Lecture 5. Likelihood and Bayesian methods. HMM rate variation models.

Joe Felsenstein

Department of Genome Sciences and Department of Biology
A simple example of likelihood – coin tossing

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

\[
L = \text{Prob} \left( D|p \right) \\
= p^5(1 - p)^6
\]

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 0, 1, and 1
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
Likelihood on trees

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,

\[
L = \text{Prob} (D|T) = \prod_{i=1}^{m} \text{Prob} \left( D^{(i)}|T \right)
\]

Lecture 5. Likelihood and Bayesian methods. HMM rate variation models. – p.6/37
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

$$\text{Prob } (D^{(i)}|T) = \sum\sum\sum\sum \text{Prob } (A, C, C, C, G, x, y, z, w|T)$$
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 $x$:

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

Summing this up, there are 256 terms in this case:

$$\sum_{x} \sum_{y} \sum_{z} \sum_{w}$$

\begin{align*}
\text{Prob} (x) & \quad \text{Prob} (y | x, t_6) & \quad \text{Prob} (A | y, t_1) \cdot \text{Prob} (C | y, t_2) \\
\text{Prob} (z | x, t_8) & \quad \text{Prob} (C | z, t_3) \\
\text{Prob} (w | z, t_7) & \quad \text{Prob} (C | w, t_4) \cdot \text{Prob} (G | w, t_5)
\end{align*}
Getting a recursive algorithm

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

\[
\begin{align*}
\text{Prob } (D^{(i)}|T) &= \\
&= \sum_x \text{Prob } (x) \\
&= \left( \sum_y \text{Prob } (y|x, t_6) \cdot \text{Prob } (A|y, t_1) \cdot \text{Prob } (C|y, t_2) \right) \\
&\quad \cdot \left( \sum_z \text{Prob } (z|x, t_8) \cdot \text{Prob } (C|z, t_3) \right) \\
&\quad \cdot \left( \sum_w \text{Prob } (w|z, t_7) \cdot \text{Prob } (C|w, t_4) \cdot \text{Prob } (G|w, t_5) \right)
\end{align*}
\]
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:

$$\sum_w \text{Prob} \ (w|z, t_7) \ L_7(w)$$
and the algorithm is:

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

$$L_{k}^{(i)}(s) = \left( \sum_{x} \text{Prob} \left( x \mid s, t_{\ell} \right) L_{\ell}^{(i)}(x) \right)$$

$$\times \left( \sum_{y} \text{Prob} \left( y \mid s, t_{m} \right) L_{m}^{(i)}(y) \right)$$
Starting and finishing the recursion

At the top of the tree the definition of the L’s specifies that they look like

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

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

\[
L^{(i)} = \sum_{x} \pi_x L_0^{(i)}(x)
\]
**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 $1 - \epsilon$ of finding the correct nucleotide, and $\epsilon/3$ of inferring each of the three other possibilities, when an A is observed, the four values should be $(1 - \epsilon, \epsilon/3, \epsilon/3, \epsilon/3)$, and when a C is observed, they should be $(\epsilon/3, 1 - \epsilon, \epsilon/3, \epsilon/3)$. 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:

\[ L_{0}^{(i)}(z) = L_{8}^{(i)}(z) \text{Prob} (z) \text{Prob} (w|z, t_6)L_{6}^{(i)}(w) \]
Finding the ML tree

So far I have just talked about the computation of the likelihood for one tree with branch lengths known.

As with the distance matrix methods, we must search the space of tree topologies, and for each one examined, we need to optimize the branch lengths to maximize the likelihood.
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 posterior distribution of possible phylogenies.
A model of variation in evolutionary rates among sites

The basic idea is that the rate at each site is drawn independently from a distribution of rates. The most widely used choice is the Gamma distribution, which has density function (if its mean is 1):

\[ f(r) = \frac{\alpha^\alpha r^{\alpha-1} e^{-\alpha r}}{\Gamma(\alpha)} \]
Gamma distributions

Gamma distributions with mean 1 and different coefficients of variation (standard deviation / mean). $\alpha = 1/CV^2$ is the “shape parameter” of the Gamma distribution.
Unrealistic aspects of the model:

- There is no reason, aside from mathematical convenience, to assume that the Gamma is the right distribution. A common variation is to assume there is a separate probability \( f_0 \) of having rate 0.
- Rates at different sites appear to be correlated, which this model does not allow.
- Rates are not constant throughout evolution – they change with time.
Rates varying among sites

If $L^{(i)}(r_i)$ is the likelihood of the tree for site $i$ given that the rate of evolution at site $i$ is $r_i$, we can integrate this over a gamma density

$$L^{(i)} = \int_0^{\infty} f(r_i; \alpha) L^{(i)}(r_i) \, dr_i$$

so that the overall likelihood is

$$L = \prod_{i=1}^{m} \left[ \int_0^{\infty} f(r_i; \alpha) L^{(i)}(r_i) \, dr_i \right]$$

Unfortunately these integrals cannot be evaluated for trees with more than a few tips as the quantities $L^{(i)}(r_i)$ complicated.
Hidden Markov Models

These are the most widely used models allowing rate variation to be correlated along the sequence.

We assume:

- There are a finite number of rates, \( m \). Rate \( i \) is \( r_i \).
- There are probabilities \( p_i \) of a site having rate \( i \).
- A process not visible to us ("hidden") assigns rates to sites. It is a Markov process working along the sequence. For example it might have transition probability \( \text{Prob}(j|i) \) of changing to rate \( j \) in the next site, given that it is at rate \( i \) in this site.
- The probability of our seeing some data are to be obtained by summing over all possible combinations of rates, weighting appropriately by their probabilities of occurrence.
Suppose that we have a way of calculating, for each possible rate at each possible site, the probability of the data at that site given that rate. This is

\[ \text{Prob} \left( D^{(i)} \mid r_j \right) \]

To get the overall probability of all data, sum over all possible paths through the array of sites \( \times \) rates, weighting each combination of rates by its probability:
A Hidden Markov Model for rates in a phylogeny

Phylogeny:

Sites:

1 2 3 4 5 6 7 8
C A C G A C G A
C G T A A C G A
C G A G A C G G
C A A A A A C G G
A A G T G C G C

Hidden Markov chain:

Rates of evolution:

10.0
2.0
0.3

Lecture 5. Likelihood and Bayesian methods. HMM rate variation models. – p.27/37
Hidden Markov Models

If there are a number of hidden rate states, with state $i$ having rate $r_i$

$$\text{Prob} \ (D \mid T) = \sum_{i_1} \sum_{i_2} \ldots \sum_{i_p} \text{Prob} \ (r_{i_1}, r_{i_2}, \ldots r_{i_p})$$

$$\times \text{Prob} \ (D \mid T, r_{i_1}, r_{i_2}, \ldots r_{i_m})$$

Evolution is independent once each site has had its rate specified

$$\text{Prob} \ (D \mid T, r_1, r_2, \ldots, r_p) =$$

$$\prod_{i=1}^{p} \text{Prob} \ (D^{(i)} \mid T, r_i).$$
Seems impossible ...

Evolution is independent once each site has had its rate specified

\[
\text{Prob} \left( \mathbf{D} \mid \mathbf{T}, r_1, r_2, \ldots, r_m \right) = \prod_{i=1}^{m} \text{Prob} \left( \mathbf{D}^{(i)} \mid \mathbf{T}, r_i \right).
\]

To compute the likelihood we sum over all ways rate states could be assigned to sites:

\[
L = \text{Prob} \left( \mathbf{D} \mid \mathbf{T} \right) = \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \ldots \sum_{i_p=1}^{m} \text{Prob} \left( r_{i_1}, r_{i_2}, \ldots, r_{i_p} \right) \\
\times \text{Prob} \left( \mathbf{D}^{(1)} \mid r_{i_1} \right) \text{Prob} \left( \mathbf{D}^{(2)} \mid r_{i_2} \right) \ldots \text{Prob} \left( \mathbf{D}^{(n)} \mid r_{i_p} \right)
\]

Problem: The number of rate combinations is very large. With 100 sites and 3 rates at each, it is \(3^{100} \approx 5 \times 10^{47}\). This makes the summation impractical.
This is an HMM

The hidden states identify the rates that applied at a site. Each rate implies (together with the tree, which is in common to all sites) a distribution of possible base patterns (\(4^n\) of them if there are \(n\) sequences on the tree). At each site one has actually occurred.

We can use the usual Forwards Algorithm to sum up likelihood over all paths through the array of rates.
The forwards-backwards algorithm.

Enables us to compute the fraction of the likelihood contributed by one of the rates at one of the sites.

Lecture 5. Likelihood and Bayesian methods. HMM rate variation models. – p.31/37
Siepel and Haussler (2004) have called the HMMs over rates (and other HMMs that operate along multiple sequence alignments and evaluate likelihoods on a tree) “phylo-HMMs”. They have applied these widely to search for conserved regions in alignments of genomes and for gene-finding.

**Fig. 5.** A screen shot from the UCSC Genome Browser [24] showing a selected region of the data set of example 2, including several exons of the MET gene (black boxes at top). The binomial-based (light gray) and parsimony-based (medium gray) conservation scores of Margulies et al. [30] are shown as tracks in the browser, as are the posterior probabilities (×1000) of state $s_1$ in the phylo-HMM (dark gray). Plots similar to this one, showing phylo-HMM-based conservation scores across the whole human genome, can be viewed online at http://genome.ucsc.edu.
References

Likelihood and Bayesian


References


HMM rates


Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* **139**: 993-1005. [Also allowing for correlated rates along the molecule]
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 make it into a PDF file, and
- displaying the slides in Adobe Acrobat Reader.

Result: nice slides using freeware.