Model Selection

GENE 606/ENTO 606/ WFSC 646

Updated 02/20/12

Suggested readings (Older)

- Ripplinger, J., and J. Sullivan. 2008. Does choice in model selection affect maximum likelihood analysis? Systematic Biology 57:76-85.
- Abdo, Z. 2005. Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation. Molecular Biology and Evolution 22: 691-703
- Sullivan, J., Z. Abdo, P. Joyce and D.L. Swofford (2005). Evaluating the performance of a successive approximations approach to parameter optimization in maximum-likelihood phylogeny estimation. Molecular Biology and Evolution 22(6): 1386-1392.
- Sullivan, J. and P. Joyce (2005). Model Selection in Phylogenetics. Annual Review of Ecology and Systematics 36: 445-466.
- Minin V. 2003. Performance-based selection of likelihood models for phylogeny estimation Systematic Biology 52: 674-683
- Posada, D. and T.R. Buckley (2004) Model selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. Systematic Biology 53(5): 793-808.
- Posada D (2008) jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution* **25**, 1253-1256.

Suggested readings (Newer)

- Kedzierska AM, Drton M, Guigó R, Casanellas M (2011) SPIn: Model Selection for Phylogenetic Mixtures via Linear Invariants. Mol Biol Evol. doi:10.1093/molbev/msr259
- Thi Nguyen MA, Gesell T, von Haeseler A (2012) ImOSM: Intermittent Evolution and Robustness of Phylogenetic Methods. Mol Biol Evol 29:663-673
- Thi Nguyen MA, Klaere S, von Haeseler A (2011) MISFITS: Evaluating the Goodness of Fit between a Phylogenetic Model and an Alignment. Mol Biol Evol 28:143-152

Review of models

- **Reversible** (GTR family): rate matrix, among-site rate variation, base frequencies
- **Non-reversible**:
  - accommodate base frequency changes in different parts of tree
  - accommodate changes in rates of sites in different parts of tree
- **Non-independence of sites**
  - Codon models (To cover in Amino-Acid Analyses Lecture)
  - rRNA models
- **Partitioned models** (To cover in Data Partitions Lecture)
  - Assume different models for each partition (e.g. different genes, codons, stem vs loop)

Why and when is the model important

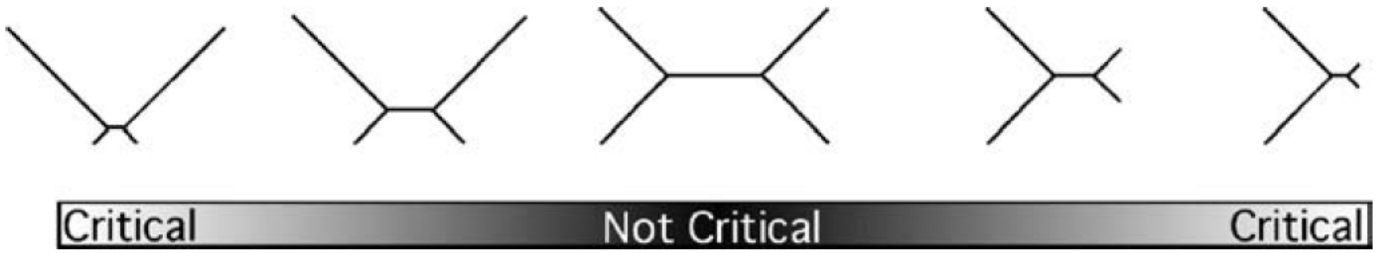


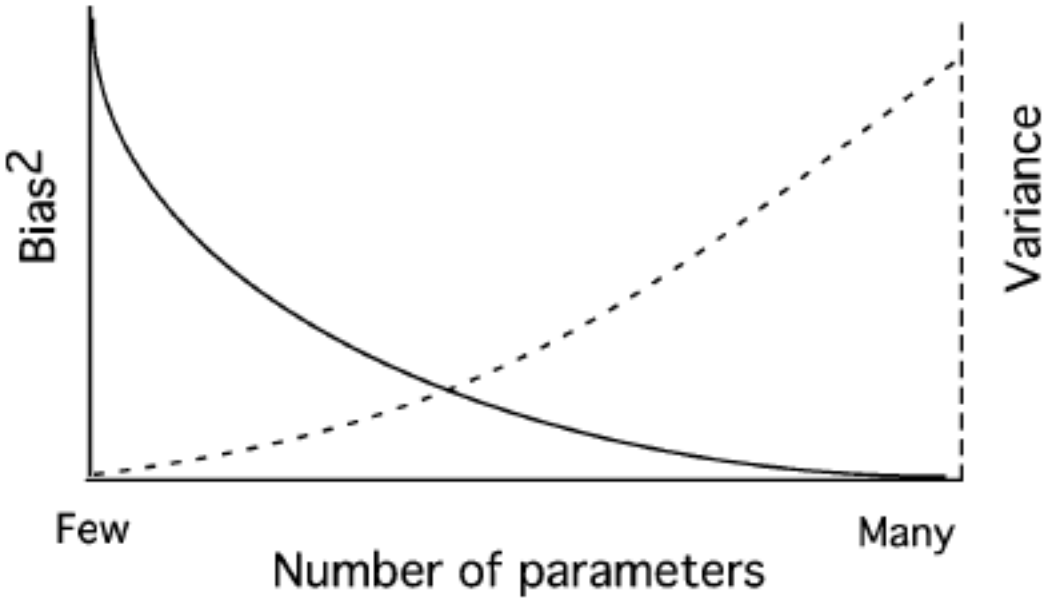
Figure 2 The effect of topology on robustness. At the center of the continuum, phylogenetics signal is strong and model choice is not critical (i.e., maximum likelihood is robust to violations of model assumptions). In the Felsenstein zone (*left*), model selection is critical, as is also the case for the inverse Felsenstein zone (*right*).

From Sullivan & Joyce (2005)

Model Selection

- All models are wrong, but some are useful (Box 1976)
- Model selection is a way of approximating, not identifying, full reality
- Statistical model selection is based on the parsimony principle; hypotheses should be kept as simple as possible
- Increasing the number of parameters will increase the fit between the model and the data (increase the likelihood), but at a cost
- Trade-off between bias and variance:
  - Bias: distance between the average estimate and truth
  - Variance: spread of the estimates around the truth

Bias vs Variance



From Posada & Buckley (2004)

Model Selection Strategies

- Likelihood Ratio Tests (jModeltest)
- Akaike Information Criterion (jModeltest)
- Bayesian Information Criterion (jModeltest)
- Performance-Based model selection (DT-ModSel and jModeltest)
- Bayes factors (“manually”)

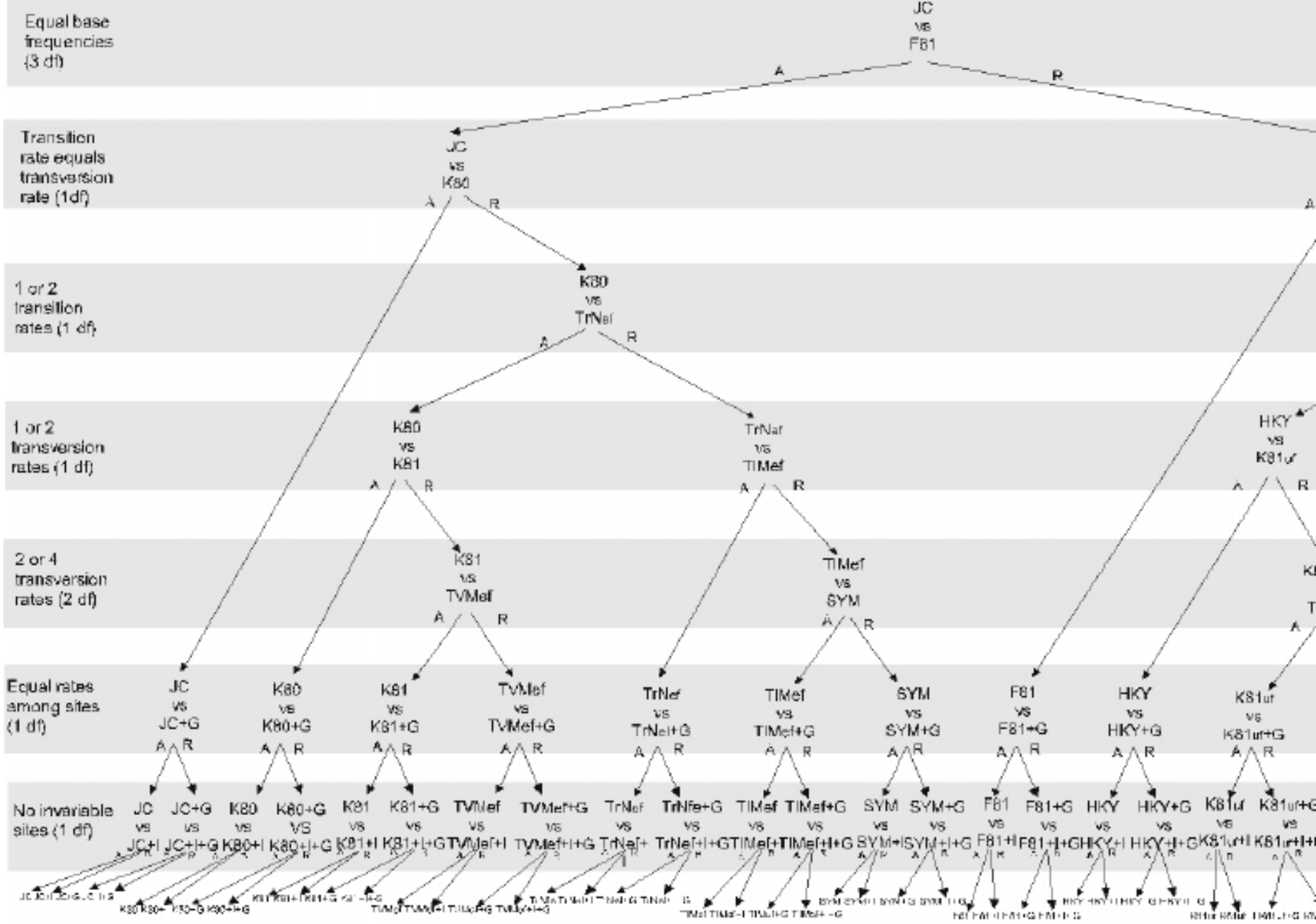
Likelihood Ratio Test (LRT)

$$\delta = 2(\ln L_1 - \ln L_0)$$

- Can **only** be used to evaluate **nested** hypotheses
- L_1 more complex model
- Test statistic evaluated under assumption of asymptotic convergence to X^2 (d.f. = diff in # parameters)

Hierarchical (h)LRT in phylogenetics

1. Infer a phylogenetic tree with another method (parsimony or distance)
2. Estimate the likelihood of that tree under different models of the GTR family (56 models in Modeltest program; 88 models in jModeltest)
3. Conduct LRT in a hierarchical fashion


Equal base frequencies (3 df)
JC vs F81
A
R
Transition rate equals transversion rate (1df)
JC vs K80
A
R
1 or 2 transition rates (1 df)
K80 vs TrNef
A
R
1 or 2 transversion rates (1 df)
K80 vs K81
A
R
TrNef vs TIMef
A
R
HKY vs K81uf
A
R
2 or 4 transversion rates (2 df)
K81 vs TVMef
A
R
TIMef vs SYM
A
R
Equal rates among sites (1 df)
JC vs JC+G
K80 vs K80+G
K81 vs K81+G
TVMef vs TVMef+G
TrNef vs TrNef+G
TIMef vs TIMef+G
SYM vs SYM+G
F81 vs F81+G
HKY vs HKY+G
K81uf vs K81uf+G
No invariable sites (1 df)
JC vs JC+I
JC+G vs JC+I+G
K80 vs K80+I
K80+G vs K80+I+G
K81 vs K81+I
K81+G vs K81+I+G
TVMef vs TVMef+I
TVMef+G vs TVMef+I+G
TrNef vs TrNef+I
TrNef+G vs TrNef+I+G
TIMef vs TIMef+I
TIMef+G vs TIMef+I+G
SYM vs SYM+I
SYM+G vs SYM+I+G
F81 vs F81+I
F81+G vs F81+I+G
HKY vs HKY+I
HKY+G vs HKY+I+G
K81uf vs K81uf+I

Potential weaknessess of hLRT

- Dependence on initial estimate of topology
  - Use of initial trees has little effect on model chosen, but
  - Very poor trees can yield very poor model estimates
- Arbitrary order of comparison can have effect on which model is selected
- Can only compare nested hypotheses

Akaike Information Criteria (AIC)

$$\mathrm{AIC}_i = -2 \ln L_i + 2k_i$$

- For a particular model *i*:
  - L_i = max log likelihood
  - k = number of parameters
- prefer the model with the smallest AIC
- provides a measure of fit between model and data and includes a penalty for overparameterization
- Small sample sizes ($n / k < 40$), use AIC_c

Bayesian Information Criteria (BIC)

$$\mathrm{BIC}_i = -2 \ln Li + k_i \ln n$$

- For a particular model *i*:
  - L_i = max log likelihood
  - *k* = number of parameters
  - *n* = sample size (# of characters ???)
- provides a measure of fit between model and data and penalizes for overparameterization (more heavily than AIC, especially with large *n*)

Performance Based-DT

- ranks models on the basis of the weighted expected error in branch-length estimates
- weights are derived from the BIC
- focuses on the fact that both the tree topology and the branch lengths (the rate of evolution × the time between each node or speciation event in the tree) are critical.
- Minin et al. 2003: incorporates some measure of phylogenetic performance. “Asses models through a penalty or loss function, related to how dissimilar the branch length estimates are across models, and pick the model with the minimum posterior loss”.

Comparison of model selection approaches

- hLRT, AIC, and BIC all use an initial topology
- hLRT can only compare nested hypotheses, while AIC and BIC can compare multiple nested and non-nested hypotheses
- AIC and BIC outcome does not depend on the order of comparisons, while hLRT does
- AIC and BIC allow assessment model selection uncertainty, and estimation of phylogenies and model parameters using all available models (model-averaged inference or multi-model inference).

Comparison of model selection approaches

Table 2
Model Selection Strategies Implemented in jModelTest

	Hierarchical Likelihood Ratio Tests	Dynamical Likelihood Ratio Tests	Akaike Information Criterion	Bayesian Information Criterion	Performance-Based Selection
Abbreviation	hLRTs	dLRTs	AIC	BIC	DT
Base tree	Fixed	Fixed	Fixed, optimized	Fixed, optimized	Fixed, optimized
Nesting requirement	Yes	Yes	No	No	No
Simultaneous comparison	No	No	Yes	Yes	Yes
Selection uncertainty	No	No	Yes	Yes	Yes[a]
Parameter importance	No	No	Yes	Yes	Yes[a]
Model averaging	No	No	Yes	Yes	Yes[a]

[a] DT weights are simply the rescaled reciprocal DT scores. This is a gross implementation very likely to change.

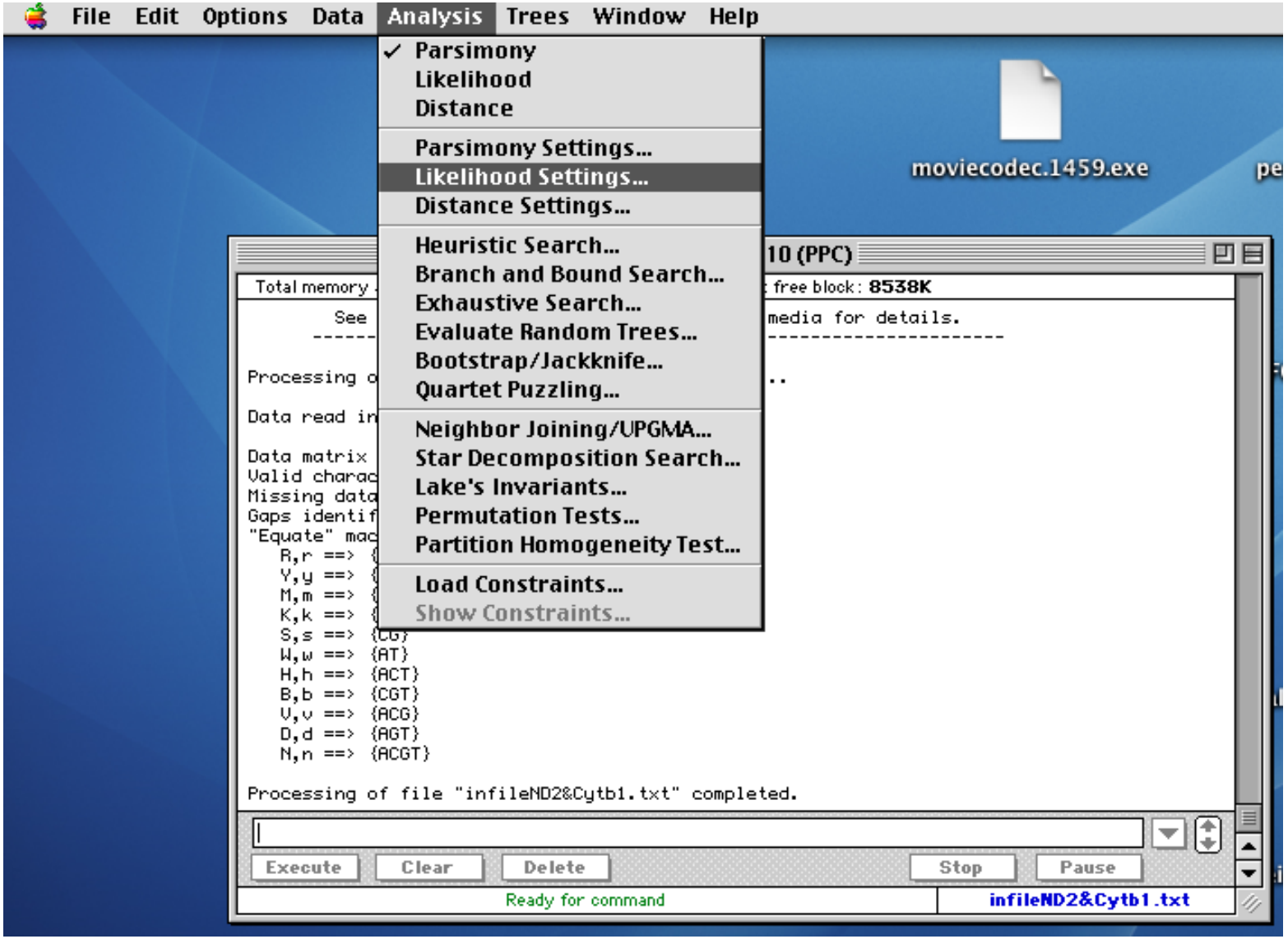
From Posada 2008

Implementation of model selection approaches

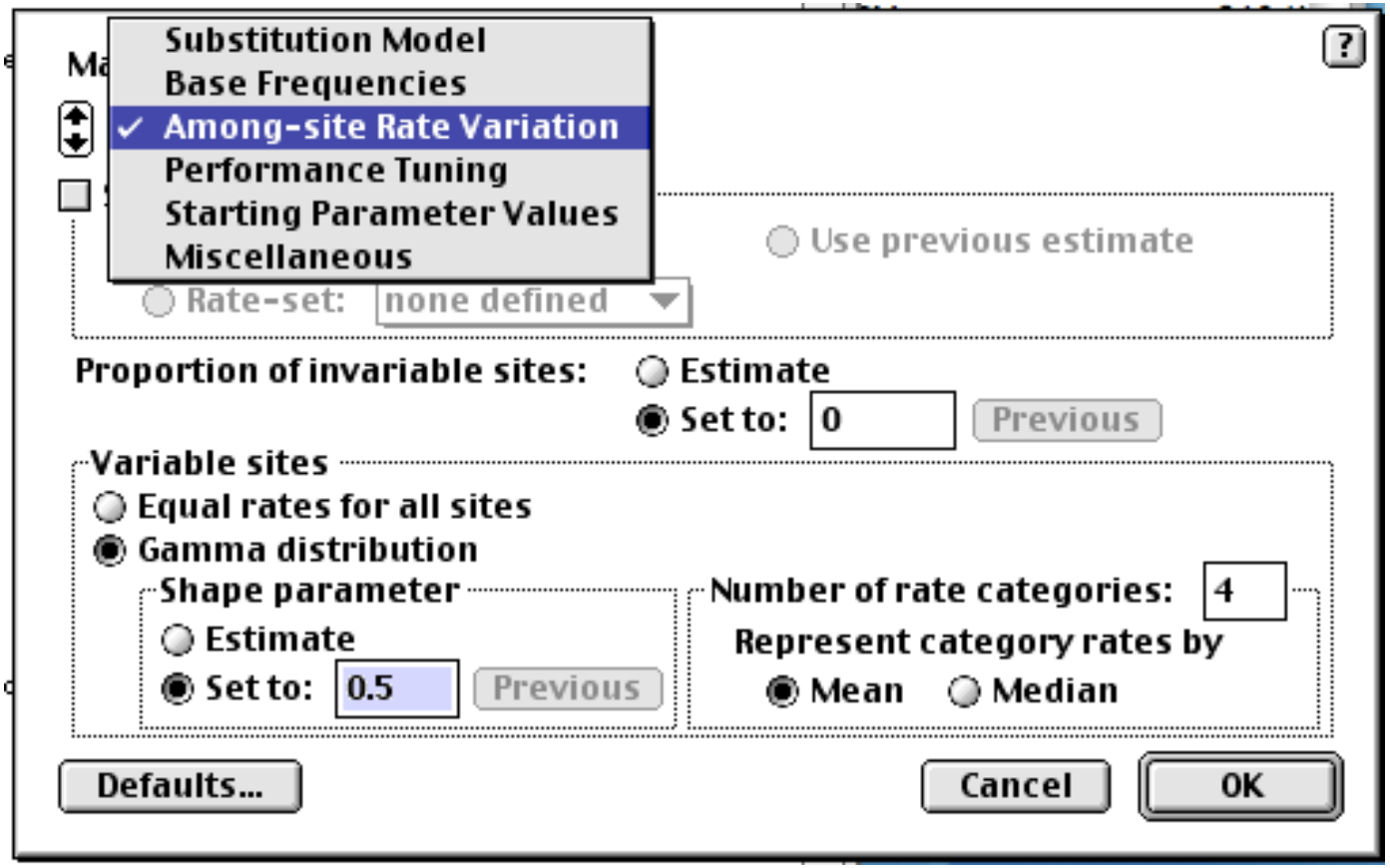
- jModeltest http://darwin.uvigo.es/software/modeltest.html
- A command file for PAUP is already available within the Modeltest package (not used any more)
  - Infers a NJ tree
  - Estimates likelihoods and parameter estimates for each of 56 models (output: model.scores)
  - Use model.scores file as input for Modeltest

- The following slides are images of the GUI windows of Paup* for selecting different models

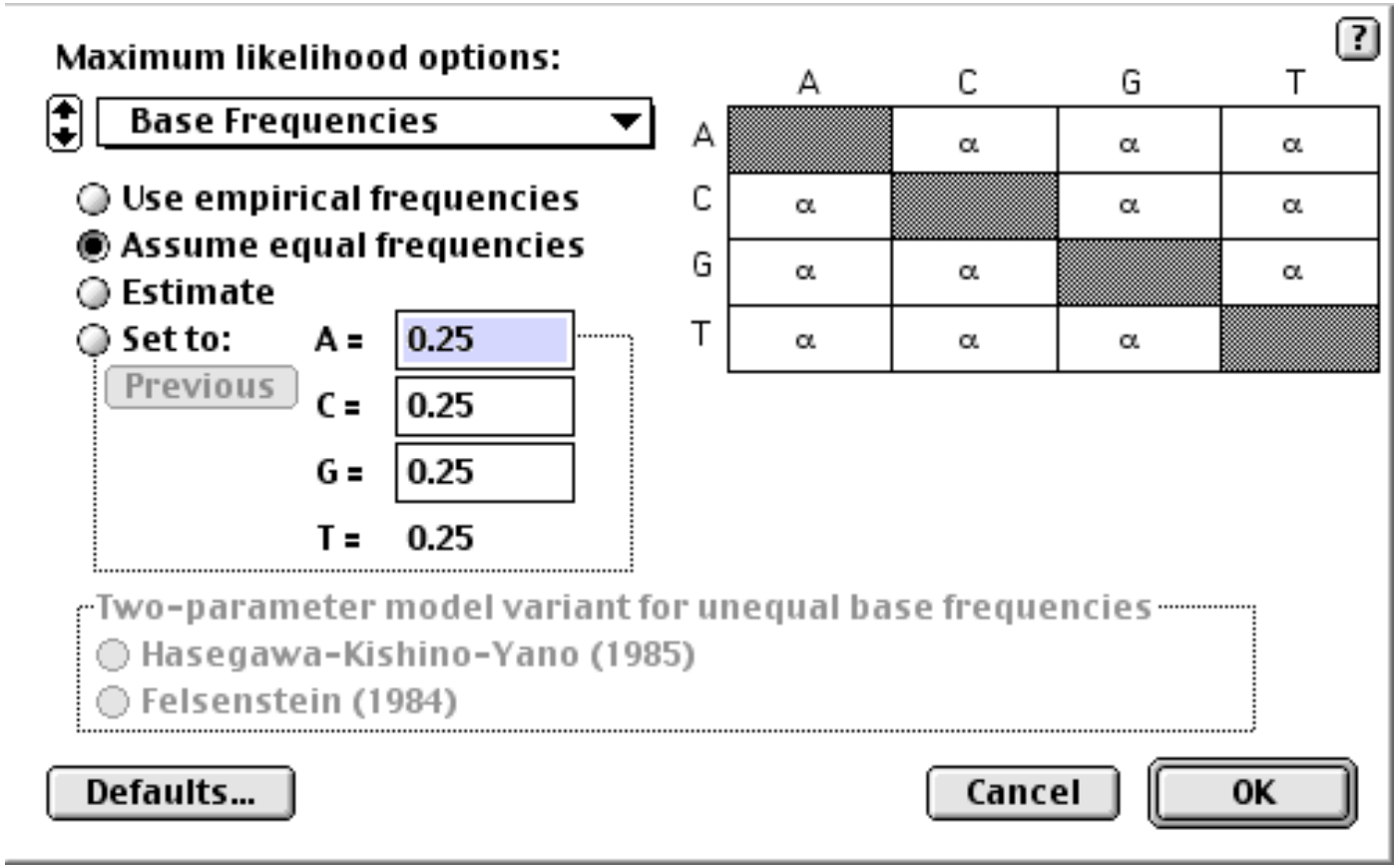
Likelihood Settings



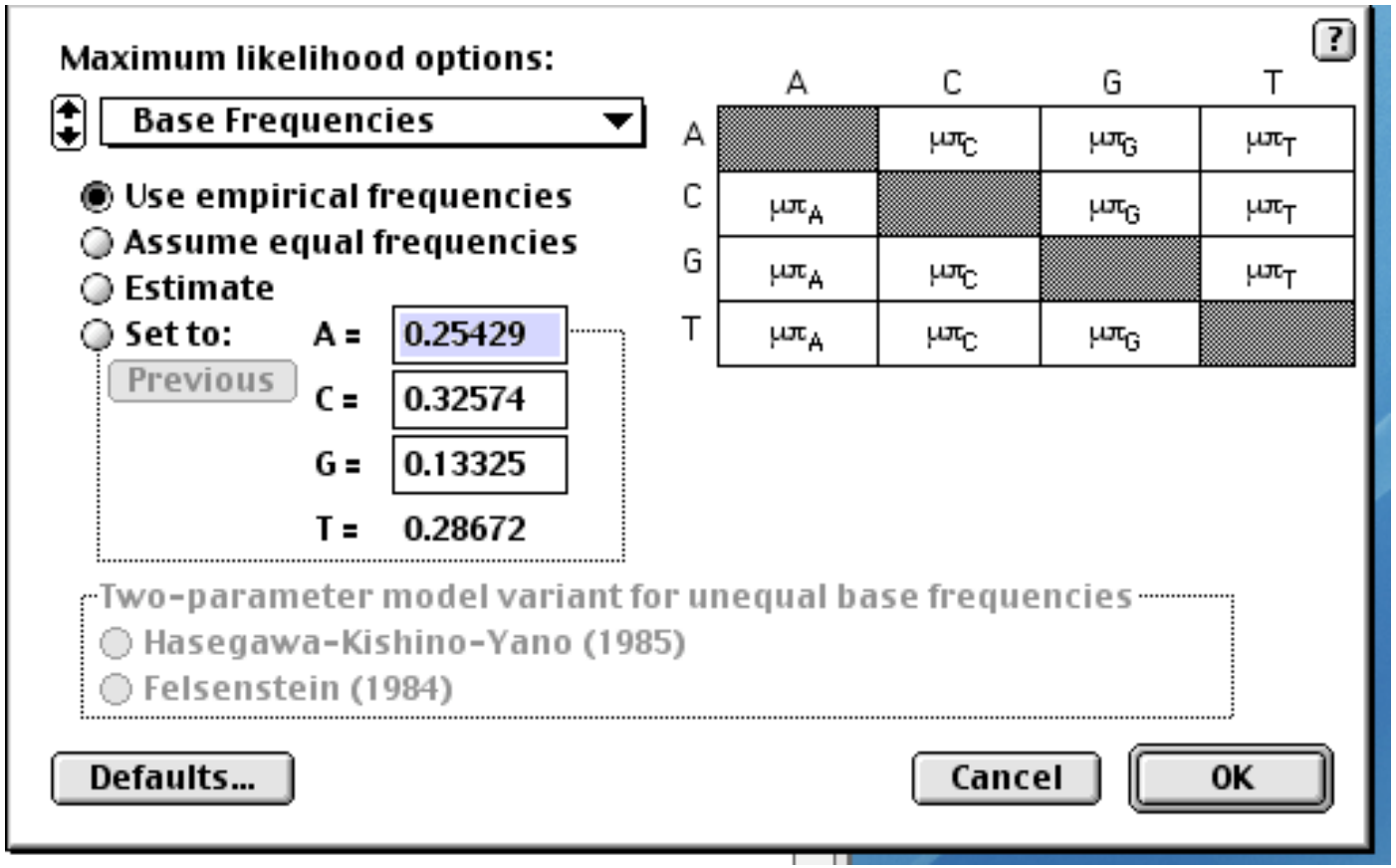
ML settings submenus



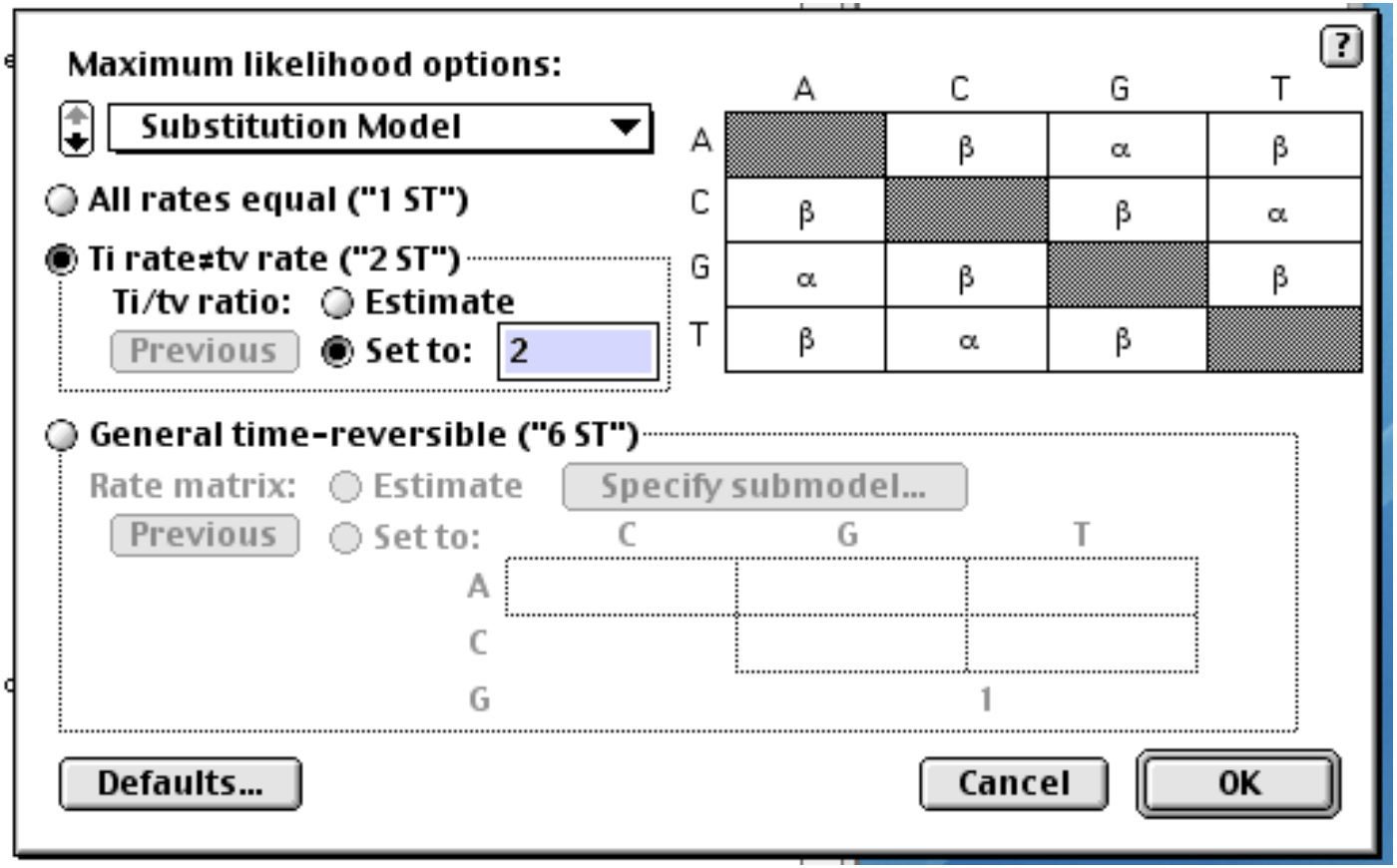
JC


Maximum likelihood options:
Base Frequencies
Use empirical frequencies
Assume equal frequencies
Estimate
Set to:
Previous
A = 0.25
C = 0.25
G = 0.25
T = 0.25
A C G T
A α α α
C α α α
G α α α
T α α α
Two-parameter model variant for unequal base frequencies
Hasegawa-Kishino-Yano (1985)
Felsenstein (1984)
Defaults...
Cancel
OK

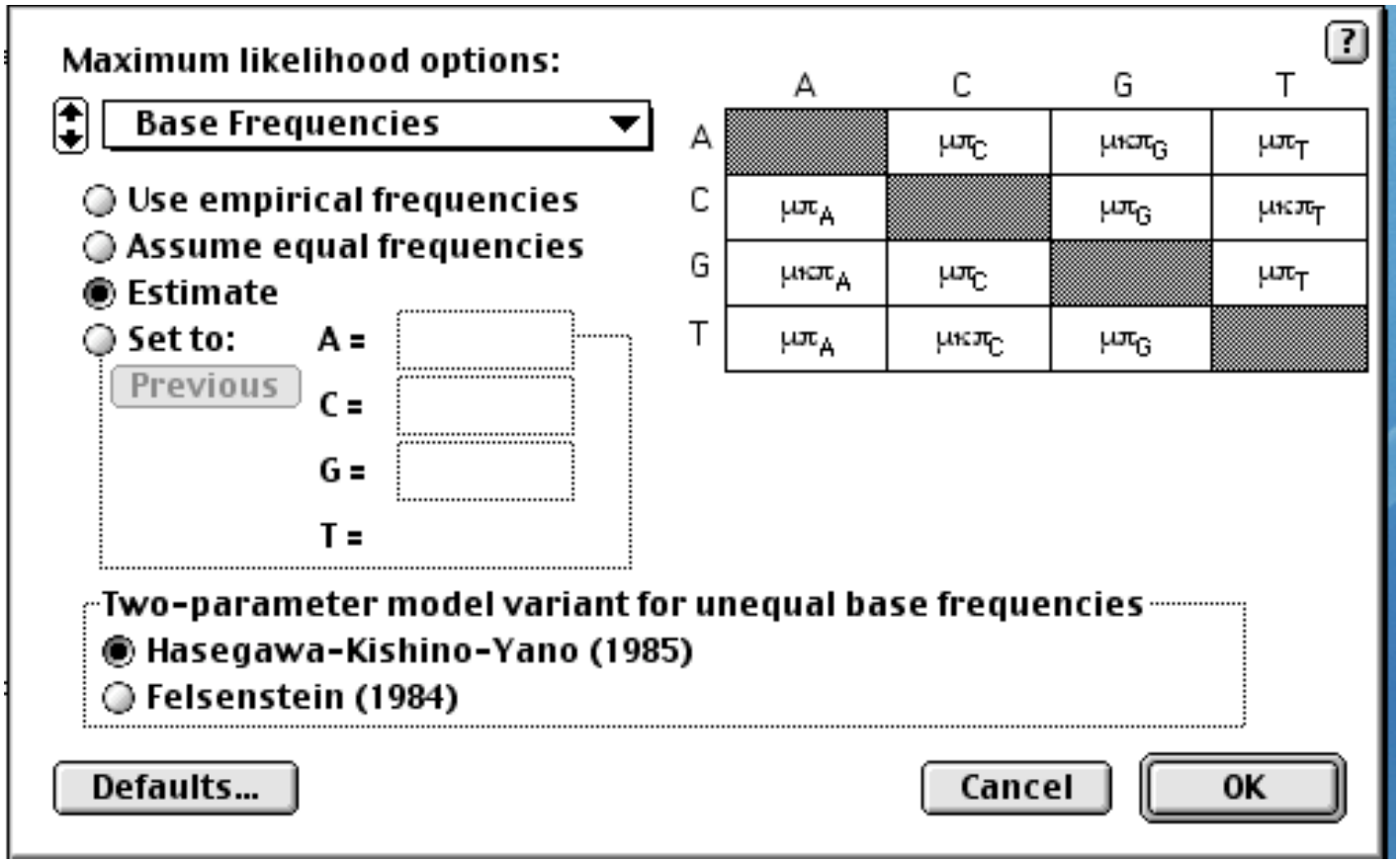
F81


Maximum likelihood options:
Base Frequencies
Use empirical frequencies
Assume equal frequencies
Estimate
Set to:
Previous
A = 0.25429
C = 0.32574
G = 0.13325
T = 0.28672
A C G T
A μπC μπG μπT
C μπA μπG μπT
G μπA μπC μπT
T μπA μπC μπG
Two-parameter model variant for unequal base frequencies
Hasegawa-Kishino-Yano (1985)
Felsenstein (1984)
Defaults...
Cancel
OK

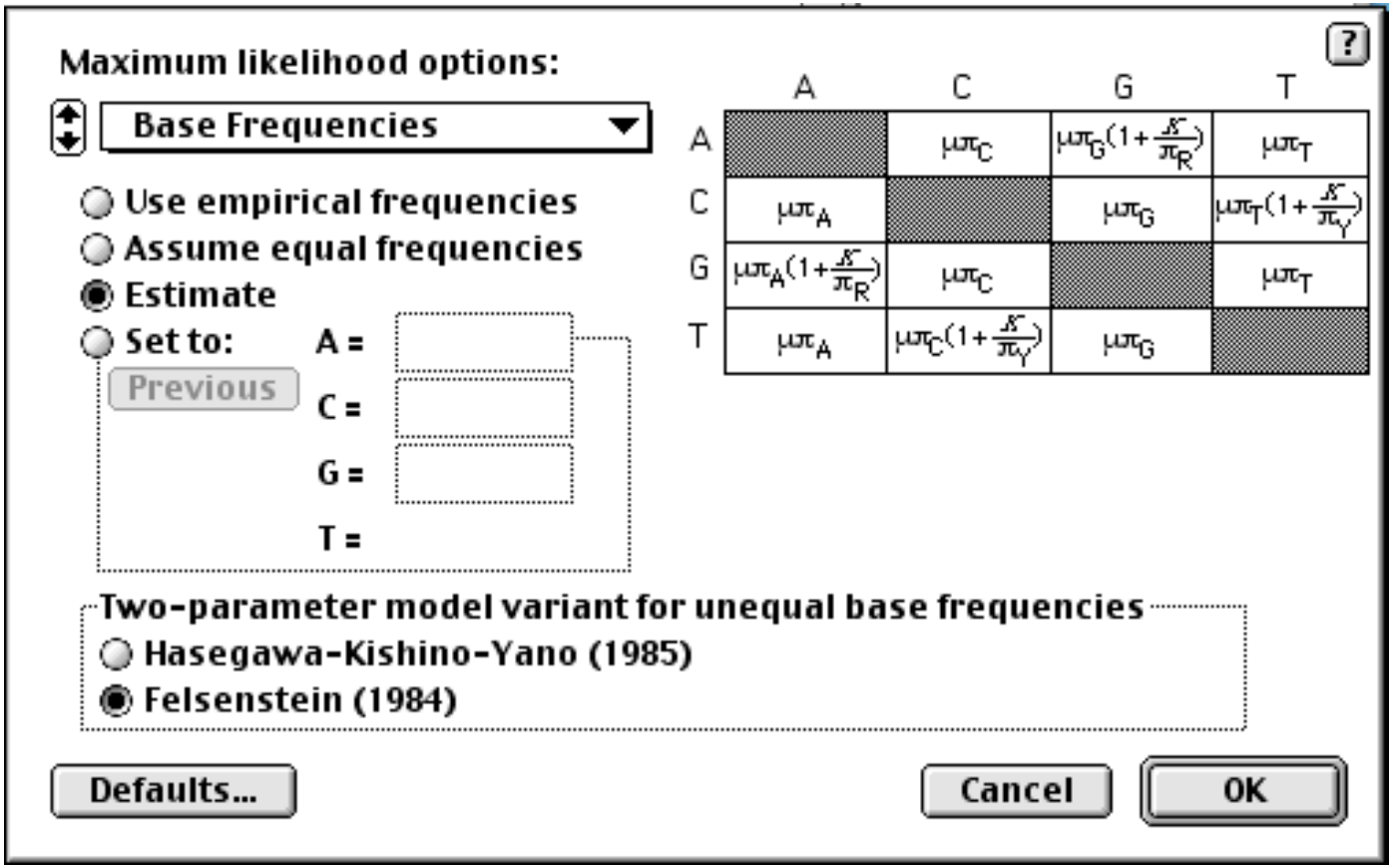
K2P


Maximum likelihood options:
Substitution Model
All rates equal ("1 ST")
Ti rate≠tv rate ("2 ST")
Ti/tv ratio:
Estimate
Previous
Set to:
2
General time-reversible ("6 ST")
Rate matrix:
Estimate
Specify submodel...
Previous
Set to:
Defaults...
Cancel
OK
A C G T
A β α β
C β β α
G α β β
T β α β
C G T
A
C
G
1

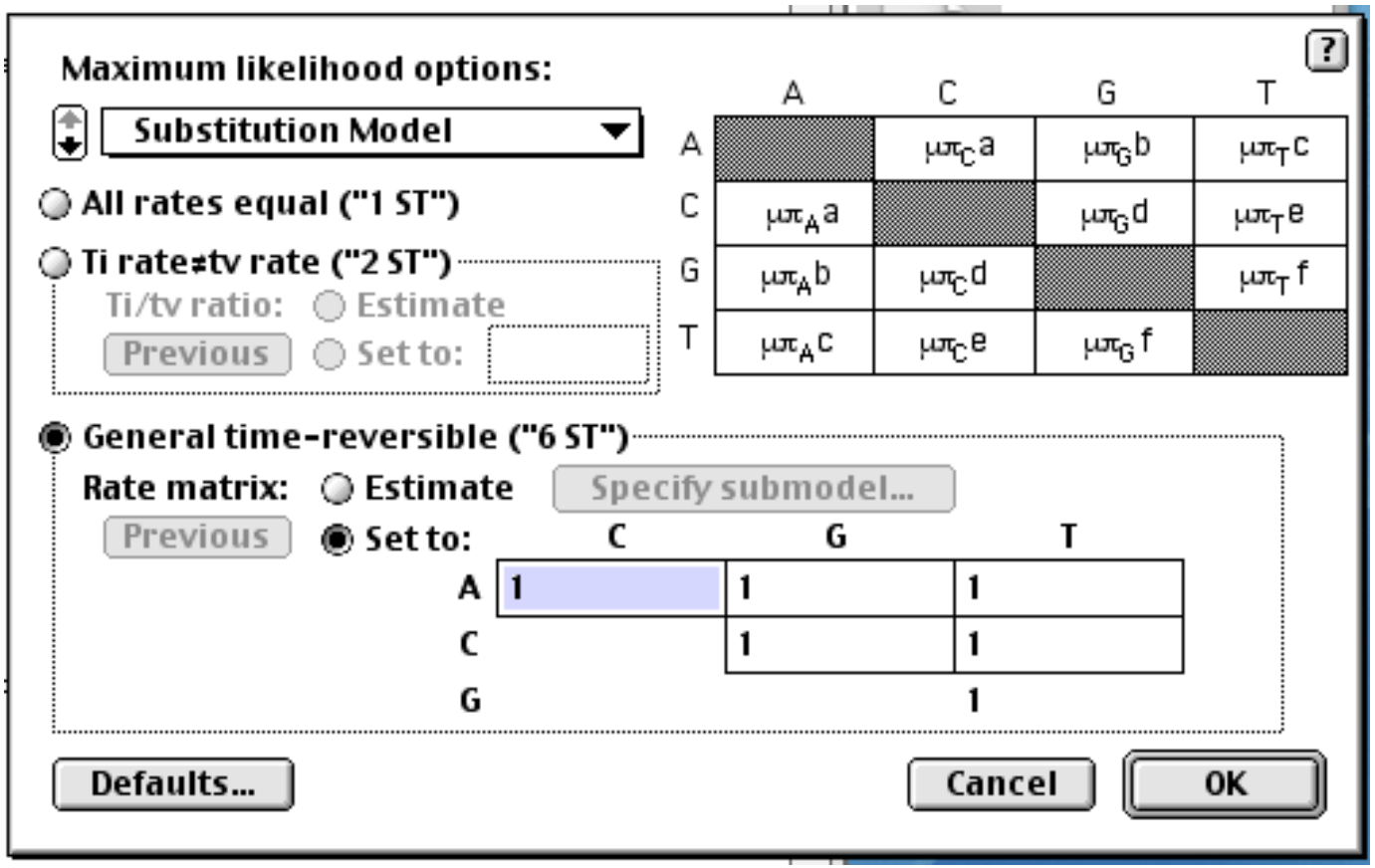
HKY85


Maximum likelihood options:
Base Frequencies
Use empirical frequencies
Assume equal frequencies
Estimate
Set to:
A =
C =
G =
T =
Previous
A C G T
A
C
G
T
Two-parameter model variant for unequal base frequencies
Hasegawa-Kishino-Yano (1985)
Felsenstein (1984)
Defaults...
Cancel
OK

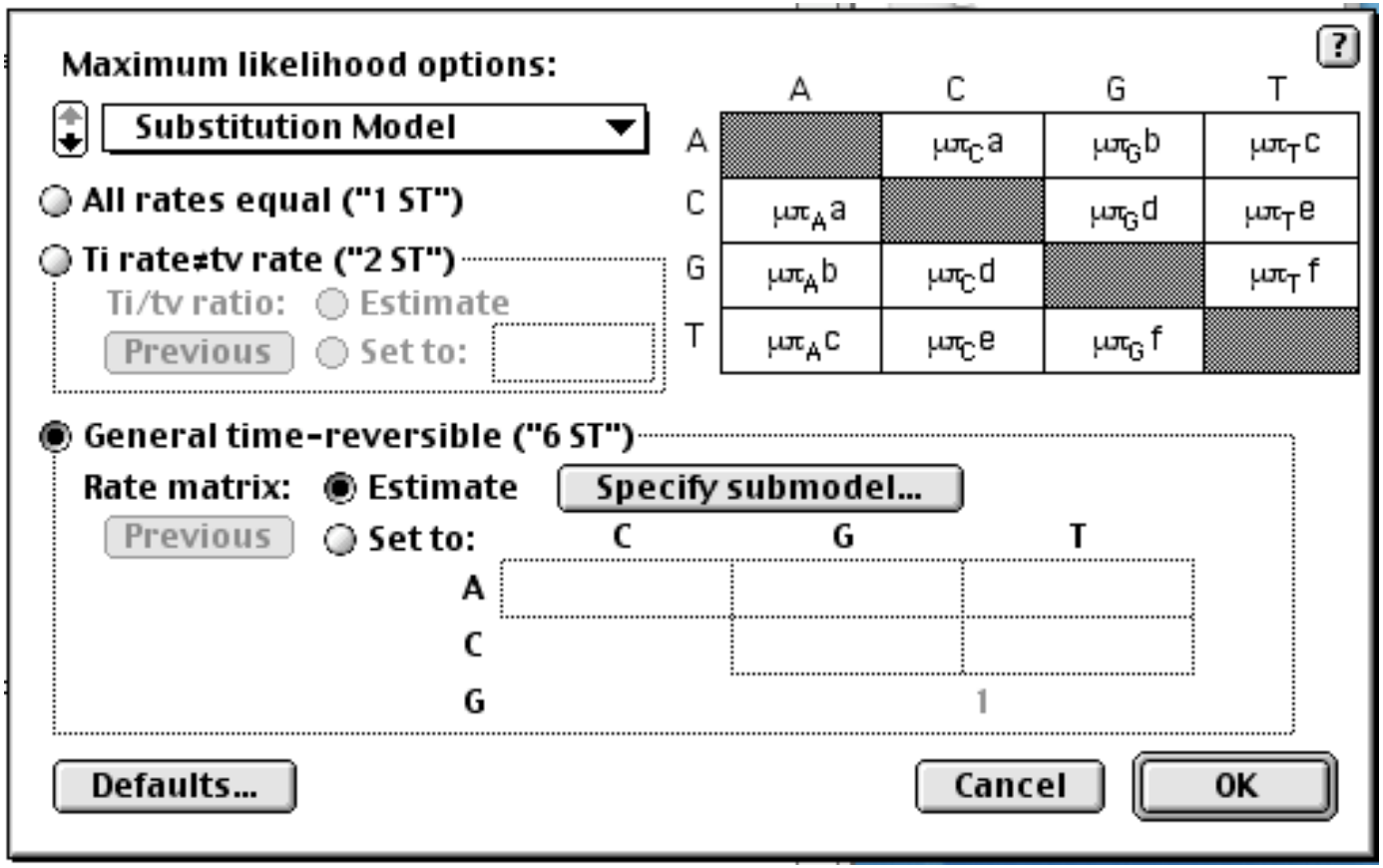
F84


Maximum likelihood options:
Base Frequencies
Use empirical frequencies
Assume equal frequencies
Estimate
Set to:
A =
C =
G =
T =
Previous
A C G T
A
C
G
T
Two-parameter model variant for unequal base frequencies
Hasegawa-Kishino-Yano (1985)
Felsenstein (1984)
Defaults...
Cancel
OK

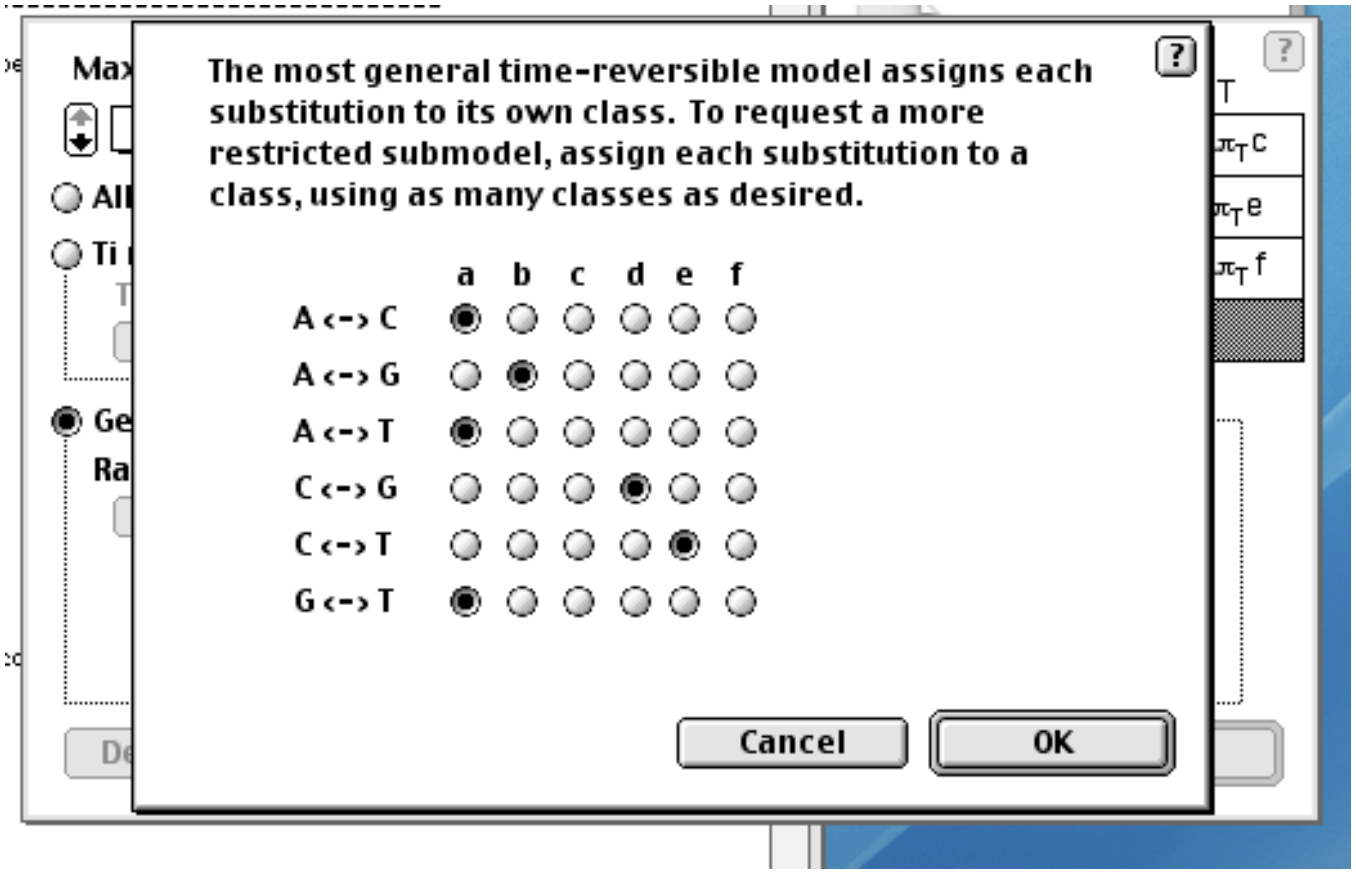
GTR


Maximum likelihood options:
Substitution Model
All rates equal ("1 ST")
Ti rate≠tv rate ("2 ST")
Ti/tv ratio: Estimate
Previous
Set to:
General time-reversible ("6 ST")
Rate matrix: Estimate
Specify submodel...
Previous
Set to:
A C G T
A 1 1 1
C 1 1
G 1
Defaults...
Cancel
OK

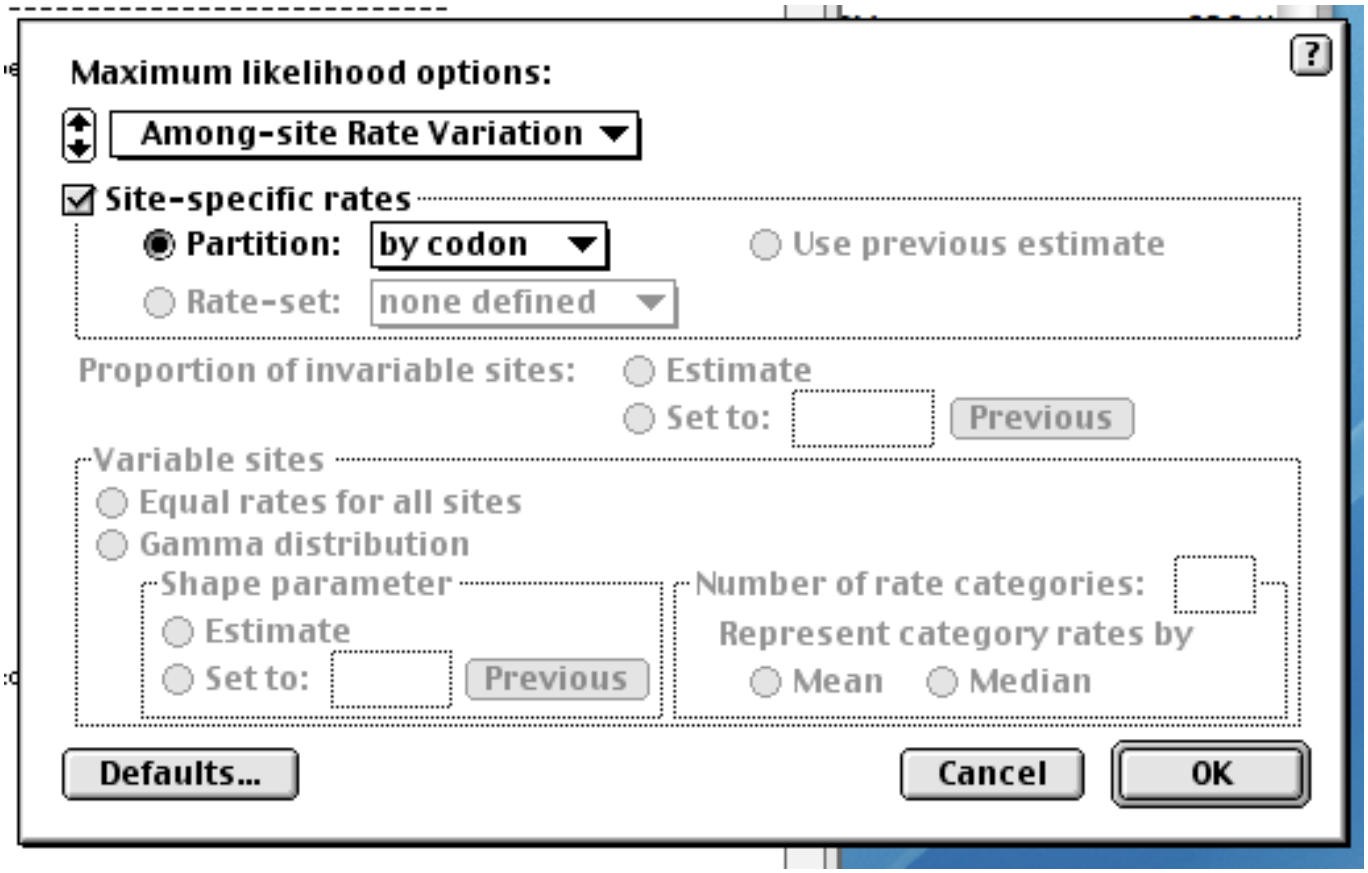
GTR


Maximum likelihood options:
Substitution Model
All rates equal ("1 ST")
Ti rate≠tv rate ("2 ST")
Ti/tv ratio:
Estimate
Previous
Set to:
A C G T
A — μπC a μπG b μπT c
C μπA a — μπG d μπT e
G μπA b μπC d — μπT f
T μπA c μπC e μπG f —
General time-reversible ("6 ST")
Rate matrix:
Estimate
Specify submodel...
Previous
Set to:
C G T
A
C
G
1
Defaults...
Cancel
OK

GTR submodels



Among Site Rate Variation (by codon)



Among Site Rate Variation (Gamma and PINVAR)