Applications of molecular phylogenies

Part 1: Introduction and Dating


Outline

- Dating
  - Molecular clocks
  - Calibration of nodes
  - Examples and problems
  - Newer methods
    - Fossilized birth-death model
    - Tip dating

- Ancestral state reconstruction:
  - Correlated trait evolution
  - Biogeography
  - Diversification analyses

- Cophylogenetic analyses
Dating phylogenies

- Zuckerkandl & Linus Pauling in 1962 first found evidence of a correlation between fossil age and number of differences in the amino acid sequence of hemoglobin.

- By assigning absolute dates to phylogenies can relate to relationships to other clades, biogeographic processes and sometimes even historical evolution.
Strict molecular clock

- Simplest model is a strict clock assumes one conversion rate from mutation rate to evolutionary time
- Trees should be **ultrametric**
But there is no universal clock

- Mutation rates differ for a huge variety of reasons:
  - Changes are stochastic (random according to a Poisson process)
  - Genes and parts of genes evolve at different rates due to different functional constraints or selective pressure
  - Species evolve at different rates:
    - Generation time
    - Population size
    - Metabolic rate
    - DNA repair efficiency
    - Selective pressure
Improvements to clock models

• Unlink clocks across gene models (Thorne & Kishino 2002).

• Local clocks
  • User-defined clades can vary in rate

• Uncorrelated relaxed clocks
  • Each branch can have its own rate
  • Uses stochastic model of evolutionary rate change and samples possible distributions of rates using Markov Chain Monte Carlo simulation (MCMC)

Extremely brief summary of Bayesian methods for phylogeny estimation

- Evaluated based on maximum likelihood calculations and parameter estimates
- Incorporates prior information
- Random walk [https://phylogeny.uconn.edu/mcmc-robot/](https://phylogeny.uconn.edu/mcmc-robot/)
- Final results integrate over parameter uncertainty

- If interested:
  - Bayes for the Uninitiated Brown 2003
  - [https://tinylink.net/ou6WV](https://tinylink.net/ou6WV)

---

Figure 4.2: The Metropolis-Hastings algorithm, which determines the acceptance probabilities of new (proposed) states. Proposed “uphill” steps are always accepted (as we are interested in regions of high probability), but “downhill” steps are accepted with a probability inversely related to the extent of the “drop” from the current state to the proposed state. Using this algorithm, drastic drops in elevation are unlikely (but not impossible). Allowing suboptimal state changes allows the chain to traverse valleys in parameter space, and therefore permits more thorough exploration.
Calibrations

• Types of calibrations
  • Fossils
  • Cophylogenetic
  • Biogeographic
  • Secondary calibration

• Ways to calibrate
  • Hard and soft bounds
  • Uniform, normal or lognormal distribution

• Which node to calibrate and how to specify prior distribution?
Fossil calibrations

- Uncertainty in node to calibrate based on morphology, **crown** vs **stem** taxa
Crown and stem taxa

- Gymnosperms
- Angiosperms
- Crown group
- Stem group
- Common ancestor of angiosperm crown group
- Common ancestor of seed plants
Fossil calibrations

• Uncertainty in node to calibrate based on morphology, crown vs stem taxa

• Should carefully justify node calibration with explicit discussion of morphology (Parham et al. 2011)

• Uncertainty in actual age of fossil

• Use of only oldest fossil per node

• Can only provide minimum bound in age, often in lognormal distribution (hard bound on minimum age and soft bound on maximum age)

• Actual distribution of prior is arbitrary

Cophylogenetic calibrations

- Many examples of closely associated taxa: specialist herbivores, parasites, predators, symbionts
- Can assume crown group tied to same host but not stem
- Provides a maximum estimate based on presence of partner
- Distribution of prior depends on how age of one partner is known

**Figs and fig wasps**

Biogeographic calibrations

- Uncertainty in age of biogeographic event
- Distribution depends on specific kind of biogeographic calibration
- Depends on accuracy of assumed calibration which is not tested, i.e., circular
  - For example, could calibrate range extension into South American based on date of isthmus of Panama formation but ignores possibility of dispersal

Dated molecular phylogenies have overturned many long held biogeographic hypotheses.
Secondary calibrations

- Uses ages estimated from other dated molecular phylogenies
- Normal distribution; depends on accuracy of first dated phylogeny
- Used in cases where there are no good fossil representatives on the clade of interest
Programs to estimated dated phylogenies

- **R8s** (Sanderson 2003)
  - Maximum likelihood
  - Used hard bound age estimates placed on specific nodes

- **BEAST** (Drummond and Rambaut 2007)
  - Most commonly used
  - Bayesian
  - Allows for incorporating uncertainty in phylogenetic reconstruction, rate and age estimates

Discrepancies between node-calibrated dating and fossils

- Frequently older ages estimated by fossil-calibrated molecular phylogenies
  - 18-101 mya gap for animals
  - ~70 mya gap for placental mammals
  - ~90 mya gap for angiosperms

- Incomplete fossil records suggests that true ages of crowns should predate fossils but....

- Some conspicuous groups with really large gaps

Does this represent a flaw in dating analyses?
Some known flaws

- Clade-specific rate heterogeneity
  - If some individual clades evolve faster rates, simulations show that they can push back estimated age of crown despite use of uncorrelated clocks (Beaulieu et al. 2015).
Some known flaws

- Even with constant diversification rate, internal branch length overestimated
- Exacerbated by not sampling all lineages
Behavior of prior

- Age immediately departs from oldest fossil age (140 mya) when run with no data at all (driven by prior).
- When run with data, marginal prior and marginal posterior have different distributions (posterior shifted younger relative to prior?)
- Marginal prior is emergent combination of all priors.
• Shifts from prior to posterior mostly affect uncalibrated nodes
  • Data overfitting?
  • Insufficient information to overrule the pseudodata in calibrations?
When run with uniform priors

- Improve separation of prior and posterior but...
  - Not really realistic
  - Hard bound on maximum age is arbitrary
  - Many distributions cluster near minimum and maximum bounds suggesting ill-fit
Take home points from Brown and Smith 2018

• Standard fossil calibration of nodes can lead to incorrect estimates of ages due to pseudodata in calibration prior information.

• Always check behavior of marginal prior when running Bayesian analyses.
Newer methods for dating phylogenies

• Two methods both allow for multiple fossils within a lineage (not just oldest one) and incorporate fossil temporal information directly

• Fossilized birth-death model (Heath et al. 2014)
  • Incorporates fossil sampling rate, $\Psi$, 
Fossilized birth death process

- Integrates over uncertainty in fossil position
Newer methods for dating phylogenies

- Two methods both allow for multiple fossils within a lineage (not just oldest one) and incorporate fossil temporal information directly

- Fossilized birth-death model (Heath et al. 2014)
  - Incorporates fossil sampling rate, $\Psi$,
  - Gavryushkina et al. 2017 finds more realistic younger age of ~12.7 mya than node-calibrated date for crown penguins.
Newer methods for dating phylogenies

- Two methods both allow for multiple fossils within a lineage (not just oldest one) and incorporate fossil temporal information directly

- Fossilized birth-death model (Heath et al. 2014)
  - Incorporates fossil sampling rate, $\Psi$,
  - Gavryushkina et al. 2017 finds more realistic younger age of $\sim$12.7 mya than node-calibrated date for crown penguins.

- Tip-dating (Ronquist et al. 2012)
  - Places fossils directly in phylogeny based on morphological data
  - Only possible when fossils have lots of morphology for reliable coding and placement on phylogeny
Summary

- Assigning absolute ages to divergence dates on a phylogeny is a difficult problem worth solving.

- Bayesian methods are most often used allowing integration over uncertainty in various parameters like mutation rates assigned to individual branches.

- Information for node calibrations can take various forms based on various sources of evidence but often suffers from arbitrary designations of prior density which can influence age estimates more than any influence of the data itself.

- Newer methods of calibrating phylogenies incorporate multiple fossils per node instead of just the oldest and seem to provide more accurate age estimates.