Maximum likelihood method in phylogenetics

Ziheng Yang (楊子恒)
Department of Biology
University College London

Plan

1. **Introduction to likelihood**
2. **Likelihood calculation on a tree**
3. **Models used in likelihood analysis**
4. **Reversibility & root of tree**
5. **Likelihood ratio test & model selection**

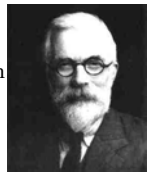
Likelihood & maximum likelihood

Likelihood is a central concept in statistics. Maximum likelihood is a major statistical methodology. (The other main competitor is Bayesian method.)

Methods discussed in a typical biostatistics course (χ^2 test, *t* test, ANOVA, *F*-test, correlation etc.) are special cases of maximum likelihood or its approximations.

Maximum likelihood is due to Fisher (1912).

Ronald A. Fisher (1890–1962)



1912: graduate, Caius College Cambridge
1919-1933: Rothamsted Agricultural Station
1925: *Statistical Methods for Research Workers* (14th Edition in 1970)
1929: Fellow of the Royal Society
1930: *Genetical Theory of Natural Selection*
1933: Galton Professor of Eugenics, UCL
1935: *The Design of Experiments* (8th Edition in 1966)
1943-1957: Balfour Professor of Genetics, Cambridge
1962 (29 July): died in Adelaide

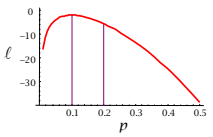
Likelihood is the probability of the data, viewed as a function of the unknown parameters.

Example 1. There are many red and blue fish in a pond. We want to estimate the proportion of red fish in the pond (p). We take a sample of $n = 100$ fish and found $x = 10$ red and $n - x = 90$ blue.

$$L(p;x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{100}{10} p^{10}(1-p)^{90}$$

$$\ell(p;x) = \log\binom{100}{10} + 10\log(p) + 90\log(1-p)$$

$$\hat{p} = 10/100 = 0.1$$



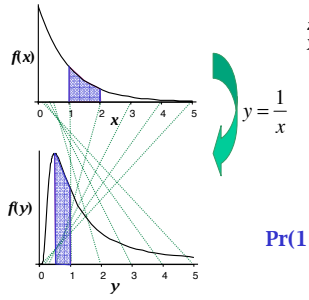
Notes

- $\binom{n}{x} = \binom{100}{10}$ is a constant and can be ignored.
- We can say that $p = 0.1$ is more likely to be true than $p = 0.2$, but prob($0.2 < p < 0.3$) is not a meaningful concept.

Probability vs. likelihood

- Probability is considered a function of the data with the parameter given (from *population* to *sample*) while likelihood is a function of the parameter when the data have been observed (from *sample* to *population*).
- Likelihood is relative, defined up to a proportionality constant. Probability sums (integrates) to one.
- The height on a likelihood curve is meaningful but the area is not. The area on a probability curve is meaningful but the height is not.

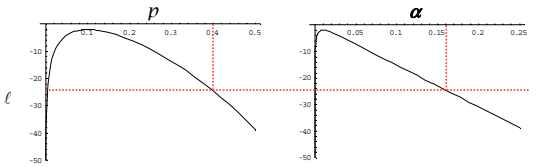
The area on a probability curve is preserved during variable transformation, but height is not.



x: exponential with mean 1
y: inverse gamma

$\Pr(1 < x < 2) = \Pr(\tfrac{1}{2} < y < 1)$

On a likelihood curve, height is preserved during parameter transformation (reparametrization). The area does not have a meaning.



$\ell(p;x) = 10\log(p) + 90\log(1-p)$ $\quad$ $\ell(\alpha;x) = 10\log(\sqrt{\alpha}) + 90\log(1-\sqrt{\alpha})$

Suppose $\alpha = p^2$. Then $\hat{\alpha} = \hat{p}$
If $\hat{p} = 0.1$ is the MLE of p, then $\hat{\alpha} = 0.01$ is the MLE of α
MLEs are known to be invariant to re-parametrizations.

Likelihood: example 2 (ABO blood groups)

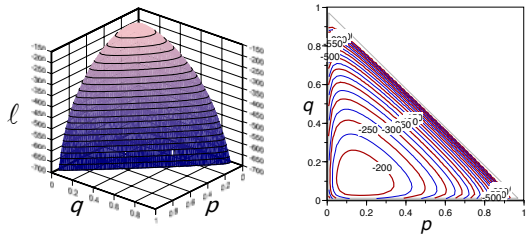
Phenotypes	Genotypes	Probability	Observed counts or freqs	
A	*AA* + *AO*	$p^2 + 2pr$	$n_A = 44$	0.26994
B	*BB* + *BO*	$q^2 + 2qr$	$n_B = 27$	0.16564
AB	*AB*	$2pq$	$n_{AB} = 4$	0.02454
O	*OO*	r^2	$n_O = 88$	0.53988

$r = 1 - p - q$

$\mathbf{n} = \{n_A, n_B, n_{AB}, n_O\}$

$$L(p,q;\mathbf{n}) = f(\mathbf{n};p,q)$$
$$= (p^2 + 2pr)^{n_A}(q^2 + 2qr)^{n_B}(2pq)^{n_{AB}}(r^2)^{n_O}$$
$$\ell(p,q;\mathbf{n}) = \log(L)$$

Likelihood: example 2 (ABO blood groups)

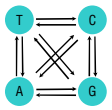


MLEs: $p = 0.1605$, $q = 0.1004$, $r = 1 - p - q = 0.7392$, $\ell = -175.448$
The 95% likelihood (confidence) region can be constructed by cutting the surface at $\ell = -175.448 - 2.995 = -178.443$.
($\chi^2_{2,95\%} = 5.99$)

Example 3: MLE of distance under JC69

The sequence distance is $d = 3\lambda t$, the expected number of substitutions per site. This is related to the proportion of different sites p by

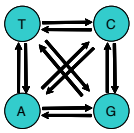
$$p = \tfrac{3}{4}[1 - \exp(-\tfrac{4}{3}d)]$$



Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21-123 *in* H. N. Munro, ed. *Mammalian protein metabolism.* Academic Press, New York.

JC69 model of substitution

$$Q = \{q_{ij}\} = \begin{bmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \alpha \\ \lambda & \lambda & \lambda & -3\lambda \end{bmatrix}$$

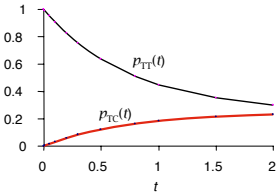


$\lambda\Delta t$ is the probability that given nucleotide T, it will change into C a very short time Δt later.

$$P(t) = \{p_{ij}(t)\} = e^{Qt} = \begin{bmatrix} \frac{1}{4}+\frac{3}{4}e^{-4\lambda t} & \frac{1}{4}-\frac{1}{4}e^{-4\lambda t} & \frac{1}{4}-\frac{1}{4}e^{-4\lambda t} & \frac{1}{4}-\frac{1}{4}e^{-4\lambda t} \\ \frac{1}{4}-\frac{1}{4}e^{-4\lambda t} & \frac{1}{4}+\frac{3}{4}e^{-4\lambda t} & \frac{1}{4}-\frac{1}{4}e^{-4\lambda t} & \frac{1}{4}-\frac{1}{4}e^{-4\lambda t} \\ \frac{1}{4}-\frac{1}{4}e^{-4\lambda t} & \frac{1}{4}-\frac{1}{4}e^{-4\lambda t} & \frac{1}{4}+\frac{3}{4}e^{-4\lambda t} & \frac{1}{4}-\frac{1}{4}e^{-4\lambda t} \\ \frac{1}{4}-\frac{1}{4}e^{-4\lambda t} & \frac{1}{4}-\frac{1}{4}e^{-4\lambda t} & \frac{1}{4}-\frac{1}{4}e^{-4\lambda t} & \frac{1}{4}+\frac{3}{4}e^{-4\lambda t} \end{bmatrix}$$

$p_{TC}(t)$ is the probability that given nucleotide T, it will change into C time t later.

Transition probability under JC69



(i) Suppose a very long sequence has T at every site and suppose all sites in the sequence evolve for a time period t. Then $\{p_{TT}(t), p_{TC}(t), p_{TA}(t), p_{TG}(t)\}$ will give the proportions of nucleotides T, C, A, and G in the sequence.

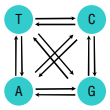
(ii) Whatever the starting nucleotide compositions, the proportions will approach ¼ when $t \to \infty$.

(iii) If the nucleotides are in proportions (¼, ¼, ¼, ¼), the proportions won't change anymore. The chain is said to be *stationary* or *in equilibrium*.

Example 3: MLE of distance under JC69

Data: $x = 10$ out of $n = 100$ sites are different.
Distance $d = 3\lambda t$, related to the proportion of different sites p by

$$p = \tfrac{3}{4}[1-\exp(-\tfrac{4}{3}d)]$$

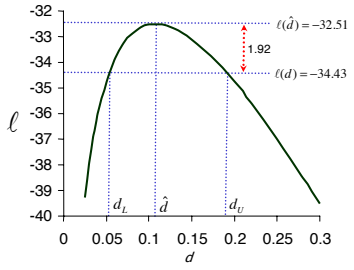


$$L(d;x) = f(x;d) = p^x(1-p)^{n-x}$$

$$\hat{d} = -\tfrac{3}{4}\log(1-\tfrac{4}{3}\times\tfrac{x}{n})$$

MLE and likelihood interval

Data:
$x = 10$ differences
$n = 100$ sites

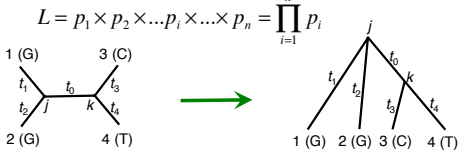


$(\chi^2_{1,95\%} = 3.84)$

Likelihood calculation on tree

Site	1	2	3	4	5	...	***i***	...			*n*
Sequence 1	C	T	C	A	T	...	**G**	...	G	T	A	A	T
Sequence 2	C	T	A	G	T	...	**G**	...	C	T	A	G	T
Sequence 3	C	T	A	G	T	...	**C**	...	G	T	A	G	T
Sequence 4	C	C	A	A	C	...	**T**	...	C	C	A	A	T
Probability	p_1	p_2				...	p_i	...					p_n

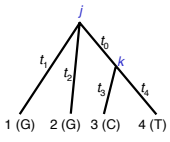
$$L = p_1 \times p_2 \times ... p_i \times ... \times p_n = \prod_{i=1}^{n} p_i$$



The probability of each site is a sum over all possible ancestral states

$$p_i = \Pr\left(\begin{matrix}T\\G\ G\ C\ T\end{matrix}\right) + \Pr\left(\begin{matrix}T\ C\\G\ G\ C\ T\end{matrix}\right) + \Pr\left(\begin{matrix}T\ A\\G\ G\ C\ T\end{matrix}\right) + \ldots + \Pr\left(\begin{matrix}G\ G\\G\ G\ C\ T\end{matrix}\right).$$

$$\Pr\left(\begin{matrix}j\\G\ G\ C\ T\end{matrix}\right) = \pi_j p_{jG}(t_1) p_{jG}(t_2) p_{jk}(t_0) p_{kC}(t_3) p_{kT}(t_4).$$



Likelihood calculation on tree: summary

To sum up, the log likelihood ℓ is a sum of the log probabilities over all sites. The probability at each site p_i is a sum over all ancestral reconstructions. For each ancestral reconstruction, the probability is a product of the transition probabilities over branches.

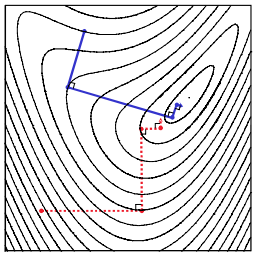
$$\ell(t_0, t_1, t_2, t_3, t_4 \mid X) = \sum_{i=1}^{n} \log(p_i)$$

ℓ is a function of the branch lengths t_0, t_1, t_2, t_3, t_4 (and substitution parameters, if any). We estimate them by maximizing ℓ. The optimum ℓ corresponding to the MLEs of parameters is the score for the tree. We repeat this process for all possible trees. The maximum likelihood tree is the one with the highest score.

Likelihood calculation on tree

Felsenstein (1981 *Journal of Molecular Evolution* 17:368-376) described an algorithm (pruning or peeling algorithm) that makes the likelihood calculation feasible.

To find the maximum likelihood estimates (MLEs), numerical optimization (nonlinear programming) algorithms are often necessary.



Ancestral reconstruction

$$p_i = \Pr\left(\begin{smallmatrix} & T & T \\ G & G & C & T \end{smallmatrix}\right) + \Pr\left(\begin{smallmatrix} & T & C \\ G & G & C & T \end{smallmatrix}\right) + \Pr\left(\begin{smallmatrix} & T & A \\ G & G & C & T \end{smallmatrix}\right) + \ldots + \Pr\left(\begin{smallmatrix} & G & G \\ G & G & C & T \end{smallmatrix}\right).$$

The assignment of states to the internal nodes of the tree (such as TT, TC, …) is called an ancestral reconstruction. The probability of each site p_i is a sum over all possible reconstructions. After the parameters are estimated, the contribution of a reconstruction to p_i gives the posterior probability for the reconstruction.

This likelihood (empirical Bayes method of ancestral reconstruction has 2 advantages over parsimony reconstruction:

(1) It uses branch lengths and relative rates.
(2) It provides a measure of accuracy.

(Yang et al. 1995. Genetics 141:1641-1650)

Ancestral reconstruction can be used to "restore" extinct proteins and to study their biochemical properties.

Pauling, L. and E. Zuckerkandl. 1963. Chemical paleogenetics: molecular "restoration studies" of extinct forms of life. *Acta Chem. Scand.* 17:S9-S16

Chang, et al. 2002. Synthetic gene technology: applications to ancestral gene reconstruction and structure-function studies of receptors. *Methods Enzymol.* **343**:274-294.

Ugalde, et al. 2004. Evolution of coral pigments recreated. *Science* 305:1433

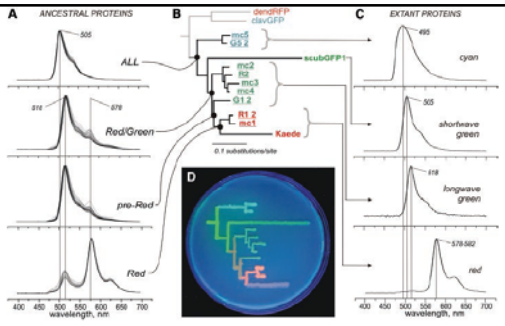


Fig. 1. (A) Fluorescence spectra of the reconstructed ancestral proteins. Multiple curves correspond to clones bearing variations at degenerate sites. (B) Phylogeny of GFP-like proteins from the great star coral *M. cavernosa* (sequence names are underlined) and closely related coral species. The red and cyan proteins from soft corals (dendRFP and clavGFP) represent an outgroup. (C) Fluorescence spectra of extant proteins. (D) Phylogenetic tree of colors from the great star coral, drawn on a petri dish with bacteria expressing extant and ancestral proteins, under ultraviolet light.

Example. ape trees for 896-bp mtDNA under K80

$\kappa = 11.4$
$\ell = -2270.5$

$\kappa = 11.1$
$\ell = -2280.6$

$\kappa = 10.7$
$\ell = -2278.6$

(Data from Brown et al. 1982. J. Mol. Evol. **18**:225-239)

Likelihood versus parsimony

- ML takes into account all ancestral state reconstructions while MP uses the most parsimonious reconstructions.
- ML weights changes differently if they occur on branches of different lengths while MP ignores branch lengths.
- ML weights different kinds of changes differently (such as transitions and transversions) while MP uses equal weighting (except for weighted parsimony).
- All assumptions under ML are explicit while the assumptions underlying MP are poorly understood.
- ML is more efficient and flexible for estimating parameters and testing hypothesis when the tree is known.
- ML is computationally much more expensive than MP.

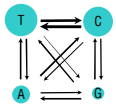
Time reversibility

Almost all models used in molecular phylogenetics, are time reversible. The Markov chain is said to be *time reversible* if and only if

$$\pi_i q_{ij} = \pi_j q_{ji}, \text{ for all } i \neq j.$$

which is the same requirement as

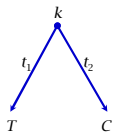
$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \text{ for all } i \neq j.$$



- The amount of flow from T to C equals the amount of flow from C to T: $\pi_T q_{TC} t = \pi_C q_{CT} t$, where $\pi_T q_{TC} t$ is the expected number of changes or "flow" from T to C over any time t.
- Reversibility does not mean symmetrical substitution rates.
- Reversibility is a mathematical convenience (Yang 1994).

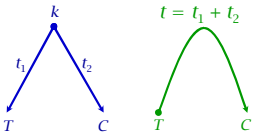
An implication of reversibility

$$\begin{aligned}
&\Pr(TC \mid t_1, t_2) \\
&= \sum_k \pi_k p_{kT}(t_1) p_{kC}(t_2) \\
&= \sum_k \pi_T p_{Tk}(t_1) p_{kC}(t_2) && \leftarrow \text{reversibility} \\
&= \pi_T \sum_k p_{Tk}(t_1) p_{kC}(t_2) \\
&= \pi_T p_{TC}(t_1 + t_2) && \leftarrow \text{Chapman-Kolmogorov theorem}
\end{aligned}$$



An implication of reversibility

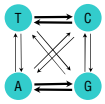
$$\Pr(TC \mid t_1, t_2) = \pi_T p_{TC}(t_1 + t_2)$$



The probability of seeing T and C at the site (or any other pair of nucleotides) is the same whether
(a) the two sequences diverged from a common ancestor and have evolved over times (distances) t_1 and t_2 along the two lineages or
(b) sequence 1 evolved over time $t = t_1 + t_2$ to become sequence 2, or
(c) wherever the root of the tree is.

Under time-reversible models and without assuming the molecular clock (constant rate over time), distance and likelihood methods cannot identify rooted trees. Only unrooted trees are estimated.

Nucleotide substitution models: K80

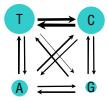


$$Q = \begin{bmatrix} . & \alpha & \beta & \beta \\ \alpha & . & \beta & \beta \\ \beta & \beta & . & \alpha \\ \beta & \beta & \alpha & . \end{bmatrix}$$

Both JC69 and K80 assume symmetrical substitution rates and equal equilibrium base frequencies (¼).

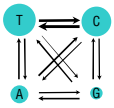
Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**:111-120.

HKY85



$$\begin{bmatrix} . & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & . & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & . & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & . \end{bmatrix}$$

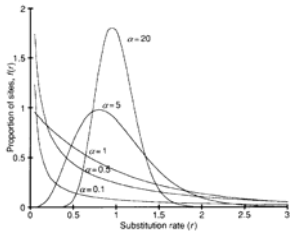
GTR (REV)



$$\begin{bmatrix} . & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & . & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & . & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & . \end{bmatrix}$$

Hasegawa, M., T. Yano, and H. Kishino. 1984. A new molecular clock of mitochondrial DNA and the evolution of Hominoids. Proc. Japan Acad. B. **60**:95-98.
Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160-174.
Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lectures in Mathematics in the Life Sciences **17**:57-86.
Yang, Z. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39**:105-111.

Likelihood models: rate variation among sites



(Yang, Z. 1996. *Trends Ecol. Evol.* **11**:367-372)

Rate variation among sites

- Gamma (Yang, Z. 1993. *Molecular Biology and Evolution* 10:1396-1401)
- Discrete-gamma (Yang, Z. 1994. *Journal of Molecular Evolution* 39:306-314)

Models like HKY85+G, GTR+G, or REV+G, using discrete gamma are in PHYLIP: dnaml & proml, PAUP4*, PAML: baseml & codeml, MrBayes.

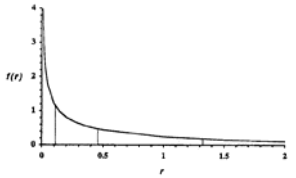


Fig. 1. Discrete approximation to the gamma distribution $G(\alpha,\beta)$, with $\alpha = \beta = 1/2$. Four categories are used to approximate the continuous distribution, with equal probability for each category. The three boundaries are 0.1015, 0.4549, and 1.3233, which are the percentage points corresponding to $p = 1/4, 2/4, 3/4$. The means of the four categories are 0.0334, 0.2519, 0.8203, 2.8944. The medians are 0.0247, 0.2389, 0.7870, 2.3535, and these are scaled to get 0.0291, 0.2807, 0.9248, and 2.7654, so that the mean of the discrete distribution is one.

Different rates for genes or site partitions (such as codon positions)

The site-partition models allow different rates, transition/transversion rate ratios, base compositions etc. for different site partitions (such as genes or codon positions). They should be useful in combined analysis of many genes.

Likelihood implementations of such models are rather primitive. Paup has so-called site-specific rate model, which allows the rates (but not other parameters) to be different among site partitions.

MrBayes uses the link and unlink command to implement flexible models, which allow some parameters of the evolutionary process to be different among site partitions and others to be the same.

(Yang Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J Mol Evol* 42:587-596)

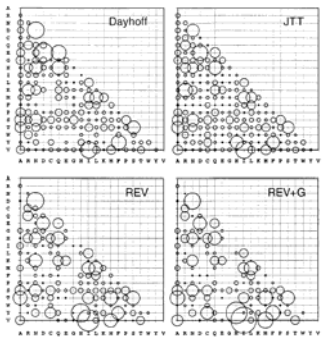
Amino acid and codon substitution models

- Models of amino acid and codon substitutions: calculations are the same as under nucleotide models except that matrices are larger (20×20 or 61×61 instead of 4 × 4) and there are more combinations of ancestral states.
- It is important to account for variable rates for amino acid models.

  Good empirical amino acid models include dayhoff+G, JTT+G, WAG+G, mtREV+G, mtmam+G.

Adachi & Hasegawa 1996. *Journal of Molecular Evolution* **42**:459-468.
Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. Pp. 345-352. *Atlas of protein sequence and structure*, Vol 5, Suppl. 3. National Biomedical Research Foundation, Washington D. C.
Jones, et al. 1992. *CABIOS* 8:275-282.
Whelan & Goldman. 2001. *Molecular Biology and Evolution* 18:691-699.
Yang, et al. 1998. *Mol. Biol. Evol.* 15:1600-1611

Empirical amino acid substitution models



Codon substitution models

Phe F TTT	Ser S TCT	Tyr Y TAT	Cys C TGT
TTC	TCC	TAC	TGC
Leu L TTA	TCA	*** * TAA	*** * TGA
TTG	TCG	TAG	Trp W TGG
Leu L CTT	Pro P CCT	His H CAT	Arg R CGT
CTC	CCC	CAC	CGC
CTA	CCA	Gln Q CAA	CGA
CTG	CCG	CAG	CGG
Thr T ACT	Thr T ACT	Asn N AAT	Ser S AGT
ACC	ACC	AAC	AGC
ACA	ACA	Lys K AAA	Arg R AGA
ACG	ACG	AAG	AGG
Val V GTT	Ala A GCT	Asp D GAT	Gly G GGT
GTC	GCC	GAC	GGC
GTA	GCA	Glu E GAA	GGA
GTG	GCG	GAG	GGG

Goldman & Yang. 1994. *Mol. Biol. Evol* **11**:725-736
Muse & Gaut. 1994. *Mol. Biol. Evol* **11**:715-724

Codon substitution models

- Codon models are natural for studying the selective pressure on the protein. Synonymous and nonsynonymous rates can be compared to detect *adaptive molecular evolution.*
- *Branch models* can be used to test for positive selection on lineages on the tree
- *Site models* can be used to test for positive selection affecting individual sites
- *Branch-site models* attempt to detect positive selection affecting a few sites on a specific lineage.

Yang, Z. 2002. Inference of selection from multiple species alignments. *Curr. Opinion Genet. Devel.* 12:688-694.
Yang, Z., and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15:496-503.
Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 24:1586-1591.

LRT & model selection (LRT, AIC, BIC, ModelTest)

Model vs. hypothesis

A model represents the background knowledge we take for granted in an analysis. It is usually not our focus of analysis, but the sensitivity (robustness) of our analysis to model assumptions is a concern.

A hypothesis represents a biological theory, which we are interested in testing.

We often use "model" to refer to "hypothesis" (as in null model and alternative model), but it is useful to make distinction.

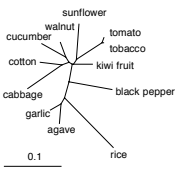
Likelihood ratio test for comparing two nested models

If the more general (alternative) model H_1 has p parameters with log likelihood ℓ_1, and the simpler (null) model H_0 has q parameters with log likelihood ℓ_0. Then twice the log likelihood difference, $2\Delta\ell = 2(\ell_1 - \ell_0)$, can be compared with the χ^2 distribution with d.f. = $p - q$ to test whether the simpler model is rejected.

Likelihood ratio test

Log likelihood values for models fitted to the data of *rbcL* genes from 12 fruits & vegetables

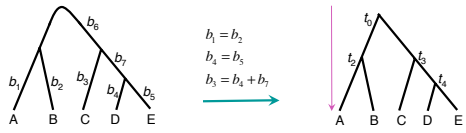
Model	p	ℓ	MLEs
JC69	21	–6,262.01	
K80	22	–6,113.86	$\kappa = 3.561$
HKY85	25	–6,101.76	$\kappa = 3.620$
JC69+G_5	22	–5,937.80	$\alpha = 0.182$
K80+G_5	23	–5,775.40	$\kappa = 4.191$, $\alpha = 0.175$
HKY85+G_5	26	–5,764.26	$\kappa = 4.296$, $\alpha = 0.175$



To compare JC69 against K80, one compares $2\Delta\ell = 2(\ell_1 - \ell_0) = 2\times148.15 = 296.3$, with $p < 1\%$.

Likelihood ratio test of the clock

The no-clock model involves $2n - 3$ parameters (the branch lengths in the unrooted tree), while the clock model involves $n - 1$ parameters (the ages of the internal nodes). Twice the log likelihood difference is thus compared with the chi square distribution with d.f. = $n - 2$ to test the clock.



H_1: no clock: $p = 2n - 3 = 7$ $n - 2 = 3$ constraints H_0: clock: $q = n - 1 = 4$

$$AIC = -2\ell + 2p$$
$$BIC = -2\ell + p \log(n)$$

p: number of parameters
n: sample size (number of sites)

Comparison of models for the mitochondrial protein sequences from 7 apes

Model	p	ℓ	LRT	AIC	BIC
DAYHOFF	11	–15,766.72		31,555.44	31,622.66
JTT	11	–15,332.90		30,687.80	30,755.02
MTMAM	11	–14,558.59		29,139.18	29,206.40
DAYHOFF+Γ_5	12	–15,618.32	296.80	31,260.64	31,333.97
JTT+ Γ_5	12	–15,192.69	280.42	30,409.38	30,482.71
MTMAM+Γ_5	12	–14,411.90	293.38	28,847.80	28,921.13

Cao, Y., *et al.* 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *J. Mol. Evol.* **47**:307-322.

Model selection and ModelTest

MODELTEST is a program for selecting the model of nucleotide substitution that best fits your data. The program chooses among 56 models, and implements three different model selection frameworks: hierarchical likelihood ratio tests (hLRTs), Akaike information criterion (AIC), and Bayesian information criterion (BIC). The program also implements the assessment of model uncertainty and tools for model averaging and calculation of parameter importance, using the AIC or the BIC.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**:817-818.

What if you don't want to use ModelTest as the referees/editors tell you to?

We note that in the literature, simple-minded use of LRT and AIC for model selection (Posada and Crandall, 1998) almost invariably led to overly complex models such as GTR+I+G. We warn against such a practice, as such parameter-rich models may not produce more reliable phylogenies. Besides the fit of the model to data, one should also consider the biological interpretations of the models and the robustness of the analysis to model assumptions...

Ren, F., H. Tanaka, and Z. Yang. 2005. An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Systematic Biology* **54**:808-818.

ML phylogenetic programs

Phylip: dnaml, dnamlk, proml (Felsenstein)
Molphy: nucml, protml (Adachi and Hasegawa 1996)
paup (Swofford)
paml (baseml & codeml) (Yang)

phyml (Guindon & Gascuel)
Raxml (Stamatakis)

Books on statistics & likelihood

DeGroot, M. H., and M. J. Schervish. 2002. *Probability and Statistics.* Addison-Wesley, Boston, USA.
Edwards AWF. 1992. *Likelihood.* John Hopkins University Press, London.
Yang, Z. 2006. *Computational Molecular Evolution.* Oxford University Press. Chapters 4.