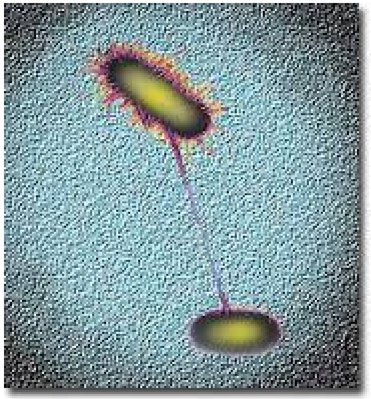
Dating species trees with lateral gene transfer

Cedric Chauve
Simon Fraser University



GJ Szollosi, Hungary

E. Tannier, P Auffret, B Boussau, P Veber, V Daubin, Lyon, France

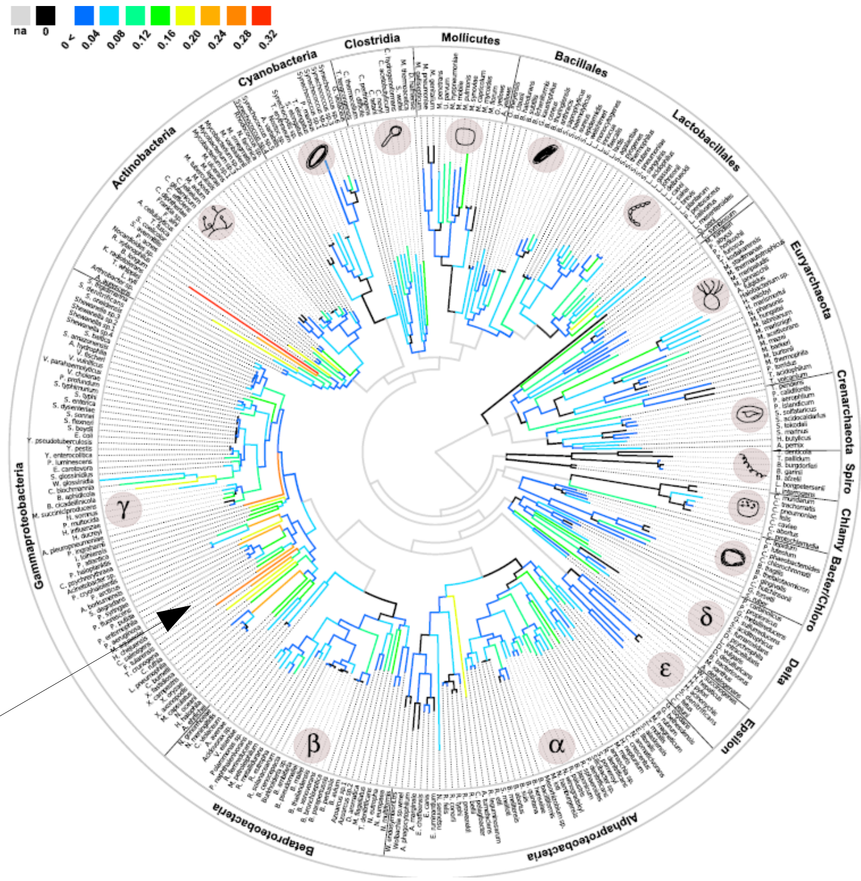
Celine Scornavacca, Montpellier, France

A Rafiey, Vancouver, Canada

This is a work-in-progress, that is being written
and will likely be made public (finally) in a month.

Banff, February 2017

Lateral gene transfer used to be seen as a phylogenetic nightmare to recover a tree, but lateral gene transfers carry invaluable evolutionary information.



This tree of life is a network
(HGT network)

- Transfers **support phylogenies**, topology and root

  Abby *et al*, Lateral gene transfer as a support for the tree of life, *PNAS* 2012
- They **provide a chronology** of diversification events

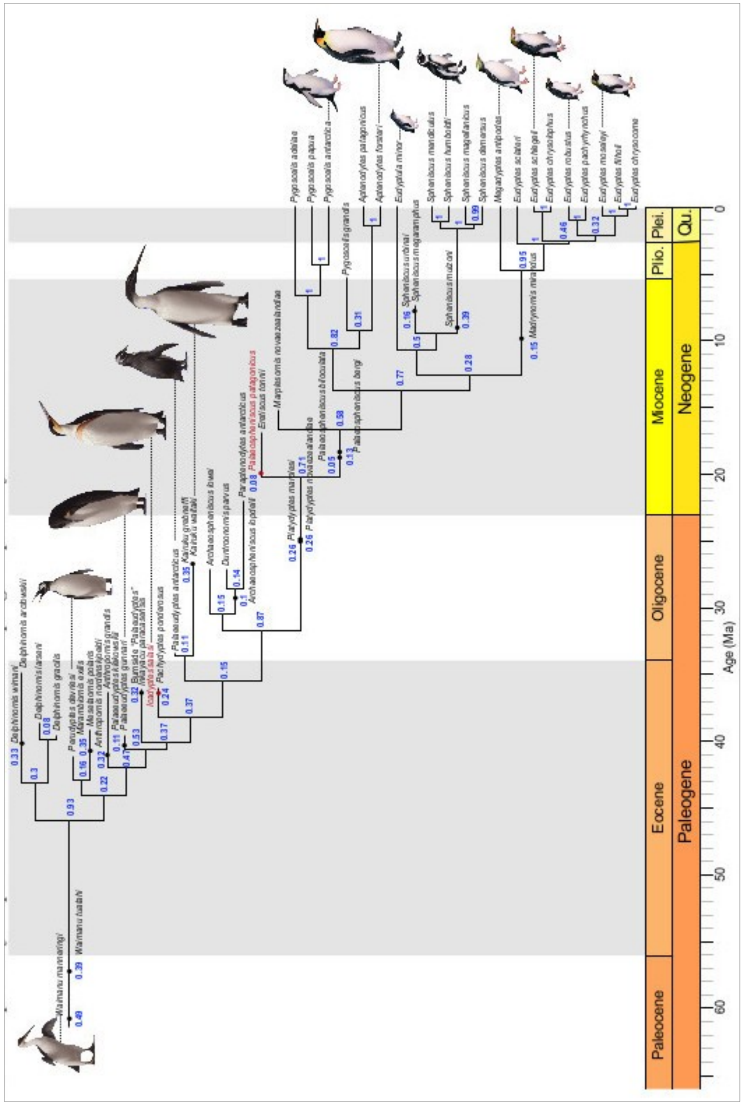
  Szöllősi *et al*, Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations, *PNAS*, 2012
- They give access to genes from **unsampled species**

  Szöllősi *et al*, Lateral gene transfer from the dead, *Sys Biol*, 2013

Dating/ranking species trees

Dating with clocks and rocks:
associating to every node of a species tree a *precise time of occurrence* using a combination of sequence and fossil data.

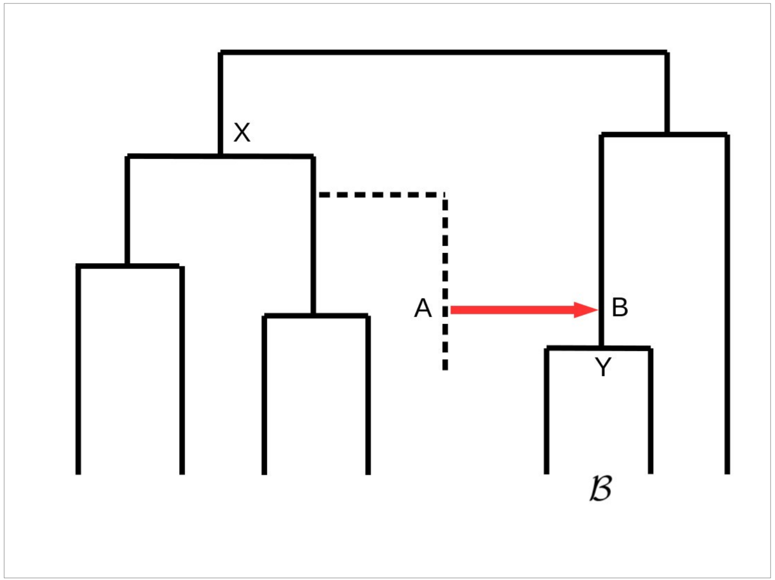
Problems:mutation rates, variability among sites and taxa, fossil calibration



Total-Evidence Dating under the Fossilized Birth-Death Process. Chi Zhang, Tanja Stadler *et al. Sys Biol* 2016.

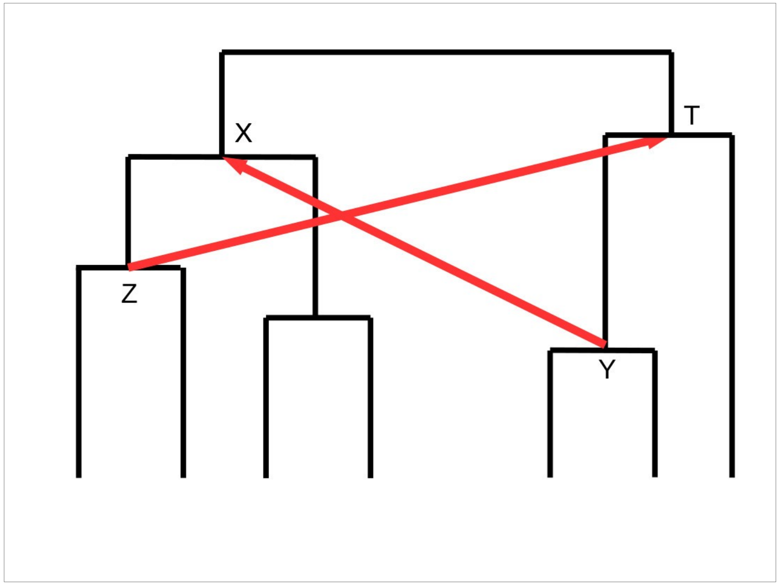
Dating with transfers:
Computing a *ranking* of the speciation events in a tree.

Dating with transfers



A transfer from A to B implies
Y=child(B) is posterior to X=parent(A)

But we can not assume that child(A) is posterior to parent(B)

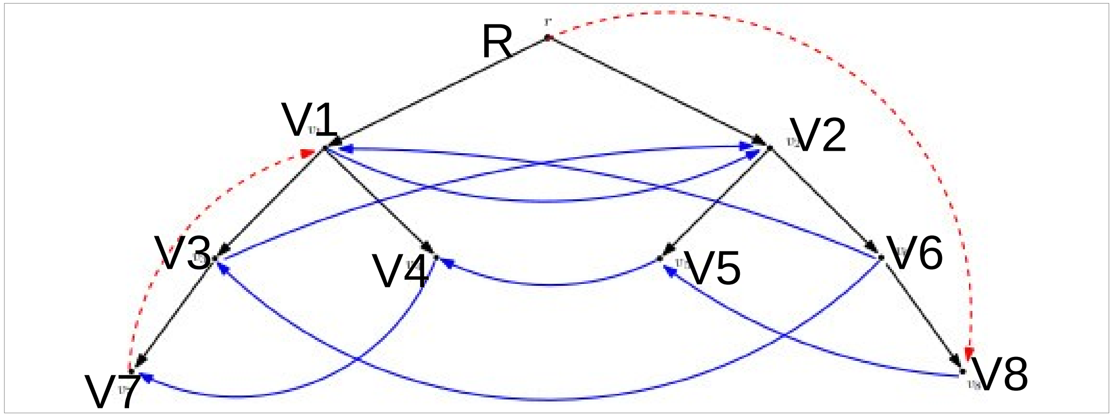


Problem :
Inferring lateral transfers is difficult and transferts inferred from many loci are noisy any conflicting.

Problem statement

Input: an undated species tree and a collection of weighted order relations between uncomparable internal nodes.

Problem: selecting a weight-maximal time-consistent set of order relations defining a (partial) order of the speciations.



Red arcs are not allowed or uniformative
Blue arcs are implied by inferred transfers
Leaves have been removed

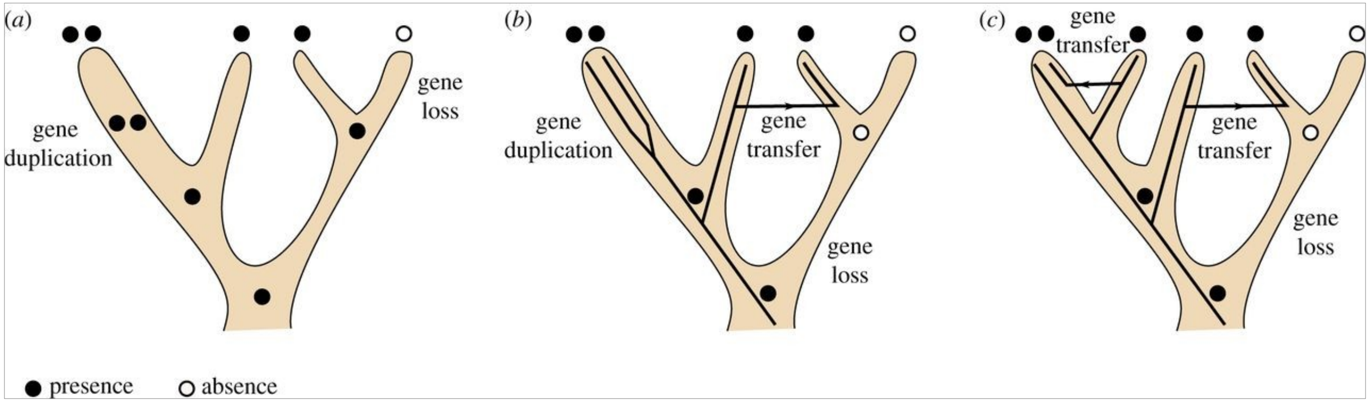
(V3,V2) and V6,V1) are time-inconsistent

Discarding (V6,V1) and (V6,V3) gives a time-consistent set of transfers and the following ranking

R < V1 < V3 < V2 < V6 < V8 < V5 < V4 < V7

Reconciliation and lateral transfers

Gene tree / Species tree Reconciliation:



Szöllősi et al, 2015

Time-consistent parsimonious scenarios : NP-hard

Tofigh et al, TCBB 2011

Potentially time-inconsistent parsimonious scenarios: O(n^2)

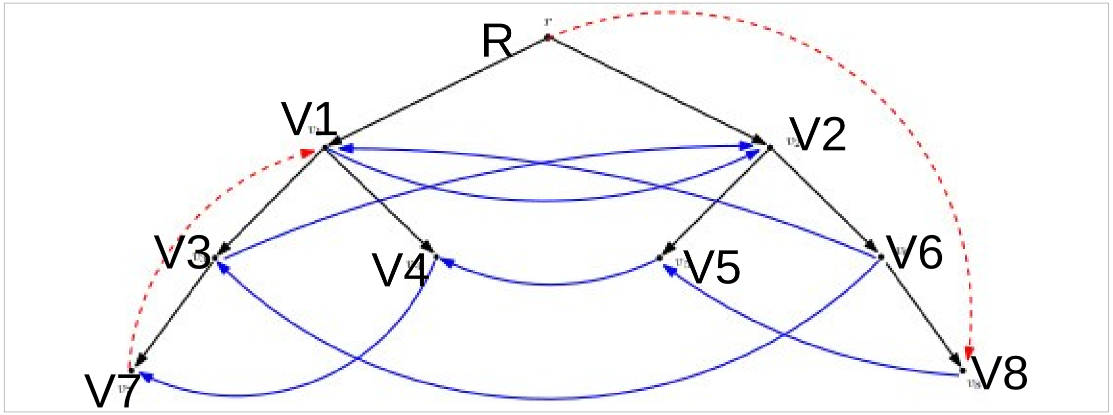
Ranger-DTL, Bansal et al, Bioinformatics 2012
Notung, Stolzer et al, J Comput Biol 2012
ecceTERA, Jacox et al, Bioinformatics 2016

Probabilistic reconciliations models:

ALE, Szöllősi et al, Sys Biol 2013
ALE_undated,Szöllősi et al, PTRSL (B) 2015
JprIME, Khan et al, BMC Bioinformatics 2015

A graph theory problem

DFAST: Directed Feedback Arc Set on a Tree

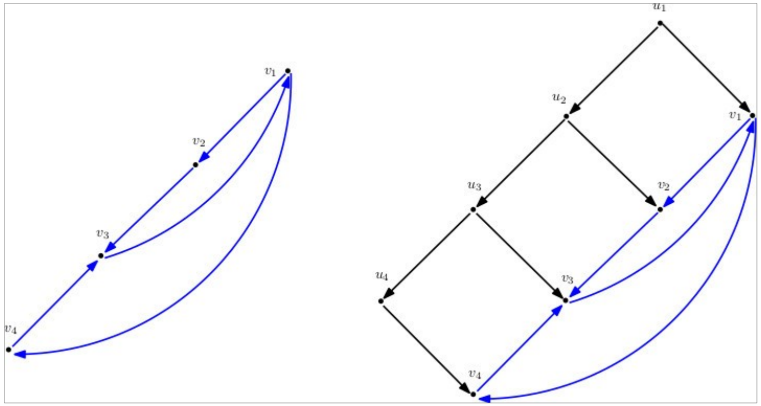


1. An arc removal problem:
Break all cycles by discarding transfer-induced arcs.

2. A vertex ordering problem:
Order the nodes of the graph in order to minimize the weight of the feedback arcs.

R < V1 < V3 < V2 < V6 < V8 < V5 < V4 < V7

Algorithmic results



DFAST is NP hard:
Reduction from the DFAS problem.

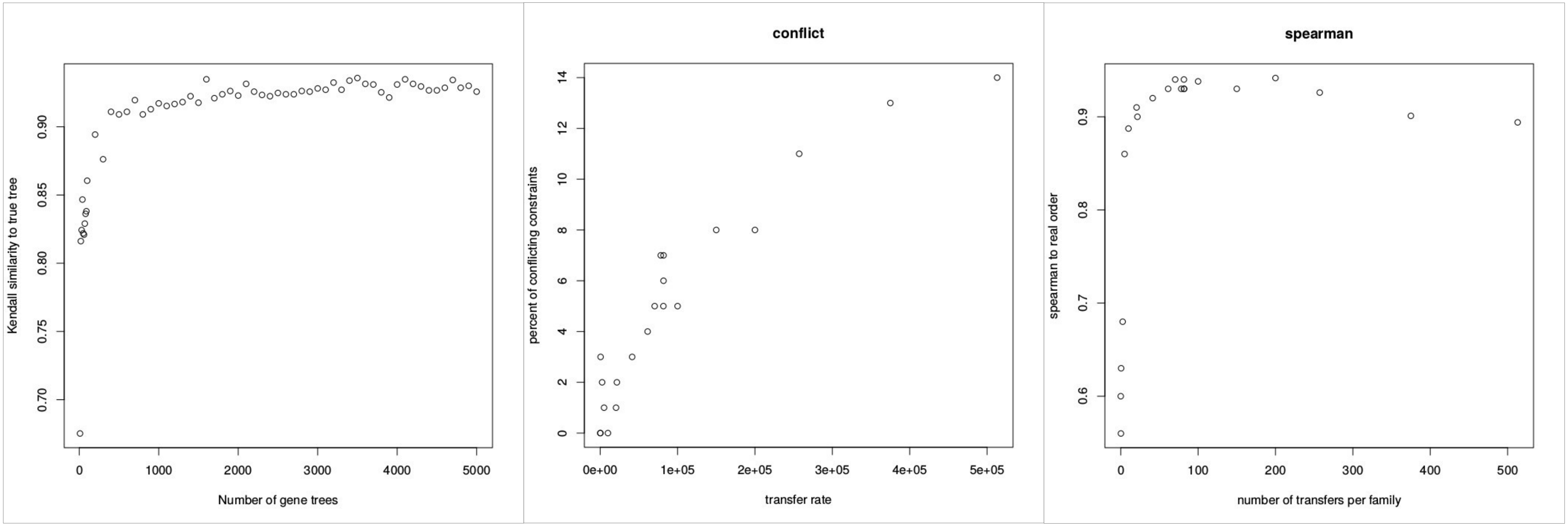
Edge removal: Greedy heuristic.

Vertex ordering:
In a top-down process:

```
    If the current subtree is small, apply a branch-and-bound
    Otherwise
        Compute an order O1 for the left subtree
        Compute an order O2 for the right subtree
        Merge O1 and O2 minimizing the weight of feedback
        arcs located between the two subtrees
        (exact Dynamic Programming algorithm)
        Try to improve through local-search
```

Experimental results

Experimental setup :
Simulated data obtained using SimPhy (Mallo *et al,* Sys. Biol. 2015)
Random dated species tree over 500 species, 100 sampled for analysis
1000 to 5000 gene trees, with no duplication but lateral transfers
Transferts rates : from 50 transfers to 500,000 transfers (over all trees)



Kendall-Tau similarity:
normalized by the (conjectured) maximum distance

Problem:
Spearman and Kendall-Tau distances

To compare the inferred ranking to the true ranking, the Kendall-Tau and Spearman footrule approaches are natural ones.

However, to measure how the inferred ranking compares to a random ranking, or to the worst ranking , one can not consider all random rankings as the tree structure imposes constraints.

We did not find anything existing about this problem and have some conjectures on how to obtain the worst-order at least.

Conclusion

Context:
Evolution with lateral transfers, or hybridization, introgression together with a tree-like underlying structure.

Principle:
Taking advantage of the information provided by transfers.
This information is likely conflicting (transfers are hard to infer).

Contribution:
A combinatorial optimization approach to clear conflicts.
Preliminary (quite simple) algorithmic results.
Encourageing preliminary experiments.

Work-in-progress / future work

Algorithms:
The algorithmic of the DFAST problem is quite open.
Considering partial orders instead of total orders is likely better.
Integrating partial orders or to consider sub-optimal solutions ?
(Gibbs-Boltzmann sampling?)

A more general approach:

1) Start from an unranked or partially ranked species tree
2) Infer transfers from reconciliation (ecceTERA, ALE)
3) DFAST, extract high confidence ranking information
4) Augment the ranking of the species tree
5) Repeat

Reconstructing ancestral gene orders using gene trees

E. Tannier, V. Daubin, W. Duchemin, Lyon, France

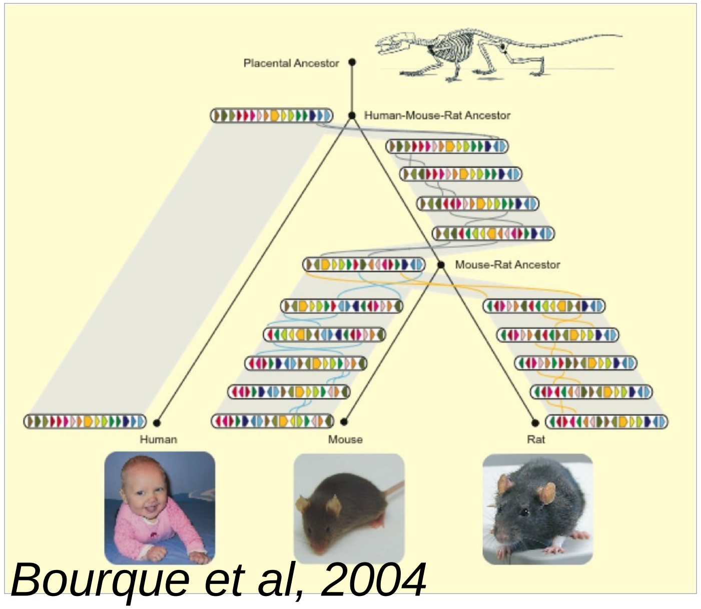
C. Scornavacca, S. Berard, A. Chateau, Y. Anselmetti, Montpellier, France

A Rajaraman, Vancouver, Canada

Y. Ponty, Polytechnique, France

M. Patterson, Milan, Italy

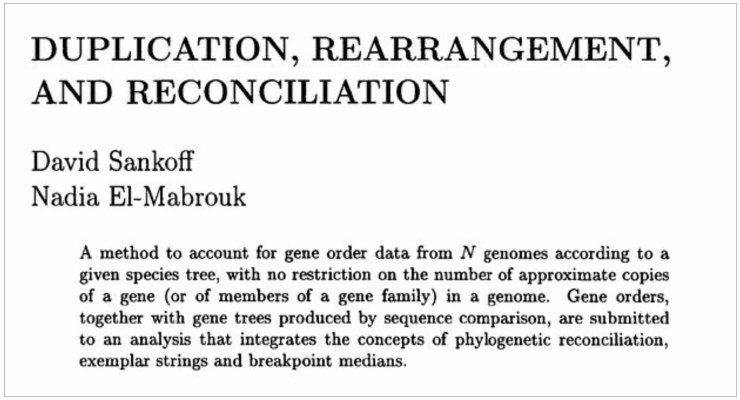
Reconstructing ancestral gene orders



Bourque et al, 2004

The « old-fahsioned » approach :
One-to-one orthologous « genes »
Signed permutations

To account for the full gene complement of a genome, we need to account for gene duplication. Reconciled gene trees then become a natural input to the problem.

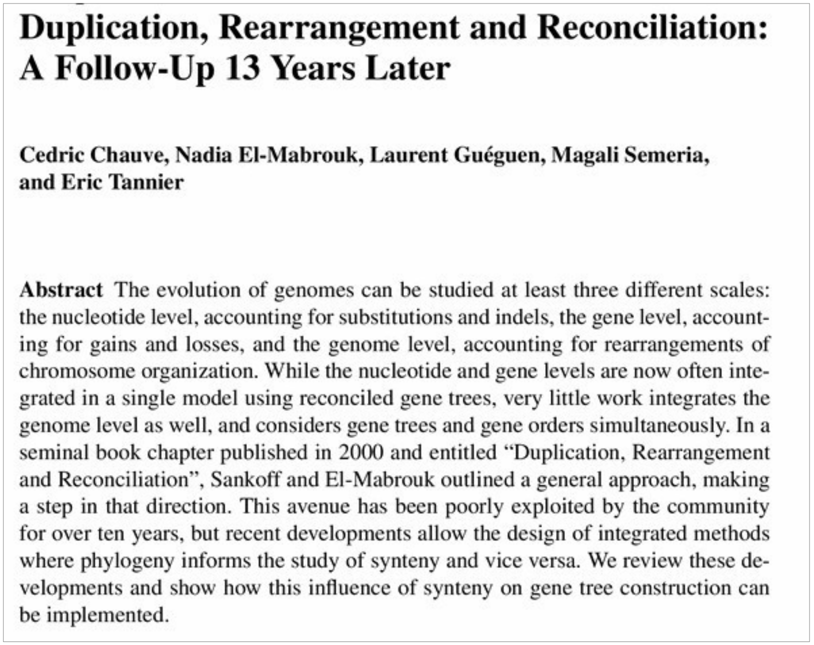


DUPLICATION, REARRANGEMENT, AND RECONCILIATION

David Sankoff
Nadia El-Mabrouk

A method to account for gene order data from N genomes according to a given species tree, with no restriction on the number of approximate copies of a gene (or of members of a gene family) in a genome. Gene orders, together with gene trees produced by sequence comparison, are submitted to an analysis that integrates the concepts of phylogenetic reconciliation, exemplar strings and breakpoint medians.

DCAF, 2000



Duplication, Rearrangement and Reconciliation: A Follow-Up 13 Years Later

Cedric Chauve, Nadia El-Mabrouk, Laurent Guéguen, Magali Semeria, and Eric Tannier

Abstract The evolution of genomes can be studied at least three different scales: the nucleotide level, accounting for substitutions and indels, the gene level, accounting for gains and losses, and the genome level, accounting for rearrangements of chromosome organization. While the nucleotide and gene levels are now often integrated in a single model using reconciled gene trees, very little work integrates the genome level as well, and considers gene trees and gene orders simultaneously. In a seminal book chapter published in 2000 and entitled "Duplication, Rearrangement and Reconciliation", Sankoff and El-Mabrouk outlined a general approach, making a step in that direction. This avenue has been poorly exploited by the community for over ten years, but recent developments allow the design of integrated methods where phylogeny informs the study of synteny and vice versa. We review these developments and show how this influence of synteny on gene tree construction can be implemented.

MAGE, 2013

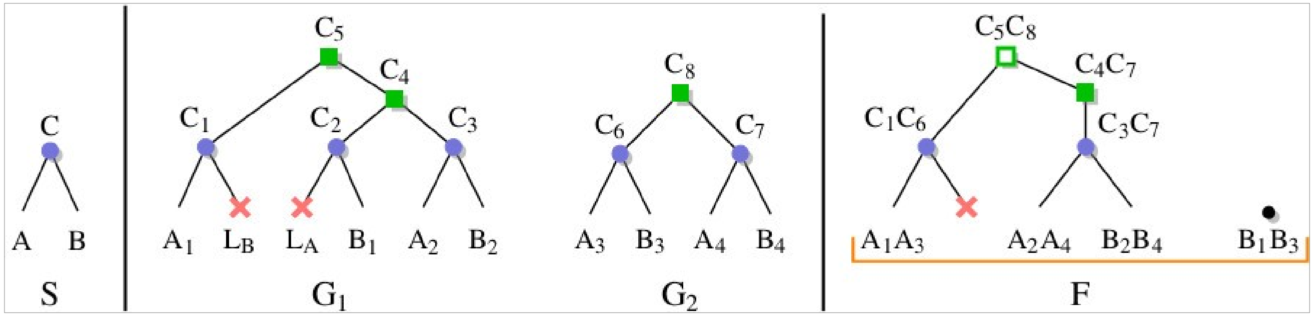
Reconstructing ancestral gene orders with reconciled gene trees

Input: a species tree, assembled and annotated extant genomes.

Step 1: clustering genes into gene families, aligning families.

Step 2a: building (rooted?) gene trees, one for each family.
Step 2b: reconciling gene trees with the species tree
(defines the gene content of each ancestral species).

Step 3: building adjacency forests, one adjacency at a time.



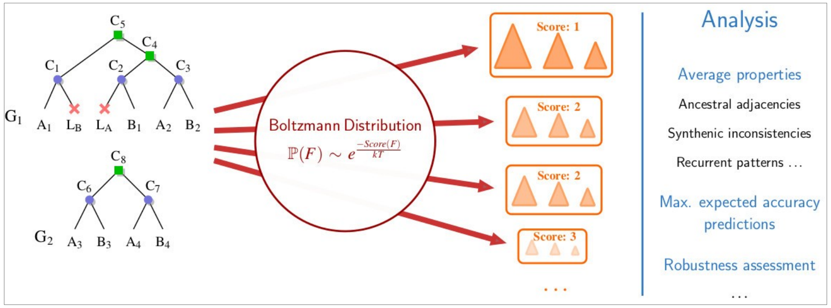
Step 4: clearing out syntenic conflicts (genes with > 2 neighbours)

DeCo*, building adjacency forests

DeCo (Berard *et al*, Bioinformatics 2012):
Parsimonious adjacency evolution scenarios,
Dynamic programming « a la Sankoff-Rousseau »

DeCoLT (Patterson *et al*, BMC Bioinformatics 2013):
Integrating lateral transfers, for dated species trees

DeClone (Zanetti *et al*, BMC Bioinformatics 2015):
From parsimony to Gibbs-Boltzmann sampling



DeCo-polytope (Rajaraman *et al*, ISBRA 2015):
From parsimony to polytopes « a la Sturmfels »

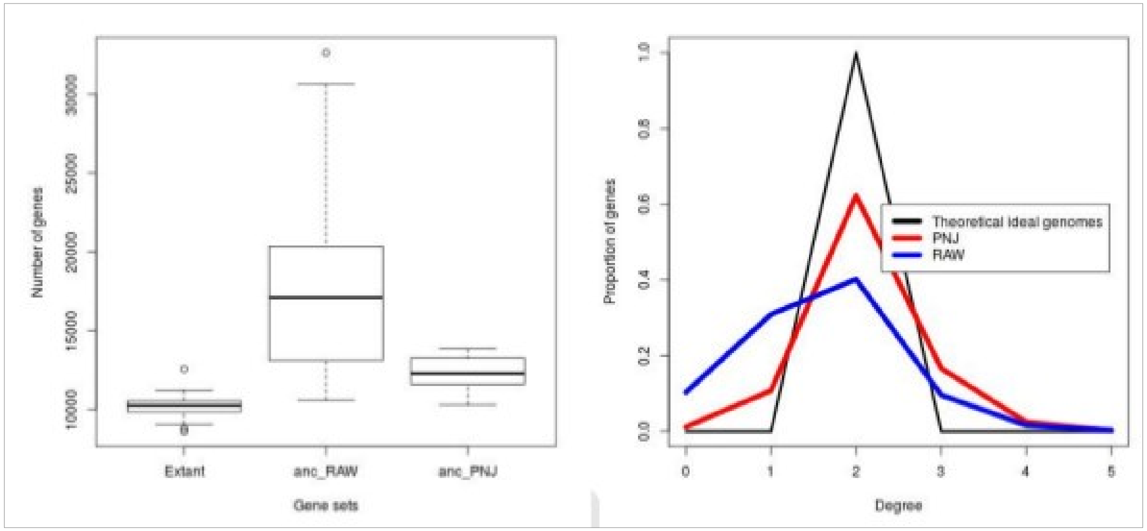
Art-DeCo (Anselmetti *et al*, BMC Genomics 2015):
Joint ancestral reconstruction / scaffolding of fragmented extant genomes

DeCo* (Duchemin *et al*, submitted to GBE, 2017) :
All in one (plus ecceTERA for computing reconciliations).

Question: improving gene trees from inconsistencies in reconstructed ancestral genomes (1)

Example 1: unrealistically large ancestral *Anopheles* genomes.

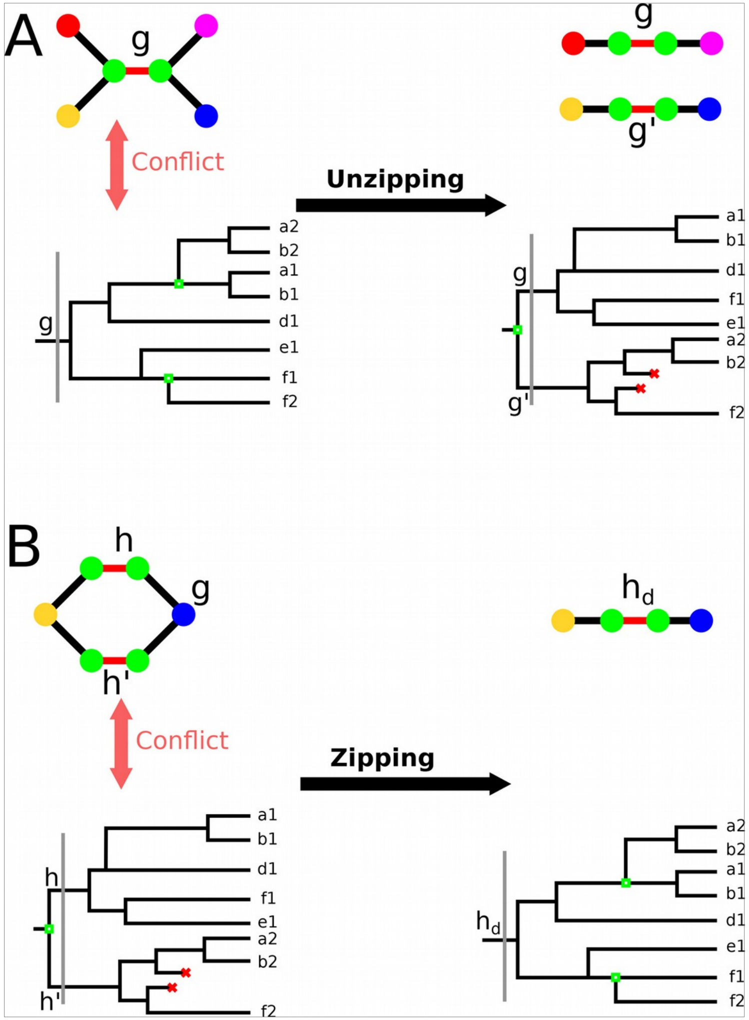
~15,000 gene families from 18 *Anopheles* genomes
Original gene trees: ML trees from VectorBase
Corrected trees: ProfileNJ (Nouathi *et al*, PloS One, 2015)
Reconciliation: ecceTERA (Jacox *et al*, Bioinformatics 2016)



Anselmetti *et al*, work-in-progress

Question: improving gene trees from inconsistencies in reconstructed ancestral genomes (2)

Example 2: zipping/unzipping duplications to correct syntenic conflicts.
Reconstruction of an ancestral *Yersinia pestis* genome



Claim: all potential errors before the syntenic conflict and that might have caused it have a non-zero probability to be true.

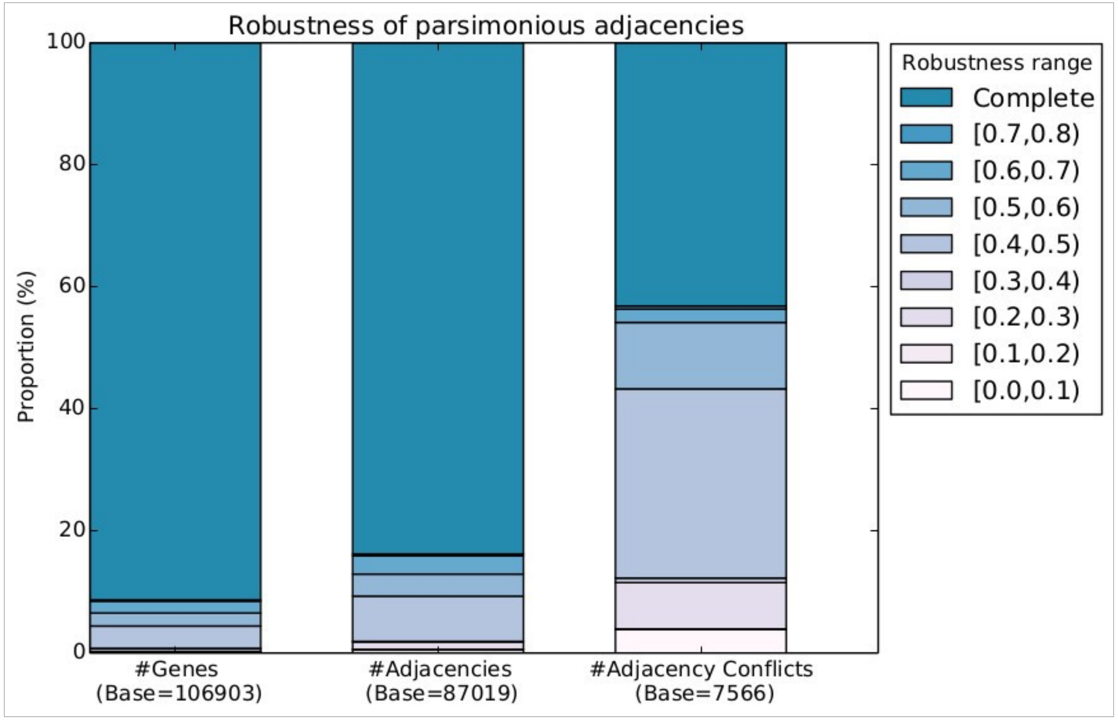
Question: Facing the hard inconsistency of syntenic conflicts, can we try to solve them by going back in the pipeline and modify, in some kind of parsiminious way, some of the underlying structures (gene families, reconciled gene trees, adjacency forests) ?

Question: improving gene trees from inconsistencies in reconstructed ancestral genomes (3)

Example 3: mammalian gene gene trees.

5,039 mammalian reconciled gene trees (Ensembl, 2012)
6,074 DeCo/DeClone instances
112,188 ancestral genes

Keeping all ancestral adjacencies that are present in all optimal solutions (DeClone) and robust to a score change (DeCo-polytope) still results in a significant level of syntenic conflict.



The conflicting adjacencies again correspond to ancestral species with gene content larger than expected.

Rajaraman *et al*, ISBRA 2015

Problem:
Species tree topology (or ranking) test f

Given two species trees (ranked if lateral gene transfer is involved), reconciled gene trees, adjacencies forests, can we use features such as

ancestral gene content,
# duplications,
# losses,
# transferts,
ILS,
# adjacency gains,
# adjacency breaks,
# syntenic conflict,

to compare both species trees ?

Aknowledgements

LBBE

isem
Institut des Sciences de l'Evolution-Montpellier

SFU
VANCOUVER

E·L·T·E
KUTATÓEGYETEM

ANR

Ancestrome

Pacific Institute *for the*
Mathematical Sciences

NSERC
CRSNG

PRACE

CINES

WESTGRID