Influence of Sites and Species on Phylogenetic Stability

M. Mariadassou A. Bar-Hen H. Kishino

Laboratoire MAP5
Université Paris Descartes

Feb. 2008
Midipile

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

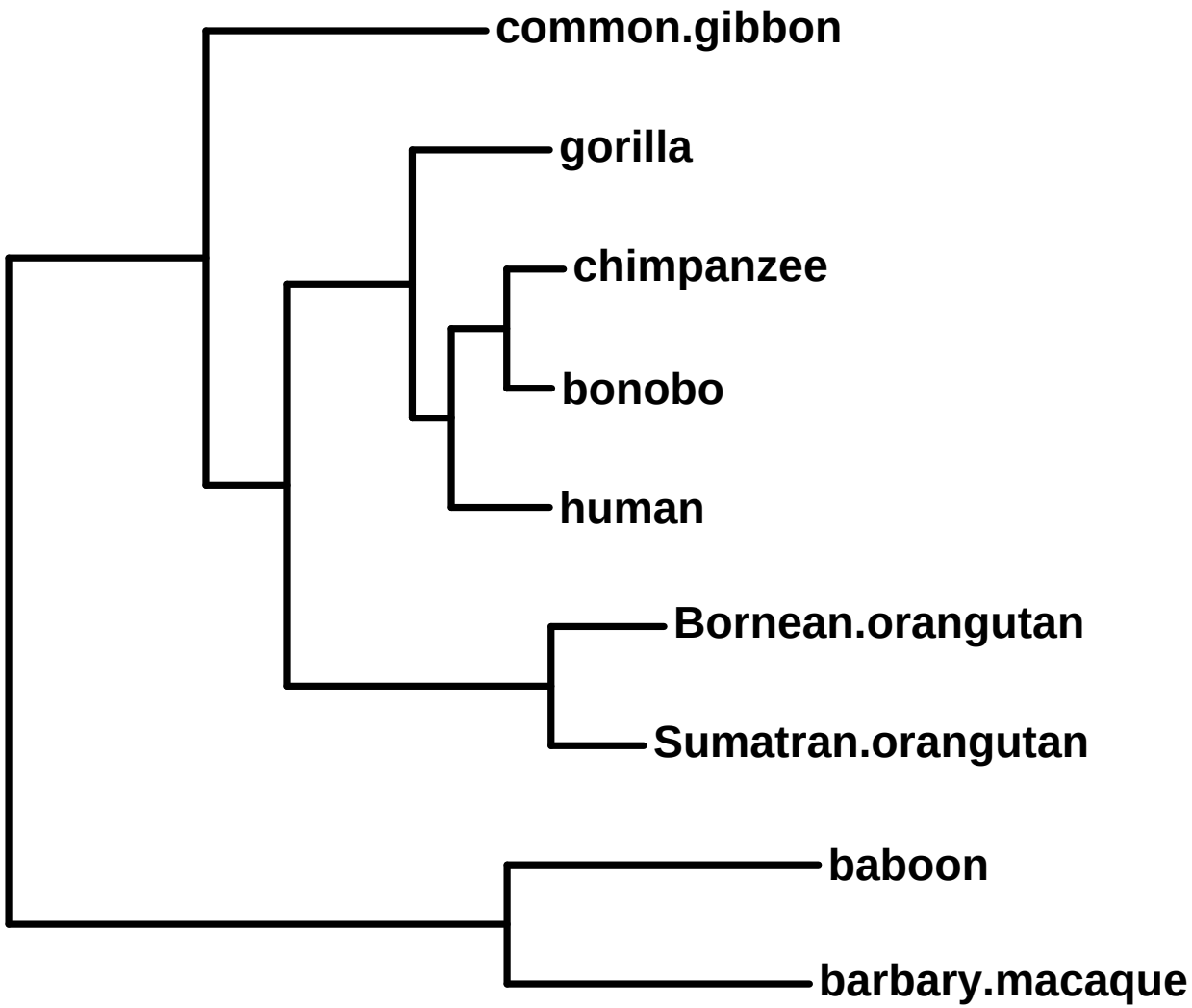
Basic Ideas:

- 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.

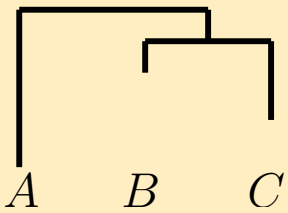
Example of A Phylogenetic Tree



Reconstructions and Limits

Two levels of reconstruction:

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



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

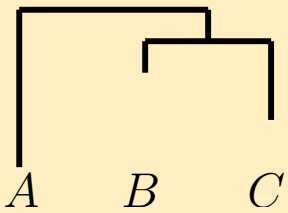
Issues:

- 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).

Issues:

- 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: $\mathbf{X}_4 = (\mathbf{NNNNND})'$,
- 2nd species (Harbor Seal): $\mathbf{X}^{(2)} = MNENLFAPFM$.

4 families of inference methods

All methods infer the tree which minimize/maximize a given criteria:

- Maximum Parsimony: minimizes the number of changes needed to explain the current data;
- Neighbor-Joining: minimizes a natural estimate of the tree length;
- Maximum Likelihood: maximizes the likelihood of the data;
- Bayesian: maximizes the posterior probability of the data.

Inference Method

Likelihood Based

- Assume $(\mathbf{X}_i)$ *i.i.d.*;
- Choose generating evolution model $M(T, \theta_T)$;
- Discrete topology T and continuous parameter model;
- Retrieve $(\hat{T}, \hat{\theta}_{\hat{T}})$ maximizing $\mathbb{P}((\mathbf{X}_i); M, T, \theta_T)$.

Inference Method

Likelihood Based

- Assume $(\mathbf{X}_i)$ *i.i.d.*;
- Choose generating evolution model $M(T, \theta_T)$;
- Discrete topology T and continuous parameter model;
- Retrieve $(\hat{T}, \hat{\theta}_{\hat{T}})$ maximizing $\mathbb{P}((\mathbf{X}_i); M, T, \theta_T)$.

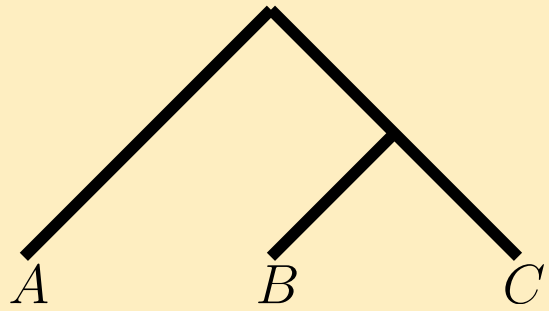
General Principle:

Alignment

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

$\longrightarrow$ ML

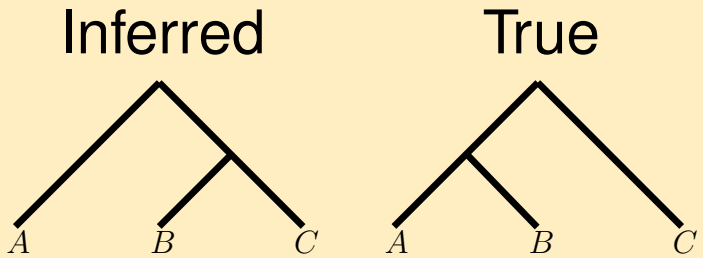
Phylogenetic tree



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:

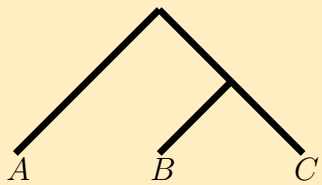
- How confident are we on the inferred tree ?
- Which parts of the tree are reliable/not reliable ?

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!

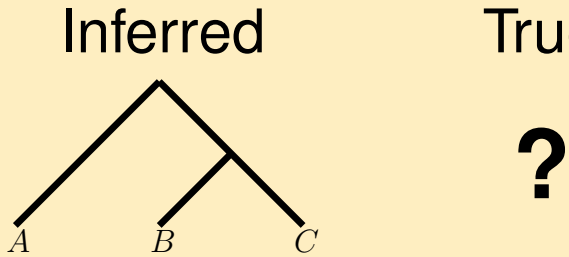
Confidence Issue:

- 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,



- 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 ?

Confidence or Robustness ?

Confidence or Robustness ?

Confidence: Another (independent) data set gives (roughly) the same inferred tree;

Robustness: Tweaking the (original) data set gives (roughly) the same inferred tree.

Robustness

- Most (if not all) available procedures are designed to check robustness, not confidence;
- The inferred tree might be far from the true tree, as long it is consistently so, we are happy.

Confidence or Robustness ?

Confidence or Robustness ?

Confidence: Another (independent) data set gives (roughly) the same inferred tree;

Robustness: Tweaking the (original) data set gives (roughly) the same inferred tree.

Robustness

- Most (if not all) available procedures are designed to check robustness, not confidence;
- The inferred tree might be far from the true tree, as long it is consistently so, we are happy.

Bootstrap Values: A Robustness 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,
- Several other ways to tweak the data.

Bootstrap Values: A Robustness 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,
- Several other ways to tweak the data.

Bootstrap Values: A Robustness 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,
- Several other ways to tweak the data.

Outlier Sites: Motivation and Goal

Motivation: Filter Data

Sites source of errors:

- Sequencing errors;
- Alignment errors;
- Presence of an atypical DNA segment;
- . . .

Goal

- Quantify the influence of each site on the tree;
- Detect outlier sites;
- Infer a robust tree.

Outlier Sites: Motivation and Goal

Motivation: Filter Data

Sites source of errors:

- Sequencing errors;
- Alignment errors;
- Presence of an atypical DNA segment;
- . . .

Goal

- Quantify the influence of each site on the tree;
- Detect outlier sites;
- Infer a robust tree.

About the Influence Function

Influence Function: Definition

Let $X_1, \ldots, X_n$ be *i.i.d.* with common d.f. F on $\mathcal{R}^d$ and $S(F)$ a functional of F. The influence function:

$$IF_{S,F}(x) = \lim_{\varepsilon \to 0} \frac{S[(1-\varepsilon)F + \varepsilon\delta_x] - S[F]}{\varepsilon}$$

measure the influence of a perturbation in direction x.

Empirical Version

For unknown S and finite size sample, $F \to F_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}$, $\varepsilon \to -1/(n-1)$:

$$\begin{aligned} IF_{S,F_n}(X_i) &= \lim_{\varepsilon \to 0} \frac{S[(1-\varepsilon)F_n + \varepsilon\delta_{X_i}] - S[F_n]}{\varepsilon} \\ &= (n-1)\big(S(F_n) - S(F_{n,-i})\big) \end{aligned}$$

where $F_{n,-i}$ is the empirical distribution on all sites but i.

About the Influence Function

Influence Function: Definition

Let $X_1, \ldots, X_n$ be *i.i.d.* with common d.f. F on $\mathcal{R}^d$ and $S(F)$ a functional of F. The influence function:

$$IF_{S,F}(x) = \lim_{\varepsilon \to 0} \frac{S[(1-\varepsilon)F + \varepsilon\delta_x] - S[F]}{\varepsilon}$$

measure the influence of a perturbation in direction x.

Empirical Version

For unknown S and finite size sample, $F \to F_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}$, $\varepsilon \to -1/(n-1)$:

$$\begin{aligned} IF_{S,F_n}(X_i) &= \lim_{\varepsilon \to 0} \frac{S[(1-\varepsilon)F_n + \varepsilon\delta_{X_i}] - S[F_n]}{\varepsilon} \\ &= (n-1)\big(S(F_n) - S(F_{n,-i})\big) \end{aligned}$$

where $F_{n,-i}$ is the empirical distribution on all sites but i.

And for Phylogenies...

Definition

Let:

- $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be the complete alignment,
- $\mathbf{X}_{-i} = \mathbf{X} \backslash \mathbf{X}_i$ all the sites but site i,
- $(\hat{T}, \hat{\theta}_{\hat{T}})$ the ML tree and associated parameters for $\mathbf{X}$,
- $(\widehat{T_{-i}}, \hat{\theta}_{\widehat{T_{-i}}})$ the ML tree and associated parameters for $\mathbf{X}_{-i}$,
- The statistic be:

$$l_{\hat{T}}(\hat{\theta}_{\hat{T}}|\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \log \mathbb{P}(\mathbf{X}_i|\hat{T}, \hat{\theta}_{\hat{T}})$$

The influence value of $\mathbf{X}_i$ is then:

$$IF_{S,F_n}(\mathbf{X}_i) = (n-1)\left(l_{\hat{T}}(\hat{\theta}_{\hat{T}}|\mathbf{X}) - l_{\widehat{T_{-i}}}(\hat{\theta}_{\widehat{T_{-i}}}|\mathbf{X}_{-i})\right)$$

And for Phylogenies...

Definition

Let:

- $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be the complete alignment,
- $\mathbf{X}_{-i} = \mathbf{X} \backslash \mathbf{X}_i$ all the sites but site i,
- $(\hat{T}, \hat{\theta}_{\hat{T}})$ the ML tree and associated parameters for $\mathbf{X}$,
- $(\widehat{T_{-i}}, \hat{\theta}_{\widehat{T_{-i}}})$ the ML tree and associated parameters for $\mathbf{X}_{-i}$,
- The statistic be:

$$l_{\hat{T}}(\hat{\theta}_{\hat{T}}|\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \log \mathbb{P}(\mathbf{X}_i|\hat{T}, \hat{\theta}_{\hat{T}})$$

The influence value of $\mathbf{X}_i$ is then:

$$IF_{S,F_n}(\mathbf{X}_i) = (n-1)\left(l_{\hat{T}}(\hat{\theta}_{\hat{T}}|\mathbf{X}) - l_{\widehat{T_{-i}}}(\hat{\theta}_{\widehat{T_{-i}}}|\mathbf{X}_{-i})\right)$$

Influence Values

Interpretation

- Positive value: enhanced support for the ML tree;
- Negative value: weakened support for the ML tree;
- Absolute value: strength of the support/disagreement;
- Many sites with small positive values and a few sites with large negative values.

Strategy towards greater stability

- Focus on outliers: sites with $IF(\mathbf{X}_i) < 0$;
- Rank them in increasing $IF(\mathbf{X}_i)$;
- Remove them one at the time until a stable tree is found.

Influence Values

Interpretation

- Positive value: enhanced support for the ML tree;
- Negative value: weakened support for the ML tree;
- Absolute value: strength of the support/disagreement;
- Many sites with small positive values and a few sites with large negative values.

Strategy towards greater stability

- Focus on outliers: sites with $IF(\mathbf{X}_i) < 0$;
- Rank them in increasing $IF(\mathbf{X}_i)$;
- Remove them one at the time until a stable tree is found.

Influence Values

Interpretation

- Positive value: enhanced support for the ML tree;
- Negative value: weakened support for the ML tree;
- Absolute value: strength of the support/disagreement;
- Many sites with small positive values and a few sites with large negative values.

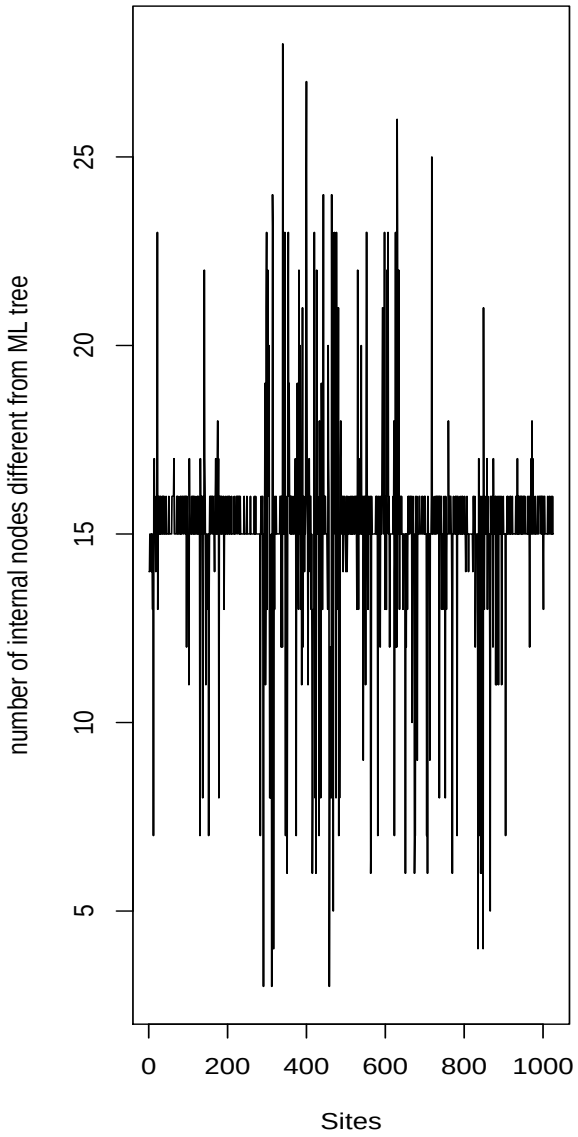
Strategy towards greater stability

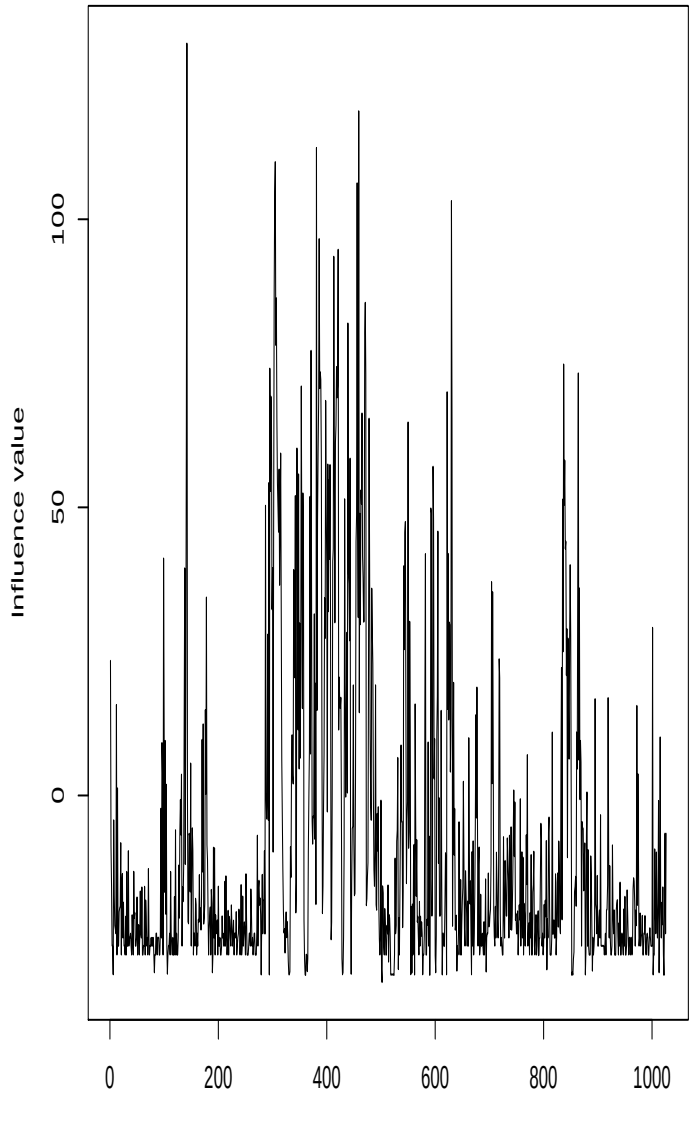
- Focus on outliers: sites with $IF(\mathbf{X}_i) < 0$;
- Rank them in increasing $IF(\mathbf{X}_i)$;
- Remove them one at the time until a stable tree is found.

Data: Zygomycetes & Chytridiomycetes

- "Lower mushrooms"
- Biology: widely unknown!
- Strong enough phylogenetic signal to correctly resolve the topology.
- 1026 sites, 158 OTUs, GTR model

Information about sites





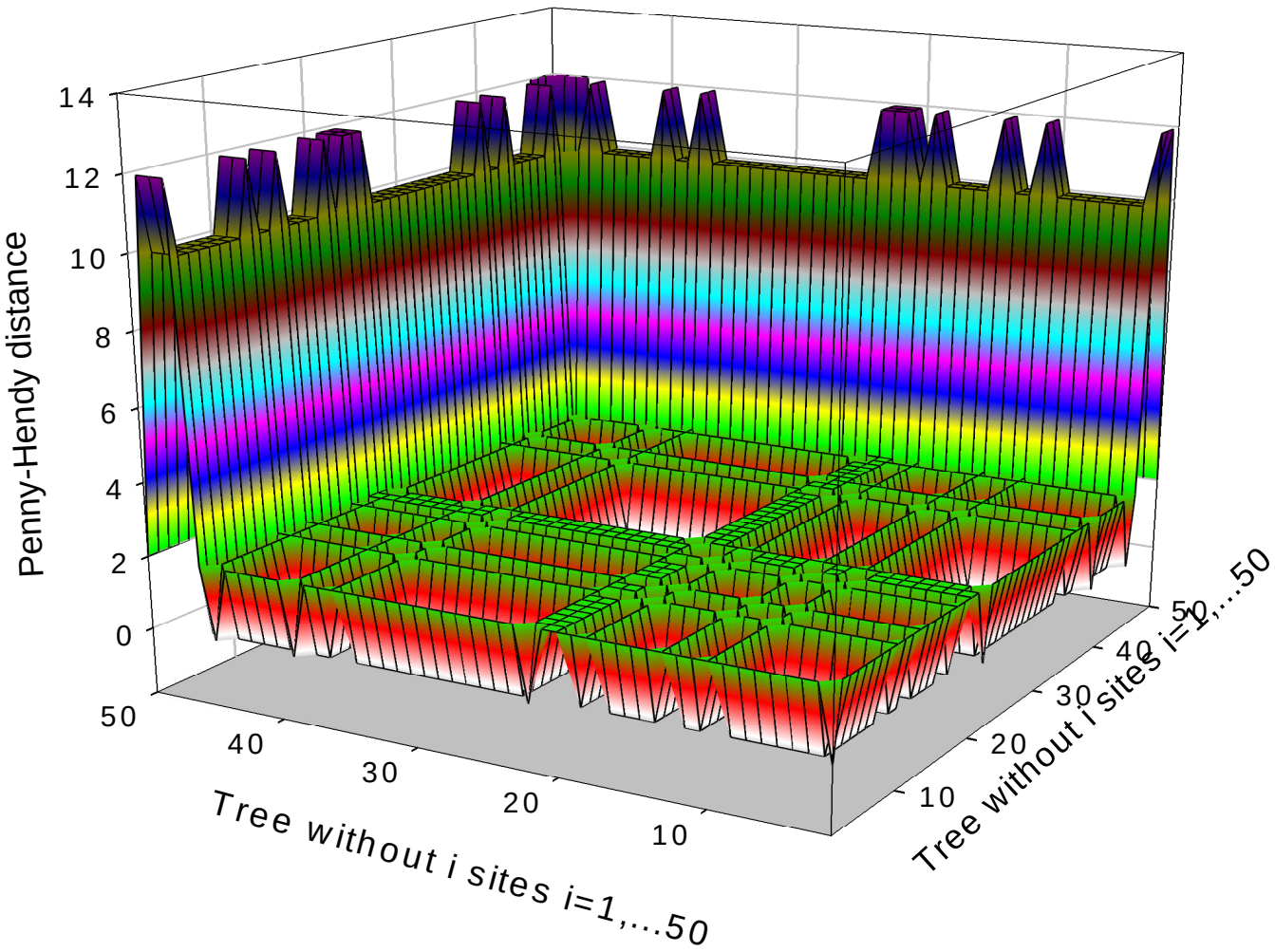
Distance between trees

0	20	18	18	18	18	18	18	18	20
20	0	2	2	2	2	2	2	2	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
20	2	2	2	2	2	2	2	2	0

T_i: trees constructed without the i most influent sites.

D_{ij}: Robinson-Foulds distance between T_i and T_j

Distance Between Trees



Bootstrap Values: A Robustness Index ?

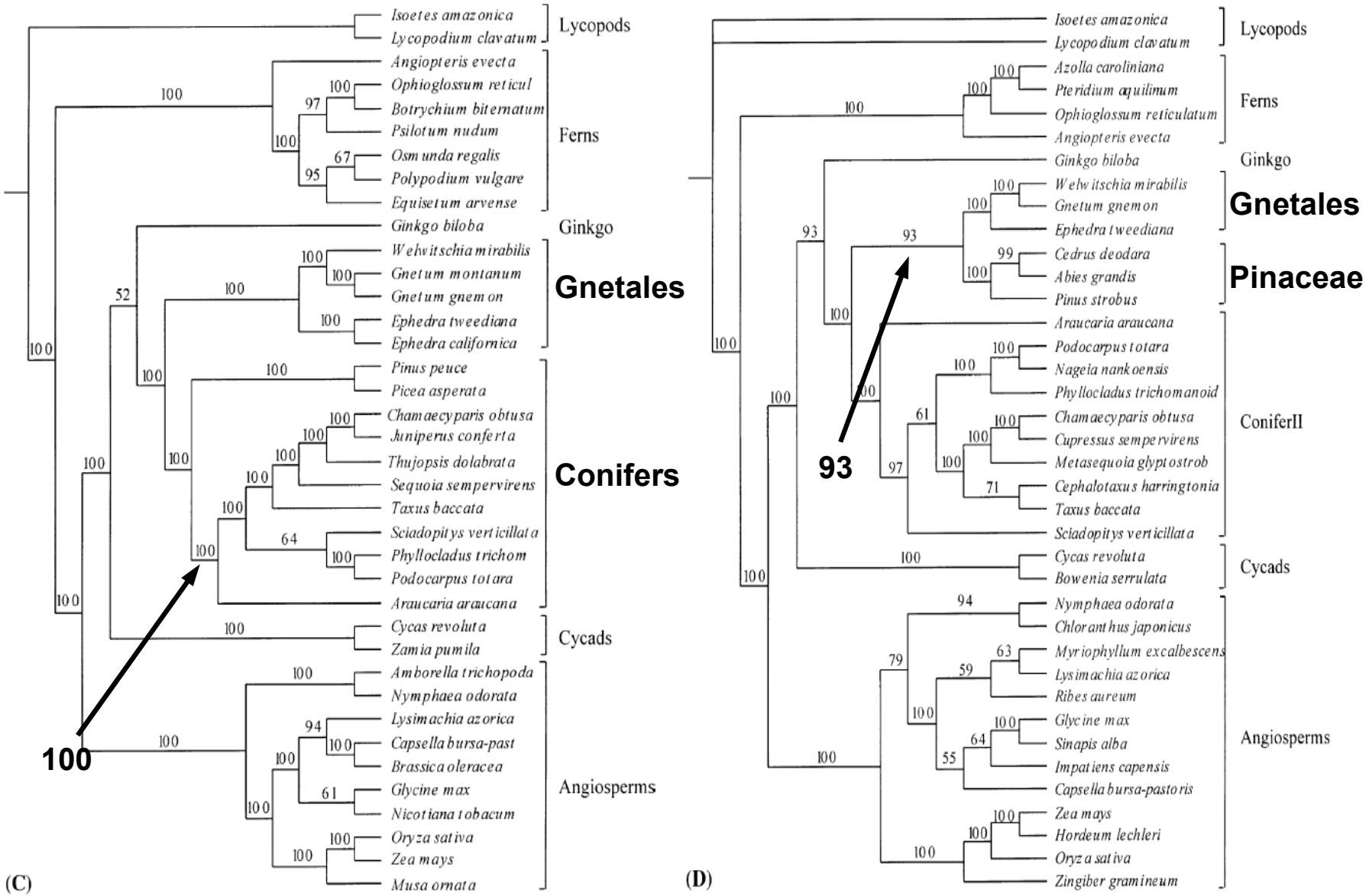
Bootstrap Strong Points:

- 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,
- Several other ways to tweak the data.

Seed Plant Phylogeny (Ridyn & al. 2002)



Species Leverage Index: Motivation and Goal

Species Leverage Index (SLI)

- **Goal:** Study the robustness 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 robustness 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 robustness 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 (SLI)

Definition

Let:

- $\mathbf{X} = \left(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(s)}\right)'$ be the complete alignment,
- $\mathbf{X}^{(-i)} = \mathbf{X} \backslash \mathbf{X}^{(i)}$ all the species but species i,
- $\widehat{T}$ the ML tree and associated parameters for $\mathbf{X}$,
- $\widehat{T}^{(-i)}$ the tree $\widehat{T}$ after pruning species i,
- $\widehat{T^{(-i)}}$ the ML tree and associated

The Species Leverage Index (SLI) of species i is:

$$SLI(i) = d(\widehat{T}^{(-i)}, \widehat{T^{(-i)}})$$

where d is any adapted distance .

Species Leverage Index (SLI)

Definition

Let:

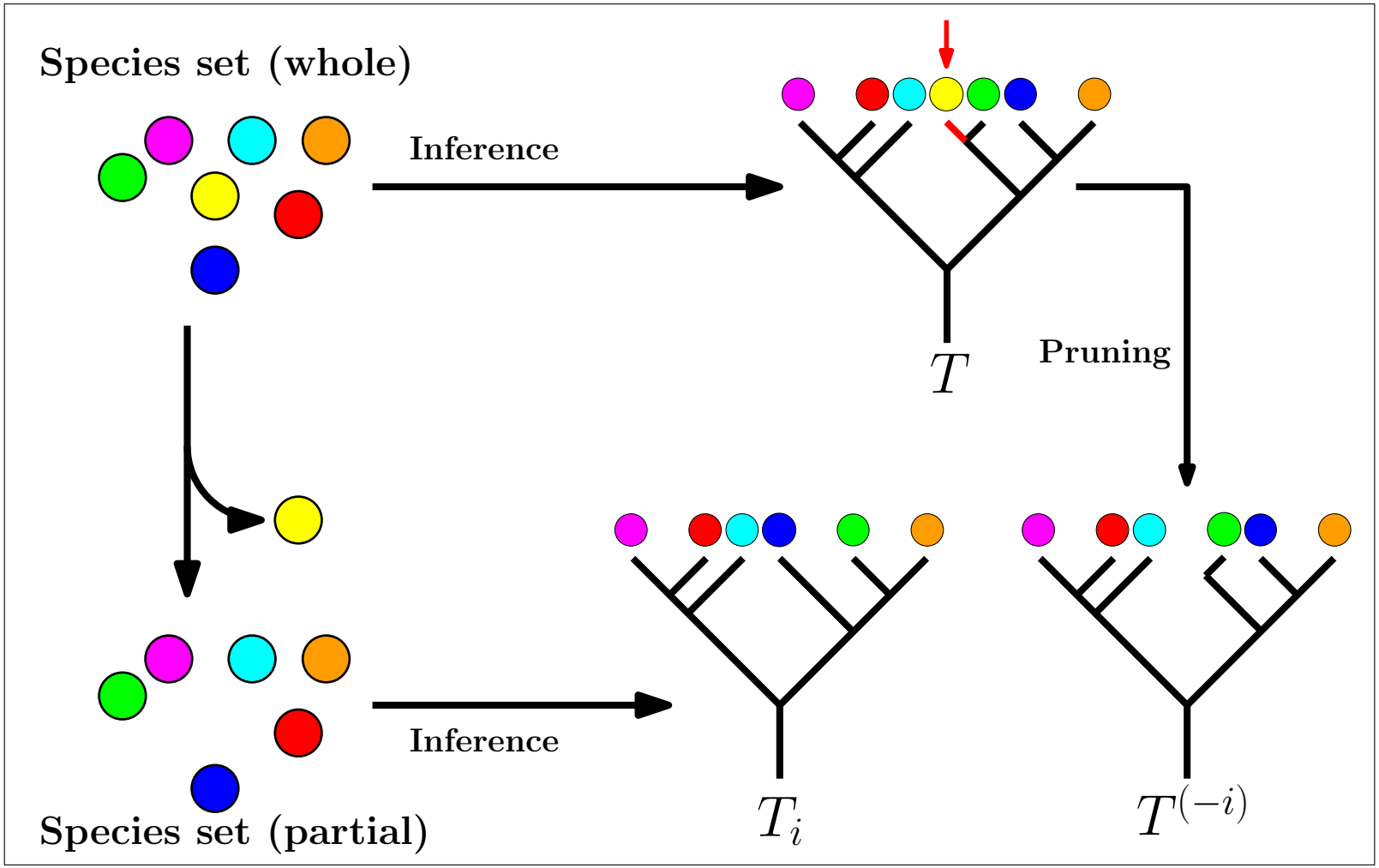
- $\mathbf{X} = \left(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(s)}\right)'$ be the complete alignment,
- $\mathbf{X}^{(-i)} = \mathbf{X} \backslash \mathbf{X}^{(i)}$ all the species but species i,
- $\widehat{T}$ the ML tree and associated parameters for $\mathbf{X}$,
- $\widehat{T}^{(-i)}$ the tree $\widehat{T}$ after pruning species i,
- $\widehat{T^{(-i)}}$ the ML tree and associated

The Species Leverage Index (SLI) of species i is:

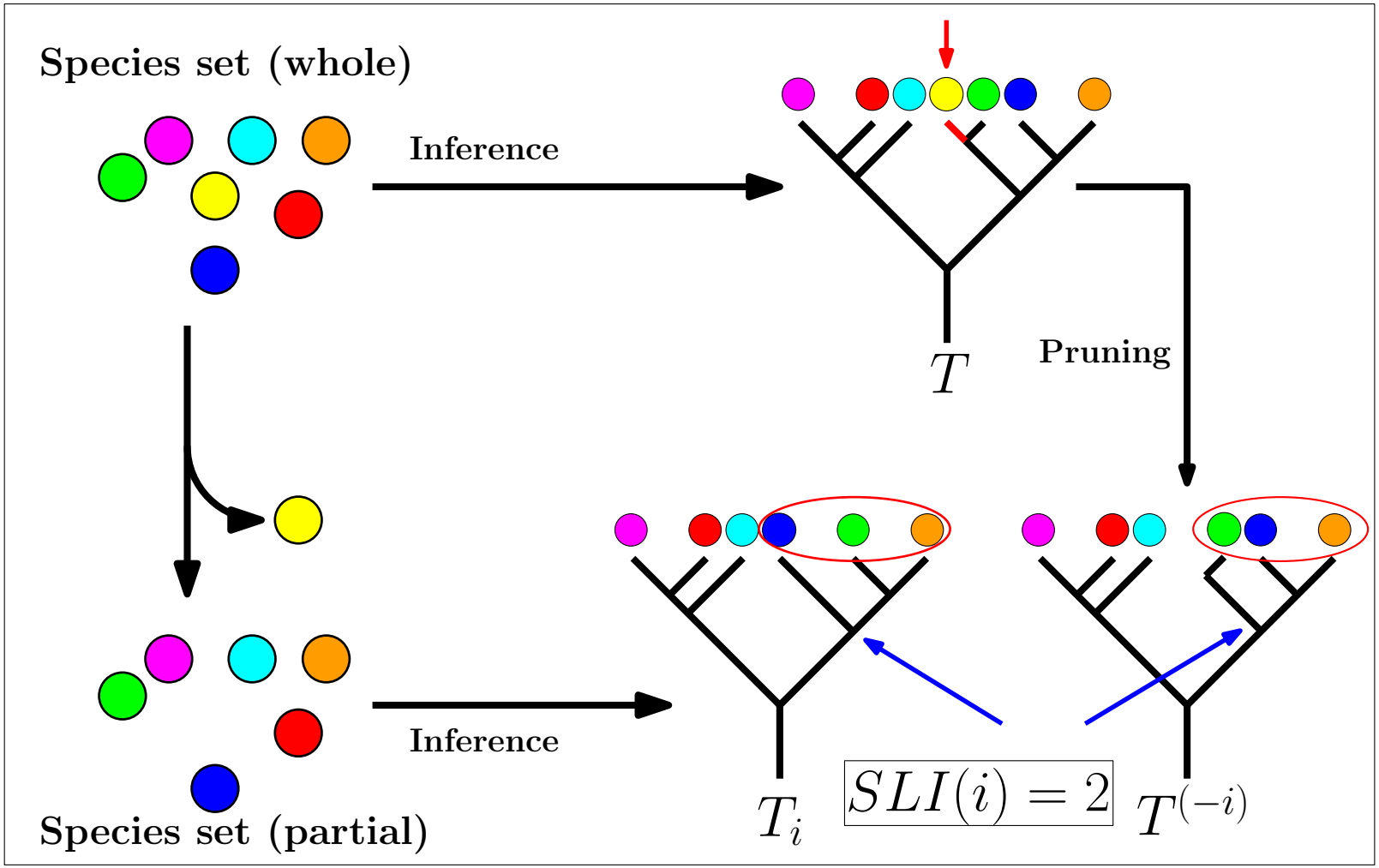
$$SLI(i) = d(\widehat{T}^{(-i)}, \widehat{T^{(-i)}})$$

where d is any adapted distance .

Method


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

Method


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

Nodes Leverage Index (NLI)

Definition

Node Leverage Index (NLI) of A: number of inferred trees in which the node is retrieved.

$$NLI(A) = \sum_{i=1}^{n} \mathbb{1}_{\widehat{T^{(-i)}}}(A)$$

Problems

- The taxa sets are different between $\widehat{T}$, $\widehat{T}^{(-i)}$ and $\widehat{T^{(-i)}}$;
- The taxa sets are different between the $\widehat{T^{(-i)}}$s;
- Some nodes *naturally* disappear when a taxon is removed;
- Find a convenient node mapping from $\widehat{T}$ to $\widehat{T}^{(-i)}$ before comparing $\widehat{T}^{(-i)}$ and $\widehat{T^{(-i)}}$.

Nodes Leverage Index (NLI)

Definition

Node Leverage Index (NLI) of A: number of inferred trees in which the node is retrieved.

$$NLI(A) = \sum_{i=1}^{n} \mathbb{1}_{\widehat{T^{(-i)}}}(A)$$

Problems

- The taxa sets are different between $\widehat{T}$, $\widehat{T}^{(-i)}$ and $\widehat{T^{(-i)}}$;
- The taxa sets are different between the $\widehat{T^{(-i)}}$s;
- Some nodes *naturally* disappear when a taxon is removed;
- Find a convenient node mapping from $\widehat{T}$ to $\widehat{T}^{(-i)}$ before comparing $\widehat{T}^{(-i)}$ and $\widehat{T^{(-i)}}$.

NLIs and SLIs

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.

NLI:
- High value: stable nodes, highly resilient to taxon sampling;
- Low value: weak nodes, highly sensitive to taxon sampling.

Strategy towards robustness

- 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.

NLIs and SLIs

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.

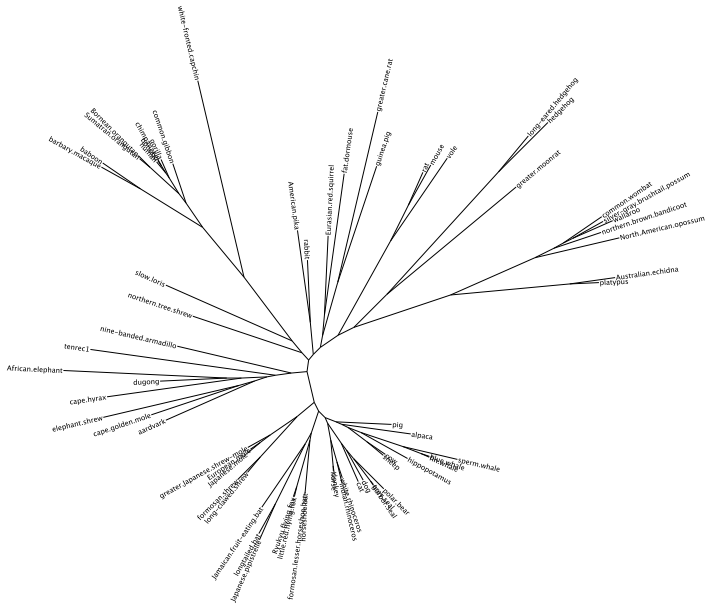
NLI:
- High value: stable nodes, highly resilient to taxon sampling;
- Low value: weak nodes, highly sensitive to taxon sampling.

Strategy towards robustness

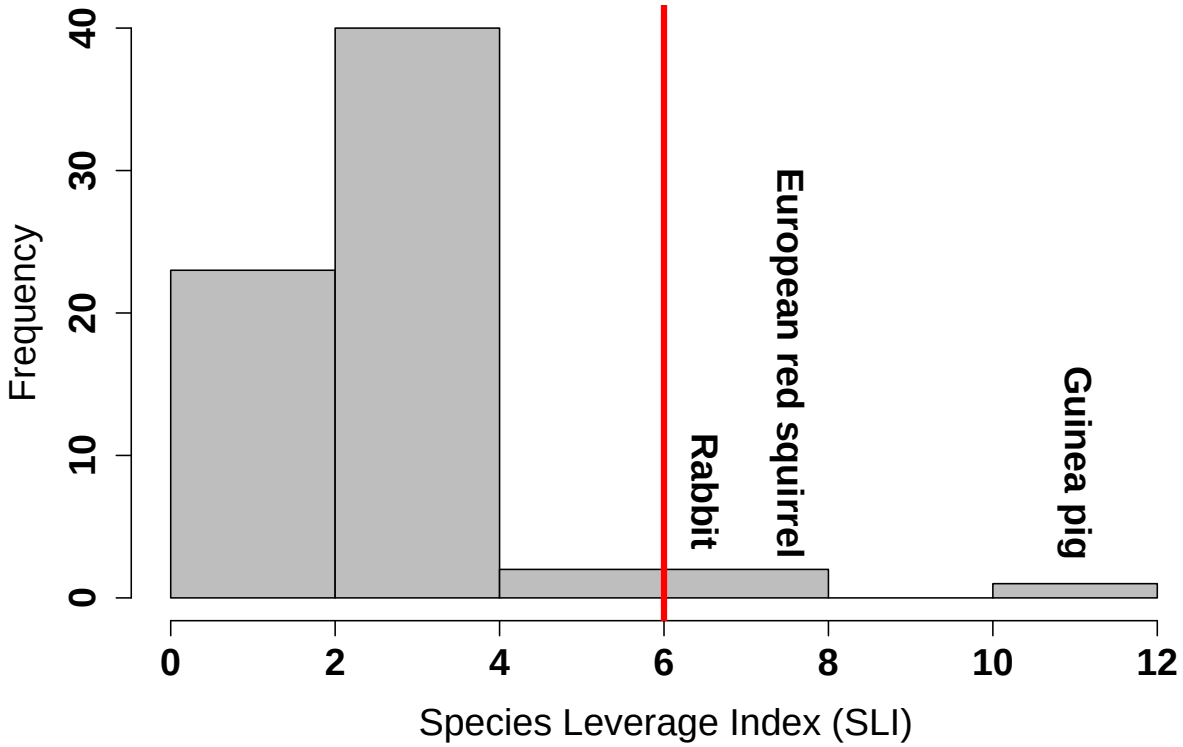
- Focus on rogues species: species with high SLI;
- Rank them in increasing SLI;
- Remove them one at the time until a stable tree is found.

Data: Placental Mammal Phylogeny

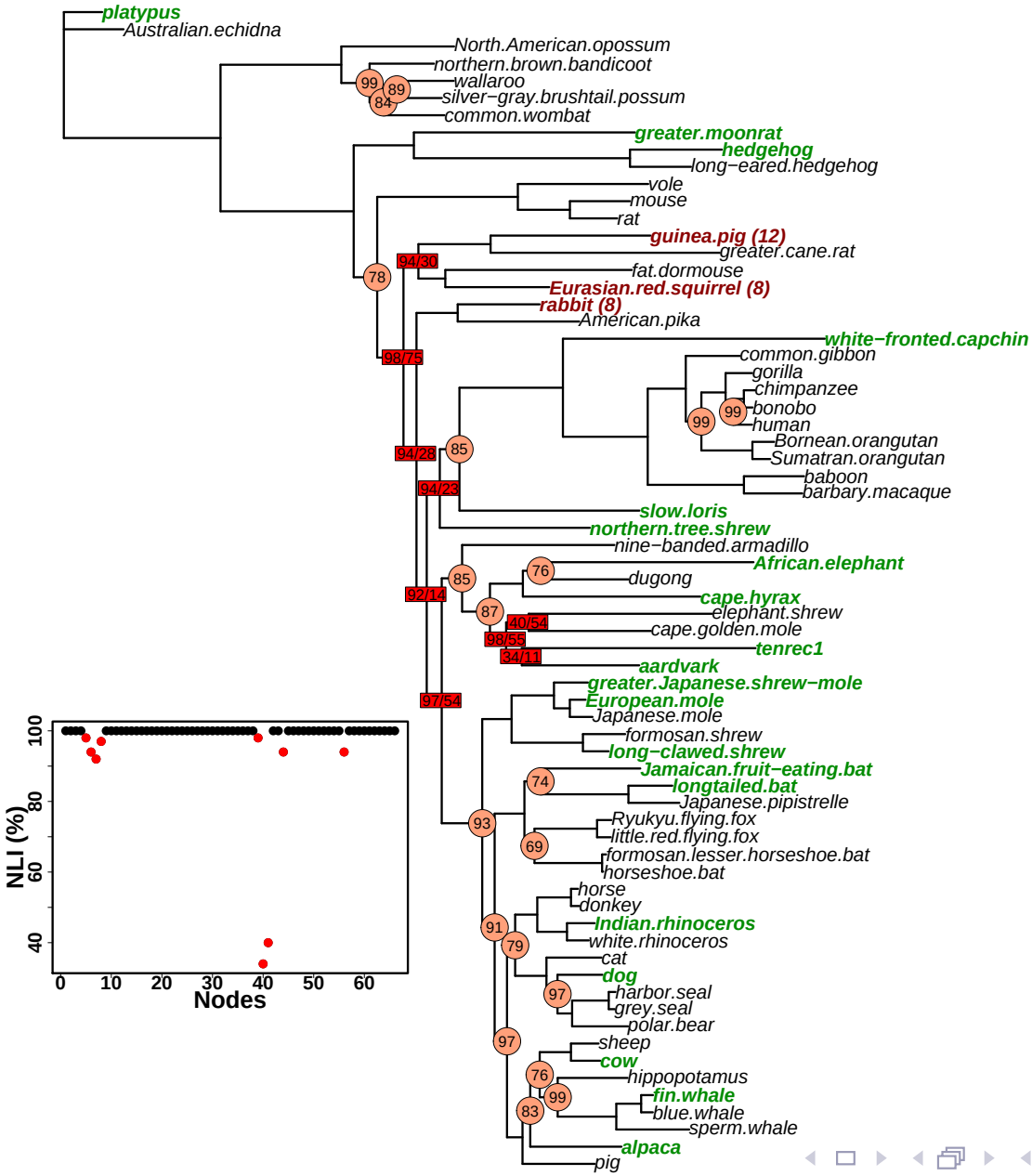
- Mitochondrial genome of 68 mammals,
- Amino Acids sequences,
- Sequences are 3658 sites long,
- Phylogeny published in Nikaido *et al.* in 2003.



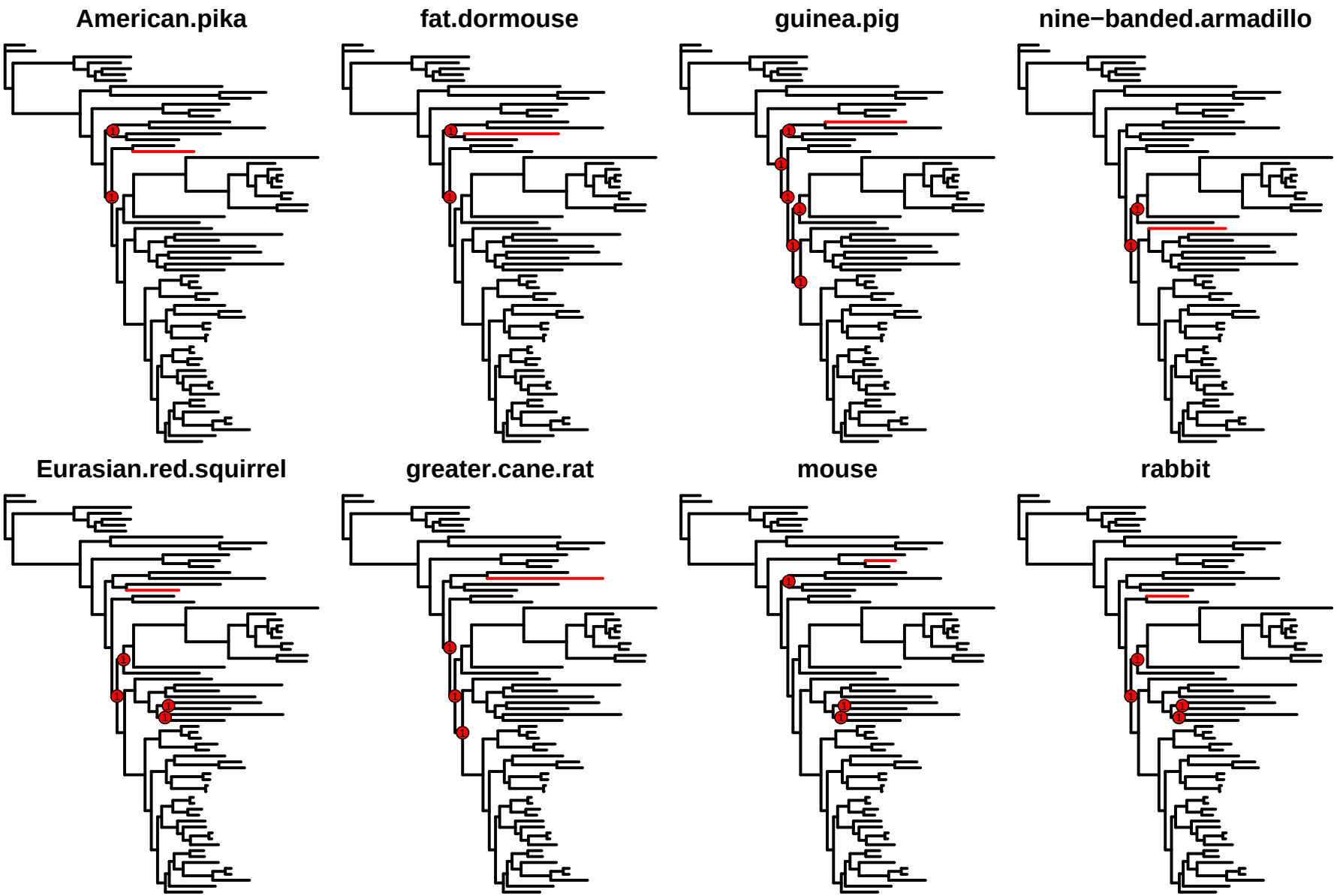
Species Leverage Index



Complete Phylogeny



Rogue Species



Summary

Two sources of uncertainties

- Outlier sites;
- Rogue species.

Two tools to detect them

- Influence functions;
- Species Leverage.

Summary

Two sources of uncertainties

- Outlier sites;
- Rogue species.

Two tools to detect them

- Influence functions;
- Species Leverage.

Conclusion and Perspectives

Conclusions

- Bootstrap: global measure of uncertainty,
- SLI,NLI are local ones to pinpoint the sources of uncertainties,
- Decompose the “black box” of bootstrap values,

Perspectives

- Impact of the evolution model,
- Statical properties of SLI,NLI.

Add-on


Homoplasy Influence on accuracy
4
8
16
32
0.1
0.05
0.025
50
75
100
125
150
300
600
RF distance
Homoplasy proportion