Lecture 6. HMMs for rates. Testing trees, bootstraps, jackknifes[sic]

Joe Felsenstein

Department of Genome Sciences and Department of Biology
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 with mean 1 and different coefficients of variation (standard deviation / mean). $\alpha = \frac{1}{CV^2}$ 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.
Likelihood with an HMM

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

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

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

Sites

Phylogeny

Sites

Phylogeny

Sites

Hidden Markov chain:

Rates

of evolution

10.0

2.0

0.3

Hidden Markov Models

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

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

Evolution is independent once each site has had its rate specified

$$
\text{Prob} \left( D \mid T, r_1, r_2, \ldots, r_p \right) = \prod_{i=1}^{p} \text{Prob} \left( D^{(i)} \mid T, r_i \right).
$$
Seems impossible ...

Evolution is independent once each site has had its rate specified

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

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

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

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.

The "forward–backward" algorithm allows us to get the probability of everything given one site’s state.

... which enables us to compute the fraction of the likelihood contributed by one of the rates at one of the sites. Alternatively, the Viterbi algorithm enables us to find the single path that contributes the most to the likelihood.
PhyloHMMs

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.
A non-phylogeny example of the bootstrap

estimate of $\theta$  
(unknown) true value of $\theta$

empirical distribution of sample

(unknown) true distribution

Distribution of estimates of parameters

Bootstrap replicates

Bootstrap sampling from a distribution (a mixture of two normals) to estimate the variance of the mean
Bootstrap sampling

To infer the error in a quantity, $\theta$, estimated from a sample of points $x_1, x_2, \ldots, x_n$ we can

- Do the following $R$ times ($R = 1000$ or so)
- Draw a “bootstrap sample" by sampling $n$ times with replacement from the sample. Call these $x_1^*, x_2^*, \ldots, x_n^*$. Note that some of the original points are represented more than once in the bootstrap sample, some once, some not at all.
- Estimate $\theta$ from the bootstrap sample, call this $\hat{\theta}_k^*$ ($k = 1, 2, \ldots, R$)
- When all $R$ bootstrap samples have been done, the distribution of $\hat{\theta}_i^*$ estimates the distribution one would get if one were able to draw repeated samples of $n$ points from the unknown true distribution.
Bootstrap sampling of phylogenies

Original Data
sequences

Bootstrap sample #1
sequences
sample same number of sites, with replacement

Bootstrap sample #2
sequences
sample same number of sites, with replacement

(and so on)

Estimate of the tree

Bootstrap estimate of the tree, #1

Bootstrap estimate of the tree, #2
More on the bootstrap for phylogenies

- The sites are assumed to have evolved independently given the tree. They are the entities that are sampled (the $x_i$).
- The trees play the role of the parameter. One ends up with a cloud of $R$ sampled trees.
- To summarize this cloud, we ask, for each branch in the tree, how frequently it appears among the cloud of trees.
- We make a tree that summarizes this for all the most frequently occurring branches.
- This is the *majority rule consensus tree* of the bootstrap estimates of the tree.
A partition on the first tree

Trees:

- AE | BCDF
- ACE | BDF
- ACEF | BD 1
- AC | BDEF
- AEF | BCD
- ADEF | BC
- ABDF | EC
- ABCE | DF

How many times each partition of species is found:
A second partition

Trees:

How many times each partition of species is found:

<table>
<thead>
<tr>
<th>Partition</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>BCDF</td>
</tr>
<tr>
<td>ACE</td>
<td>BDF</td>
</tr>
<tr>
<td>ACEF</td>
<td>BD</td>
</tr>
<tr>
<td>AC</td>
<td>BDEF</td>
</tr>
<tr>
<td>AEF</td>
<td>BCD</td>
</tr>
<tr>
<td>ADEF</td>
<td>BC</td>
</tr>
<tr>
<td>ABDF</td>
<td>EC</td>
</tr>
<tr>
<td>ABCE</td>
<td>DF</td>
</tr>
</tbody>
</table>
A third partition

Trees:

How many times each partition of species is found:

<table>
<thead>
<tr>
<th>Partition</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>BCDF</td>
</tr>
<tr>
<td>ACE</td>
<td>BDF</td>
</tr>
<tr>
<td>ACEF</td>
<td>BD</td>
</tr>
<tr>
<td>AC</td>
<td>BDEF</td>
</tr>
<tr>
<td>AEF</td>
<td>BCD</td>
</tr>
<tr>
<td>ADEF</td>
<td>BC</td>
</tr>
<tr>
<td>ABDF</td>
<td>EC</td>
</tr>
<tr>
<td>ABCE</td>
<td>DF</td>
</tr>
</tbody>
</table>
The second tree and its partitions

Trees:

How many times each partition of species is found:

- AE | BCDF 1
- ACE | BDF 2
- ACEF | BD 1
- AC | BDEF 1
- AEF | BCD
- ADEF | BC
- ABDF | EC
- ABCE | DF 1
The third tree and its partitions

Trees:

- [Diagram of tree AE | BCDF]
- [Diagram of tree ACE | BDF]
- [Diagram of tree ADEF | BC]
- [Diagram of tree ABDF | EC]
- [Diagram of tree ABCE | DF]

How many times each partition of species is found:

- AE | BCDF: 2
- ACE | BDF: 2
- ACEF | BD: 1
- AC | BDEF: 1
- AEF | BCD: 1
- ADEF | BC: 1
- ABDF | EC: 1
- ABCE | DF: 1
The fourth tree and its partitions

Trees:

- **Tree 1**
  - E
  - A
  - C
  - F
  - D
  - B

- **Tree 2**
  - E
  - C
  - A
  - D
  - F

- **Tree 3**
  - E
  - A
  - B
  - C
  - D
  - F

- **Tree 4**
  - E
  - C
  - A
  - D
  - F

How many times each partition of species is found:

- \( AE | BCDF \) 3
- \( ACE | BDF \) 2
- \( ACEF | BD \) 1
- \( AC | BDEF \) 1
- \( AEF | BCD \) 1
- \( ADEF | BC \) 2
- \( ABDF | EC \)
- \( ABCE | DF \) 2
The fifth tree and its partitions

Trees:

How many times each partition of species is found:

- AE | BCDF: 3
- ACE | BDF: 3
- ACEF | BD: 1
- AC | BDEF: 1
- AEF | BCD: 1
- ADEF | BC: 2
- ABDF | EC: 1
- ABCE | DF: 3
The trees and their partitions

Trees:

AE | BCDF 3
ACE | BDF 3
ACEF | BD 1
AC | BDEF 1
AEF | BCD 1
ADEF | BC 2
ABDF | EC 1
ABCE | DF 3
Majority rule consensus trees

Trees:

- $E \quad A \quad C \quad B \quad D \quad F$
- $E \quad A \quad F \quad C \quad D$
- $E \quad A \quad B \quad C \quad D \quad F$
- $E \quad B \quad C \quad D \quad F$
- $E \quad C \quad D \quad A \quad F$

How many times each partition of species is found:

- $AE \mid BCDF = 3$
- $ACE \mid BDF = 3$
- $ACEF \mid BD = 1$
- $AC \mid BDEF = 1$
- $AEF \mid BCD = 1$
- $ADEF \mid BC = 2$
- $ABDF \mid EC = 1$
- $ABCE \mid DF = 3$

Majority-rule consensus tree of the unrooted trees:
An example of bootstrap sampling of trees

Bovine
Mouse
Squir Monk
Chimp
Human
Gorilla
Orang
Gibbon
Rhesus Mac
Jpn Macaq
Crab−E.Mac
BarbMacaq
Tarsier
Lemur

232 nucleotide, 14-species mitochondrial D-loop analyzed by parsimony, 100 bootstrap replicates
Potential problems with the bootstrap

1. Sites may not evolve independently
2. Sites may not come from a common distribution (but can consider them sampled from a mixture of possible distributions)
3. If do not know which branch is of interest at the outset, a “multiple-tests" problem means P values are overstated
4. P values are biased (too conservative)
5. Bootstrapping does not correct biases in phylogeny methods
Other resampling methods

- Delete-half jackknife. Sample a random 50% of the sites, \textit{without} replacement.
- Delete-\(1/e\) jackknife (Farris et. al. 1996) (too little deletion from a statistical viewpoint).
- Reweighting characters by choosing weights from an exponential distribution.
- In fact, reweighting them by any exchangeable weights having coefficient of variation of 1
- Parametric bootstrap – simulate data sets of this size assuming the estimate of the tree is the truth
- (to correct for correlation among adjacent sites) (Künsch, 1989)
  Block-bootstrapping – sample \(n/b\) blocks of \(b\) adjacent sites.
Delete half jackknife on the example
Parametric bootstrapping

The Parametric Bootstrap (Efron, 1985)

Suppose we have independent observations drawn from a known distribution:

\[ x_1, x_2, x_3, \ldots, x_n \rightarrow \theta \]

and take the distribution of the \[ \theta \]

and a parameter, \( \theta \), calculated from this.

To infer the variability of \( \theta \)

Use the current estimate, \( \hat{\theta} \)

Use the distribution that has that as its true parameter

Sample \( R \) data sets from that distribution, each having the same sample size as the original sample

\[ x_1^*, x_2^*, x_3^*, \ldots, x_n^* \rightarrow \hat{\theta}_1 \]

\[ x_1^*, x_2^*, x_3^*, \ldots, x_n^* \rightarrow \hat{\theta}_2 \]

\[ \vdots \]

\[ x_1^*, x_2^*, x_3^*, \ldots, x_n^* \rightarrow \hat{\theta}_R \]

and take the distribution of the \( \hat{\theta}_i \)

as the estimate of the distribution

from which it is drawn

Lecture 6. HMMs for rates. Testing trees, bootstraps, jackknifes[sic] – p.31/50
The parametric bootstrap for phylogenies

An example of the parametric bootstrap

Likelihood ratio confidence limits on Ts/Tn ratio

Transition / transversion ratio for the 14-species primate data set
Likelihoods in tree space – a 3-species clock example
The constraints for a molecular clock

Constraints for a clock

\[ v_1 = v_2 \]

\[ v_4 = v_5 \]

\[ v_1 + v_6 = v_3 \]

\[ v_3 + v_7 = v_4 + v_8 \]
Testing for a molecular clock

To test for a molecular clock:

- Obtain the likelihood with no constraint of a molecular clock (For primates data with $T_s/T_n = 30$ we get $\ln L_1 = -2616.86$)
- Obtain the highest likelihood for a tree which is constrained to have a molecular clock: $\ln L_0 = -2679.0$
- Look up $2(\ln L_1 - \ln L_0) = 2 \times 62.14 = 124.28$ on a $\chi^2$ distribution with $n - 2 = 12$ degrees of freedom (in this case the result is significant)
Two trees to be tested by paired sites tests

Tree I
- Mouse
- Bovine
- Gibbon
- Orang
- Gorilla
- Chimp
- Human

Tree II
- Mouse
- Bovine
- Gibbon
- Orang
- Gorilla
- Chimp
- Human
## Differences in log likelihoods site by site

<table>
<thead>
<tr>
<th>Tree</th>
<th>site</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>231</th>
<th>232</th>
<th>ln L</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td></td>
<td>2.971</td>
<td>−4.483</td>
<td>−5.673</td>
<td>−5.883</td>
<td>−2.691</td>
<td>−8.003</td>
<td>⋯</td>
<td>2.971</td>
<td>−2.691</td>
</tr>
<tr>
<td>II</td>
<td></td>
<td>2.983</td>
<td>−4.494</td>
<td>−5.685</td>
<td>−5.898</td>
<td>−2.700</td>
<td>−7.572</td>
<td>⋯</td>
<td>2.987</td>
<td>−2.705</td>
</tr>
<tr>
<td>Diff</td>
<td></td>
<td>+0.012</td>
<td>+0.111</td>
<td>+0.013</td>
<td>+0.015</td>
<td>+0.010</td>
<td>−0.431</td>
<td>⋯</td>
<td>+0.012</td>
<td>+0.010</td>
</tr>
</tbody>
</table>
Histogram of log likelihood differences

Difference in log likelihood at site
Paired sites tests

- Winning sites test (Prager and Wilson, 1988). Do a sign test on the signs of the differences.
- $z$ test (me, 1993 in PHYLIP documentation). Assume differences are normal, do $z$ test of whether mean (hence sum) difference is significant.
- $t$ test. Swofford et. al., 1996: do a $t$ test (paired)
- Wilcoxon ranked sums test (Templeton, 1983).
- RELL test (Kishino and Hasegawa, 1989 per my suggestion). Bootstrap resample sites, get distribution of difference of totals.
In this example

- Winning sites test. 160 of 232 sites favor tree I. $P < 3.279 \times 10^{-9}$
- $z$ test. Difference of log-likelihood totals is 0.948104 standard deviations from 0, $P = 0.343077$. Not significant.
- $t$ test. Same as $z$ test for this large a number of sites.
- Wilcoxon ranked sums test. Rank sum is 4.82805 standard deviations below its expected value, $P = 0.000001378765$
- RELL test. 8,326 out of 10,000 samples have a positive sum, $P = 0.3348$ (two-sided)
Bayesian methods

Bayesian methods have gained popularity, some of it because they can be computationally faster than bootstrapping (in my view you should use them if you agree with them, not just because of speed).

In the Bayesian framework, one can avoid the separate calculation of confidence intervals. The posterior distribution of trees shows us how much credence to give different trees (for example, it assigns probabilities to different tree topologies).

The interesting issue is how to summarize this posterior distribution in the best way. In this respect Bayesian methods leave you in a situation analogous to having the cloud of bootstrap-sampled trees without yet having summarized them.

Clade probabilities, computed in the same way as bootstrap probabilities from the posterior cloud of trees, are a popular way of summarizing this. They are used in the popular Bayesian program MrBayes.
References, page 1


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]
References


bias in the bootstrap


Sanderson, M. J. 1995. Objections to bootstrapping phylogenies: a critique. *Systematic Biology* **44**: 299-320. [Good but he accepts a few criticisms I would not have accepted]
variations on the bootstrap

Harshman, J. 1994. The effect of irrelevant characters on bootstrap values. *Systematic Zoology* 43: 419-424. [Not much effect on bootstrap support with parsimony whether or not you include invariant characters]


paired sites tests


paired sites, and miscellaneous


Yang, Z. 2006. *Computational Molecular Evolution*. Oxford University Press, Oxford. [material is in chapters 1, 2, 4, 5, section 6.4, sections 7.1 and 7.2]