Why should we care about cograph heuristics? Phylogenomics with paralogs

Marc Hellmuth

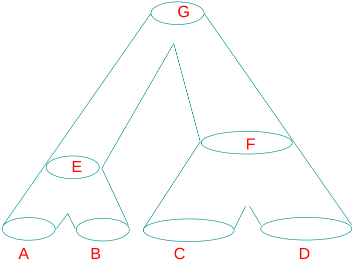
Faculty of Mathematics and Computer Science
Saarland University, Germany

Joint work with:
Nicolas Wieseke, Peter F Stadler, Martin Middendorf
Maribel Hernandez Rosales (U Leipzig)
Katherina Huber, Vincent Moulton (U East Anglia),
Hans-Peter Lenhof (U Saarland), Marcus Lechner (U Marburg)

30TH TBI WINTERSEMINAR, BLED 2015

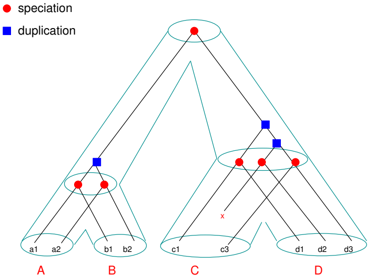
1. Phylogeny and Basics
2. Orthology, Paralogy and Gene Trees **- Cograph Editing**
3. Inferring Species Trees
4. Results

Phylogenetics



Phylogenetics

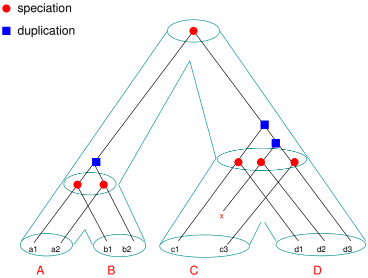
- species are characterized by its genome: a “bag of genes”
- “Genes” evolve along a rooted tree
- unique event labeling
  $t : V^0 \rightarrow M = \{\bullet, \blacksquare\}$
  two types of branching events:



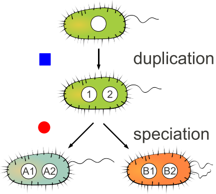
Phylogenetics

- species are characterized by its genome: a "bag of genes"
- "Genes" evolve along a rooted tree
- unique event labeling $t : V^0 \to M = \{\bullet, \blacksquare\}$

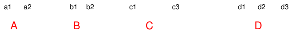
  two types of branching events:



1. Gene duplication: an offspring has two copies of a single gene of its ancestor
2. Speciation: two offspring species inherit the entire genome of their common ancestor

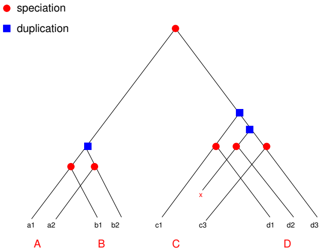


The Problem in Practice



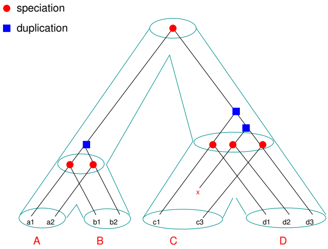
- Only the subset of leaves of the gene tree corresponding to genes in extant (currently living) species is observable.
- All internal nodes and the event labelling *t* in the gene tree must be inferred from data.
- The events and the topology of the gene tree can be used (under several constraints) to infer the species tree (Reconciliation)

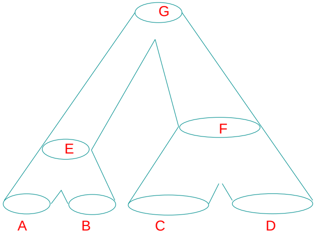
The Problem in Practice



- Only the subset of leaves of the gene tree corresponding to genes in extant (currently living) species is observable.
- All internal nodes and the event labelling *t* in the gene tree must be inferred from data.
- The events and the topology of the gene tree can be used (under several constraints) to infer the species tree (Reconciliation)

The Problem in Practice





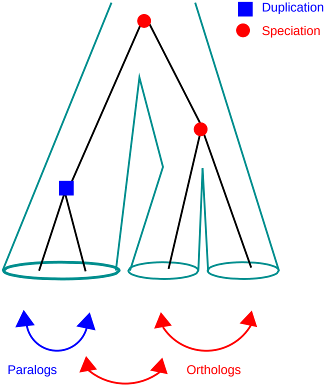
- Only the subset of leaves of the gene tree corresponding to genes in extant (currently living) species is observable.
- All internal nodes and the event labelling t in the gene tree must be inferred from data.
- The events and the topology of the gene tree can be used (under several constraints) to infer the species tree (Reconciliation)

Orthologs and Paralogs

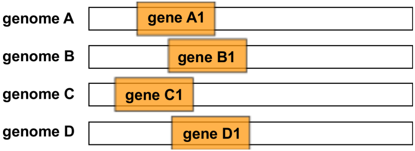
Orthology and paralogy are important concepts in evolutionary biology and are defined in terms of the pair (T, t).

Two genes x and y are

- orthologs if $t(\text{lca}(x,y)) =$ ● =speciation
- paralogs if $t(\text{lca}(x,y)) =$ ■ =duplication

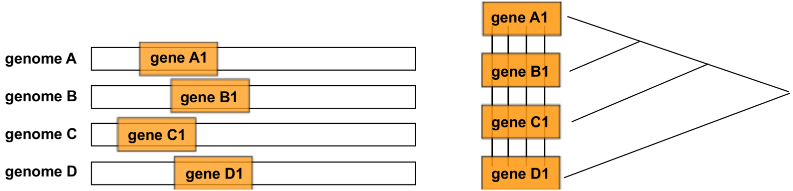


State-of-the-Art Tree Reconstruction



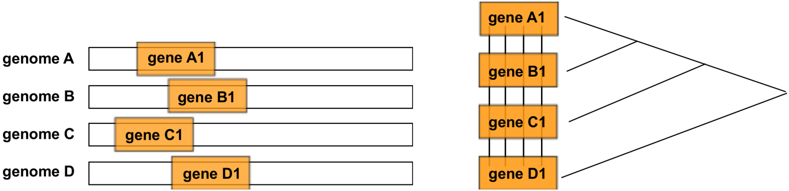
- Find 1:1-orthologs.
    - Paralogs = dangerous nuisance that has to be detected and removed.
    - Select families of genes that rarely exhibit duplications (e.g. rRNAs, ribosomal proteins)

State-of-the-Art Tree Reconstruction



- Find 1:1-orthologs.
    - Paralogs = dangerous nuisance that has to be detected and removed.
    - Select families of genes that rarely exhibit duplications (e.g. rRNAs, ribosomal proteins)
- Alignments of protein or DNA sequences and standart techniques yield evolutionary history that is believed to be congruent to that of the respective species.

State-of-the-Art Tree Reconstruction



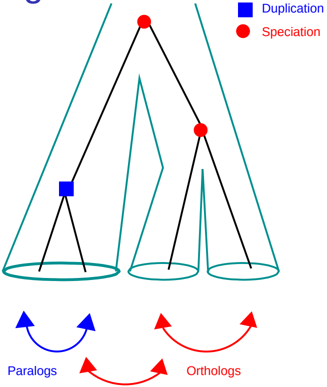
Pitfalls:

- The set of usable gene sets is strongly restricted ($\leq 10\%$).
- Information of evolutionary events as paralogs or xenologs is ignored.
- It is often mistakenly assumed that the orthology relation is transitive.

Orthologs and Paralogs

Two genes x and y are

- orthologs if
  $t(\text{lca}(x,y)) = $ ● $=$speciation
- paralogs if
  $t(\text{lca}(x,y)) = $ ■ $=$duplication



$\Longrightarrow$ orthology relation Θ can be estimated directly from the data, without constructing either gene or species trees
e.g. with ProtheinOrtho or its extension PoFF

ProteinOrtho: Detection of (Co)orthologs in large-scale analysis., Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ, BMC Bioinformatics, 2011

Estimating Θ directly from the data

The relation $\widehat{\Theta}$ is only an estimate of a “correct” orthology relation Θ.

Aim: Correct initial estimate $\widehat{\Theta}$ to the “closest” orthology relation Θ that fits the data and build corresponding gene and species trees.

$\Longrightarrow$ What is a “closest” orthology relation Θ?

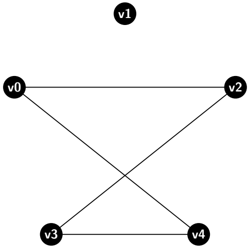
Characterization of Θ

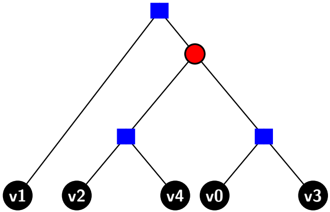
Question: When does the initial estimate $\widehat{\Theta}$ fit the data?

Equivalently we can ask for a "symbolic representation":

For a given $\widehat{\Theta}$ when does there exist a tree T with event labeling t s.t.

- $t(\mathrm{lca}(x,y)) =$ ● $=$ *speciation* for all $(x,y) \in \widehat{\Theta}$ and
- $t(\mathrm{lca}(x,y)) =$ ■ $=$ *duplication* for all $(x,y) \notin \widehat{\Theta}$?





$G_{\widehat{\Theta}}$ with edge set $\widehat{\Theta} = \{(v0,v2),(v0,v4),(v2,v3),(v3,v4)\}$

Characterization of Θ

Question: When does the initial estimate $\widehat{\Theta}$ fit the data?

Equivalently we can ask for a "symbolic representation":

For a given $\widehat{\Theta}$ when does there exist a tree T with event labeling t s.t.

- $t(\text{lca}(x,y)) = \bullet =$ *speciation* for all $(x,y) \in \widehat{\Theta}$ and
- $t(\text{lca}(x,y)) = \blacksquare =$ *duplication* for all $(x,y) \notin \widehat{\Theta}$?

We used results by Böcker & Dress (1998) on "symbolic ultrametrics":

Theorem

The following conditions are equivalent

- *There is a symbolic representation for $\widehat{\Theta}$.*
- $G_{\widehat{\Theta}}$ ***is a Cograph.***

Recovering Symbolically Dated, Rooted Trees from Symbolic Ultrametrics, Böcker & Dress, Adv. Math., 1998

Orthology Relations, Symbolic Ultrametrics, and Cographs, Hellmuth M, H.-Rosales M, Huber K, Moulton V, Stadler PF, Wieseke N, J. Math. Biol., 2012

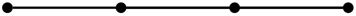
Cograph (=Complement reducible graph)

Corneil et al., 1981:

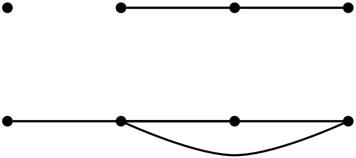
Cographs are defined recursively (Def. omitted)

G is Cograph IFF G is "induced P_4-free"

Forbidden:



Allowed:



Complement reducible graphs[1]

[1] **Complement reducible graphs**, Corneil DG, Lerchs H, Steward Burlingham L, Discr. Appl. Math., 1981

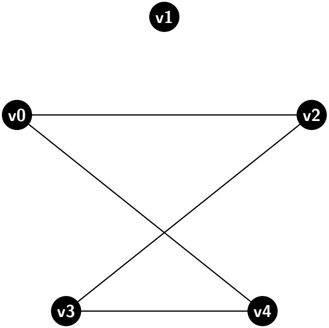
Cograph (=Complement reducible graph)

Corneil et al., 1981:

Cographs are defined recursively (Def. omitted)

G is Cograph IFF *G* is “induced P_4-free”

Every Cograph is associated with a unique Cotree.



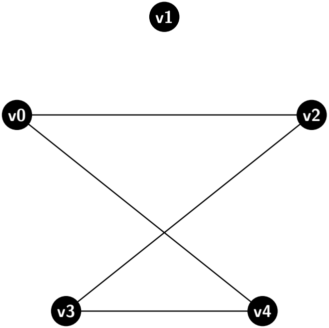
Cograph (=Complement reducible graph)

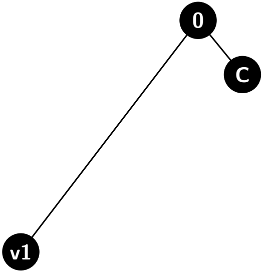
Corneil et al., 1981:

Cographs are defined recursively (Def. omitted)

G is Cograph IFF G is “induced P_4-free”

Every Cograph is associated with a unique Cotree.





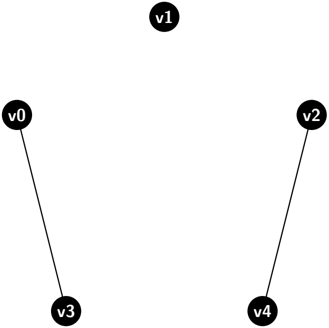
Cograph (=Complement reducible graph)

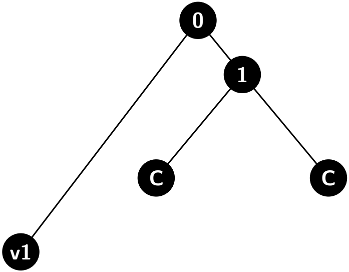
Corneil et al., 1981:

Cographs are defined recursively (Def. omitted)

G is Cograph IFF G is “induced P_4-free”

Every Cograph is associated with a unique Cotree.





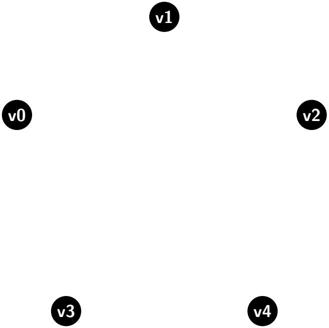
Cograph (=Complement reducible graph)

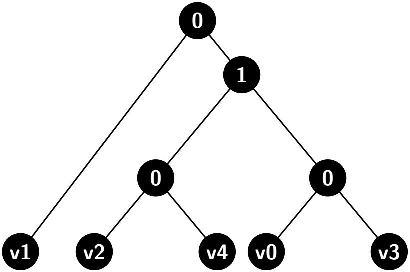
Corneil et al., 1981:

Cographs are defined recursively (Def. omitted)

G is Cograph IFF G is “induced P_4-free”

Every Cograph is associated with a unique Cotree.





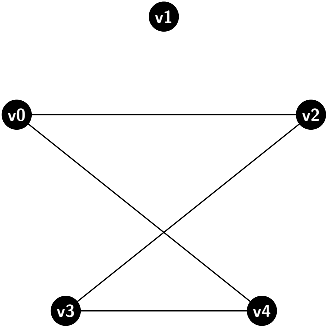
Cograph (=Complement reducible graph)

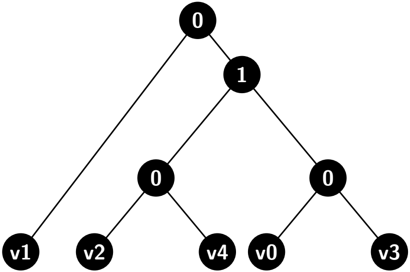
Corneil et al., 1981:

Cographs are defined recursively (Def. omitted)

G is Cograph IFF G is “induced P_4-free”

Every Cograph is associated with a unique Cotree.





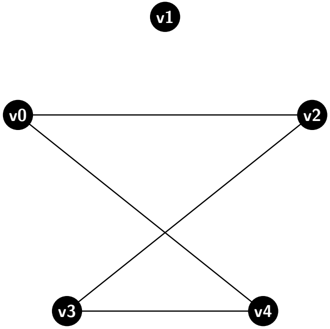
Cograph (=Complement reducible graph)

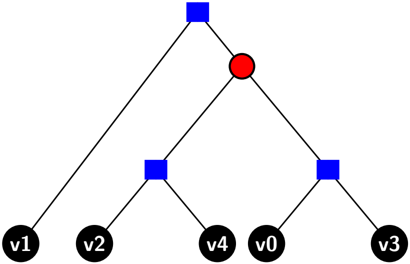
Corneil et al., 1981:

Cographs are defined recursively (Def. omitted)

G is Cograph IFF G is “induced P_4-free”

Every Cograph is associated with a unique Cotree.





$(x,y) \in E(G) = \Theta$ if and only if lca$(x,y) = 1 =$ ●

Characterization of Θ

Idea: Correct the initial estimate $\widehat{\Theta}$ to the “closest” orthology relation Θ that fits the data.

Theorem

There is a symbolic representation (T, t) for $\widehat{\Theta} \iff G_{\widehat{\Theta}}$ is a Cograph.

There is a symbolic representation (T, t) for any symbolic relation (=colored graph G) $\iff$ each monochromatic subgraph is a Cograph and on each triangle in G at most 2 colors are used.

Orthology Relations, Symbolic Ultrametrics, and Cographs, Hellmuth M, H.-Rosales M, Huber K, Moulton V, Stadler PF, Wieseke N, J. Math. Biol., 2012

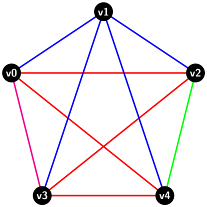
Characterization of Θ

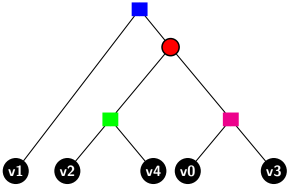
Idea: Correct the initial estimate $\widehat{\Theta}$ to the “closest” orthology relation Θ that fits the data.

Theorem

There is a symbolic representation (T,t) for $\widehat{\Theta}$ $\Longleftrightarrow$ $G_{\widehat{\Theta}}$ is a Cograph.

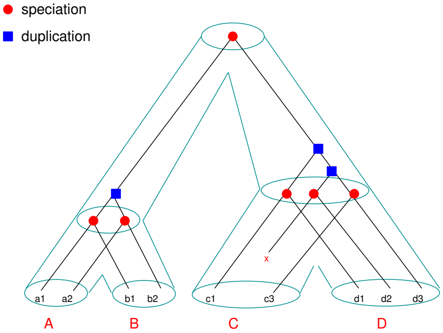
There is a symbolic representation (T,t) for any symbolic relation (=colored graph G) $\Longleftrightarrow$ each monochromatic subgraph is a Cograph and on each triangle in G at most 2 colors are used.



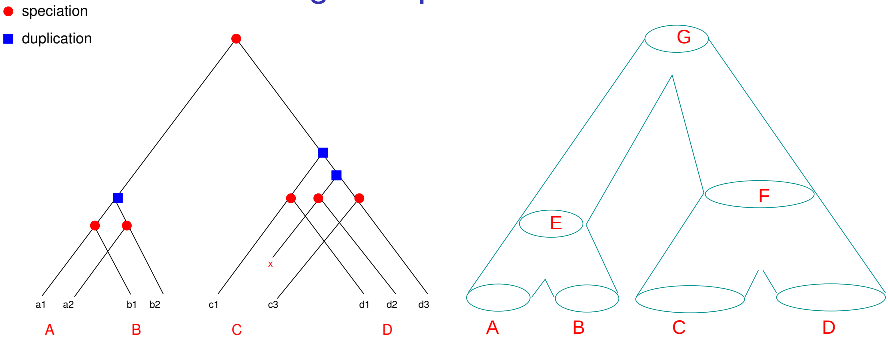


Orthology Relations, Symbolic Ultrametrics, and Cographs, Hellmuth M, H.-Rosales M, Huber K, Moulton V, Stadler PF, Wieseke N, J. Math. Biol., 2012

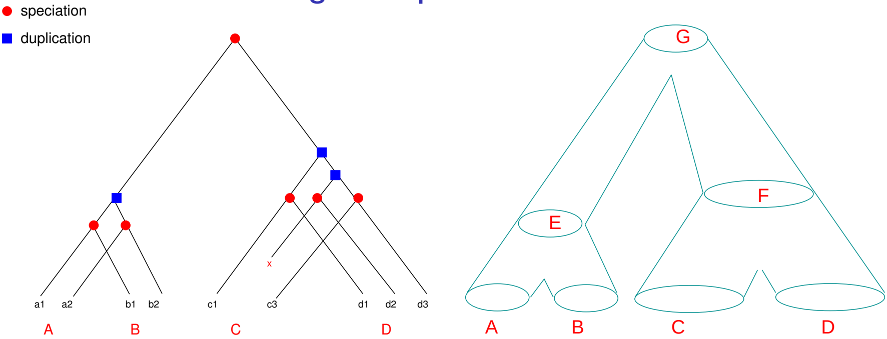
Finding the species trees



Finding the species trees



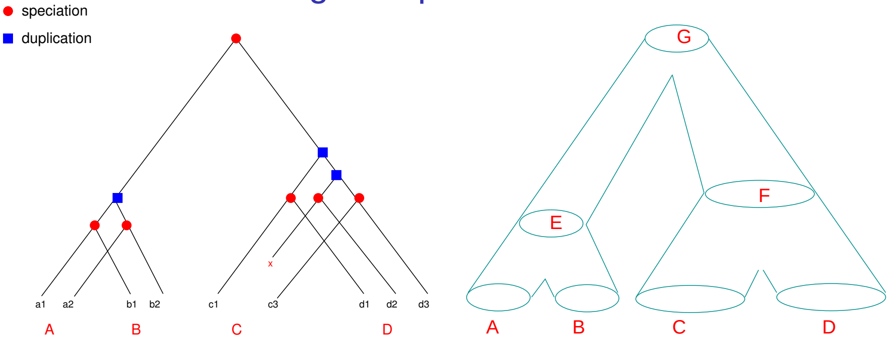
Finding the species trees



Infer local topologies of the species tree from the gene tree:
$\mathbb{S} = \{\mathrm{AB}|\mathrm{C}, \mathrm{AB}|\mathrm{D}, \mathrm{CD}|\mathrm{A}, \mathrm{CD}|\mathrm{B}\}$

From Event-Labeled Gene Trees to Species Trees., H.-Rosales M, Hellmuth M, Huber K, Moulton V, Wieseke N, Stadler PF, BMC Bioinformatics, 2012

Finding the species trees

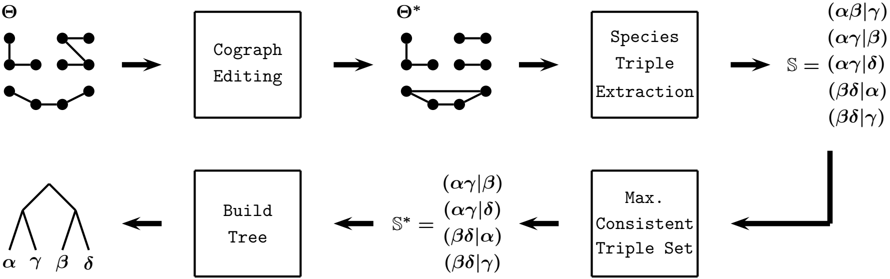


Infer local topologies of the species tree from the gene tree:
$\mathbb{S} = \{AB|C, AB|D, CD|A, CD|B\}$

Theorem. *Based on $\mathbb{S}$ it can be verified in polynomial time if there is a species tree where the gene tree can be embedded into.*
If there is a species tree for the gene tree, the species tree & embedding can be computed in polynomial time.

From Event-Labeled Gene Trees to Species Trees., H.-Rosales M, Hellmuth M, Huber K, Moulton V, Wieseke N, Stadler PF, BMC Bioinformatics, 2012

Workflow



Θ

Cograph Editing

Θ^*

Species Triple Extraction

$\mathbb{S} = (\alpha\beta|\gamma)\ (\alpha\gamma|\beta)\ (\alpha\gamma|\delta)\ (\beta\delta|\alpha)\ (\beta\delta|\gamma)$

Max. Consistent Triple Set

$\mathbb{S}^* = (\alpha\gamma|\beta)\ (\alpha\gamma|\delta)\ (\beta\delta|\alpha)\ (\beta\delta|\gamma)$

Build Tree

α γ β δ

We formulated all NP-hard problems (CE, MCT, LRT) as Integer Linear Program (ILP):

$$\min F(x) \text{ s.t. } Ax \leq b$$

Phylogenomics with Paralogs, Hellmuth M, Wieseke N, Lechner M, Lenhof HP, Middendorf M, Stadler PF, PNAS, 2015

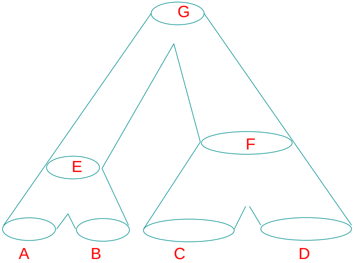
Results - Simulation

The entire worflow as ILP is implemented in the Software **`ParaPhylo`** using IBM ILOG CPLEX™ Optimizer 12.6.

It is freely available from
`pacosy.informatik.uni-leipzig.de/paraphylo`

Artificial data generated with ALF:

- generate binary species tree
- simulate dupl./loss/HGT history of gene sequences



ALF-a simulation framework for genome evolution., Dalquen et al., Mol. Biol. Evol., 2012

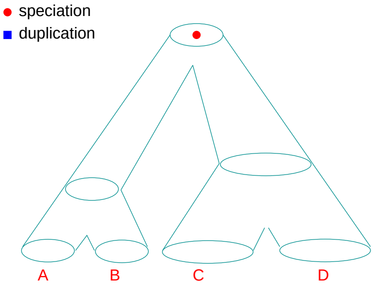
Results - Simulation

The entire worflow as ILP is implemented in the Software **`ParaPhylo`** using IBM ILOG CPLEX™ Optimizer 12.6.

It is freely available from **`pacosy.informatik.uni-leipzig.de/paraphylo`**

Artificial data generated with ALF:

- generate binary species tree
- simulate dupl./loss/HGT history of gene sequences



ALF-a simulation framework for genome evolution., Dalquen et al., Mol. Biol. Evol., 2012

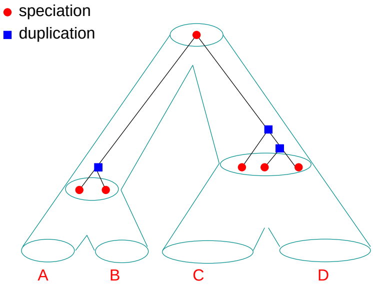
Results - Simulation

The entire worflow as ILP is implemented in the Software **`ParaPhylo`** using IBM ILOG CPLEX™ Optimizer 12.6.

It is freely available from **`pacosy.informatik.uni-leipzig.de/paraphylo`**

Artificial data generated with ALF:

- generate binary species tree
- simulate dupl./loss/HGT history of gene sequences



ALF-a simulation framework for genome evolution., Dalquen et al., Mol. Biol. Evol., 2012

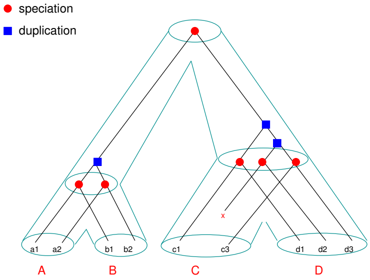
Results - Simulation

The entire worflow as ILP is implemented in the Software **`ParaPhylo`** using IBM ILOG CPLEX™ Optimizer 12.6.

It is freely available from
`pacosy.informatik.uni-leipzig.de/paraphylo`

Artificial data generated with ALF:

- generate binary species tree
- simulate dupl./loss/HGT history of gene sequences



ALF-a simulation framework for genome evolution., Dalquen et al., Mol. Biol. Evol., 2012

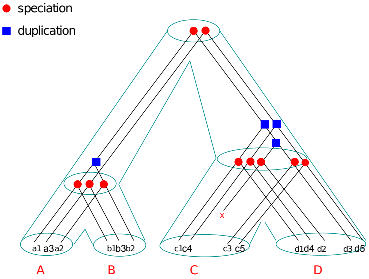
Results - Simulation

The entire worflow as ILP is implemented in the Software **`ParaPhylo`** using IBM ILOG CPLEX™ Optimizer 12.6.

It is freely available from
`pacosy.informatik.uni-leipzig.de/paraphylo`

Artificial data generated with ALF:

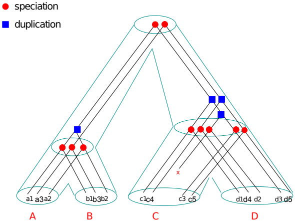
- generate binary species tree
- simulate dupl./loss/HGT history of gene sequences



ALF-a simulation framework for genome evolution., Dalquen et al., Mol. Biol. Evol., 2012

Results - Simulation 1

ALF (no HGT)



⟶ The cograph $G_{\ominus}$ is directly accessible

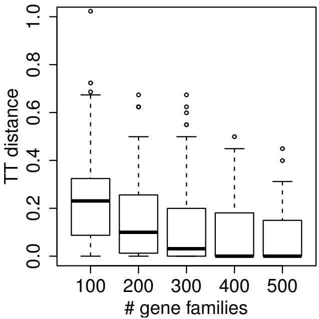
⟶ Compute cotree of $G_{\ominus}$

⟶ Extract the species triples set $\mathbb{S}$ (consistent)

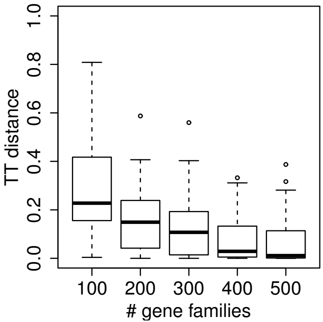
⟶ Compute least resolved species tree and compare it with initial species tree

Results - Simulation 1

Accuracy of reconstructed species trees as function of number of independent gene families:



10 species



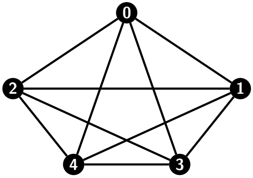
20 species

Simulation with ALF with duplication/loss rate 0.005 ($\sim$ 8% duplications) and no HGT.

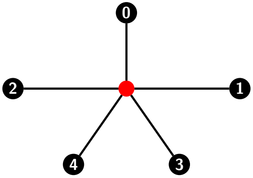
TT distance $\widehat{=}$ “num different triples in initial and reconstructed species tree”

Phylogenomics with Paralogs

In our model: $(x,y) \notin \Theta$ iff the distinct genes x and y are paralogs



G_Θ



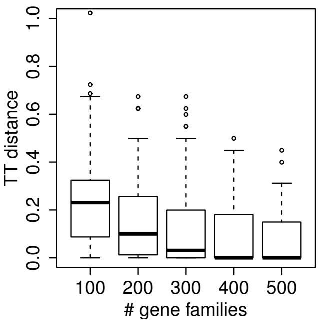
(T,t)

If $\nexists$ paralogs → G_Θ is a clique → gene tree is a star → no species triples can be inferred.

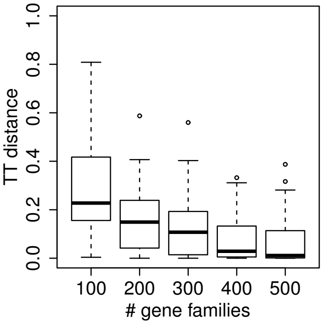
To obtain fully resolved species trees, a sufficient number of gene duplications must have occurred, since the phylogenetic information utilized by our approach is entirely contained in the duplication events.

Results - Simulation 1

Accuracy of reconstructed species trees as function of number of independent gene families:



10 species

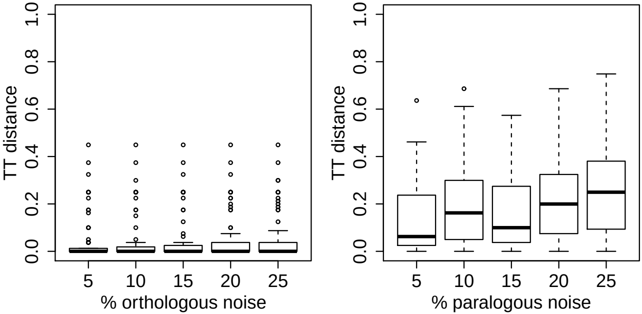


20 species

Average TT distance always smaller than 0.09 for more than 300 gene families, independent from the number of species.

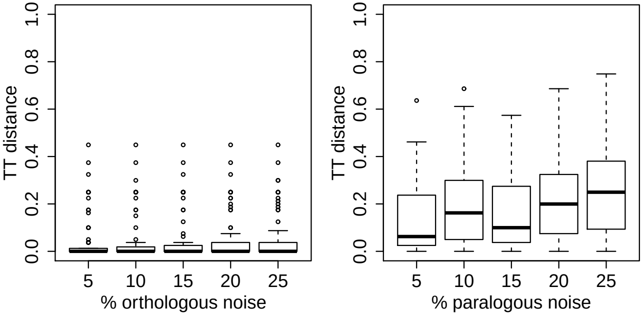
Deviations from perfect reconstructions are exclusively explained by a lack of perfect resolution.

Results - Simulation - Noise



- ALF (10 species and 1000 gene families) - $G_{\ominus}$ as before - add noise - start ILP-pipeline (CE→MCS→LRT).
- orthologous noise (overpredicting): flip paralogs with prob. p
- paralogous noise (underpredicting): flip orthologs with prob. p
- $p \in [0.05, 0.25]$

Results - Simulation - Noise



orthologous noise: additional edges in $G_\ominus$

$\longrightarrow$ $G_\ominus$ becomes more clique-alike

$\longrightarrow$ less species triples can be inferred and thus, less wrong species triples

paralogous noise: remove edges from $G_\ominus$

$\longrightarrow$ $G_\ominus$ becomes less clique-alike

$\longrightarrow$ more species triples can be inferred and thus, more more wrong species triples

Results - Runtime

Table: Running time in seconds on 2 Six-Core AMD Opteron™ Processors with 2.6GHz for individual sub-tasks: **CE** cograph editing, **MCS** maximal consistent subset of triples, **LRT** least resolved tree.

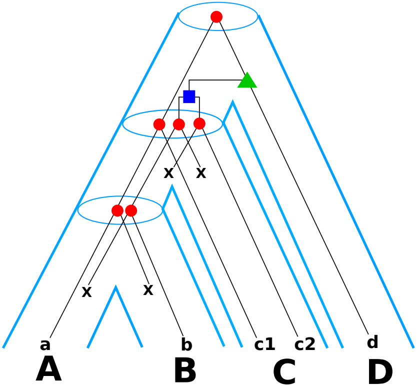
Data	**CE**	**MCS**	**LRT**	**Total**[a]
Simulations[b]	125[c]	< 1	< 1[d]	126
Aquificales[e]	34	< 1	< 1 (6)[g]	34
Enterobacteriales[f]	2673	2	< 1 (1749)[g]	2676

[a] Total time includes triple extraction, parsing input, and writing output files.
[b] Average of 2000 simulations with ALF, 10 species, 1000 gene families. 100 runs for each 4 noise models with different $p \in \{0.05, 0.1, 0.15, 0.2, 0.25\}$
[c] 2,000,000 cographs, 41 not optimally solved within time limit of 30 min.
[d] In 95.95% of the simulations the LRT could be found using `BUILD`.
[e] 11 Aquificales species with 2887 gene families.
[f] 19 Enterobacteriales species with 8308 gene families.
[g] A unique tree was obtained using `BUILD`. Second value indicates running time with ILP solving enforced.

Horizontal gene transfer (HGT)

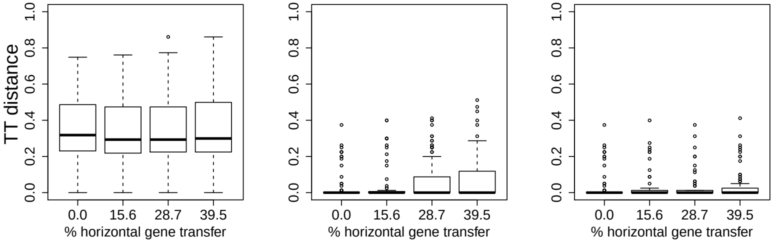
HGT refers to the transfer of genes between organisms in a manner other than traditional reproduction (sexual or asexual reproduction) and across different species (e.g. as in bacteria).

- ● speciation
- ■ duplication
- ▲ horizontal gene transfer



Horizontal gene transfer (HGT)

Dependence on the intensity of horizontal gene transfer:



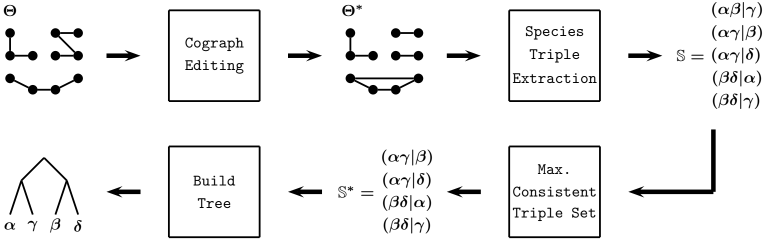
`ProteinOrtho`

perfect orthology knowledge

perfect paralogy knowledge

ALF: 10 species, 1000 gene families, duplication/loss rate 0.005 and HGT rate ranging from 0.0 to 0.0075.

Conclusion

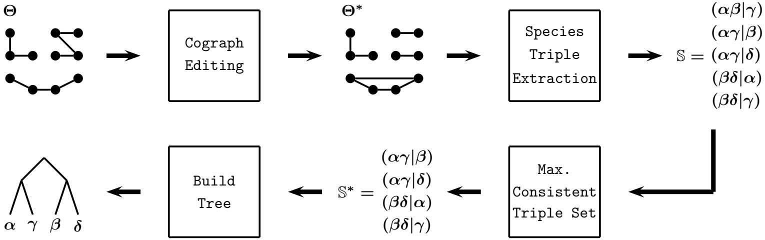

Θ
Cograph Editing
Θ^*
Species Triple Extraction
$\mathbb{S} = (\alpha\beta|\gamma)$ $(\alpha\gamma|\beta)$ $(\alpha\gamma|\delta)$ $(\beta\delta|\alpha)$ $(\beta\delta|\gamma)$
Max. Consistent Triple Set
$\mathbb{S}^* = (\alpha\gamma|\beta)$ $(\alpha\gamma|\delta)$ $(\beta\delta|\alpha)$ $(\beta\delta|\gamma)$
Build Tree
α γ β δ

In “classical standart” approaches, paralogs are treated as a dangerous nuisance that has to be detected and removed.

However, paralogy is the key!

Summary of Results here:
Phylogenomics with Paralogs. Hellmuth, Wieseke, Lechner, Lenhof, Middendorf, Stadler, PNAS, 2015

Conclusion



Cograph Editing

Species Triple Extraction

$\mathbb{S} = (\alpha\beta|\gamma)\ (\alpha\gamma|\beta)\ (\alpha\gamma|\delta)\ (\beta\delta|\alpha)\ (\beta\delta|\gamma)$

Max. Consistent Triple Set

$\mathbb{S}^* = (\alpha\gamma|\beta)\ (\alpha\gamma|\delta)\ (\beta\delta|\alpha)\ (\beta\delta|\gamma)$

Build Tree

1. Improve orthology inference tools.
2. Develop paralogy inference tools.
3. Efficient heuristics for the cograph editing and least resolved tree P.
4. On parts in G_Θ that are cliques incorporate “classical” approaches.
5. Generalization of mathematical phylogenetic framework to deal exactly with *HGT* and with phylogenetic *networks*.

THANK YOU!

Symbolic Ultrametrics

The map $\delta : X \times X \to M^{\odot}$ is said to be a symbolic ultrametric (on X) if the following conditions are satisfied

(U0) $\delta(x,y) = \odot$ if and only if $x = y$.

(U1) $\delta(x,y) = \delta(y,x)$ for all $x,y \in X$.

(U2) $|\{\delta(x,y), \delta(x,z), \delta(y,z)\}| \leq 2$ for all $x,y,z \in X$; and

(U3) there are no four pairwise distinct elements x, y, u, and v of X such that

$$\delta(x,y) = \delta(y,u) = \delta(u,v) \neq \delta(y,v) = \delta(x,v) = \delta(x,u)$$

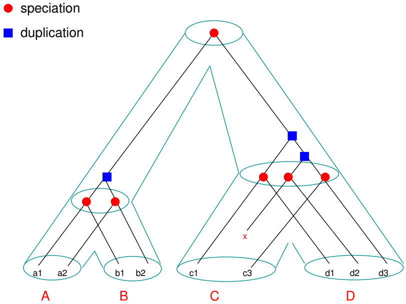
Note: every ultrametric induces a symbolic ultrametric.

Sketch: Estimating Θ directly from the Data

- We know the assignment of genes to species and we can measure similarity $s(x,y)$ of two genes using sequence alignments and `blast` bit scores
- $y \in B$ is a (putative) ortholog of $x \in A$, in symbols $(x,y) \in \widehat{\Theta}$, if
  1. $A \neq B$,

     orthologs are never found in the same species
  2. $s(x,y) \approx \max_{z \in B} s(x,z)$,

     if x and y are orthologs, then they do not have (much) closer relatives in the two species.



The relation $\widehat{\Theta}$ is only an estimate of a “correct” orthology relation: $(x,y) \in \Theta \text{ iff } t(x,y) = \bullet = \text{speciation}$

ILP - Cograph Editing

$$\min \sum_{(x,y)\in\mathfrak{G}\times\mathfrak{G}} (1-\Theta_{xy})E_{xy} + \sum_{(x,y)\in\mathfrak{G}\times\mathfrak{G}} \Theta_{xy}(1-E_{xy})$$

$$E_{xy} = 0 \text{ for all } x,y \in \mathfrak{G} \text{ with } \sigma(x) = \sigma(y)$$

$$\begin{aligned} &E_{wx} + E_{xy} + E_{yz} - E_{xz} - E_{wy} - E_{wz} \leq 2 \\ &\forall \text{ ordered tuples } (w,x,y,z) \text{ of distinct } w,x,y,z \in \mathfrak{G} \end{aligned}$$

This requires, $O(|\mathfrak{G}|^2)$ binary variables and $O(|\mathfrak{G}|^4)$ constraints; $\mathfrak{G} =$ gene set.

ILP - Max. Consistent Triple Set

$$\max \textstyle\sum_{(\alpha\beta|\gamma)\in\mathbb{S}} T'_{(\alpha\beta|\gamma)}$$

$$T'_{(\alpha\beta|\gamma)} + T'_{(\alpha\gamma|\beta)} + T'_{(\beta\gamma|\alpha)} = 1$$

$$2T'_{(\alpha\beta|\gamma)} + 2T'_{(\alpha\delta|\beta)} - T'_{(\beta\delta|\gamma)} - T'_{(\alpha\delta|\gamma)} \leq 2$$

$$0 \leq T'_{(\alpha\beta|\gamma)} + T_{(\alpha\beta|\gamma)} - 2T^{*}_{(\alpha\beta|\gamma)} \leq 1$$

Weighted version:

$$\max \textstyle\sum_{(\alpha\beta|\gamma)\in\mathbb{S}} T'_{(\alpha\beta|\gamma)} * w(\alpha\beta|\gamma)$$

Rooted species triples:
$T_{(\alpha\beta|\gamma)} = 1$ iff $(\alpha\beta|\gamma) \in \mathbb{S}$

Max. consistent subset $\mathbb{S}^* \subset \mathbb{S}$:
$T^{*}_{(\alpha\beta|\gamma)} = 1$ iff $(\alpha\beta|\gamma) \in \mathbb{S}^*$

Auxiliary consistent strict dense species triples $\mathbb{S}'$ with $\mathbb{S}^* \subseteq \mathbb{S}'$:
$T'_{(\alpha\beta|\gamma)} = 1$ iff $(\alpha\beta|\gamma) \in \mathbb{S}'$

Thus maximizing $|\mathbb{S} \cap \mathbb{S}'|$ maximizes $|\mathbb{S}^*|$ since $\mathbb{S}^* = \mathbb{S} \cap \mathbb{S}'$

The ILP formulation that uses $O(|\mathfrak{S}|^3)$ variables and $O(|\mathfrak{S}|^4)$ constraints;
$\mathfrak{S}$ =species set.

Theorem

A strictly dense triple set R on L with $|L| \geq 3$ is consistent if and only if $cl(\tilde{R}) \subseteq R$ holds for all $\tilde{R} \subseteq R$ with $|\tilde{R}| = 2$.

ILP - Least Resolved Tree

$\min \sum_p Y_p$

$0 \le Y_p|\mathfrak{S}| - \sum_{\alpha\in\mathfrak{S}} M_{\alpha p} \le |\mathfrak{S}| - 1$

$0 \le M_{\alpha p} + M_{\beta p} - 2N_{\alpha\beta,p} \le 1$

$1 - |\mathfrak{S}|(1 - T^*_{(\alpha\beta|\gamma)}) \le$
$\sum_p N_{\alpha\beta,p} - \frac{1}{2}N_{\alpha\gamma,p} - \frac{1}{2}N_{\beta\gamma,p}$

$C_{p,q,01} \ge -M_{\alpha p} + M_{\alpha q}$
$C_{p,q,10} \ge M_{\alpha p} - M_{\alpha q}$
$C_{p,q,11} \ge M_{\alpha p} + M_{\alpha q} - 1$
$C_{p,q,01} + C_{p,q,10} + C_{p,q,11} \le 2 \ \forall p,q$

Set of clusters $M_{\alpha p}$:
$M_{\alpha p} = 1$ iff $\alpha \in \mathfrak{S}$ is contained in cluster $p \in \{1,\dots,|\mathfrak{S}|-2\}$.

Cluster p contains both species α and β ($N_{\alpha\beta,p}$):
$N_{\alpha\beta,p} = 1$ iff $M_{\alpha p} = 1$ and $M_{\beta p} = 1$

Compatibility (3-gamete condition):
$C_{p,q,\Gamma\Lambda} = 1$ iff cluster p and q have gamete $\Gamma\Lambda \in \{01,10,11\}$

Y_p Non-trivial clusters: Y_p=1 iff cluster $p \ne \emptyset$.

This requires $O(|\mathfrak{S}|^3)$ variables and constraints; $\mathfrak{S}$ = species set.

“partial” hierarchy: for p and q holds $p \cap q \in \{p,q,\emptyset\}$. ($p,q$ compatible)
p and q are incompatible if there are (not necessarily distinct) species $\alpha,\beta,\gamma \in \mathfrak{S}$ with $\alpha \in p\setminus q$ and $\beta \in q \setminus p$, and $\gamma \in p \cap q$.
Then $(M_{\alpha p}, M_{\alpha q}) = (1,0)$, $(M_{\beta p}, M_{\beta q}) = (0,1)$, $(M_{\gamma p}, M_{\gamma q}) = (1,1)$.

Results

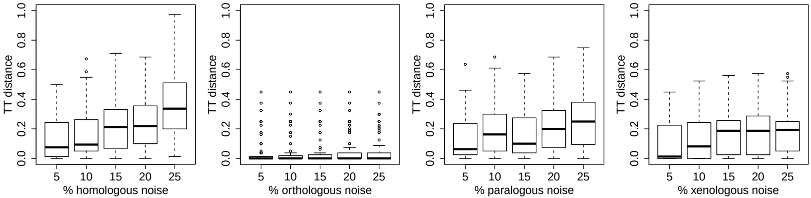
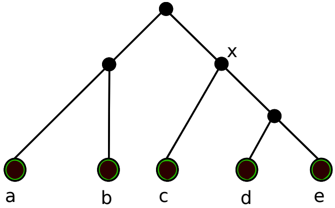
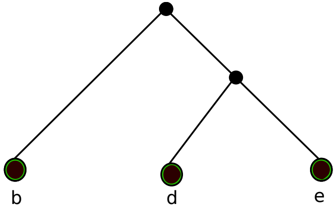


Figure: Accuracy of reconstructed species trees as function of noise level ($p = 5 - 25\%$) and noise type in the raw orthology data Θ. Tree distance is measured by the triple metric (TT) for 100 reconstructed phylogenetic trees with ten species.

Trees and triples





For three leaves a, b, c in T we write ab|c if the path from a to b does not intersect the path from c to the root.

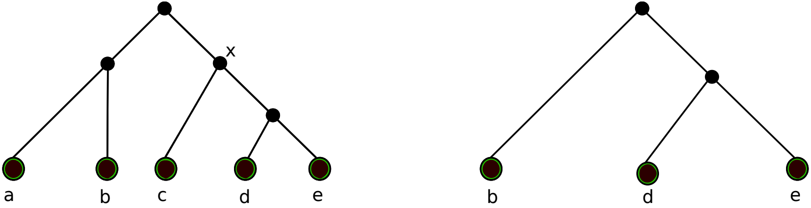
Right Tree:

$\mathscr{R}(T) = \{\text{de}|\text{b}\}$

Left Tree:

$\mathscr{R}(T) = \{\text{ab}|\text{c}, \text{ab}|\text{d}, \text{ab}|\text{e}, \text{de}|\text{a}, \text{de}|\text{b}, \text{de}|\text{c}, \text{cd}|\text{a}, \text{cd}|\text{b}, \text{ce}|\text{a}, \text{ce}|\text{b}\}$

Trees and triples



For three leaves a, b, c in T we write ab|c if the path from a to b does not intersect the path from c to the root.

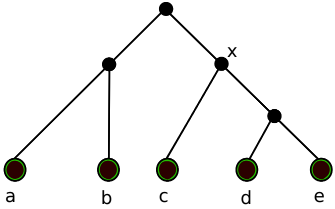
Right Tree:
$\mathscr{R}(T) = \{\text{de}|\text{b}\}$

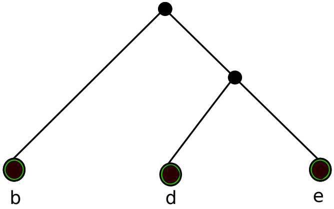
Left Tree:
$\mathscr{R}(T) = \{\text{ab}|\text{c}, \text{ab}|\text{d}, \text{ab}|\text{e}, \text{de}|\text{a}, \text{de}|\text{b}, \text{de}|\text{c}, \text{cd}|\text{a}, \text{cd}|\text{b}, \text{ce}|\text{a}, \text{ce}|\text{b}\}$

An arbitrary set of triples $\mathscr{R}$ is consistent,
if there is a tree that displays all triples in $\mathscr{R}$

Exmpl: $\mathscr{R}(T)$ is consistent. $\mathscr{R}(T) \cup \{\text{eb}|\text{d}\}$ is not consistent.

Trees and triples





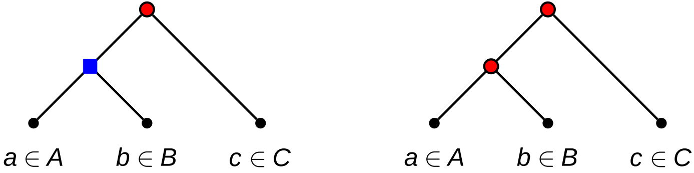
For three leaves a, b, c in T we write ab|c if the path from a to b does not intersect the path from c to the root.

Right Tree:
$\mathscr{R}(T) = \{\mathrm{de|b}\}$

Left Tree:
$\mathscr{R}(T) = \{\mathrm{ab|c}, \mathrm{ab|d}, \mathrm{ab|e}, \mathrm{de|a}, \mathrm{de|b}, \mathrm{de|c}, \mathrm{cd|a}, \mathrm{cd|b}, \mathrm{ce|a}, \mathrm{ce|b}\}$

Theorem [Aho, Sagiv, Szymanski, Ullman - 1981, Semple & Steel - 2003]
There is a polynomial time algorithm – called `BUILD` – that constructs a tree for a given set of triples $\mathscr{R}$ or recognizes $\mathscr{R}$ as inconsistent.

Triples for inferring the species tree



Given an event-labeled gene tree (T,t) and $\mathrm{ab}|\mathrm{c} \in \mathscr{R}(T)$.
We write $\mathrm{ab}|\mathrm{c}^{\bullet}$ if

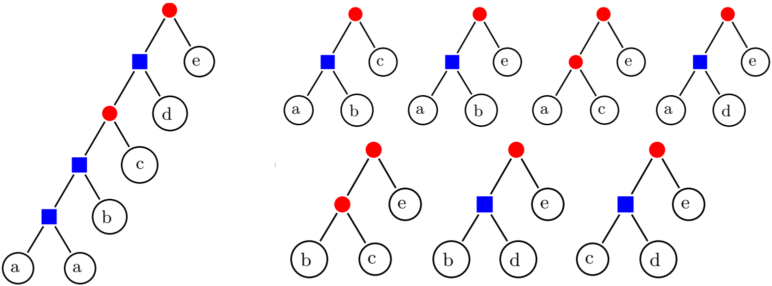
$$t(\mathrm{lca}(a,b,c)) = \bullet = \text{“}\textit{speciation}\text{''}$$

We know the assignment of genes to the species in which they occur.
This gives us the triple set:

$$\mathbb{S} = \{(\mathrm{AB}|\mathrm{C} : \exists\ \mathrm{ab}|\mathrm{c}^{\bullet} \text{ with } a \in A, b \in B, c \in C\}$$

Triples for inferring the species tree

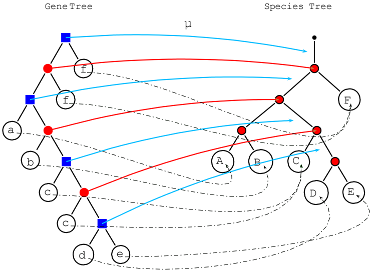
$$\mathbb{S} = \{(\mathsf{AB}|\mathsf{C} : \exists\ \mathsf{ab}|\mathsf{c}^{\bullet} \text{ with } a \in A, b \in B, c \in C\}$$



$$\mathbb{S} = \{\mathsf{AB}|\mathsf{C}, \mathsf{AB}|\mathsf{E}, \mathsf{AC}|\mathsf{E}, \mathsf{AD}|\mathsf{E}, \mathsf{BC}|\mathsf{E}, \mathsf{BD}|\mathsf{E}, \mathsf{CD}|\mathsf{E}\}$$

Triples for inferring the species tree

$$\mathbb{S} = \{(\mathrm{AB}|\mathrm{C} : \exists\ \mathrm{ab}|\mathrm{c}^{\bullet} \text{ with } a \in A, b \in B, c \in C\}$$



Theorem

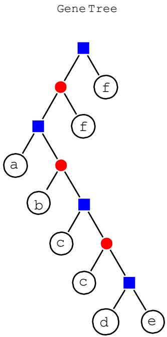
There is a species tree for the gene tree (T,t), i.e., for the symbolic representation of $\Theta \iff$ the triple set $\mathbb{S}$ is consistent.

A reconciliation map μ from (T,t) to the species tree S can be constructed in polynomial time.

From Event-Labeled Gene Trees to Species Trees., H.-Rosales M, Hellmuth M, Huber K, Moulton V, Wieseke N, Stadler PF, BMC Bioinformatics, 2012

Inferring the Species Tree in $O(|\mathfrak{G}||\mathfrak{S}|)$ time

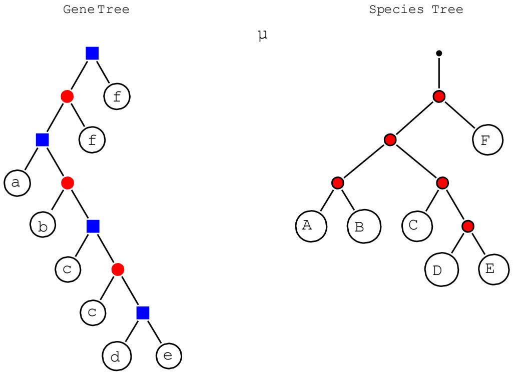
Given: Gene tree $(T,t) = ((V,E),t)$, Gene set $\mathfrak{G} \subseteq V$
Consistent triple set $\mathbb{S}$ Species set $\mathfrak{S}$
map $\sigma : \mathfrak{G} \to \mathfrak{S}$ from genes to its respective species.



Inferring the Species Tree in $O(|\mathfrak{G}||\mathfrak{S}|)$ time

Given: Gene tree $(T,t) = ((V,E),t)$, Gene set $\mathfrak{G} \subseteq V$
Consistent triple set $\mathbb{S}$ Species set $\mathfrak{S}$
map $\sigma : \mathfrak{G} \to \mathfrak{S}$ from genes to its respective species.

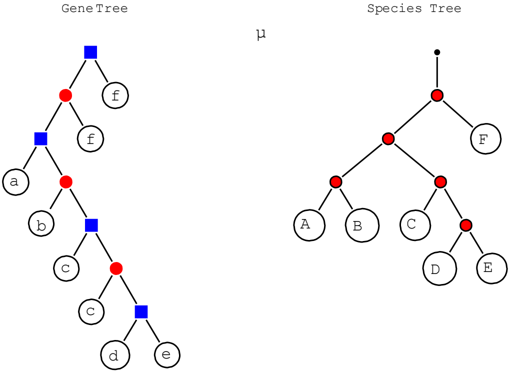
1. Construct a species tree S=(W,F) from $\mathbb{S}$ (e.g. with `Build`).



Inferring the Species Tree in $O(|\mathfrak{G}||\mathfrak{S}|)$ time

Given: Gene tree $(T,t) = ((V,E),t)$, Gene set $\mathfrak{G} \subseteq V$
Consistent triple set $\mathbb{S}$ Species set $\mathfrak{S}$
map $\sigma : \mathfrak{G} \to \mathfrak{S}$ from genes to its respective species.

1. Construct a species tree S=(W,F) from $\mathbb{S}$ (e.g. with `Build`).
2. Construct the reconciliation map $\mu : V \to W \cup F$ as follows:

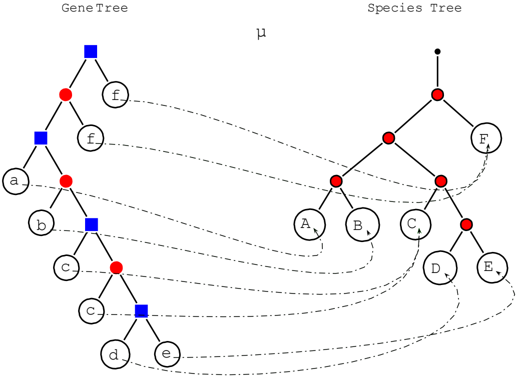


Inferring the Species Tree in $O(|\mathfrak{G}||\mathfrak{S}|)$ time

Given: Gene tree $(T,t) = ((V,E),t)$, Gene set $\mathfrak{G} \subseteq V$
Consistent triple set $\mathbb{S}$ Species set $\mathfrak{S}$
map $\sigma : \mathfrak{G} \to \mathfrak{S}$ from genes to its respective species.

1. Construct a species tree S=(W,F) from $\mathbb{S}$ (e.g. with `Build`).
2. Construct the reconciliation map $\mu : V \to W \cup F$ as follows:

- $\mu(x) = \sigma(x)$ for all genes $x \in \mathfrak{G}$.

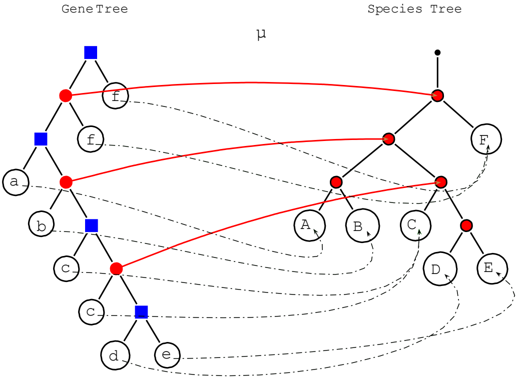


Inferring the Species Tree in $O(|\mathfrak{G}||\mathfrak{S}|)$ time

Given: Gene tree $(T,t) = ((V,E),t)$, Gene set $\mathfrak{G} \subseteq V$
Consistent triple set $\mathbb{S}$ Species set $\mathfrak{S}$
map $\sigma : \mathfrak{G} \to \mathfrak{S}$ from genes to its respective species.

1. Construct a species tree S=(W,F) from $\mathbb{S}$ (e.g. with `Build`).
2. Construct the reconciliation map $\mu : V \to W \cup F$ as follows:

- $\mu(x) = \sigma(x)$ for all genes $x \in \mathfrak{G}$.
- $\mu(x) = \mathrm{lca}_S(\sigma(L(x)))$ if $t(x) = \bullet =$ *speciation*

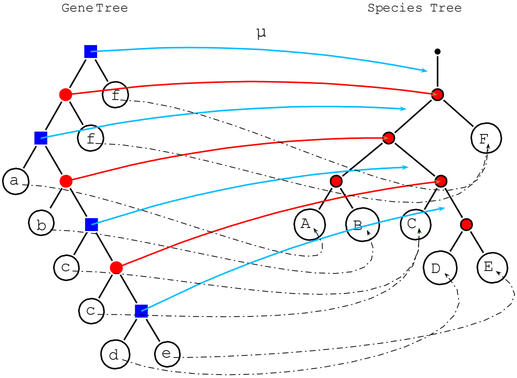


Inferring the Species Tree in $O(|\mathfrak{G}||\mathfrak{S}|)$ time

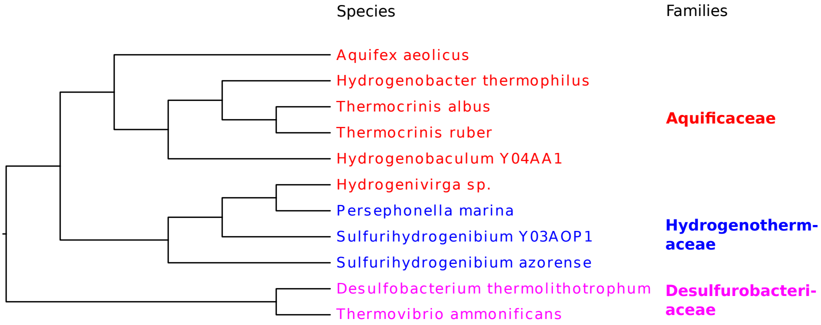
Given: Gene tree $(T,t) = ((V,E),t)$, Gene set $\mathfrak{G} \subseteq V$
Consistent triple set $\mathbb{S}$ Species set $\mathfrak{S}$
map $\sigma : \mathfrak{G} \to \mathfrak{S}$ from genes to its respective species.

1. Construct a species tree S=(W,F) from $\mathbb{S}$ (e.g. with `Build`).
2. Construct the reconciliation map $\mu : V \to W \cup F$ as follows:

- $\mu(x) = \sigma(x)$ for all genes $x \in \mathfrak{G}$.
- $\mu(x) = \mathrm{lca}_S(\sigma(L(x)))$ if $t(x) = \bullet =$ *speciation*
- $\mu(x) = [u, \mathrm{lca}_S(\sigma(L(x)))]$ if $t(x) = \blacksquare =$ *duplication*



Results - Real Life Data



- Class of bacteria that live in harsh environmental settings, e.g., hot springs, sulfur pools, and thermal ocean vents.
- 11 Aquificales species with 2887 gene families
  (1372 - 3809 genes per species)
- `ProteinOrtho` → ILP-pipeline (CE→MCS→LRT).