Week 6: Protein sequence models, likelihood, hidden Markov models

Genome 570

February, 2016

Variation of rates of evolution across sites

The basic way all these models deal with rates of evolution is using the fact that a rate r times as fast is exactly equivalent to evolving along a branch that is r times as long, so that it is easy to calculate the transition probabilities with rate r:

$$P_{ij}(r,\ t)\ =\ P_{ij}(r\,t)$$

The likelihood for a pair of sequences averages over all rates at each site, using the density function $f(r)$ for the distribution of rates:

$$L(t)\ =\ \prod_{i=1}^{\text{sites}}\ \left(\int_0^{\infty}\ f(r)\ \pi_{n_i}\ P_{m_i n_i}(r\,t)\ dr\right)$$

The Gamma distribution

$$f(r) = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}}\, x^{\alpha-1}\, e^{-\frac{x}{\beta}}$$

$$E[x] = \alpha\,\beta$$

$$\mathrm{Var}[x] = \alpha\,\beta^2$$

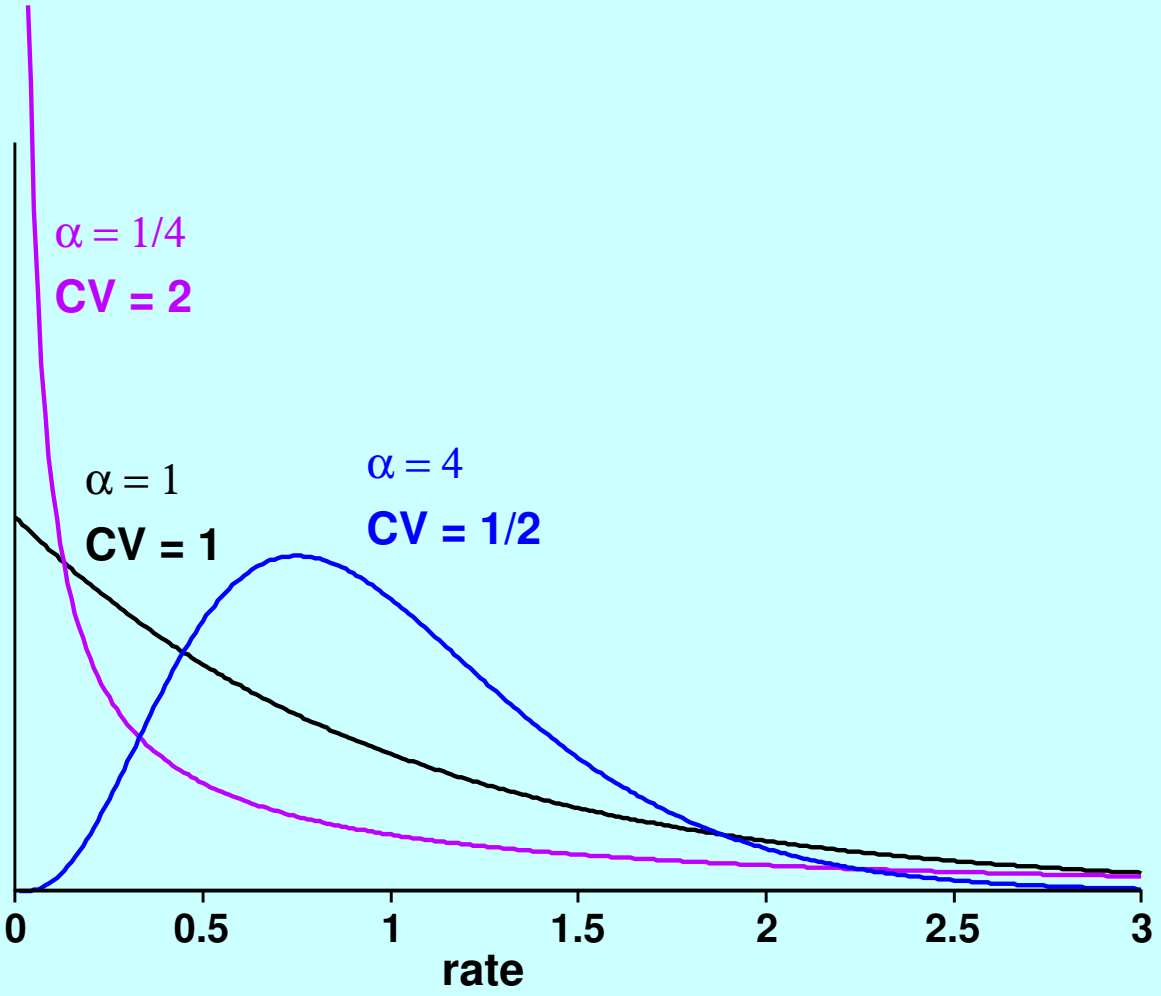
To get a mean of 1, set $\beta = 1/\alpha$ so that

$$f(r) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)}\, r^{\alpha-1}\, e^{-\alpha r}.$$

so that the squared coefficient of variation is $1/\alpha$.

Gamma distributions

Here are density functions of Gamma distributions for three different values of α:



Gamma rate variation in the Jukes-Cantor model

For example, for the Jukes-Cantor distance, to get the fraction of sites different we do

$$D_S = \int_0^\infty f(r) \frac{3}{4}\left(1 - e^{-\frac{4}{3} r\, u t}\right) dr$$

leading to the formula for D as a function of D_S

$$D = -\frac{3}{4}\alpha\left[1 - \left(1 - \frac{4}{3}D_S\right)^{-1/\alpha}\right]$$

Gamma rate variation in other models

For many other distances such as the Tamura-Nei family, the transition probabilites are of the form

$$P_{ij}(t) \;=\; A_{ij} \;+\; B_{ij}\, e^{-bt} \;+\; C_{ij}\, e^{-ct}$$

and integrating termwise we can make use of the fact that

$$E_r\left[e^{-b\,r\,t}\right] = \left(1 + \frac{1}{\alpha} b\, t\right)^{-\alpha}$$

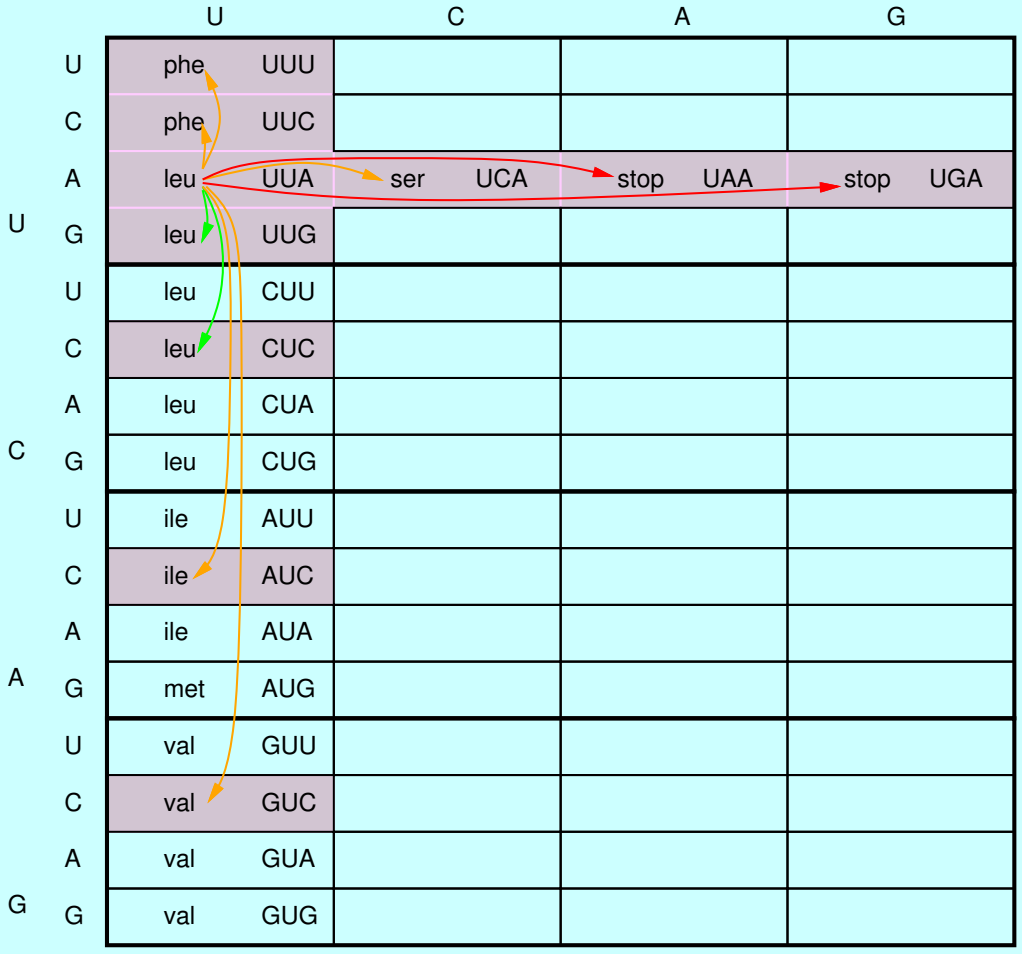
and just use that to replace e^{-bt} in the formulas for the transition probabilities.

Dayhoff's PAM001 matrix

		A ala	R arg	N asn	D asp	C cys	Q gln	E glu	G gly	H his	I ile	L leu	K lys	M met	F phe	P pro	S ser	T thr	W trp
A	ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0
R	arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8
N	asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1
D	asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0
C	cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0
Q	gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0
E	glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0
G	gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0
H	his	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1
I	ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0
L	leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4
K	lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0
M	met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0
F	phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3
P	pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0
S	ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5
T	thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0
W	trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	99
Y	tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2
V	val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0

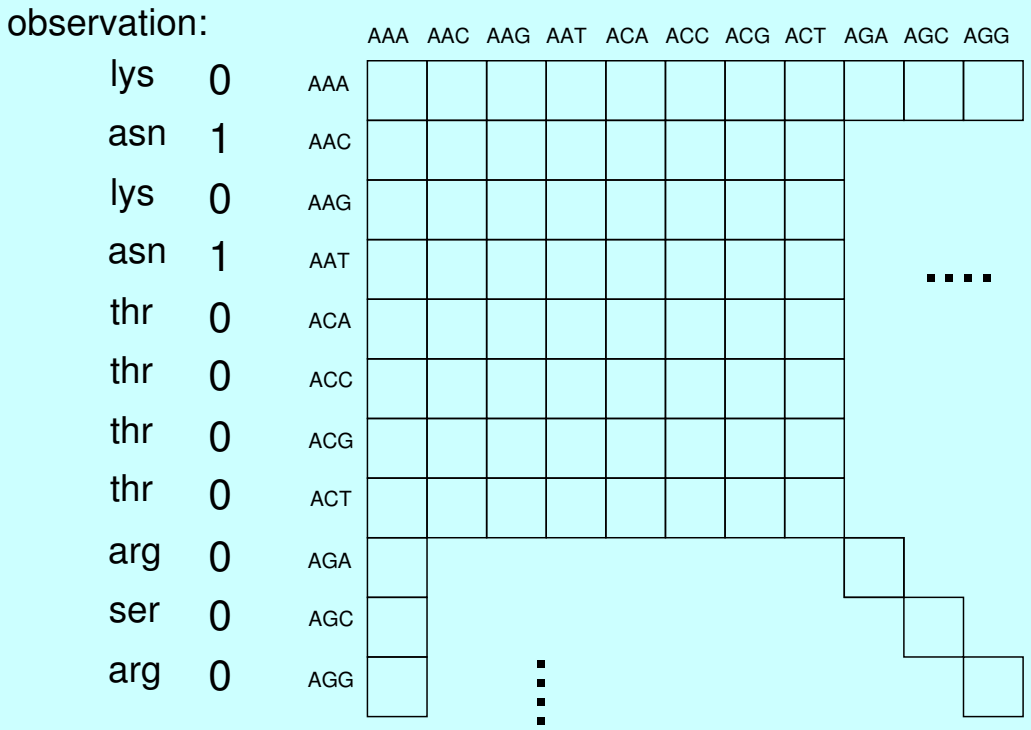
The codon model

Goldman & Yang, MBE 1994; Muse and Weir, MBE 1994



Probabilities of change vary depending on whether amino acid is changing, and to what

A codon-based model of protein evolution



In each cell:

$$P_{ij}(v)\, a_{ij}$$

where $P_{ij}(v)$ is the probability of codon change

and a_{ij} is the probability that the change is accepted

Considerations for a protein model

Making a model for protein evolution (a not-very-practical approach)

- Use a good model of DNA evolution.
- Use the appropriate genetic code.
- When an amino acid changes, accept it with probability that declines as the amino acids become more different.
- Fit this to empirical information on protein evolution.
- Take into account variation of rate from site to site.
- Take into account correlation of rates in adjacent sites.
- How about protein structure? Secondary structure? 3D structure?

(the first four steps are the "codon model" of Goldman and Yang, 1994 and Muse and Gaut, 1994, both in *Molecular Biology and Evolution*. The next two are the rate variation machinery of Yang, 1995, 1996 and Felsenstein and Churchill, 1996).

Likelihood ratios: the odds-ratio form of Bayes' theorem

$$\underbrace{\frac{\text{Prob}\,(H_1|D)}{\text{Prob}\,(H_2|D)}}_{\text{posterior odds ratio}} = \underbrace{\frac{\text{Prob}\,(D|H_1)}{\text{Prob}\,(D|H_2)}}_{\text{likelihood ratio}} \underbrace{\frac{\text{Prob}\,(H_1)}{\text{Prob}\,(H_2)}}_{\text{prior odds ratio}}$$

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:

$$\begin{aligned} & \text{Prob}\,(D|H_i) \\ = \;& \text{Prob}\,(D^{(1)}|H_i)\;\; \text{Prob}\,(D^{(2)}|H_i)\;\; \ldots \;\; \text{Prob}\,(D^{(n)}|H_i) \end{aligned}$$

so we can rewrite the odds-ratio formula as

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

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 `HHTTHTHHTTT`, the likelihood is:

$$\begin{aligned} L &= \text{Prob}\,(D|p) \\ &= p\,p\,(1-p)\,(1-p)\,p\,(1-p)\,p\,p\,(1-p)\,(1-p)\,(1-p) \\ &= p^5(1-p)^6 \end{aligned}$$

Solving for the maximum likelihood estimate for p by finding the maximum:

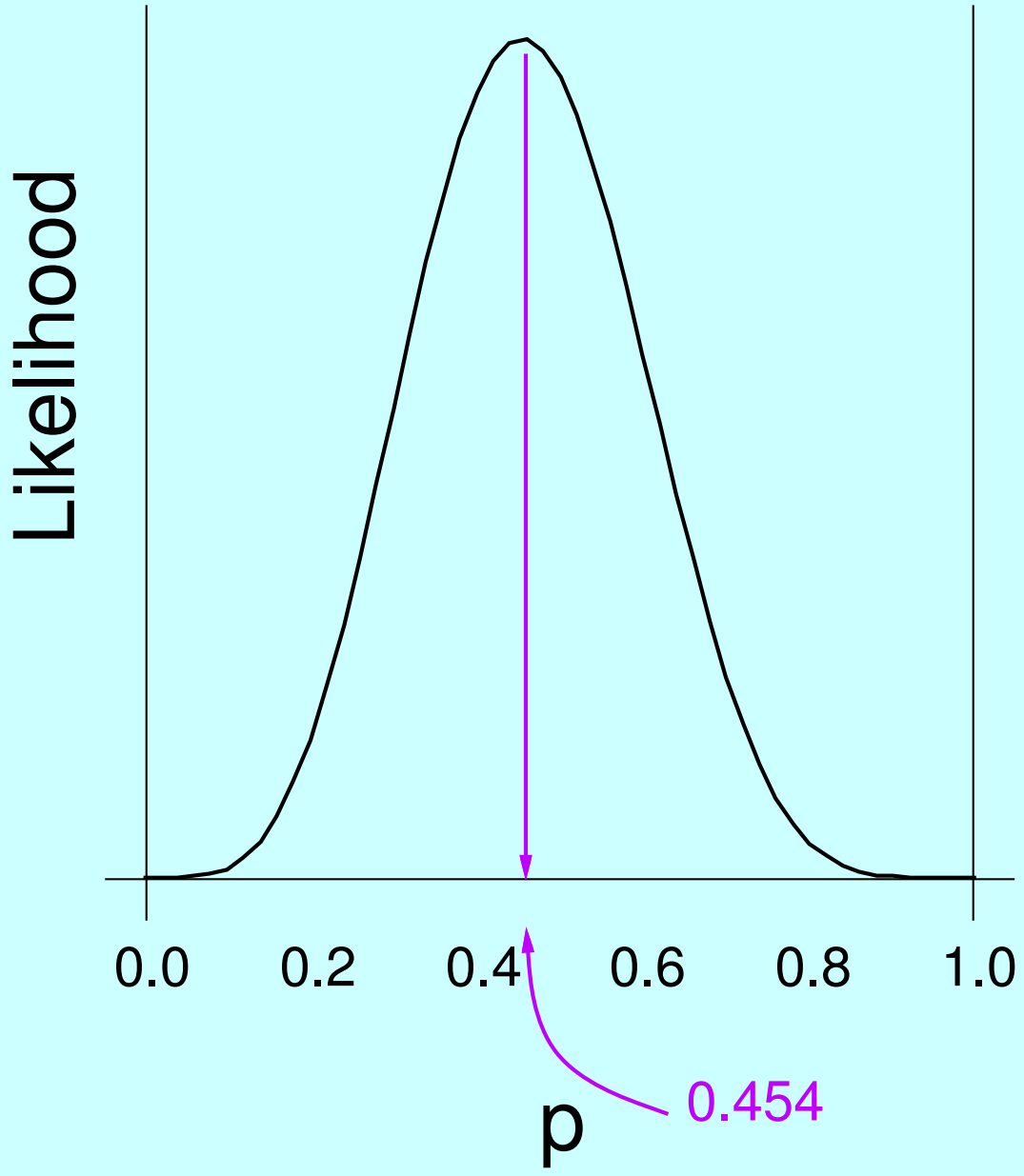
$$\frac{dL}{dp} = 5p^4(1-p)^6 - 6p^5(1-p)^5$$

and equating it to zero and solving:

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

gives $\widehat{p} = 5/11$

Likelihood as function of p for the 11 coin tosses



Maximizing the likelihood

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

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

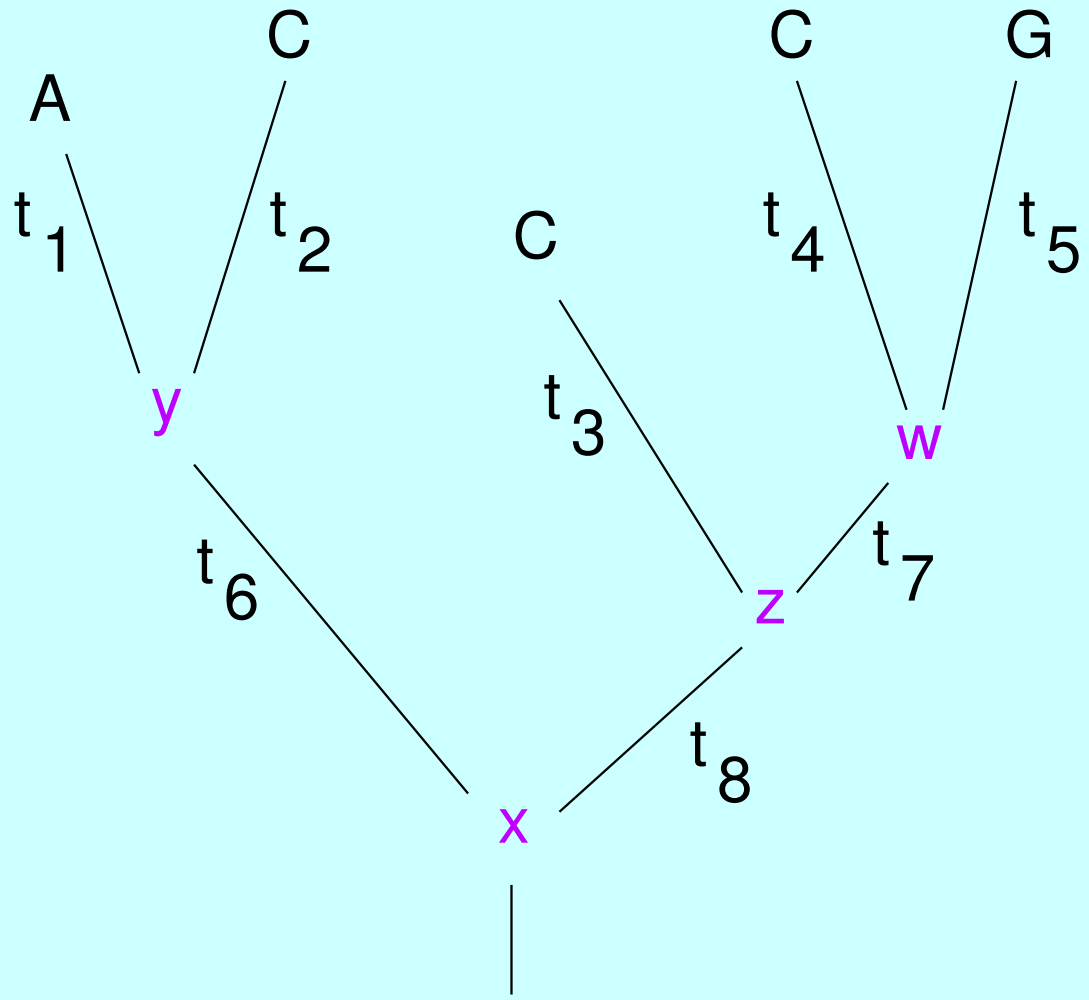
and

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

so that, again,

$$\widehat{p} \;=\; 5/11$$

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):

$$\text{Prob}\,(\mathsf{A},\mathsf{C},\mathsf{C},\mathsf{C},\mathsf{G},\mathsf{x},\mathsf{y},\mathsf{z},\mathsf{w}|\mathsf{T}) =$$

$$\text{Prob}\,(\mathsf{x})\ \text{Prob}\,(\mathsf{y}|\mathsf{x},\mathsf{t}_6)\ \text{Prob}\,(\mathsf{A}|\mathsf{y},\mathsf{t}_1)\ \text{Prob}\,(\mathsf{C}|\mathsf{y},\mathsf{t}_2)$$

$$\text{Prob}\,(\mathsf{z}|\mathsf{x},\mathsf{t}_8)\ \text{Prob}\,(\mathsf{C}|\mathsf{z},\mathsf{t}_3)$$

$$\text{Prob}\,(\mathsf{w}|\mathsf{z},\mathsf{t}_7)\ \text{Prob}\,(\mathsf{C}|\mathsf{w},\mathsf{t}_4)\ \text{Prob}\,(\mathsf{G}|\mathsf{w},\mathsf{t}_5)$$

The result looks hard to compute

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

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

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:

$$\begin{aligned}
&\text{Prob}\ \left(\mathsf{D}^{(\mathsf{i})}|\mathsf{T}\right) \quad = \\
&\sum_{\mathsf{x}} \text{Prob}\ (\mathsf{x}) \left(\sum_{\mathsf{y}} \text{Prob}\ (\mathsf{y}|\mathsf{x}, \mathsf{t}_6)\ \text{Prob}\ (\mathsf{A}|\mathsf{y}, \mathsf{t}_1)\ \text{Prob}\ (\mathsf{C}|\mathsf{y}, \mathsf{t}_2) \right) \\
&\qquad \left(\sum_{\mathsf{z}} \text{Prob}\ (\mathsf{z}|\mathsf{x}, \mathsf{t}_8)\ \text{Prob}\ (\mathsf{C}|\mathsf{z}, \mathsf{t}_3) \right. \\
&\qquad \left. \left(\sum_{\mathsf{w}} \text{Prob}\ (\mathsf{w}|\mathsf{z}, \mathsf{t}_7)\ \text{Prob}\ (\mathsf{C}|\mathsf{w}, \mathsf{t}_4)\ \text{Prob}\ (\mathsf{G}|\mathsf{w}, \mathsf{t}_5) \right) \right)
\end{aligned}$$

The pattern of parentheses parallels the structure of the tree:

$$(\mathsf{A}, \mathsf{C})\,(\mathsf{C}, (\mathsf{C}, \mathsf{G}))$$

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:

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

Another is the term including that:

$$L_8(z) = \text{Prob}\,(C|z, t_3)\left(\sum_w \text{Prob}\,(w|z, t_7)\ \text{Prob}\,(C|w, t_4)\ \text{Prob}\,(G|w, t_5)\right)$$

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: $(1,\ 0,\ 1,\ 0)$

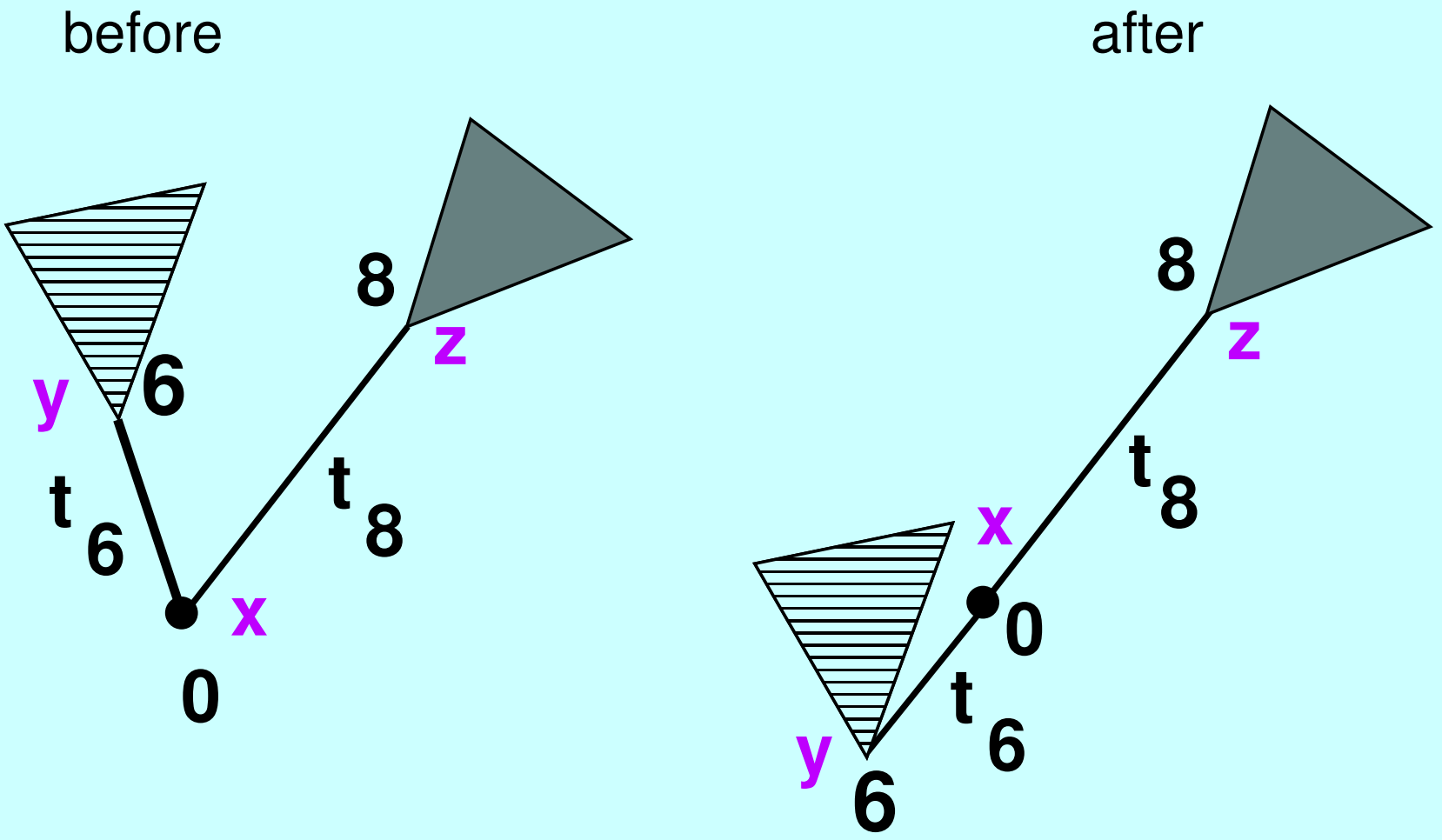
Note – do *not* do something like this: $(0.5,\ 0,\ 0.5,\ 0)$. It is *not* the probability of *being* that base but the probability of the observation *given* that base.

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



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

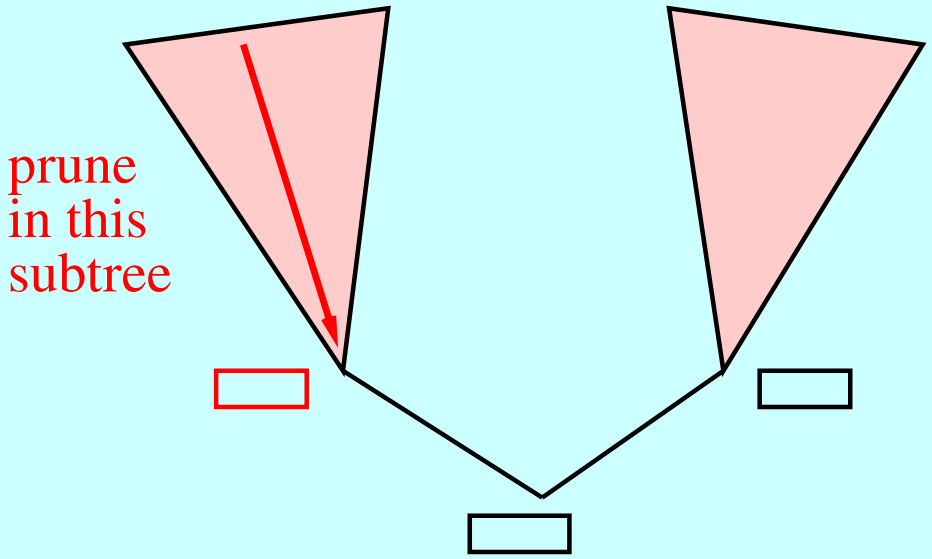
$$\text{Prob}\,(x)\ \text{Prob}\,(y|x, t_6) = \text{Prob}\,(y)\ \text{Prob}\,(x|y, t_6).$$

Substituting that in:

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

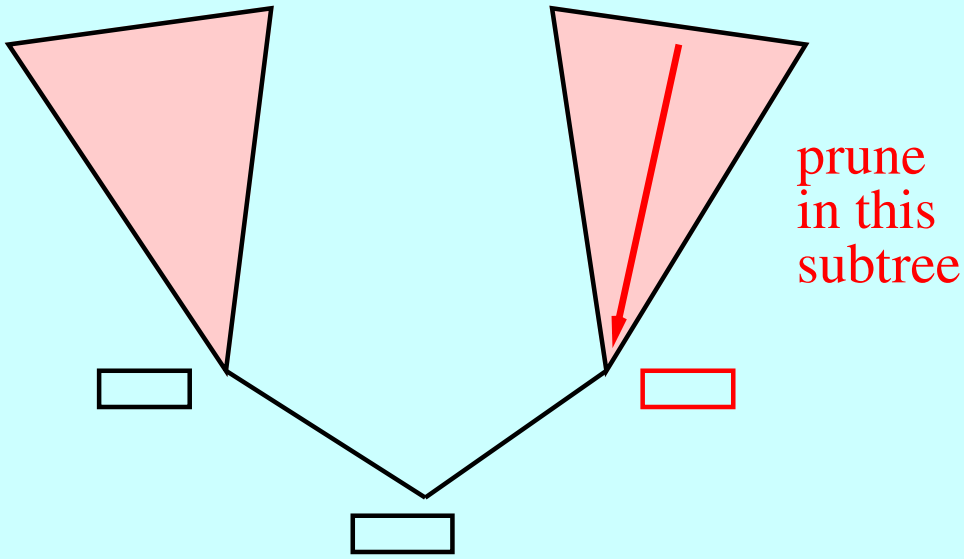
... which means that (if the model of change is a reversible one) the likelihood does not depend on where the root is placed in the unrooted tree.

To compute likelihood when the root is in one branch



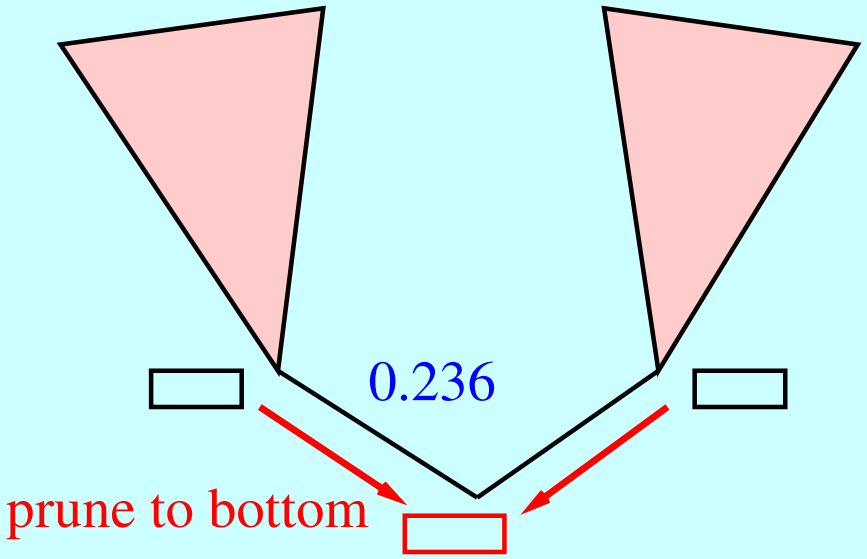
First we get (for the site) the conditional likelihoods for the left subtree ...

To compute likelihood when the root is in one branch



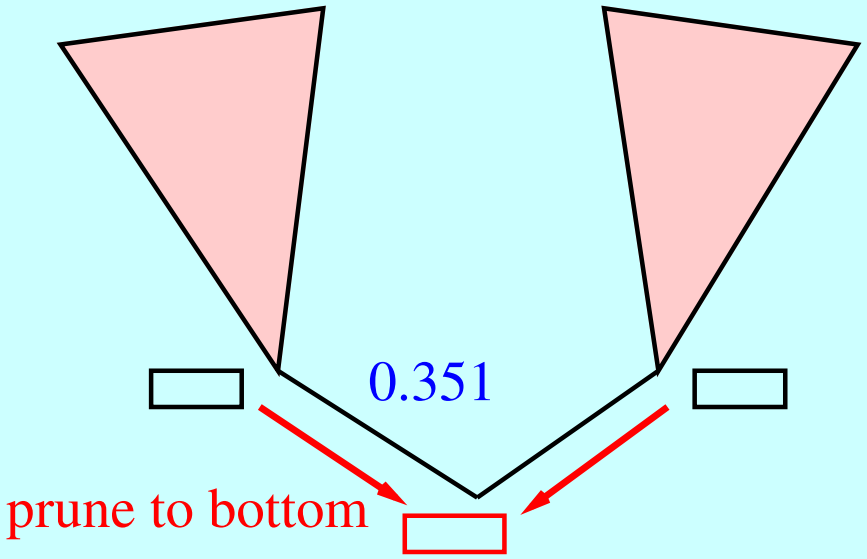
... then we get the conditional likelihoods for the right subtree

To compute likelihood when the root is in one branch



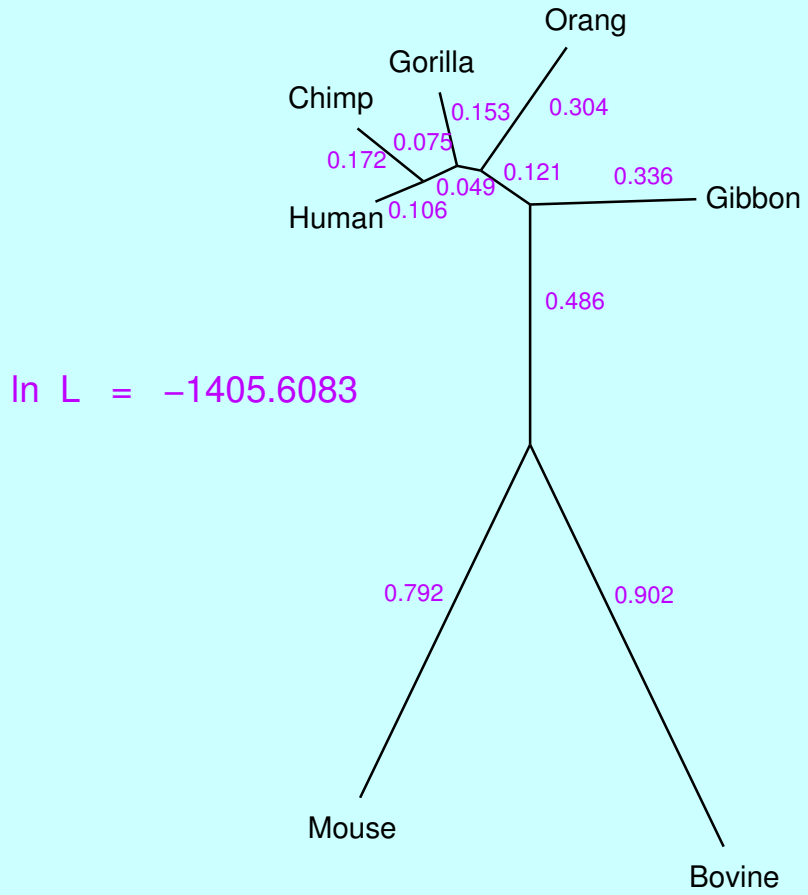
... and finally we get the conditional likelihoods for the root, and from that the likelhoods for the site. (Note that it does not matter where in that branch we place the root, as long as the sum of the branch lengths on either side of the root is 0.236).

To do it with a different branch length ...



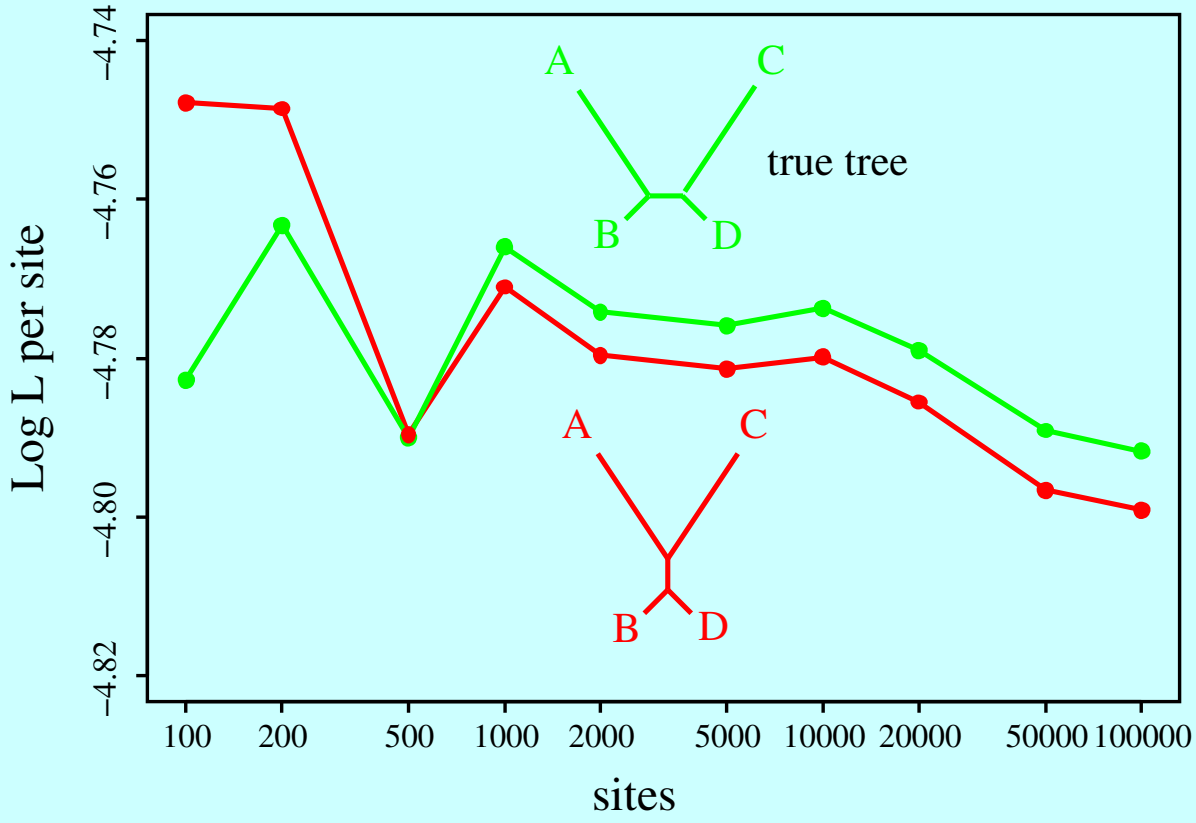
... we can just do the last step, but now with the new branch length. The other conditional likelihoods were already calculated, have not changed, and so we can re-use them. This really speeds things up for calculating how the likelihood depends on the length of one branch – just put the root there!

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.

A simulation showing consistency of the ML tree



As more and more sites are added the true tree topology is favored. Since the vertical scale here is $\log(L)$ per site, the difference of log-likelihoods becomes very great.

Rate variation among sites

Evolution is independent once each site has had its rate specified

$$\text{Prob}\,(\mathsf{D} \mid \mathsf{T}, \mathsf{r}_1, \mathsf{r}_2, \ldots, \mathsf{r}_p) \;=$$

$$\prod_{i=1}^{p} \text{Prob}\,(\mathsf{D}^{(i)} \mid \mathsf{T}, \mathsf{r}_i).$$

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:

$$L^{(i)} = \int_0^{\infty} f(r;\alpha)\, L^{(i)}(r)\, dr \simeq \sum_{j=1}^{k} w_k L^{(i)}(r_k)$$

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 $\mathrm{Prob}\ (j|i)$ 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

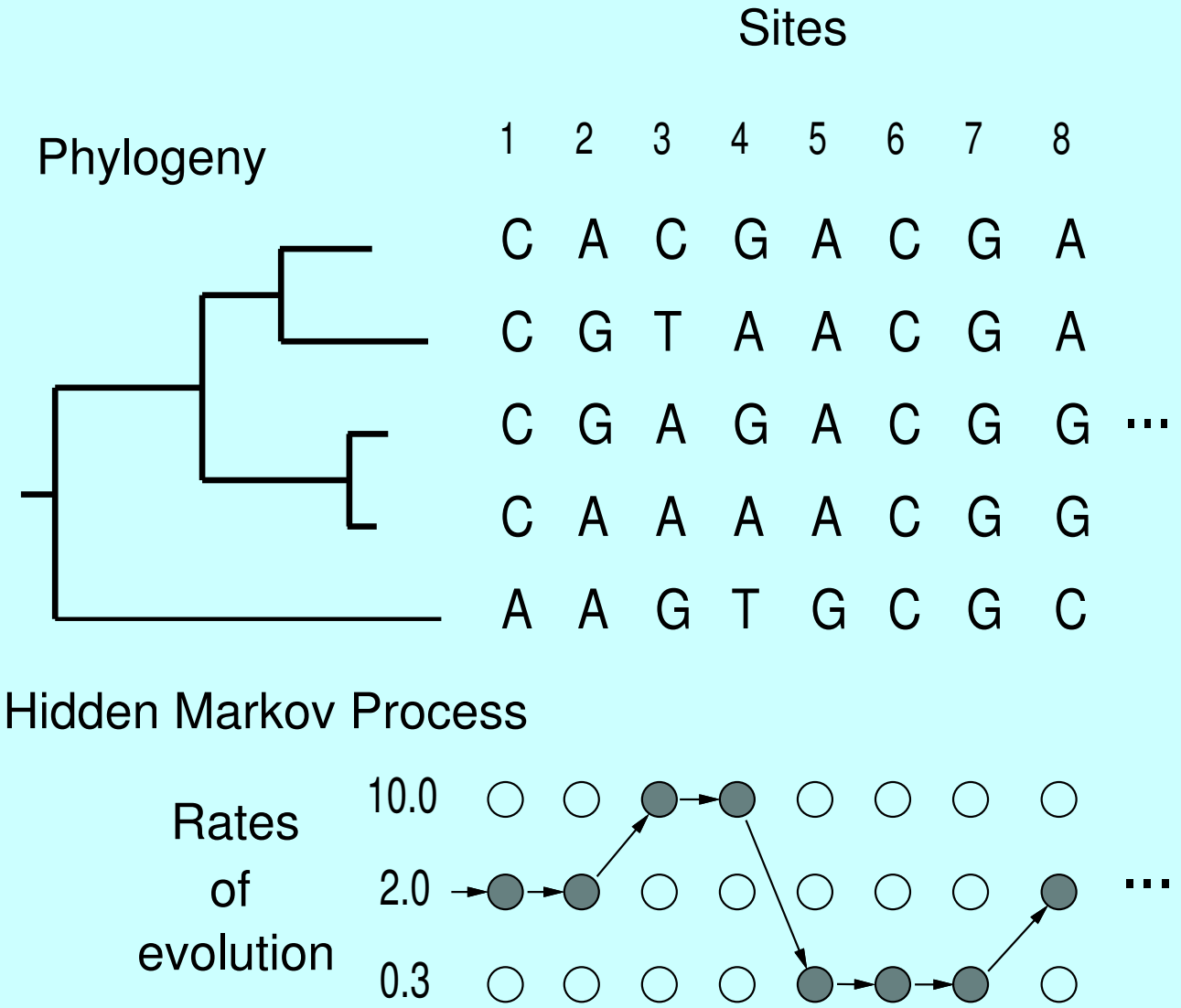
Suppose that we have a way of calculating, for each possible rate at each possible site, the probability of the data at that site (i) given that rate (r_j). This is

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

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



Hidden Markov Models

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

$$\begin{aligned}\text{Prob}\,(\mathsf{D} \mid \mathsf{T}) \;&=\; \sum_{\mathsf{i}_1}\sum_{\mathsf{i}_2}\cdots\sum_{\mathsf{i}_p}\text{Prob}\,(\mathsf{r}_{\mathsf{i}_1},\mathsf{r}_{\mathsf{i}_2},\ldots\mathsf{r}_{\mathsf{i}_p}) \\ &\qquad\times\text{Prob}\,(\mathsf{D} \mid \mathsf{T},\mathsf{r}_{\mathsf{i}_1},\mathsf{r}_{\mathsf{i}_2},\ldots\mathsf{r}_{\mathsf{i}_m})\end{aligned}$$

Evolution is independent once each site has had its rate specified

$$\text{Prob}\,(\mathsf{D} \mid \mathsf{T},\mathsf{r}_1,\mathsf{r}_2,\ldots,\mathsf{r}_p) \;=$$

$$\prod_{\mathsf{i}=1}^{\mathsf{p}}\text{Prob}\,(\mathsf{D}^{(\mathsf{i})} \mid \mathsf{T},\mathsf{r}_\mathsf{i}).$$

Seems impossible ...

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

$$
\begin{aligned}
L &= \text{Prob}\,(D \mid T) \\
&= \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \cdots \sum_{i_p=1}^{m} \text{Prob}\,\left(r_{i_1}, r_{i_2}, \ldots, r_{i_p}\right) \\
&\quad \times \text{Prob}\,\left(D^{(1)} \mid r_{i_1}\right) \text{Prob}\,\left(D^{(2)} \mid r_{i_2}\right) \ldots \text{Prob}\,\left(D^{(n)} \mid r_{i_p}\right)
\end{aligned}
$$

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:

$$
\begin{aligned}
L &= \text{Prob}\,(D \mid T) \\
&= \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \cdots \sum_{i_p=1}^{m} \text{Prob}\,(i_1)\,\text{Prob}\,\left(D^{(1)} \mid r_{i_1}\right) \\
&\qquad \times \text{Prob}\,(i_2 \mid i_1)\,\text{Prob}\,\left(D^{(2)} \mid r_{i_2}\right) \\
&\qquad \times \text{Prob}\,(i_3 \mid i_2)\,\text{Prob}\,\left(D^{(3)} \mid r_{i_3}\right) \\
&\qquad \vdots \\
&\qquad \times \text{Prob}\,(i_p \mid i_{p-1})\,\text{Prob}\,\left(D^{(p)} \mid r_{i_p}\right)
\end{aligned}
$$

Using Horner's Rule

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

$$
\begin{aligned}
L \;=\; & \sum_{i_1=1}^{m} \mathrm{Prob}\,(i_1)\,\mathrm{Prob}\,\left(D^{(1)} \mid r_{i_1}\right) \\
& \sum_{i_2=1}^{m} \mathrm{Prob}\,(i_2 \mid i_1)\,\mathrm{Prob}\,\left(D^{(2)} \mid r_{i_2}\right) \\
& \sum_{i_3=1}^{m} \mathrm{Prob}\,(i_3 \mid i_2)\,\mathrm{Prob}\,\left(D^{(3)} \mid r_{i_3}\right) \\
& \vdots \\
& \sum_{i_p=1}^{m} \mathrm{Prob}\,(i_p \mid i_{p-1})\,\mathrm{Prob}\,\left(D^{(p)} \mid r_{i_p}\right)
\end{aligned}
$$

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 $\mathcal{P}_i(j)$ 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

$$\mathcal{P}_p(i_p) \;=\; \mathrm{Prob}\,\left(D^{(p)} \mid r_{i_p}\right)$$

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

Recursive calculation

More generally, for site $\ell < \mathsf{p}$ and its rates i_ℓ

$$\mathcal{P}_\ell(\mathsf{i}_\ell) \;=\; \mathrm{Prob}\,\left(\mathsf{D}^{(\ell)} \mid \mathsf{r}_{\mathsf{i}_\ell}\right) \sum_{\mathsf{i}_\ell=1}^{\mathsf{m}} \mathrm{Prob}\,(\mathsf{i}_{\ell+1} \mid \mathsf{i}_\ell)\, \mathcal{P}_{\ell+1}(\mathsf{i}_{\ell+1})$$

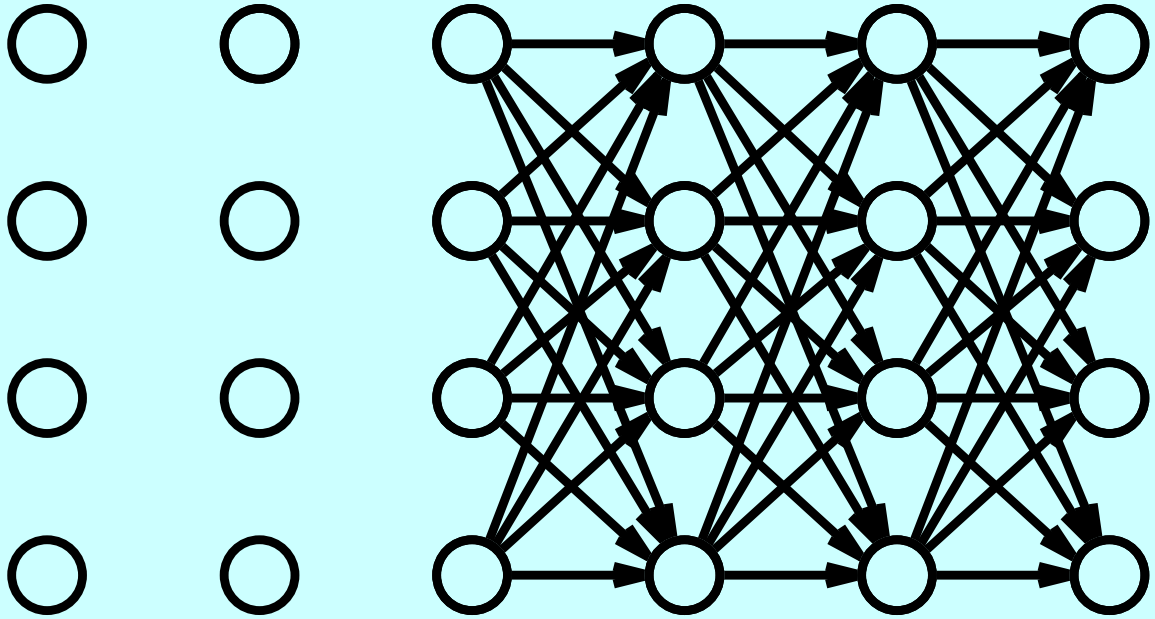
We can compute the $\mathcal{P}$'s recursively using this, starting with the last site and moving leftwards down the sequence. Finally we have the $\mathcal{P}_1(\mathsf{i}_1)$ for all m states. These are simply weighted by the equilibrium probabilities of the Markov chain of rate categories:

$$\mathsf{L} \;=\; \mathrm{Prob}\,(\mathsf{D} \mid \mathsf{T}) \;=\; \sum_{\mathsf{i}_1=1}^{\mathsf{m}} \mathrm{Prob}\,(\mathsf{i}_1)\, \mathcal{P}_1(\mathsf{i}_1)$$

An entirely similar calculation can be done from left to right, remembering that the transition probabilities $\mathrm{Prob}\,(\mathsf{i}_\mathsf{k}|\mathsf{i}_{\mathsf{k}+1})$ 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

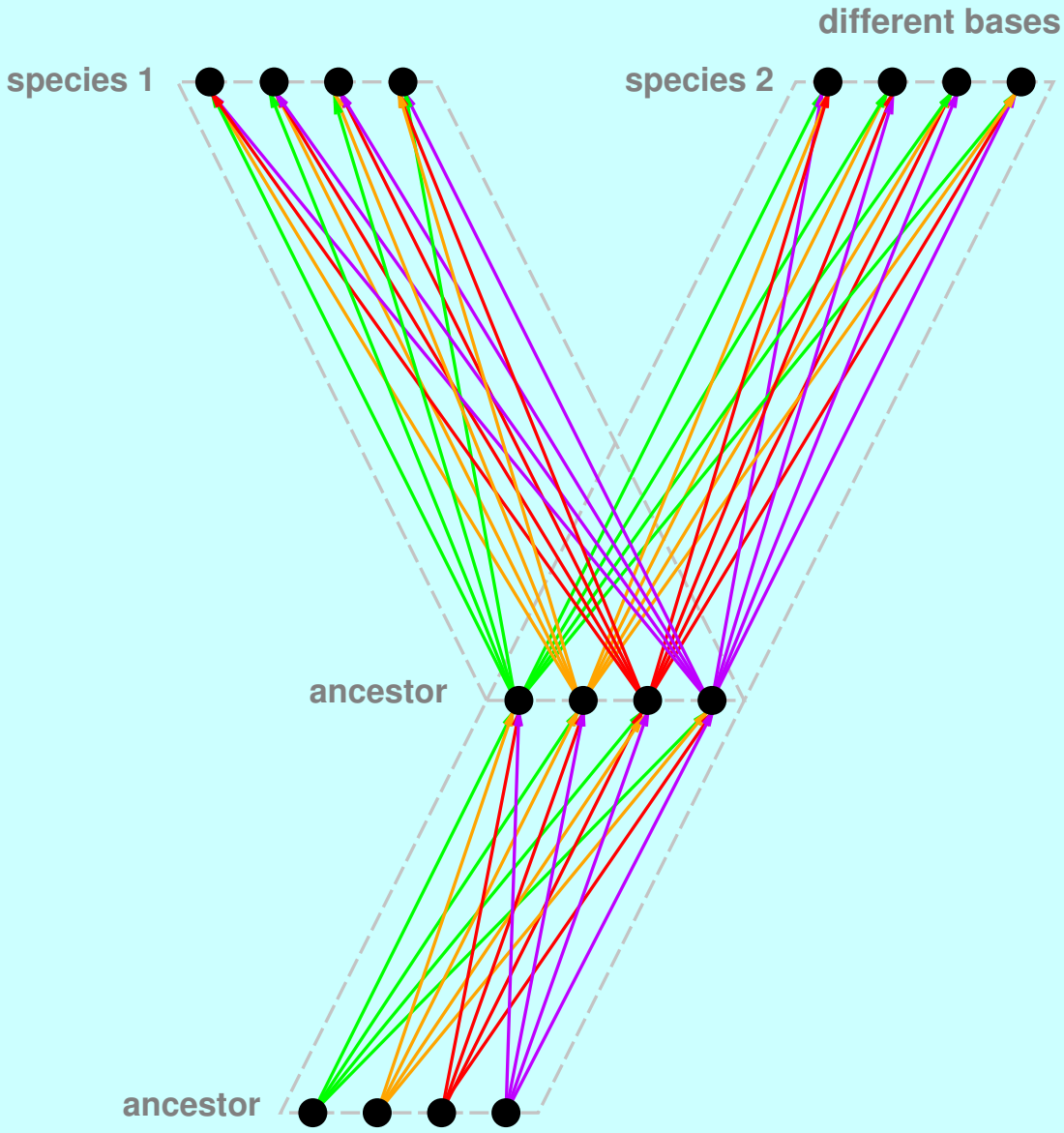
At the end, at site m:

$$\text{Prob}\,(D^{[m]}|T, r_{i_m}) \;=\; \text{Prob}\,(D^{(m)}|T, r_{i_m})$$

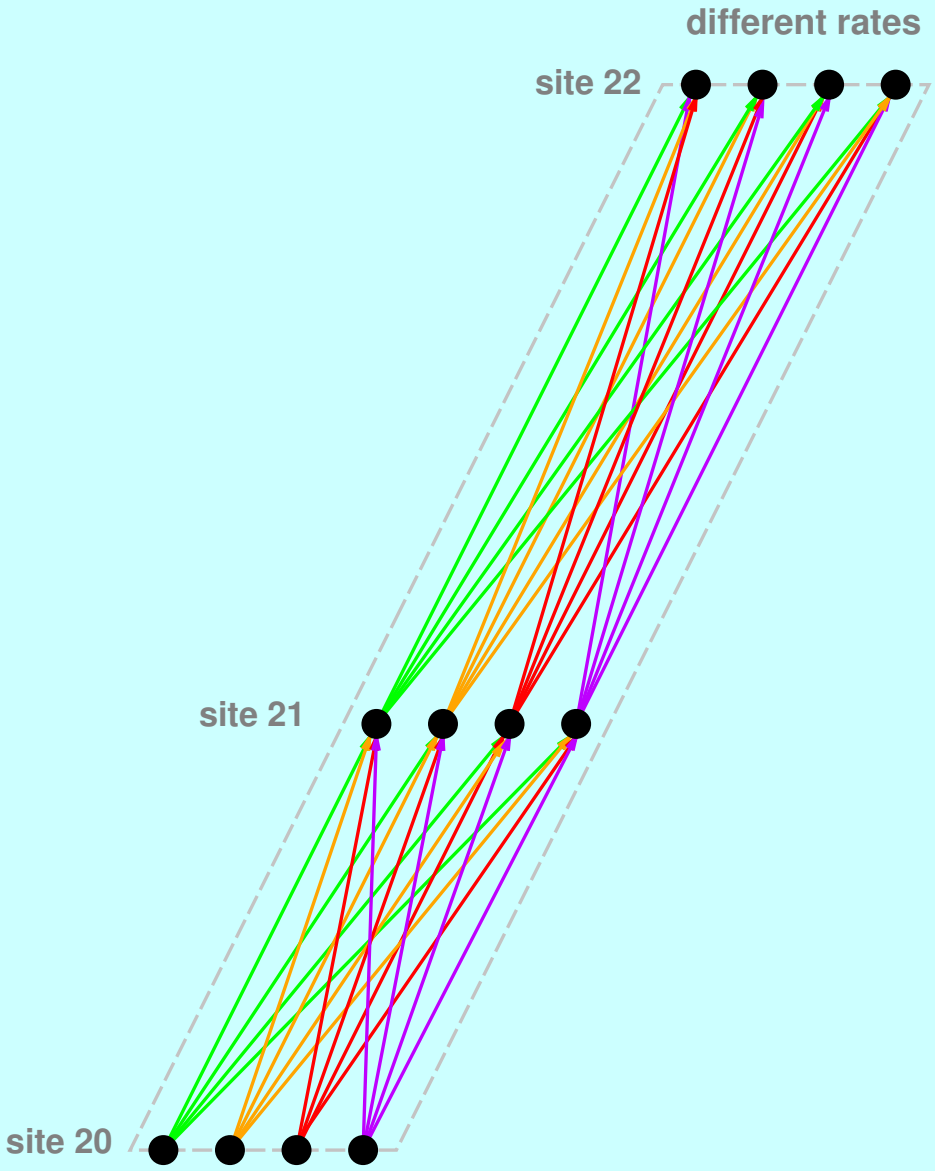
and once we get to site 1, we need only use the prior probabilities of the rates r_i to get a weighted sum:

$$\text{Prob}\,(D|T) \;=\; \sum_{i_1} \pi_{i_1}\, \text{Prob}\,(D^{[1]}|T, r_{i_1})$$

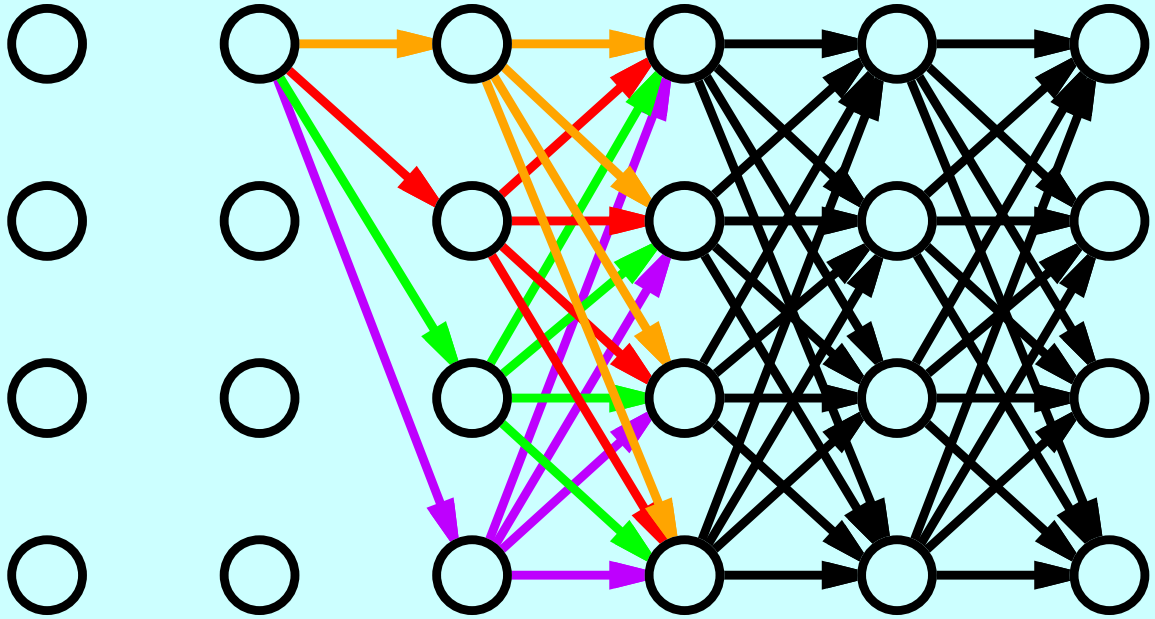
The pruning algorithm is just like



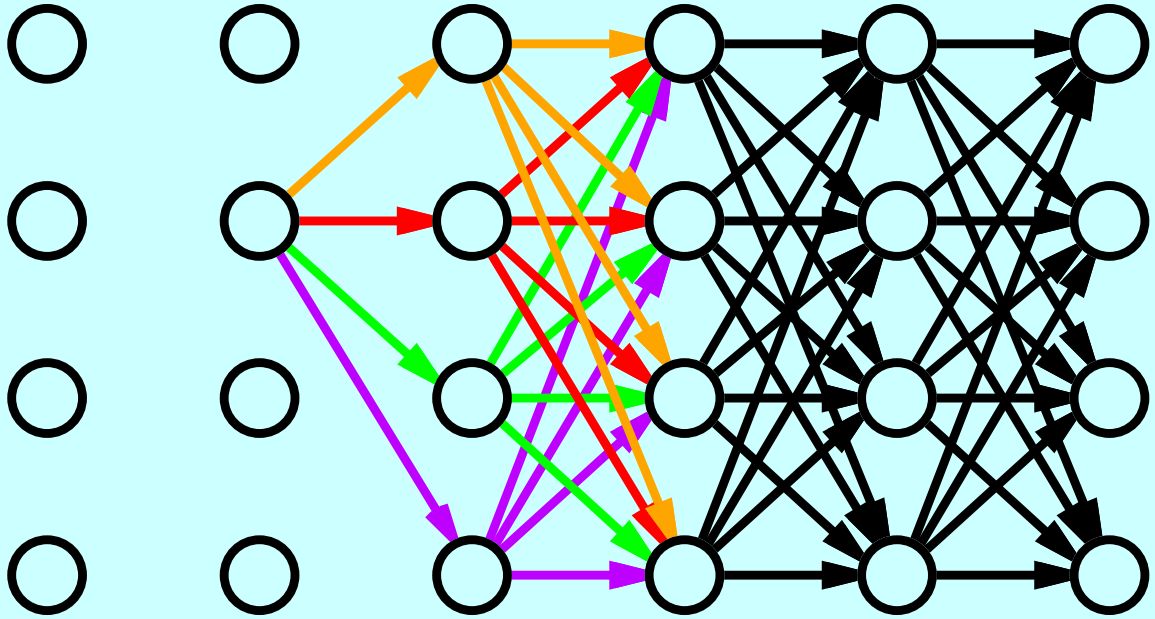
the forward algorithm



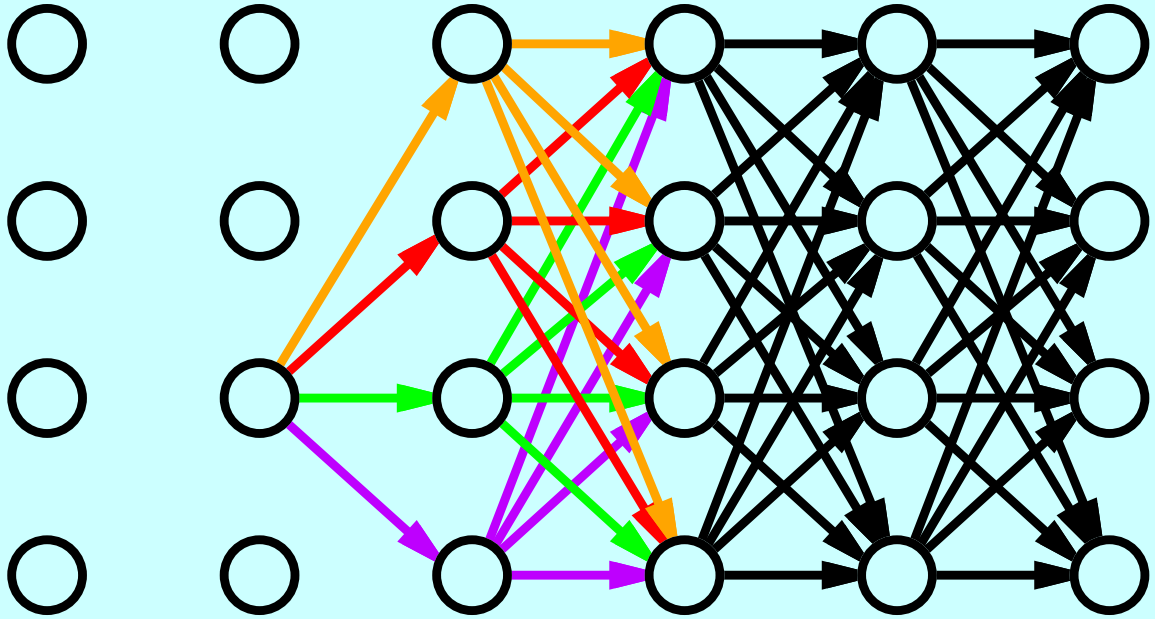
Paths from one state in one site



Paths from another state in that site

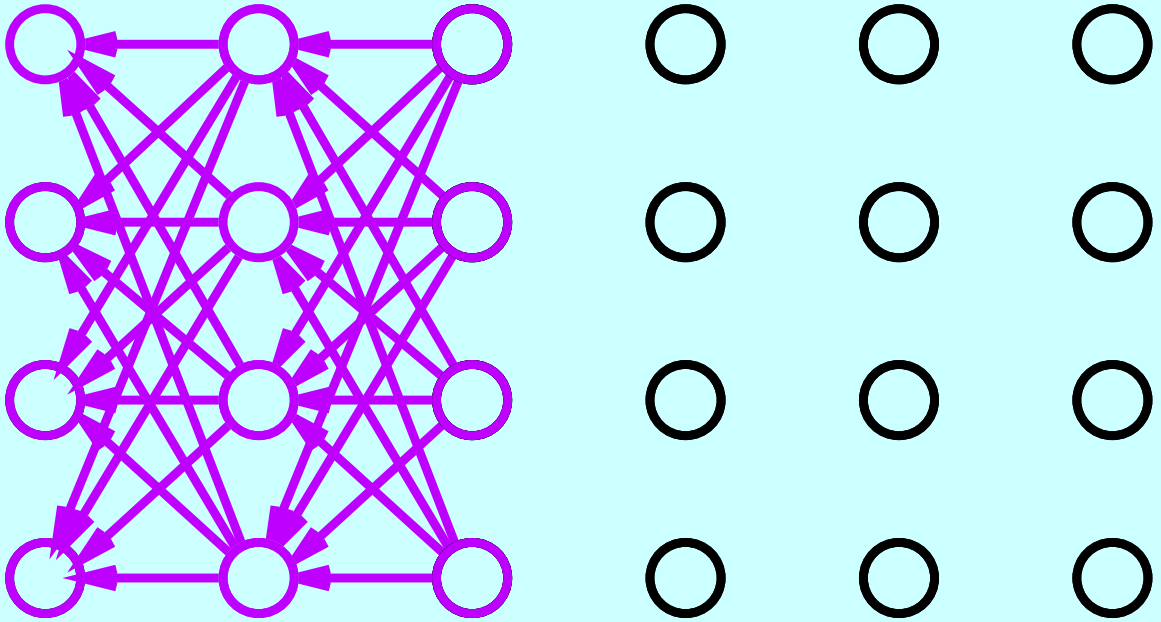


Paths from a third state



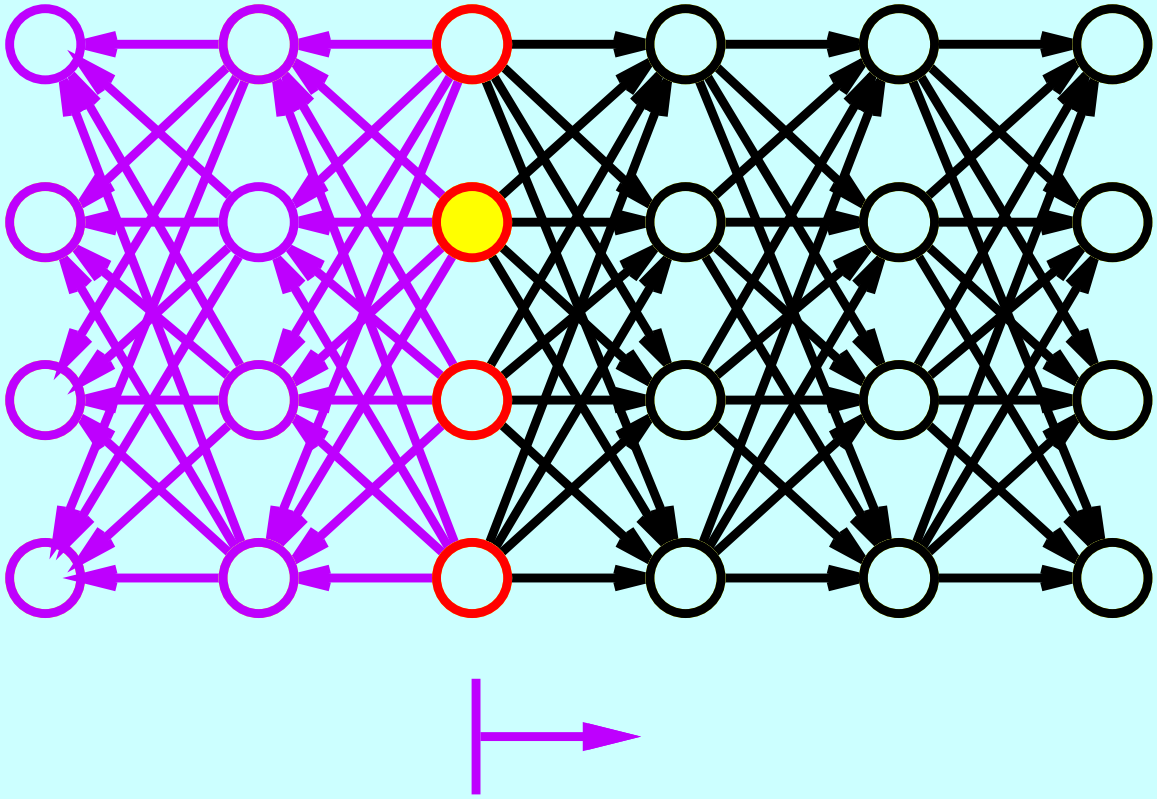
We can also use the backward algorithm

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:

$$\text{Prob}\,(r_i|r_j) \;=\; (1-\lambda)\,\delta_{ij} + \lambda\,\pi_i$$

A numerical example. Cytochrome B

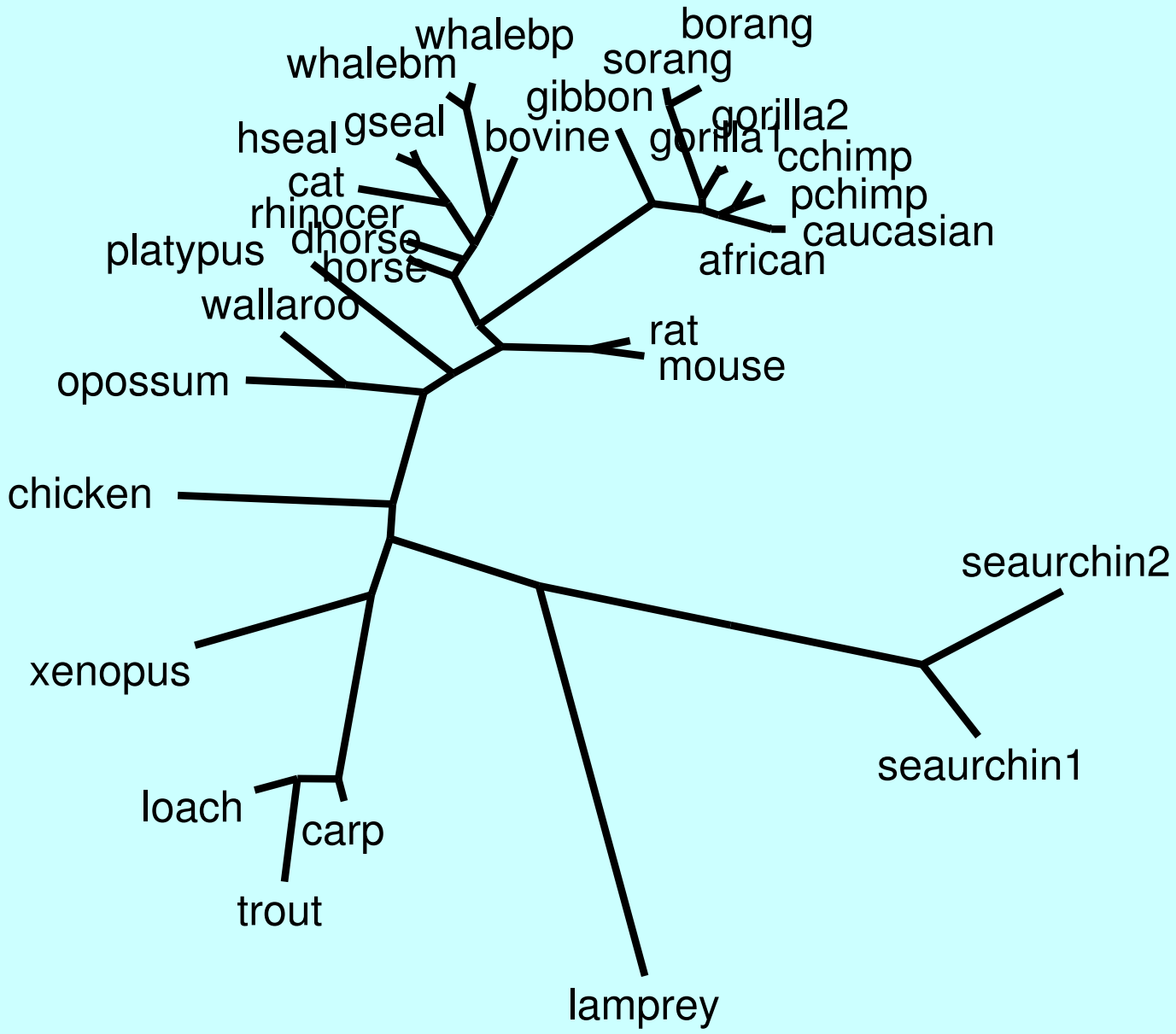
We analyze 31 cytochrome B sequences, aligned by Naoko Takezaki, using the Proml protein maximum likelihood program. Assume a Hidden Markov Model with 3 states, rates:

category	rate	probability
1	0.0	0.2
2	1.0	0.4
3	3.0	0.4

and expected block length 3.

We get a reasonable, but not perfect, tree with the best rate combination inferred to be

Phylogeny for Takezaki cytochrome B



Rates inferred from Cytochrome B

```
                1333333311 3222322313 3321113222 2133111111 1331133123 1122111

african         M-----TPMR KINPLMKLIN HSFIDLPTPS NISAWWNFGS LLGACLILQI TTGLFLA
caucasian       .......... .........R .......... .......... ..T....... .......
cchimp          .......T.. .......... .......... .......... .......... .......
pchimp          .......T.. .......... .......... ..T....... .......... .......
gorilla1        .......... T...A..... .......... ..T....... .......... .......
gorilla2        .......... T...A..... .......... ..T....... .......... .......
borang          .......... T......... .L........ .......... ......I.TI .......
sorang          ......ST.. T......... .L........ .......... ......I... .......
gibbon          .......L.. T......... .L....A... ..M....... ..........I .......
bovine          ......NI.. SH....IV.N A.....A... ..S....... ..I......L .......
whalebm         ......NI.. TH....I..D A......... ..S....... ..L...V..L .......
whalebp         ......NI.. TH....IV.D A.V....... ..S....... ..L...M..L .......
dhorse          ......NI.. SH..I.I... ......A... ..S....... ..I......L .......
horse           ......NI.. SH..I.I... .......... ..S....... ..I......L .......
rhinocer        ......NI.. SH..V.I... .......... ..S....... ..I......L .......
cat             ......NI.. SH..I.I... ......A... .......... ..V..T...L .......
gseal           ......NI.. TH....I..N .......... .......... ..I......L .......
hseal           ......NI.. TH....I..N .......... .......... ..I......L .......
mouse           ......N.... TH..F.I... ......A... ..S....... ..V..MV..I .......
rat             ......NI.. SH..F.I... ......A... ..S....... ..V..MV..L .......
platypus        .....NNL.. TH..I.IV.. .......... ..S....... ..L...I..L .......
wallaroo        ......NL.. SH..I.IV.. ......A... .......... ......I..L .......
opossum         ......NI.. TH....I..D .......... .......... ..V...I..L .......
chicken         ....APNI.. SH..L.M..N .L....A... .......... .AV..MT..L ...L...
xenopus         ....APNI.. SH..I.I..N .......... ..SL...... ..V...A..I .......
carp            ....A-SL.. TH..I.IA.D ALV....... .......... ..L...T..L .......
loach           ....A-SL.. TH..I.IA.D ALV...A... ..V....... ..L...T..L .......
trout           ....A-NL.. TH..L.IA.D ALV...A... ..V....... ..L..AT..L .......
lamprey         .SHQPSII.. TH..LS.G.S MLV...S.A. .......... .SL......I ...I...
seaurchin1      -...LG.L.. EH.IFRIL.S T.V...L... L.I....... ..L...T..L .......
seaurchin2      -...AG.L.. EH.IFRIL.S T.V...L... L.M....... ..L..I.LI ..I....
```

Rates inferred from Cytochrome B

```
                2223311112 2222222222 2222232112 2222222223 1222221112 3333111

african         PDASTAFSSI AHITRDVNYG WIIRYLHANG ASMFFICLFL HIGRGLYYGS FLYSETW
caucasian       .......... .......... .......... .......... .......... .......
cchimp          .......... .......... .......... .......... .......... ...L...
pchimp          .......... .......... .......... ...L...... .V........ ...L...
gorilla1        .......... .......... .T........ .......... .......... ..HQ...
gorilla2        .......... .......... .T........ .......... .......... ..HQ...
borang          ...T...... .......... .M..H..... ...L...... .......... .THL...
sorang          .......... .......... .M..H..... .......... .......... .THL...
gibbon          .........V .......... .......... .......... .......... ...L...
bovine          S.TT.....V T..C...... .....M.... ........YM .V........ YTFL...
whalebm         ..TM.....V T..C...... .V........ ........YA .M........ HAFR...
whalebp         ..TT.....V T..C...... .......... ........YA .M........ YAFR...
dhorse          S.TT.....V T..C...... .......... .........I .V........ YTFL...
horse           S.TT.....V T..C...... .......... .........I .V........ YTFL...
rhinocer        ..TT.....V T..C...... .M........ .........I .V........ YTFL...
cat             S.TM.....V T..C...... .......... ........YM .V...M.... YTF....
gseal           S.TT.....V T..C...... .......... ........YM .V........ YTFT...
hseal           S.TT.....V T..C...... .......... ........YM .V........ YTFT...
mouse           S.TM.....V T..C...... .L...M.... .......... .V........ YTFM...
rat             S.TM.....V T..C...... .L....Q... .......... .V........ YTFL...
platypus        S.T......V ...C...... .L...M.... ..L..M.I.. .......... YTQT...
wallaroo        S.TL.....V ...C...... .L..N..... .....M.... .V...I.... Y..K...
opossum         S.TL.....V ...C...... .L..NI.... .....M.... .V...I.... Y..K...
chicken         A.T.L....V ..TC.N.Q.. .L..N..... ..F....I.. .......... Y..K...
xenopus         A.T.M....V ...CF..... LL..N..... L.F....IY. .......... ...K...
carp            S.I......V T..C...... .L..NV.... ..F....IYM ..A....... Y..K...
loach           S.I......V ...C...... .L..NI.... ..F.....Y. ..A....... Y..K...
trout           S.I......V C..C...S.. .L..NI.... ..F....IYM ..A....... Y..K...
lamprey         ANTEL....V M..C....N. .LM.N..... ........IYA .....I.... Y..K...
seaurchin1      A.I.L....A S..C...... .LL.NV.... ..L....MYC .........G SNKI...
seaurchin2      A.INL....V S..C...... .LL.NV...C ..L....MYC .........L TNKI...
```

PhyloHMMs: used in the UCSC Genome Browser

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

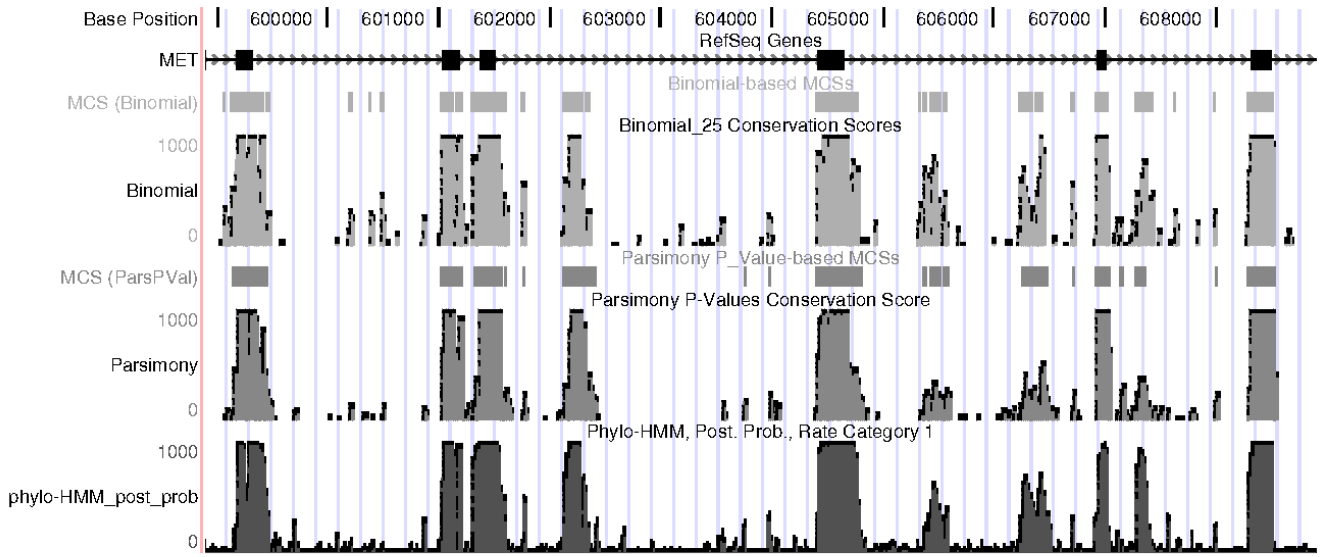


Fig. 5. A screen shot from the UCSC Genome Browser [24] showing a selected region of the data set of example 2, including several exons of the *MET* gene (black boxes at top). The binomial-based (light gray) and parsimony-based (medium gray) conservation scores of Margulies et al. [30] are shown as tracks in the browser, as are the posterior probabilities ($\times 1000$) 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.