Theoretical Population Biology 98 (2014) 48–58

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Markovian approximation to the finite loci coalescent with recombination along multiple sequences

Asger Hobolth [a,*], Jens Ledet Jensen [b]

[a] Bioinformatics Research Center, Aarhus University, Denmark
[b] Department of Mathematics, Aarhus University, Denmark

ARTICLE INFO

Article history:
Available online 28 January 2014

Keywords:
Coalescent with recombination
Conditional approach
Distance between segregating sites
Markov along sequences

ABSTRACT

The coalescent with recombination process has initially been formulated backwards in time, but simulation algorithms and inference procedures often apply along sequences. Therefore it is of major interest to approximate the coalescent with recombination process by a Markov chain along sequences. We consider the finite loci case and two or more sequences. We formulate a natural Markovian approximation for the tree building process along the sequences, and derive simple and analytically tractable formulae for the distribution of the tree at the next locus conditioned on the tree at the present locus. We compare our Markov approximation to other sequential Markov chains and discuss various applications.

1. Introduction

For the analysis of the ancestral relationship between a pair of sequences Li and Durbin (2011) use a hidden Markov model where the hidden variable describes a tree process. The background for this formulation originates from Wiuf and Hein (1999) where it is described how the ancestral recombination graph can be obtained as a tree building process along the sequences. In Wiuf and Hein (1999) the space along the sequences is continuous. The tree building process is not Markovian along the sequences, but a number of Markov approximations have been introduced, namely the SMC model of McVean and Cardin (2005) and the SMC' model of Marjoram and Wall (2006) (with an extension to a higher order Markov model given by Chen et al., 2009).

For the finite loci version of the model there is a natural procedure for making a Markov approximation along the sequences: namely using the transition kernel obtained from the conditional distribution of the tree at locus $(n+1)$ given the tree at locus n. In this paper we derive this distribution.

The study of the coalescent with recombination for two sequences and two loci is of considerable interest and is treated in detail in the first part of this paper. The process is described in Simonsen and Churchill (1997) as a Markov process backwards in time with nine states. Simonsen and Churchill (1997) provide the conditional distribution of the tree height at the right locus given the tree height at the left locus. However, this formula is not feasible for routine use because it consists of an infinite sum where each term requires numerical integration. In Section 2 we provide a very simple formula for the conditional distribution of tree heights for two sequences.

Section 3 is concerned with the coalescent with recombination for multiple sequences. We provide a general description of the Markov process backwards in time for the next tree conditioned on the current tree. For more than four sequences the state space of the general Markov approximation for trees along the sequences becomes rather large. In Section 3.3 of the paper we discuss how to further approximate the general Markov process (denoted the 'natural Markov approximation') by limiting the number of recombination events between adjacent trees.

In Section 4 we compare our natural Markov approximation to the SMC and SMC' models, extend to variable population size, apply the model to the distribution of the length between segregating sites, and consider the information gained by using four sequences instead of only two. The SMC and SMC' models can be viewed as further approximations of the natural Markov approximation. For two sequences and two loci we investigate in detail when the SMC and SMC' models are appropriate (Sections 4.1 and 4.2). We confirm the expectation that the SMC model should not be used for inferring the recombination rate. Furthermore we find that the SMC' model is appropriate if the coalescence rate is higher or of the same magnitude as the recombination rate between two sites, but less appropriate when the coalescence rate is a magnitude smaller than the recombination rate between two sites.

Our results for the transitions between tree heights can easily be extended to incorporate variable population size. We demonstrate that population size changes can have a dramatic effect on the

* Correspondence to: Bioinformatics Research Center, Aarhus University, C.F. Møllers Alle, Building 1110, DK-8000 Aarhus C, Denmark.
E-mail addresses: asger@birc.au.dk (A. Hobolth), jlj@imf.au.dk (J.L. Jensen).

0040-5809/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.tpb.2014.01.002

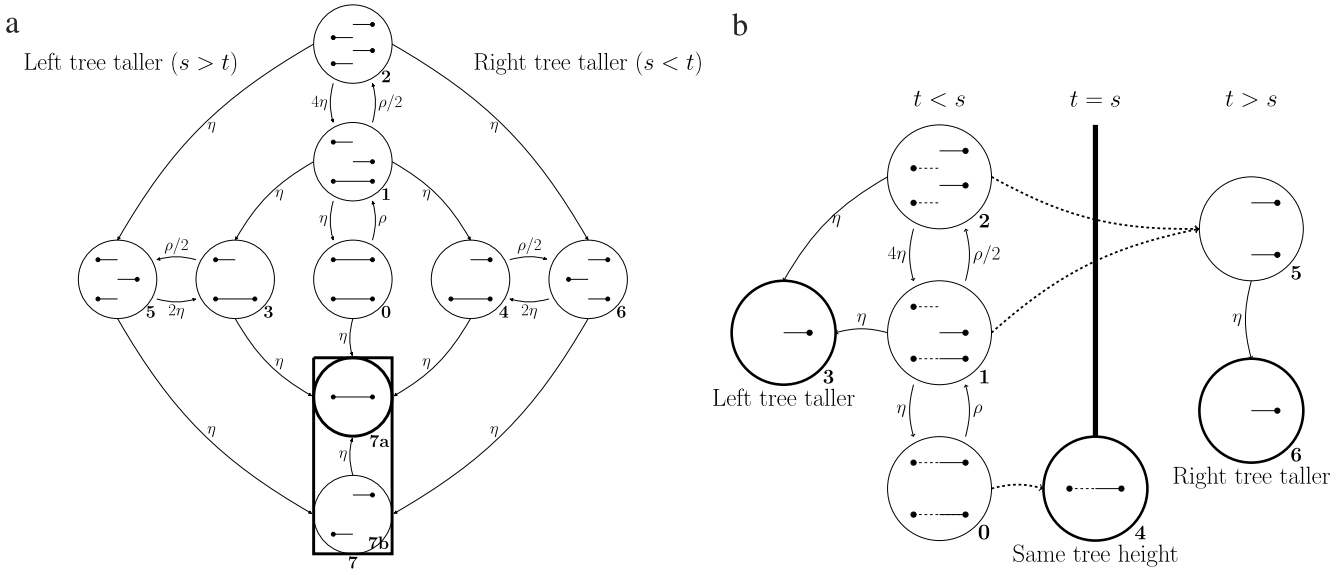

Fig. 1. (a) Complete ancestral recombination graph for two loci and two sequences described as a Markov chain backwards in time. A line with a bullet at both ends signifies a paired sequence (ancestral material to the sample at both loci), whereas a line with a bullet at one end only is a broken sequence with ancestral material at one locus only. (b) State diagram and rates for the conditional Simonsen–Churchill model, i.e. the Markov chain describing the right tree conditional on the height s of the left tree.
Source: The figure is adapted from Fig. 3 in Simonsen and Churchill (1997).

transition probabilities. These effects can also be seen in the distribution of the distance between segregating sites, a summary statistics recently suggested for pairs of sequences by Harris and Nielsen (2013). We discuss variable population sizes and distances between segregating sites in Sections 4.3 and 4.4. Finally, in a simulation study we quantify the information gained by using four sequences instead of only two (Section 4.5).

2. Two sequences

In this section we describe a Markov approximation of the coalescent with recombination process for two loci and two sequences. In particular we specify the probability of the right and left tree heights being identical conditional on the left tree height, and the conditional density of the right tree height conditional on the left tree height and a change in tree height.

2.1. Simonsen and Churchill model

The coalescent with recombination for two loci and two sequences is described in Simonsen and Churchill (1997) as a Markov process backwards in time with nine states. The nine states are shown in the left part of Fig. 1 together with the transition rates. A sequence is either *paired* (●——●), meaning that it contains material ancestral to the sample at both loci, or it is *broken* (●— or —●) when containing ancestral material to the sample at only one locus. The rate for a coalescence event is η for any two sequences (broken or paired), and the rate for a recombination is $\rho/2$ on any paired sequence. The chain begins in state 0 with two paired sequences. The height of the left tree, denoted by S, is the time at which the process enters one of the states 4, 6, 7a or 7b, and the height of the right tree, T, is the time at which one of the states 3, 5, 7a or 7b is entered. When state 7a is entered from state 0 the two tree heights are identical. State 7a is absorbing and corresponds to the grand most recent common ancestor. When only the tree heights are of interest states 7a and 7b can be joined to a common state 7.

It is evident from Fig. 1(a) that marginally the left and the right tree follows an ordinary coalescence process so that the density of the tree height is $\eta e^{-\eta t}$. When considering the Markov chain of tree heights along the sequences obtained from the conditional distribution of the right tree height T given the left tree height $S = s$ the stationary density is therefore $\eta e^{-\eta t}$.

To find a very simple description of the Markov chain, we now turn to a direct formulation of the conditional coalescent with recombination for the right tree given the left tree height $S = s$. The latter is again given as a Markov process backwards in time. The states and transition rates are shown in Fig. 1(b). Before time s there are 4 states: both of the right sequences are paired (state 0), one is paired and one is broken (state 1), both are broken (state 2), and only one sequence is left corresponding to the right tree being smaller than the left tree (absorbing state 3 in Fig. 1(b)). Each paired sequence has rate $\rho/2$ for breaking, and a broken sequence coalesces with any other lineage present at rate η. If, at time s, both right sequences are paired, the right tree ends up having the same height as the left tree. If, instead, one or both sequences are broken, the development for the right tree for $t > s$ is given by a Markov process with two states only (states 5 and 6), corresponding to the ordinary coalescence process.

For $t < s$ the rate matrix of the Markov process is

$$\Omega = \begin{pmatrix} -\rho & \rho & 0 & 0 \\ \eta & -(2\eta + \rho/2) & \rho/2 & \eta \\ 0 & 4\eta & -5\eta & \eta \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{1}$$

where we denote the states by 0, 1, 2 and 3. We use the notation $[A]_{ij}$ for the entries of a matrix.

Theorem 1 (*Conditional Simonsen–Churchill model*)**.** *The conditional probability of no change from the left to the right tree is*

$$P(T = s|S = s) = [e^{s\Omega}]_{00}, \tag{2}$$

and the conditional density $q(t|s)$ of T given $S = s$ and given $T \neq S$ is

$$q(t|s) = \begin{cases} \dfrac{\eta([e^{t\Omega}]_{01} + [e^{t\Omega}]_{02})}{1 - [e^{s\Omega}]_{00}} & t < s, \\ \dfrac{\eta e^{-(t-s)\eta}([e^{s\Omega}]_{01} + [e^{s\Omega}]_{02})}{1 - [e^{s\Omega}]_{00}} & t > s. \end{cases} \tag{3}$$

Table 1
The different events and corresponding rates of the coalescent with recombination process.

Event	Rate
$L_j + L_i \to L_j,\ i < j$	η
$L_i + P_{uv} \to P_{wv},\ w = \max\{i, u\}$	η
$P_{uv} + P_{ij} \to P_{uv},\ (i, j) < (u, v)$	η
$R_j + R_i \to R_j,\ i < j$	η
$R_j + L_i \to P_{ij}$	η
$R_i + P_{uv} \to P_{uw},\ w = \max\{i, v\}$	η
$P_{ij} \to (L_i, R_j)$	$\rho/2$

Proof. The conditional process needs to be in state 0 at time $T = s$ in order to have $T = S$. This gives the first result (2).

For $t < s$ we must at time t have a transition to state 3 from either state 1 or state 2. For $t > s$ we need to be in state 1 or state 2 at time s, and then follow an ordinary coalescence process from time s to time t. This gives the second result (3). □

We note that the rate matrix Ω is analytically tractable. In Lemma 3 in Appendix A we provide analytical expressions for the eigenvalues, eigenvectors, and matrix exponential entries $[e^{t\Omega}]_{0j}$.

The simplicity of the conditional density in Theorem 1 should be contrasted to the formula stated in Simonsen and Churchill (1997). There the joint density of the two tree heights (S, T) is given by

$$p_{S,T}(s, t) = \sum_v \Psi(s, t; v)Q(v),$$

where the sum is over an infinite set, $Q(v)$ is given explicitly and $\Psi(s, t; v)$ involves fourfold convolution of gamma densities.

In Theorem 1 the approach is via the conditional process. Using instead the joint process, described in Fig. 1(a), we obtain the following result.

Lemma 2. *Let Λ denote the rate matrix from the left part of Fig. 1 with states* 7a *and* 7b *joined together. The conditional probability of no change from the left to the right tree is*

$$P(T = s|S = s) = e^{\eta s}[e^{\Lambda s}]_{00}, \tag{4}$$

and the conditional density $q(t|s)$ of T given $S = s$ and given $T \neq S$ is

$$q(t|s) = \begin{cases} \eta e^{-(s-t)\eta}\dfrac{[e^{\Lambda t}]_{01} + [e^{\Lambda t}]_{02}}{e^{-\eta s} - [e^{\Lambda s}]_{00}} & t < s, \\ \eta e^{-(t-s)\eta}\dfrac{[e^{\Lambda s}]_{01} + [e^{\Lambda s}]_{02}}{e^{-\eta s} - [e^{\Lambda s}]_{00}} & t > s. \end{cases} \tag{5}$$

A proof is provided in Appendix B.

3. More than two sequences

3.1. General description of process for two loci

The coalescent with recombination on a finite set of loci is defined as a Markov process backwards in time where any two lineages have rate η to coalesce and the rate for a recombination between any two loci is $\rho/2$. A recombination event that leads to a sequence without ancestral material is disregarded.

For two loci and k sequences at time zero we use the following notation. The ancestral left material at time zero is numbered $L_1, L_2, \ldots, L_k$ and the ancestral right material is numbered $R_1, R_2, \ldots, R_k$. When L_i and L_j, $i < j$, coalesce the result is numbered L_j. Similarly for the coalescence of right material. Sequences with left ancestral material only, or sequences with right ancestral material only, are numbered according to the ancestral material. A sequence having left ancestral material L_i and right ancestral material R_j is denoted as P_{ij}. For the latter we say that $(i, j) < (u, v)$ if $j < v$ or if $j = v$ and $i < u$. All the different events, their rates and the resulting sequences are summarized in Table 1.

We now give a description of the coalescent with recombination that allows for an identification of the marginal process of the left tree and the conditional process of the right tree conditional on the events of the left tree. The marginal coalescence process for the left tree is given by the coalescence rates of the left sequences L_i and paired sequences P_{ij}, corresponding to the upper part in Table 1. Conditionally on the events in the left tree the right tree is described by following the right sequences R_i and the recombination process, corresponding to the lower part in Table 1. The right ancestral material can either be paired or broken, corresponding to P-sequences and R-sequences. A paired sequence becomes broken at the rate $\rho/2$. The right coalescence process is given by all the coalescence events of the R-sequences. Thus, a broken sequence can coalesce with any other broken sequence or with any sequence in the left tree at the rate η.

We now describe the conditional process of the right tree in terms of a continuous time finite state Markov process. We divide the time axis into intervals according to the coalescence events in the left tree. Consider a time interval $[T_1, T_2]$ where in the left tree there are ℓ lineages present and at time T_2 the two left lineages c_1 and c_2 coalesce. Let the remaining left sequences be $L_{i_3}, \ldots, L_{i_\ell}$. Let the number of right lineages at time T_1 be r. As an example, in Fig. 3 we could consider $T_1 = 0$, $T_2 = u_1$ and $\ell = r = 4$, or $T_1 = u_1$, $T_2 = u_3$ and $\ell = r = 3$. At any point in time we need to keep track of which right sequences are paired to left sequences. To this end, let A be the subset of right sequences paired to either L_{c_1} or L_{c_2}, and let B_j be the subset of right sequences paired to the left sequence L_{i_j}, $j = 3, \ldots, \ell$. Clearly, $|A| \leq 2$ and $|B_j| \leq 1$. The set A is needed to handle the possibility of a coalescence of two right sequences at the time T_2 where there is a coalescence in the left tree. Similarly, the sets $B_3, \ldots, B_\ell$ are used to identify the new state right after T_2. The state space $\mathcal{S}(\ell, r)$ of the process from T_1 until T_2, or until a coalescence of two right sequences, consists of the different possibilities of $(A, B_3, \ldots, B_\ell)$, and an absorbing state D (D stands for death) that is entered on coalescence of two right-ancestral lineages. If the process enters state D before time T_2 there is a change to the new state space $\mathcal{S}(\ell, r - 1)$. As an example, in Fig. 3 there is a change from state space $\mathcal{S}(3, 3)$ to $\mathcal{S}(3, 2)$ at time u_2.

The different transitions and their rates for the Markov chain with state space $\mathcal{S}(\ell, r)$ are as follows: recombination decreases one of the sets $A, B_3, \ldots, B_\ell$ by one, and the rate for such an event is $\rho/2$. Coalescence of a right sequence and a left sequence $(R_j + L_i \to P_{ij})$ increases one of $A, B_3, \ldots, B_\ell$ by one, and this type of event has rate $(2 - |A|)\eta$ or $(1 - |B_j|)\eta$. Finally, a coalescence of two right-ancestral lineages, corresponding to entering state D, has rate $\{s(r-s) + (r-s)(r-s-1)/2\}\eta$, where $s = |A| + |B_3| + \cdots + |B_\ell|$ is the number of paired sequences. Let $\Omega(\ell, r)$ be the rate matrix on $\mathcal{S}(\ell, r)$ defined by the above rates. For $\ell = r = 3$ the rate matrix $\Omega(\ell, r)$ is illustrated in Fig. 2.

The number of states for the Markov process with rate matrix $\Omega(\ell, r)$ can be calculated by dividing according to the number of elements in the set A. First define $B(s, k)$ to be the number of possible ways for placing at most one numbered item in each of k numbered boxes out of s items:

$$B(s, k) = \sum_{j=0}^{k} \binom{s}{j}\binom{k}{j} j!$$

Dividing according to $|A| = 0$, $|A| = 1$ or $|A| = 2$ we find

$$|\mathcal{S}(\ell, r)| = B(r, \ell - 2) + rB(r - 1, \ell - 2) + \frac{r(r-1)}{2}B(r - 2, \ell - 2) + 1.$$

The number of elements in $\mathcal{S}(\ell, r)$ for ℓ and r at most six is displayed in Table 2.

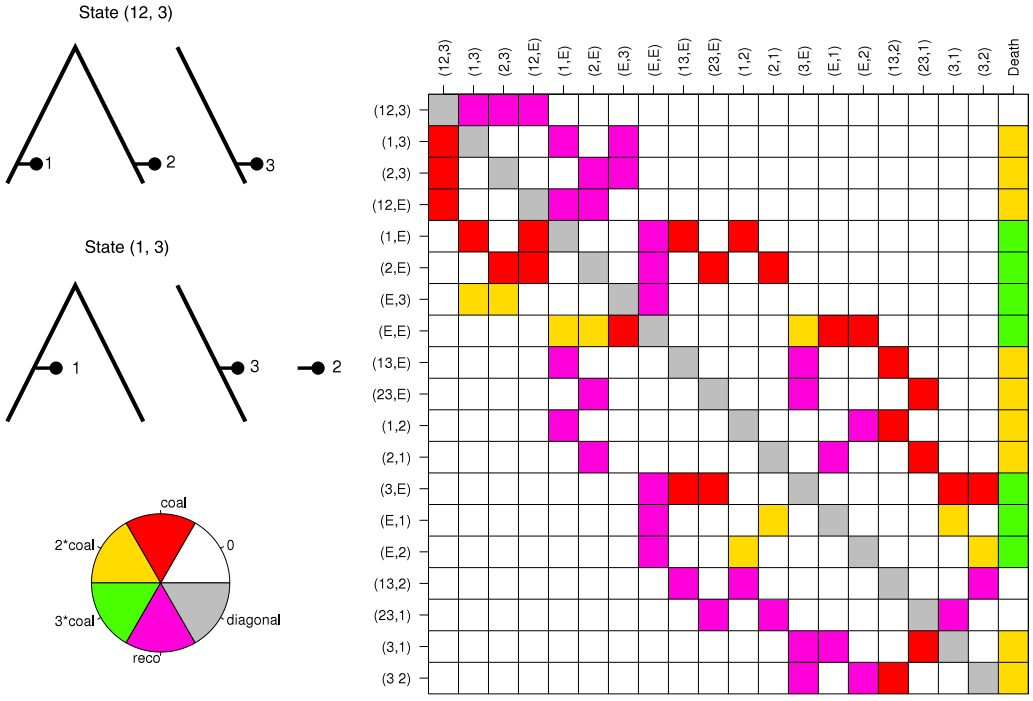

Fig. 2. Illustration of the state space $\mathcal{S}(\ell, r)$ for $\ell = r = 3$ and the corresponding rate matrix. The state space consists of the death state D and states of the form (A, B), where A and B are disjoint subsets of $\{1, 2, 3\}$. The size of the state space is $|\mathcal{S}(\ell, r)| = 20$, and E denotes the empty set. We also illustrate the states $(12, 3)$ (short for $(\{1, 2\}, 3)$) and $(1, 3)$.

Table 2
The number of states $|\mathcal{S}(\ell, r)|$ of the Markov chain describing the right ancestral tree when the left tree has ℓ lineages and the right tree has r lineages.

ℓ	$r = 2$	$r = 3$	$r = 4$	$r = 5$	$r = 6$
2	5	8	12	17	23
3	9	20	40	72	119
4	15	44	116	267	545
5	23	86	288	842	2141
6	33	152	628	2277	7187

The calculation of the density of the right tree given the left tree is done iteratively, splitting the calculation according to the coalescence events. Let (ℓ_i, r_i) be the number of left and right sequences after the i'th coalescence event with $(\ell_0, r_0) = (k, k)$. When there is a coalescence event in the left or the right tree only, ℓ or r is reduced by one. When there is a simultaneous coalescence event in both trees both of ℓ and r are reduced by one. The corresponding time points are denoted by u_i and we let $K + 1$ be the first instance of $r_{K+1} = 1$. Let

$$P(a, b; \ell_i, r_i) = \exp\{(u_{i+1} - u_i)\Omega(\ell_i, r_i)\}_{ab} \tag{6}$$

be the transition probability from state a to state b within $\mathcal{S}(\ell_i, r_i)$. When $\ell_i = 1$ the state space $\mathcal{S}(1, r)$ has two states only, one of which is the absorbing state D and the other we denote $\emptyset$. The transition probability in this case is $P(\emptyset, \emptyset; 1, r) = \exp\{-\eta[r(r-1)/2](u_{i+1} - u_i)\}$. Let $I_i \in \mathcal{S}(\ell_i, r_i)$ be the set of possible starting states at time u_i and $E_i \in \mathcal{S}(\ell_i, r_i)$ be the set of possible ending states just prior to time u_{i+1}, that is, those states that correspond to the event at time u_{i+1}. Formally, we let T be a transfer function that translates E_i to $I_{i+1} = T(E_i)$. Finally, for a vector h indexed by E_i we let $h * S$ be the vector indexed by I_{i+1} with

$$(h * S)(d) = \sum_{e \in E_i : T(e) = d} h(e).$$

We can now calculate the conditional density iteratively as follows. Let $P_0 = 1$ and I_0 be the initial state at time zero. Then we define for $i = 0, \ldots, K$:

$$\begin{aligned} P^0_{i+1} &= \eta^{1(\ell_{i+1} = \ell_i)} P_i P(I_i, E_i; \ell_i, r_i), \\ I_{i+1} &= T(E_i), \\ P_{i+1} &= P^0_{i+1} * S, \end{aligned} \tag{7}$$

with P_{K+1} being the conditional density. Here P_i is generally a vector and $P(I_i, E_i; \ell_i, r_i)$ a matrix. In the first line η enters when a coalescence of two right-ancestral lineages takes place.

3.2. Four sequences

Let us now describe the ending states E_i and the transfer function T in more detail for the particular case $k = 4$ sequences. It is helpful to have a concrete example in mind and for this purpose we use the left and right trees shown in Fig. 3.

The set of starting states I_i and ending states E_i are as follows:

$$(\ell_0, r_0) = (4, 4)$$
$$I_0 = \{(R_1R_2, R_3, R_4)\}$$
$$E_0 = \{(R_1R_2, R_3, R_4)_a, (R_1R_2, R_4, R_3)_a, (R_1R_2, R_3, \emptyset)_b, (R_1R_2, \emptyset, R_3)_b, (R_1R_2, R_4, \emptyset)_c, (R_1R_2, \emptyset, R_4)_c, (R_1R_2, \emptyset, \emptyset)_d\}$$
$$(\ell_1, r_1) = (3, 3)$$
$$I_1 = \{(R_3R_4, R_2)_a, (R_3, R_2)_b, (R_4, R_2)_c, (\emptyset, R_2)_d\}$$
$$E_1 = \{(R_2R_4, \emptyset)_a, (R_3R_4, \emptyset)_a, (R_2, R_4)_b, (R_3, R_4)_b, (R_2, \emptyset)_c, (R_3, \emptyset)_c, (R_4, R_2)_d, (R_4, R_3)_d, (R_4, \emptyset)_e, (\emptyset, R_2)_f, (\emptyset, R_3)_f, (\emptyset, R_4)_g, (\emptyset, \emptyset)_h\}$$
$$(\ell_2, r_2) = (3, 2)$$
$$I_2 = \{(R_3R_4, \emptyset)_a, (R_3, R_4)_b, (R_3, \emptyset)_c, (R_4, R_3)_d, (R_4, \emptyset)_e, (\emptyset, R_3)_f, (\emptyset, R_4)_g, (\emptyset, \emptyset)_h\}$$
$$E_2 = \{(R_3, R_4)_a, (R_4, R_3)_a, (R_3, \emptyset)_b, (R_4, \emptyset)_b, (\emptyset, R_3)_c, (\emptyset, R_4)_c, (\emptyset, \emptyset)_d\}$$
$$(\ell_3, r_3) = (2, 2)$$
$$I_3 = \{(R_3R_4)_a, (R_3)_b, (R_4)_c, (\emptyset)_d\}$$

$E_3 = \{(R_3), (R_4), (\emptyset)\}$
$(\ell_4, r_4) = (2, 1)$.

Here the subscript on each state is a label to indicate which states in E_i are transferred to a particular state in I_{i+1}.

The first event at time u_1 is a coalescent in both trees with L_1 and L_2 coalescing together with R_1 and R_2. The corresponding ending states E_0, of the form (A, B_3, B_4), are those with $A = \{R_1, R_2\}$. There are seven such states. Since the next coalescent event in the left tree is between L_3 and L_4, the starting states I_1, of the form $(\tilde{A}, \tilde{B}_3)$, are derived from E_0 by taking $\tilde{A} = B_3 \cup B_4$ and $\tilde{B}_3 = \{R_2\}$. This gives four states.

The second event at time u_2 is a coalescent in the right tree only between R_2 and R_3. The corresponding ending states E_1, of the form (A, B_3), are those with $A \cup B_3$ containing at most one of R_2 and R_3. There are thirteen such states. The starting states I_2 are states from E_1 with R_2 replaced by R_3.

The third event at time u_3 is a coalescent in the left tree only between L_3 and L_4. The corresponding ending states E_2, of the form (A, B_3), are those where A has at most one element. There are seven such states. Since the next coalescent event in the left tree is between L_2 and L_4, the starting states I_3, of the form $(\tilde{A})$, are derived from E_2 by taking $\tilde{A} = A \cup B_3$.

Finally, the fourth event at time u_4 is a coalescent in the right tree only between R_3 and R_4. The corresponding ending states E_3, of the form (A), are those with A containing at most one element. There are four such states.

3.3. Approximation for more than four sequences

We now make restrictions on the possible recombination events in order to make the state space smaller in the conditional process of the right tree given the left tree. Let $\alpha = r - |A| - |B_3| - \cdots - |B_\ell|$ be the number of broken (right) lineages, and let $\beta = \ell - |A| - |B_3| - \cdots - |B_\ell|$ be the number of broken left sequences. Here α and β are functions of time. We require that at all times $\max\{\alpha, \beta\} \leq \kappa$ for some chosen constant κ. Furthermore we require that there is at most one recombination event between two consecutive coalescence times u_i and u_{i+1}, and the recombination has to come prior to any coalescence events between u_i and u_{i+1}. This means that in the case $\max\{\alpha, \beta\} = \kappa$ only coalescences are allowed between u_i and u_{i+1}. If $\max\{\alpha, \beta\} < \kappa$ a recombination is allowed before any coalescence events. Instead of being able to move from any state in $\mathcal{S}(l, r)$ (except the absorbing state D) to any other state, one is now restricted to a much smaller set starting from a particular state. Thus, the information in the transition matrix $P(\ell, r)$ in (6) can be calculated from a set of much smaller matrices.

Let us consider the choice $\kappa = 2$ in more detail. If $\alpha = \beta = 2$ the reduced state space consists of the initial state $S_0 = (A, B_3, \ldots, B_\ell)$ together with all 6 states obtained from a coalescence of broken right sequences to broken left lineages, and the absorbing state D corresponding to a coalescence of two right lineages. A total of 8 states.

If instead $\alpha = \beta = 1$ (and $\kappa = 2$) the reduced state space consists of the initial state $S_0 = (A, B_3, \ldots, B_\ell)$, all $\ell - \alpha$ states obtained by a recombination, all states obtained by one or two coalescence events after the recombination, and the absorbing state D corresponding to a coalescence of two right lineages. A coalescence event is also possible directly from S_0, but this gives the same state as a recombination followed by two suitable coalescent events. The number of states is $6(\ell - \alpha) + 3$.

For the case $\ell = r = 2$ (and $\kappa \geq 2$) our reduced process disallows more than one recombination. This corresponds to removing state 2 in Fig. 1(b) and letting the transition from state 1 to state 0 be a transition to state 0^* instead, where one can move to state 4 only from the latter state. Thus this process is different from the SMC and SMC' defined below.

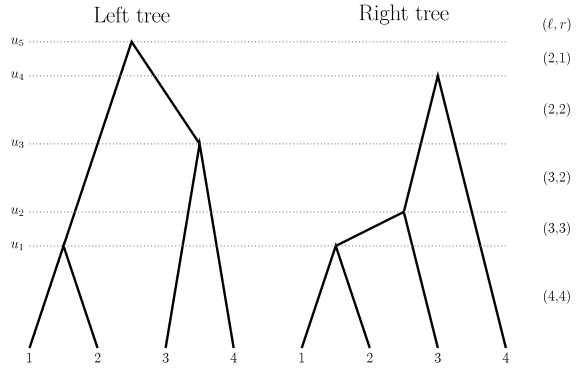

Fig. 3. Left tree: L_1 and L_2 coalesce at time u_1, L_3 and L_4 coalesce at time u_3, and L_2 and L_4 coalesce at time u_5. Right tree: R_1 and R_2 coalesce at time u_1, R_2 and R_3 coalesce at time u_2, and R_3 and R_4 coalesce at time u_4.

4. Applications

4.1. Comparison to other Markov models

In Theorem 1 we formulated a simple formula for the natural Markovian approximation to the spatial version of the ancestral recombination graph in the case of a finite set of loci and two sequences. In the continuous site model there exist a number of different Markovian approximations.

In the continuous site model the coalescent with recombination is defined by any two lineages having rate η for coalescence, any sequence has rate $\rho/2$ for a recombination event and the latter is uniformly distributed along the sequence. The two sites version of this model is the one given through the rates in Table 1. For the continuous site model McVean and Cardin (2005) introduced a restriction in order to obtain a Markov structure along the sequence. Their restriction says that two lineages can coalesce only if they share the ancestral material. The two sites version of this model therefore corresponds to removing the events $R_j + L_i \to P_{ij}$ from Table 1 since an R sequence and an L sequence do not share the ancestral material. For the two loci joint model in Fig. 1(a) this corresponds to removing the transitions $1 \to 0, 2 \to 1, 5 \to 3, 6 \to 4$, and $7b \to 7a$. For the conditional version in Fig. 1(b) one removes the transitions $1 \to 0$ and $2 \to 1$. To study the process of tree heights in the latter conditional process states 1 and 2 can be joined as both have the same rate for entering state 3, and both are transferred to state 5 at time s. We thus end up with a conditional model with three states, corresponding to 0, $1 + 2$ and 3, and with rate matrix

$$\Omega_{\text{SMC}} = \begin{pmatrix} -\rho & \rho & 0 \\ 0 & -\eta & \eta \\ 0 & 0 & 0 \end{pmatrix} \tag{8}$$

prior to time s. In this model we have that the conditional probability of no change in the tree height is $P(T = s|S = s) = e^{-\rho s}$, and the conditional density $q(t|s)$ of T given $S = s$ and given $T \neq S$ is

$$q(t|s) = \begin{cases} \dfrac{\eta\rho(e^{-\rho t} - e^{-\eta t})}{(\eta - \rho)(1 - e^{-\rho s})} & t < s, \\ \dfrac{\eta\rho e^{-\eta(t-s)}(e^{-\rho s} - e^{-\eta s})}{(\eta - \rho)(1 - e^{-\rho s})} & t > s. \end{cases}$$

The above two sites model appears in Paul et al. (2011).

In the continuous site model another Markovian approximation is the SMC' model of Marjoram and Wall (2006). This model is an extension of the SMC model where back-coalescence events of shared material are allowed. The two sites version of the model therefore corresponds to removing the events $R_j + L_i \to P_{ij}$ if $i \neq j$ and keeping the events $R_i + L_i \to P_{ii}$. In Fig. 1(b) this corresponds to

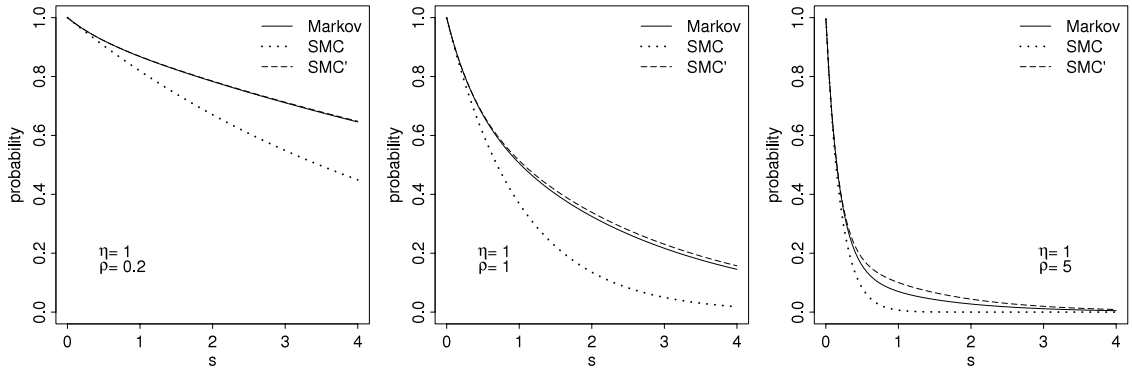

Fig. 4. Probability for not changing the tree for four three discrete space models: natural Markov, SMC/Li–Durbin and SMC'. Recall that the SMC and Li–Durbin change probabilities are identical.

allowing only half of the transitions from state 2 to state 1. To stay within the framework of Fig. 1(b) we consider the model obtained by removing transitions between state 2 and state 1. As above we get three states, 0, 1 + 2 and 3, and the rate matrix becomes

$$\Omega_{\mathrm{SMC}'} = \begin{pmatrix} -\rho & \rho & 0 \\ \eta & -2\eta & \eta \\ 0 & 0 & 0 \end{pmatrix}. \tag{9}$$

Also for this approximation explicit formulas can be given. We term this model SMC' and use this name since some backcoalescence events are allowed.

The Li and Durbin (2011) model that originally motivated our result in Theorem 1 is not one of the models described above. The holding time in the Li and Durbin (2011) model is $P(T = s|S = s) = e^{-\rho s}$, as in the discrete space SMC model. The transition density

$$q(t|s) = \begin{cases} \frac{1}{s}(1 - e^{-\eta t}) & \text{if } t < s \\ \frac{1}{s}(1 - e^{-\eta s})e^{-\eta(t-s)} & \text{if } t > s, \end{cases}$$

is based on ideas from the continuous site model, with the unfortunate implication that the marginal tree height no longer follows the distribution from the coalescent model, that is, the density is not $\pi(t) = \eta e^{-\eta t}$. In fact the stationary distribution of the Li and Durbin (2011) model is $\sigma(t) \propto te^{-\eta t}/(1 - e^{-\rho t})$. This result can easily be verified from the detailed balance condition; with $s \neq t$ we have

$$q(t|s)(1 - e^{-\rho s})\sigma(s) = q(s|t)(1 - e^{-\rho t})\sigma(t).$$

Also note that the transition density in the Li–Durbin model does not depend on the recombination rate.

We now compare the three approximate models (SMC, SMC' and Li–Durbin) to the full natural Markov approximation in terms of the probability for a tree change and the transition densities conditional on a tree change. In Fig. 4 we show the probability for a tree change for $\eta = 1$ and $\rho = 0.2, 1, 5$. We find that the SMC approximation does not perform very well, but the SMC' is a very reasonable approximation except when ρ is a magnitude larger than η.

In Fig. 5 we show the transition densities for $\eta = 1, \rho = 0.2, 1, 5$ and $s = 0.2, 1, 2$. We observe that the SMC model is a very reasonable approximation of the natural Markov model for small values of s. For large values of s the SMC approximation is less appropriate. The SMC' model is a very good approximation for all values of ρ, η and s considered in this comparison. The Li–Durbin model does not perform very well for large values of the recombination rate.

We tentatively conclude that the SMC' is a good approximation to the natural Markov model, but we also recall that the conditional description of the natural Markov model is so simple that very little computational advantage is gained from the approximation. Furthermore The SMC' model is superior to the SMC and Li–Durbin model.

4.2. Parameter estimation from simulated trees

In Fig. 6 we compare the performance of the three approximations of the ancestral recombination graph in terms of ability to recover the true coalescence and recombination rates. Using Hudson's simulator (Hudson, 2002) we simulated tree heights from the ancestral recombination graph for four different values of the recombination rate, namely $\rho = 0.2, 1.0, 6.0$ and 20.0, and a fixed value of the coalescence rate $\eta = 1.0$. We thus consider situations where the recombination rate is an order of magnitude smaller, of similar magnitude, and an order of magnitude larger than the coalescence rate. The number of simulated sites is $I = 20{,}000, 4000, 700$ and 200, respectively. These numbers ensure approximately the same number of tree changes for the different values of the recombination rate. For each value of ρ we simulated 50 independent samples and numerically maximized the likelihood arising from the Markov approximation

$$L(\eta, \rho; (t_1, \ldots, t_I)) = \pi(t_1) \prod_{i=1}^{I-1} L(\eta, \rho; t_i, t_{i+1}), \tag{10}$$

where $\pi(t) = \eta \exp(-\eta t)$ is the stationary distribution for the process and

$$L(\eta, \rho; s, t) = \begin{cases} [e^{t\Omega}]_{00} & t = s \\ \left([e^{t\Omega}]_{01} + [e^{t\Omega}]_{02}\right)\eta & t < s, \\ \left([e^{t\Omega}]_{01} + [e^{t\Omega}]_{02}\right)\eta e^{-\eta(t-s)} & t > s \end{cases} \tag{11}$$

is the likelihood for the tree height t at the next locus conditional on the tree height s at the current locus. The likelihood for the SMC and SMC' models are also given by (10) and (11) except that Ω is substituted by Ω_{SMC} from (8) and $\Omega_{\mathrm{SMC}'}$ from (9). In the natural Markov model the estimating equation obtained from the derivative of the log likelihood is based on the exact conditional distribution for two tree heights, and therefore the estimates are asymptotically unbiased (Cox, 1993) in that case. This is *not* the case for the SMC or SMC' models.

From Fig. 6 we observe that the SMC model described above always underestimates the recombination rate. Perhaps more interestingly, we also see that when the recombination rate is of similar or smaller magnitude than the coalescence rate, then the Markov and SMC' models described above are very similar. When the recombination rate is a magnitude larger than the coalescence rate we find, however, that the SMC' model overestimates the recombination rate.

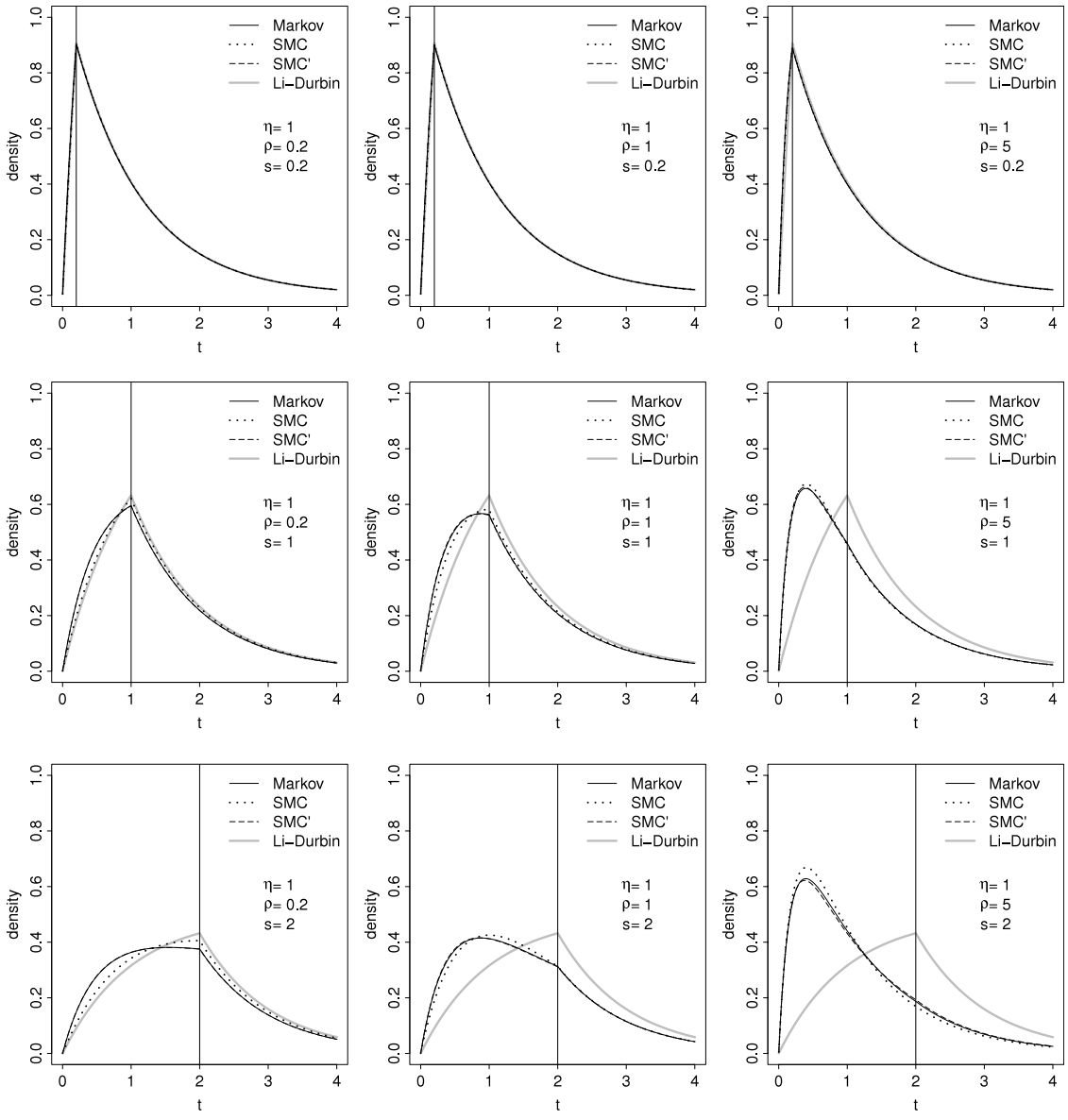

Fig. 5. Transition densities for the four discrete space models: natural Markov, SMC, SMC' and Li–Durbin.

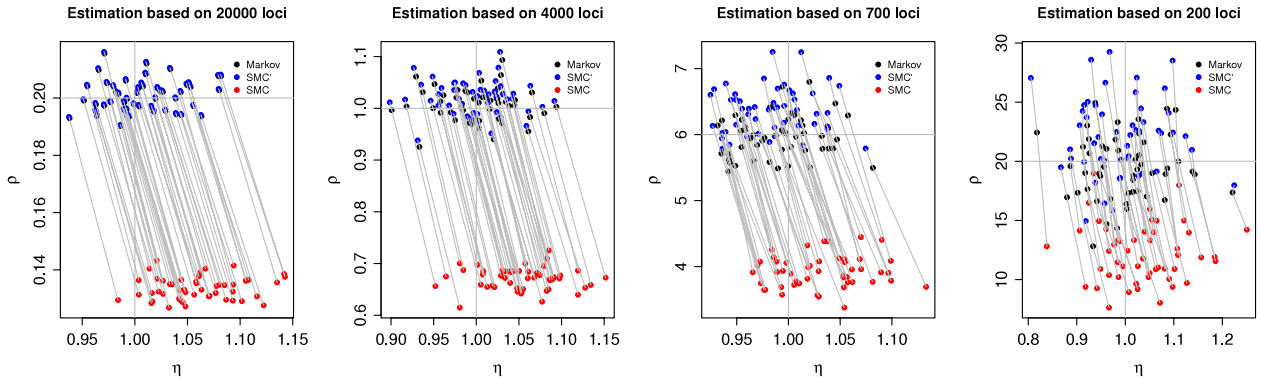

Fig. 6. Comparison between full Markov, SMC and SMC' approximations.

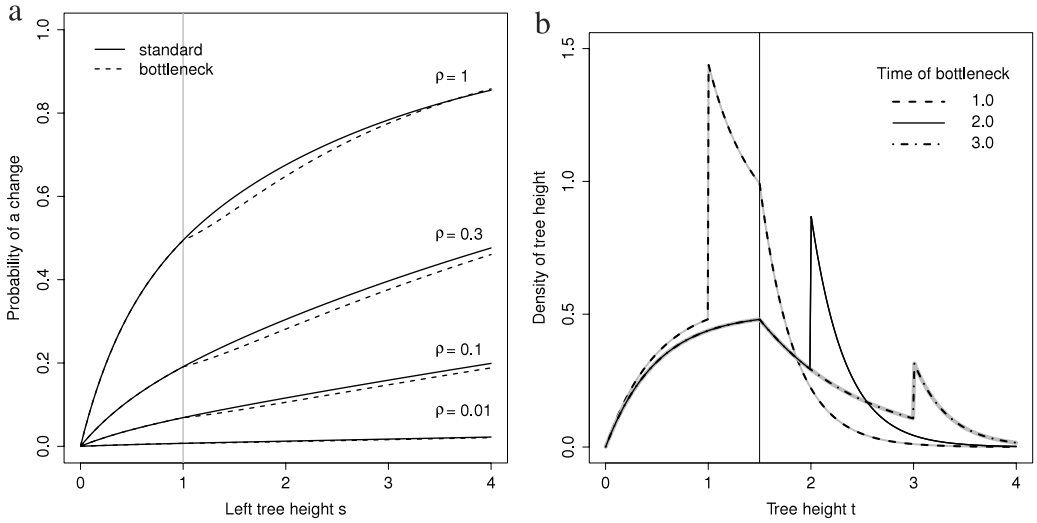

Fig. 7. (a) Probability for changing the tree from locus n to locus $(n+1)$ conditional on the tree height s at locus n. Solid lines: standard (constant size) coalescent model with varying recombination rates. Dashed lines: population size variability according to a bottleneck (see text for details). (b) Density for a new tree height at locus $(n+1)$ conditional on the tree height $s = 1.5$ at locus n and a tree change. The recombination rate is $\rho = 0.01$, and the population size varies according to a bottleneck.

4.3. Extension to variable population size

In the case of a variable population size the rate matrix Ω in (3) is no longer constant because the coalescent rate depends on time t. We let the coalescent rate be scaled by a function $\gamma(t)$ at time t, and for $t < s$ the rate matrix of the Markov process becomes

$$\Omega(u) = \rho \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \eta\gamma(u) \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & -2 & 0 & 1 \\ 0 & 4 & -5 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \rho A_\rho + \eta\gamma(u) A_\eta, \quad (12)$$

with obvious definition of the matrices A_ρ and A_η. We must therefore replace $t\Omega$ by $t\rho A_\rho + \eta\Gamma(t)A_\eta$ and s and t by $\Gamma(s)$ and $\Gamma(t)$, where $\Gamma(t) = \int_0^t \gamma(u)du$, and replace the initial η in (3) by $\eta\gamma(t)$ for $t < s$ and by $\eta\gamma(s)$ for $t > s$.

In order to illustrate the dynamics we first consider the probability for changing the tree. The standard coalescent (constant population size) with varying recombination rates is shown as the solid lines in Fig. 7(a). The coalescence rate is $\eta = 1$ and the recombination rate is indicated in the figure. As expected, the probability for changing the tree increases with the recombination rate and the previous tree height. The dashed lines in the figure show the situation when the population size changes according to a bottleneck. The bottleneck occurs at time $t_b = 1.0$ and reduces the population size to a third of the size before the bottleneck, i.e. $\gamma(t) = 1$ if $t \le t_b$, and $\gamma(t) = 3$ if $t > t_b$. It is evident from the figure that if the tree in the left locus is smaller than the bottleneck, then the probability of a tree change remains the same, but if the tree is higher than the bottleneck, then the probability of a tree change is different.

We also consider the density of the new tree height conditional on a change, but where the population size changes according to a bottleneck taking place at either $t_b = 1.0$, $t_b = 2.0$ or $t_b = 3.0$. After the bottleneck the population size is reduced by a factor of three. We condition on the previous tree having height $s = 1.5$ and a recombination rate $\rho = 0.01$. The density for the three situations is shown in Fig. 7(b). We note that the bottleneck has a very dramatic effect on the density for the new tree.

4.4. Time discretization and distance between segregating sites

Li and Durbin (2011) and Mailund et al. (2011) analyse pairs of sequences using a hidden Markov model (HMM). The hidden states are tree heights (times to the most recent common ancestor). The tree height is discretized to obtain a finite hidden state space. The emissions are alignment columns with probabilities corresponding to a substitution process on the tree.

We now describe how we discretize time for the case of two sequences considered in Section 2.1. The discrete version of the Markov process is used to build a finite Markov chain along the two sequences. When the finite Markov chain is combined with a substitution process we obtain a HMM as in Li and Durbin (2011). Instead of investigating the full HMM here we choose to focus on the distribution of the distance between segregating sites. This summary statistics has recently been used by Harris and Nielsen (2013) for demographic inference.

Let the discrete time points (backwards in time) of the Markov chain be $d_0 = 0 < d_1 < d_2 < \cdots < d_{M-1} < d_M = \infty$ and denote the corresponding states by $1, 2, \ldots, M$. State m ($m \in \{1, \ldots, M\}$) corresponds to tree heights in the interval between d_{m-1} and d_m. The continuous stationary distribution is $\pi(t) = \eta \exp(-\eta t)$, and therefore the discrete times are chosen such that $1 - \exp(-\eta d_m) = m/M$, or $d_m = -\log(1 - m/M)/\eta$, where we define $\log(0) = -\infty$.

We choose the discretization

$$P(T = k|S = j) = \begin{cases} [e^{\Omega d_k}]_{03} - [e^{\Omega d_{k-1}}]_{03} & \text{if } k < j, \\ [e^{\Omega d_k}]_{00} + \sum_{i\in\{0,1,2\}} [e^{\Omega d_{k-1}}]_{0i}[e^{\Omega(d_k - d_{k-1})}]_{i3} & \text{if } k = j, \\ \left([e^{\Omega d_j}]_{01} + [e^{\Omega d_j}]_{02}\right) e^{-\eta(d_{k-1} - d_j)} \left(1 - e^{-\eta(d_k - d_{k-1})}\right) & \text{if } k > j, \end{cases}$$

for $(k, j) \in \{1, \ldots, M\}^2$. The rationale for the discretization is as follows. The condition $S = j$ means that the left coalescence happens in the time interval $[d_{j-1}, d_j[$. The case $k < j$ is simply the probability that the Markov chain enters state 3 in the time interval $[d_{k-1}, d_k[$. In order to have $k > j$ the Markov chain should be in state 1 or 2 at time d_j, and coalesce in the time interval $[d_k, d_{k-1}[$. The case $k = j$ is the probability that the chain is in state 0 at time d_k *or* that the chain is in state 0,1 or 2 at time d_{k-1} and in state 3 at time d_k.

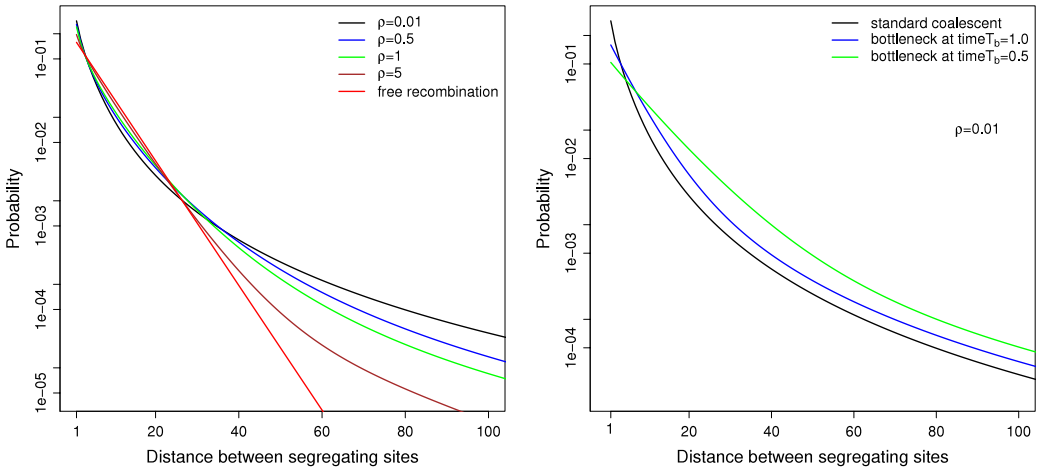

Fig. 8. Distribution of the distance between segregating sites. (a) Standard coalescent with various recombination rates. (b) Population size varies according to a bottleneck.

Turning to the distribution of the distance between two segregating sites we let $\mathcal{M}_i$ denote the event that a segregating site occurs at position i. In the Jukes–Cantor mutation model and with a tree of height d_m we have

$$P(\mathcal{M}|m) = \frac{3}{4}\Big(1 - \exp(-8\lambda d_m/3)\Big),$$

where λ is the mutation rate. The probability that the site is non-segregating is given by

$$P(\mathcal{M}^c|m) = \frac{1}{4} + \frac{3}{4}\exp(-8\lambda d_m/3)$$

(here superscript c stands for complement). Let μ be the vector with entry $\mu(m) = P(\mathcal{M}|m)$, $m = 1, \ldots, M$.

Let φ denote the stationary distribution of the Markov chain with transition probabilities given by the $M \times M$ matrix $P^{(S,T)}$ with (j, k) entry $P^{(S,T)}(j, k) = P(T = k|S = j)$. Now consider the stationary distribution of tree height conditional on a mutation

$$P(m|\mathcal{M}) = \frac{P(m, \mathcal{M})}{P(\mathcal{M})} = \frac{\mu(m)\varphi(m)}{\sum_m \mu(m)\varphi(m)}.$$

Let ϕ be the corresponding vector with entry $\phi(m) = P(m|\mathcal{M})$. We are now in a position to calculate the distribution between segregating sites. The probability for two adjacent segregating sites is

$$\begin{aligned} P(\mathcal{M}_1|\mathcal{M}_0) &= \sum_j \sum_k P(j|\mathcal{M}_0)P(T = k|S = j)P(\mathcal{M}_1|k) \\ &= \phi^T P^{(S,T)}\mu. \end{aligned}$$

Similarly

$$P(\mathcal{M}_2, \mathcal{M}_1^c|\mathcal{M}_0) = \phi^T P^{(S,T)}(I - \mathrm{diag}(\mu))P^{(S,T)}\mu,$$

where $\mathrm{diag}(\mu)$ is the $M \times M$ matrix with μ on the diagonal. Generally we have

$$\begin{aligned} &P(\mathcal{M}_k, \mathcal{M}_{k-1}^c, \ldots, \mathcal{M}_1^c|\mathcal{M}_0) \\ &= \phi^T\Big[P^{(S,T)}(I - \mathrm{diag}(\mu))\Big]^{k-1}P^{(S,T)}\mu. \end{aligned} \tag{13}$$

In Fig. 8 we show how the distribution of the distance between segregating sites depends on the recombination rate and changes in population size. Fig. 8(a) shows that for small recombination rates the tail of the distribution is more heavy than for large recombination rates. The reason is that for small recombination rates the Markov chain tends to stay longer in trees with a small height. We also included the case of free recombination ($\rho \to \infty$) where the waiting time is geometrically distributed with the probability of a mutation (i.e. success probability) equal to

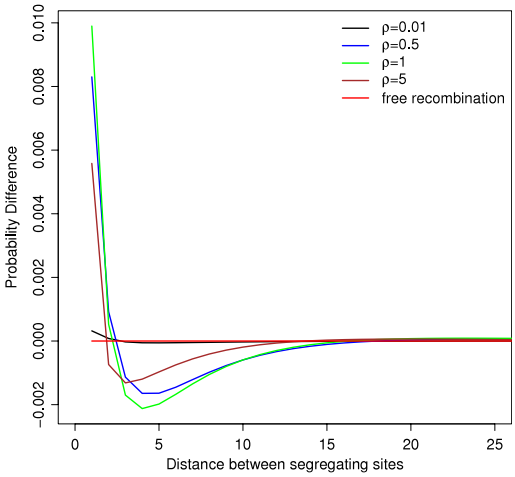

Fig. 9. Difference in distribution of the distance between segregating sites between the natural Markov and the SMC model.

$$\begin{aligned} P(\mathcal{M}) &= \int_0^\infty P(\mathcal{M}|t)\pi(t)dt = \int_0^\infty \frac{3}{4}\Big(1 - e^{(-8\lambda t/3)}\Big)\eta e^{-\eta t}dt \\ &= \frac{3}{4}\,\frac{8\lambda/3}{\eta + 8\lambda/3}. \end{aligned}$$

In Fig. 9 we show the difference in the tract length distribution between the natural Markov and the SMC model. We observe that for small or large values of the recombination rate, the tract length distributions are the same for the two models, but for intermediate values the difference can be rather large. In particular the SMC model predicts too many short distances and therefore also too few long distances between segregating sites. This result is perhaps not surprising in the light of Fig. 4 where we showed that the SMC model changes tree too often compared to the natural Markov model.

Harris and Nielsen (2013) in their calculation of the distribution of the distance between segregating sites use the SMC model. They do not discretize time, but derive an iterative procedure for calculating the probability (13). A similar derivation can be made with the model in Theorem 1 using Lemma 3. However, the discretization and recursive procedure suggested here provide a more simple implementation than the one provided in Claim 1 page 13 in Harris and Nielsen (2013).

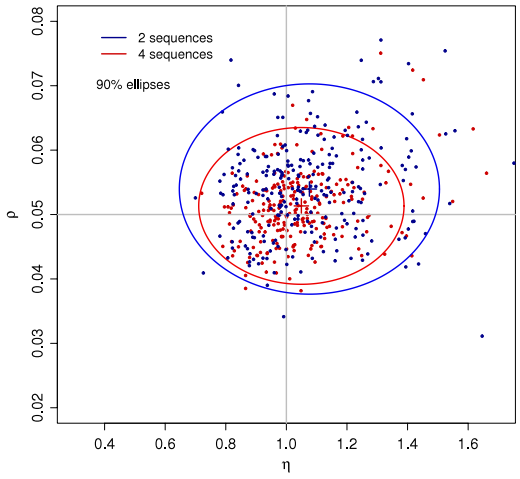

Fig. 10. Maximum likelihood estimates of (η, ρ) based on two sequences (blue points) or four sequences (red points). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.5. Information gained by using more sequences

In order to better understand the information gained by increasing the number of sequences from two to four, and to confirm that our Markovian approximation of the ancestral recombination graph can be used for parameter estimation, we conducted a similar simulation study as in Section 4.1.

We again used ms (Hudson, 2002) to simulate trees from the ancestral recombination graph with four sequences, and calculated the likelihood arising from the Markov approximation

$$L(\eta, \rho; (\tau_1, \ldots, \tau_l)) = \pi(\tau_1) \prod_{i=1}^{l-1} L(\eta, \rho; \tau_i, \tau_{i+1}),$$

where τ_i are the trees and $\pi(\tau_1)$ is the stationary distribution of the first tree. Here the likelihood $L(\eta, \rho; \tau_i, \tau_{i+1})$ is the likelihood for the tree τ_{i+1} at the next locus conditional on the tree τ_i at the current locus. The likelihood is given by (7).

We simulated 250 independent and identically distributed samples with $I = 2000$ loci, coalescence rate $\eta = 1.0$ and recombination rate $\rho = 0.05$ between loci. In Fig. 10 we show the resulting parameter estimates based on the trees from all four sequences (red points) and the tree heights from the first two sequences only (blue points). We find that the standard deviation of the estimates for η decreases from 0.200 to 0.157, and the standard deviation of the estimates for ρ decreases from 0.0076 to 0.0057. Analyzing twice as much independent data reduces the standard deviation by a factor of $\sqrt{2} = 1.41$. The standard deviation is reduced by $0.200/0.157 = 1.27$ for η and $0.0076/0.0057 = 1.33$ for ρ, so analyzing two independent two-sequence data sets gives more accurate parameter estimates than analyzing a single four-sequence data set.

We emphasize, however, that in order to handle more complex demographic scenarios such as e.g. two populations (isolated or with migration) with one individual from each population we must be able to handle four sequences.

5. Discussion

In this paper, we have derived a Markov description of the coalescent with recombination process along two or more sequences. For two sequences we obtain the situation described in Simonsen and Churchill (1997). The Simonsen and Churchill (1997) model is the setup needed in Li and Durbin (2011), where the hidden states are the times to the most recent common ancestor and the observed data are the observed sequences. Using the natural Markov approximation the artifacts of the Li and Durbin (2011) model are avoided. The Simonsen and Churchill (1997) description of the two loci and two sequences situation is in terms of the joint distribution of the two tree heights, as shown in Fig. 1(a). Our description is in terms of the conditional distribution of the right tree conditional on the left tree, as shown in Fig. 1(b). This new description has a smaller state space and is easy to manipulate, as demonstrated in terms of calculating e.g. the distribution between segregating sites as in Section 4.4.

For three and four sequences we have discussed the conditional natural Markov approximation in detail. The state space becomes rather large, but is still tractable. One advantage of the natural Markov approximation is that it allows us to re-visit and better understand the further approximations of the SMC and SMC' models. In particular we find that the SMC' model is appropriate provided that the recombination rate is not too large compared to the coalescent rate.

For five or more sequences the state space is large, and when combining with a discretization of the trees the resulting hidden state space becomes enormous. Rasmussen and Siepel (2013) describe a promising new method for the simulation-based inference of the ancestral recombination graph for multiple sequences. Their method is a Gibbs sampling procedure, where in each iteration the ancestry for a single sequence is updated conditional on the ancestry of the remaining sequences and the observed sequences. Rasmussen and Siepel (2013) use the SMC model along the sequences. It would be interesting to better understand the SMC approximation for multiple sequences in the light of the natural Markov approximation advocated in this paper. In particular even for humans the probability for two or more recombination events between any two base pairs could become so large that such events should not be ignored.

Acknowledgments

We thank the reviewers and editor for many constructive and helpful comments on earlier versions of this article.

Appendix A. Analysis of rate matrix for $(\ell, r) = (2, 2)$

Lemma 3 (*Eigenvalue Decomposition and Exponential of Ω.*). *The rate matrix Ω in (1) has an eigenvalue $\lambda_0 = 0$ with corresponding eigenvector $(1, 1, 1, 1)$. The remaining eigenvalues, λ_i, $i = 1, 2, 3$, are the three solutions to*

$$\lambda^3 + \lambda^2\left(7\eta + \frac{3}{2}\rho\right) + \lambda\left(10\eta^2 + \frac{13}{2}\eta\rho + \frac{1}{2}\rho^2\right) + 5\eta^2\rho + \frac{1}{2}\eta\rho^2 = 0,$$

with corresponding eigenvectors

$$(1, x_i, y_i, 0) = \left(1, 1 + \frac{\lambda_i}{\rho}, \lambda_i^2\frac{2}{\rho^2} + 2\lambda_i\frac{2\eta + \frac{3}{2}\rho}{\rho^2}\frac{2\eta + \rho}{\rho}, 0\right).$$

The transition probabilities can be written as

$$[e^{t\Omega}]_{0j} = a_1(j)e^{t\lambda_1} + a_2(j)e^{t\lambda_2} + a_3(j)e^{t\lambda_3}$$

with

$$a_1(0) = \frac{x_3 - y_3(x_2 - x_3)}{(y_1 - y_3)(x_2 - x_3) - (y_2 - y_3)(x_1 - x_3)},$$

$$a_2(0) = \frac{-x_3 - (x_1 - x_3)a_1(0)}{x_2 - x_3},$$

$$
\begin{aligned}
a_3(0) &= 1 - a_1(0) - a_2(0),\\
a_1(1) &= \frac{y_2 - y_3}{(y_2 - y_3)(x_1 - x_3) - (y_1 - y_3)(x_2 - x_3)},\\
a_2(1) &= \frac{-(y_1 - y_3)a_1(1)}{y_2 - y_3},\\
a_3(1) &= -a_1(1) - a_2(1),\\
a_1(2) &= \frac{x_2 - x_3}{(y_1 - y_3)(x_2 - x_3) - (y_2 - y_3)(x_1 - x_3)},\\
a_2(2) &= \frac{-(x_1 - x_3)a_1(2)}{x_2 - x_3},\\
a_3(2) &= -a_1(2) - a_2(2).
\end{aligned}
$$

Appendix B. Proof of Lemma 2

Proof. Marginally, the left tree follows an ordinary coalescence process and so the density of the tree height S is $\eta e^{-\eta s}$. In order to have $S = T = s$ we enter state 7 from state 0 at time s. The density of this traversal is $\eta[e^{\Lambda s}]_{00}$ where Λ is the rate matrix from the left part of Fig. 1 with states 7a and 7b joined together. Thus,

$$P(T = s|S = s) = \frac{\eta[e^{\Lambda s}]_{00}}{\eta e^{-\eta s}} = e^{\eta s}[e^{\Lambda s}]_{00}.$$

The joint density $f_{S,T}$ of (S, T) at the point (s, t) with $t < s$ is obtained by entering either state 3 from state 1, or state 5 from state 2 at time t, and then following an ordinary coalescence process from time t to time s:

$$f_{S,T}(s, t) = \eta([e^{\Lambda t}]_{01} + [e^{\Lambda t}]_{02})\eta e^{-(s-t)\eta}, \quad t < s.$$

Similarly, the joint density when $t > s$ is

$$f_{S,T}(s, t) = \eta([e^{\Lambda s}]_{01} + [e^{\Lambda s}]_{02})\eta e^{-(t-s)\eta}, \quad t > s.$$

The density of S at the height s on the set where $T \neq S$ is obtained by entering state 4 from state 1, state 6 from state 2, or state 7 from either 3 or 5 at time s

$$
\begin{aligned}
f_S(s; T \neq S) &= \eta([e^{\Lambda s}]_{01} + [e^{\Lambda s}]_{02} + [e^{\Lambda s}]_{03} + [e^{\Lambda s}]_{05})\\
&= \eta(e^{-\eta s} - [e^{\Lambda s}]_{00}),
\end{aligned}
$$

where $\exp(-\eta s)$ is the probability of no coalescence in the left tree, i.e. the probability of being in one of the states 0, 1, 2, 3 or 5. On dividing $f_{S,T}(s, t)$ by $f_S(s; T \neq S)$ we obtain the result of the lemma. □

References

Chen, G.K., Marjoram, P., Wall, J.D., 2009. Fast and flexible simulation of DNA sequence data. Genome Res. 19, 136–142.

Cox, D.R., 1993. Unbiased estimating equations derived from statistics that are functions of a parameter. Biometrika 80, 905–909.

Harris, K., Nielsen, R., 2013. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet 9 (6), e1003521.

Hudson, R.R., 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18, 337–338.

Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475, 493–496.

Mailund, T., Dutheil, J.Y., Hobolth, A., Lunter, G., Schierup, M.H., 2011. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. PLoS Genet 7 (3), e1001319.

Marjoram, P., Wall, J.D., 2006. Fast 'coalescent' simulation. BMC Genet. 7. Article 16.

McVean, G.A.T., Cardin, N.J., 2005. Approximating the coalescent with recombination. Philos. Trans. R. Soc. B. 360, 1387–1393.

Paul, J.S., Steinrücken, M., Song, Y.S., 2011. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. Genetics 187, 115–1128.

Rasmussen, M.D., Siepel, A., 2013. Genome-wide inference of ancestral recombination graphs. Manuscript. ArXiv:1306.5110.

Simonsen, K.L., Churchill, G.A., 1997. A Markov chain model of coalescence with recombination. Theor. Popul. Biol. 52, 43–59.

Wiuf, C., Hein, J., 1999. Recombination as a point process along sequences. Theor. Popul. Biol. 55, 248–259.