Lecture 5. Rate variation, protein models, likelihood and Bayesian methods. HMMs.

Joe Felsenstein

Department of Genome Sciences and Department of Biology

Rate variation among sites

- In reality, rates of evolution are not constant among sites.

Rate variation among sites

- In reality, rates of evolution are not constant among sites.
- Fortunately, in the transition probability formulas, rates come in as simple multiples of times

$$\mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},\mathsf{u},\mathsf{t}) \;=\; \mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},1,\mathsf{ut})$$

Rate variation among sites

- In reality, rates of evolution are not constant among sites.
- Fortunately, in the transition probability formulas, rates come in as simple multiples of times

$$\mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},\mathsf{u},\mathsf{t}) \;=\; \mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},1,\mathsf{ut})$$

- Thus if we know the rates at two sites, we can compute the probabilities of change by simply, for each site, multiplying all branch lengths by the appropriate rate

(continued ...)

- If we don't know the rates, we can imagine averaging them over a distribution $f(u)$ of rates. Usually the Gamma distribution is used

$$\mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},\mathsf{t}) \;=\; \int_0^\infty \mathsf{f(u)}\,\mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},\mathsf{u},\mathsf{t})\,\mathsf{du}$$

(continued ...)

- If we don't know the rates, we can imagine averaging them over a distribution $f(u)$ of rates. Usually the Gamma distribution is used

$$\mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},\mathsf{t}) \;=\; \int_0^\infty \mathsf{f(u)}\,\mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},\mathsf{u},\mathsf{t})\,\mathsf{du}$$

- In practice a discrete histogram of rates approximates the integration

(continued ...)

- If we don't know the rates, we can imagine averaging them over a distribution $f(u)$ of rates. Usually the Gamma distribution is used

$$\mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},\mathsf{t}) \;=\; \int_0^\infty \mathsf{f(u)}\,\mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},\mathsf{u},\mathsf{t})\,\mathsf{du}$$

- In practice a discrete histogram of rates approximates the integration
- (For the Gamma it seems best to use Generalized Laguerre Quadrature to pick the rates and frequencies in the histogram).

(continued ...)

- If we don't know the rates, we can imagine averaging them over a distribution $f(u)$ of rates. Usually the Gamma distribution is used

$$\mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},\mathsf{t}) \;=\; \int_0^\infty \mathsf{f(u)}\,\mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},\mathsf{u},\mathsf{t})\,\mathsf{du}$$

- In practice a discrete histogram of rates approximates the integration
- (For the Gamma it seems best to use Generalized Laguerre Quadrature to pick the rates and frequencies in the histogram).
- Also, there are actually autocorrelations with neighboring sites having similar rates of change.

(continued ...)

- If we don't know the rates, we can imagine averaging them over a distribution $f(u)$ of rates. Usually the Gamma distribution is used

$$\mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},\mathsf{t}) \;=\; \int_0^\infty \mathsf{f(u)}\,\mathrm{Prob}\,(\mathsf{i}\,|\,\mathsf{j},\mathsf{u},\mathsf{t})\,\mathsf{du}$$

- In practice a discrete histogram of rates approximates the integration
- (For the Gamma it seems best to use Generalized Laguerre Quadrature to pick the rates and frequencies in the histogram).
- Also, there are actually autocorrelations with neighboring sites having similar rates of change.
- This can be handled by Hidden Markov Models, which we cover later.

A pioneer of protein evolution

Margaret Dayhoff, about 1966

Models of amino acid change in proteins

There are a variety of models put forward since the mid-1960's:

1. Amino acid transition matrices
   - Dayhoff (1968) model. Tabulation of empirical changes in closely related pairs of proteins, normalized. The PAM100 matrix, for example, is the expected transition matrix given 1 substitution per position.
   - Jones, Taylor and Thornton (1992) recalculated PAM matrices (the JTT matrix) from a much larger set of data.
   - Jones, Taylor, and Thornton (1994a, 1994b) have tabulated a separate mutation data matrix for transmembrane proteins.
   - Koshi and Goldstein (1995) have described the tabulation of further context-dependent mutation data matrices.
   - Henikoff and Henikoff (1992) have tabulated the BLOSUM matrix for conserved motifs in gene families.
2. Goldman and Yang (1994) pioneered codon-based models (see next screen).

Approaches to protein sequence models

- Use a good model of DNA evolution.

Approaches to protein sequence models

- Use a good model of DNA evolution.
- Use the appropriate genetic code.

Approaches to protein sequence models

- Use a good model of DNA evolution.
- Use the appropriate genetic code.
- When an amino acid changes, accept that change with a probability that is smaller, the more different the two amino acids are in their chemical properties (size, hydrophobicity etc.)

Approaches to protein sequence models

- Use a good model of DNA evolution.
- Use the appropriate genetic code.
- When an amino acid changes, accept that change with a probability that is smaller, the more different the two amino acids are in their chemical properties (size, hydrophobicity etc.)
- Fit this to empirical information on protein evolution.

Approaches to protein sequence models

- Use a good model of DNA evolution.
- Use the appropriate genetic code.
- When an amino acid changes, accept that change with a probability that is smaller, the more different the two amino acids are in their chemical properties (size, hydrophobicity etc.)
- Fit this to empirical information on protein evolution.
- Take into account variation of rates among sites, by allowing variation in rates of acceptance of changes.

Approaches to protein sequence models

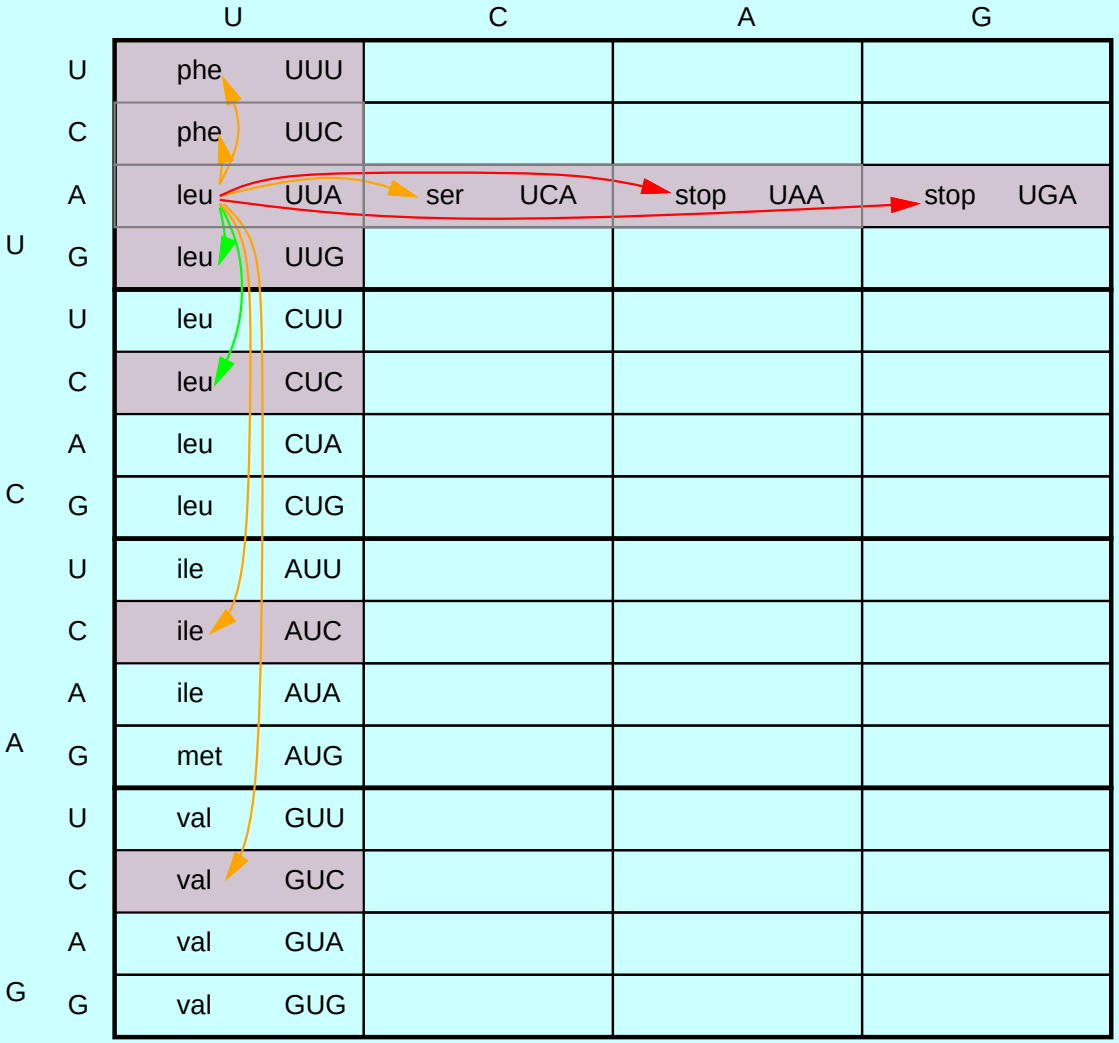
- Use a good model of DNA evolution.
- Use the appropriate genetic code.
- When an amino acid changes, accept that change with a probability that is smaller, the more different the two amino acids are in their chemical properties (size, hydrophobicity etc.)
- Fit this to empirical information on protein evolution.
- Take into account variation of rates among sites, by allowing variation in rates of acceptance of changes.
- Take into account correlation of rates of change in neighboring sites by having the acceptance rate change by an HMM.

Approaches to protein sequence models

- Use a good model of DNA evolution.
- Use the appropriate genetic code.
- When an amino acid changes, accept that change with a probability that is smaller, the more different the two amino acids are in their chemical properties (size, hydrophobicity etc.)
- Fit this to empirical information on protein evolution.
- Take into account variation of rates among sites, by allowing variation in rates of acceptance of changes.
- Take into account correlation of rates of change in neighboring sites by having the acceptance rate change by an HMM.
- How about protein structure? (as secondary structure? as 3D structure?)

Codon models

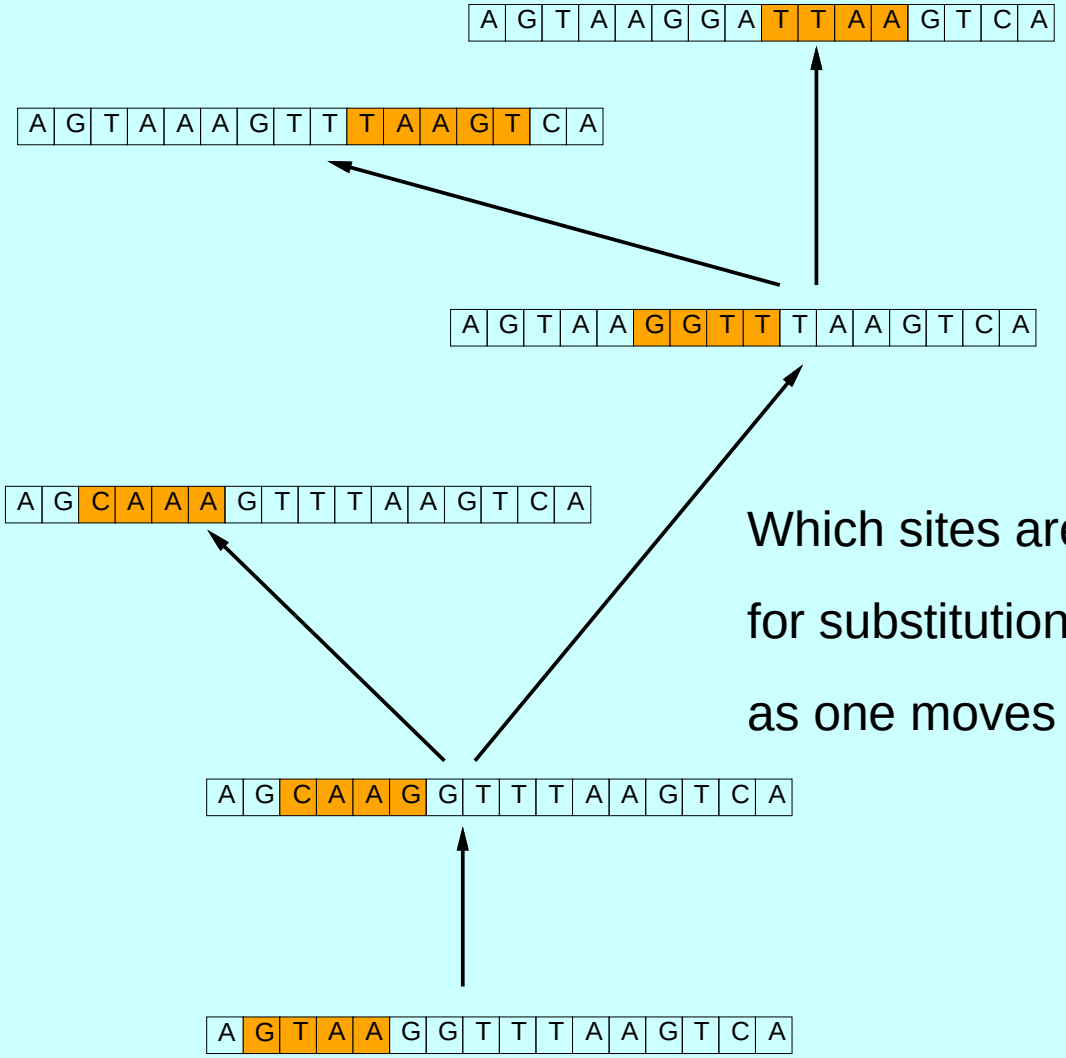
Goldman & Yang, MBE 1994; Muse and Weir, MBE 1994



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

Likelihoods and odds ratios

Bayes' Theorem relates prior and posterior probabilities of an hypothesis H:

$$\begin{aligned} \text{Prob}\,(\mathsf{H}|\mathsf{D}) &= \text{Prob}\,(\mathsf{H}\text{ and }\mathsf{D})/\,\text{Prob}\,(\mathsf{D}) \\ &= \text{Prob}\,(\mathsf{D}|\mathsf{H})\,\text{Prob}\,(\mathsf{H})/\,\text{Prob}\,(\mathsf{D}) \end{aligned}$$

The ratios of posterior probabilities of two hypotheses, H_1 and H_2, put this into its "odds ratio" form ($\text{Prob}\,(\mathsf{D})$ cancels):

$$\frac{\text{Prob}\,(\mathsf{H}_1|\mathsf{D})}{\text{Prob}\,(\mathsf{H}_2|\mathsf{D})} = \frac{\text{Prob}\,(\mathsf{D}|\mathsf{H}_1)}{\text{Prob}\,(\mathsf{D}|\mathsf{H}_2)}\,\frac{\text{Prob}\,(\mathsf{H}_1)}{\text{Prob}\,(\mathsf{H}_2)}$$

Note that this says that the posterior odds in favor of H_1 over H_2 are the product of the prior odds and a likelihood ratio. The likelihood of the hypothesis H is the probability of the observed data given it, ($\text{Prob}\,(\mathsf{D}\mid\mathsf{H})$). This is *not* the same as the probability of the hypothesis given the data. That is the posterior probability of H and requires that we also have a believable prior probability ($\text{Prob}\,(\mathsf{H})$)

Rationale of likelihood inference

If the data consists of n items that are conditionally independent given the hypothesis i,

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

and we can then write the likelihood ratio as a product of ratios:

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

If the amount of data is large the likelihood ratio terms will dominate and push the result towards the correct hypothesis. This can console us somewhat for the lack of a believable prior.

Properties of likelihood inference

Likeihood inference has (usually) properties of

- Consistency. As the number of data items n gets large, we converge to the correct hypothesis with probability 1.
- Efficiency. Asymptotically, the likelihood estimate has the smallest possible variance (it need not be best for any finite number n of data points).

A simple example – coin tossing

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

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

so that trying to maximize it we get

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

finding the ML estimate

and searching for a value of p for which the slope is zero:

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

which has roots at 0, 1, and $5/11$

Log likelihoods

Alternatively, we could maximize not L but its logarithm.

This turns products into sums:

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

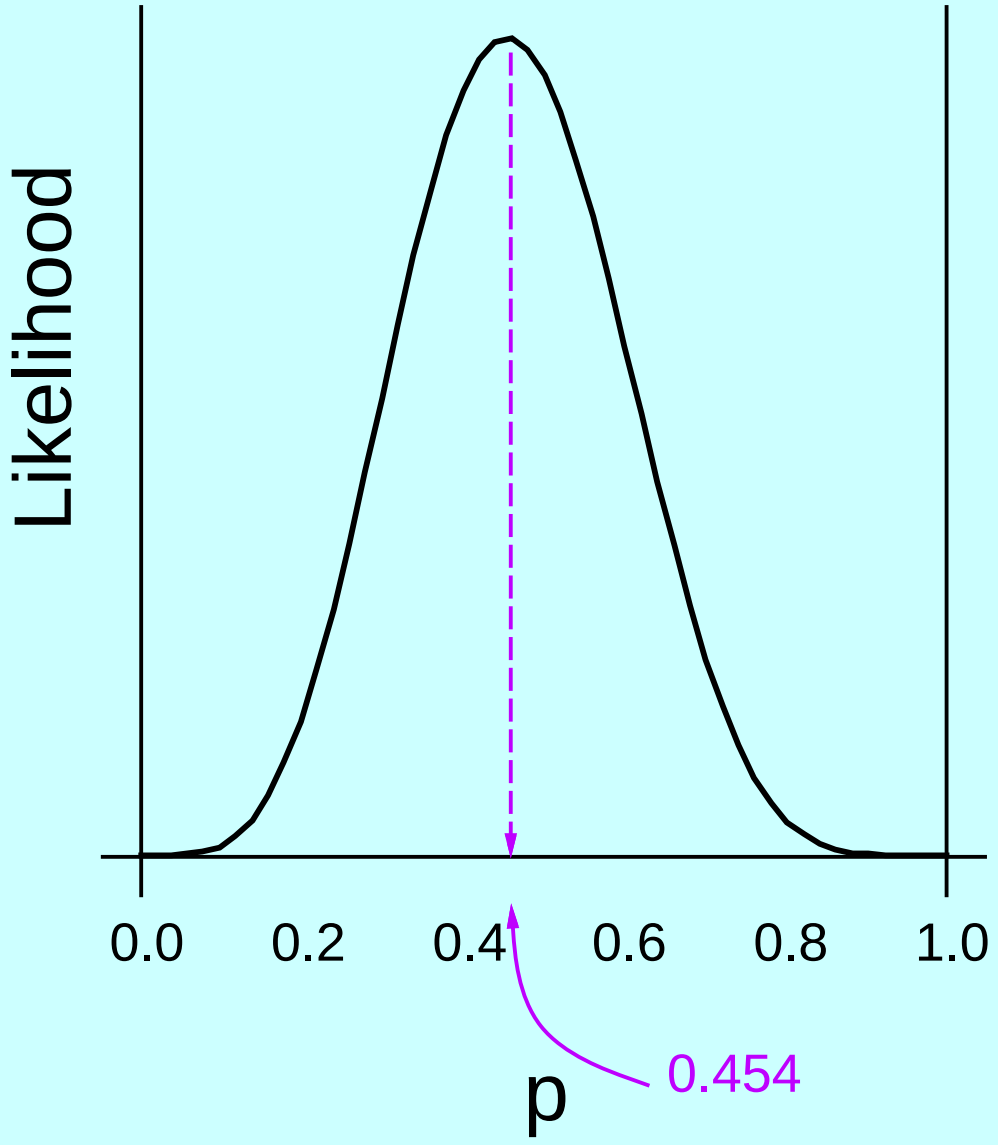
whereby

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

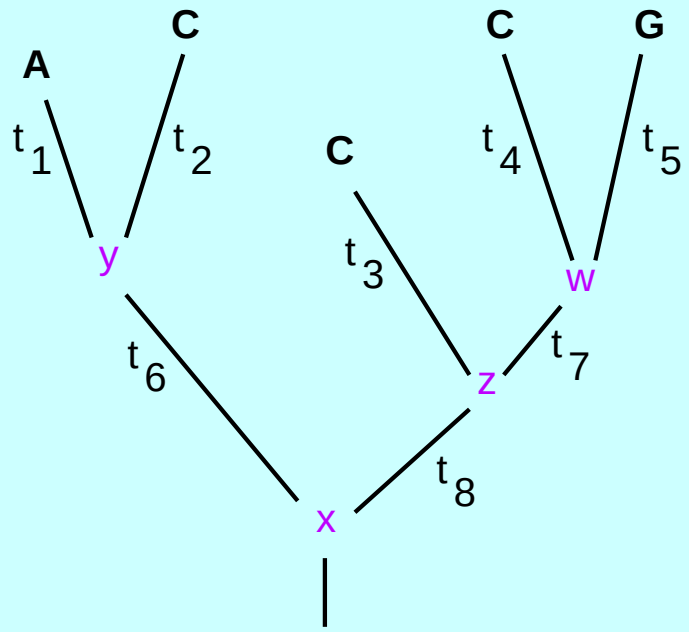
so that finally

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

Likelihood curve for coin tosses



Likelihood on trees



A tree, with branch lengths, and the data at a single site This example is used to describe calculation of the likelihood

Since the sites evolve independently on the same tree,

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

Likelihood at one site on a tree

We can compute this by summing over all assignments of states x, y, z and w to the interior nodes

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

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

Computing the terms

For each combination of states, the Markov process allows us to express it as a product of probabilities of a series of changes, with the probability that we start in state x:

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

Computing the terms

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

$$\sum_x \sum_y \sum_z \sum_w$$

$$\begin{aligned}
&\text{Prob}\,(x) \quad \text{Prob}\,(y|x,t_6) \quad \text{Prob}\,(A|y,t_1)\,\text{Prob}\,(C|y,t_2)\\
&\qquad\quad \text{Prob}\,(z|x,t_8) \quad \text{Prob}\,(C|z,t_3)\\
&\qquad\quad\qquad\quad \text{Prob}\,(w|z,t_7)\,\text{Prob}\,(C|w,t_4)\,\text{Prob}\,(G|w,t_5)
\end{aligned}$$

Getting a recursive algorithm

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

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

$$\sum_{\mathsf{x}} \text{Prob}\left(\mathsf{x}\right)$$

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

$$\left(\sum_{\mathsf{z}} \text{Prob}\left(\mathsf{z}|\mathsf{x},\mathsf{t}_8\right) \quad \text{Prob}\left(\mathsf{C}|\mathsf{z},\mathsf{t}_3\right)\right.$$

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

The pruning algorithm

Note that the pattern of parentheses in the previous expression is the

$$(A, C)\,(C, (C, G))$$

If $L_k^{(i)}(s)$ is the probability of everything that is observed from node k on the tree on up, at site i, conditional on node k having state s, we can express

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

as:

$$\sum_w \text{Prob}\,(w|z, t_7)\ L_7(w)$$

and the algorithm is:

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

$$\mathsf{L}_{\mathsf{k}}^{(\mathsf{i})}(\mathsf{s}) = \left(\sum_{\mathsf{x}} \text{Prob}\ (\mathsf{x}|\mathsf{s}, \mathsf{t}_\ell)\, \mathsf{L}_\ell^{(\mathsf{i})}(\mathsf{x})\right) \times \left(\sum_{\mathsf{y}} \text{Prob}\ (\mathsf{y}|\mathsf{s}, \mathsf{t}_\mathsf{m})\, \mathsf{L}_\mathsf{m}^{(\mathsf{i})}(\mathsf{y})\right)$$

Starting and finishing the recursion

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

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

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

$$L^{(i)} = \sum_x \pi_x L_0^{(i)}(x)$$

Ambiguity and error in the sequences

Ambiguity. If a tip has an ambiguity state such as R (purine, either A or G) we use

$$\mathsf{L}^{(\mathsf{i})} = (1, 0, 1, 0)$$

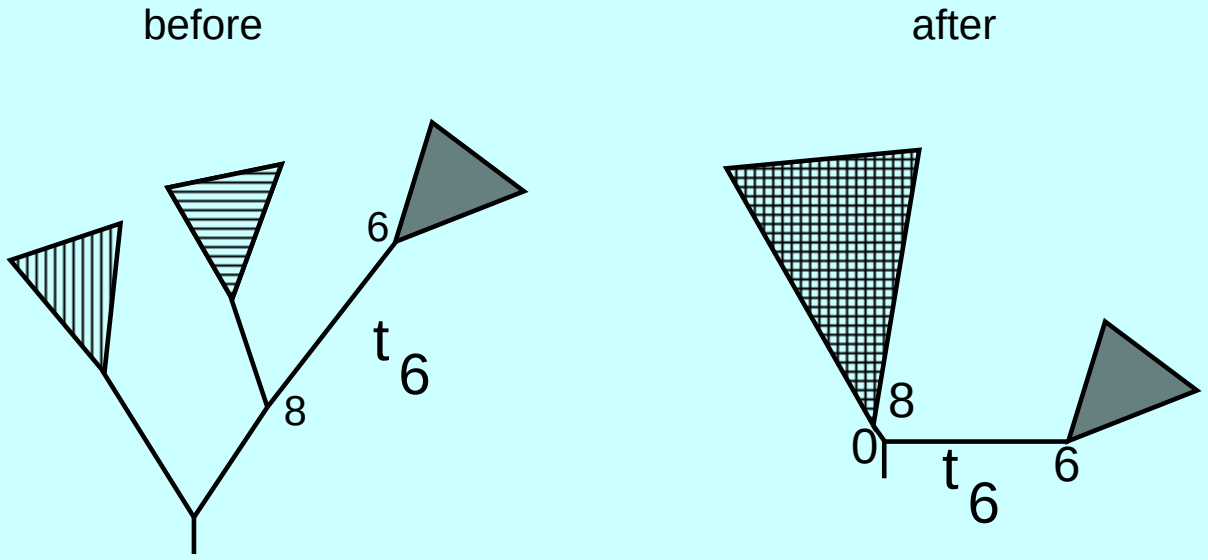
and if it has an unknown nucleotide ("N")

$$\mathsf{L}^{(\mathsf{i})} = (1, 1, 1, 1)$$

This handles ambiguities naturally.

Error. If our sequencing has probability $1 - \epsilon$ of finding the correct nucleotide, and $\epsilon/3$ of inferring each of the three other possibilities, when an A is observed, the four values should be $(1 - \epsilon, \epsilon/3, \epsilon/3, \epsilon/3)$, and when a C is observed, they should be $(\epsilon/3, 1 - \epsilon, \epsilon/3, \epsilon/3)$.
The result is a simple handling of sequencing error, provided it occurs independently in different bases.

The tree is effectively unrooted



The region around nodes 6 and 8 in the tree, when a new root (node 0) is placed in that branch
The subtrees are shown as shaded triangles

For the tree on the left of the figure above,

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

using reversibility ...

Reversibility of the substitution process guarantees us that

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

Substituting, we get

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

Finally we see that this is the same as the likelihood for a tree rooted at node 8:

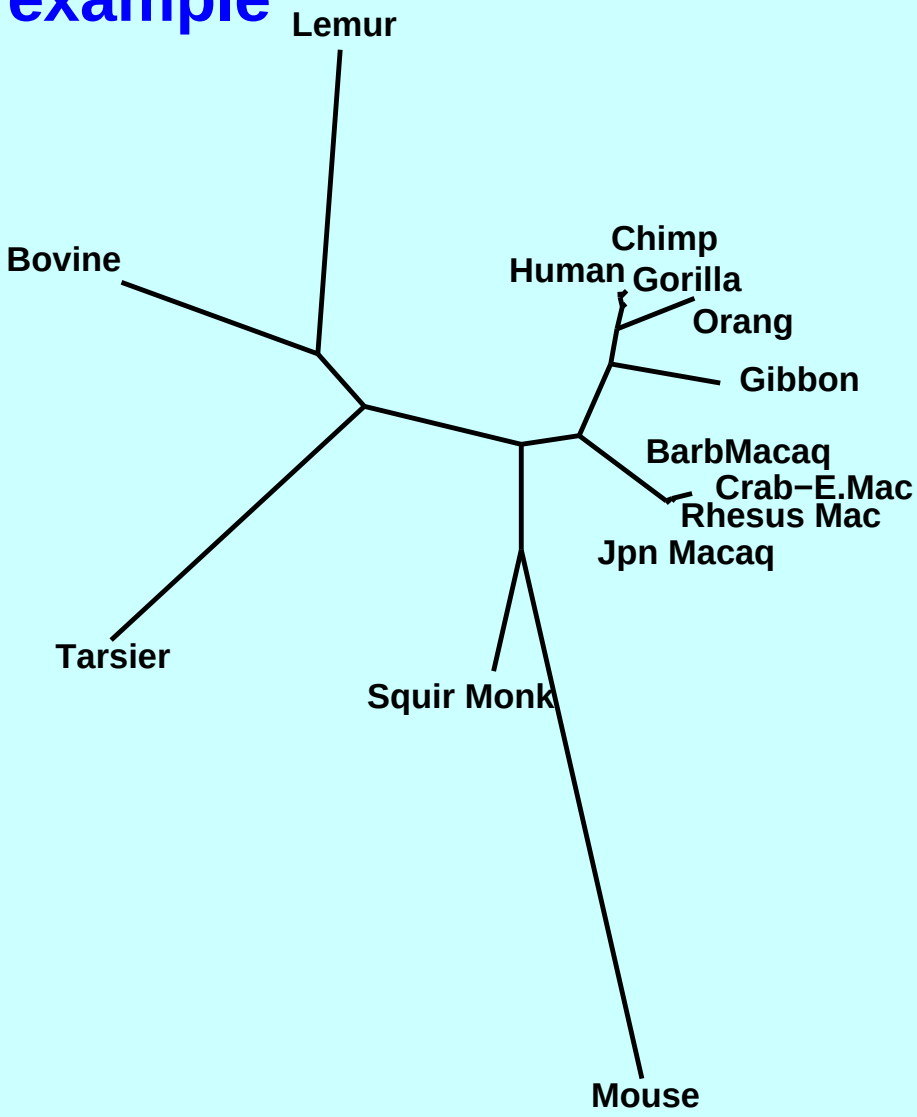
$$L_0^{(i)}(z) = L_8^{(i)}(z)\,\text{Prob}\,(z)\,\text{Prob}\,(w|z,t_6)L_6^{(i)}(w)$$

Finding the ML tree

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

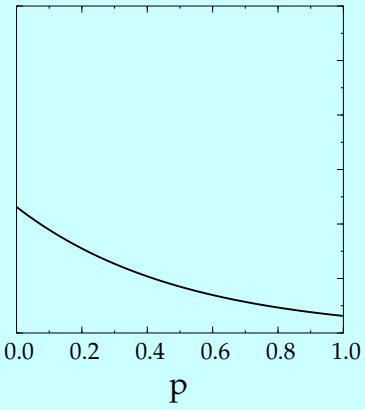
As with the distance matrix methods, we must search the space of tree topologies, and for each one examined, we need to optimize the branch lengths to maximize the likelihood.

A numerical example



A 232-nucleotide mitochondrial noncoding region data set over 14 species gives this ML tree with $\ln \mathrm{L} = -2616.86$ with a transition/transversion ratio of 30

Bayesian inference with coin tossing:



Bayesian methods

An example of Bayesian inference with coin-tossing. The probability of heads is assumed to have a prior (top) which is a truncated exponential with mean 0.34348 on the interval (0,1). The likelihood curve (middle) and the posterior on the probability of heads (bottom) are shown, when there are 11 tosses with 5 heads.

Bayesian phylogeny methods

Bayesian inference has been applied to inferring phylogenies (Rannala and Yang, 1996; Mau and Larget, 1997; Li, Pearl and Doss, 2000).

- All use a prior distribution on trees. The prior has enough influence on the result that its reasonableness should be a major concern. In particular, the depth of the tree may be seriously affected by the distribution of depths in the prior.
- All use Markov Chain Monte Carlo (MCMC) methods (we will introduce these in our discussion of coalescents) They sample from the posterior distribution.
- When these methods make sense they not only get you a point estimate of the phylogeny, they get you a posterior distribution of possible phylogenies.

References, page 1

Dayhoff, M. O. and R. V. Eck. 1968. *Atlas of Protein Sequence and Structure 1967-1968.* National Biomedical Research Foundation, Silver Spring, Maryland. [Dayhoff's PAM modelfor proteins]

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer Applcations in the Biosciences (CABIOS)* **8:** 275-282. [JTT model for proteins]

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1994a. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33:** 3038-3049. [JTT membrane protein model]

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1994b. A mutation data matrix for transmembrane proteins. *FEBS Letters* **339:** 269-275 . [JTT membrane protein model]

Henikoff, S. and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences, USA* **89:** 10915-10919. [BLOSUM protein model]

References, page 2

Koshi, J. M. and R. A. Goldstein. 1995. Context-dependent optimal substitution matrices. *Protein Engineering* **8:** 641-645. [Generating other kinds of protein model matrices]

Fitch, W. M. and E. Markowitz. 1970. An improved method for determi ning codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics* **4:** 579-593. [The first suggestion of a covarion model]

Muse, S. V. and B S. Gaut. 1994. A likelihood method for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11:** 715-724. [One of the two introductions of the codon model]

Goldman, N. and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11:** 725-736 [One of the two introductions of the codon model]

Fisher, R. A. 1912. On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* **41:** 155-160. [First modern paper introducing likelihood]

References, page 3

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A* **222:** 309-368. [Likelihood in generality]

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

Kashyap, R. L., and S. Subas. 1974. Statistical estimation of parameters in a phylogenetic tree using a dynamic model of the substitutional process. *Journal of Theoretical Biology* **47:** 75-101. [Second paper applying likelihood to molecular sequences]

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 Publ. No. 6, London. [First paper on likelihood for phylogenies]

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]

References, page 4

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17:** 368-376. [Made likelihood practical for n species]

Li, S., D. Pearl, and H. Doss. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American Statistical Association* **95:** 493-508. [Bayesian inference of phylogenies by MCMC]

Mau, B., M. A. Newton, and B. Larget. 1997. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Molecular Biology and Evolution* **14:** 717-724. [Bayesian inference of phylogenies by MCMC]

Rannala, B. and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Molecular Evolution* **43:** 304-311. [Bayesian inference of phylogenies by MCMC]

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts. [material is in chapters 13, 14, 16]

Yang, Z. 2006. *Computational Molecular Evolution*. Oxford University Press, Oxford. [material is in pages 89-93, and chapters 1, 2]

Semple, C. and M. Steel. 2003. *Phylogenetics*. Oxford University Press, Oxford. [Material is in pages 145-160 and chapter 8]