7. Hypothesis testing and model selection

EEOB563

Spring 2020

"All models are wrong but some are useful" Box, 1976

1 Preliminary considerations

In ML we choose the hypothesis that maximizes the probability of the data [*e.g.*, $L(H|D) = P(D|H)$]. For the phylogenetic analysis of aligned sequences, our hypothesis consists of two components: a phylogenetic tree and a description of the way individual sequences evolve by nucleotide or amino acid replacement along the branches of that tree. These replacements are usually described as the products of chance mutation events, and their occurrence at each sequence site is mathematically modeled by a Markov process. Typically, the same Markov process model is applied to each sequence site (the IID assumption). It is generally agreed that more complex models, often describing replacement rates in terms of a variety of biological phenomena, give a statistically better fit to observed patterns of sequence evolution and more accurate and robust estimates both of phylogeny and the statistical confidence in the phylogeny (Fig. 1)

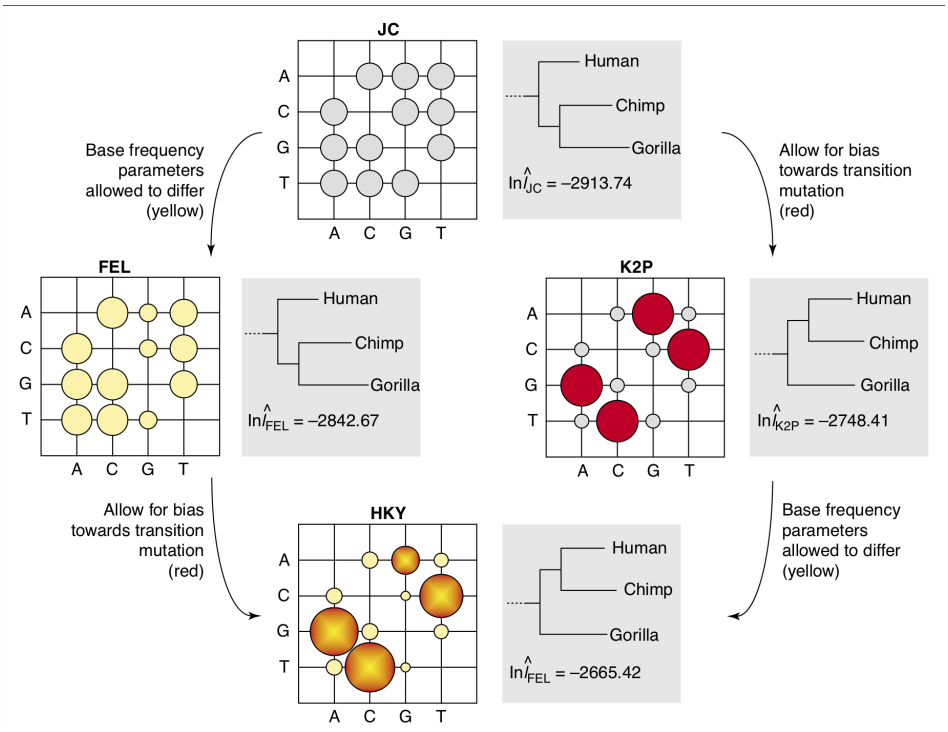

Figure 1: Change in likelihood value and inferred phylogeny as the result of using a more complex model of nucleotide substitutions. From: Whelan et. al. 2001, TIG 17:262

There are two approaches to building models of sequence evolution. One approach is to build models **empirically** using properties calculated through comparisons of large numbers of observed

sequences. Empirical models result in fixed parameter values, which are estimated only once and then assumed to be applicable to all datasets. The alternative approach is for models to be built **parametrically** on the basis of the chemical or biological properties of DNA and amino acids. Parametric models allow the parameter values to be derived from the dataset in each particular analysis. All DNA substitution models we used so far are parametric models. Both methods result in Markov process models, defined by matrices containing the relative rates of occurrence of all possible replacements. From these are calculated the probabilities of change from any nucleotide to any other nucleotide (or any amino acid to any other amino acid) over any period of evolutionary time (e.g. from one end of a branch to the other) at any site. However, one question that needs to be addressed is how to choose the "best" model for an analysis.

2 Likelihood Ratio tests

When we use a model to approximate reality, we have to make a tradeoff between bias (distance between the average estimate and truth) and variance (spread around the truth). When we use a more complex model, we decrease bias, but increase variance. We can use several tests to find a happy compromise between the two.

Likelihood provides a natural means of hypothesis testing. The Likelihood Ratio Test (LRT) statistic for comparing two hypotheses (Λ) is defined as

$$\Lambda = \frac{\max[L(\text{null hypothesis}|\text{data})]}{\max[L(\text{alternative hypothesis}|\text{data})]}$$

When nested hypotheses are examined, Λ will always be < 1 and $-2log\Lambda$ is approximately χ^2 distributed under the null hypothesis with q degrees of freedom, where q is the difference in the number of free parameters between the null and alternative hypotheses. We can use LRT to select a model of sequence evolution. Here is a table showing parameters for some of these models:

Model	Rate params	Base frequencies	# of free params	Reference
JC	$a=b=c=d=e=f$	$\pi_A=\pi_C=\pi_G=\pi_T$	1	Jukes & Cantor, 1969
K80, K2P	$a=c=d=f, b=e$	$\pi_A=\pi_C=\pi_G=\pi_T$	2	Kimura, 1980
TrNef	$a=c=d=f, b, e$	$\pi_A=\pi_C=\pi_G=\pi_T$	3	Tamura & Nei, 1993
K81, K3ST	$a=f, b=e, c=d$	$\pi_A=\pi_C=\pi_G=\pi_T$	3	Kimura, 1981
SYM	a, b, c, d, e, f	$\pi_A=\pi_C=\pi_G=\pi_T$	6	Zharkikh, 1994
F81	$a=b=c=d=e=f$	$\pi_A, \pi_C, \pi_G, \pi_T$	4	Felsenstein, 1981
HKY	$a=c=d=f, b, e$	$\pi_A, \pi_C, \pi_G, \pi_T$	5	Hasegawa et al., 1985
TrN	$a=c=d=f, b, e$	$\pi_A, \pi_C, \pi_G, \pi_T$	6	Tamura & Nei, 1993
GTR, REV	a, b, c, d, e, f	$\pi_A, \pi_C, \pi_G, \pi_T$	9	Lanave et al., 1984; Rodrigues et al., 1990

There are many other questions that can be tested with LRT (Fig. 2)

Table 1. Biological questions involving phylogeny that have been addressed using LRTs.

Question	Assumptions	Results
Are DNA substitution rates constant among lineages [that is, does a molecular clock exist (*10*)]?	H_0: Assume that DNA substitution rates are equal among lineages. H_1: Allow substitution rates to vary among lineages.	A molecular clock is most often rejected, suggesting that there is rate variation among lineages.
Is a DNA substitution model adequate to explain the data (*16*)?	H_0: Assume a particular model of DNA substitution. H_1: Assume a multinomial distribution for the frequencies of site patterns.	Current models of DNA substitution fit the observed data poorly. Sequences from pseudogenes show the best fit.
Are DNA substitution rates biased for different nucleotides (*16*)?	H_0: Assume that substitution rates are equal among nucleotides (for example, the transition rate equals the transversion rate). H_1: Allow transition rate–transversion rate bias.	The addition of unequal rate parameters to the substitution matrix usually provides an improved fit of the model.
Are DNA substitution rates constant among sites (*27*)?	H_0: Assume equal rates among sites. H_1: Allow among-site rate heterogeneity.	The addition of parameters allowing among-site rate variation typically provides a significant improvement to the fit of the model.
Are DNA substitution rates constant among genomic regions [that is, in different genes or different codon positions (*21*)]?	H_0: Assume that substitution rates are the same in all data partitions (regions). H_1: Assume an independent substitution rate for each partition (region).	Rates vary significantly among genomic regions (for example, at different codon positions).
Is the DNA substitution process identical among lineages (*22*)?	H_0: Assume a homogeneous substitution process among lineages. H_1: Allow parameters of the substitution model to vary among lineages.	Base frequencies and the transition rate–transversion rate bias varied significantly among four of the major lineages that gave rise to present-day life forms (*22*).
Are the substitutions in stem regions of ribosomal DNA sequences correlated (*34*)?	H_0: Assume that substitution is independent among sites. H_1: Allow correlated changes in nucleotide duplets in stem regions.	A model that allows for correlated substitutions at pair-bonded stem sites of ribosomal DNA sequences provides an improved fit of the model (*34*).
Is the DNA substitution process identical among genomic regions (*21*)?	H_0: Assume that the substitution parameters are the same among genomic regions. H_1: Allow substitution parameters to vary among genomic regions.	Base frequencies and transition rate–transversion rate bias significantly varied in first, second, third, and transfer RNA partitions of mitochondrial data (*21*).
Is a prespecified taxonomic group monophyletic (*35*)?	H_0: Assume that a group is monophyletic. H_1: Relax the constraint of monophyly.	Analysis of partial HIV sequences from the patients of a dentist supported the idea of multiple sources of infection for one of the patients (*39*).
Are phylogenies estimated from different data congruent (*31*)?	H_0: Assume that the same phylogeny underlies all data partitions. H_1: Allow different phylogenies to underlie different data partitions.	This test has not been widely applied.
Are the phylogenies for hosts and parasites consistent with a common history (*25*)?	H_0: Assume an identical phylogeny for associated hosts and parasites. H_1: Allow different phylogenies for hosts and parasites.	For 13 species of gophers and their associated lice, the phylogenies appear different; for a subset of these species, the phylogeny of hosts and parasites appears identical (*25*).
Are the speciation times for hosts and parasites the same (*25*)?	H_0: Assume that hosts and associated parasites speciated at the same time. H_1: Allow speciation times to vary independently in hosts and parasites.	For five species of cospeciating gophers and lice, the speciation times appear to be identical (*25*).

Figure 2: Biological questions that have been addressed using LRTs. From Huelsenbeck and Rannala 1997, Science 276:227.

3 Other methods of hypotheses testing

While hLRTs are among the most popular methods for model selection, they have several limitations: 1) assume that one of the models is correct 2) when the models are not nested the χ^2 approximation is no longer valid 3) designed to test hypothesis; hypothesis testing & model selection are not equivalent + problems with the step-wise procedure

Approaches like Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

offer important advantages.

3.1 Akaike Information Criterion (AIC)

The AIC (originally "an information criterion" by Hirotugu Akaike) is an asymptotically unbiased estimation of the expected relative Kullback-Leibler information quantity, which represents the amount of information lost when we use model g to approximate model f. Designed to estimate the predictive accuracy of competing hypotheses.

The AIC for a given model is a function of its maximized log-likelihood (l) and the number of estimable parameters (K):

$$AIC = -2l + 2K$$

We can think of it as the amount of information lost when we use a model to approximate the real process of evolution. Hence, prefer the model with smallest AIC.

When n is small compared to K ($n/K < 40$), where n is the sample size, the use of a second-order AIC, AICc is recommended:

$$AIC = -2l + 2K + \frac{2K(K+1)n}{n-k-1}$$

Akaike suggested that the approximates the relative likelihood of the models given the data, which can be normalized to obtain a set of Akaike weights (w). Can use them to establish 95% confidence set of models.

Can use Akaike weights to obtain a model-averaged estimate of any parameter (including phylogenies).

3.2 Bayesian framework

Model selection is an integral part of Bayesian estimation; different strategies exist

Bayes Factors (analogue of the LRT). Contrast evidence provided by the data for two competing models:

$$B_{ij} = \frac{P(D|M_i)}{P(D|M_j)}$$

Compare model likelihoods rather than joint likelihood. Very strong if $B_{ij} > 150$; Strong if $12 < B_{ij} < 150$; barely worth mentioning if $1 < B_{ij} < 3$.

3.2.1 Bayesian information criterion (BIC)

The BIC for a given model is a function of its maximized log-likelihood (l), the number of estimable parameters (K), and the sample size (n), approximated by # of characters in the alignment:

$$BIC = -2l + Klog(n)$$

Given equal priors, choosing a model with the smallest BIC is equivalent to selecting the model with the max. posterior probability.

Note, there there is a fundamental difference in the way we calculate likelihood (P(D—M)) for Bayes Factors and BIC. In the former, it is calculated by integrating across parameters of the model, rather than by fixing them at the ML values.

4 Is model testing necessary?

In a recent study Abadi et al. suggested that although incongruence regarding the selected model is frequent, choosing one criterion over another is not crucial when topologies and ancestral sequence reconstruction are the desired output. However, these results have been challenged here.