DOI: 10.1534/genetics.110.124164

# Letter to the Editors

## On Computing the Coalescence Time Density in an Isolation-With-Migration Model With Few Samples

**Asger Hobolth,[1] Lars Nørvang Andersen and Thomas Mailund**

*Bioinformatics Research Centre, Aarhus University, 8000 Aarhus, Denmark*

In a recent article, Wang and Hey (2009) consider estimation of the parameters in an isolation-with-migration (IM) model for two species. For each locus, the data $X$ consist of two samples, and therefore the probability of the data depends only on the time to the most recent common ancestor (MRCA), and we can write the likelihood for a single locus as

$$L(\Theta|X) = P(X|\Theta) = \int_0^\infty P(X|t) f(t|\Theta)\, dt,$$

where $f(t|\Theta)$ is the probability density of the coalescent time. Assuming free recombination between loci, the full likelihood is a product of each locus likelihood.

To determine the likelihood we must determine the density $f(t|\Theta)$ for coalescent of two samples in the IM model. Wang and Hey (2009) find the time to the MRCA by explicitly integrating over all possible sample paths in the system. The purpose of this letter is to demonstrate that the time to the MRCA is easy to compute from a matrix exponential. Furthermore, the matrix exponential framework has the advantage that it generalizes to more than two samples. We describe the IM model, briefly describe the solution to computing the coalescence time density from Wang and Hey (2009), and finally present an approach that computes the density through matrix exponentials.

The description of the model is taken from Wang and Hey (2009) and formulates the IM model as a continuous time Markov chain. Before time $T$ and for two samples the system is in one of the following five states: $S_{11}$, both genes are in population 1; $S_{22}$, both genes are in population 2; $S_{12}$, one gene is in population 1 and the other is in population 2; $S_1$, the genes have coalesced and the single gene is in population 1; and $S_2$, the genes have coalesced and the single gene is in population 2. The rates between the states can be described by the instantaneous rate matrix $Q$ given by

$$Q = \begin{array}{c} \\ S_{11} \\ S_{12} \\ S_{22} \\ S_1 \\ S_2 \end{array} \begin{array}{c} \begin{array}{ccccc} S_{11} & S_{12} & S_{22} & S_1 & S_2 \end{array} \\ \begin{pmatrix} . & 2m_1 & 0 & 2/\theta_1 & 0 \\ m_2 & . & m_1 & 0 & 0 \\ 0 & 2m_2 & . & 0 & 2/\theta_2 \\ 0 & 0 & 0 & . & m_1 \\ 0 & 0 & 0 & m_2 & . \end{pmatrix} \end{array}, \quad (1)$$

where the diagonals are such that each row sums to zero. After time $T$, the system only has two states: $S_{AA}$ corresponding to two genes in the ancestral population and $S_A$ corresponding to one single gene in the ancestral population. The rate of going from state $S_{AA}$ to state $S_A$ is $2/\theta_A$. The model parameters are thus $\Theta = (\theta_1, \theta_2, \theta_A, m_1, m_2, T)$, where $\theta_1$, $\theta_2$, and $\theta_A$ are the scaled population sizes, $m_1$ and $m_2$ are the migration rates and $T$ is the speciation time. We refer to Wang and Hey (2009, Figure 1) for an illustration of the model and for more details on the model parameters.

Now consider the sample path $z = \{z(s) : 0 \le s \le t\}$ shown in Figure 1, where the coalescent happens at time $t < T$ and in population 2. The density for this sample path is given by

$$f(z|\Theta) = \frac{2}{\theta_2}(2m_1)^y m_2^y m_1^{x-y+1}(2m_2)^{x-y} \times \exp\left[-2\left(\frac{1}{\theta_1} + m_1\right)V - (m_1+m_2)U - 2\left(\frac{1}{\theta_2}+m_2\right)W\right],$$

where $x$ and $y$ are the number of transitions to the states $S_{12}$ and $S_{11}$, respectively. The number of transitions from $S_{11}$ to $S_{12}$ and from $S_{12}$ to $S_{11}$ is therefore $y$ and, since the starting state is $S_{12}$, the number of transitions from $S_{12}$ to $S_{22}$ and $S_{22}$ to $S_{12}$ is $x - y + 1$ and $x - y$. The variables $U$, $V$, and $W$ are the total amount of time spent in the states $S_{12}$, $S_{11}$, and $S_{22}$, respectively.

To determine the density $f_2(t|\Theta)$ for coalescent in population 2 at time $t < T$, Wang and Hey (2009) explicitly integrate all possible sample paths that find a MRCA at time $t$. The integration is performed by first integrating all sample paths that share the same values

Available freely online through the author-supported open access option.

[1]*Corresponding author:* Bioinformatics Research Centre, Aarhus University, 8000 Aarhus, Denmark. E-mail: asger@birc.au.dk

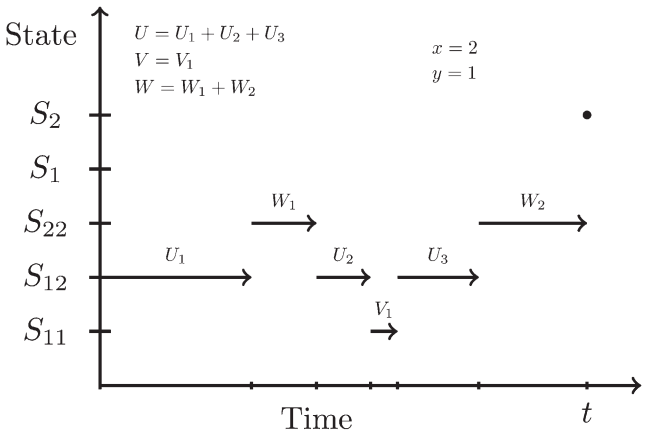


FIGURE 1.—Illustration of a sample path with coalescent in population 2 at time $t$.

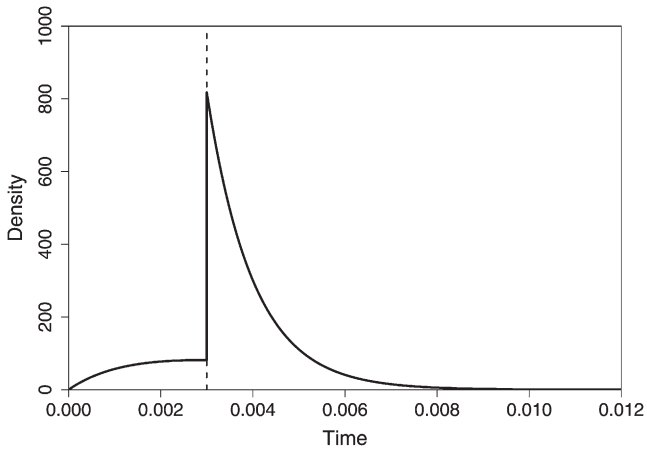


FIGURE 2.—Illustration of the density (4) and (5) for coalescent in the two species isolation with migration model.

for the five variables $(x, y, U, V, W)$; this density is termed $p(x, y, U, V, W|\Theta)$. Second, this expression is summed over variables $(x, y)$ and integrated over variables $(U, V, W)$ with the constraint $U+V+W=t$:

$$f_2(t\,|\,\Theta) = \oiint_{U+V+W=t} \sum_{x=1}^{\infty}\sum_{y=0}^{x} p(x, y, U, V, W\,|\,\Theta).$$

The inner summation becomes a Bessel $I_\alpha(x)$ function and the two-dimensional planar integration is handled numerically.

It is possible, however, to take immediate advantage of the continuous time Markov chain representation (1) and to solve the system of ordinary differential equations analytically. The two samples are either from the same species (corresponding to the starting state being $S_{11}$ or $S_{22}$) or the two samples are from each of the species (in which case the starting state is $S_{12}$). Starting in state $a$, the density for coalescent in population 1, for $t < T$, is given by

$$f_1(t|\Theta) = (e^{Qt})_{aS_{11}}(2/\theta_1), \tag{2}$$

where $e^A$ is the matrix exponential $e^A = \sum_{i=0}^{\infty} A^i/i!$ and $(e^A)_{jk}$ is entry $(j, k)$ in $e^A$. Calculating matrix exponentials has a very long history (*e.g.*, MOLER and VAN LOAN 2003), and implementations are available in most standard computer languages. The density for coalescent in population 2 at time $t < T$ is

$$f_2(t\,|\,\Theta) = (e^{Qt})_{aS_{22}}(2/\theta_2), \tag{3}$$

and the total density for a coalescent at time $t < T$ is

$$f(t\,|\,\Theta) = f_1(t\,|\,\Theta) + f_2(t\,|\,\Theta). \tag{4}$$

Similarly, the density for coalescent in the ancestral population at time $t > T$ is

$$f(t\,|\,\Theta) = \left[(e^{QT})_{aS_{11}} + (e^{QT})_{aS_{12}} + (e^{QT})_{aS_{22}}\right]\frac{2}{\theta_A}\, e^{-(2/\theta_A)(t-T)}. \tag{5}$$

In Figure 2 we illustrate the coalescent density in the two species IM model. We use the same parameters as in the simulation study in WANG and HEY (2009, Table 6): $\theta_1 = 0.005$, $\theta_2 = 0.003$, $\theta_A = 0.002$, $m_1 = 50$, $m_2 = 100$, and $T = 0.003$ (the vertical line).

Multiple analytical approaches for computing the coalescence time density can be found in the literature. Variations of the IM model have been analyzed using Laplace transforms in, *e.g.*, LATTER (1973), TAKAHATA (1995), and WILKINSON-HERBOTS (1998). Their results can also be derived using matrix exponentiation. In WAKELEY (1996), a spectral decomposition was used to obtain a continuous-time approximation to a discrete time model. Generalizations of the IM model dealing with two loci with recombination between loci are analyzed using expressions for continuous time Markov chains in SLATKIN and POLLACK (2006) and SIMONSEN and CHURCHILL (1997).

We finally emphasize that the matrix exponential framework described above generalizes to more than two samples and more than two populations. For any coalescence system that can be expressed as a finite-state homogeneous continuous time Markov chain, we can compute the density of coalescence times using matrix exponentiation. Expressing a coalescence process as such a system is straightforward although a major complication is an explosion in the number of states when the number of samples and populations increase.

If the Markov chain is not homogeneous, that is, the rate matrix $Q$ depends on the time parameter $t$, simple matrix exponentiation is no longer a solution to the coupled set of differential equations. The model of INNAN and WATANABE (2006), for instance, consists of

the same set of states and almost the same rate matrix, but has the migration rates depend linearly on the time variable $t$. For this system, the approach described above cannot be applied.

## LITERATURE CITED


Innan, H., and H. Watanabe, 2006 The effect of gene flow on the coalescent time in the human–chimpanzee ancestral population. Mol. Biol. Evol. **23:** 1040–1047.

Latter, B. D. H., 1973 The island model of population differentiation: a general solution. Genetics **73:** 147–157.

Moler, C., and C. Van Loan, 2003 Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev. **45:** 3–49.

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

Slatkin, M., and J. L. Pollack, 2006 The concordance of gene trees and species trees at two linked loci. Genetics **172:** 1979–1984.

Takahata, J., 1995 A genetic perspective on the origin and history of humans. Annu. Rev. Ecol. Syst. **26:** 343–372.

Wakeley, J., 1996 Pairwise differences under a general model of population subdivision. Genetics **75:** 81–89.

Wang, Y., and J. Hey, 2009 Estimating divergence parameters with small samples from a large number of loci. Genetics **184:** 363–379.

Wilkinson-Herbots, H. M., 1998 Genealogy and subpopulation differentiation under various models of population structure. J. Math. Biol. **37:** 535–585.


Communicating editor: L. Excoffier