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	ATGTTCATAAACCGG...
Acinonyx_jubatus____cheetah	ATGTTCATAATCCGC...
Neofelis_nebulosa____clouded_leopard	ATGTTCATAAACCGC...
Uncia_uncia____snow_leopard	ATGTTCATAAACCGC...
Panthera_pardus____leopard	ATGTTCATAAACCGC...
Panthera_tigris____tiger	ATGTTCATAAACCGC...
Canis_lupus_familiaris____domestic_dog	ATGTTCATTAACCGA...
Canis_lupus____gray_wolf	ATGTTCATTAACCGA...
Canis_latrans____coyote	ATGTTCATTAACCGA...
Cuon_alpinus____dhole	ATGTTCATTAACCGA...
Vulpes_vulpes____red_fox	ATGTTCATTAATCGA...
Nyctereutes_procyonoides____raccoon_dog	ATGTTCATTAACCGA...

But why?

- To conserve phylogenetic diversity of plant communities[1], which in turn could lead to feature diversity

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3440956/

But why?

- To conserve phylogenetic diversity of plant communities[1], which in turn could lead to feature diversity
- A certain natural product might have desired qualities, but come with unwanted side effects.

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3440956/

But why?

- To conserve phylogenetic diversity of plant communities[1], 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 outbreaks[2]

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3440956/
[2] https://www.ncbi.nlm.nih.gov/pubmed/24037268

But why?

- To conserve phylogenetic diversity of plant communities[1], 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 outbreaks[2]
- Looking for new species

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3440956/
[2] https://www.ncbi.nlm.nih.gov/pubmed/24037268

But why?

- To conserve phylogenetic diversity of plant communities[1], 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 outbreaks[2]
- Looking for new species
- Forensics; for example finding evidence that victims were infected by same strain of HIV[3]

[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.

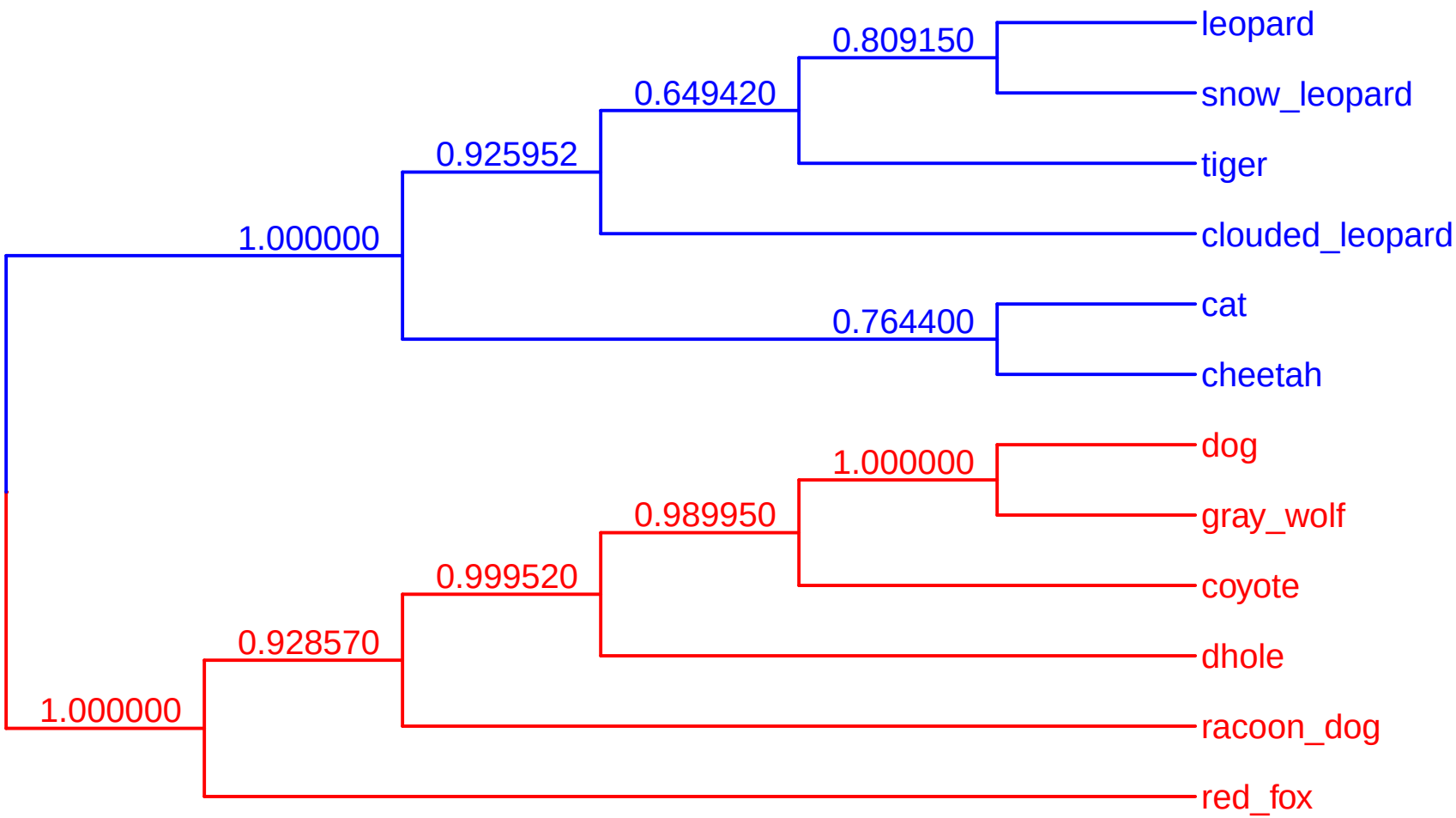
But how?

1) Based on the observed data, fit a model.
2) Sample from the model.

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

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.

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)

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 $\Longrightarrow$ estimated probability is 0. Reasonable?

So... what's the problem?

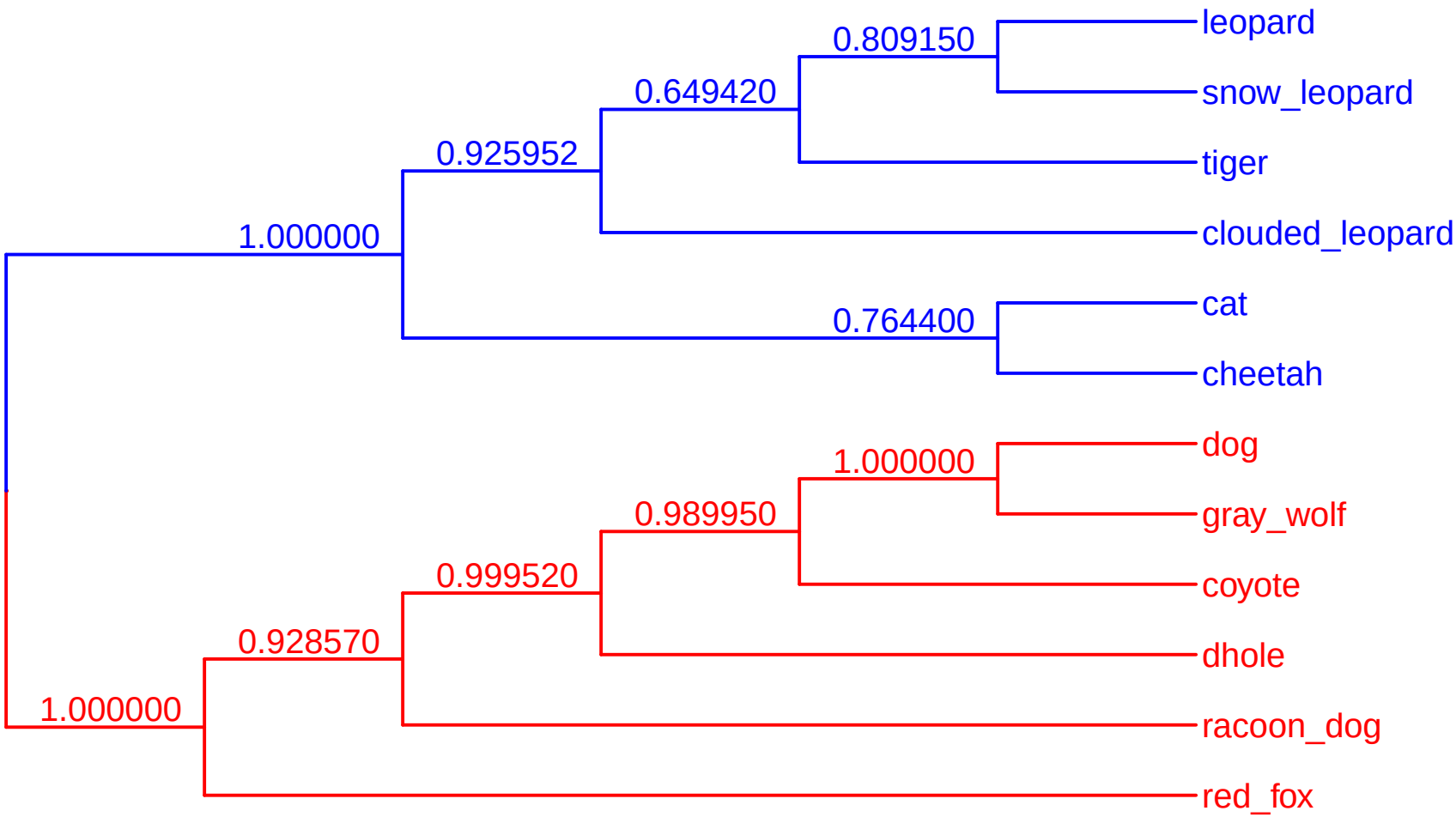
- The sample space is HUGE! 654,729,075 possible trees with only 12 taxa
- When doing inference, we consider the probability of a tree/subclades.
  - Estimated using simple relative frequencies (SRF)
  - Unsampled tree $\Longrightarrow$ estimated probability is 0. Reasonable?
- Using Conditional Clade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

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


leopard
snow_leopard
tiger
clouded_leopard
cat
cheetah
dog
gray_wolf
coyote
dhole
racoon_dog
red_fox
0.809150
0.649420
0.925952
1.000000
0.764400
1.000000
0.989950
0.999520
0.928570
1.000000

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

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

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

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?

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!

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

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?

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...

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.

Done so far:

- Research.
- Learned about phylogenetic trees

Done so far:

- Research.
- Learned about phylogenetic trees
- Tools: MrBayes, Bistro, ccdprobs

Done so far:

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

Plan of attack:

- simulate trees in Julia

Plan of attack:

- simulate trees in Julia
- use MrBayes/bistro and ccdprobs to get CCDs for bootstrap and MCMC samples

Plan of attack:

- simulate trees in Julia
- use MrBayes/bistro and ccdprobs to get CCDs for bootstrap and MCMC samples
- compare

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