Lecture 33. Phylogeny methods, part 5 (Likelihood methods)

Joe Felsenstein

Department of Genome Sciences and Department of Biology
Likelihoods and odds ratios

Bayes’ Theorem relates prior and posterior probabilities of an hypothesis $H$:

$$
\text{Prob} \ (H|D) = \frac{\text{Prob} \ (H \text{ and } D)}{\text{Prob} \ (D)} = \frac{\text{Prob} \ (D|H) \ \text{Prob} \ (H)}{\text{Prob} \ (D)}
$$

The ratios of posterior probabilities of two hypotheses, $H_1$ and $H_2$ can be written, putting this into its “odds ratio” form ($\text{Prob} \ (D)$ cancels):

$$
\frac{\text{Prob} \ (H_1|D)}{\text{Prob} \ (H_2|D)} = \frac{\text{Prob} \ (D|H_1)}{\text{Prob} \ (D|H_2)} \frac{\text{Prob} \ (H_1)}{\text{Prob} \ (H_2)}
$$

Note that this says that the posterior odds in favor of $H_1$ over $H_2$ are the product of prior odds and a likelihood ratio. The likelihood of the hypothesis $H$ is the probability of the observed data given it, $\text{Prob} \ (D|H)$. This is not the same as the probability of the hypothesis given the data. That is the posterior probability of $H$ and requires that we also have a believable prior probability $\text{Prob} \ (H)$.
Rationale of likelihood inference

If the data consists of $n$ items that are conditionally independent given the hypothesis $H_i$,

$$\text{Prob}\ (D|H_i)$$

$$= \text{Prob}\ (D^{(1)}|H_i) \ \text{Prob}\ (D^{(2)}|H_i) \ ... \ \text{Prob}\ (D^{(n)}|H_i).$$

and we can then write the likelihood ratio $\text{Prob}\ (D|H_1) / \text{Prob}\ (D|H_2)$ as a product of ratios:

$$\frac{\text{Prob}\ (D|H_1)}{\text{Prob}\ (D|H_2)} = \left( \prod_{i=1}^{n} \frac{\text{Prob}\ (D^{(i)}|H_1)}{\text{Prob}\ (D^{(i)}|H_2)} \right)$$

If the amount of data is large the likelihood ratio terms will dominate and push the result towards the correct hypothesis. This can console us somewhat for the lack of a believable prior.
Properties of likelihood inference

Likelihood inference has (usually) properties of

- Consistency. As the number of data items $n$ gets large, we converge to the correct hypothesis with probability 1.

- Efficiency. Asymptotically, the likelihood estimate has the smallest possible variance (it need not be best for any finite number $n$ of data points).
A simple example – coin tossing

If we toss a coin which has heads probability $p$ and get HHTTHTHHTTT the likelihood is

$$L = \text{Prob } (D|p)$$


$$= p^5(1 - p)^6$$

so that trying to maximize it we get

$$\frac{dL}{dp} = 5p^4(1 - p)^6 - 6p^5(1 - p)^5$$
finding the ML estimate

and searching for a value of $p$ for which the slope is zero:

$$\frac{dL}{dp} = p^4(1 - p)^5 (5(1 - p) - 6p) = 0$$

which has roots at $p = 0$, $p = 1$, and $p = 5/11$
Log likelihoods

Alternatively, we could maximize not $L$ but its logarithm. This turns products into sums:

$$\ln L = 5 \ln p + 6 \ln(1 - p)$$

whereby

$$\frac{d(\ln L)}{dp} = \frac{5}{p} - \frac{6}{1 - p} = 0$$

so that finally

$$\hat{p} = \frac{5}{11}$$
Likelihood curve for coin tosses

\[ p = 0.454 \]
A tree, with branch lengths, and the data at a single site. This example is used to describe calculation of the likelihood. Since the sites evolve independently on the same tree, the likelihood is given by:

\[ L = \text{Prob} (D|T) = \prod_{i=1}^{m} \text{Prob} \left( D^{(i)}|T \right) \]
Likelihood at one site on a tree

We can compute this by summing over all assignments of states $x$, $y$, $z$ and $w$ to the interior nodes

$$\text{Prob} \ (D^{(i)}|T) = \sum_{x} \sum_{y} \sum_{z} \sum_{w} \text{Prob} \ (A, C, C, C, G, x, y, z, w|T)$$
Computing the terms

For each combination of states, the Markov process allows us to express it as a product of probabilities of a series of changes, with the probability that we start in state $x$:

$$
\text{Prob} \ (A, C, C, C, G, x, y, z, w | T) = \\
\text{Prob} \ (x) \ \text{Prob} \ (y | x, t_6) \ \text{Prob} \ (A | y, t_1) \ \text{Prob} \ (C | y, t_2) \\
\text{Prob} \ (z | x, t_8) \ \text{Prob} \ (C | z, t_3) \\
\text{Prob} \ (w | z, t_7) \ \text{Prob} \ (C | w, t_4) \ \text{Prob} \ (G | w, t_5)
$$
Computing the terms

Summing this up, there are \(4^4 = 256\) terms in this case:

\[
\text{Prob } (D^{(i)}|T) = \sum \sum \sum \sum \\
\text{Prob } (x) \text{ Prob } (y|x, t_6) \text{ Prob } (A|y, t_1) \text{ Prob } (C|y, t_2) \text{ Prob } (z|x, t_8) \text{ Prob } (C|z, t_3) \text{ Prob } (w|z, t_7) \text{ Prob } (C|w, t_4) \text{ Prob } (G|w, t_5)
\]
Getting a recursive algorithm

This seems hopeless, but when we move the summation signs as far right as possible

\[
\text{Prob } (D^{(i)}|T) = \\
\sum_x \text{Prob } (x) \\
\left( \sum_y \text{Prob } (y|x, t_6) \text{ Prob } (A|y, t_1) \text{ Prob } (C|y, t_2) \right) \\
\left( \sum_z \text{Prob } (z|x, t_8) \text{ Prob } (C|z, t_3) \right) \\
\left( \sum_w \text{Prob } (w|z, t_7) \text{ Prob } (C|w, t_4) \text{ Prob } (G|w, t_5) \right)
\]
The pruning algorithm

Note that the pattern of parentheses in the previous expression is the

$$(A, C') (C, (C, G))$$

If $L_k^{(i)}(s)$ is the probability of everything that is observed from node $k$ on the tree on up, at site $i$, conditional on node $k$ having state $s$, we can express

$$\left( \sum_w \text{Prob} \ (w | z, t_7) \text{Prob} \ (C | w, t_4) \text{Prob} \ (G | w, t_5) \right)$$

as:

$$\left( \sum_w \text{Prob} \ (w | z, t_7) L_k^{(i)}(w) \right)$$
and the algorithm is:

Continuing with this we find that the following algorithm computes the $L_k$'s from the $L_\ell$ and $L_m$ above them,

$$L_k^{(i)}(s) = \left( \sum_x \text{Prob } (x|s, t_\ell) L_\ell^{(i)}(x) \right) \times \left( \sum_y \text{Prob } (y|s, t_m) L_m^{(i)}(y) \right)$$
Starting and finishing the recursion

At the top of the tree the definition of the \(L\)’s specifies that they look like this

\[
\left( L^{(i)}(A), L^{(i)}(C), L^{(i)}(G), L^{(i)}(T) \right) = (1, 0, 0, 0)
\]

and at the bottom the likelihood for the whole site can be computed simply by weighting by the equilibrium state probabilities

\[
L^{(i)} = \sum_{x} \pi_{x} L^{(i)}_{0}(x)
\]
Ambiguity and error in the sequences

Ambiguity. If a tip has an ambiguity state such as R (purine, either A or G) we use

\[ L^{(i)} = (1, 0, 1, 0) \]

and if it has an unknown nucleotide ("N")

\[ L^{(i)} = (1, 1, 1, 1) \]

This handles ambiguities naturally.

Error. If our sequencing has probability \(1 - \varepsilon\) of finding the correct nucleotide, and \(\varepsilon/3\) of inferring each of the three other possibilities, when an A is observed, the four values should be \((1 - \varepsilon, \varepsilon/3, \varepsilon/3, \varepsilon/3)\), and when a C is observed, they should be \((\varepsilon/3, 1 - \varepsilon, \varepsilon/3, \varepsilon/3)\)

The result is a simple handling of sequencing error, provided it occurs independently in different bases.
The tree is effectively unrooted

The region around nodes 6 and 8 in the tree, when a new root (node 0) is placed in that branch. The subtrees are shown as shaded triangles.

For the tree on the left of the figure above,

\[ L^{(i)} = \sum_{y} \sum_{z} \sum_{x} \text{Prob}(x) \text{Prob}(y|x, t_6) \text{Prob}(z|x, t_8). \]
using reversibility ...

Reversibility of the substitution process guarantees us that

\[ \text{Prob} (x) \text{Prob} (y|x, t_6) = \text{Prob} (y) \text{Prob} (x|y, t_6). \]

Substituting, we get

\[ L^{(i)} = \sum_y \sum_z \sum_x \text{Prob} (y) \text{Prob} (x|y, t_6) \text{Prob} (z|x, t_8) \]

Finally we see that this is the same as the likelihood for a tree rooted at node 8:

\[ L_0^{(i)} (z) = L_8^{(i)} (z) \text{Prob} (z) \text{Prob} (w|z, t_6) L_6^{(i)} (w) \]
A numerical example

A 232-nucleotide mitochondrial noncoding region data set over 14 species gives this ML tree with $\ln L = -2616.86$ with a transition/transversion ratio of 30.
Bayesian inference with coin tossing:

Bayesian methods

An example of Bayesian inference with coin-tossing. The probability of heads is assumed to have a prior (top) which is a truncated exponential with mean 0.34348 on the interval (0,1). The likelihood curve (middle) and the posterior on the probability of heads (bottom) are shown, when there are 11 tosses with 5 heads.
Bayesian phylogeny methods

Bayesian inference has been applied to inferring phylogenies (Rannala and Yang, 1996; Mau and Larget, 1997; Li, Pearl and Doss, 2000).

- All use a prior distribution on trees. The prior has enough influence on the result that its reasonableness should be a major concern. In particular, the depth of the tree may be seriously affected by the distribution of depths in the prior.

- All use Markov Chain Monte Carlo (MCMC) methods (we will introduce these in our discussion of coalescents) They sample from the posterior distribution.

- When these methods make sense they not only get you a point estimate of the phylogeny, they get you a distribution of possible phylogenies.
References


References


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

Result: nice slides using freeware.