

Phylogenetic trees – can we bootstrap?

Ralph Møller Trane

10/21/2018

Phylo-what trees...?

Given a number of aligned DNA sequences, determine how they are related.

Our toy example: cats and dogs!

Taxa	Aligned DNA sequence
Felis_catus____domestic_cat	ATGTTTCATAAACCGG...
Acinonyx_jubatus____cheetah	ATGTTTCATAATCCGC...
Neofelis_nebulosa____clouded_leopard	ATGTTTCATAAACCGC...
Uncia_uncia____snow_leopard	ATGTTTCATAAACCGC...
Panthera_pardus____leopard	ATGTTTCATAAACCGC...
Panthera_tigris____tiger	ATGTTTCATAAACCGC...
Canis_lupus_familiaris____domestic_dog	ATGTTTCATTAACCGA...
Canis_lupus____gray_wolf	ATGTTTCATTAACCGA...
Canis_latrans____coyote	ATGTTTCATTAACCGA...
Cuon_alpinus____dhole	ATGTTTCATTAACCGA...
Vulpes_vulpes____red_fox	ATGTTTCATTAATCGA...
Nyctereutes_procyonoides____raccoon_dog	ATGTTTCATTAACCGA...

But why?

- To conserve phylogenetic diversity of plant communities¹, which in turn could lead to feature diversity
- A certain natural product might have desired qualities, but come with unwanted side effects.
- Phylogenetic trees can be related to transmission trees of infectious disease outbreaks²
- Looking for new species
- Forensics; for example finding evidence that victims were infected by same strain of HIV³

¹<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3440956/>

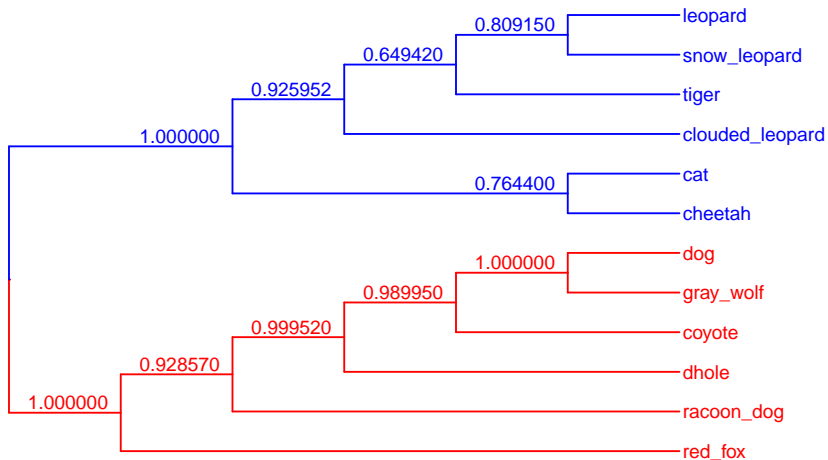
²<https://www.ncbi.nlm.nih.gov/pubmed/24037268>

³<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4072429/>

But how?

- 1) Based on the observed data, fit a model.
- 2) Sample from the model.
- 3) Results often represented using a single tree annotated with posterior probabilities of clades.

Example - cats and dogs

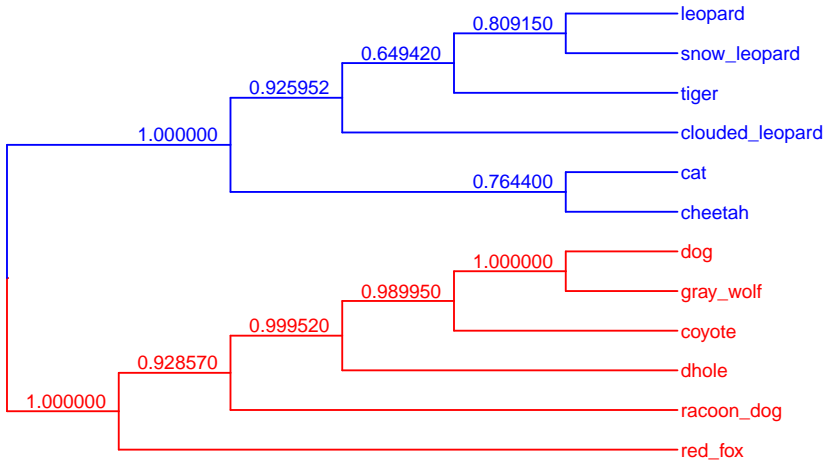


So... what's the problem?

- The sample space is HUGE! 654,729,075 possible trees with only 12 taxa
- When doing inference, we consider the probability of a tree/subclades.
 - Estimated using simple relative frequencies (SRF)
 - Unsampled tree \implies estimated probability is 0. Reasonable?
- Using Conditional Clade Probabilities (CCD) we can estimate the probability of a tree based on subclades.

Conditional Clade Probabilities?

- Given a split, further evolution in one subclade independent on the other
- Estimate probability of a tree using bayes and conditional independence



Conditional Clade Probabilities?

- Given a split, further evolution in one subclade independent on the other
- Estimate probability of a tree using bayes and conditional independence
- SRF based vs CCD based probabilities
 - similar when we want them to be
 - CCDs non-zero for unsampled trees

Where are we going with this?

- Don't need to directly estimate tree probabilities, can “simply” estimate CCD
- Maybe we can do this using bootstrap samples?
 - pro: much faster!
 - con: not as accurate
- Question: if it is less accurate, why bother?
 - Answer: actually works...
 - ... sometimes

Goal:

- Investigate when it works/doesn't work

Done so far:

- Research.
- Learned about phylogenetic trees
- Tools: MrBayes, Bisto, ccdprobs
- implemented method to read in a tree in Julia

Plan of attack:

- simulate trees in Julia
- use MrBayes/bistro and ccdprobs to get CCDs for bootstrap and MCMC samples
- compare
- ~~hopefully~~ find a pattern