Molecular phylogenetics

Genetic distance

Define genetic distance between a pair of ‘homologous’ sequences *x* and *y* as the number of substitutions that have occurred (per alignment site) since *x* and *y* diverged from their common ancestor

Genetic distance

Given the following sequence alignment, infer the genetic distance

```
A C G T T C A T T - - T G
A G - T C C C T G G G G G
```

Simplification: Ignore alignment positions with gaps

Model of evolution

Continuous-time process over {A,C,G,T}

$$p_{ij}(t) = \left(e^{-Gt}\right)_{ij}$$

$$L(A \mid t) = \prod_i p_{A_i^1 A_i^2}(t)$$

$$\hat{t} = \arg\max_t L(A \mid t)$$

Some standard models of nucleotide evolution: Jukes and Cantor

$$
\begin{array}{cccccc}
 & & A & T & C & G \\
 & A & -3\alpha & \alpha & \alpha & \alpha \\
 & T & \alpha & -3\alpha & \alpha & \alpha \\
 & C & \alpha & \alpha & -3\alpha & \alpha \\
 & G & \alpha & \alpha & \alpha & -3\alpha
\end{array}
$$

Kimura two-parameter model (1980)

Models a difference in the rate of transitions and transversions

$$\begin{array}{cccccc} & A & T & C & G \\ A & - & \beta & \beta & \alpha \\ T & \beta & - & \alpha & \beta \\ C & \beta & \alpha & - & \beta \\ G & \alpha & \beta & \beta & - \end{array}$$

With $q_{ii} = -\sum_{i \neq j} q_{ij}$

Recall (exercises 2, question 6):

$$c_i p_{ij} = c_j p_{ji} \qquad \text{for all } i,j$$

$$\sum c_i = 1$$

imply c_i are the limiting probabilities of the chain and the chain is reversible

Analogous result applies to the g_{ij} of a continuous-time chain

Kishino-Hasegawa-Yano (1985)

Includes parameters g_k for the equilibrium nucleotide frequencies

	A	T	C	G
A	–	βg_T	βg_C	αg_G
T	βg_A	–	αg_C	βg_G
C	βg_A	αg_T	–	βg_G
G	αg_A	βg_T	βg_C	–

General Time-Reversible Model (Simon Tavaré 1986)

	A	T	C	G
A	–	αg_T	βg_C	δg_G
T	αg_A	–	γg_C	εg_G
C	βg_A	γg_T	–	ηg_G
G	δg_A	εg_T	ηg_C	–

Generator matrix (or equivalently time) scaled so that one substitution expected in one unit of time

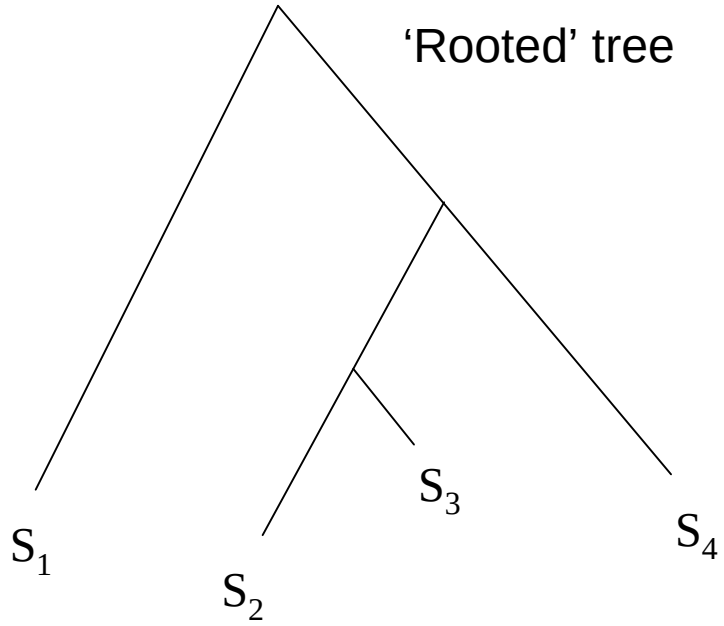
$$\sum_{i \neq j} \pi_i g_{ij} = 1$$

Commonly used evolutionary models also allow heterogeneity in rate of evolution across alignment sites (typically modeled with discretized gamma distribution)

In general, simpler nucleotide substitution models are nested within successively more complex models – standard model comparison techniques can be used to select an appropriate model.

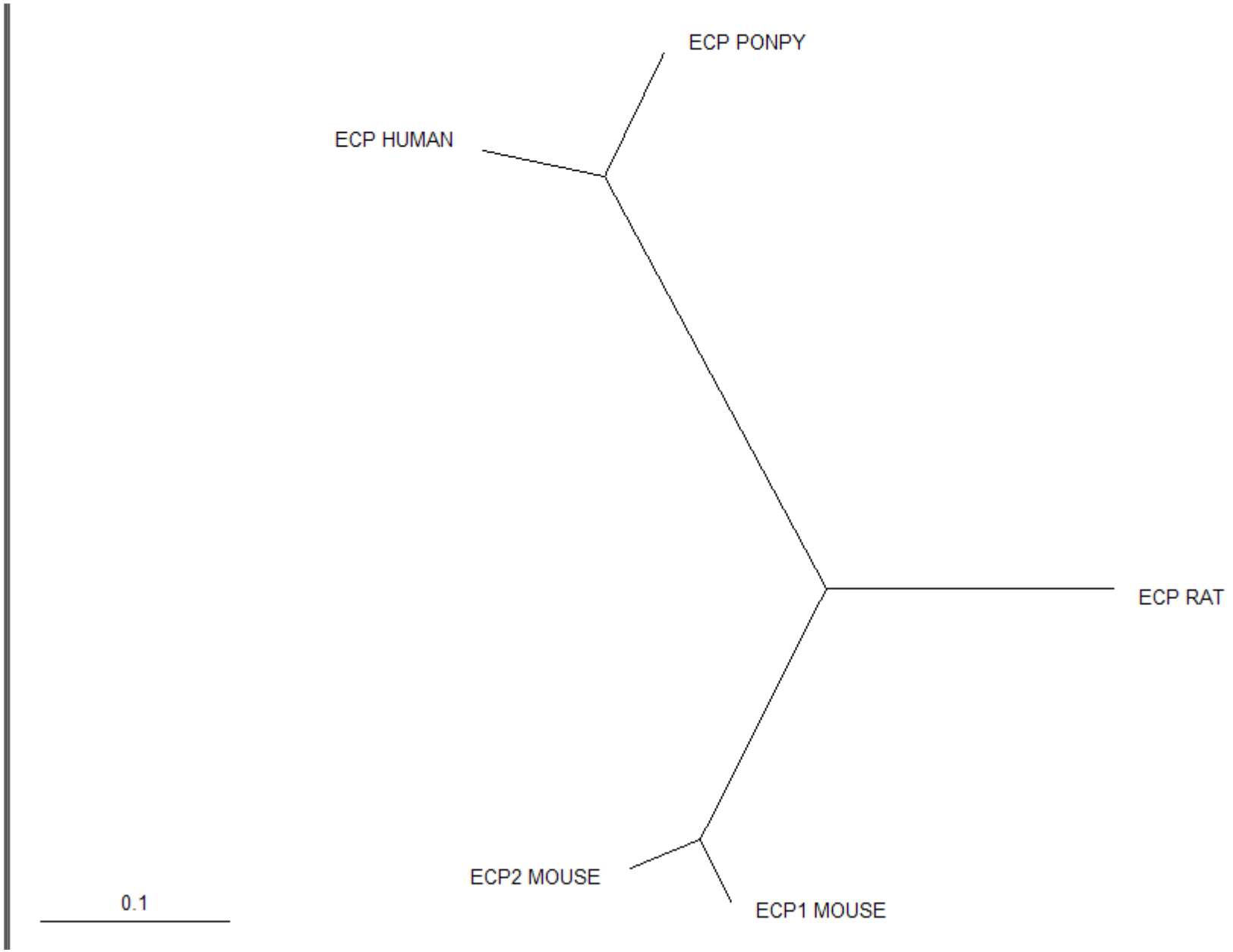
Phylogenetic tree: binary tree with edges representing genetic distance

An evolving sequence can bifurcate (e.g. speciation), giving rise to two daughter sequences



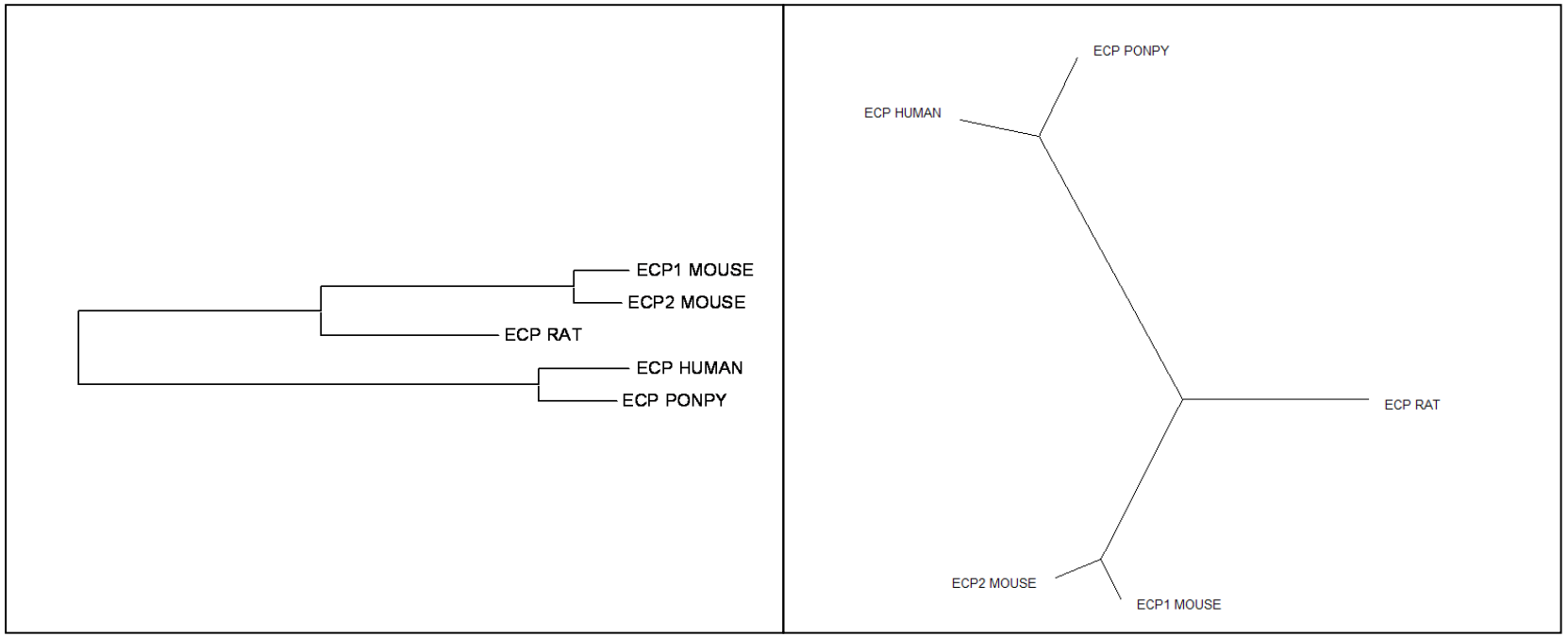
Branch length represents genetic distance between sequences (or hypothetisized sequences) at the nodes

‘Unrooted’ tree

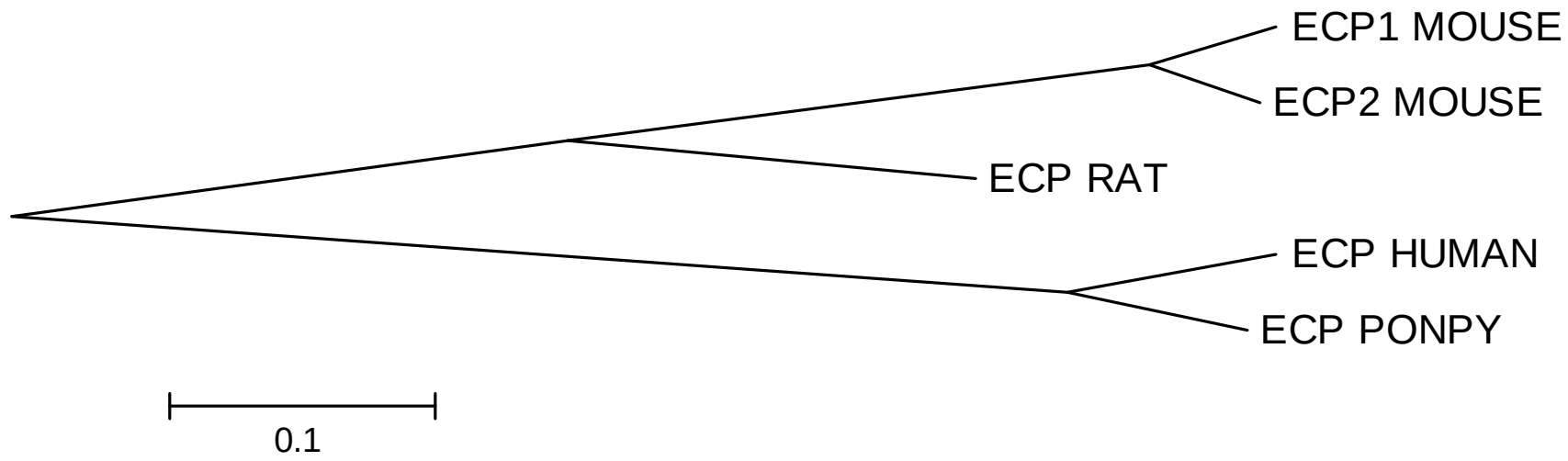


Rooted versus unrooted tree

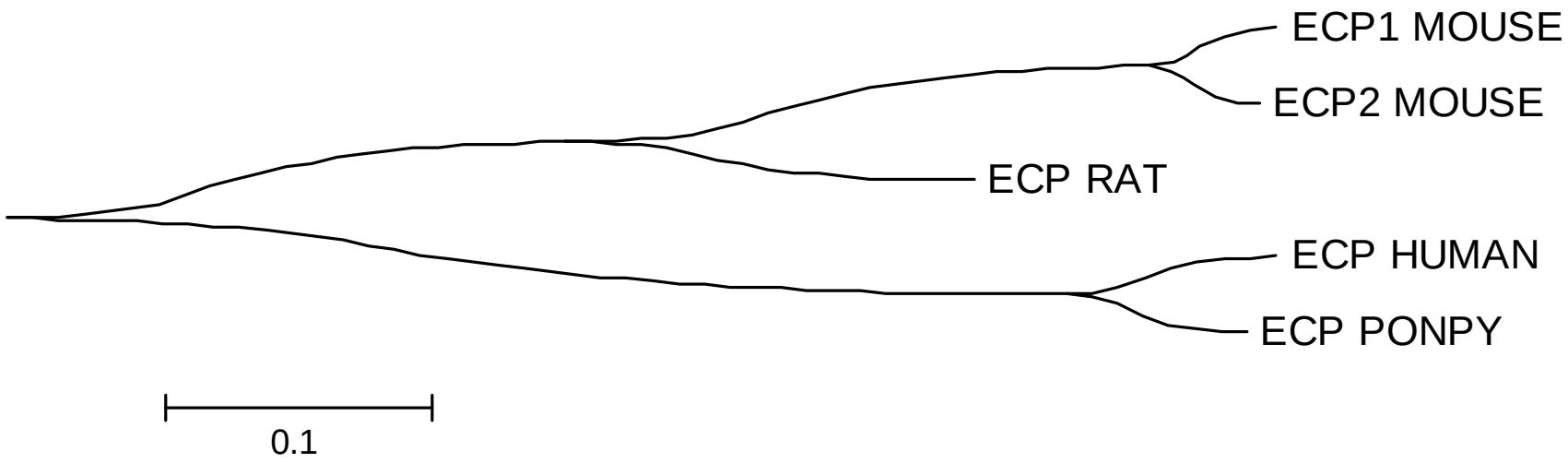
Most substitution models are reversible (see previous slides). Therefore the models cannot distinguish the time-direction of evolution. External information is usually incorporated to decide the position of the root (hypothetical ancestor of all of the sequences represented in the alignment)



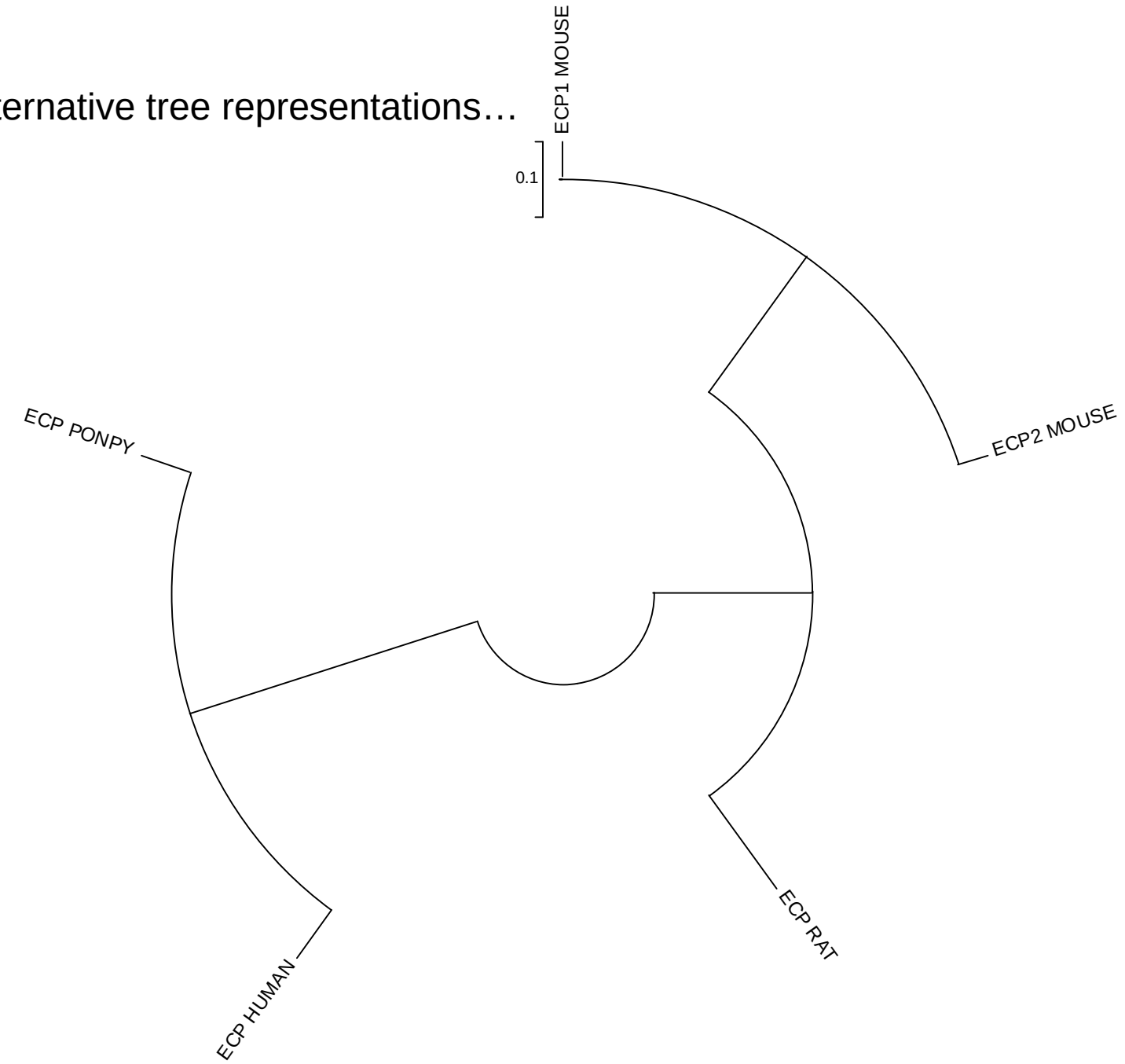
Alternative tree representations…



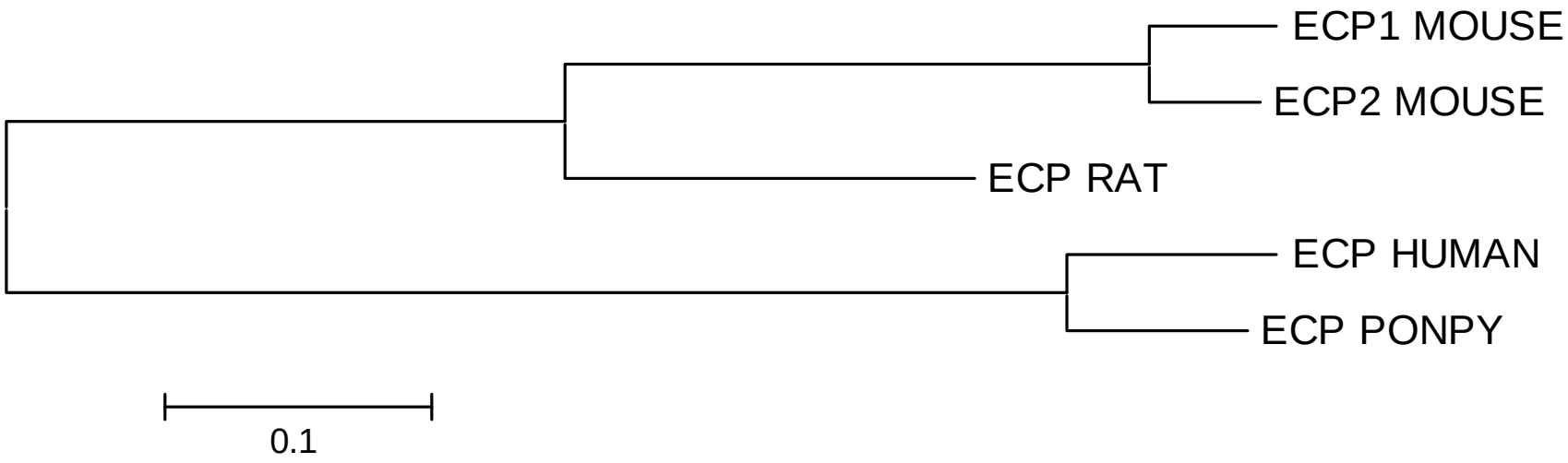
Alternative tree representations…



Alternative tree representations…



Most conventional representation of a ‘rooted’ phylogenetic tree



The phylogeny problem

Given a set of aligned DNA or amino acid sequences, infer the phylogenetic tree representing the evolution of the set of taxa

Requires:

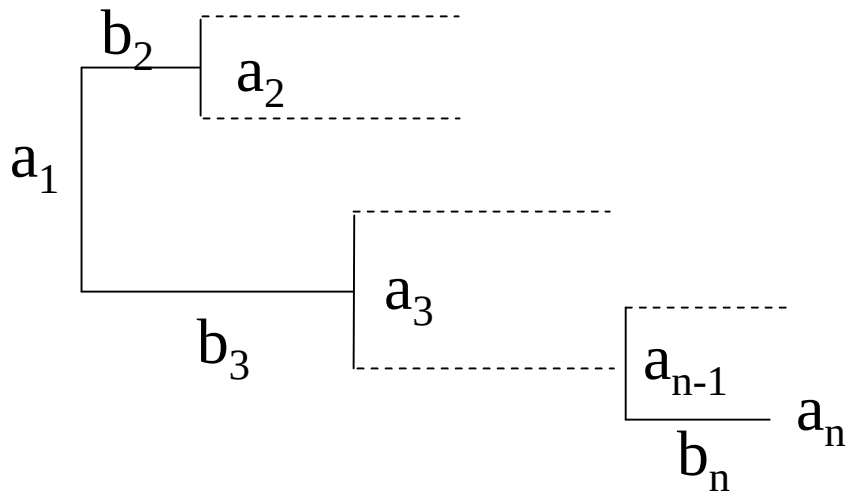
- An optimality criterion (what constitutes the ‘best’ tree)
- Search algorithm

Commonly applied optimality criteria are

- Minimum evolution (tree with shortest sum of branch lengths)
- Maximum parsimony (tree requiring smallest number of steps to explain the data)
- Maximum Likelihood
- Maximum *a posteriori* Probability (MAP)

The likelihood of a tree:

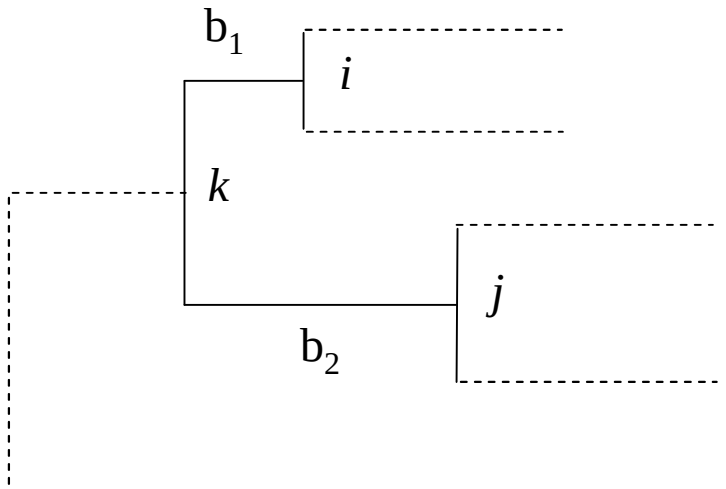
$$P(D|T) = \sum_{a_1}\sum_{a_2}\sum_{a_3}\ldots\sum_{a_n} P(a_1)P(a_2|b_2,a_1)P(a_3|b_3,a_1)\ldots P(a_{n-1}|b_n,a_n)$$



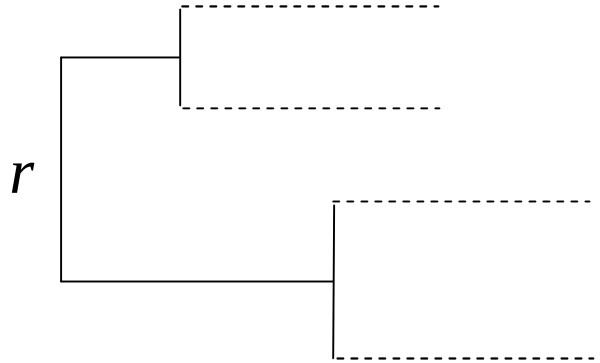
A recursive algorithm is used to avoid doing all the summations (Felsenstein's Pruning Algorithm)

Let L_r^k be the likelihood of the *subtree* decended from node *k*, given that the character at node *k*, is *m* then

$$L_m^k = \left(\sum_s L_s^i P(s \mid m, b_1) \right) \left(\sum_s L_s^j P(s \mid m, b_2) \right)$$



$$P(T \mid D) = \sum_{s} L_s^r \, p(s)$$

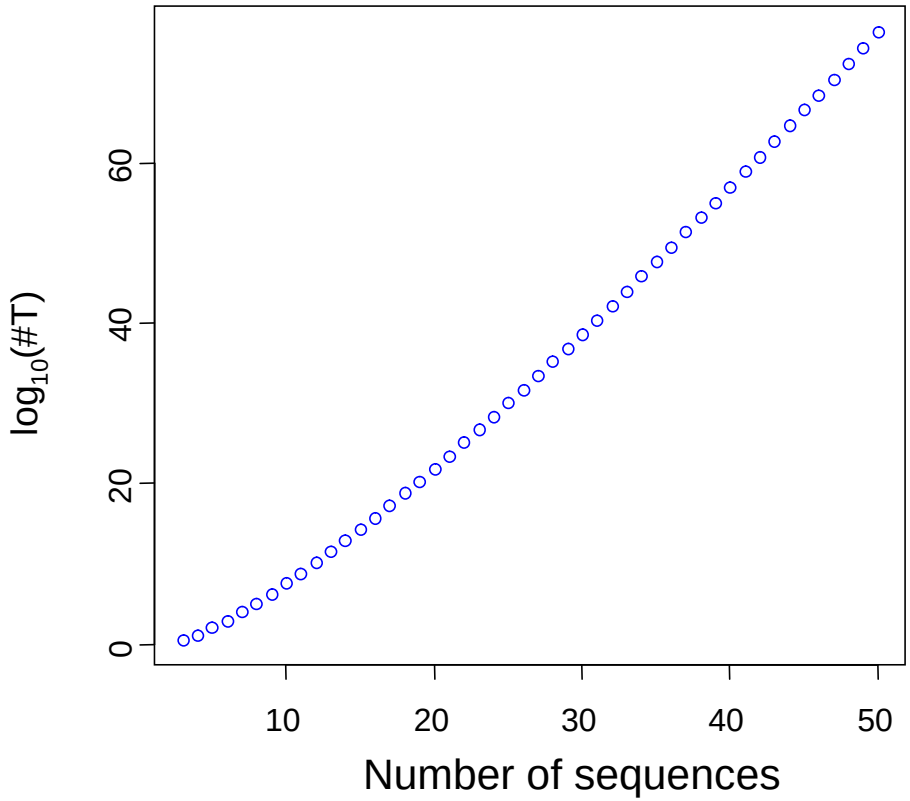


Complexity: O(n . m . k^2)

(n = # sequences; m = sequence length; k = alphabet size)

How many trees?

$$(2n - 3)!/(2^{n-2}\,(n - 2)!)$$



Searching for optimal trees

Branch & Bound

Heuristics – usually local perturbations with hill-climbing

Markov-Chain Monte Carlo (MC^3)

Genetic algorithms

etc.

Bayesian MCMC in phylogenetics

Prior over trees (often flat)

Starting tree (star decomposition, step-wise addition, NJ)

Proposals: tree perturbations

Acceptance depends on ratio of posterior probabilities (= ratio of likelihoods)

Determine burn-in and convergence

In molecular phylogenetics the prior is usually ‘flat’

Why bother?

1. We get the answer as a probability

2. We get to use MCMC to sample over trees/search for ‘best’ tree

3. Integrates over nuisance parameters rather than using their optimal values

Recall Markov Chain Monte Carlo: a Markov chain is specified with the desired limiting probabilities. Only ratios of limiting probabilities were required

$$p_{ij} = q_{ij} \min\left(1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right) \quad \text{for } i \neq j; \quad (p_{ii} \text{ s.t. } \sum_j p_{ij} = 1)$$

$$P(T, \Theta) = \frac{P(D|T, \Theta)P(T, \Theta)}{P(D)}$$

$P(D)$ is difficult to calculate, except for very small numbers of taxa, but cancels in the context of MCMC.

Running a phylogenetic MCMC

Generate long chain of trees/parameters sampled according to their joint posterior probability

The number of times the chain visits tree X is proportional to the probability of tree X

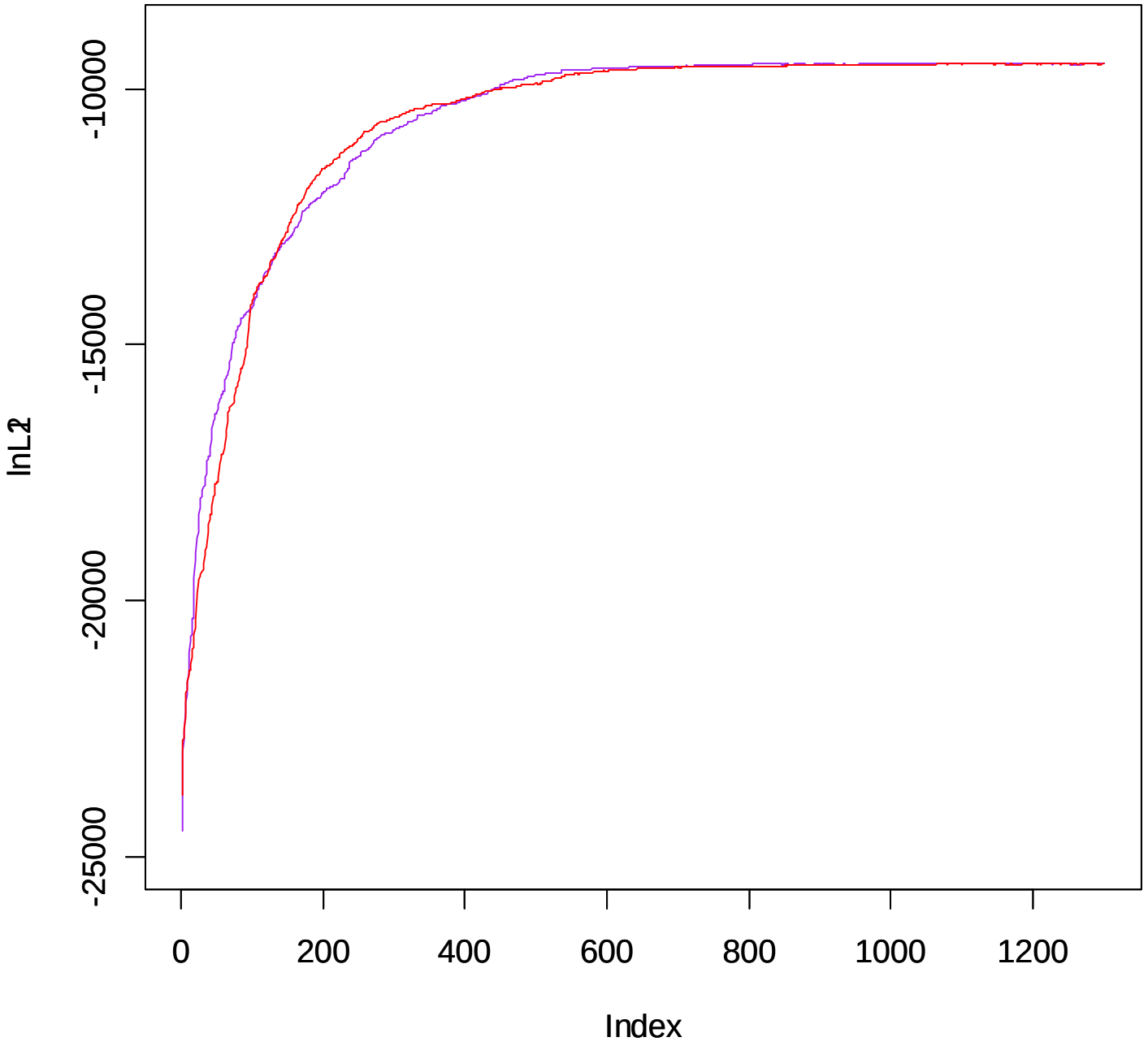
The number of times a specific branch is sampled can be used to estimate the posterior probability that the ‘clade’ of taxa specified by the branch is correct

Multiple chains may be used (Metropolis coupled MCMC = MC^3)

- Only one chain is sampled
- The other chains are heated (i.e. they can take bigger steps)
- Chains can swap states
- Allows crossing of valleys

Burnin

- From an arbitrary starting point the chain can take some time to equilibriate
- Consequently, the chain takes some time before samples are obtaiend according to their posterior probabilites
- Initially probability of trees increases with time
- Programmes allowed to run until the probabilities are fluctuating randomly about a constant mean
- Data generated before the chain equilibriates are discarded


lnL2
-10000
-15000
-20000
-25000
0
200
400
600
800
1000
1200
Index

Advantages of Bayesian methods

- relatively fast
- easily interpretable
- often very accurate

Disadvantages of Bayesian methods

- can be difficult to be sure of convergence (this has improved with availability of better diagnostics)
- still controversial in molecular phylogenetics – thought by some to exaggerate confidence

Software: e.g. MrBayes

Real world sequences evolve with insertions and deletions

The problem of inferring the tree under realistic models that include insertion and deletion mutations is an area of active research (e.g. Redelings & Suchard, 2007)

Models of codon sequence evolution

‘Universal’ genetic code is degenerate => natural classification of mutations as:

Nonsynonymous: amino acid changing (rate - dN)

Synonymous: no amino acid change (rate - dS)

ω: dN/dS

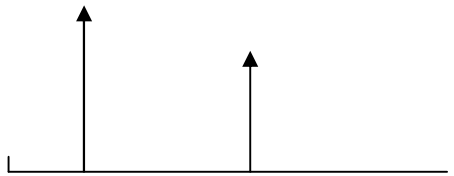
$\omega > 1$ => adaptive evolution (actually ‘diversifying selection)

Codon Substitution model

$$p_{ij} = \begin{cases} 0, & \text{codons differing at more than one position} \\ \pi_j, & \text{synonymous transversion} \\ \kappa\pi_j, & \text{synonymous transition} \\ \omega\pi_j, & \text{nonsynonymous transversion} \\ \omega\kappa\pi_j, & \text{nonsynonymous transition} \end{cases}$$

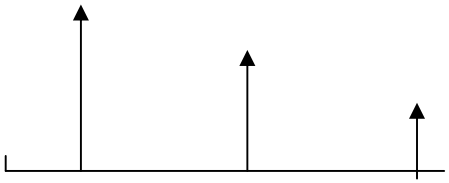
Example: Analysis of selection (random effects model)

Neutral model:



$\omega_- < 1 \quad p_{\omega-} \quad \omega_N = 1$

Selection model:



$\omega_- < 1 \quad p_{\omega-} \quad \omega_N = 1$

$\omega_+ > 1 \quad p_{\omega+}$

Questions of interest

Is there a subset of sites with $\omega > 1$

- model comparison techniques

Which sites are evolving adaptively (empirical Bayes method)

- fix all parameter values to their ML estimates

- using ML estimates as priors, calculate posterior probabilities of belonging to selection site class

$$P(\omega_i > 1) \propto L(D \mid \omega_i > 1) P(\omega_i > 1)$$

Analysis of selection (random effects model)

- Infer a phylogenetic tree

- Obtain ML estimates of all parameters

- Use LRT (or other model comparison method) to evaluate evidence for selection

- Use empirical Bayes method (or a variant) to estimate posterior probabilities of belonging to the selection site class

$$P(\omega_i = \omega_+ \mid D) = \frac{P(D \mid \omega_i = \omega_+)\, p_{\omega_+}}{\sum_k P(\omega_i = \omega_+)\, p_{\omega_k}}$$