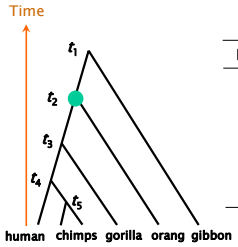
Estimation of species divergence times

Ziheng Yang
University College London

Use of molecular clock to date species divergences



Node	Distance	Time (MY)
1	0.0627	15.8±1.5
2	0.0517	13
3	0.0307	7.7 ±1.0
4	0.0201	5.0 ± 0.9
5	0.0093	2.4 ± 0.6

(Mutation rate = 0.0517/13)

Some difficulties of dating

- The molecular clock is often violated, and assumptions about rates affect time estimation.
- Fossil calibrations involve uncertainties (errors in dating a fossil and errors in assigning a fossil on the phylogeny).
- Rates and patterns of substitution are different at different loci.

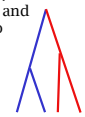
Strategies to improve molecular dating

- Use multiple genes which may be evolving in different ways.
- Use multiple fossil calibrations to constrain the rates.
- Use good estimation methods

Likelihood

Branches are grouped into rate classes, and rates and times are estimated jointly.

Kishino & Hasegawa (1990 *Methods Enzymol.* **183**, 550-570) assigned transition and transversion rates to branches on a phylogeny and estimated rates and times simultaneously. A normal approximation to the numbers of transitional and transversional differences between sequences is used to calculate the likelihood.



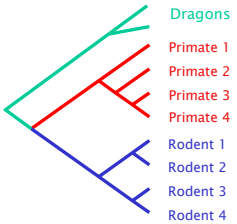
Quartet-Dating (Rambaut & Bromham 1998 *Mol. Biol. Evol.* **15,** 442-448) assigns two rates on a tree of four species. Likelihood calculated on sequence alignment.

This was extended to an arbitrary tree (Yoder & Yang 2000 *Mol. Biol. Evol.* **17,** 1081-1090) and to multiple genes and multiple calibrations (Yang & Yoder 2003 *Syst. Biol.* **52,** 705-716).

Likelihood local-clock model

assigns rate parameters for pre-specified branches

$r_0 = 1$
$r_P = 1.41421$
$r_R = 3.14159$



Likelihood method

Advantages

- Multiple gene loci can be analyzed simultaneously, with their differences accounted for.
- Multiple fossil calibrations can be used simultaneously.

Disadvantages

- Assignments of rates to branches are arbitrary.
- Calibration node ages are assumed to be constants, known without error. (The penalized-likelihood method (r8s) uses constrained minimization to incorporate fossil uncertainties, which is problematic.)

Bayesian methods

Rate drifts over time, described by a probabilistic model.

A geometric Brownian motion model is used to model rate changes by Thorne, Kishino & colleagues (Thorne, Kishino, & Painter 1998 *Mol. Biol. Evol.* **15,** 1647-1657; Kishino, Thorne & Bruno 2001 *Mol. Biol. Evol.* **18,** 352-361; Thorne & Kishino 2002 *Syst. Biol.* **51,** 689-702). Fossil calibrations are specified as constraints, that is, minimum and/or maximum ages of nodes on the tree.

Yang & Rannala (2006. *Mol. Biol. Evol.* **23**:212-226) developed "soft bounds" to accommodate fossil uncertainties. The molecular clock assumption is relaxed by Rannala & Yang (2007. Syst. Biol. 56:453-466)

Drummond et al. (2006. PLoS Biology 4:e88) developed a similar MCMC program (called BEAST) that can use statistical distributions to accommodate fossil uncertainties.

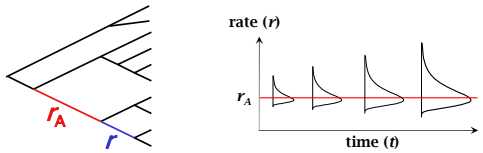
Bayesian MCMC algorithm for date estimation

$$f(\mathbf{t},\mathbf{r},\theta \mid D) = \frac{f(D \mid \mathbf{t},\mathbf{r})\, f(\mathbf{r} \mid \theta,\mathbf{t})\, f(\mathbf{t} \mid \theta)\, f(\theta)}{f(D)}$$

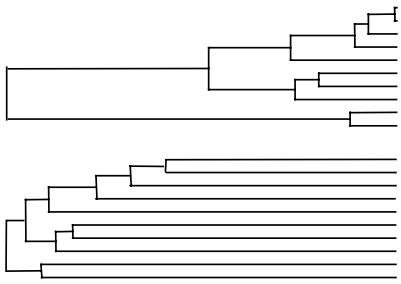
t: times
r: rates
θ: parameters
D: data

Prior model of rate drift (geometric Brownian motion)

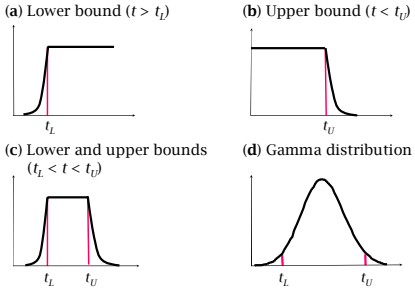
The rate r of a branch (node) is a random variable centred around the ancestral rate r_A. The variance σ^2 determines how variable the rates are on the tree.



Prior for times specified using the birth–death process with species sampling

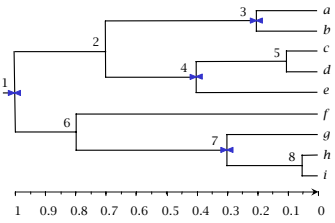


Soft bounds to account for fossil uncertainties

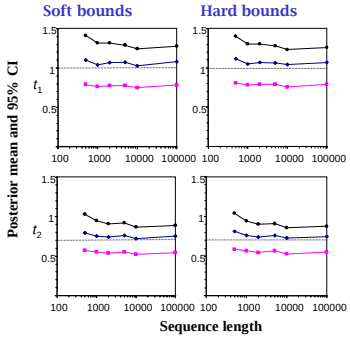


Hard vs. soft bounds: a simulation study

t_1: (0.5, 1.5) or (3.5, 4.5)
t_3: (0.1, 0.3)
t_4: (0.3, 0.5)
t_2: (0.2, 0.4)

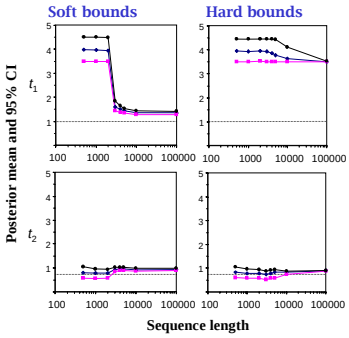


Good fossil: t_1: (0.5, 1.5)



When there is no conflict (between fossils, and between fossils and sequences), hard and soft bounds produce identical results.

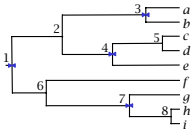
Bad fossil: t_1: (3.5, 4.5)

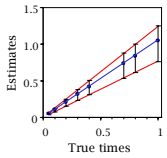


When there are conflicts (e.g., if a fossil is grossly wrong), hard and soft bounds behave differently. Hard bounds gave **incorrect time estimates with high confidence**, while soft bounds are better.

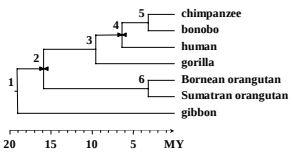
Infinite-sites theory

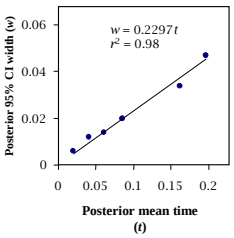
- When the amount of sequence data approaches infinity
  - the posterior CIs get narrower.
  - the posterior means, and the lower and upper CI limits fall on straight lines.
- Even with an infinite amount of sequence data, time estimates involve substantial uncertainties.





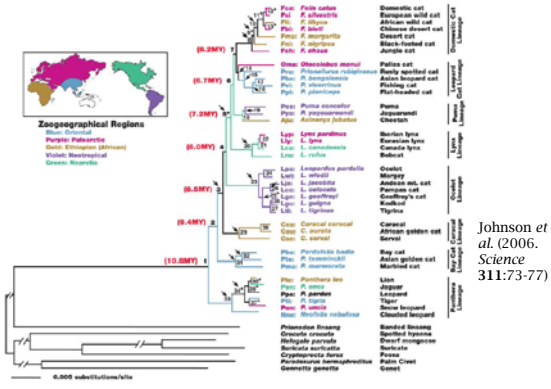
Infinite-sites theory & plot





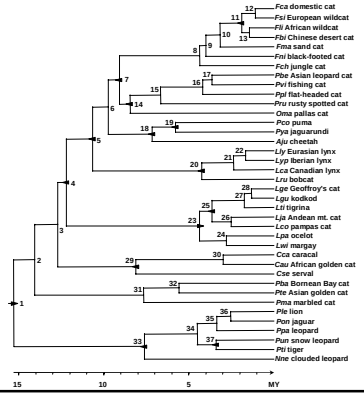
Calibrations (gamma priors)
Node 4: 6-8MY
Node 2: 12-16 MY
Data: mt cDNA, 3331bp, (Cao *et al.* 1998)
Model: HKY85+C+Γ_5

It is now more profitable to dig than to sequence!



Johnson *et al.* (2006. *Science* **311**:73-77)

Ages of cats



Data from Johnson et al. (2006. *Science* **311**:73-77)

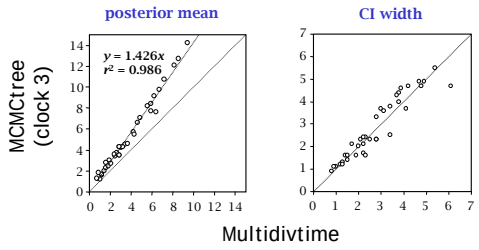
19,984bp, lots of '?'.
Model: HKY85+Γ_5

Posterior mean and 95% CIs of divergence times (MY)

Node	Fossil	Johnson et al.	39s-clock 3	38s-clock 1	38s-clock 3
1	<16	10.8 (8.4, 14.5)	14.0 (10.1, 16.7)	15.2 (12.2, 17.1)	15.3 (12.4, 17.1)
2		9.4 (7.4, 12.8)	12.2 (8.7, 15.0)	14.1 (11.2, 16.2)	14.2 (10.9, 16.4)
3		8.5 (6.7, 11.6)	11.1 (7.9, 13.6)	12.8 (10.1, 14.6)	12.7 (9.7, 14.6)
4	>5	8.1 (6.3, 11.0)	10.5 (7.5, 13.0)	12.2 (9.7, 14.1)	12.1 (9.1, 14.0)
5	>5.3	7.2 (5.6, 9.8)	9.4 (6.8, 11.8)	10.8 (8.5, 12.4)	10.7 (7.8, 12.5)
6		6.7 (5.3, 9.2)	8.7 (6.3, 10.9)	10.0 (7.9, 11.5)	9.8 (6.9, 11.5)
7	>4.2	6.2 (4.8, 8.6)	8.1 (5.8, 10.2)	9.3 (7.3, 10.8)	9.1 (6.2, 10.6)
11	>1	1.4 (0.9, 2.2)	1.7 (1.2, 2.3)	2.3 (1.7, 2.9)	2.0 (1.3, 2.6)
14	>1	5.9 (4.5, 8.2)	7.6 (5.5, 9.6)	8.7 (6.8, 10.1)	8.4 (5.7, 10.0)
18	>3.8	4.9 (3.9, 6.9)	6.2 (4.4, 8.1)	7.0 (5.5, 8.3)	7.1 (5.0, 8.7)
19	>1.8	4.2 (3.2, 6.0)	5.0 (3.4, 6.7)	5.6 (4.3, 6.8)	5.7 (3.9, 7.2)
20	>2.5	3.2 (2.5, 4.7)	3.8 (2.6, 5.3)	4.6 (3.5, 5.7)	4.3 (3.1, 5.5)
23	<5	2.9 (2.0, 4.2)	3.7 (2.7, 4.9)	4.4 (3.4, 5.0)	4.3 (3.3, 5.0)
25	>1	2.4 (1.7, 3.6)	3.1 (2.2, 4.1)	3.7 (2.8, 4.4)	3.6 (2.7, 4.3)
29	>3.8	5.6 (4.1, 7.9)	7.2 (4.9, 9.4)	8.3 (6.4, 9.8)	8.2 (6.1, 10.1)
33	>3.8	6.4 (4.5, 9.3)	7.2 (4.8, 10.3)	7.7 (6.1, 9.0)	7.6 (5.8, 10.5)
37	>1	2.9 (1.8, 4.6)	3.2 (2.0, 5.0)	3.5 (2.7, 4.4)	3.5 (2.5, 4.8)

Model: HKY85+Γ_5. clock1: global clock; clock3: correlated rates

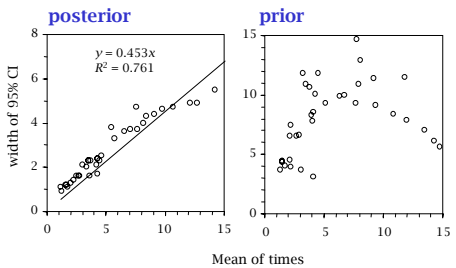
Estimates from the two programs have systematic differences



It is not very clear which differences are important.

	Multidivtime	MCMCtree	
Calibrations	Hard bounds	Soft bounds	➨
Prior on rates & times	Geometric Brownian motion for rates, recursive algorithm for times	Same, plus independent rates; birth-death on times	➜
Likelihood	Approximate	Exact & approximate	➜
Data	Outgroup	No outgroup	➨

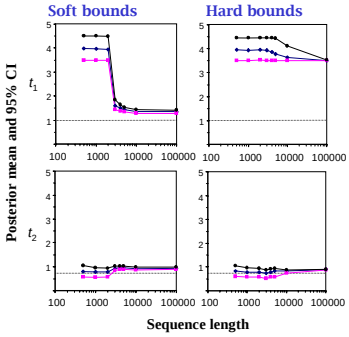
Infinite-sites plot, cat data



Summary

- Bayesian MCMC methods provide a natural framework for assembling different sources of information (uncertainties), from fossils, sequences, etc.
- Many factors may affect divergence time estimation, and their relative importance is not well-understood:
  - violation of the clock
  - prior on rates and times
  - substitution model
- More work is needed to model fossil uncertainties reliably.

Bad fossil: t_1: (3.5, 4.5)



When there are conflicts (e.g., if a fossil is grossly wrong), hard and soft bounds behave differently. Hard bounds gave **incorrect time estimates with high confidence**, while soft bounds are better.

A model of errors in fossils

Every fossil has a probability p_E of being erroneous (and excluded from the analysis), with the prior $p_E \sim$ beta(p, q). The MCMC algorithm then generates a posterior distribution of p_E and a posterior probability that each fossil is in error.

A simple way of implementing the model is to have an indicator variable for each fossil, and update the indicators in the MCMC.

Suppose there are $c = 3$ fossils.

- **U** = $U_1U_2U_3$ = 111: all three are used. Pr = $(1 - p_E)^3$
- **U** = 011: first not in use, the other two in use. Pr = $p_E(1-p_E)^2$
- **U** = 000 is not allowed.

U can take $2^c - 1$ values for c fossils.

Constraint on the root age

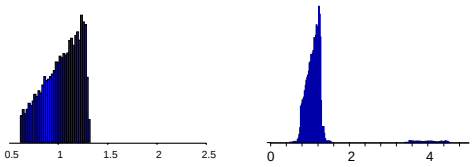
We require that the root age be constrained, whether or not a fossil is present at the root.

- Simple upper bound: $t_1 < t_{max}$
- Lower & upper bounds: $t_{min} < t_1 < t_{max}$

The constraints are implemented as soft bounds, and are used if there is no fossil at the root, or if there is a fossil at the root but it is considered erroneous (as specified by **U**).

Simulation under the clock: 9s, 200K bp

Posterior distribution of root age t_1



2 good fossils.

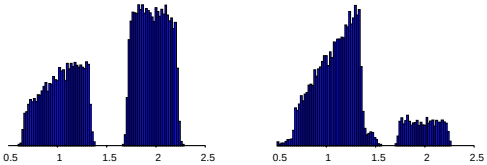
t_2: 0.2-0.4
t_3: 0.1-0.3
((a, b) '>0.1<0.3', ((c, d), e)), (f, (g, (h, i)) '>0.2<0.4'));

3 good fossils
1 bad fossil

As in Rannala & Yang (2006)

Simulation under the clock: 9s, 200K bp

Posterior distribution of root age t_1



One good fossil t_2: 0.2-0.4
One bad fossil t_4: 0.7-0.9 ($t_1 \sim 2.0$).
(((a, b), ((c, d), e) '>0.7<0.9'), (f, (g, (h, i)) '>0.2<0.4'));

2 good fossils t_2: 0.2-0.4, t_3: 0.1-0.3
1 bad fossil t_4: 0.7-0.9 ($t_1 \sim 2.0$).
(((a, b) '>0.1<0.3', ((c, d), e) '>0.7<0.9'), (f, (g, (h, i)) '>0.2<0.4'));

Real dataset

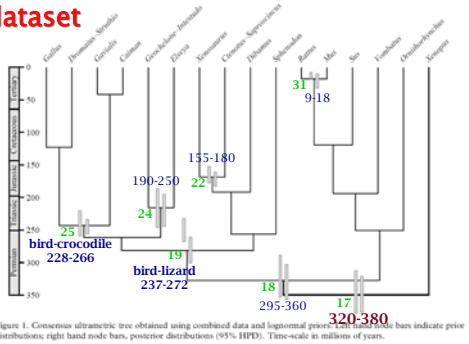


Figure 1. Consensus ultrametric tree obtained using combined data and lognormal priors. Error bars on node bars indicate prior distributions; right hand node bars, posterior distributions (95% HPD). Time-scale in millions of years.

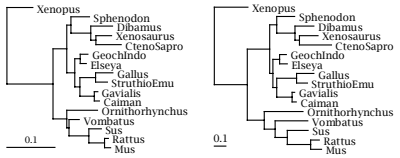
Sanders & Lee (2007 *Biology Letters* **3**: 275-279) used molecules to evaluate the reliability of fossils.

Sequence data statistics (MLEs under HKY+Γ_5)

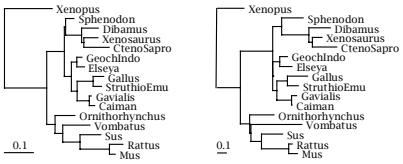
Partition	bp	tree length	$\hat{\kappa}$	$\hat{\alpha}$
Rag 1&2	2106	1.1	2.9	0.26
Rag 3	1053	5.2	6.1	3.82
Cmos 1&2	738	1.9	2.3	0.65
Cmos 3	369	59.9	5.9	11.06

ML branch lengths without the clock (HKY+Γ_5)

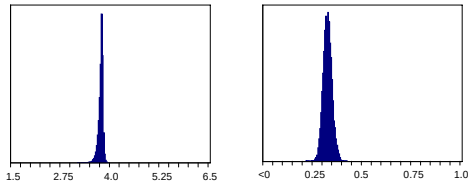
Rag



cMos



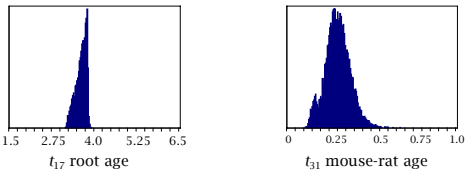
Posterior estimates of times under the clock



RootAge >3.2<3.8, p_E~beta(1,10).

Fossil	Error
Node 18:	0.001
Node 19:	1.000
Node 22:	1.000
Node 24:	1.000
Node 25:	1.000
Node 31:	0.999

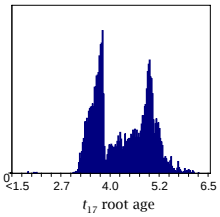
Posterior estimates of times under relaxed clock model (clock3: auto-correlated)



RootAge >3.2<3.8, p_E~beta(1,10).

Fossil	Error
Node 18:	0.001
Node 19:	1.000
Node 22:	1.000
Node 24:	1.000
Node 25:	1.000
Node 31:	0.999

Posterior of times under clock3



RootAge<5
C17 (3.2, 3.8) treated as fossil

Fossil	error
C17 (3.2, 3.8)	0.658
C18 (2.95, 3.60)	0.598
C19 (2.37, 2.72)	0.539
C22 (1.55, 1.80)	0.88
C24 (1.90, 2.50)	0.778
C25 (2.28, 2.66)	0.86
C31 (0.09, 0.18)	0.953

Constraint on root age is important.

Posterior estimates of times

Node	Calibration	Clock	Independent rates	Autocorrelated rates
17 root	(320, 380)	376 (361, 384)	362 (326, 381)	360 (324, 381)
18	(295, 360)	285 (269, 301)	305 (243, 332)	332 (296, 374)
19 bird-lizard	(237, 272)	162 (150, 175)	169 (130, 204)	230 (185, 263)
22	(155, 180)	86 (77, 94)	76 (55, 99)	93 (68, 118)
24	(190, 250)	39 (33, 46)	48 (29, 74)	130 (46, 190)
25 bird-crocodile	(228, 266)	103 (94, 113)	107 (79, 136)	156 (123, 194)
31 mouse-rat	(9, 18)	32 (28, 36)	32 (17, 45)	30 (15, 45)

p_E ~ beta(0.1, 10), with mean 0.01.

Summary

- The model performs well in simulated data.
- In the real data set, there seems to be much conflict, and the Bayesian method appears place high confidence on whatever results it ends up with. More tests are needed.
- Bayesian statistics provides a natural framework for incorporating different sources of information (uncertainties), from fossils, sequences, etc.
- However, when the models are complex, and it is not always easy to understand the effects of the different components. The prior becomes more important with the increase of model complexity.

References

Drummond et al. 2006. *PLoS Biology* 4:e88.
Kishino et al. 2001. *Mol. Biol. Evol.* 18:352-361.
Thorne et al. 1998. *Mol. Biol. Evol.* 15:1647-1657.
Yang & Rannala 2006. *Mol. Biol. Evol.* 23:212-226.
Rannala & Yang 2007. *Syst. Biol.* 56:453-466.
Yang, Z., 2006. *Computational Molecular Evolution.* OUP, Chapter 7.