Likelihood and phylogenies

Joe Felsenstein

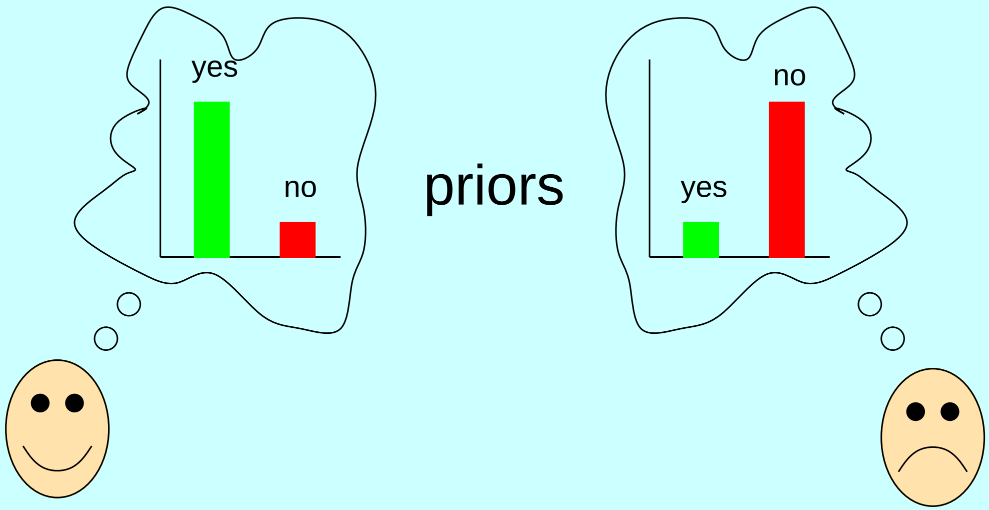
Depts. of Genome Sciences and of Biology, University of Washington

Odds ratio justification for maximum likelihood

D	**the data**
H_1	**Hypothesis 1**
H_2	**Hypothesis 2**
$\mid$	**the symbol for “given”**

$$\underbrace{\frac{\mathrm{Prob}\,(\mathsf{H}_1)}{\mathrm{Prob}\,(\mathsf{H}_2)}}_{\text{Prior odds ratio}} \quad \underbrace{\frac{\mathrm{Prob}\,(\mathsf{D}\mid\mathsf{H}_1)}{\mathrm{Prob}\,(\mathsf{D}\mid\mathsf{H}_2)}}_{\text{Likelihood ratio}} \quad = \quad \underbrace{\frac{\mathrm{Prob}\,(\mathsf{H}_1\mid\mathsf{D})}{\mathrm{Prob}\,(\mathsf{H}_2\mid\mathsf{D})}}_{\text{Posterior odds ratio}}$$

If a space probe finds no Little Green Men on Mars



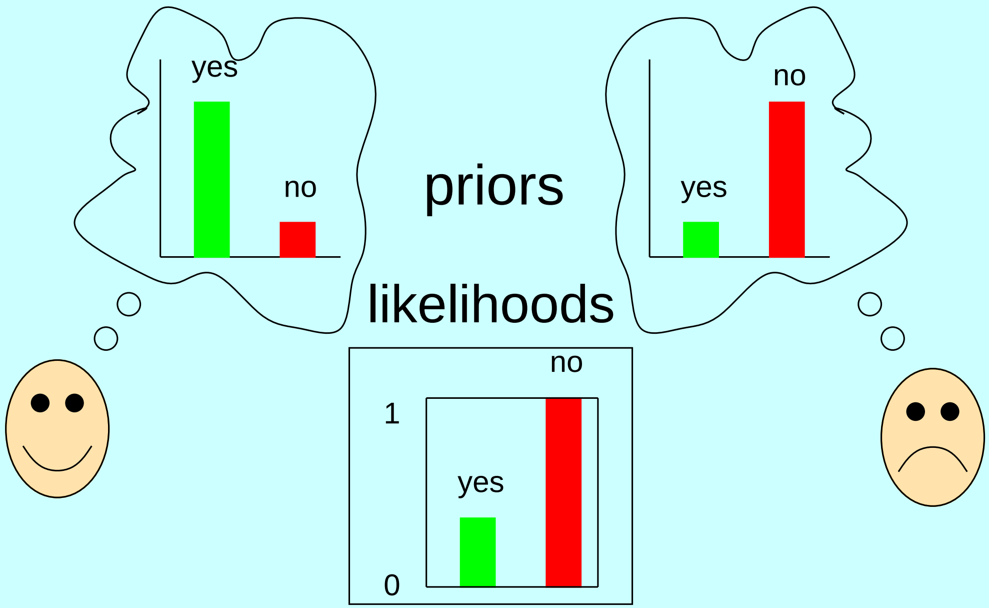
an optimist

a pessimist

$$\frac{4}{1}$$

$$\frac{1}{4}$$

If a space probe finds no Little Green Men on Mars



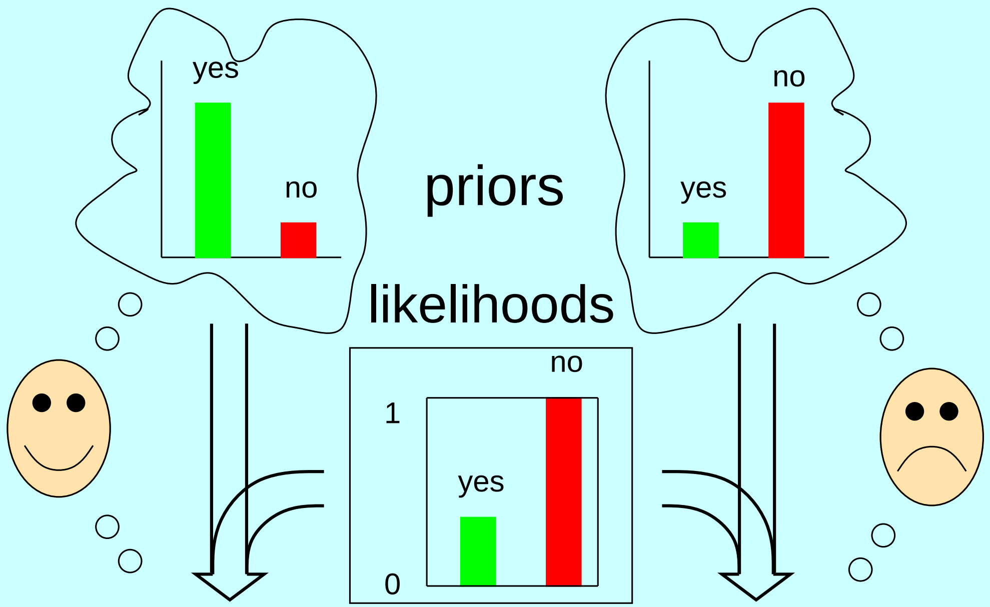
an optimist

a pessimist

$$\frac{4}{1}$$

$$\frac{1}{4}$$

If a space probe finds no Little Green Men on Mars



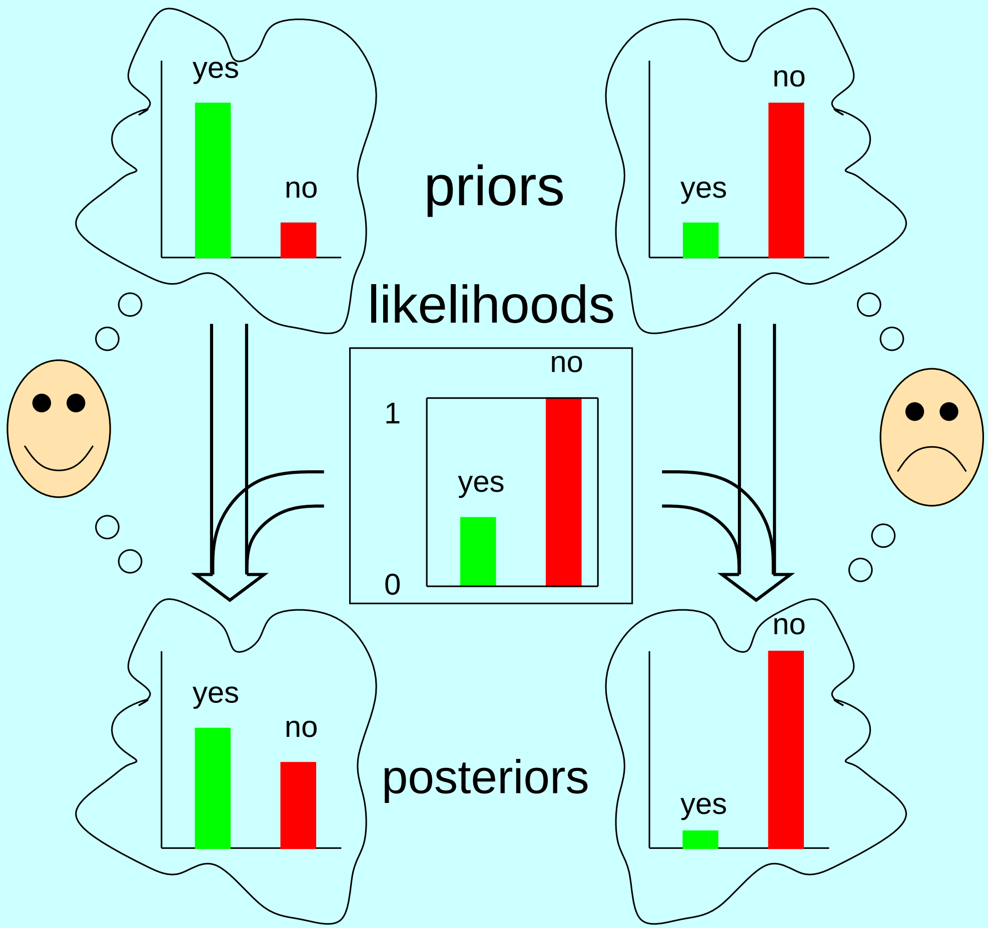
an optimist

a pessimist

$$\frac{4}{1} \times \frac{1/3}{1}$$

$$\frac{1}{4} \times \frac{1/3}{1}$$

If a space probe finds no Little Green Men on Mars



an optimist

a pessimist

$$\frac{4}{1} \times \frac{1/3}{1} = \frac{4}{3}$$

$$\frac{1}{4} \times \frac{1/3}{1} = \frac{1}{12}$$

The likelihood ratio term ultimately dominates

If we see one Little Green Man, the likelihood calculation does the right thing:

$$\frac{1}{4} \times \frac{2/3}{0} = \frac{\infty}{1}$$

(put this way, this is OK but not mathematically kosher)

If we send n space probes and keep seeing none, the likelihood ratio term is

$$\left(\frac{1}{3}\right)^n$$

It dominates the calculation, overwhelming the prior.

Thus even if we don't have a prior we can believe in, we may be interested in knowing which hypothesis the likelihood ratio is recommending ...

Likelihood in Simple Coin-Tossing

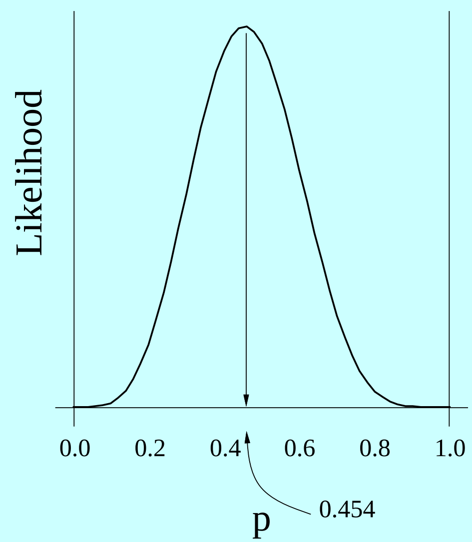
Tossing a coin n times, with probability p of heads, the probability of outcome HHTHTTTTHTTH is

$$pp(1-p)p(1-p)(1-p)(1-p)(1-p)p(1-p)(1-p)p$$

which is

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

Plotting L against p to find its maximum:



Differentiating to find the maximum:

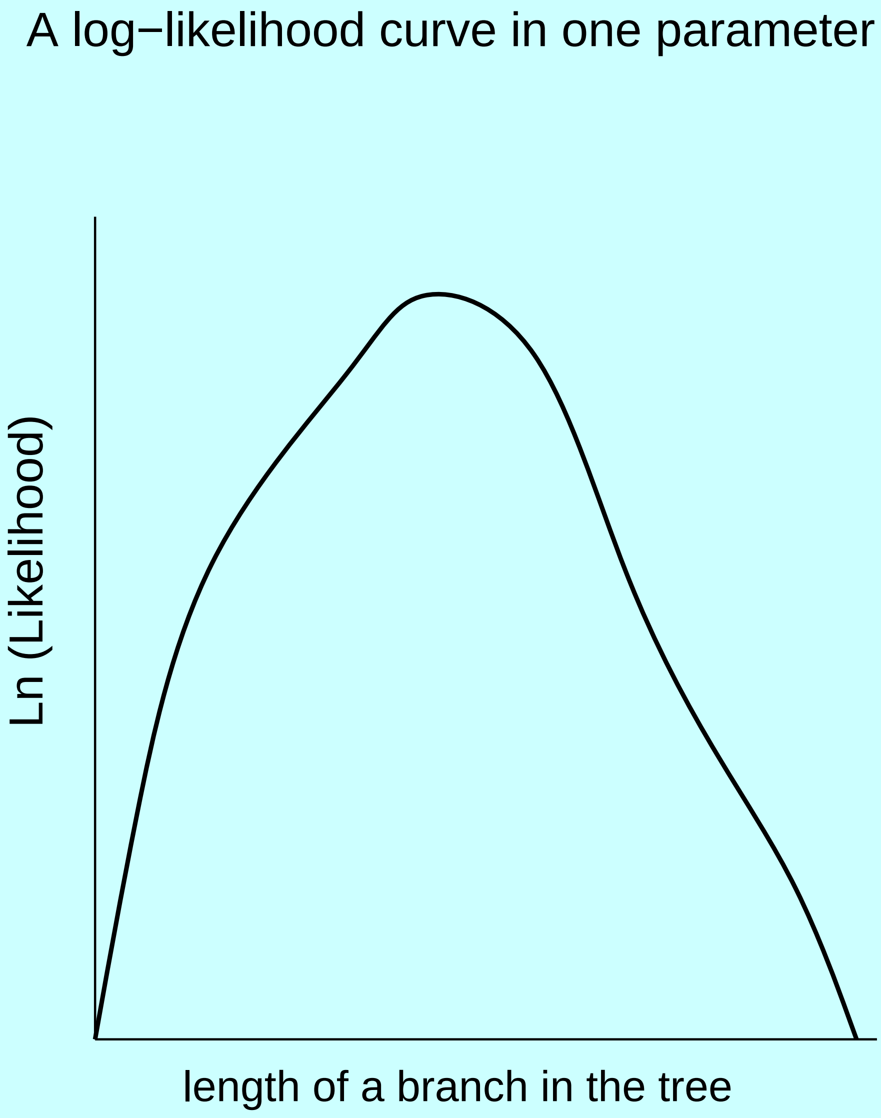
Differentiating the expression for L with respect to p and equating the derivative to 0, the value of p that is at the peak is found (not surprisingly) to be $p = 5/11$:

$$\frac{\partial L}{\partial p} = \left(\frac{5}{p} - \frac{6}{1-p}\right) p^5 (1-p)^6 = 0$$

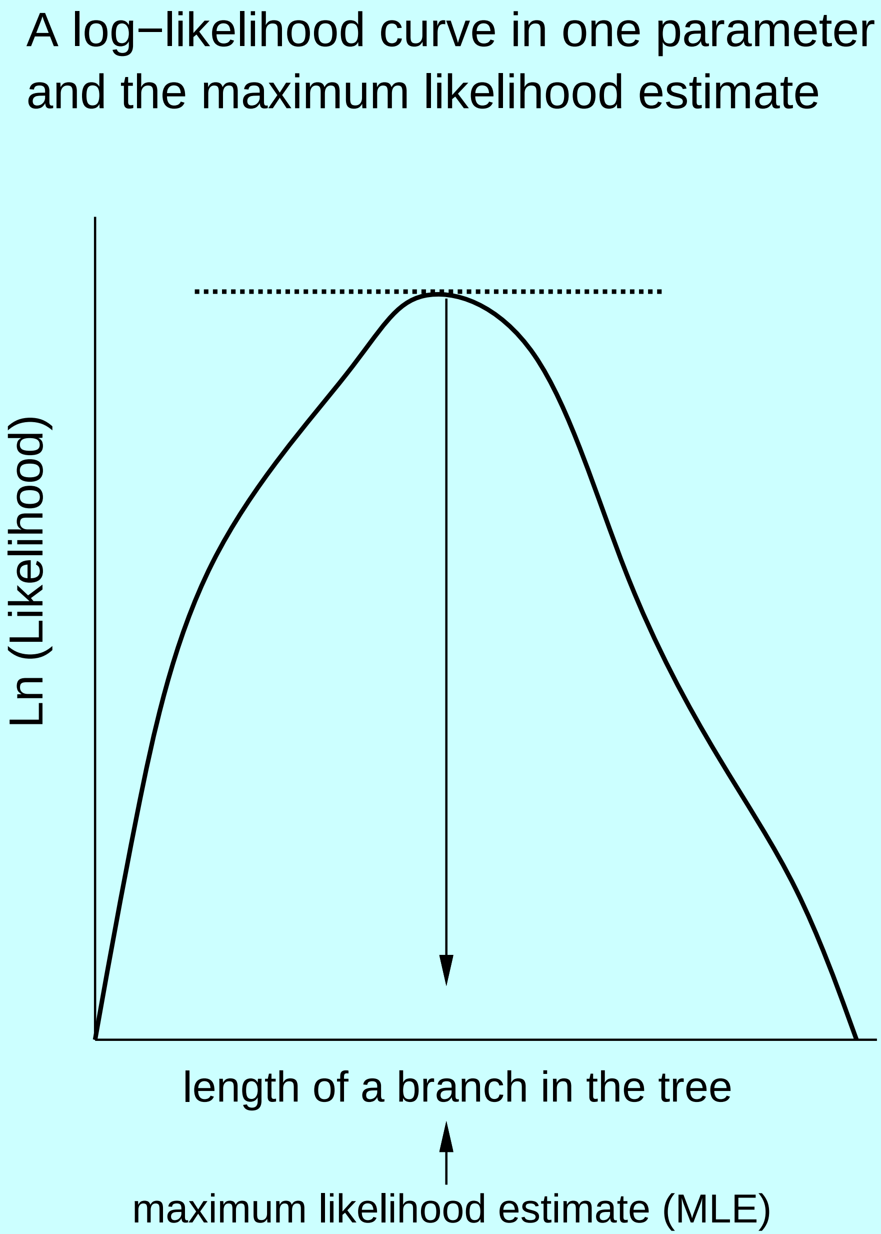
$$5 - 11\,p = 0$$

$$\hat{p} = \frac{5}{11}$$

A log-likelihood curve

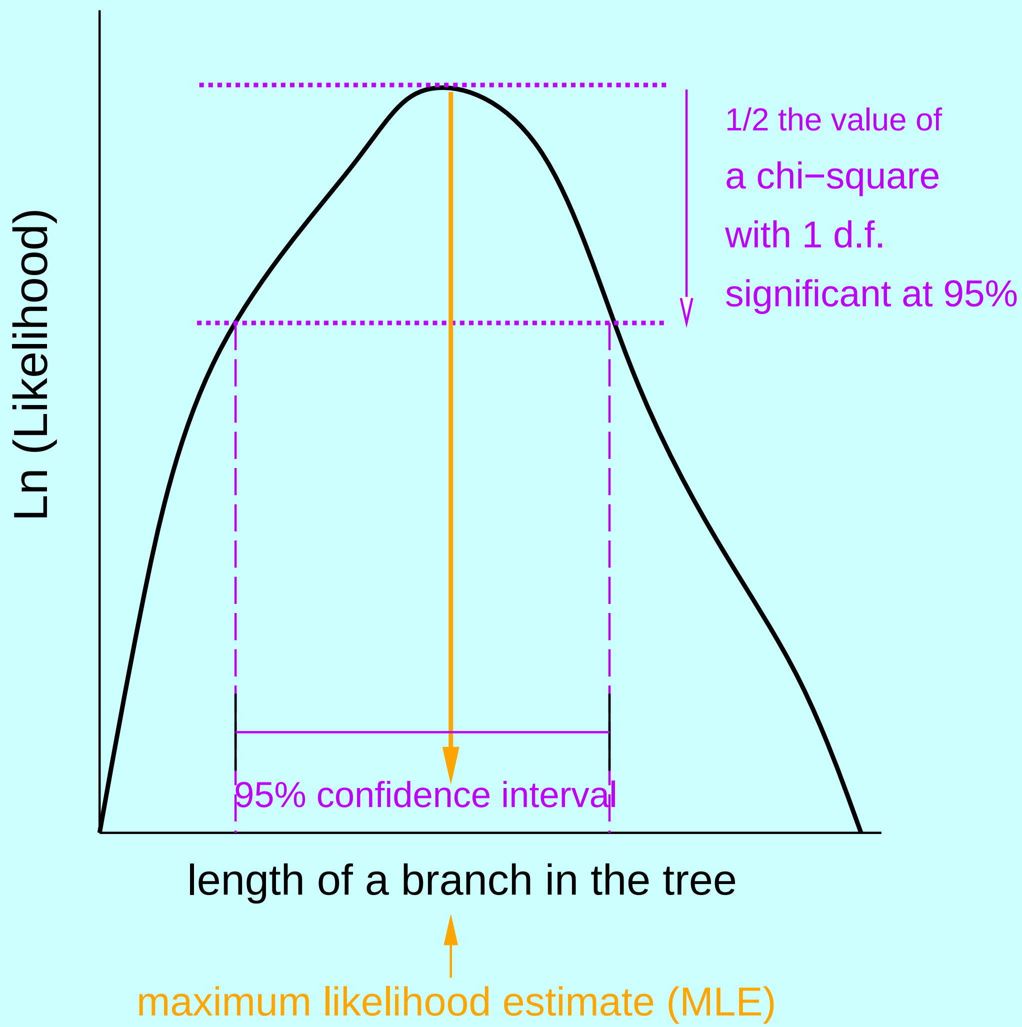


Its maximum likelihood estimate

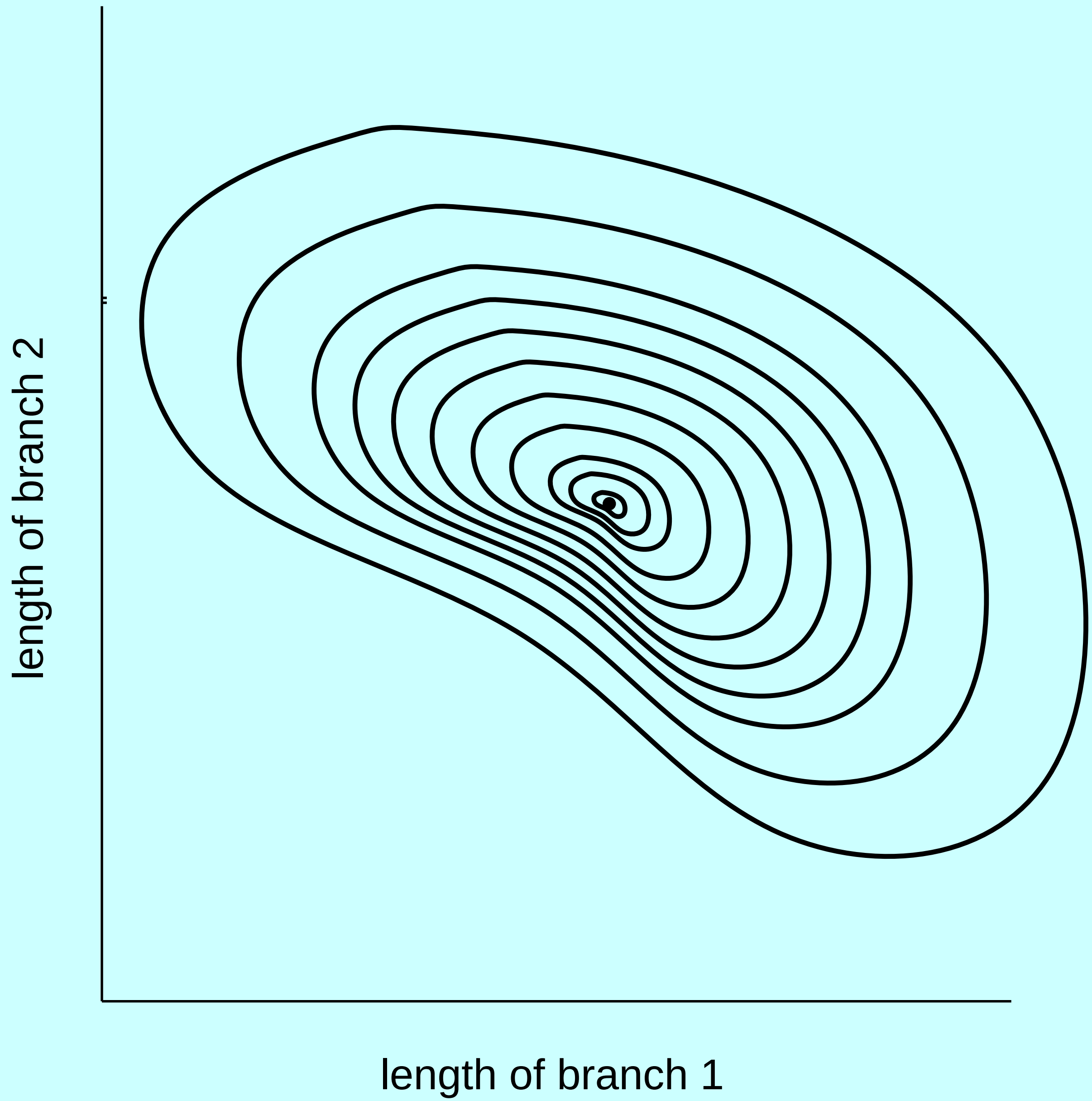


The (approximate, asymptotic) confidence interval

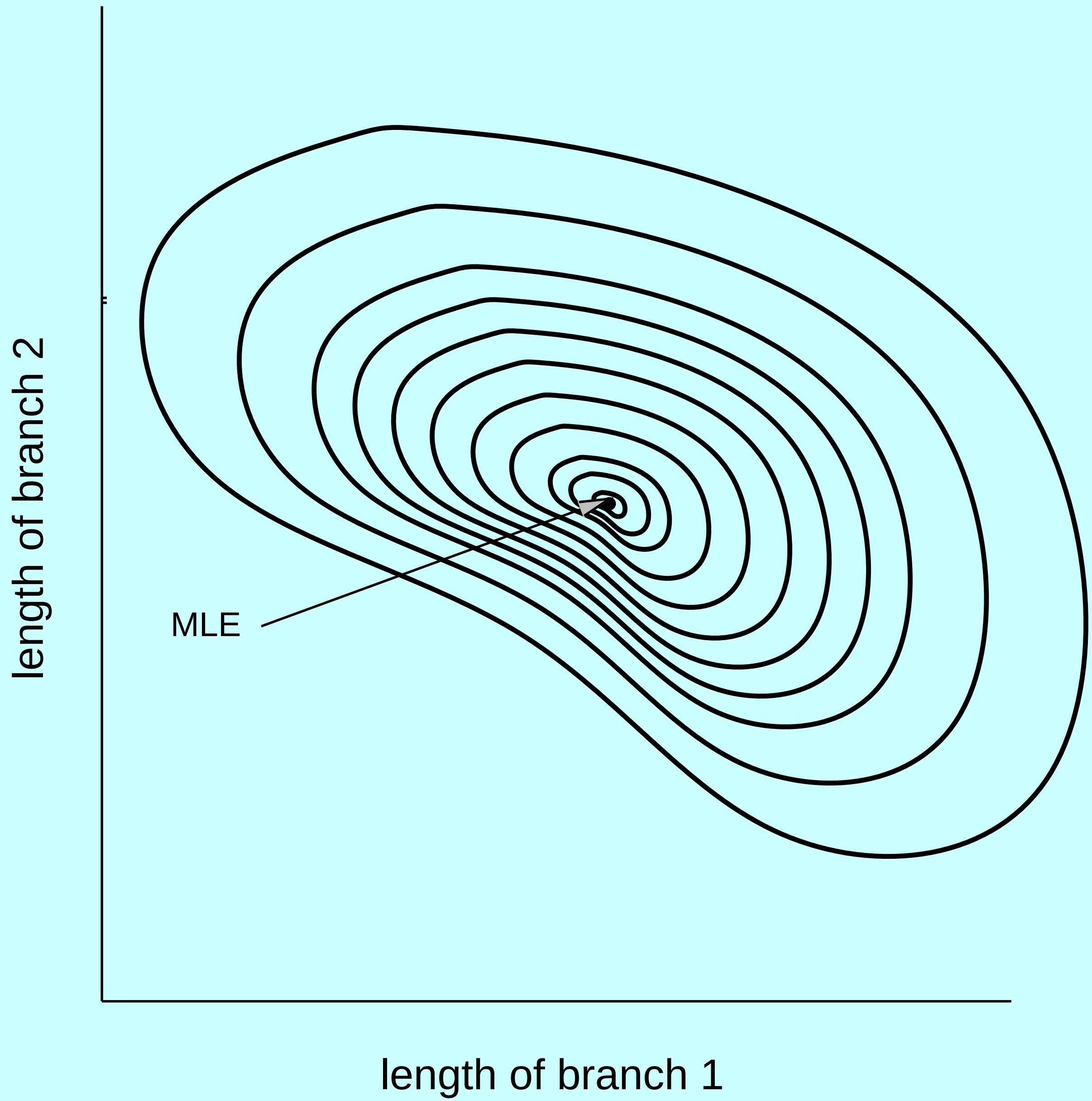
A log−likelihood curve in one parameter
and the maximum likelihood estimate and
confidence interval derived from it



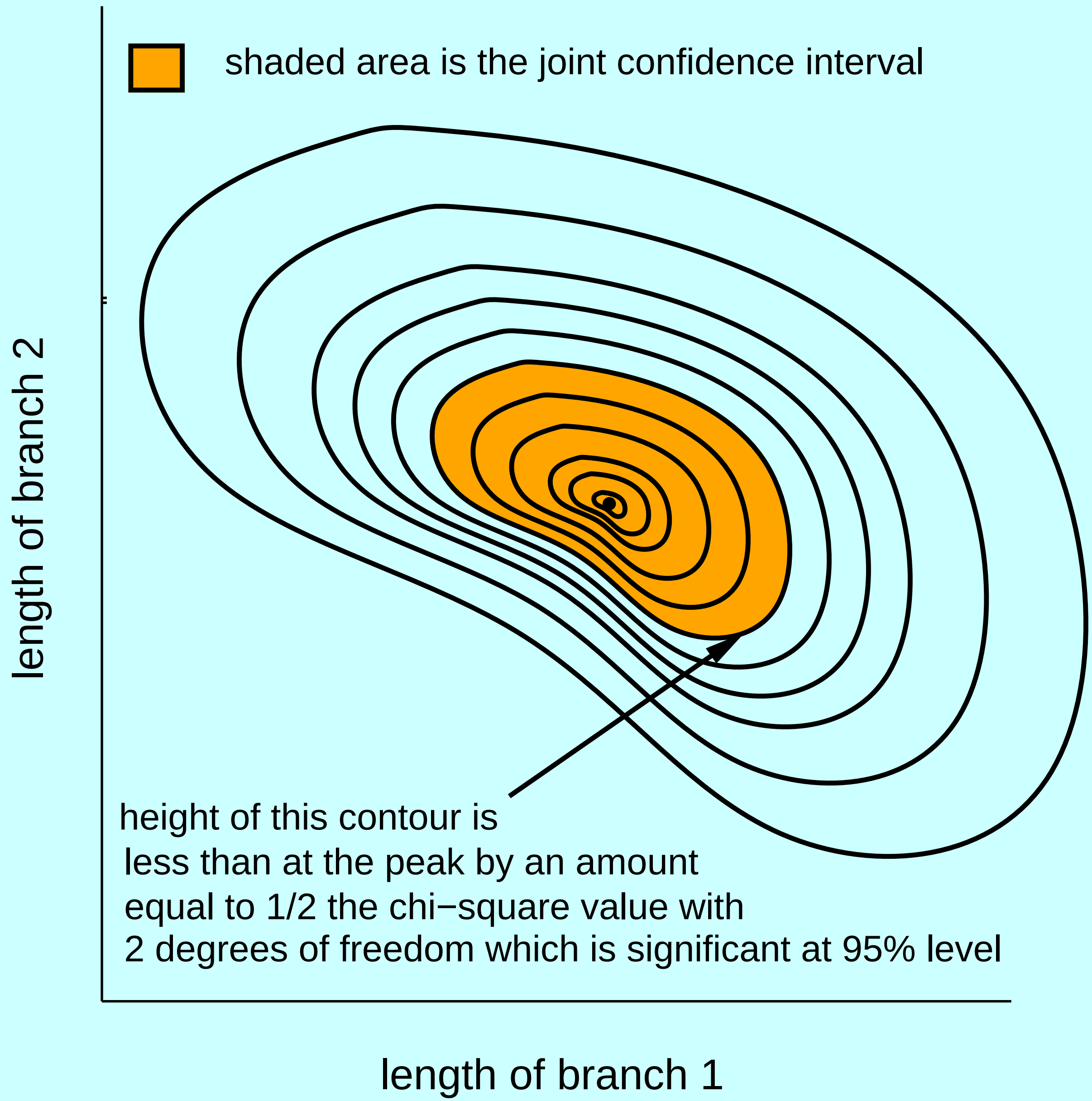
Contours of a log-likelihood surface in two dimensions



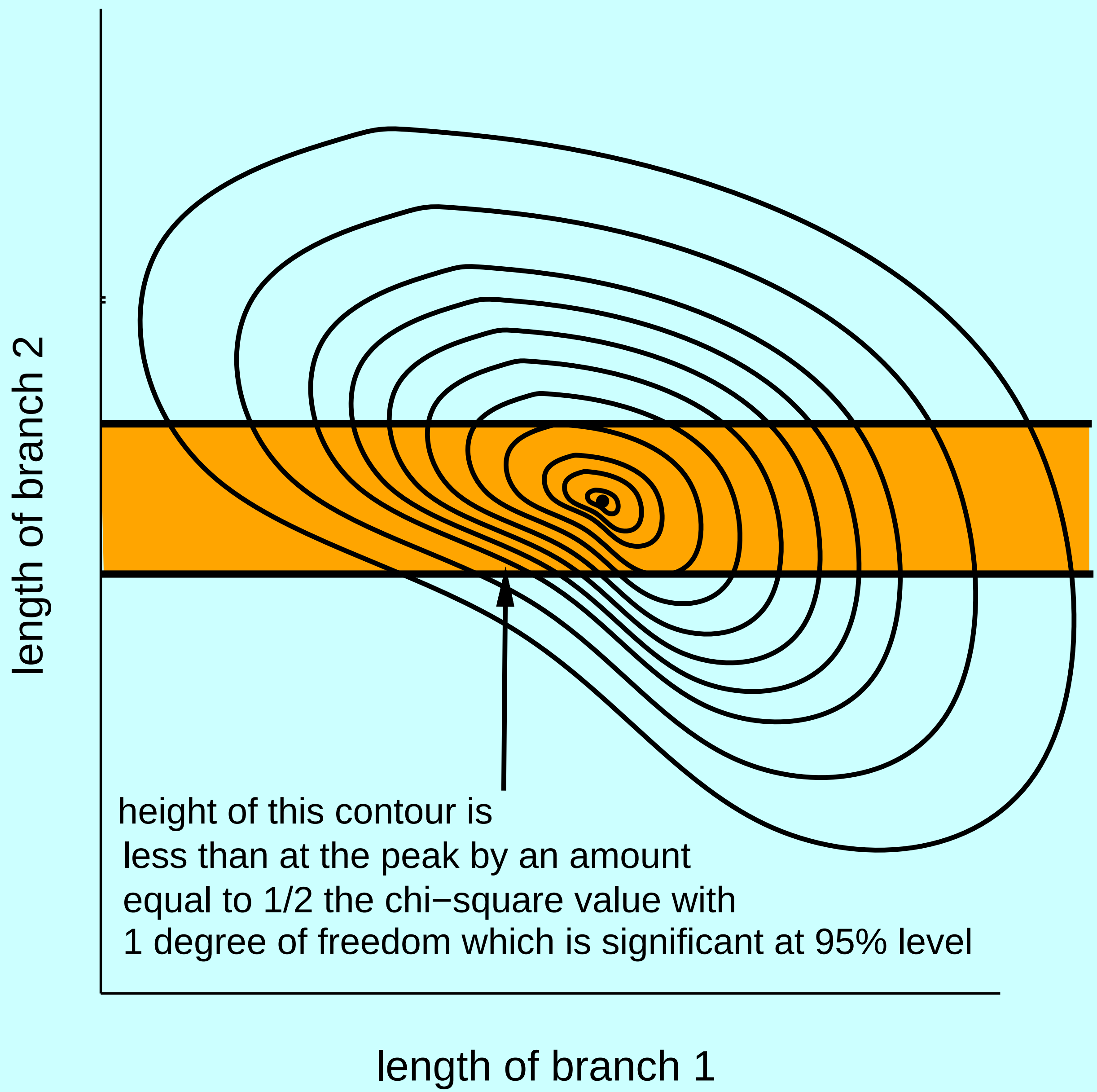
Contours of a log-likelihood surface in two dimensions



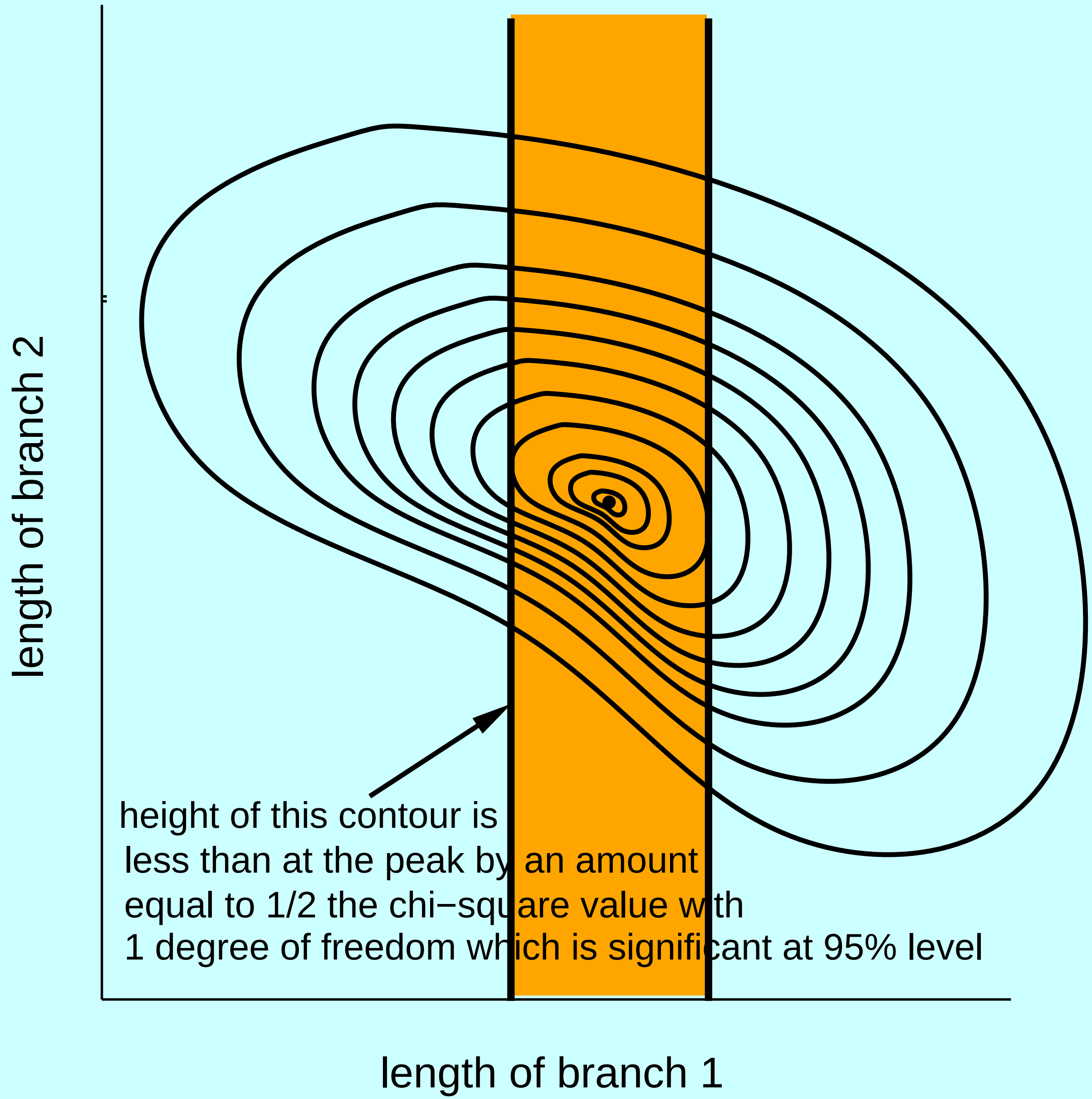
Log-likelihood-based confidence set for two variables



Confidence interval for one variable



Confidence interval for the other variable



Calculating the likelihood of a tree

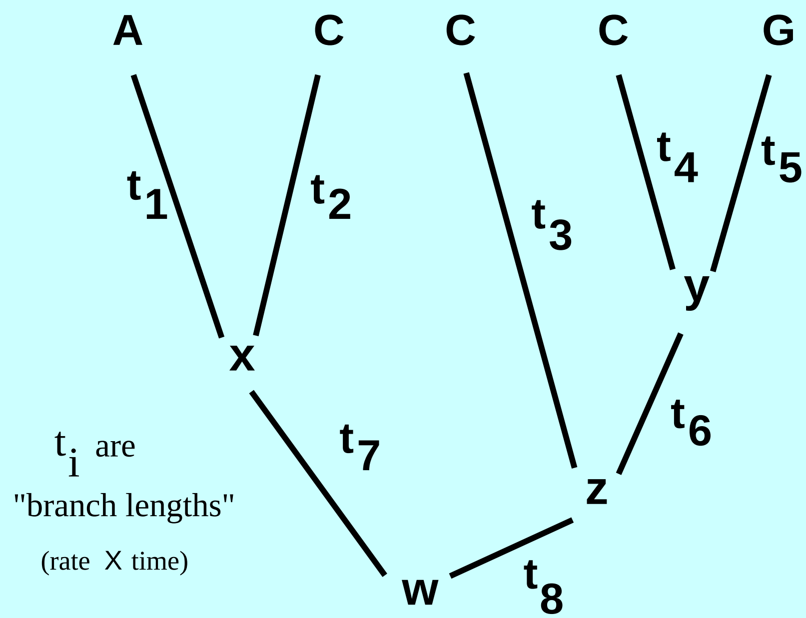
If we have molecular sequences on a tree, the likelihood is the product over sites of the data $D^{[i]}$ for each site (if those evolve independently):

$$L \;=\; \mathrm{Prob}\,(D \mid T) \;=\; \prod_{i=1}^{\text{sites}} \mathrm{Prob}\,(D^{[i]} \mid T)$$

With log-likelihoods, the product becomes a sum:

$$\ln L \;=\; \ln\;\mathrm{Prob}\,(D \mid T) \;=\; \sum_{i=1}^{\text{sites}} \ln\;\mathrm{Prob}\,(D^{[i]} \mid T)$$

Calculating the likelihood for site i on a tree



Sum over all possible states (bases) at interior nodes:

$$
\begin{aligned}
\mathsf{L}^{(\mathsf{i})} \quad = \quad & \sum_{\mathsf{x}} \sum_{\mathsf{y}} \sum_{\mathsf{z}} \sum_{\mathsf{w}} \mathrm{Prob}\,(\mathsf{w}) \;\; \mathrm{Prob}\,(\mathsf{x}\,|\,\mathsf{w}, \mathsf{t}_7)\; \mathrm{Prob}\,(\mathsf{A}\,|\,\mathsf{x}, \mathsf{t}_1)\; \mathrm{Prob}\,(\mathsf{C}\,|\,\mathsf{x}, \mathsf{t}_2) \\
& \times \mathrm{Prob}\,(\mathsf{z}\,|\,\mathsf{w}, \mathsf{t}_8)\; \mathrm{Prob}\,(\mathsf{C}\,|\,\mathsf{z}, \mathsf{t}_3) \\
& \times \mathrm{Prob}\,(\mathsf{y}\,|\,\mathsf{z}, \mathsf{t}_6)\; \mathrm{Prob}\,(\mathsf{C}\,|\,\mathsf{y}, \mathsf{t}_4)\; \mathrm{Prob}\,(\mathsf{G}\,|\,\mathsf{y}, \mathsf{t}_5)
\end{aligned}
$$

Calculating the likelihood for site i on a tree

We use the conditional likelihoods: $\mathsf{L}_j^{(i)}(\mathsf{s})$

These compute the probability of everything at site i at or above node j on the tree, given that node j is in state s. Thus it assumes something (s) that we don't know in practice – so we compute these for all states s.

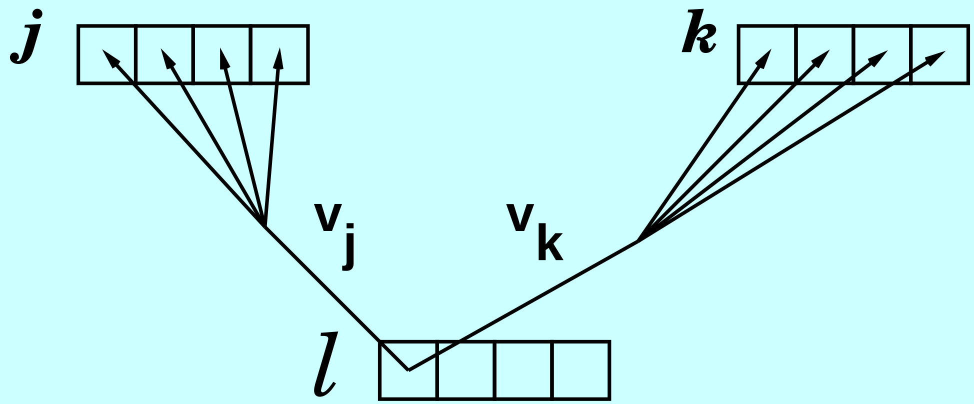
At the tips we can define these quantities: if the observed state is (say) C, the vector of L's is

$$(0, 1, 0, 0)$$

.
If we observe an ambiguity, say R (purine), they are

$$(1, 0, 1, 0), \qquad \textit{not} \quad (1/2, 0, 1/2, 0)$$

The “pruning" algorithm:



$$
\begin{aligned}
L_\ell^{(i)}(s) \quad &= \quad \left[\sum_{s_j} \mathrm{Prob}\left(s_j \mid s, v_j\right) L_j^{(i)}(s_j)\right] \\
&\quad \times \left[\sum_{s_k} \mathrm{Prob}\left(s_k \mid s, v_k\right) L_k^{(i)}(s_k)\right]
\end{aligned}
$$

(Felsenstein, 1973; 1981).

and at the bottom of the tree:

$$L_0^{(i)} = \sum_s \pi_s \, L_0^{(i)}(s)$$

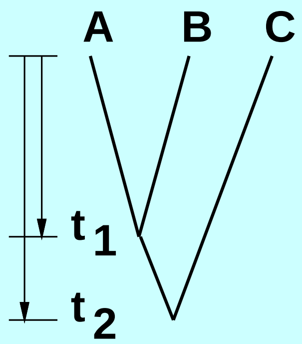
(Felsenstein, 1973, 1981)

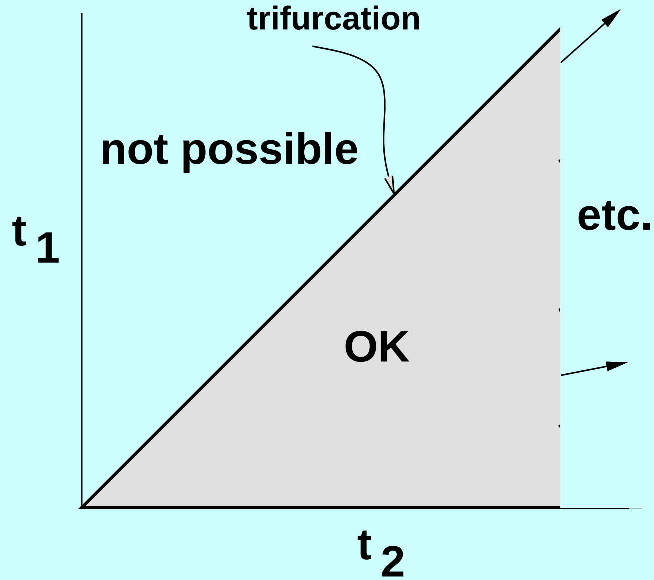
and having gotten the likelihoods for each site:

$$L = \prod_{i=1}^{\text{sites}} L_0^{(i)}$$

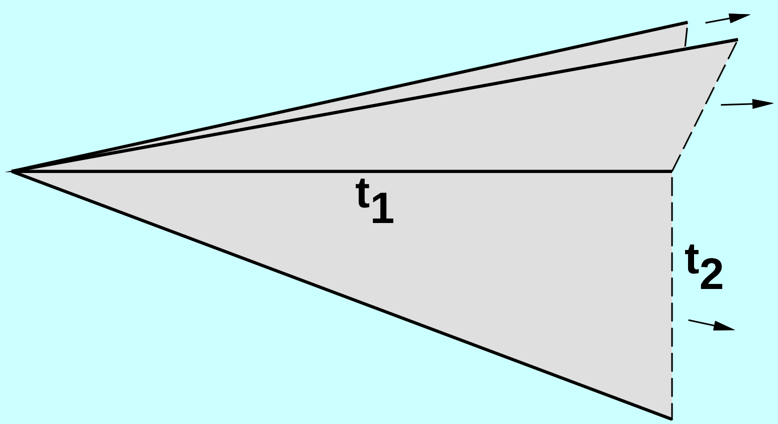
What does “tree space" (with branch lengths) look like?

an example: three species with a clock



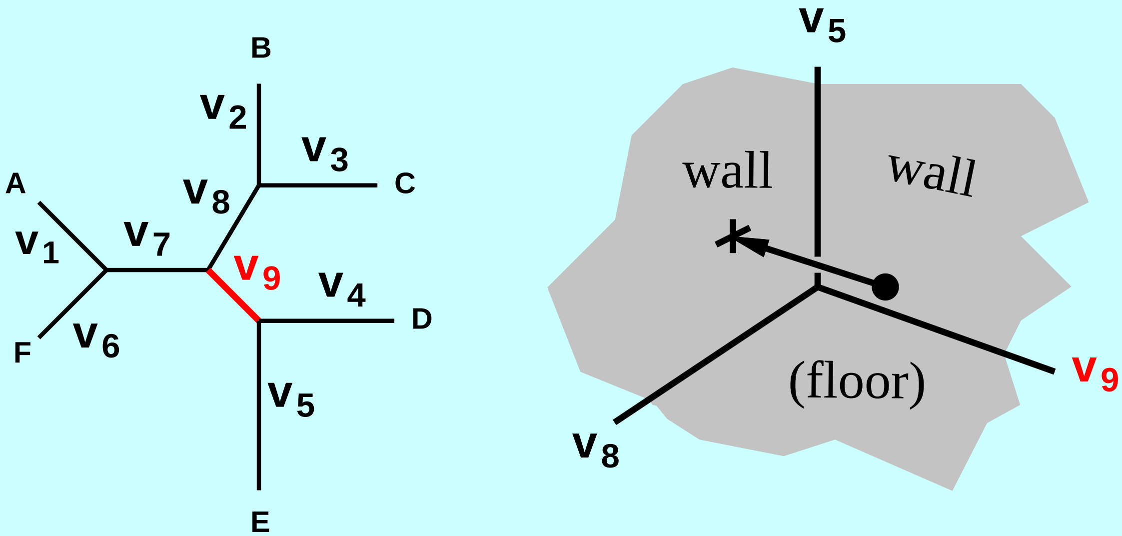


when we consider all three possible topologies, the space looks like:



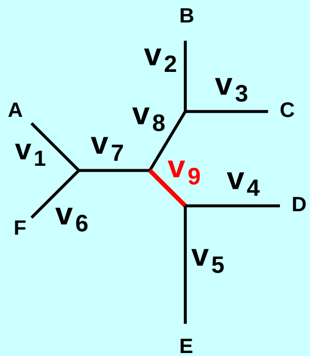
For one tree topology

The space of trees varying all $2n - 3$ branch lengths, each a nonegative number, defines an “orthant" (open corner) of a $(2n - 3)$-dimensional real space:



Through the looking-glass

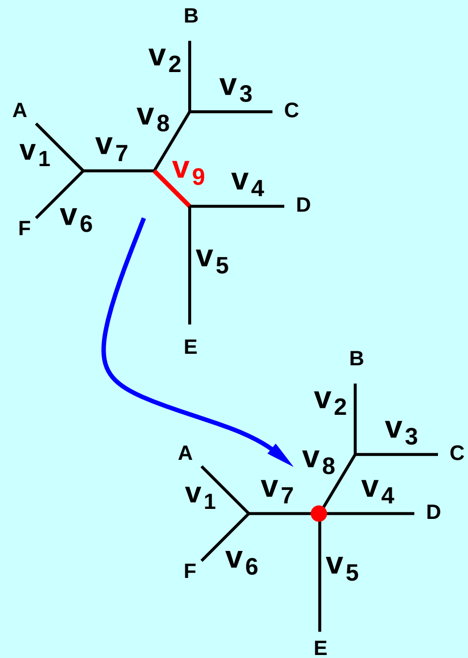
Shrinking one of the $n-1$ interior branches to 0, we arrive at a trifurcation:



Here, as we pass “through the looking glass" we are also touch the space for two other tree topologies, and we could enter either.

Through the looking-glass

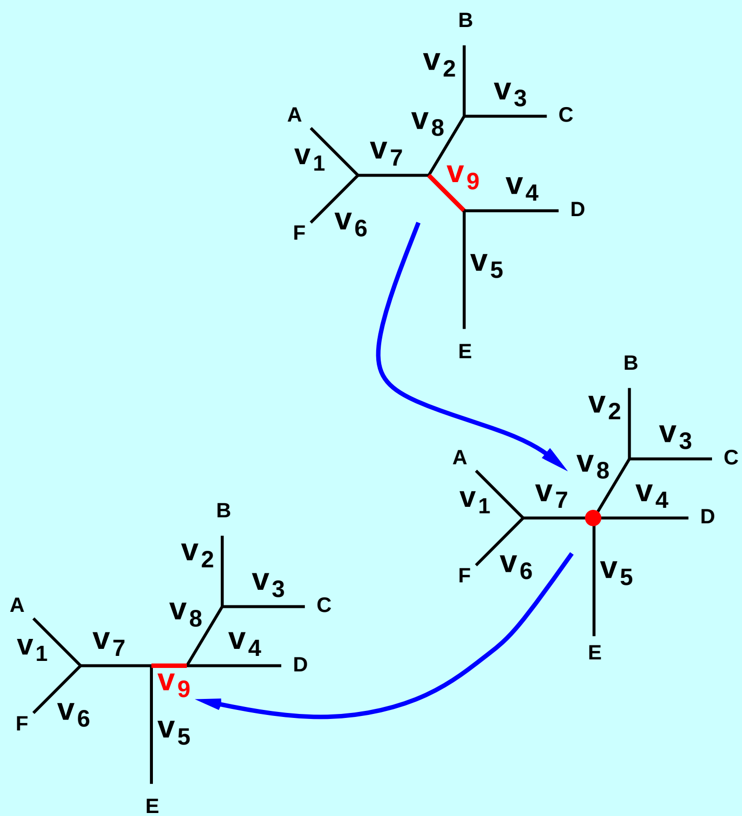
Shrinking one of the $n - 1$ interior branches to 0, we arrive at a trifurcation:



Here, as we pass “through the looking glass" we are also touch the space for two other tree topologies, and we could enter either.

Through the looking-glass

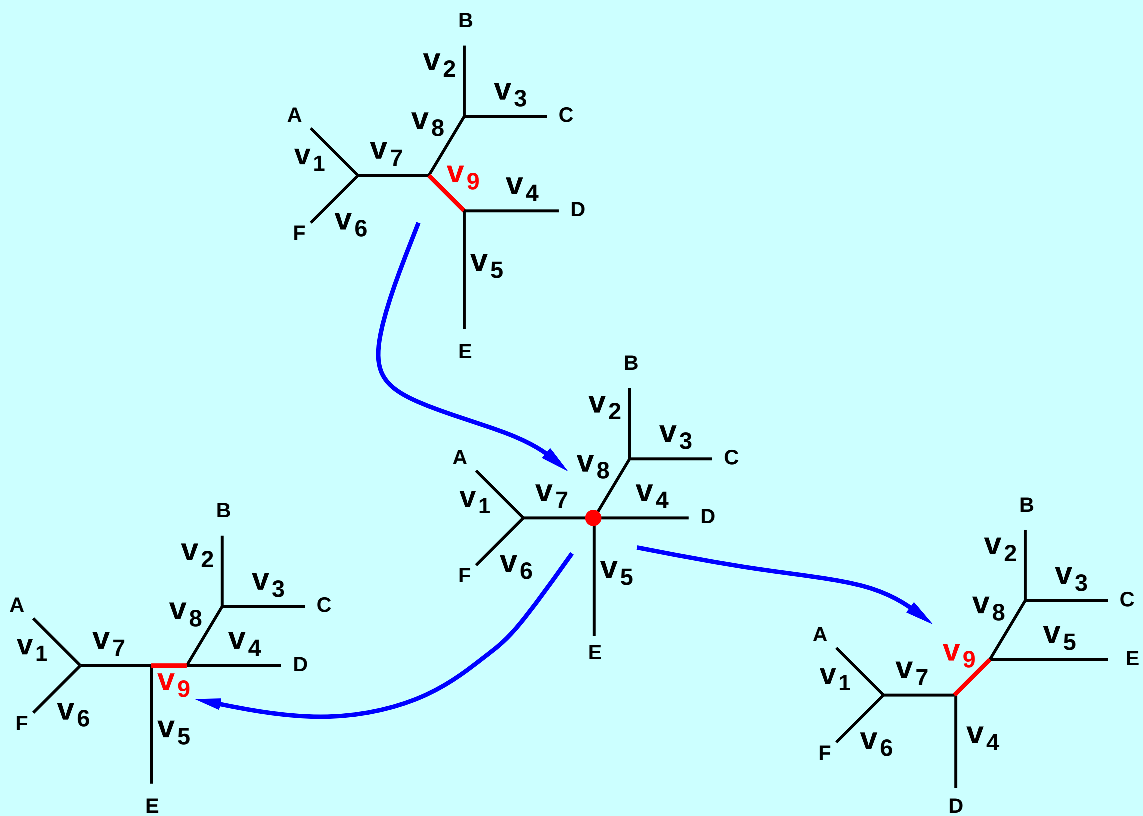
Shrinking one of the $n - 1$ interior branches to 0, we arrive at a trifurcation:



Here, as we pass “through the looking glass" we are also touch the space for two other tree topologies, and we could enter either.

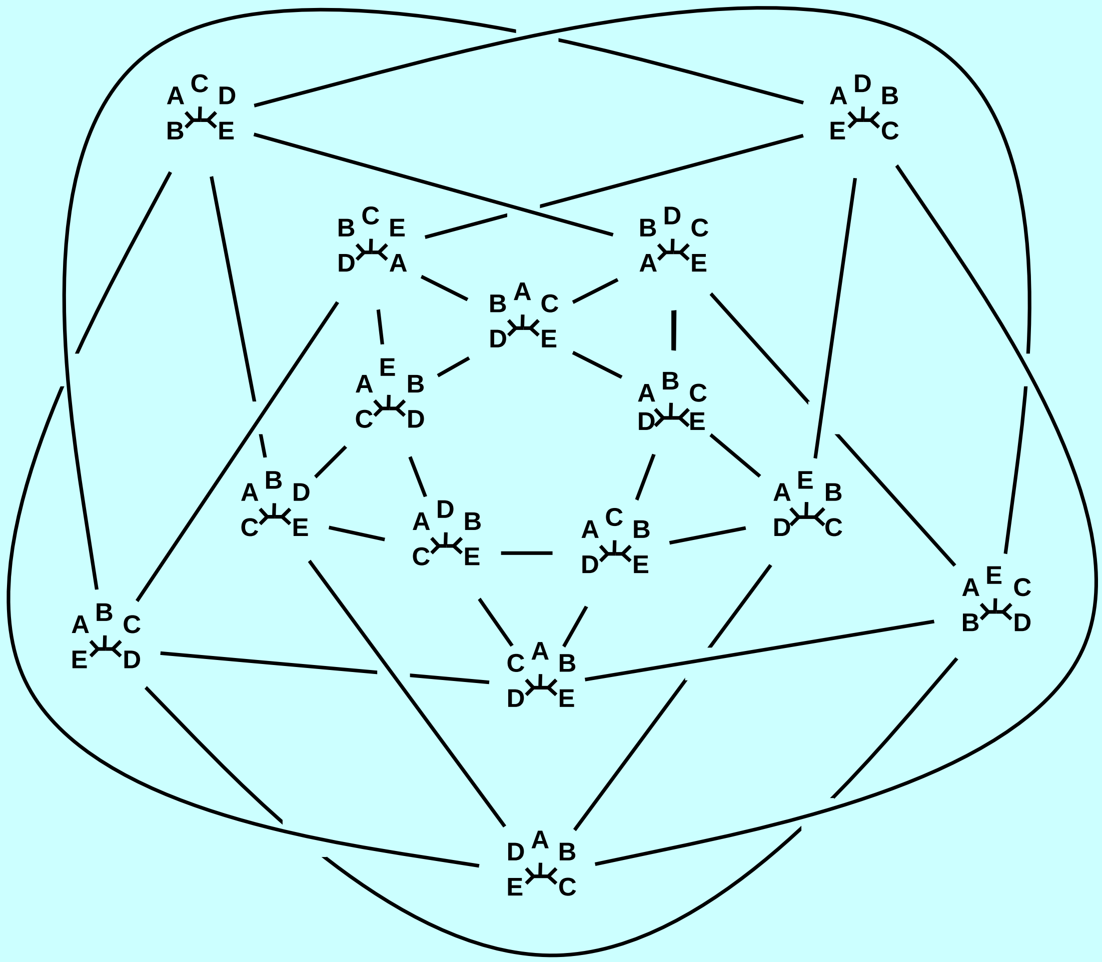
Through the looking-glass

Shrinking one of the $n-1$ interior branches to 0, we arrive at a trifurcation:



Here, as we pass "through the looking glass" we are also touch the space for two other tree topologies, and we could enter either.

The graph of all trees of 5 species



The Schoenberg graph (all 15 trees of size 5 connected by NNI's)

A data example: mitochondrial D-loop sequences

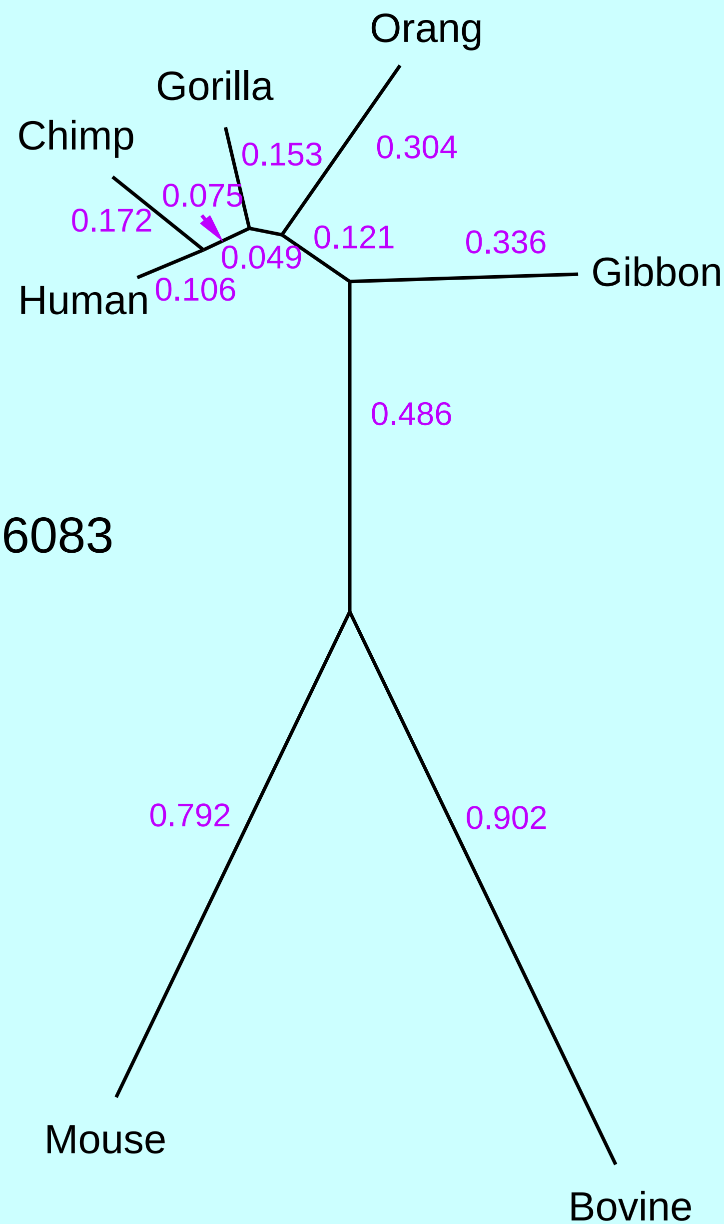
```
Bovine     CCAAACCTGT CCCCACCATC TAACACCAAC CCACATATAC AAGCTAAACC AAAAATACCA
Mouse      CCAAAAAAAC ATCCAAACAC CAACCCCAGC CCTTACGCAA TAGCCATACA AAGAATATTA
Gibbon     CTATACCCAC CCAACTCGAC CTACACCAAT CCCCACATAG CACACAGACC AACAACCTCC
Orang      CCCCACCCGT CTACACCAGC CAACACCAAC CCCCACCTAC TATACCAACC AATAACCTCT
Gorilla    CCCCATTTAT CCATAAAAAC CAACACCAAC CCCCATCTAA CACACAAACT AATGACCCCC
Chimp      CCCCATCCAC CCATACAAAC CAACATTACC CTCCATCCAA TATACAAACT AACAACCTCC
Human      CCCCACTCAC CCATACAAAC CAACACCACT CTCCACCTAA TATACAAATT AATAACCTCC

           TACTACTAAA AACTCAAATT AACTCTTTAA TCTTTATACA ACATTCCACC AACCTATCCA
           TACAACCATA AATAAGACTA ATCTATTAAA ATAACCCATT ACGATACAAA ATCCCTTTCG
           CACCTTCCAT ACCAAGCCCC GACTTTACCG CCAACGCACC TCATCAAAAC ATACCTACAA
           CAACCCCTAA ACCAAACACT ATCCCCAAAA CCAACACACT CTACCAAAAT ACACCCCCAA
           CACCCTCAAA GCCAAACACC AACCCTATAA TCAATACGCC TTATCAAAAC ACACCCCCAA
           CACTCTTCAG ACCGAACACC AATCTCACAA CCAACACGCC CCGTCAAAAC ACCCCTTCAG
           CACCTTCAGA ACTGAACGCC AATCTCATAA CCAACACACC CCATCAAAGC ACCCCTCCAA

           CACAAAAAAA CTCATATTTA TCTAAATACG AACTTCACAC AACCTTAACA CATAAACATA
           TCTAGATACA AACCACAACA CACAATTAAT ACACACCACA ATTACAATAC TAAACTCCCA
           CACAAACAAA TGCCCCCCCA CCCTCCTTCT TCAAGCCCAC TAGACCATCC TACCTTCCTA
           TTCACATCCG CACACCCCCA CCCCCCCTGC CCACGTCCAT CCCATCACCC TCTCCTCCCA
           CATAAACCCA CGCACCCCCA CCCCTTCCGC CCATGCTCAC CACATCATCT CTCCCCTTCA
           CACAAATTCA TACACCCCTA CCTTTCCTAC CCACGTTCAC CACATCATCC CCCCCTCTCA
           CACAAACCCG CACACCTCCA CCCCCCTCGT CTACGCTTAC CACGTCATCC CTCCCTCTCA

           CCCCAGCCCA ACACCCTTCC ACAAATCCTT AATATACGCA CCATAAATAA CA
           TCCCACCAAA TCACCCTCCA TCAAATCCAC AAATTACACA ACCATTAACC CA
           GCACGCCAAG CTCTCTACCA TCAAACGCAC AACTTACACA TACAGAACCA CA
           ACACCCTAAG CCACCTTCCT CAAAATCCAA AACCCACACA ACCGAAACAA CA
           ACACCTCAAT CCACCTCCCC CCAAATACAC AATTCACACA AACAATACCA CA
           ACATCTTGAC TCGCCTCTCT CCAAACACAC AATTCACGCA AACAACGCCA CA
           ACACCTTAAC TCACCTTCTC CCAAACGCAC AATTCGCACA CACAACGCCA CA
```

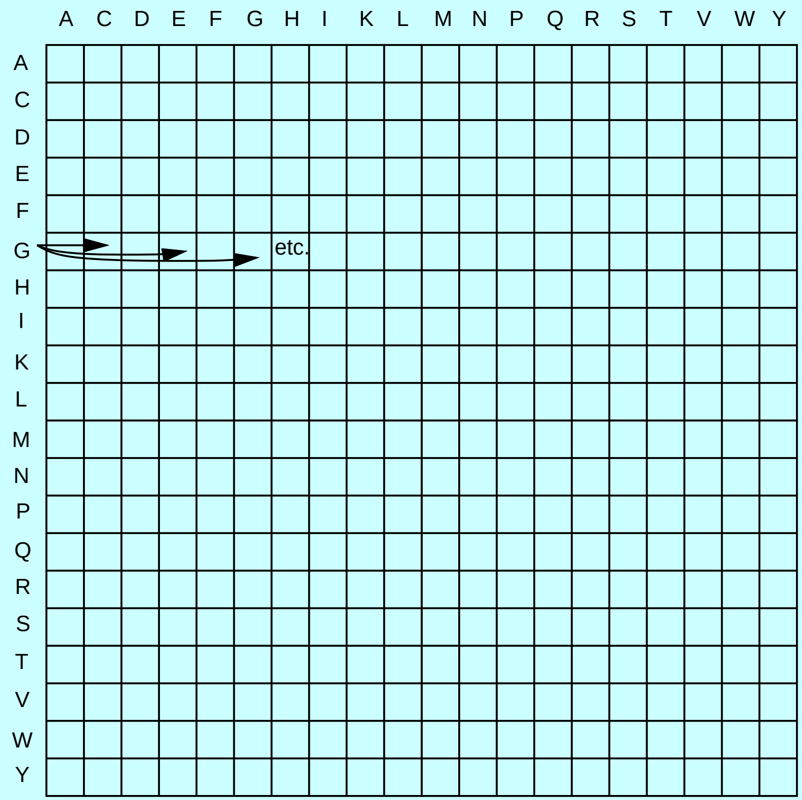
which gives the ML tree



ln L = −1405.6083

Maximum likelihood tree for the Hasegawa 232-site mitochondrial D-loop data set, with Ts/Tn set to 2, analyzed with maximum likelihood (DNAML)

Models with amino acids



Dayhoff PAM model

Jones−Taylor−Thornton model

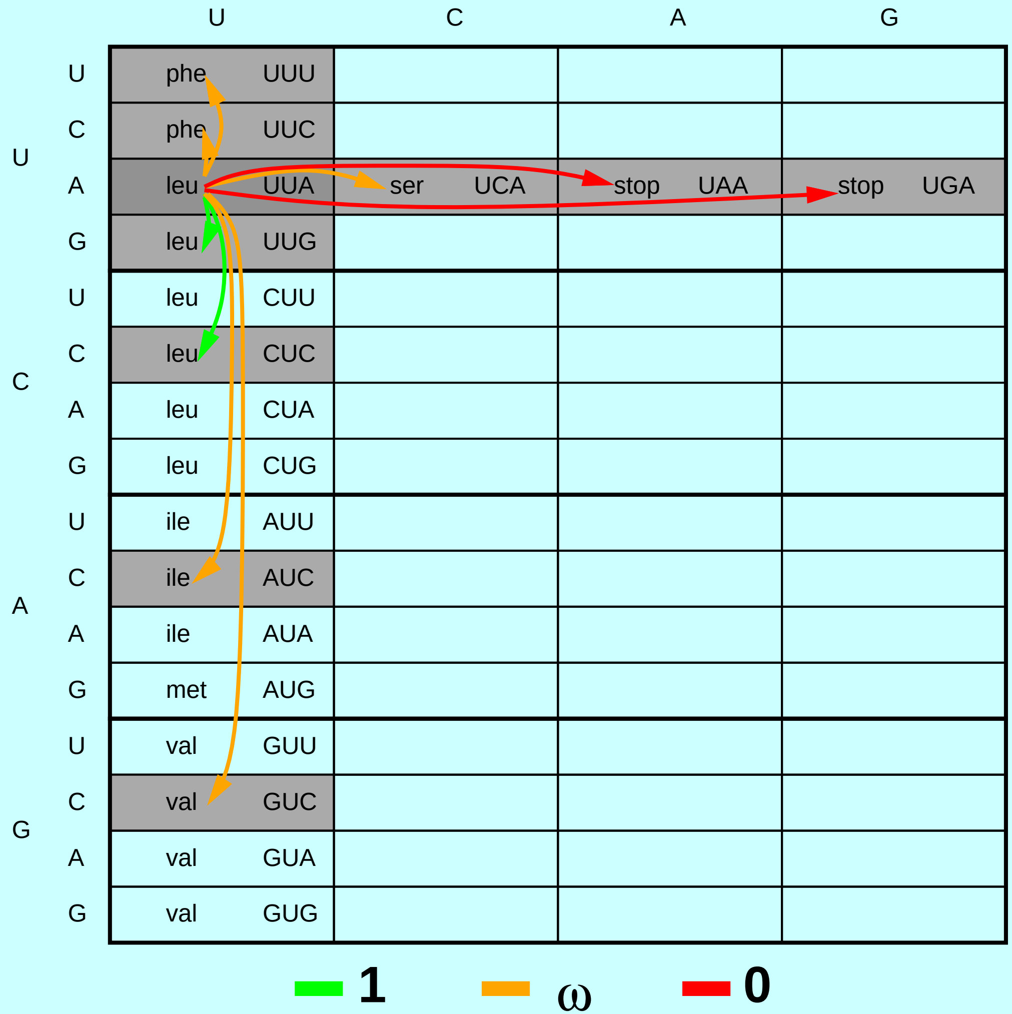
specific models for secondary−structure contexts or membrane proteins

Models adapted from Henikoff BLOSUM scoring

But ... how to take DNA sequence into account? Constraints of code?

Codon models

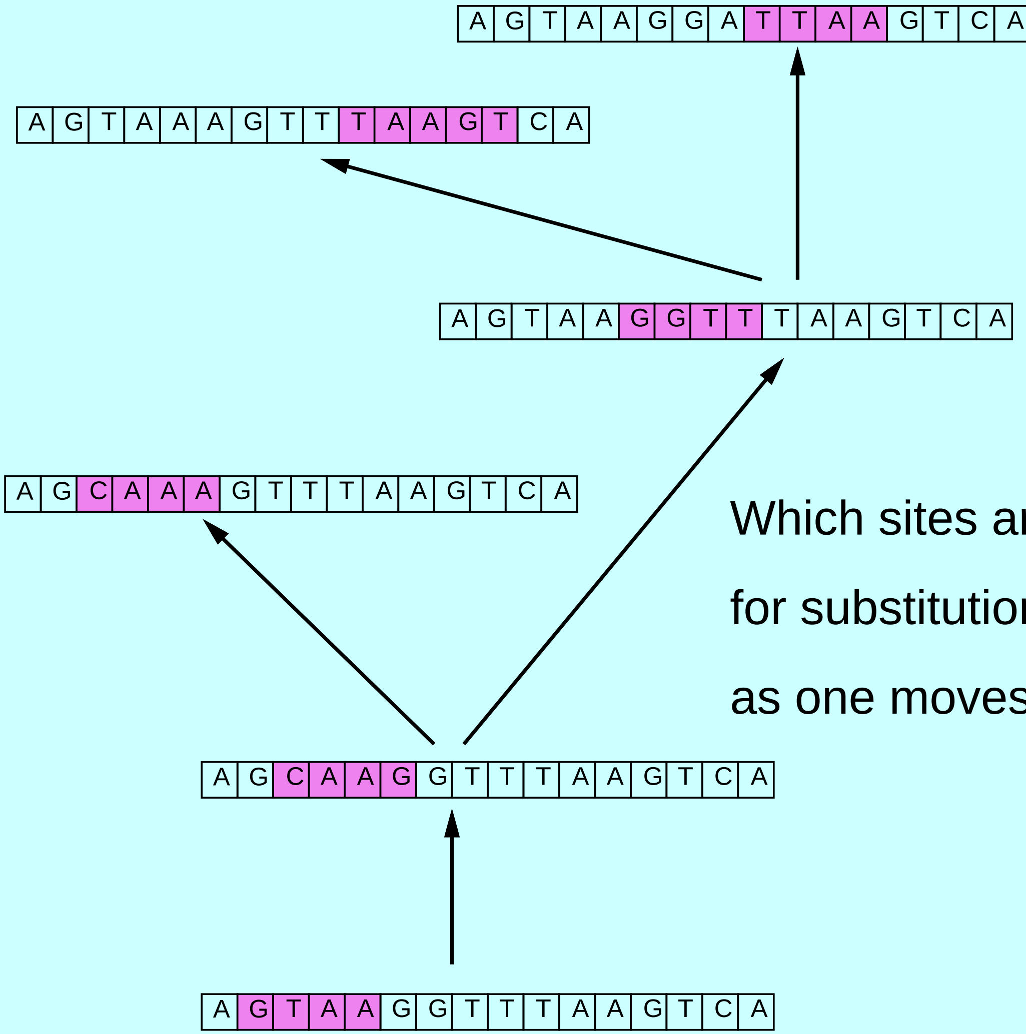
Goldman & Yang, 1994; Muse & Gaut, 1994)



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

Covarion models?

(Fitch and Markowitz, 1970)



Which sites are available

for substitutions changes

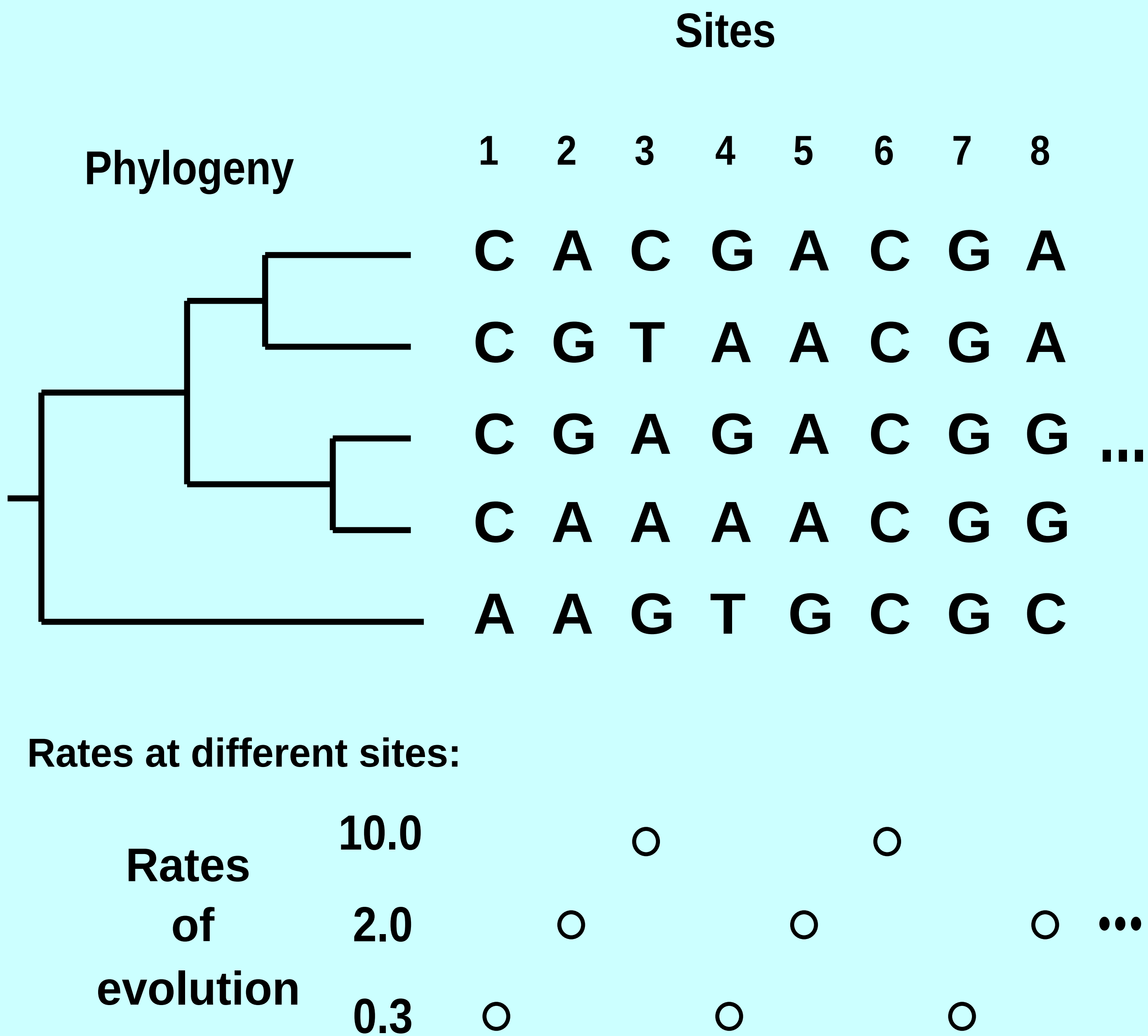
as one moves along the tree

How to calculate likelihood with rate variation

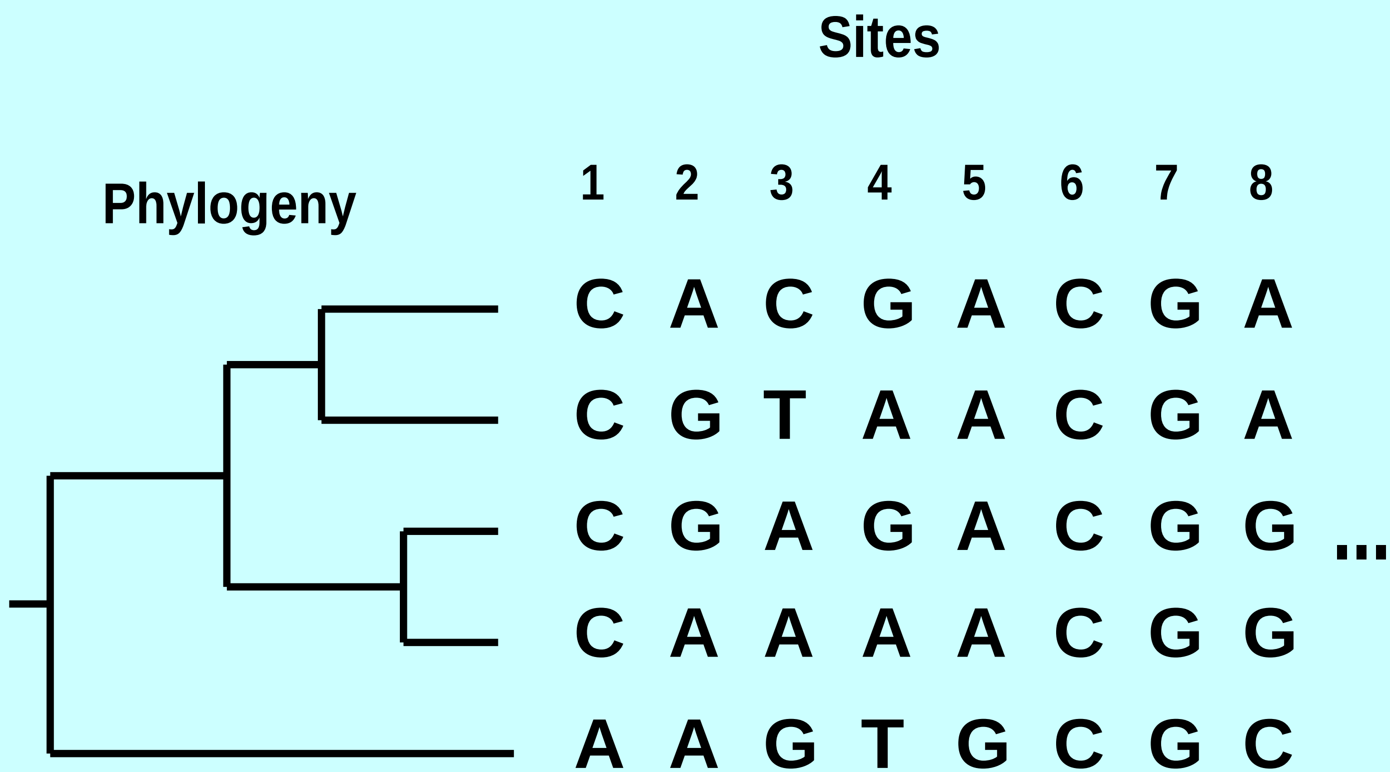
Easy! Since branch lengths always come into transition probability formulas as $r \times t$, can just multiply lengths of branches by the appropriate factor to calculate the likelihood for a site.

(Branch lengths are usually scaled by assuming a rate of 1.)

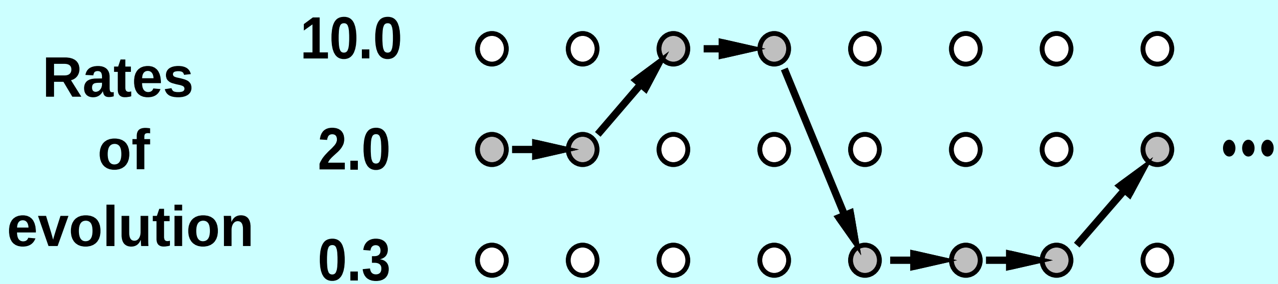
Rate variation among sites



Hidden Markov Model of rate variation among sites



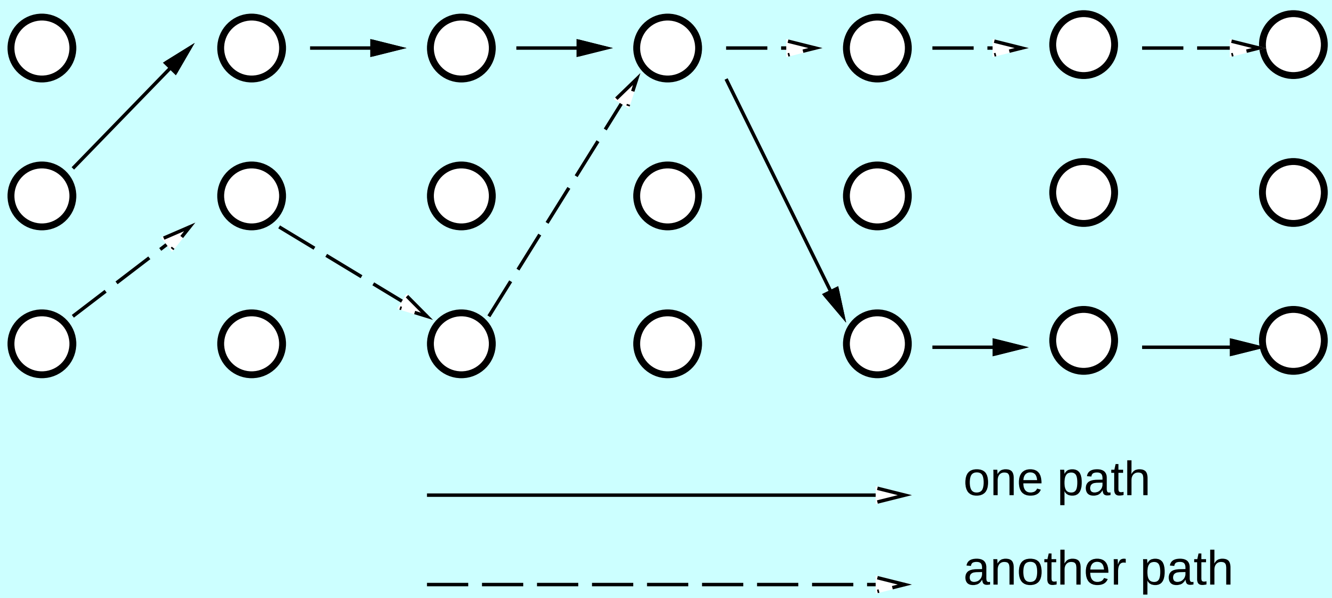
Hidden Markov chain that assigns rates:



Hidden Markov Models sum up over all paths

The Hidden Markov Chain method sums up likelihoods over all possible paths through the states:

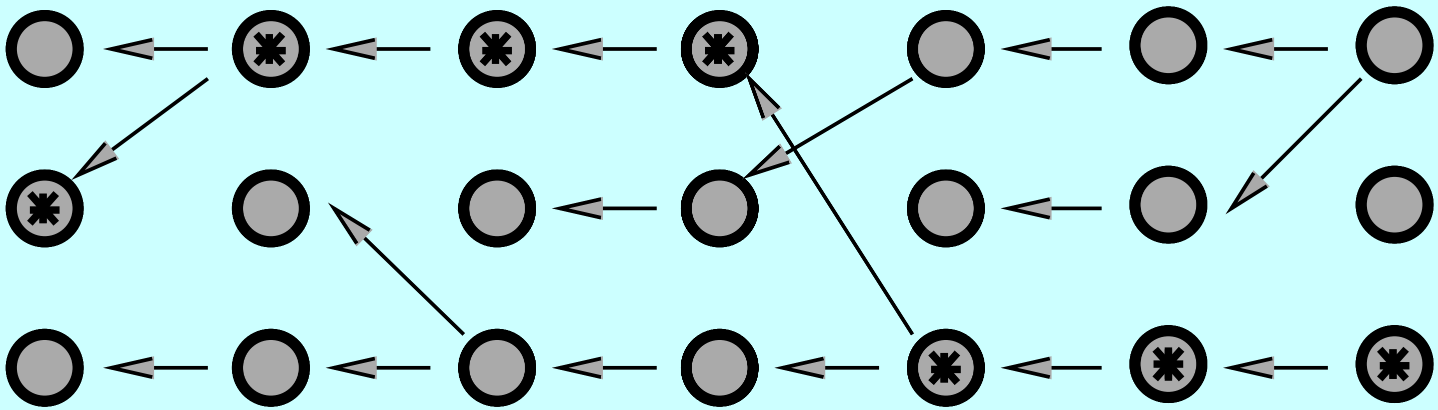
$$\text{Prob (Data | tree)} = \sum_{\text{paths}} \text{Prob(Data| tree, path)}\ \ \text{Prob(path)}$$



The rate combination contributing the most:

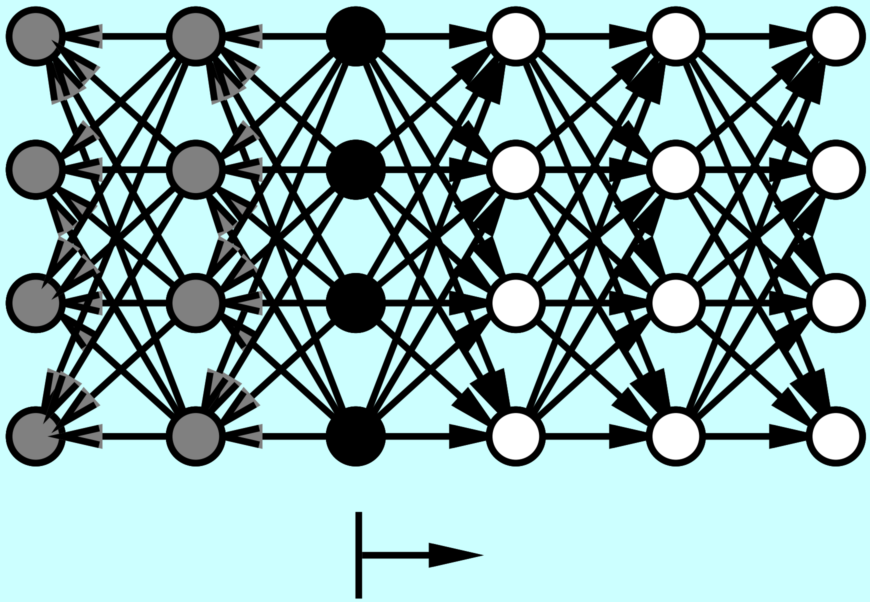
We can leave behind pointers that allow us to backtrack

This can be done by a dynamic programming algorithm called the Viterbi Algorithm, well-known in the HMM literature.



(Of course, this one might account for only 0.001 of the likelihood)

Forwards-Backwards algorithm (marginal probabilities)



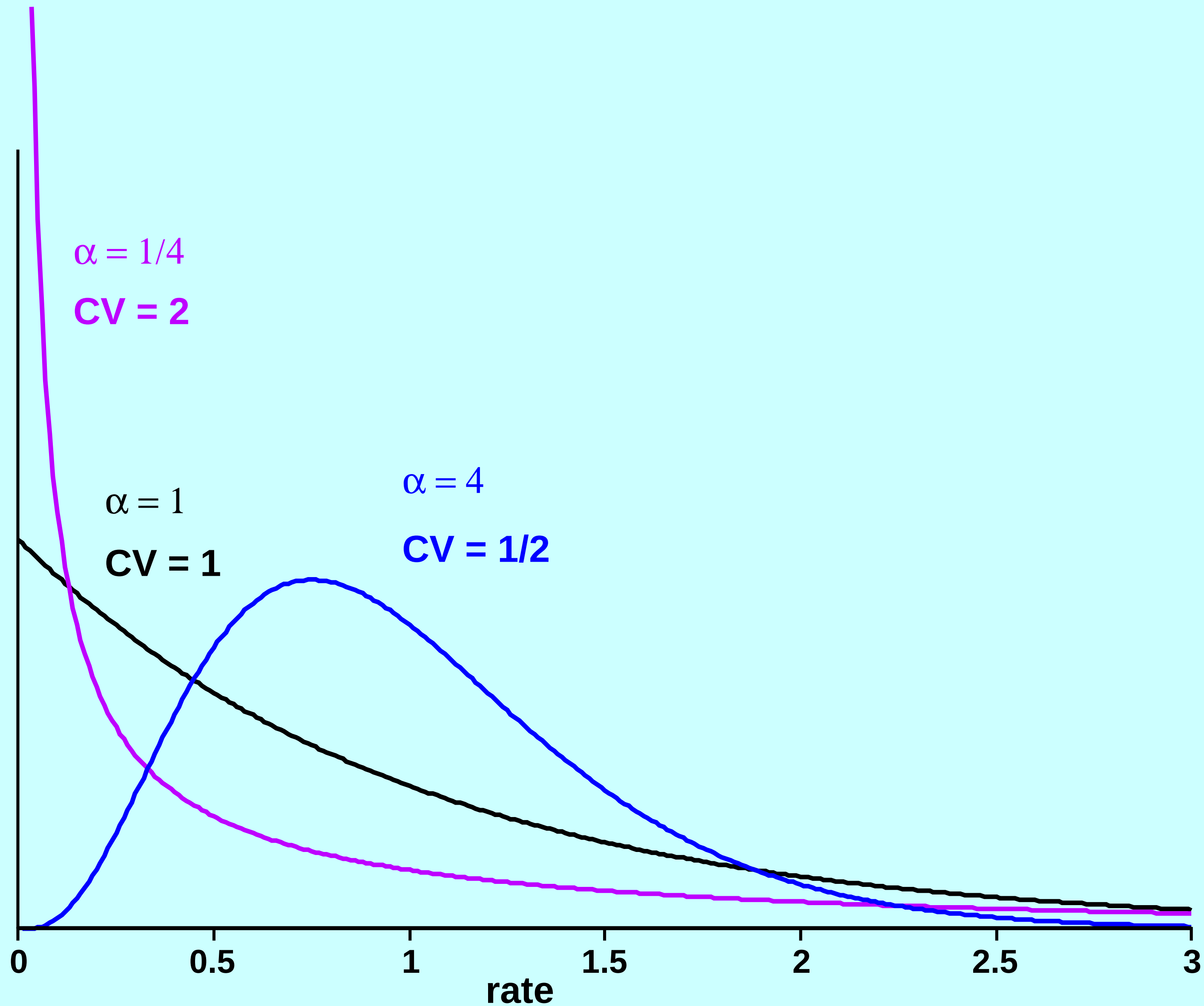
The Forwards−Backwards algorithm

can calculate the contribution of one rate

at a given site to the overall likelihood

(a little different from the Viterbi calculation)

The Gamma distribution, used for rates



A numerical example. Cyochrome B

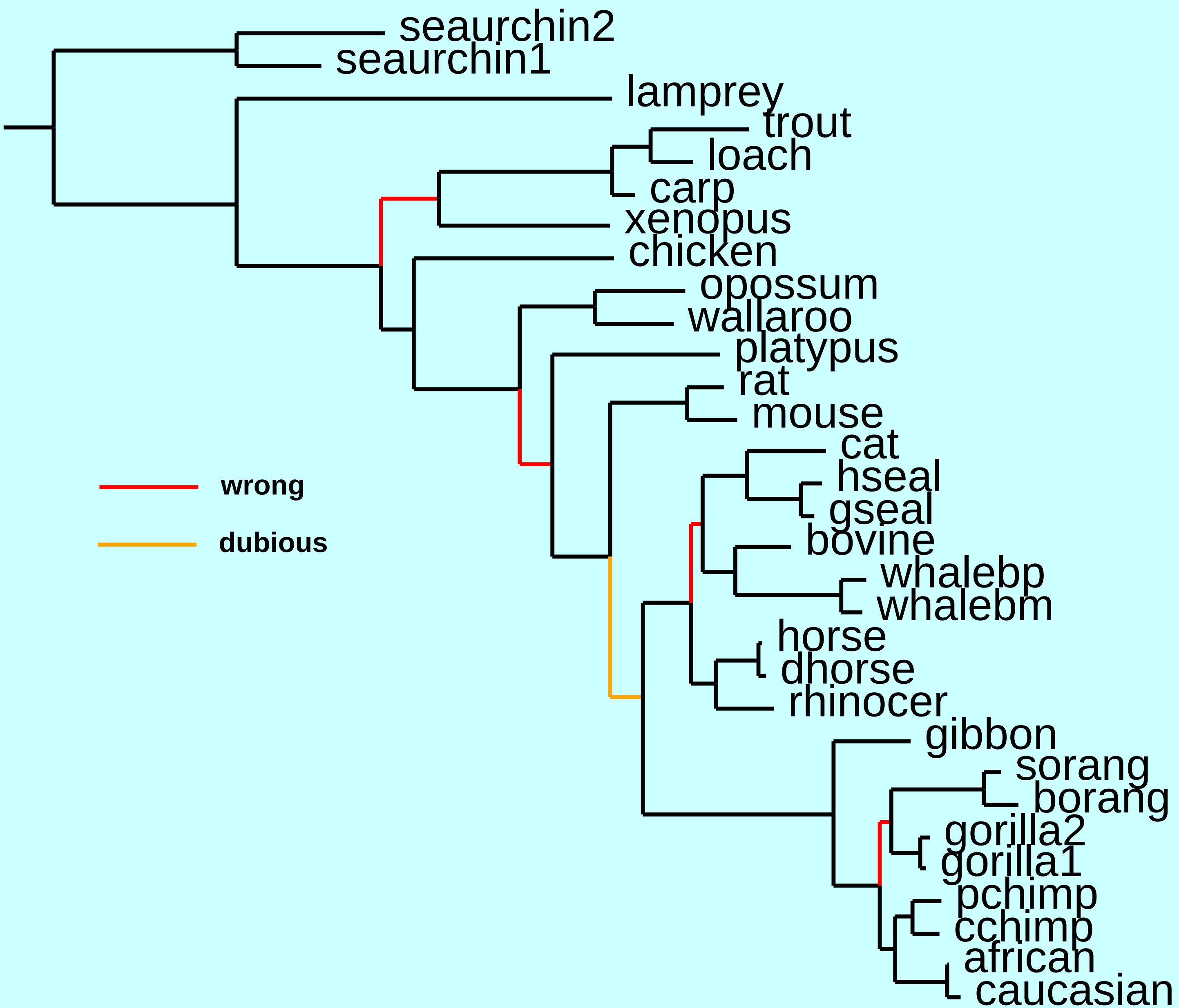
We analyze 31 cytochrome B sequences, aligned by Naoko Takezaki, using the Proml protein maximum likelihood program. Assume a Hidden Markov Model with 3 states, rates:

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

The cytochrome B tree from the above run



(It's not perfect).

Rates inferred from Cytochrome B

```
             1333333311 3222322313 3321113222 2133111111 1331133123 112211111
african      M-----TPMRK INPLMKLINH SFIDLPTPSN ISAWWNFGSL LGACLILQIT TGLFLAMH
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 333311112
african        PDASTAFSSI AHITRDVNYG WIIRYLHANG ASMFFICLFL HIGRGLYYGS FLYSETWNI
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....T
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....V
seaurchin1     A.I.L....A S..C...... .LL.NV.... ..L....MYC .........G SNKI....V
seaurchin2     A.INL....V S..C...... .LL.NV...C ..L....MYC .........L TNKI....V
```

References

Likelihood

Edwards, A. W. F. and L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. pp. 67-76 in *Phenetic and Phylogenetic Classification*, ed. V. H. Heywood and J. McNeill. Systematics Association Publication No. 6. Systematics Association, London. **[The founding paper for parsimony and likelihood for phylogenies, using gene frequencies]**

Jukes, T. H. and C. Cantor. 1969. Evolution of protein molecules. pp. 21-132 in *Mammalian Protein Metabolism,* ed. M. N. Munro. Academic Press, New York. **[The Jukes-Cantor model, in one formula and a couple of sentences]**

Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems. In *Statistical Decision Theory and Related Topics,* ed. S. S. Gupta and J. Yackel, pp. 1-27. New York: Academic Press. **[First paper on likelihood for molecular sequences. Neyman was a famous statistician.]**

Felsenstein, J. 1973. Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* **22:** 240-249. **[The pruning algorithm, parsimony is not same as likelihood]**

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17:** 368-376. **[Making likelihood useable for molecular sequences]**

(more references)

Yang, Z. 1994. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* **10:** 1396-1401. **[Use of gamma distribution of rate variation in ML phylogenies]**

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39:** 306-314. **[Approximating gamma distribution in ML phylogenies by an HMM]**

Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* **139:** 993-1005. **[Allowing for autocorrelated rates along the molecule using an HMM for ML phylogenies]**

Felsenstein, J. and G. A. Churchill. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution *Molecular Biology and Evolution* **13:** 93-104. **[HMM approach to evolutionary rate variation]**

Thorne, J. L., N. Goldman, and D. T. Jones. 1996. Combining protein evolution and secondary structure. *Molecular Biology and Evolution* **13** 666-673. **[HMM for secondary structure of proteins, with phylogenies]**

(more references)

General reading

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts. **[Book you and all your friends must rush out and buy]**

Yang, Z. 2006. *Computational Molecular Evolution*. Oxford University Press, Oxford. **[Well-thought-out book on molecular phylogenies]**

Semple, C. and M. Steel. 2003. *Phylogenetics*. Oxford University Press, Oxford. **[Good for a mathematical audience]**