J. Math. Biol. (2010) 61:455–473
DOI 10.1007/s00285-009-0308-1

Mathematical Biology

Asymptotic behavior of a Moran model with mutations, drift and recombination among multiple loci

Adam Bobrowski · Tomasz Wojdyła · Marek Kimmel

Received: 19 November 2008 / Revised: 22 October 2009 / Published online: 11 November 2009

Abstract In this paper, we extend the theoretical treatment of the Moran model of genetic drift with recombination and mutation, which was previously introduced by us for the case of two loci, to the case of n loci. Recombination, when considered in the Wright–Fisher model, makes it considerably less tractable. In the works of Griffiths, Hudson and Kaplan and their colleagues important properties were established using the coalescent approach. Other more recent approaches form a body of work to which we would like to contribute. The specific framework used in our paper allows finding close-form relationships, which however are limited to a set of distributions, which jointly characterize allelic states at a number of loci at the same or different chromosome(s) but which do not jointly characterize allelic states at a single locus on two or more chromosomes. However, the system is sufficiently rich to allow computing, albeit

This research was supported by the Polish Government research fund for years 2005–2008, Grant 1 P03A 044 29 (0356/P03/2005/29).

A. Bobrowski (✉)
Department of Mathematics, Faculty of Electrical Engineering,
Technical University of Lublin, Nadbystrzycka 38A, 20-618 Lublin, Poland
e-mail: a.bobrowski@pollub.pl

T. Wojdyła
Institute of Computer Science, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland
e-mail: tpwojdyla@polsl.pl

M. Kimmel
Department of Statistics, Rice University, P. O. Box 1892, Houston, TX 77251, USA
e-mail: kimmel@stat.rice.edu

M. Kimmel
Systems Engineering Group, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland

in general numerically, all possible multipoint linkage disequilibria under recombination, mutation and drift. We explore the algorithms enabling construction of the transition probability matrices of the Markov chain describing the process. We find that asymptotically the effects of recombination become indistinguishable, at least as characterized by the set of distributions we consider, from the effects of mutation and drift. Mathematically, the results are based on the foundations of the theory of semigroups of operators. This approach allows generalization to any Markov-type mutation model. Based on these fundamental results, we explore the rates of convergence to the limit distribution, using Dobrushin's coefficient and spectral gap.

Keywords Recombination · Moran model · Markov chain · Mutation · Genetic drift

Mathematics Subject Classification (2000) 92D25 · 60J05 · 60J35

1 Introduction

The Wright–Fisher model (Ewens 2004) provides an established simplified description of the dynamics of populations of individuals endowed with genomes, under the action of genetic drift, mutation, selection and recombination, including various demographic scenarios such as bottlenecks, expansions and migrations. Even in the simplified framework, the number of possible scenarios is enormous and many of them pose almost intractable difficulties. The essential reason of these complications is that the Wright–Fisher model is a Markov chain with an enormous state space. For example, if there are $2N$ genomes involved, each of them with K variants, the dimension of the chain is equal to K^{2N}, which is unmanageable even for modest $2N$ and K values. These difficulties gave rise to a large number of approximate or asymptotic mathematical techniques, usually assuming large population size N. Of these techniques, the best developed are diffusion approximations in forward time and coalescent processes in reverse time.

Recombination, when considered as one of the genetic forces active in the Wright–Fisher model, makes it considerably less tractable. In the framework of coalescence, there exist effective computational techniques to trace the ancestry of a sample, when recombination is considered. In the works of Griffiths such as Griffiths (1981) and other authors, important properties were established, such as the distribution of the size of the ancestral graph in the coalescent with recombination. The important result concerning the correlation of the times to the most recent common ancestors of two recombining loci, has been found by Hudson (1983), based on results obtained by Griffiths (1981). Last but not least, in forward time, it is now possible to efficiently simulate the Wright–Fisher model with recombination (Kimmel and Peng 2005) even for large-scale problems.

There has been recently at least three monographs, which thoroughly explain recombination, mostly in the context of the coalescent in reverse time. These are Durret (2002), Hein et al. (2006) and Wakeley (2008). These monographs also have large reference lists. From the large quantity of recently published papers, we may idiosyncratically quote Barton et al. (2004) and Baake and Herms (2008).

However, there are some aspects of population genetics under recombination that are still to be clarified. One of them is the question of identifiability, i.e., if the population can reach a stage at which it is indistinguishable from the population evolving solely under drift and mutation. In this paper we approach and solve this problem using a mathematical relative of the Wright–Fisher model, known as the Moran model (Ewens 2004). Our approach is mostly forward-time.

The simplest version of our model, which assumes only two recombining loci per haploid individual (genome), has been previously published in Bobrowski and Kimmel (2003) and Kimmel and Polańska (1999). Therefore, we depart here from the next least complicated case, i.e., that of three loci, before proceeding to the general case of n loci. We derive semigroups of operators describing the evolution of the distributions characterizing the recombinant status of all loci in each of the individuals in the sample. This leads to a characterization of the limit distributions of the process.

2 The case of three loci

We start with the description of our model in the case of three loci, where more explicit formulas are available; for the case of two loci see Bobrowski and Kimmel (2003).

2.1 Mathematical notations

Throughout the paper $\mathbb{A}$ is a countable set of allele types: typically it is $\mathbb{Z}$ or one of its subsets. l^1 stands for the space of absolutely summable sequences $(\xi_i)_{i\in\mathbb{A}}$, with the norm $\|(\xi_i)_{i\in\mathbb{A}}\| = \sum_{i\in\mathbb{A}} |\xi_i|$. The elements $e_j = (\delta_{ij})_{i\in\mathbb{A}}$, $j \in \mathbb{A}$, where δ_{ij} is the Kronecker delta, form a basis of l^1; any $(\xi_i)_{i\in\mathbb{A}} \in l^1$ may be represented as $(\xi_i)_{i\in\mathbb{A}} = \sum_{i\in\mathbb{A}} \xi_i e_i$. Vector $(\xi_i)_{i\in\mathbb{A}} \in l^1$ is said to be a distribution iff $\xi_i \geq 0$, $i \in \mathbb{A}$, and $\sum_{i\in\mathbb{A}} \xi_i = 1$. $\mathcal{M}_n$ where n is an integer, is the space of absolutely summable n-dimensional matrices $\mathrm{m} = (\mu_{i_1,\ldots,i_n})_{i_1,\ldots,i_n\in\mathbb{A}}$ with the norm $\sum_{i_1,\ldots,i_n\in\mathbb{A}} |\mu_{i_1,\ldots,i_n}|$. A matrix $\mathrm{m} \in \mathcal{M}_n$ is termed a distribution iff its entries are non-negative and add up to 1. Distributions in $\mathcal{M}_n$ are distributions of n-tuples of $\mathbb{A}$-valued random variables.

$\mathcal{M}_n$ may be viewed as a tensor product of n copies of l^1 : $\mathcal{M}_n = (l^1)^{n\otimes}$; the tensor product of $(\xi_{i,j})_{i\in\mathbb{A}} \in l^1$, $j = 1,\ldots,n$, is $(\xi_{i_1,1}\xi_{i_2,2}\cdots\xi_{i_n,n})_{i_1,\ldots,i_n\in\mathbb{A}}$. More specifically, $\mathcal{M}_n$ is isometrically isomorphic to the completion of the algebraic tensor product with respect to the projective norm (Defant and Floret 1993; Ryan 2002).

If $A_i, i = 1,\ldots,n$ are bounded linear operators in l^1, then their tensor product is an operator in $\mathcal{M}_n$ defined as $A_1 \otimes \cdots \otimes A_n(\mu_{i_1,\ldots,i_n})_{i_1,\ldots,i_n\in\mathbb{A}} = \sum_{i_1,\ldots,i_n\in\mathbb{A}} \mu_{i_1,\ldots,i_n} A_1 e_{i_1} \otimes A_2 e_{i_2} \otimes \cdots \otimes A_n e_{i_n}$. We have, $\|A_1 \otimes \cdots \otimes A_n\| = \prod_{i=1}^n \|A_i\|$.

A family of operators $\{S(t), t \geq 0\}$ in a Banach space $\mathbb{X}$ is said to be a strongly continuous semigroup (Engel and Nagel 2000) iff $S(t)S(s) = S(t+s)$, $S(0)$ is the identity operator, and $\lim_{t\to 0} S(t)x = x$ (strongly, i.e., in the $\mathbb{X}$-space norm), for all $x \in \mathbb{X}$. A strongly continuous semigroup in l^1 is termed a Markov semigroup iff all $S(t)$ map distributions into distributions; in particular we must have $\|S(t)\| = 1$. If $\{S_i(t), t \geq 0\}, i = 1,\ldots,n$ are strongly continuous semigroups of Markov operators

in l^1, then $\{S_1(t) \otimes \cdots \otimes S_n(t), t \geq 0\}$ is a strongly continuous Markov semigroup in $\mathcal{M}_n$. This semigroup is called the tensor product of $\{S_i(t), t \geq 0\}$, $i = 1, \ldots, n$.

The Cartesian product $\mathcal{M}_n^m$ of m copies of $\mathcal{M}_n$ provides a way of gathering information on distributions of m n-tuples of $\mathbb{A}$-valued variables. This space may be seen as a direct sum of its m subspaces, with the jth of them being the subspace of m-tuples $(m_i)_{i=1,\ldots,m}$ where all but m_j are zero. On the other hand, all these subspaces may be identified with $\mathcal{M}_n$, and so we may see $\mathcal{M}_n^m$ as a direct sum of m copies of $\mathcal{M}_n$. The norm in this space is $\|(m_i)_{i=1,\ldots,m}\| = \sum_{i=1}^m \|m_i\|_{\mathcal{M}_n}$. We say that $x \in \mathcal{M}_n^m$ is a distribution iff it is a convex combination of m distributions in $\mathcal{M}_n$. A Markov operator in $\mathcal{M}_n^m$ is an operator mapping distributions into distributions. If $\{T_i(t), t \geq 0\}, i = 1, \ldots, m$ are Markov semigroups in $\mathcal{M}_n$, then $\{\bigoplus_{i=1}^m T_i(t), t \geq 0\}$, defined as $\bigoplus_{i=1}^m T_i(t)(\sum_{i=1}^c m_i) = \sum_{i=1}^m T_i(t)m_i$ is a Markov semigroup in $\mathcal{M}_n^m$. The domain of the infinitesimal generator $\mathcal{G}$ of this semigroup is the Cartesian product of the domains of the generators G_i of $\{T_i(t), t \geq 0\}$ and we have $\mathcal{G}(m_i)_{i=1,\ldots,m} = (G_i m_i)_{i=1,\ldots,m}$ for $(m_i)_{i=1,\ldots,m}$ in this domain.

2.2 The model

Consider a population of $2N$ individuals. Each individual is represented as a triple of $\mathbb{A}$-valued random variables (where $\mathbb{A}$ is the set of allelic types, see Sect. 2.1) describing three loci on a chromosome in linear order; the ith individual being a triple (X_i, Y_i, Z_i), $i = 1, \ldots, 2N$. We assume that these triples are exchangeable, that each of them evolves in time as a triple of independent, non-explosive Markov chains, independent of the other ones, but with the same transition probabilities; this models mutation at three loci of a chromosome in each individual. The process of mutation at the first locus in each individual is described by means of a strongly continuous semigroup $\{S_X(t), t \geq 0\}$ of Markov operators in l^1. This means that if $x \in l^1$ is the distribution of allele types at time 0 then $S_X(t)x$ is the distribution of allele types at time t. The process of mutation on the second locus is governed by a semigroup $\{S_Y(t), t \geq 0\}$, and the process of mutation on the third locus is governed by a semigroup $\{S_Z(t), t \geq 0\}$. The tensor product semigroup $\{S(t), t \geq 0\}$, $S(t) = S_X(t) \otimes S_Y(t) \otimes S_Z(t)$, describes evolution of distributions at three loci, provided mutations at these loci occur independently.

We also incorporate recombination and genetic drift in the model by assuming that each individual's life-length is an exponential random variable with parameter $\lambda/2$ and that at the moment of an individual's death, the triple by which it is represented is replaced by another triple in the following manner. With probability $1 - r$, where $r = r_1 + r_2$ with $r_1, r_2 \in [0, 1]$ such that $r \in (0, 1]$, are given parameters, one of the triples (X_j, Y_j, Z_j), $j = 1, \ldots, 2N$ is drawn at random to replace the deceased one. With probability r_1 the X-variable is drawn at random first and the par (Y, Z) next, independently of the result of the first draw (this models recombination after the first locus). With probability r_2, the pair (X, Y) is drawn first and the variable Z next, independently of the first draw (this models recombination after the second locus). As a consequence, a new triple becomes one of the already existing triples (including the one just deceased) (X_j, Y_j, Z_j), each of them with probability $(1-r)\frac{1}{2N} + r\frac{1}{(2N)^2}$, or

one of the two types of "mixed ones": either (X_j, Y_k, Y_k), $j \neq k$ each with probability $r_1 \frac{1}{(2N)^2}$ or (X_j, Y_j, Z_k), $j \neq k$ each with probability $r_2 \frac{1}{(2N)^2}$.

If the triples (X_i, Y_i, Z_i) are exchangeable, then, because of the sampling scheme, it is obvious that so are the newly formed triples immediately after an individual's death. This fact follows from Lemma 1 in Bobrowski and Kimmel (2003) if we note that either the pair (X, Y) or the pair (Y, Z) can be treated as a single compound locus. Therefore, exchangeability is preserved in the model.

2.3 Relations between partial distributions of the population immediately before and immediately after an individual's death

Let (X_a, Y_a, Z_a) and $(\tilde{X}_a, \tilde{Y}_a, \tilde{Z}_a)$, $a = 1, \ldots, 2N$ be the triples representing individuals in the population immediately before and immediately after an individual's death. By exchangeability of (X_a, Y_a, Z_a), $a \in 1, \ldots, 2N$, the distribution of (X_a, Y_a, Z_b) where $a \neq b$ does not depend on the choice of a and b; we will denote it by D_{112}. The same is true of the distributions of (X_a, Y_a, Z_a), (X_a, Y_b, Z_a), (X_a, Y_b, Z_b) and (X_a, Y_b, Z_c) where a, b and c are distinct numbers; we denote these distributions by D_{111}, D_{121}, D_{122} and D_{123}, respectively. The corresponding Ds with tilde denote distributions in the population immediately after an individual's death.

As we shall see shortly, all $\tilde{D}$s are convex combinations of Ds, so that there exists a transition matrix Θ of a Markov chain such that

$$\tilde{D} = \Theta D, \tag{1}$$

where $\tilde{D}$ and D are the (column-)vectors with coordinates $\tilde{D}_{111}$, $\tilde{D}_{112}$, $\tilde{D}_{121}$, $\tilde{D}_{122}$, $\tilde{D}_{123}$ and D_{111}, D_{112}, D_{121}, D_{122}, D_{123}, respectively (note the lexicographic order). We will write Θ as a convex combination

$$\Theta = (1 - r)\Theta_0 + r_1\Theta_1 + r_2\Theta_2 \tag{2}$$

of three transition matrices, corresponding to the cases of no recombination, and of recombination after the first and after the second locus, respectively.

To this end, we note that if there was no recombination, none of D_{112}, D_{121} and D_{122} has changed unless $i = 1, j = 2$ or $i = 2, j = 1$. In this last case $\tilde{D}_{112} = \tilde{D}_{121} = \tilde{D}_{122} = D_{111}$. Hence,

$$\Theta_0 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{2}{(2N)^2} & 1 - \frac{2}{(2N)^2} & 0 & 0 & 0 \\ \frac{2}{(2N)^2} & 0 & 1 - \frac{2}{(2N)^2} & 0 & 0 \\ \frac{2}{(2N)^2} & 0 & 0 & 1 - \frac{2}{(2N)^2} & 0 \\ 0 & \frac{2}{(2N)^2} & \frac{2}{(2N)^2} & \frac{2}{(2N)^2} & 1 - \frac{6}{(2N)^2} \end{pmatrix}, \tag{3}$$

where the form of the last row is justified as follows. If $i = 1, j = 2$, then $(\tilde{X}_1, \tilde{Y}_2, \tilde{Z}_3) = (X_2, Y_2, Z_3)$ and so $\tilde{D}_{123} = D_{112}$; similarly we show that this equality

Table 1 Calculation of Θ_1

	$i \neq 1, 2$	$i = 1$, $j \neq 2$, $k \neq 2$	$i = 1$, $j = 2$, $k \neq 2$	$i = 1$, $j = 2$, $k = 2$	$i = 1$, $j \neq 2$, $k = 2$	$i = 2$, $k \neq 1$	$i = 2$, $k = 1$
$\tilde{D}_{112}$	D_{112}	$D_{112}, j = k$ $D_{123}, j \neq k$	D_{121}	D_{111}	D_{122}	D_{112}	D_{111}
$\tilde{D}_{121}$	D_{121}	$D_{121}, j = k$ $D_{123}, j \neq k$	D_{112}	D_{111}	D_{122}	D_{121}	D_{111}
$\tilde{D}_{122}$	D_{122}	D_{122}	D_{111}	D_{111}	D_{122}	D_{122}	D_{111}

If $i = 1$ and recombination took place after the first locus, $(\tilde{X}_1, \tilde{Y}_1, \tilde{Z}_1) = (X_j, Y_k, Z_k)$ and so $(\tilde{X}_1, \tilde{Y}_2, \tilde{Z}_1) = (X_j, Y_2, Z_k)$. Considering all possible cases for j and k we obtain four entries in the middle row of table. The remaining entries in the table are obtained similarly

is true when $i = 2$ and $j = 1$. Analogously, $\tilde{D}_{123} = D_{122}$ if either $i = 2, j = 3$ or $i = 3, j = 2$, and $\tilde{D}_{123} = D_{121}$ if either $i = 3, j = 1$ or $i = 1, j = 3$. In the remaining cases $\tilde{D}_{123} = D_{123}$.

To find the three rows in the middle of Θ_1 we consider recombination between the first two loci, by listing the possible cases in Table 1. This gives Θ_1 in the form:

$$\begin{pmatrix} 1-\frac{2N-1}{(2N)^2} & 0 & 0 & \frac{2N-1}{(2N)^2} & 0 \\ \frac{2N+1}{(2N)^3} & \frac{2N-2}{2N}+\frac{2N-1}{(2N)^3}+\frac{2N-1}{(2N)^2} & \frac{2N-1}{(2N)^3} & \frac{2N-1}{(2N)^3} & \frac{(2N-1)(2N-2)}{(2N)^3} \\ \frac{2N+1}{(2N)^3} & \frac{2N-1}{(2N)^3} & \frac{2N-2}{2N}+\frac{2N-1}{(2N)^3}+\frac{2N-1}{(2N)^2} & \frac{2N-1}{(2N)^3} & \frac{(2N-1)(2N-2)}{(2N)^3} \\ \frac{2}{(2N)^2} & 0 & 0 & 1-\frac{2}{(2N)^2} & 0 \\ 0 & \frac{2}{(2N)^2} & \frac{2}{(2N)^2} & \frac{2}{(2N)^2} & 1-\frac{6}{(2N)^2} \end{pmatrix}. \quad (4)$$

Obtaining the first row here is straightforward, and the last row is obtained by noting that: (a) for $i = 1$, $\tilde{D}_{123} = D_{112}$ provided $j = 2$, $\tilde{D}_{123} = D_{121}$ provided $j = 3$, and $\tilde{D}_{123} = D_{123}$ in the remaining cases, (b) for $i = 2$, $\tilde{D}_{123} = D_{112}$ provided $k = 1$, $\tilde{D}_{123} = D_{122}$ provided $k = 3$, and $\tilde{D}_{123} = D_{123}$ in the remaining cases, (c) for $i = 3$, $\tilde{D}_{123} = D_{121}$ provided $k = 1$, $\tilde{D}_{123} = D_{122}$ provided $k = 2$, and $\tilde{D}_{123} = D_{123}$ in the remaining cases, and (d) for $i \geq 4$, $\tilde{D}_{123} = D_{123}$.

To cover the case of recombination after the second locus we note that our model is symmetric with respect to numbering loci. More specifically, if the loci were numbered from the last one to the first, the distributions D_{111}, D_{112}, D_{121}, D_{122}, D_{123} would have become D_{111}, D_{122}, D_{121}, D_{112}, D_{123}, which amounts to transposition of D_{112} and D_{122}. Since such a numbering transposes recombination loci, the matrix Θ_2 may be obtained from Θ_1 by interchanging columns 2 and 4 and, next, interchanging rows 2 and 4.

2.4 Evolution of D in time

Distributions D_{111}, D_{112}, D_{121}, D_{122} and D_{123} form a complete system in that their evolution in time depends merely on their initial conditions, the matrix Θ and the

semigroup $\{S(t), t \geq 0\}$ in $\mathcal{M}_3$. For, if we let G be the generator of $\{S(t), t \geq 0\}$ and $\mathcal{G}$ be the generator of the Cartesian product $\{\mathcal{S}(t), t \geq 0\}$ of five copies of $\{S(t), t \geq 0\}$ in $\mathcal{M}_3^5$, then writing $D(t)$ for the (column-) vector of $D_{111}(t)$, $D_{112}(t)$, $D_{121}(t)$, $D_{122}(t)$ and $D_{123}(t)$ we have

$$\frac{\mathrm{d}D(t)}{\mathrm{d}t} = \mathcal{G}D(t) + \lambda N\Theta D(t) - \lambda N D(t), \quad t \geq 0 \tag{5}$$

provided that $D(0)$, the initial state of the distributions belongs to $\mathcal{D}(\mathcal{G})$, i.e., if all of its coordinates belong to $\mathcal{D}(G)$. In other words, $D(t) = \mathcal{T}(t)D(0)$ where the semigroup $\{\mathcal{T}(t), t \geq 0\}$ is generated by $\mathcal{G} + \lambda N\Theta - \lambda N$. The proof of these facts is analogous to that given in Bobrowski and Kimmel (2003) where the case of two loci is treated.

The result is intuitively clear: in the absence of genetic drift, where the members of the population evolve without influencing one another, the behavior of D is governed by (5) with $\lambda = 0$; in this case (5) is an uncoupled system of five independent equations ($\lambda = 0$ gives infinite life-time of an individual; see also (9) later on). The process of birth-death events is then treated as a perturbation of the uncoupled system; and these events occur at an exponential rate λN (since there are $2N$ individuals, each of them having independent exponentially distributed life-lengths with parameter $\lambda/2$).

If the transition matrix $\Theta = (\theta_{ij})$ where i, j are in $\{1, \ldots, \kappa\}$, where κ is a natural number, is ergodic, we define Π as a $\kappa \times \kappa$ matrix with all rows equal to a stationary distribution of the matrix Θ.

Lemma 1 *For the model with three loci the transition matrix Θ is ergodic.*

Proof For a transition matrix $\Phi = (\phi_{ij})$ where i, j are, say, in $\{1, \ldots, \kappa\}$, where κ is a natural number, we define the Dobrushin's ergodicity coefficient α as

$$\alpha = \alpha(\Phi) = \min_{1 \leq i,j \leq \kappa} \sum_{k=1}^{\kappa} \min(\phi_{ik}, \phi_{jk}) = 1 - \frac{1}{2} \max_{1 \leq i,j \leq \kappa} \sum_{k=1}^{\kappa} |\phi_{i,k} - \phi_{j,k}|, \tag{6}$$

see Iosifescu (1980, Sections 1.11.3–1.11.5). This coefficient provides an efficient way of studying asymptotic behavior of Φ in that if $\alpha(\Phi) > 0$ then the matrix Φ is ergodic and there exists a probability vector $(\pi_i)_{i=1,\ldots,\kappa}$ such that

$$\|\Phi^n - \Pi\| \leq (1 - \alpha(\Phi))^n, \tag{7}$$

where $\|\cdot\|$ denotes the maximum of all absolute values of entries of a matrix (Iosifescu 1980, Section 4.1.3).

To estimate the Dobrushin's coefficient of Θ we note that $\beta = 2 - 2\alpha$ (which is the maximum appearing in (6)) is a convex function of Φ, and so α is concave. Therefore, $\alpha(\Theta) \geq (1-r)\alpha(\Theta_0) + r_1\alpha(\Theta_1) + r_2\alpha(\Theta_2)$. Since, for Θ_0 this maximum is attained for $i = 1$ and $j = 5$, and equals 2, $\alpha(\Theta_0) = 0$. Similarly, the maximum for Θ_1 is attained simultaneously for $(i, j) = (1, 2), (1, 3), (1, 5), (2, 4), (3, 4)$ and $(4, 5)$ (provided $2N \geq 3$), and equals $2 - \frac{1}{N^2}$. Hence, $\alpha(\Theta_1) = \frac{2}{(2N)^2}$. Finally, since interchanging rows and columns does not influence the value of α, $\alpha(\Theta_2) = \alpha(\Theta_1)$. Hence, $\alpha(\Theta) \geq r\alpha(\Theta_1) \geq \frac{2r}{(2N)^2} > 0$.

The ergodicity of the transition matrix Θ leads to the following theorem:

Theorem 1 *For the model with three loci*:

$$\lim_{t\to\infty} \|\mathcal{T}(t) - \mathcal{S}(t)\Pi\| = 0, \tag{8}$$

and the speed of convergence is exponential.

Proof The semigroup generated by $\mathcal{G}$ acts on a vector $m = (m_i)_{i=1,\ldots,5}$ of five elements of $\mathcal{M}_3$ as follows:

$$\mathcal{S}(t)m = (S(t)m_i)_{i=1,\ldots,5}. \tag{9}$$

This implies that $\mathcal{S}(t)$ commutes with Θ and so we have $\mathcal{T}(t) = \mathcal{S}(t)\mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Theta}$ (cf. Theorem 3 in Griego and Hersh 1971). On the other hand, by (7) and $\Pi^n = \Pi$, $n \geq 2$, we have $\|\mathrm{e}^{\lambda Nt\Theta} - \mathrm{e}^{\lambda Nt\Pi}\| \leq \sum_{n=1}^{\infty} \frac{(\lambda Nt)^n \|\Theta^n - \Pi\|}{n!} \leq \sum_{n=1}^{\infty} \frac{(\lambda Nt)^n (1-\alpha(\Theta))^n}{n!} \leq \mathrm{e}^{\lambda Nt(1-\alpha)}$. Thus, $\|\mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Theta} - \mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Pi}\| \leq \mathrm{e}^{-\lambda Nt\alpha(\Theta)}$. Using $\mathrm{e}^{\lambda Nt\Pi} = I + \Pi(\mathrm{e}^{\lambda Nt} - 1)$, we get

$$\begin{aligned}
\|\mathcal{T}(t) - \mathcal{S}(t)\Pi\| &\leq \|\mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Theta} - \Pi\| \\
&\leq \|\mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Theta} - \mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Pi}\| + \|\mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Pi} - \Pi\| \\
&\leq \mathrm{e}^{-\lambda Nt\alpha(\Theta)} + \mathrm{e}^{-\lambda Nt} \leq \mathrm{e}^{-\frac{\lambda r}{2N}t} + \mathrm{e}^{-\lambda Nt},
\end{aligned} \tag{10}$$

as desired.

The most interesting practical consequence of Theorem 1 is that for large t, the distribution $D_{111}(t)$ in the model with drift and recombination is asymptotically the same as that in the model *without drift and recombination* provided the initial condition in the latter is the appropriate convex combination involving the stationary distribution of the matrix Θ:

$$D_{111}(t) \sim S(t)(\pi_1 D_{111}(0) + \pi_2 D_{112}(0) + \pi_3 D_{121}(0) + \pi_4 D_{122}(0) + \pi_5 D_{123}(0)).$$

In other words, recombination influences the model merely through this stationary distribution, and this is regardless of the way mutations are modeled.

Example 1 An explicit form of π may be found using *Mathematica 5.2* but the formula is long and non-informative. However, if in (4) we disregard starting from the second and the third row of the matrix all the terms of (N^{-2}) order, which for large populations are insignificant, and assume $r_1 = r_2$, the formula simplifies and yields,

$$\pi_1 = a^2 \frac{s(3s-5) - 3a(2-3s+s^2)}{(as-a-s)(6a^2+5as-9a^2s+s^2-4as^2+3a^2s^2)},$$
$$\pi_2 = \frac{as(a-1)[3a(s-1)-2s]}{(as-s-2a)(as-a-s)(3as-3a-s)},$$
$$\pi_3 = \frac{as^2(a-1)}{3a^3(s-2)(s-1)^2 - s^3 + as^2(5s-6) + a^2s(18s-7s^2-11)},$$
$$\pi_4 = \frac{as(a-1)[3a(s-1)-2s]}{(as-a-s)(6a^2+5as-9a^2s+s^2-4as^2+3a^2s^2)},$$
$$\pi_5 = \frac{(a-1)s^2(3a-1)}{as(5-4s)+s^2+3a^2(2-3s+s^2)},$$

where $s = \frac{r}{2}$, and $a = \frac{1}{2N}$. In particular, if $\frac{1}{2N} \ll s$, then $(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5) \approx (0, 0, 0, 0, 1)$ while if $2Ns \to c$, then $(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5)$ is approximately equal to $\frac{1}{(c+1)(c+2)(c+3)}(5c+6, c(3+2c), c^2, c(3+2c), c(c+1))$.

3 The general case

3.1 The model

In the general model there are n loci, and hence $n-1$ possible recombination sites. The population is composed of $2N$ individuals represented by n-tuples of $\mathbb{A}$-valued (see Sect. 2.1) random variables $(X_{a,b})_{b=1,\dots,n}$, $a = 1, \dots, 2N$, a being the individual number and b being the locus number. In between death/recombination events each locus in each individual evolves independently with mutation on the bth locus described by a Markov semigroup $\{S_b(t), t \geq 0\}$. The time to such an event is exponential with parameter λN. At the moment of death three numbers i, j and k are chosen with replacement from $1, \dots, 2N$, and then

- with the probability $1-r$ (no recombination), the whole n-tuple $(X_{i,l})_{l=1,\dots,n}$ is replaced by the n-tuple $(X_{j,l})_{l=1,\dots,n}$
- for $m = 1, \dots, n-1$, with the probability r_m (recombination after locus m), the $X_{i,l}, l = 1, \dots, m$ are replaced by the $X_{j,l}, l = 1, \dots, m$, and the $X_{i,l}, l = m+1, \dots, n$ are replaced by $X_{k,l}, l = m+1, \dots, n$.

Above, the positive numbers $r_m, m = 1, \dots, n-1$ are such that $\sum_{m=1}^{n-1} r_m = r \in (0, 1]$. Arguing as in Sect. 2.2 we see that such a procedure does not lead out of the class of exchangeable n-tuples.

3.2 The matrix Θ

The distributions involved in the model are labeled as $D_{i_1,\dots,i_n}$ where the multi-indexes $(i_1, \dots, i_n)$ satisfy the following properties:

1. i_1 is 1,
2. $i_\alpha \leq \max(i_1, \dots, i_{\alpha-1}) + 1, \alpha \geq 2$;

such multi-indexes will be called regular. There are ϖ_n regular multi-indexes, where ϖ_n is the Bell number, the number of ways to partition a set of n elements into subsets Graham et al. (1994). For every partition we have a natural order of its elements (subsets) where the first subset is the one containing the element 1 and the kth is the one containing the smallest number not included in the previous $k-1$ subsets (provided such number exists). To such naturally ordered partition we assign the regular multi-index by labelling elements of the kth subset with label k, and this map is injective. On the other hand, given a regular multi-index, we obtain a partition by collecting all numbers with the same index into one subset. Such assignment of a partition is injective, since the multi-index agrees with the labeling obtained from the natural order.

We arrange all the distributions $D_{i_1,\ldots,i_n}$ in the lexical order, thus forming vector D. Similarly, we form the vector $\tilde{D}$ of the distributions $\tilde{D}_{i_1,\ldots,i_n}$, and consider the way a coalescence/recombination event influences it. Suppose the recombination occurred after the sth locus, we are interested in $\tilde{D}_{i_1,\ldots,i_n}$ and we know that the ith individual died to be replaced partly by the jth and partly by the kth individual. Then, $\tilde{D}_{i_1,\ldots,i_n}$ equals $D_{j_1,\ldots,j_n}$ where the multi-index $(j_1,\ldots,j_n)$ is formed as follows. First, all occurrences of i at up to and including the sth place in $(i_1,\ldots,i_n)$ are replaced by j, and all the remaining occurrences are replaced by k. Then, the newly formed multi-index is transformed into a regular multi-index as follows. First, we change all occurrences of i_1 to 1, if the first condition of regularity is not yet met. Next, we look for the first place, say i_α, where the second requirement is not met. If there is no such place, we are done. Otherwise, we replace i_α and all its occurrences by the smallest integer larger than all i_β preceding i_α, and we continue this procedure until the multi-index is regular.

As a result, $\tilde{D}_{i_1,\ldots,i_n}$ is a convex combination of all possible $D_{j_1,\ldots,j_n}$'s; each choice of i, j and k leading from $\tilde{D}_{i_1,\ldots,i_n}$ to $D_{j_1,\ldots,j_n}$ adds the term $\frac{1}{(2N)^3} D_{j_1,\ldots,j_n}$ to this combination (all choices of i, j and k are equally likely). All coefficients of this combination are, themselves, linear combinations of $1, a, a^2$ and a^3 where $a = (2N)^{-1}$.

Hence, there exists a $\varpi_n \times \varpi_n$ matrix Θ such that

$$\tilde{D} = \Theta D.$$

Since ϖ_n is a fast growing sequence ([Graham et al. 1994, p. 693], e.g., $\varpi_4 = 15$, $\varpi_5 = 52$, and $\varpi_9 = 21147$), finding an explicit form of Θ by hand is not advisable. However, we have the following fundamental lemma.

Lemma 2 *For any number of loci the transition matrix Θ is ergodic.*

Proof The index $(1,\ldots,1)$ is an aperiodic state for Θ, since the one-step transition probability from this state to itself is positive. Hence, Θ being finite, it suffices to show that all other states communicate with $(1,\ldots,1)$. Let $(i_1,\ldots,i_n)$ be an arbitrary regular multi-index. Consider a recombination event: let s be the recombination site number, i be the number of the individual to be replaced, j be the number of the individual supplying the loci with numbers 1 through s, and k be the number of the individual supplying the loci with numbers $s+1$ through n. Taking $s = 1$, $i = j = 1$

and $k = i_2$ we jump from $(1, \ldots, 1)$ to $(i_1, i_2, \ldots, i_2)$, i_1 being equal to 1 by assumption. After arriving at $(i_1, i_2, \ldots, i_l, \ldots, i_l), l \geq 2$ we choose $s = l, i = j = i_l$, $k = i_{l+1}$ to jump to $(i_1, i_2, \ldots, i_l, i_{l+1}, \ldots, i_{l+1})$. Hence, after at most $n-1$ jumps, we arrive at $(i_1, \ldots, i_n)$.

Conversely, starting from $(i_1, \ldots, i_n)$, we choose $s = n-1, i = j = i_n, k = i_{n-1}$ to jump to $(i_1, \ldots, i_{n-2}, i_{n-1}, i_{n-1})$. After arriving at $(i_1, \ldots, i_{n-l-1}, i_{n-l}, \ldots, i_{n-l})$, $1 \leq l \leq n-2$ we choose $s = n-l-1, i = j = i_{n-l}$ and $k = i_{n-l-1}$, to jump to $(i_1, \ldots, i_{n-l-2}, i_{n-l-1}, \ldots, i_{n-l-1})$. Hence, after at most $n-1$ jumps we arrive at $(i_1, \ldots, i_1) = (1, \ldots, 1)$, proving our claim.

3.3 Evolution of D in time

Using the Markov semigroups $\{S_b(t), t \geq 0\}$ in l^1, introduced in Sect. 3.1, we define their tensor product semigroup $\{S(t), t \geq 0\}$ in $\mathcal{M}_n$, and denote its generator by G.

Also, let $\mathcal{G}$ be the generator of the Cartesian product $\{\mathcal{S}(t), t \geq 0\}$ of $m = m(n) = \varpi_n$ copies of $\{S(t), t \geq 0\}$ in $\mathcal{M}_n^m$. Then, as a function of time, the vector $D(t)$ of distributions involved in the model satisfies Eq. (5) provided that $D(0)$ belongs to $\mathcal{D}(\mathcal{G})$. Equivalently,

$$D(t) = \mathcal{T}(t)D(0)$$

for any $D(0) \in \mathcal{M}_n^m$ where the semigroup $\{\mathcal{T}(t), t \geq 0\}$ is generated by $\mathcal{G} + \lambda N\Theta - \lambda N$. Furthermore, $\mathcal{T}(t) = \mathcal{S}(t)\mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Theta}$, Θ commuting with $\mathcal{S}(t)$. Finally, by Lemma 2, there exists a probability vector $(\pi_i)_{i=1,\ldots,\varpi_n}$ such that $\lim_{t\to\infty} \|\mathrm{e}^{-\lambda Nt}\mathrm{e}^{Nt\Theta} - \Pi\| = 0$ where Π is a $\varpi_n \times \varpi_n$ matrix with all rows equal to $(\pi_i)_{i=1,\ldots,\varpi_n}$. As a result we obtain the following theorem:

Theorem 2 *For any number of loci $n \geq 3$:*

$$\lim_{t\to\infty} \|\mathcal{T}(t)D(0) - \mathcal{S}(t)\Pi D(0)\| = 0.$$

As a consequence

$$D_1(t) \sim S(t) \sum_{\iota=1}^{\varpi_n} \pi_\iota D_\iota(0),$$

where instead of $D_{i_1,\ldots,i_n}$ we write D_ι where $\iota = \iota(i_1, \ldots, i_n)$ denotes the position of $D_{i_1,\ldots,i_n}$ in D; for example $\iota(1, \ldots, 1) = 1$ and $\iota(1, 2, \ldots, n) = \varpi_n$.

3.4 Algorithmic determination of Θ

Already for $n = 4$ the Θ is too large to be manageable by hand. We wrote a computer program that does the job for $n \leq 9$ (for $n \geq 10$ the size of Θ makes the task unmanageable even for a computer). Below, we explain how it works.

We have

$$\Theta = (1-r)\Theta_0 + \sum_{s=1}^{n} r_s \Theta_s,$$

where Θ_0 corresponds to no recombination, and Θ_s describes recombination after locus s, so the task reduces to calculating the latter matrices. To this end, first, all possible distribution types are generated and arranged in the lexical order; each distribution is labeled by a regular multi-index $(i_1, \ldots, i_n)$ or, equivalently, by its position $\iota = \iota(i_1, \ldots, i_n)$ in the vector D. We note that the off-diagonal entries in Θ_s are linear combinations of three numbers: a, a^2 and a^3 where $a = (2N)^{-1}$. The probabilities on the diagonal are equal to 1 plus a linear combination of the type described above, and may also be described by specifying coefficients of this linear combination. Hence, we initialize the matrix Θ_s of size $\varpi_n \times \varpi_n$ with entries being three dimensional vectors, by assigning zero vectors to all of its entries. The entries are denoted $\Theta_s[\tilde{\iota}, \iota]$. Their actual values are computed in ϖ_n iterations, each iteration leading to one row of the matrix ($\tilde{\iota}$ fixed, ι varies) as follows.

Each of $(2N)^3$ triples (i, j, k), describing the recombination event in which the deceased ith member of the population was replaced partly by j and partly by k, leading from $\tilde{D}_{\tilde{\iota}}$ to D_ι, adds the probability $(2N)^{-3}$ to the $\Theta_s[\tilde{\iota}, \iota]$. However, we do not need to run the algorithm through all possible triples. To see this, let μ be the number of distinct characters in the multi-index corresponding to $\tilde{\iota}$. Note that if $i > \mu$, then $\tilde{D}_{\tilde{\iota}}$ is equal to $D_{\tilde{\iota}}$. Next, consider the recombination event described by (i, j, k) where $i \le \mu$, $j > \mu$ and $j \ne k$, and assume that it leads from $\tilde{D}_{\tilde{\iota}}$ to D_ι. Then, any recombination event (i, j', k) where $j' > \mu$ and $j' \ne k$ also leads from $\tilde{D}_{\tilde{\iota}}$ to D_ι. The same is true if j and k are interchanged. Also if $i \le \mu$, $j = k > \mu$ and $D_{\tilde{\iota}}$ leads to D_ι, then $D_{\tilde{\iota}}$ leads to D_ι for all other events (i, j', k') where $j' = k' > \mu$. Hence, the triples (i, j, k) naturally split into six classes, and the computation of Θ_s may be performed in the following six steps:

1. (The case where $i > \mu$.) Set $\Theta_s[\tilde{\iota}, \tilde{\iota}]$ to $[-\mu, 0, 0]$. (There are $2N - \mu$ choices for $i > \mu$.)
2. (The case where $1 \le i, j, k \le \mu$.) Run through all $i, j, k \le \mu$ to increase $\Theta_s[\tilde{\iota}, \iota]$ by $(0, 0, 1)$ each time $D_{\tilde{\iota}}$ leads to D_ι.
3. (The case where $1 \le i, j \le \mu$ and $k > \mu$.) Set $k = \mu + 1$ and run through all $i, j \le \mu$ to increase $\Theta_s[\tilde{\iota}, \iota]$ by $(0, 1, -\mu)$ each time $D_{\tilde{\iota}}$ leads to D_ι. (There are $2N - \mu$ choices for $k > \mu$.)
4. (The case where $1 \le i, k \le \mu$ and $j > \mu$.) Set $j = \mu + 1$ and run through all $i, k \le \mu$ to increase $\Theta_s[\tilde{\iota}, \iota]$ by $(0, 1, -\mu)$ each time $D_{\tilde{\iota}}$ leads to D_ι.
5. (The case where $1 \le i \le \mu$, and $j = k > \mu$.) Set $j = k = \mu + 1$ and run through all $i \le \mu$ to increase $\Theta_s[\tilde{\iota}, \iota]$ by $(0, 1, -\mu)$ each time $D_{\tilde{\iota}}$ leads to D_ι.
6. (The case where $1 \le i \le \mu$, $j, k > \mu$ and $j \ne k$.) Set $j = \mu + 1$, $k = \mu + 2$ and run through all $i \le \mu$ to increase $\Theta_s[\tilde{\iota}, \iota]$ by $(1, -2\mu - 1, \mu(\mu + 1))$ each time $D_{\tilde{\iota}}$ leads to D_ι.

3.5 Complexity of the algorithm

The memory complexity $M(n)$ is the sum total of the space used by the matrixes used in the algorithm. The first of these, termed *symbolic*, is build of ϖ_n^2 triples of coefficients of 4 bytes numbers, and the second, termed the *value matrix*, is obtained from the first one by multiplying these coefficients by the consecutive powers of $\frac{1}{2N}$ (with 1 added on the main diagonal). Hence, $M(n) = M(SymbolMatrix) + M(ValueMatrix) = 12\varpi_n^2 + 8\varpi_n^2 = 20\varpi_n^2[Byte]$, each number in the value matrix using 8 bytes. For example $M(8) = 340\,\text{MB}$ and $M(9) = 8.5\,\text{GB}$.

In calculating each row of the matrix, we perform n^3 iterations (actually, μ^3 iterations) and must use n iterations to transform a multi-index involved into its regular form. Hence, time-complexity of calculating each row is of the order n^4. Taking into account the initialization process, we obtain that time complexity is $n^4\varpi_n + \varpi_n^2$.

4 Numerical results

In this section, we explore how the distributions of the model depend on its parameters. We also use the notion of the spectral gap to understand the rate of convergence of the solutions of our model to equilibrium. Finally we comment on the linkage disequilibria produced by the model.

4.1 Stationary distributions

The stationary distribution π is calculated using our program. We iteratively multiply the transition matrix Θ by itself until we reach the matrix with each row being almost equal to any other row of the matrix. The matrix after the kth iteration is equal to Θ^{2^k}. Two rows of the matrix are considered equal if all the differences of the values of the corresponding entries of both rows are lower than the chosen precision (usually equal to 10^{-6}).

The numerical calculations show that: (1) As $r = \sum_{i=1}^{n-1} r_i$ increases and $2N$ is fixed, the role of π_1 in the stationary distribution decreases to 0, while that of π_{ϖ_n} increases to 1 (Fig. 1). (2) With the growth of r, each π_i where $1 < i < \varpi_n$ initially increases to a maximal value, and then decreases to zero. If $r_1 = r_2 = \cdots = r_{n-1}$, tuples of the distributions related by symmetry, such as D_{11122} and D_{11222}, reach the maximal value at the same time (Fig. 2). This suggests that with the growth of r, the probability mass tends to concentrate close to π_{ϖ_n}; this intuition is supported in Fig. 3 where the expected number of recombination events ER is shown to grow with r. The value is calculated as follows: $ER = \sum_{i=1}^{\varpi_n} \pi_i \gamma_i$, where γ_i is the number of recombination events needed to obtain the ith distribution. In this case we assume that after each locus only one recombination may take place. Then, for each distribution, the number of recombination events leading to it may be easily calculated as the number of the consecutive pairs of loci descended from the different individuals. For example, to obtain the distribution D_{1223} exactly two recombination events are required (after the first and after the third locus). However, the role of distributions close to the last

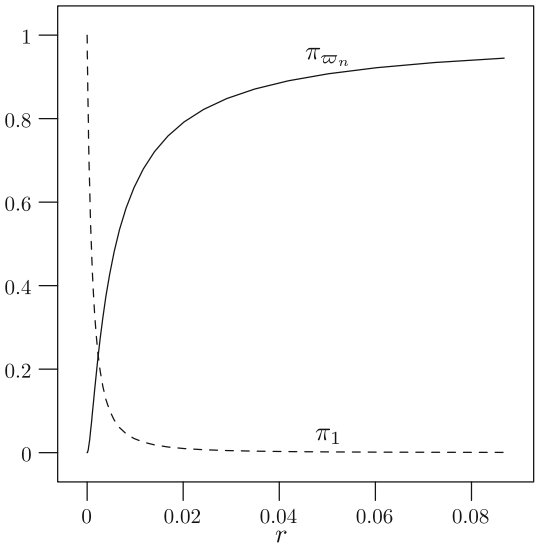


Fig. 1 Values of the first (π_1) and the last (π_{ϖ_n}) entry of the stationary distribution for the model with five loci as a function of the recombination rate with constant population size $2N = 1{,}000$

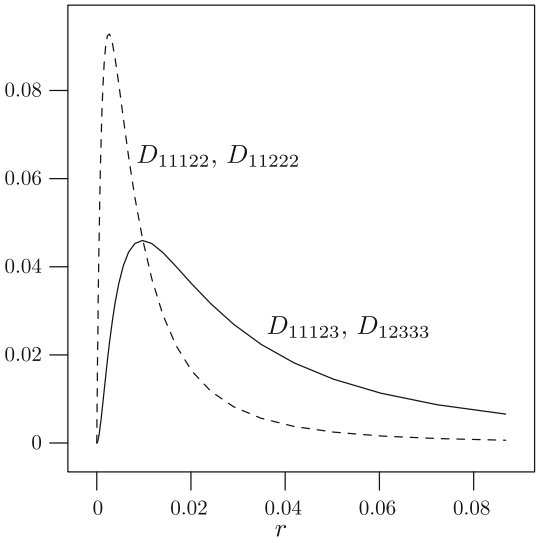


Fig. 2 Examples of values of entries of the stationary distribution as a function of the recombination rate for constant population size $2N = 1{,}000$. Since we assume $r_1 = r_2 = \cdots = r_{n-1}$ the entries for the distributions related by symmetry (such as D_{11123} and D_{12333}) are equal

one in the lexical order may decrease quite slowly (see Fig. 4). Finally, the speed (based on the number of discrete generations) of reaching the stationary distribution of Θ is of the order of the population size (Fig. 5). We assume that the matrix reaches the stationary distribution when all its entries differ from the corresponding entries of the previously calculated stationary distribution by less than 10^{-6}.

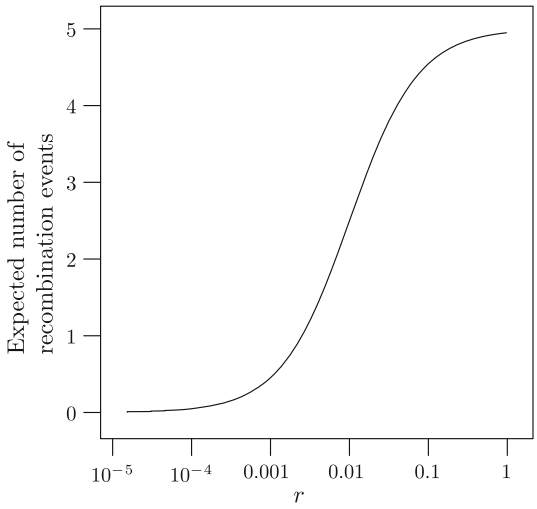


Fig. 3 Expected number of recombination events for the model with six loci as a function of the recombination rate with constant population size $2N = 1{,}000$. For more details see Sect. 4.1

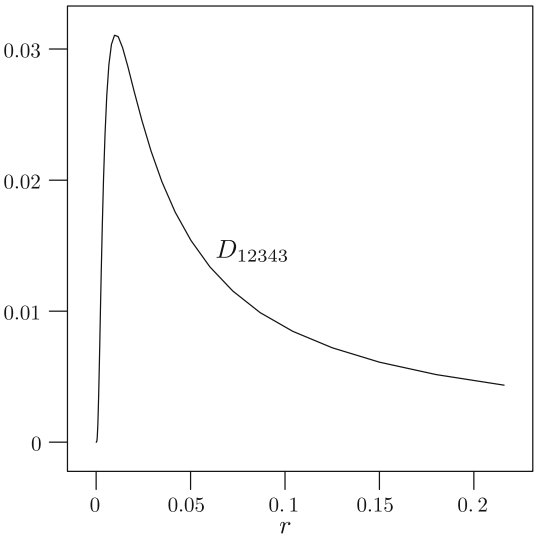


Fig. 4 An entry D_{12343} of the stationary distribution close to the last distribution in the lexical order for five loci as a function of the recombination rate for constant population size $2N = 1{,}000$

4.2 Spectral gap of the matrix Θ

Important information about the behavior of our model in time may be obtained by calculating the spectral gap of the matrix Θ. In Lemma 2 we have proved that the iterates of the matrix Θ asymptotically converge to a unique equilibrium. A higher value of the spectral gap is indicative of faster convergence. Therefore, if we compute the spectral gap for populations of chromosomes with different numbers of loci or different population sizes, the spectral gap provides us with the means to characterize the influence of these parameters on the rate of convergence.

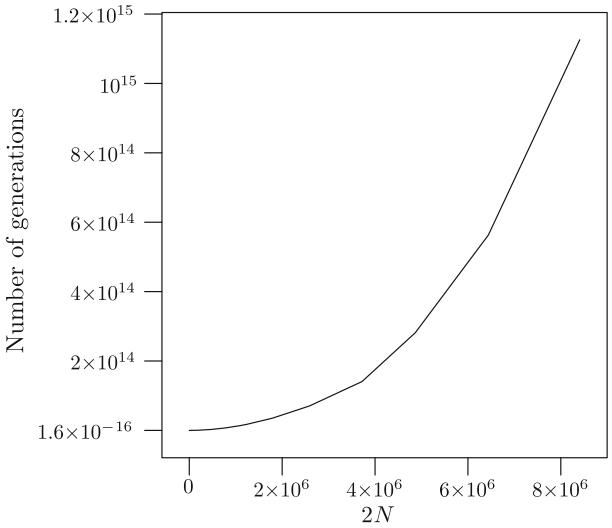


Fig. 5 Number of discrete generations required for the transition matrix Θ to reach, with given precision, the stationary distribution in each row, as a function of the population size. We assume that the matrix reaches the stationary distribution when all its entries differ from the corresponding entries of the stationary distribution by less than 10^{-6}

By definition (Saloff-Coste 1997), the spectral gap of Θ is equal to the smallest nonzero eigenvalue of the matrix $Q = I - \frac{1}{2}(\Theta + \Theta^*)$, where I is an identity matrix and Θ^* is the transition matrix of the time-reversed process. Each entry of the matrix Θ^*_{xy} is defined as: $\Theta^*_{xy} = \Theta_{yx}\frac{\pi_y}{\pi_x}$, where π is the stationary distribution of the matrix Θ.

To calculate the stationary distribution π, we apply the program described in the previous section. Eigenvalues of the matrix Q are obtained by the computer program based on the QR algorithm (Golub and Van Loan 1996). The results obtained by us are intuitively clear; the speed of convergence decreases when the number of loci (Fig. 6) or population size (Fig. 7) increases.

4.3 Linkage disequilibria

In Bobrowski and Kimmel (2003), which concerned the special case of two loci, it has been shown that when there were two alleles at each of the loci 1 and 2 and no mutation, the two-point linkage disequilibrium (in a modified notation) $L_{12} = p_{11} - p_{1\cdot}p_{\cdot 1} = \Pr[X_1 = 1,\ Y_1 = 1] - \Pr[X_1 = 1]\Pr[Y_1 = 1]$ dissolved with time, and for large N, it followed the expression

$$L_{12}(t) \approx \frac{1}{4}\left[\frac{1}{Nr-1} + \frac{Nr}{Nr-1}\exp(-\lambda r t/2)\right].$$

This latter expression is consistent with the known behavior of L_{12} in the absence of drift ($N \to \infty$). Distributions derived in the current paper allow obtaining dissolution

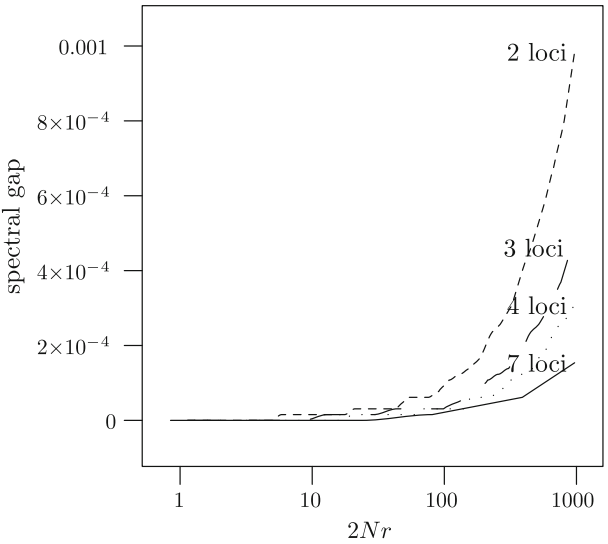
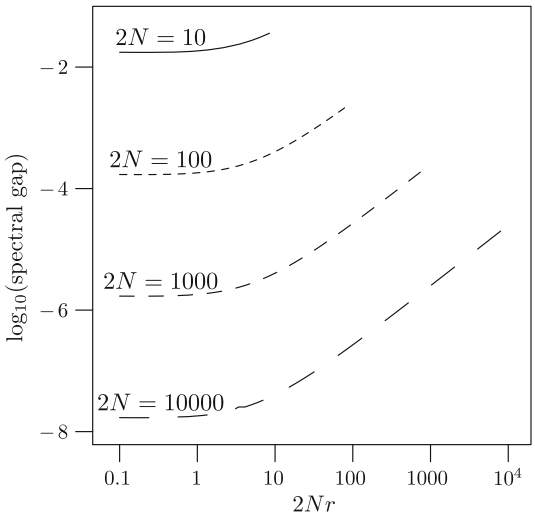


Fig. 6 The spectral gap as a function of $2Nr$ coefficient calculated for the models with different number of loci and constant population size $2N = 1{,}000$

Fig. 7 The spectral gap for the model with 5 loci as a function of $2Nr$, for various population sizes. Notice that the population size has a significant influence on the value of the spectral gap. Increasing the population size ten times results in decreasing the value of the spectral gap by about a hundred times



patterns for the multipoint equilibria in the presence of drift. For example, a three-point equilibrium has the form of Weir (1966, Equation (3.13))

$$L_{123} = p_{111} - p_{1\cdot\cdot}L_{23} - p_{\cdot1\cdot}L_{13} - p_{\cdot\cdot1}L_{12} - p_{1\cdot\cdot}p_{\cdot1\cdot}p_{\cdot\cdot1},$$

where

$$
\begin{aligned}
p_{111} &= \Pr[X_1 = 1,\ Y_1 = 1,\ Z_1 = 1],\\
p_{1\cdot\cdot} &= \Pr[X_1 = 1],\\
p_{\cdot 1\cdot} &= \Pr[Y_1 = 1],\\
p_{\cdot\cdot 1} &= \Pr[Z_1 = 1],\\
L_{12} &= p_{11\cdot} - p_{1\cdot\cdot}p_{\cdot 1\cdot},\\
L_{13} &= p_{1\cdot 1} - p_{1\cdot\cdot}p_{\cdot\cdot 1},\\
L_{23} &= p_{\cdot 11} - p_{\cdot 1\cdot}p_{\cdot\cdot 1},
\end{aligned}
$$

are further expressed in the terms of distributions of the type of D_{111}, D_{11} and D_1.

However, the dimensionality of the problem increases considerably. In the case of two loci, the special case of two alleles at each locus and no mutation, led to a system of eight ordinary differential equations (ODEs). Indeed, only two types of distributions, p_{11} and p_{12} were present. For each of these, there were 2^2 possible states at the two loci: (1, 1), (1, 2), (2, 1), and (2, 2). This leads to $2 \times 2^2 = 8$ variables. In the case of three loci, the special case of two alleles at each locus and no mutation, leads to a system of 40 ODEs. Indeed, now five types of distributions, p_{111}, p_{112}, p_{121}, p_{122}, and p_{123}, are present. For each of these, there are 2^3 possible states at the three loci. This leads to $5 \times 2^3 = 40$ variables. Therefore, an algebraic derivation of the expression for dissolution of a three-point disequilibrium seems to present difficulties, unless symbolic software is efficiently used. As for more general multi-point disequilibria, the task seems very involved.

5 Discussion

In this paper, we extend the theoretical treatment of the Moran model of genetic drift with recombination and mutation, which was introduced in Bobrowski and Kimmel (2003) for the case of two loci, to the case of n-loci. The specific framework used in our paper allows to find close-form relationships, which however are limited to a set of distributions, which jointly characterize allelic states at a number of loci at the same or different chromosome(s) but which do not jointly characterize allelic states at a single locus on two or more chromosomes. As an example, probabilities such as $\Pr[X_1 = x_1,\ Y_1 = y_1,\ Z_2 = z_2]$ are included in the system, whereas probabilities such as $\Pr[X_1 = x_1,\ X_2 = x_2,\ Z_2 = z_2]$ are not. However, the system is sufficiently rich to allow computing all possible multipoint linkage disequilibria under recombination, mutation and drift, as well as their variances and covariances (c.f. Chapter 3 of Weir's book, Weir 1966).

We explore the algorithms enabling construction of the transition probability matrices of the Markov chain describing the process. We find that asymptotically the effects of recombination become indistinguishable, at least as characterized by the set of distributions we consider, from the effects of mutation and drift (Theorems 1 and 2). Mathematically, the results are based on the theory of semigroups of operators. This approach allows generalization to any Markov-type mutation model. Based on these

fundamental results, we explore the rates of convergence to the limit distribution, using Dobrushin's coefficient and spectral gaps.

As it can be seen any examples involving application of our model for more than three loci require serious computations. These specialized applications will be the subject of future papers.

6 Availability of the computer program

Our program may be downloaded from http://sun.aei.polsl.pl/~twojdyla/genpop/ in two versions, both working well for $n \leq 8$. In the first, a simpler one, by setting $r_1, \ldots, r_{n-1}$ and $2N$, one obtains the matrices Θ_i and the stationary distribution $(\pi_i)_{i=1,\ldots,\varpi_n}$. The second version is a programming library. The website resources also include a few examples of how to use our program.

Acknowledgments We thank the Referees for their remarks which have improved significantly the final version of this paper.

References

Baake E, Herms I (2008) Single-crossover dynamics: finite versus infinite populations. Bull Math Biol 70:603–624

Barton NH, Etheridge AM, Sturm AK (2004) Coalescence in a random background. Ann Appl Probab 14(2):754–785

Bobrowski A, Kimmel M (2003) A random evolution related to a Fisher–Wright–Moran model with mutation, recombination and drift. Math Methods Appl Sci 26:1587–1599

Defant A, Floret K (1993) Tensor norms and operator ideals. North Holland, Amsterdam

Durret R (2002) Probability models for DNA sequence evolution. Springer, New York

Engel K-J, Nagel R (2000) One-parameter semigroups for linear evolution equations. Springer, Berlin

Ewens WJ (2004) Mathematical population genetics. Springer, Berlin

Golub GH, Van Loan CF (1996) Matrix computations, 3rd edn. Johns Hopkins University Press, Baltimore

Graham RL, Knuth DE, Patashnik O (1994) Concrete mathematics, 2nd edn. Addison-Wesley, Reading

Griego RJ, Hersh R (1971) Theory of random evolutions with applications to partial differential equations. Trans Am Math Soc 156:405–418

Griffiths RC (1981) Neutral two-locus multiple allele models with recombination. Theor Popul Biol 19:169–186

Hein J, Schierup MH, Wiuf C (2006) Gene genealogies, variation and evolution. Oxford university press, Oxford

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23:183–201

Iosifescu M (1980) Finite Markov processes and their applications. Wiley, New York

Kimmel M, Peng B (2005) simuPOP: a forward-time population genetics simulation environment. Bioinformatics 21(18):3686–3687

Kimmel M, Polańska J (1999) A model of dynamics of mutation, genetic drift and recombination in DNA-repeat genetic loci. Arch Control Sci 9(XVL, 1–2):143–157

Ryan RA (2002) Introduction to tensor products of banach spaces. Springer, Berlin

Saloff-Coste L (1997) Lectures on finite Markov chains. In: Lectures on probability theory and statistics, Lecture notes in mathematics, vol 1665. Springer, Berlin, pp 301–413

Wakeley J (2008) Coalescent theory. Ben Roberts, Greenwood Village

Weir BS (1966) Genetic data analysis II. Sinauer Associates, Sunderland