Lecture 8. Coalescent likelihoods and introduction to MCMC

Joe Felsenstein

Department of Genome Sciences and Department of Biology
Some typical data with within-population variation

\[ L = \text{Prob (CAGTTTCAGCGTCC, CAGTTTCAGCGTCC, ...)} = ?? \]

To infer parameters of evolutionary-genetic models, we need to compute the likelihood for a set of genotypes sampled from a population. With few exceptions, no expressions for this likelihood exist.
If we knew the genealogical tree, we could ...

... as we know from work on phylogenies how to compute the joint probability at the tips.

so we can compute

$$\text{Prob}(\text{CAGTTTCAGCGTCC}, \text{CAGTTTCAGCGTCC}, \ldots \mid \text{Genealogy})$$

but how to compute the overall likelihood from this?
Two sources of variation

1. Randomness of mutation
   - Affected by the mutation rate $u$
   - Can reduce variance of number of mutations per site per branch by examining more sites

2. Randomness of coalescence of lineages
   - Affected by effective population size $N_e$
   - Coalescence times allow estimation of $N_e$
   - Can reduce variability by looking at
     - (i) more gene copies, or
     - (ii) more loci
The basic equation for coalescent likelihoods

In the case of a single population with parameters

\( N_e \) effective population size

\( \mu \) mutation rate per site

and assuming \( G' \) stands for a coalescent genealogy and \( D \) for the sequences,

\[
L = \text{Prob} (D \mid N_e, \mu)
\]

\[
= \sum_{G'} \text{Prob} (G' \mid N_e) \text{Prob} (D \mid G', \mu)
\]

\underbrace{\text{Kingman's prior}} \quad \underbrace{\text{likelihood of tree}}
Rescaling branch lengths ...

Rescaling branch lengths of $G'$ so that branches are given in expected mutations per site, $G = \mu G'$, we get (if we let $\Theta = 4N_e \mu$)

$$L = \sum_{G'} \text{Prob} (G \mid \Theta) \text{Prob} (D \mid G)$$

as the fundamental equation. For more complex population scenarios one simply replaces $\Theta$ with a vector of parameters.
In a simple example we can compute likelihood curve

If two sequences of length 1000 differ by 0.5% under a Jukes-Cantor model

\[ L(N, \mu) = \text{Prob (5 different } | \mu, t) \]

\[ = \int_0^\infty \frac{1}{2N} e^{-\frac{t}{2N}} \left(\frac{1000}{5}\right) \left(\frac{3}{4} \left(1 - e^{-\frac{4}{3}\mu(2t)}\right)\right)^5 \left(\frac{1}{4} \left(1 + 3 e^{-\frac{4}{3}\mu(2t)}\right)\right)^{995} \text{dt} \]

In the product, the left two terms are the exponential density from the Kingman coalescent time density, the right three terms are the binomial probability which is the likelihood for a tree whose branch length is \(2\mu t\).

Replacing \(\mu t\) by a new variable \(u\), this is a function only of \(4N\mu\)

\[ \int_0^\infty \frac{2}{4N\mu} e^{-\frac{2u}{4N\mu}} \left(\frac{1000}{5}\right) \left(\frac{3}{4} \left(1 - e^{-\frac{8}{3}u}\right)\right)^5 \left(\frac{1}{4} \left(1 + 3 e^{-\frac{8}{3}u}\right)\right)^{995} \text{du} \]

and can be easily evaluated numerically to get \(L(4N\mu)\).
If two sequences with 1000 bases 0.5% different

This is the likelihood (right) term in the preceding equation, the binomial probability that 5 sites would differ out of 1000, as a function of the branch length $2\mu t$
If two sequences with 1000 bases 0.5% different

For a given value of $\Theta$, 0.005, we plot the two-tip Kingman coalescent density function of the branch length (twice the time to coalescence), and we can integrate the product of that with the likelihood curve. The right graph shows that integral plotted against $\Theta$. 

0.000 0.005 0.010 0.015 0.020 0.025 0.030

2$\mu$t

0.000 0.005 0.010 0.015 0.020 0.025 0.030

$\Theta$

$L$

0 5 x 10^{-620} 1.0 x 10^{-619} 1.5 x 10^{-619} 2.0 x 10^{-619} 2.5 x 10^{-619} 3.0 x 10^{-619}
If two sequences with 1000 bases 0.5% different

On the right is the same integral, but now using instead the Kingman coalescent density for $\Theta = 0.01$. 
If two sequences with 1000 bases 0.5% different

Finally, we can do it also for $\Theta = 0.02$, and connect the three points in the right-hand graph by a curve (this graph shows the curve we would get if we did a finer grid of points).
Labelled histories

(Edwards, 1970; Harding, 1971)

Trees that differ in the time-ordering of their nodes

These two are different:

A  B  C  D
A  B
    C
     D

These two are the same:

A  B  C  D
A  B
    C
     D
The number of labelled histories

The labelled history is essentially a list of the pairs of lineages that coalesce, in order. So the number of these is

\[ \frac{n(n-1)}{2} \frac{(n-1)(n-2)}{2} \frac{(n-2)(n-3)}{2} \cdots \frac{2 \times 1}{2} \]

\[ = \frac{n!(n-1)!}{2^{n-1}} \]
The number of these rises rapidly:

<table>
<thead>
<tr>
<th>Tips</th>
<th>Labelled histories</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>180</td>
</tr>
<tr>
<td>6</td>
<td>2700</td>
</tr>
<tr>
<td>7</td>
<td>56,700</td>
</tr>
<tr>
<td>8</td>
<td>1,587,600</td>
</tr>
<tr>
<td>9</td>
<td>57,153,600</td>
</tr>
<tr>
<td>10</td>
<td>2,571,912,000</td>
</tr>
</tbody>
</table>
Monte Carlo integration

To get the area under a curve, we can either evaluate the function \( f(x) \) at a series of grid points and add up heights \( \times \) widths:

or we can sample at random the same number of points, add up height \( \times \) width:
The importance sampling formula

Expectation of a function $h(x)$ over a distribution whose density function is $g(x)$:

$$E_g[h(x)] = \int x h(x) g(x) \, dx$$
Importance Sampling

The function we integrate:

\[ f(x) \]

We sample from this density:

\[ g(x) \]
The integral can be computed as follows:

\[ \int f(x) \, dx = \int \frac{f(x)}{g(x)} \, g(x) \, dx \]

\[ = E_g \left[ \frac{f(x)}{g(x)} \right] \]

\[ \approx \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)} \]

(where the sample points \( x_i \) are drawn from density \( g(x) \))

In effect, each point sampled is taken to be a histogram bar whose width is less if it sampled from an area where \( g(x) \) is larger, so that more samples get taken from that area.
Accuracy is improved by importance sampling

Grid
Monte Carlo
Importance
Transition probabilities that achieve a given distribution

If we desire a particular equilibrium distribution $\pi_i$ then one way to achieve it is to run a Markov chain that has transition probabilities that achieve \textit{detailed balance}, so that for each pair of states the fraction of cases that move from $i$ to $j$ is the same as the fraction that move from $j$ to $i$. If $P_{ij}$ is the conditional probability of going from $i$ to $j$ then we achieve this if:

$$\pi_i \cdot P_{ij} = \pi_j \cdot P_{ji}$$
So if $g_i$ is proportional to the desired distribution,

$$\frac{P_{ij}}{P_{ji}} = \frac{g_j}{g_i}$$

Any choice of $P$’s that satisfies this is OK. To move around as fast as possible, suppose $g_j > g_i$. Then when $j$ is proposed from $i$, accept it always. When $i$ is proposed from $j$, accept it with probability $g_i/g_j$. So we use $P_{ij} = 1$ and $P_{ji} = g_i/g_j$. 
MCMC: The Metropolis-Hastings method

To draw a sample \( G_1, \ldots, G_n \) from a distribution proportional to a function \( g(G) \):

1. Draw a change in \( G \) from some “proposal distribution”: \( x \rightarrow y \)

2a. (Metropolis et. al., 1953):
   Accept the change if a uniformly-distributed random number \( R \) satisfies
   \[ R < \frac{g(y)}{g(x)} \]

2b. Hastings (*Biometrika*, 1970) corrected for biases toward some \( y \)’s in the proposal distribution by using instead
   \[ R < \frac{\text{Prob}(x|y)}{\text{Prob}(y|x)} \frac{g(y)}{g(x)} \]

Repeat many times. If we do this long enough, and various niceness conditions hold, then \( G_1, \ldots, G_m \) will be a sample from the right distribution.
An aside – the Gibbs Sampler

If the proposal distribution is proportional to the relative probabilities $g(y)$ of all states obtained by removing one coordinate $z_i$ from $(z_1, z_2, \ldots, z_n)$ and then sampling proportional to the relative probabilities of all possible values that could be put there ...

Then the Hastings ratio for going from $x$ to $y$ is $g(x)/g(y)$, which leads to a perfect cancellation – you should always accept!

When you can do this it is generally a very fast method.
Computing coalescent likelihoods by MCMC

We want to compute \( \int_G \text{Prob} \ (G|\Theta)\text{Prob} \ (D|G)\,dG \). We use an importance sampling density proportional to the interior of the integral at some trial value \( \Theta_0 \) of the parameter. Then it is

\[
g(G) = \frac{\text{Prob} \ (G|\Theta_0)\text{Prob} \ (D|G) \,dG}{\int_G \text{Prob} \ (G|\Theta_0)\text{Prob} \ (D|G) \,dG}
\]

whose denominator is

\[
L(\Theta_0) = \int_G \text{Prob} \ (G \mid \Theta_0)\text{Prob} \ (D \mid G) \,dG
\]
The integral is:

\[ L(\Theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{f(G_i)}{g(G_i)} \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \frac{\text{Prob}(G|\Theta)\text{Prob}(D|G)}{\text{Prob}(G|\Theta_0)\text{Prob}(D|G)/L(\Theta_0)} \]

This leads to

\[ \frac{L(\Theta)}{L(\Theta_0)} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(G_i)}{g(G_i)} = \frac{1}{n} \sum_{i=1}^{n} \frac{\text{Prob}(G_i|\Theta)}{\text{Prob}(G_i|\Theta_0)} \]

Note that in the computation of the likelihood ratio, the terms for the probability of the data have disappeared! Where is the data? Its influence is on the importance sampling, where in tree space the trees are sampled from.
Tree rearrangements proposed:

A conditional coalescent rearrangement strategy
more ...

First pick a random node (interior or tip) and remove its subtree
Then allow this node to re-coalesce with the tree
and finally we get:

The resulting tree proposed by this process
Being left out of the story:

- We choose the rearrangements so that the proposal distribution is a “conditional coalescent”.
- We do a Hastings correction given this.
- The end result is a perfect cancellation (which is pleasant rather than essential).
- This leaves us with the rule that we use $\text{Prob}(D | G)$ as the only function in the Metropolizing.
One ends up with a curve that might look like this:

Results of analysing a data set with 50 sequences of 500 bases which was simulated with a true value of $\Theta = 0.01$
A coalescent and a species tree
Integrating over all coalescents

Evaluating fit of multiple loci to a species tree
Summing over all possible genealogies at each locus

\[ L = \text{Prob (Data} \mid \text{Species tree}) \]

\[ = \prod_{\text{loci}} \sum_{\text{coalescent trees}} \text{Prob (coalescent} \, i \mid \text{Species tree}) \times \text{Prob (Data} \, i \mid \text{coalescent} \, i) \]

This involves integrating over all coalescent trees, separately at each locus (if the loci are not closely linked). If they are closely linked then one would jointly integrate them over a coalescent which has recombinations.
New genetic tools being deployed

Likelihood or Bayesian inference using sampling methods with coalescents

Mig = Migration, Rec = Recombination, Grow = Population growth, Split = Splittings, Bayes = Bayesian

<table>
<thead>
<tr>
<th>Program Name</th>
<th>Mig?</th>
<th>Rec?</th>
<th>Grow?</th>
<th>Split?</th>
<th>Bayes?</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAMARC (Kuhner, Yamato et al.)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>n</td>
<td>y/n</td>
</tr>
<tr>
<td>BEAST (Drummond, Rambaut)</td>
<td>n</td>
<td>n</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Genetree (Griffiths and Bahlo)</td>
<td>Y</td>
<td>n</td>
<td>Y</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Migrate (Beerli)</td>
<td>Y</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>Y</td>
</tr>
<tr>
<td>IM, Ima (Nielsen &amp; Hey)</td>
<td>Y</td>
<td>n</td>
<td>y/n</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Batwing (Wilson and Balding)</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>
References


References

Felsenstein, J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research* **59**: 139-147. [Suggests using the coalescents]


References


References


References

Nielsen, R. 1997. Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theoretical Population Biology* **53**: 143-151. [The first coalescent likelihood paper with more than one species]

