Estimation of species divergence times incorporating fossil and molecular information

Ziheng Yang
Department of Genetics, Evolution & Environment
University College London
Outline

- Molecular clock dating of species divergences
- Bayesian estimation of divergence times
  - Prior on rates and rate-drift models
  - Prior on times and fossil calibrations
  - User-specified and effective time prior
- Infinite-sites theory & plot
- Limits of divergence time estimation
Knowledge of absolute (geological) ages of species divergences is useful.

- It satisfies our curiosity;
- It may allow us to place the speciation events into the correct geological or palaeo-environmental contexts, and to understand speciation & dispersal mechanisms;
- It allows us to estimate evolutionary rate;
- Dating divergences of viral sequences may allow us to understand viral transmission mechanisms;
- etc.

**Fig. 1.** Molecular, cladistic, and biogeographic evidence indicates that the Cambrian evolutionary 'explosion' may have been underpinned by a long period of Precambrian diversification. The undoubted diversification at the base of the Cambrian accompanied the acquisition of hard parts – and increase in size and fossilization potential of 'Phanerozoic style' metazoans – but the ground plans of these clades probably had a much longer history than has been claimed. Thick solid lines: known range. Shaded lines: ghost range implied cladistically. Thin solid lines: implied pattern of Precambrian cladogenesis. Animals minute, asterisks denote divergences dated by molecular data. Illustration by M.A. Wilks.
Origin of mammals at the K-T boundary

Fig. 2. Phylogenies of modern mammals and birds plotted against time, based on morphological\textsuperscript{7} and molecular data\textsuperscript{8,15}. The mammalian phylogeny is supported by recent research using long molecular sequences\textsuperscript{8,14}. Cladistic and biogeographical data indicate the Early Tertiary 'explosion' of mammalian and avian diversity was preceded by a period of unobserved diversification, which molecular data suggests was surprisingly extensive. Thick solid lines: known range. Shaded lines: ghost range implied cladistically. Thin solid lines: ranges calculated from molecular data. Ovals: divergences dated by molecular data. Dotted lines: not accurately dated.
Molecular clock means that the DNA or protein sequences accumulate changes (stochastically) at a constant rate, so that genetic sequences keep a record of time in their changes like a clock does.

We have sequences from the modern species and can calculate distances between them (either pairwise comparisons or likelihood). Then if the age of one node on the tree is known, all other ages can be estimated. We assume the phylogeny is known.
Complications

- The clock may be violated, and assumptions about rates affect time estimation.
- *Fossil calibrations* involve uncertainties (errors in dating a fossil and errors in assigning a fossil onto a branch of the phylogeny, etc.).
- Rates and patterns of evolution may differ at different gene loci.
Bayesian statistics is currently the only framework that can integrate different sources of information/uncertainties in a clock dating analysis.


The Bayesian framework for estimating species divergence times

Posterior $\propto$ prior $\times$ likelihood

$$f(t, r, \theta | D) \propto f(\theta)f(t | \theta)f(r | \theta, t)f(D | \theta, t, r)$$

- $t$: times;
- $r$: rates;
- $\theta$: substitution parameters;
- $D$: sequence data

Parameters in the model: $\Theta = \{t, r, \theta\}$
Rate prior: A model of rate drift to deal with violation of the clock (geometric Brownian motion): $f(r|\theta,t)$

Rate $r$ for a branch (node) has a log-normal distribution around the ancestral rate $r_A$. The variance parameter $\nu$ determines how variable the rates are on the tree.

$$f(r | r_A) = \frac{1}{r\sqrt{2\pi\nu t}} \exp\left\{ -\frac{1}{2\nu} (\log r - \log r_A + \frac{1}{2} t\nu)^2 \right\}$$

Or one can assume independent rates for branches

The rate $r$ of a branch is a random variable from a common distribution such as the log-normal or gamma.

There are conflicting reports concerning which of those prior models fit the data better. A more important issue is the robustness of the posterior time estimates.

Time prior: the birth–death–sampling process is used to specify the time prior \( f(t|\theta) \).

Model parameters:

\( \lambda \), per-lineage birth (speciation) rate

\( \mu \), per-lineage death (extinction) rate

\( \rho \), sampling fraction of modern species

Time prior also incorporates fossil calibration information.

By using different \( \lambda, \mu, \rho \), one can examine the impact of the time prior. There is great interest in estimating \( \lambda \) and \( \mu \), but it is unclear whether there is so much information in the data.
Soft bounds & arbitrary distributions to account for fossil uncertainties

(a) Lower (min) bound \((t > t_L)\)

(b) Upper (max) bound \((t < t_U)\)

(c) Lower and upper bounds \((t_L < t < t_U)\)

(d) Gamma distribution

Probabilistic modeling and statistical analysis of fossil data (occurrences and morphological measurements) can generate statistical distributions of node ages on the tree, which can then be used in molecular clock dating.

Figure 1: An illustration of the stochastic model of fossil finds. Bases of five stratigraphic intervals at $T_1, \ldots, T_5$ Myr ago are shown along the x-axis. The temporal gap between the base of the final interval and the point at which the two founding species originate is denoted by $\tau$. Thick lines indicate species day.

A Bayesian analysis of the fossil occurrence data generated the posterior distributions for two node ages, which are then used as priors in a molecular clock dating analysis.

Divergence times of mammals

Red: intraordinal
Green: interordinal

(Springer, et al. 2003 PNAS 100:1056-1061)
User-specified prior and effective prior

- Specified prior is the calibration density specified by the user. This may not satisfy the requirement that ancestors are older than descendants.
- Effective prior is the prior used by the program after the truncation. This is generated by running the program without data.
User-specified prior (calibrations) and effective prior

Calibrations: red, green, blue:  
Prior: purple  
Posterior: black:

Limits of divergence time estimation

distance \( (d) = \text{time} \ (t) \times \text{rate} \ (r) \)

**Data:** An animal ran from A to B, and the distance is \( d = 100 \) meters.

**Question:** How long did the animal run \((t)\)?

This is a hard question!

**Useful information?**
- The distance is measured extremely precisely: \( d = 100.000 \).
- Animal is a tortoise not a hare, so one can assign a prior on the rate/speed \( r \).

**Insight:** Even if \( d \) is measured with no errors, our estimate of \( t \) (and \( r \)) will be imprecise and the prior on \( t \) and \( r \) will matter.
Time and rate are confounded as sequences provide information about distances only.

Likelihood depends on distance $d = 2t \times r$. Time unit = 100 Myrs.
$t \sim G(2, 2)$ with mean 1
$r \sim G(2, 4)$ (with mean 0.5, $0.5 \times 10^{-8}$ changes/site/year).
Infinite-sites theory & plot

The *infinite-sites* theory gives the posterior distribution of times and rates etc. when the amount of sequence data approaches infinity (and when uncertain fossil calibrations are fixed).

Under the clock, the posterior converges to one dimension, with the posterior means, and the lower and upper CI limits becoming proportional, when the number of sites approaches infinity.

When the amount of sequence data approaches infinity, $R \to 1$. Here it is still useful to generate more sequence data.

The slope reflects the precision of fossil calibrations. Here every 1 MY of divergence adds 0.45MY to the CI.
Infinite-sites plot

8,708,584 sites of codon positions 1&2
The confounding effect of rates and times means that the prior is always important.

Use of multiple loci helps to reduce the uncertainty, and is becoming more and more and common.

The i.i.d. prior on rates for loci used in current dating programs makes a strong statement about the average rate over all loci, with two consequences.

(a) False precision: Posterior time estimates are too certain.

(b) Undue influence: Posterior time estimates are sensitive to the rate prior and may be wrong if the prior is unreasonable.

If \( \text{var}(\mu_i) = \nu \) for rate at locus \( i \), \( \bar{\mu} = \frac{1}{L} \sum_{i=1}^{L} \mu_i \) has \( \text{var}(\bar{\mu}) = \frac{\nu}{L} \to 0 \), as \( L \to \infty \).

Impact of the prior on rates for gene loci

human-chimpanzee divergence time under i.i.d. prior

(1) Fast rate $G(2,2)$, $10^{-8}$/site/year
(2) Medium rate $G(2,20)$, $10^{-9}$/site/year
(3) Slow rate $G(2,200)$, $10^{-10}$/site/year
Summary

• Bayesian MCMC methods provide a natural framework for integrating different sources of information, from fossils, sequences, etc.

• The confounding effects of rates and times cause many problems. Even with whole genomes, the posterior estimates of times and rates will have uncertainties and will be sensitive to the prior.

• Different programs (BEAST, MrBayes, MCMCtree, etc.) can generate very different priors from the same calibrations and thus posteriors on times. The user should inspect the prior used by the program (run the program without data).

• More effort is needed to compile and analyse fossil data to generate calibration densities for clock dating.
Acknowledgments

Bruce Rannala
Mario dos Reis
Phil Donoghue

BBSRC

http://abacus.gene.ucl.ac.uk/software/