5) are in one-to-one correspondence with the equilibria of (6). (In the sequel, we use the term fixed point for both discrete and continuous dynamics.)

As mentioned above, our main interest at present is in processes that conserve the total copy number in each event, i.e., $T_{ij,k\ell} > 0$ for $i + j = k + \ell$ only. More general scenarios are possible, and also interesting, but already the concept of an equilibrium gets rather involved, whence we do not go into further details here. Together with the normalisation (3) and the symmetry condition from above, this yields

$$\sum_{i,j\geq 0} i\,T_{ij,k\ell} = \sum_{i,j\geq 0} \frac{i+j}{2}\,T_{ij,k\ell} = \sum_{i,j\geq 0} T_{ij,k\ell}\,\frac{k+\ell}{2} = \frac{k+\ell}{2}, \tag{7}$$

the second equality of which is an alternative condition that can replace the strict preservation of the copy number as follows.

LEMMA 1. *Let $\mathcal{R}$ be defined by* (2), *with $T_{ij,k\ell} \geq 0$ subject to the normalisation* (3) *and the symmetry conditions* (4). *If also the second equality in* (7) *is satisfied, for all $k,\ell \in \mathbb{N}_0$, the mean copy number in the population is preserved.*

Proof. This is a simple calculation,

$$\sum_{i\geq 0} i\,\mathcal{R}(\boldsymbol{p})_i = \sum_{i,j,k,\ell\geq 0} i\,T_{ij,k\ell}\,p_k\,p_\ell = \sum_{k,\ell\geq 0} \frac{k+\ell}{2}\,p_k\,p_\ell = \sum_{k\geq 0} k\,p_k,$$

which shows the claim, provided that the mean $\mathfrak{m} := \sum_i i p_i$ is well-defined. ∎

From now on, we use the symbol $\mathfrak{m}$ for the mean, in order not to confuse it with summation indices and the like.

3. Markov chains for comparison. Let us take a brief detour to look at the linear counterpart, a countable state Markov chain, in the deterministic limit of the forward equation for the time evolution of its probability distribution. To this end, consider again probability vectors $\boldsymbol{p}$ on $\mathbb{N}_0$ and define

$$\mathfrak{M}(\boldsymbol{p})_k := \sum_{\ell=0}^{\infty} M_{k\ell}\,p_\ell,$$

for $k \in \mathbb{N}_0$, where all $M_{k\ell} \geq 0$ together with $\sum_{k=0}^{\infty} M_{k\ell} = 1$ for all $\ell \in \mathbb{N}_0$. This also makes the above sums well-defined on all elements of ℓ^1. Note that the matrix $M = (M_{k\ell})_{k,\ell\in\mathbb{N}_0}$ is transposed in comparison with the standard convention for Markov chains [21], because we are using it here in a dynamical systems context, with action of the matrix to the column vector on the right. The time evolution now either reads

$$\boldsymbol{p}(t+1) = \mathfrak{M}(\boldsymbol{p}(t)) \qquad \text{(in discrete time)} \tag{8}$$

or

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{p} = (\mathfrak{M} - \mathbb{1})(\boldsymbol{p}) \qquad \text{(in continuous time)}, \tag{9}$$

where the rate constant is again assumed to be 1, compare the remark after Eq. (6).

The iteration of the discrete version (8) on ℓ^1 is well-defined, while uniqueness of the solution of the initial value problem for the continuous time counterpart (9) on the same space follows from its global Lipschitz property,

$$\|\mathfrak{M}(\boldsymbol{p}) - \mathfrak{M}(\boldsymbol{q})\|_1 = \|\mathfrak{M}(\boldsymbol{p} - \boldsymbol{q})\|_1 \leq \sum_{k,\ell\geq 0} M_{k\ell}\, |p_\ell - q_\ell| = \sum_{\ell\geq 0} |p_\ell - q_\ell| = \|\boldsymbol{p} - \boldsymbol{q}\|_1,$$

which holds for all $\boldsymbol{p}, \boldsymbol{q} \in \ell^1$. The properties of the matrix M guarantee that the positive cone as well as the simplex of probability vectors are preserved in forward time. Consequently, one can consider (8) and (9) as dynamical systems on ℓ^1. As the latter is a Banach space of infinite dimension, the unit ball is no longer compact in the norm topology, whence some extra care is needed for the results.

As before, fixed points of (8) line up with equilibria of (9), so that we speak of fixed points in both cases. Their existence is provided by Perron-Frobenius theory for countable state Markov matrices, see [12, Ch. 7.1] or [21, Ch. 5] for a detailed account. Irreducibility, aperiodicity and primitivity are defined as in the finite-dimensional case without difficulty. However, for meaningful results on eigenvalues and eigenvectors, one additionally needs the concept of *recurrence*, see [12, p. 197 f.] for a nice summary.

The Perron value λ emerges from the radius of convergence, ρ, of the power series $T(z) = \sum_{n\geq 0} (zM)^n$ via $\rho = 1/\lambda$. Clearly, we have $\rho \geq 1$ for a Markov matrix. If one diagonal entry (and then any) of $T(z)$ diverges at 1 (so that $\rho = 1$ in this case, compare [21, Thm. 6.6]), the countable state Markov matrix M is called *recurrent*, where the behaviour of the diagonal element $T(z)_{ii}$, as $z \to 1$, reflects the expected number of recurrences to i, which is infinite in this case. Moreover, a unique normalised and strictly positive (right) eigenvector $\boldsymbol{p} \in \mathcal{M}_1^+$ exists with $\mathfrak{M}(\boldsymbol{p}) = \boldsymbol{p}$, see [21, Thm. 5.4]. This probability vector has the meaning of the unique equilibrium distribution and is the desired fixed point of the dynamics.

Assume for a moment, in addition to the above conditions on $\mathfrak{M}$, that

$$\sum_{i\geq 0} i M_{ij} = j, \quad \text{for all } j. \tag{10}$$

As before, this is a sufficient condition for the mean to be preserved under the dynamics, because one has

$$\sum_{i\geq 0} i\, \mathfrak{M}(\boldsymbol{p})_i = \sum_{i\geq 0} \sum_{j\geq 0} i M_{ij} p_j = \sum_{j\geq 0} \sum_{i\geq 0} i M_{ij} p_j = \sum_{j\geq 0} j p_j,$$

with the interchange of summation being permissible due to absolute convergence of the sums involved, provided that $\mathfrak{m} = \sum_j j p_j$ exists. However, a condition of type (10) is usually too restrictive for a linear system, wherefore we do not impose it here. As we shall see, the mean copy number can be preserved without it.

A probability vector $\boldsymbol{p}$ is called *reversible* for $\mathfrak{M}$ when the detailed balance equation

$$M_{k\ell}\, p_\ell = M_{\ell k}\, p_k \tag{11}$$

holds for all $k, \ell \in \mathbb{N}_0$. An important consequence is that any reversible $\boldsymbol{p}$ is automatically a fixed point of $\mathfrak{M}$:

$$\mathfrak{M}(\boldsymbol{p})_k = \sum_{\ell \geq 0} M_{k\ell}\, p_\ell = \sum_{\ell \geq 0} M_{\ell k}\, p_k = p_k .$$

Reversibility often provides a simpler way to actually calculate a specific fixed point than the defining matrix eigenvalue equation.

Since the Perron-Frobenius eigenvalue λ need not be isolated in the spectrum of M, the convergence properties are more subtle than in the finite-dimensional situation. Under certain extra conditions (e.g., if λ *is* isolated), the time evolution of an arbitrary initial condition converges exponentially fast towards the fixed point. However, when the matrix M is not only recurrent, but also positive recurrent, one has at least convergence of the discrete iteration, see [21] for details. Here, positive recurrence means that the expected time for a return to the state i is *finite*, which is clearly stronger than mere recurrence.

The standard geometric distribution with parameter $\alpha \in (0,1)$ is a discrete probability distribution on $\mathbb{N}_0$, defined by the probability vector $\boldsymbol{p}$ with

$$p_n := \alpha(1-\alpha)^n, \quad \text{for } n \in \mathbb{N}_0. \tag{12}$$

Clearly, $p_n > 0$ and $\sum_{n \geq 0} p_n = 1$, while $\mathfrak{m} = \sum_{n \geq 0} n p_n = (1-\alpha)/\alpha$, so that $\alpha = 1/(\mathfrak{m}+1)$. If we define the matrix $M = (M_{ij})_{i,j \geq 0}$ by $M_{ij} = p_i$, one has

$$(M\boldsymbol{p})_i = \sum_j M_{ij} p_j = p_i \sum_j p_j = p_i ,$$

so that $M\boldsymbol{p} = \boldsymbol{p}$. One clearly has $M^n = M$ for all $n \in \mathbb{N}$. Consequently, each entry of $T(z)$ is a geometric series of the form $M_{ij}(1 + z + z^2 + \ldots)$, which thus diverges at $z = 1$. In particular, M is (positive) recurrent.

The matrix M does not satisfy Eq. (10). Nevertheless, the mean copy number is preserved in the following sense. Let $\boldsymbol{a}$ be an arbitrary probability vector with mean $\mathfrak{m}$, and $\boldsymbol{p}$ the geometric distribution according to (12) with the same mean. With the corresponding matrix M, one then finds

$$\sum_{i,j} i M_{ij} a_j = \sum_{i,j} i p_i a_j = \sum_i i p_i \sum_j a_j = \mathfrak{m},$$

which results in the mean preservation, provided one starts with an initial condition $\boldsymbol{a}$ of mean $\mathfrak{m}$. Otherwise, the iteration maps $\boldsymbol{a}$ to an image of mean $\mathfrak{m}$ in the first step, and preserves $\mathfrak{m}$ in all subsequent iterations.

Further eigenvectors of M are given by $\boldsymbol{q}^{(\ell)} := \boldsymbol{e}_0 - \boldsymbol{e}_\ell$ for $\ell \in \mathbb{N}$, where $\boldsymbol{e}_i$ is the standard basis vector with 1 in coordinate i and 0 otherwise. All these extra vectors belong to the eigenvalue 0, which is the only other eigenvalue of M. In fact, M is diagonalisable, and it is not difficult to see that an arbitrary vector $\boldsymbol{a} = (a_0, a_1, a_2, \ldots) \in \ell^1$ can be written as a convergent expansion, $\boldsymbol{a} = \beta\boldsymbol{p} + \sum_{\ell\geq 1}(\beta p_\ell - a_\ell)\boldsymbol{q}^{(\ell)}$, where $\beta = \sum_{i\geq 0} a_i$. Consequently, the chosen eigenvectors of M form a basis of ℓ^1. If $U = (\boldsymbol{p}, \boldsymbol{q}^{(1)}, \boldsymbol{q}^{(2)}, \ldots)$ denotes the matrix that columnwise consists of the eigenvectors of M, one has

$$M = U\operatorname{diag}(1,0,0,\ldots)\, U^{-1},$$

which makes the relation $M^n = M$ for $n \in \mathbb{N}$ particularly transparent. Moreover, one sees that M commutes with all matrices N of the form $N = UAU^{-1}$ where A has the block form

$$A = \begin{pmatrix} a & \boldsymbol{0}^t \\ \boldsymbol{0} & A' \end{pmatrix}$$

with an arbitrary matrix A'. Restricting N so that $M + N$ is still Markov, one can find multi-parameter families of Markov matrices that share the given stationary geometric distribution $\boldsymbol{p}$. The same stationary probability vector $\boldsymbol{p}$ can thus arise from many other Markov chains as well.

Let us now return to the bilinear counterpart to see which of these structural properties possess an analogue, and to describe the setting of our later analysis.

4. General structure of the bilinear system. Consider the crossover dynamics as defined by (2). Let us begin by stating the following general fact.

PROPOSITION 1. *If the recombinator $\mathcal{R}$ of* (2) *satisfies the normalisation conditions* (3), *one has the global Lipschitz condition*

$$\|\mathcal{R}(\boldsymbol{x}) - \mathcal{R}(\boldsymbol{y})\|_1 \leq C\|\boldsymbol{x} - \boldsymbol{y}\|_1,$$

with constant $C = 3$ on ℓ^1, respectively $C = 2$ if $\boldsymbol{x}$, $\boldsymbol{y} \in \mathcal{M}_r$.

Proof. Let $\boldsymbol{x}$, $\boldsymbol{y} \in \ell^1$ be non-zero (otherwise the statement is trivial). Then, one has

$$\begin{aligned}
\|\mathcal{R}(\boldsymbol{x}) - \mathcal{R}(\boldsymbol{y})\|_1 &= \sum_{i\geq 0}\biggl|\sum_{j,k,\ell\geq 0} T_{ij,k\ell}\biggl(\frac{x_k\,x_\ell}{\|\boldsymbol{x}\|_1} - \frac{y_k\,y_\ell}{\|\boldsymbol{y}\|_1}\biggr)\biggr| \\
&\leq \sum_{k,\ell\geq 0}\biggl|\frac{x_k\,x_\ell}{\|\boldsymbol{x}\|_1} - \frac{y_k\,y_\ell}{\|\boldsymbol{y}\|_1}\biggr|\sum_{i,j\geq 0} T_{ij,k\ell} = \sum_{k,\ell\geq 0}\biggl|\frac{x_k\,x_\ell}{\|\boldsymbol{x}\|_1} - \frac{x_k\,y_\ell}{\|\boldsymbol{x}\|_1} + \frac{x_k\,y_\ell}{\|\boldsymbol{x}\|_1} - \frac{y_k\,y_\ell}{\|\boldsymbol{y}\|_1}\biggr| \\
&\leq \sum_{k,\ell\geq 0}\biggl(\frac{|x_k|}{\|\boldsymbol{x}\|_1}|x_\ell - y_\ell| + |y_\ell|\biggl|\frac{x_k}{\|\boldsymbol{x}\|_1} - \frac{y_k}{\|\boldsymbol{y}\|_1}\biggr|\biggr) = \|\boldsymbol{x} - \boldsymbol{y}\|_1 + \frac{\bigl\|\|\boldsymbol{y}\|_1\boldsymbol{x} - \|\boldsymbol{x}\|_1\boldsymbol{y}\bigr\|_1}{\|\boldsymbol{x}\|_1}.
\end{aligned}$$

The last term becomes

$$\frac{1}{\|\boldsymbol{x}\|_1}\bigl\|\|\boldsymbol{y}\|_1\boldsymbol{x} - \|\boldsymbol{x}\|_1\boldsymbol{y}\bigr\|_1 = \frac{1}{\|\boldsymbol{x}\|_1}\bigl\|(\|\boldsymbol{y}\|_1 - \|\boldsymbol{x}\|_1)\boldsymbol{x} + \|\boldsymbol{x}\|_1(\boldsymbol{x} - \boldsymbol{y})\bigr\|_1 \leq 2\|\boldsymbol{x} - \boldsymbol{y}\|_1,$$

from which $\|\mathcal{R}(\boldsymbol{x}) - \mathcal{R}(\boldsymbol{y})\|_1 \leq 3\|\boldsymbol{x} - \boldsymbol{y}\|_1$ follows for $\boldsymbol{x}$, $\boldsymbol{y} \in \ell^1$. If $\boldsymbol{x}$, $\boldsymbol{y} \in \mathcal{M}_r$, one has $\|\boldsymbol{x}\|_1 = \|\boldsymbol{y}\|_1$ and the above calculation simplifies to $\|\mathcal{R}(\boldsymbol{x}) - \mathcal{R}(\boldsymbol{y})\|_1 \leq 2\|\boldsymbol{x} - \boldsymbol{y}\|_1$. ■

In continuous time, this is a sufficient condition for the existence of a unique solution of the initial value problem (6), compare [1, Thms. 7.6 and 10.3].

It is instructive to generalise the notion of reversibility. We call a probability vector $\boldsymbol{p} \in \mathcal{M}_1^+$ *reversible* for a recombinator $\mathcal{R}$ of the form (2) if, for all $i, j, k, \ell \geq 0$,

$$T_{ij,k\ell}\, p_k\, p_\ell = T_{k\ell,ij}\, p_i\, p_j. \tag{13}$$

Though this set of equations for detailed balance is much more restrictive than its linear counterpart in Eq. (11), the relevance of this concept is evident from the following property.

Lemma 2. *If $\boldsymbol{p} \in \mathcal{M}_1^+$ is reversible for $\mathcal{R}$, it is also a fixed point of $\mathcal{R}$.*

Proof. Assume $\boldsymbol{p}$ to be reversible for $\mathcal{R}$. Then, by (3),

$$\mathcal{R}(\boldsymbol{p})_i = \sum_{j,k,\ell \geq 0} T_{ij,k\ell}\, p_k\, p_\ell = \sum_{j,k,\ell \geq 0} T_{k\ell,ij}\, p_i\, p_j = p_i \sum_{j \geq 0} p_j = p_i,$$

for all $i \in \mathbb{N}_0$, which shows the claim. ■

Returning to the original question of the existence of fixed points, we now recall the following facts, compare [6, 22] for details and proofs, which are needed for some general statements in the fixed point discussion.

Proposition 2 ([27, Cor. to Thm. V.1.5]). *Assume the sequence $\big(\boldsymbol{p}^{(n)}\big)$ in $\mathcal{M}_1^+$ to converge in the weak-$*$ topology (i.e., pointwise, or vaguely) to some $\boldsymbol{p} \in \mathcal{M}_1^+$, i.e.,*

$$\lim_{n\to\infty} p_k^{(n)} = p_k \quad \text{for all } k \in \mathbb{N}_0, \qquad \text{with } p_k \geq 0 \text{ and } \textstyle\sum_{k\geq 0} p_k = 1.$$

Then, it also converges weakly (in the probabilistic sense) and in total variation, i.e., $\lim_{n\to\infty} \|\boldsymbol{p}^{(n)} - \boldsymbol{p}\|_1 = 0$. ■

Recall from [6] that a set of measures $\mathcal{M} \subset \mathcal{M}_1^+$ is called *tight* when, for every $\varepsilon > 0$, there is an $m \in \mathbb{N}_0$ such that $\sum_{k\geq m} p_k < \varepsilon$, simultaneously for all $\boldsymbol{p} \in \mathcal{M}$. This is a uniformity condition which serves as a condition for the compactness needed later on.

Proposition 3. *Assume that the recombinator $\mathcal{R}$ from (2) satisfies the normalisation (3) and possesses a convex, weak-$*$ closed invariant set $\mathcal{M} \subset \mathcal{M}_1^+$, i.e., $\mathcal{R}(\mathcal{M}) \subset \mathcal{M}$, that is tight. Then, $\mathcal{R}$ has a fixed point in $\mathcal{M}$.*

Proof. Prohorov's theorem [22, Thm. III.2.1] states that tightness and relative compactness in the weak-$*$ topology are equivalent (see also [6, Chs. 1.1 and 1.5]). In our case, $\mathcal{M}$ is tight and weak-$*$ closed, therefore, due to Proposition 2, norm compact. Further, $\mathcal{M}$ is convex by assumption, and $\mathcal{R}$ is (norm) continuous by Proposition 1. Thus, the claim follows from the Leray–Schauder–Tychonov fixed point theorem [20, Thm. V.19]. ■

For several explicit models, we shall see that such compact invariant subsets indeed exist. On the other hand, once again due to the infinite-dimensional nature of the dynamical system, their identification and use for the various proofs is essential.

5. Takahata's model. An early and now classic example was given by Takahata [24]. In our terminology, he used a recombinator based upon the transition probabilities

$$T_{ij,k\ell} := \frac{1}{k+\ell+1}\, \delta_{i+j,k+\ell}, \tag{14}$$

for $i, j, k, \ell \in \mathbb{N}_0$. Observing $\operatorname{card}\{(i,j) \mid i,j \in \mathbb{N}_0,\, i+j = k+\ell\} = k+\ell+1$, it is clear that T just describes a recombination with uniform distribution of the copy number pairs (k,ℓ) on each (finite) block of possibilities with $k+\ell$ fixed. One can also check, via Eq. (7) and Lemma 1, that the mean $\mathfrak{m}$ is preserved. On the basis of Eq. (14), the action of the recombinator from Eq. (2) on probability vectors $\boldsymbol{p}$ simplifies to

$$\mathcal{R}(\boldsymbol{p})_i = \sum_{\substack{k,\ell \geq 0 \\ k+\ell \geq i}} \frac{1}{k+\ell+1}\, p_k p_\ell . \tag{15}$$

Though this model is mathematically rather transparent, it lacks a good intuitive justification on the level of the biological processes. Nevertheless, its properties seem to be in acceptable agreement with at least some of the observations, compare [13, 3], though other results, as those shown in [9], indicate that also other types of equilibria appear in experiment.

Proposition 4. *The probability vector* $\boldsymbol{p}$ *defined by*

$$p_n = \frac{1}{\mathfrak{m}+1}\left(\frac{\mathfrak{m}}{\mathfrak{m}+1}\right)^n, \quad n \in \mathbb{N}_0,$$

is a reversible equilibrium with mean $\mathfrak{m}$ *for the dynamics based on* T *of* (14).

Proof. Using standard identities with geometric series and their derivatives, it is easy to check that $\boldsymbol{p}$ indeed defines a probability vector on $\mathbb{N}_0$ with mean $\mathfrak{m}$. Detailed balance follows from a simple calculation,

$$\begin{aligned} T_{ij,k\ell}\, p_k p_\ell &= \frac{\delta_{i+j,k+\ell}}{k+\ell+1}\left(\frac{1}{\mathfrak{m}+1}\right)^2 \left(\frac{\mathfrak{m}}{\mathfrak{m}+1}\right)^{k+\ell} \\ &= \frac{\delta_{k+\ell,i+j}}{i+j+1}\left(\frac{1}{\mathfrak{m}+1}\right)^2 \left(\frac{\mathfrak{m}}{\mathfrak{m}+1}\right)^{i+j} = T_{k\ell,ij}\, p_i p_j , \end{aligned}$$

thus completing the claim by means of Lemma 2. ∎

These equilibria are geometric distributions as also discussed above in the Markov context. However, in view of some experimental findings reported in [9] and further arguments put forward in [23], one would like to see an initial rise, and perhaps also a maximum in the vicinity of $n \approx \mathfrak{m}$. One should note that measurements often skip the entries for small copy numbers (which seem to be rather unreliable), so that a graph with a power law decay need not indicate the absence of some (weak) form of a maximum. As the methods for the further analysis of Takahata's model are similar to what we need later on for alternative models, we first continue to investigate Takahata's model.

Theorem 1. *If the initial condition, with mean* $\mathfrak{m}$, *satisfies a certain tightness condition* ($\limsup_{k\to\infty} \sqrt[k]{p_k(0)} < 1$), *the dynamics, both in discrete and in continuous time, converges to the equilibrium vector* $\boldsymbol{p}$ *from Proposition* 4, *with* $\lim_{t\to\infty} \|\boldsymbol{p}(t) - \boldsymbol{p}\|_1 = 0$.

The proof of this theorem, quite appropriately for the present context, uses an approach via generating functions and then relies on Banach's contraction principle. It requires several preparatory steps.

Let α and δ be fixed, with $0 < \alpha \leq \delta < \infty$, and consider the space

$$X_{\alpha,\delta} := \{\boldsymbol{a} = (a_k)_{k\in\mathbb{N}_0} \mid a_0 = 1,\ a_1 = \alpha,\ \text{and } 0 \leq a_k \leq \delta^k \text{ for all } k \geq 2\}. \tag{16}$$

If equipped with the metric

$$d(\boldsymbol{a}, \boldsymbol{b}) = \sum_{k\geq 0} d_k \, |a_k - b_k|, \tag{17}$$

where $d_k = (\gamma/\delta)^k$ for some $0 < \gamma < \frac{1}{3}$, the space $X_{\alpha.\delta}$ is compact [19, Prop. 5].

Let us define a new vector, $b(\boldsymbol{p})$, for suitable $\boldsymbol{p}$, by

$$b(\boldsymbol{p})_k := \sum_{\ell\geq k} \binom{\ell}{k} p_\ell, \tag{18}$$

which is certainly well-defined for all $\boldsymbol{p}$ with $\limsup_{k\to\infty} \sqrt[k]{p_k} < 1$, by an application of [19, Prop. 6]. This proposition also clarifies the connection with the space $X_{\alpha,\delta}$ for suitable parameters α and δ. As we shall see, $X_{\alpha,\delta}$ is an example of a compact, convex space that is invariant under the recombinator dynamics. It is easy to check that one has $b(\boldsymbol{p})_0 = 1$ and $b(\boldsymbol{p})_1 = \mathfrak{m}$, so that we need $X_{\alpha,\delta}$ with $\alpha = \mathfrak{m}$ and $\delta \geq \mathfrak{m}$.

Lemma 3. *For any* $\boldsymbol{p}$ *with* $\limsup_{k\to\infty} \sqrt[k]{p_k} < 1$, *one has the convolution identity*

$$b\big(\mathcal{R}(\boldsymbol{p})\big)_k = \frac{1}{k+1} \sum_{m=0}^{k} b(\boldsymbol{p})_m b(\boldsymbol{p})_{k-m}.$$

Proof. Let $\boldsymbol{p}$ be an arbitrary probability vector with $\limsup_{k\to\infty} \sqrt[k]{p_k} < 1$, so that the mapping b is well-defined. The left hand side leads to

$$\begin{aligned} b\big(\mathcal{R}(\boldsymbol{p})\big)_k &= \sum_{\ell\geq k} \binom{\ell}{k} \mathcal{R}(\boldsymbol{p})_\ell = \sum_{\ell\geq k} \binom{\ell}{k} \sum_{\substack{r,s\geq 0\\ r+s\geq \ell}} \frac{p_r p_s}{r+s+1} \\ &= \sum_{\substack{r,s\geq 0\\ r+s\geq k}} \frac{p_r p_s}{r+s+1} \sum_{\ell=k}^{r+s} \binom{\ell}{k} = \frac{1}{k+1} \sum_{\substack{r,s\geq 0\\ r+s\geq k}} \binom{r+s}{k} p_r p_s, \end{aligned}$$

where a standard identity on binomial coefficients was used in the last step.

On the other hand, one finds

$$\begin{aligned} \sum_{m=0}^{k} b(\boldsymbol{p})_m b(\boldsymbol{p})_{k-m} &= \sum_{\substack{m,n\geq 0\\ m+n=k}} \Big(\sum_{r\geq m} \binom{r}{m} p_r\Big)\Big(\sum_{s\geq n} \binom{s}{n} p_s\Big) \\ &= \sum_{\substack{m,n\geq 0\\ m+n=k}} \sum_{\substack{r,s\geq 0\\ r+s\geq k}} \binom{r}{m}\binom{s}{n} p_r p_s = \sum_{\substack{r,s\geq 0\\ r+s\geq k}} p_r p_s \sum_{\substack{m,n\geq 0\\ m+n=k}} \binom{r}{m}\binom{s}{n} \\ &= \sum_{\substack{r,s\geq 0\\ r+s\geq k}} \binom{r+s}{k} p_r p_s, \end{aligned}$$

again using a standard identity, together with the fact that $\binom{n}{m} = 0$ for $m > n$ when n is an integer. A comparison of the two calculations establishes the claim. ∎

The further relevance of Lemma 3 stems from the following property of the generating function of $\boldsymbol{p}$, defined by $\psi(z) = \sum_{\ell\geq 0} p_\ell z^\ell$. When rewritten as a Taylor series around 1 rather than around 0, one obtains

$$\psi(z) = \sum_{\ell\geq 0} p_\ell z^\ell = \sum_{k\geq 0} \Biggl(\sum_{\ell\geq k} \binom{\ell}{k} p_\ell \Biggr) (z-1)^k = \sum_{k\geq 0} b(\boldsymbol{p})_k \, (z-1)^k. \tag{19}$$

Under the assumptions on $\boldsymbol{p}$, the radius of convergence of $\psi(z)$ is larger than 1, so that this calculation is on firm grounds. Lemma 3 now tells us that we may study the recombination action on the level of the expansion coefficients.

Let us therefore define the induced recombination operator $\widetilde{\mathcal{R}}$ on any space of type $X_{\alpha,\delta}$, with $\delta \geq \alpha$, by $\widetilde{\mathcal{R}}(b(\boldsymbol{a})) = b(\mathcal{R}(\boldsymbol{a}))$, which establishes a commuting diagram of the mappings $\mathcal{R}$ and $\widetilde{\mathcal{R}}$ in the obvious way. More precisely, one first restricts the action of $\mathcal{R}$ to a suitable subspace of $\mathcal{M}_1^+$, so that the mapping b is well-defined. If $\boldsymbol{p}$ satisfies the condition of Lemma 3, so that the radius of convergence of ψ exceeds 1, the probability vector $\boldsymbol{p}$ is also completely determined by its moments, compare [22, Thm. II.12.7] together with the observation that $\psi(e^{it})$ is the (convergent) moment generating function of $\boldsymbol{p}$. As all moments, in turn, are specified by the entries of $b(\boldsymbol{p})$, the latter uniquely determines $\boldsymbol{p}$ in this situation.

It is easy to check that the vector $(1, \alpha, \alpha^2, \ldots)$ is a fixed point of $\widetilde{\mathcal{R}}$ in $X_{\alpha,\delta}$, for any $\delta \geq \alpha$. Choosing $\alpha = \mathfrak{m}$, this vector is the image of the probability vector $\boldsymbol{p}$ from Proposition 4 under the mapping b.

Proposition 5. *On $X_{\alpha,\delta}$, the map defined by $\widetilde{\mathcal{R}}$ is a contraction. In particular, it is a globally Lipschitz continuous mapping of $X_{\alpha,\delta}$ into itself.*

Proof. Let $\delta \geq \alpha > 0$ be given, as well as arbitrary $\boldsymbol{a}, \boldsymbol{b} \in X_{\alpha,\delta}$. Clearly, we have $\widetilde{\mathcal{R}}(\boldsymbol{a})_0 = 1$ and $\widetilde{\mathcal{R}}(\boldsymbol{a})_1 = \alpha$. For $k \geq 2$, one finds $\widetilde{\mathcal{R}}(\boldsymbol{a})_k = \frac{1}{k+1} \sum_{\ell=0}^{k} a_\ell a_{k-\ell} \leq \delta^k$. This proves that $\widetilde{\mathcal{R}}$ maps $X_{\alpha,\delta}$ into itself.

The space $X_{\alpha,\delta}$ is equipped with the metric d from (17). Since, due to $\boldsymbol{b} \in X_{\alpha,\delta}$, also $\widetilde{\mathcal{R}}(\boldsymbol{b})_0 = 1$ and $\widetilde{\mathcal{R}}(\boldsymbol{b})_1 = \alpha$, the contraction estimate reads as follows.

$$\begin{aligned} d(\widetilde{\mathcal{R}}(\boldsymbol{a}), \widetilde{\mathcal{R}}(\boldsymbol{b})) &= \sum_{k\geq 2} \frac{d_k}{k+1} \Bigl| \sum_{\ell=0}^{k} (a_\ell a_{k-\ell} - b_\ell b_{k-\ell}) \Bigr| = \sum_{k\geq 2} \frac{d_k}{k+1} \Bigl| \sum_{\ell=0}^{k} (a_\ell - b_\ell)(a_{k-\ell} + b_{k-\ell}) \Bigr| \\ &\leq \sum_{k\geq 2} \frac{2\, d_k}{k+1} \sum_{\ell=2}^{k} \delta^{k-\ell} \, |a_\ell - b_\ell| = \sum_{\ell\geq 2} d_\ell \, |a_\ell - b_\ell| \sum_{k\geq \ell} \frac{2}{k+1} \, \delta^{k-\ell} \, \frac{d_k}{d_\ell}. \end{aligned}$$

With the choice $d_k = (\gamma/\delta)^k$, where we had $\gamma < \frac{1}{3}$, we can now find, for $\ell \geq 2$, an upper bound for the inner sum,

$$\sum_{k\geq \ell} \frac{2}{k+1} \, \delta^{k-\ell} \, \frac{d_k}{d_\ell} \leq \frac{2}{3} \sum_{k\geq \ell} \gamma^{k-\ell} = \frac{2}{3-3\gamma} =: C < 1,$$

which, together with the previous calculation, proves the contraction property,

$$d(\widetilde{\mathcal{R}}(\boldsymbol{a}), \widetilde{\mathcal{R}}(\boldsymbol{b})) \leq C \, d(\boldsymbol{a}, \boldsymbol{b}),$$

with contraction constant $C < 1$. Clearly, this also means that $\widetilde{\mathcal{R}}$ is globally Lipschitz continuous. ∎

This shows that, in discrete time, we have exponentially fast convergence of the sequence $(\widetilde{\mathcal{R}}^n(\boldsymbol{a}))_{n\geq 1}$, with $\boldsymbol{a} \in X_{\alpha,\delta}$, to a unique fixed point in $X_{\alpha,\delta}$. It is specified by the mean copy number $\mathfrak{m}$ of the probability vector $\boldsymbol{p}$ that underlies $\boldsymbol{a} = b(\boldsymbol{p})$, via $\alpha = \mathfrak{m}$, see above. Clearly, this fixed point (in $X_{\alpha,\delta}$) is the image (under b) of the equilibrium vector $\boldsymbol{p} \in \mathcal{M}_1^+$ calculated earlier in Proposition 4, as the mapping b is invertible in this situation. The claim of Theorem 1 for discrete time is now clear, with exponentially fast convergence to the equilibrium, from any initial condition as specified there.

For the slightly more involved treatment of the continuous time case, we refer to [19]. It is based on the construction of a Lyapunov function, similar to that of [19, Prop. 13].

6. Internal crossover. Another rather obvious model is based on the assumption that the shorter of the two sequences (or stretches) can align with any connected block of the longer sequence, but without any overhang. This situation has been coined internal unequal crossover, or *internal crossover* for short. Here, restricting to probability measures on $\mathbb{N}_0$, the recombinator (2) simplifies to

$$\mathcal{R}_0(\boldsymbol{p})_i = \sum_{\substack{k,\ell\geq 0\\ k\wedge\ell\leq i\leq k\vee\ell}} \frac{p_k p_\ell}{1+|k-\ell|}, \tag{20}$$

where $k \wedge \ell$ ($k \vee \ell$) stands for the minimum (the maximum) of k and ℓ, see [23, 19, 18] for details on this model. We choose the notation $\mathcal{R}_0$ for reasons that will become clear later on.

In our search for fixed points, it is again useful to look for probability vectors that are reversible for $\mathcal{R}_0$. Since both forward and backward transition probabilities are simultaneously non-zero only when $\{i,j\} = \{k,\ell\} \subset \{n, n+1\}$ for some n, the components p_k may only be positive on this small set as well. By the following proposition, this indeed characterises all fixed points of this case.

Proposition 6. *A probability measure $\boldsymbol{p} \in \mathcal{M}_1^+$ is a fixed point of $\mathcal{R}_0$ if and only if its mean copy number $\mathfrak{m} = \sum_{k\geq 0} k\, p_k$ is finite, together with $p_{\lfloor \mathfrak{m}\rfloor} = \lfloor \mathfrak{m}\rfloor + 1 - \mathfrak{m}$, $p_{\lceil \mathfrak{m}\rceil} = \mathfrak{m} + 1 - \lceil \mathfrak{m}\rceil$, and $p_k = 0$ for all other k. This includes the case that $\mathfrak{m}$ is a non-negative integer, where $p_{\lfloor \mathfrak{m}\rfloor} = p_{\lceil \mathfrak{m}\rceil} = p_{\mathfrak{m}} = 1$.*

Proof. The 'if' follows easily by insertion into (13) and Lemma 2. For the 'only if' part, let i denote the smallest integer such that $p_i > 0$. Then,

$$\mathcal{R}(\boldsymbol{p})_i = p_i^2 + 2p_i \sum_{\ell\geq 1} \frac{p_{i+\ell}}{1+\ell} = p_i\biggl(p_i + p_{i+1} + \sum_{\ell\geq 2} \frac{2}{\ell+1} p_{i+\ell}\biggr) \leq p_i,$$

where the last step follows since $\frac{2}{\ell+1} < 1$ in the last sum. One has equality precisely when $p_k = 0$ for all $k \geq i+2$. This implies $\mathfrak{m} < \infty$ and the uniqueness of $\boldsymbol{p}$ (given $\mathfrak{m}$) with the non-zero frequencies as claimed. ∎

In this case, one may select a compact subset within the probability vectors by demanding the existence of the centred r-th moment, for some fixed $r > 1$. More precisely, with

$$\mu_s(\boldsymbol{p}) := \sum_{\ell \geq 0} |\ell - \mathfrak{m}|^s p_\ell,$$

one considers the set

$$\mathcal{M}^+_{1,\mathfrak{m},C} := \{\boldsymbol{p} \in \mathcal{M}^+_1 \mid \textstyle\sum_k k p_k = \mathfrak{m} \text{ and } \mu_r(\boldsymbol{p}) \leq C\} \tag{21}$$

for an arbitrary, but fixed $C < \infty$, equipped with our usual metric as introduced before in (1). This gives a compact and convex space [19, Lemma 2]. Moreover, one has

Lemma 4. *Let $r > 1$ be fixed and consider the space $\mathcal{M}^+_{1,\mathfrak{m},C}$ of (21). Then, both μ_1 and μ_r satisfy*

$$\mu_s(\mathcal{R}_0(\boldsymbol{p})) \leq \mu_s(\boldsymbol{p}),$$

with equality if and only if $\boldsymbol{p}$ is a fixed point of $\mathcal{R}_0$.

Moreover, $\mu_1 : \mathcal{M}^+_{1,\mathfrak{m},C} \longrightarrow \mathbb{R}_{\geq 0}$ is continuous and defines a Lyapunov function for the dynamics in continuous time.

Proof. To show the first claim, consider

$$\begin{aligned}\mu_s(\mathcal{R}_0(\boldsymbol{p})) &= \sum_{i \geq 0} \sum_{\substack{k,\ell \geq 0 \\ k\wedge\ell \leq i \leq k\vee\ell}} \frac{|i - \mathfrak{m}|^s}{1 + |k-\ell|} p_k p_\ell \\ &= \sum_{k,\ell \geq 0} \frac{p_k p_\ell}{1 + |k-\ell|} \frac{1}{2} \sum_{i=k\wedge\ell}^{k\vee\ell} (|i - \mathfrak{m}|^s + |k + \ell - i - \mathfrak{m}|^s).\end{aligned} \tag{22}$$

For notational convenience, let $j = k + \ell - i$. We now show

$$|i - \mathfrak{m}|^s + |k + \ell - i - \mathfrak{m}|^s \leq |k - \mathfrak{m}|^s + |\ell - \mathfrak{m}|^s. \tag{23}$$

If $\{k, \ell\} = \{i, j\}$, then (23) holds with equality. Otherwise, assume without loss of generality that $k < i \leq j < \ell$. If $\mathfrak{m} \leq k$ or $\mathfrak{m} \geq \ell$, we have equality for $s = 1$, but a strict inequality for $s = r$ due to the convexity of $x \mapsto x^r$. (For $s = 1$, this describes the fact that a recombination event between two sequences that are both longer or both shorter than the mean does not change their averaged distance to the mean copy number.) In the remaining cases, the inequality is strict as well. Hence, $\mu_s(\mathcal{R}_0(\boldsymbol{p})) \leq \mu_s(\boldsymbol{p})$ with equality if and only if $\boldsymbol{p}$ is a fixed point of $\mathcal{R}_0$, since otherwise the sum in (22) contains at least one term for which (23) holds as a strict inequality.

To see that μ_1 is continuous, consider a convergent sequence $(\boldsymbol{p}^{(n)})$ in $\mathcal{M}^+_{1,\mathfrak{m},C}$ and the random variables $H^{(n)} = |K^{(n)} - \mathfrak{m}|$, where the $K^{(n)}$ are independent $\mathbb{N}_0$-valued random variables with laws $\boldsymbol{p}^{(n)}$. Due to the structure of $\mathcal{M}^+_{1,\mathfrak{m},C}$, the random variables $H^{(n)}$ are uniformly integrable, which implies the convergence of the corresponding expectation values by [5, Thm. 25.12]. This, in turn, is nothing but the continuity of μ_1. Since $\mu_1(\boldsymbol{p})$ is linear in $\boldsymbol{p}$ and thus infinitely differentiable, so is the solution $\boldsymbol{p}(t)$ for every initial condition $\boldsymbol{p}_0 \in \mathcal{M}^+_{1,\mathfrak{m},C}$, compare [1, Thm. 9.5 and Remark 9.6(b)]. Therefore, we have

$$\dot{\mu}_1(\boldsymbol{p}_0) = \liminf_{t \to 0^+} \frac{\mu_1(\boldsymbol{p}(t)) - \mu_1(\boldsymbol{p}_0)}{t} = \mu_1(\mathcal{R}_0(\boldsymbol{p}_0)) - \mu_1(\boldsymbol{p}_0) \leq 0,$$

again with equality if and only if $\boldsymbol{p}_0$ is a fixed point. Thus, μ_1 is a Lyapunov function as claimed. ∎

Finally, this gives the following convergence result, the proof of which is given in [19] and not repeated here.

THEOREM 2. *Assume that, for the initial condition* $\boldsymbol{p}(0)$ *and fixed* $r > 1$*, the* r*-th moment exists,* $\mu_r(\boldsymbol{p}) < \infty$*. Then,* $\mathfrak{m} = \sum_\ell \ell p_\ell$ *is finite and, both in discrete and in continuous time,* $\lim_{t\to\infty} \|\boldsymbol{p}(t) - \boldsymbol{p}\|_1 = 0$ *with the appropriate fixed point* $\boldsymbol{p}$ *from Proposition* 6. ∎

Let us mention that, for $q = 0$, the recombinator can be expressed in terms of explicit frequencies $\pi_{k,\ell}$ of fragment pairs before concatenation (with copy numbers k and ℓ) as $\mathcal{R}_0(\boldsymbol{p})_i = \sum_{j=0}^{i} \pi_{j,i-j}$. It is as yet an open question whether this can be used to simplify the above treatment.

7. Random crossover. This model deviates from the previous one in that it admits arbitrary overhangs, up to the case where, after the crossover, one sequence got it all while the other lost everything. The possible alignments for any pair are supposed to be equally likely, so that the recombinator (2), again restricted to the probability measures, now reads

$$\mathcal{R}_1(\boldsymbol{p})_i = \sum_{\substack{k,\ell\geq 0\\ k+\ell\geq i}} \frac{1+\min\{k,\ell,i,k+\ell-i\}}{(k+1)(\ell+1)}\, p_k\, p_\ell. \tag{24}$$

As for our previous two examples, using Lemma 2 once again, the reversibility condition,

$$\frac{p_k}{k+1}\frac{p_\ell}{\ell+1} = \frac{p_i}{i+1}\frac{p_j}{j+1}, \quad \text{for all } k+\ell = i+j,$$

leads to an expression for fixed points. In fact, these relations have $p_k = C(k+1)x^k$ as a solution, with appropriate parameter x and normalisation constant C. Again, it turns out that all fixed points are given this way, as was originally noticed (in a different way) in [23, Thm. A.2].

PROPOSITION 7. *Every fixed point* $\boldsymbol{p} \in \mathcal{M}_1^+$ *of* $\mathcal{R}_1$ *has finite mean* $\mathfrak{m} = \sum_k k p_k$*, and is uniquely specified by the value of* $\mathfrak{m}$*. Explicitly, one has*

$$p_k = \left(\frac{2}{\mathfrak{m}+2}\right)^2 (k+1)\left(\frac{\mathfrak{m}}{\mathfrak{m}+2}\right)^k$$

with $k \in \mathbb{N}_0$. ∎

One can verify this in several ways, one being a direct calculation via induction. Interestingly, this equilibrium is the convolution of two geometric distributions (of equal mean $\mathfrak{m}/2$), and hence also of geometric type according to our terminology (which follows that of [23]). It might be interesting to explore this observation a little further in the future.

At this point, one can define, very much in analogy to the situation in Takahata's model above, an induced recombinator, $\widetilde{\mathcal{R}}_1$, acting once more on spaces of the form $X_{\alpha,\delta}$. It is given as

$$\widetilde{\mathcal{R}}_1(\boldsymbol{p}) = a\bigl(\mathcal{R}_1(\boldsymbol{p})\bigr)$$

where the mapping a is defined by

$$a(\boldsymbol{p})_k = \frac{1}{k+1} \sum_{\ell \geq k} \binom{\ell}{k} p_\ell = \frac{1}{k+1}\, b(\boldsymbol{p})_k .$$

It is thus closely related to our above mapping b.

The main result on this model, proved in detail in [19, 18], reads as follows.

Theorem 3. *Assume that* $\limsup_{k\to\infty} \sqrt[k]{p_k(0)} < 1$. *Then, both in discrete and in continuous time,* $\lim_{t\to\infty} \|\boldsymbol{p}(t) - \boldsymbol{p}\|_1 = 0$, *where* $\boldsymbol{p}$ *is the corresponding fixed point according to Proposition* 7.

Proof. The proof is very similar to the one used above for the Takahata model, and employs once again Banach's contraction principle for the induced action of $\widetilde{\mathcal{R}}_1$ on $X_{\alpha,\delta}$. Since all details have been given in [19], we omit them here. ∎

The fixed points of Proposition 7 are of the expected geometric type, and are perhaps more realistic than those of the Takahata model, at least for cases where a maximum is present in the repeat distribution. However, one should note that the experimental situation is not completely convincing at present, so that it seems advantageous to have a versatile model class at hand.

8. An interpolation. When considering the recombinators $\mathcal{R}_0$ and $\mathcal{R}_1$ in comparison, one would like to find further models that share properties of both of them, or interpolate between them in a suitable way. In particular, $\mathcal{R}_0$ is unrealistic due to the complete confinement of the shorter bit within the range of the longer one, while $\mathcal{R}_1$ poses no restriction at all for any kind of overhang. One such interpolation was initially investigated in discrete time by Atteson and Shpak in [23], based on preceding work by Ohta [17] and Walsh [26], see also [19, 18] for more. The interpolation employs a penalty function idea for overhangs of the shorter sequence, and leads (in the above language) to a recombinator $\mathcal{R}_q$ with $0 \leq q \leq 1$. The latter is based upon the transition probabilities

$$T^{(q)}_{ij,k\ell} = C^{(q)}_{k\ell}\, \delta_{i+j,k+\ell}\, (1 + \min\{k,\ell,i,j\})\, q^{0 \vee (k\wedge\ell - i\wedge j)}, \tag{25}$$

where $k \vee \ell := \max\{k,\ell\}$, $k \wedge \ell := \min\{k,\ell\}$, and $0^0 = 1$. The normalisation constants $C^{(q)}_{k\ell}$ are chosen such that (3) holds, i.e., $\sum_{i,j\geq 0} T^{(q)}_{ij,k\ell} = 1$. These constants are symmetric in k and ℓ and read explicitly

$$C^{(q)}_{k\ell} = \frac{(1-q)^2}{(k\wedge\ell+1)(|k-\ell|+1)(1-q)^2 + 2q(k\wedge\ell - (k\wedge\ell+1)q + q^{k\wedge\ell+1})} .$$

Note further that the total number of units is indeed conserved in each event and that the process is symmetric within both pairs. Hence (7) is satisfied.

Unfortunately, the situation with the fixed points is a lot more complicated due to the following result.

Proposition 8. *For parameter values* $q \in (0,1)$, *any fixed point* $\boldsymbol{p} \in \mathcal{M}_1^+$ *of the recombinator* $\mathcal{R}_q$, *given by* (2) *and* (25), *satisfies* $p_k > 0$ *for all* $k \geq 0$ (*unless it is the trivial fixed point* $\boldsymbol{p} = (1,0,0,\ldots)$ *we excluded*). *None of these extra fixed points is reversible.*

Proof. Let a non-trivial fixed point $\boldsymbol{p}$ be given and choose any $n > 0$ with $p_n > 0$. Observe that $T^{(q)}_{n+1\; n-1,nn} > 0$ for $0 < q < 1$ and hence

$$p_{n\pm1} = \mathcal{R}_q(\boldsymbol{p})_{n\pm1} = \sum_{j,k,\ell\geq 0} T^{(q)}_{n\pm1\; j,k\ell} p_k\, p_\ell \;\geq\; T^{(q)}_{n+1\; n-1,nn}\, p_n\, p_n > 0.$$

The first statement now follows by induction. For the second statement, evaluate the reversibility condition (13) for all combinations of i, j, k, ℓ with $i + j = k + \ell \leq 4$. This leads to four independent equations. Three of them can be transformed to the recursion

$$p_k = \frac{(k+1)q}{2(k-1)+2q}\,\frac{p_1}{p_0}\, p_{k-1}, \qquad k \in \{2, 3, 4\},$$

from which one derives explicit equations for all p_k with $k \in \{2, 3, 4\}$ in terms of p_0 and p_1. Inserting the one for p_2 into the remaining equation yields another equation for p_4 in terms of p_0 and p_1, which contradicts the first equation for all $q \in (0, 1)$, as is easily verified. ■

Nevertheless, the dynamics is well defined, and respects the compact subsets defined above in forward time, compare [19, Thm. 4]. Based upon the analysis in [18, 19], and further numerical work on the fixed points, it is plausible that, given the mean copy number $\mathfrak{m}$, never more than one fixed point for $\mathcal{R}_q$ exists. Due to the global convergence results at $q = 0$ and $q = 1$, any non-uniqueness in the vicinity of these parameter values could only come from a bifurcation, not from an independent source. Numerical investigations indicate that no bifurcation is present, but this needs to be analysed further.

Moreover, the Lipschitz constant for the corresponding induced recombinator $\widetilde{\mathcal{R}}_q$ can be expected to be continuous in the parameter q, hence to remain strictly less than 1 on the sets $X_{\alpha,\delta}$ in the neighbourhood of $q = 1$. So, at least locally, the contraction property should be preserved. For further progress, it seems advantageous [11] to use a rather different approach based on the analysis of similar problems in evolutionary game theory. Here, one would aim to establish a slightly weaker type of convergence result for all $0 < q < 1$, and probably even on the larger compact set $\mathcal{M}^{+}_{1,\mathfrak{m},C}$ from Eq. (21).

9. Open problems and outlook. The results for the various models presented here show that initial configurations, subject to some specific conditions that are no restriction in practice, converge to one of the known fixed points. These results apply to the deterministic dynamics of the infinite population limit.

In view of the biological applications, one is also interested in possible deviations from this picture on the level of large, but finite, populations, i.e., for the underlying stochastic process, e.g., a variant of the Moran model with unequal crossover. In this model class, however, important deviations seem unlikely, due to the known convergence results for the infinite population limit, see [2] for more.

Since the above equilibrium distributions have finite support or are exponentially small for large copy numbers, one can also expect these systems to behave very much like ones with only finitely many types. In this sense, the results are typical, and the more general setting with probability vectors on $\mathbb{N}_0$ is adequate. This is also supported by several simulations [18].

Still, an open question is a more complete understanding of the regime $q \in (0,1)$ in Section 8. Due to the loss of reversibility of the fixed points, the analysis becomes rather involved. Preliminary investigations [18] have not given any hint on values of q where convergence fails or where alternative stable fixed points show up, though this is presently only based on numerical experiments and perturbative arguments. It might be advantageous (and perhaps also more realistic) to search for other ways to interpolate between the cases $q = 0$ and $q = 1$, preferably ones that maintain the reversibility of the equilibria. This question certainly deserves further attention.

References

[1] H. Amann, *Ordinary Differential Equations*, de Gruyter, Berlin, 1990.

[2] E. Baake and I. Hildebrandt, *Single-crossover dynamics: finite versus infinite populations*, Bull. Math. Biol. 70 (2008), 603–624.

[3] D. Bachtrog, S. Weiss, B. Zangerl, G. Brem and C. Schlötterer, *Distribution of dinucleotide microsatellites in the* Drosophila melanogaster *genome*, Mol. Biol. Evol. 16 (1999), 602–610.

[4] C. Bank, *Diskrete Rekombinationsdynamik für repetitive Strukturen*, Diplomarbeit, Univ. Bielefeld, 2007.

[5] P. Billingsley, *Probability and Measure*, 3rd ed., Wiley, New York, 1995.

[6] P. Billingsley, *Convergence of Probability Measures*, 2nd ed., Wiley, New York, 1999.

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

[8] C. D. Bustamante, *Population genetics of molecular evolution*, in [16], pp. 63–99.

[9] P. Calabrese and R. Sainudiin, *Models of microsatellite evolution*, in [16], pp. 289–305.

[10] D. Graur and W.-H. Li, *Fundamentals of Molecular Evolution*, 2nd ed., Sinauer, Sunderland, 2000.

[11] J. Hofbauer, private communication, 2003.

[12] B. Kitchens, *Symbolic Dynamics—One-sided, Two-sided and Countable State Markov Chains*, Springer, Berlin, 1998.

[13] S. Kruglyak, R. T. Durrett, M. D. Schug and C. F. Aquadro, *Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations*, Proc. Natl. Acad. Sci. USA 95 (1998), 10774–10778.

[14] M. Lachowicz, *General population systems. Macroscopic limit of a class of stochastic semigroups*, J. Math. Anal. Appl. 307 (2005), 585–605.

[15] M. Lachowicz, *Micro and meso scales of description corresponding to a model of tissue invasion by solid tumors*, Math. Models Meth. Appl. Sci. 15 (2005), 1667–1683.

[16] R. Nielsen (ed), *Statistical Methods in Molecular Evolution*, Springer, New York, 2005.

[17] T. Ohta, *On the evolution of multigene families*, Theor. Pop. Biol. 23 (1983), 216–240.

[18] O. Redner, *Models for mutation, selection, and recombination in infinite populations*, Dissertation, Univ. Greifswald; Shaker, Aachen, 2003.

[19] O. Redner and M. Baake, *Unequal crossover dynamics in discrete and continuous time*, J. Math. Biol. 49 (2004), 201–226; arXiv:math.DS/0402351.

[20] M. Reed and B. Simon, *Methods of Modern Mathematical Physics I: Functional Analysis*, Academic Press, San Diego, 1980.

[21] E. Seneta, *Non-negative Matrices and Markov Chains*, rev. printing, Springer, New York, 2006.

[22] A. N. Shiryaev, *Probability*, 2nd ed., Springer, New York, 1996.

[23] M. Shpak and K. Atteson, *A survey of unequal crossover systems and their mathematical properties*, Bull. Math. Biol. 64 (2002), 703–746.

[24] N. Takahata, *A mathematical study on the distribution of the number of repeated genes per chromosome*, Genet. Res. 38 (1981), 97–102.

[25] J. Wakeley, *Coalescent Theory: An Introduction*, Roberts and Company, Greenwood Village, CO, 2008.

[26] J. B. Walsh, *Persistence of tandem arrays: Implications for satellite and simple-sequence DNAs*, Genetics 115 (1987), 553–567.

[27] K. Yosida, *Functional Analysis*, 6th ed., Springer, Berlin, 1980.