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

(ANR-08-EMER-011)

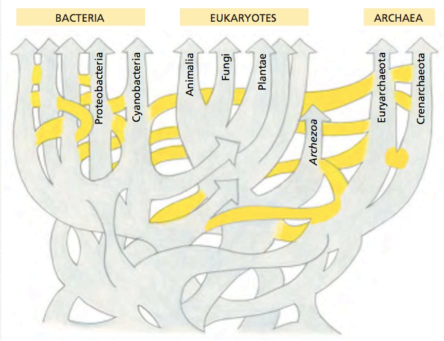
Jean-Philippe Doyon [1] Celine Scornavacca[2] Gergely J. Szöllősi[3]
Vincent Ranwez[4] Vincent Berry[1]

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.

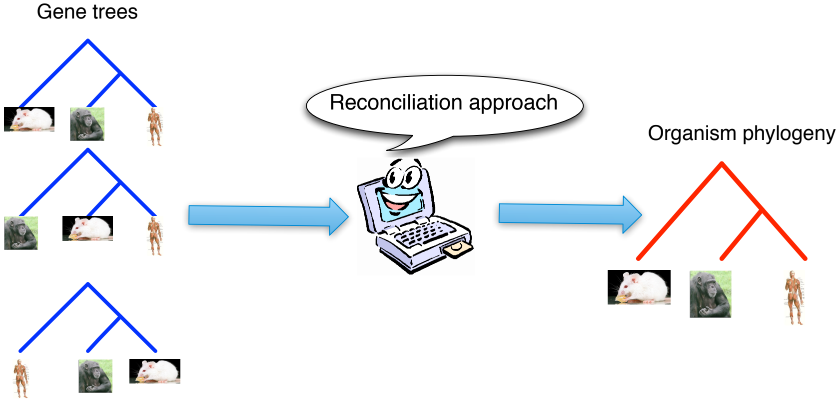
JOBIM, Montpellier
September 2010

Concept of the Tree of Life

- **For**: Kurland et al. 2003, Puigbò et al. 2009
- **Moderate:** Galtier et Daubin 2008
- **Against**: Bapteste et al. 2005; Koonin 2007



Inferring a Tree of Life by retrieving signals from gene trees



Gene trees

Reconciliation approach

Organism phylogeny

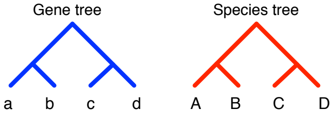
Reconciliation

- Parsimony and probabilistic app. [Lagergren et al.; Gorecki et al.]
- Used to identify orthologous sequences (functional annotation)
- Similar concepts in Ecology and Biogeography [Page et al.]

Reconciliation depicts coevolution

(Some) Macro Evolutionary Events

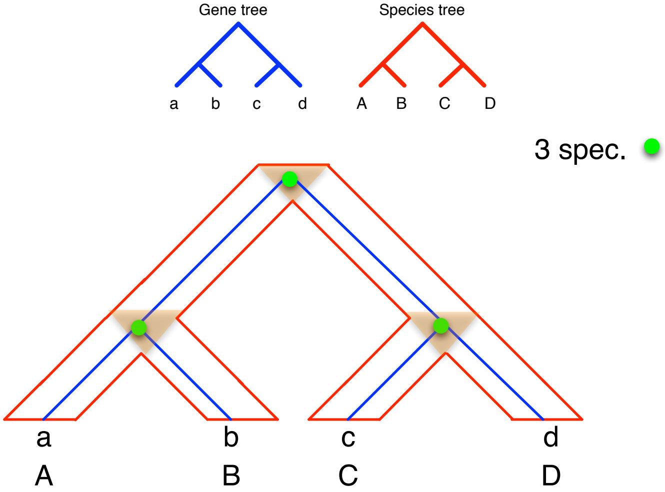
Speciation ($\mathbb{S}$), Duplication ($\mathbb{D}$), Transfer ($\mathbb{T}$), and Loss ($\mathbb{L}$)



Reconciliation depicts coevolution

(Some) Macro Evolutionary Events

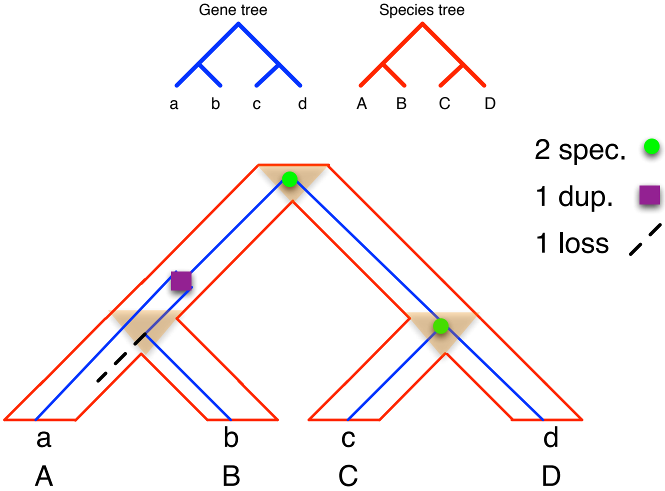
Speciation ($\mathbb{S}$), Duplication ($\mathbb{D}$), Transfer ($\mathbb{T}$), and Loss ($\mathbb{L}$)



Reconciliation depicts coevolution

(Some) Macro Evolutionary Events

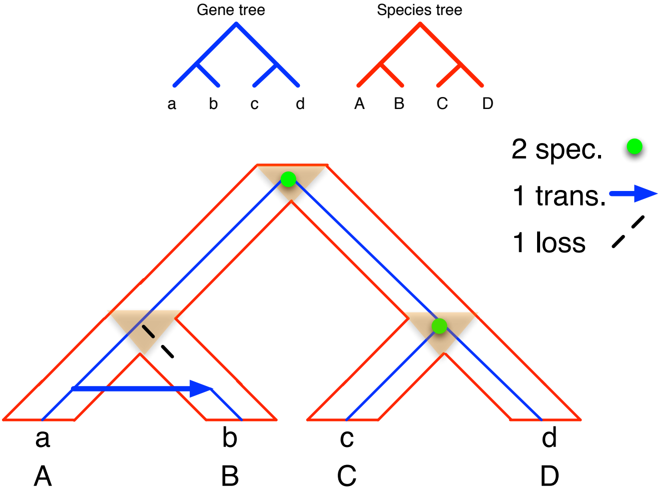
Speciation ($\mathbb{S}$), Duplication ($\mathbb{D}$), Transfer ($\mathbb{T}$), and Loss ($\mathbb{L}$)



Reconciliation depicts coevolution

(Some) Macro Evolutionary Events

Speciation ($\mathbb{S}$), Duplication ($\mathbb{D}$), Transfer ($\mathbb{T}$), and Loss ($\mathbb{L}$)



The Most Parsimonious Reconciliation problem

- Input: **costs** for each event ($\mathbb{D}$, $\mathbb{T}$, $\mathbb{L}$, $\mathbb{S}$) and gene / species **trees**
- Output: a reconciliation that has a **Minimal cost** and is **Time consistent**

(May be more than one optimal reconciliation)

The Most Parsimonious Reconciliation problem

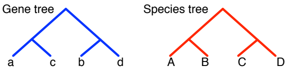
- Input: **costs** for each event ($\mathbb{D}$, $\mathbb{T}$, $\mathbb{L}$, $\mathbb{S}$) and gene / species **trees**
- Output: a reconciliation that has a **Minimal cost** and is **Time consistent**

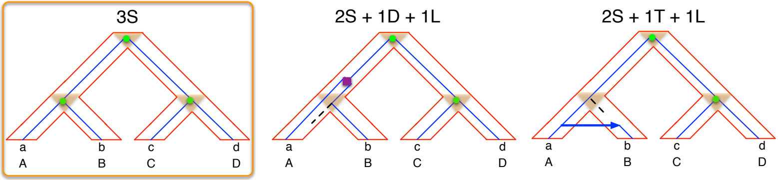
(May be more than one optimal reconciliation)

The Most Parsimonious Reconciliation problem

- Input: **costs** for each event ($\mathbb{D}$, $\mathbb{T}$, $\mathbb{L}$, $\mathbb{S}$) and gene / species **trees**
- Output: a reconciliation that has a **Minimal cost** and is **Time consistent**

(May be more than one optimal reconciliation)

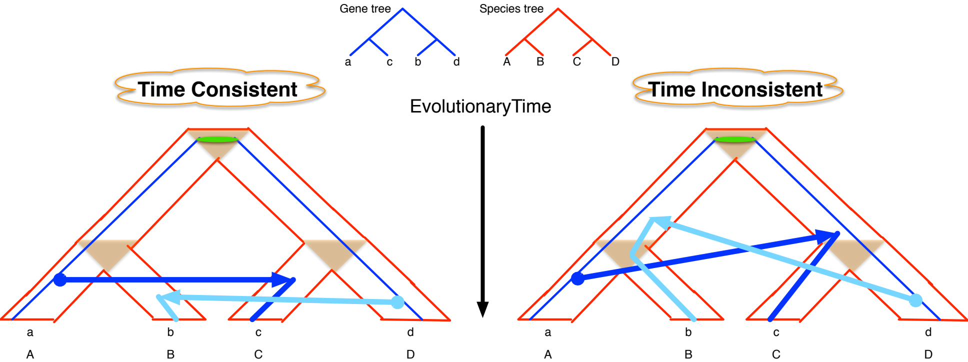




The Most Parsimonious Reconciliation problem

- Input: **costs** for each event ($\mathbb{D}$, $\mathbb{T}$, $\mathbb{L}$, $\mathbb{S}$) and gene / species **trees**
- Output: a reconciliation that has a **Minimal cost** and is **Time consistent**

(May be more than one optimal reconciliation)



Previous approaches & models

Species Graph [Gorecki]

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

Inconvenients of 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]
- Don't guarante optimality!!!

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.

Inconvenients of 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]
- Don't guarante optimality!!!

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

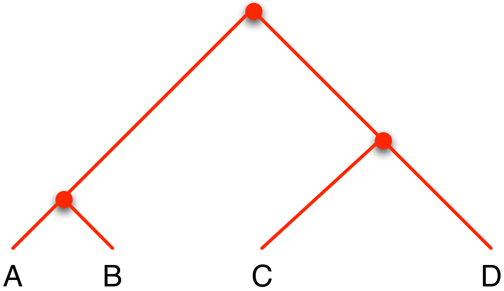
- Based on a small subdivision S' of S
- Fast: runs in time $O(|S'| \cdot |G|)$
- Previous algorithms in $O(|S|^4 \cdot |G|^4)$ and $O(|S'|^3 \cdot |G|)$

Experimental results for the relevance of parsimony

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

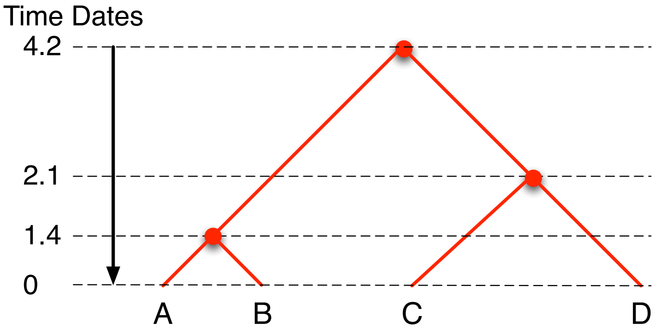
An Efficient Reconciliation Model

- Date species tree [LARTILLOT ET AL. 2004, AKERBORG ET AL. 2008, ETC.].
- Discretize evolutionary time into slices.
- Allow transfers within the same time slice.



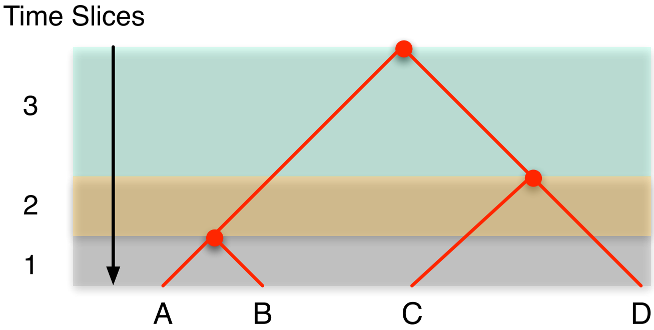
An Efficient Reconciliation Model

- Date species tree [LARTILLOT ET AL. 2004, AKERBORG ET AL. 2008, ETC.].
- Discretize evolutionary time into slices.
- Allow transfers within the same time slice.



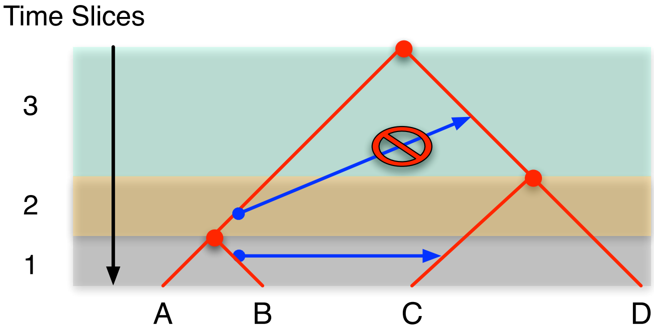
An Efficient Reconciliation Model

- Date species tree [Lartillot et al. 2004, Akerborg et al. 2008, etc.].
- Discretize evolutionary time into slices.
- Allow transfers within the same time slice.



An Efficient Reconciliation Model

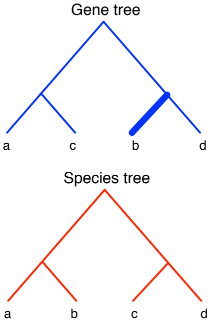
- Date species tree [LARTILLOT ET AL. 2004, AKERBORG ET AL. 2008, ETC.].
- Discretize evolutionary time into slices.
- Allow transfers within the same time slice.

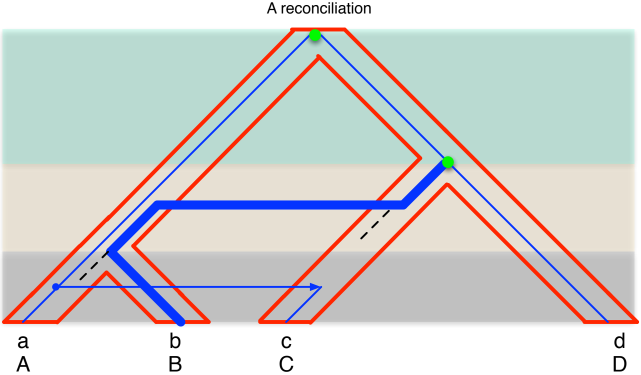


An Efficient Reconciliation Model

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 $\mathbb{S}$, $\mathbb{D}$, $\mathbb{T}$, and $\mathbb{L}$ events

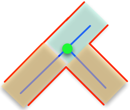




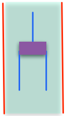
An Efficient Reconciliation Model

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

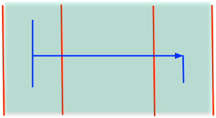
Speciation ($\mathbb{S}$)



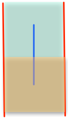
Duplication ($\mathbb{D}$)



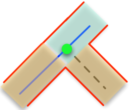
Transfer ($\mathbb{T}$)



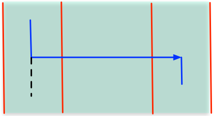
No event ($\varnothing$)



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)
- Gene Tree + Real Reconciliation generated (with rates $\mathbb{L}_R$, $\mathbb{T}_R$ and $\mathbb{D}_R$)
- Based on realistic loss rates [CSUROS AND MIKLOS]
- Gene trees have between 59 and 93 leaves

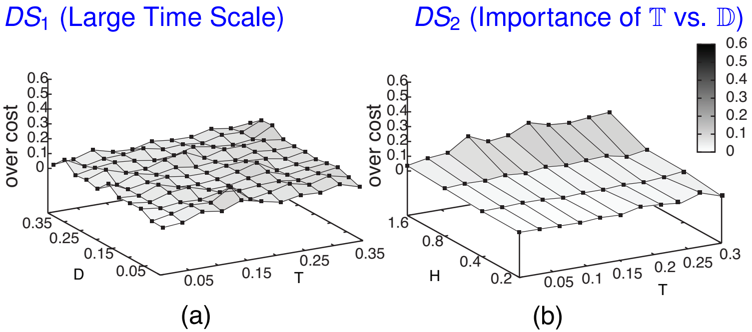
DS_1: "Simulate" a relatively Large Time Scale (archaean or bacterial phylum)

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

DS_2: "Simulate" Different Time Scales and Vary the Importance of $\mathbb{T}$ vs. $\mathbb{D}$

- 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.
- $8,800\ G = (20\ G) \times (10\ S) \times (4 \times 11\ \textit{rate pairs})$

Efficiency of parsimony according to costs

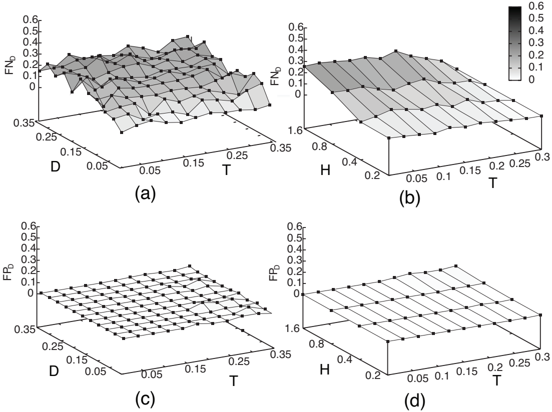


Over cost of the real scenario w.r.t. MPR

- 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 **Great!**

Accuracy of parsimony to retrieve $\mathbb{D}$ events

DS_1 (Large Time Scale) DS_2 (Importance of $\mathbb{T}$ vs. $\mathbb{D}$)



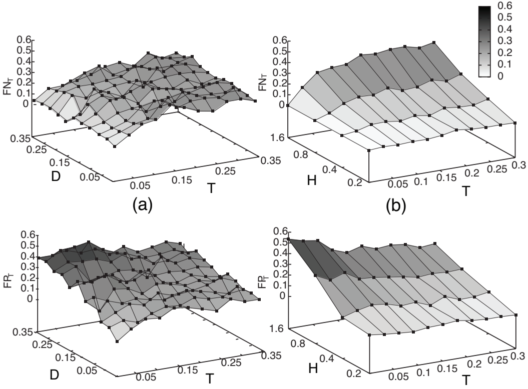
False Negatives / Positives: Node of G + Branche in S'

- Reasonnably few forgotten duplications (homoplasy and several MPRs?)
- *Very* few False Positives **Not bad!**

Accuracy of parsimony to retrieve $\mathbb{T}$ events

DS$_1$ (Large Time Scale)

DS$_2$ (Importance of $\mathbb{T}$ vs. $\mathbb{D}$)



(a) (b)

(c) (d)

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

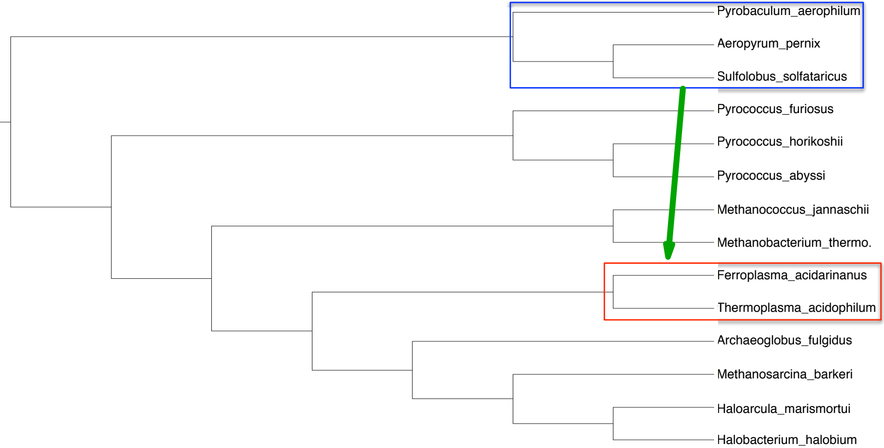
- Dated species tree: 14 archaeal (53 ribosomal proteins)
  3 dates for (Ferroplasma A., Thermoplasma A.) clade. [TIMETREE]
- Gene tree: ribosomal proteins
  2 roots. [MATTE ET AL. 2002; TOFIGH ET AL. 2010])
- 6 cases

Our approach vs. *Tofigh et al.* (May propose Time Inconsistent transfer)

- Our approach: 5 transfers + 3 losses
- *Tofigh et al.:* 5 transfers / duplications (Losses = “a posteriori”)

What is the relevancy of these transfers?

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

Conclusion

Our Most Parsimonious Reconciliation algorithm

- Proposes Time-Consistent transfers;
- Directly account for losses (discriminate among different scenarios).
- Much faster (cpu / complexity) than previous ones [Gorbunov et al. 09].
- Program available: **www.lirmm.fr/phylariane/**

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. [Doyon et al. 2009]
- Links between MPR and ML reconciliations
- Polytomous trees (low supported clades) [Vernot et al. 2008]

Conclusion

Our Most Parsimonious Reconciliation algorithm

- Proposes Time-Consistent transfers;
- Directly account for losses (discriminate among different scenarios).
- Much faster (cpu / complexity) than previous ones [GORBUNOV ET AL. 09].
- Program available: **www.lirmm.fr/phylariane/**

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).

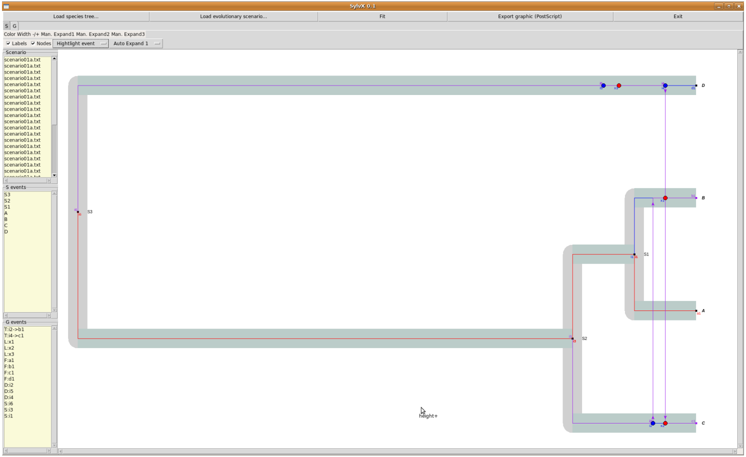
What next?

- Enumerating and counting MPRs. [DOYON ET AL. 2009]
- Links between MPR and ML reconciliations
- Polytomous trees (low supported clades) [VERNOT ET AL. 2008]

Phyl-ARIANE project

SylvX - reconciliation viewer/editor

- Automatic computation and manual modifications of reconciliations
- Graphical operators & analyses: re-rooting, zooming, filtering events



Scenario
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt
scenario01a.txt

S events
S3
S2
S1
A
B
C
D

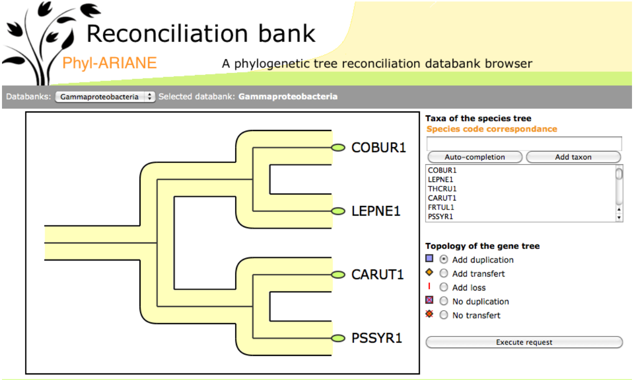
G events
T:i2->b1
T:i4->c1
L:x1
L:x2
L:x3
F:a1
F:b1
F:c1
F:d1
D:i2
D:i5
D:i4
S:i6
S:i3
S:i1

Phyl-ARIANE project

Reconciliation bank browser (gamma and beta proteobacteria)

Query a databank for

- evolutionary events (duplications, losses, and transfers);
- reconciliation patterns.



Reconciliation bank

Phyl-ARIANE

A phylogenetic tree reconciliation databank browser

Databanks: Gammaproteobacteria

Selected databank: **Gammaproteobacteria**

Taxa of the species tree

Species code correspondance

COBUR1
LEPNE1
THCRU1
CARUT1
FRTUL1
PSSYR1

Topology of the gene tree

● Add duplication
○ Add transfert
○ Add loss
○ No duplication
○ No transfert

Acknowledgment

Phyl-ARIANE

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

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

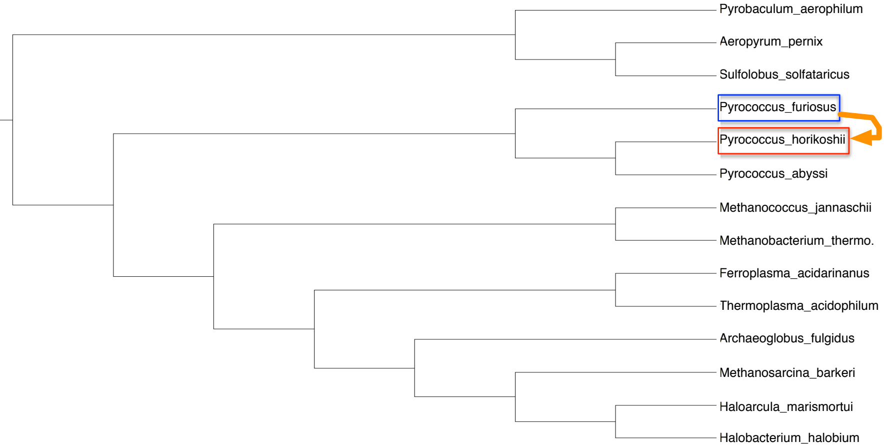
Thanks

- 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

Funding

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

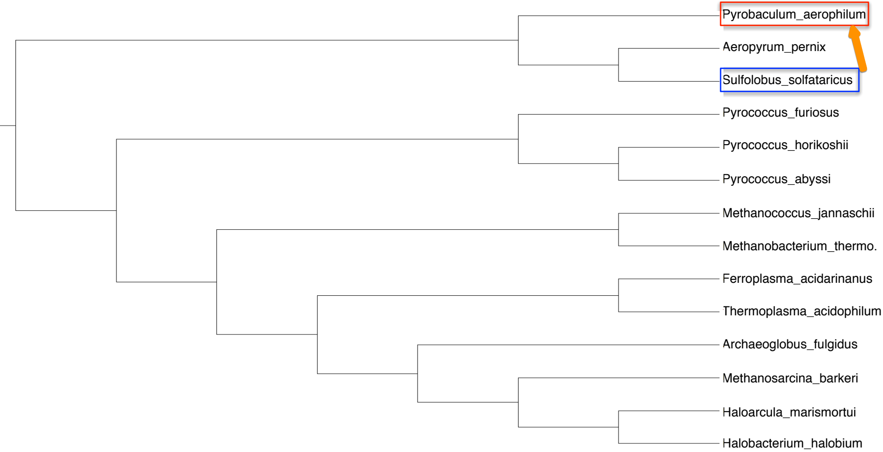
From Pyrococcus furiosus to Pyrococcus horikoshii



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

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

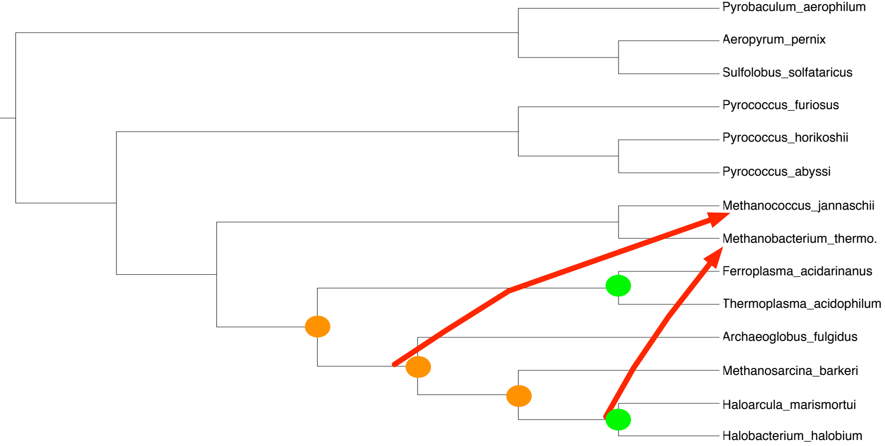
From Sulfolobus solfataricus to Pyrobaculum aerophilum



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

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

Reconciling trees with lack of resolution

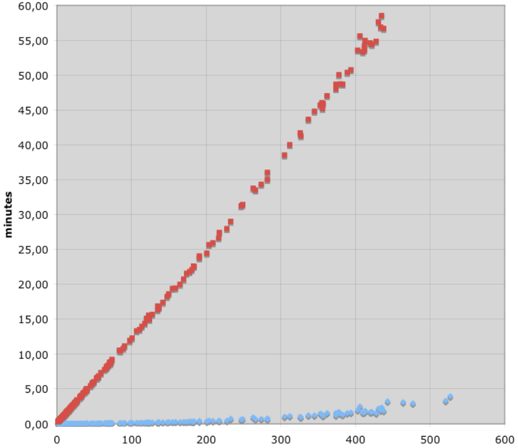


Artifactual transfers (probably)

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

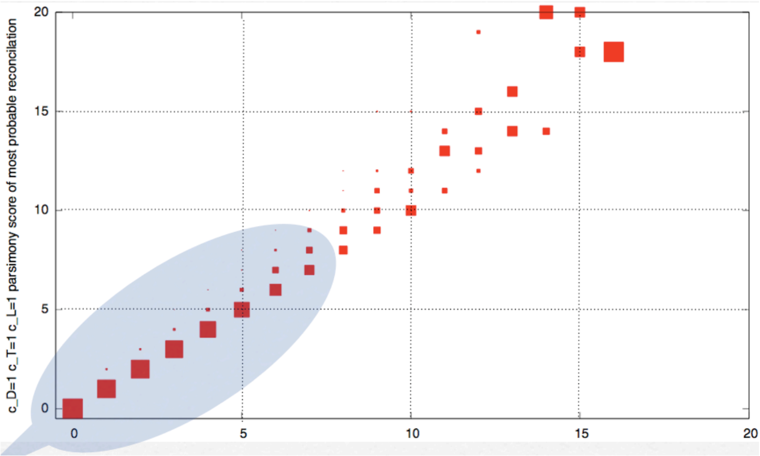
Running times

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



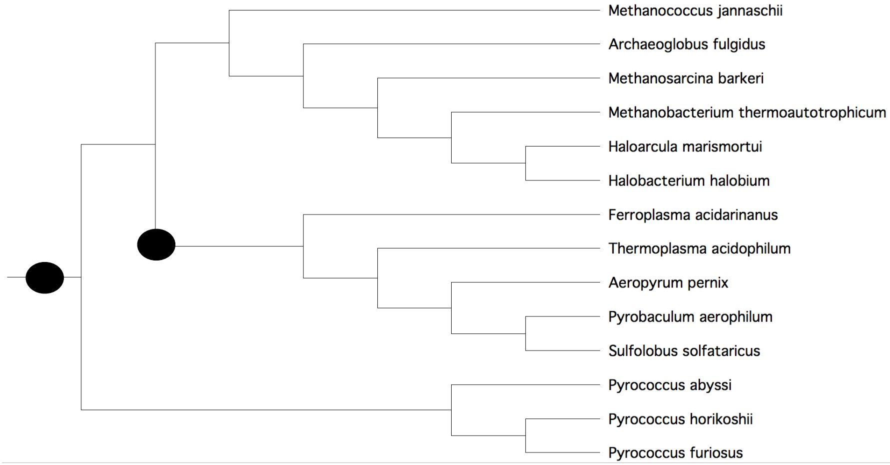
nb of leaves in *G*

Relationship between the MP and ML criteria



cost of MP reconciliation

Two roots for the (rpl12e) ribosomal proteins



Dynamic Programming Algorithm

Properties

- $\mathbb{SL}$ is an optimal scenario where one gene goes extinct after an $\mathbb{S}$ event (Idem for $\mathbb{TL}$ and $\mathbb{T}$)
- Any $\mathbb{TL}$ 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 $\{G_1, \ldots, G_n\}$, 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.

Whole Genome Duplications