An ABC example: estimating the divergence time of primates

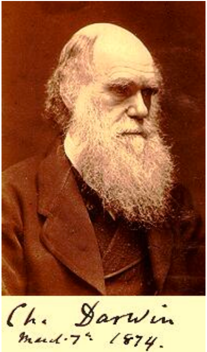
A short course on ABC lecture

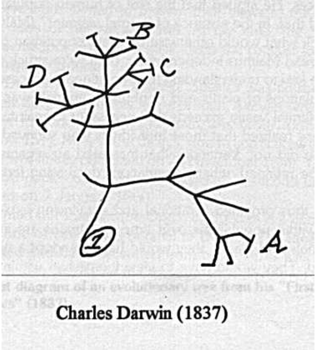
DAMTP February 20 2017

Statistical inference on trees: timescales

- Introduction
- Primate fossil record
- Dating splits by ABC
- Today's posterior is tomorrow's prior: molecular data
- Conclusions

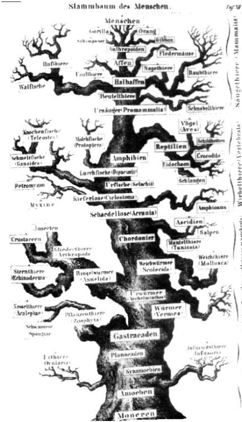
Charles Darwin (1809 - 1882)



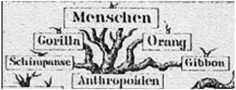


Charles Darwin (1837)

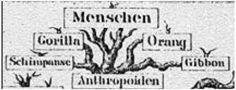
Ernst Haeckel (1834-1919)

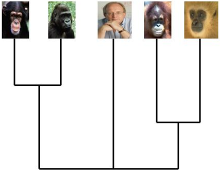


Haeckel tree of life (detail)

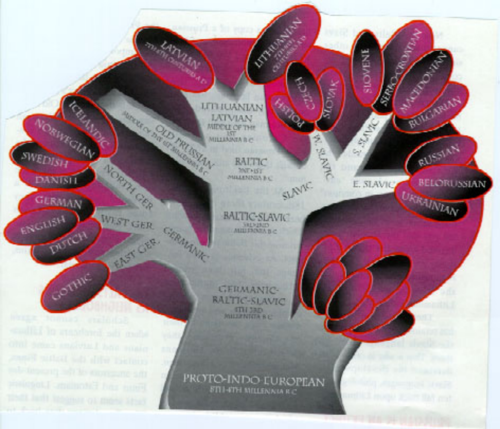


Haeckel tree of life (detail)



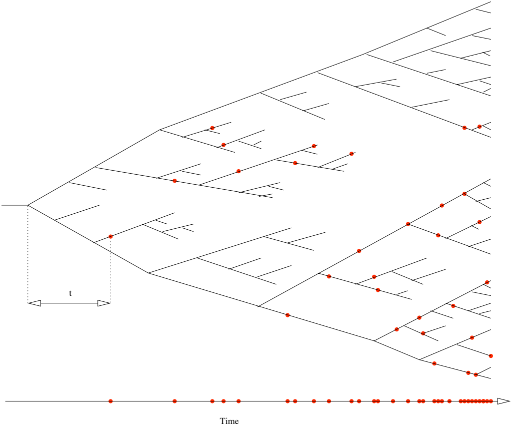


August Schleicher (1821-1868)



The Primates

Primate Evolution



Reconciling molecular and fossil records?

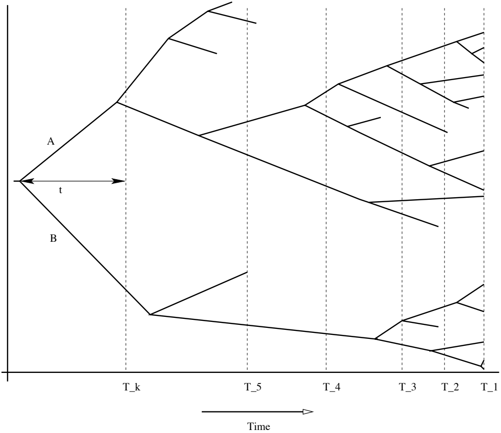
- Extant primates are strepsirrhines (lemurs and lorises) and haplorhines (tarsiers and anthropoids)
- Molecular estimate of time of divergence is approximately 90 mya
- Fossil record suggests 60-65 mya
- Fossil record is patchy

Problem: Use the fossil record to estimate the age of the last common ancestor of extant primates

Primate Data

Epoch	k	T_k	Observed number of species (D_k)
Late Pleistocene	1	0.15	19
Middle Pleistocene	2	0.9	28
Early Pleistocene	3	1.8	22
Late Pliocene	4	3.6	47
Early Pliocene	5	5.3	11
Late Miocene	6	11.2	38
Middle Miocene	7	16.4	46
Early Miocene	8	23.8	36
Late Oligocene	9	28.5	4
Early Oligocene	10	33.7	20
Late Eocene	11	37.0	32
Middle Eocene	12	49.0	103
Early Eocene	13	54.8	68
Pre-Eocene	14		0

The evolutionary process



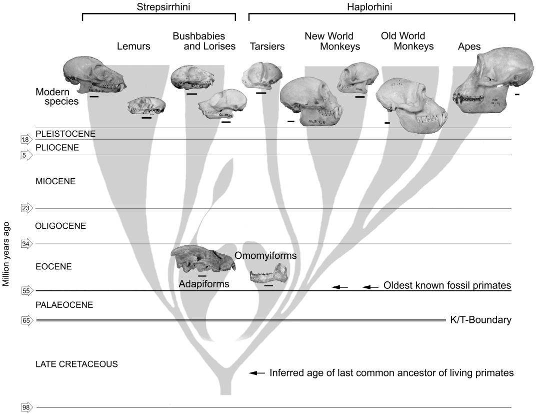
What happened?

- Average sampling fraction of 5.7%
  - upper 95% limit 7.4%
- Estimated divergence time 81.5 mya
  - 95% CI (72.0, 89.6) mya

Tavaré, Marshall, Will, Soligo & Martin *Nature*, 2002

- Pravda, Times, BBC, . . . , assorted religious fanatics, . . .

Primate Evolution



Why more?

- Bayesian approach more natural
- Allows us to incorporate prior information
- Sampling fractions
  - probability of finding a fossil in bin i is α_i
  - $\boldsymbol{\alpha} = \alpha\boldsymbol{p}$, $\boldsymbol{p}$ known
  - reasonable?
- Other models for finds?
- Allowing for dinosaur extinction at K/T boundary?

Fossil record: ABC approach

Data can be thought of in two parts:

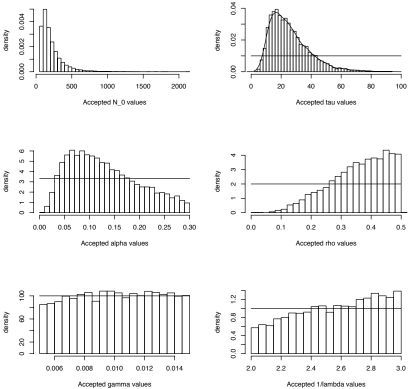
(a) the observed number of fossils $F_{\rm obs}$ found

(b) the proportions $p_{j,{\rm obs}}$ found in jth bin

A suitable metric might be

$$\left|\frac{F}{F_{\rm obs}} - 1\right| + \sum_{j=1}^{k+1} |p_j - p_{j,{\rm obs}}|$$

Results $\epsilon = 0.1$



Some ABC technicalities

Hybrid ABC schemes

Sensitivity: Exploring Other Models

One advantage of ABC – it is easy to change the input . . .

- Choice of ρ
- Demography
- Sampling fractions
- K/T crash 65 mya
    - the time of origin of primates is even further back in the Cretaceous
- Poisson sampling scheme: length in bin matters
- Dating other split points

Hybrid ABC schemes: ABC-Gibbs

J1 If currently at $\boldsymbol{\theta} = (\theta_1, \theta_2)$, draw θ_1' from $\pi(\theta_1 | \mathcal{D}, \theta_2)$ and set $\boldsymbol{\theta} = (\theta_1', \theta_2)$.

J2 Draw θ_2' from $\pi(\theta_2)$ and simulate data $\mathcal{D}'$ using parameter $\boldsymbol{\theta} = (\theta_1', \theta_2')$.

J3 If $\mathcal{D} = \mathcal{D}'$, set $\boldsymbol{\theta} = (\theta_1', \theta_2')$ and return to step J1. Otherwise stay at $\boldsymbol{\theta} = (\theta_1', \theta_2)$ and return to step J2.

Steps J2 and J3 above are the mechanical version of the rejection algorithm which gives samples from $\pi(\theta_2 | \mathcal{D}, \theta_1)$.

By replacing step J3 with

J3$'$ If $\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon$, set $\boldsymbol{\theta} = (\theta_1', \theta_2')$ and return to step J1. Otherwise stay at $\boldsymbol{\theta} = (\theta_1', \theta_2)$ and return to step J2.

we can generate approximate draws from $\pi(\theta_2 | \mathcal{D}, \theta_1)$.

- Could also use Approximate Metropolis-within-Gibbs and other variants

Dealing with Sampling Fractions

$$f(\boldsymbol{\lambda}, \tau, \mathcal{N}, \boldsymbol{\alpha}|\mathcal{D}) \propto \mathbb{P}(\mathcal{D}|\boldsymbol{\alpha}, \boldsymbol{\lambda}, \tau, \mathcal{N})\mathbb{P}(\mathcal{N}|\tau, \boldsymbol{\lambda})f(\tau)f(\boldsymbol{\lambda})f(\boldsymbol{\alpha})$$

where

- $\boldsymbol{\lambda} = (\lambda, \gamma, \rho)$ growth parameters,
- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{14})$ sampling fractions
- $\mathcal{N}$ is the underlying tree structure

Give sampling fractions independent Beta(a, b) priors

Gibbs-ABC Example

Split the random variable into two parts:
$\boldsymbol{\alpha}$ and $(\boldsymbol{\lambda}, \tau, \mathcal{N})$

Sample from the two conditional distributions

- $f(\boldsymbol{\alpha} \mid \mathcal{D}, \boldsymbol{\lambda}, \tau, \mathcal{N})$
- $f(\tau, \boldsymbol{\lambda}, \mathcal{N} \mid \mathcal{D}, \boldsymbol{\alpha})$

Conditional distribution of α

$$\begin{aligned}
&f(\boldsymbol{\alpha} \mid \mathcal{D}, \boldsymbol{\lambda}, \tau, \mathcal{N}) \\
&\quad \propto \quad f(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \tau, \mathcal{N} \mid \mathcal{D}) \\
&\quad \propto \quad \mathbb{P}(\mathcal{N} \mid \tau, \lambda) f(\tau) f(\lambda) f(\boldsymbol{\alpha}) \mathbb{P}(\mathcal{D} \mid \tau, \boldsymbol{\lambda}, \mathcal{N}, \boldsymbol{\alpha}) \\
&\quad \propto \quad f(\boldsymbol{\alpha}) \mathbb{P}(\mathcal{D} \mid \mathcal{N}, \boldsymbol{\alpha}) \\
&\quad \propto \quad \Pi_{i=1}^{14} \alpha_i^{d_i} (1-\alpha_i)^{N_i - d_i} \alpha_i^{a-1} (1-\alpha_i)^{b-1} \\
&\quad \propto \quad \Pi f_{\beta}(\alpha_i \,; d_i + a, N_i - d_i + b)
\end{aligned}$$

Posterior mean of $\alpha_i = \frac{a+d_i}{N_i+a+b} \approx \frac{d_i}{N_i}$

Conditional distribution of $(\tau, \boldsymbol{\lambda}, \mathcal{N})$

$$\begin{aligned} f(\tau, \boldsymbol{\lambda}, \mathcal{N} | \mathcal{D}, \boldsymbol{\alpha}) &\propto f(\boldsymbol{\lambda}, \tau, \mathcal{N}, \boldsymbol{\alpha} | \mathcal{D}) \\ &\propto \mathbb{P}(\mathcal{D} | \boldsymbol{\lambda}, \boldsymbol{\alpha}, \mathcal{N}, \alpha) \mathbb{P}(\mathcal{N} | \tau, \lambda) f(\tau) f(\boldsymbol{\lambda}) \end{aligned}$$

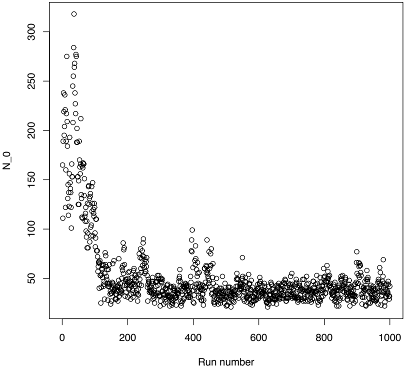
Simulate from this using ABC: accept $(\boldsymbol{\lambda}, \tau, \mathcal{N})$ if $\rho(\mathcal{D}, \mathcal{D}') < \epsilon$, where $\mathcal{D}'$ represents the simulated data

Metric and Priors

$$\begin{aligned}
\tau &\sim U[0,100] \\
\alpha &\sim U[0,0.6] \\
\rho &\sim U[0,0.8] \\
\gamma &\sim U[0.005,0.015] \\
1/\lambda &\sim U[2,3] \\
a &= 0.1 \\
b &= 1 \\
\epsilon &= 0.2
\end{aligned}$$

Same metric as before

No free lunches



Tweak metric

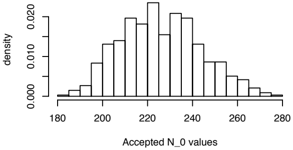
- The observed N_0 values are too small
  - require $N_0 > 235$
  - change the metric

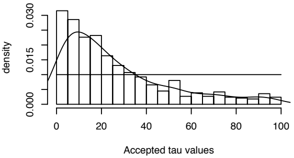
$$\rho(\mathcal{D}, \mathcal{D}') = \sum_{i=1}^{k} \left| \frac{D_i}{D_+} - \frac{D'_i}{D'_+} \right| + \left| \frac{D'_+}{D_+} - 1 \right| + \left| \frac{N'_0}{N_0} - 1 \right|$$

- Penalises trees with N_0 values far from 235

Results: $\epsilon = 0.3$

	min	LQ	Median	mean	UQ	Max
N_0	184	212	224	226	238	279
τ	0.0	8.0	18.6	26.3	36.8	99.5





Sensitivity: Exploring Other Models

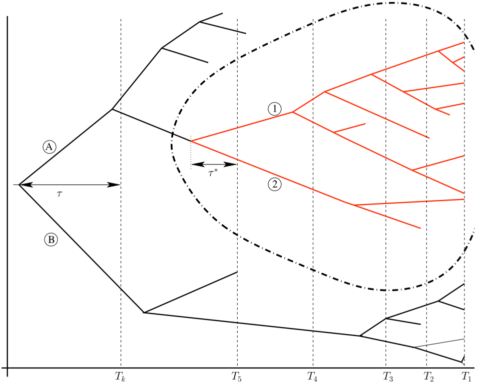
One advantage of ABC – it is easy to change the input . . .

- Choice of ρ
- Demography
- Sampling fractions
- K/T crash 65 mya
  - the time of origin of primates is even further back in the Cretaceous
- Poisson sampling scheme: length in bin matters
- Dating other split points

Old World/New World Split

Epoch	k	T_k	Hap/Strep number of species (D_k)	Plat/Cat number of species (D_k^*)
Late Pleistocene	1	0.15	19	19
Middle Pleistocene	2	0.9	28	28
Early Pleistocene	3	1.8	22	22
Late Pliocene	4	3.6	47	44
Early Pliocene	5	5.3	11	10
Late Miocene	6	11.2	38	33
Middle Miocene	7	16.4	46	43
Early Miocene	8	23.8	36	30
Late Oligocene	9	28.5	4	3
Early Oligocene	10	33.7	20	6
Late Eocene	11	37.0	32	2
Middle Eocene	12	49.0	103	0
Early Eocene	13	54.8	68	
Pre-Eocene	14		0	

Dating Two Splits



Details

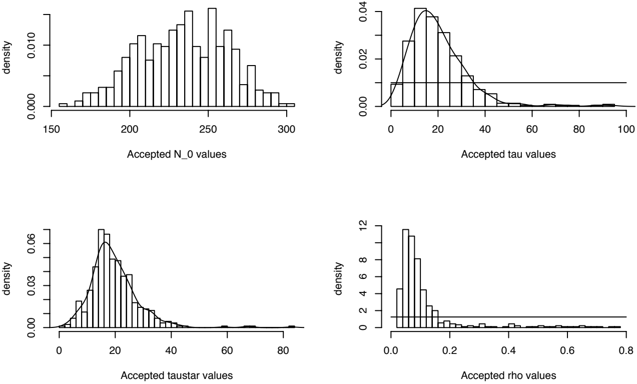
- $N_0 = 235$ species for the Strep/Hap,
- $\epsilon = 0.4$ for both metrics

	min	LQ	Median	mean	UQ	Max
N_0	159	212	234	233	254	303
τ	0.9	12.1	17.6	20.1	25.3	94.5
τ^*	1.6	14.5	18.2	19.6	23.5	82.9

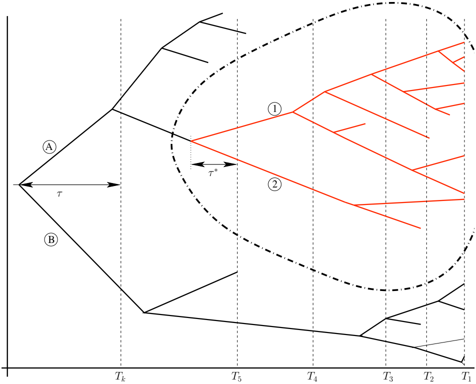
The median posterior sampling fractions ($\times 100$)

α_1	α_2	α_4	α_5	α_6	α_8	α_9	α_{10}	α_{11}	α_{12}	α_{13}	α_{14}
8	10	12	3	6	7	1	8	22	41	80	1
8	8	8	4	4	4	1	4	8	8	8	1

Posteriors


density
0.000
0.010
150
200
250
300
Accepted N_0 values
density
0.00
0.02
0.04
0
20
40
60
80
100
Accepted tau values
density
0.00
0.03
0.06
0
20
40
60
80
Accepted taustar values
density
0
2
4
6
8
12
0.0
0.2
0.4
0.6
0.8
Accepted rho values

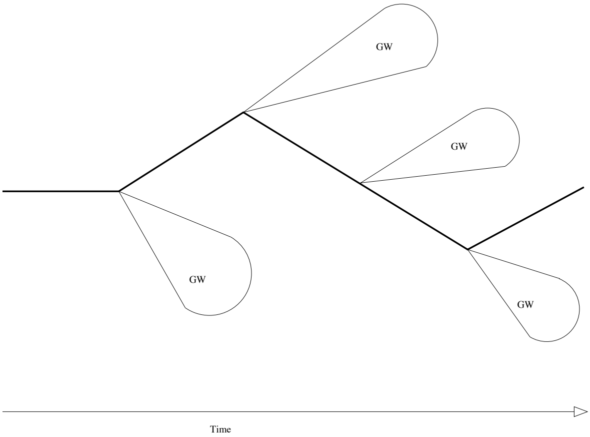
Dating Two Splits, revisited



The structure of branching processes

- Our approach to inferring multiple split points is heuristic
- What other approaches might work?
- Consider conditioning the process on a split at a fixed time
  - leads to a size-biassed GW process
  - For ABC, need to be able to simulate the process
  - Can use rejection . . .

A(nother) fishbone process



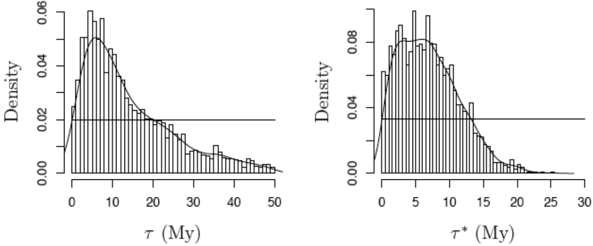
Which metric?

$$\rho(\mathcal{D}, X) = \sum_{i=1}^{14} \left| \frac{D_i}{D_+} - \frac{X_i}{X_+} \right| + \left| \frac{X_+}{D_+} - 1 \right| + \left| \frac{X_0}{N_0} - 1 \right|.$$

Match up:

- Proportions of fossils observed in each bin
- Total number of fossils observed
- Number of extant species

What happened?

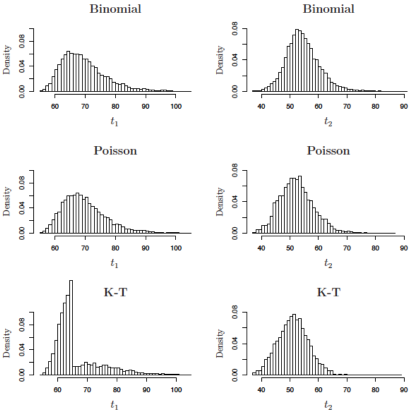


Combining fossil record with molecular data

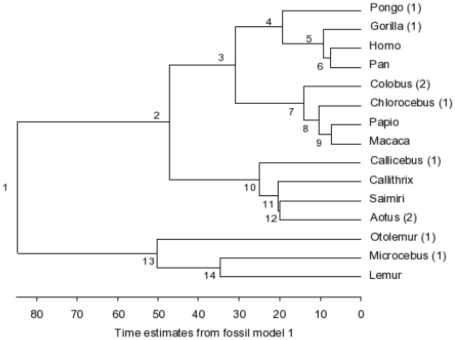
Yesterday's posterior is tomorrow's prior . . .

- Estimate posterior for two primate divergence times
- Use as prior for dating nodes from molecular data ($mcmctree$)
- Data are updated from earlier analysis

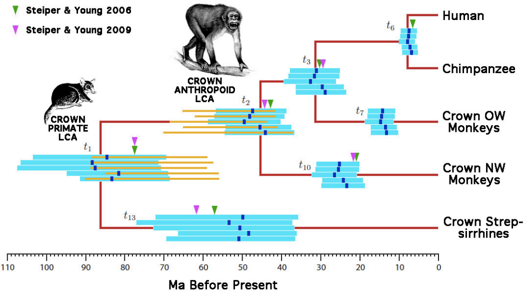
The posteriors



The phylogeny of the species (Poisson model)



The molecular data



References

- ST, Marshall C, Will O, Soligo C & Martin R (2002) Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature*, **416**, 726–729

- Wilkinson R & ST (2009) Estimating primate divergence times by using conditioned birth-and-death processes. *Theoretical Population Biology*, **75**, 278–295

- Wilkinson R, Steiper M, Soligo C, Martin R, Yang Z & ST (2011) Dating primate divergences through an integrated analysis of palaeontological and molecular data. *Systematic Biology*, **60**, 16–31.