Lecture 29. Coalescents, part 2. (Likelihoods and introduction to MCMC)

Joe Felsenstein

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

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.

\[ L = \text{Prob} \left( \text{CAGTTTCAGCGTCC} , \text{CAGTTTCAGCGTCC} , \ldots \right) = ?? \]
But there is a way to compute it

However if we knew the genealogical tree connecting the haplotypes we know from work on phylogenies (evolutionary trees) how to compute the probability of the sample at the tips of that tree

so we can compute

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

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

Two levels of variability

(1) Randomness of mutation

\( u \)

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

\( N_e \)

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

\[ N_e \quad \text{effective population size} \]
\[ \mu \quad \text{mutation rate per site} \]

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

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

\[ \text{Kingman's prior} \quad \text{likelihood of tree} \]
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 $\Theta = 4N_e\mu$

\[
L = \sum_G \text{Prob} \left( G \mid \Theta \right) \text{Prob} \left( D \mid G \right)
\]

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

The likelihood calculation in a sample of two gene copies

The product of the prior on \( t \),

\[
\Theta_1, \Theta_2, \Theta_3
\]

times the likelihood of that \( t \) from the data,

when integrated over all possible \( t \)'s, gives the likelihood for the underlying parameter \( \Theta \)
Labelled histories

Labelled Histories (Edwards, 1970; Harding, 1971)

Trees that differ in the time-ordering of their nodes

These two are different:

These two are the same:
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

\[
\frac{n(n-1)}{2} \cdot \frac{(n-1)(n-2)}{2} \cdot \frac{(n-2)(n-3)}{2} \cdot \cdots \cdot \frac{2 \times 1}{2} = \frac{n!(n-1)!}{2^{n-1}}
\]
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

We sample from this density

\[ f(x) \]

\[ g(x) \]
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 \(g(x)\))
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 *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:

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

$$P_{ij}/P_{ji} = g_j/g_i$$

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

$$R < \frac{\text{Prob}(x|y) \cdot g(y)}{\text{Prob}(y|x) \cdot g(x)}$$

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.
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

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

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

A conditional coalescent rearrangement strategy
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 $\text{Prob } (D \mid G)$ 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 $\Theta = 0.01$
Griffiths’ and Tavaré’s (1994) method

They sample sequences of events (coalescences, mutations, etc.) – these have no times but show mutations explicitly
Griffiths and Tavaré’s method as importance sampling

\( D \) the data (sequences)

\( \beta \) the parameters (\( 4N_e \mu \) and such)

\( H_i \) the \( i \)-th of all possible histories of events

\( h_{ij} \) the \( j \)-th event in history \( H_i \)

\( a_{ijk}(\beta) \) the probability (rate) of the \( k \)-th of the possible events that could have happened at stage \( j \) of history \( i \) (ignoring the data).
Some definitions

\( b_{ij}(\beta) \) the probability (rate) of the one that did happen at stage \( j \) of history \( i \)

\( c_{ijk}(\beta) \) the probability (rate) of the \( k \)-th of the possible events that could have happened at stage \( j \) of history \( i \) (counting only those that are compatible with the data).
The Griffiths-Tavaré method

\[ L = \sum_H \text{Prob} \left( H \mid \beta \right) \text{Prob} \left( D \mid H \right) = E_f \text{Prob} \left( D \mid H \right) \]

The distribution \( f \) is:

\[ \text{Prob} \left( H_i \mid g \right) = \prod_j \frac{b_{ij}(\beta)}{\sum_k a_{ijk}(\beta)} = \frac{\prod_j b_{ij}(\beta)}{\prod_j (\sum_k a_{ijk}(\beta))} \]

(the distribution \( g \) is the same but with \( c \)'s instead of \( a \)'s).

\[ L(\beta) = E_f \left[ \text{Prob} \left( D \mid H \right) \right] = E_g \left[ \frac{f}{g} \text{Prob} \left( D \mid H \right) \right] \]
We end up with

\[
L(\beta) = \mathbb{E}_g \left[ \frac{\Pi_j b_{ij}(\beta)}{\Pi_j \left( \sum_k a_{ijk}(\beta) \right)} \right] \\
= \mathbb{E}_g \left[ \Pi_j \left( \frac{b_{ij}(\beta)}{b_{ij}(\beta_0)} \right) \left( \frac{\sum_k c_{ijk}(\beta_0)}{\sum_k a_{ijk}(\beta)} \right) \right]
\]
A coalescent and a species tree

Gene tree and Species tree

N_1, N_2, N_3, N_4, N_5

t_1, t_2
Integrating over all coalescents

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

\[ L = \text{Prob} \left( \text{Data} \mid \text{Tree} \right) \]

\[ = \prod_{\text{loci}} \sum_{\text{trees}} \text{Prob} \left( \text{coalescent } i \mid \text{Species tree} \right) \times \text{Prob} \left( \text{Data } i \mid \text{coalescent } i \right) \]
## 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, Beerli et al.)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>BEAST (Drummond, Rambaut, Pybus)</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>Batwing (Wilson and Balding)</td>
<td>n</td>
<td>n</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>MDIV (Nielsen)</td>
<td>Y</td>
<td>n</td>
<td>n</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>
The LAMARC package

LAMARC

Our LAMARC package of coalescent likelihood programs can be found at

http://evolution.gs.washington.edu/lamarc.html

The original program, released in 1995, was:

**COALESCE** Estimated 4Nu in a single population of constant size

There are at present four programs in distribution

**FLUCTUATE** Estimates 4Nu and g in an exponentially growing population

**MIGRATE** Estimates 4Nu and 4Nm in an n-population case

**RECOMBINE** Estimates 4Nu and 4Nr with recombination and one population

**LAMARC** The new combined program

These are available as generic C source code

and PowerMac and Windows executables
Now available ... LAMARC (the program)

Problem: too many combinations of forces for us to write one program for each combination.
Also, using one program with all forces present may require too many resources.

Solution?
Use an object-oriented language (C++ or Java)
Build a program that can "self-assemble" in response to the user's choices
Only these features needed will be used at run time

- Pop. growth
- Migration
- Recombination
- Selection
- Splitting of pops.

User 1

User 2

User 3

User n
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]
How it was done

This projection produced as a PDF, not a PowerPoint file, and viewed using the Full Screen mode (in the View menu of Adobe Acrobat Reader):

- using the *prosper* style in LaTeX,
- using Latex to make a `.dvi` file,
- using `dvips` to turn this into a Postscript file,
- using `ps2pdf` to mill it into a PDF file, and
- displaying the slides in Adobe Acrobat Reader.

Result: nice slides using freeware.