Relaxed Bayesian phylogenetics

Molecular Clocks and Calibration

Alexei Drummond, alexei@cs.auckland.ac.nz
University of Auckland

Workshop on Molecular Evolution
Cesky Krumlov, 30th Jan 2015

1. BEAST
2. Tree Space
3. Bayesian phylogenetics
4. Clocks and calibrations
5. Relaxed phylogenetics
6. Total evidence and fossilized birth-death tree prior

BEAST

BEAST

BEAST focuses on **time-trees** (phylochronologies); both species trees and gene trees

Currently useful for

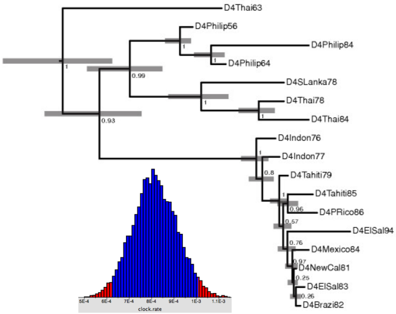
- Divergence time dating
- Estimating phylogenies under relaxed clock models
- Single population coalescent reconstruction
- Estimation of rates from viruses or ancient DNA
- Co-estimation of species trees and gene trees
- Automatic partitioning and substitution model selection

Working on

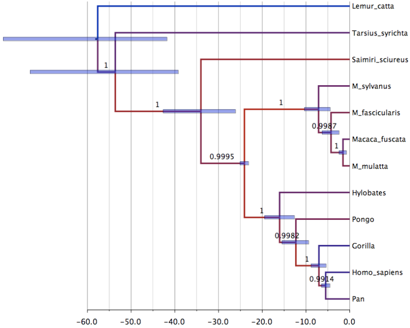
- More tree priors, more clock models, more substitution models
- More efficient tree sampling techniques (HMC)
- Phylodynamical models, host-pathogen co-phylogeny models

BEAST 1.4.8

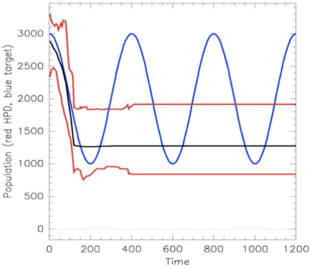
Rates/dates from serially sampled data



Relaxed phylogenetics



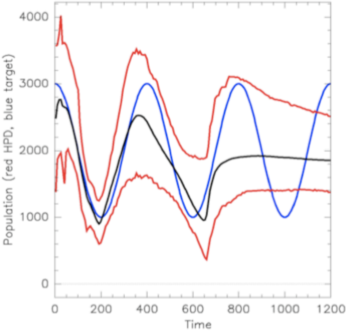
Estimating population size and changes

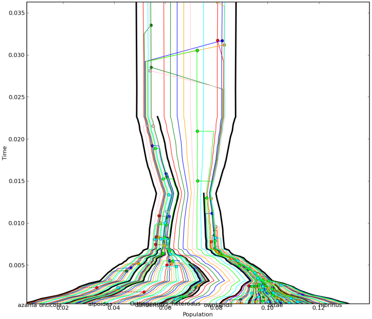


BEAST 1.6

Bayesian skyline plots and coalescent models with multiple loci

Coestimation of species tree and gene trees



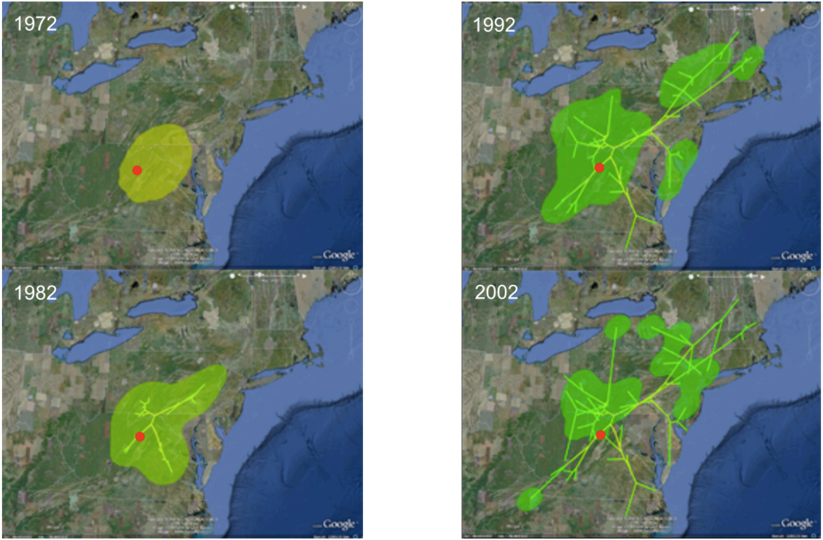


Bayesian skyride

Generalized partitioning

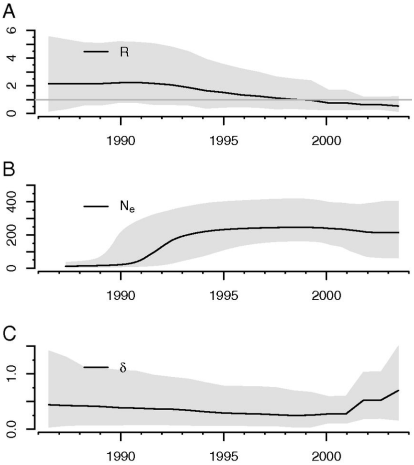
BEAST 1.8

Phylogeographic models

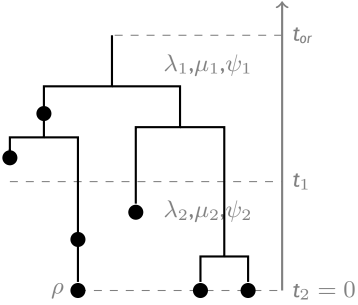


BEAST 2.2

Birth-death-skyline models



Fossilized birth-death models

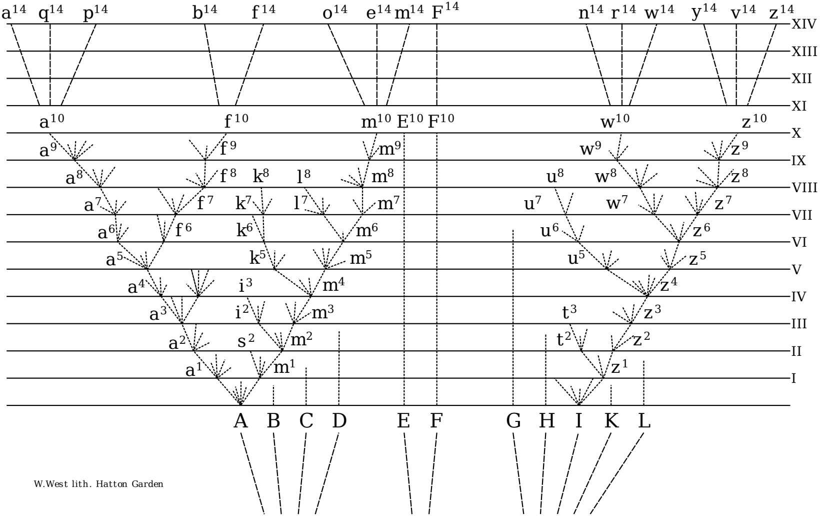


and lots of others (e.g. *Dirichlet process site partition model averaging*)

Tree Space

Darwin's Tree of Life

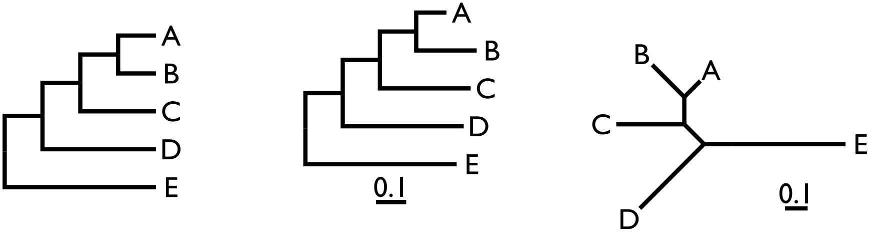
The only illustration in the *Origin of Species* (Darwin, 1859)


a^{14} q^{14} p^{14} b^{14} f^{14} o^{14} e^{14} m^{14} F^{14} n^{14} r^{14} w^{14} y^{14} v^{14} z^{14}
XIV
XIII
XII
XI
a^{10} f^{10} m^{10} E^{10} F^{10} w^{10} z^{10}
X
a^{9} f^{9} m^{9} w^{9} z^{9}
IX
a^{8} f^{8} k^{8} l^{8} m^{8} u^{8} w^{8} z^{8}
VIII
a^{7} f^{7} k^{7} l^{7} m^{7} u^{7} w^{7} z^{7}
VII
a^{6} f^{6} k^{6} m^{6} u^{6} z^{6}
VI
a^{5} k^{5} m^{5} u^{5} z^{5}
V
a^{4} i^{3} m^{4} z^{4}
IV
a^{3} i^{2} m^{3} t^{3} z^{3}
III
a^{2} s^{2} m^{2} t^{2} z^{2}
II
a^{1} m^{1} z^{1}
I
A B C D E F G H I K L
W.West lith. Hatton Garden

Types of phylogenies and representations

rooted trees

unrooted tree



(a) cladogram

(b) phylogram

(c) unrooted tree

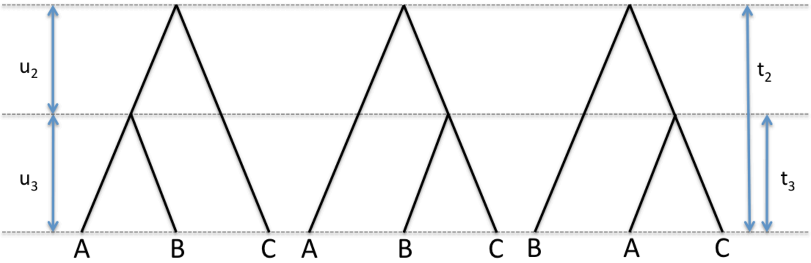
((((A, B), C), D), E);

((((A:0.1, B:0.2):0.12, C:0.3):0.123, D:0.4):0.1234, E:0.5);

branches (edges) and their lengths, nodes, tips (leaves)

The tip-labeled time-tree

A tip-labeled time-tree is described by a *tip-labeled ranked topology* of size k and *coalescent times*, $\mathbf{u} = \{u_2, \ldots, u_k\}$.



These time-trees of size 3 can be interpreted as describing the possible alternative evolutionary histories for three species or (uniparental) ancestries of the three individuals represented by the labeled tips.

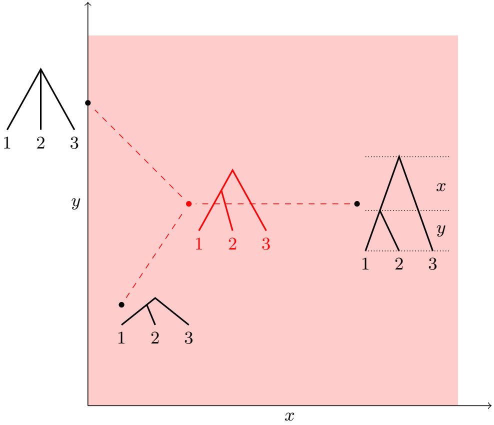


Figure: A Euclidean two-dimensional space representing the space of all possible time-trees for the topology ((1,2),3). There are two parameters, *x* and *y*, one for each of the two inter-coalescent intervals, the sum of which is the age of the root ($t_{root} = x + y$). Three trees are displayed, along with their arithmetic mean tree, also called the *centroid*. The dashed lines show the path connecting each of the three trees to the mean tree by the shortest distance (i.e. their deviations from the mean).

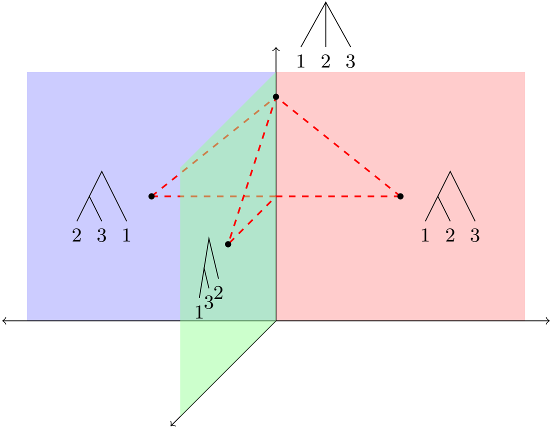


Figure: τ_3, the simplest non-trivial tree space (for time-trees), representing the space of time-trees for $n = 3$ taxa sampled contemporaneously. Each of the three non-degenerate tree topologies is represented by a two-dimensional Euclidean space (as illustrated in Figure 1) and these subspaces meet at a single shared edge representing the star tree, which is a one-dimensional subspace and thus has a single parameter (the age of the root). The dashed lines shows the paths of shortest distance between the four displayed trees.

Another space of tip-labeled time-trees of size 3

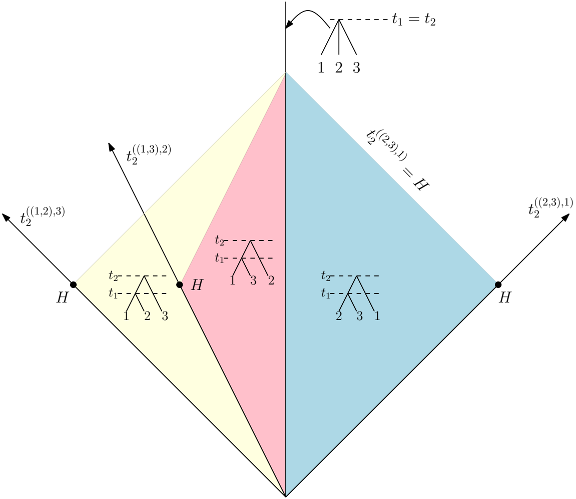


Figure: Space $\mathbb{T}_3$.

Figure from Gavruskin & Drummond (2014)

A space of tip-labeled time-trees of size 4

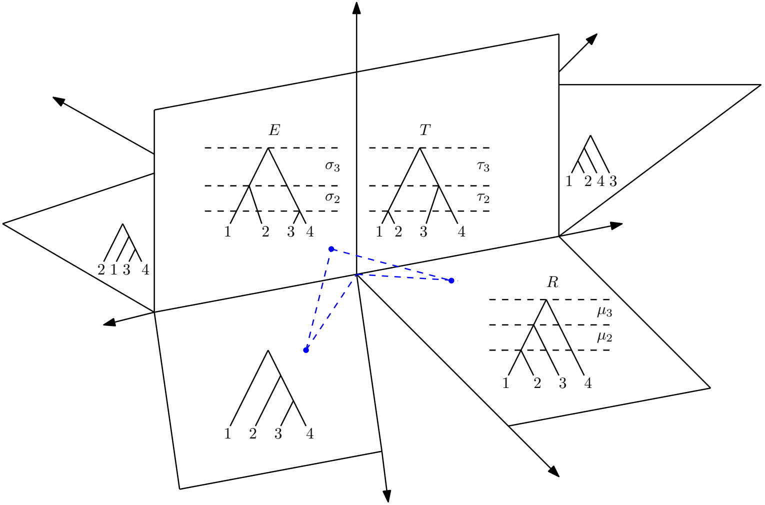
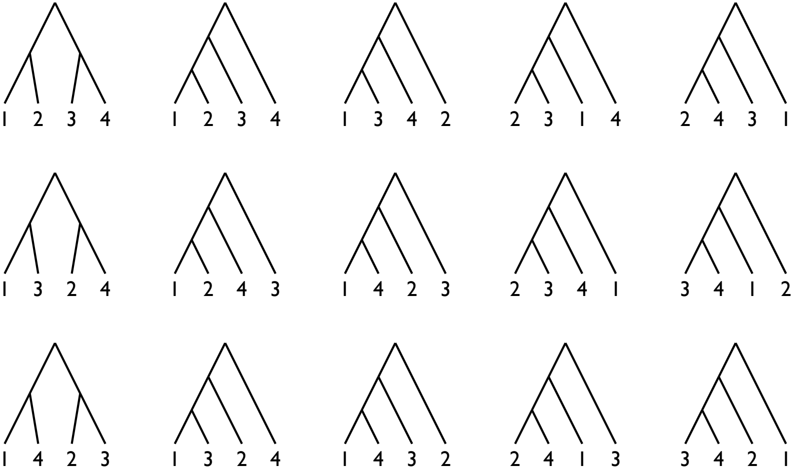


Figure: Three-dimensional projection of 4-dimensional τ-space 4.

Figure from Gavruskin & Drummond (2014)

Unranked tree topologies of size 4



How many trees are there?

For n species there are

$$T_n = 1 \times 3 \times 5 \times \cdots \times (2n-3) = \frac{(2n-3)!}{(n-2)!2^{n-2}}$$

rooted, tip-labelled binary trees:

n	#trees	
4	15	enumerable by hand
5	105	enumerable by hand on a rainy day
6	945	enumerable by computer
7	10395	still searchable very quickly on computer
8	135135	about the number of hairs on your head
9	2027025	greater than the population of Auckland
10	34459425	≈ upper limit for exhaustive search
20	8.20×10^{21}	≈ upper limit of branch-and-bound searching
48	3.21×10^{70}	≈ the number of particles in the Universe
136	2.11×10^{267}	number of trees to choose from in the "Out of Africa" data (Vigilant *et al.* 1991)

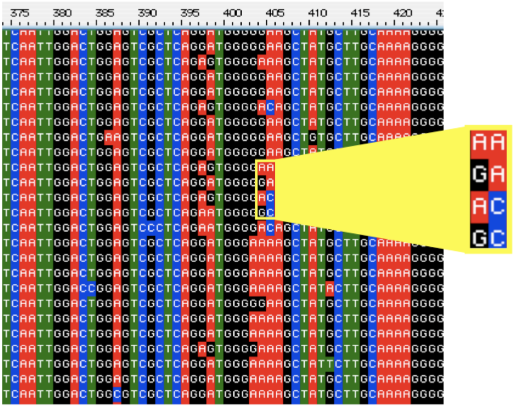
Counting different types of rooted trees

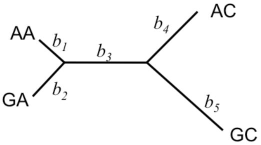
n	#shapes	#trees, $\vert\mathcal{T}_n\vert$	#ranked trees	#fully ranked trees
2	1	1	1	1
3	1	3	3	4
4	2	15	18	34
5	3	105	180	496
6	6	945	2700	11056
7	11	10395	56700	349504
8	23	135135	1587600	14873104
9	46	2027025	57153600	819786496
10	98	34459425	2571912000	56814228736

Table: The number of unlabeled rooted tree shapes, the number of labelled rooted trees, the number of labelled ranked trees (on contemporaneous tips), and the number of fully-ranked trees (on distinctly-timed tips) as a function of the number of taxa, *n*.

Bayesian phylogenetics

Felsenstein's likelihood (1981)





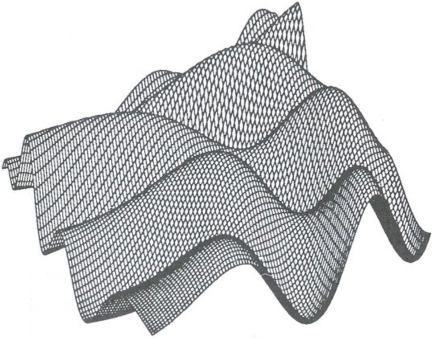
$$L(T) = Pr\{D|T, Q\}$$

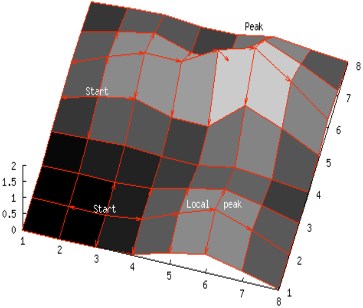
The probability of the data, $Pr\{D|T, Q\}$ can be efficiently calculated given a phylogenetic tree (T), and a **probabilistic model** of molecular evolution (Q).

In statistical phylogenetics, branch lengths are traditionally unconstrained.

Tree space as a hilly landscape

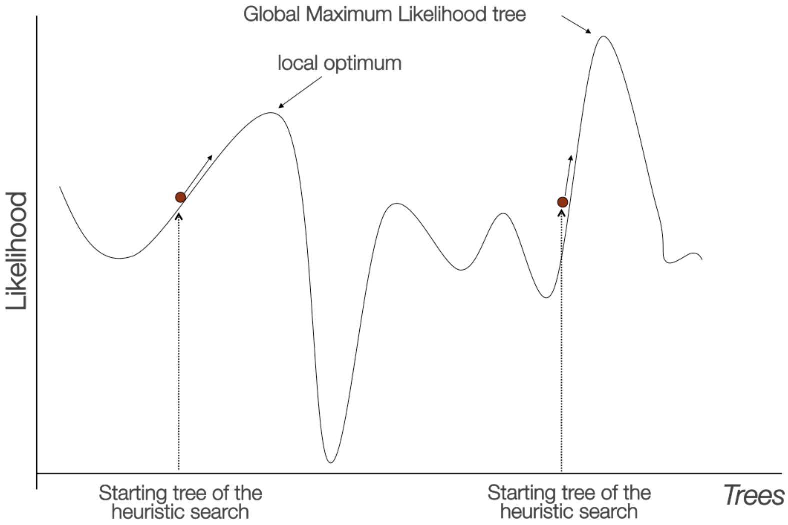
The space of all possible trees can be visualized as a hilly landscape. Nearby points in this landscape represent similar trees, and the height of the landscape is the probability of the tree at that point.





- This space can be **sampled** in a Bayesian analysis with MCMC
- The peak can be identified by a **search algorithm** in the context of maximum likelihoods

Local tree search and multiple optima


Global Maximum Likelihood tree
local optimum
Likelihood
Starting tree of the heuristic search
Starting tree of the heuristic search
Trees

Bayes rule in statistics

$$Pr(\theta|D) = \frac{Pr(D|\theta)Pr(\theta)}{Pr(D)}$$

where

- $P(D|\theta)$ is the likelihood,
- $Pr(\theta)$ is the prior distribution and
- $Pr(\theta|D)$ is the posterior distribution.
- $Pr(D)$ is the marginal likelihood of the data.

Bayes rule in phylogenetics

$$p(T, Q|D) = \frac{Pr\{D|T, Q\}p(T)p(Q)}{Pr\{D\}}$$

where

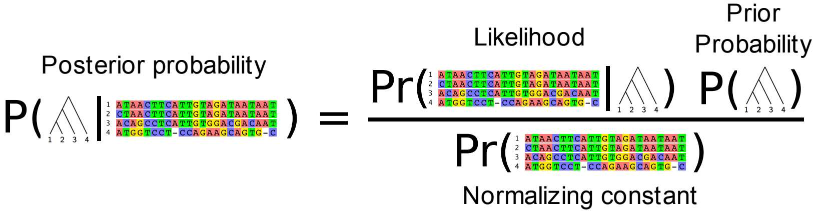
- $Pr(D|T, Q)$ is Felsenstein's likelihood,
- $p(T)$ is the prior distribution on phylogenetic trees,
- $p(Q)$ is the prior distribution on the model of evolution and
- $p(T, Q|D)$ is the posterior distribution
- $Pr(D)$ is the marginal likelihood of the data.

Bayesian reconstruction of phylogenetic trees

Yang & Rannala (1997), Mau, Newton & Larget (1998)

In the context of Bayesian phylogenetics, what we want to compute is the **probability of the tree** given the data.

We can compute that from the **likelihood** using **Bayes Theorem**:

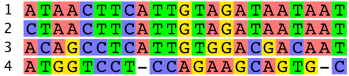

Posterior probability
Likelihood
Prior Probability
P(
Pr(
1 ATAACTTCATTGTAGATAATAAT
2 CTAACTTCATTGTAGATAATAAT
3 ACAGCCTCATTGTGGACGACAAT
4 ATGGTCCT-CCAGAAGCAGTG-C
1 2 3 4
=
Normalizing constant

This is known as the **Posterior probability** of the tree. Another method of reconstructing the evolutionary history is then to find the tree that has the **Maximum Posterior probability**.

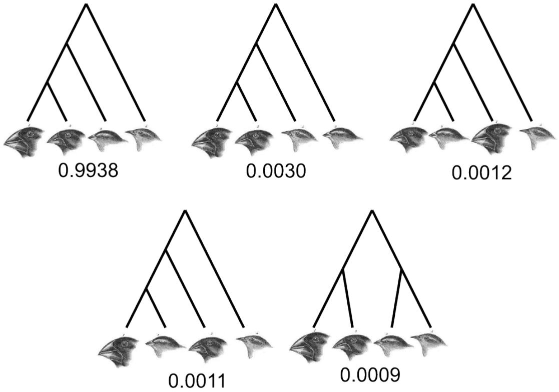
Bayesian Phylogenetics

- 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.
- Often a more specific hypothesis 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.

The Posterior Distribution on Darwin’s Finches

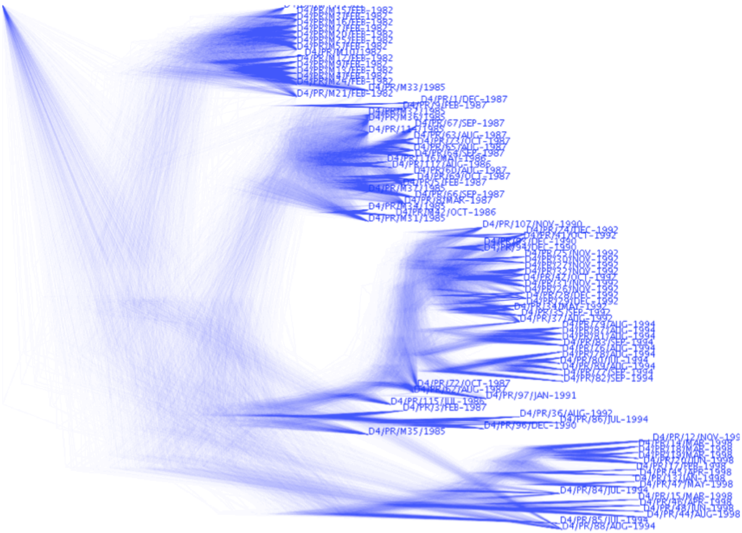


```
1 ATAACTTCATTGTAGATAATAAT
2 CTAACTTCATTGTAGATAATAAT
3 ACAGCCTCATTGTGGACGACAAT
4 ATGGTCCT-CCAGAAGCAGTG-C
```



This posterior probability distribution was computed using an algorithm called **Markov chain Monte Carlo** implemented in the BEAST software package (Drummond & Rambaut, 2007).

The posterior distribution for larger trees



Elaborating the model

Basic model: (posterior proportional to likelihood × prior)

$$p(T|D) \propto \Pr\{D|T\}p(T)$$

Substitution model estimation:

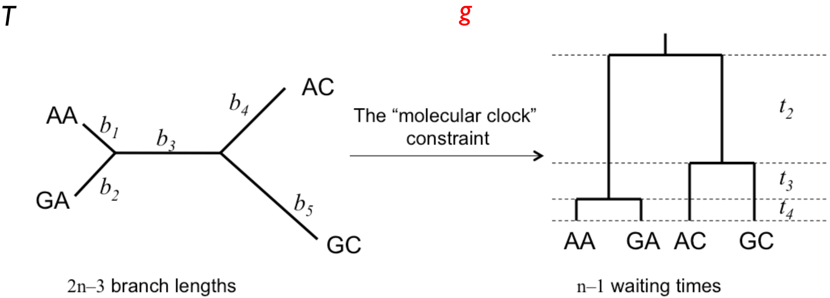
$$p(T, Q|D) \propto \Pr\{D|T, Q\}p(T)p(Q)$$

Substitution model and parametric tree prior:

$$p(T, Q, \theta|D) \propto \Pr\{D|T, Q\}p(T|\theta)p(Q)p(\theta)$$

Clocks and calibrations

The molecular clock constraint

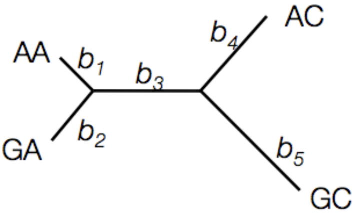


Standard BEAST model:

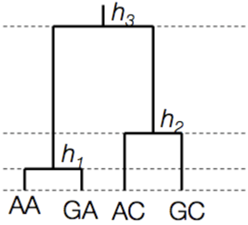
$$p(g, Q, \theta|D) \propto \Pr\{D|g, Q\}p(g|\theta)p(Q)p(\theta)$$

The joint posterior probability of the **rooted** time-tree (*g*) the substitution matrix (*Q*) and the tree prior parameters (θ) is sampled using Markov chain Monte Carlo (Drummond *et al*, 2002; 2006)

Model assumptions



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



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

Calibration via a global molecular clock

Basic model: (Tree in expected substitutions per site)

$$p(g, \theta|D) \propto \Pr\{D|g\}p(g|\theta)p(\theta)$$

Fix (i.e. condition on) the global rate to μ:

$$p(g, \theta|D) \propto \Pr\{D|\mu \times g\}p(g|\theta)p(\theta)$$

Estimate the global rate:

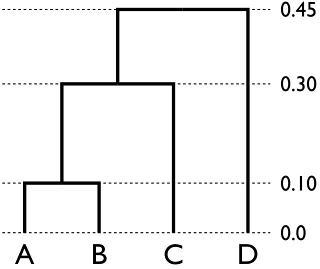
$$p(g, \mu, \theta|D) \propto \Pr\{D|\mu \times g\}p(g|\theta)p(\theta)p(\mu)$$

In the models above the parameters related to the details of the substitution process (Q) have been suppressed for simplicity.

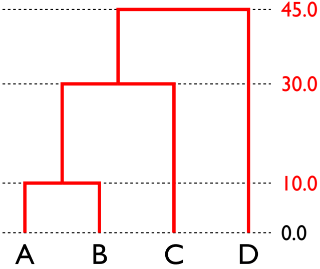
Genetic distance = rate × time

Strict molecular clock

$$T = \mu \times g$$



$= \quad 0.01 \quad \times$

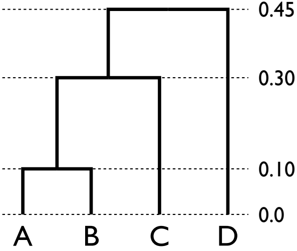


“substitution tree”

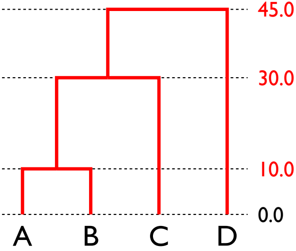
evolutionary rate
substitutions / site / unit time

time tree

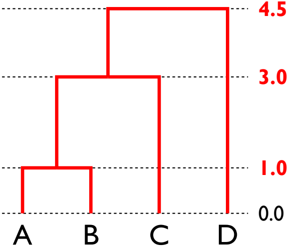
Non-identifiability of rate and times



$= 0.01 \times$



$= \mathbf{0.1} \times$



“substitution tree”

evolutionary rate
substitutions / site / unit time

time tree

A simple calibration is not simple

Consider the simplest type of calibration to admit uncertainty: the placement of an upper and a lower limit on the age of a single calibrated divergence (h_C) in the tree:

$$f(h_C) = \begin{cases} 1/(u-l) & l \leq h_C \leq u \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

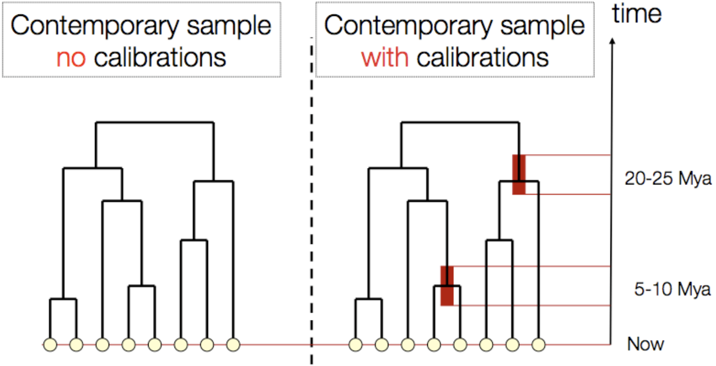
This calibration already has two quite distinct interpretations. One interpretation is that the resulting marginal prior distribution on the calibrated divergence should obey the tree process prior (f_G, e.g. Yule or Birth-death) but be **constrained** to be within the upper and lower bounds:

$$\rho_G(\mathbf{g}|\theta) \propto f_G(\mathbf{g}|\theta) f(h_C), \tag{2}$$

Alternatively, the marginal prior of h_C is uniform and conditioned on:

$$\rho_G(\mathbf{g}|\theta) \propto f_G(\mathbf{g}_{-h_C}|\theta, h_C) f(h_C), \tag{3}$$

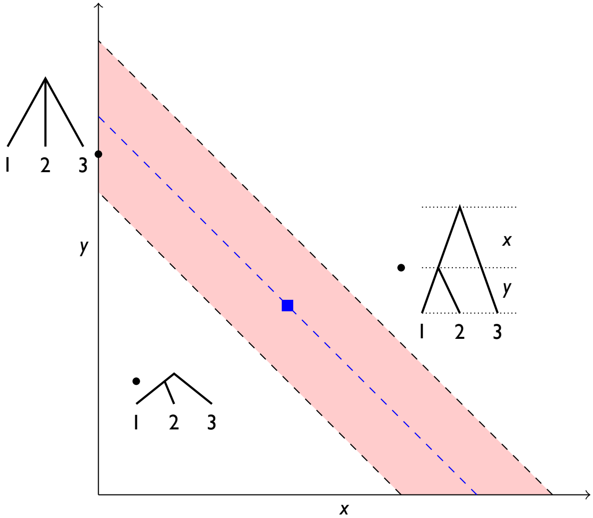
Absolute time via calibrations



Let $\rho_G(g|\theta)$ be "calibrated" $f_G(g|\theta)$ and estimate the rate, μ:

$$p(\mu, g, \theta|D) \propto \Pr\{D|\mu \times g\}\rho_G(g|\theta)f_N(\theta)f_M(\mu)$$

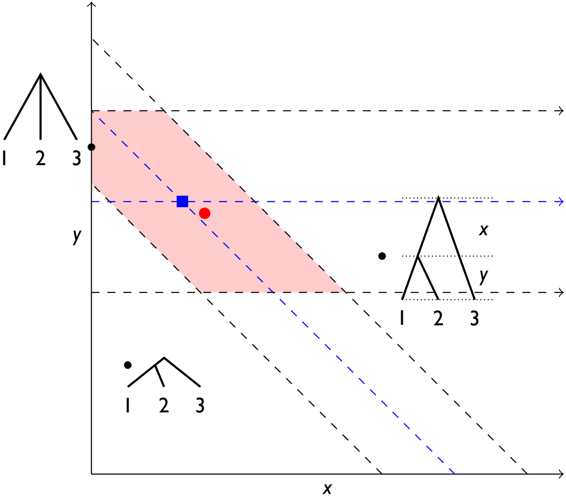
Calibrating tree space



Single calibration on the root height: $8 < x + y < 12$

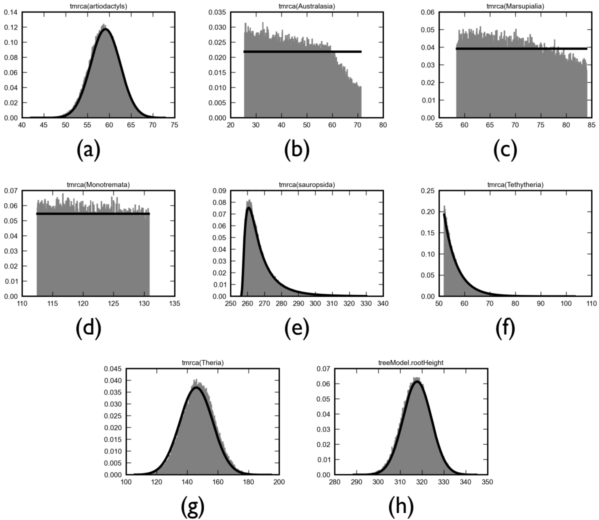
Calibrating tree space

Two calibrations is even less simple!



First calibration: $8 < x + y < 12$

Second calibration: $5 < y < 10$

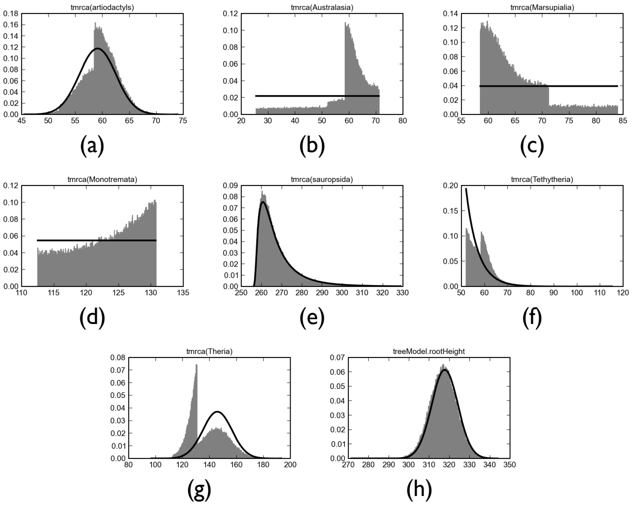


(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure: A simple construction of calibrated tree prior: $\rho_G(g) \propto f_G(g) \times \prod_{i=1}^{k} f_i(s_i)$. Where $f_i()$ is the univariate "calibration density" for the divergence time of the i'th calibrated node in the tree. Monophyly is enforced for each calibrated node.



(a) (b) (c)

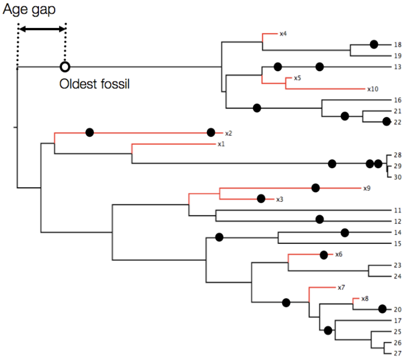
(d) (e) (f)

(g) (h)

Figure: The marginal prior distributions that result from BEAST (gray) versus calibration densities (black) specified for the calibrated nodes from **[?]**. The marginal prior distributions were obtained from a MCMC run using the prior only.

How do I pick the calibration density?

Modeling the Fossil Age Gap



What is the probability distribution of the age gap?

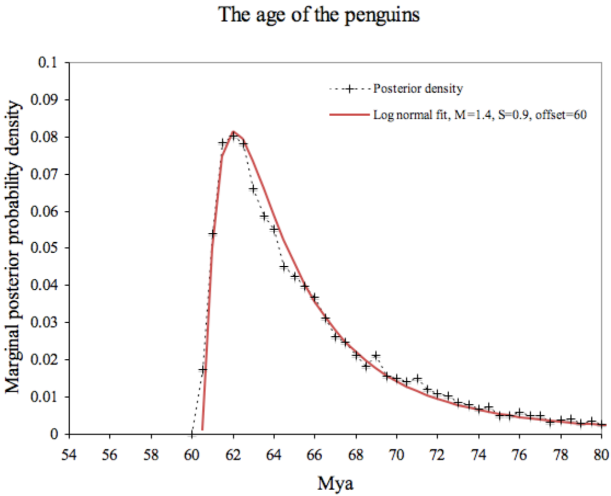
60-61.5 Myr penguin

Prof. Ewan Fordyce with reconstruction of *Waimanu tuatahi*

Current day penguin species: 20

Number of independent penguin fossils with good geological age from all ages: 20-60

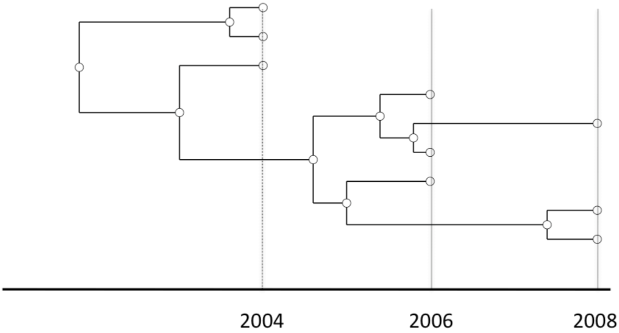
The posterior estimate of the age of penguins



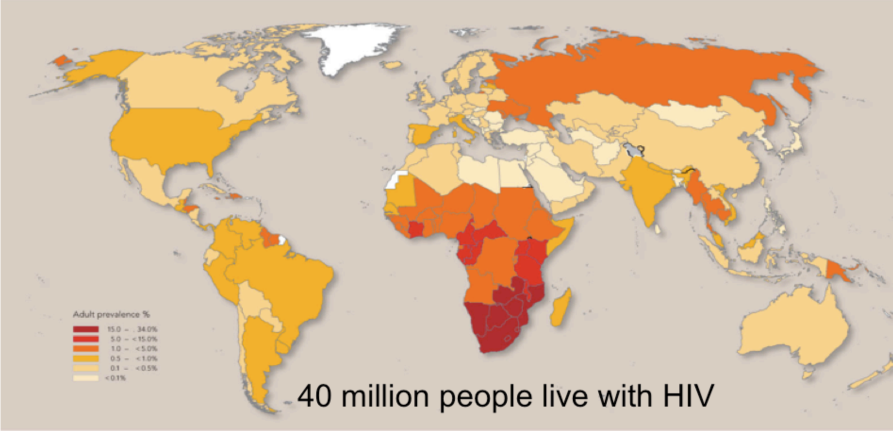
Evolution is happening right now!

Rodrigo and Felsenstein, 1999; Drummond *et al*, 2002

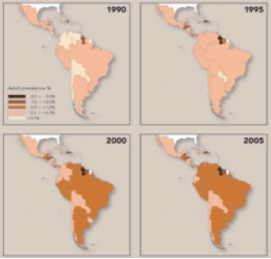
Many pathogens, such as HIV, Hepatitis C and Influenza A, evolve very rapidly, so that samples of the virus population from different times directly reveal evolutionary change.

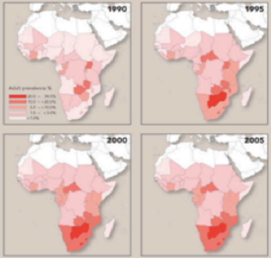


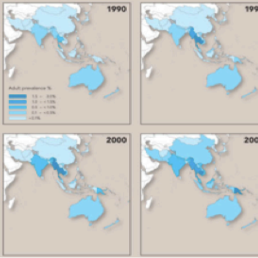
In fact it becomes possible to **calibrate** the tree and thus place the tree on a time scale - by constraining the tips to known sampling times



40 million people live with HIV

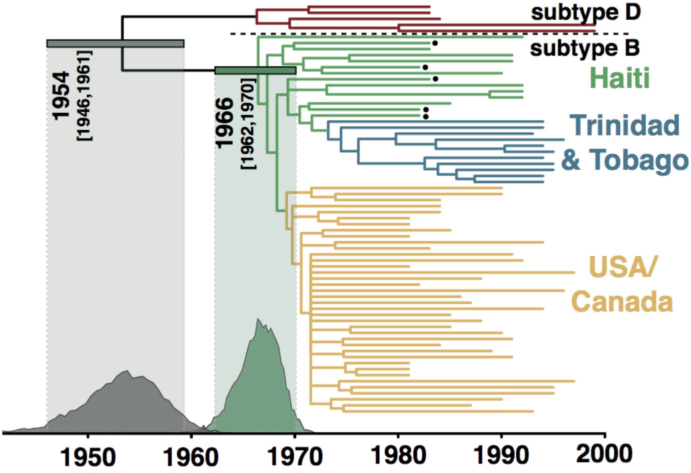






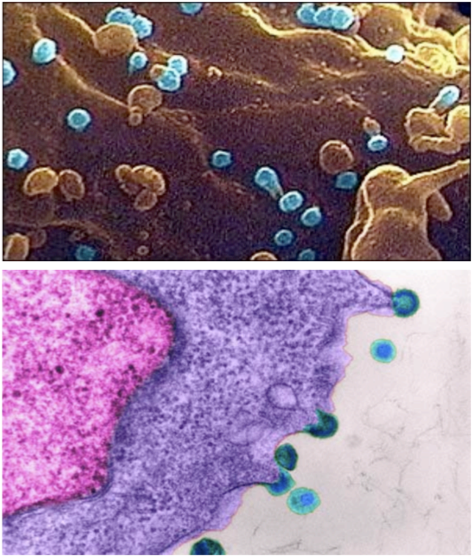
A calibrated phylogenetic inference

Origin of HIV Epidemic in the Americas, Gilbert *et al* (2007)



A phylogenetic reconstruction of samples of HIV-1 virus. Each degree one node represents a single infected individual from whom a blood sample has been taken.

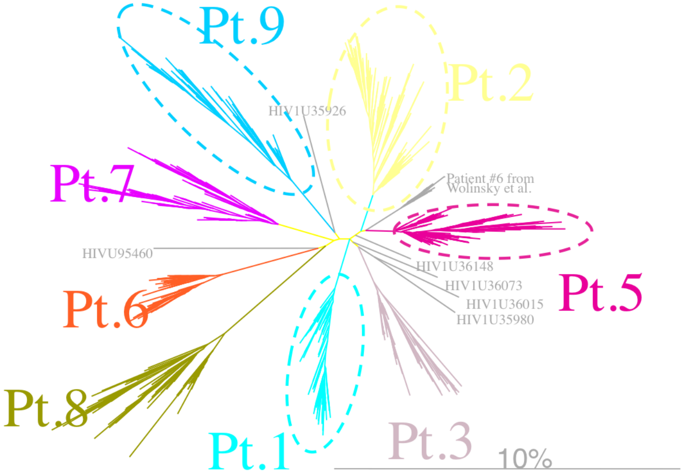
Human immunodeficiency virus type 1(HIV-1)



A single HIV-1 infected person has at least $10^7 - 10^8$ infected cells, with each infected cell producing $\sim 10^3$ viral particles during its life time.

A tree of HIV sequences from 9 infected patients

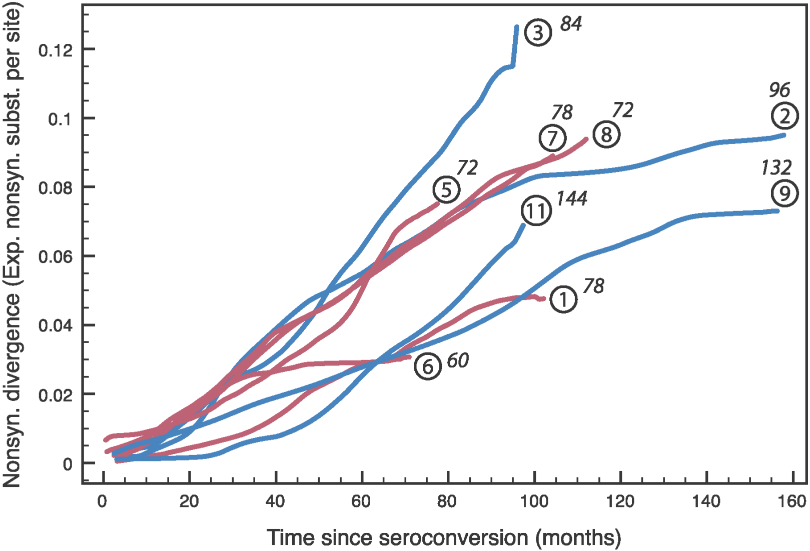
Shankarappa *et al* (1999)



A phylogenetic reconstruction of samples of HIV-1 virus. Each degree one node represents a single virus particle isolated from a blood sample of one of 9 patients.

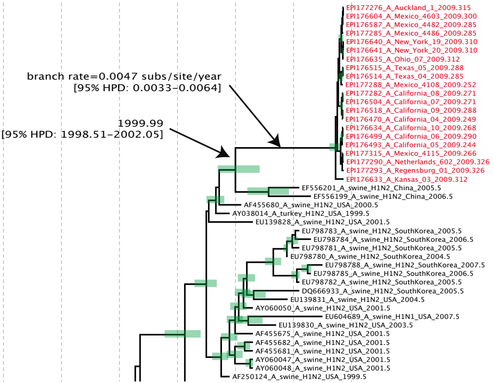
Estimated accumulation of evolutionary change

Lemey *et al* (2008)

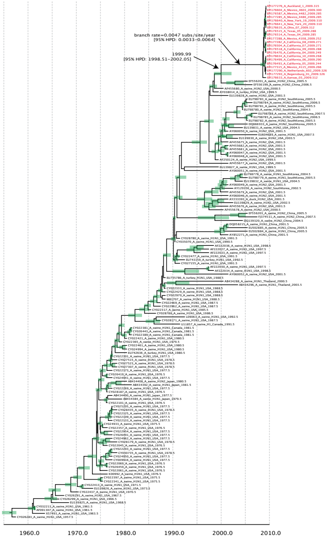


On the Origin of 2009 H1N1 Swine Flu outbreak

http://tree.bio.ed.ac.uk/groups/influenza/



EPI177276_A_Auckland_1_2009.315
EPI176604_A_Mexico_4603_2009.300
EPI176587_A_Mexico_4482_2009.285
EPI177285_A_Mexico_4486_2009.285
EPI176640_A_New_York_19_2009.310
EPI176641_A_New_York_20_2009.310
EPI176635_A_Ohio_07_2009.312
EPI176515_A_Texas_05_2009.288
EPI176514_A_Texas_04_2009.285
EPI177288_A_Mexico_4108_2009.252
EPI177282_A_California_08_2009.271
EPI176504_A_California_07_2009.271
EPI176518_A_California_09_2009.288
EPI176470_A_California_04_2009.249
EPI176634_A_California_10_2009.268
EPI176499_A_California_06_2009.290
EPI176493_A_California_05_2009.244
EPI177315_A_Mexico_4115_2009.266
EPI177290_A_Netherlands_602_2009.326
EPI177293_A_Regensburg_01_2009.326
EPI176633_A_Kansas_03_2009.312
branch rate=0.0047 subs/site/year
[95% HPD: 0.0033–0.0064]
1999.99
[95% HPD: 1998.51–2002.05]
EF556201_A_swine_H1N2_China_2005.5
EF556199_A_swine_H1N2_China_2006.5
AF455680_A_swine_H1N2_USA_2000.5
AY038014_A_turkey_H1N2_USA_1999.5
EU139828_A_swine_H1N2_USA_2001.5
EU798783_A_swine_H1N2_SouthKorea_2005.5
EU798784_A_swine_H1N2_SouthKorea_2006.5
EU798781_A_swine_H1N2_SouthKorea_2005.5
EU798780_A_swine_H1N2_SouthKorea_2004.5
EU798788_A_swine_H1N2_SouthKorea_2007.5
EU798785_A_swine_H1N2_SouthKorea_2006.5
EU798782_A_swine_H1N2_SouthKorea_2005.5
DQ666933_A_swine_H1N2_SouthKorea_2005.5
EU139831_A_swine_H1N2_USA_2004.5
AY060050_A_swine_H1N2_USA_2001.5
EU604689_A_swine_H1N1_USA_2007.5
EU139830_A_swine_H1N2_USA_2003.5
AF455675_A_swine_H1N2_USA_2001.5
AF455682_A_swine_H1N2_USA_2001.5
AF455681_A_swine_H1N2_USA_2001.5
AY060047_A_swine_H1N2_USA_2001.5
AY060048_A_swine_H1N2_USA_2001.5
AF250124_A_swine_H1N2_USA_1999.5

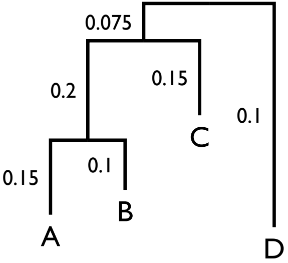


Relaxed phylogenetics

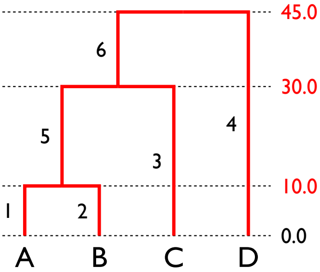
Genetic distance = rate × time

Relaxed molecular clock

$$T = \vec{\mu} \star \mathbf{g}$$



$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$



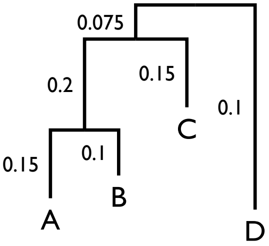
“substitution tree”

evolutionary rates

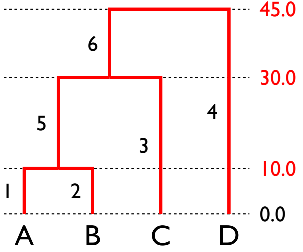
substitutions / site / unit time

time tree

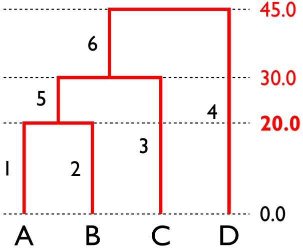
Nonidentifiability in the relaxed clock



$$= \begin{pmatrix}0.015\\0.01\\0.005\\0.01\\0.01\\0.005\end{pmatrix} \star$$



$$= \begin{pmatrix}\mathbf{0.0075}\\\mathbf{0.005}\\0.005\\0.01\\\mathbf{0.02}\\0.005\end{pmatrix} \star$$

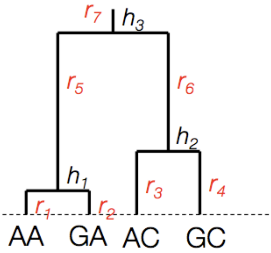


“substitution tree”

evolutionary rates
substitutions / site / unit time

time tree

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

$$r \sim Exp(\lambda)$$

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

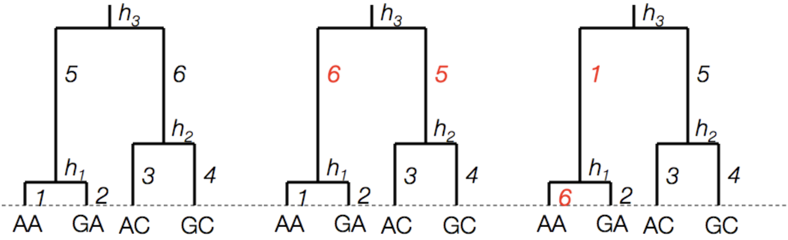
$$r \sim Gamma(\alpha, \beta)$$

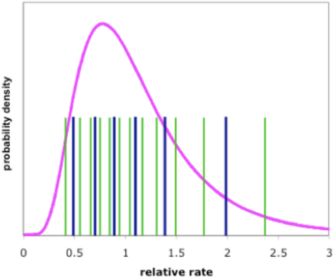
We introduce a relaxed clock model in which there is no prior correlation between child and parent rates

“Un-correlated” or “memory-less” relaxed clocks

ML

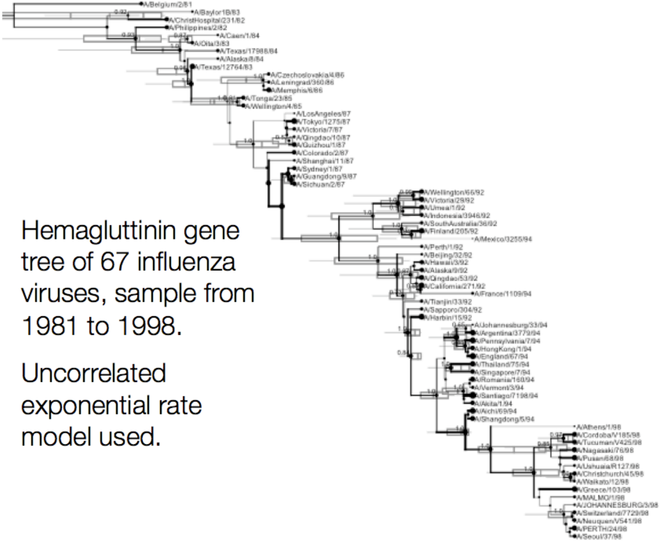
Sampling branch rates using MCMC





1. Rates are summarized into 2n-2 rate categories (e.g. blue is 6 categories; green is 12 categories).
2. Rates categories are sampled during MCMC by two operators:
    1. Random walk operator
    2. Swap operator
3. For purposes of topology changes, rate categories are associated with child node.

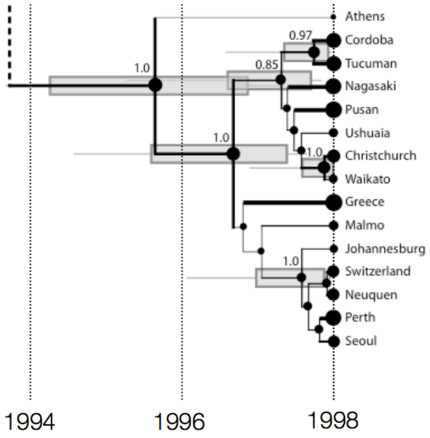
Influenza A gene tree estimated by relaxed molecular clock



1. Hemagluttinin gene tree of 67 influenza viruses, sample from 1981 to 1998.

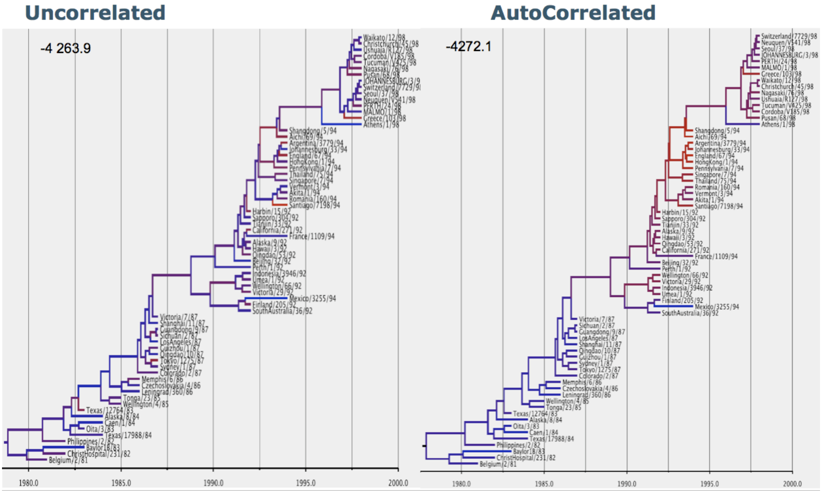
2. Uncorrelated exponential rate model used.

Influenza A gene tree estimated by relaxed molecular clock



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

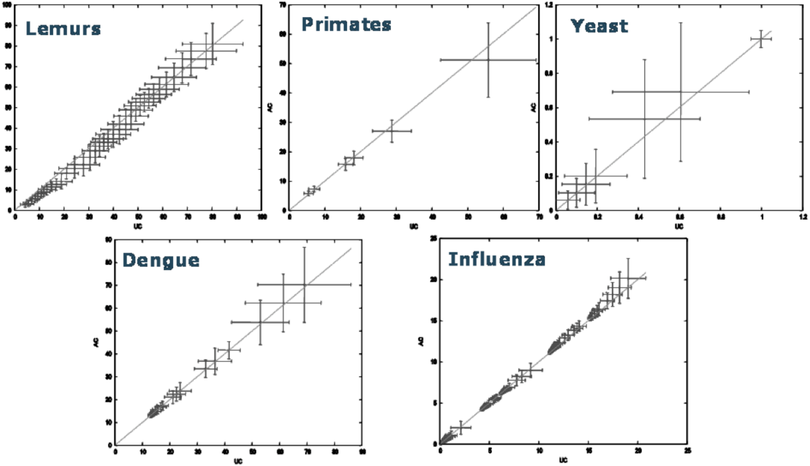
Influenza trees under different relaxed clock models

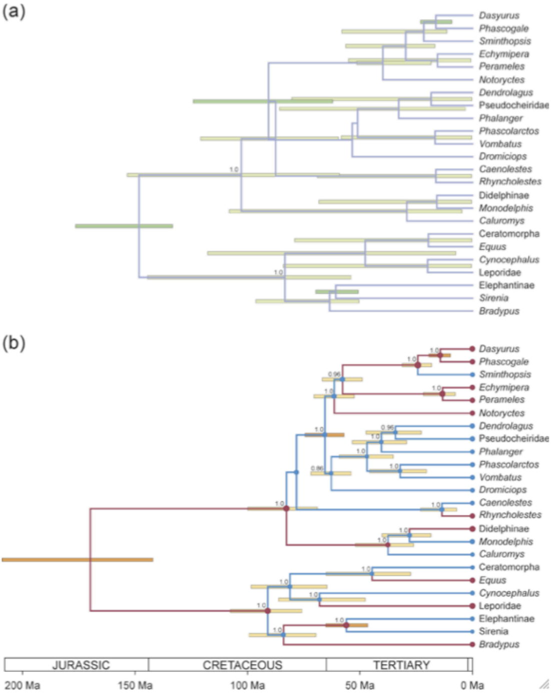


Uncorrelated

AutoCorrelated

UC versus AC on five data sets



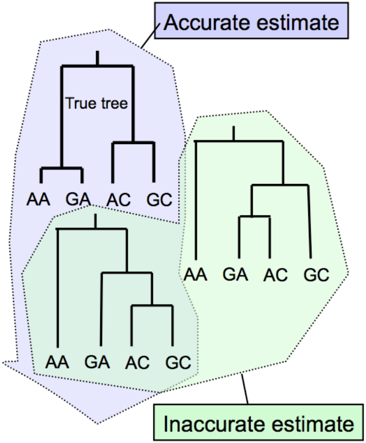


Prior versus Posterior

Marsupials example (24 taxa, 5658 nucleotides)

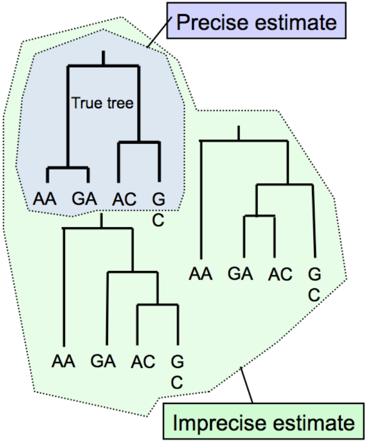
Accuracy in Bayesian Phylogenetics

- 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?



Precision in Bayesian Phylogenetics

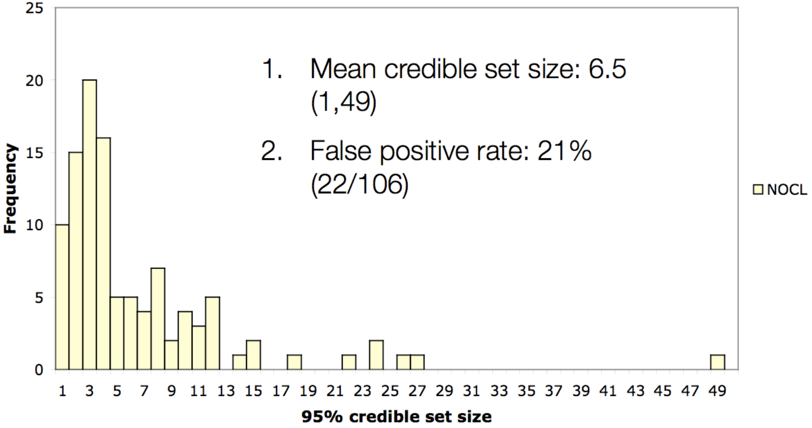
- The precision of an estimate can be described by how much is excluded.
- How small is the 95% credible set of trees?



Testing Accuracy and Precision with **real data**

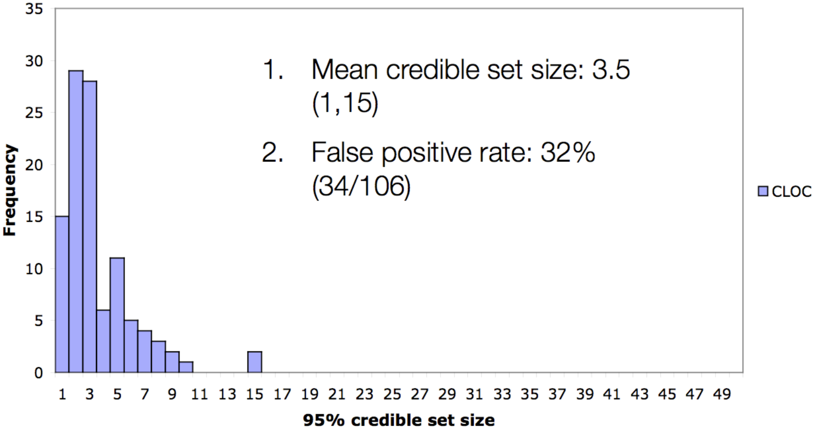
- Used 106 genes from 8 species of yeast (Rokas *et al*, 2003) and 4 other “phylogenomic” data sets
- For each gene used both MrBayes and BEAST to estimate phylogeny and 95% credible set
- Assumed true tree is the tree estimated using all the concatenated data set.
- Tabulated number of trees in credible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

Rokas data: MrBayes tree estimates



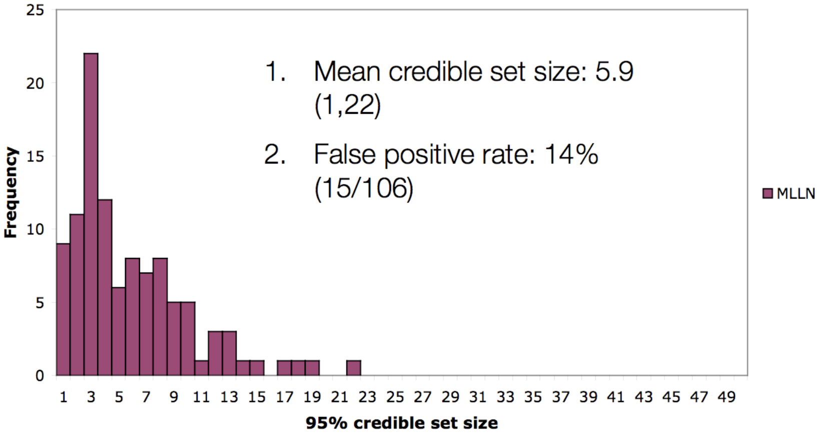
1. Mean credible set size: 6.5 (1,49)
2. False positive rate: 21% (22/106)

Rokas data: Strict clock tree estimates from BEAST



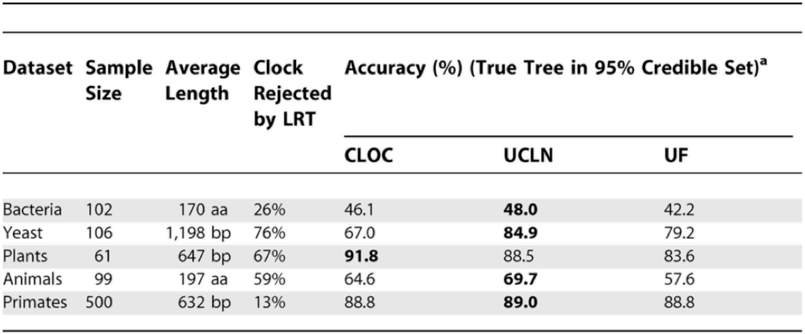
1. Mean credible set size: 3.5 (1,15)
2. False positive rate: 32% (34/106)

Rokas data: Relaxed clock tree estimates from BEAST



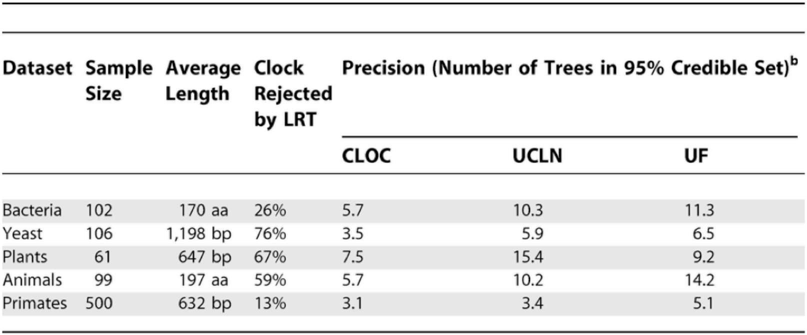
1. Mean credible set size: 5.9 (1,22)
2. False positive rate: 14% (15/106)

Summary of Bayesian Accuracy on five large data sets



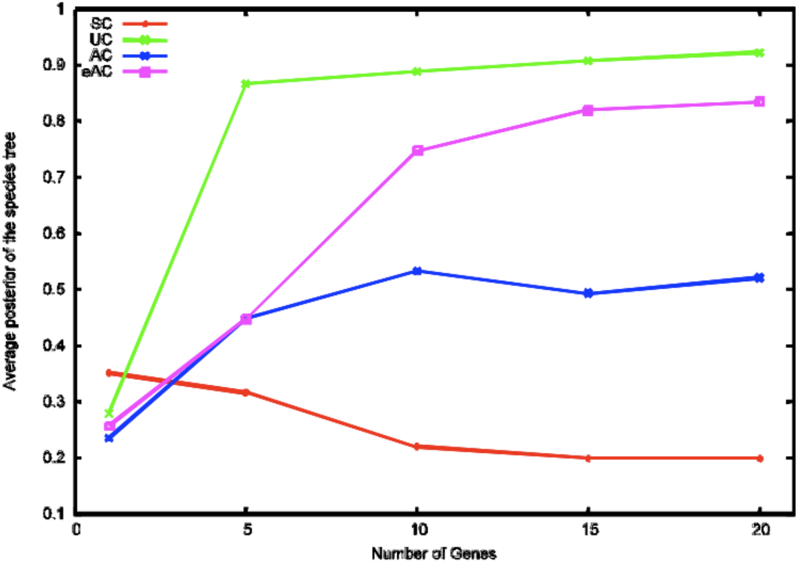
Dataset	Sample Size	Average Length	Clock Rejected by LRT	Accuracy (%) (True Tree in 95% Credible Set)[a]		
				CLOC	UCLN	UF
Bacteria	102	170 aa	26%	46.1	**48.0**	42.2
Yeast	106	1,198 bp	76%	67.0	**84.9**	79.2
Plants	61	647 bp	67%	**91.8**	88.5	83.6
Animals	99	197 aa	59%	64.6	**69.7**	57.6
Primates	500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

Summary of Bayesian Precision on five large data sets

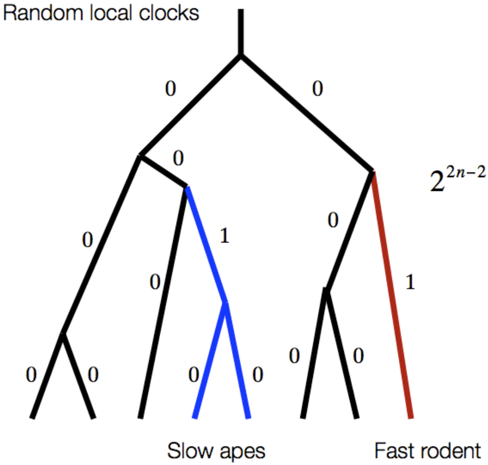


Dataset	Sample Size	Average Length	Clock Rejected by LRT	Precision (Number of Trees in 95% Credible Set)[b]		
				CLOC	UCLN	UF
Bacteria	102	170 aa	26%	5.7	10.3	11.3
Yeast	106	1,198 bp	76%	3.5	5.9	6.5
Plants	61	647 bp	67%	7.5	15.4	9.2
Animals	99	197 aa	59%	5.7	10.2	14.2
Primates	500	632 bp	13%	3.1	3.4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

Increasing the length of the sequence



Random local molecular clocks



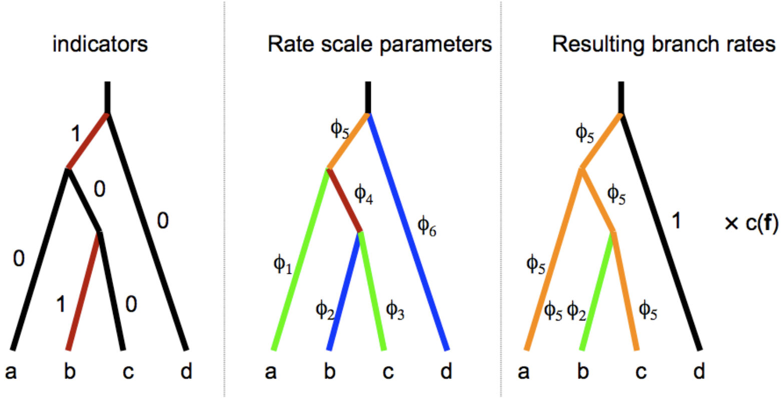
Random local clocks

Slow apes

Fast rodent

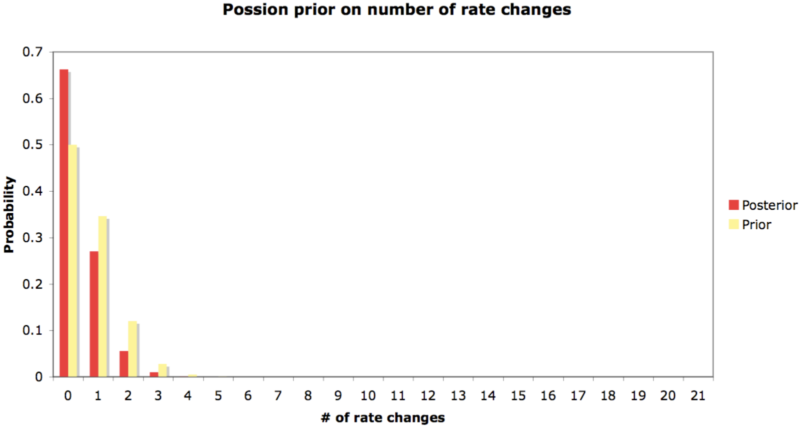
2^{2n-2} local molecular clock models

Random local molecular clocks

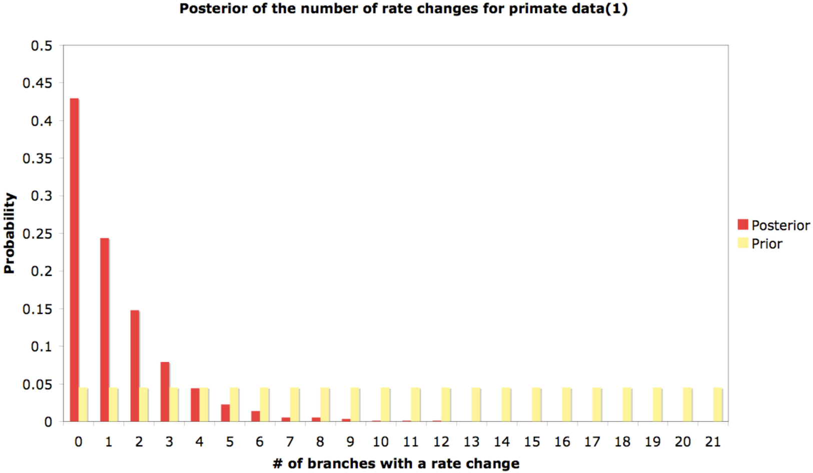


Red/Orange fast, Green/Blue slow

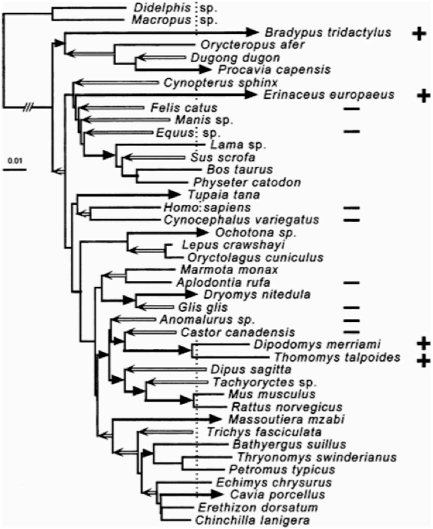
Primate data set (Poisson prior on # rate changes)



Primate data set (Uniform prior on # rate changes)



Rodents (1+2 codon positions from 3 nuclear genes)

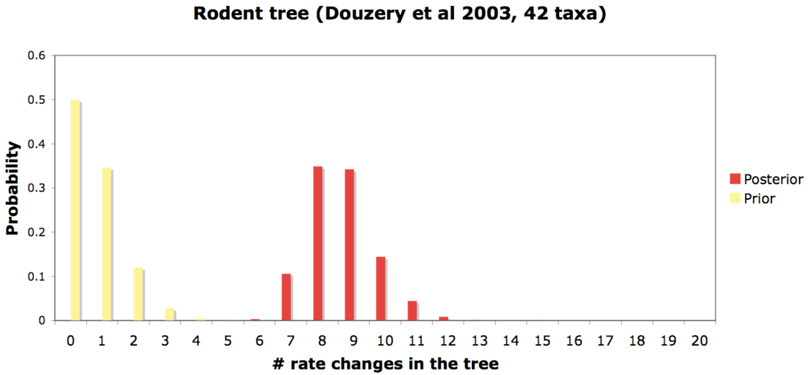


82 branches

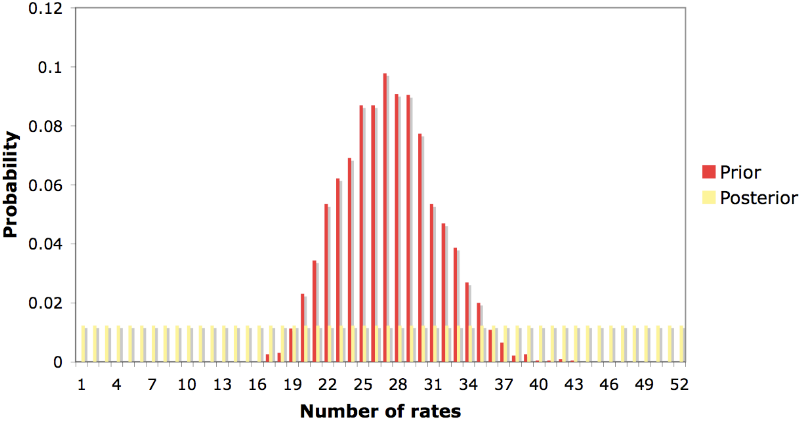
38 rate changes according to Douzery et al 2003

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).

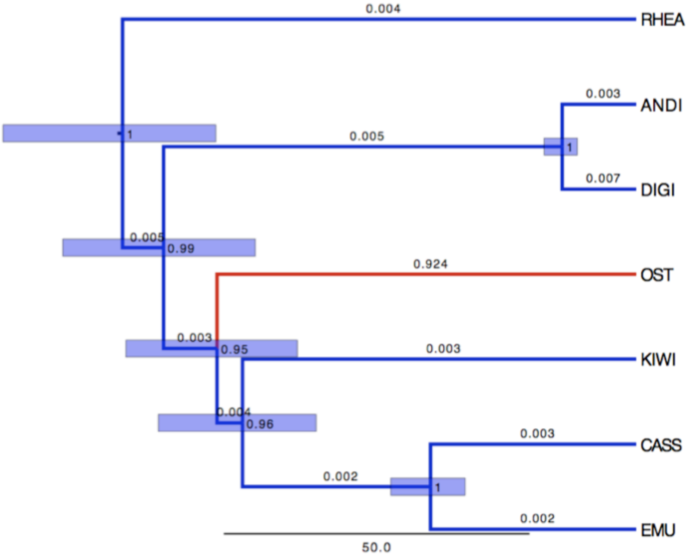
Rodent data set (Poisson prior on # rate changes)



Rodents data set (Uniform prior on # rate changes)



Ratite relaxed clock on full mitochondrial sequences



Fossilized birth-death trees

Slides by Alexandra Garyushkina
sasha.gavryushkina@auckland.ac.nz

Total evidence dating with fossilized birth-death tree prior

- We have molecular data of extant species, morphological data of extant and fossil species and geological ages of fossils (or geological age intervals).
- We want to utilise all this data to learn about evolutionary history of organisms, divergence dates and macroevolutionary parameters.
- Our preferred method is Bayesian phylogenetic inference.

The calibration method

- An important problem that still requires attention is the divergence dating.
- A common practice in estimating divergence times is using calibration methods when times of internal nodes in the tree are calibrated using fossil records.
- This method has a few drawbacks:
    - usually only the oldest fossil in the clade is used,
    - ad hoc calibration densities,
    - using calibration densities modifies the tree prior,
    - due to computational problems just a few calibration nodes can be used (Heled and Drummond 2012).

Alternative methods

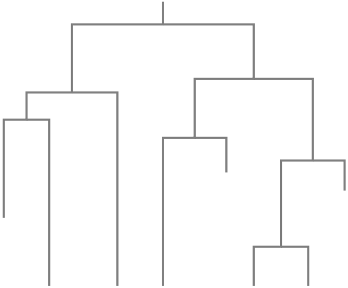
Two main features of the new methods:

- explicit modelling of fossilization events as a part of the tree branching process (Pyron 2011, Heath *et al* 2014),
- utilising all existing data (total evidence) in a joint inference (Pyron 2011, Ronquist *et al* 2012).

Total evidence

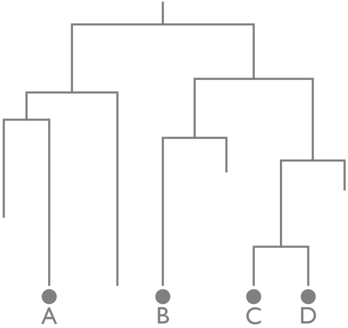
- Pyron 2011 uses Lewis MK model to model evolution of morphological characters.
- Ronquist *et al* 2012, Wood *et al* 2012.
- They used Yule or Uniform tree prior.

Birth-death-sampling models



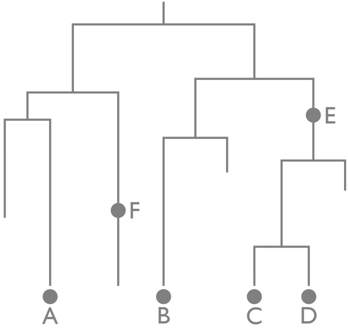
Full tree

Birth-death-sampling models



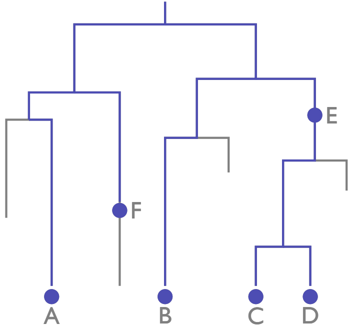
Full tree

Birth-death-sampling models



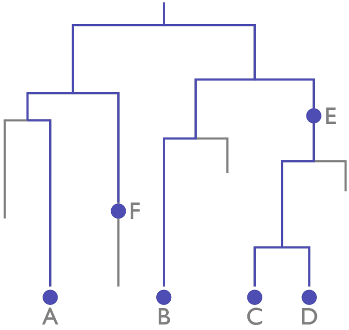
Full tree

Birth-death-sampling models

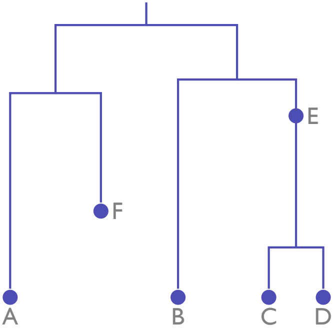


Full tree

Birth-death-sampling models



Full tree



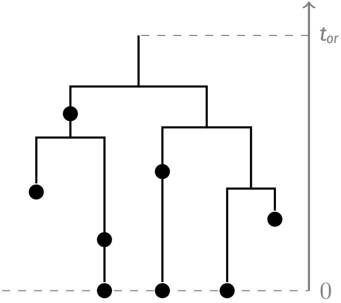
Sampled tree

Fossilized birth-death model (FBD)

Stadler 2010, Heath *et al* 2014.

The process starts at time $t_{or} > 0$ and ends at time zero (present time).

- birth rate λ
- death rate μ
- sampling rate ψ
- sampling at present probability ρ



Sampled tree

Model parameters: $\eta = (t_{or}, \lambda, \mu, \psi, \rho)$.
All the parameters are identifiable.

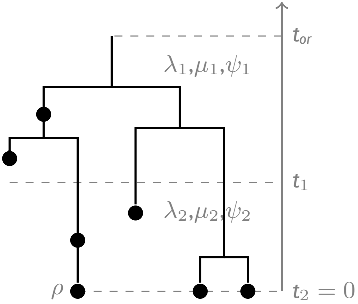
Fossilized birth-death skyline model (FBD skyline)

Stadler and Kühnert *et al* 2012, Gavryushkina *et al* 2014.

There are k time intervals and parameters remain constants within the intervals but may vary from one interval to another

- birth rates $\lambda_1, \ldots, \lambda_k$
- death rates $\mu_1, \ldots, \mu_k$
- sampling rates $\psi_1, \ldots, \psi_k$
- sampling at time t_k (present) probability ρ

Model parameters:
$\eta = (t_{or}, \bar{\lambda}, \bar{\mu}, \bar{\psi}, \rho)$



Sampled tree

Heath *et al* approach and its extensions

Heath *et al* 2014 first used FBD model to infer divergence times of bears in a Bayesian MCMC framework:

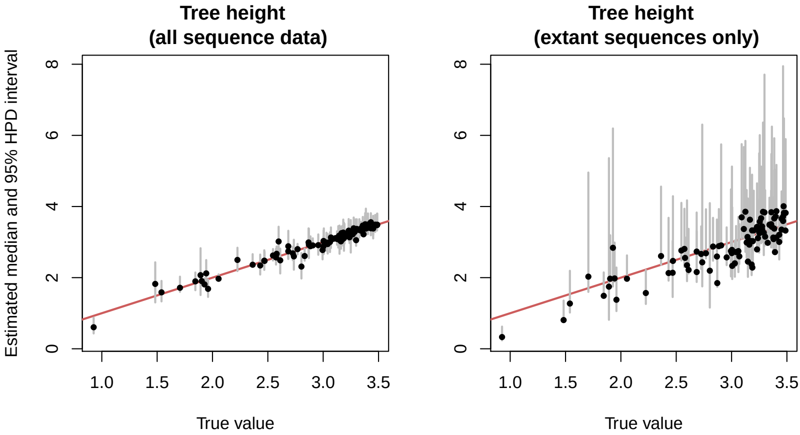
- the topology of extant species is fixed
- they only had occurrence dates of fossil samples
- they fixed ρ to the truth in the inference

We extend this model in two ways:

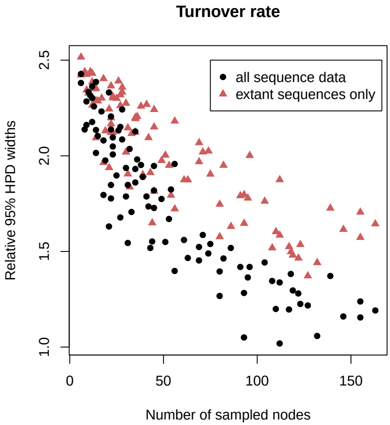
- sampling sampled ancestor trees (Gavryushkina *et al* 2014)
- incorparating morphological data

SA and Morph-models packages for BEAST2 (`www.beast2.org`)

Comparative data of fossil tips vs only occurrence dates

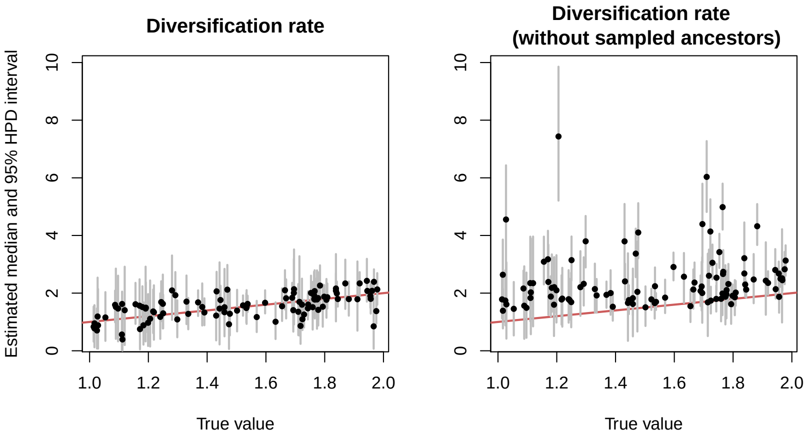


Comparative data of fossil tips vs only occurrence dates



turnover rate = $\frac{\mu}{\lambda}$

Biased estimates when not modelling sampled ancestors



diversification rate = $\lambda - \mu$

Analysis of penguin morphological data

Penguin dataset (Ksepka *et al* 2012) consisting of morphological data of:

- all extant penguins (19 species)
- 37 fossil species assigned to stratigraphic intervals

Models:

- Lewis MK vs MKv (Lewis 2001)
- Partitions vs single alignment
- Rate variation across sites and across partitions
- Tree prior: FBD and Skyline FBD
- Different tree model parameterisations

Parameterisations

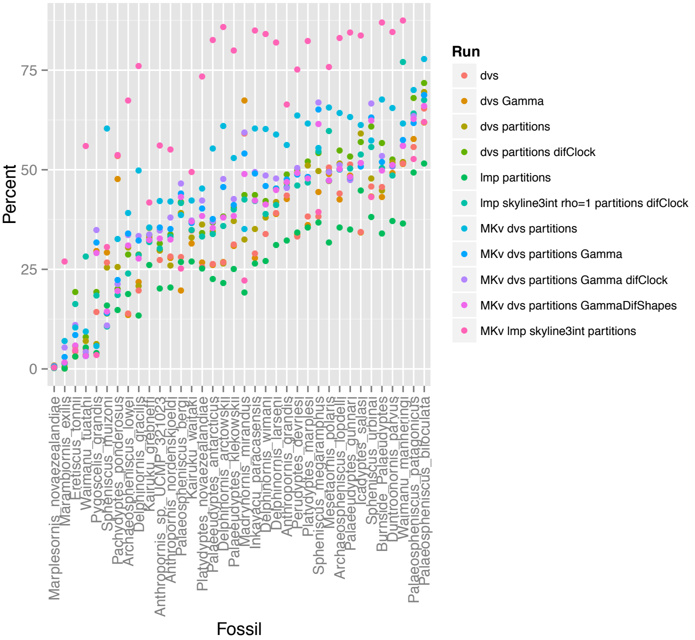
lmp:

birth rate	λ
death rate	μ
sampling rate	ψ
other parameters	t_{or}, ρ

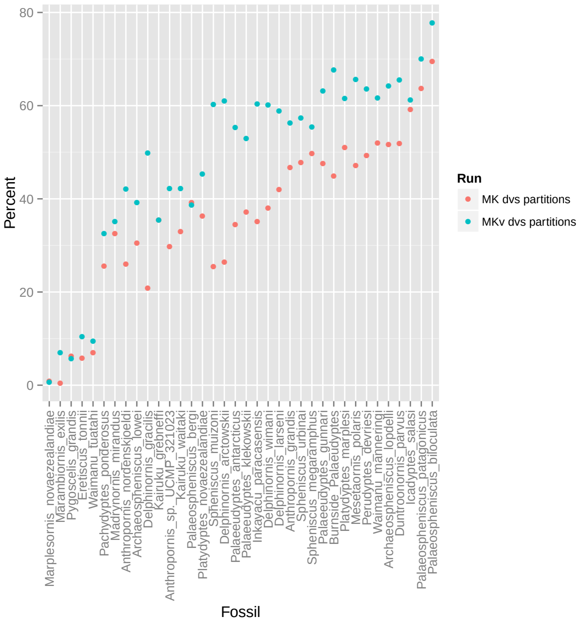
dvs:

net diversification rate	$d = \lambda - \mu$
turnover rate	$\nu = \frac{\mu}{\lambda}$
sampling proportion	$s = \frac{\psi}{\mu+\psi}$
other parameters	t_{or}, ρ

Posterior probabilities that fossils are sampled ancestors

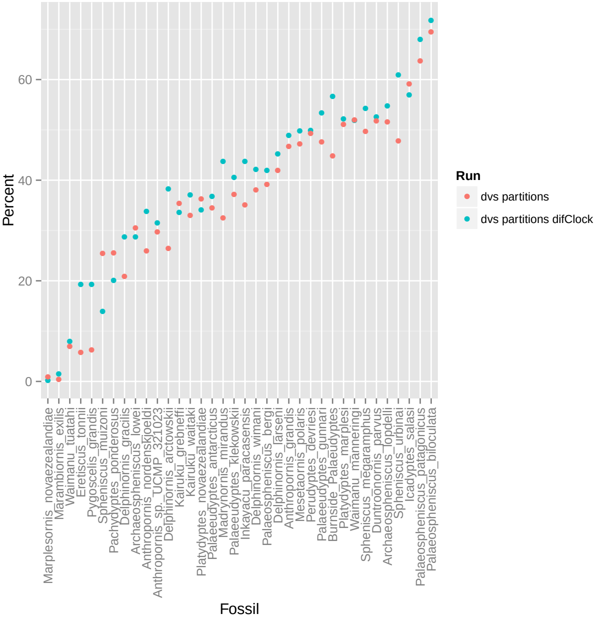


MK vs MKv

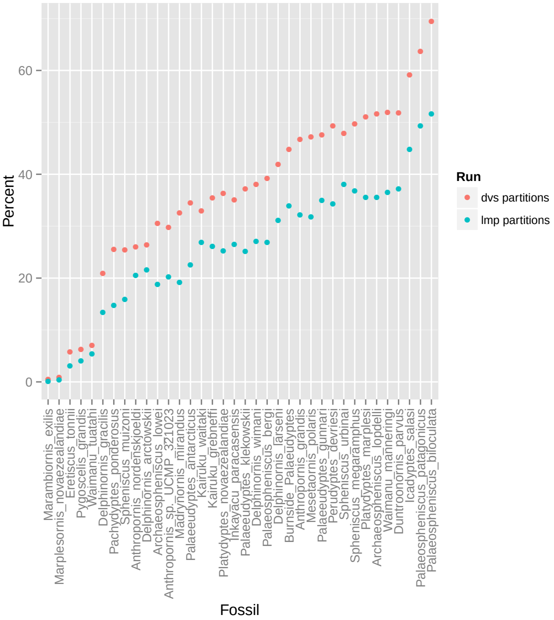


Percent
80
60
40
20
0
Marplesornis_novaezealandiae
Marambiornis_exilis
Pygoscelis_grandis
Eretiscus_tonnii
Waimanu_tuatahi
Pachydyptes_ponderosus
Madrynornis_mirandus
Anthropornis_nordenskjoeldi
Archaeospheniscus_lowei
Delphinornis_gracilis
Kairuku_grebneffi
Anthropornis_sp._UCMP_321023
Kairuku_waitaki
Palaeospheniscus_bergi
Platydyptes_novaezealandiae
Spheniscus_muizoni
Delphinornis_arctowskii
Palaeeudyptes_antarcticus
Palaeeudyptes_klekowskii
Inkayacu_paracasensis
Delphinornis_wimani
Delphinornis_larseni
Anthropornis_grandis
Spheniscus_urbinai
Spheniscus_megaramphus
Palaeeudyptes_gunnari
Burnside_Palaeudyptes
Platydyptes_marplesi
Mesetaornis_polaris
Perudyptes_devriesi
Waimanu_manneringi
Archaeospheniscus_lopdelli
Duntroonornis_parvus
Icadyptes_salasi
Palaeospheniscus_patagonicus
Palaeospheniscus_biloculata
Fossil
Run
MK dvs partitions
MKv dvs partitions

One clock rate vs different clock rates



lmp vs dvs



Percent
60
40
20
0
Fossil
Marambiornis_exilis
Marplesornis_novaezealandiae
Eretiscus_tonnii
Pygoscelis_grandis
Waimanu_tuatahi
Delphinornis_gracilis
Pachydyptes_ponderosus
Spheniscus_muizoni
Anthropornis_nordenskjoeldi
Delphinornis_arctowskii
Archaeospheniscus_lowei
Anthropornis_sp._UCMP_321023
Madrynornis_mirandus
Palaeeudyptes_antarcticus
Kairuku_waitaki
Kairuku_grebneffi
Platydyptes_novaezealandiae
Inkayacu_paracasensis
Palaeeudyptes_klekowskii
Delphinornis_wimani
Palaeospheniscus_bergi
Delphinornis_larseni
Burnside_Palaeudyptes
Anthropornis_grandis
Mesetaornis_polaris
Palaeeudyptes_gunnari
Perudyptes_devriesi
Spheniscus_urbinai
Spheniscus_megaramphus
Platydyptes_marplesi
Archaeospheniscus_lopdelli
Waimanu_manneringi
Duntroonornis_parvus
Icadyptes_salasi
Palaeospheniscus_patagonicus
Palaeospheniscus_biloculata
Run
dvs partitions
lmp partitions

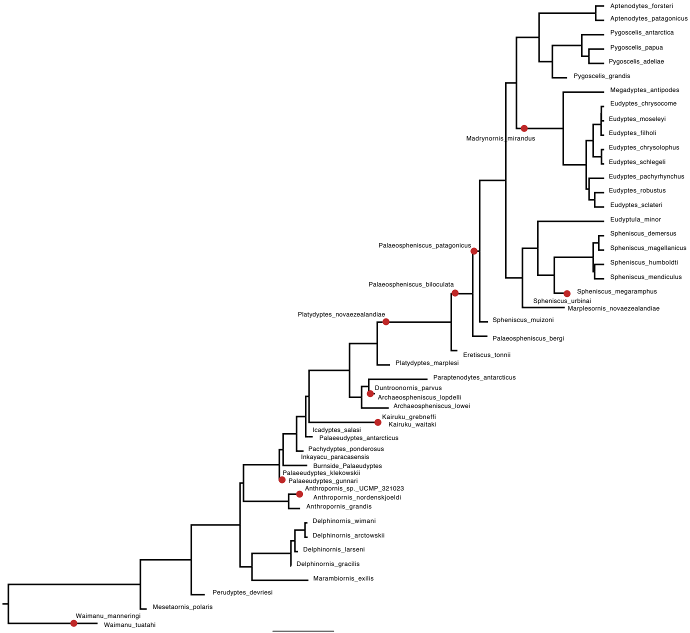
One analysis of penguin morphological data

Model settings:

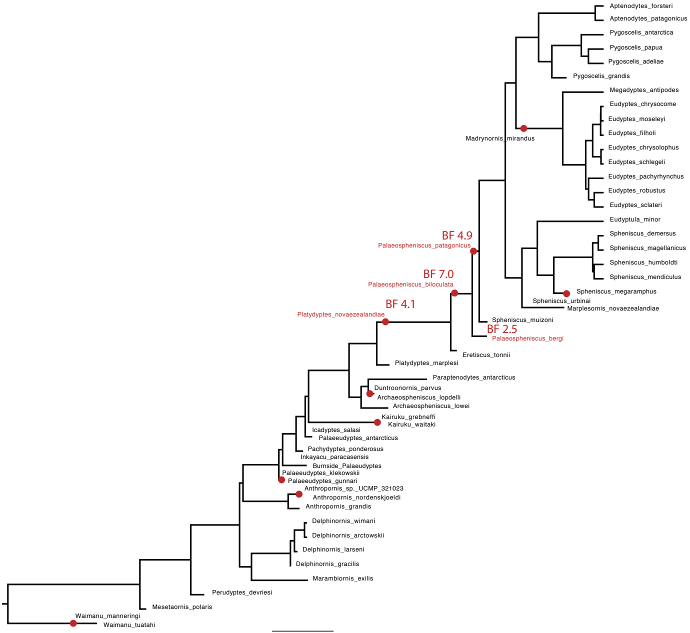
- MK model with partitions,
- Substitution rate variable across partitions,
- FBD Skyline model with 3 equidistant intervals,
- ρ fixed to one.

Summary tree: maximum sampled ancestor clade credibility tree with target heights.

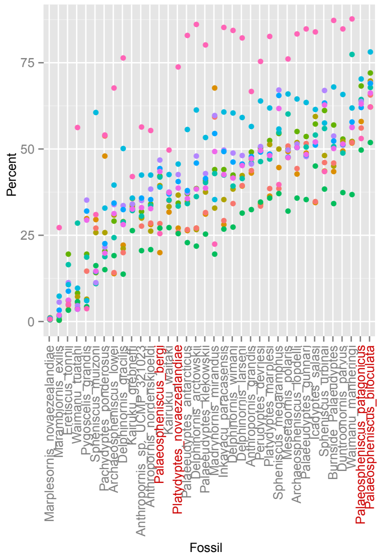
MCC tree from penguin morphological data analysis



MCC tree from penguin morphological data analysis



Posterior probabilities that fossils are sampled ancestors



Completing the total evidence analysis of penguins

Remains:

- add molecular data to the analysis,
- use a relaxed clock model to estimate divergence times.

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 position, without the need for an outgroup
  - They automatically provide estimates of relative divergence dates, or absolute divergence dates when calibration information is available
- Calibration is hard and interesting
  - Specifying natural means of calibrating phylogenies is subtle
  - Recent methods for including fossil evidence include new tree priors, and opportunities for total evidence dating.
- The geometry of (time) is understudied and its study could lead to new methods for doing phylogenetic inference and posterior post-processing and summary.