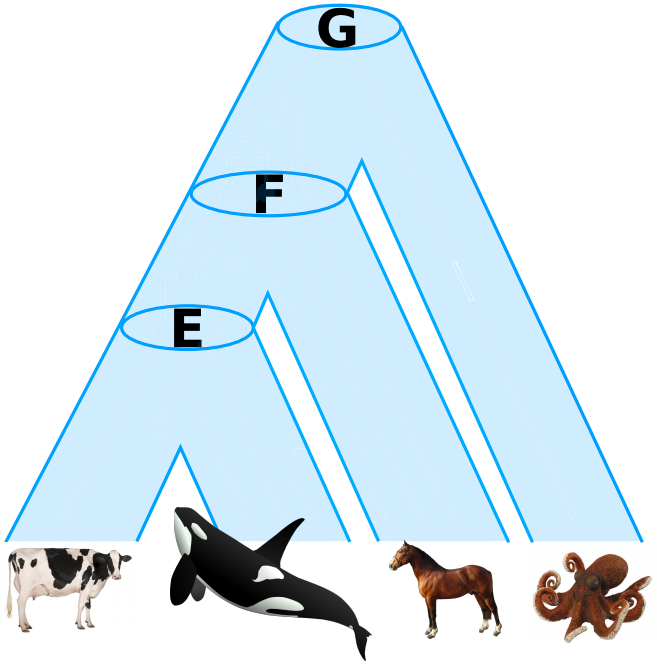
Computational Biology

Comparative Genomics and Phylogenies

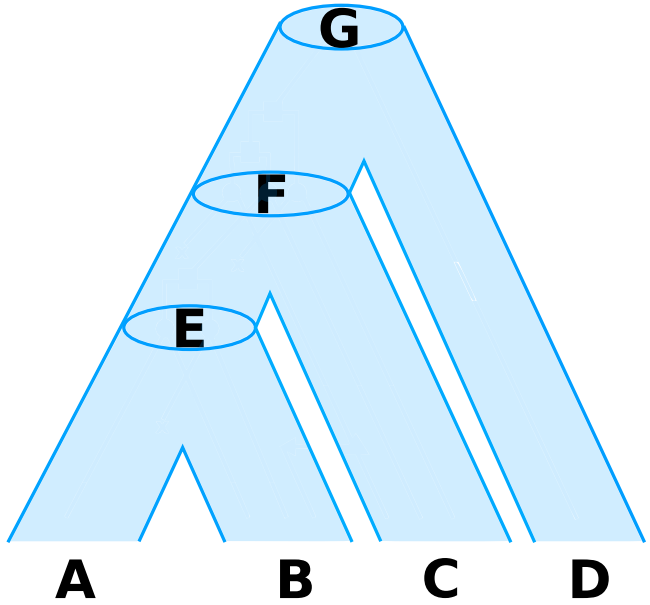
Marc Hellmuth

Department of Mathematics
Stockholm University

The “true” evolutionary history

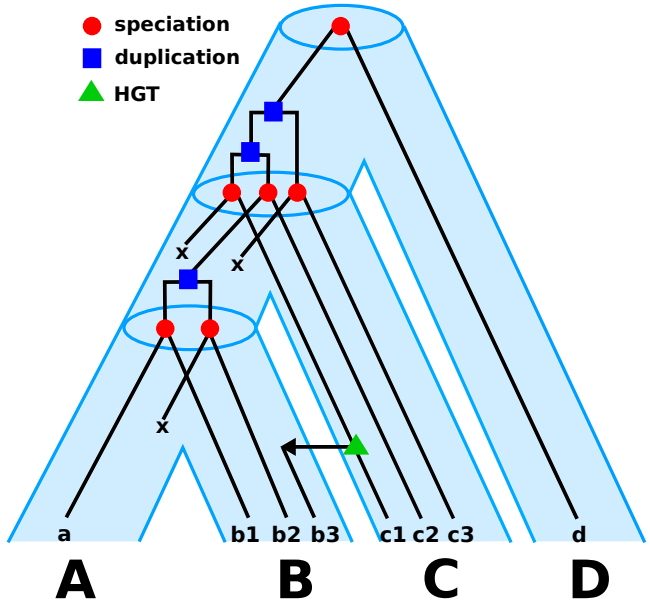


The “true” evolutionary history



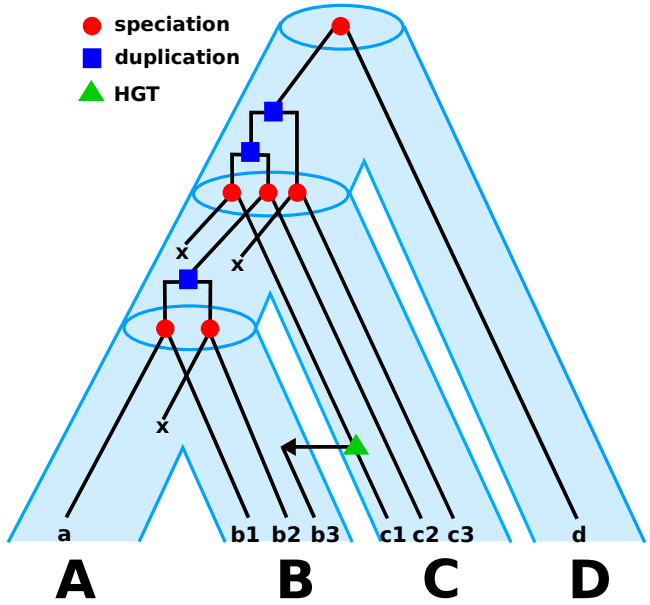
The “true” evolutionary history

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

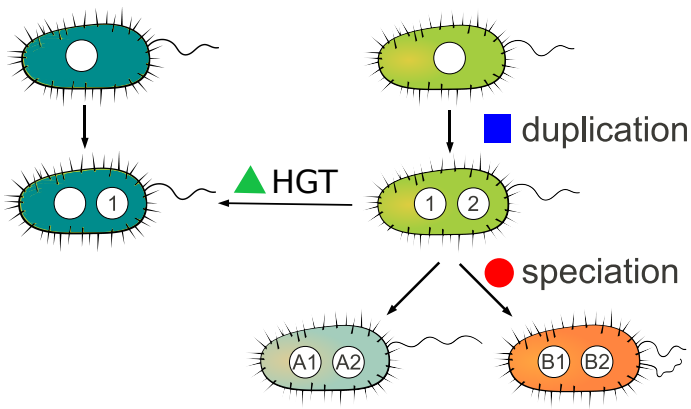


The “true” evolutionary history

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

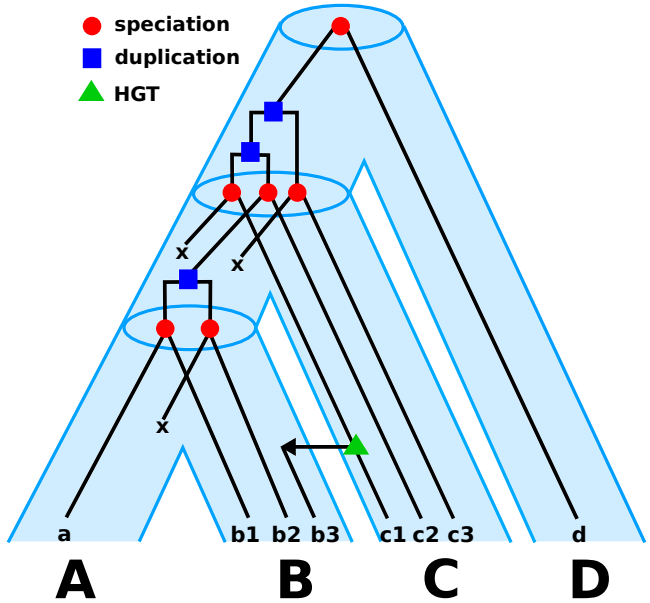


- ■ Gene duplication : an offspring has two copies of a single gene of its ancestor
- • Speciation : two offspring species inherit the entire genome of their common ancestor
- ▲ HGT : transfer of genes between organisms in a manner other than traditional reproduction and across different species



The “true” evolutionary history

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



Homology Relations [binary relations between genes]

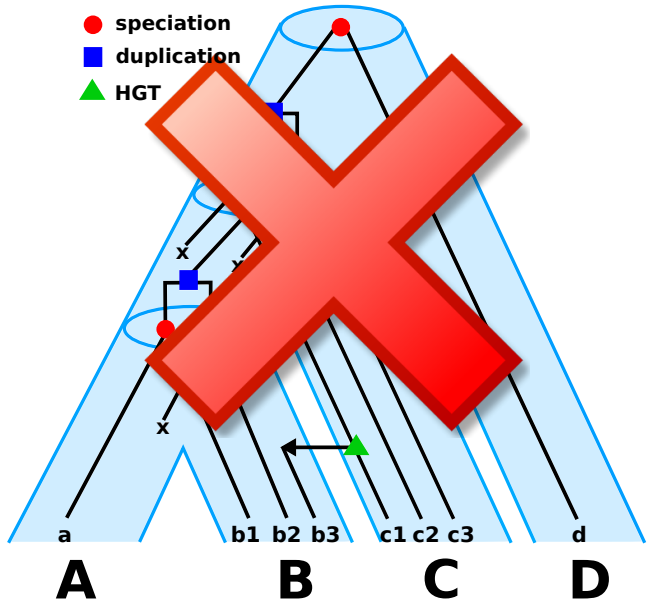
Orthology and **Paralogy** (via vertex colors in T)

Best Matches (evolutionary closest relatives)

Xenology (HGT on the path between two genes)

The “true” evolutionary history

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



Homology Relations [binary relations between genes]

Orthology and **Paralogy** (via vertex colors in T)

Best Matches (evolutionary closest relatives)

Xenology (HGT on the path between two genes)

Knowledge of Hom. Rel. is important !

gene functions & genome annotation $\to$ medicine, drug development, . . .

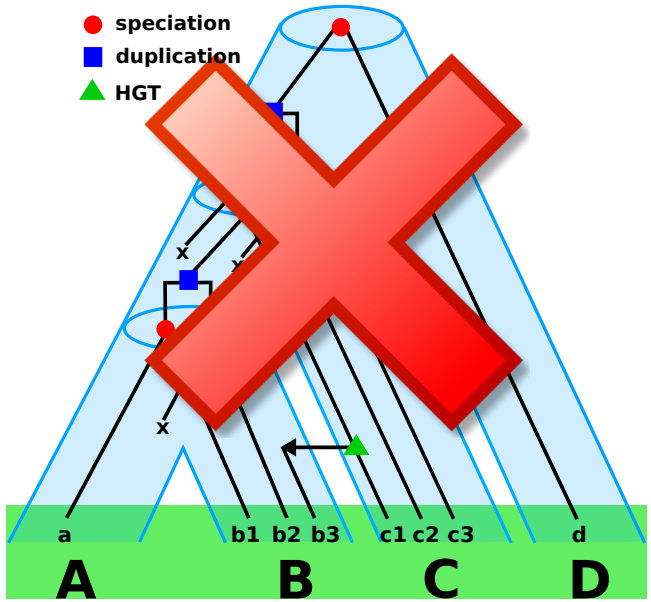
understanding mechanisms that act on genes

reconstruction of species trees

. . .

The “true” evolutionary history

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



Homology Relations [binary relations between genes]

Orthology and **Paralogy** (via vertex colors in T)

Best Matches (evolutionary closest relatives)

Xenology (HGT on the path between two genes)

Knowledge of Hom. Rel. is important !

gene functions & genome annotation $\to$ medicine, drug development, . . .

understanding mechanisms that act on genes

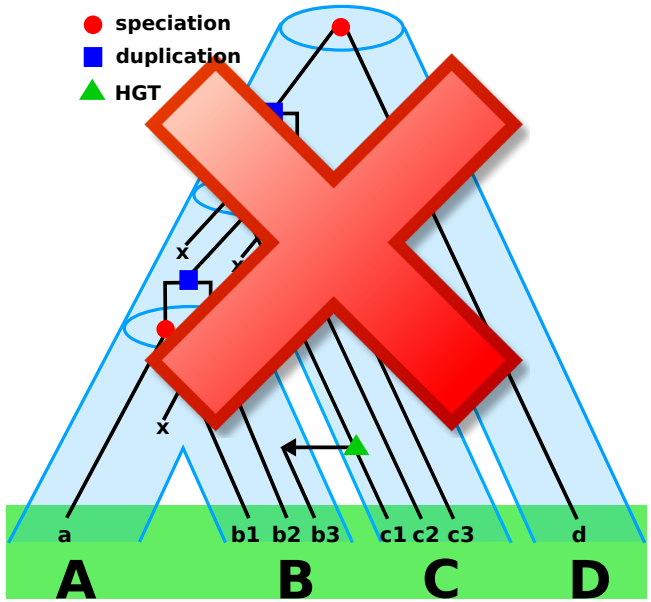
reconstruction of species trees

. . .

Plenty of homology relations exist and are defined in terms of the true history

The “true” evolutionary history

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

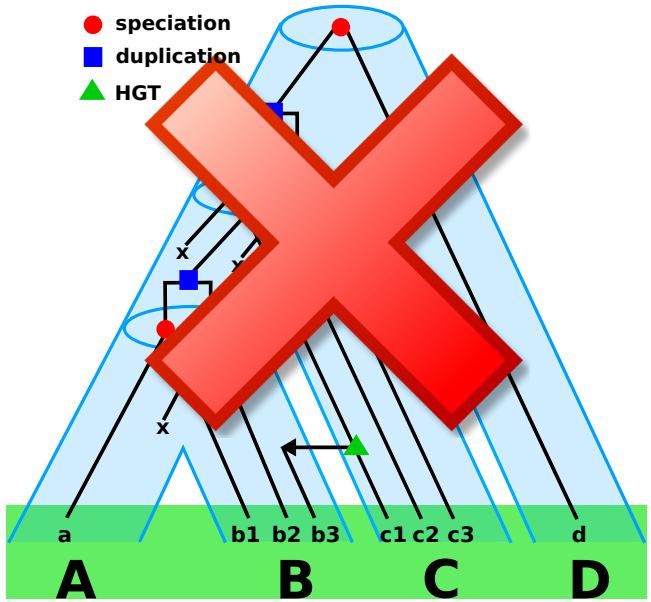


PROBLEM:

- Homology Relations are defined by the true evolutionary scenario!

The “true” evolutionary history

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

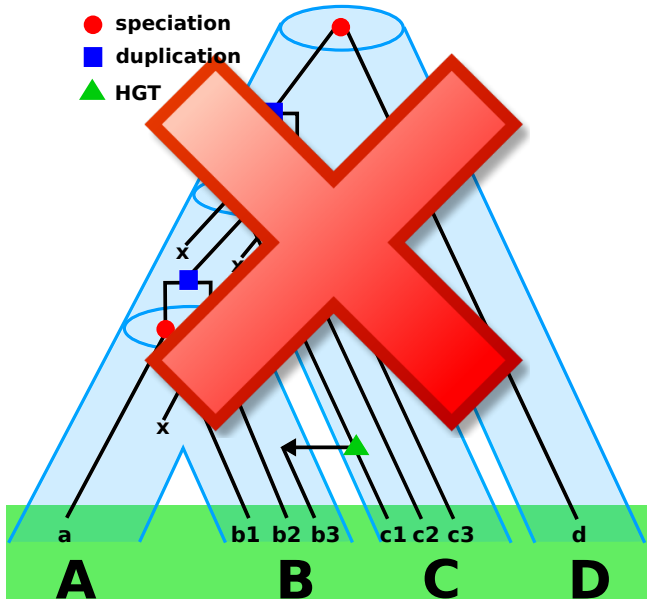


PROBLEM:

- Homology Relations are defined by the true evolutionary scenario!
- However, we don’t know and will never know the truth,
  since we cannot observe the past!

The “true” evolutionary history

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

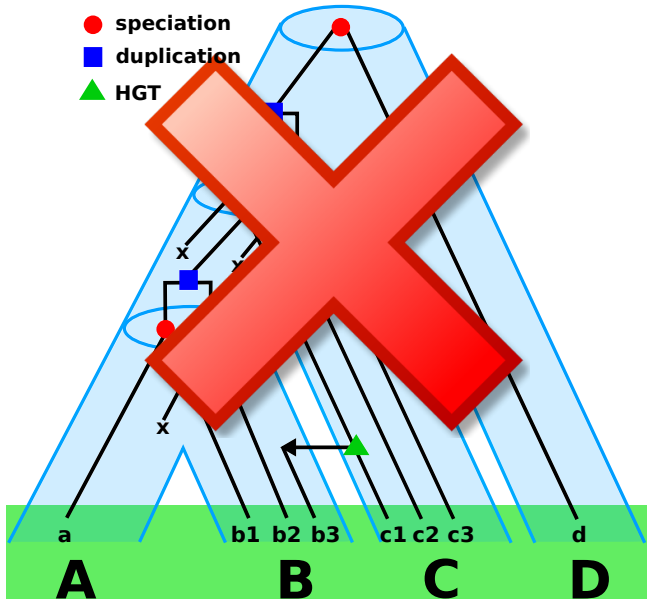


PROBLEM:

- Homology Relations are defined by the true evolutionary scenario!
- However, we don’t know and will never know the truth,
  since we cannot observe the past!
- But, we want to know the Homology Relations of the genes of extant species ("green box")

The “true” evolutionary history

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



PROBLEM:

- Homology Relations are defined by the true evolutionary scenario!
- However, we don’t know and will never know the truth,
  since we cannot observe the past!
- But, we want to know the Homology Relations of the genes of extant species ("green box")

What now?

Homology Relations: Orthology and Xenology

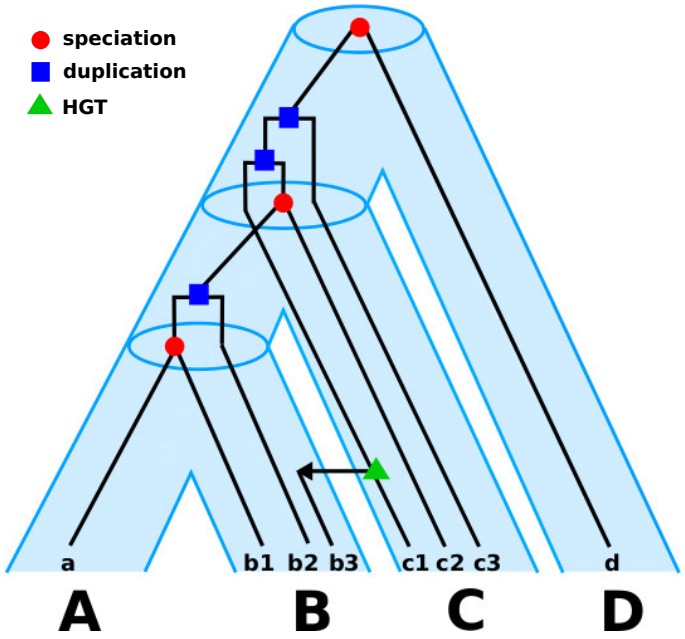
In the following, we will have closer look to two fundamental homology relations:

Orthology and **Xenology**

Orthology

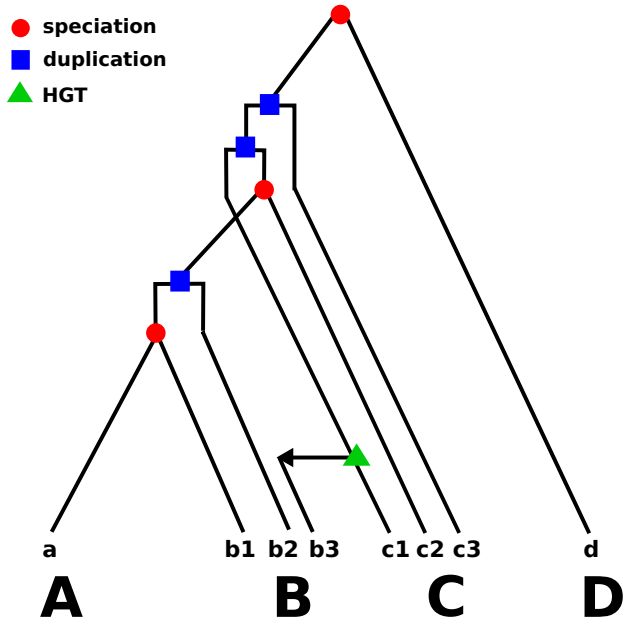
defined in terms of vertex-labels.

Orthology



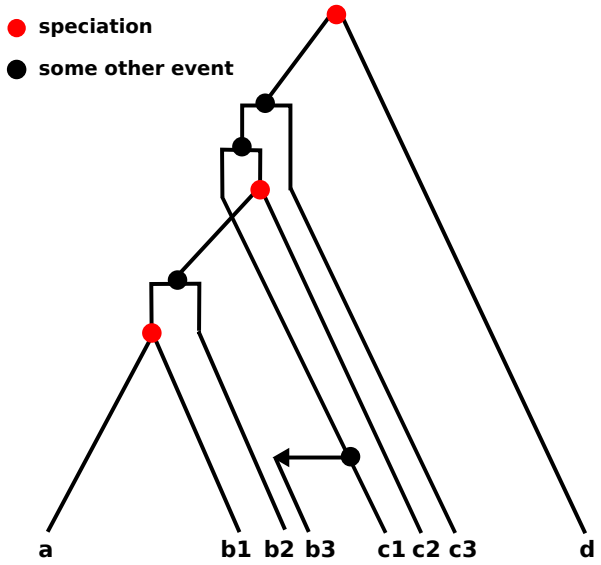
Two genes are homologs if they share a common ancestor in the ***true*** history.
Two genes x and y are orthologs if they were separated by a “speciation” event in the ***true*** history.

Orthology



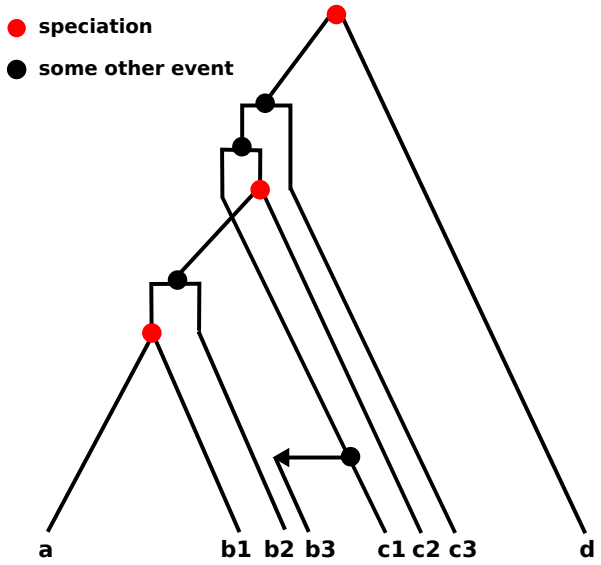
Two genes are homologs if they share a common ancestor in the ***true*** history.
Two genes x and y are orthologs if they were separated by a "speciation" event in the ***true*** history.

Orthology



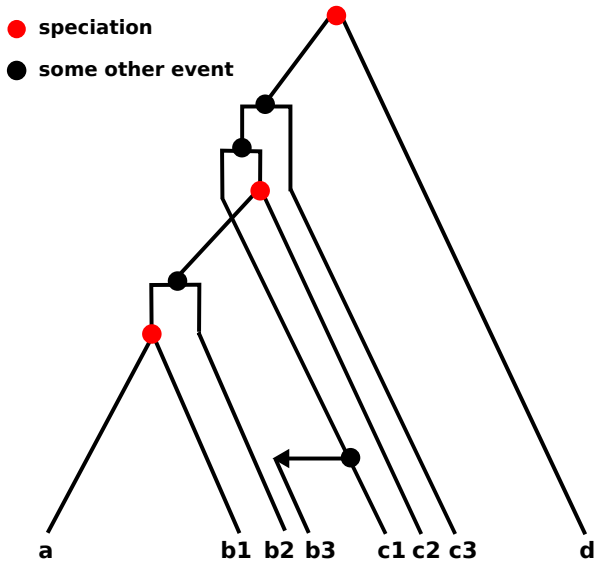
Two genes are homologs if they share a common ancestor in the ***true*** history.
Two genes x and y are orthologs if they were separated by a “speciation” event in the ***true*** history.

Orthology



Two genes are homologs if they share a common ancestor in the ***true*** history.
Two genes x and y are orthologs if they were separated by a “speciation” event in the ***true*** history.

Orthology

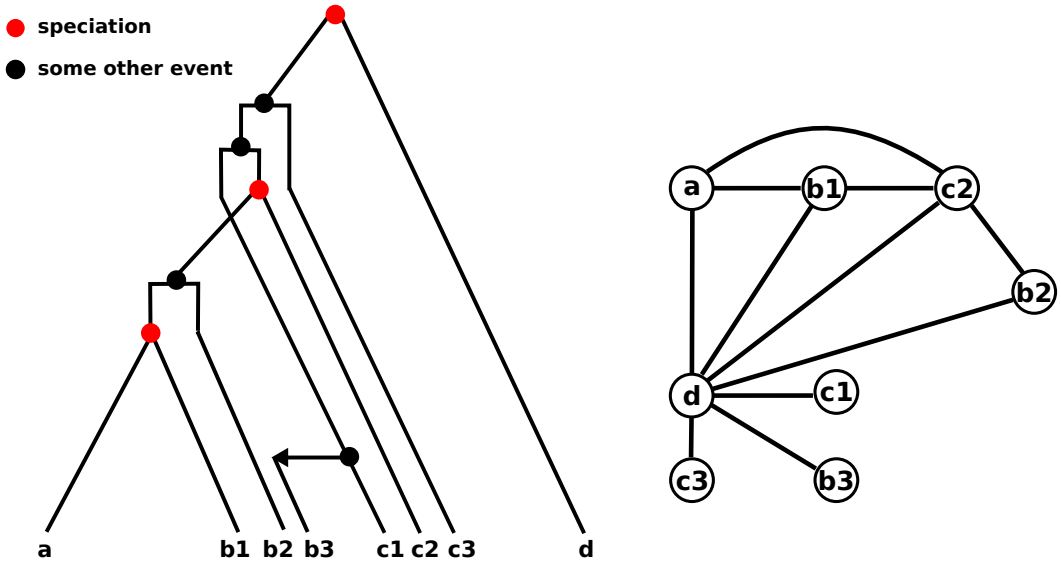


Two genes are homologs if they share a common ancestor in the ***true*** history.
Two genes x and y are orthologs if they were separated by a "speciation" event in the ***true*** history.

Mathematical Translation:

Given the ***true*** gene tree $T = (V, E)$ with coloring $t\colon V^0 \to \{\bullet, \bullet\}$.
Two leaves x and y of T are orthologs, if $t(\mathrm{lca}_T(x, y)) = \bullet$.

Orthology

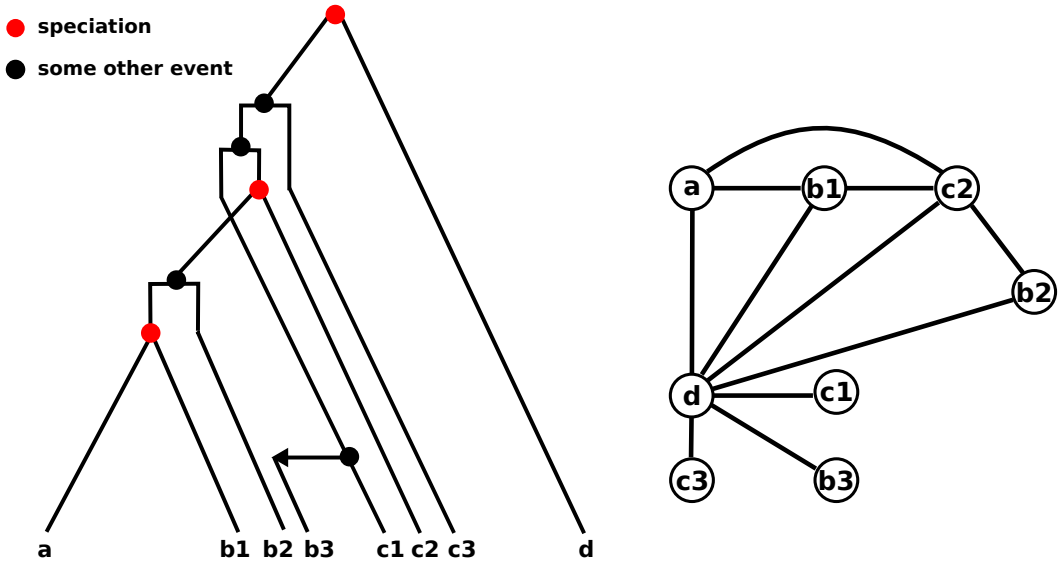


Two genes are homologs if they share a common ancestor in the ***true*** history.
Two genes x and y are orthologs if they were separated by a "speciation" event in the ***true*** history.

Mathematical Translation:

Given the ***true*** gene tree $T = (V, E)$ with coloring $t\colon V^0 \to \{\bullet, \bullet\}$.
Two leaves x and y of T are orthologs, if $t(\mathrm{lca}_T(x, y)) = \bullet$.

Orthology



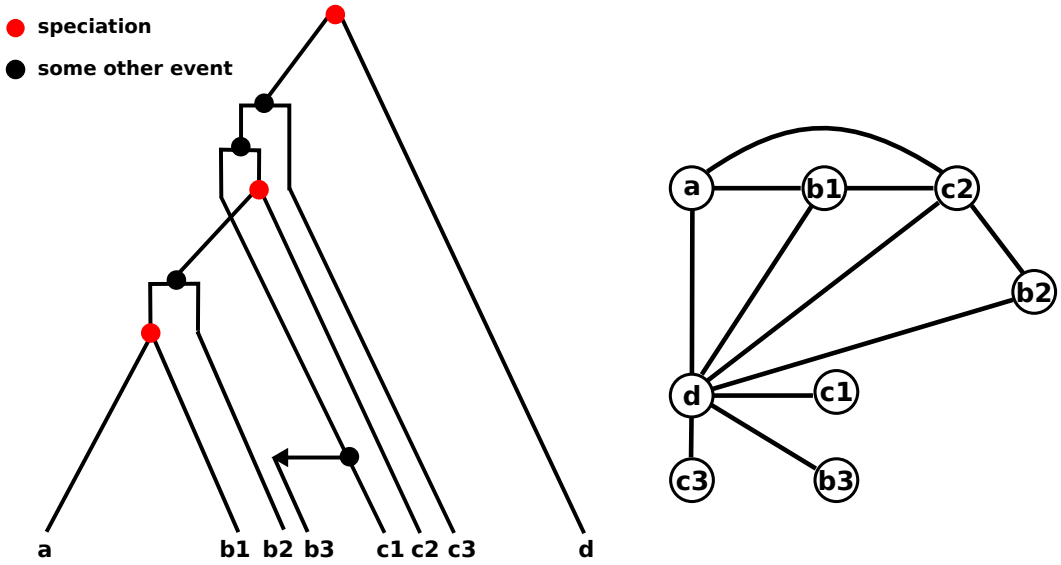
Two genes are homologs if they share a common ancestor in the ***true*** history.
Two genes x and y are orthologs if they were separated by a "speciation" event in the ***true*** history.

Mathematical Translation:

Given the ***true*** gene tree $T = (V, E)$ with coloring $t\colon V^0 \to \{\bullet, \bullet\}$.
Two leaves x and y of T are orthologs, if $t(\mathrm{lca}_T(x, y)) = \bullet$.

The orthology relation $R_\bullet$ comprises all pairs (x, y) of orthologous genes.

Orthology



Two genes are homologs if they share a common ancestor in the ***true*** history.
Two genes x and y are orthologs if they were separated by a "speciation" event in the ***true*** history.

Mathematical Translation:

Given the ***true*** gene tree $T = (V, E)$ with coloring $t\colon V^0 \to \{\bullet, \bullet\}$.
Two leaves x and y of T are orthologs, if $t(\mathrm{lca}_T(x, y)) = \bullet$.

The orthology relation $R_\bullet$ comprises all pairs (x, y) of orthologous genes.

Observation: $R_\bullet$ has a precise mathematical definition
in terms of the ***true*** history – *which is unkown!*

Orthology-Inference: 2 classical methods

Tree-based inference

- construct gene and species trees and find reconciliation map μ between them
- based on the placing of vertices in gene tree to species tree on infers speciation events

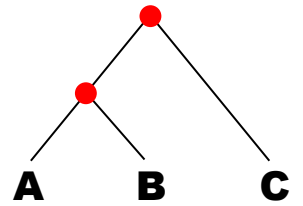
Graph-based inference

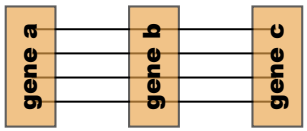
- no tree required

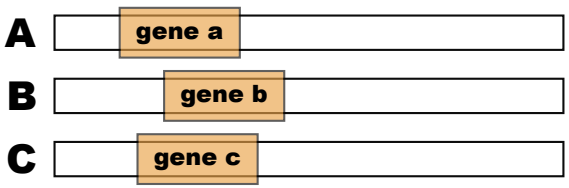
Typically run in two phases:

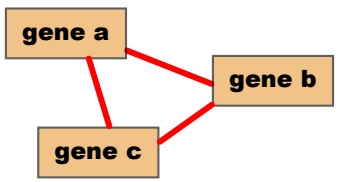
- a graph construction phase, in which pairs of orthologous genes are inferred and connected by edges
- a clustering/clean-up phase, in which (groups of) orthologous genes are constructed/extracted based on the structure of the graph

Orthology-Inference: Tree-based









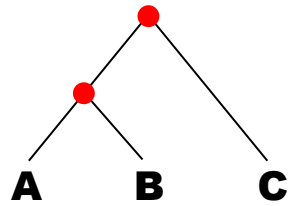
Compute Species Tree:

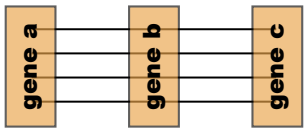
- Find 1:1-orthologs

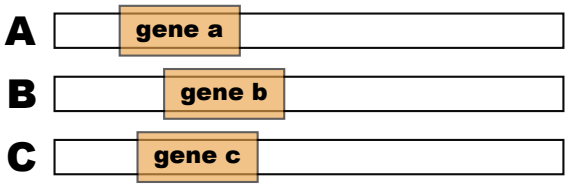
  = collection of genes such that from each species one gene and each gene is ortholog to all other genes in this collection

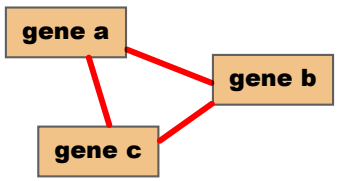
  - Select families of genes that rarely exhibit duplications (e.g. rRNAs, ribosomal proteins)

Orthology-Inference: Tree-based









Compute Species Tree:

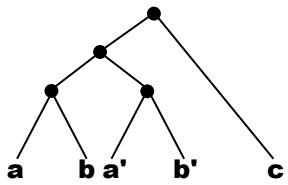
- Find 1:1-orthologs

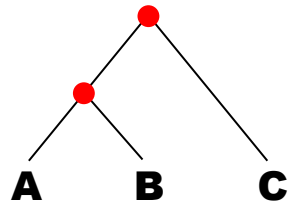
  = collection of genes such that from each species one gene and each gene is ortholog to all other genes in this collection

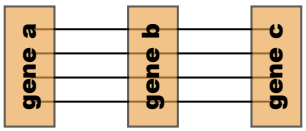
  - Select families of genes that rarely exhibit duplications (e.g. rRNAs, ribosomal proteins)

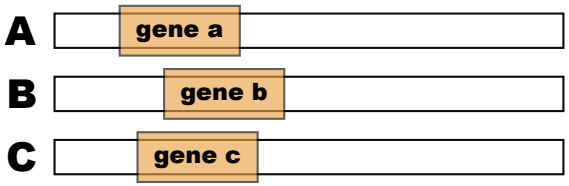
- Alignments of protein or DNA sequences and standart techniques yield gene tree with speciation-events only
  This history is believed to be congruent to that of the respective species.

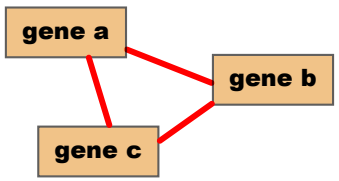
Orthology-Inference: Tree-based







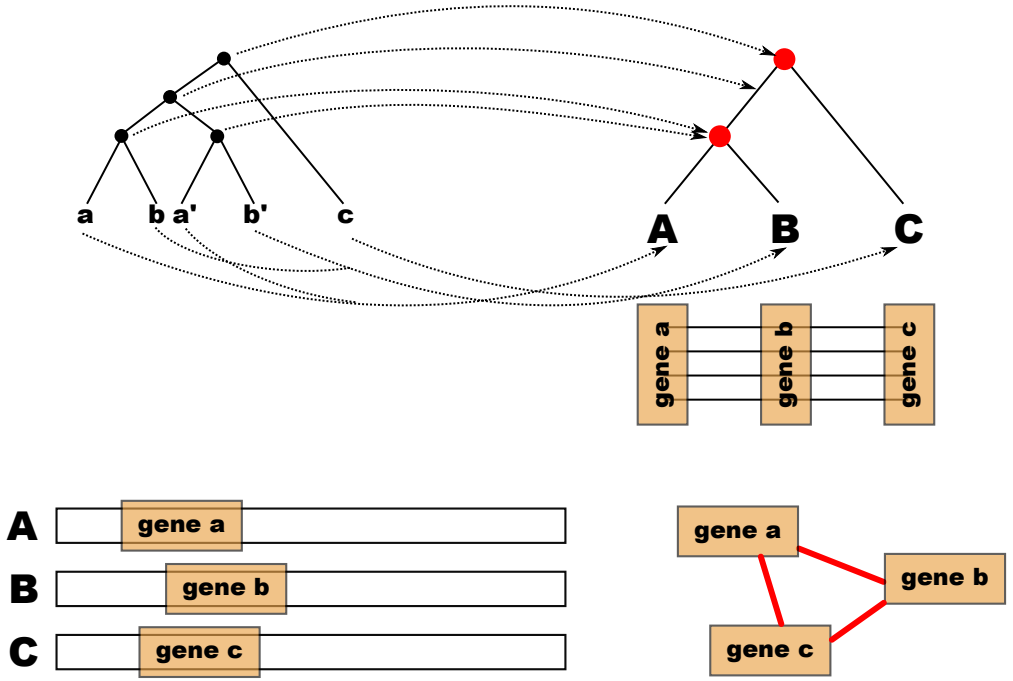




Compute Gene Tree *without events*:

- Alignments of protein or DNA sequences and standart techniques

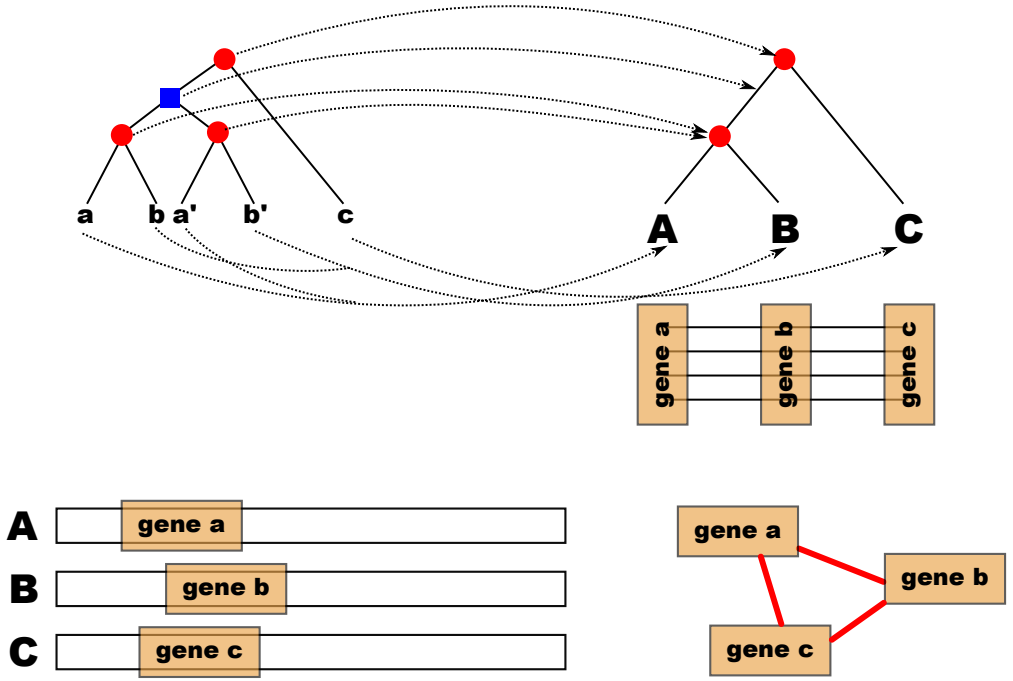
Orthology-Inference: Tree-based



Compute Events of Gene Tree:

- Find reconciliation map μ w.r.t. certain optimization criteria (e.g. parsimony = minimize number of losses and duplications)

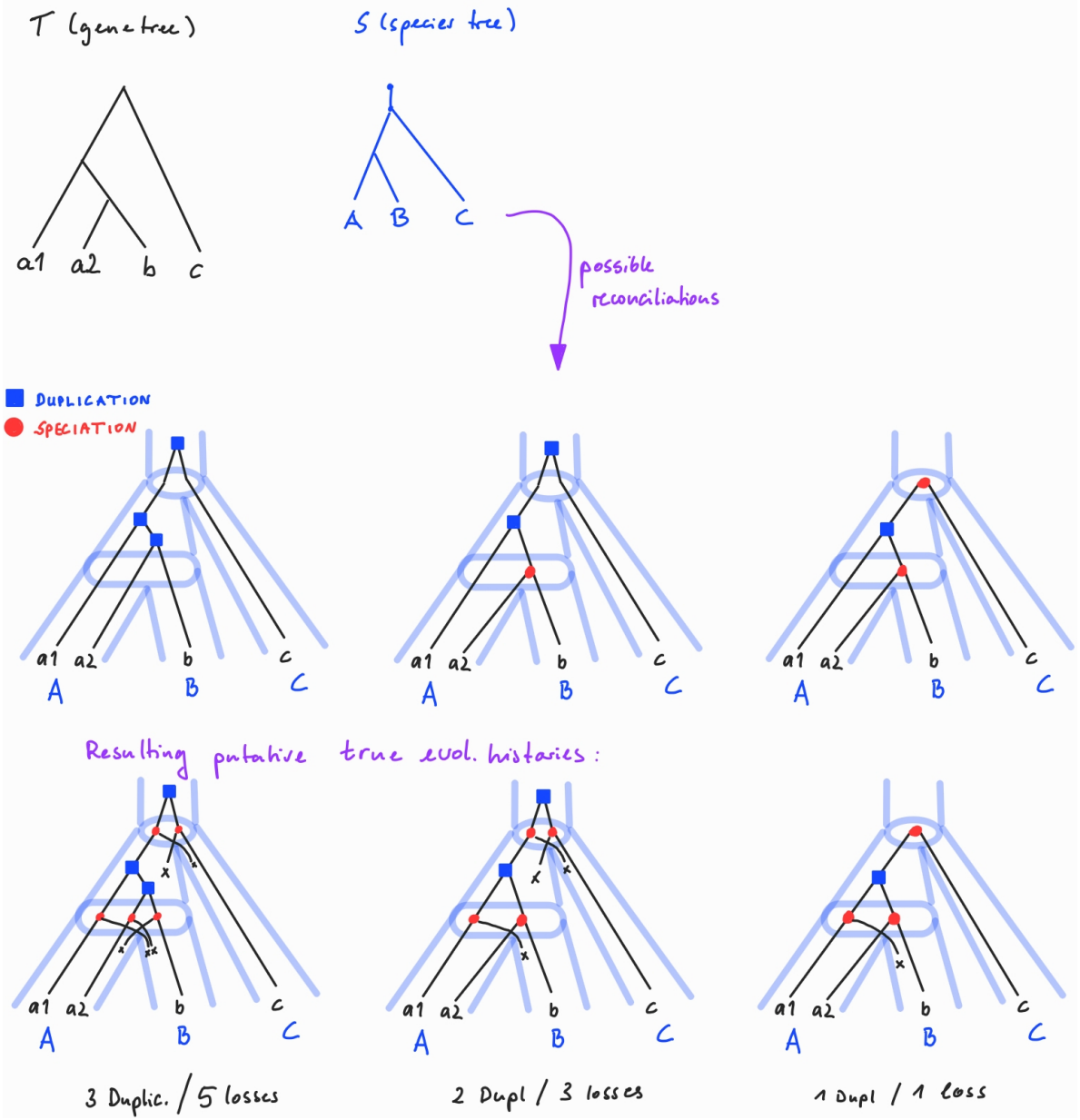
Orthology-Inference: Tree-based



Compute Events of Gene Tree:

- Find reconciliation map μ w.r.t. certain optimization criteria (e.g. parsimony = minimize number of losses and duplications)
- Use μ to infer the events (and thus orthology, paralogs, . . .)

Orthology-Inference: Tree-based



Observation

- **Compute Species Tree**
  - some orthologs must already be known!
  - since only 1:1 orthologs are used, $\sim$ 90% of the genetic sequence material remains unused
- **Compute Gene Tree + Reconciliation**
  - Methods that allow to reconstruct the history of arbitrary genes rely on “restrictive” evolutionary models (e.g. event probabilities, maximum parsimony)

Observation

- **Compute Species Tree**
  - some orthologs must already be known!
  - since only 1:1 orthologs are used, $\sim$ 90% of the genetic sequence material remains unused
- **Compute Gene Tree + Reconciliation**
  - Methods that allow to reconstruct the history of arbitrary genes rely on "restrictive" evolutionary models (e.g. event probabilities, maximum parsimony)

Observation

- **Compute Species Tree**
  - some orthologs must already be known!
  - since only 1:1 orthologs are used, $\sim$ 90% of the genetic sequence material remains unused
- **Compute Gene Tree + Reconciliation**
  - Methods that allow to reconstruct the history of arbitrary genes rely on "restrictive" evolutionary models (e.g. event probabilities, maximum parsimony)

This reveals a circular problem:
Reconstruction of species trees requires identifying events of the family evolution
Reconstruction of event-labeled gene trees requires a known species trees

Accuracy strongly depends on the predicted gene tree and the used methods (together with underlying evolutionary model) to reconcile gene and species tree.

Observation

- **Compute Species Tree**
  - some orthologs must already be known!
  - since only 1:1 orthologs are used, $\sim$ 90% of the genetic sequence material remains unused
- **Compute Gene Tree + Reconciliation**
  - Methods that allow to reconstruct the history of arbitrary genes rely on "restrictive" evolutionary models (e.g. event probabilities, maximum parsimony)

This reveals a circular problem:
Reconstruction of species trees requires identifying events of the family evolution
Reconstruction of event-labeled gene trees requires a known species trees

Accuracy strongly depends on the predicted gene tree and the used methods (together with underlying evolutionary model) to reconcile gene and species tree.

Orthology-Inference: Graph-based

Typically run in two phases:

- a graph construction phase, in which pairs of orthologous genes are inferred and connected by edges
- a clustering/clean-up phase, in which (groups of) orthologous genes are constructed/extracted based on the structure of the graph

Orthology-Inference: Graph-based

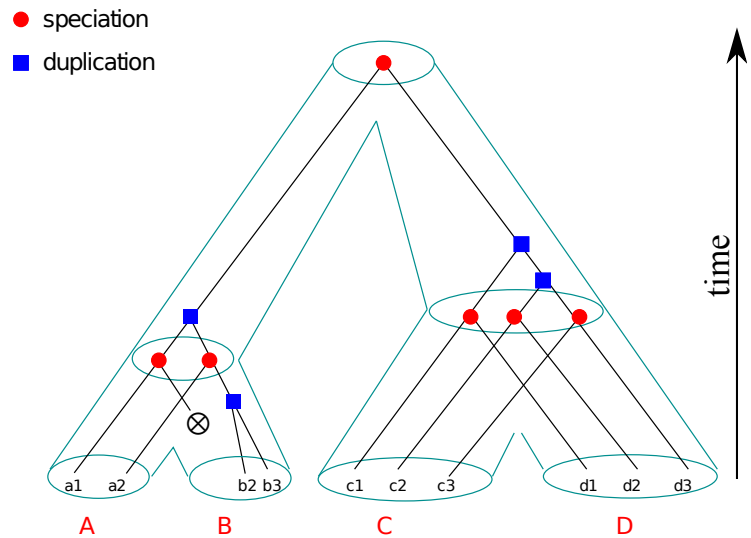
A perfect Example (no HGT):

- T gene tree, S species tree
- $t_S(X, Y)$ = divergence time of species X, Y.

Orthology-Inference: Graph-based

A perfect Example (no HGT):

- T gene tree, S species tree
- $t_S(X, Y)$ = divergence time of species X, Y.



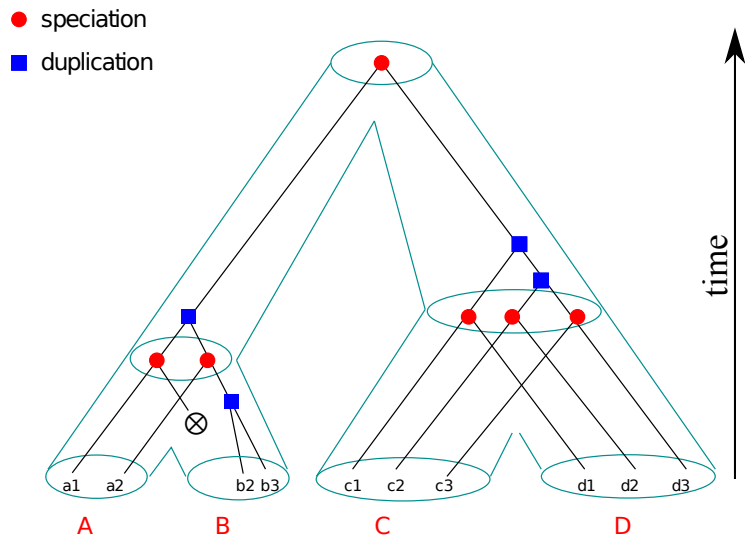
Orthologs tend to be the homologs that diverged least. Why?

If no HGT, orthologs branched by definition at the latest possible time point—the speciation between the two genomes in question.

Orthology-Inference: Graph-based

A perfect Example (no HGT):

- T gene tree, S species tree
- $t_S(X, Y)$ = divergence time of species X, Y.
- $y \in Y$ is orthologous to $x \in X$, if
  1. $X \neq Y$, orthologs are never found in the same species
  2. $t_T(x, y) \simeq t_S(X, Y)$, divergence time of x and y in $T \simeq t_S(X, Y)$.



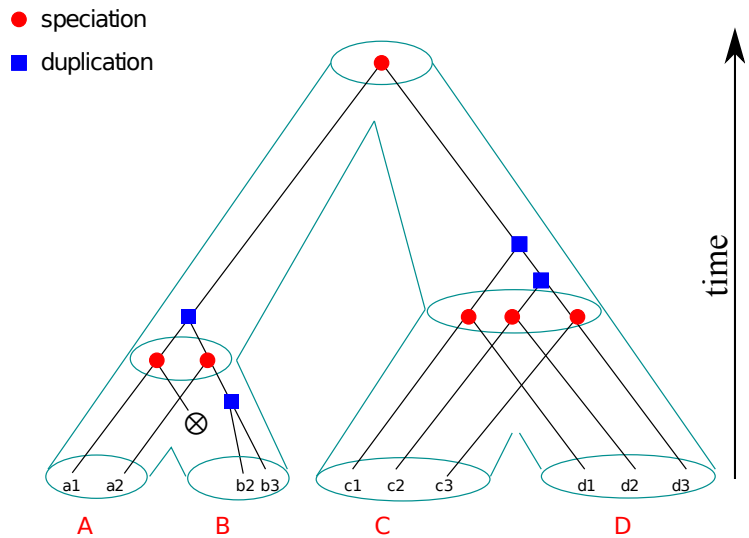
Orthologs tend to be the homologs that diverged least. Why?

If no HGT, orthologs branched by definition at the latest possible time point—the speciation between the two genomes in question.

Orthology-Inference: Graph-based

A perfect Example (no HGT):

- T gene tree, S species tree
- $t_S(X, Y)$ = divergence time of species X, Y.
- $y \in Y$ is orthologous to $x \in X$, if
  1. $X \neq Y$, orthologs are never found in the same species
  2. $t_T(x, y) \simeq t_S(X, Y)$, divergence time of x and y in $T \simeq t_S(X, Y)$.



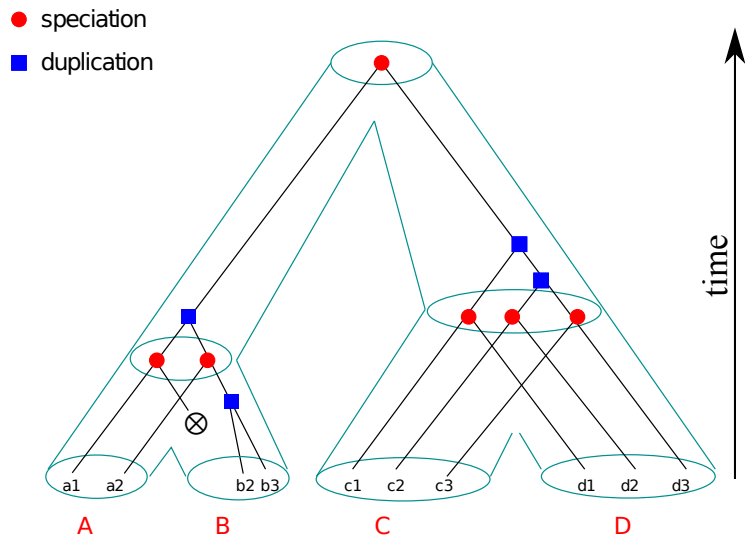
True divergence times of genes/species not known!

BUT: Sequence similarity $sim(x, y)$ can be measured.

Orthology-Inference: Graph-based

A perfect Example (no HGT):

- T gene tree, S species tree
- $t_S(X, Y)$ = divergence time of species X, Y.
- $y \in Y$ is orthologous to $x \in X$, if
  1. $X \neq Y$, orthologs are never found in the same species
  2. $t_T(x, y) \simeq t_S(X, Y)$, divergence time of x and y in $T \simeq t_S(X, Y)$.



Not too weired mutation rates: $t_T(x, y) \leq t_T(x, y') \iff sim(x, y) \geq sim(x, y')$

(No const. mol. clock assumption!)

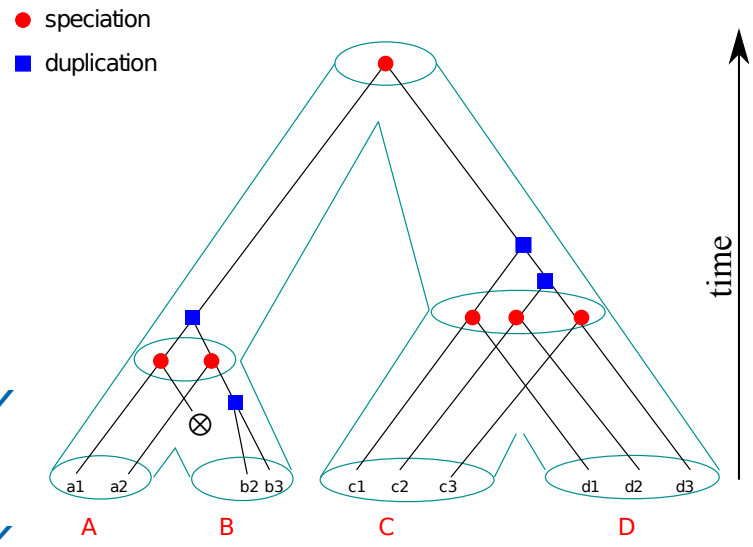
"closer related, higher similarity"

Orthology-Inference: Graph-based

In Practice (e.g. `ProteinOrtho`):

- T gene tree, S species tree
- ~~$t_S(X, Y)$ = divergence time of species X, Y.~~
- $x \in X$ and $y \in Y$ are "estimated" orth., if
  1. $X \neq Y$,
     orthologs are never found in the same species
  2. $sim(x, y) \gtrsim sim(x, y') \; \forall \; y' \in Y$ and
     $sim(y, x) \gtrsim sim(y, x') \; \forall \; x' \in X$.

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



Not too weired mutation rates: $t_T(x, y) \leq t_T(x, y') \iff sim(x, y) \geq sim(x, y')$

(No const. mol. clock assumption!)

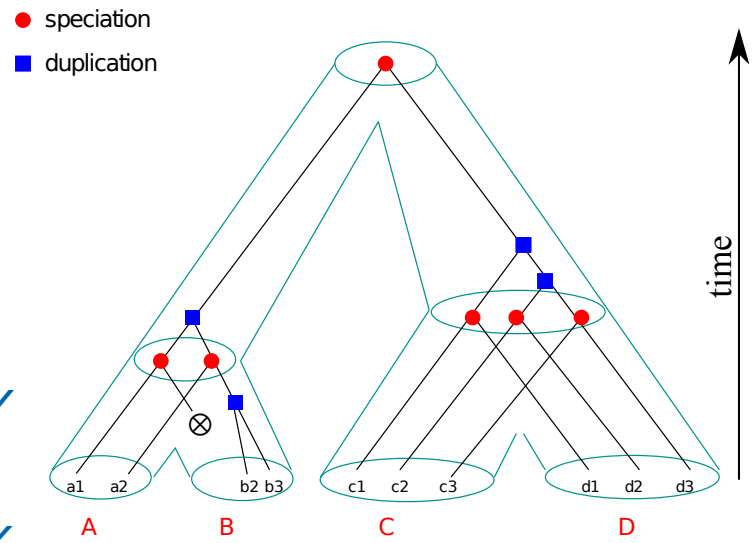
"closer related, higher similarity"

Orthology-Inference: Graph-based

In Practice (e.g. `ProteinOrtho`):

- T gene tree, S species tree
- ~~$t_S(X, Y)$ = divergence time of species X, Y.~~
- $x \in X$ and $y \in Y$ are "estimated" orth., if
  1. $X \neq Y$,
     orthologs are never found in the same species
  2. $sim(x, y) \gtrsim sim(x, y') \; \forall \; y' \in Y$ and
     $sim(y, x) \gtrsim sim(y, x') \; \forall \; x' \in X$.

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



Exmpl: $sim(a2, b2) \geq sim(a2, b3)$ ($b2$ is one of the genes in B that is "closest" to $a2$)

$sim(b2, a2) \geq sim(b2, a1))$ ($a2$ is one of the genes in A that is "closest" to $b2$)

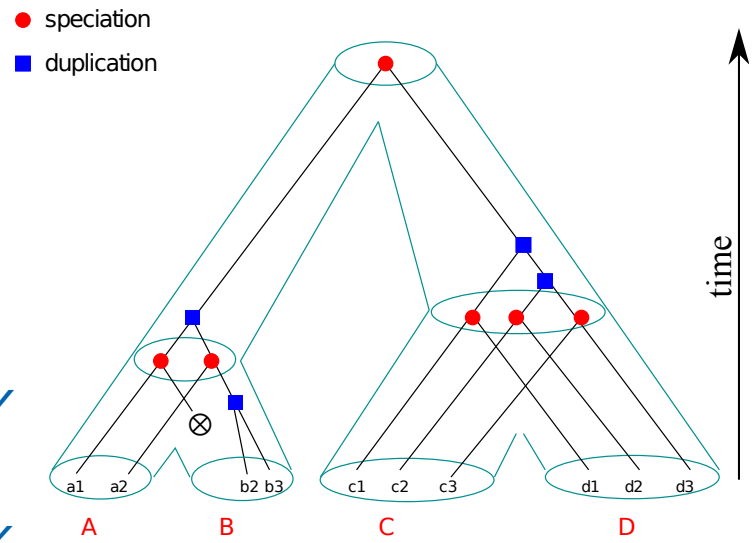
$\Longrightarrow$ $a2$ and $b2$ are "estimated" orthologs

Orthology-Inference: Graph-based

In Practice (e.g. `ProteinOrtho`):

- T gene tree, S species tree
- ~~$t_S(X, Y)$ = divergence time of species X, Y.~~
- $x \in X$ and $y \in Y$ are "estimated" orth., if
  1. $X \neq Y$,
     orthologs are never found in the same species
  2. $sim(x, y) \gtrsim sim(x, y') \; \forall \; y' \in Y$ and
     $sim(y, x) \gtrsim sim(y, x') \; \forall \; x' \in X$.

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



Exmpl: $sim(a1, b2) \geq sim(a1, b3)$ ($b2$ is one of the genes in B that is "closest" to $a1$)

$sim(b2, a1) \not\geq sim(b2, a2))$ ($a1$ is not "closest" to $b2$ among the genes in A)

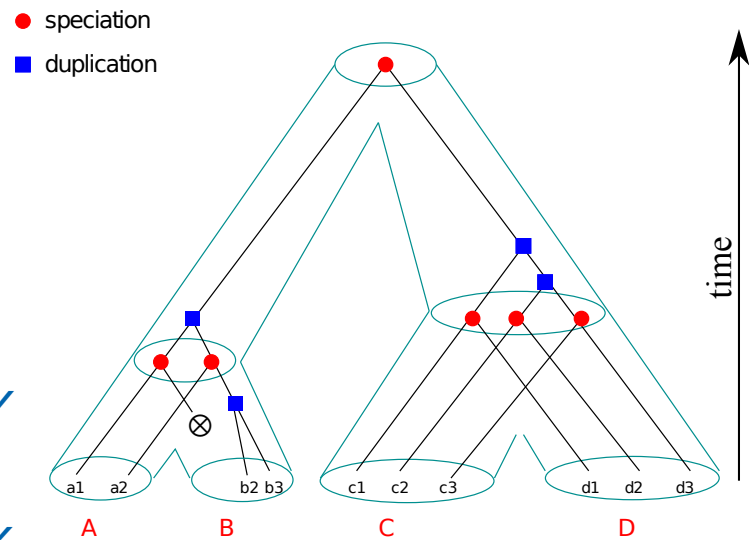
$\Longrightarrow$ $a1$ and $b2$ will not be estimated as orthologs

Orthology-Inference: Graph-based

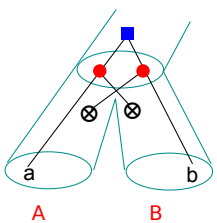
In Practice (e.g. `ProteinOrtho`):

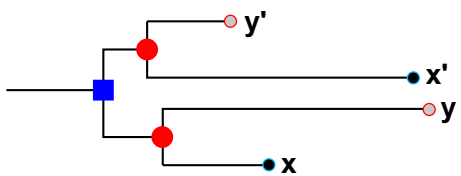
- T gene tree, S species tree
- ~~$t_S(X, Y)$ = divergence time of species X, Y.~~
- $x \in X$ and $y \in Y$ are "estimated" orth., if
  1. $X \neq Y$,
     orthologs are never found in the same species
  2. $sim(x, y) \gtrsim sim(x, y') \; \forall \; y' \in Y$ and
     $sim(y, x) \gtrsim sim(y, x') \; \forall \; x' \in X$.

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



This cannot work perfectly:



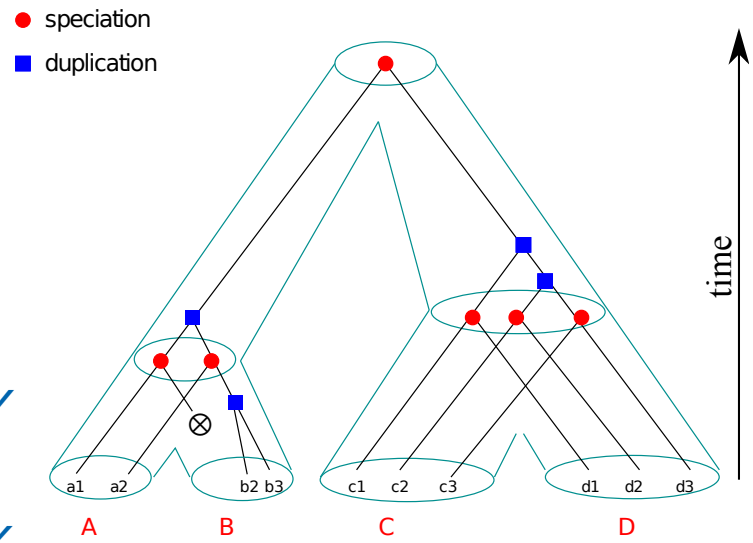


Orthology-Inference: Graph-based

In Practice (e.g. `ProteinOrtho`):

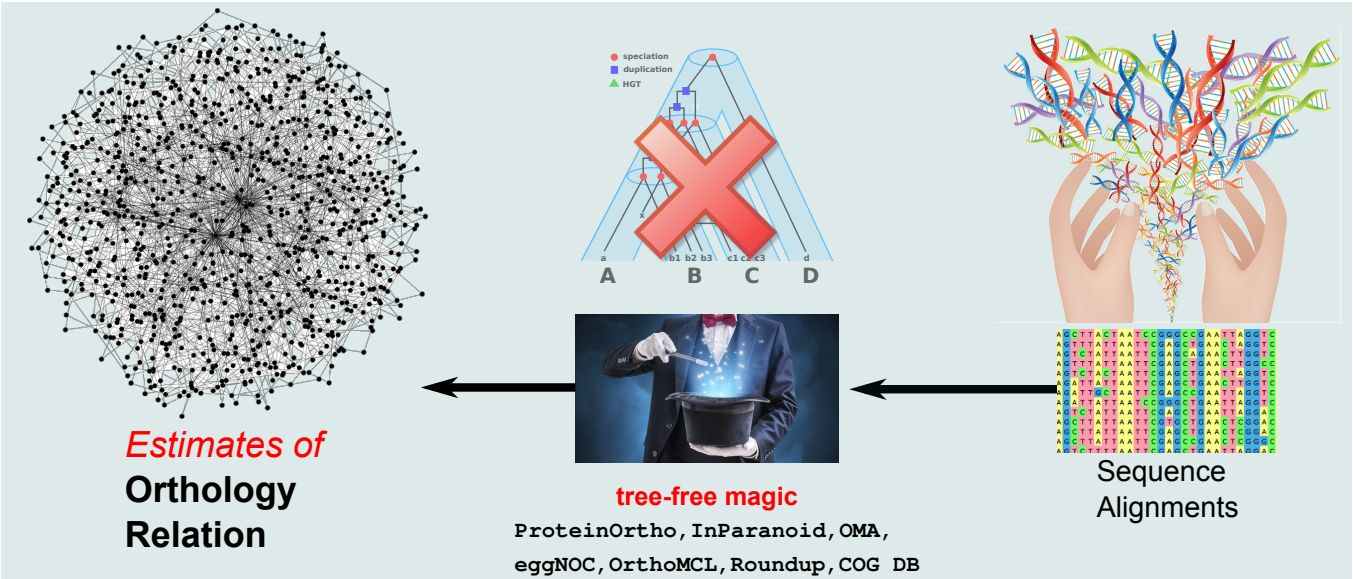
- T gene tree, S species tree
- ~~$t_S(X, Y)$ = divergence time of species X, Y.~~
- $x \in X$ and $y \in Y$ are "estimated" orth., if
  1. $X \neq Y$,
     orthologs are never found in the same species
  2. $sim(x, y) \gtrsim sim(x, y') \; \forall \; y' \in Y$ and
     $sim(y, x) \gtrsim sim(y, x') \; \forall \; x' \in X$.

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



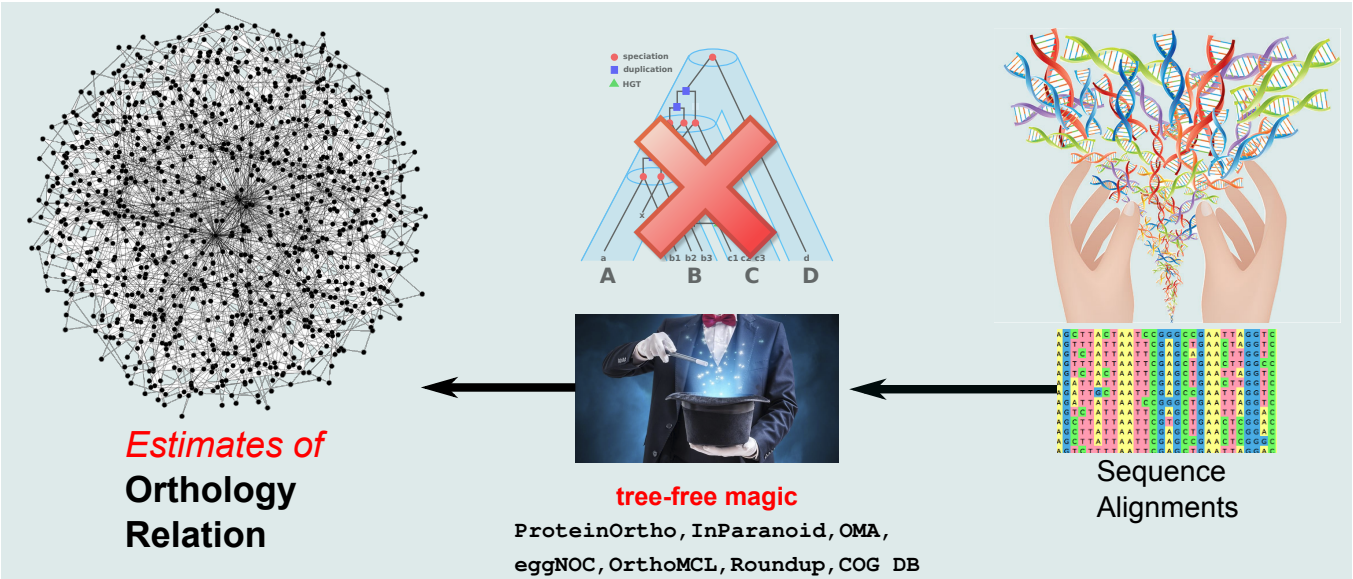
Estimates of orthologs rely on reciprocal best match (RBM) heuristics.

Orthology-Inference: Graph-based



How can we trust such estimates $\widehat{R}_\bullet$ of the *true* $R_\bullet$?

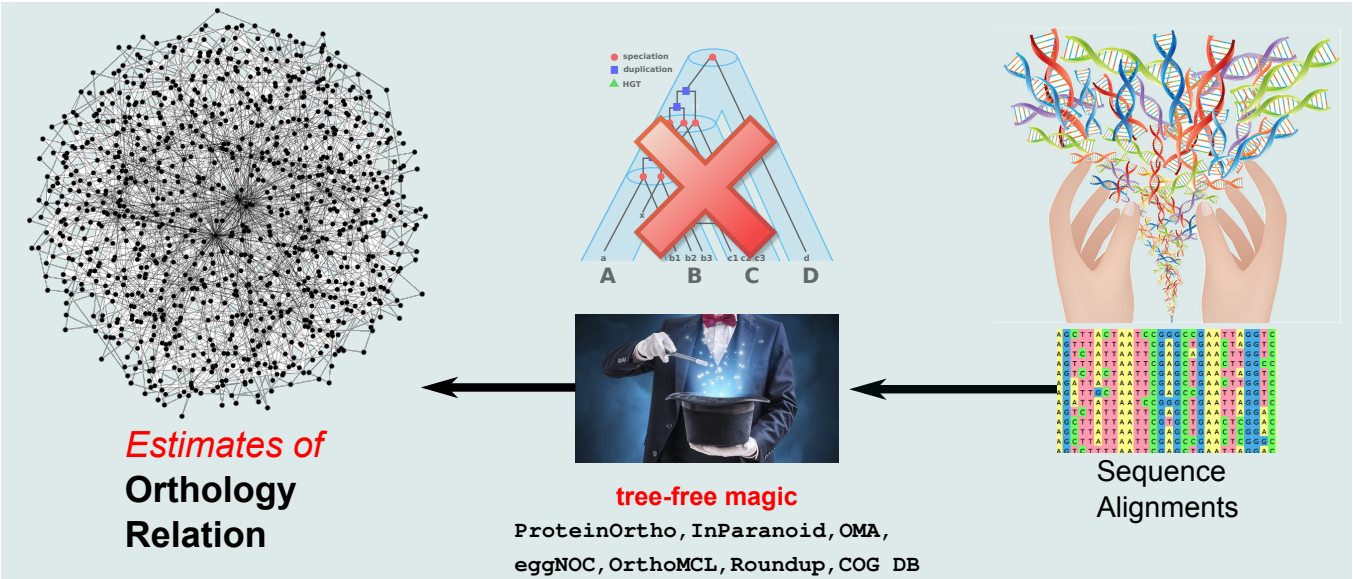
Orthology-Inference: Graph-based



How can we trust such estimates $\widehat{R}_\bullet$ of the *true* $R_\bullet$?

The least task we can do:
Ask for an event-labeled gene tree that supports our observation.

Orthology-Inference: Graph-based



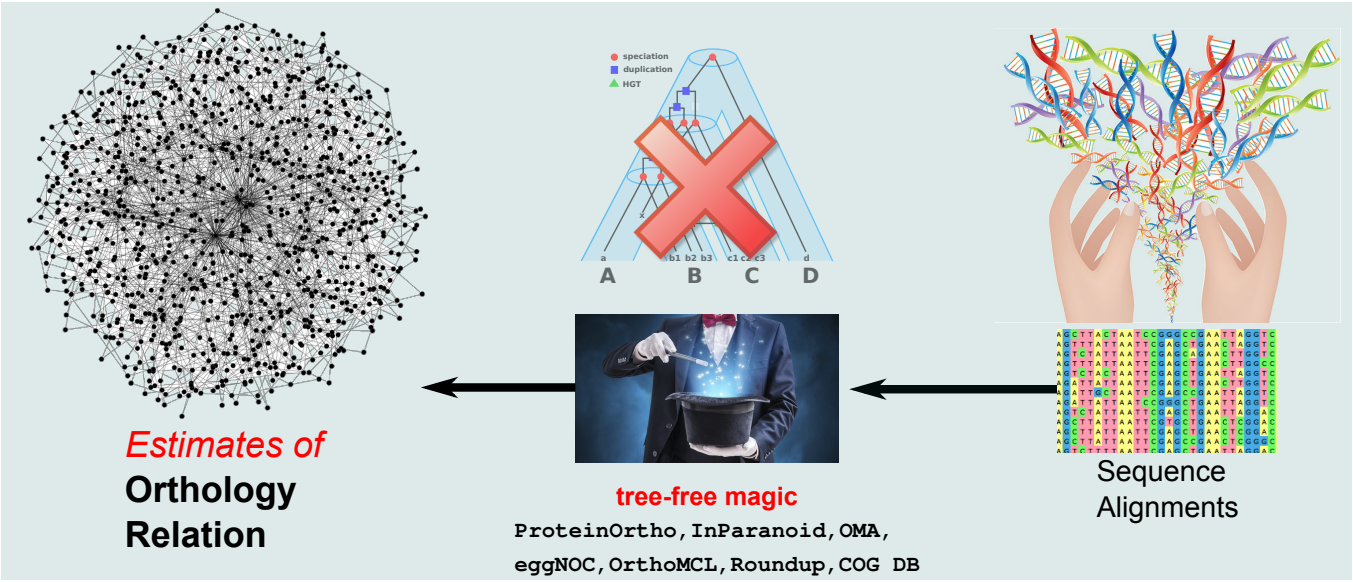
How can we trust such estimates $\widehat{R}_{\bullet}$ of the *true* $R_{\bullet}$?

The least task we can do:
Ask for an event-labeled gene tree that supports our observation.

An estimated orthology relation $\widehat{R}_{\bullet}$ is **feasible** if
there is a tree $T = (V, E)$ with coloring $t\colon V^0 \to \{\bullet, \bullet\}$ such that

$$t(\mathrm{lca}_T(x, y)) = \bullet \iff (x, y) \in \widehat{R}_{\bullet} \text{ for all distinct } x, y \in X.$$

Orthology-Inference: Graph-based



How can we trust such estimates $\widehat{R}_{\bullet}$ of the *true* $R_{\bullet}$?

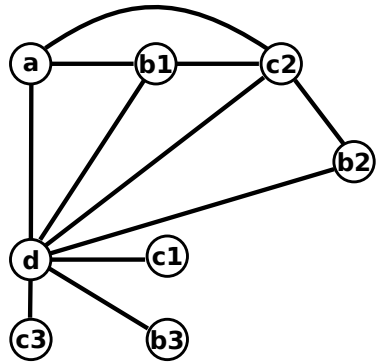
The least task we can do:
Ask for an event-labeled gene tree that supports our observation.

An estimated orthology relation $\widehat{R}_{\bullet}$ is **feasible** if
there is a tree $T = (V, E)$ with coloring $t\colon V^0 \to \{\bullet, \bullet\}$ such that

$$t(\mathrm{lca}_T(x, y)) = \bullet \iff (x, y) \in \widehat{R}_{\bullet} \text{ for all distinct } x, y \in X.$$

Can we mathematically characterize **feasible** estimates $\widehat{R}_{\bullet}$?

Orthology-Inference: Graph-based



How can we trust such estimates $\widehat{R}_{\bullet}$ of the *true* $R_{\bullet}$?

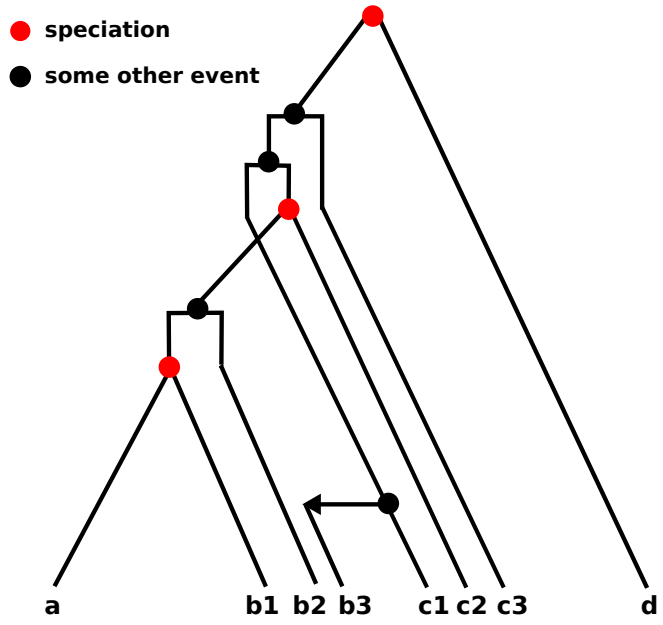
The least task we can do:
Ask for an event-labeled gene tree that supports our observation.

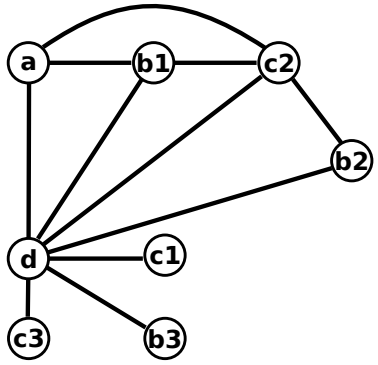
An estimated orthology relation $\widehat{R}_{\bullet}$ is **feasible** if
there is a tree $T = (V, E)$ with coloring $t\colon V^0 \to \{\bullet, \bullet\}$ such that

$$t(\mathrm{lca}_T(x, y)) = \bullet \iff (x, y) \in \widehat{R}_{\bullet} \text{ for all distinct } x, y \in X.$$

Can we mathematically characterize **feasible** estimates $\widehat{R}_{\bullet}$?

Orthology-Inference: Graph-based





How can we trust such estimates $\widehat{R}_{\bullet}$ of the *true* $R_{\bullet}$?

The least task we can do:
Ask for an event-labeled gene tree that supports our observation.

An estimated orthology relation $\widehat{R}_{\bullet}$ is **feasible** if
there is a tree $T = (V, E)$ with coloring $t\colon V^0 \to \{\bullet, \bullet\}$ such that

$$t(\mathrm{lca}_T(x, y)) = \bullet \iff (x, y) \in \widehat{R}_{\bullet} \text{ for all distinct } x, y \in X.$$

Can we mathematically characterize **feasible** estimates $\widehat{R}_{\bullet}$?

The Structure of Orthology

Theorem

The following statements are equivalent:

1. *An (estimated) orthology relation is feasible.*
2. *Its graph-representation does not contain induced P_4s.*
3. *Its graph-representation is a cograph.*

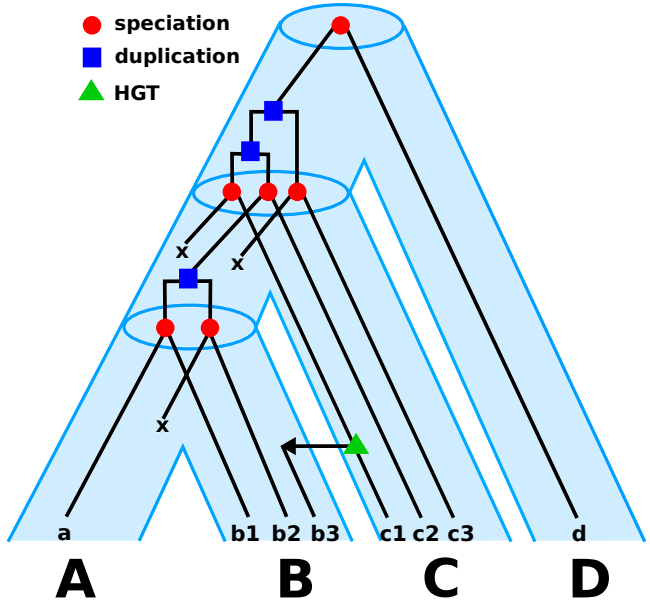
terminology + proof in whiteboard

Xenology

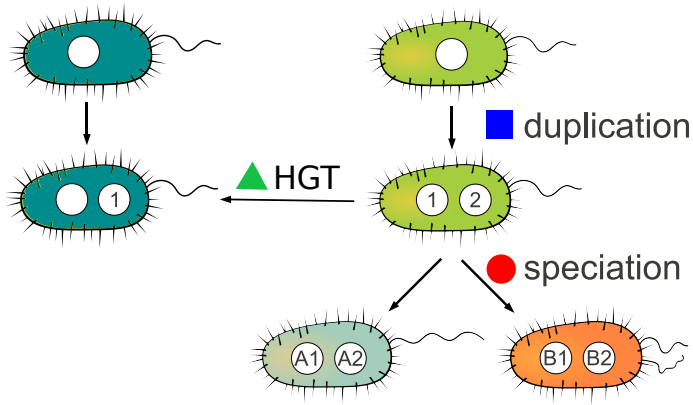
defined in terms of edge-labels (HGT).

Xenology

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

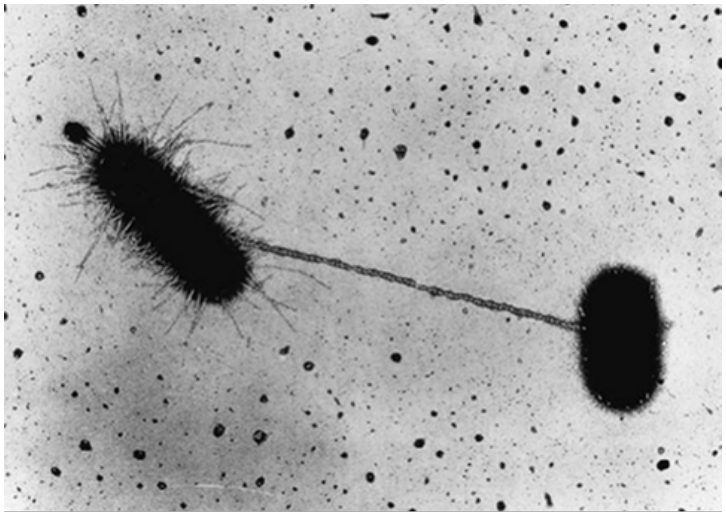


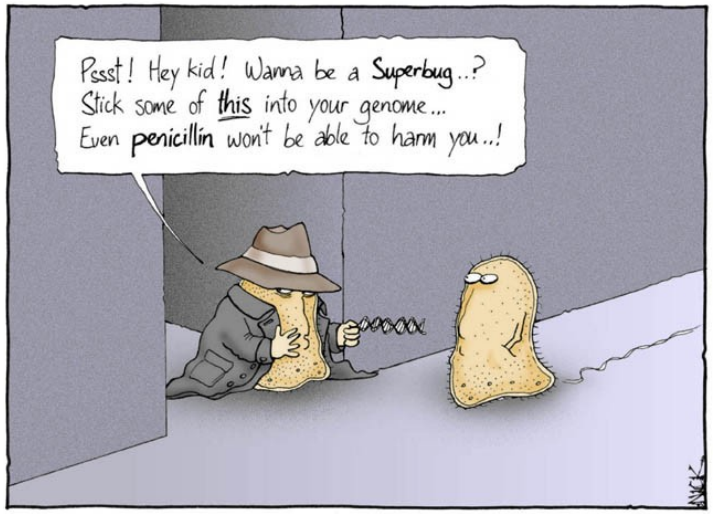
- ■ Gene duplication : an offspring has two copies of a single gene of its ancestor
- ● Speciation : two offspring species inherit the entire genome of their common ancestor
- ▲ HGT : transfer of genes between organisms in a manner other than traditional reproduction and across different species



Examples of HGT

Bacteria to Bacteria:





HGT is a significant cause of increased drug resistance when one bacterial cell acquires resistance, and the resistance genes are transferred to other species.

Barlow, **What antimicrobial resistance has taught us about horizontal gene transfer**, *Methods in Molecular Biology. 532: 397-411*, 2009

Hawkey and Jones, **The changing epidemiology of resistance**, *Journal of Antimicrobial Chemotherapy. 64 (Suppl 1): i3-10.*, 2009

Stearns and Hoekstra, **Evolution: An introduction (2nd ed.)**, *Oxford Univ. Press*, 2005

Examples of HGT

Bacteria to Animals:

A bacterial gene discovered in the genome of the *coffee berry borer beetle*, a major pest, allows the beetle to occupy a unique ecological niche and feed exclusively on coffee beans.

The transferred gene, which lets the beetle break down complex sugars in the coffee bean, came from the beetle's gut bacteria.

Acuna et al. , **Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee**, *PNAS. 109 (11): 4197-4202*, 2012

Phillips, **Bacterial gene helps coffee beetle get its fix**, *Nature News*, 2012

Examples of HGT

Fungi to Animals:

In the *pea aphid* (Erbsenlaus) red and green color insects frequently coexist in natural populations.

Color polymorphism in the pea aphid is determined by carotenoid genes that were transferred from a fungus.

Pea aphids are the only animals the can synthesize carotenoid and thus, to produce the red pigment carotin. Due to a symbiosis with a bacteria, some of the pea aphids are colored green.

Moran and Jarvik, **Lateral Transfer of Genes from Fungi Underlies Carotenoid Production in Aphids**, *Science. 328 (5978): 624-627*, 2010

Examples of HGT

Fungi to Animals:

In the *pea aphid* (Erbsenlaus) red and green color insects frequently coexist in natural populations.

Color polymorphism in the pea aphid is determined by carotenoid genes that were transferred from a fungus.

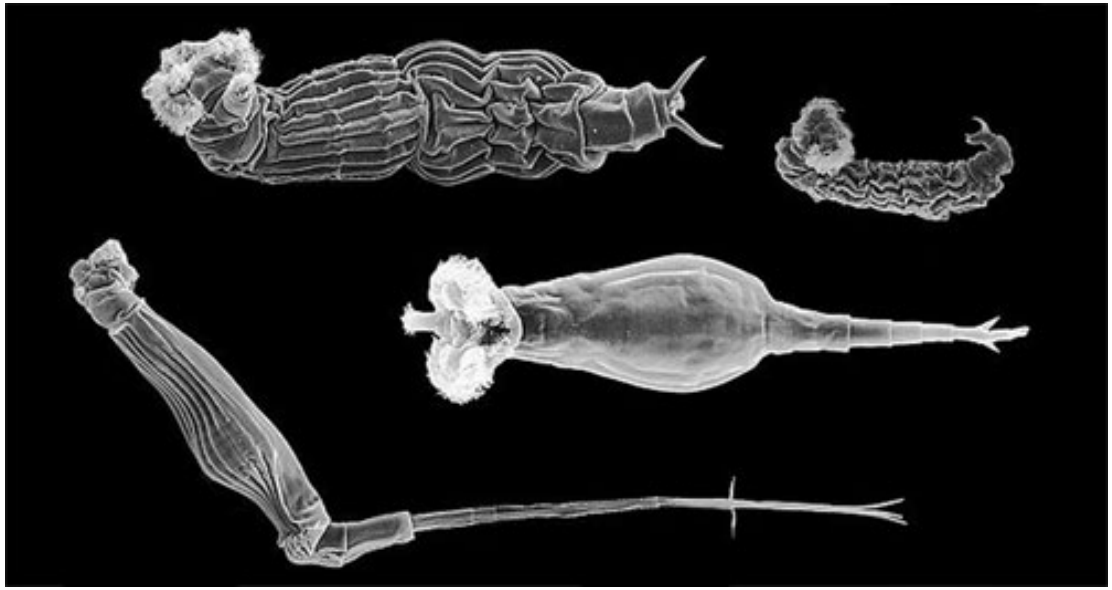
Natural enemies: *lady beetles* preferentially attack red aphids on green plants,
parasitoid wasps deposit eggs in green aphids more frequently.

HYP: Opposite predation and parasitism pressures maintain the color variation in the aphid populations.

Moran and Jarvik, **Lateral Transfer of Genes from Fungi Underlies Carotenoid Production in Aphids**, *Science. 328 (5978): 624-627*, 2010

Examples of HGT

Bacteria/Fungi to Animals:



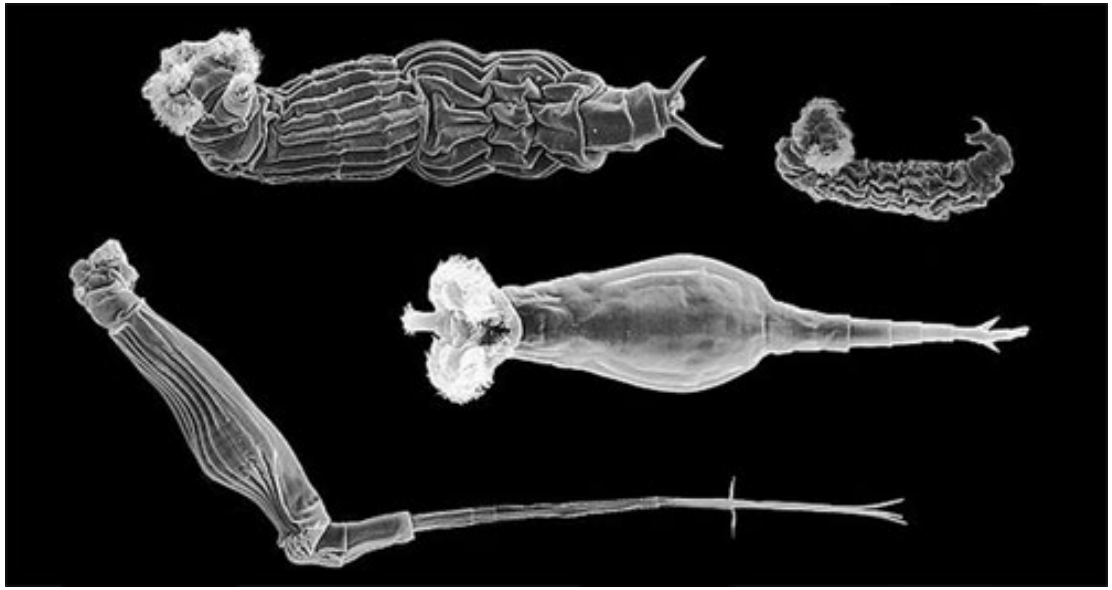
Bdelloid rotifers currently hold the 'record' for HGT in animals with $\sim$8% of their genes from bacterial origins.

Watson, **Bdelloids Surviving on Borrowed DNA**, *Science/AAAS News*, 2012

Crisp et al. , **Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes**, *Genome Biol. 16: 50*, 2015

Examples of HGT

Bacteria/Fungi to Animals:



Bdelloid rotifers currently hold the 'record' for HGT in animals with $\sim$8% of their genes from bacterial origins.

A study found the genomes of 40 animals (including 10 primates, four Caenorhabditis worms, and 12 Drosophila insects) contained genes which had been transferred from bacteria and fungi by HGT.

Watson, **Bdelloids Surviving on Borrowed DNA**, *Science/AAAS News*, 2012

Crisp et al. , **Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes**, *Genome Biol. 16: 50*, 2015

Examples of HGT

Plant to Plant

Ferns have the neochrome-gene that allows them to “produce” an unconventional photoreceptor that senses both blue *and* red light, affording ferns a unique advantage in forests shaded by flowering plants.

This neochrome-gene is not part of any other “higher” plant.

There is strong evidence that Ferns acquired the neochrome-gene from the moss-like plant *Hornwort* via HGT.

Li et al. , **Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns**, *PNAS 111:18, 6672-6677*, 2014

Zimmer, **Plants That Practice Genetic Engineering**, *New York Times*, 2015

Examples of HGT

Artificial HGT (genetic engineering)

Examples of HGT

Artificial HGT (genetic engineering)

1973 - 1982: Normally insulin is produced in the pancreas, but in people with type-1 diabetes there is a problem with insulin production and thus, they have to inject insulin to control their blood sugar levels.

Genetic engineering has been used to produce a type of insulin, very similar to our own, from yeast and bacteria like E. coli.

This genetically modified insulin, “Humulin” was licensed for human use in 1982.

Cohen et al. , **Construction of biologically functional bacterial plasmids in vitro**, *PNAS 70: 3240-3244*, 1973

Examples of HGT

Artificial HGT (genetic engineering)

2001: Enviropig (Frankenswine) - “greener” and cheaper pig.

Golovan et al. , **Pigs expressing salivary phytase produce low-phosphorus manure**, *Nature Biotechnology 19(8): 741-5*, 2001

Examples of HGT

Artificial HGT (genetic engineering)

2011: Glow-in-the-dark cats; Scientists in South Korea altered the DNA (using jellyfish genes) of a kitty so that its fur would glow in the dark.

Wongsrikeao et al. , **Antiviral restriction factor transgenesis in the domestic cat**, *Nature Methods 8, 853-859*, 2011

Examples of HGT

Artificial HGT (genetic engineering)

August 2017: For the first time, scientists corrected a heart-disease-causing mutation in early stage human embryos with gene editing.

Using the CRISPR-Cas9 method, they corrected the mutation within the embryo and so, the defect would also not be passed on to future generations.

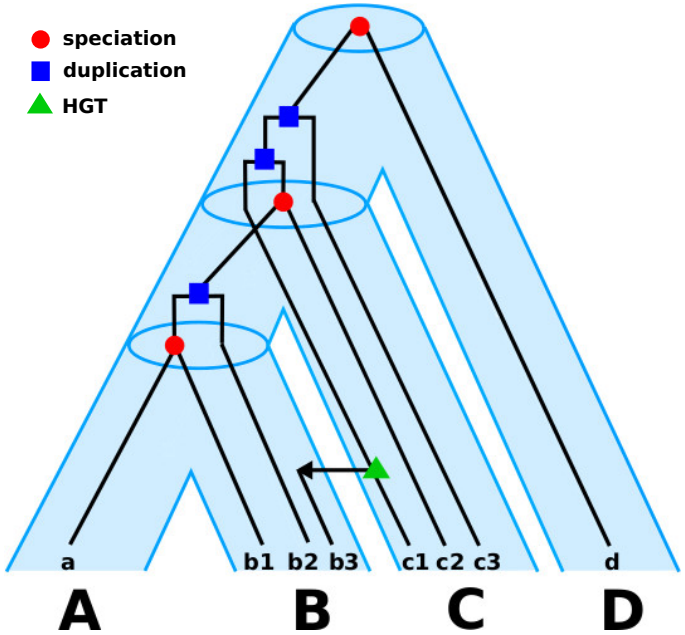
Hong Ma et al. , **Correction of a pathogenic gene mutation in human embryos**, *Nature 548, 413-419*, Aug. 2017

Examples of HGT

Artificial HGT (genetic engineering)

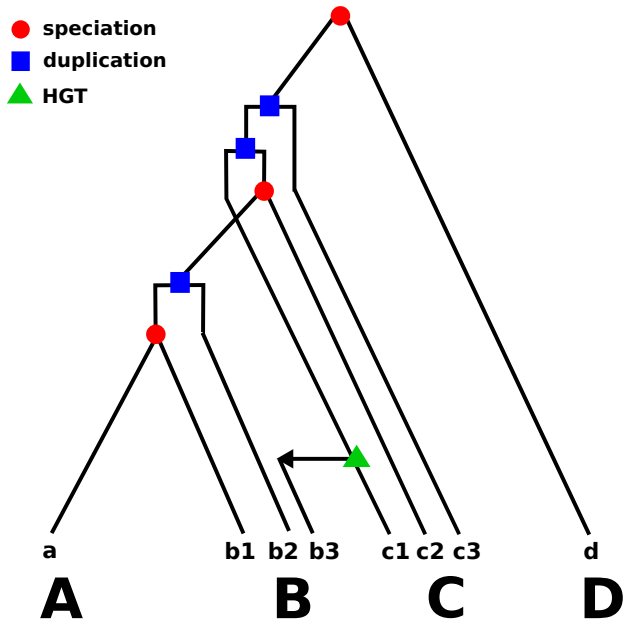
We consider here non-artificial HGT.

Xenology



Two genes are homologs if they share a common ancestor in the ***true*** history.
Two genes x and y are xenologs if there was a transfer along the path between x, y in the ***true*** history.

Xenology

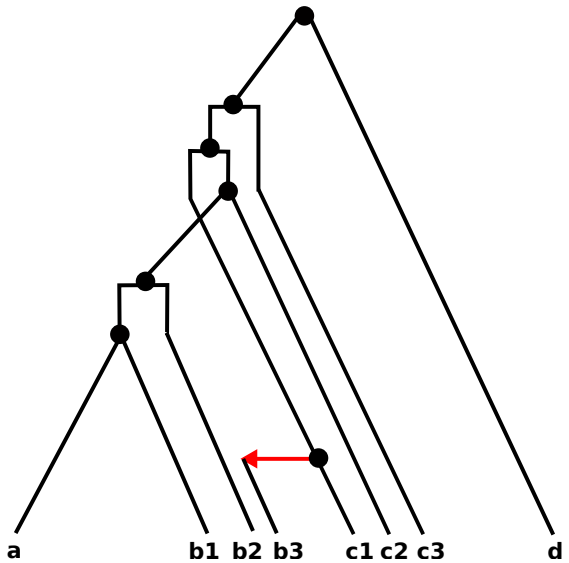


Two genes are homologs if they share a common ancestor in the ***true*** history.
Two genes x and y are xenologs if there was a transfer along the path between x, y in the ***true*** history.

Mathematical Translation:

Given the ***true*** gene tree $T = (V, E)$ with coloring $\lambda \colon E \to \{0, 1\}$.
Two leaves x and y of T are xenologs, if there is an edge e with $\lambda(e) = 1$ in the path connecting x, y in T

Xenology



Two genes are homologs if they share a common ancestor in the ***true*** history.
Two genes x and y are xenologs if there was a transfer along the path between x, y in the ***true*** history.

Mathematical Translation:

Given the ***true*** gene tree $T = (V, E)$ with coloring $\lambda\colon E \to \{0, 1\}$.
Two leaves x and y of T are xenologs, if there is an edge e with $\lambda(e) = 1$ in the path connecting x, y in T
The xenology relation $\mathcal{X}$ comprises all pairs (x, y) of xenologous genes.

Xenology-Inference (sketch of methods)

Tree-based inference

- construct gene and species trees and find reconciliation map μ between them
- based on the placing of vertices in gene tree to species tree on infers HGT-edges

Parametric inference

- no tree required
- use certain characteristics of the genome sequences under consideration

  If some fragment or gene of the genome significantly deviates from the characteristics, this is a sign for putative HGT
- requires description of what defines a “typical” gene in terms of parameters such as nucleotide composition (e.g. the GC-content), oligonucleotide frequencies, or other structural features.

Implicit phylogenetic inference

- no tree required
- based on sequence similarities and evolutionary distances.

Xenology-Inference (sketch of methods)

Tree-based inference

- construct gene and species trees and find reconciliation map μ between them
- based on the placing of vertices in gene tree to species tree on infers HGT-edges

Parametric inference

- no tree required
- use certain characteristics of the genome sequences under consideration

  If some fragment or gene of the genome significantly deviates from the characteristics, this is a sign for putative HGT
- requires description of what defines a “typical” gene in terms of parameters such as nucleotide composition (e.g. the GC-content), oligonucleotide frequencies, or other structural features.

Implicit phylogenetic inference

- no tree required
- based on sequence similarities and evolutionary distances.

Xenology-Inference (sketch of methods)

Tree-based inference

- construct gene and species trees and find reconciliation map μ between them
- based on the placing of vertices in gene tree to species tree on infers HGT-edges

Parametric inference

- no tree required
- use certain characteristics of the genome sequences under consideration

  If some fragment or gene of the genome significantly deviates from the characteristics, this is a sign for putative HGT
- requires description of what defines a “typical” gene in terms of parameters such as nucleotide composition (e.g. the GC-content), oligonucleotide frequencies, or other structural features.

Implicit phylogenetic inference

- no tree required
- based on sequence similarities and evolutionary distances.

Xenology-Inference (sketch of methods)

Tree-based inference

- construct gene and species trees and find reconciliation map μ between them
- based on the placing of vertices in gene tree to species tree on infers HGT-edges

Parametric inference

- no tree required
- use certain characteristics of the genome sequences under consideration

  If some fragment or gene of the genome significantly deviates from the characteristics, this is a sign for putative HGT
- requires description of what defines a “typical” gene in terms of parameters such as nucleotide composition (e.g. the GC-content), oligonucleotide frequencies, or other structural features.

Implicit phylogenetic inference

- no tree required
- based on sequence similarities and evolutionary distances.

The Structure of Xenology

Theorem

The following statements are equivalent:

1. *An (estimated) xenology relation is feasible.*
2. *Its graph-representation does not contain induced $K_1 + K_2$s.*
3. *Its graph-representation is a complete multipartite graph.*

terminology + proof in whiteboard