

Influence of Sites and Species on Phylogenetic Stability

M. Mariadassou A. Bar-Hen H. Kishino

Laboratoire MAP5
Université Paris Descartes

24 June 2009
LBBE Lab Seminar

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

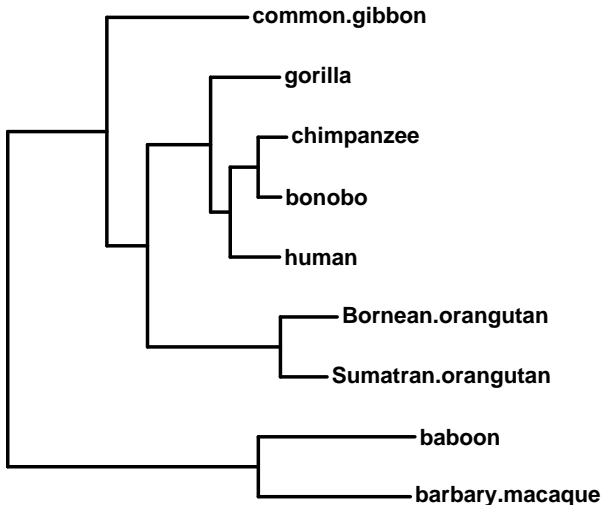
Basic Ideas:

- 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**.

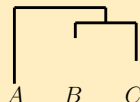
Example of A Phylogenetic Tree



Reconstructions and Limits

Two levels of reconstruction:

- Reconstruct the phylogeny:
 - Topology;
 - Branchs lengths.
- Reconstruct ancestral sequences.



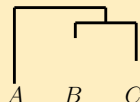
Issues:

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

Reconstructions and Limits

Two levels of reconstruction:

- Reconstruct the phylogeny:
 - Topology;
 - Branchs lengths.
- Reconstruct ancestral sequences.



Issues:

- Evolution is a **unique** event,
- Genetic information available only for **extant** species,
- Almost no direct observations or information about 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	<i>M</i>	<i>N</i>	<i>E</i>	N	<i>L</i>	<i>F</i>	<i>A</i>	<i>P</i>	<i>F</i>
Blue Whale	<i>M</i>	<i>N</i>	<i>E</i>	N	<i>L</i>	<i>F</i>	<i>A</i>	<i>P</i>	<i>F</i>
Chimpanzee	<i>M</i>	<i>N</i>	<i>E</i>	N	<i>L</i>	<i>F</i>	<i>A</i>	<i>S</i>	<i>F</i>
Bonobo	<i>M</i>	<i>N</i>	<i>E</i>	N	<i>L</i>	<i>F</i>	<i>A</i>	<i>S</i>	<i>F</i>
Gorilla	<i>M</i>	<i>N</i>	<i>E</i>	N	<i>L</i>	<i>F</i>	<i>A</i>	<i>S</i>	<i>F</i>
Bornean Orangutan	<i>M</i>	<i>N</i>	<i>E</i>	D	<i>L</i>	<i>F</i>	<i>T</i>	<i>P</i>	<i>F</i>

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

Several inference methods

Scan all trees and select the one which optimizes a given criteria:

- Maximum Parsimony: minimizes the number of changes needed to explain the current data;
- Maximum Likelihood: maximizes the likelihood of the data;
- Bayesian Analysis: maximizes the posterior probability of the data;
- etc.

Likelihood Based

- Assume (\mathbf{X}_i) *i.i.d.*;
- Choose an **evolution model** M (e.g. GTR+I+ Γ 4);
- **Discrete** topology T , **continuous** branch lengths b_T and model parameters θ_M ;
- Retrieve $(\hat{T}, \hat{b}_{\hat{T}}, \hat{\theta}_M)$ maximizing $\mathbb{P}((\mathbf{X}_i); M, \theta_M, T, b_T)$.

Likelihood Based

- Assume (\mathbf{X}_i) *i.i.d.*;
- Choose an **evolution model** M (e.g. GTR+I+ Γ 4);
- **Discrete** topology T , **continuous** branch lengths b_T and model parameters θ_M ;
- Retrieve $(\hat{T}, \hat{b}_{\hat{T}}, \hat{\theta}_{\hat{M}})$ maximizing $\mathbb{P}((\mathbf{X}_i); M, \theta_M, T, b_T)$.

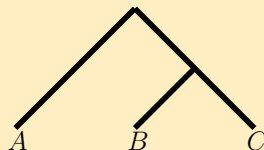
General Principle:

Alignment

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

→
ML

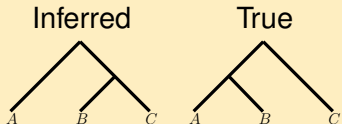
Phylogenetic tree



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!

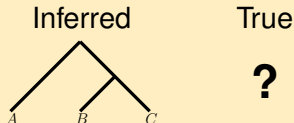
Confidence Issue:

- 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,



- **But the true tree is not available!**

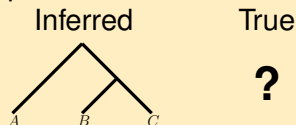
Confidence Issue:

- 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

- When no independent data set is available, we can only 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

- When no independent data set is available, we can only 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 uncertainties:
 - Finite sequence lengths (sampling errors);
 - Poor alignment quality (influent sites);
 - Poor taxon sampling (rogue taxa);
 - ...
- **Global** measure of uncertainty.

Bootstrap Weak Points:

- **Global** measure of uncertainty;
- Unable to breakdown the uncertainty;
- Unable to pinpoint **local** sources of uncertainty;
- Several other ways to tweak the data.

Bootstrap Values: A Robustness Index ?

Bootstrap Strong Points:

- Many potential causes for uncertainties:
 - Finite sequence lengths (sampling errors);
 - Poor alignment quality (influent sites);
 - Poor taxon sampling (rogue taxa);
 - ...
- **Global** measure of uncertainty.

Bootstrap Weak Points:

- **Global** measure of uncertainty;
- Unable to breakdown the uncertainty;
- Unable to pinpoint **local** sources of uncertainty;
- Several other ways to tweak the data.

Bootstrap Values: A Robustness Index ?

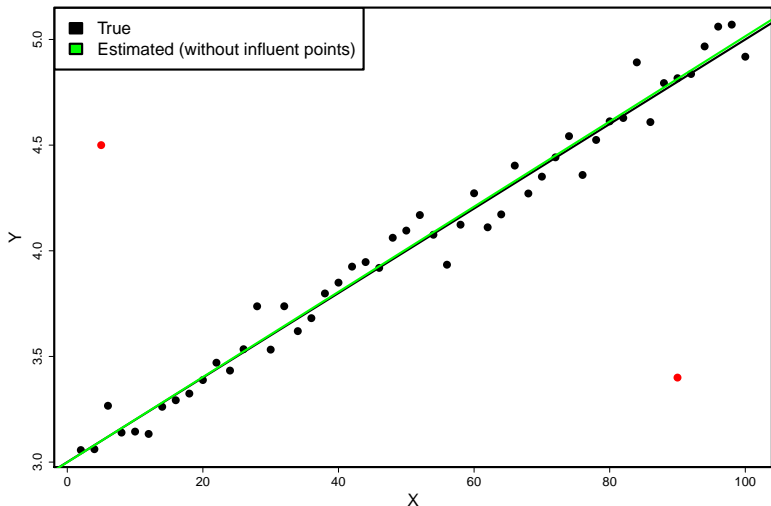
Bootstrap Strong Points:

- Many potential causes for uncertainties:
 - Finite sequence lengths (sampling errors);
 - **Poor alignment quality (influent sites);**
 - Poor taxon sampling (rogue taxa);
 - ...
- **Global** measure of uncertainty.

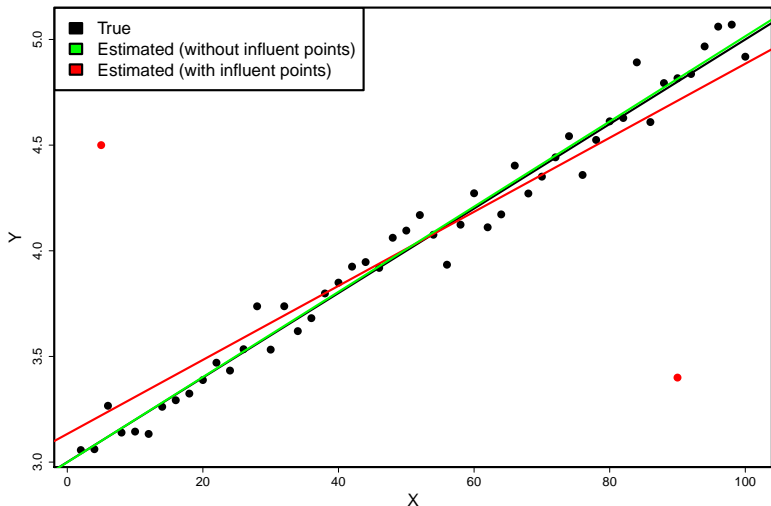
Bootstrap Weak Points:

- **Global** measure of uncertainty;
- Unable to breakdown the uncertainty;
- Unable to pinpoint **local** sources of uncertainty;
- Several other ways to tweak the data.

A Simple Example: Linear Regression



A Simple Example: Linear Regression



Outlier Sites: Motivation and Goal

Motivation: Filter Data

Source of errors for **sites**:

- 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

Source of errors for **sites**:

- 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, \dots, 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**:

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

measure the **influence** of a perturbation in direction x .

Empirical Version

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

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

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

About the Influence Function

Influence Function: Definition

Let X_1, \dots, 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**:

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

measure the **influence** of a perturbation in direction x .

Empirical Version

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

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

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

And for Phylogenies...

Definition

Let:

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

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

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

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

And for Phylogenies...

Definition

Let:

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

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

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

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

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.

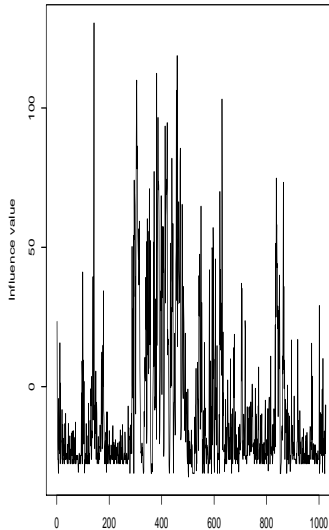
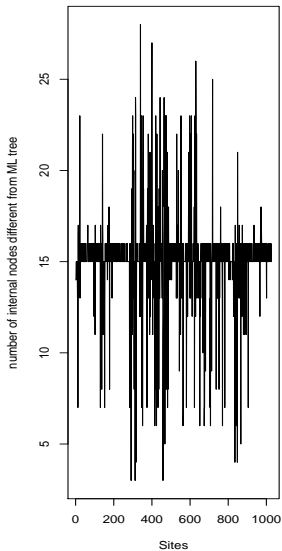
Strategy towards greater stability

- 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

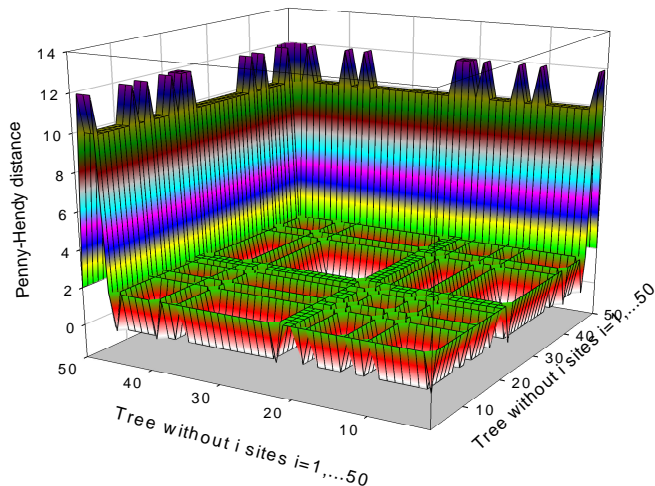
0	20	18	18	18	18	18	18	18	20
20	0	2	2	2	2	2	2	2	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
20	2	2	2	2	2	2	2	2	0

T_i : trees constructed without the i most influential sites.

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

Distance Between Trees

Distance between trees



Bootstrap Values: A Robustness Index ?

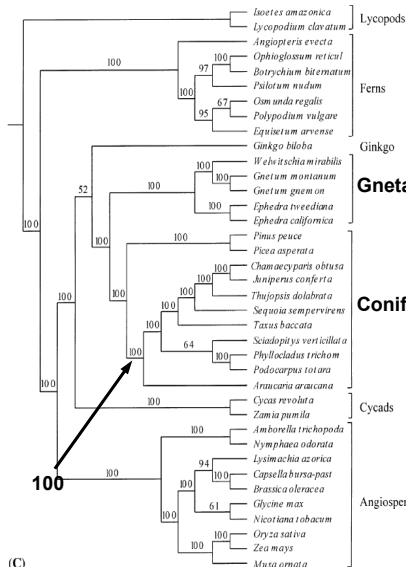
Bootstrap Strong Points:

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

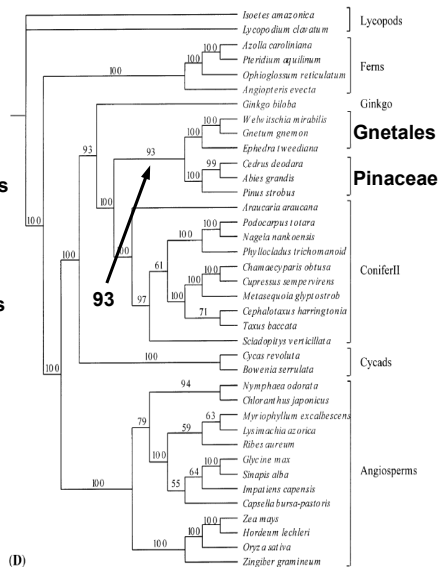
Bootstrap Weak Points:

- **Global** measure of uncertainty,
- Unable to breakdown the uncertainty,
- Unable to pinpoint **local** sources of uncertainty,
- Several other ways to tweak the data.

Seed Plant Phylogeny (Ridyn & al. 2002)



(C)



(D)

Taxon Influence Index: Motivation and Goal

Taxon Influence Index (TII)

- **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 taxa sample;
 - Remove species one at the time and infer a new tree $T^{(-i)}$ on the smaller species set;
 - Quantify difference between T and $T^{(-i)}$.

Taxon Influence Index: Motivation and Goal

Taxon Influence Index (TII)

- **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 taxa sample;
 - Remove species one at the time and infer a new tree $T^{(-i)}$ on the smaller species set;
 - Quantify difference between T and $T^{(-i)}$.

Taxon Influence Index: Motivation and Goal

Taxon Influence Index (TII)

- **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 taxa sample;
 - Remove species one at the time and infer a new tree $T^{(-i)}$ on the smaller species set;
 - Quantify difference between T and $T^{(-i)}$.

Taxon Influence Index (TII)

Definition

Let:

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

The **Taxon Influence Index (TII)** of species i is:

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

where d is any adapted distance.

Taxon Influence Index (TII)

Definition

Let:

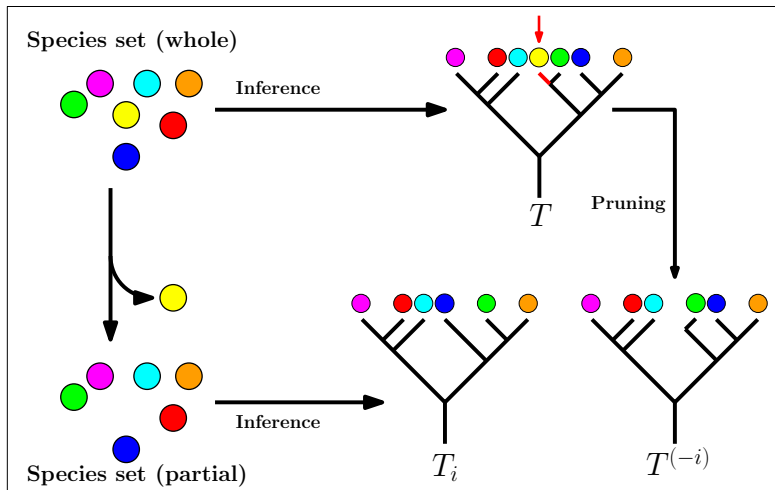
- $\mathbf{X} = (\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(s)})'$ be the complete alignment,
- $\mathbf{X}^{(-i)} = \mathbf{X} \setminus \mathbf{X}^{(i)}$ all the species but species i ,
- \hat{T} the ML tree and associated parameters for \mathbf{X} ,
- $\hat{T}^{(-i)}$ the tree \hat{T} after pruning species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

The **Taxon Influence Index (TII)** of species i is:

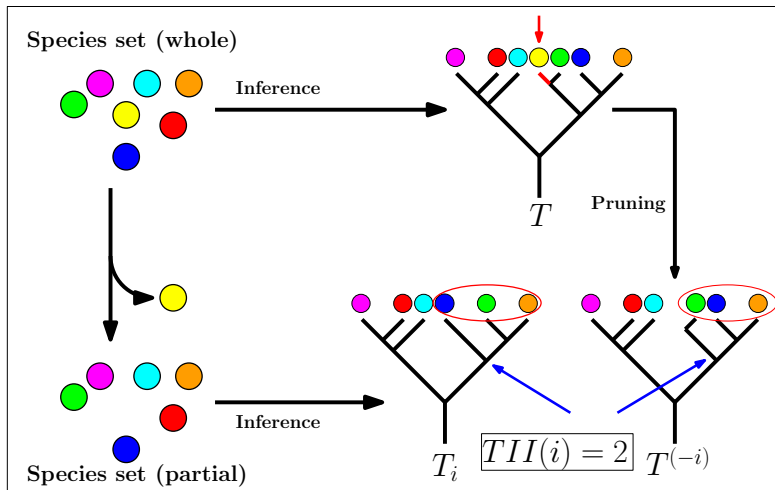
$$TII(i) = d(\hat{T}^{(-i)}, \widehat{T^{(-i)}})$$

where d is any adapted distance.

Method



Method



Branch Stability scores (BS scores)

Definition

Branch Stability score (BS score) of b : number of inferred trees in which the node is retrieved,

$$BS(b) = \sum_{i=1}^s \mathbb{1}_{\widehat{T^{(-i)}}}(b)$$

Problems

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

Branch Stability scores (BS scores)

Definition

Branch Stability score (BS score) of b : number of inferred trees in which the node is retrieved,

$$BS(b) = \sum_{i=1}^s \mathbb{1}_{\widehat{T^{(-i)}}}(b)$$

Problems

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

TII and BS scores

Interpretation

- TII:**
- Low value: adding/removing the taxon from the data set has (almost) impact on the tree;
 - High value: Influential taxon, adding/removing it greatly affects the tree, may be a rogue taxon.
- BS:**
- High value: stable branch, highly resilient to taxon sampling;
 - Low value: weak branch, highly sensitive to taxon sampling.

Strategy towards robustness

- Focus on **rogue taxa**: taxa with high TII;
- Rank them in increasing TII;
- Remove them one at the time until a stable tree is found.

TII and BS scores

Interpretation

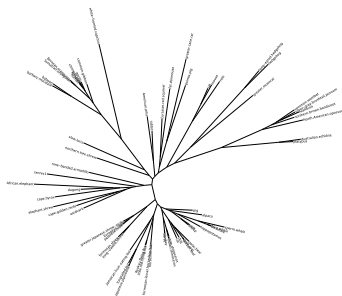
- TII:**
- Low value: adding/removing the taxon from the data set has (almost) impact on the tree;
 - High value: Influential taxon, adding/removing it greatly affects the tree, may be a rogue taxon.
- BS:**
- High value: stable branch, highly resilient to taxon sampling;
 - Low value: weak branch, highly sensitive to taxon sampling.

Strategy towards robustness

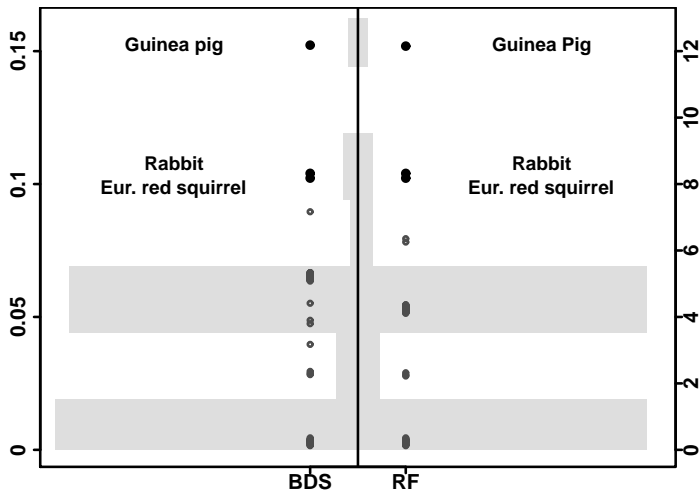
- Focus on **rogue taxa**: taxa with high TII;
- Rank them in increasing TII;
- Remove them one at the time until a stable tree is found.

Data: Placental Mammal Phylogeny

- Mitochondrial genome of 68 mammals;
- Amino Acids sequences;
- Sequences are 3658 AA long;
- MtMam + I + Γ 4 model;
- Phylogeny published in Nikaido *et al.* in 2003.

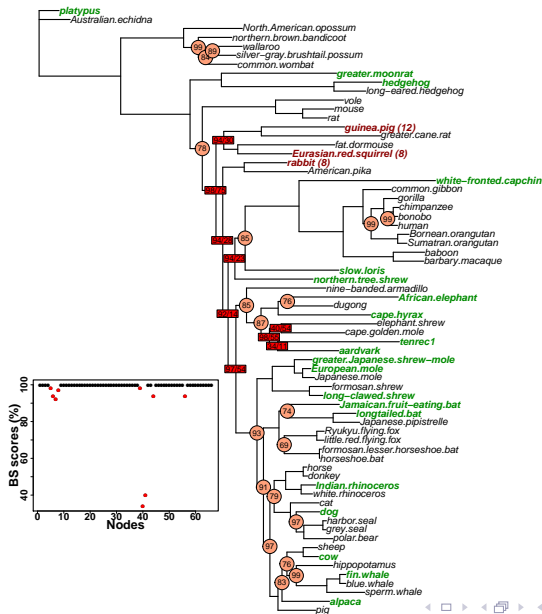


Taxon Influence Index

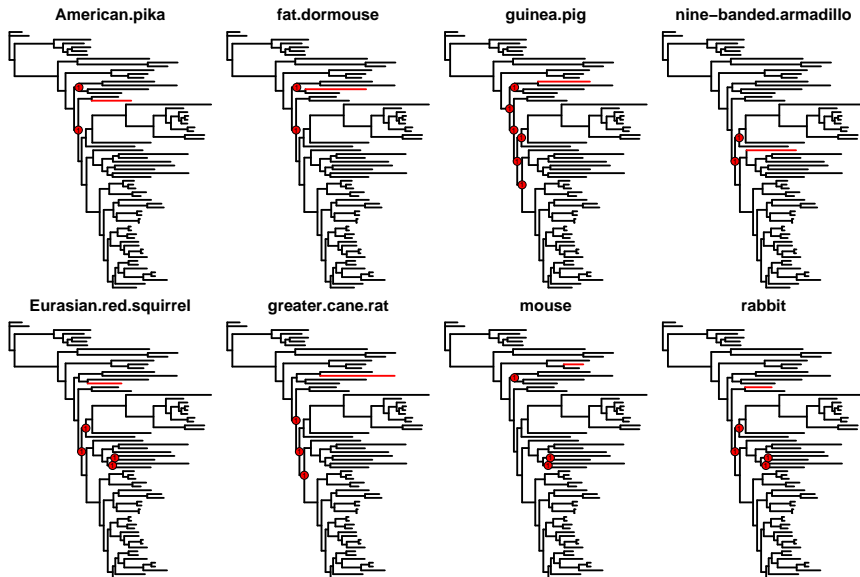


BSD: Branch-score distance / RF: Robinson-Foulds distance

Complete Phylogeny



Influential Species



Summary

Two local sources of instability

- Outlier sites;
- Rogue species.

Two tools to detect them

- Influence functions;
- Taxon Influence Index.

Summary

Two local sources of instability

- Outlier sites;
- Rogue species.

Two tools to detect them

- Influence functions;
- Taxon Influence Index.

Conclusion and Perspectives

Conclusions

- Bootstrap: global measure of uncertainty,
- TII, BS and IF scores and are local ones to pinpoint the sources of uncertainties,

Perspectives

- Impact of the evolution model / inference method;
- Statical properties of the indices.

Further Reading



A. Bar-Hen, M. Mariadassou, M.-A. Poursat and
P. Vandenkoornhuyse.

Influence function for robust phylogenetic reconstructions.

Mol. Biol. Evol. 25:869–873, 2008



M. Mariadassou, A. Bar-Hen and H. Kishino.

Taxon influence: assessing taxon-induced incongruities in
phylogenetic inference

submitted to *Syst. Biol.*