

Easy Mathematics of Phylogenetic Trees

Olivier Gascuel

Directeur de Recherche au CNRS

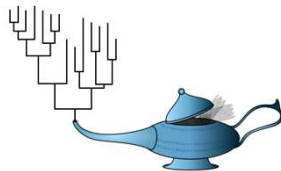
**Institut de Systématique, Evolution, Biodiversité (ISYEB)
Muséum National d'Histoire Naturelle**

Académie des Sciences

olivier.gascuel@mnhn.fr

<https://isyeb.mnhn.fr/fr/annuaire/olivier-gascuel-7496>

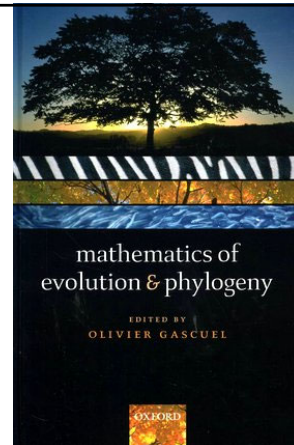
1



PhyML

Phylogenetics

- Tree building algorithms (BioNJ, PhyML, NGPhylogeny.fr...)
- Tree combinatorics (minimum evolution, supertree ...)
- Statistical modeling (AA substitution, LG)
- Branch testing, model selection (aLRT, bootstrap, SMS)
- Ancestral reconstructions (math, dating, phylogeography...)



2

-
- The diagram illustrates the HIV-1 genome and its four reading frames. The top part shows the genome with LTRs, gag, pol, and env genes, and accessory genes. The bottom part shows the four reading frames: rev (frame +2), tat (frame +1), pol (frame 0), and env (frame -2).

- 4

Numbers

Number of edges in a binary tree with n taxa:

2 tax: 1, 3 tax: 3, 4 tax: 5, 5 tax : 7 ...

n tax: $e(n) = e(n-1) + 2 = 2n - 3$



5

Numbers

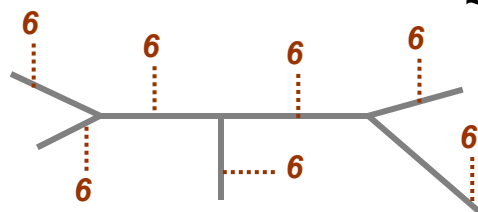
Number of unrooted binary trees with n taxa: :

2 tax: 1, 3 tax: 1, 4 tax: 3, 5 tax : 15, 5 tax : 105 ...

53 tax: $\approx 10^{80} \approx$ atoms in the universe

n tax: $t(n) = t(n-1) \times e(n-1) = (2n-5)(2n-7) \dots$

$\approx n^n$



→ hard optimization problems!

6

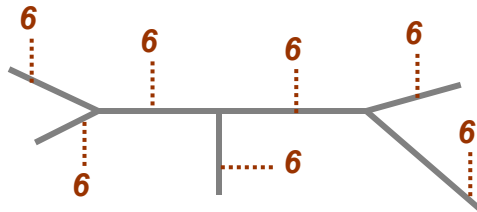
Numbers

Number of unrooted binary trees

2 tax: 1, 3 tax: 1, 4 tax: 3, 5 tax: 1

53 tax: $\approx 10^{80} \approx$ atoms in the universe

n tax: $t(n) = t(n-1) \times e(n-1) = (2n-5)(2n-7) \dots$
 $\approx n^n$



7

How many edges in a rooted tree ?

How many rooted trees ?

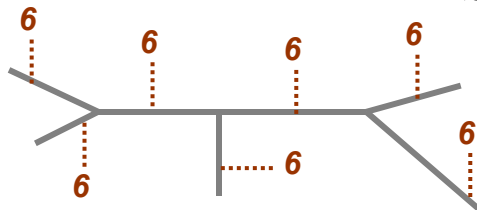
Numbers

Number of unrooted binary trees

2 tax: 1, 3 tax: 1, 4 tax: 3, 5 tax: 1

53 tax: $\approx 10^{80} \approx$ atoms in the universe

n tax: $t(n) = t(n-1) \times e(n-1) = (2n-5)(2n-7) \dots$
 $\approx n^n$



8

How many edges in a rooted tree ?

$2n-2$

How many rooted trees ?

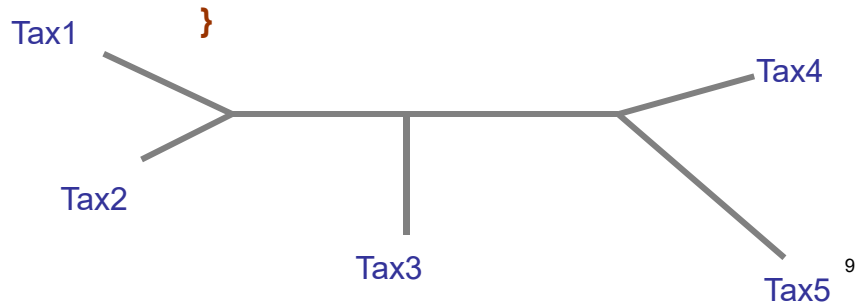
$t(n+1)$

Bipartitions (binary characters, splits)

Topology →

{ {Tax1, Tax2} | {Tax3, Tax4, Tax5}
 {Tax1, Tax2, Tax3} | {Tax4, Tax5}
 {Tax1} | {Tax2, Tax3, Tax4, Tax5}
 {Tax2} | L - {Tax2}, {Tax3} | L - {Tax3}
 {Tax4} | L - {Tax4}, {Tax5} | L - {Tax5}

Tree and network
 building
 Topology comparison

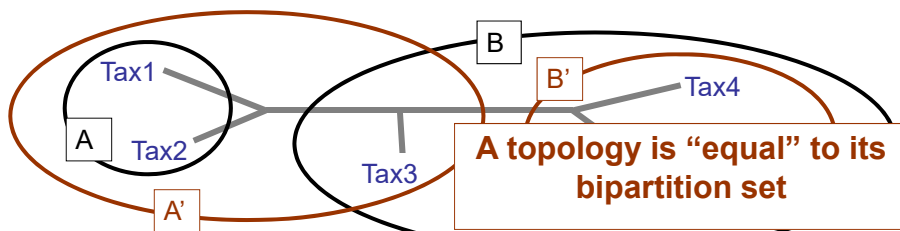


9

Bipartitions (binary characters, splits)

- A topology defines a bipartition set
- Given a bipartition set, it's easy to check that it defines a unique topology, using a local condition:

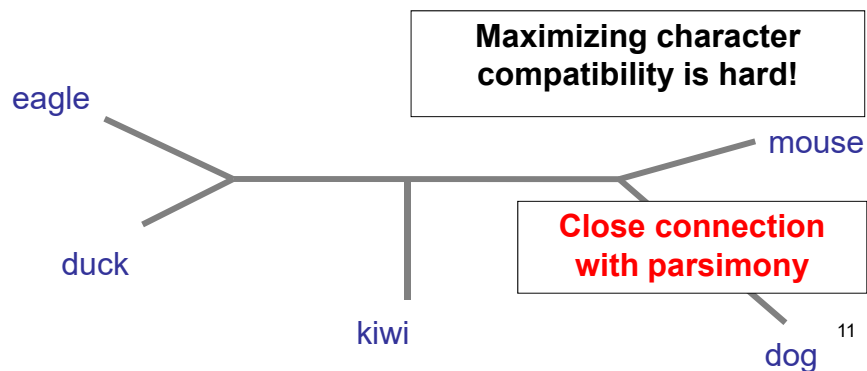
$A | B$ and $A' | B'$ are tree compatible iff
 one of $A \cap A'$, $A \cap B'$, $B \cap A'$, $B \cap B'$ is empty



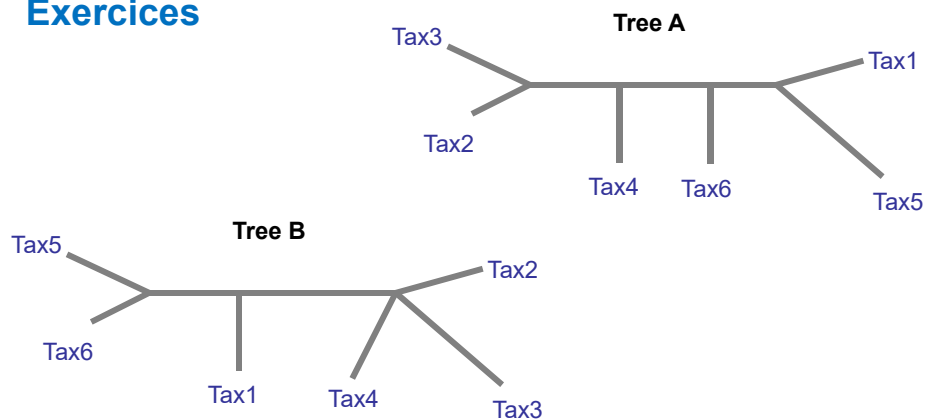
10

Bipartitions (binary characters, splits)

- $L = \{\text{eagle, duck, dog, mouse, kiwi}\}$
- $\text{wings} = \{\text{eagle, duck, kiwi}\} \mid \{\text{mouse, dog}\}$
- $\text{fly} = \{\text{eagle, duck}\} \mid \{\text{kiwi, mouse, dog}\}$
- $\{\text{eagle, duck}\} \cap \{\text{mouse, dog}\} = \emptyset$

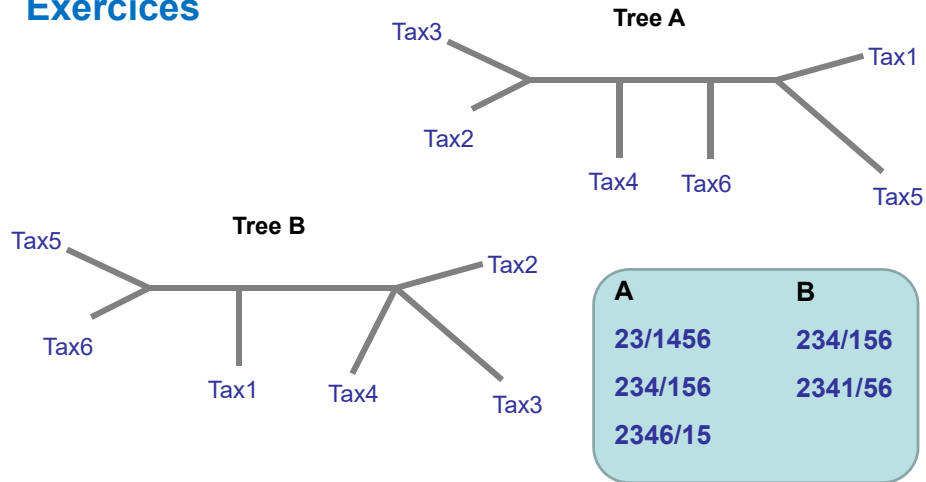


Exercises



Bipartitions of tree A ? Of tree B ?
Which ones are tree compatible ?

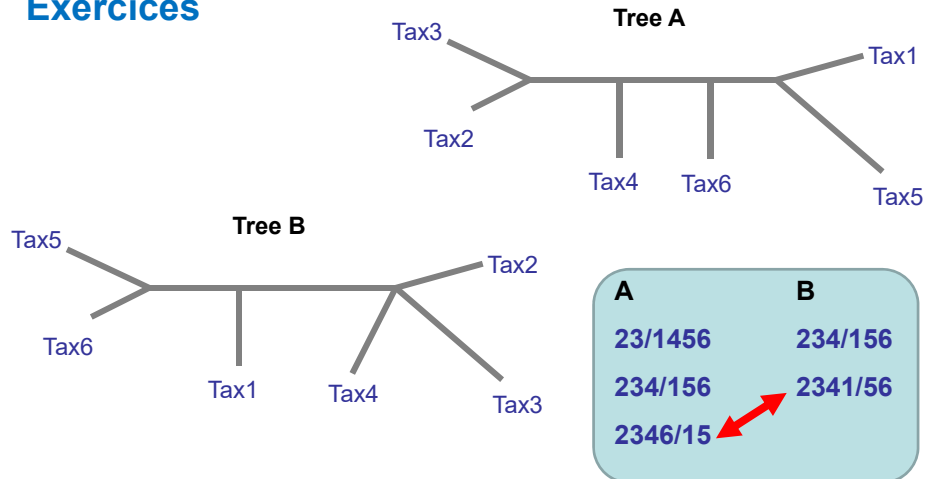
Exercices



**Bipartitions of tree A ? Of tree B ?
Which ones are tree compatible ?**

13

Exercices

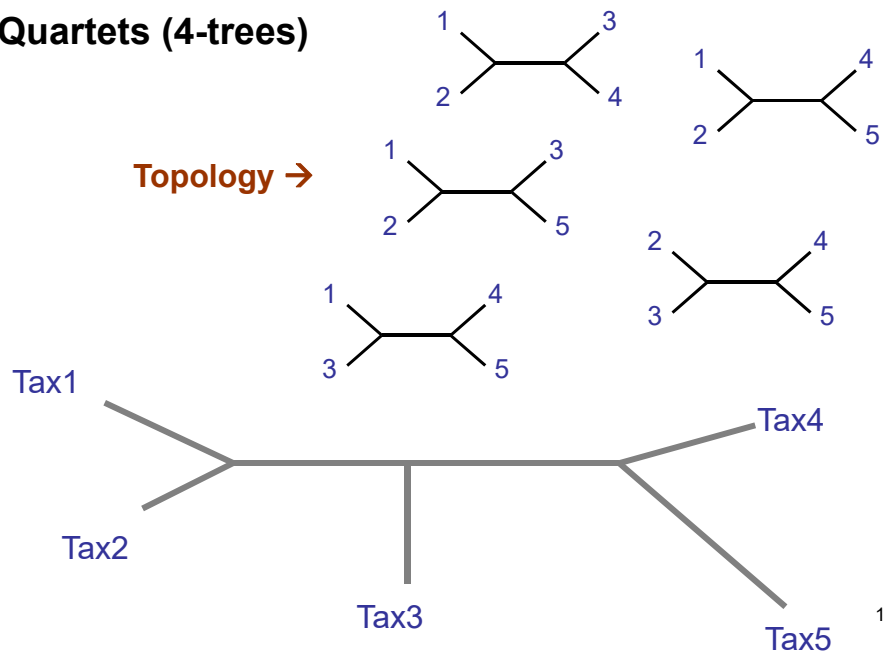


**Bipartitions of tree A ? Of tree B ?
Which ones are tree compatible ?**

14

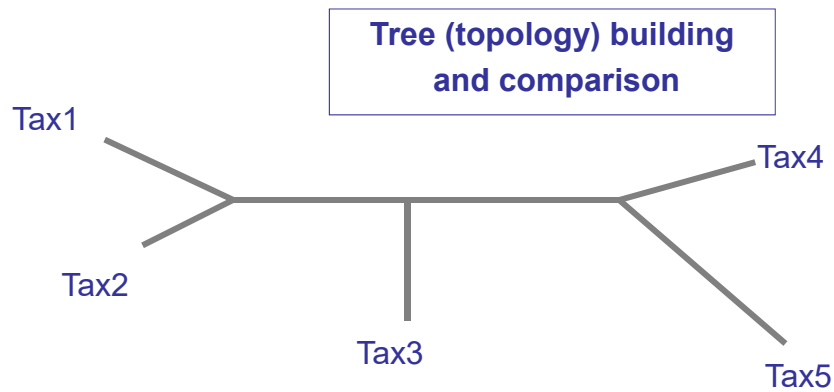
Quartets (4-trees)

Topology →



Quartets (4-trees)

Topology → { 12|34, 12|35, 12|45, 13|45, 23|45 }



Quartets (4-trees)

- A complete quartet set: for every quadruple $\{i, j, k, l\}$ we have one resolved 4-tree, e.g. $ij \mid kl$
- A binary topology defines a complete quartet set
- It is easy to check that a complete quartet set is tree compatible, and then defines a **unique** tree.

A tree is “equal”
to its quartet set.

17

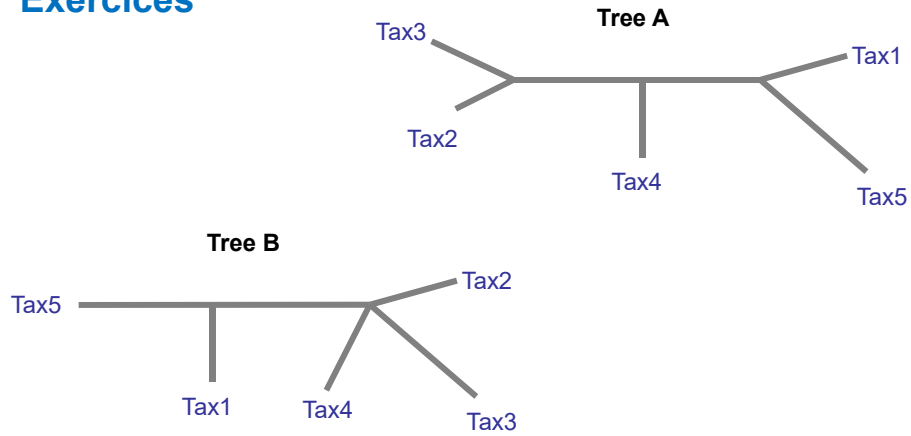
Quartets (4-trees)

- It's easy to infer 4-trees for all quadruples (eg ML)
- **But: 4-trees are not reliable**
- It is computationally hard to check that an incomplete quartet set is tree compatible
- It is computationally hard to select the maximum number of compatible 4-trees

Heuristics needed!

18

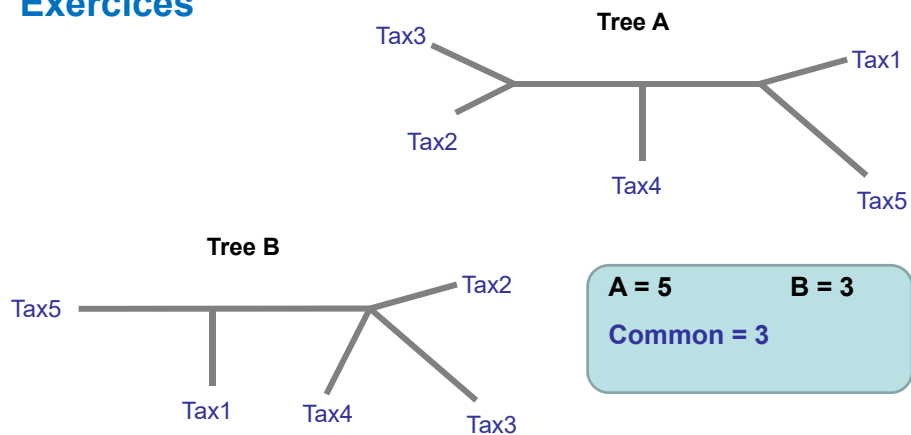
Exercices



**How many quartets are induced by tree A ?
By tree B ? How many in common? Why?**

19

Exercices



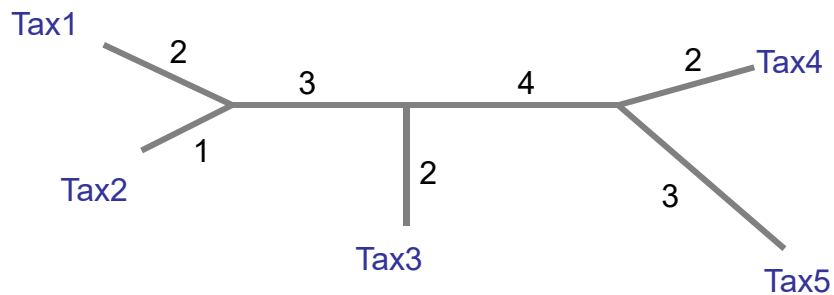
**How many quartets are induced by tree A ?
By tree B ? How many in common? Why?**

20

Additive distances

Tree with branch lengths →

Tree building
(and comparison)

$$\begin{pmatrix} 0 & 3 & 7 & 11 & 12 \\ 3 & 0 & 6 & 10 & 11 \\ 7 & 6 & 0 & 8 & 9 \\ 11 & 10 & 8 & 0 & 5 \\ 12 & 11 & 9 & 5 & 0 \end{pmatrix}$$


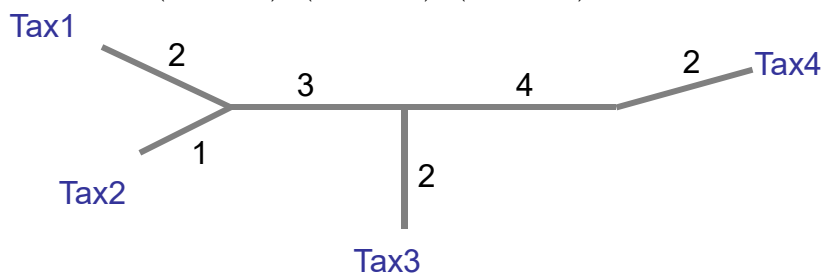
21

Additive distances

$i = 1, j = 2, k = 3, l = 4$

- A tree with lengths defines an “additive distance”.
- A distance is additive iff it satisfies the local “4-point” condition:

For every quadruple i, j, k, l , the two largests of $(\delta_{ij} + \delta_{kl}), (\delta_{ik} + \delta_{jl}), (\delta_{il} + \delta_{jk})$ are equal



22

Additive distances

- A tree with lengths defines an additive distance.
- A distance is additive iff it satisfies the local 4-point condition, which is easily checked.
- An additive distance defines a **unique** tree, which is easily built.

A tree is “equal” to its path length distance

23

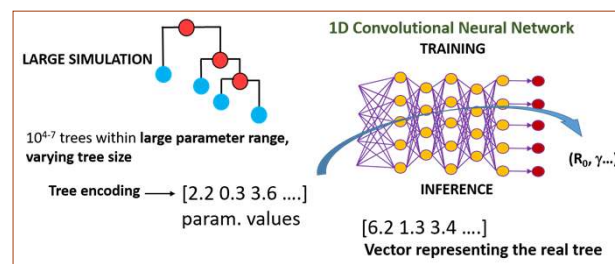
Additive distances

- Estimating evolutionary distances between all taxon pairs is easy (ML)
- **But these distances are never 100% additive**
- This induces hard optimization problems
- Numerous approaches and heuristics

24

Tree encoding for deep learning

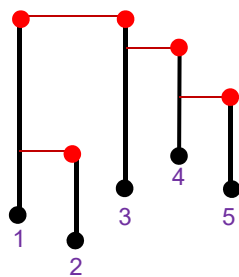
- A phylogenetic tree T inferred from sequences
- A mathematical model of trees M (epidemiological, diversification...) with parameters (e.g. R_0 , speciation rate...) to be estimated from T
- Simulate many trees using M and a large range of parameter values
- Train a deep neural network NN to predict the parameter values, based on the **encoding** of input simulated trees
- Use the trained NN to predict the parameters associated to T 's **encoding**



25

Tree encoding for deep learning

- Our trees are rooted with branch lengths
- Distance matrices do not work, too many weights in deep NN :(
- An unrooted tree can be defined by $(2n-3)$ well chosen distances



Taxa	1	2	3	4	5
Top	0	4	0	1	3
Bottom	6	7	5	4	5
Country	Afr	Afr	Car	US	US

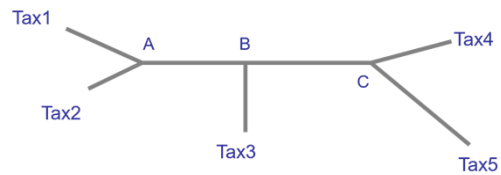
Encoded by $2n$ entries
(+ taxon ordering)

Epidemiology... Voznica... Gascuel, Nature Communications 2022
 Diversification... Lambert Voznica Morlon, Systematic Biology 2023
 Phylogeography... Thompson ... Landis, Systematic Biology 2024

26

Others

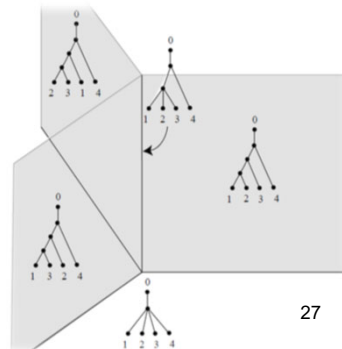
- **Graphs**



- **Mathematical expressions – Newick format**

((Tax1:1, Tax2:1):1, Tax3:1, (Tax5:1, Tax5:1):1);

- **Points in BHV space (Billera-Holmes-Vogtmann)**



Summary

- **Many definitions of trees (bipartitions, quartets, pairwise distances, BHV, combinatorial,)**
- **Used to represent, compare and analyze trees**
- **These definitions involve easy (polynomial) algorithms to recognize trees and change of representation**
- **But hard problems to infer trees from data**

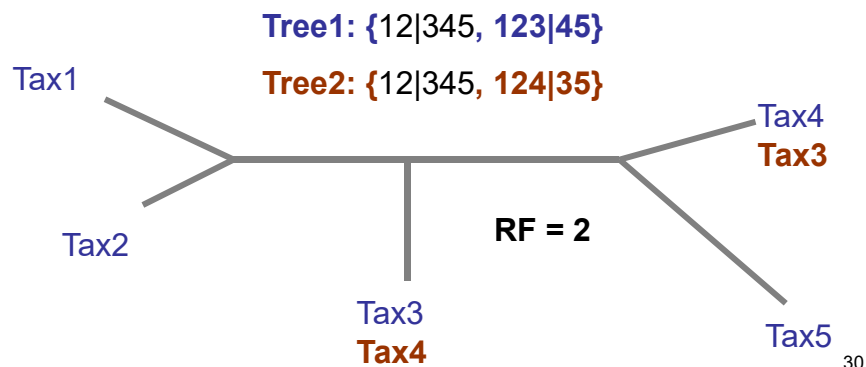
Topological (Tree) distances

- Measure the distance between two topologies with the same taxon set (e.g. two gene trees)
- To analyze alternative trees (e.g. with parsimony)
- To compare reconstruction methods with simulated data
- To infer horizontal gene transfers (gene / species trees)
- ...

29

Bipartition distance (Robinson & Foulds – RF)

- Number of bipartitions in one tree but not the other (easy to compute)

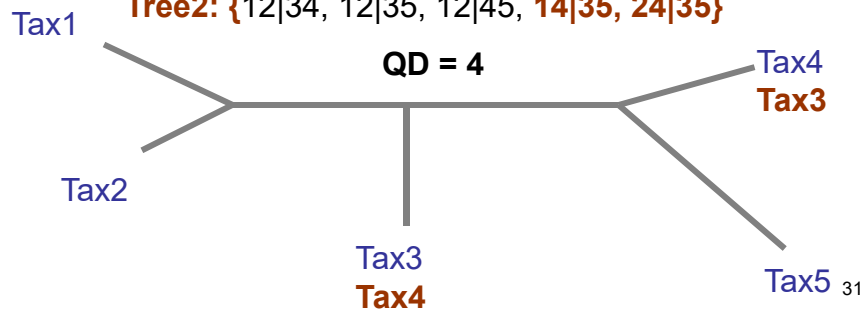


Quartet distance

- Number of 4-trees in one tree but not the other
- Easy to compute, more refined than RF distance

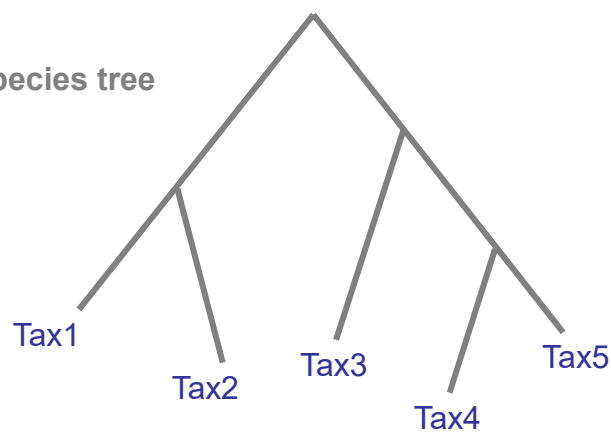
Tree1: {12|34, 12|35, 12|45, **13|45, 23|45**}

Tree2: {12|34, 12|35, 12|45, **14|35, 24|35**}



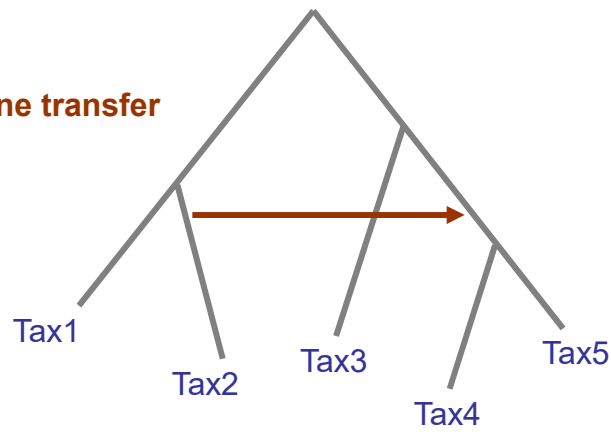
Horizontal gene transfers and SPR distance

Species tree



Horizontal gene transfers and SPR distance

Gene transfer



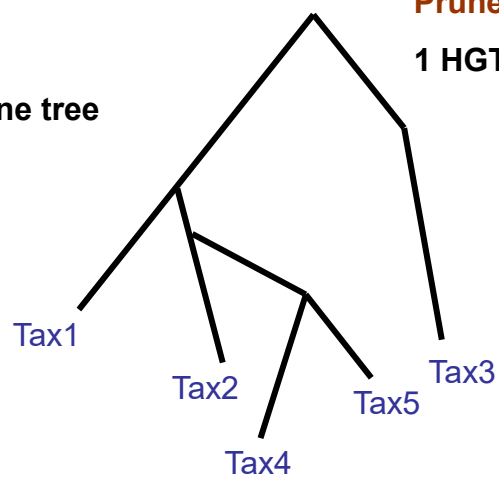
33

Horizontal gene transfers and SPR distance

Subtree (Tax4, Tax5) is
Pruned and Regraft

1 HGT → 1 SPR

Gene tree



34

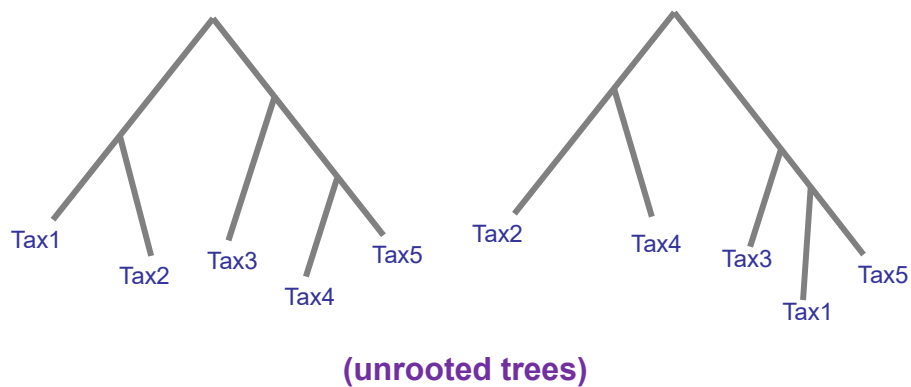
Horizontal gene transfers and SPR distance

- **SPR distance: minimum number of SPR moves required to transform one tree into the other.**
- **Biologically relevant: $\approx$ number of HGTs**
- **Very hard to compute!**

35

Exercise

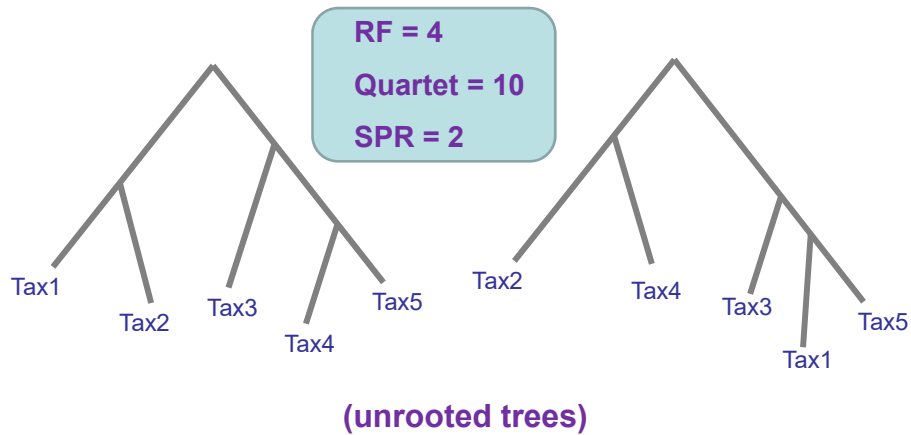
Compute the RF, quartet and SPR distances between:



36

Exercise

Compute the RF, quartet and SPR distances between:



Others

- **Matching distance (allows errors, more refined than RF)**

every branch in tree 1 is matched to a branch in tree 2

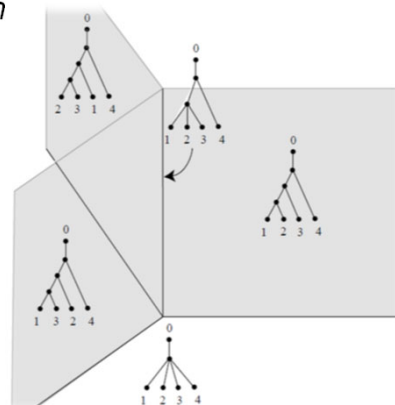
= numbers of errors in best match

- **Weighted RF**

accounts for branch lengths

- **Geodesic distance in BHV**

accounts for branch lengths



Summary

Many distances between trees

Some:

- Easy to compute (e.g. RF, quartet, geodesic)
- With biological interpretation (e.g. SPR)
- Refined (e.g. matching, quartet)
- Accounts for branch lengths (e.g. geodesic, weighted RF)

39

Consensus

- We aim at estimating the consensus of a family of trees with the same taxon set.
- Most consensus problems are hard
- But it's easy to define and compute the majority rule consensus tree

40

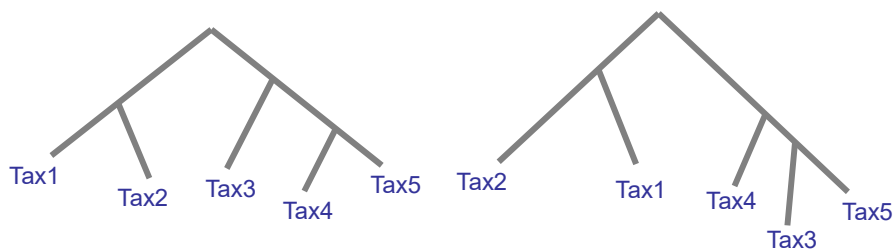
Majority rule consensus tree

- n trees with the same taxon set
- every tree t defines a bipartition set $B_t = \{b\}$
- collect $B = \{b \text{ seen in } > n/2 \text{ sets } B_t\}$
- any pair b, b' is seen in at least one common set (tree) B_t
- therefore, b and b' are tree compatible
- and B defines a unique tree!

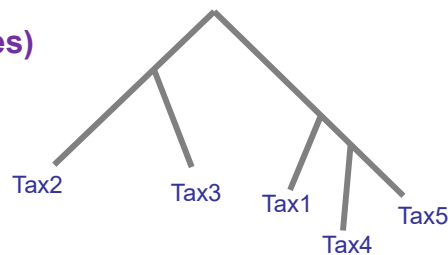
As this tree may be poorly resolved, we often use the extended version (e.g. CONSENSE from Phylip)

41

Exercise: Compute the majority consensus tree between

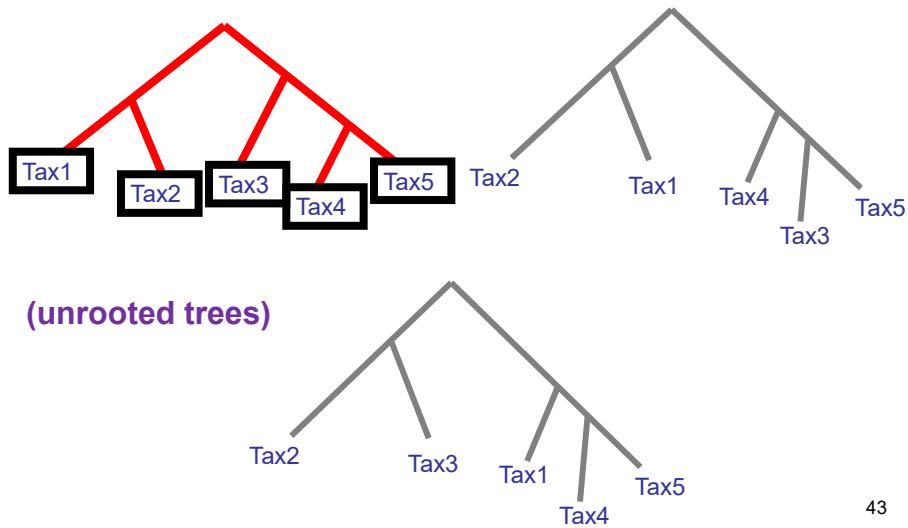


(unrooted trees)



42

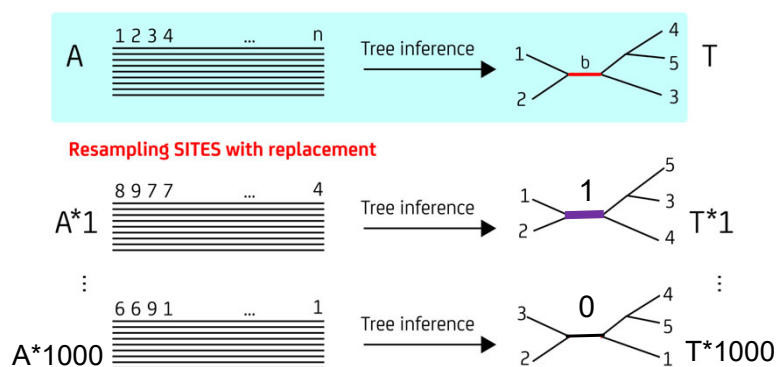
Exercise: Compute the majority consensus tree between



43

The phylogenetic bootstrap

Joe Felsenstein 1985 ~50,000 citations

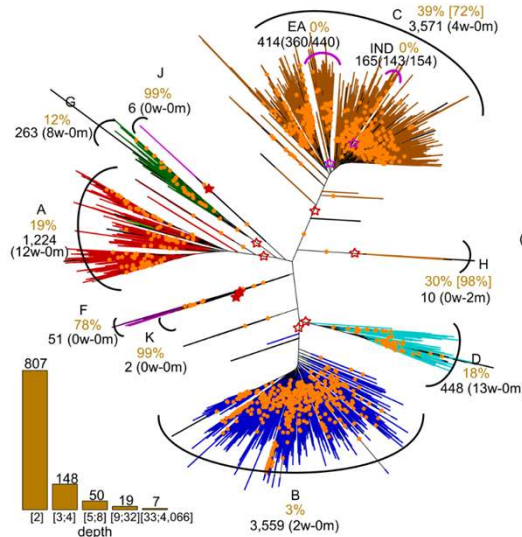


- **FBP = Proportion of bootstrap trees** containing the exact same **SPLIT**
- For a given T^* , b is either **present** or **absent**: **binary 0/1 function**

44

FBP does not work with large trees

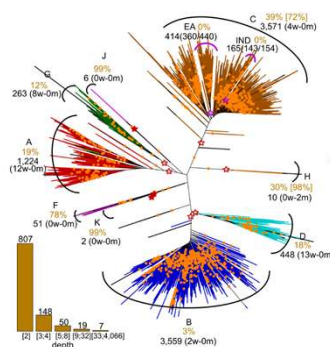
- 9,147 HIV *pol* sequences
- 9 subtypes coloured using jpHMM
- Strong signal re. subtypes
- Orange tablets: FBP > 70%
- No support for subtypes (deep branches)



45

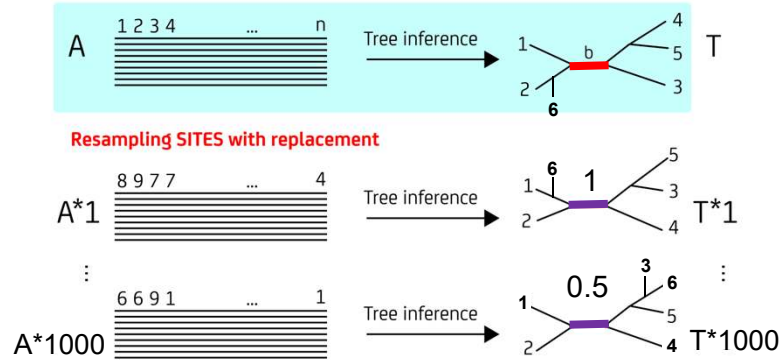
Why FBP fails with large number of taxa?

- A single « rogue taxon » is enough to drastically reduce FBP
- Recombination, convergence...
- Partial sequences, contamination...
- Reconstruction errors...
- The more taxa, the higher the probability of « rogue taxa »
- A strong impact on deep branches



46

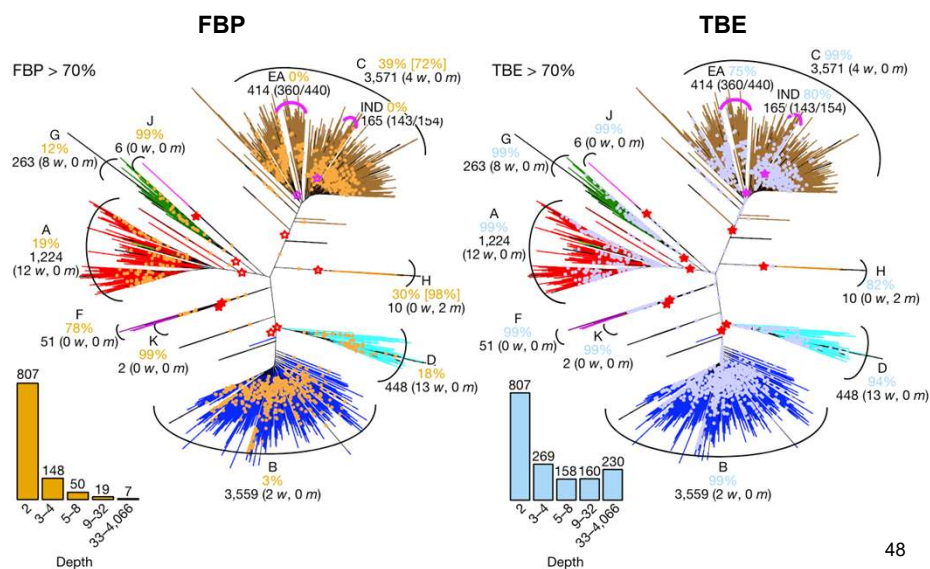
Transfer Bootstrap Expectation (*Nature* 2018) - Key Idea!



- We replace the 0/1 function of FBP by a **"continuous" function in [0,1]**, which measures the presence of b in T^* and **allows for errors**.
- We estimate the **expectation of this continuous function** using bootstrap trees

47

The 9 HIV subtypes are highly supported by TBE



48

Renewing Felsenstein's phylogenetic bootstrap in the era of big data

F. Lemoine^{1,2}, J.-B. Domelevo Entfellner^{3,4}, E. Wilkinson⁵, D. Correia¹, M. Davila Felipe¹, T. De Oliveira^{5,6} & O. Gascuel^{1,7*}

The resulting supports are higher and do not induce falsely supported branches. The application of our method to large mammal, HIV and simulated datasets reveals their phylogenetic signals, whereas Felsenstein's bootstrap fails to do so.

Big means many taxa!

49

Syst. Biol. XX(X):1–16, 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of the Society of Systematic Biologists. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

<https://doi.org/10.1093/sysbio/syad052>

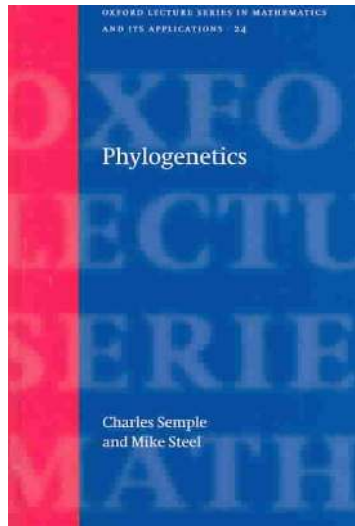
Advance Access Publication XXXX XX, XXXX

Robustness of Felsenstein's Versus Transfer Bootstrap Supports With Respect to Taxon Sampling

PAUL ZAHARIAS^{1,*} , FRÉDÉRIC LEMOINE^{2,3}  AND OLIVIER GASCUEL^{1,*} 

Our results show that the main critique of TBE stands in extreme cases with shallow branches and highly unbalanced sampling among clades, but that TBE is still robust in most cases, while FBP is inescapably negatively impacted by high taxon sampling.

50



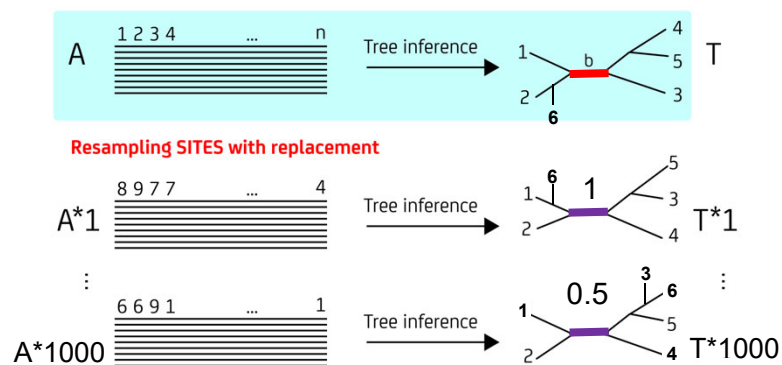
Mike Steel



Charles Semple

51

Transfer Bootstrap Expectation (*Nature* 2018) - Key Idea!

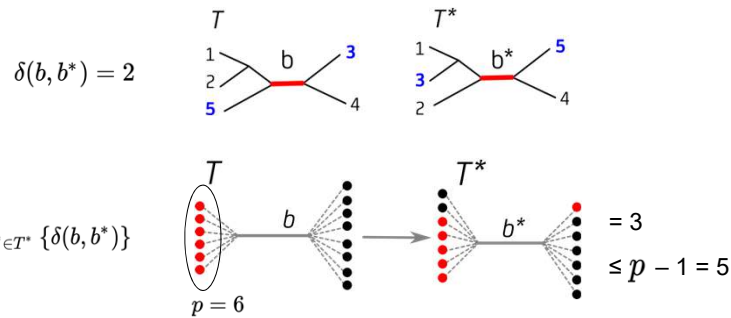


- We replace the 0/1 function of FBP by a **"continuous" function in [0,1]**, which measures the presence of b in T^* .
- We estimate the **expectation of this continuous function** using bootstrap trees

52

The Transfer Distance and Index

Davila Felipe et al.
J. Math. Biol. 2019



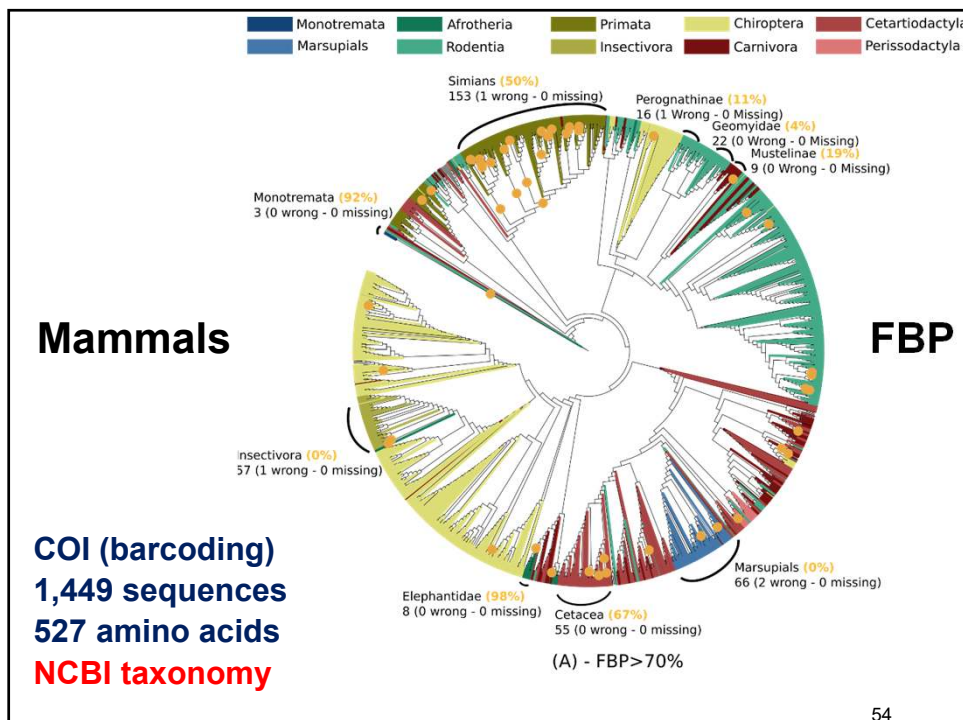
$$TBE = 1 - \left(\frac{\phi(b, T^*)}{p-1} \right)$$

$$0 \leq TBE \leq 1$$

Negligible CPU
 $O(n^2 \times \text{\#boots})$

- $TBE(b) = 1$ iff b belongs to all bootstrap trees T^*
- $TBE(b) \approx 0$ with random “bootstrap” trees
- $TBE(b) \geq FBP(b) \gg$ with large p = with cherries ($p = 2$)
- TBE is nearly insensitive to the presence of rogue taxa
- Conflictual branches are not supported (Mammals/Simul.)
- Easily interpreted as a number of taxa to be transferred

53



54

