

Phylogeny Robustness: Influence of Species

M. Mariadassou A. Bar-Hen H. Kishino

MAP5 Lab
Université Paris Descartes

January 2009
SMPGD'09

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

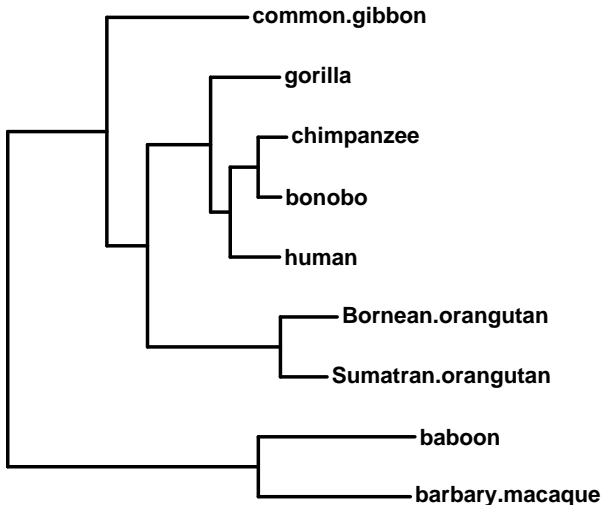
Basic Ideas:

- 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

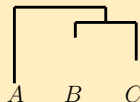


Reconstructions and Limits

Two levels of reconstruction:

- Reconstruct the phylogeny:

- Topology,
- Branchs lengths.



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

Issues:

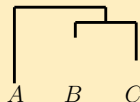
- Evolution is a **unique** event,
- Genetic information available only for **extant** species,
- Almost no direct observations or results on the evolutionary process.

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	<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): $\mathbf{X}^{(2)} = \mathbf{MNENLFAPFM}$.

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 $\mathbb{P}((\mathbf{X}_i); M, T, \theta_T)$.

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 $\mathbb{P}((\mathbf{X}_i); M, T, \theta_T)$.

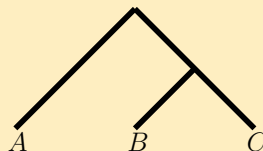
General Principle:

Alignment

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

→
ML

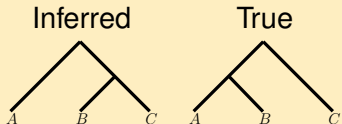
Phylogenetic tree



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!

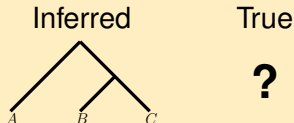
Confidence Issue:

- 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!**

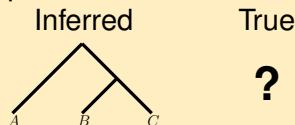
Confidence Issue:

- 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

- 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;

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;

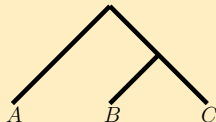
Bootstrap Values: the Theory

Original Dataset:

Alignment

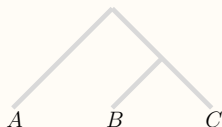
A	A	C	T	T
B	G	G	A	T
C	G	G	C	C

Phylogenetic tree

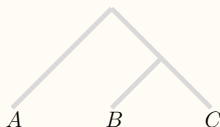


Bootstrap Datasets:

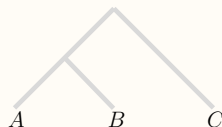
A	A	C	T	C
B	G	G	A	G
C	G	G	C	G



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



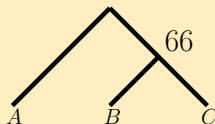
Bootstrap Values: the Theory

Original Dataset:

Alignment

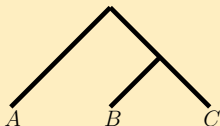
A	A	C	T	T
B	G	G	A	T
C	G	G	C	C

Phylogenetic tree

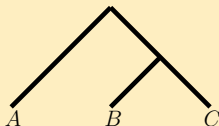


Bootstrap Datasets:

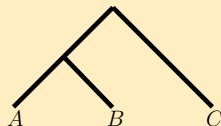
A	A	C	T	C
B	G	G	A	G
C	G	G	C	G



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



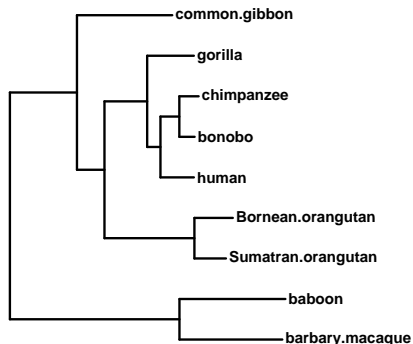
A	T	T	T	T
B	A	T	A	T
C	C	C	C	C



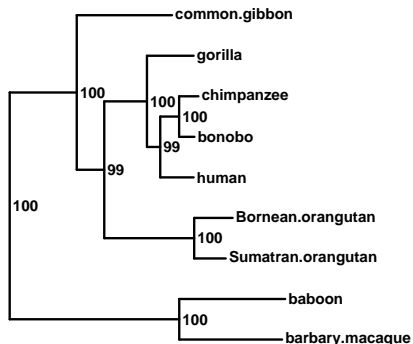
Bootstrap Values: An Example

Bootstrap Values

Without



With



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 ?

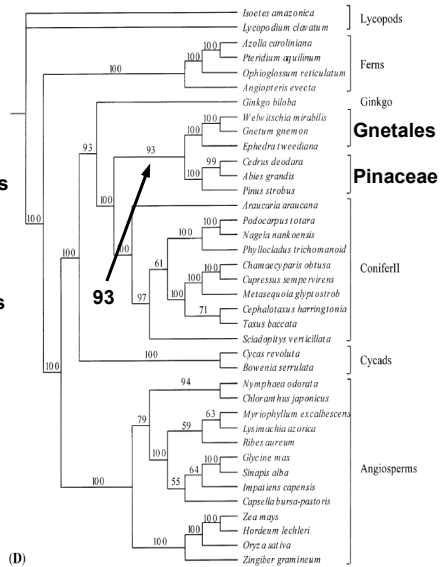
Bootstrap Strong Points:

- 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.

12 / 23



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:

- $\mathbf{X} = (\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(s)})'$ 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 Species Leverage Index (SLI) of species i is:

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

where d is any adapted distance .

Species Leverage Index (SLI)

Definition

Let:

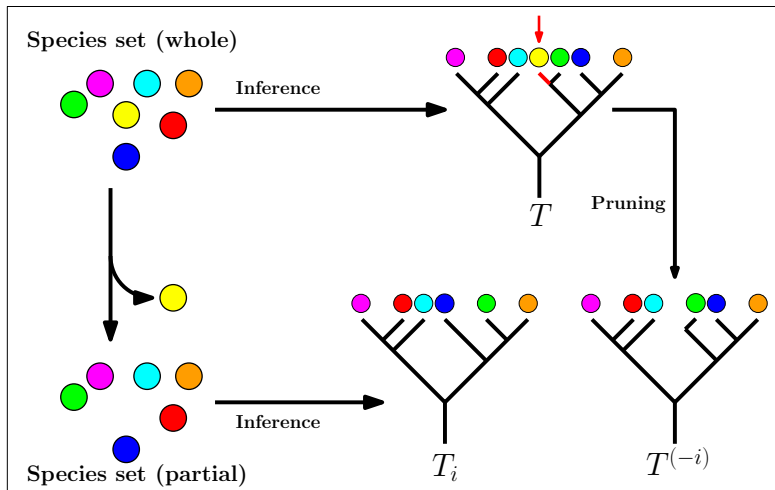
- $\mathbf{X} = (\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(s)})'$ be the complete alignment,
- $\mathbf{X}^{(-i)} = \mathbf{X} \setminus \mathbf{X}^{(i)}$ all the species but species i ,
- \hat{T} the ML tree and associated parameters for \mathbf{X} ,
- $\hat{T}^{(-i)}$ the tree \hat{T} after pruning species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

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

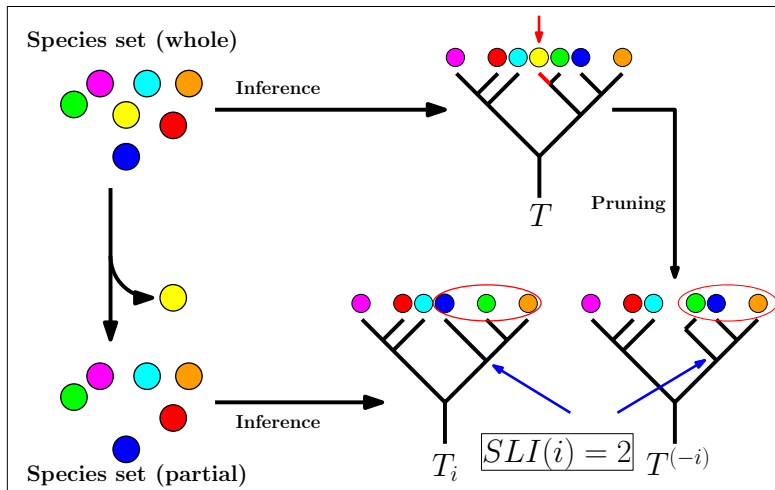
$$SLI(i) = d(\hat{T}^{(-i)}, \widehat{T^{(-i)}})$$

where d is any adapted distance .

Method



Method



Nodes Leverage Index (NLI)

Definition

Let:

- $\mathbf{X}, \mathbf{X}^{(-i)}, \hat{T}, \hat{T}^{(-i)}, \widehat{T^{(-i)}}$ defined as before,
- A an internal node of \hat{T} ,

The **Nodes Leverage Index (NLI)** of A is:

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

with $\mathbb{1}_{\widehat{T^{(-i)}}}(A)$ being 1 if A is present in $\widehat{T^{(-i)}}$ and 0 otherwise.

Nodes Leverage Index (NLI)

Definition

Let:

- $\mathbf{X}, \mathbf{X}^{(-i)}, \hat{T}, \hat{T}^{(-i)}, \widehat{T^{(-i)}}$ defined as before,
- A an internal node of \hat{T} ,

The **Nodes Leverage Index (NLI)** of A is:

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

with $\mathbb{1}_{\widehat{T^{(-i)}}}(A)$ being 1 if A is present in $\widehat{T^{(-i)}}$ and 0 otherwise.

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

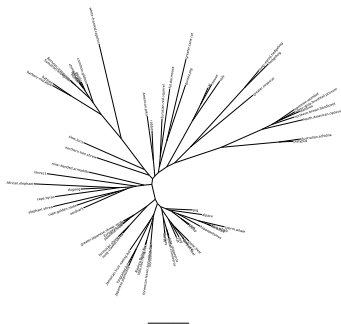
- SLI:**
- Low value: adding/removing the species from the dataset has (almost) impact on the tree;
 - High value: “rogue” species, adding/removing it greatly affects the tree.
- NLI:**
- High value: stable nodes, highly resilient to taxon sampling;
 - Low value: weak nodes, highly sensitive to taxon sampling.

Strategy towards robustness

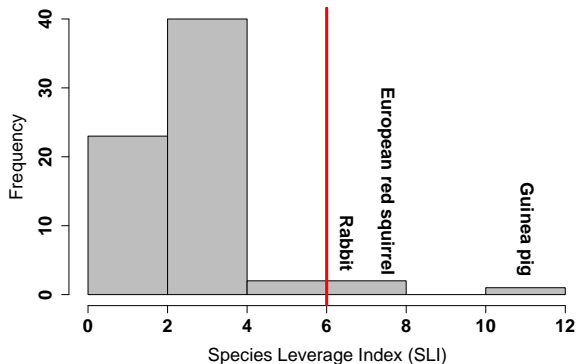
- Focus on **rogues species**: species with high SLI;
- Rank them in increasing SLI;
- Remove them one at the time until a stable tree is found.

Data: Placental Mammal Phylogeny

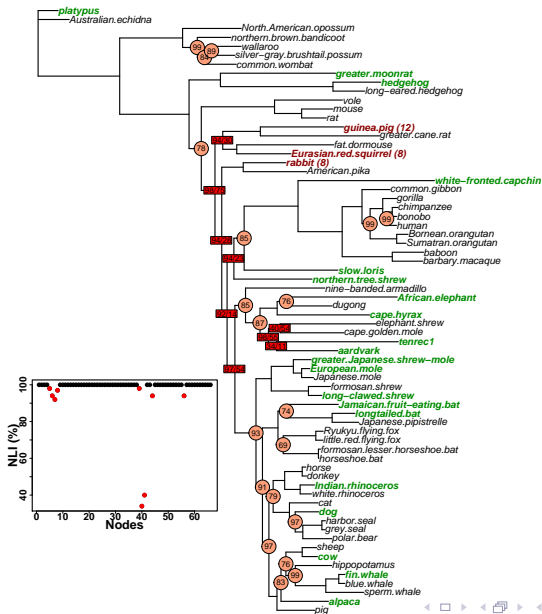
- Mitochondrial genome of 68 mammals,
- Amino Acids sequences,
- Sequences are 3658 sites long,
- Phylogeny published in Nikaido *et al.* in 2003.



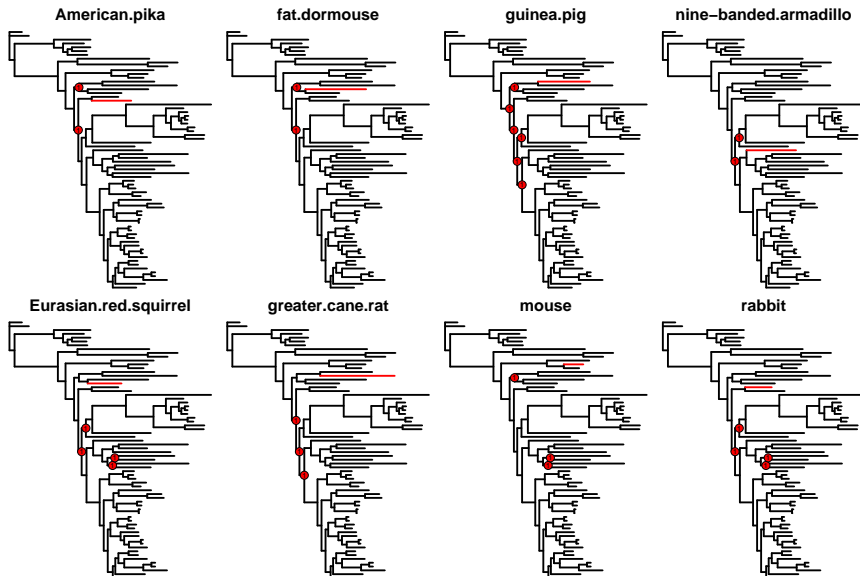
Species Leverage Index



Complete Phylogeny



Rogue Species



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,
- Anything else I can think about.

The End

Thank you for your attention!