relaxed phylogenetics and dating with confidence

Alexei Drummond

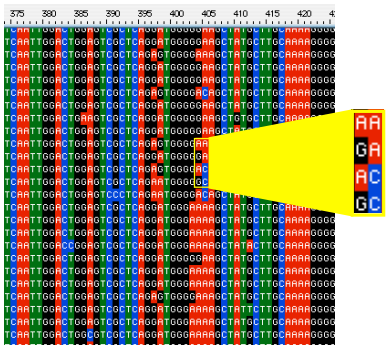
Overview

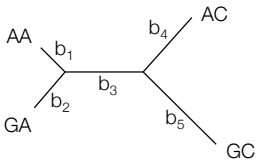
- Bayesian phylogenetics
- Relaxed molecular clocks in BEAST 1.4
- New relaxed molecular clock model

Review of Bayesian evolutionary inference

- The output of a Bayesian evolutionary analysis is a **probability distribution** on trees and parameter values.
- For **phylogenetics** the tree topology is the object of interest. The substitution parameters and tree prior parameters are a **nuisance** that we average over using MCMC and then ignore.
- For **population genetics** the tree and substitution parameters are a **nuisance** that we average over and then ignore, focusing instead on the population parameters.
- Sometimes an evolutionary hypothesis more specific than a full tree topology is of interest (like "Did this adaptive radiation predate the Miocene?") and then the result of the analysis should be the testing of this hypothesis, averaged over all trees and parameter values, weighted by their probability given the data.

Molecular evolutionary model: Felsenstein's likelihood (1981)



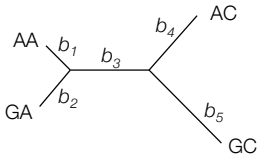


The probability of the sequence alignment,

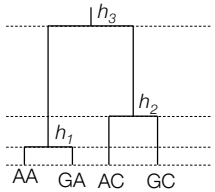
$$\Pr\{D \mid T, Q\}$$

can be efficiently calculated given a tree and branch lengths (T), and a probabilistic model of mutation represented by an instantaneous rate matrix (Q). **In phylogenetics, branch lengths are usually unconstrained**.

Model assumptions

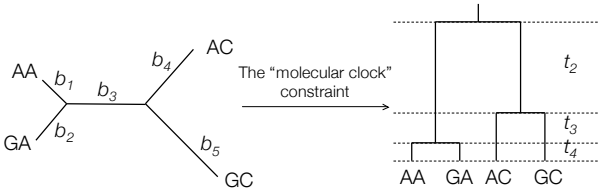


1. Product of rate and time (branch length) is independent and identically distributed among branches.
2. The root of the tree could be anywhere with equal probability.
3. Topology implies nothing about branch lengths.



1. Rate of evolution is the same on all branches.
2. The root of the tree is equidistant from all tips.
3. Topology constrains branch lengths (e.g. two branches in a cherry must be of equal length)

Combining the clock-constraint with Felsenstein's likelihood



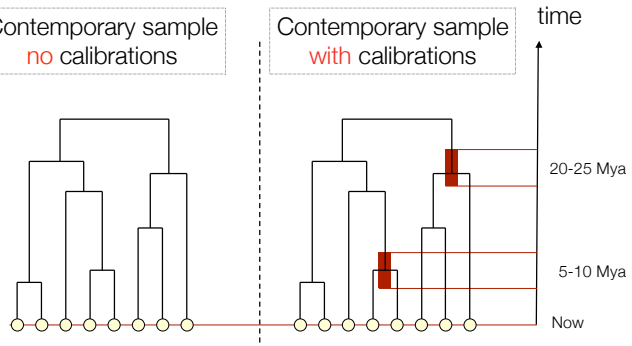
2n–3 branch lengths

n–1 waiting times

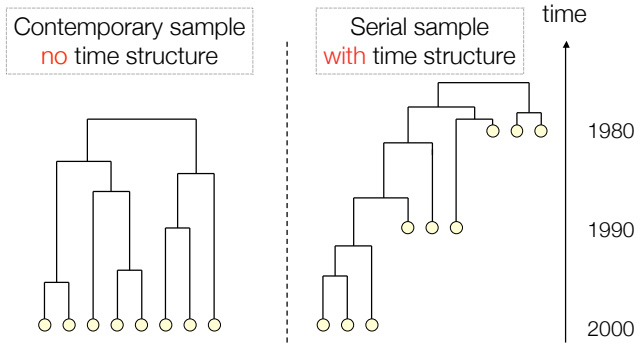
$$p(N,g,Q \mid D) \propto \Pr\{D \mid \mu g,Q\} f_G(g \mid N) f_N(N) f_Q(Q)$$

The joint posterior probability of the the tree (*g*), the tree prior parameters (N) and the substitution matrix (*Q*) are estimated using Markov chain Monte Carlo (Drummond et al, Genetics, 2002)

Time structure via calibrations



Time structure in samples themselves



Full Bayesian Model

Probability (density) of what we don't know given what we do know.

Likelihood function

other priors

$$P(g, \mu, N_e, Q \mid D) = \frac{1}{Z} P(D \mid g, \mu, Q) f_G(g \mid N_e) f_m(\mu) f_N(N_e) f_Q(Q)$$

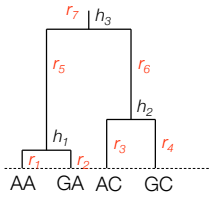
Unknown normalizing constant

tree prior

Q = substitution parameters
N_e = tree prior parameters
g = tree
μ = overall substitution rate

In the software package BEAST, MCMC integration can be used to provide a chain of samples from this density.

Relaxing the molecular clock



In the field of **divergence time estimation** auto-correlated relaxed clocks have been considered.

e.g. Thorne et al, 1998:

$$r_i \sim LogNormal(r_{A(i)}, \sigma^2 \Delta t_i)$$

AC

$r \sim Exp(\lambda)$

$r \sim LogNormal(\mu, \sigma^2)$

$r \sim Gamma(\alpha, \beta)$

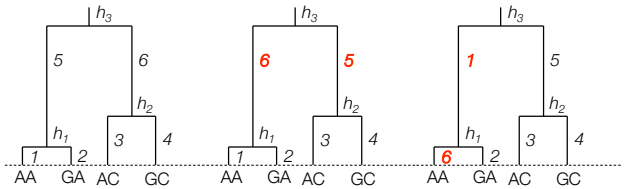
We introduce a relaxed clock model in which there is no prior correlation between child and parent rates

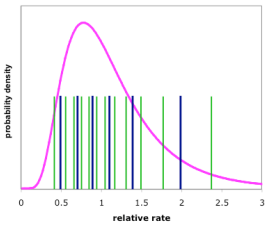
"Un-correlated" or "memory-less" relaxed clocks

ML

Drummond et al, available in BEAST

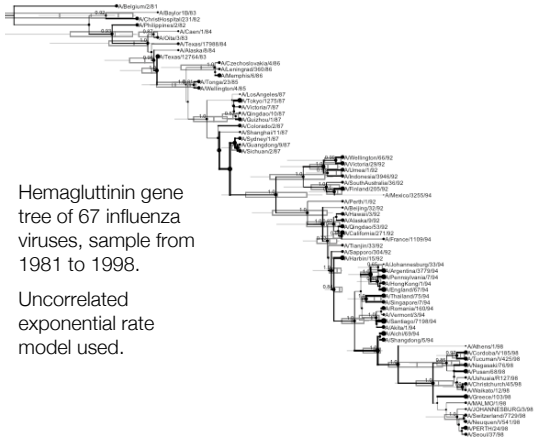
Sampling branch rates using MCMC





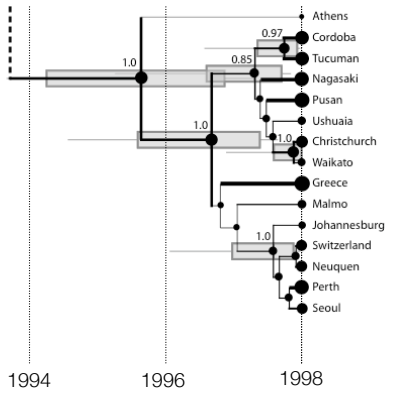
1. Rates are summarized into 2n-2 rate categories (e.g. blue is 6 categories; green is 12 categories).
2. Random pairs of rates categories are swapped during MCMC.
3. For purposes of topology changes, rate categories are associated with child node.

Relaxed influenza gene tree



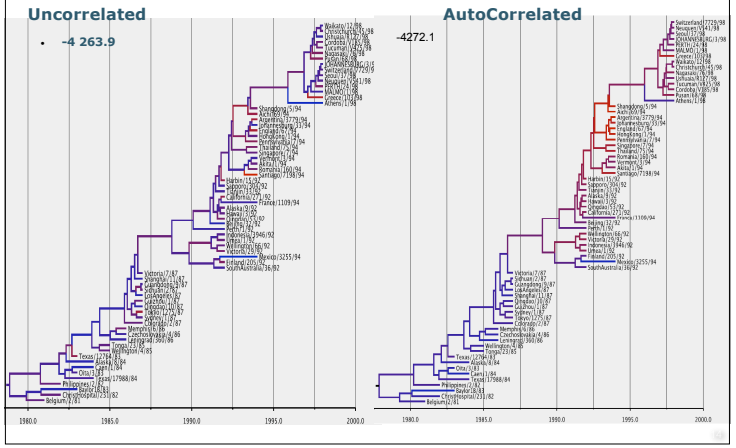
1. Hemagluttinin gene tree of 67 influenza viruses, sample from 1981 to 1998.
2. Uncorrelated exponential rate model used.

Relaxed influenza gene tree (cont)

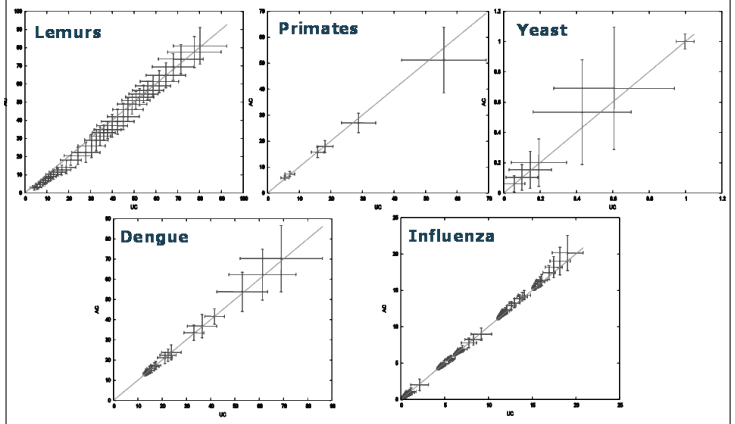


1. Box-and-whisker plots show uncertainty in divergence times (only for splits with posterior probability > 0.5)
2. Node size and branch thickness proportional to substitution rate.

Influenza dataset consensus trees

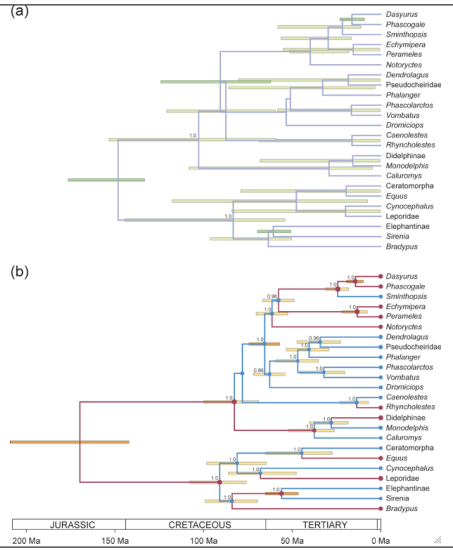


Divergence times of UC and AC

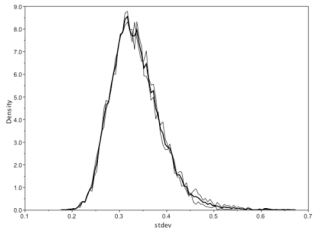


Prior versus Posterior

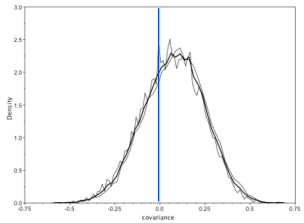
Marsupials example (24 taxa, 5658 nucleotides)



Posterior distribution of rates across branches



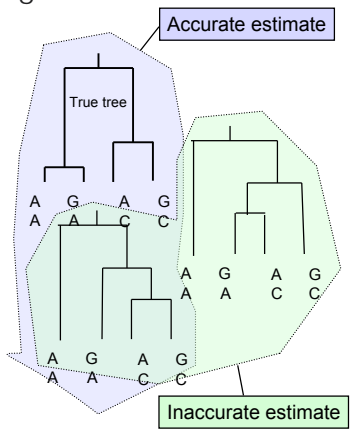
Coefficient of variation (0.23,0.41)



Correlation of parent and child branch rates (-0.26, 0.4)

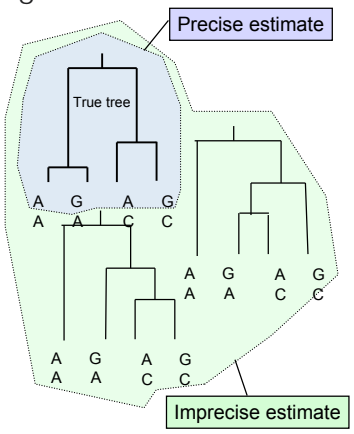
Accuracy in Bayesian phylogenetics

- Phylogenetics is an estimation problem, in which the phylogenetic tree topology is the object we wish to estimate.
- The error associated with this estimation can be described by the **95% credible set** of trees: the smallest set of trees including 95% of the posterior probability.
- A standard measure of accuracy is the **false positive rate**. How often do we exclude the true tree from the **95% credible set**? Ideally it would be 5%…



Precision in Bayesian phylogenetics

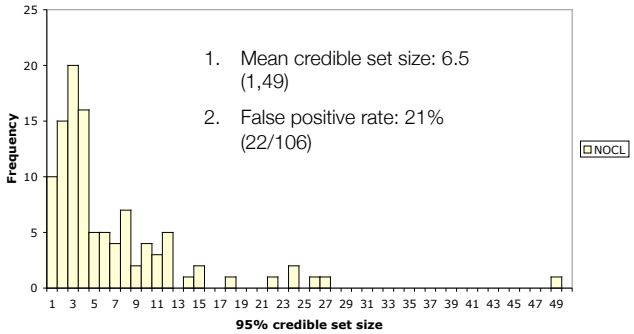
- The precision of an estimate can be described by how much is excluded.
- How small is the **95% credible set** of trees?



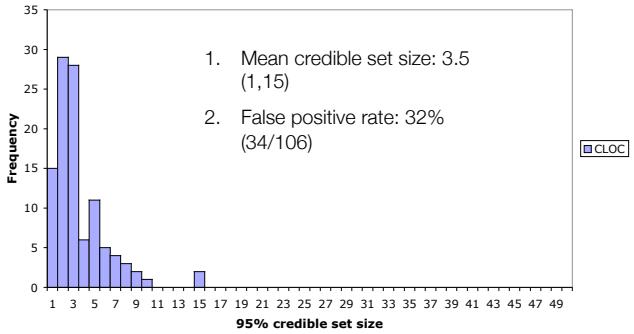
Testing accuracy and precision with **real data**

- Used 106 genes from 8 species of yeast (Rokas et al, 2003) and 4 other “phylogenomic” data sets
- For each gene used both MrBayes and BEAST to estimate phylogeny and 95% credible set
- Assumed true tree is the tree estimated using all the concatenated data set.
- Tabulated number of trees in credible set and whether the true tree was in credible set for MrBayes (unconstrained) and BEAST (MLLN and CLOC models)

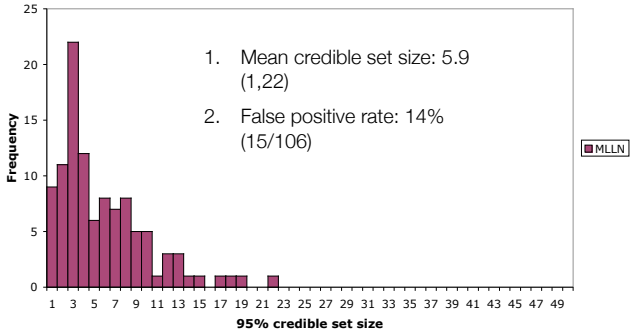
MrBayes results



Beast:CLOC results



Beast:UCLN results



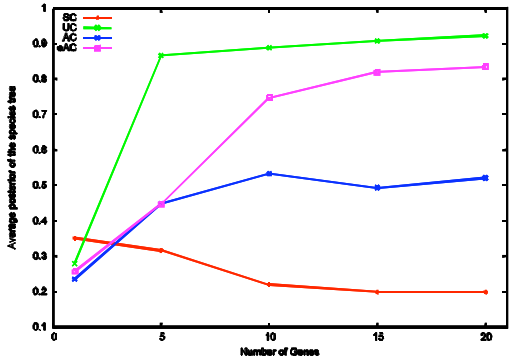
Summary of tree accuracy results

Dataset	Sample Size	Average Length	Clock Rejected by LRT	Accuracy (%) (True Tree in 95% Credible Set)[a]		
				CLOC	UCLN	UF
Bacteria	102	170 aa	26%	46.1	**48.0**	42.2
Yeast	106	1,198 bp	76%	67.0	**84.9**	79.2
Plants	61	647 bp	67%	**91.8**	88.5	83.6
Animals	99	197 aa	59%	64.6	**69.7**	57.6
Primates	500	632 bp	13%	88.8	**89.0**	88.8

Summary of tree precision results

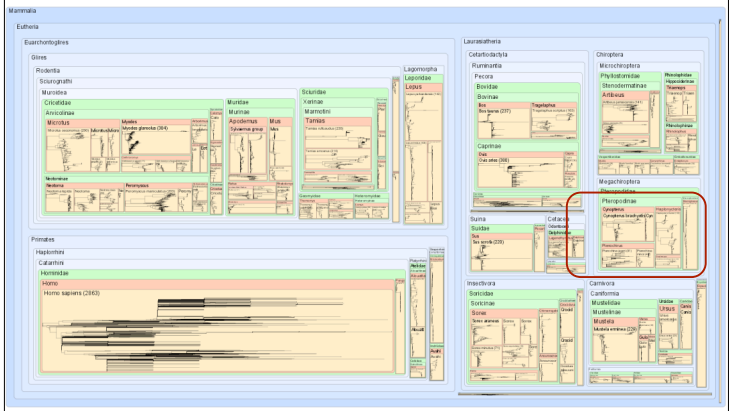
Dataset	Sample Size	Average Length	Clock Rejected by LRT	Precision (Number of Trees in 95% Credible Set)[b]		
				CLOC	UCLN	UF
Bacteria	102	170 aa	26%	5.7	10.3	11.3
Yeast	106	1,198 bp	76%	3.5	5.9	6.5
Plants	61	647 bp	67%	7.5	15.4	9.2
Animals	99	197 aa	59%	5.7	10.2	14.2
Primates	500	632 bp	13%	3.1	3.4	5.1

Increasing the length of the sequence



new relaxed clock model

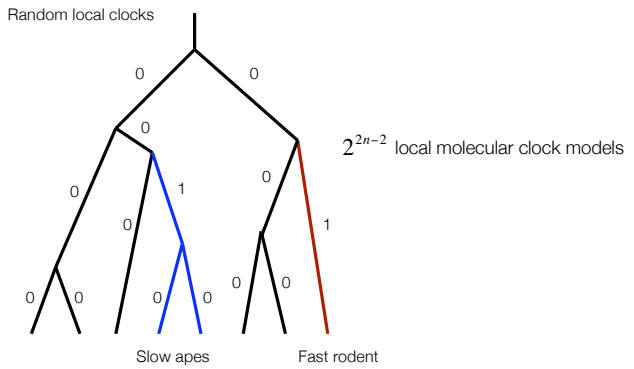
Cytochrome b gene trees for >100 Mammals in Genbank



Check this BLAST view and more at **http://www.geneious.com**

When is different really different?

A new model for relaxing the clock



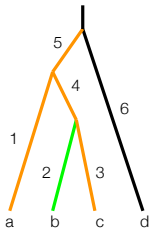
A new model for relaxing the clock

$\mathbf{d} = \{\delta_1, \delta_2, ..., \delta_{2n-2}\}$
$\delta_k \in \{0,1\}$
Indicators

$\mathbf{f} = \{\phi_1, \phi_2, ..., \phi_{2n-2}\}$
$\phi_i \in (0, \infty)$
Rate scale parameters

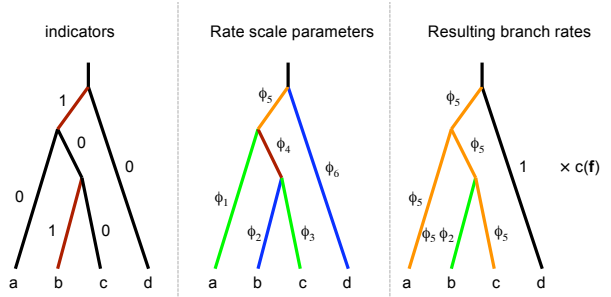
$r = \{r_1, r_2, ..., r_{2n-2}\}$
$r_i = \phi_i^{\delta_i}$
Rate scales

$\mathbf{m} = \{\mu_1, \mu_2, ..., \mu_{2n-2}\}$
$\mu_i = c(\mathbf{f}) r_i r_{parent(i)}$
Branch rates



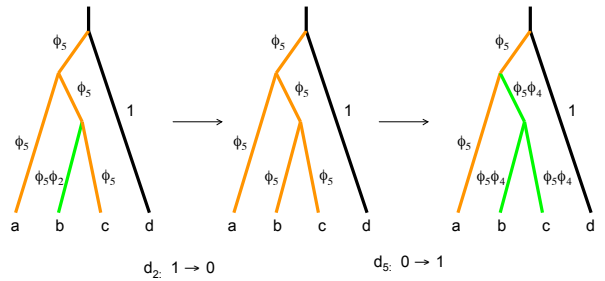
$\mathbf{d} = \{0,1,0,0,1,0\}$
$\mathbf{f} = \{\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6\}$
$\mathbf{r} = \{1, \phi_2, 1, 1, \phi_5, 1\}$
$\mathbf{m} = c(\mathbf{f})\{\phi_5, \phi_2\phi_5, \phi_5, \phi_5, \phi_5, 1\}$

A new model of relaxed clock

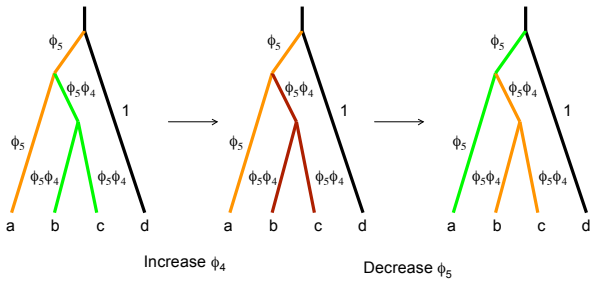


Red/Orange fast, Green/Blue slow

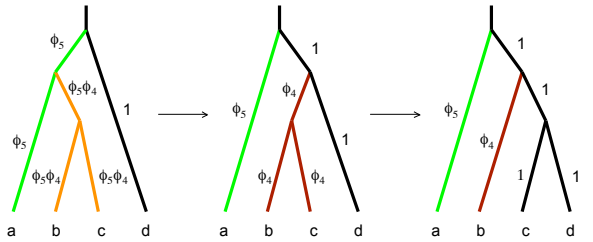
Sampling the indicators, **d**



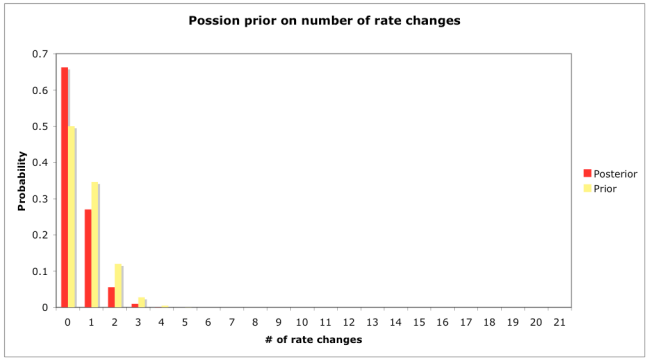
Sampling the rate parameters, **f**



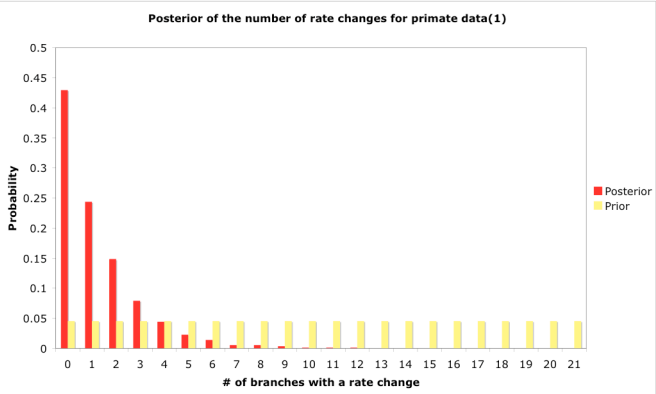
Sampling the trees



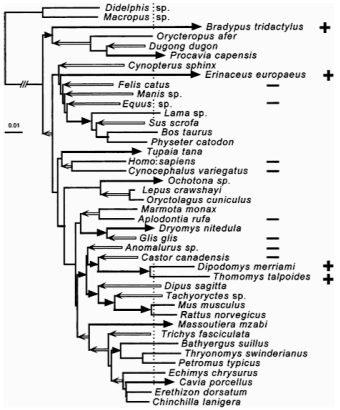
Primate.nex (Poisson prior)



Primate.nex (Uniform prior)



Rodents (1+2 pos of 3 nuclear genes)

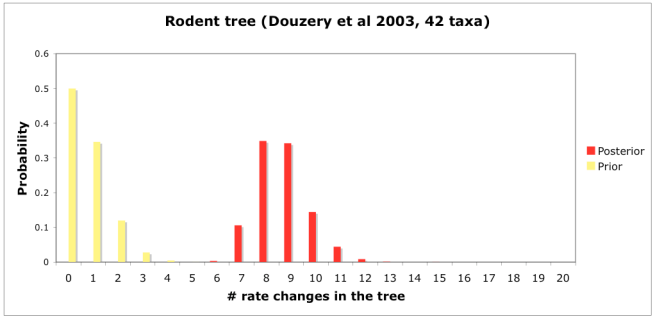


82 branches

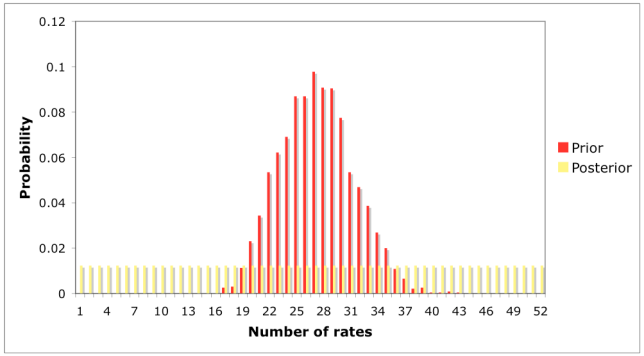
38 rate changes according to Douzery et al 2003

Fig. 1. Extensive nucleotide substitution rate variations in the first two codon positions of the ADRA2B + IRBP + vWF nuclear genes between placental mammals. The *vertical dashed line* indicates the mean value of the root-to-tip distance of the 40 placental taxa. Significantly faster- or slower-evolving species are indicated, respectively, by a + or a − as evidenced by the branch-length test. Significantly faster- and slower-evolving branches as evidenced by the two-cluster test are indicated, respectively, by *filled arrows* pointing right and *open arrows* pointing left. The scale unit corresponds to the expected number of nucleotide substitutions per site. The log-likelihood of this tree is ln*L* = −26,054.36, and its AIC is 52282.78. In the clock-like constrained model—with a single global clock—a significant loss of log-likelihood is observed (ln*L* = −26,222.37, AIC = 52,538.74).

Douzery et al, 2003 (Poisson prior)

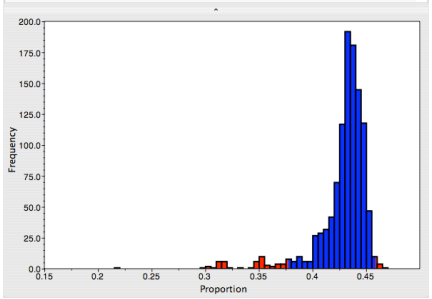


Douzery et al, 2003 (Uniform prior)

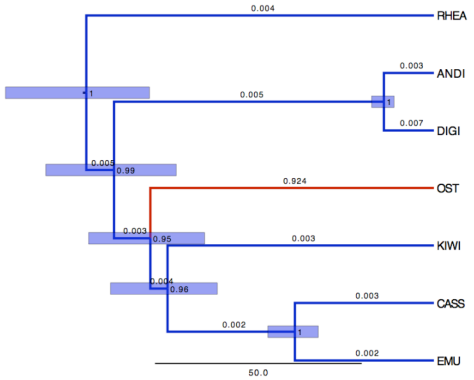


Proportion of tree occupied by largest clock - (Poisson prior)

Summary Statistic	Proportion
mean	0.428
stdev of mean	1.71E-3
median	0.434
95% HPD lower	0.377
95% HPD upper	0.459
auto-correlation time (ACT)	5.225
effective sample size (ESS)	210.527

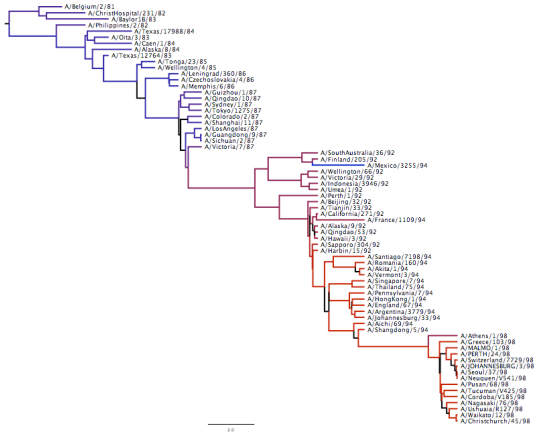


Ratite tree example

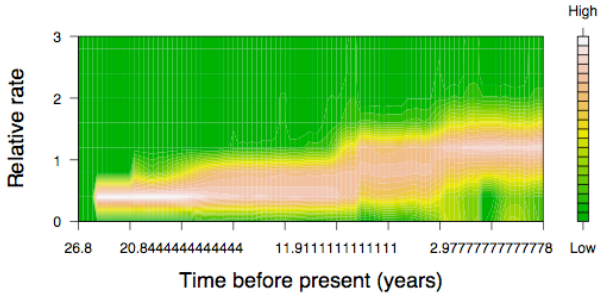


A single strongly supported rate change is found

Influenza A (69 sequences, 20 years)



Influenza A



Conclusions

- Relaxed molecular clocks have many benefits over unconstrained models for phylogenetic inference
  - They appear to estimate the phylogenetic tree more accurately on real data sets
  - They automatically provide estimates of a root, without the need for an outgroup
  - They automatically provide estimates of relative divergence dates, or absolute divergence dates when calibration information is available
- Future directions
  - Phylogenetic inference with autocorrelated rates (done, not published)
  - Estimation of correlations in rate variation across multiple genes

the end

Phylogenetics: One tree to rule them all

- Majority consensus tree
  (what branch lengths?)
  Might not exist in the MCMC sample
- Highest posterior density state
  tree/branches/parameters
  Might just have very good branch lengths, but otherwise be a relatively unlikely topology
- Highest posterior density tree
  (averaging branches/parameters)
  Hard to estimate if there are many different trees in the 95% credible interval of trees
- Median tree?
  Select the tree in the sample that minimizes the distance to the other trees using some metric.
- Maximum credibility tree?