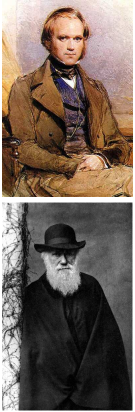
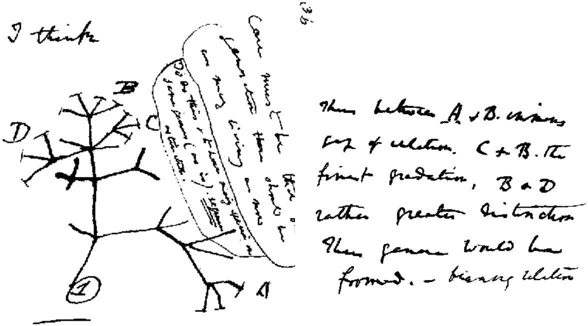
Bioinformatics

(Phylogenetic Tree Reconstruction)

Marc Hellmuth

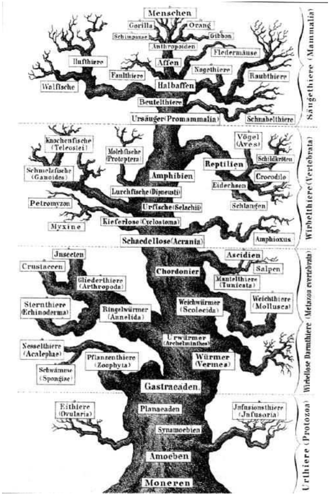
Phylogenetic Reconstruction




36
I think
1
A
B
C
D
Thus between A & B. immense gap of relation. C & B. the finest gradation, B & D rather greater distinction
Thus genera would be formed. — bearing relation

“I think” by Charles Darwin (1837) - One of the first evolutionary trees.

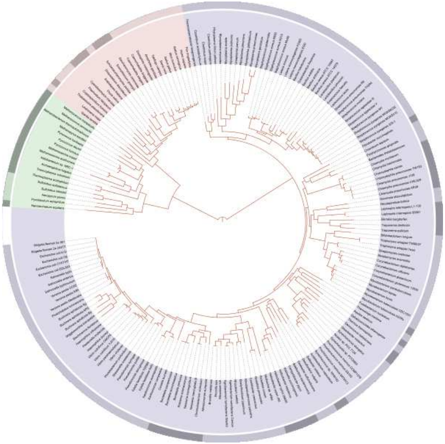
Tree of Live - A Better Picture



Ernst Haeckel, 1879

Tree of Live - A Better Picture[1]

Relationship between species with sequenced genomes as of 2006.



center = last universal ancestor of all life on earth.

three domains of life:

eukaryota (animals, plants and fungi);

bacteria;

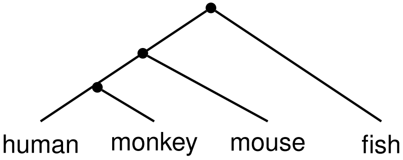
archaea.

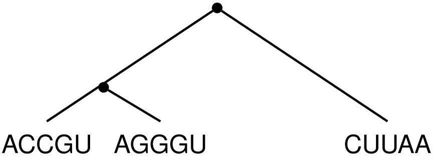
[1] Ciccarelli, FD (2006). "Toward automatic reconstruction of a highly resolved tree of life.". Science; Letunic, I (2007). "Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation.". Bioinformatics

Aim: Assemble a tree representing a hypothesis about the evolutionary history of a set of genes, species or other taxa.

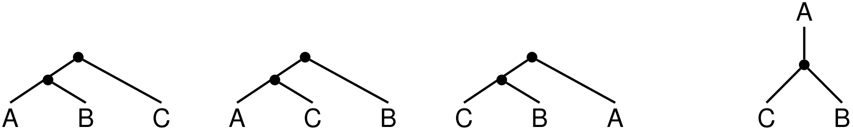
Trees are "good" approximation (does not work if one considers e.g. horizontal gene transfers)

A phylogenetic tree on set of taxa X is tupel (T, λ) s.t. $T = (V, E)$ is unordered tree with unique labels $\lambda(v) \in X$ for all leaves $v \in L \subseteq V$.

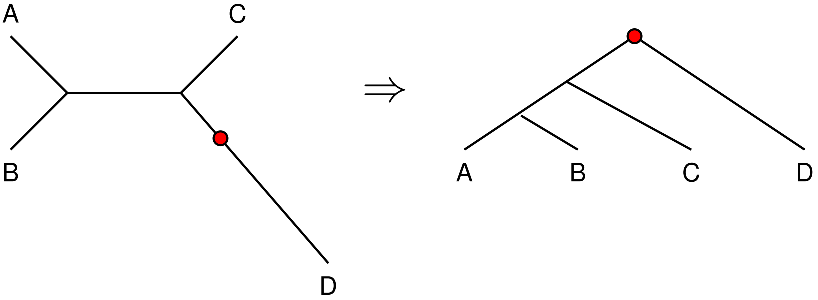




Rooted vs. Unrooted



Unrooted tree (right) “displays” all three rooted trees on three leaves.



Depending on the application, phylogenetic trees may:

- be rooted or unrooted
- have weighted or unweighted edges
- have bounded degree
  (maximum nr of children of each internal node)

The problem in practise:

- Inference of the gene or species tree T is a classical problem of molecular phylogenetics.
  In practice it can only be solved approximately.
- Only the subset of leaves of the species or gene tree corresponding to extant (currently living) species or genes in extant (currently living) species is observable.
- All internal nodes (and the event labeling t) in the gene tree must be inferred from data.
  events: duplication, speciation (Later!)

Lemma

There are $(2n-3)!! = 1 \cdot 3 \cdot \dots \cdot (2n-3)$ *rooted trees with n leaves, and* $(2n-5)!!$ *unrooted trees with n leaves*

	n	3	4	5	6	10	20
Exmpl:	unrooted	1	3	15	105	2’027’025	$2.22 \cdot 10^{20}$
	rooted	3	15	105	945	34’459’425	$8.20 \cdot 10^{21}$

Aim: Assemble a tree representing a hypothesis about the evolutionary history of a set of genes, species or other taxa.

Methods:

- Distance Based e.g.:
  - Ultrametric Tree Reconstruction
  - Additive Tree Reconstruction
- Character Based e.g.:
  - Parsimony Methods
  - Maximum Likelihood
- Consensus Methods e.g.:
  - BUILD

UPGMA

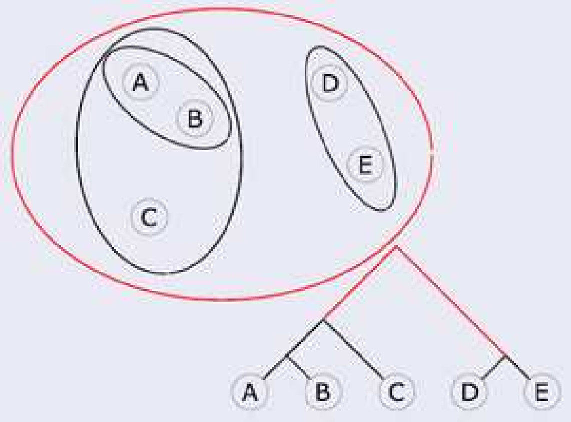
Unweighted **P**air **G**roup **M**ethod with **A**rithmetic Mean

- Assume “constant moleculare clock”:
  one assumes that mutations always appear with the same probability independent from time, location, kind of mutation (mutation = bygone past time)
- The two sequences with with the shortest evolutionary distance between them are assumed to have been the last that diverged, and represented by the most recent internal node.
- Cluster the data and at each step merge clusters.
- Distances between clusters:

$$D(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} D_{x,y}$$

- Moreover, compute “ultrametric trees”.

UPGMA - Idea

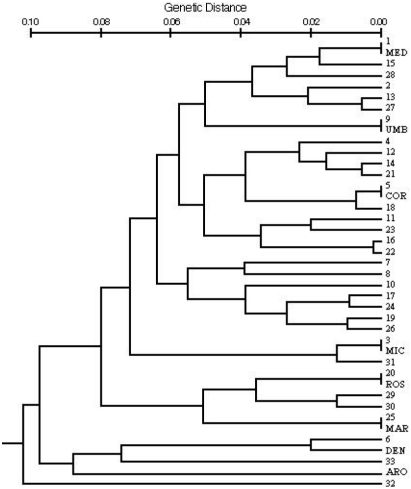


It works correctly, if the underlying “distance-matrix” is an ultrametric

A metric D on $M = \{1, \ldots, n\}$ is an ultrametric if for all $x, y, z \in M$ holds

$$D_{xy} \leq \max\{D_{xz}, D_{zy}\}.$$

Example: Ultrametric Tree [2]



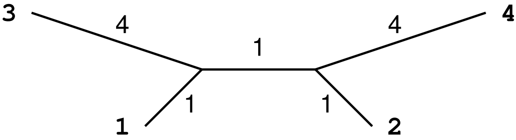
[2] taken from: Evolution of polyploid agamic complexes with examples from Antennaria (Asteraceae), RJ Bayer, Opera Bot, 1996

Neighbor Joining and Additive Tree

For a given $n \times n$ distance matrix D an additive tree T for D is an unrooted tree with

1. T is binary, having n leaves (bijectively labeled by $1, \ldots, n$)
2. each edge (x, y) of T is (positive) weighted with branch length b_{xy}
3. For any pair of leaves i, j it holds: $D_{ij} =$ sum of edge weights b_{xy} along path from i to j in T.

$$D = \begin{pmatrix} 0 & 3 & 5 & 6 \\ & 0 & 6 & 5 \\ & & 0 & 9 \\ & & & 0 \end{pmatrix}$$

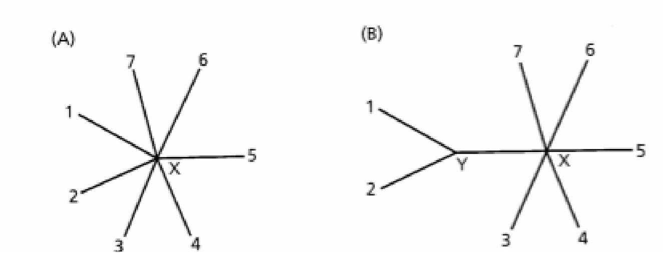


Neighbor Joining (NJ)

NJ does not assume constant molecular clock.

Basis of NJ is concept of minimum evolution, that is, the “true” tree will be that for which the total branch length is shortest.

Idea: Start with “star” tree and separate stepwisely vertices that are together “quite” close and also “quite” far away from the rest until a fully resolved tree has been built. (Note, these two vertices are not necessarily the nearest ones).



It works correctly, if the underlying “distance-matrix” is additive
A metric D on $M = \{1, \dots, n\}$ is additive if for all $x, y, a, b \in M$ holds

$$D_{xy} + D_{ab} \leq \max\{D_{xa} + D_{yb}, D_{xb} + D_{ya}\}.$$

Consensus Methods[3]

Assume a set T of phylogenetic trees has already been constructed.
Aim: Summarize the information in T in the “best way”.
“best way” := find largest subtree, find supertree, ...

[3] parts of this section are based on talk by Jesper Jansson (2010 MSP Annual Convention)

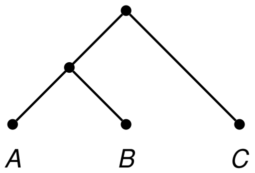
Supertree

Aim: Merge a given set of (possibly conflicting) phylogenetic trees into one tree. Keep as much branching information as possible!
Motivation:

- Combine many trees constructed from different data sets.
  $\rightarrow$ more reliable answers.
- Computationally expensive methods can yield highly accurate trees for small, overlapping subsets of the objects.
- Most individual studies investigate relatively few species.
  Supertrees allow us to deduce new evolutionary relationships.

Rooted Triples

Rooted triplet= rooted binary phylogenetic tree with exactly three leaves.



For three leaves A, B, C in T we write $((A, B), C)$ if the path from A to B does not intersect the path from C to the root ρ.

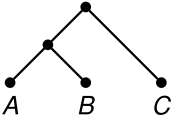
That is the unique rooted triplet with

$$lca(A, B) \prec lca(A, C) = lca(B, C)$$

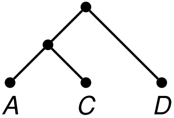
Any rooted phylogenetic tree can be represented by a set of rooted triplets.

Combining Rooted Triples

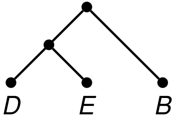
$((A,B)C)$



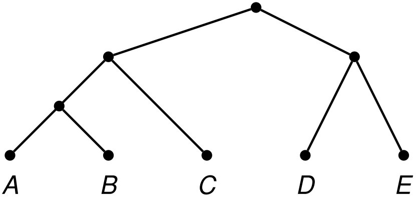
$((A,C)D)$



$((D,E)B)$

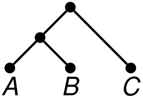


Consensus Tree “displays” all rooted triples:

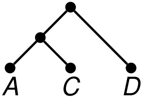


Combining Rooted Triples

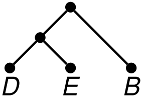
$((A,B)C)$



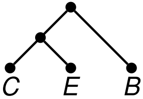
$((A,C)D)$



$((D,E)B)$

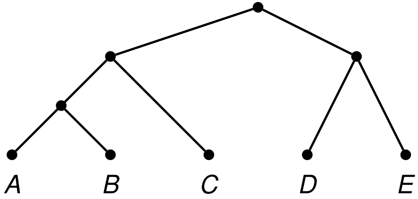


$((C,E)B)$



Consensus Tree does not always exist!!

Consistence



For three leaves A, B, C in T we write $((A, B), C)$ if the path from A to B does not intersect the path from C to the root ρ.

That is the unique rooted triplet with

$$lca(A, B) \prec lca(A, C) = lca(B, C)$$

T and an arbitrary triple $((A, B), C)$ are consistent iff

$$lca(A, B) \prec lca(A, C) = lca(B, C)$$

T displays $((A, B), C)$.

BUILD

Theorem (Aho, Sagiv, Szymanski, Ullman - 1981; Semple & Steel - 2003)

Let $\mathscr{R}$ by a collection of rooted triples with leaf set $\mathscr{L}$. Then there is an $O(|\mathscr{R}||\mathscr{L}|)$ time algorithm – called $\texttt{BUILD}$ – that either

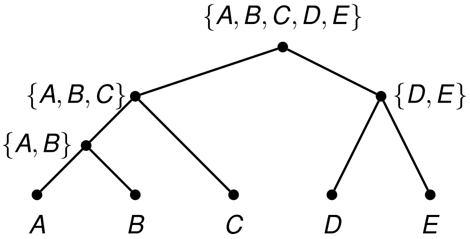
- *constructs a phylogenetic tree $T_{|\mathscr{R}}$ that displays each member of $\mathscr{R}$*

  or

- *recognizes $\mathscr{R}$ as inconsistent.*

BUILD

Idea of this recursive, top-down approach: Partition $\mathscr{L}$ into blocks according to $\mathscr{R}$. Output a tree consisting of a root whose children are roots of the trees obtained by recursing on each block.



BUILD

Let $\mathscr{R}$ be a set of triples defined on a leaf set $\mathscr{L}$.

For any $L \subseteq \mathscr{L}$ define $\mathscr{R}_{|L} = \{((x,y)z) \in \mathscr{R} \mid x,y,z \in L\}$.

To find blocks use auxiliary graph $G(\mathscr{R}_{|L}, L) = (L, E)$ with $(x,y) \in E$ iff there is a triple $((x,y)z) \in \mathscr{R}_{|L}$

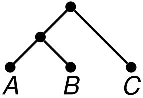
BUILD

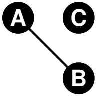
Let $\mathscr{R}$ be a set of triples defined on a leaf set $\mathscr{L}$.

For any $L \subseteq \mathscr{L}$ define $\mathscr{R}_{|L} = \{((x,y)z) \in \mathscr{R} \mid x,y,z \in L\}$.

To find blocks use auxiliary graph $G(\mathscr{R}_{|L}, L) = (L, E)$ with $(x,y) \in E$ iff there is a triple $((x,y)z) \in \mathscr{R}_{|L}$

Exmpl: $L = \{A,B,C\}$, $\mathscr{R} = ((A,B)C)$, $G(\mathscr{R}_{|L}, L)$





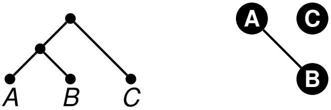
BUILD

Let $\mathscr{R}$ be a set of triples defined on a leaf set $\mathscr{L}$.

For any $L \subseteq \mathscr{L}$ define $\mathscr{R}_{|L} = \{((x,y)z) \in \mathscr{R} \mid x,y,z \in L\}$.

To find blocks use auxiliary graph $G(\mathscr{R}_{|L}, L) = (L, E)$ with $(x,y) \in E$ iff there is a triple $((x,y)z) \in \mathscr{R}_{|L}$

Exmpl: $L = \{A,B,C\}$, $\mathscr{R} = ((A,B)C)$, $G(\mathscr{R}_{|L}, L)$



Crucial observation: If ((xy)z) is consistent with a tree T then the leaves labeled by x and y cannot descend from two different children of the root of T, i.e., x and y must belong to the same block.

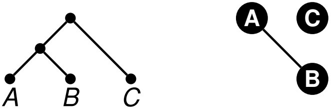
BUILD

Let $\mathscr{R}$ be a set of triples defined on a leaf set $\mathscr{L}$.

For any $L \subseteq \mathscr{L}$ define $\mathscr{R}_{|L} = \{((x,y)z) \in \mathscr{R} \mid x,y,z \in L\}$.

To find blocks use auxiliary graph $G(\mathscr{R}_{|L}, L) = (L, E)$ with $(x,y) \in E$ iff there is a triple $((x,y)z) \in \mathscr{R}_{|L}$

Exmpl: $L = \{A,B,C\}$, $\mathscr{R} = ((A,B)C)$, $G(\mathscr{R}_{|L}, L)$



Crucial observation: If ((xy)z) is consistent with a tree T then the leaves labeled by x and y cannot descend from two different children of the root of T, i.e., x and y must belong to the same block.

Therefore, the algorithm defines the partition of $L \subseteq \mathscr{L}$ by:
Blocks of leaves iff connected components in $G(\mathscr{R}_{|L}, L)$

BUILD

Lemma (Aho, Sagiv, Szymanski, Ullman (1981), Bryant & Steel (1995))

A given triple set $\mathscr{R}$ on a leaf set $\mathscr{L}$ is consistent if and only if for all $L \subseteq \mathscr{L}$ with $|L| > 1$ the graph $G(\mathscr{R}_{|L}, L)$ is disconnected.

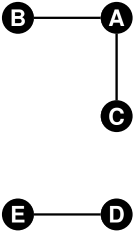
BUILD

```
INPUT: Set of triples in $\mathscr{R}$, leaf set $\mathscr{L}$.
OUTPUT: A rooted, phylog. tree distinctly leaf-labeled by $\mathscr{L}$
  consistent with all rooted triplets in $\mathscr{R}$, if one exists; otherwise null.
compute $G(\mathscr{R},\mathscr{L})$
compute connected components $C_1,\ldots,C_s$ of $G(\mathscr{R},\mathscr{L})$
if $s = 1$ and $|\mathscr{L}| = 1$ then
    return tree $\simeq K_1$
else if $s = 1$ and $|\mathscr{L}| > 1$ then
    return null
else
    for i = $1,\ldots s$ do
        $T_i$ = BUILD($\mathscr{R}_{|V(C_i)}, V(C_i)$)
    end for
    if $T_i \neq null$ for all $i = 1,\ldots s$ then
        attach all of these trees to a common parent node and let $T$ be
          the resulting tree; else $T = null$.
    end if
end if
```

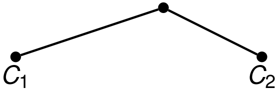
BUILD - Example

$\mathscr{R} = \{((AB)C), ((AC)D), ((DE)B)\}$

$G(\mathscr{R}, \mathscr{L})$:



$\texttt{BUILD}(\mathscr{R}, \mathscr{L} = \{A, B, C, D, E\}$



$C_1 := \texttt{BUILD}(\mathscr{R}_{|\mathscr{L}}, \mathscr{L} = \{A, B, C\})$

$C_2 := \texttt{BUILD}(\mathscr{R}_{|\mathscr{L}}, \mathscr{L} = \{D, E\})$

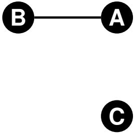
$\texttt{BUILD}$ - Example

$$\mathscr{R} = \{((AB)C), ((AC)D), ((DE)B)\}$$

$$C_1 := \texttt{BUILD}(\mathscr{R}_{|\mathscr{L}}, \mathscr{L} = \{A, B, C\})$$
$$\mathscr{R}_1 := \{((AB)C)\}$$

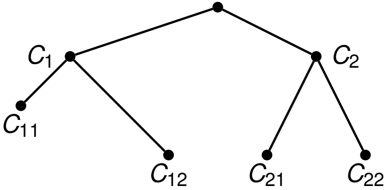
$$C_2 := \texttt{BUILD}(\mathscr{R}_{|\mathscr{L}}, \mathscr{L} = \{D, E\})$$
$$\mathscr{R}_2 := \emptyset$$

$G(\{A, B, C\})$:



$G(\{D, E\})$:

$\texttt{BUILD}(\mathscr{R}, \mathscr{L} = \{A, B, C, D, E\})$



$\texttt{BUILD}$ - Example

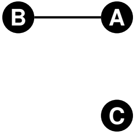
$$\mathscr{R} = \{((AB)C), ((AC)D), ((DE)B)\}$$

$$C_1 := \texttt{BUILD}(\mathscr{R}_{|\mathscr{L}}, \mathscr{L} = \{A, B, C\})$$

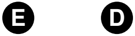
$$C_2 := \texttt{BUILD}(\mathscr{R}_{|\mathscr{L}}, \mathscr{L} = \{D, E\})$$

$$C_{11} := \texttt{BUILD}(\mathscr{R}_{|\mathscr{L}}, \mathscr{L} = \{A, B\})$$
$$C_{12} := \texttt{BUILD}(\emptyset, \{C\})$$
$$C_{21} := \texttt{BUILD}(\emptyset, \{D\})$$
$$C_{22} := \texttt{BUILD}(\emptyset, \{E\})$$

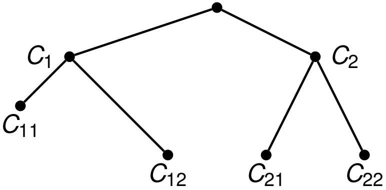
$G(\{A, B, C\})$:



$G(\{D, E\})$:

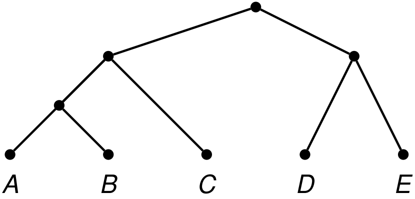


$\texttt{BUILD}(\mathscr{R}, \mathscr{L} = \{A, B, C, D, E\})$



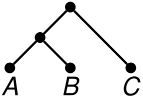
BUILD - Example

$\texttt{BUILD}(\mathscr{R}, \mathscr{L} = \{A, B, C, D, E\})$

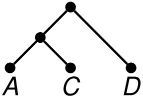


BUILD - Example

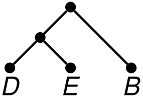
$((A, B)C)$



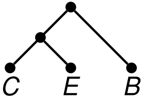
$((A, C)D)$



$((D, E)B)$

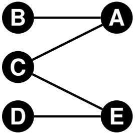


$((C, E)B)$



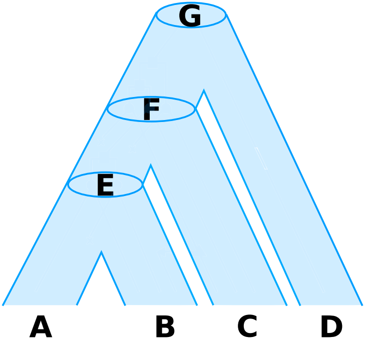
Consensus Tree does not always exist!!

$G(\mathscr{R}, \mathscr{L})$:



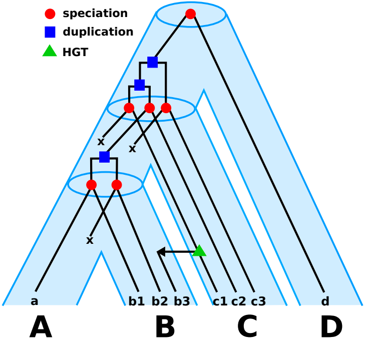
Phylogenetics with Evolutionary Event Relations

The “true” evolutionary History



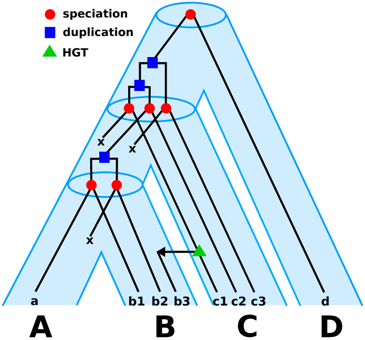
The “true” evolutionary History

- species are characterized by its genome: a “bag of genes”
- “Genes” evolve along a *rooted* tree with unique *event labeling* $t : V^0 \rightarrow M = \{●, ■, ▲\}$

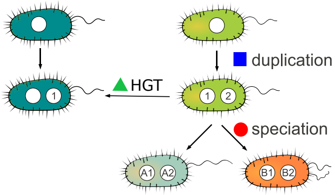


The “true” evolutionary History

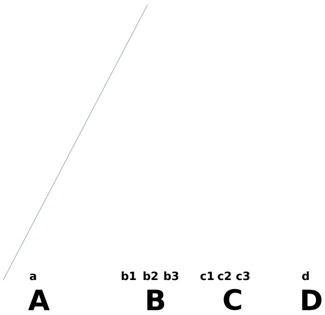
- species are characterized by its genome: a “bag of genes”
- “Genes” evolve along a *rooted* tree with unique *event labeling* $t : V^0 \rightarrow M = \{\bullet, \blacksquare, \blacktriangle\}$



- ■ Gene duplication : an offspring has two copies of a single gene of its ancestor
- ● Speciation : two offspring species inherit the entire genome of their common ancestor
- ▲ HGT : transfer of genes between organisms in a manner other than traditional reproduction and across different species

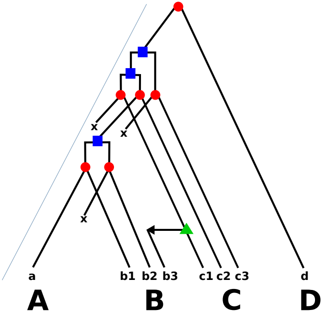


The Problem in Practice



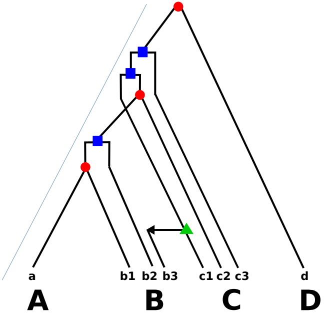
- Only the subset of leaves of the gene tree corresponding to genes in extant (currently living) species is observable.
- All internal nodes and the event labelling *t* in the gene tree must be inferred from data.
- We cannot observe and reconstruct all events (losses).
- The events and the topology of the gene tree can be used (under several constraints) to infer the species tree (Reconciliation)

The Problem in Practice



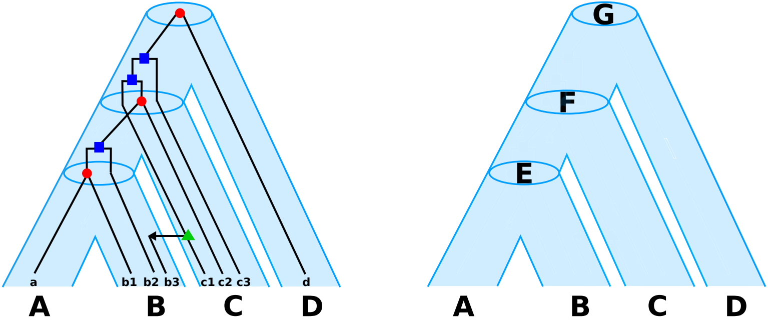
- Only the subset of leaves of the gene tree corresponding to genes in extant (currently living) species is observable.
- All internal nodes and the event labelling *t* in the gene tree must be inferred from data.
- We cannot observe and reconstruct all events (losses).
- The events and the topology of the gene tree can be used (under several constraints) to infer the species tree (Reconciliation)

The Problem in Practice



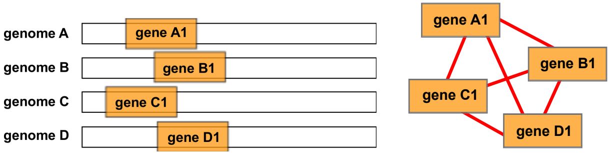
- Only the subset of leaves of the gene tree corresponding to genes in extant (currently living) species is observable.
- All internal nodes and the event labelling *t* in the gene tree must be inferred from data.
- We cannot observe and reconstruct all events (losses).
- The events and the topology of the gene tree can be used (under several constraints) to infer the species tree (Reconciliation)

The Problem in Practice



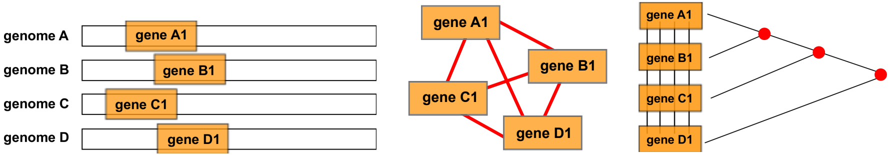
- Only the subset of leaves of the gene tree corresponding to genes in extant (currently living) species is observable.
- All internal nodes and the event labelling *t* in the gene tree must be inferred from data.
- We cannot observe and reconstruct all events (losses).
- The events and the topology of the gene tree can be used (under several constraints) to infer the species tree (Reconciliation)

State-of-the-Art Tree Reconstruction



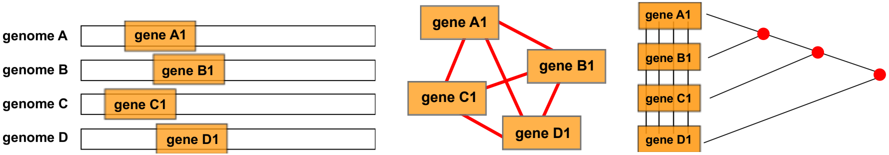
- Find 1:1-orthologs.
  - Paralogs = dangerous nuisance that has to be detected and removed.
  - Select families of genes that rarely exhibit duplications (e.g. rRNAs, ribosomal proteins)

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.

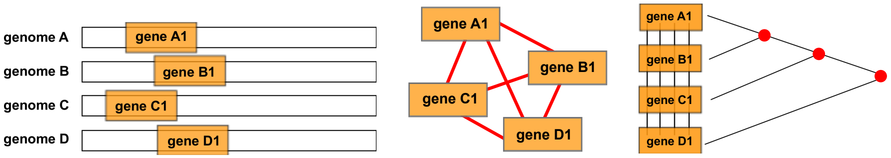
State-of-the-Art Tree Reconstruction



Pitfalls:

- Information of evolutionary events as paralogs or xenologs is ignored, although they might contain valuable information about the evolutionary history of the species.
- The set of usable gene sets is strongly restricted ($\leq 10\%$).

State-of-the-Art Tree Reconstruction

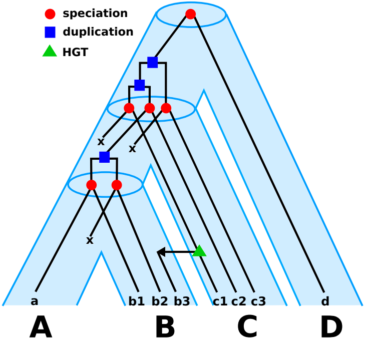


Pitfalls:

- Information of evolutionary events as paralogs or xenologs is ignored, although they might contain valuable information about the evolutionary history of the species.
- The set of usable gene sets is strongly restricted ($\leq 10\%$).

Thus, to get a better picture of the species evolution we try to include also the information of paralogs and xenologs.

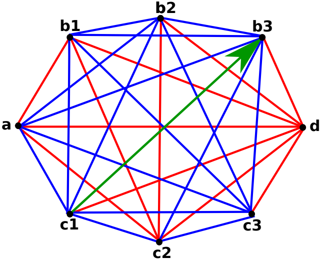
Tree-Representable Sets of Binary Relations

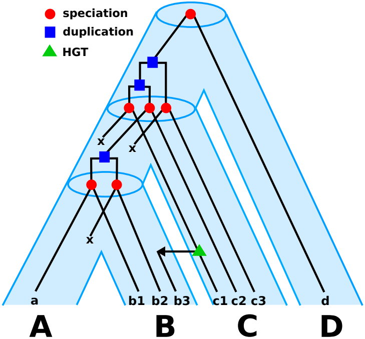


An ordered pair (x, y) of two genes comprises

- orthologs if $\text{lca}(x, y) =$ ● $=$*speciation*
- paralogs if $\text{lca}(x, y) =$ ■ $=$*duplication*
- xenologs if $\text{lca}(x, y) =$ ▲ $=$*HGT* and ▲ “points from” x to y in T

Tree-Representable Sets of Binary Relations

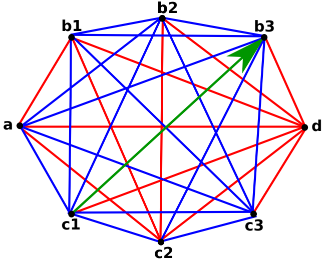


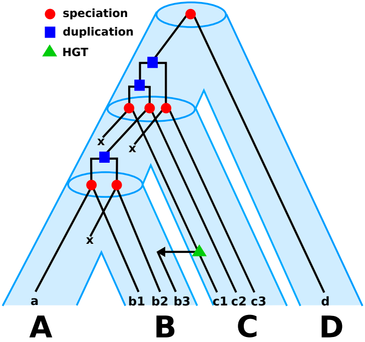


The gene-tree determines three distinct relations

- $R_\bullet$, the orthologs ($\mathrm{lca}(x,y) = \bullet$)
- $R_\blacksquare$, the paralogs ($\mathrm{lca}(x,y) = \blacksquare$)
- $R_\blacktriangle$, the xenologs ($\mathrm{lca}(x,y) = \blacktriangle$, $\blacktriangle$ “points from” x to y in T)

Tree-Representable Sets of Binary Relations





Orthologs, Paralogs (and to some extent HGT) can be estimated without inferring a gene- or species trees.

Assume we have *estimated* binary relations $R_1, \ldots, R_k$ s.t.

$$(xy) \in R_i \text{ iff } \mathrm{lca}(xy) = i \text{ in ordered tree } T$$

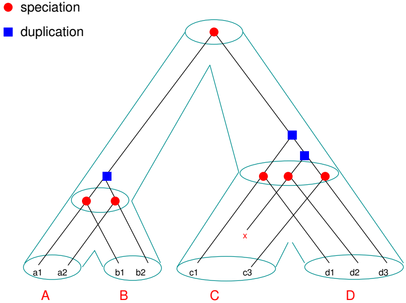
Thus, it is important to understand, when these estimates $R_1, \ldots, R_k$ can be "represented" in a single tree — thus, the edge-colored graph-representation.

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. $s(x,y) \approx \max_{x\in A, z\in B} s(x,z)$,

     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: $(x,y) \in \Theta$ iff $t(x,y) =$ ● $=$ speciation

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?

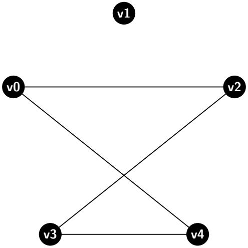
Characterization of Θ

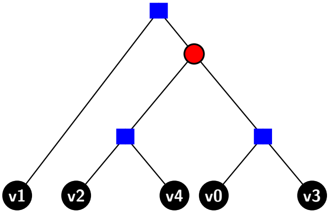
Question: When does the initial estimate $\widehat{\Theta}$ fit the data?

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.

- $t(\mathrm{lca}(x,y)) = \bullet =$ *speciation* for all $(x,y) \in \widehat{\Theta}$ and
- $t(\mathrm{lca}(x,y)) = \blacksquare =$ *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.

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

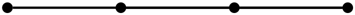
Cograph (=Complement reducible graph)

Corneil et al., 1981:

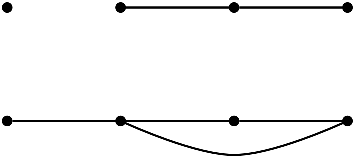
Cographs are defined recursively (Def. omitted)

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

Forbidden:



Allowed:



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

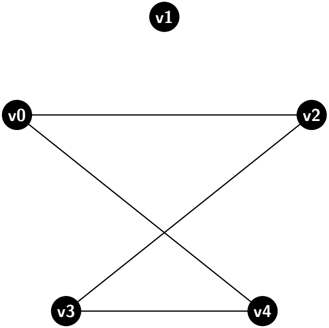
Cograph (=Complement reducible graph)

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.



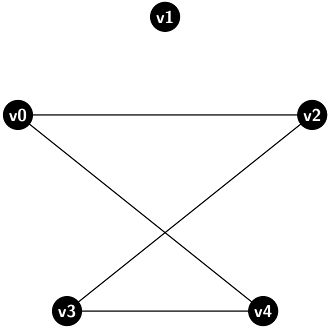
Cograph (=Complement reducible graph)

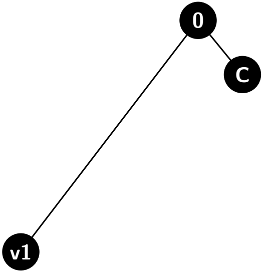
Corneil et al., 1981:

Cographs are defined recursively (Def. omitted)

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

Every Cograph is associated with a unique Cotree.





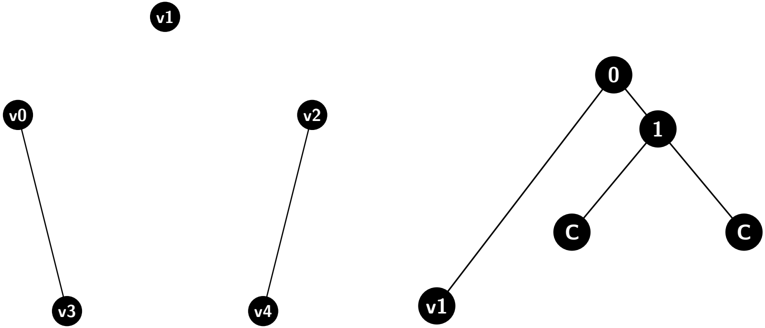
Cograph (=Complement reducible graph)

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.



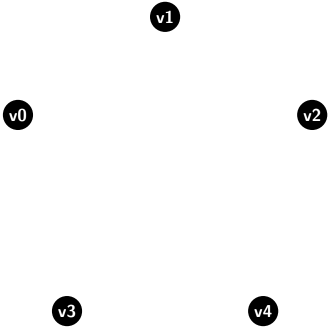
Cograph (=Complement reducible graph)

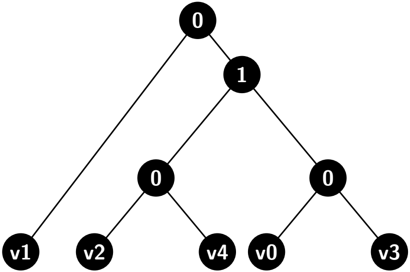
Corneil et al., 1981:

Cographs are defined recursively (Def. omitted)

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

Every Cograph is associated with a unique Cotree.





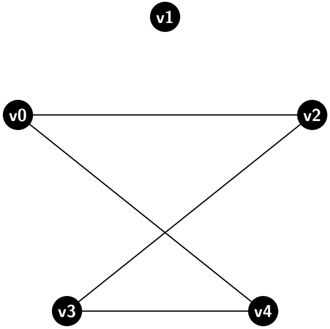
Cograph (=Complement reducible graph)

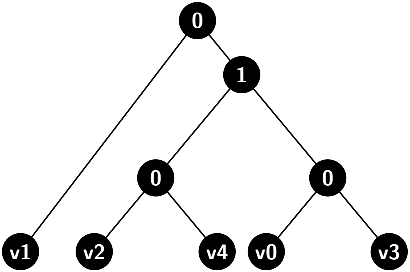
Corneil et al., 1981:

Cographs are defined recursively (Def. omitted)

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

Every Cograph is associated with a unique Cotree.





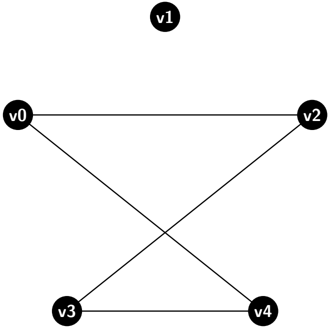
Cograph (=Complement reducible graph)

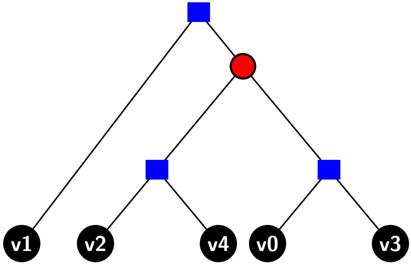
Corneil et al., 1981:

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 $\mathrm{lca}(x, y) = 1 =$ ●

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

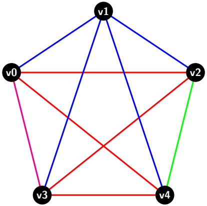
Characterization of Θ

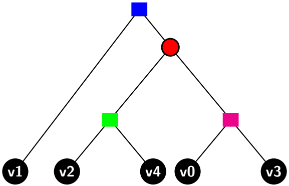
Idea: Correct the initial estimate $\widehat{\Theta}$ to the "closest" orthology relation Θ that fits the data.

Theorem

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

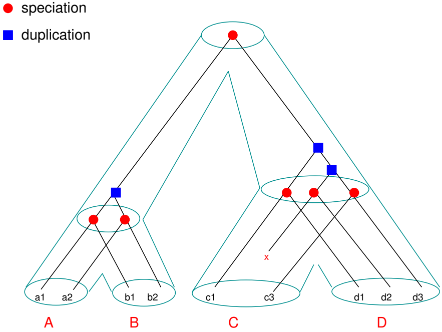
QUESTION

Assume we have a valid orthology relation.

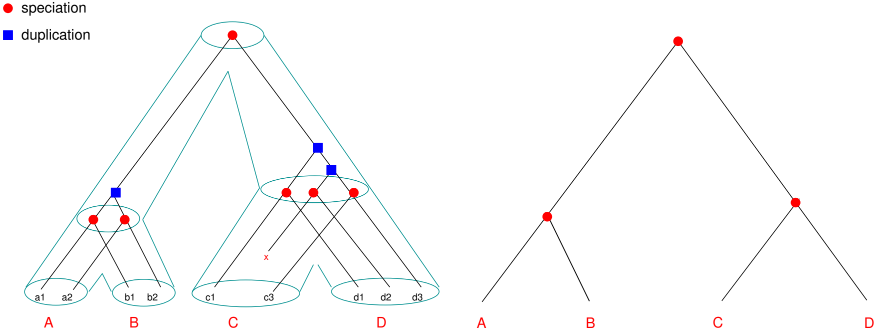
Therefore, we obtain an event-labeled gene tree.

How can we infer the species tree?

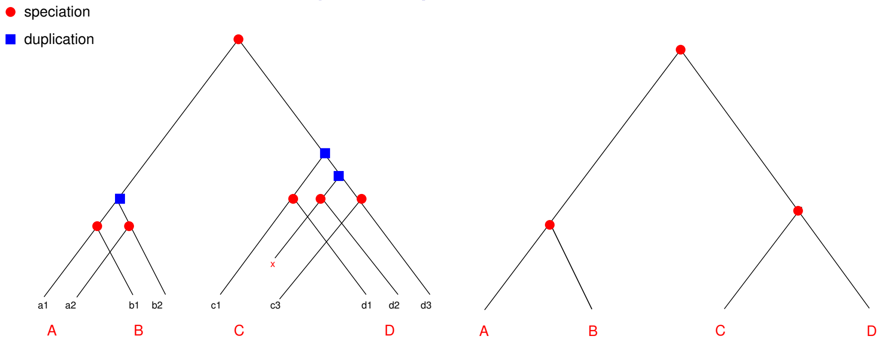
Finding the species trees



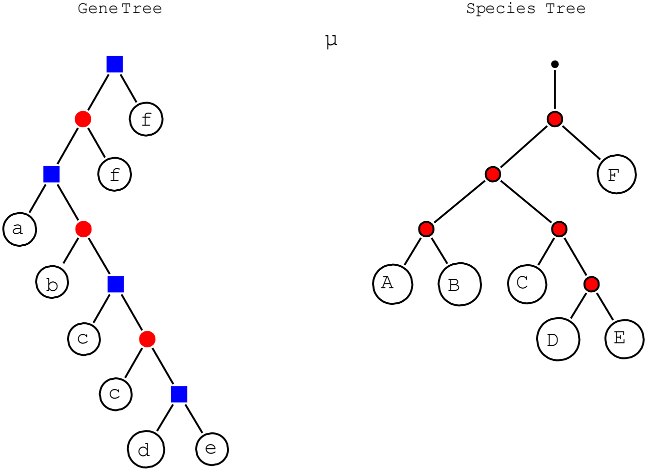
Finding the species trees



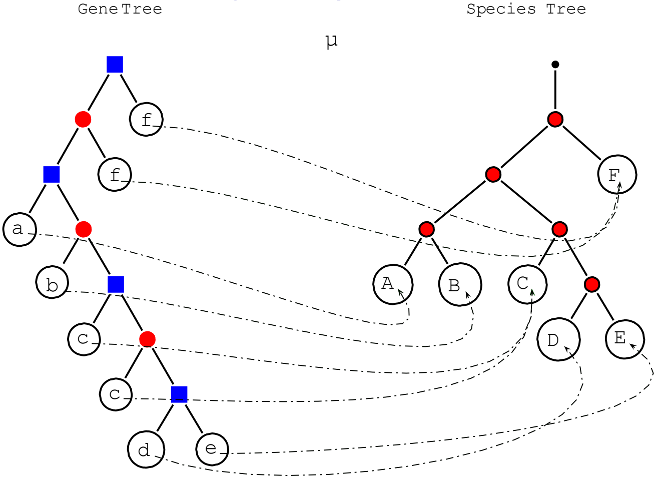
Finding the species trees



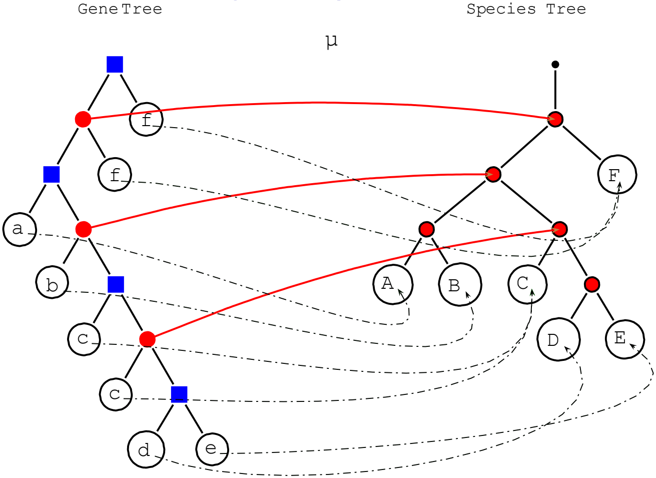
Finding the species trees



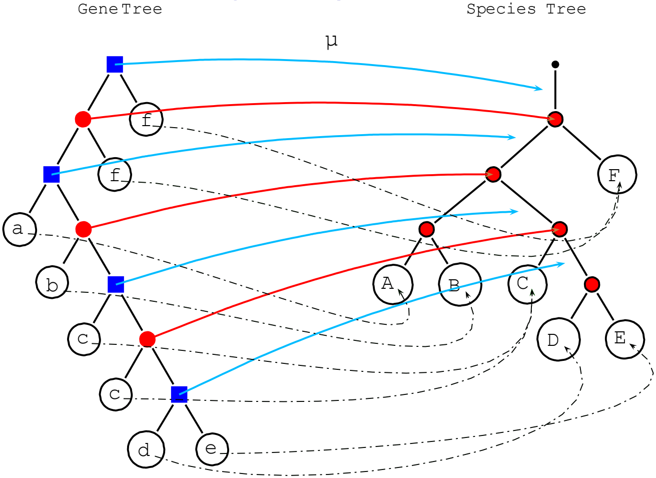
Finding the species trees



Finding the species trees

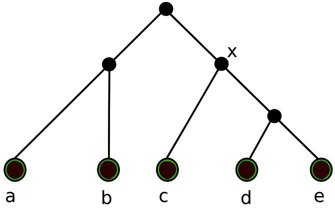


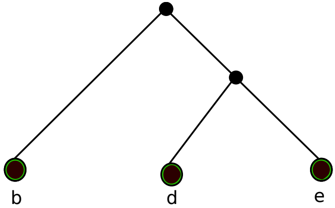
Finding the species trees



Question: When does there exist a species tree for a given gene tree and a reconciliation map μ between them?

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.

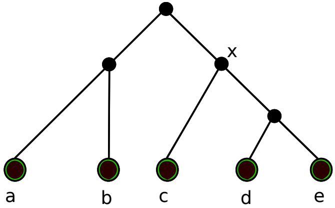
Right Tree:

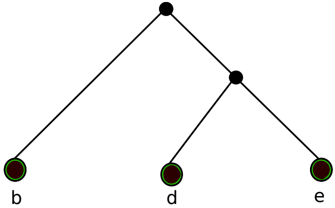
$\mathscr{R}(T) = \{\text{de}|\text{b}\}$

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.

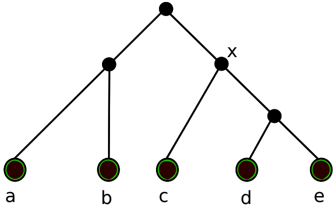
Right Tree:
$\mathscr{R}(T) = \{\text{de|b}\}$

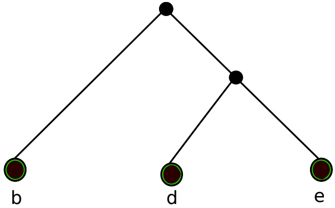
Left Tree:
$\mathscr{R}(T) = \{\text{ab|c}, \text{ab|d}, \text{ab|e}, \text{de|a}, \text{de|b}, \text{de|c}, \text{cd|a}, \text{cd|b}, \text{ce|a}, \text{ce|b}\}$

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

Trees and triples





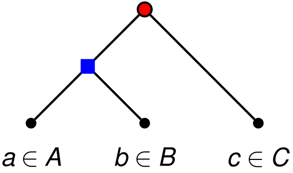
For three leaves a, b, c in T we write ab|c if the path from a to b does not intersect the path from c to the root.

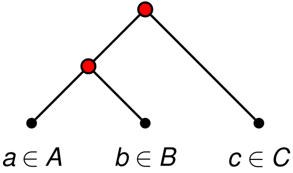
Right Tree:
$\mathscr{R}(T) = \{\text{de}|\text{b}\}$

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

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 $\mathrm{ab|c} \in \mathscr{R}(T)$.
We write $\mathrm{ab|c}^{\bullet}$ if

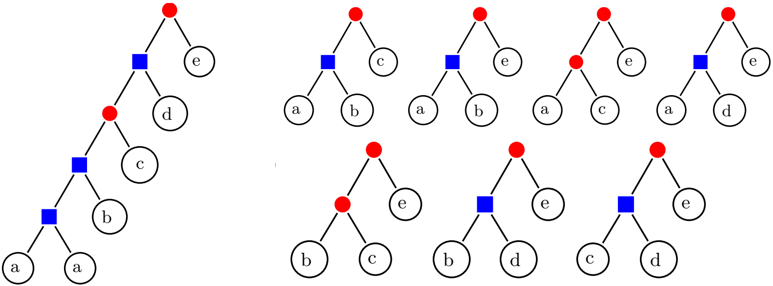
$$t(\mathrm{lca}(a,b,c)) = \bullet = \text{``}\textit{speciation}''$$

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

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

Triples for inferring the species tree

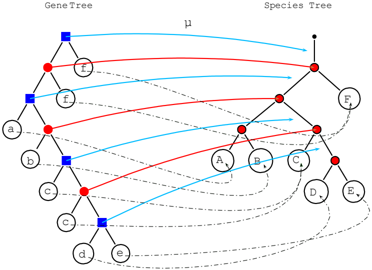
$$\mathbb{S} = \{(\mathsf{AB}|\mathsf{C} : \exists\ \mathsf{ab}|\mathsf{c}^{\bullet} \text{ with } a \in A, b \in B, c \in C\}$$



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

Triples for inferring the species tree

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



Theorem

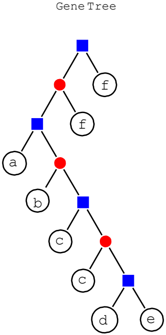
There is a species tree for the gene tree (T, t), i.e., for the symbolic representation of $\Theta \iff$ the triple set $\mathbb{S}$ is consistent.

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

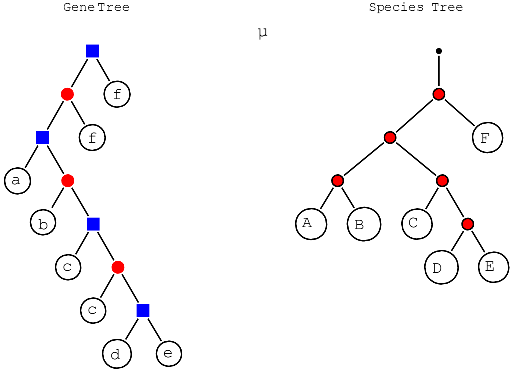
Given: Gene tree $(T,t) = ((V,E),t)$, Gene set $\mathfrak{G} \subseteq V$
Consistent triple set $\mathbb{S}$ Species set $\mathfrak{S}$
map $\sigma : \mathfrak{G} \to \mathfrak{S}$ from genes to its respective species.



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.

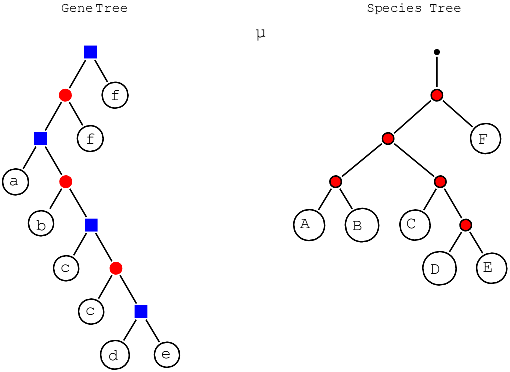
1. Construct a species tree S=(W,F) from $\mathbb{S}$ (e.g. with `Build`).



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:

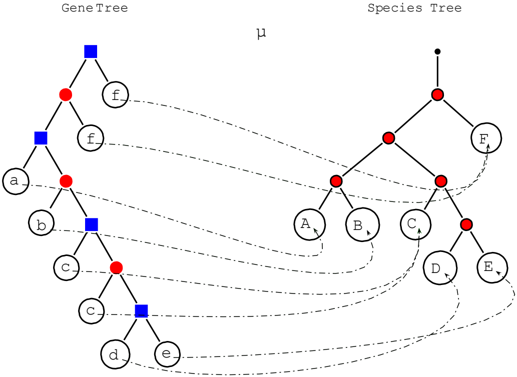


Inferring the Species Tree in $O(|\mathfrak{G}||\mathfrak{S}|)$ time

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 $x \in \mathfrak{G}$.

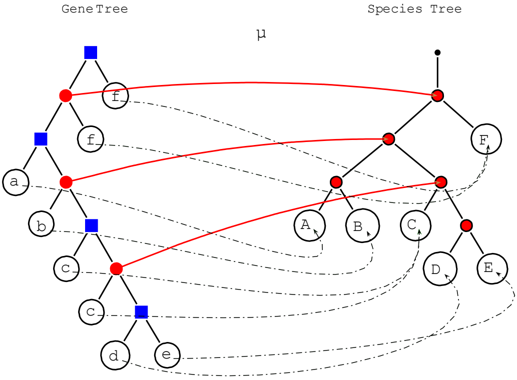


Inferring the Species Tree in $O(|\mathfrak{G}||\mathfrak{S}|)$ time

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*

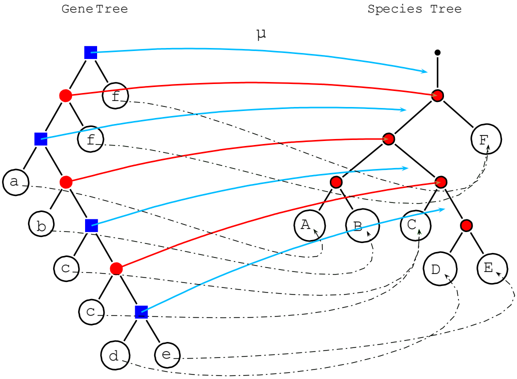


Inferring the Species Tree in $O(|\mathfrak{G}||\mathfrak{S}|)$ time

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.

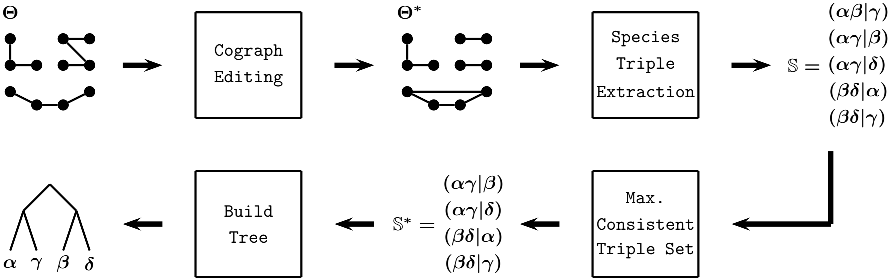
1. Construct a species tree S=(W,F) from $\mathbb{S}$ (e.g. with `Build`).
2. Construct the reconciliation map $\mu : V \to W \cup F$ as follows:

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



Workflow **ParaPhylo**

Given a binary relation Θ comprising e.g. the estimated orthologs or paralogs.



Cograph Editing

Species Triple Extraction

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

Max. Consistent Triple Set

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

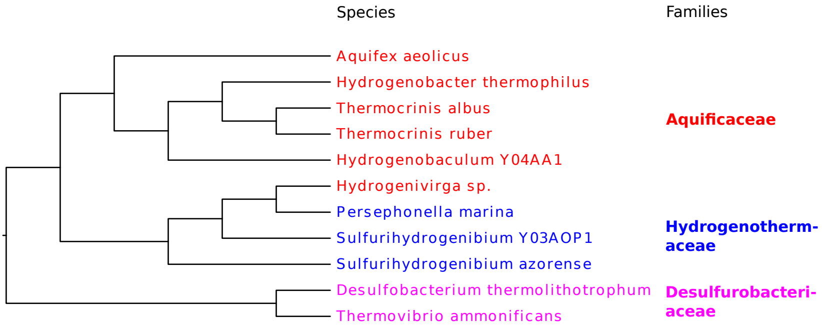
Build Tree

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

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

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

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).

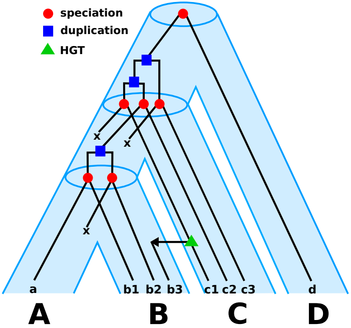
Results: Simulation

Artificial data generated with `ALF`:[0]

Simulation of “true” evol. history

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

Output: Species tree with embedded gene trees and gene-sequences



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

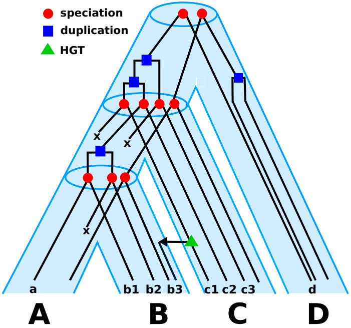
Results: Simulation

Artificial data generated with `ALF`:

Simulation of “true” evol. history

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

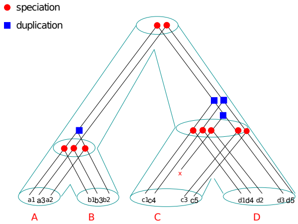
Output: Species tree with embedded gene trees and gene-sequences



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

Results - Simulation without HGT

ALF (no HGT)



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

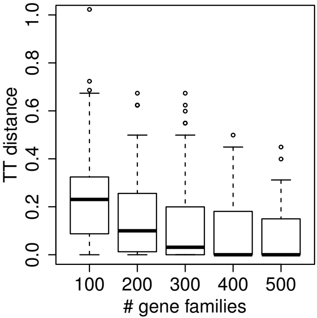
⟶ Compute cotree of $G_{\ominus}$

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

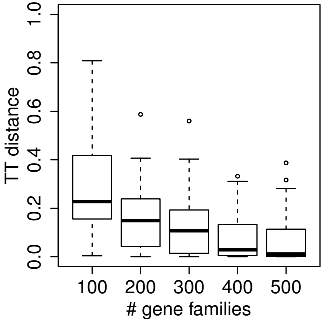
⟶ Compute least resolved species tree and compare it with initial species tree

Results - Simulation without HGT

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



10 species



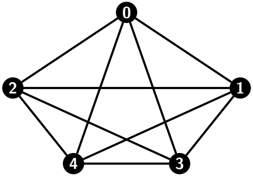
20 species

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

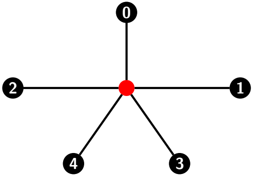
TT distance $\widehat{=}$ “num different triples in initial and reconstructed species tree”

Phylogenomics with Paralogs

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



G_Θ



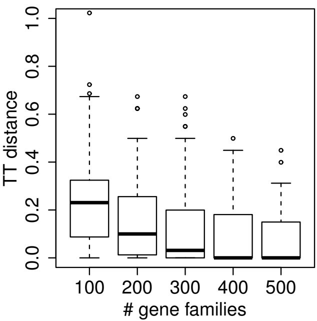
(T, t)

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

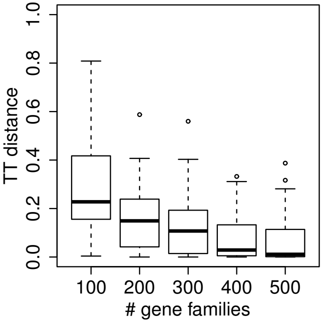
To obtain fully resolved species trees, a sufficient number of gene duplications must have occurred, since the phylogenetic information utilized by our approach is entirely contained in the duplication events.

Results - Simulation without HGT

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



10 species

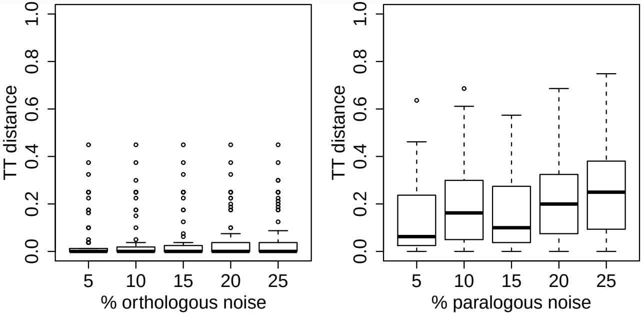


20 species

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

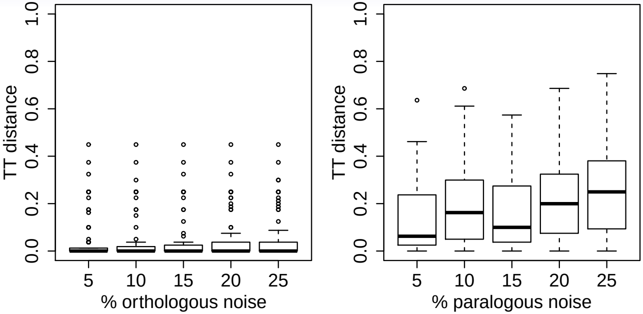
Deviations from perfect reconstructions are exclusively explained by a lack of perfect resolution.

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.

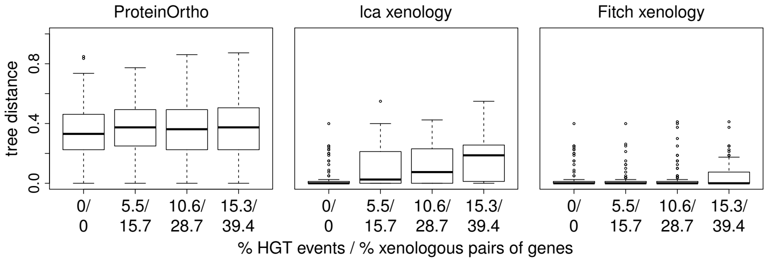
[d] In 95.95% of the simulations the LRT could be found using BUILD.

[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.

Results - HGT



left Θ = “estim.” orthologs via `ProteinOrtho`

middle Θ = orthologs + lca-xenologs

(orthology-overprediction / all paralogs are correctly identified)

right Θ = orthologs + all pairs of genes having at least one HGT event on their path

(orthology-overprediction / all paralogs that are not disturbed by HGT on their paths are correctly identified)

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

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

Summary of the Results

Results:

- We don't need to restrict the dataset to 1:1 orthologs!
- More genefamilies (*incl. paralogs*) → more accurate species trees.
- Don't worry to much about HGT.
- accurate species trees from real data for up to 20 species with $\sim$10000 gene families

Summary of the Results

Results:

- We don't need to restrict the dataset to 1:1 orthologs!
- More genefamilies (*incl. paralogs*) → more accurate species trees.
- Don't worry to much about HGT.
- accurate species trees from real data for up to 20 species with $\sim$10000 gene families

Open Problems and TODO's:

- More accurate orthology prediction methods are needed, or even methods to predict paralogs and xenologs.
- ILP allows to compute *exact* solutions for the NP-hard problems, however, for larger species sets the runtime dramatically increases → We need reliable and efficient heuristics.