Ancestral inference from gene trees in subdivided populations.

Robert C. Griffiths

Department of Statistics
University of Oxford
1 South Parks Rd
Oxford OX1 3TG, UK
griff@stats.ox.ac.uk

1. Coalescent trees and gene trees.

The ancestry of a sample of \( n \) genes can be described by the coalescent process (Kingman 1982). The number of ancestors of the sample looking back in time, \( \{A_n(t), t \geq 0\} \) is a death process with rates \( \mu_k = (k) \), \( k = n, \ldots, 2 \). Time is measured in units of \( N \) generations, where \( N \) is the gene population size, and the model is a limit from a Wright-Fisher model as \( N \to \infty \). The analogue of the coalescent process in a subdivided population model with \( g \) subpopulations is \( \{l_i(t), t \geq 0\} \), where \( l_i(t) \), \( i = 1, \ldots, g \) denotes the number of ancestor lineages in subpopulation \( i \) of an initial sample \( l(0) \). This process has been studied by Notohara (1990), Herbots (1997). A coloured coalescent tree describes ancestry. Associate a colour \( c \) with subpopulation \( c, c = 1, \ldots, g \), and colour the edges of the coalescent tree according to which subpopulation an edge is in. The tree evolves back in time, with coalescence rates within subpopulations of \( \left( \begin{array}{c} n \\ 2 \end{array} \right) q_i^{-1}, i = 1, \ldots, g \), \( \sum_{i=1}^{g} q_i = 1 \), where relative subpopulation sizes are \( q_i, i = 1, \ldots, g \) in the limit as subpopulation sizes tend to infinity. Edges change colour from \( c \) to \( d \) at rate \( m_{cd}, c \neq d, c, d = 1, \ldots, g \). \( M = (m_{cd}) \), with \( m_{cc} = -\sum_{d \neq c} m_{cd} \), is the matrix of backward migration rates. Denoting \( N_1, \ldots, N_g \) as the population sizes, time is measured in units of \( N = N_1 + \cdots + N_g \) generations. Mutations occur along the edges of the coalescent tree at rate \( \theta/2 \). A coalescent tree is illustrated in Figure 1. There are three subpopulations, \( A, B, C \) with \( (2, 3, 2) \) genes in respective subpopulations initially. Different line types represent colours and show which subpopulations lineages are in back in time. Bullets \( \bullet \) show mutations on the edges of the tree.

![Figure 1: Coalescent tree](image1.png)

![Figure 2: Gene tree](image2.png)

In the infinitely-many-sites model of mutation each gene is considered to be an infinitely long DNA sequence, with a mutation always segregating a new site, never before segregating in the population. Thus a mutation occurring on an edge in the coalescent tree appears in all the descendents of that edge.
The observed gene tree corresponding to the coalescent tree in Figure 1 is shown in Figure 2. Vertices of the tree are now mutations, rather than coalescences as in the coalescent tree. A gene tree with multiplicities included is exactly equivalent to a sample configuration of sequences in the infinitely-many-sites model, and its construction from sequences is a perfect phylogeny problem (Gusfield 1991). The gene tree is unique up to permutations of site labels along single edges. For example sites 3 and 4 cannot be distinguished as to which mutation occurred first, but to obtain a definite age ordering we suppose that 3,4 is the age order of mutations.

Some questions of interest are: constructing a gene tree from sequences; determining the root of the tree by maximum likelihood; maximum likelihood estimation of migration rates $M$ and mutation rate $\theta$; detection of population growth by likelihood techniques; determining the distribution of the time to the most recent common ancestor (TMRCA) of a sample of sequences; determining the distribution of the age of the mutations on the gene tree; determining subpopulation ancestors, where they were and times to them; and determining in which subpopulations mutations occurred. Ancestral inference techniques in a panmictic population can be addressed using the methods of Griffiths and Tavaré (1994a,b,c;1998b). Prior studies of maximum likelihood estimation of parameters in subdivided population models are: Beerli and Felsenstein (1998) who use a MCMC approach to estimate rates in molecular data, and Nath and Griffiths (1996) where methods of Griffiths and Tavaré (1994a) are used to estimate rates from allele data.

2. Likelihood surfaces for gene trees.

Let $T$, $n$ denote a gene tree with multiplicities of the lineages $n$ and $p(T, n; \theta, M)$ be the probability of the labelled tree without regard to the order of the sequences in the sample. It is straightforward to simulate coalescent trees such as in Figure 1, however this will not evaluate the likelihood of a specific gene tree which requires an advanced simulation technique. An algorithm of Griffiths and Tavaré (1994a,b,c;1998b) can be used to compute simulated likelihood surfaces $p(T, n; \theta, M)$ with respect to $\theta, M$. The algorithm in its simplest form is described as follows. A recursive equation for the probability of a gene tree can be expressed in the form

\begin{equation}
 p(T, n; \theta, M) = f(T, n; \theta, M) \sum P(T', n' | T, n; \theta, M)p(T, n'; \theta, M),
\end{equation}

where \{P(T', n' | T, n; \theta, M)\} are transition functions of a Markov chain constructed by scaling coefficients to add to 1, and $f(T, n; \theta, M)$ is a scale constant. The gene trees $(T', n')$ are the possible parent trees of the sample formed by either coalescence, removal of a singleton mutation, or migration as the last event back in time. The Markov chain has a tree state space which has absorbing states of singleton trees, corresponding to the ancestor, in respective subpopulations. Denoting $\tau$ as the step at which the chain is absorbed, starting from $(T, n)$ and $(T(j), n(j)), j = 0, \ldots, \tau$ the sample path of the chain, (1) implies a representation

\begin{equation}
 p(T, n; \theta, M) = E_{(T, n)}(F), \quad \text{where} \quad F = \prod_{j=1}^{\tau-1} f(T(j), n(j); \theta, M).
\end{equation}

The representation makes it possible to estimate the likelihood by independently, repeatedly, simulating the Markov chain and taking the average of the functional values

\begin{equation}
 \hat{p}(T, n; \theta, M) = F,
\end{equation}

where $F_1, F_2, \ldots$ are functionals evaluated on respective simulation runs. Entire likelihood surfaces can be generated by importance sampling, running the chain at fixed parameter values $\theta_0, M_0$. Likelihood surfaces generated by the average functional over repeated simulation runs
using importance sampling are smooth, in contrast to a surface built up of single point estimates from (2). The evolutionary variance is decreased by using the same coalescent tree on each run for the likelihood surface, which is however only locally accurate around \( \theta_0, M_0 \).

Felsenstein calls the embedded Markov chain where events occur in the coalescent process (coalescence, migration or mutation) a history and relates the representation (2) to importance sampling from histories. Current research exploits this idea to construct efficient algorithms for likelihood computations, particularly in models of DNA sequences when there is possibly back mutation at sites, Stephens (1999).

### 3. Ancestral inference from a gene tree.

The algorithm for computing \( p(T, n) \) can be extended to compute functionals on the coalescent tree which include time information, for example the TMRCA and ages of mutations. The stochastic process with a gene tree state space is then a Markov process with jump rate

\[
\sum_{i=1}^{g} \left( \frac{l_i}{2} \right) q_i^{-1} + \sum_{i=1}^{g} a_i |m_{kk}| + n\theta/2.
\]

Transitions are made according to an embedded Markov chain with transition probabilities \( P(T', n'|T, n; \theta, M) \). In the case of the TMRCA and ages of mutations each simulation run generates an estimate of the likelihood \( F \) (in equation 2), together with ages of mutations \( a = (a_1, \ldots, a_k) \), and the TMRCA \( \xi \). An empirical discrete distribution of ages and TMRCA from \( r \) simulation runs is then \((a_1, \xi_1, p_1), \ldots,(a_r, \xi_r, p_r)\) where \( p_i = F_i / \sum_{j=1}^{r} F_j, i = 1, \ldots, r \). The age and TMRCA distribution is interpreted as being conditional on the topology of the gene tree without time information. Details of the algorithms mentioned here are developed in detail with examples in Bahlo and Griffiths (1999).

In the example tree suppose that \( \theta = 2.5 \) and the migration rate is 2.0 between each pair of subpopulations. The gene tree in Figure 2 is drawn to scale with expected ages of mutations. The mean TMRCA is 1.8 with standard deviation 0.6. The MRCA is very close to having occurred with equal probability in each subpopulation. Because of the high migration rate mutations higher up in the tree have almost equal probability of occurring in each subpopulation, but the mutations lower in the tree at sites 1 and 6 are more interesting. These have respective probabilities of occurring in \( A, B, C \) of 0.26, 0.25, 0.49 for site 1, and 0.14, 0.71, 0.15 at site 6.

Griffiths and Tavaré (1998), Stephens (1999), Wiuf and Donnelly (1999) study the distribution of the age of a mutation at a site which subtends \( b \) mutant copies and \( n - b \) non-mutant copies in a sample of \( n \) genes and the distribution of the TMRCA in such a sample.

An extension to a gene tree is to consider the marginal distribution of particular sites only. This is appropriate in sequence data which has just been sequenced for particular sites, and can be done to consider the geographical information from particular mutant sites. Then the probability of a genetree \( (T, n) \), given \( k \) particular segregating sites is

\[
\lim_{\theta_1, \ldots, \theta_k \to 0} \frac{p(T, n; \theta_1, \ldots, \theta_k, M)}{E\left(1 - e^{-\theta_i L/2}\right) \cdots \left(1 - e^{-\theta_k L/2}\right)}
\]

where \( \theta_1, \ldots, \theta_k \) are rates at the \( k \) sites, and \( L \) is the total edge length of the coalescent tree.

In the example tree the mutation at site 3 occurs respectively in \( A, B, C \) 2, 1, 1 times. Conditional on this configuration the mean and standard deviation of the age is 1.6, 1.2 and for the TMRCA 3.0, 1.7. The times are much older than in the full gene tree where the mean age is 0.46 and mean TMRCA 1.8. The probabilities that the mutation occurred respectively in \( A, B, C \) are 0.35, 0.32, 0.33. The simulated densities of the age distribution of the mutation at site 3 in the complete gene tree with \( \theta = 2.5 \), and the marginal density just assuming the configuration of the single mutation are shown in Figure 3.
A program GENETREE written in portable C code implements the algorithms mentioned in this paper. It is available from the Mathematical Genetics page at www.stats.ox.ac.uk.

REFERENCES


