Easy Mathematics of Phylogenetic Trees

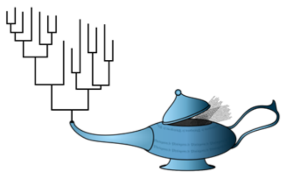
Olivier Gascuel

Directeur de Recherche au CNRS

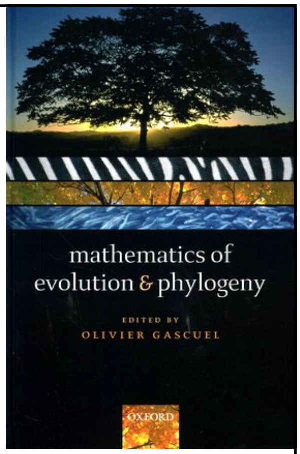
Institut de Systématique, Evolution, Biodiversité (ISYEB)
Muséum National d'Histoire Naturelle

Académie des Sciences

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



PhyML

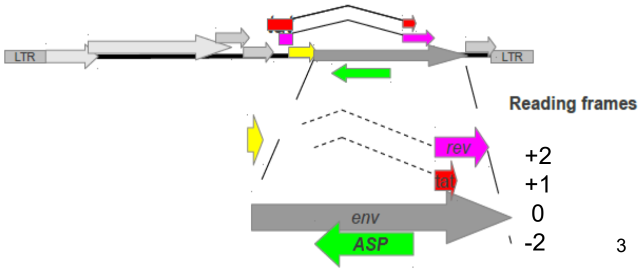


Phylogenetics

- 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…)

Pathogens and epidemiology

- Origin and functional genomics of *Plasmodium falciparum*
- History of HIV-1 subtypes and CRFs (PLOS Path 2021)
- HIV-1 drug resistance mutations (AIDS, Viruses 2023)
- The 10th gene of HIV-1 M (PNAS, 2016)
- Phylodynamics (Nature Comm 2022)



Easy Mathematics of Phylogenetic Trees

- **Numbers**
  - **Number of branches in a tree**
  - **Number of trees**
- **Many definitions of trees**
  - **Trees as bipartitions**
  - **Trees as quartets**
  - **Trees as distance matrices**
  - **Tree encoding for deep learning**
  - **Others…**
- **Topological distances**
  - **Bipartition distance**
  - **Quartet distance**
  - **SPR distance**
  - **Others…**
- **Consensus of trees**
- **Bootstrap branch supports**
  - **Felsenstein's bootstrap proportion (FBP)**
  - **Transfer bootstrap expectation (TBE)**

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:** $e(n) = e(n\text{-}1) + 2 = 2n - 3$

Numbers

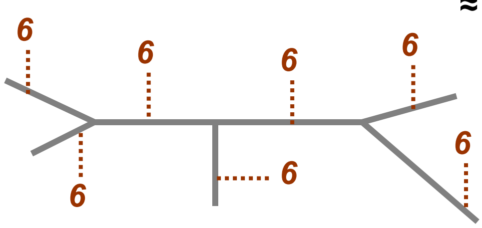
Number of unrooted binary trees with *n* taxa: :

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:** $t(n) = t(n - 1) \times e(n - 1) = (2n – 5)(2n – 7) \ldots$

$\approx n^n$



→ hard optimization problems!

Numbers

How many edges in a rooted tree ?

How many rooted trees ?

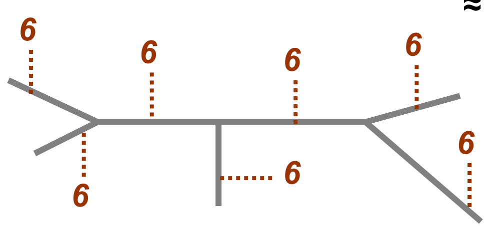
Number of unrooted binary trees

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

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

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

$\approx n^n$



7

Numbers

How many edges in a rooted tree ?

2n-2

How many rooted trees ?

t(*n*+1)

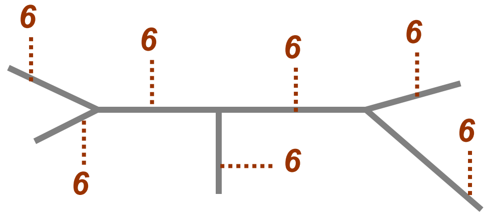
Number of unrooted binary trees

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

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

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

$\approx n^n$



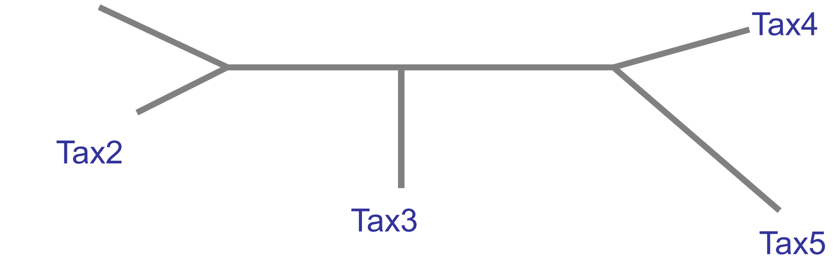
8

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}
}

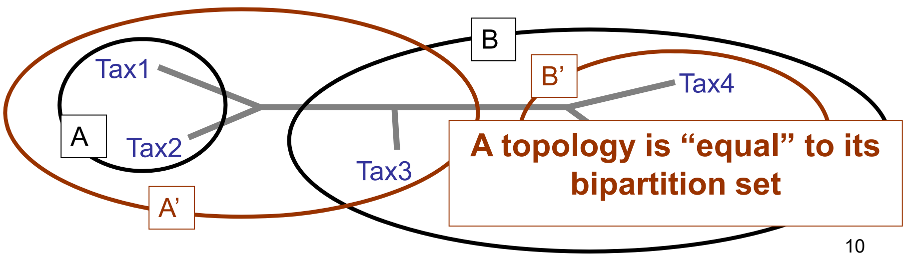
Tree and network building
Topology comparison



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:

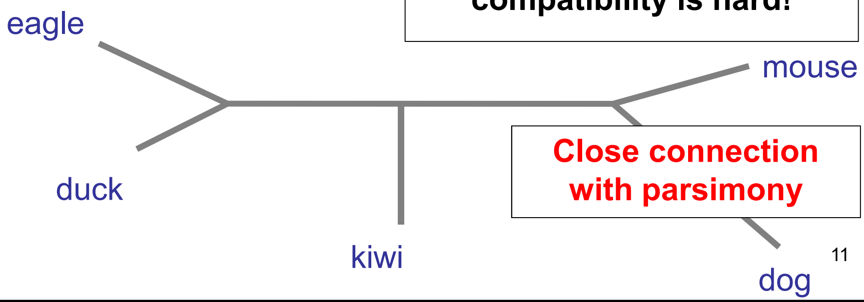
A | B and A' | B' are tree compatible iff
one of $A \cap A'$, $A \cap B'$, $B \cap A'$, $B \cap B'$ is empty



Bipartitions (binary characters, splits)

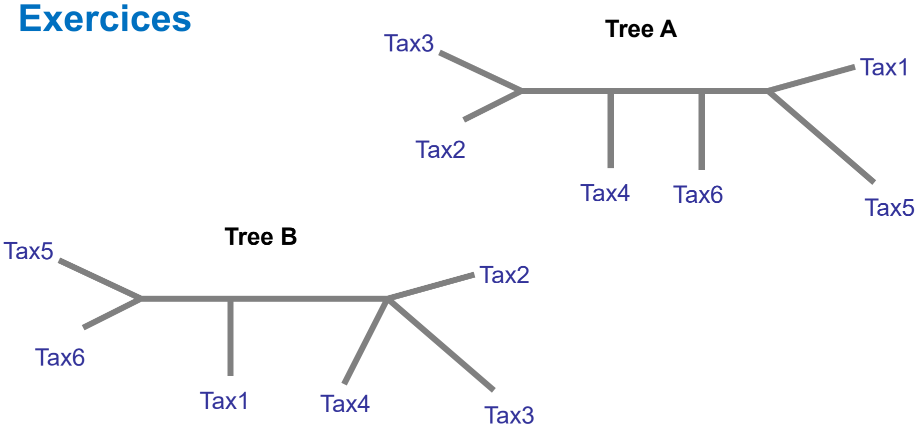
- L = {eagle, duck, dog, mouse, kiwi)
- **wings = {eagle, duck, kiwi} | {mouse, dog}**
- fly = {eagle, duck} | {kiwi, mouse, dog}
- {eagle, duck} ∩ **{mouse, dog}** = ∅

Maximizing character compatibility is hard!



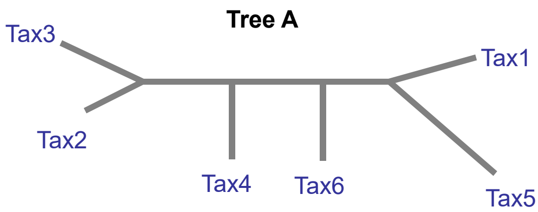
Close connection with parsimony

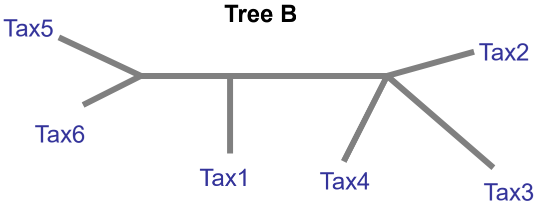
Exercices



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

Exercices



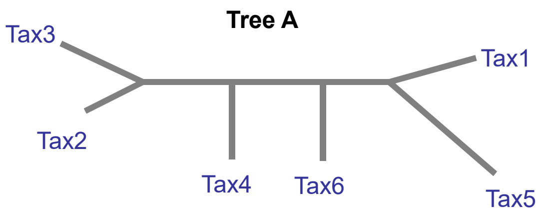


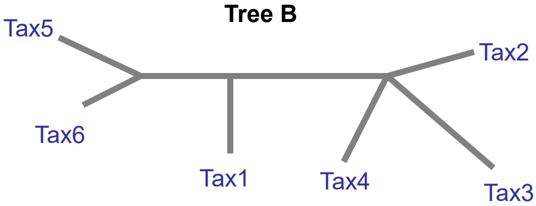
A	B
23/1456	234/156
234/156	2341/56
2346/15	

Bipartitions of tree A ? Of tree B ?

Which ones are tree compatible ?

Exercices





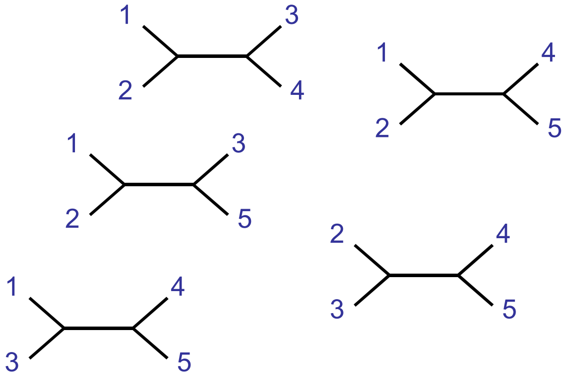
A	B
23/1456	234/156
234/156	2341/56
2346/15	

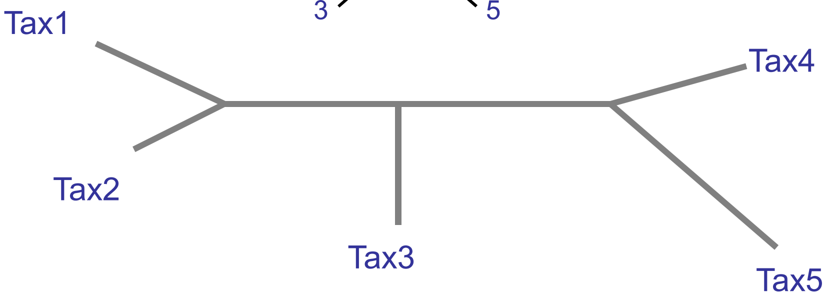
Bipartitions of tree A ? Of tree B ?

Which ones are tree compatible ?

Quartets (4-trees)

Topology →

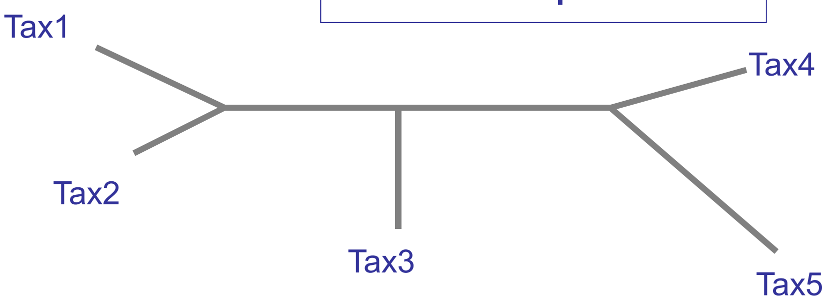




Quartets (4-trees)

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

Tree (topology) building and comparison



Quartets (4-trees)

- A complete quartet set: for every quadruple {i, j, k, l} we have one resolved 4-tree, e.g. ij | kl
- A binary topology defines a complete quartet set
- It 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.

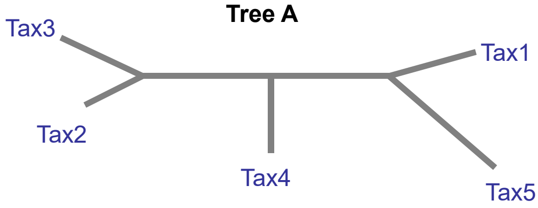
Quartets (4-trees)

- 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

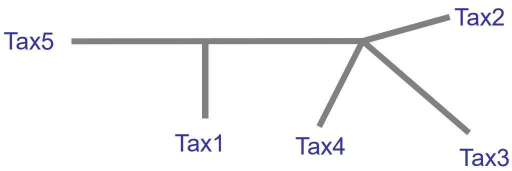
Heuristics needed!

Exercices

Tree A



Tree B



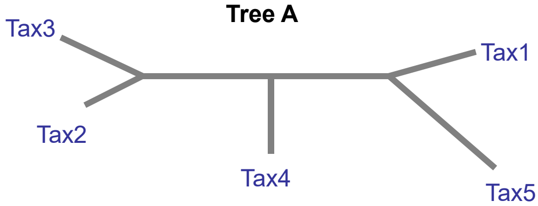
How many quartets are induced by tree A ?
By tree B ? How many in common? Why?

19

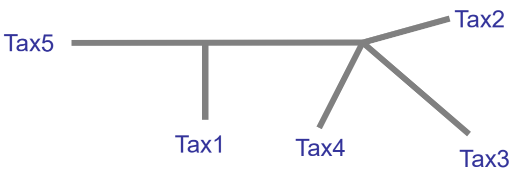
Olivier Gascuel – CNRS – Muséum National d'Histoire Naturelle – Académie des Sciences – Phylogenomics-Cesky-2024

Exercices

Tree A



Tree B



A = 5 **B = 3**

Common = 3

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

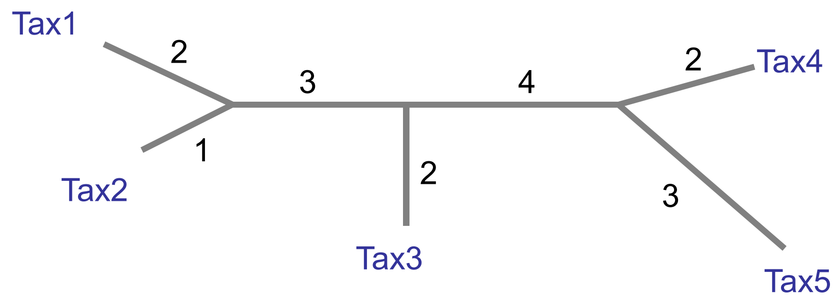
20

Additive distances

Tree with branch lengths →

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

Tree building (and comparison)



21

Additive distances

17

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

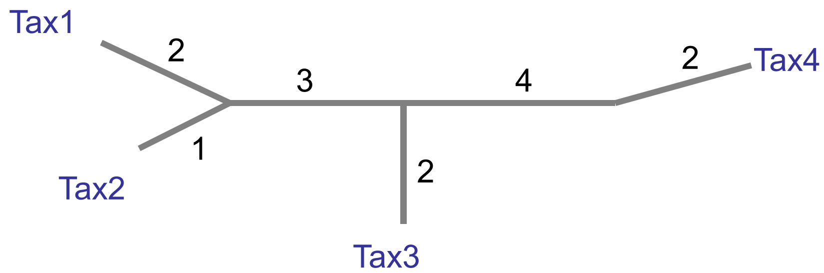
- **A tree with lengths defines an "additive distance".**
- **A distance is additive iff it satisfies the local "4-point" condition:**

11

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

17



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.

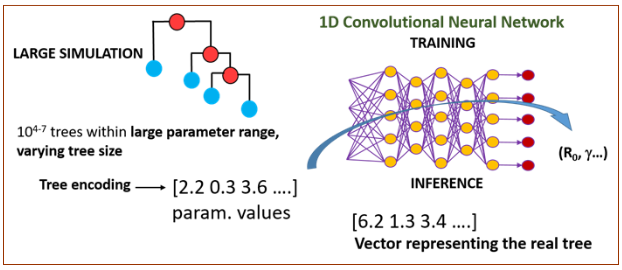
A tree is "equal" to its path length distance

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

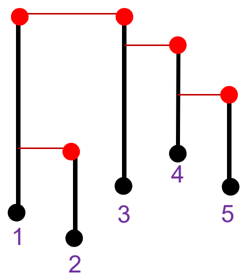
Tree encoding for deep learning

- 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**



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 (2*n*-3) well chosen distances



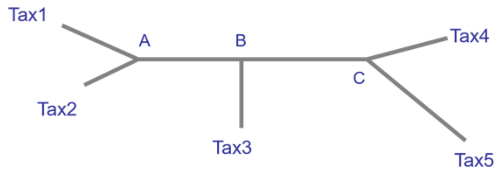
Taxa	1	2	3	4	5
Top	0	4	0	1	3
Bottom	6	7	5	4	5
Country	Afr	Afr	Car	US	US

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

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

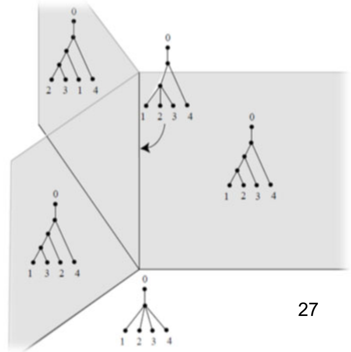
Others

- 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

Topogical (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)
- …

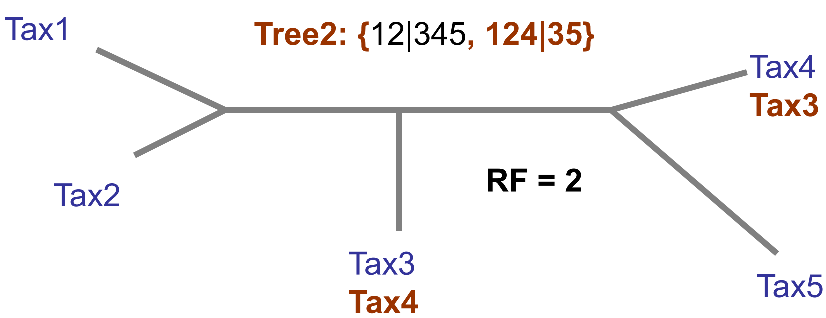
29

Bipartition distance (Robinson & Foulds – RF)

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

Tree1: {12|345, **123|45}**

Tree2: {12|345, **124|35}**



30

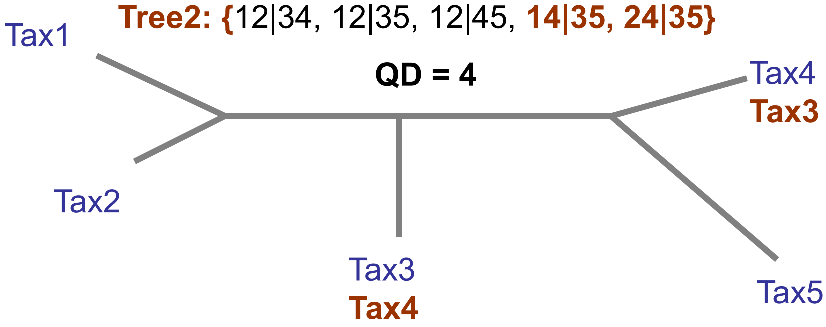
Quartet distance

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

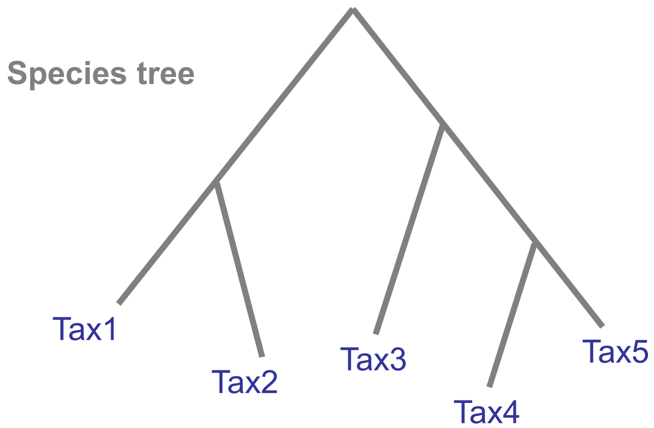
Tree1: {12|34, 12|35, 12|45, **13|45, 23|45}**

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

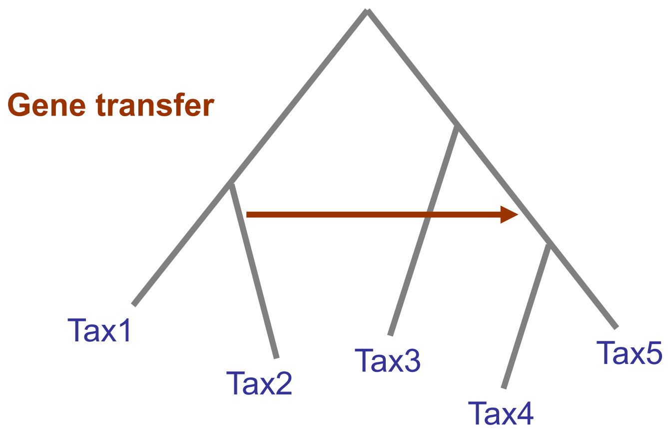
QD = 4



Horizontal gene transfers and SPR distance



Horizontal gene transfers and SPR distance

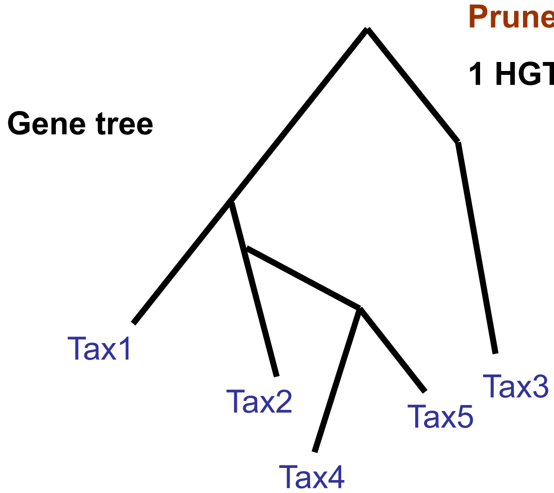


Olivier Gascuel – CNRS – Muséum National d'Histoire Naturelle – Académie des Sciences – Phylogenomics-Cesky-2024

Horizontal gene transfers and SPR distance

Subtree (Tax4, Tax5) is Pruned and Regraft

1 HGT → 1 SPR

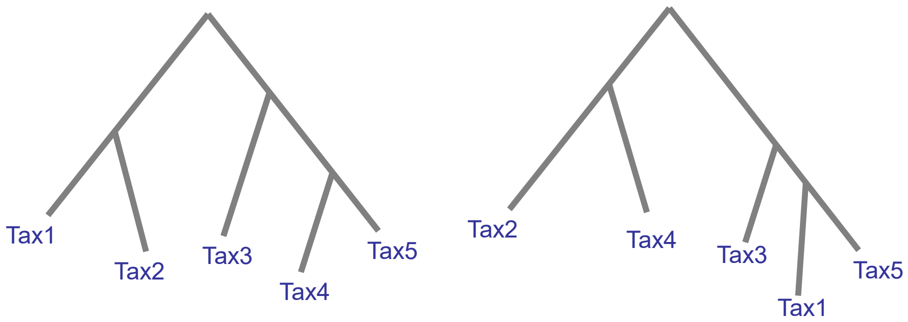


Horizontal gene transfers and SPR distance

- SPR distance: minimum number of SPR moves required to tansform one tree into the other.
- Biologically relevant: ≈ number of HGTs
- Very hard to compute!

Exercice

Compute the RF, quartet and SPR distances between:



(unrooted trees)

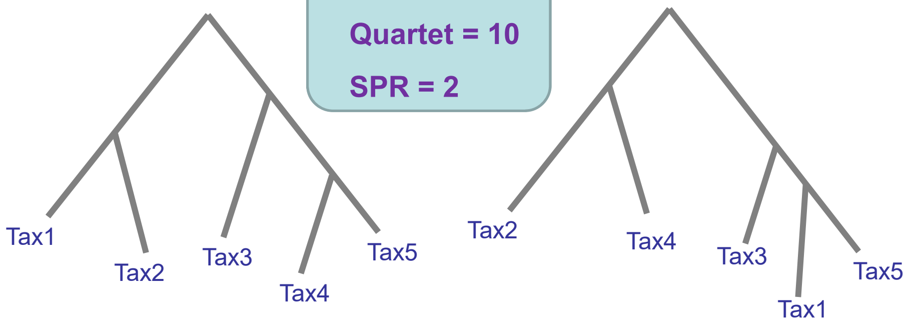
Exercice

Compute the RF, quartet and SPR distances between:

RF = 4

Quartet = 10

SPR = 2



(unrooted trees)

Others

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

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

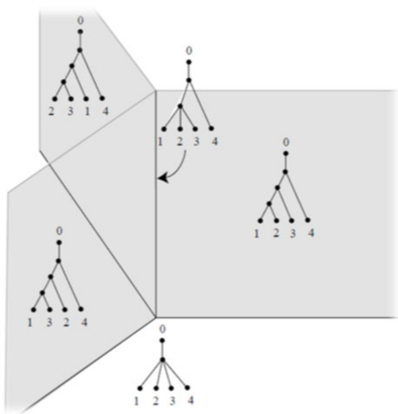
  = numbers of errors in best match

- **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)

Consensus

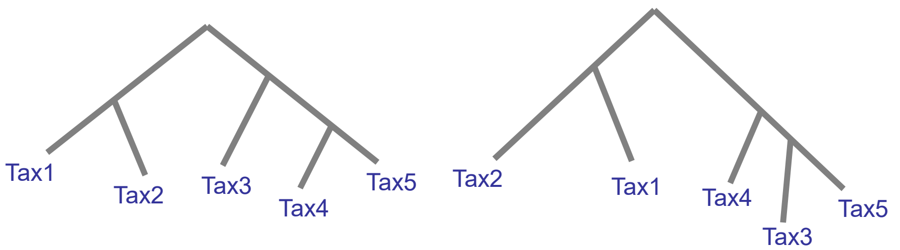
- We aim at estimating the consensus of a family of trees with the same taxon set.
- Most consensus problems are hard
- But it's easy to define and compute the majority rule consensus tree

Majority rule consensus tree

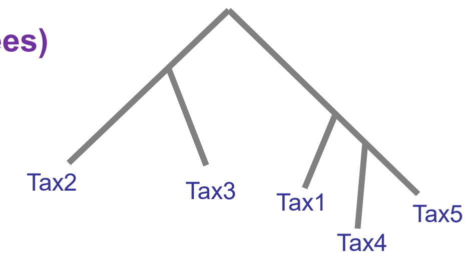
- *n* trees with the same taxon set
- every tree *t* defines a bipartition set B*t* = {b}
- collect B = { b seen in > *n*/2 sets B*t* }
- any pair b, b' is seen in at least one common set (tree) B*t*
- therefore, b and b' are tree compatible
- and B defines a unique tree!

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

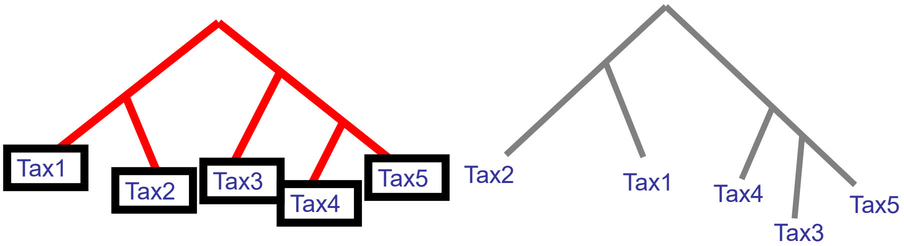
Exercice: Compute the majority consensus tree between



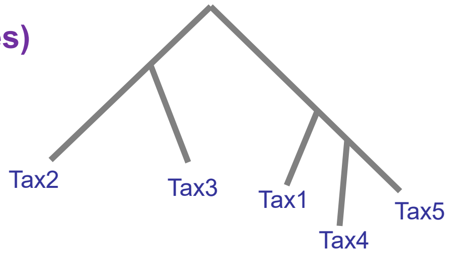
(unrooted trees)



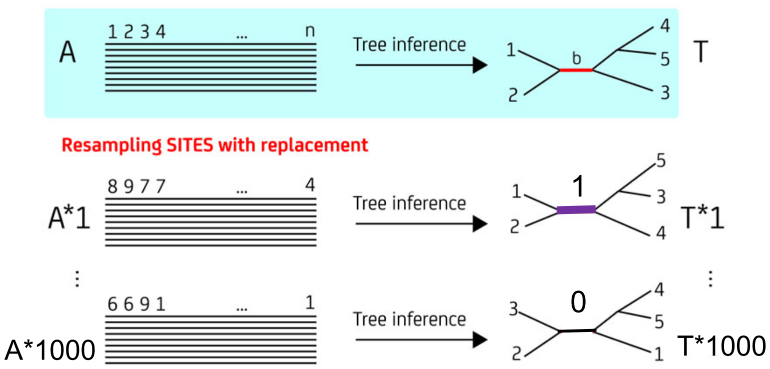
Exercice: **Compute the majority consensus tree between**



(unrooted trees)



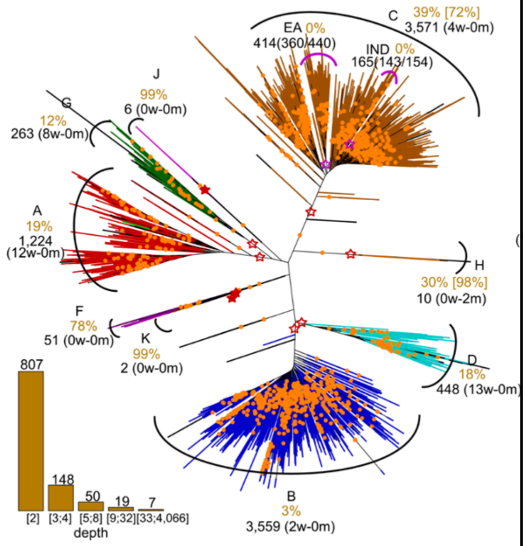
The phylogenetic bootstrap
Joe Felsenstein 1985 ~50,000 citations



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

FBP does not work with large trees

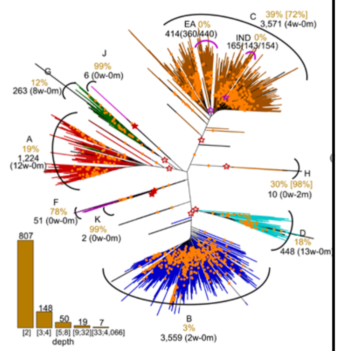
- 9,147 HIV *pol* sequences
- 9 subtypes coloured using jpHMM
- Strong signal re. subtypes
- Orange tablets: FBP > 70%
- No support for subtypes (deep branches)



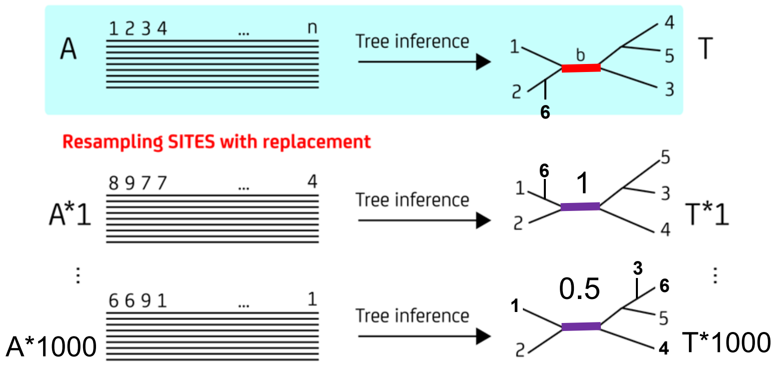
Olivier Gascuel – CNRS – Muséum National d'Histoire Naturelle – Académie des Sciences – Phylogenomics-Cesky-2024

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**

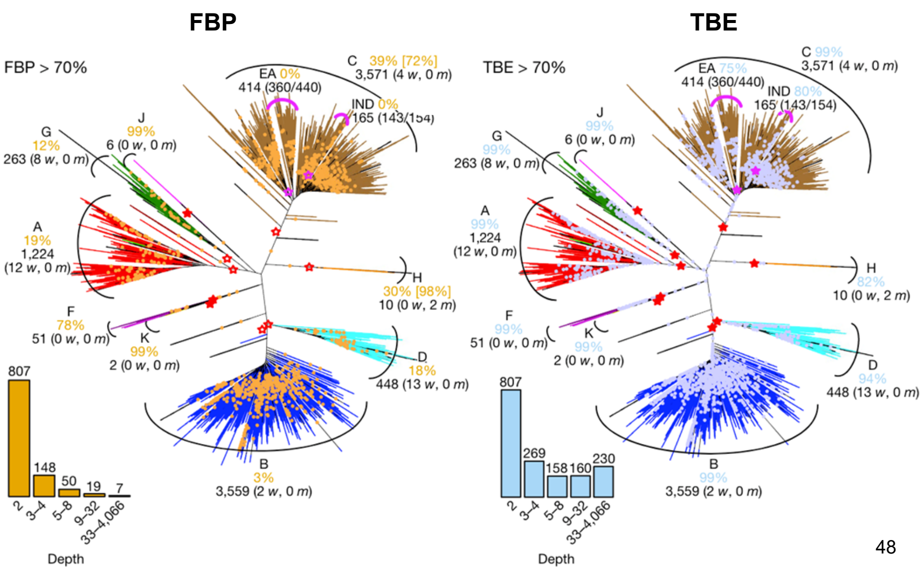


Transfer Bootstrap Expectation (*Nature* 2018) - Key Idea!



- 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

The 9 HIV subtypes are highly supported by TBE



ARTICLE **Nature 2018**

https://doi.org/10.1038/s41586-018-0043-0

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

F. Lemoine[1,2], J.-B. Domelevo Entfellner[3,4], E. Wilkinson[5], D. Correia[1], M. Dávila Felipe[1], 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

Systematic Biology 2023

Syst. Biol. XX(XX):1–16, 2023

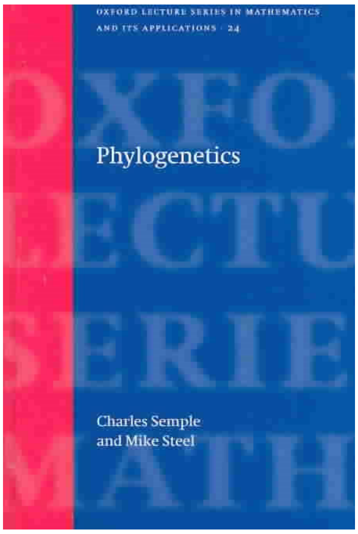
https://doi.org/10.1093/sysbio/syad052
Advance Access Publication XXXX XX, XXXX

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.

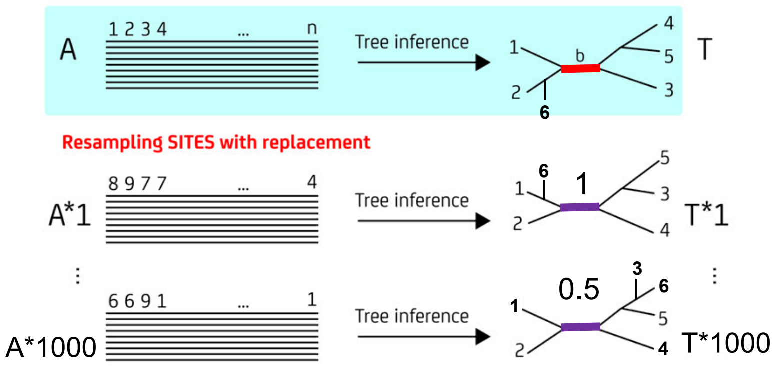
50



Mike Steel

Charles Semple

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



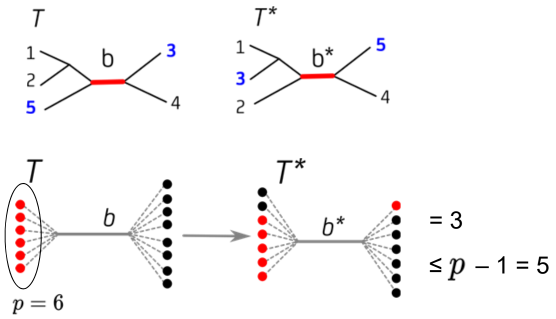
- We replace the 0/1 function of FBP by a "**continuous" function in [0,1],** which measures the presence of b in T*.
- We estimate the **expectation of this continuous function** using bootstrap trees

The Transfer Distance and Index

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

$\delta(b, b^*) = 2$

$\phi(b, T^*) = min_{b^* \in T^*} \{\delta(b, b^*)\}$



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

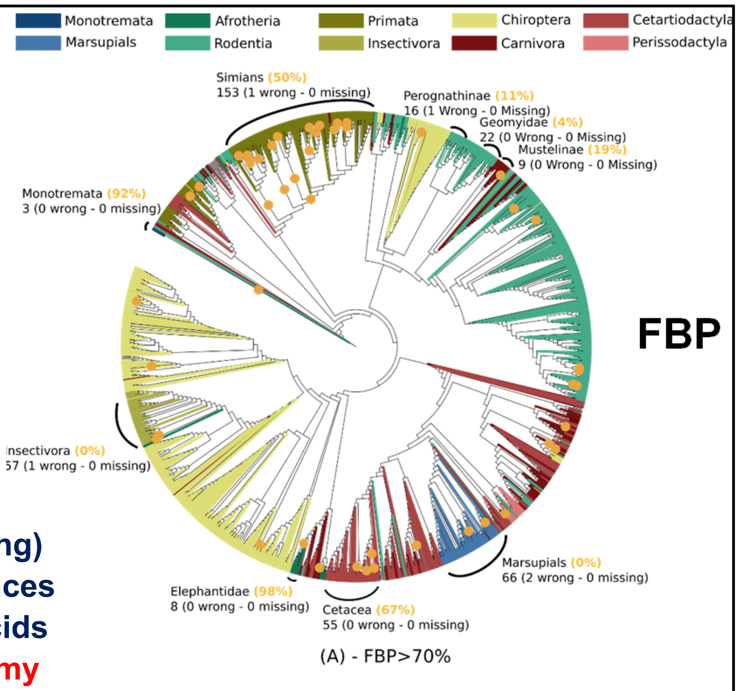
$0 \leq TBE \leq 1$

Negligible CPU
O(n[2] x #boots)

- TBE(b) = 1 iff b belongs to all bootstrap trees T*
- TBE(b) ≈ 0 with random "bootstrap" trees
- TBE(b) ≥ FBP(b) >> 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

Mammals



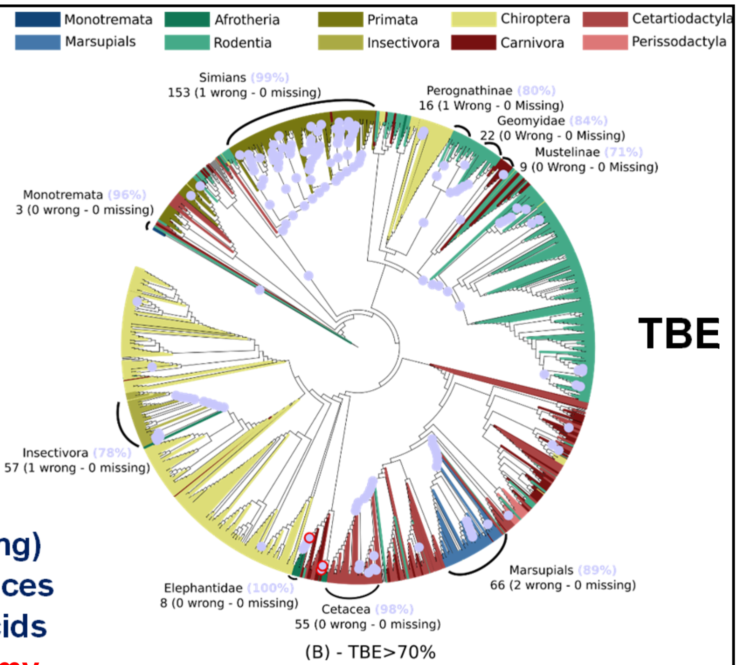
(A) - FBP>70%

FBP

COI (barcoding)
1,449 sequences
527 amino acids
NCBI taxonomy

54

Mammals



TBE

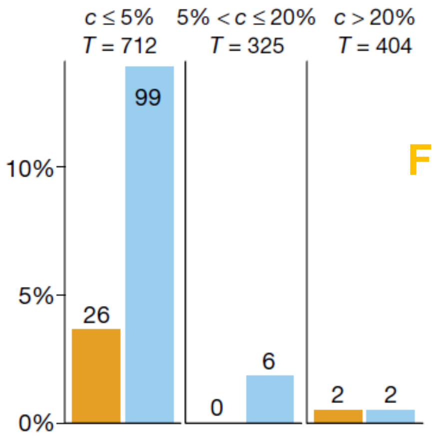
COI (barcoding)
1,449 sequences
527 amino acids
NCBI taxonomy

55

No support for conflictual branches

Quartet conflicts **low** **moderate** **high**

Mammals



FBP vs TBE

COI (barcoding)
1,449 sequences
527 amino acids
NCBI taxonomy
Same results with simulations

56