Week 6: Restriction sites, RAPDs, microsatellites, likelihood, hidden Markov models

Genome 570

February, 2012
Restriction sites and microsatellite loci

Note – owing to time pressure, we are going to skip the restriction sites and microsatellites material in the lectures. The slides are still included here.
Restriction sites

GGG\textcolor{red}{AATCC}CAGTCAGGTTCCGTTAGT\textcolor{red}{GAATCCGTTACC}CGTAAT

GGG\textcolor{red}{CAATCCC}CAGTCAGGTTCCGTTAGT\textcolor{red}{GAATCCGTTACC}CGTAAT

GGG\textcolor{red}{CAATCCC}CAGTCAGGTTCCGTTAGT\textcolor{red}{GAATCCGTTACC}CGTAAT

GGG\textcolor{red}{CAATCCC}CAGTCAGGTTCCGTTAGT\textcolor{red}{GAATCCGTTACC}CGTAAT

CCC\textcolor{red}{GGTTACC}CGTAAT
Number of sequences in a restriction site location

\[ N_k = \binom{r}{k} 3^k. \]

<table>
<thead>
<tr>
<th>(k)</th>
<th>number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>135</td>
</tr>
<tr>
<td>3</td>
<td>540</td>
</tr>
<tr>
<td>4</td>
<td>1215</td>
</tr>
<tr>
<td>5</td>
<td>1458</td>
</tr>
<tr>
<td>6</td>
<td>729</td>
</tr>
</tbody>
</table>
Transition probabilities for restriction sites, aggregated

\[
P_{k\ell} = \sum_{m=a}^{b} \binom{r-k}{\ell-k+m} p^{\ell-k+m} (1 - p)^{r-k-(\ell-k+m)}
\]

\[
\times \binom{k}{m} \left( \frac{p}{3} \right)^m \left( 1 - \frac{p}{3} \right)^{k-m}
\]

\[
P_{k\ell} = \min[k, r-\ell] \sum_{m=\max[0,k-\ell]}^{\min[k,r-\ell]} \binom{r-k}{\ell-k+m} p^{\ell-k+m} (1 - p)^{r-\ell+m}
\]

\[
\times \binom{k}{m} \left( \frac{p}{3} \right)^m \left( 1 - \frac{p}{3} \right)^{k-m}
\]
Presence/absence of restriction sites

\[ \pi_0 = \left( \frac{1}{4} \right)^r \]

\[ P_{++} = \pi_0 (1 - p)^r \]

\[ P_{++} = \left( \frac{1}{4} \right)^r \left[ \frac{1 + 3e^{-\frac{4}{3}t}}{4} \right]^r \]

\[ P_{+-} = P_{-+} = \left( \frac{1}{4} \right)^r \left( 1 - \left[ \frac{1 + 3e^{-\frac{4}{3}t}}{4} \right]^r \right) \]

\[ P_{--} = 1 - \left( \frac{1}{4} \right)^r \left( 2 - \left[ \frac{1 + 3e^{-\frac{4}{3}t}}{4} \right]^r \right) \]
Conditioning on the site not being entirely absent

\[
\text{Prob} \ (++ \mid \text{not} \ - -) = \frac{P_{++}}{P_{++} + P_{+-} + P_{-+}}
\]

\[
\text{Prob} \ (-+ \mid \text{not} \ - -) = \frac{P_{--}}{P_{++} + P_{+-} + P_{-+}}
\]

\[
\text{Prob} \ (+- \mid \text{not} \ - -) = \frac{P_{+-}}{P_{++} + P_{+-} + P_{-+}}
\]
Likelihoods for two sequences with restriction sites

\[ L = \left( \frac{P_{++}}{P_{++} + P_{+-} + P_{-+}} \right)^{n_{++}} \left( \frac{P_{+-}}{P_{++} + P_{+-} + P_{-+}} \right)^{n_{+-} + n_{-+}} \]

is of the form

\[ L = x^{n_{++}} \left( \frac{1}{2} (1 - x) \right)^{n_{+-} + n_{-+}} \]

where

\[ x = \frac{P_{++}}{P_{++} + P_{+-} + P_{-+}} \]
Maximizing the likelihood

The maximum likelihood is when

\[ x = \frac{n_{++}}{n_{++} + n_{+-} + n_{--}} \]

using the equations for the P’s and for \( x \)

\[ \left( \frac{1 + 3e^{-\frac{4}{3}t}}{4} \right)^r = \frac{n_{++}}{n_{++} + \frac{1}{2} (n_{+-} + n_{--})} \]

which can be solved for \( t \) to give the distance

\[ D = \hat{t} = -\frac{3}{4} \ln \left( \frac{4}{3} \left[ \frac{n_{++}}{n_{++} + \frac{1}{2} (n_{+-} + n_{--})} \right]^{1/r} - \frac{1}{3} \right) \]
Are shared restriction sites parallelisms?

![Graph showing the probability of an ancestor being a plus symbol over time. The graph includes a two-state 0-1 model and a restriction site (6 bases).]
A one-step microsatellite model

Computing the probability that there are a net of $i$ steps by having $i + k$ steps up and $k$ steps down,

$$\text{Prob (there are } i + 2k \text{ steps)} = e^{-\mu t} (\mu t)^{i+2k} / (i + 2k)!$$

$$\text{Prob (i + k of these are steps upwards)} = \binom{i+2k}{i+k} \left(\frac{1}{2}\right)^{i+k} \left(\frac{1}{2}\right)^k$$

$$= \binom{i+2k}{i+k} \left(\frac{1}{2}\right)^{i+2k}$$

... we put these together to and sum over $k$ to get the result ...
Transition probabilities in the microsatellite model

Transition probability of \( i \) steps net change in branch of length \( \mu t \) is then

\[
p_i(\mu t) = \sum_{k=0}^{\infty} e^{-\mu t} \frac{(\mu t)^{i+2k}}{(i+2k)!} \binom{i+2k}{i+k} \left( \frac{1}{2} \right)^{i+2k}
\]

\[
= e^{-\mu t} \sum_{k=0}^{\infty} \left( \frac{\mu t}{2} \right)^{i+2k} \frac{1}{(i+k)!k!}
\]
Why the \((\delta \mu)^2\) distance works

approximately

\(2N_e\) generations

approximately

\(2N_e\) generations

\(t\) generations
$$ (\delta \mu)^2 \text{ distance for microsatellites} $$

$$ (\delta \mu)^2 = (\mu_A - \mu_B)^2 $$
Transition probabilities in a one-step model

\[ \mu_t = 1 \]

\[ \mu_t = 5 \]

change in number of repeats
A Brownian motion approximation

\[ \text{Prob (} m + k | m \text{)} \simeq \min \left[ 1, \frac{1}{\sqrt{\mu t} \sqrt{2\pi}} \exp \left( -\frac{1}{2} \frac{k^2}{\mu t} \right) \right] \]
Is it accurate? well ...

![Graph showing probability and branch length relationship]
Likelihood ratios: the odds-ratio form of Bayes’ theorem

\[
\frac{\text{Prob} (H_1|D)}{\text{Prob} (H_2|D)} = \frac{\text{Prob} (D|H_1)}{\text{Prob} (D|H_2)} \quad \frac{\text{Prob} (H_1)}{\text{Prob} (H_2)}
\]

posterior odds ratio  likelihood ratio  prior odds ratio

Given prior odds, we can use the data to compute the posterior odds.
With many sites the likelihood wins out

Independence of the evolution in the different sites implies that:

\[
\text{Prob} \left( D \mid H_i \right) \\
= \text{Prob} \left( D^{(1)} \mid H_i \right) \text{Prob} \left( D^{(2)} \mid H_i \right) \ldots \text{Prob} \left( D^{(n)} \mid H_i \right)
\]

so we can rewrite the odds-ratio formula as

\[
\frac{\text{Prob} \left( H_1 \mid D \right)}{\text{Prob} \left( H_2 \mid D \right)} = \left( \prod_{i=1}^{n} \frac{\text{Prob} \left( D^{(i)} \mid H_1 \right)}{\text{Prob} \left( D^{(i)} \mid H_2 \right)} \right) \frac{\text{Prob} \left( H_1 \right)}{\text{Prob} \left( H_2 \right)}
\]

This implies that as \( n \) gets large the likelihood-ratio part will dominate.
An example – coin tossing

If we toss a coin 11 times and get HHTTHHTHHTTT, the likelihood is:

\[
\]

Solving for the maximum likelihood estimate for \( p \) by finding the maximum:

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

and equating it to zero and solving:

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

gives \( \hat{p} = 5/11 \)
Likelihood as function of $p$ for the 11 coin tosses

$p = 0.454$
Maximizing the likelihood

Maximizing the likelihood is the same as maximizing the log likelihood, because its log increases as the number increases, so:

\[
\ln L = 5 \ln p + 6 \ln(1 - p),
\]

and

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

so that, again,

\[
\hat{p} = \frac{5}{11}
\]
An example with one site

We want to compute the probability of these data at the tips of the tree, given the tree and the branch lengths. Each site has an independent outcome, all on the same tree.
Likelihood sums over states of interior nodes

The likelihood is the product over sites (as they are independent given the tree):

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

For site i, summing over all possible states at interior nodes of the tree:

\[ \text{Prob}\left(D^{(i)}|T\right) = \sum_x \sum_y \sum_z \sum_w \text{Prob}(A, C, C, C, G, x, y, z, w|T) \]
... the products over events in the branches

By independence of events in different branches (conditional only on their starting states):

\[
\text{Prob} \ (A, C, C, C, G, x, y, z, w | T) = \\
\text{Prob} \ (x) \ \text{Prob} \ (y | x, t_6) \ \text{Prob} \ (A | y, t_1) \ \text{Prob} \ (C | y, t_2) \\
\text{Prob} \ (z | x, t_8) \ \text{Prob} \ (C | z, t_3) \\
\text{Prob} \ (w | z, t_7) \ \text{Prob} \ (C | w, t_4) \ \text{Prob} \ (G | w, t_5)
\]
The result looks hard to compute

Summing over all \( x, y, z, \) and \( w \) (each taking values \( A, G, C, T \)):

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

This could be hard to do on a larger tree. For example, with 20 species there are 19 interior nodes and thus the number of outcomes for interior nodes is \( 4^{19} = 274,877,906,944 \).
... but there’s a trick ...

We can move summations in as far as possible:

$$\text{Prob} \left( D^{(i)} \mid T \right) =$$

$$\sum_x \text{Prob} \left( x \right) \left( \sum_y \text{Prob} \left( y \mid x, t_6 \right) \text{Prob} \left( A \mid y, t_1 \right) \text{Prob} \left( C \mid y, t_2 \right) \right)$$

$$\left( \sum_z \text{Prob} \left( z \mid x, t_8 \right) \text{Prob} \left( C \mid z, t_3 \right) \right)$$

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

The pattern of parentheses parallels the structure of the tree:

$$(A, C) \ (C, (C, G))$$
Conditional likelihoods in this calculation

Working from innermost parentheses outwards is the same as working down the tree. We can define a quantity

\[ L_j^{(i)}(s) \]
the conditional likelihood at site \( i \) of everything at or above point \( j \) in the tree, given that point \( j \) have state \( s \)

One such is the term:

\[ L_7(w) = \text{Prob} (C|w, t_4) \cdot \text{Prob} (G|w, t_5) \]

Another is the term including that:

\[ L_8(z) = \text{Prob} (C|z, t_3) \left( \sum_w \text{Prob} (w|z, t_7) \cdot \text{Prob} (C|w, t_4) \cdot \text{Prob} (G|w, t_5) \right) \]
The pruning algorithm

This follows the recursion down the tree:

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

\[ \times \left( \sum_{y} \text{Prob } (y|s, t_{m}) L_{m}^{(i)}(y) \right) \]

At a tip the quantity is easily seen to be like this (if the tip is in state A):

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

At the bottom we have a weighted sum over all states, weighted by their prior probabilities:

\[ L^{(i)} = \sum_{x} \pi_{x} L_{0}^{(i)}(x) \]

We can do that because, if evolution has gone on for a long time before the root of the tree, the probabilities of bases there are just the equilibrium probabilities under the DNA model (or whatever model we assume).
Handling ambiguity and error

If a base is unknown: use (1, 1, 1, 1). If known only to be a purine: (1, 0, 1, 0)

Note – do not do something like this: (0.5, 0, 0.5, 0). It is not the probability of being that base but the probability of the observation given that base.

If there is sequencing error use something like this:

\[ (1 - \varepsilon, \varepsilon/3, \varepsilon/3, \varepsilon/3) \]

(assuming an error is equally likely to be each of the three other bases).
Rerooting the tree

before

after
Unrootedness

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

Reversibility of the Markov process guarantees that

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

Substituting that in:

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

... which means that (if the model of change is a reversible one) the likelihood does not depend on where the root is placed in the unrooted tree.
To compute likelihood when the root is in one branch

First we get (for the site) the conditional likelihoods for the left subtree ...
To compute likelihood when the root is in one branch

... then we get the conditional likelihoods for the right subtree
To compute likelihood when the root is in one branch

... and finally we get the conditional likelihoods for the root, and from that the likelihoods for the site. (Note that it does not matter where in that branch we place the root, as long as the sum of the branch lengths on either side of the root is 0.236).
To do it with a different branch length ...

... we can just do the last step, but now with the new branch length. The other conditional likelihoods were already calculated, have not changed, and so we can re-use them. This really speeds things up for calculating how the likelihood depends on the length of one branch – just put the root there!
A data example: 7-species primates mtDNA

These are noncoding or synonymous sites from the D-loop region (and some adjacent coding regions) of mitochondrial DNA, selected by Masami Hasegawa from sequences done by S. Hayashi and coworkers in 1985.
Rate variation among sites

Evolution is independent once each site has had its rate specified

$$\text{Prob} \left( D \mid T, r_1, r_2, \ldots, r_p \right) = \prod_{i=1}^{p} \text{Prob} \left( D^{(i)} \mid T, r_i \right).$$
Coping with uncertainty about rates

Using a Gamma distribution independently assigning rates to sites:

\[ L = \prod_{i=1}^{m} \left[ \int_{0}^{\infty} f(r; \alpha) L^{(i)}(r) \, dr \right] \]

Unfortunately this is hard to compute on a tree with more than a few species. Yang (1994a) approximated this by a discrete histogram of rates:

\[ L^{(i)} = \int_{0}^{\infty} f(r; \alpha) L^{(i)}(r) \, dr \approx \sum_{j=1}^{k} w_k L^{(i)}(r_k) \]

Felsenstein (J. Mol. Evol., 2001) has suggested using Gauss-Laguerre quadrature to choose the rates \( r_i \) and the weights \( w_i \).
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 a[n] 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 \((i)\) given that rate \((r_j)\). This is

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

This can be done because the probabilities of change as a function of the rate \(r\) and time \(t\) are (in almost all models) just functions of their product \(rt\), so a site that has twice the rate is just like a site that has branches twice as long.

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:
Hidden Markov Model of rate variation

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 Process

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 $r_i$

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

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

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 ...

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

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

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

Fortunately, the terms can be reordered:

\[ L = \text{Prob} \left( \frac{D}{T} \right) \]

\[ = \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \cdots \sum_{i_p=1}^{m} \text{Prob} \left( i_1 \right) \text{Prob} \left( \frac{D^{(1)}}{r_{i_1}} \right) \]

\[ \times \text{Prob} \left( i_2 \mid i_1 \right) \text{Prob} \left( \frac{D^{(2)}}{r_{i_2}} \right) \]

\[ \times \text{Prob} \left( i_3 \mid i_2 \right) \text{Prob} \left( \frac{D^{(3)}}{r_{i_3}} \right) \]

\[ \vdots \]

\[ \times \text{Prob} \left( i_p \mid i_{p-1} \right) \text{Prob} \left( \frac{D^{(p)}}{r_{i_p}} \right) \]
Using Horner’s Rule

and the summations can be moved each as far rightwards as it can go:

\[ L = \sum_{i_1=1}^{m} \text{Prob} \left( i_1 \right) \text{Prob} \left( D^{(1)} \mid r_{i_1} \right) \]

\[ \sum_{i_2=1}^{m} \text{Prob} \left( i_2 \mid i_1 \right) \text{Prob} \left( D^{(2)} \mid r_{i_2} \right) \]

\[ \sum_{i_3=1}^{m} \text{Prob} \left( i_3 \mid i_2 \right) \text{Prob} \left( D^{(3)} \mid r_{i_3} \right) \]

\[ \vdots \]

\[ \sum_{i_p=1}^{m} \text{Prob} \left( i_p \mid i_{p-1} \right) \text{Prob} \left( D^{(p)} \mid r_{i_p} \right) \]
**Recursive calculation of HMM likelihoods**

The summations can be evaluated innermost-outwards. The same summations appear in multiple terms. We can then evaluate them only once. A huge saving results. The result is this algorithm:

Define $P_i(j)$ as the probability of everything at or to the right of site $i$, given that site $i$ has the $j$-th rate.

Now we can immediately see for the last site that for each possible rate category $i_p$

$$P_p(i_p) = \text{Prob} \left( D^{(p)} \mid r_{i_p} \right)$$

(as "at or to the right of" simply means "at" for that site).
Recursive calculation

More generally, for site $\ell < p$ and its rates $i_\ell$

$$P_\ell(i_\ell) = \text{Prob} \left( D^{(\ell)} \mid r_{i_\ell} \right) \sum_{i_{\ell+1}=1}^{m} \text{Prob} \left( i_{\ell+1} \mid i_\ell \right) P_{\ell+1}(i_{\ell+1})$$

We can compute the $P$’s recursively using this, starting with the last site and moving leftwards down the sequence. Finally we have the $P_1(i_1)$ for all $m$ states. These are simply weighted by the equilibrium probabilities of the Markov chain of rate categories:

$$L = \text{Prob} \left( D \mid T \right) = \sum_{i_1=1}^{m} \text{Prob} \left( i_1 \right) P_1(i_1)$$

An entirely similar calculation can be done from left to right, remembering that the transition probabilities $\text{Prob} \left( i_k \mid i_{k+1} \right)$ would be different in that case.
All paths through the array

array of conditional probabilities of everything at or to the right of that site, given the state at that site
Starting and finishing the calculation

At the end, at site $m$:

$$\text{Prob} \left( D^{[m]} | T, r_{im} \right) = \text{Prob} \left( D^{(m)} | T, r_{im} \right)$$

and once we get to site 1, we need only use the prior probabilities of the rates $r_i$ to get a weighted sum:

$$\text{Prob} \left( D | T \right) = \sum_{i_1} \pi_{i_1} \text{Prob} \left( D^{[1]} | T, r_{i_1} \right)$$
The pruning algorithm is just like
the forward algorithm
Paths from one state in one site
Paths from another state in that site
Paths from a third state
We can also use the backward algorithm

you can also do it the other way
Using both we can find likelihood contributed by a state

the "forward–backward" algorithm allows us to get the probability of everything given one site’s state
A particular Markov process on rates

There are many possible Markov processes that could be used in the HMM rates problem. I have used:

\[ \text{Prob} \left( r_i | r_j \right) = (1 - \lambda) \delta_{ij} + \lambda \pi_i \]
A numerical example. Cytochrome B

We analyze 31 cytochrome B sequences, aligned by Naoko Takezaki, using the Proml protein maximum likelihood program. Assume a Hidden Markov Model with 3 states, rates:

<table>
<thead>
<tr>
<th>category</th>
<th>rate</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>3.0</td>
<td>0.4</td>
</tr>
</tbody>
</table>

and expected block length 3.

We get a reasonable, but not perfect, tree with the best rate combination inferred to be
### Rates inferred from Cytochrome B

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence 1</th>
<th>Sequence 2</th>
<th>Sequence 3</th>
<th>Sequence 4</th>
<th>Sequence 5</th>
<th>Sequence 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>African</td>
<td>M-----TPMR</td>
<td>KINPLMKLIN</td>
<td>HSFIDLPTPS</td>
<td>NISAWWNFGS</td>
<td>LLGACILQLI</td>
<td>TTGLFLAMHYS</td>
</tr>
<tr>
<td>Caucasian</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
</tr>
<tr>
<td>Gorilla1</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
</tr>
<tr>
<td>Gorilla2</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
</tr>
<tr>
<td>Borang</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
</tr>
<tr>
<td>Sorang</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
</tr>
<tr>
<td>Gibbon</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
<td>...........</td>
</tr>
<tr>
<td>Bovine</td>
<td>NI SH IVN A A S I L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whalebm</td>
<td>NI TH ID A S L V L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whalebp</td>
<td>NI TH IVD AV S L M L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dhorse</td>
<td>NI SH I I A S I L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horse</td>
<td>NI SH I I S I L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhinocer</td>
<td>NI SH V I S I L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cat</td>
<td>NI SH I I A V T L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gseal</td>
<td>NI TH I N I L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hseal</td>
<td>NI TH I N I L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse</td>
<td>N TH F I A S V MV I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rat</td>
<td>NI SH F I A S V MV L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platypus</td>
<td>NNL TH I IV S L I L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wallaroo</td>
<td>NL SH I IV A I L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opossum</td>
<td>NI TH I D V I L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chicken</td>
<td>APNI SH LMN L A AV MT L L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xenopus</td>
<td>APNI SH I I N SL V A I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carp</td>
<td>A-SL TH IA D ALV L T L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loach</td>
<td>A-SL TH IA D ALV A V L L T L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trout</td>
<td>A-NL TH L IA D ALV A V L AT L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lamprey</td>
<td>SHQPSTI TH LS G S MLV S A SL I I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seaurchin1</td>
<td>LGL L EH IFRIL S TV L L I L T L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seaurchin2</td>
<td>AG L EH IFRIL S TV L L M I I L T L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Week 6: Restriction sites, RFLPs, microsatellites, likelihood, hidden Markov models – p.61/63*
### Rates inferred from Cytochrome B

<table>
<thead>
<tr>
<th></th>
<th>2223311112</th>
<th>2222222222</th>
<th>2222232112</th>
<th>2222222223</th>
<th>1222221112</th>
<th>3333111112</th>
</tr>
</thead>
<tbody>
<tr>
<td>african</td>
<td>PDASTAFSSI</td>
<td>AHITRDVNYG</td>
<td>WIIRYLFHANG</td>
<td>ASMFFICLFL</td>
<td>HIGRGLYYG</td>
<td>FLYSETWNIG</td>
</tr>
<tr>
<td>caucasian</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>cchimp</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>pchimp</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>gorilla1</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>gorilla2</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>borang</td>
<td>T.................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>sorang</td>
<td>M H.............</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>gibbon</td>
<td>..........................</td>
<td>V..................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>bovine</td>
<td>S TT..............</td>
<td>T V C.............</td>
<td>............ M.........</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>whalebm</td>
<td>. TM............</td>
<td>V T C.............</td>
<td>............ V.........</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>whalebp</td>
<td>. TT............</td>
<td>V T C.............</td>
<td>............ V.........</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>dhorse</td>
<td>S TT............</td>
<td>V T C.............</td>
<td>............ I.........</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>horse</td>
<td>S TT............</td>
<td>V T C.............</td>
<td>............ I.........</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>rhinocer</td>
<td>. TT............</td>
<td>V T C.............</td>
<td>............ M.........</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>cat</td>
<td>S TM............</td>
<td>V T C.............</td>
<td>............ YM V M</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>gseal</td>
<td>S TT............</td>
<td>V T C.............</td>
<td>............ YM V M</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>hseal</td>
<td>S TT............</td>
<td>V T C.............</td>
<td>............ YM V M</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>mouse</td>
<td>S TM............</td>
<td>V T C.............</td>
<td>............ L M</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>rat</td>
<td>S TM............</td>
<td>V T C.............</td>
<td>............ L Q</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>platypus</td>
<td>S T............</td>
<td>V C.............</td>
<td>............ L M I</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>wallaroo</td>
<td>S TL...........</td>
<td>V C.............</td>
<td>............ L N M</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>opossum</td>
<td>S TL...........</td>
<td>V C.............</td>
<td>............ L NI M</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>chicken</td>
<td>A T L........</td>
<td>V TC N Q......</td>
<td>............ L N F I</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>xenopus</td>
<td>A T M........</td>
<td>V CF............</td>
<td>............ LL N L F IY</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>carp</td>
<td>S I.............</td>
<td>V T C.............</td>
<td>............ L NV F IYM A</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>loach</td>
<td>S I.............</td>
<td>V C.............</td>
<td>............ L NI F Y A A</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>trout</td>
<td>S I.............</td>
<td>V C S............</td>
<td>............ L NI F IYM A</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>lamprey</td>
<td>ANTEL........</td>
<td>V M C N........</td>
<td>............ L MN IYA I</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>seaurchin1</td>
<td>A I L........</td>
<td>A S C............</td>
<td>............ LL NV L MYC G</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>seaurchin2</td>
<td>A INL........</td>
<td>V S C............</td>
<td>............ LL NV L MYC L</td>
<td>..........................</td>
<td>..........................</td>
<td>..........................</td>
</tr>
</tbody>
</table>

*Week 6: Restriction sites, RAPDs, microsatellites, likelihood, hidden Markov models – p.62/63*
**PhyloHMMs: used in the UCSC Genome Browser**

The conservation scores calculated in the Genome Browser use PhyloHMMs, which is just these HMM methods.

---

**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.