Robustesse des Phylogénies

M. Mariadassou A. Bar-Hen H. Kishino

Laboratoire MAP5
Université Paris Descartes

Novembre 2008
GDR Statistique et Santé

Molecular Phylogeny

Main Goal:

Use biological macromolecules (DNA, proteins) to unravel the evolutionary history of a set of species

Basic Ideas:

- Closely related species: highly similar molecules,
- Distantly related species: not so similar molecules,
- Use similarity information to reconstruct probable evolution,

Results:

- Evolution is assumed to be tree-like,
- Results are displayed as a phylogenetic tree.

Molecular Phylogeny

Main Goal:

Use biological macromolecules (DNA, proteins) to unravel the evolutionary history of a set of species

Basic Ideas:

- Closely related species: highly similar molecules,
- Distantly related species: not so similar molecules,
- Use similarity information to reconstruct probable evolution,

Results:

- Evolution is assumed to be tree-like,
- Results are displayed as a phylogenetic tree.

Molecular Phylogeny

Main Goal:

Use biological macromolecules (DNA, proteins) to unravel the evolutionary history of a set of species

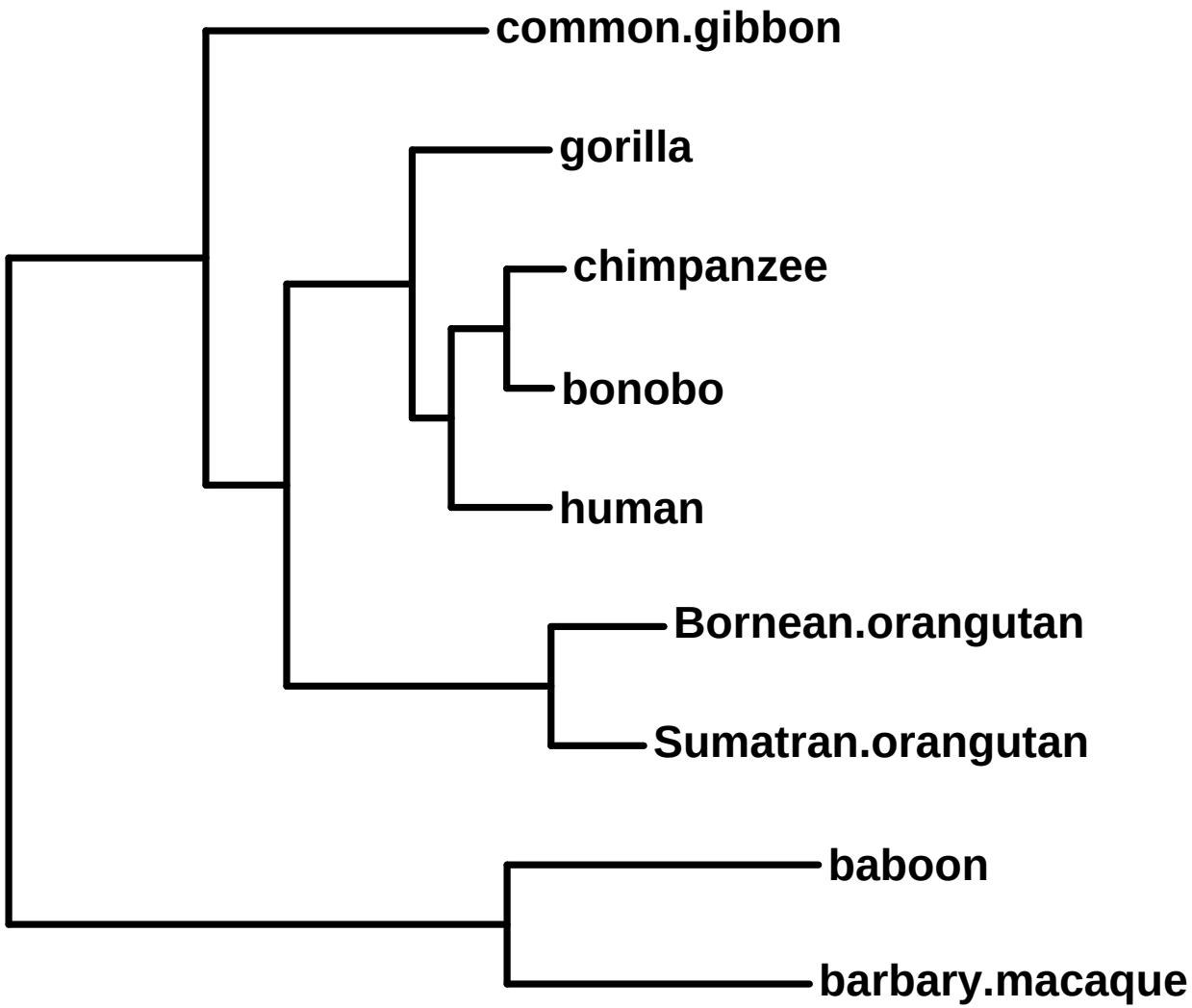
Basic Ideas:

- Closely related species: highly similar molecules,
- Distantly related species: not so similar molecules,
- Use similarity information to reconstruct probable evolution,

Results:

- Evolution is assumed to be tree-like,
- Results are displayed as a phylogenetic tree.

Example of A Phylogenetic Tree



Phylogenies: A Widely Used Tool

Prominent Role in Bioinformatics:

- Distinguish orthologous genes from paralogous ones,
- Phylogenomic profiling,
- Phylogenetic footprinting.

Species-Level Studies:

- Establish Tree of Life, phylogeny of all living species,
- Barcoding to identify new species and their relation to previously existing species,
- Natural way to measure biodiversity.

Phylogenies: A Widely Used Tool

Prominent Role in Bioinformatics:

- Distinguish orthologous genes from paralogous ones,
- Phylogenomic profiling,
- Phylogenetic footprinting.

Species-Level Studies:

- Establish Tree of Life, phylogeny of all living species,
- Barcoding to identify new species and their relation to previously existing species,
- Natural way to measure biodiversity.

Phylogenies and Epidemiology

Fundamental in Modern Epidemiology:

Powerful tool to

- Identify and classify rapidly evolving pathogens,
- Trace the history of infections,
- Predict outbreaks.

Recent Examples:

- Crucial in identifying SARS,
- Useful to study relationships between virulence and genetic evolution (HIV, influenza)

Phylogenies and Epidemiology

Fundamental in Modern Epidemiology:

Powerful tool to

- Identify and classify rapidly evolving pathogens,
- Trace the history of infections,
- Predict outbreaks.

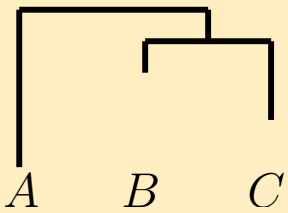
Recent Examples:

- Crucial in identifying SARS,
- Useful to study relationships between virulence and genetic evolution (HIV, influenza)

Reconstructions and Limits

Two levels of reconstruction:

- Reconstruct the phylogeny:
  - Topology,
  - Branchs lengths.



- Reconstruct ancestral states (*e.g.* gene of ancestor).

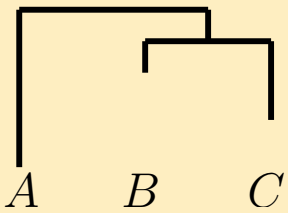
Problems:

- Evolution is a unique event,
- Genetic information available only for extant species,
- Almost no direct observations or results on the evolutionary process.

Reconstructions and Limits

Two levels of reconstruction:

- Reconstruct the phylogeny:
  - Topology,
  - Branchs lengths.



- Reconstruct ancestral states (*e.g.* gene of ancestor).

Problems:

- Evolution is a unique event,
- Genetic information available only for extant species,
- Almost no direct observations or results on the evolutionary process.

Data Structure

Collection: Select gene/protein shared by all species, sequence it and align the sequences.

Example:

- Alignment $\mathcal{X} = (X_{ij})$ of size $s \times n$ (6 species $\times$ 10 sites)

`Fin Whale`	M	N	E	**N**	L	F	A	P	F
`Blue Whale`	M	N	E	**N**	L	F	A	P	F
`Chimpanzee`	M	N	E	**N**	L	F	A	S	F
`Bonobo`	M	N	E	**N**	L	F	A	S	F
`Gorilla`	M	N	E	**N**	L	F	A	S	F
`Bornean Orangutan`	M	N	E	**D**	L	F	T	P	F

- $\mathcal{X}_{24} =$ **N**,
- 4th site: $\mathbf{X}_4 =$ (**NNNNND**)$'$,
- 2^{nd} species (Harbor Seal): $\mathbf{X}^{(2)} = MNENLFAPFM$.

Inference Methods

Three Families of Methods:

Distance based: (U/W)PGMA, Neighbor-Joining (NJ),

Parsimony based: (un)corrected parsimony (MP),

Likelihood based: Bayesian inference (BI), Maximum Likelihood (ML).

Inference Methods

Three Families of Methods:

Distance based: (U/W)PGMA, Neighbor-Joining (NJ),

Parsimony based: (un)corrected parsimony (MP),

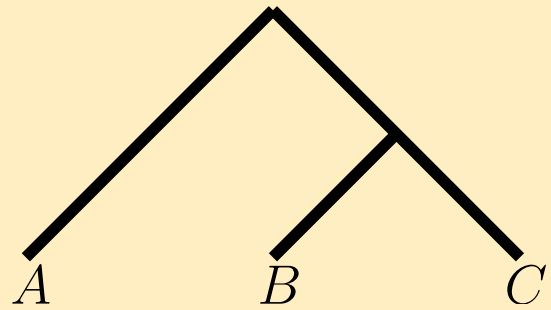
Likelihood based: Bayesian inference (BI), Maximum Likelihood (ML).

General Principle:

Alignment $\longrightarrow$ **Phylogenetic tree**

A	C	C	T	T
B	G	G	A	A
C	G	G	A	C

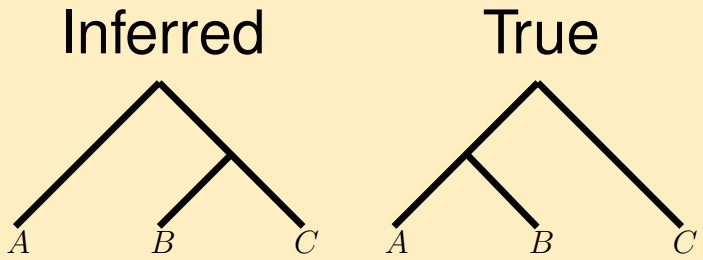
NJ, MP,
ML, BI,
...



End of The Story ?

Inference Problems:

- Compare inferred tree to true tree to assess how good it is,



- But the true tree is not available!

Confidence Issue:

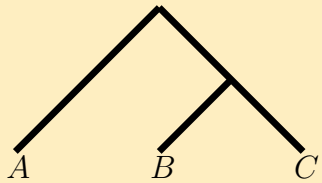
- How confident are we on the inferred tree ?
- Which parts of the tree are reliable/not reliable ?

End of The Story ?

Inference Problems:

- Compare inferred tree to true tree to assess how good it is,

Inferred



True

?

- But the true tree is not available!

Confidence Issue:

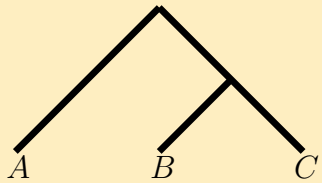
- How confident are we on the inferred tree ?
- Which parts of the tree are reliable/not reliable ?

End of The Story ?

Inference Problems:

- Compare inferred tree to true tree to assess how good it is,

Inferred



True

?

- But the true tree is not available!

Confidence Issue:

- How confident are we on the inferred tree ?
- Which parts of the tree are reliable/not reliable ?

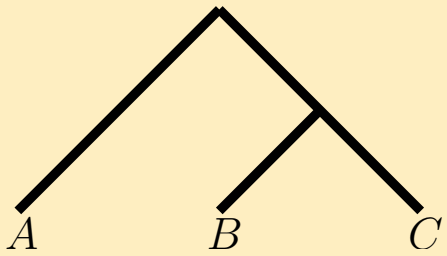
Bootstrap Values: the Theory

Original Dataset:

Alignment

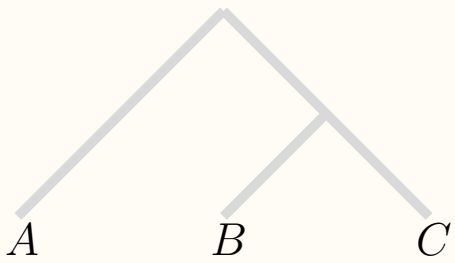
A	A	C	T	T
B	G	G	A	T
C	G	G	C	C

Phylogenetic tree

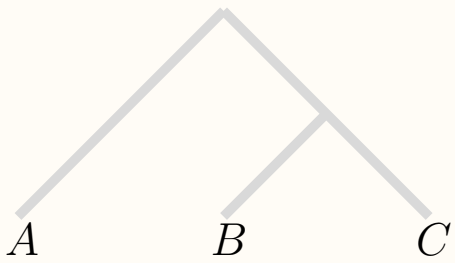


Bootstrap Datasets:

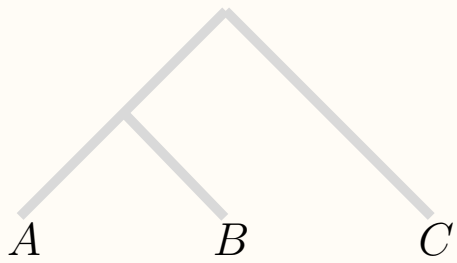
A	A	C	T	C
B	G	G	A	G
C	G	G	C	G



A	C	A	T	A
B	G	G	A	G
C	G	G	C	G



A	T	T	T	T
B	A	T	A	T
C	C	C	C	C



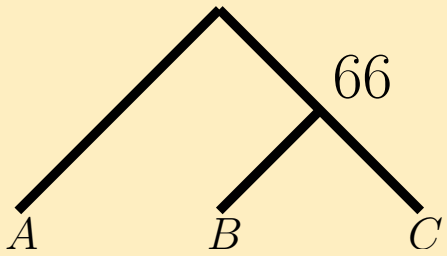
Bootstrap Values: the Theory

Original Dataset:

Alignment

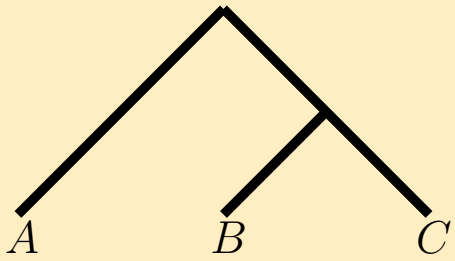
A	*A*	*C*	*T*	*T*
B	*G*	*G*	*A*	*T*
C	*G*	*G*	*C*	*C*

Phylogenetic tree

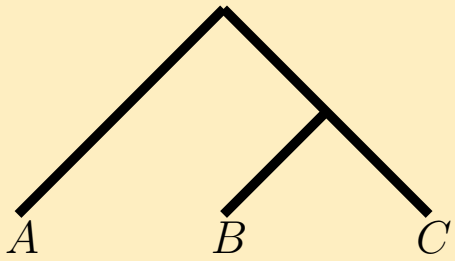


Bootstrap Datasets:

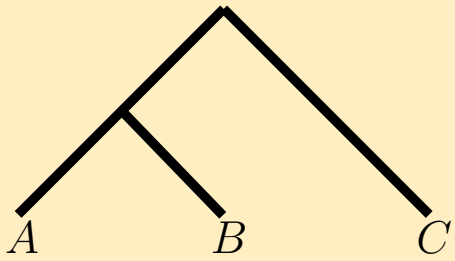
A	*A*	*C*	*T*	*C*
B	*G*	*G*	*A*	*G*
C	*G*	*G*	*C*	*G*



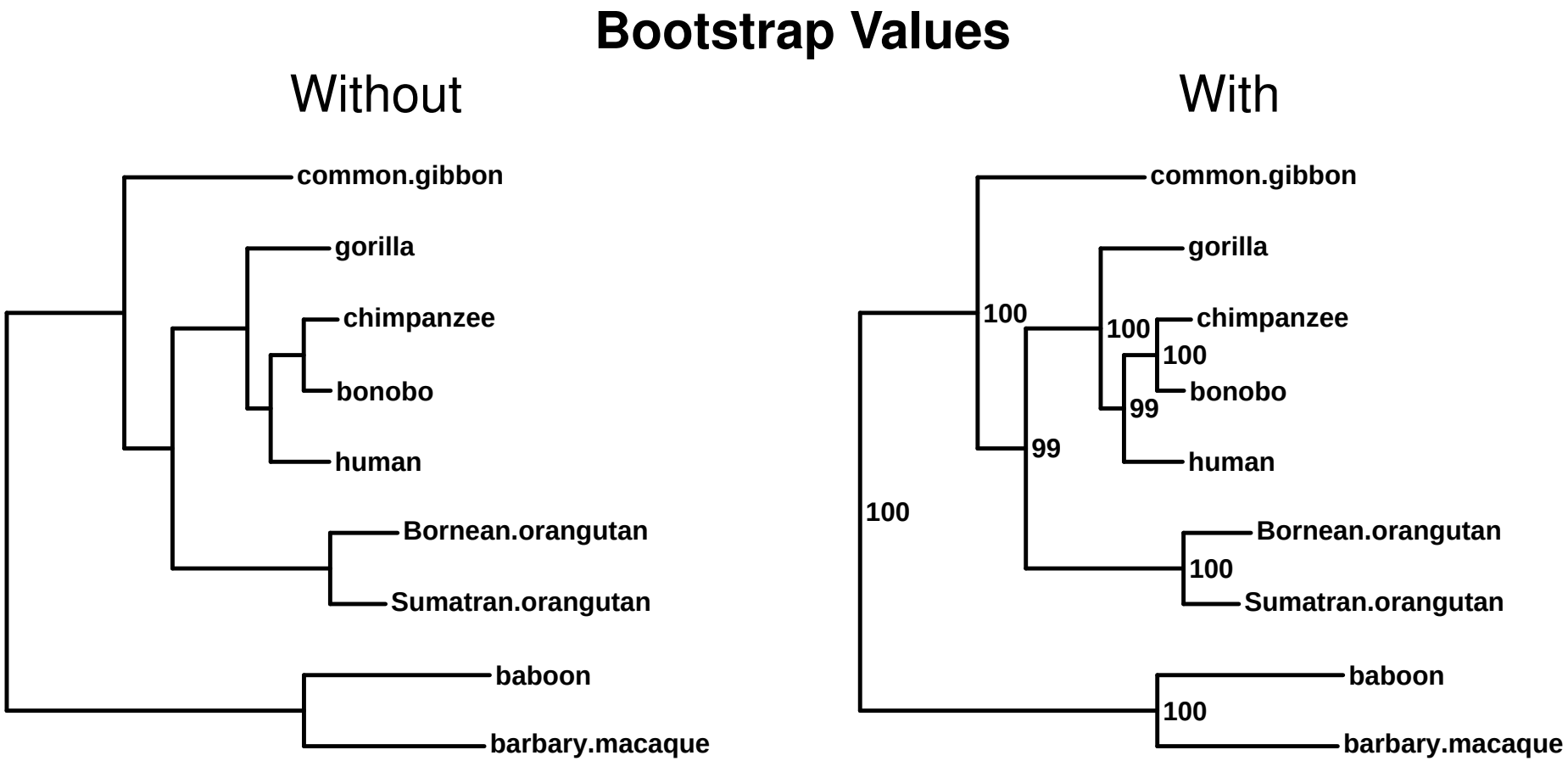
A	*C*	*A*	*T*	*A*
B	*G*	*G*	*A*	*G*
C	*G*	*G*	*C*	*G*



A	*T*	*T*	*T*	*T*
B	*A*	*T*	*A*	*T*
C	*C*	*C*	*C*	*C*



Bootstrap Values: An Example



Bootstrap Values: A Confidence Index ?

Bootstrap Strong Points:

- Many potential causes for uncertainty:
  - Finite sequence lengths,
  - Poor alignment quality (outlier sites),
  - Poor species sampling (rogue species),
  - Model misspecification,
  - . . .
- Global measure of uncertainty,

Bootstrap Weak Points:

- Global measure of uncertainty,
- Unable to breakdown the uncertainty,
- Unable to pinpoint local sources of uncertainties.

Bootstrap Values: A Confidence Index ?

Bootstrap Strong Points:

- Many potential causes for uncertainty:
  - Finite sequence lengths,
  - Poor alignment quality (outlier sites),
  - Poor species sampling (rogue species),
  - Model misspecification,
  - . . .
- Global measure of uncertainty,

Bootstrap Weak Points:

- Global measure of uncertainty,
- Unable to breakdown the uncertainty,
- Unable to pinpoint local sources of uncertainties.

Bootstrap Values: A Confidence Index ?

Bootstrap Strong Points:

- Many potential causes for uncertainty:
  - Finite sequence lengths,
  - Poor alignment quality (outlier sites),
  - Poor species sampling (rogue species),
  - Model misspecification,
  - ...
- Global measure of uncertainty,

Bootstrap Weak Points:

- Global measure of uncertainty,
- Unable to breakdown the uncertainty,
- Unable to pinpoint local sources of uncertainties.

Species Leverage Index: Motivation and Goal

Species Leverage Index (SLI)

- **Goal:** Study the stability of the tree with respect to the species,
- **Motivation:** Thanks to strange evolutionary features not taken into account by the inference method, some species may exert a strong pull toward a biased estimated phylogeny,
- **Method:**
  - Infer the phylogeny T with the whole species set,
  - Remove species one at the time and infer a new tree T_i on the smaller species set,
  - Quantify difference between T and T_i.

Species Leverage Index: Motivation and Goal

Species Leverage Index (SLI)

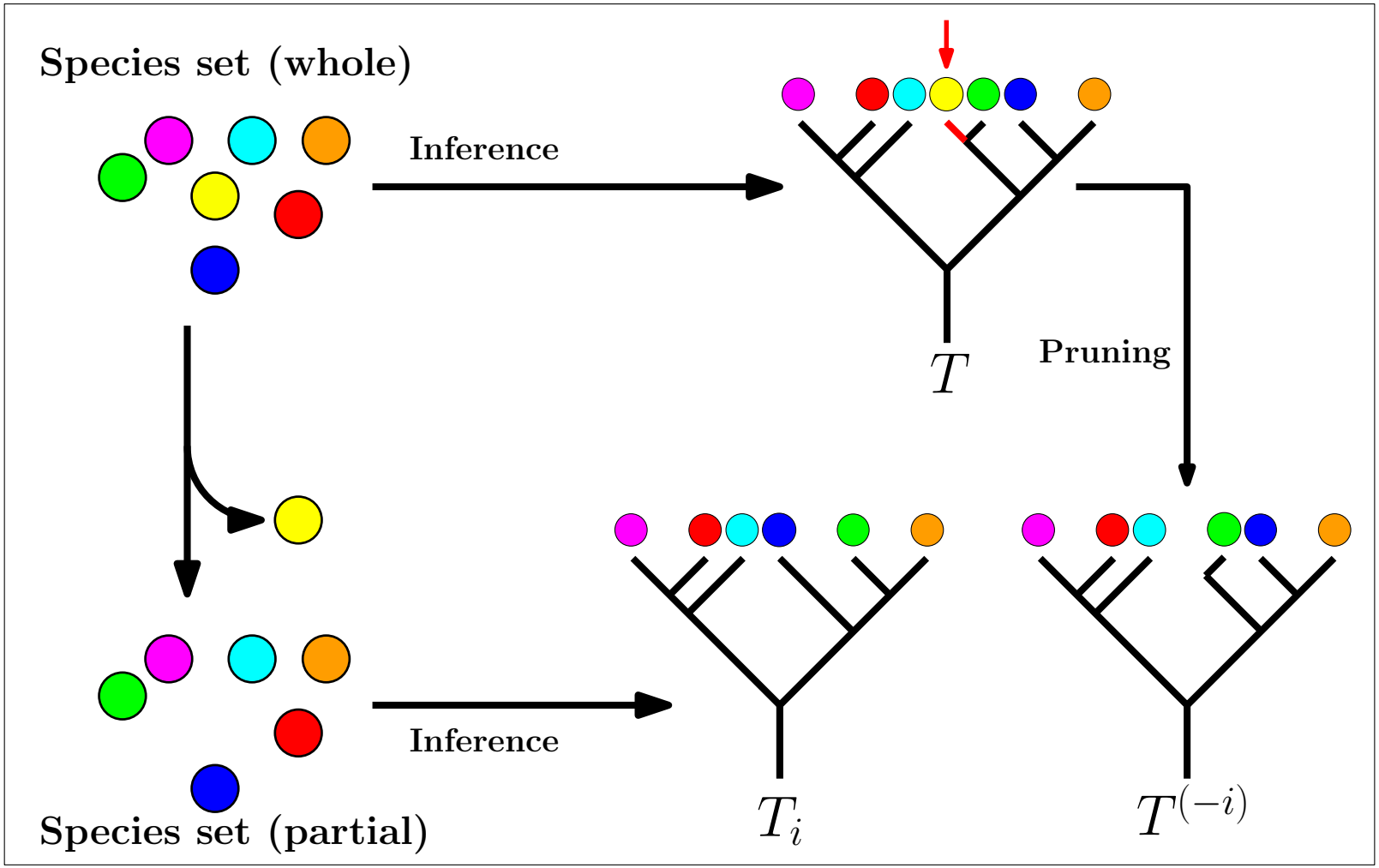
- **Goal:** Study the stability of the tree with respect to the species,
- **Motivation:** Thanks to strange evolutionary features not taken into account by the inference method, some species may exert a strong pull toward a biased estimated phylogeny,
- **Method:**
  - Infer the phylogeny T with the whole species set,
  - Remove species one at the time and infer a new tree T_i on the smaller species set,
  - Quantify difference between T and T_i.

Species Leverage Index: Motivation and Goal

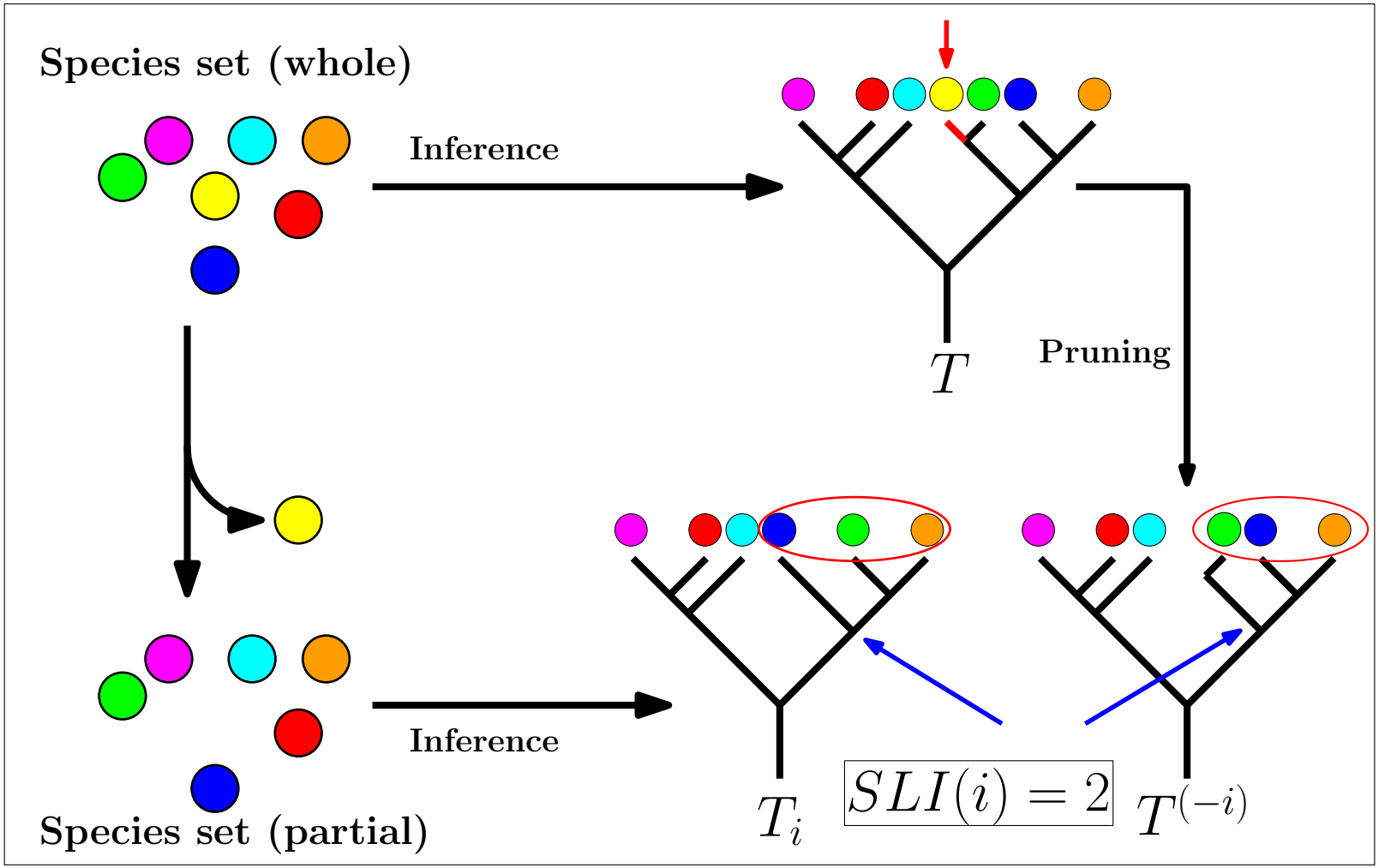
Species Leverage Index (SLI)

- **Goal:** Study the stability of the tree with respect to the species,
- **Motivation:** Thanks to strange evolutionary features not taken into account by the inference method, some species may exert a strong pull toward a biased estimated phylogeny,
- **Method:**
  - Infer the phylogeny T with the whole species set,
  - Remove species one at the time and infer a new tree T_i on the smaller species set,
  - Quantify difference between T and T_i.

Method


Species set (whole)
Inference
T
Pruning
Inference
Species set (partial)
T_i
$T^{(-i)}$

Method


Species set (whole)
Inference
T
Pruning
Inference
Species set (partial)
T_i
$SLI(i) = 2$
$T^{(-i)}$

Species Leverage Index Use

Interpretation:

SLI:

- Low value: adding/removing the species from the dataset has (almost) impact on the tree,
- High value: “rogue” species, adding/removing it greatly affects the tree.

Strategy towards greater stability

- Focus on rogues species: species with high SLI,
- Rank them in increasing SLI,
- Remove them one at the time until a stable tree is found.

Species Leverage Index Use

Interpretation:

SLI:

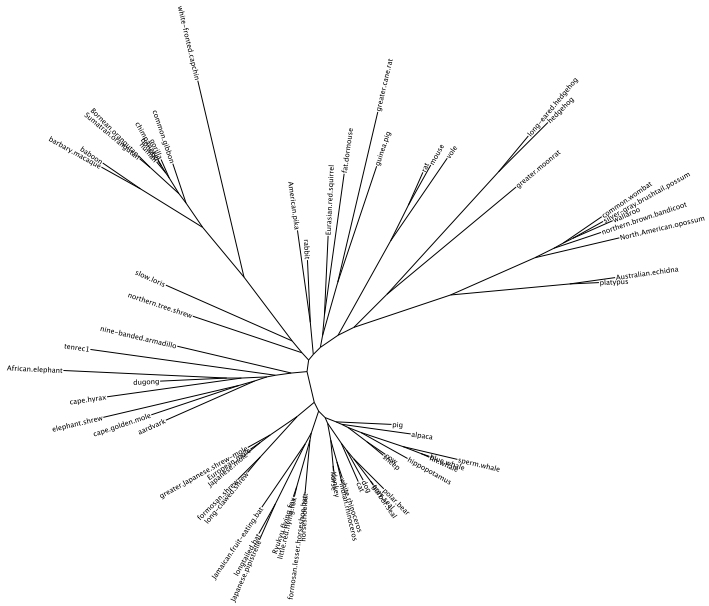
- Low value: adding/removing the species from the dataset has (almost) impact on the tree,
- High value: “rogue” species, adding/removing it greatly affects the tree.

Strategy towards greater stability

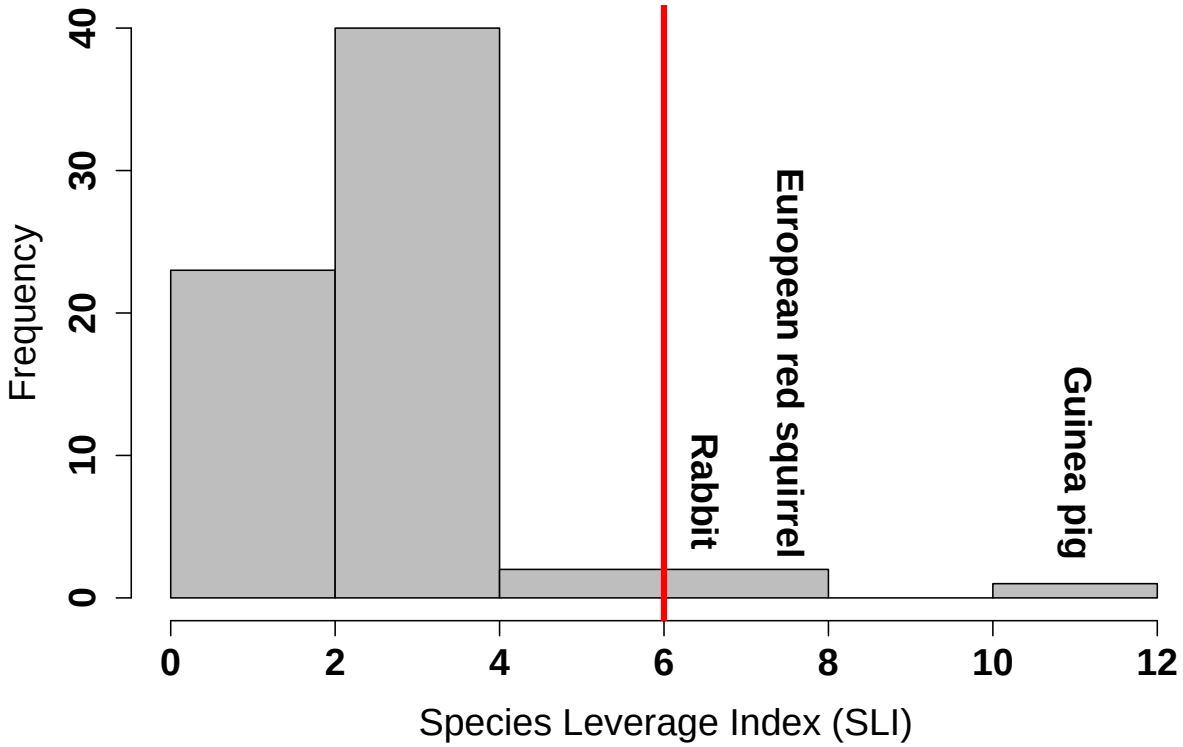
- Focus on rogues species: species with high SLI,
- Rank them in increasing SLI,
- Remove them one at the time until a stable tree is found.

Data: Placental Mammal Phylogeny

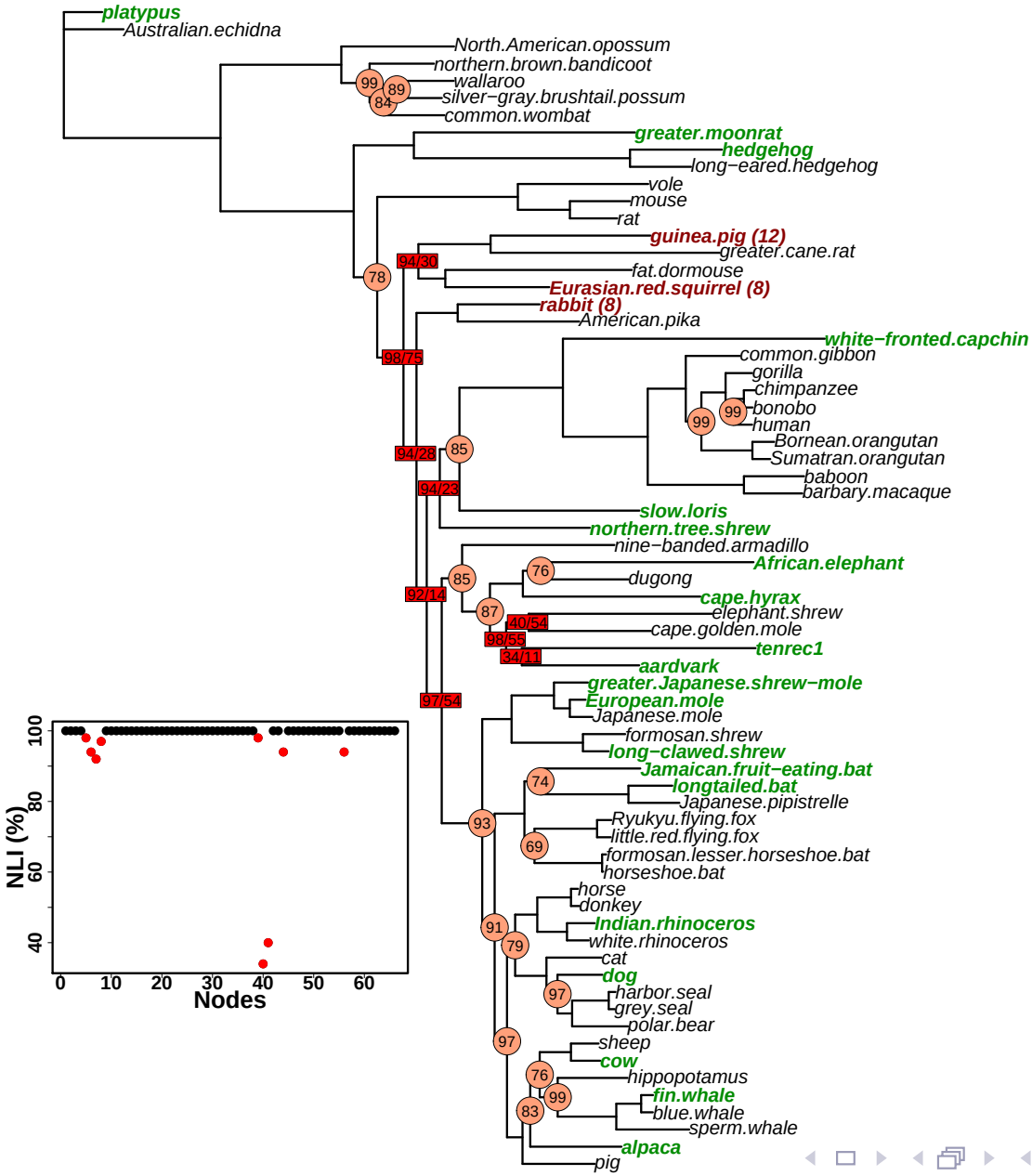
- Mitochondrial genome of 68 mammals,
- Amino Acids sequences,
- Sequences are 3658 sites long,
- Phylogeny published in Nikaido *et al.* in 2003.



Species Leverage Index



Complete Phylogeny



Rogue Species