

An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications, and transfers

Jean-Philippe Doyon¹ Celine Scornavacca² Gergely J. Szöllősi³
Vincent Ranwez⁴ Vincent Berry¹

1 - LIRMM, CNRS - Univ. Montpellier 2, France.

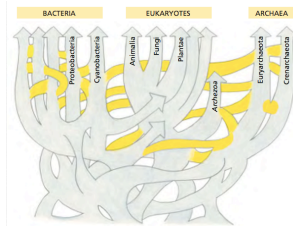
2 - Center for Bioinformatics (ZBIT), Tuebingen Univ., Germany.

3- LBBE, CNRS - Univ. Lyon 1, France.

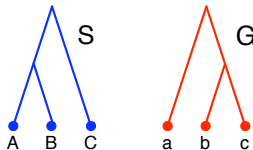
4- ISEM, CNRS - Univ. Montpellier 2, France.

SMBE, Lyon
July 2010

Inferring the Tree of Life



- Vertical signal from the early stages of life is still visible in current genomes [DAUBIN ET AL 03, KURLAND ET AL 03, ...]
- A ToL might be obtained by retrieving signals from gene trees.
- **RECONCILIATION** explains incongruence between G and S by postulating macro events. **A CYCLE!!**

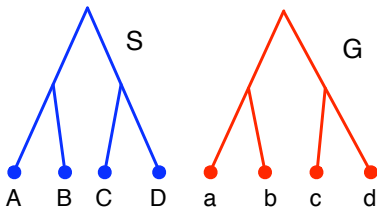


The MPR problem

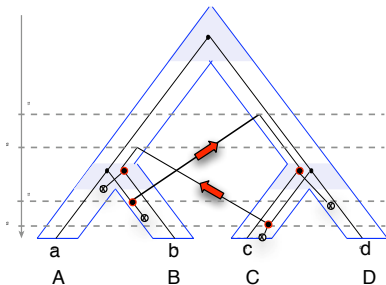
The Most Parsimonious Reconciliation problem

- Speciation (S), Duplication (D), Transfer (T), and Loss (L) events (A cost for each event).
- Compute a reconciliation with a minimal cost

Gene and Species trees



Time inconsistent reconciliation



Previous approaches & models

Species Graph

[Gorecki]

Locations of (possible) transfers are defined in advance in S .

Some existing reconciliation models

- Don't **directly** account for losses [HALLETT & LAGERGREN 04]
- Can lead to **time inconsistent reconciliations**
(Tarzan & Jane software) [MERKLE ET AL 05-10]

Dated species tree S

[LAGERGREN'S GROUP 09-10, LYUBETSKY ET AL 09, MERKLE ET AL 05-10, GORBUNOV ET AL 09, LIBESKIND-HADAS 09]

Previous approaches & models

Species Graph

[Gorecki]

Locations of (possible) transfers are defined in advance in S .

Some existing reconciliation models

- Don't **directly** account for losses [HALLETT & LAGERGREN 04]
- Can lead to **time inconsistent reconciliations**
(Tarzan & Jane software) [MERKLE ET AL 05-10]

Dated species tree S

[LAGERGREN'S GROUP 09-10, LYUBETSKY ET AL 09, MERKLE ET AL 05-10, GORBUNOV ET AL 09, LIBESKIND-HADAS 09]

Our contribution

An efficient model for MPR problem

- Considering a **dated** species tree S .
- Relying on **6 atomic events**, each one being fast to investigate

A dynamic programming algorithm

- Based on a **small subdivision S'** of S
- **Fast**: runs in time $O(|S'| \cdot |G|)$
- Previous algorithms in $O(|S|^4 \cdot |G|^4)$ and $O(|S'|^3 \cdot |G|)$

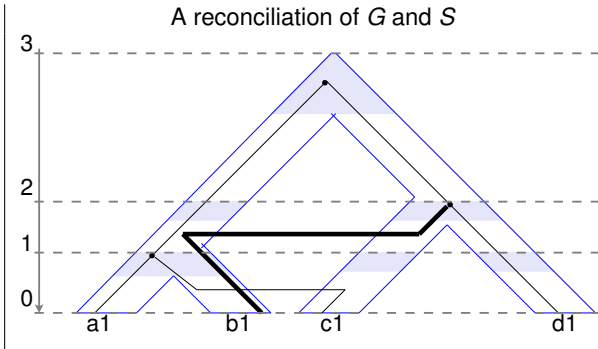
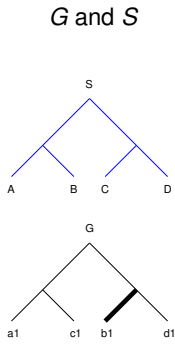
Experimental results for the relevance of parsimony

Is parsimony relevant to infer the evolutionary scenario of a gene family?

An Efficient Reconciliation Model

A reconciliation between a gene tree G and a species tree S

- Maps each edge of G onto an ordered sequence of branches of S' .
- Induces speciation (\mathbb{S}), duplication (\mathbb{D}) and transfer (\mathbb{T}).

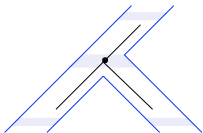


Transfers between branches within the same time slice

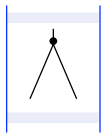
An Efficient Reconciliation Model

Six *Atomic events*, where losses are implicitly considered (*Parsimony*)

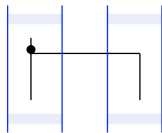
Speciation (\mathbb{S})



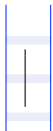
Duplication (\mathbb{D})



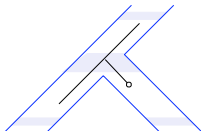
Transfer (\mathbb{T})



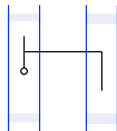
No event (\emptyset)



Speciation + Loss (\mathbb{SL})



Transfer+Loss (\mathbb{TL})



Theorem

A Most Parsimonious Reconciliation is computed in time $\Theta(|G| \cdot |S'|)$

Two Datasets DS_1 and DS_2

Details of the simulation process

- 10 species trees on 100 species (Birth and Death with a ratio = 1.25)
- DTL scenarios generated (Poisson process with rates \mathbb{L}_R , \mathbb{T}_R and \mathbb{D}_R)
- Gene trees have between 59 and 93 leaves

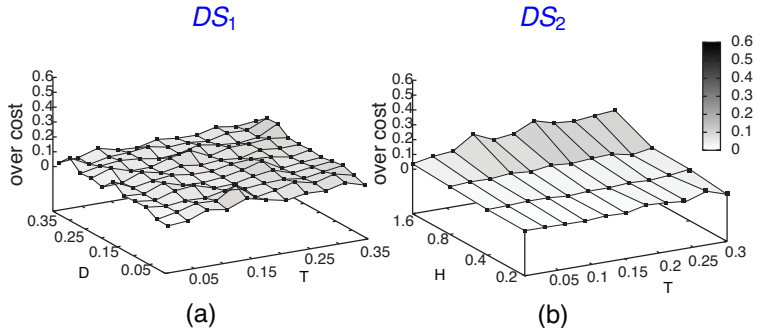
DS_1 : “Simulate” a relatively large time scale (archaean or bacterial phylum)

- Fixed rate $\mathbb{L}_R = 0.7$ and tree height $h = 1$
- 11 values for \mathbb{T}_R and \mathbb{D}_R in $[0.01, 0.35]$
- 6,050 $G = (5 G) \times (10 S) \times (11 \times 11 \text{ rate pairs})$

DS_2 : “Simulate” different phylogenetic time scales

- Four different tree heights $h \in [0.2, 0.4, 0.8, 1.6]$
- Fixed ratio $\mathbb{L}_R / (\mathbb{D}_R + \mathbb{T}_R + \mathbb{L}_R) = 0.7$ [CSUROS AND MIKLOS]
- 11 values for $\mathbb{T}_R \in [0, 0.3]$ and $\mathbb{D}_R = 0.3 - \mathbb{T}_R$ fixed (varying the importance of \mathbb{T} versus \mathbb{D}).
- 8,800 $G = (20 G) \times (10 S) \times (4 \times 11 \text{ rate pairs})$

Efficiency of parsimony according to costs

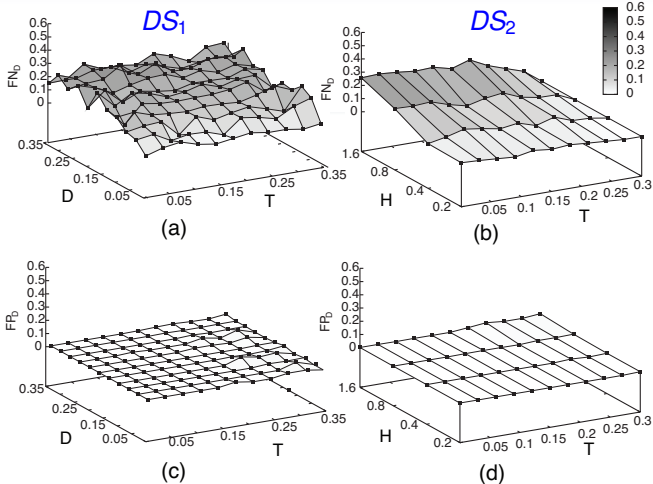


Over cost of the real scenario w.r.t. MPR

- Small for all **D** and **T** rates (DS_1)
- Increases with the **height** of the gene trees (DS_2)
- Parsimony might be considered as a **credible criterion** for reconciliations

Great!

Accuracy of parsimony to retrieve \mathbb{D} events

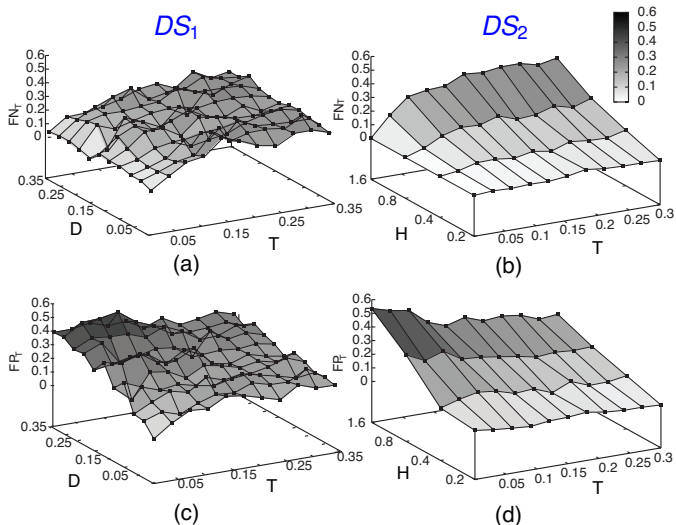


False Negatives / Positives: Node of G + Branche in S'

- Reasonably few forgotten duplications (homoplasy and several MPRs?)
- *Very* few False Positives

Not bad!

Accuracy of parsimony to retrieve \mathbb{T} events



False Negatives / Positives: Node of $G + 2$ Branches in S'

Large number of \mathbb{D} leads to non-trivial errors in \mathbb{T} prediction

Huh huh... :(

Transfers among archaeal genomes

Input data

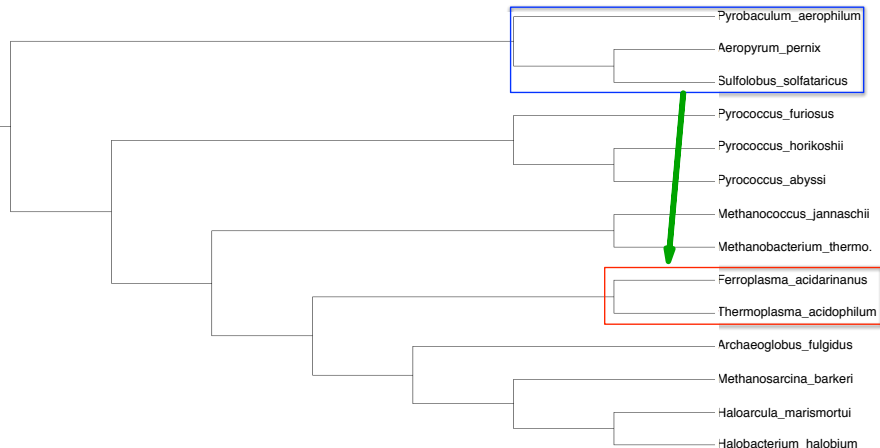
- Dated species tree: 14 archaeal (53 ribosomal proteins)
3 dates for (*Ferroplasma* A., *Thermoplasma* A.) clade.
- Gene tree: ribosomal proteins
2 roots. [MATTE ET AL. 2002; TOFIGH ET AL. 2010]
- 6 cases

MPR against *Tofigh et al.* (May propose Time Inconsistent transfer)

- MPR: 5T + 3L
- *Tofigh et al.*: 5 D/T (Losses are considered “a posteriori”)

What is the relevancy of the 5 transfers?

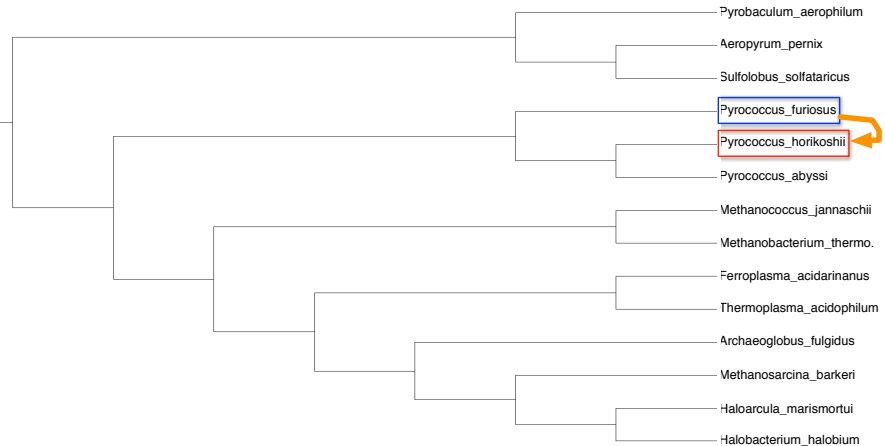
#1: From Crenarchaeota to the plasma



Apparently correct (both roots of G , \neq direction)

- Other transfers proposed in the same dir. and with different methods
- Same ecological niche

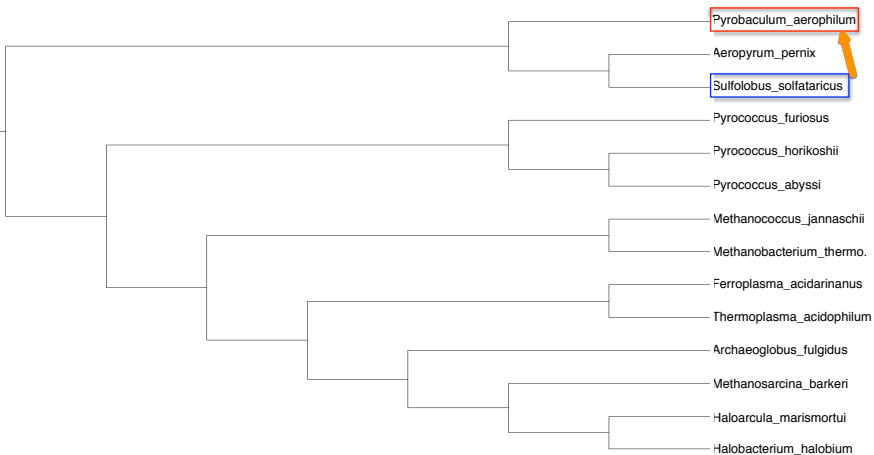
#2: From *Pyrococcus furiosus* to *Pyrococcus horikoshii*



Seem to be correct (both roots of *G*)

- High bootstrap values in species and gene trees.
- But small sequences and branch lengths (gene tree).

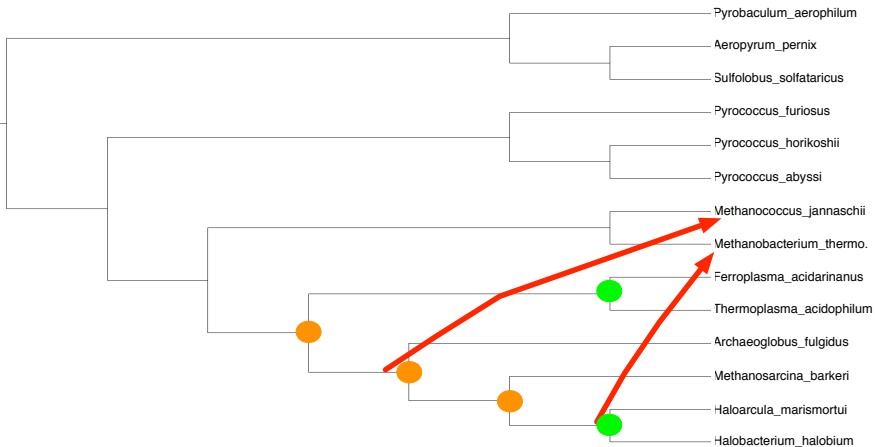
#3: From *Sulfolobus solfataricus* to *Pyrobaculum aerophilum*



Seem to be correct (both roots of G)

- High bootstrap values in species and gene trees.
- More studies to do.

#4,5: Reconciling trees with lack of resolution



Artifactual transfers (probably)

- Low bootstrap values in species and gene trees.
- Collapsing unsupported nodes erases discrepancies between trees.

Our Most Parsimonious Reconciliation algorithm

- Proposes Time-Consistent transfers;
- Directly account for losses (discriminate among different scenarios).
- Much faster (cpu and complexity) than previous ones

Experimental conclusions

- Parsimony cost fits nicely with real one.
- Few duplications not recovered and almost no incorrect ones predicted.
- Transfers less correctly predicted ($\approx 20 - 30\%$ errors).

What next?

- Enumerating and counting MPRs.
- Links between MPR and ML reconciliations [DOYON ET AL 09]
- Polytomous trees (as in Notung) [VERNOT ET AL 08]
- Use synteny information

Our Most Parsimonious Reconciliation algorithm

- Proposes Time-Consistent transfers;
- Directly account for losses (discriminate among different scenarios).
- Much faster (cpu and complexity) than previous ones

Experimental conclusions

- Parsimony cost fits nicely with real one.
- Few duplications not recovered and almost no incorrect ones predicted.
- Transfers less correctly predicted ($\approx 20 - 30\%$ errors).

What next?

- Enumerating and counting MPRs.
- Links between MPR and ML reconciliations [DOYON ET AL 09]
- Polytomous trees (as in Notung) [VERNOT ET AL 08]
- Use synteny information

Acknowledgment



Phyl-ARIANE

Phylogenomics: integrated algorithms and visualizations for analyzing the evolution of life

<http://www.lirmm.fr/phylariane/>

Thanks

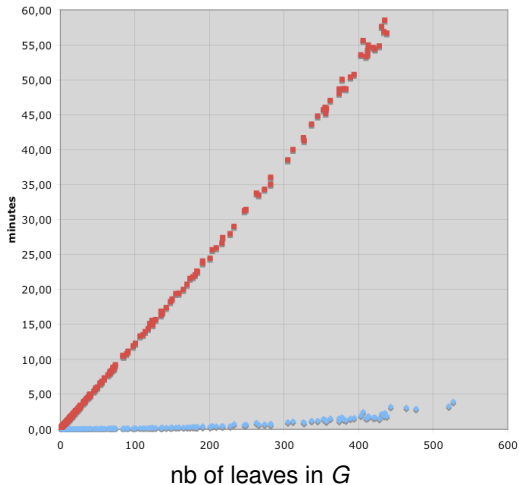
- Vincent Berry, LIRMM, Montpellier (FR)
- Celine Scornavacca, Univ. Tuebingen (GE)
- Vincent Ranwez, ISE-M, Montpellier (FR)
- Gergely Szöllözi Eric Tannier & Vincent Daubin, Lyon (FR)
- Céline Brochier for the gene tree and her help on the Archaeal dataset
- Mukul S. Bansal for the dataset of Guigo et al. 1996
- Vassily Lyubetsky and Kostya Gorbunov, Moscow (RU)

Funding

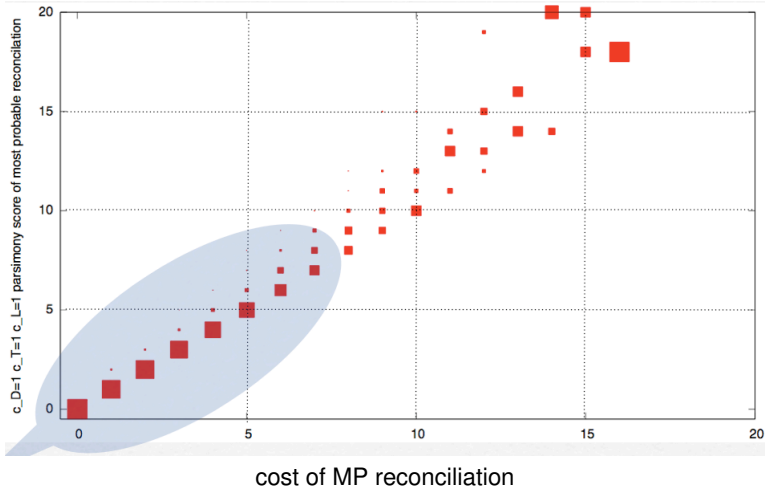
Phylariane ANR project, Région LR, CNRS, ...

Running times

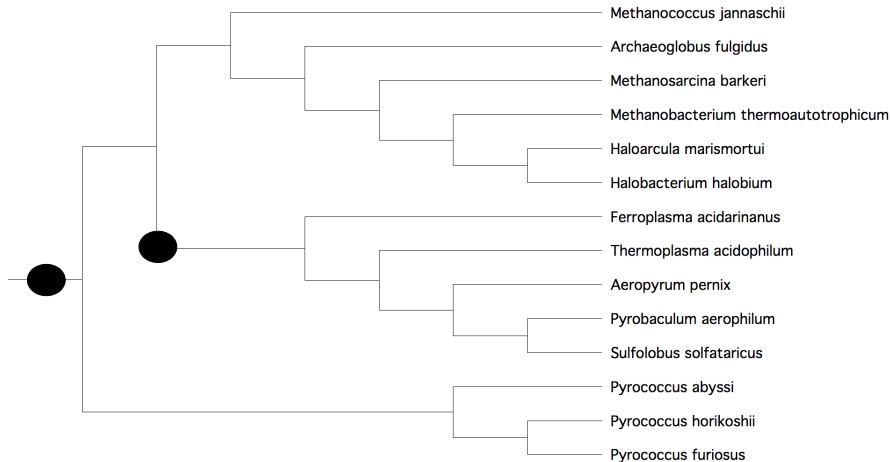
Comparison with an implementation of [Gorbunov et al 09]:
from dozens minutes to less than 2 sec (between 1.09s and 1.38s)



Relationship between the MP and ML criteria



Two roots for the (rpl12e) ribosomal proteins



Dynamic Programming Algorithm

Properties

- $\mathbb{S}\mathbb{L}$ is an optimal scenario where one gene goes extinct after an \mathbb{S} event (Idem for $\mathbb{T}\mathbb{L}$ and \mathbb{T})
- Any $\mathbb{T}\mathbb{L}$ event is (possibly) followed by a different event.
- The model allows to progress either in S' (its time) or in G .
- The best landing place is independent of the donor branch.

Maximum Likelihood approach

Similar algorithm applies to ML

[SZÖLLÖZI ET AL IN PREP.]

Vertebrates: Whole Genome Duplications

Episode Clustering Problem (without transfer)

Given S and $\{G_1, \dots, G_n\}$, minimize the number of locations in S where all duplications can be placed.

53 gene trees form 16 vertebrates

[GUIGO ET AL. 1996]

	# Dup	# Spots of S	# WGD	MPR wrt. Guigo
Guigo et al.	46	4	5	
MPR ($\mathbb{D}_C, \mathbb{L}_C \geq 1$)	46	6	9	80%
MPR ($\mathbb{D}_C = 1, \mathbb{L}_C = 0$)	46	6	9	95%

Episode Clustering Problem

On this dataset, **Whole-Genome-Duplications** can be retrieved and located solely with **Most-Parsimonious-Reconciliation** classical approaches.

