Model Selection Using Model Test

Laura Salter Kubatko
Departments of Statistics and
Evolution, Ecology, and Organismal Biology
The Ohio State University
lkubatko@stat.ohio-state.edu

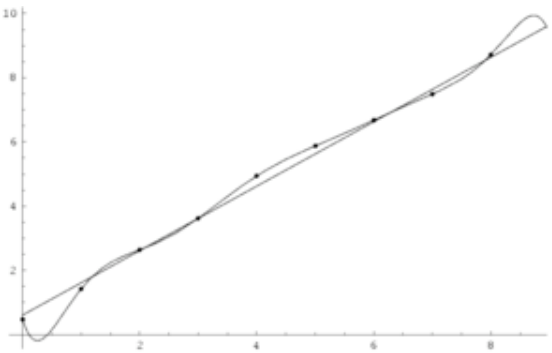
April 7, 2010

Model Selection

- The likelihood framework allows us to assess the "fit" of models to a particular data set.
- The goal is to find a model that is complex enough to capture the processes at work in the data without overfitting.
- There is a trade-off between bias and variance here – by adding parameters to a model we obtain an improvement in fit, but parameter estimates become "worse" because there are more parameters to estimate using a fixed amount of data.

Model Selection

- The following figure illustrates the issue for the simpler problem of fitting a curve to a data set:



- The points are the data.
- Two models are fit – a straight line and a polynomial.
- The polynomial passes through every data point and has more parameters.
- The line has fewer parameters and avoids modeling unlikely fluctuations in the extremes of the data.

Model Selection

- Several criteria are commonly used for model selection
  - Likelihood ratio tests
  - Aikaike information criterion (AIC)
  - Bayesian information criterion (BIC)
  - Other possibilities
- These all use the likelihood function in some way.

Likelihood Ratio Test (LRT)

- The LRT can be applied to compare nested models – pairs of models for which one is a special case of the other.
- The test statistic is $\Delta = 2(lnL_1 - lnL_0)$, where lnL_0 is the maximum log likelihood under the null model and lnL_1 is the maximum log likelihood under the alternative model.
- If the null model can be viewed as a special case of the alternative model, then statistical theory allows use of the χ^2 distribution to compute a p-value.
- Simulation can be used to compare non-nested models.

AIC and BIC

- Both of these criteria use the value of the likelihood function (larger likelihoods mean better fit of the model to the data), but include a penalty for using more parameter-rich models.
- The AIC is

$$AIC = -2lnL + 2p$$

where p is the number of parameters in the model being considered.

- The BIC is

$$BIC = -2lnL + p \times log(n)$$

where n is the number of sites in the sequence.

- Other Bayesian approaches are possible – see Posada and Buckley, *Syst. Biol.* 53: 793-808, 2004.

Modeltest

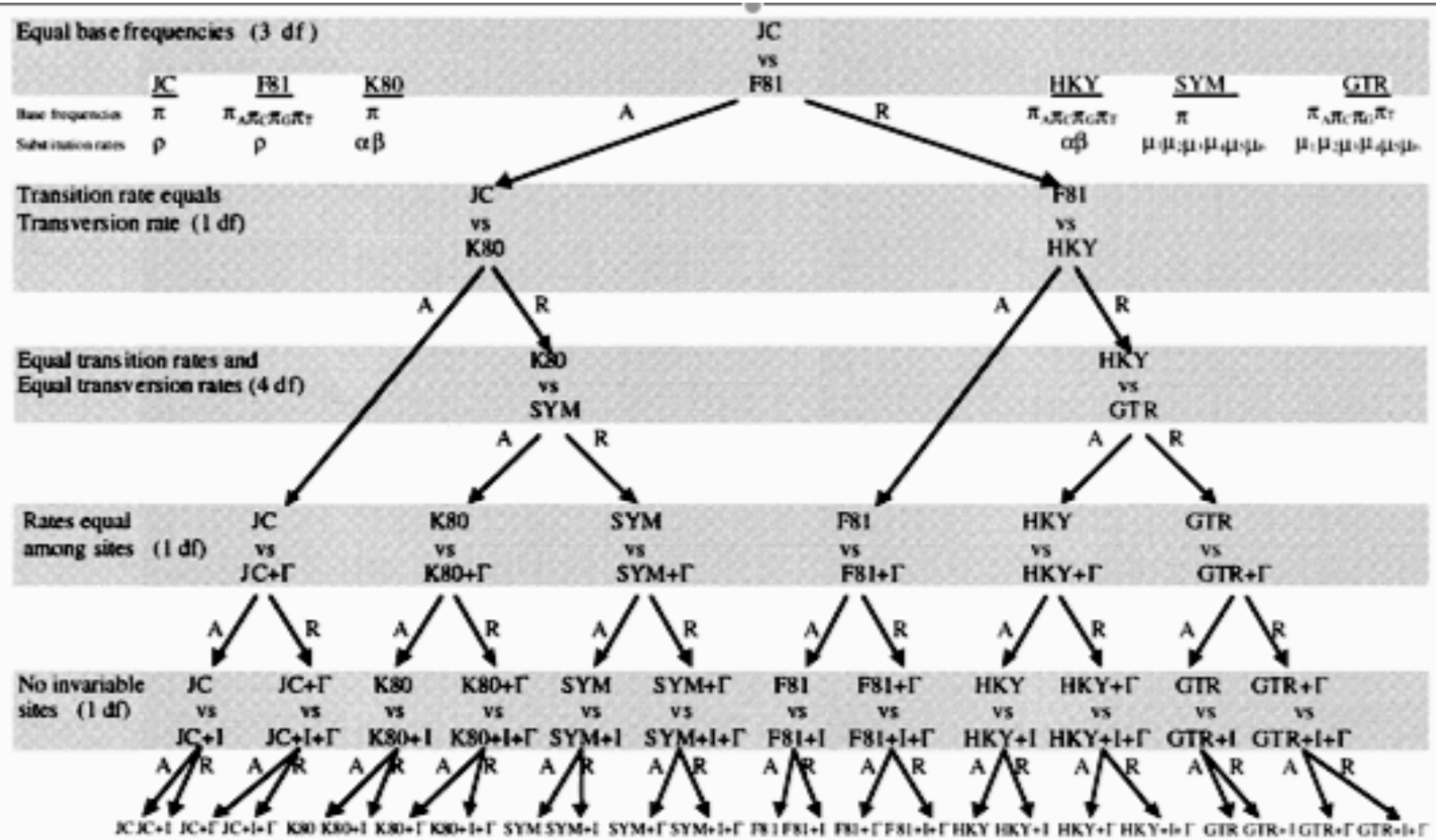


Fig. 1. Hierarchical hypothesis testing in MODELTEST. At each level the null hypothesis (upper model) is either accepted (A) or rejected (R). The models of DNA substitution are: JC (Jukes and Cantor, 1969), K80 (Kimura, 1980), SYM (Zharkikh, 1994), F81 (Felsenstein, 1981), HKY (Hasegawa *et al.*, 1985), and GTR (Rodríguez *et al.*, 1990). Γ: shape parameter of the gamma distribution; I: proportion of invariable sites. df: degrees of freedom. 1: equal base frequencies (0.25), π_A: frequency of adenine, π_C: frequency of cytosine, π_G: frequency of guanine, π_T: frequency of thymine. ρ: equal substitution rate, α: transition rate, β: transversion rate; μ_1: A⇒C rate, μ_2: A⇒G rate, μ_3: A⇒T rate, μ_4: C⇒G rate, μ_5: C⇒T rate, μ_6: G⇒T rate.