Department of Mathematics
Aarhus University

A state space model for mutations along two genomes

State space model: Simonsen-Churchill framework

The coalescent with recombination for two loci and two sequences is originally described in Simonsen and Churchill (1997) as a continuous time Markov chain backwards in time with eight states as shown in Figure 1. This Markov chain is given a careful treatment in the textbooks by Durrett (2008, Section 3.1.1) and Wakeley (2009, Section 7.2.4), and we therefore only briefly explain the basic properties of the model here.

A single sequence is either linked ($\bullet\!\!-\!\!\bullet$, $\times\!\!-\!\!\bullet$, $\bullet\!\!-\!\!\times$ or $\times\!\!-\!\!\times$) meaning that it contains material ancestral to the sample at both loci, or it is unlinked ($\bullet$–, –$\bullet$, –$\times$ or $\times$–) when it contains material ancestral to the sample at only one locus. The coalescent rate is one for any two sequences, and the recombination rate is $\rho/2$ for any linked sequence. The chain begins at time zero in state 1 with two linked sequences. After an exponential waiting time with rate $1+\rho$ the chain enters state 8 with probability $1/(1+\rho)$ or state 2 with probability $\rho/(1+\rho)$. The transition from state 1 to state 8 is a coalescent event, and the left and right tree heights are identical. The transition from state 1 to state 2 is a recombination event that breaks apart one of the two sequences. All other transitions have similar interpretations. Common ancestry for a locus is marked with a $\times$, so the transition from e.g. state 1 to state 8 is a transition to the state $\times\!\!-\!\!\times$.

The height S of the left tree is the first entry time to one of the states 5, 7 or 8 (states with a left $\times$), and the height T of the right tree is the first entry time to one of the states 4, 6 or 8 (states with a right $\times$). When state 8 is entered from state 1 the two tree heights are identical. State 8 is absorbing because only the tree heights are of interest.

The following Theorem from Hobolth and Jensen (2014) provides the two key ingredients for the state space model, namely the conditional probability for staying in a state $P(T=s|S=s)$ and the conditional density $q(t|s)$ of a new tree height t conditional on a change and a previous tree height s.

Theorem 1. *Let Λ denote the 8×8 rate matrix from Figure 1. The conditional probability of no change from the left to the right tree is*

$$P(T=s|S=s) = e^s[e^{\Lambda s}]_{11}, \tag{1}$$

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

$$q(t|s) = \begin{cases} e^{-(s-t)}\frac{[e^{\Lambda t}]_{12}+[e^{\Lambda t}]_{13}}{e^{-s}-[e^{\Lambda s}]_{11}} & t<s, \\ \\ e^{-(t-s)}\frac{[e^{\Lambda s}]_{12}+[e^{\Lambda s}]_{13}}{e^{-s}-[e^{\Lambda s}]_{11}} & t>s. \end{cases} \tag{2}$$

Proof. An application of Bayes formula gives the conditional probability of no change

$$P(T=s|S=s) = \frac{P(S=s,T=s)}{P(S=s)} = \frac{P(S=T=s)}{P(S=s)} = \frac{[e^{\Lambda s}]_{11}}{e^{-s}} = e^s[e^{\Lambda s}]_{11}.$$

It is evident from Figure 1 that marginally the left and the right tree follows an ordinary coalescence process so that the marginal density of a tree height is exponential with rate one. For future reference we note the consequence that

$$P(S\geq s) = e^{-s} = [e^{\Lambda s}]_{11}+[e^{\Lambda s}]_{12}+[e^{\Lambda s}]_{13}+[e^{\Lambda s}]_{14}+[e^{\Lambda s}]_{16}. \tag{3}$$

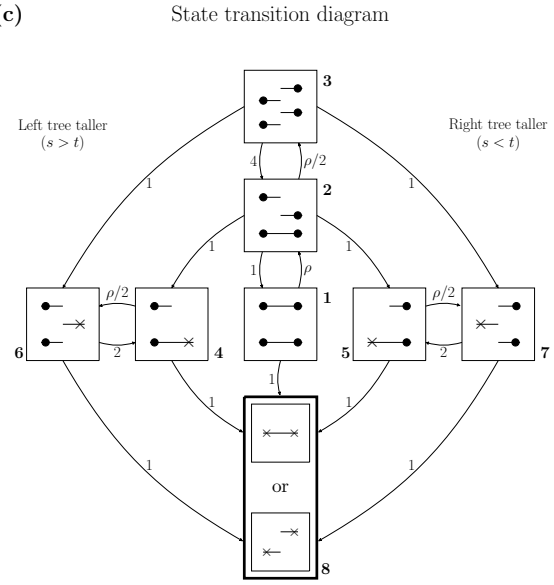

Figure 1: Complete ancestral recombination graph for two loci and two sequences described as a continuous time Markov chain backwards in time. The figure is adapted from Figure 7.7 in Wakeley (2009). A line with a bullet or a cross at both ends is a linked sequence (ancestral material to the sample at both loci), whereas a line with a bullet or a cross at one end only is a sequence with ancestral material at one locus only.

The conditional density $q(t|s)$ is given by

$$q(t|s) = P(T = t|S = s, s \neq t) = \frac{P(S = s, T = t, s \neq t)}{P(S = s, S \neq T)}.$$

We first consider the denominator, and note that the density of $S = s$ on the set where $S \neq T$ is given by either entering state 5 from state 2, entering state 7 from state 3 or entering state 8 from state 4 or 6 at time s:

$$P(S = s, S \neq T) = [e^{\Lambda s}]_{12} + [e^{\Lambda s}]_{13} + [e^{\Lambda s}]_{14} + [e^{\Lambda s}]_{16} = e^{-s} - [e^{\Lambda s}]_{11},$$

where in the last equality we used equation (3). We second consider the numerator, i.e. the joint density for the tree heights $S = s$ and $T = t$. For $s < t$ this event is achieved by entering state 5 from state 2 or state 7 from state 3 at time s, and then enter state 8 from state 5 or 7 at time t. The density for the latter event is exponential with rate 1, so we get

$$P(S = s, T = t, s < t) = \{[e^{\Lambda s}]_{12} + [e^{\Lambda s}]_{13}\}e^{-(t-s)}. \tag{4}$$

The case $s > t$ is obtained by a symmetry argument. □

Wakeley (2009, Section 7.2.4) noted that the transitions between state 4 and 6 and the transitions between state 5 and 7 can be removed from the chain if we are only interested in the tree heights. Actually, even more transitions can be removed from the chain. Note from eqn. (1) and eqn. (2) that we only need the entries $(1, 1), (1, 2)$ and $(1, 3)$ in $e^{\Lambda t}$ for calculating the probability of the same tree height in the next position and the transition density conditional on a change. These entries can be found from a reduced rate matrix where state 4, 5, 6 and 7 are removed and the rate from state 2 and 3

to a new absorbing state equals 2. In other words: Define the reduced rate matrix

$$\tilde{\Lambda} = \begin{pmatrix} -(1+\rho) & \rho & 0 & 1 \\ 1 & -(3+\rho/2) & \rho/2 & 2 \\ 0 & 4 & -6 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

where states are numbered 1, 2, 3 and 4. The holding time and transition density for the model are now given by eqn. (1) and (2) in our Theorem above with Λ substituted by $\tilde{\Lambda}$.

In the left plot in Figure 2 we illustrate the probability (1) of the same tree, and in the right plot of Figure 2 we illustrate the density (2) of a new tree height conditional on a change.

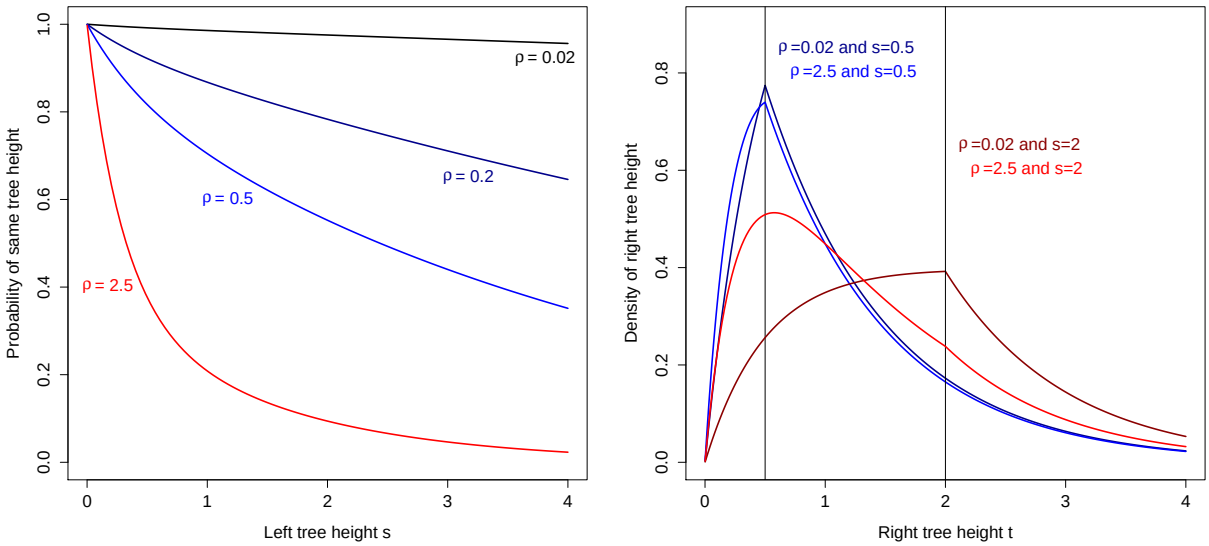

Figure 2: Probability of same tree height (left) and density for new tree height conditional on a new tree height (right).

Time discretization: Setting up the finite state HMM

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), and 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 (see Figure 3).

We now describe how we discretize time for the case of two sequences considered in the previous section. 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).

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 ($m \in \{1, \ldots, M\}$) then corresponds to a tree height in the interval between d_{m-1} and d_m. The continuous stationary distribution is $\pi(t) = \exp(-t)$, and therefore the discrete times are chosen such that $1 - \exp(-d_m) = m/M$, or $d_m = -\log(1 - m/M)$, where we define $\log(0) = -\infty$.

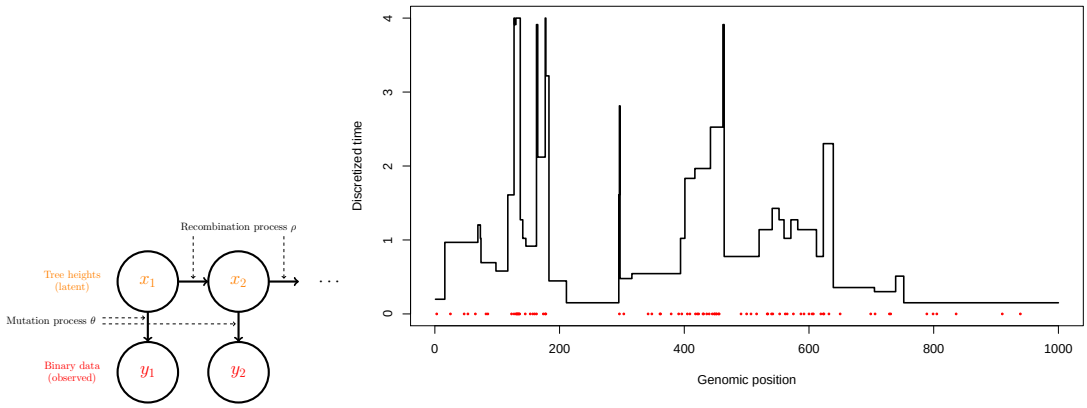

Figure 3: Left: Graphical structure of the hidden Markov Model. Right: Simulation from the hidden Markov model.

We now get for $1 \leq \ell, r \leq M$ the joint density

$$P(L = \ell, R = r) \tag{5}$$
$$= \begin{cases} \sum_{k\in\{5,7\}} \sum_{j\in\{5,7\}} \sum_{i\in\{1,2,3\}} [e^{\Lambda d_{\ell-1}}]_{1i} [e^{\Lambda(d_\ell - d_{\ell-1})}]_{ij} [e^{\Lambda(d_{r-1} - d_\ell)}]_{jk} [e^{\Lambda(d_r - d_{r-1})}]_{k8} & \text{if } \ell < r \\ \sum_{i\in\{1,2,3\}} [e^{\Lambda d_{\ell-1}}]_{0i} [e^{\Lambda(d_\ell - d_{\ell-1})}]_{i8} & \text{if } \ell = r \\ P(L = r, R = \ell) & \text{if } \ell > r. \end{cases}$$

The reason for the first case is that in order for the left tree height to be in state $\ell < r$ it must be in state 1, 2 or 3 at time $d_{\ell-1}$ and in state 5 or 7 at time d_ℓ (i.e. there has been no coalescent events before time $d_{\ell-1}$ and a left coalescent event between time $d_{\ell-1}$ and d_ℓ), and similarly it must still be in state 5 or 7 at time d_{r-1} and in state 8 at time d_r (i.e. there has been no coalescent events between time d_ℓ and time d_{r-1} and a right coalescent event between time d_{r-1} and time d_r). The next case corresponds to no coalescent events before time $d_{\ell-1}$ and both a left and a right coalescent event between time $d_{\ell-1}$ and d_ℓ. The last case is due to symmetry of the chain.

From the joint tree states (ℓ, r) we easily get the conditional tree states

$$P_{(\ell,r)} = P(r|\ell) = P(R = r|L = \ell) = \frac{P(L = \ell, R = r)}{P(L = \ell)},$$

where $P(L = \ell) = \sum_r P(R = r, L = \ell)$. These probabilities are used in the HMM.

Careful treatment of mutation process

A careful treatment of the mutation process allows for a more coarse binning procedure and is needed to avoid biasing the results. In continuous time the probability for a mutation given a tree height t is given by $\mu(t) = 1 - \exp(-\theta t)$, and the stationary tree height distribution is $\pi(t) = \exp(-t)$. The

probability of a mutation conditionally on the hidden state m becomes

$$\begin{aligned}
\mu_m &= p(y_i = 1|x_i = m) \\
&= p(y_i = 1|t \in (d_{m-1}, d_m)) = \frac{p(y_i = 1|t \in (d_{m-1}, d_m))}{p(t \in (d_{m-1}, d_m)} \\
&= \frac{\int_{d_{m-1}}^{d_m} p(y_i = 1|t)\pi(t)dt}{\int_{d_{m-1}}^{d_m} \pi(t)dt} = \frac{\int_{d_{m-1}}^{d_m} (1 - e^{-\theta t})e^{-t}dt}{\int_{d_{m-1}}^{d_m} e^{-t}dt} \\
&= 1 - e^{-\theta d_{m-1}} \frac{\left(1 - e^{-(1+\theta)(d_m - d_{m-1})}\right)}{(1+\theta)\left(1 - e^{-(d_m - d_{m-1})}\right)}. \qquad (6)
\end{aligned}$$

Note that with a fine discretization we have that the interval $d_m - d_{m-1}$ is small and the first-order Taylor expansion $\exp(-az) \approx 1 - az$ for z small gives

$$p(y_i = 1|x_i = m) \approx 1 - e^{-\theta d_{m-1}},$$

as perhaps expected. We are, however, discretizing the interval $[0, \infty[$, so it is not possible to avoid one or more large bins. Generally we have found that a careful treatment of the mutation process is crucial for accurate inference.

Simulation study

Here we choose to focus on three inference methods for estimating the recombination rate. The first method is based on the full likelihood obtained from the classical forward (or backward) algorithm for HMMs. The second is based on the distribution of the distance between segregating sites. This summary statistics was used in Harris and Nielsen (2013) for demographic inference. This summary statistics is sometimes also described as the distribution of the distance between heterozygote sites, runs of homozygosity, or the nearest-neighbour distribution. The third summary statistics is the probability that two sites a certain distance apart are both heterozygote sites. This probability is closely related to the pair correlation function from spatial statistics (Nielsen, Simonsen and Hobolth, 2016) and to the zygosity correlation introduced in Lynch *et al.* (2014).

Summary statistics: Runs of Homozygosity and Pair Correlation

Recall that in continuous time the probability for a mutation given a tree height t is given by $\mu(t) = 1 - \exp(-\theta t)$, and the stationary tree height distribution is $\pi(t) = \exp(-t)$. The marginal probability for a mutation is therefore given by

$$\int_0^\infty \mu(t)\pi(t)dt = \theta/(1+\theta). \qquad (7)$$

We also get the stationary distribution

$$\phi(t) = \frac{\mu(t)\pi(t)}{\int_0^\infty \mu(t)\pi(t)dt} = \frac{1+\theta}{\theta} e^{-t}\left(1 - e^{-\theta t}\right)$$

for a tree height t conditional on a mutation. Figure 4(a) shows $\phi(t)$ for different values of θ. Note that small mutation rates implies a higher tree height when we condition on a mutation. In discrete time the probability for a mutation given a tree height m was given by (6). Let $\mu = (\mu_1, \ldots, \mu_M)$ be the vector

of mutation probabilities. The stationary distribution $\phi = (\phi_1, \ldots, \phi_M)$ for a state m conditional on a mutation is given by

$$\phi_\ell = \frac{\mu_\ell \pi_\ell}{\sum_{m=1}^{M} \mu_m \pi_m},$$

where $\pi_m = 1/M$ because this is how the time discretization was chosen.

The probability for a mutation at a distance r from a typical mutation is then given by

$$\kappa(r) = \phi' P^r \mu,$$

where $'$ denotes vector transpose. In Figure 4(b) we show $\kappa(r)$ as a function of ρ and θ. Note that the curves converge to $\theta/(1+\theta)$ and that the behaviour for small r is determined by the recombination rate.

The distribution of runs of homozygosity is given by

$$\nu(r) = \phi' \Big[P \mathrm{diag}(e - \mu) \Big]^{r-1} P\mu.$$

Here $e = (1, \ldots, 1)$ is the vector of length M with 1 in every entry and $\mathrm{diag}(e - \mu)$ is the diagonal matrix with $e - \mu$ on the diagonal. In Figure 4(c) we show $\nu(r)$ as a function of ρ and θ.

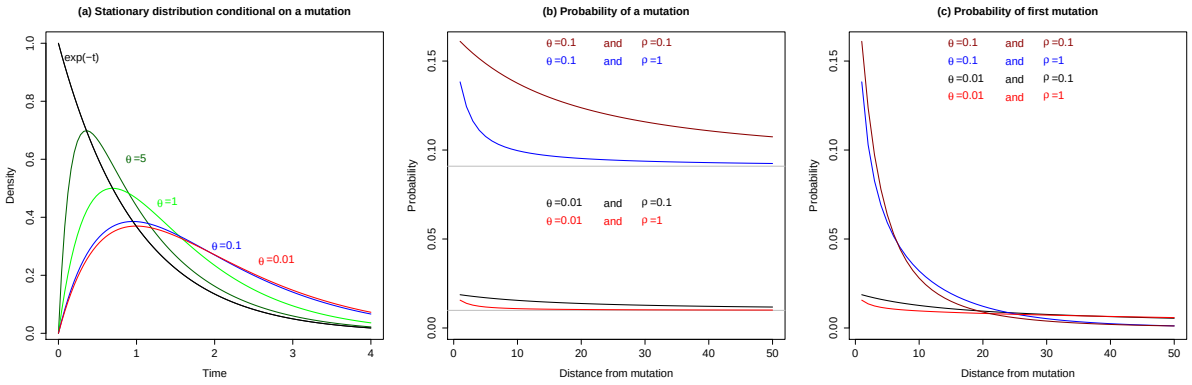

Figure 4: Left: Stationary distribution of tree height conditional on a mutation. Middle: Probability of a mutation at various distances away from a mutation. Right: Probability of the first mutation at various distances away from a mutation.

Parameter estimation

We estimate the mutation rate using an estimating equation based on the marginal probability for a mutation (7). If the observed frequency of a mutation is $\hat{p}$, then the mutation rate is $\hat{\theta} = \hat{p}/(1-\hat{p})$ (see left plot in Figure 5). The recombination rate is estimated using maximum likelihood for the HMM and goodness of fit for the pair correlation (see middle plot in Figure 5) and runs of homozygosity (see right plot in Figure 5).

We simulated 50 sequences of length 20.000 base pairs and with mutation rate $\theta = 0.1$ and recombination rate $\rho = 0.1$. We estimated the mutation rate using the estimating equation and the recombination rate using maximum likelihood and the HMM, and goodness of fit for the pair correlation and nearest neighbour. As expected the HMM procedure shows the best results because here we are using all the available information. It seems, however, that we are not loosing to much power when

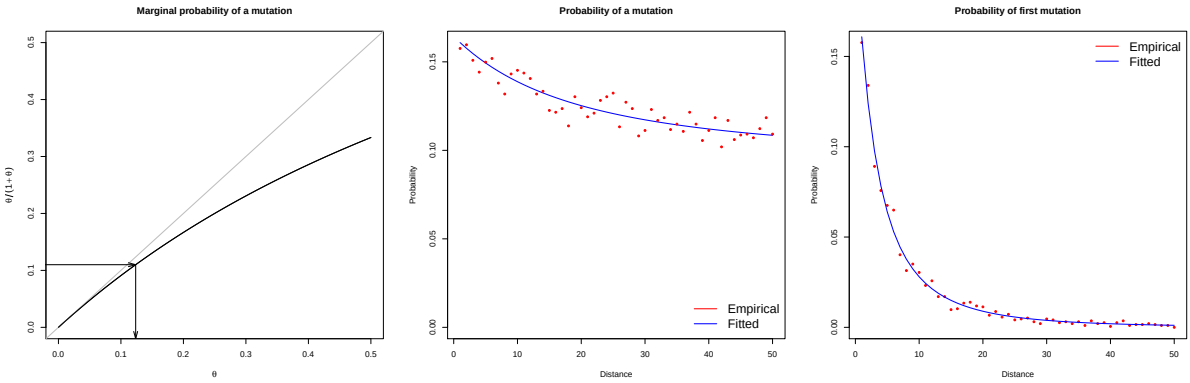

Figure 5: Parameter estimation for summary statistics. Left: The mutation rate θ is estimated from the observed number of mutations and length of the region. Middle: The recombination rate ρ is estimated using the empirical distribution of a mutation at various distances from a mutation. Right: The recombination rate is estimated using the empirical distribution of the first mutation from a mutation.

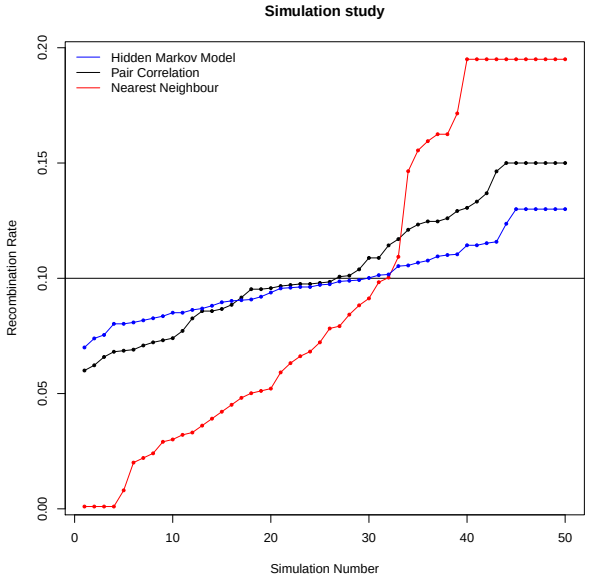

Figure 6: Results of parameter estimation for simulation study. The pair correlation summary performs rather well compared to the full HMM data analysis. Nearest neighbour is a poor summary statistics.

applying the pair correlation function. This is in contrast to the nearest neighbour summary statistics that perform much worse than the other two methods.

We have provided a detailed treatment of the main components involved in an analysis of pair of DNA sequences based on an HMM derived from coalescent theory. Applications and extensions include Li and Durbin (2011) who also consider a model for a pairs of sequences and allows variability in the coalescent rate. In Schiffels and Durbin (2014) the Li-Durbin model is extended to more than two sequences and Terhorst, Kamm and Song (2017) develop a model where hundreds of sequences can be analysed. Rasmussen *et al.* (2014) provide a method for simulating from the posterior distribution of

ancestral recombination graphs.

References

Hobolth, A and Jensen, JL (2014). Markovian approximation to the finite loci coalescent with recombination along multiple sequences. Theoretical Population Biology, 98, 48-58.

Lynch, M, Xu, S, Maruki, T, Jiang, X, Pfaffelhuber, P and Haubold, B (2014). Genome-wide linkage-disequilibrium profiles from single individuals. Genetics, 198, pages 269-81.

Nielsen, SV, Simonsen, S and Hobolth, A (2016). Inferring population genetic parameters: Particle Filtering, HMM, Ripley's K-function or Runs of Homozygosity? In: M Frith and CNS Pedersen (Eds.) WABI 2016, LNBI 9838, pages 234-245.