An Efficient Algorithm for Gene/Species Trees Parsimonious Reconciliation with Losses, Duplications, and Transfers

(ANR-08-EMER-011 - PhyI-ARIANE)

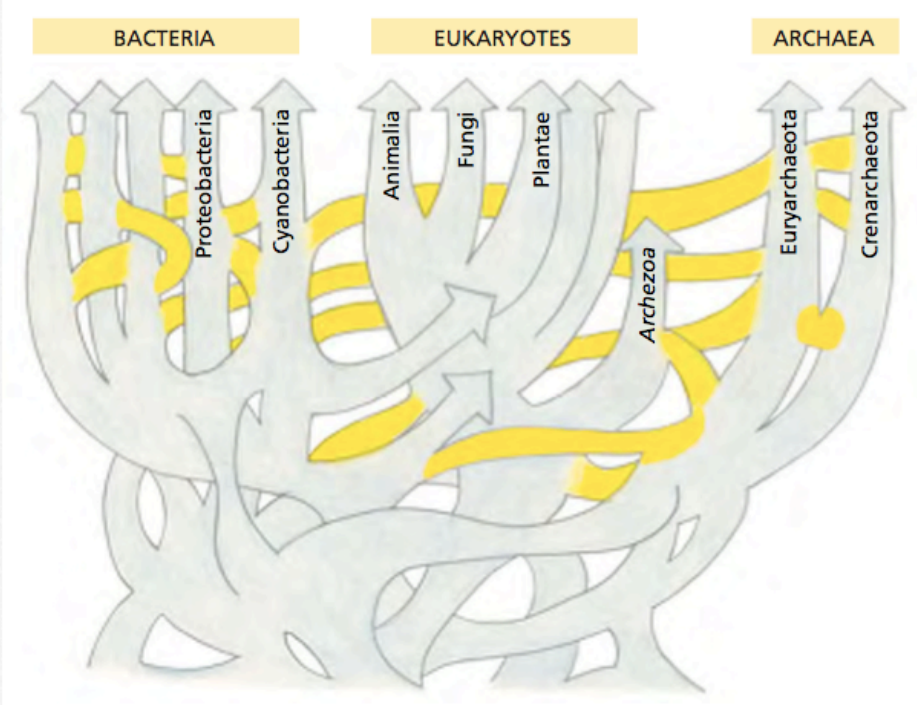
Jean-Philippe Doyon [1] Celine Scornavacca[2] K. Yu. Gorbunov [3]
Gergely J. Szöllősi[4] Vincent Ranwez[5] Vincent Berry[1]

1 - LIRMM, CNRS - Univ. Montpellier 2, France.
2 - Center for Bioinformatics (ZBIT), Tuebingen Univ., Germany.
3- Kharkevich IITP, Russian Academy of Sciences, Moscow.
4- LBBE, CNRS - Univ. Lyon 1, France.
5- ISEM, CNRS - Univ. Montpellier 2, France.

RECOMB Comparative Genomics, Ottawa
October 2010

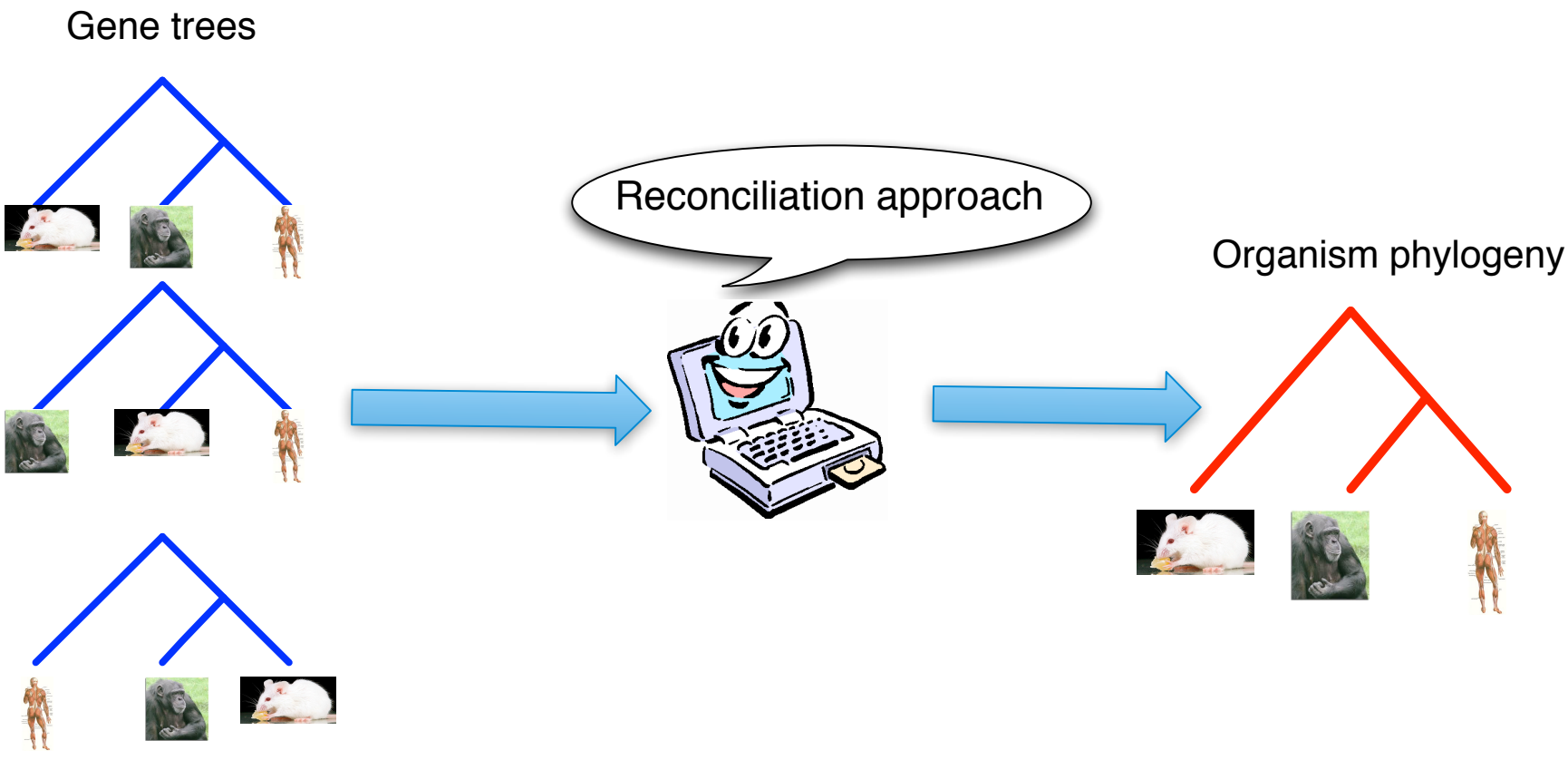
Concept of the Tree of Life

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



We assume that such a ToL exists

Inferring a Tree of Life by retrieving signals from gene trees



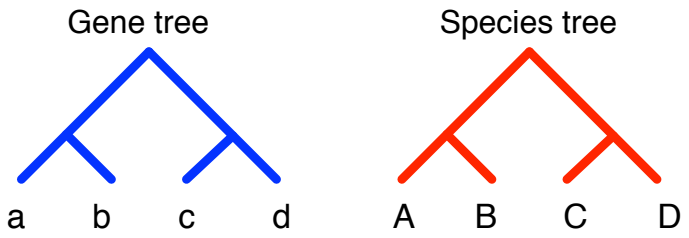
Reconciliation

- 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

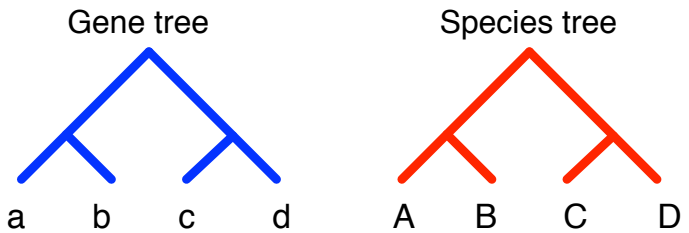
Speciation ($\mathbb{S}$), Duplication ($\mathbb{D}$), Transfer ($\mathbb{T}$), and Loss ($\mathbb{L}$)

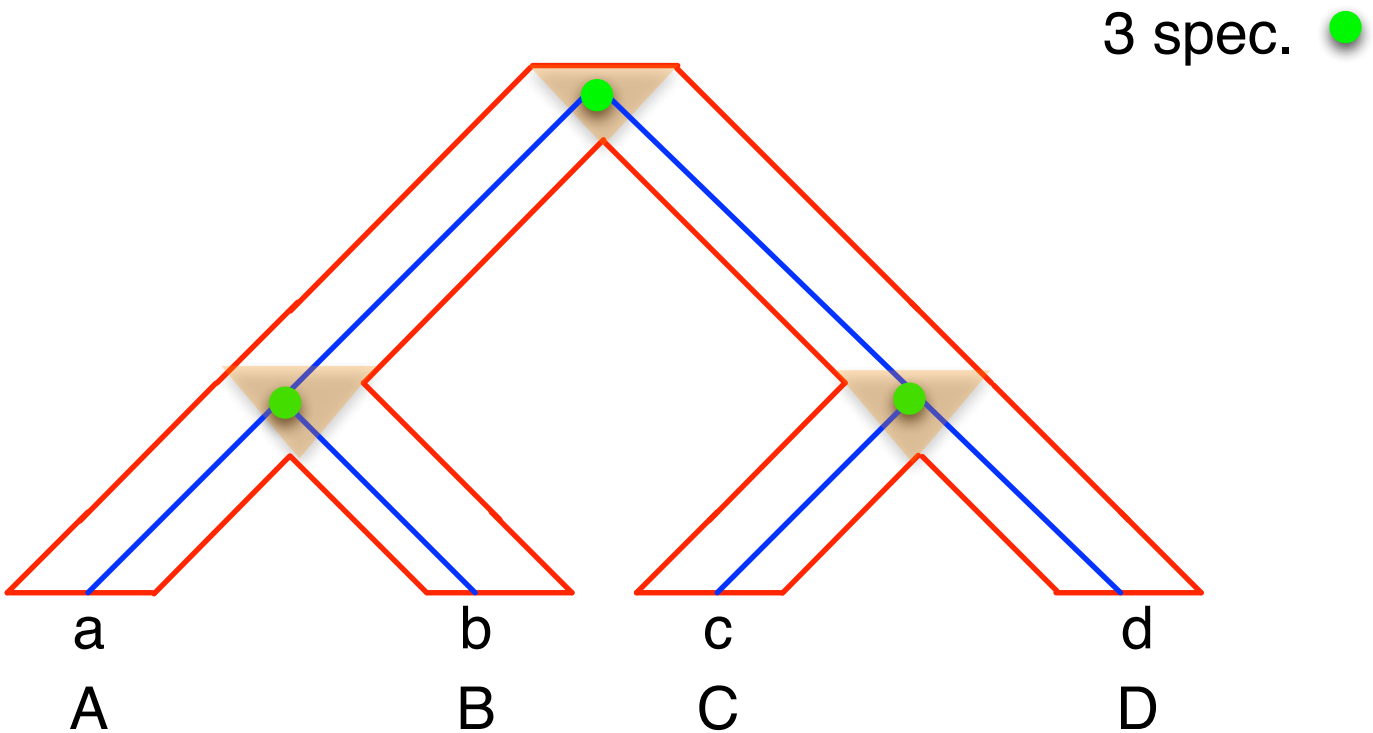


Reconciliation depicts coevolution

(Some) Macro Evolutionary Events

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

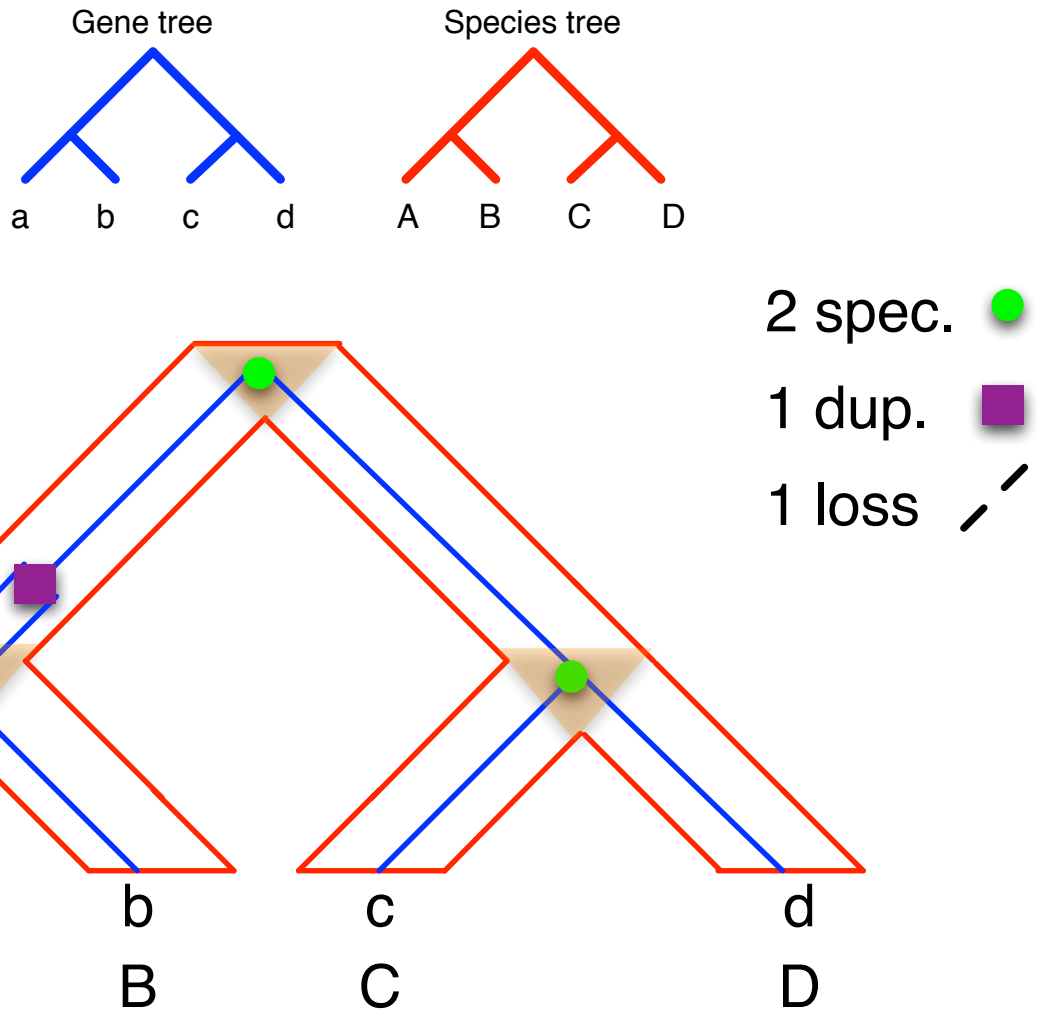




Reconciliation depicts coevolution

(Some) Macro Evolutionary Events

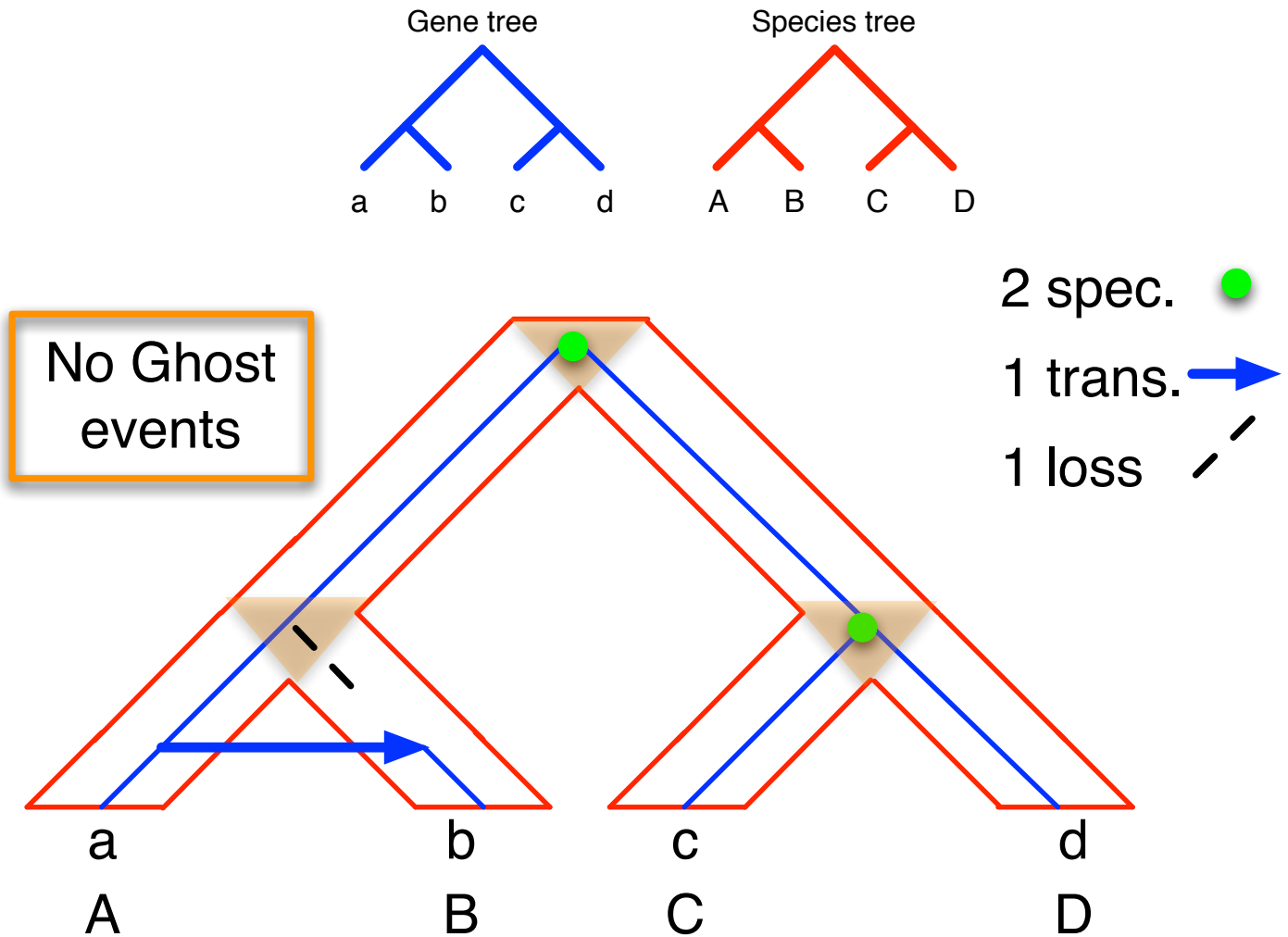
Speciation ($\mathbb{S}$), Duplication ($\mathbb{D}$), Transfer ($\mathbb{T}$), and Loss ($\mathbb{L}$)



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 MPR)

The Most Parsimonious Reconciliation problem

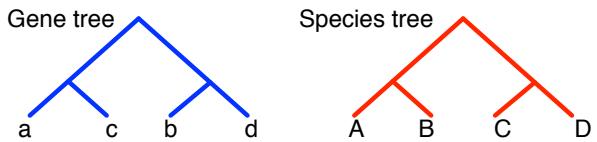
- Input: **costs** for each event ($\mathbb{D}$, $\mathbb{T}$, $\mathbb{L}$, $\mathbb{S}$) and gene / species **trees**
- Output: a reconciliation that has a **Minimal cost** and is **Time consistent**

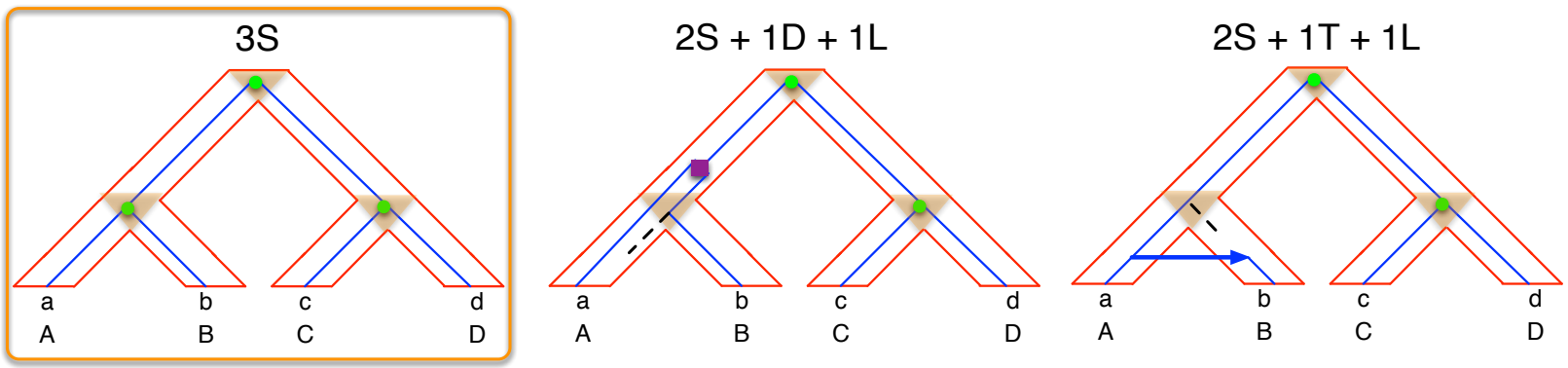
(May be more than one MPR)

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 MPR)

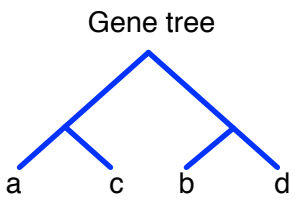


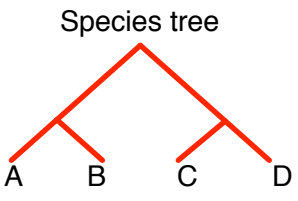


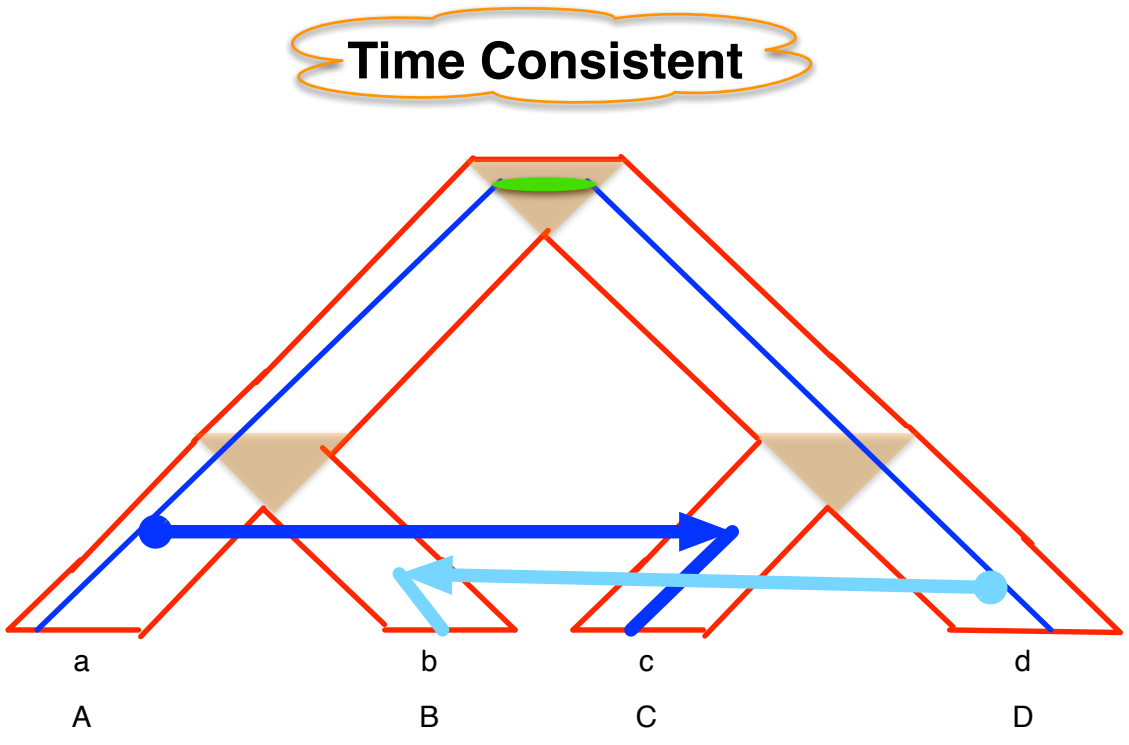
The Most Parsimonious Reconciliation problem

- 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 MPR)





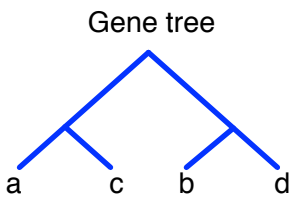


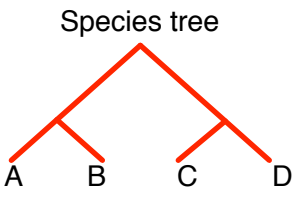
Local and Global Time Consistencies

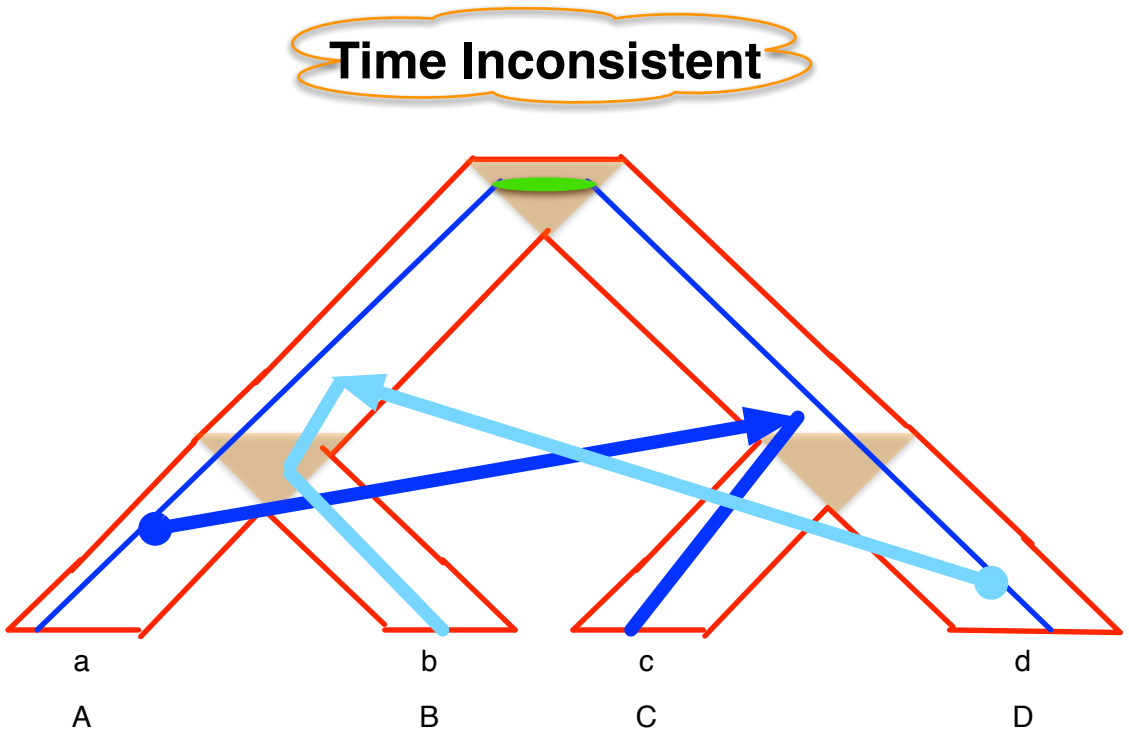
The Most Parsimonious Reconciliation problem

- 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 MPR)







Local and Global Time Consistencies

Previous approaches & models

Species Graph [Gorecki]

- Locations of (possible) transfers are defined in advance.
- A MPR is computed in $O(n^3 \cdot |G|)$ ($n =$ size of the given species graph).
- Computing a species graph that induces a MPR is NP-complete.

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.
- A MPR is computed in $O(n^3 \cdot |G|)$ ($n =$ size of the given species graph).
- Computing a species graph that induces a MPR is NP-complete.

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

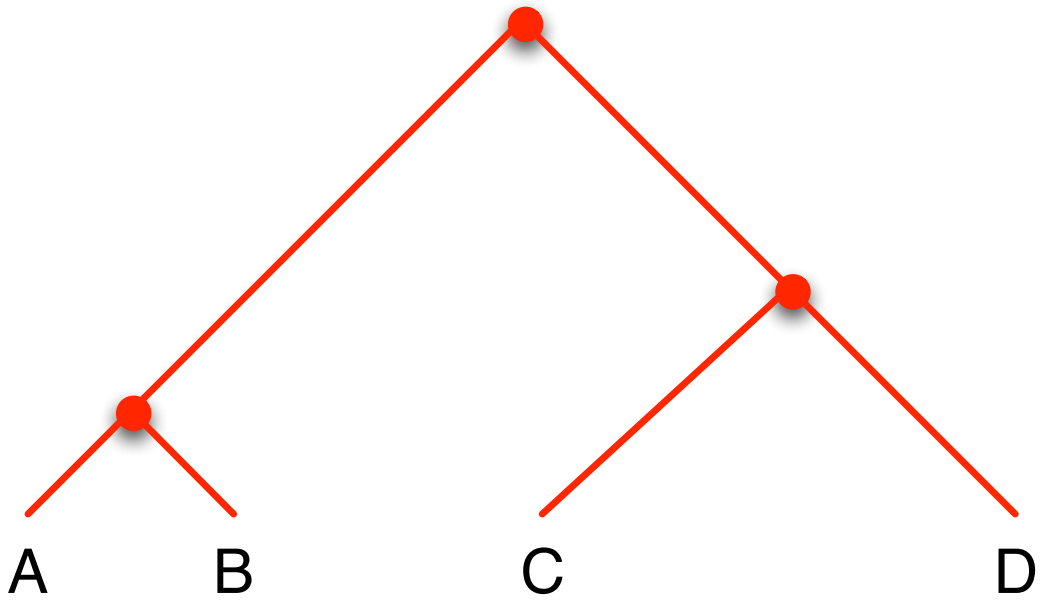
- Based on a small subdivision S' of S
- Fast: runs in time $O(|S'| \cdot |G|)$
- Previous algorithms in $O(|S|^4 \cdot |G|^4)$ and $O(|S'|^3 \cdot |G|)$

Experimental results for the relevance of parsimony

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

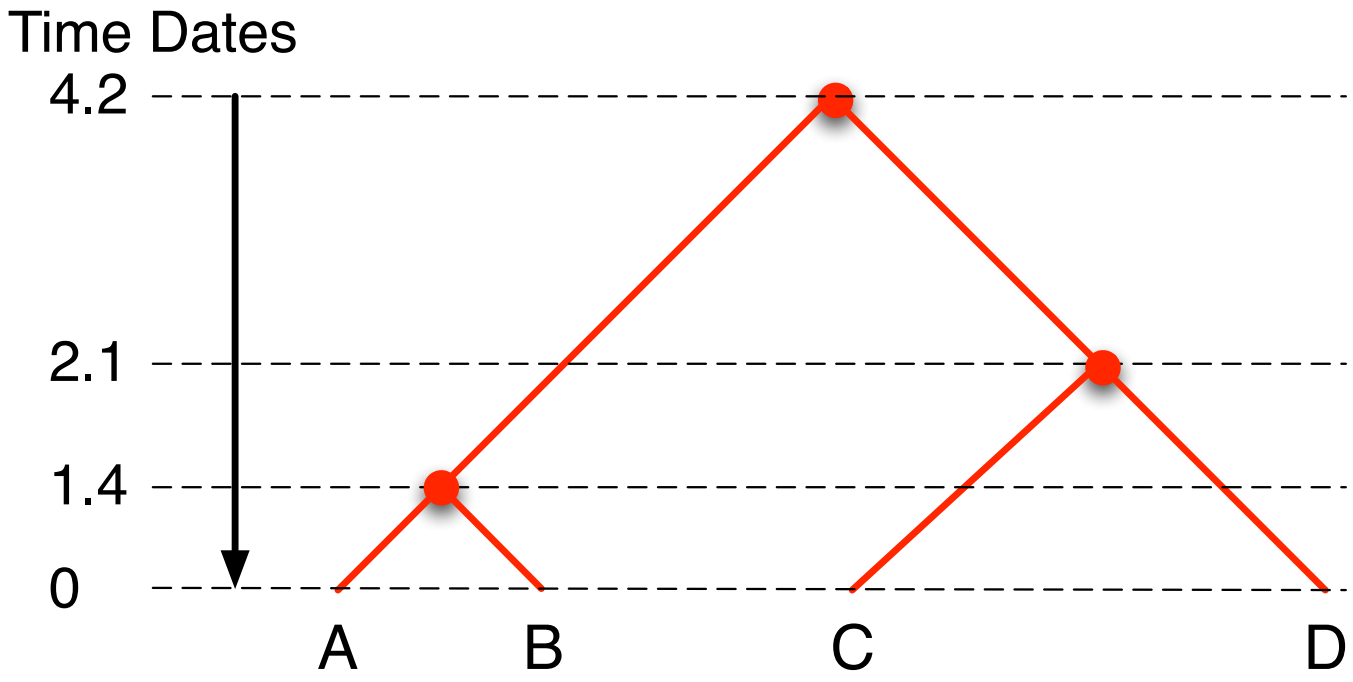
An Efficient Reconciliation Model

- Assign (relative) dates to species [LARTILLOT, 2004; AKERBORG, 2008].
- Discretize evolutionary time into slices: subdivision S' of S.
- Allow transfers within the same time slice.



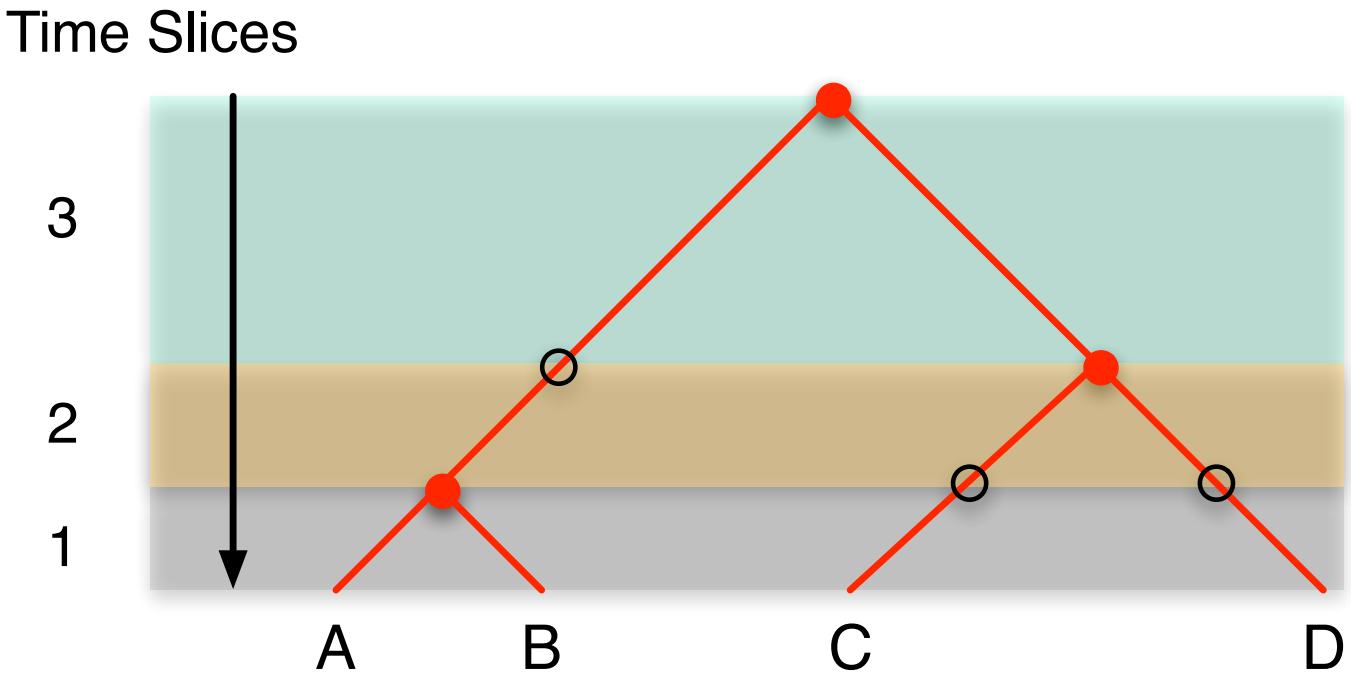
An Efficient Reconciliation Model

- Assign (relative) dates to species [LARTILLOT, 2004; AKERBORG, 2008].
- Discretize evolutionary time into slices: subdivision S' of S.
- Allow transfers within the same time slice.



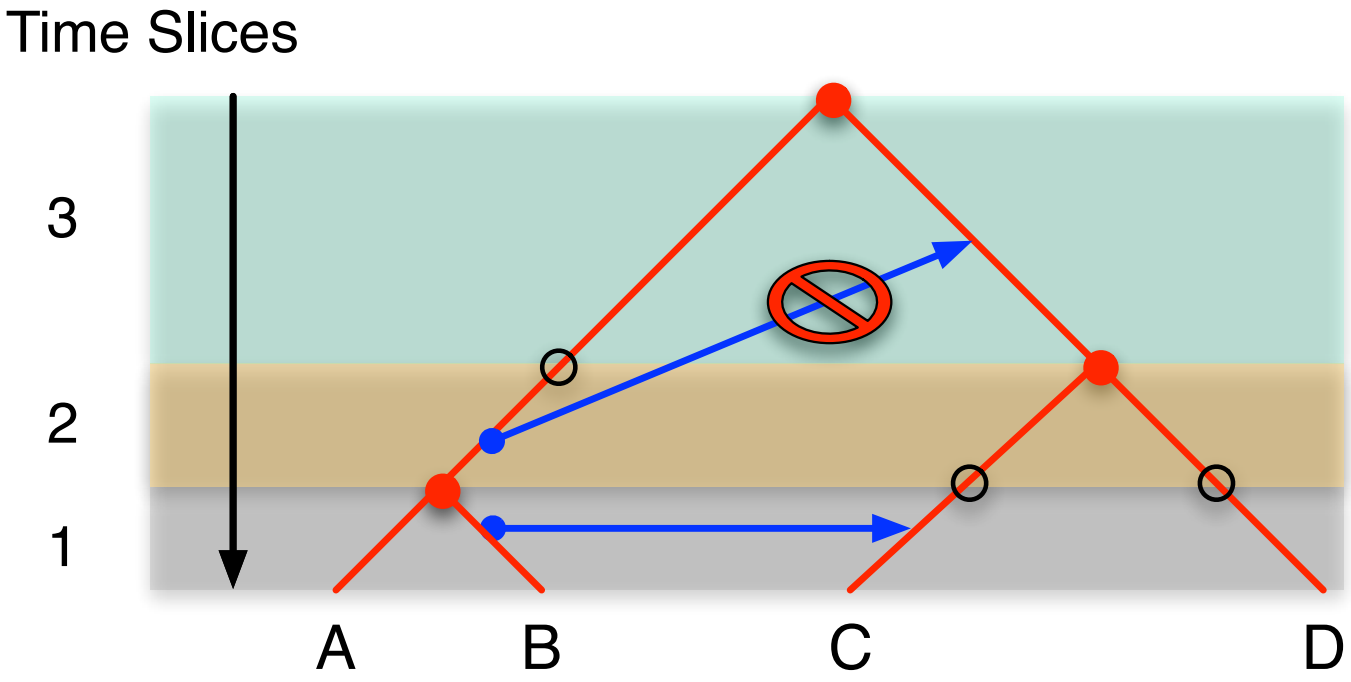
An Efficient Reconciliation Model

- Assign (relative) dates to species [LARTILLOT, 2004; AKERBORG, 2008].
- Discretize evolutionary time into slices: subdivision S' of S.
- Allow transfers within the same time slice.



An Efficient Reconciliation Model

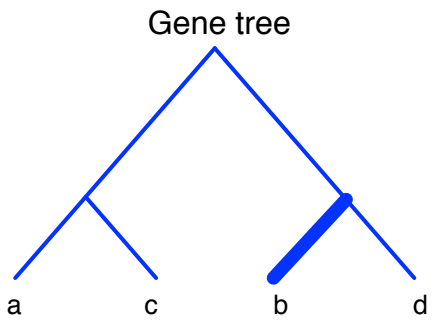
- Assign (relative) dates to species [LARTILLOT, 2004; AKERBORG, 2008].
- Discretize evolutionary time into slices: subdivision S' of S.
- Allow transfers within the same time slice.

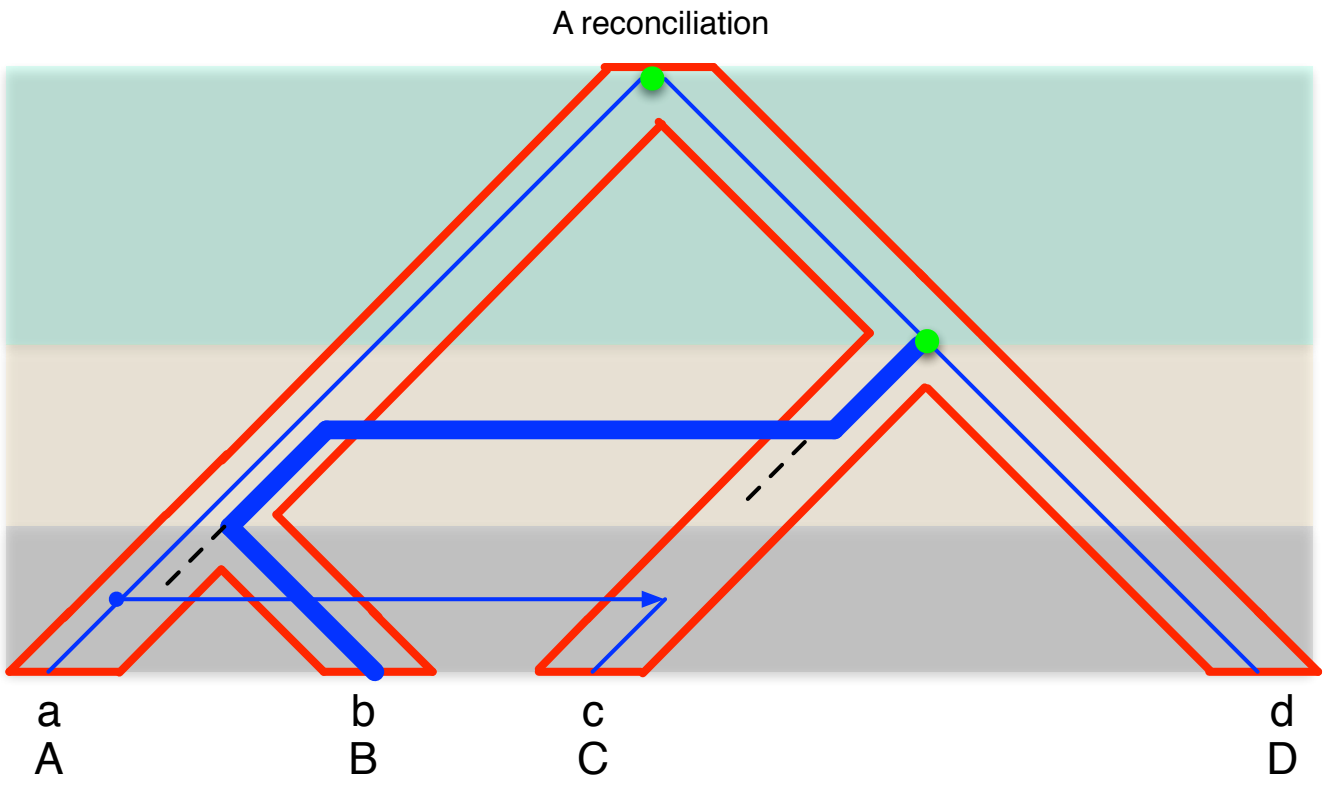


An Efficient Reconciliation Model

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

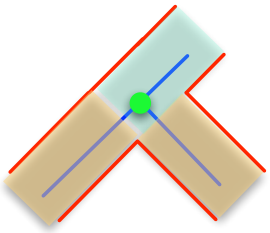




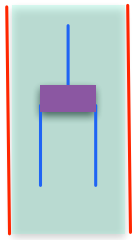
An Efficient Reconciliation Model

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

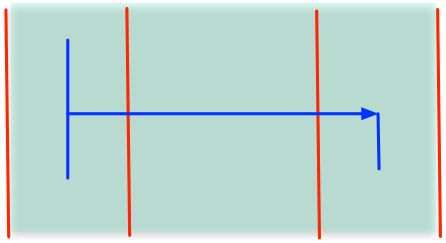
Speciation ($\mathbb{S}$)



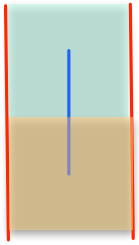
Duplication ($\mathbb{D}$)



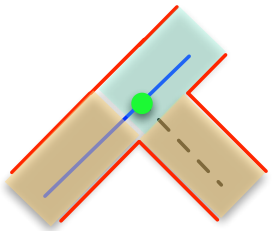
Transfer ($\mathbb{T}$)



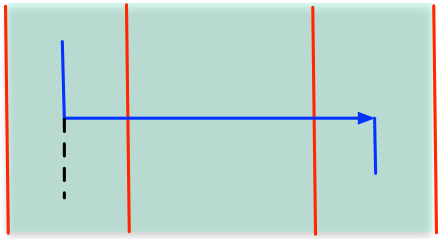
No event ($\varnothing$)



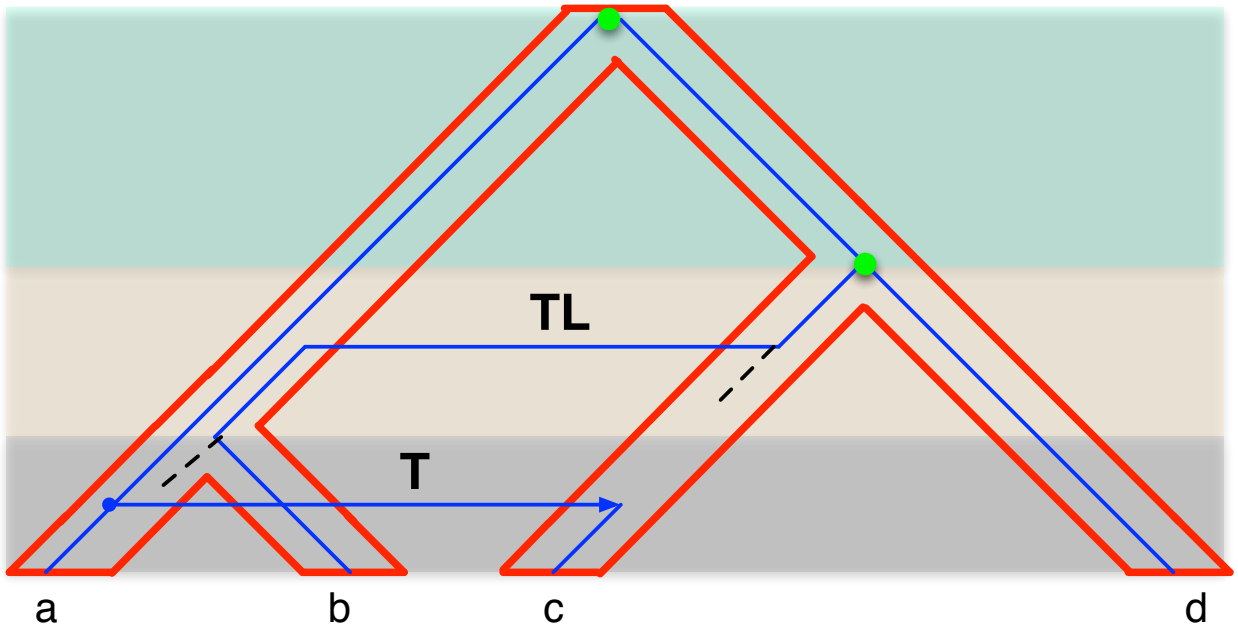
Speciation + Loss ($\mathbb{SL}$)



Transfer+Loss ($\mathbb{TL}$)



An Efficient Reconciliation Model



Properties

1. $\mathbb{SL}$ and $\mathbb{TL}$ are **parsimonious event associations**.
   (There is no need to examine scenarios where the gene lineage goes extinct.)
2. The **local and global consistencies** for $\mathbb{T}$ and $\mathbb{TL}$ events are ensured by the discretization of S.

Dynamic Programming Algorithm

Remark

The complexity of previous algorithms is hampered by the time to examine transfers between branches of S'

Property 4

Given an edge (u_p, u) of G, computing the optimal recipients for the branches at time t of S' can be factorized.

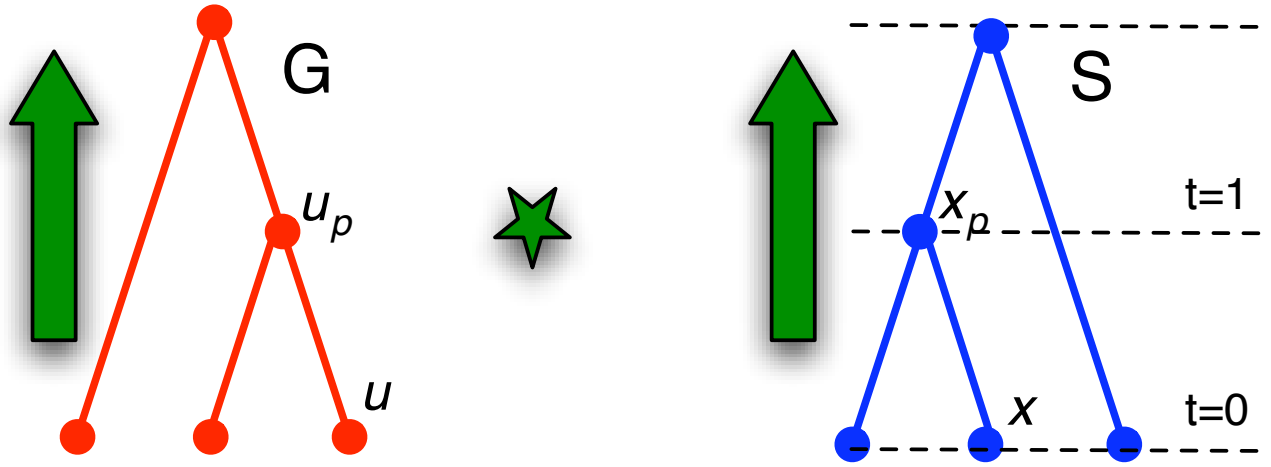
Property 3

For each edge of G, more than one TL event at time t is not parsimonious.

Dynamic Programming Algorithm

Principle

1: **for all** edge (u_p, u) of G following a bottom-up traversal **do**
2: **for all** time t of S' in backward time order **do**
3: Compute the first 2 optimal recipients for u_1-u_2 at time t (Optimization)
4: **for all** branch (x_p, x) of S' located at time t **do**
Compute the costs for $\mathbb{S}$, $\mathbb{D}$, $\mathbb{SL}$, $\varnothing$, and $\mathbb{T}$ events
5: Compute the first two optimal recipients for u at time t (Optimization)
6: **for all** branch (x_p, x) of S' located at time t **do**
Compute the costs for $\mathbb{TL}$ event



Dynamic Programming Algorithm

Theorem 1

Any Most Parsimonious scenario is considered by the reconciliation model.

Theorem 2

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

Same optimization applies to ML version of the DP algorithm [SZÖLLÖZI ET AL IN PREP.]

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)

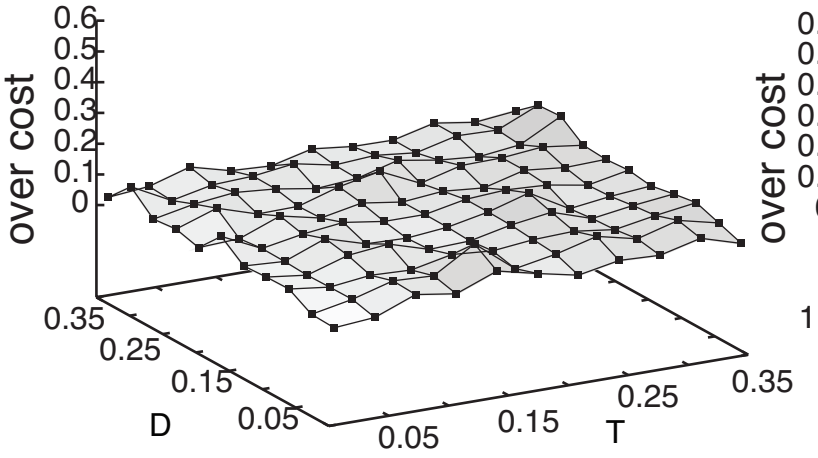
- Fixed rate $\mathbb{L}_R = 0.7$ and tree height $h = 1$
- 11 values for $\mathbb{T}_R$ and $\mathbb{D}_R$ in [0.01, 0.35]
- $6,050\ G = (10\ S) \times (11 \times 11\ \textit{rate pairs}) \times (5\ G)$

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

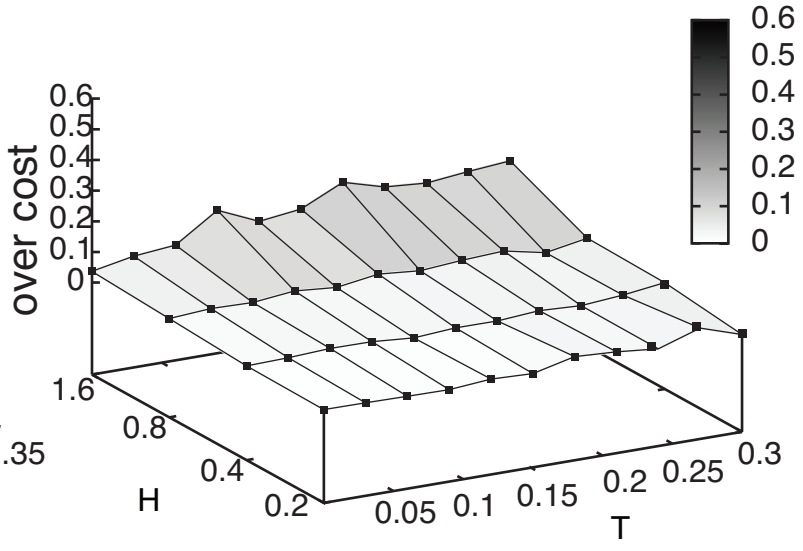
Efficiency of parsimony according to costs

DS_1 (Large Time Scale)



(a)

DS_2 (Importance of $\mathbb{T}$ vs. $\mathbb{D}$)



(b)

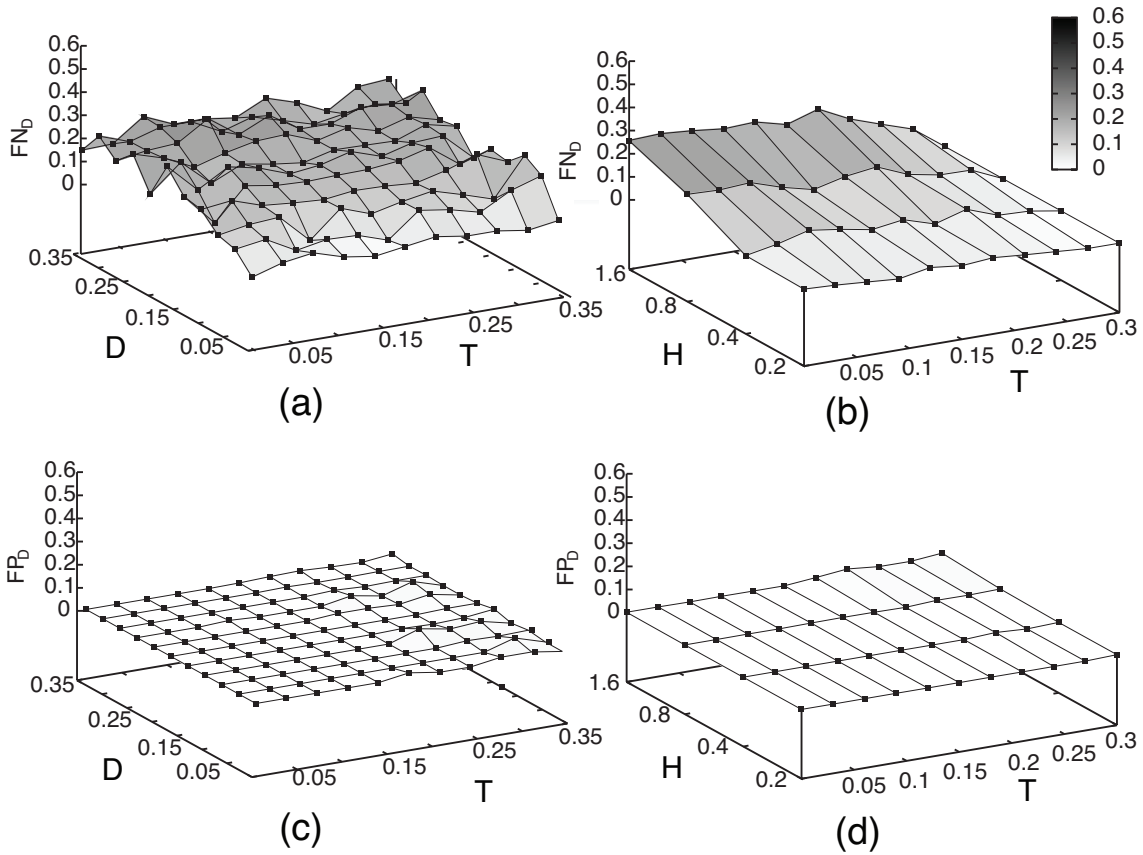
Over cost of the Real Scenario w.r.t. MPR

- Small for all $\mathbb{D}$ and $\mathbb{T}$ 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. $\mathbb{D}$)



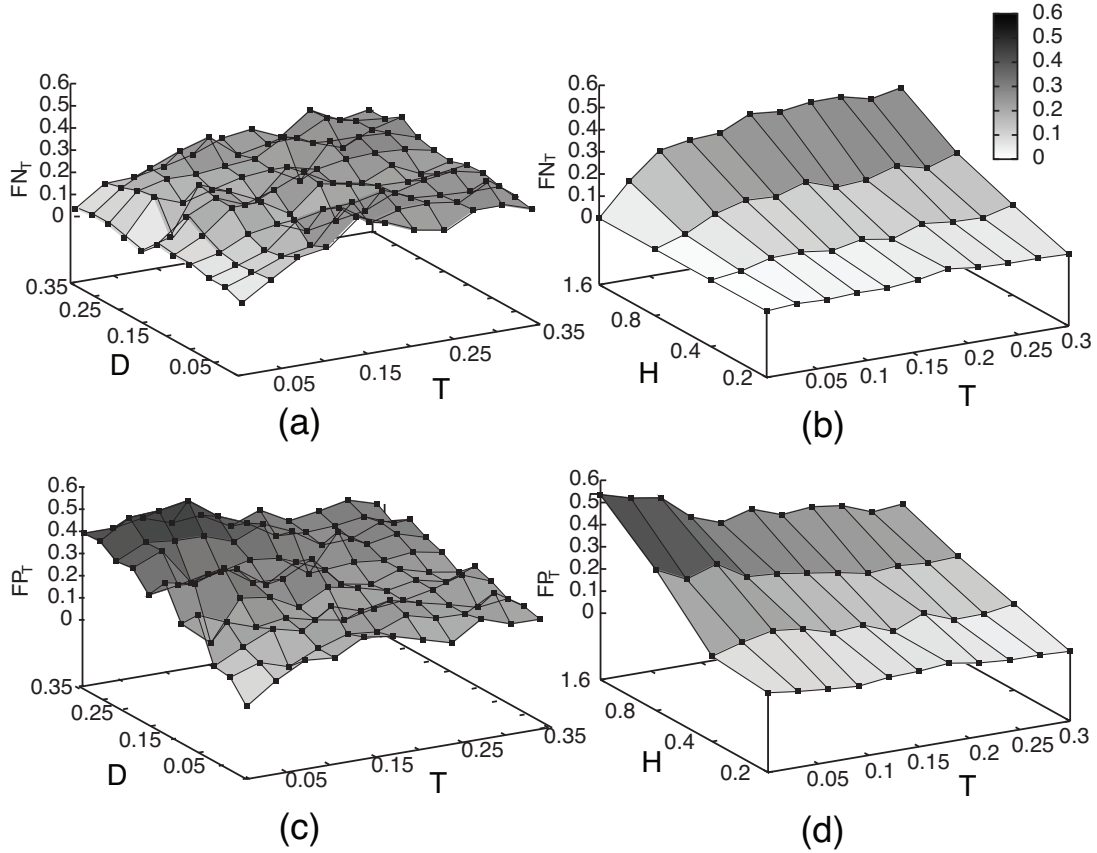
False Negatives / Positives: Node of G + Branche in S'

- 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. $\mathbb{D}$)



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... :(**

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]

Acknowledgment

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

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

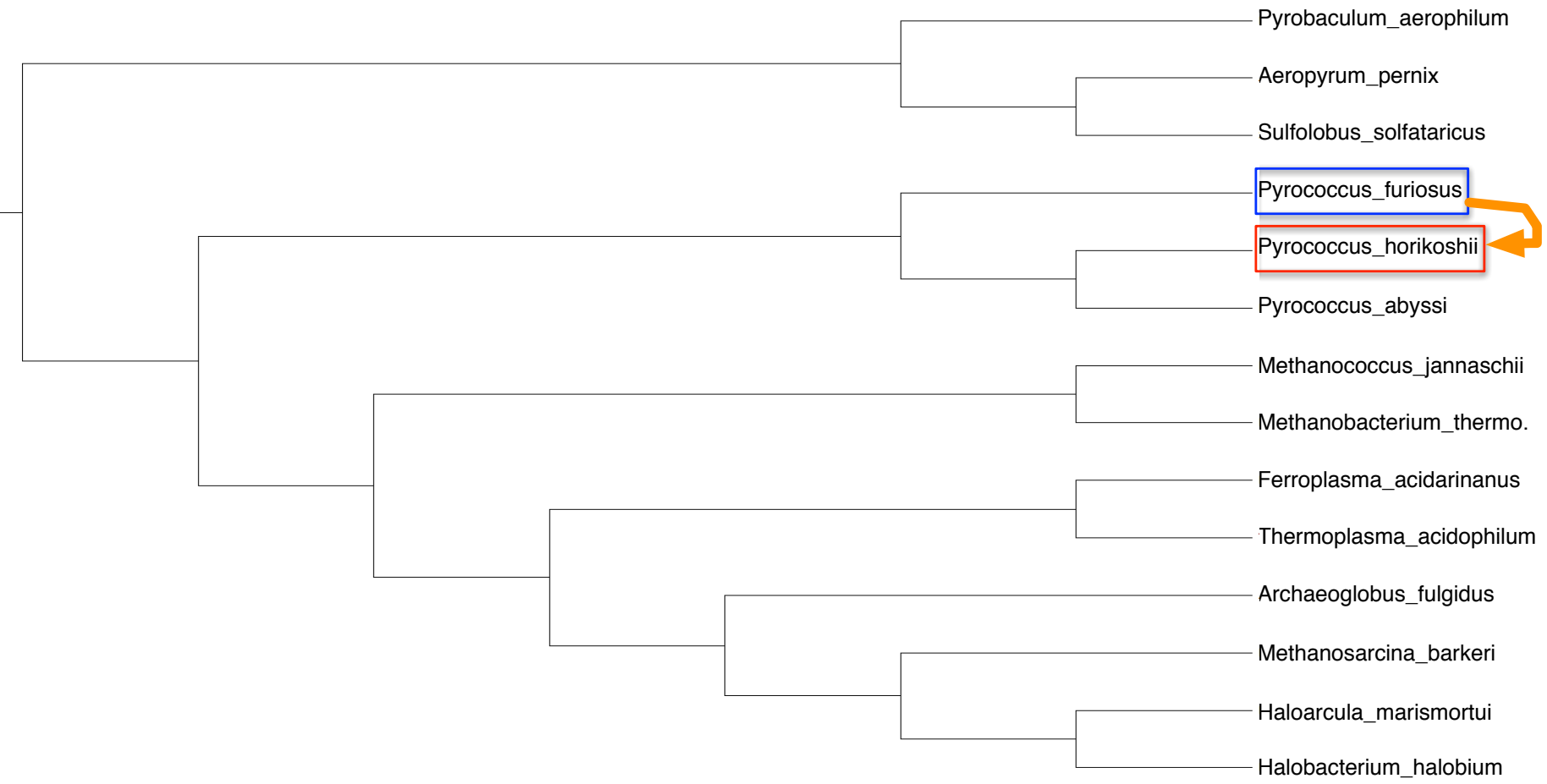
Thanks

- 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, ...

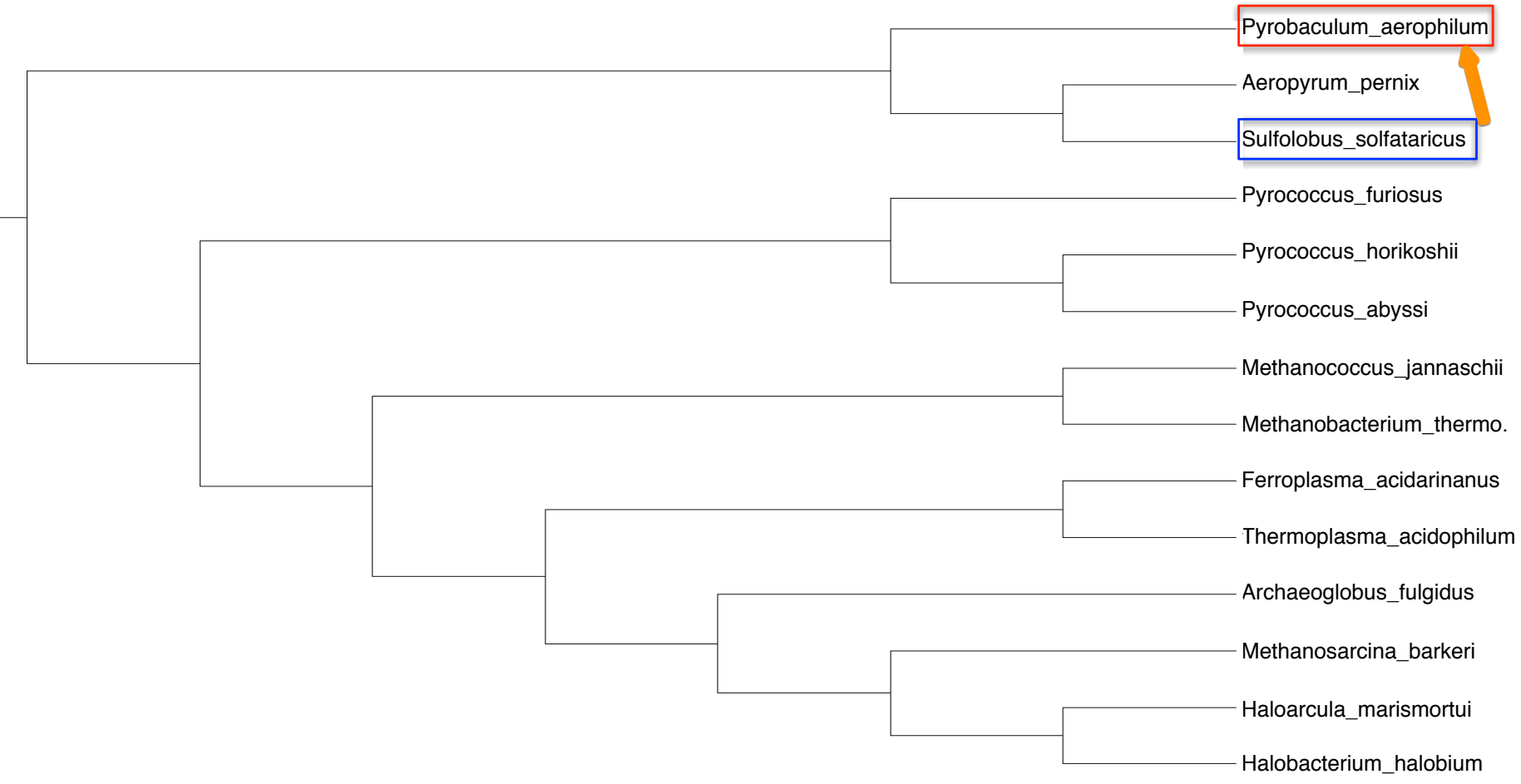
From Pyrococcus furiosus to Pyrococcus horikoshii



Seem to be correct (both roots of G)

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

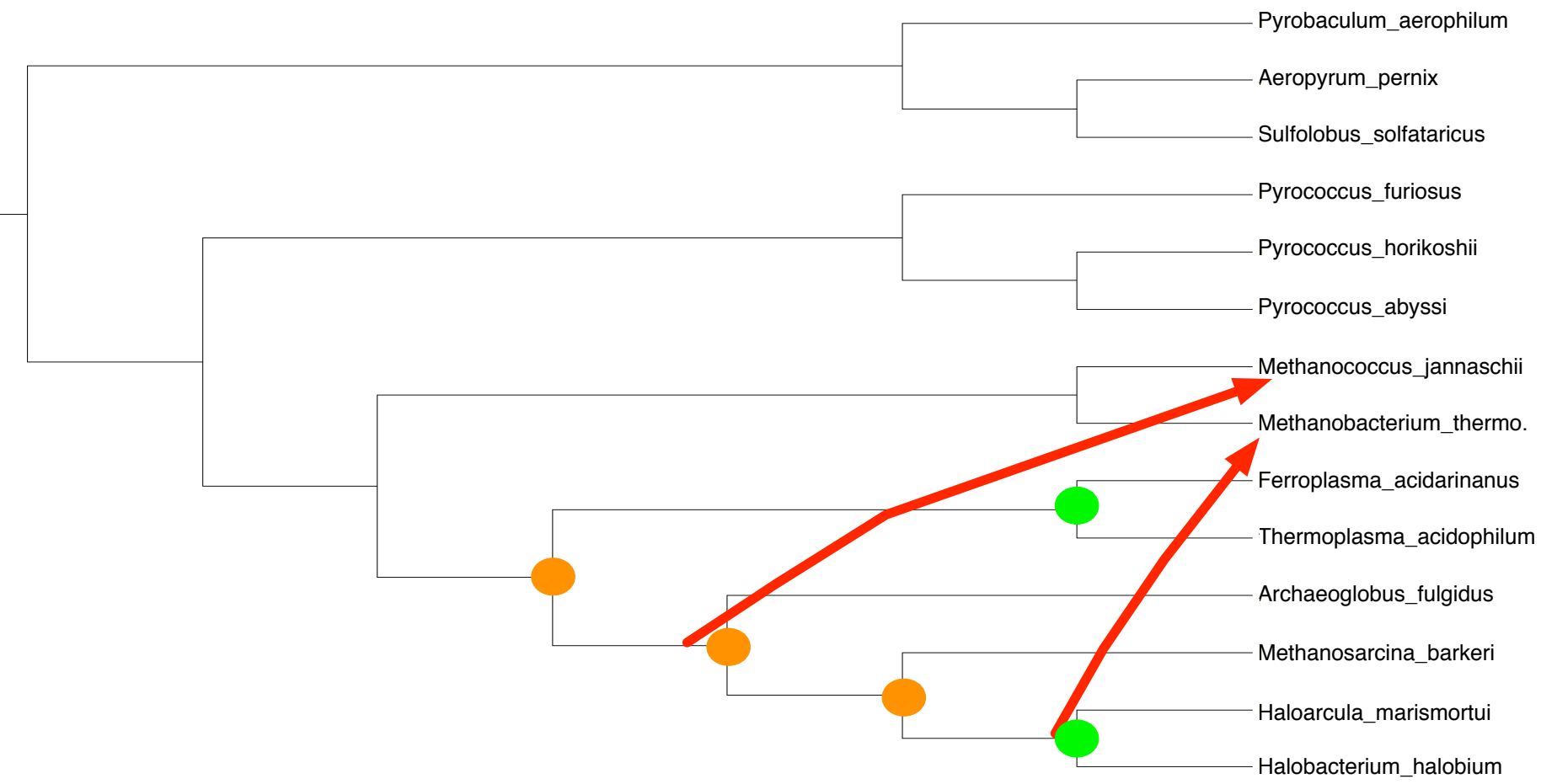
From Sulfolobus solfataricus to Pyrobaculum aerophilum



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

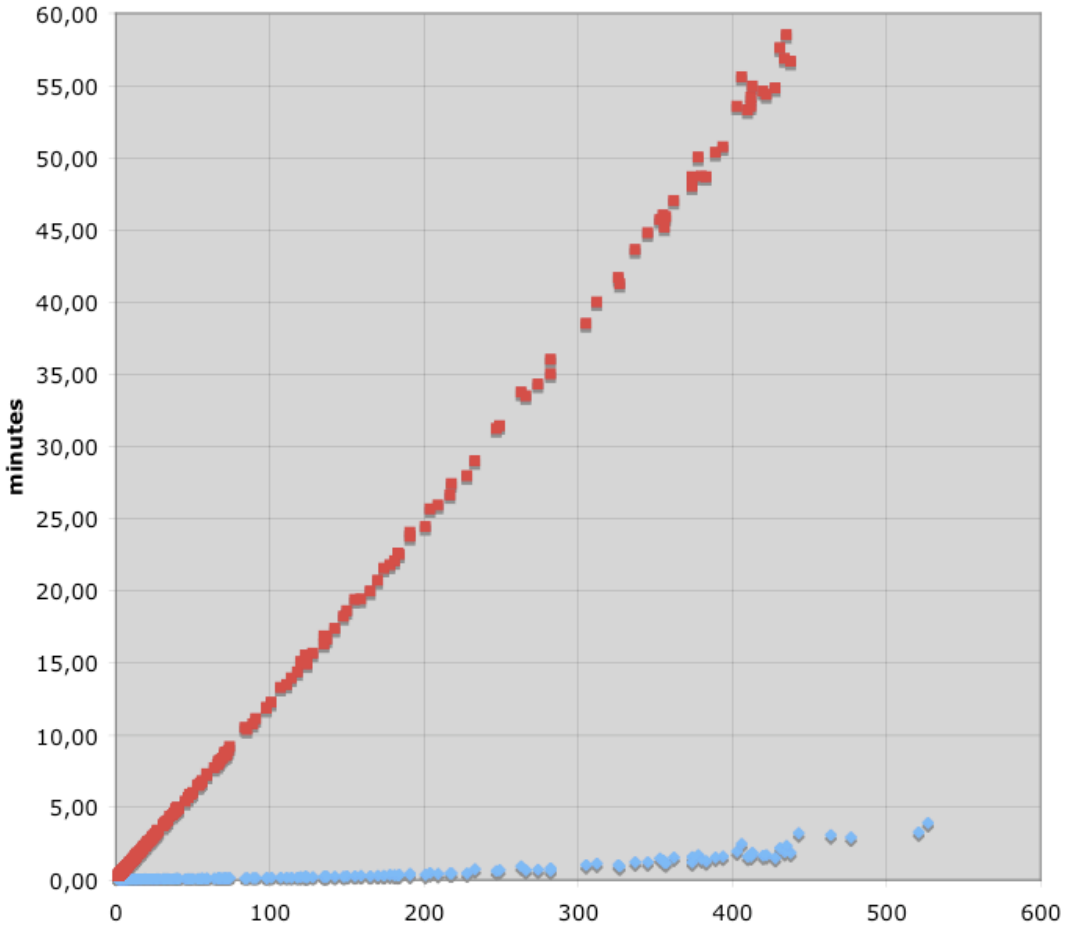


Artifactual transfers (probably)

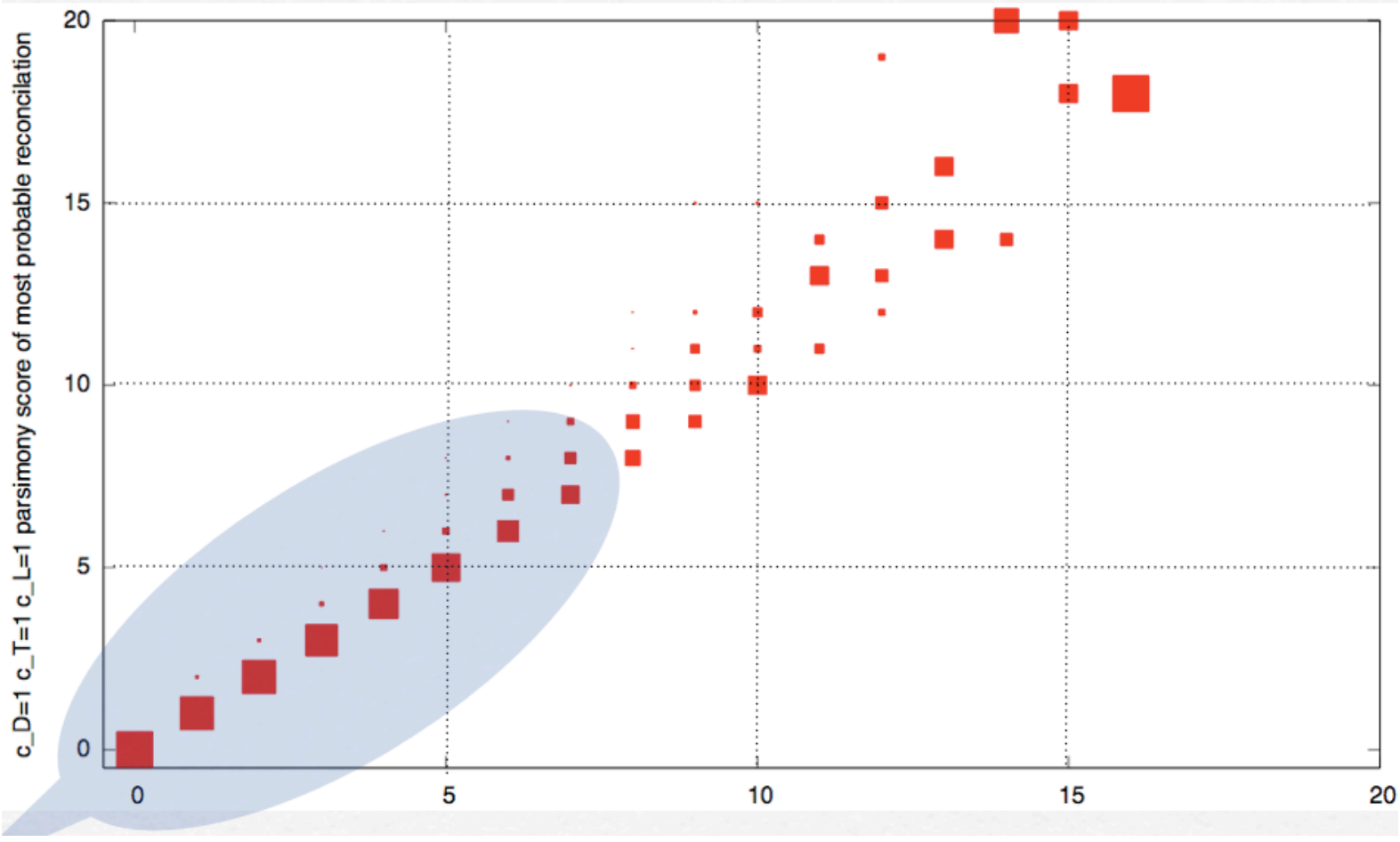
- Low bootstrap values in species and gene trees.
- Collapsing unsupported nodes erases discrepancies between trees.

Running times

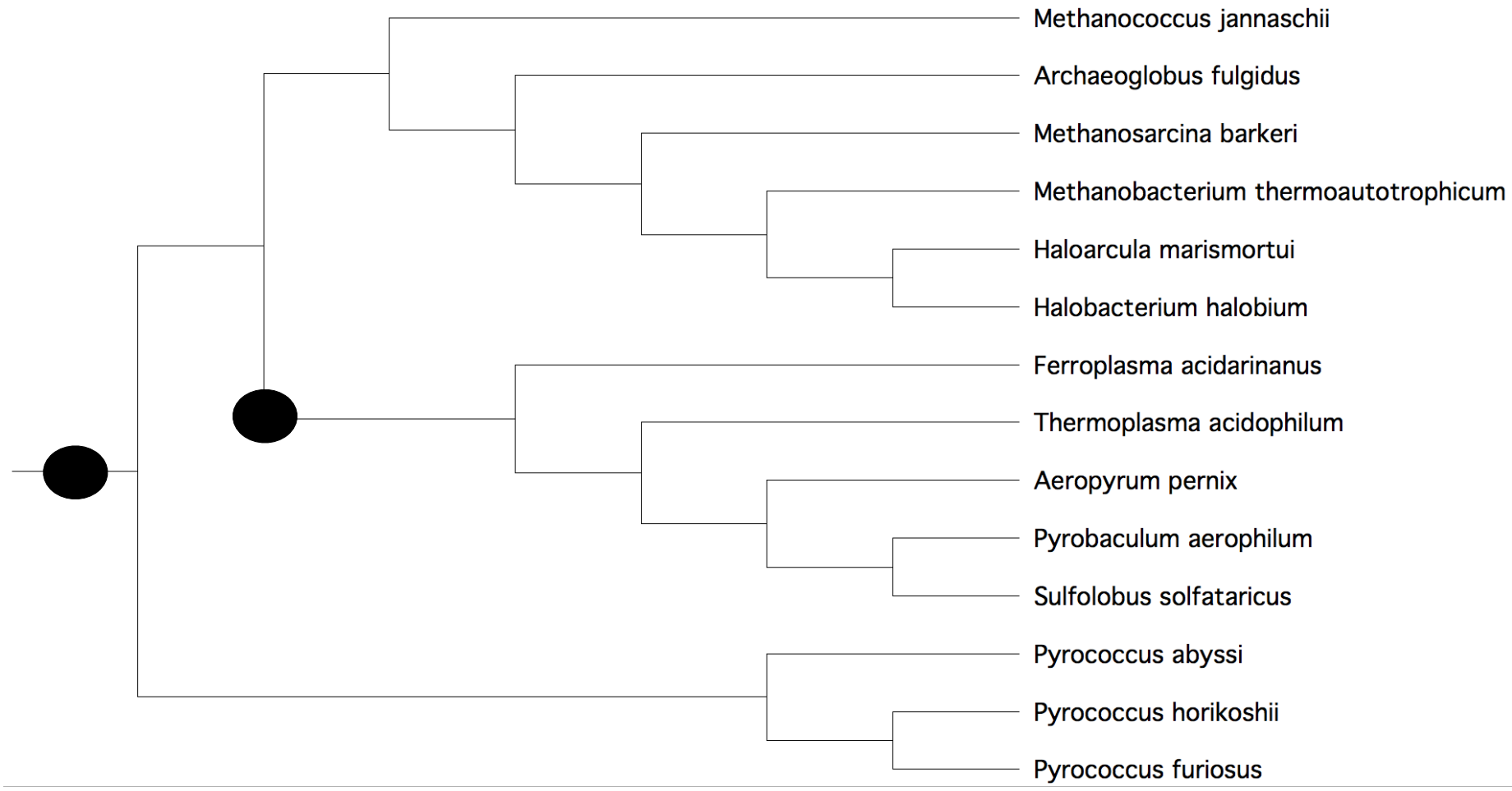
Comparison with an implementation of [Gorbunov et al 09]:
from dozens minutes to less than 2 sec (between 1.09s and 1.38s)



Relationship between the MP and ML criteria



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.]

Transfers among archaeal genomes

Input data

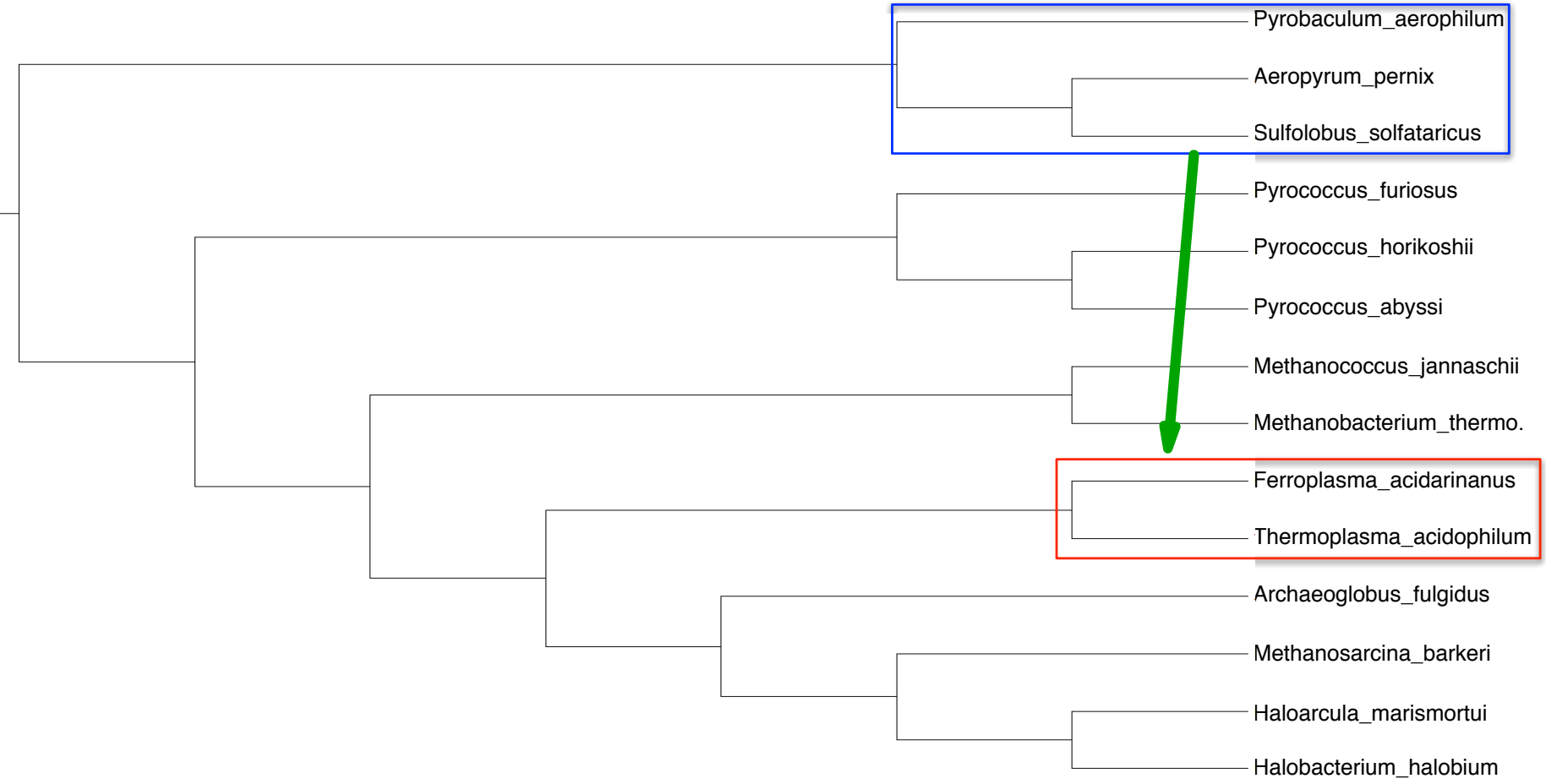
- Dated species tree: 14 archaeal (53 ribosomal proteins)
  3 dates for (Ferroplasma A., Thermoplasma A.) clade. [TIMETREE]
- Gene tree: ribosomal proteins
  2 roots. [MATTE ET AL. 2002; TOFIGH ET AL. 2010])
- 6 cases

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

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