Integrative Biology 200A "PRINCIPLES OF PHYLOGENETICS" Spring 2014
University of California, Berkeley Kipling Will- 14 Feb*

Maximum Likelihood:

Maximum likelihood is a general statistical method for estimating unknown parameters of a probability model. A parameter is some descriptor of the model. A familiar model might be the normal distribution of a population with two parameters: the mean and variance. In phylogenetics there are many parameters, including rates, differential transformation costs, and, most important, the tree itself.

Likelihood is defined to be a quantity proportional to the *probability of observing the data given the model*, P(D|M). Thus, if we have a model (i.e. the tree and parameters), we can calculate the probability the observations would have actually been observed as a function of the model. We then examine this likelihood function to see where it is greatest, and the value of the parameter of interests (usually the tree and/or branch lengths) at that point is the maximum likelihood estimate of the parameter.

Simple Coin Flip example:

The likelihood for heads with probability *p* for a series of 11 tosses assumed to be independent-

HHTTHTHHTTT 5 heads, 6 tails

Assuming a fair coin what is the likelihood of this series results?
$L=(n!/k!(n-k)!)p^{k}(1-p)^{n-k}= 0.22559$ [for p=0.5]

Where **n** is the number of tosses, **k** is the number coming up heads, **p** is the probability of heads. **L** is maximized (0.23609) when p= 0.45454, intuitively 5/11. Thinking of **L** as a function dependent on the number of tosses (**n**), number of heads observed (**k**) and the true probability (**p**) of getting heads in a single toss. Since **n** and **k** are observed you can try various values for **p** to find the one that maximizes **L**. In other words, this can be plotted by brute force determination, or calculated by taking the derivative of the plot and looking for where the slope =0.

Maximum Likelihood can be used as an optimality measure for choosing a preferred tree or set of trees. It evaluates a hypothesis (branching pattern), which is a proposed evolutionary history, in terms of the probability that the implemented model and the hypothesized history would have given rise to the observed data set. Essentially a pattern that has a higher probability is preferred over one with lower probability. *Use of an optimality criterion to select trees is a feature shared with parsimony methods.*

Advantages and disadvantages of maximum likelihood methods:

<u>Supposed advantages</u>.
- Appropriate for simple data like DNA sequences, where we can reasonably model the largely stochastic processes, i.e. a statistical description of the stochastic processes.
-lower variance than other methods (i.e. estimation method least affected by sampling error)
-robust to many violations of the assumptions in the evolutionary model, even with very short sequences it may outperform alternative methods such as parsimony or distance methods.

-the method is statistically well understood
-has explicit model of evolution that you can make fit the data
-evaluate different tree topologies (vs. NJ)
-use all the sequence information (vs. Distance)
-better accounting for branch lengths, e.g. incorporates "multiple hits" thereby providing more realistic branch length and reducing the region of LBA. Also, information is derived from sites that would be uninformative under parsimony.

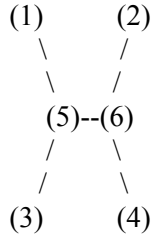
Supposed disadvantages.
-very computationally intensive and so slow (though this is becoming much less of an issue)
-Apparently susceptible to asymmetrical presence of data in partitions *(see Simmons, M.P., 2011. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. Cladistics. 27:1-15.)*
-the result is dependent on the model used and information is derived from sites that are uninformative under parsimony is only due to the model used.
-questionably applicable to complex data like morphology given the difficulty of modeling the numerous processes
-philosophically less well established, especially in terms the applicability of probabilities and statistical measures of unique historical events (vs. Parsimony as a general principle). This is a fundamental distinction between reconstruction and estimation, e.g. *"Although the true phylogeny maybe "unknowable" it can nonetheless be estimated…"* Phylogenetic Inference", Swofford, Olsen, Waddell, and Hillis, in Molecular Systematics, 2nd ed., Sinauer Ass., Inc., 1996, Ch. 11.

Simple tree example:

Assume that we have the aligned nucleotide sequences for four taxa:

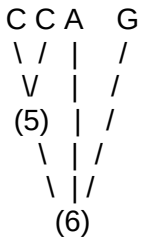
	1	**j**	....N
(1)	A G G C T	**C**	C A A ....A
(2)	A G G T T	**C**	G A A ....A
(3)	A G C C C	**A**	G A A.... A
(4)	A T T T C	**G**	G A A.... C

and we want to evaluate the likelihood of the unrooted tree represented by the nucleotides of site **j** in the sequence and shown below:



What is the probability that this tree would have generated the data presented in the sequence under the chosen model?

Since most of the models currently used are **time-reversible**, the likelihood of the tree is generally independent of the position of the root. Therefore it is convenient to root the tree at an arbitrary internal node as done below,



Under the assumption that nucleotide sites evolve independently (the Markovian model), we can calculate the likelihood for each site separately and combine the likelihood into a total value at the end. To calculate the likelihood for site **j**, we have to consider all the possible scenarios by which the nucleotides present at the tips of the tree could have evolved. So the likelihood for a particular site is the summation of the probabilities of every possible reconstruction of ancestral states, given some model of base substitution. So in this specific case all possible nucleotides A, G, C, and T occupying nodes (5) and (6), or 4 x 4 = 16 possibilities.

Since any one of these scenarios could have led to the nucleotide configuration at the tip of the tree, we must calculate the probability of each and sum them to obtain the total probability for each site **j**.

The likelihood for the full tree then is the product of the likelihood at each site.

$$L = L(1) \times L(2) \ldots\ldots \times L(N) = \prod_{j=1}^{N} L(j)$$

Since the individual likelihoods are extremely small numbers it is convenient to sum the log likelihoods at each site and report the likelihood of the entire tree as the log likelihood.

$$\ln L = \ln L(1) + \ln L(2) \ldots\ldots + \ln L(N) = \sum_{j=1}^{N} \ln L(j)$$

Models:

The basic form is a matrix: Q =

	A	C	G	T
A	$-\mu(a\pi_C+b\pi_G+c\pi_T)$	$\mu a\pi_C$	$\boldsymbol{\mu b\pi_G}$	$\mu c\pi_T$
C	$\mu g\pi_A$	$-\mu(g\pi_A+d\pi_G+e\pi_T)$	$\mu d\pi_G$	$\boldsymbol{\mu e\pi_T}$
G	$\boldsymbol{\mu h\pi_A}$	$\mu i\pi_C$	$-\mu(h\pi_C+j\pi_G+f\pi_T)$	$\mu f\pi_T$
T	$\mu i\pi_A$	$\boldsymbol{\mu k\pi_C}$	$\mu l\pi_G$	$-\mu(i\pi_C+k\pi_G+l\pi_T)$

Where μ= mean instantaneous substitution rate, where time is measured as substitutions μ=1
a,b,c,…l = rate parameter for each possible transformation of one base to another
π_A= frequency of bases A, C,G, & T
transitions in bold

Nearly all substitution models are rate tables that are variation of this general form, e.g. JC sets all rates equal to 1 (a…l =1) and frequency are equal (π_A, π_C, π_G, π_T all equal ¼), K2P where the observation that transitions and tranversions occur at different rates (b, e, h, k are adjusted by constant K).

Among-Site Rate Variation (Γ)
The starting hypothesis is that all sites are assumed to have equal rates of substitution. This assumption can be relaxed, allowing rates to differ across sites by having rates drawn from a gamma distribution. The gamma is useful as its shape parameter (ά) has a strong influence on the values in the distribution.

Choosing a model:

As you might imagine, there are many models already available (ModelTest discussed below looks at 56!!) and an effectively infinite number are possible. How can one choose?

The program ModelTest (Posada & Crandal 1998) uses log likelihood scores to establish the model that best fits the data. Goodness of fit is tested using the likelihood ratio score.

max [L_0 (simpler model) | Data]
max [L_1 (more complex model) | Data]

This is a nested comparison (i.e. L_0 is a special case of L_1)

Adding additional parameters will always result in a higher likelihood score. However, at some point adding additional parameters is no longer justified in terms of significant improvement in fit of a model to a particular dataset. Over parameterizing simply fits the model to noise in the data.

A simple example:

HKY85 -lnL = 1787.08
GTR -lnL = 1784.82

Then, LR= 2*(lnL1-lnL2) ; LR = 2 (1787.08 - 1784.82) = 4.52

degrees of freedom = 4 (GTR adds 4 additional parameters to HKY85)
critical value (P = 0.05) = 9.49

The added parameters are not justified by this significance test.

Some typical models.

- JC, Jukes & Cantor (1969): all substitutions are equal and all base frequencies are equal. Most restrictive.
- F81, Felsenstein (1981): all substitutions are equal, base frequencies can vary.
-K2P, Kimura 2 parameter, Kimura (1980): Transitions and transversions have different substitution rates, base frequencies are assumed equal.
-HKY85, Hasegawa-Kishino-Yano (1985): Transitions and transversions have different substitution rates, base frequencies can vary.
-GTR: General Time Reversible (Lanave et al. 1984) Six classes of substitutions, base frequencies vary.

*some of the text is taken from online notes provided by M. Sanderson or the text book by Swofford et al. and other sources online.

		H	H	T	T	H	T	H	H	T	T	T	(p^5*(1-p)^6	L
h=p	p=0	0	0	1	1	0	1	0	0	1	1	1	0.000000	0.00000
t=1-p	p=0.1	0.1	0.1	0.9	0.9	0.1	0.9	0.1	0.1	0.9	0.9	0.9	0.000005	0.00246
	p=0.2	0.2	0.2	0.8	0.8	0.2	0.8	0.2	0.2	0.8	0.8	0.8	0.000084	0.03876
	p=0.3	0.3	0.3	0.7	0.7	0.3	0.7	0.3	0.3	0.7	0.7	0.7	0.000286	0.13208
	p=0.4	0.4	0.4	0.6	0.6	0.4	0.6	0.4	0.4	0.6	0.6	0.6	0.000478	0.22072
	p=0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.000488	0.22559
	p=0.6	0.6	0.6	0.4	0.4	0.6	0.4	0.6	0.6	0.4	0.4	0.4	0.000319	0.14715
	p=0.7	0.7	0.7	0.3	0.3	0.7	0.3	0.7	0.7	0.3	0.3	0.3	0.000123	0.05661
	p=0.8	0.8	0.8	0.2	0.2	0.8	0.2	0.8	0.8	0.2	0.2	0.2	0.000021	0.00969
	p=0.9	0.9	0.9	0.1	0.1	0.9	0.1	0.9	0.9	0.1	0.1	0.1	0.000001	0.00027
	p=1	1	1	0	0	1	0	1	1	0	0	0	0.000000	0.00000

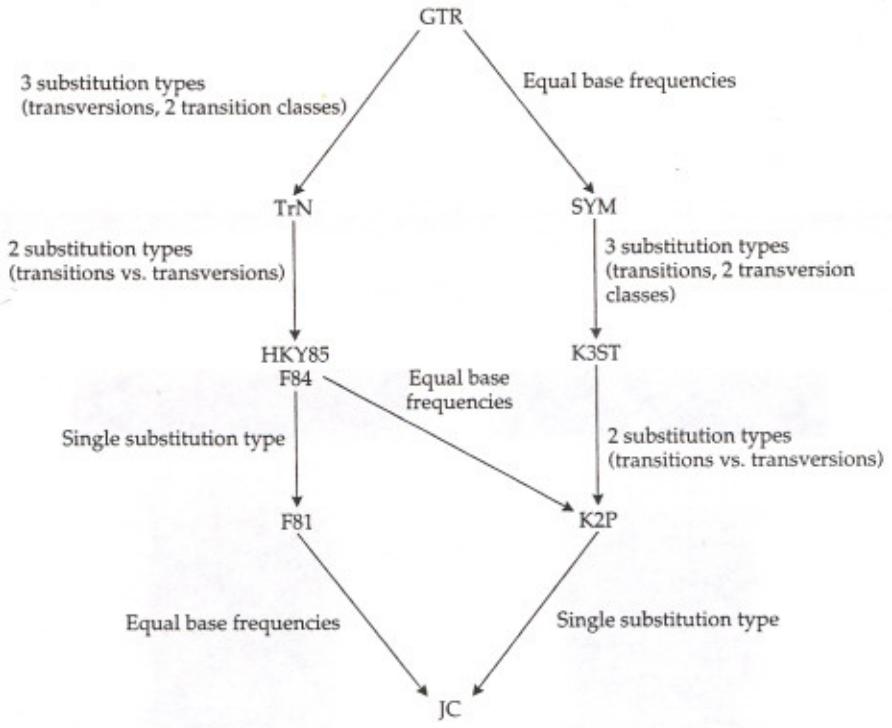


Figure 11 Relationship between special cases of the general time-reversible family of substitution models. Arrow labels indicate restrictions that convert a more general model to a more specific one. Model abbreviations: F81, model of Felsenstein, 1981a (equivalent to the "equal input" model of Tajima and Nei, 1982); F84, model used in versions 2.6 and later of PHYLIP (Felsenstein, 1993; Kishino and Hasegawa, 1989); GTR, General time-reversible (Lanave et al., 1984; Tavaré, 1986; Rodríguez et al., 1990); HKY85, Hasegawa-Kishino-Yano model (Hasegawa et al., 1985b); JC, Jukes and Cantor (1969) model; K2P, Kimura (1980) two-parameter model; K3ST, Kimura (1981) three-substitution-type model; SYM, model described by Zharkikh (1994); TrN, Tamura and Nei (1993) model.

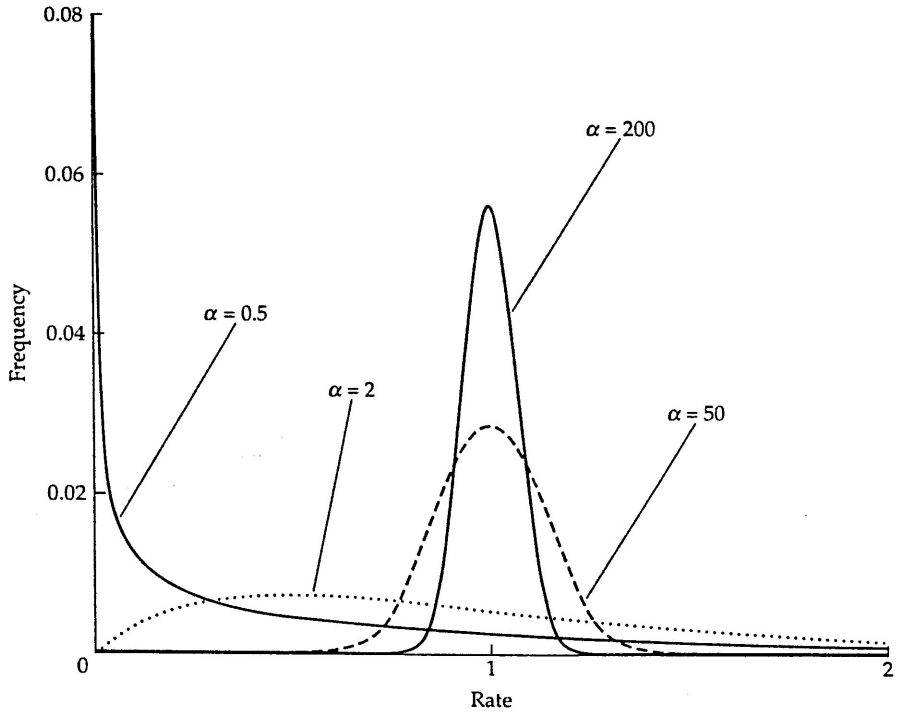


Figure 13 The gamma distribution for four different values of the shape parameter (α). When α is small, most of the sites evolve very slowly, but a few sites have moderate-to-high rates. As α increases, the distribution becomes more peaked and symmetrical about a mean rate of 1.0. When α is infinity, all sites have relative rate 1.0, so that an equal-rates model can be obtained as a special case of the gamma model.