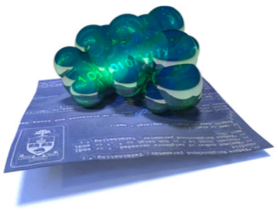
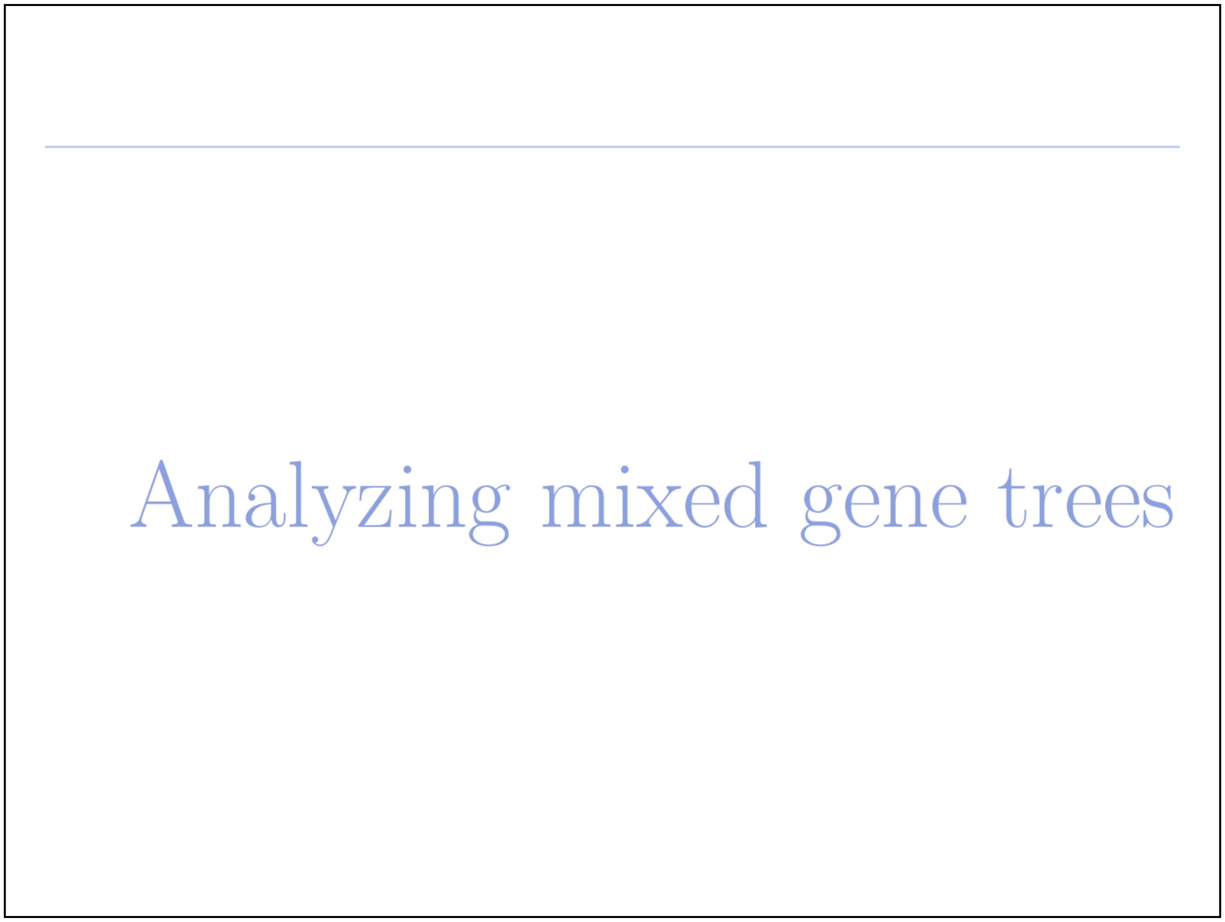
A
Bioinformatics
Course

Analyzing Phylogenetic Trees

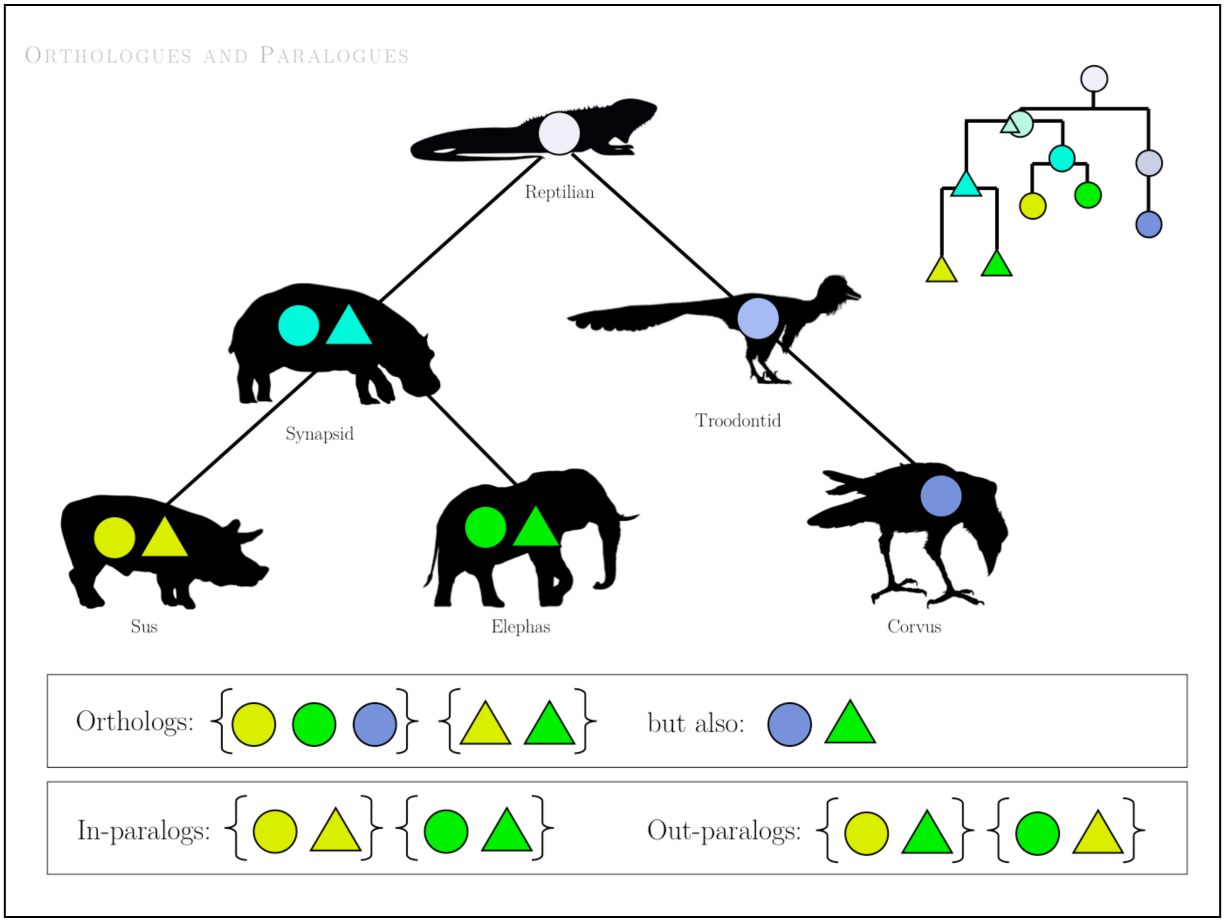


Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto



Analyzing mixed gene trees



Orthologues and Paralogues

Orthologs: {● ● ●} {▲ ▲} but also: ● ▲

In-paralogs: {● ▲} {● ▲} Out-paralogs: {● ▲} {● ▲}

The **interpretation** of phylogenetic trees involves assigning all branchpoints to *speciation* or *duplication* events, and reconciling the inferred topology and assignment with the (known) speciation tree of the organisms that are represented.

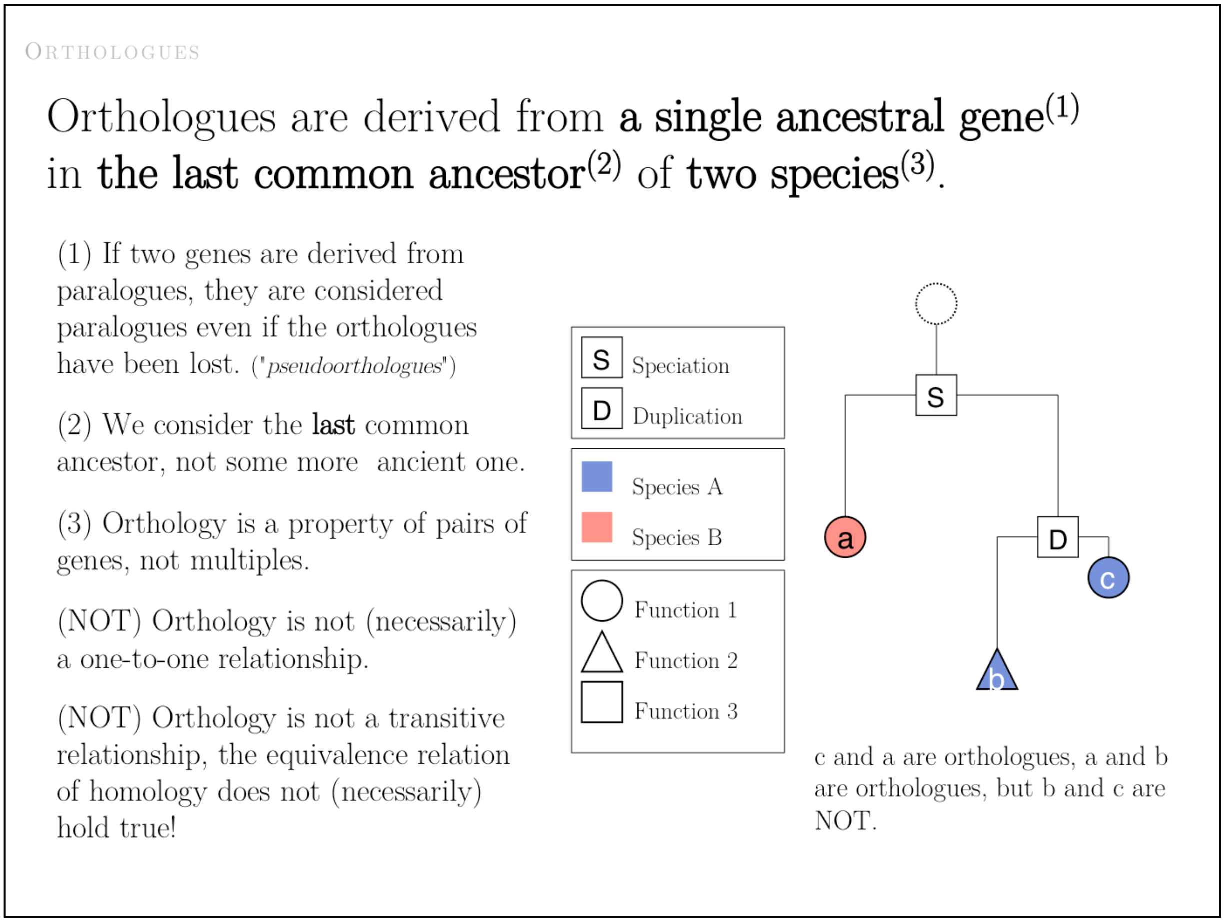
Definitions:

- **orthologues** arise through **speciation;**
- **paralogues** arise through **duplicaion.**

cf. Koonin EV. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet.* **39**:309-338.

The literature distinguishes between in-paralogues (within one species) and out paralogues (between species). However there is no distinct term for the **RBM gene** (or BBH – Bidirectional Best Hit), one just assumes that RBMs **are** orthologues although even in bacteria this is true in (only) about 95% of the cases.

cf. Wolf & Koonin (2012) A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol Evol.* **4**:1286-94.



Orthologues

Orthologues are derived from **a single ancestral gene**[1] in **the last common ancestor**[2] of **two species**[3].

(1) If two genes are derived from paralogues, they are considered paralogues even if the orthologues have been lost. ("*pseudoorthologues*")

(2) We consider the **last** common ancestor, not some more ancient one.

(3) Orthology is a property of pairs of genes, not multiples.

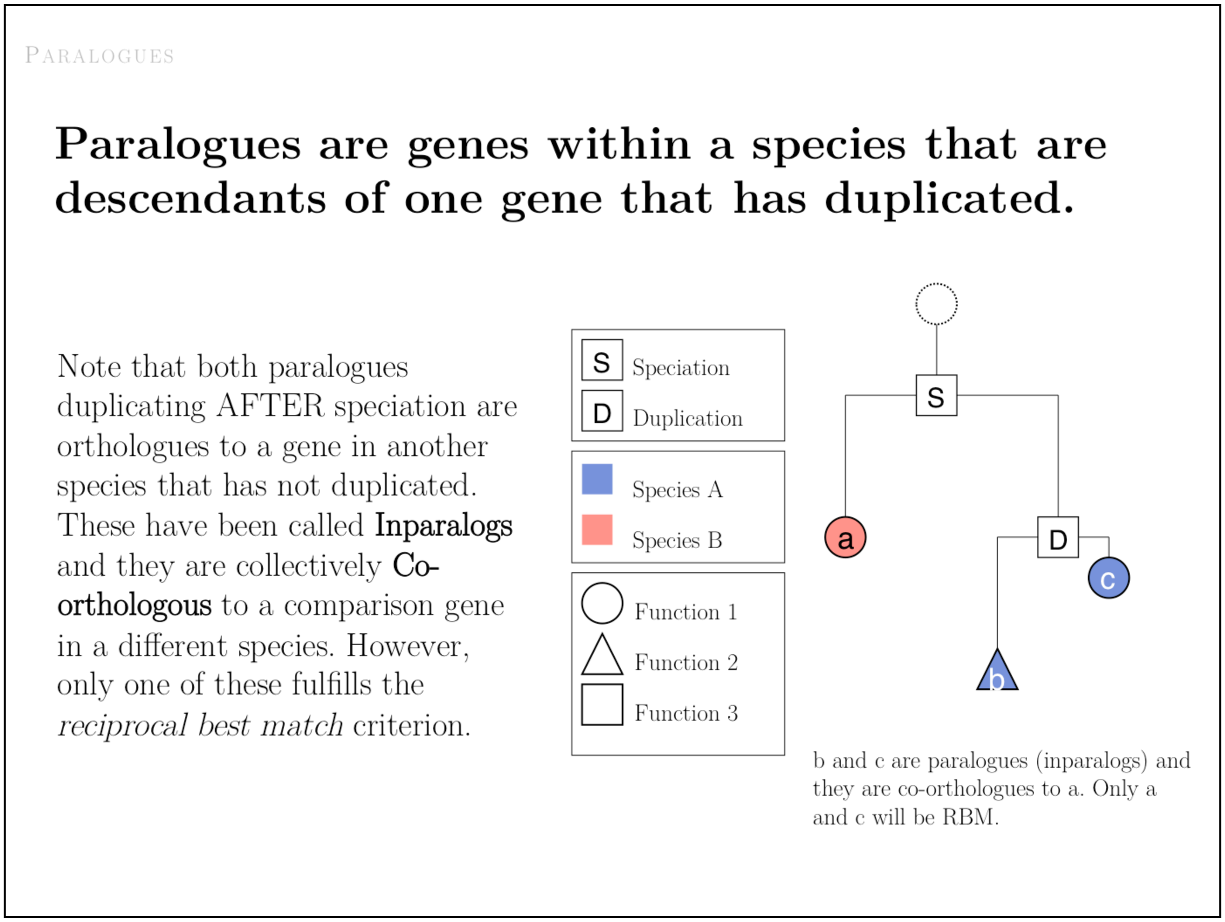
(NOT) Orthology is not (necessarily) a one-to-one relationship.

(NOT) Orthology is not a transitive relationship, the equivalence relation of homology does not (necessarily) hold true!

c and a are orthologues, a and b are orthologues, but b and c are NOT.

To analyze mixed gene trees, we evaluate each branch point and determine whether it represents a speciation or a duplication event.

Speciation events give rise to orthologues, duplication events give rise to paralogues.



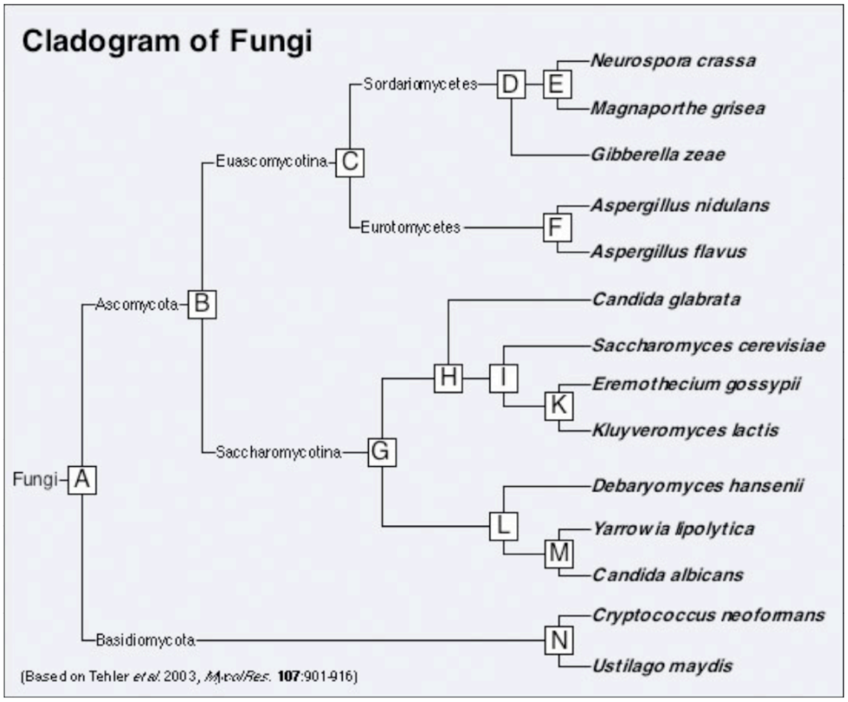
Paralogues are genes within a species that are descendants of one gene that has duplicated.

Note that both paralogues duplicating AFTER speciation are orthologues to a gene in another species that has not duplicated. These have been called **Inparalogs** and they are collectively **Co-orthologous** to a comparison gene in a different species. However, only one of these fulfills the *reciprocal best match* criterion.

b and c are paralogues (inparalogs) and they are co-orthologues to a. Only a and c will be RBM.

Gene Tree (Orthologues)

For orthologous genes, the gene tree should recapitulate the sequence of speciation events.



(Based on Tehler *et al.* 2003, *Mycol Res.* **107**:901-916)

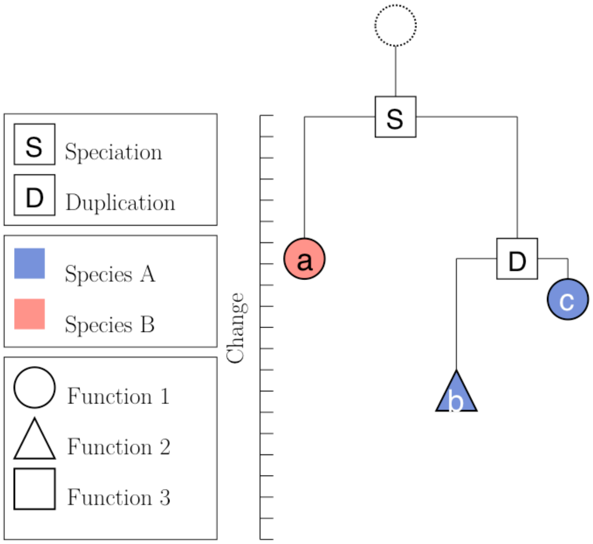
Reciprocal Best Match (RBM)

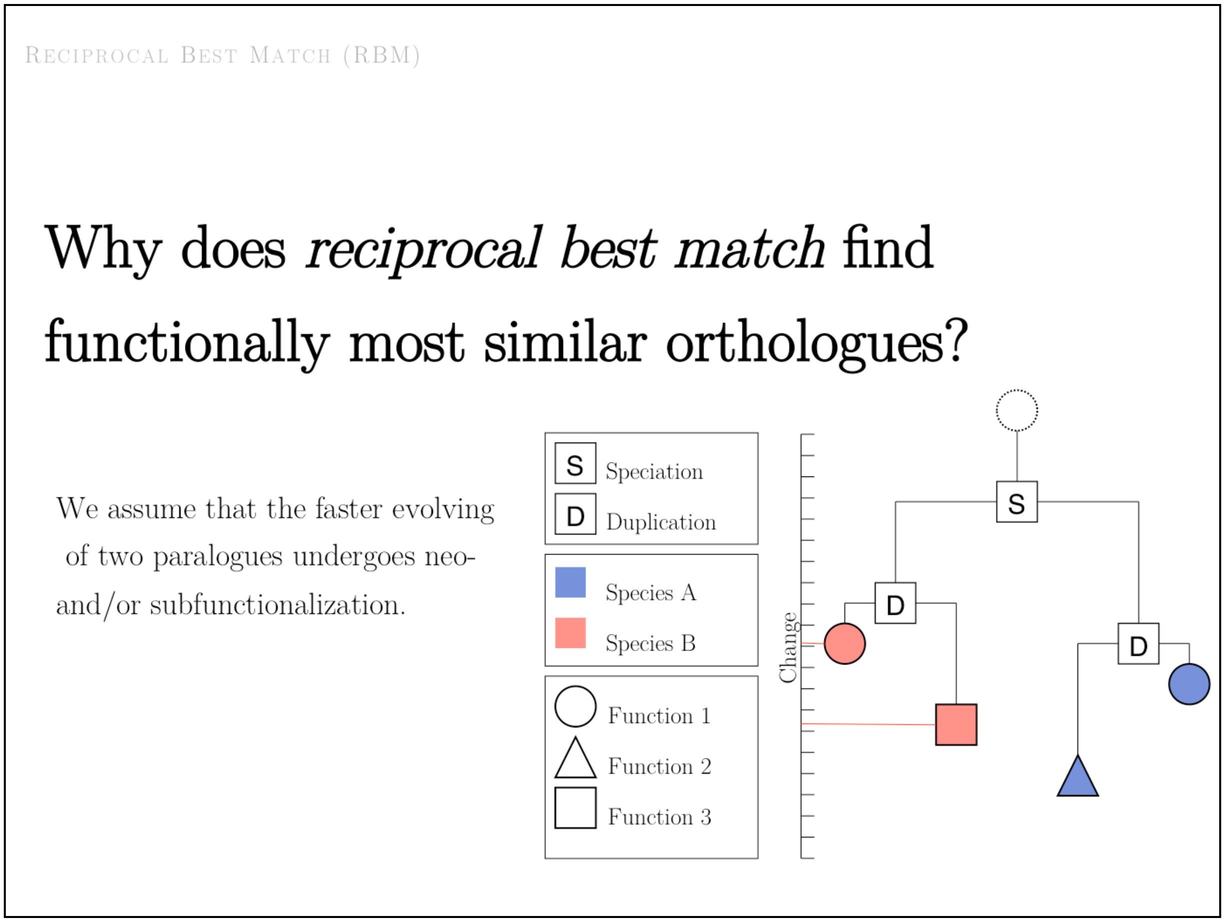
Reciprocal best match finds the respectively most similar genes in a genome. If the acquisition of a new function involves a period of accelerated evolution, duplicated genes that retain their function will be more similar to each other than those that change.

Why does *reciprocal best match* find orthologues?

Because we assume that after gene duplication the two descendants evolve at different rates. One paralogue will evolve at a faster rate from the ancestor than the other.

Reciprocal best match finds the respectively most similar genes in a genome.





Reciprocal Best Match (RBM)

Why does *reciprocal best match* find functionally most similar orthologues?

We assume that the faster evolving of two paralogues undergoes neo- and/or subfunctionalization.

S Speciation

D Duplication

🟦 Species A

🟥 Species B

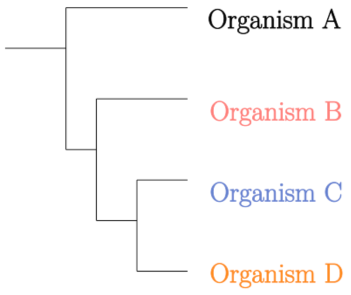
○ Function 1

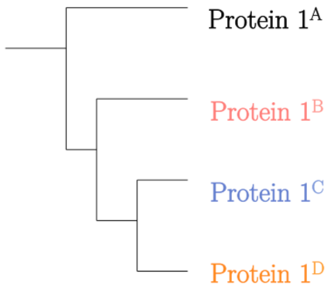
△ Function 2

□ Function 3

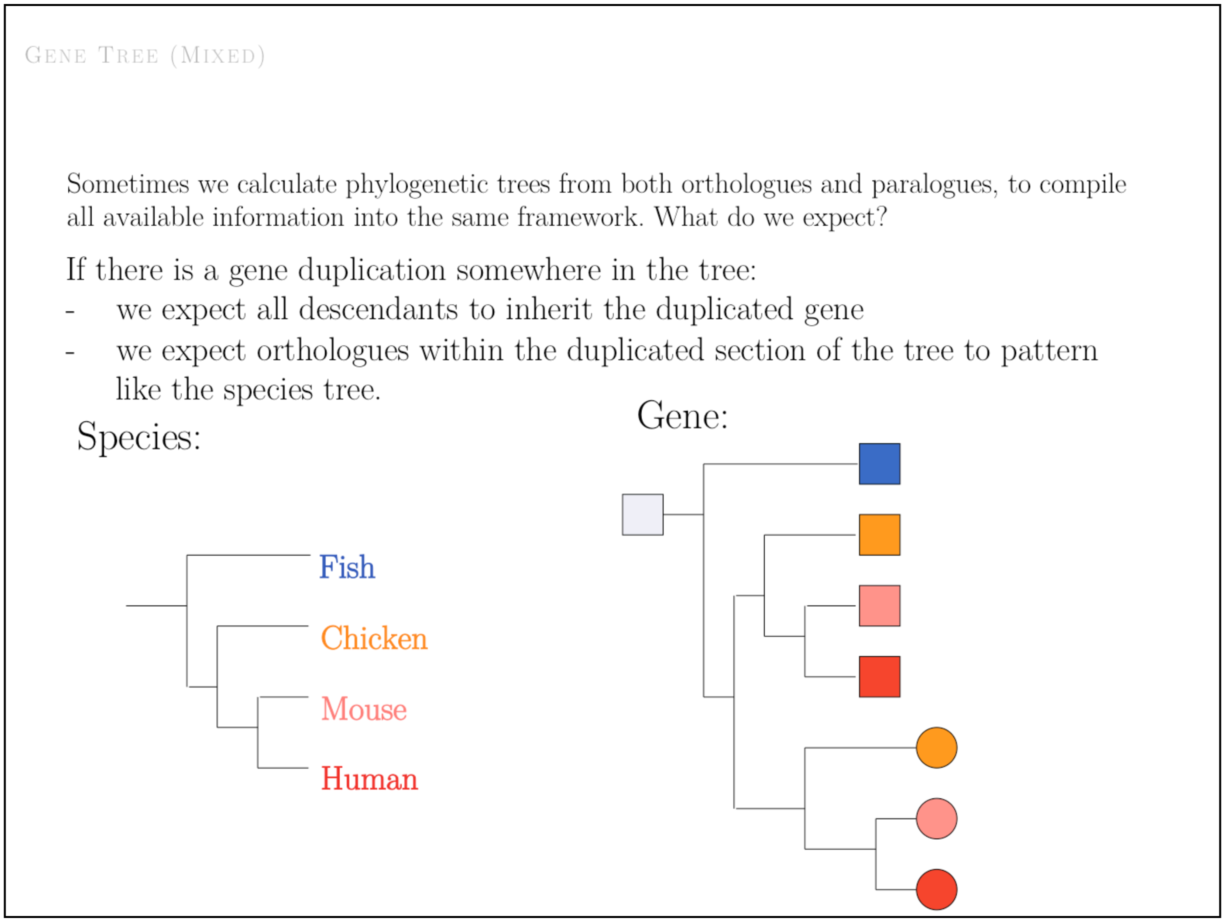
Simply add up the vertical path lengths in this tree between all pairs across the speciation node to show that RBM indeed finds the two genes with conserved Function 1, and not the pairs with either of the novel Functions 2 or 3.

Gene Tree (Orthologues)





If you consider a single, orthologous gene, you should expect that the gene tree recapitulates at least the topology of the cladogram, hopefully also the branch lengths of the phylogram.



Gene Tree (Mixed)

Sometimes we calculate phylogenetic trees from both orthologues and paralogues, to compile all available information into the same framework. What do we expect?

If there is a gene duplication somewhere in the tree:

- we expect all descendants to inherit the duplicated gene
- we expect orthologues within the duplicated section of the tree to pattern like the species tree.

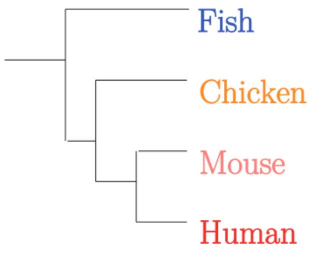
Species:

Gene:

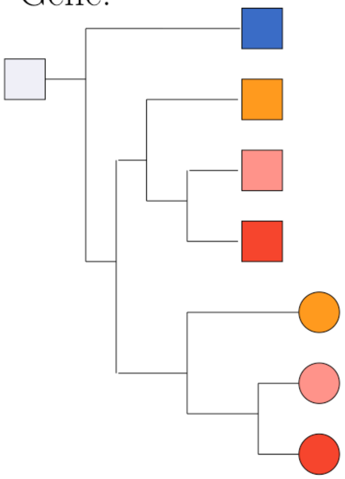
Gene Tree (Mixed)

The duplication point lies **before the cenancestor of the included species** and **after the cenancestor of included/excluded species.**

Species:



Gene:

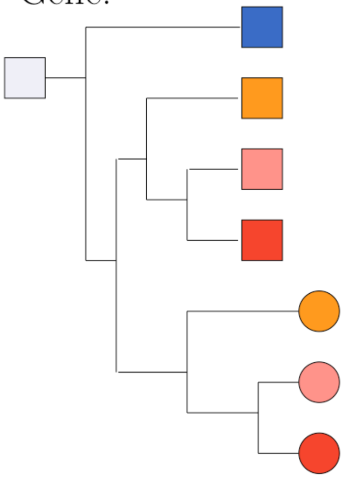


Gene-loss, lack of resolution of branching points and HGT can complicate the analysis.

To analyse a mixed gene tree:

- we need a species tree for reference
- we need to keep track of species in the tree
- we need to keep track of groups of orthologues

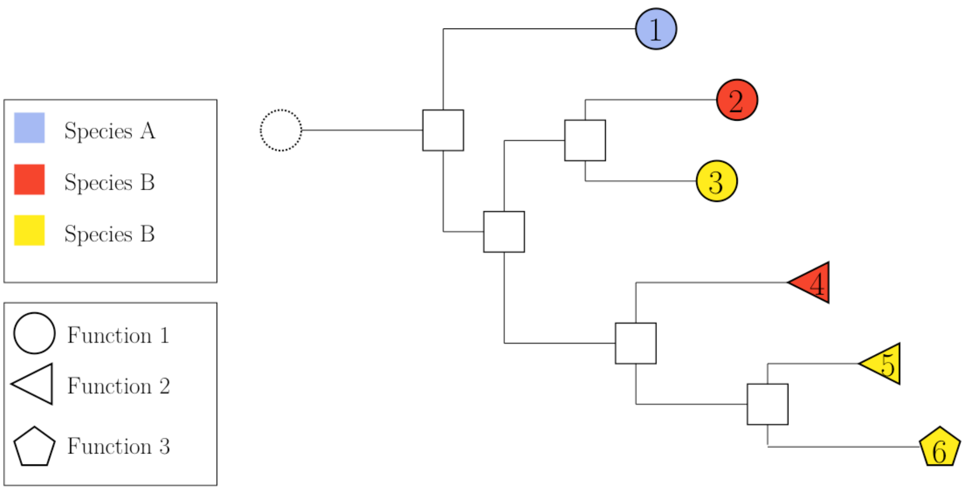
Gene:



Gene Tree (Mixed)

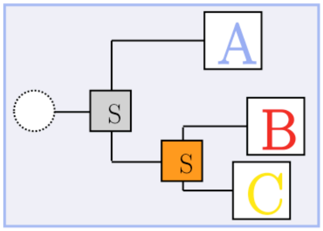
Assume we have the following mixed gene tree.

How did it evolve? What Speciations and duplications led to this tree?



Gene Tree (Mixed)

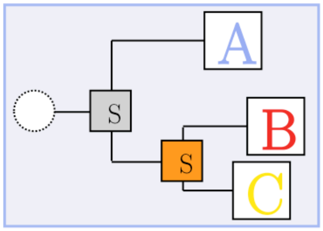
Always start from a reference model - the phylogram (or at least cladogram) of species. Whatever else happens over time, all variation is constrained by the periods of *joint* or *independent* evolution that is due to the *sequence of speciation events*, represented by the branching topology of the tree.



- 🟦 Species A
- 🟥 Species B
- 🟨 Species B

Gene Tree (Mixed)

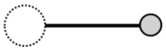
Some gene evolves through the mists of time, part of– and passenger in– an organism that contains it.



🟦 Species A

🟥 Species B

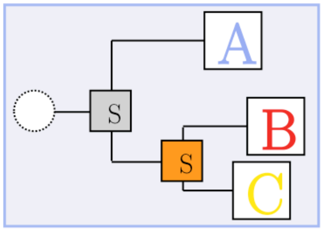
🟨 Species B



○ Function 1

Gene Tree (Mixed)

A Speciation generates two orthologues that evolve further by *separately* accumulating mutations.



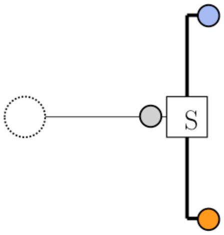
S Speciation

🟦 Species A

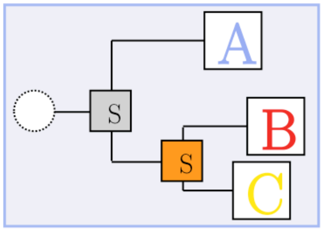
🟥 Species B

🟨 Species B

○ Function 1



Gene Tree (Mixed)



Further speciation events generate all OTUs . Obviously we expect the phylogram of orthologues to closely resemble the species tree.

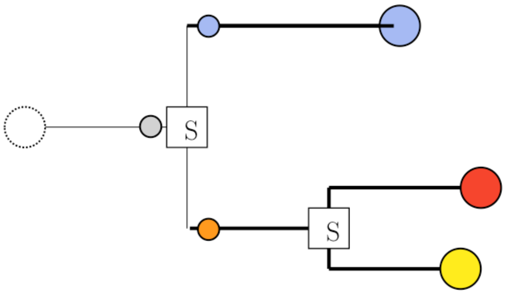
S Speciation

🟦 Species A

🟥 Species B

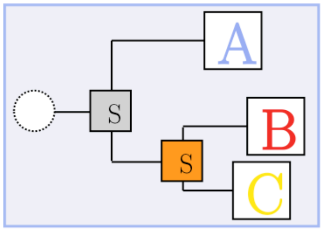
🟨 Species B

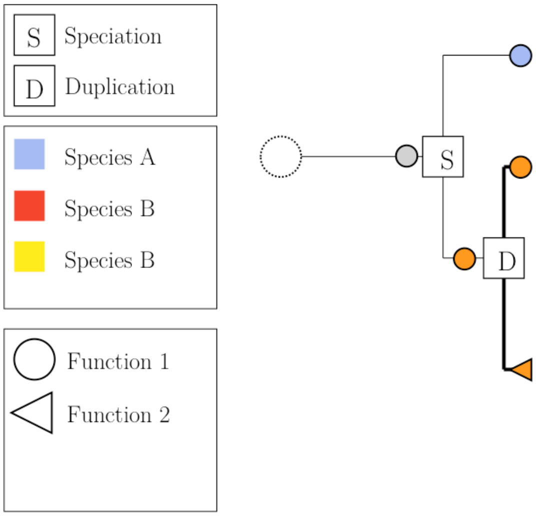
○ Function 1



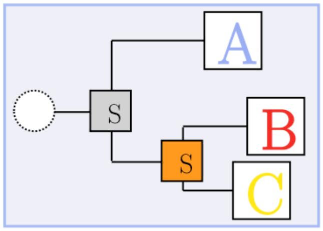
Gene Tree (Mixed)

Hoever, if there is a duplication event in the tree – in this case *before* the speciation event we just discussed, the story changes: a duplication event generates a **copy** of a gene *in one organism*. Typically, one of the copies may acquire a new function.





Gene Tree (Mixed)



Consequently, after the duplication, nothing is changed for *one of the duplicates* as the species separate...

S Speciation

D Duplication

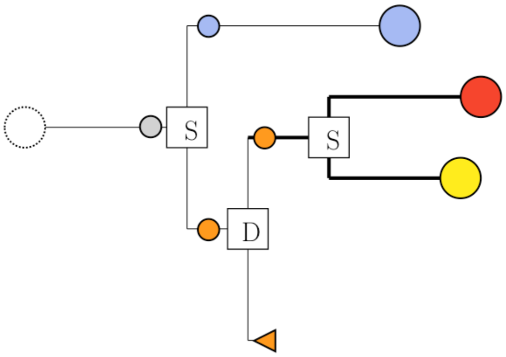
🟦 Species A

🟥 Species B

🟨 Species B

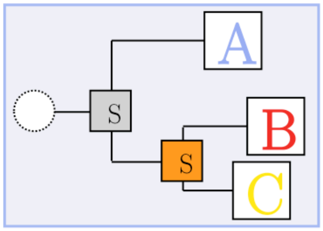
○ Function 1

◁ Function 2



Gene Tree (Mixed)

... while the *other duplicate* evolves according to the same speciation pattern as its sibling (albeit, typically with different rates). **Every duplication creates a copy of the branching pattern of the species tree.**



S Speciation

D Duplication

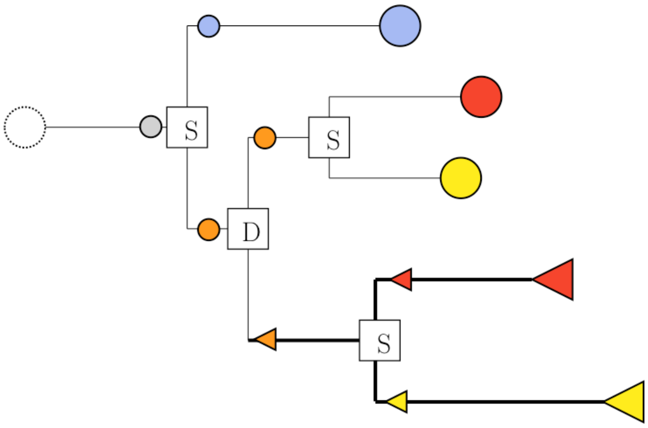
🟦 Species A

🟥 Species B

🟨 Species B

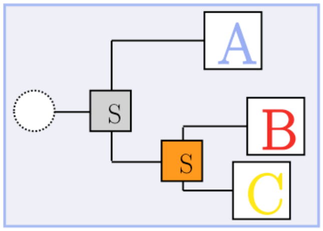
○ Function 1

◁ Function 2



Gene Tree (Mixed)

The final tree can be decomposed into full or partial copies of the species tree. In real life, *uncertain branching order* and *gene loss* complicates the tree. But we can always look for the best match to the model.



S Speciation

D Duplication

🟦 Species A

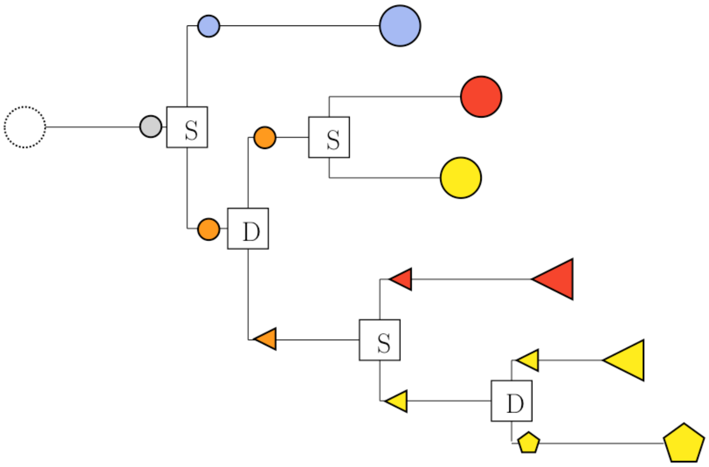
🟥 Species B

🟨 Species B

○ Function 1

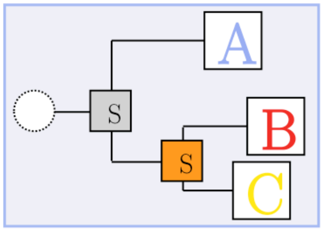
◁ Function 2

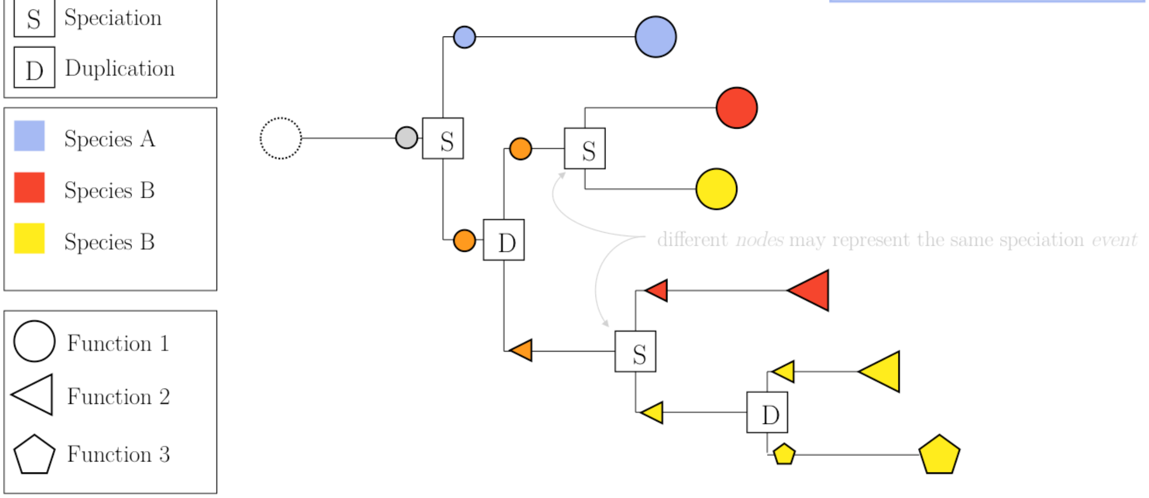
⬠ Function 3



Gene Tree (Mixed)

Note: the number of *duplication nodes* corresponds exactly to the number of duplication events. The number of *speciation nodes* is determined by the number and size of branches that have been inserted.



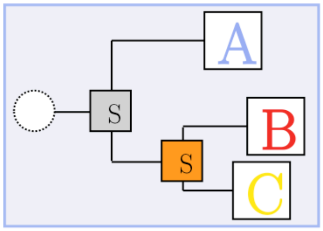


different *nodes* may represent the same speciation *event*

Gene Tree (Mixed)

To *analyze* a tree, you perform this process in reverse!

Label all speciation and duplication events in this tree. Define sets of most similar orthologues and of paralogues.



S Speciation

D Duplication

🟦 Species A

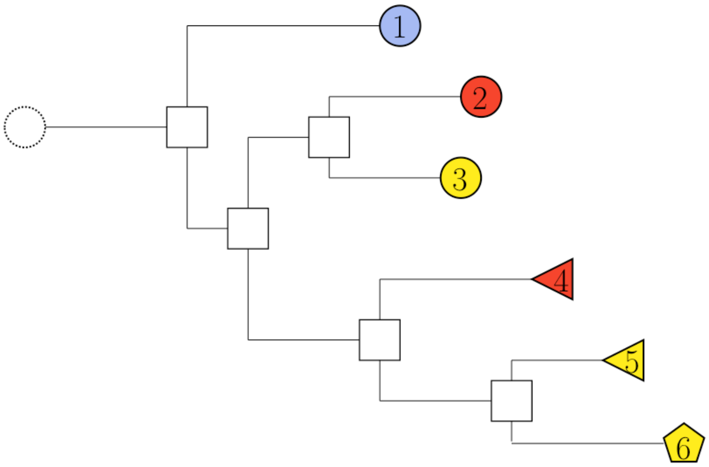
🟥 Species B

🟨 Species B

○ Function 1

◁ Function 2

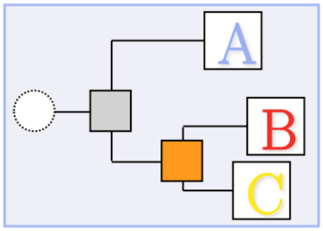
⬠ Function 3



Gene Tree (Mixed)

To *analyze* a tree, you perform this process in reverse!

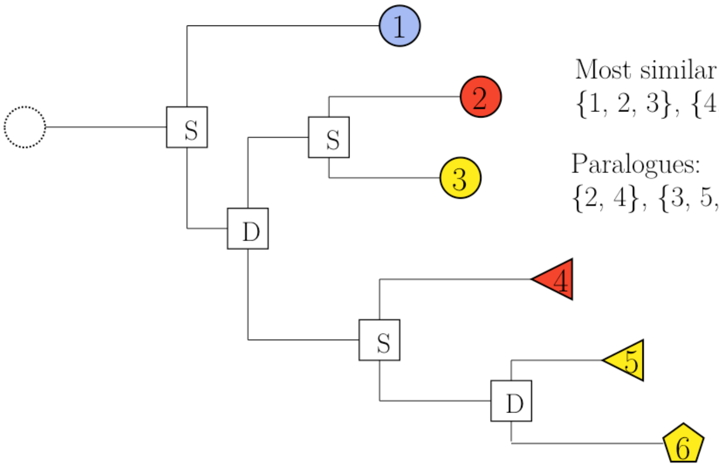
Label all speciation and duplication events in this tree. Define sets of most similar orthologues and of paralogues.



S	Speciation
D	Duplication

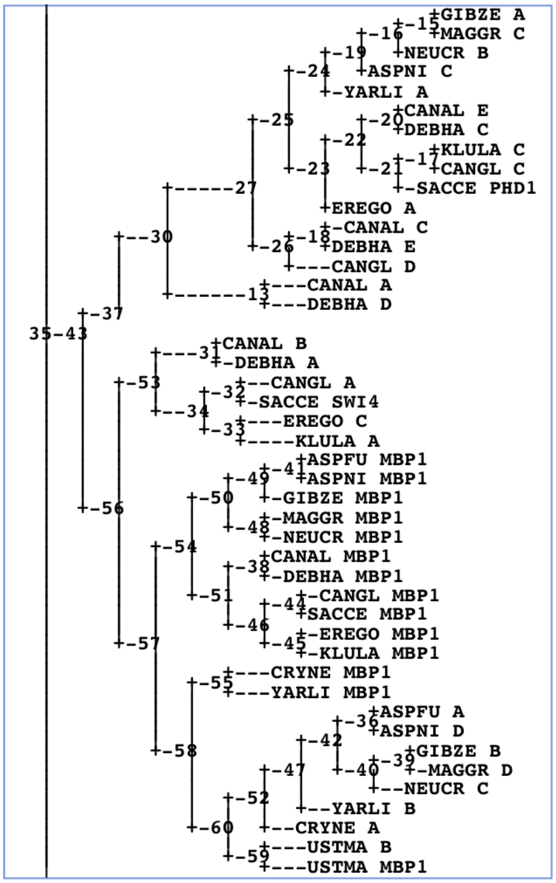
🟦	Species A
🟥	Species B
🟨	Species B

○	Function 1
◁	Function 2
⬠	Function 3



Most similar orthologues:
{1, 2, 3}, {4, 5}

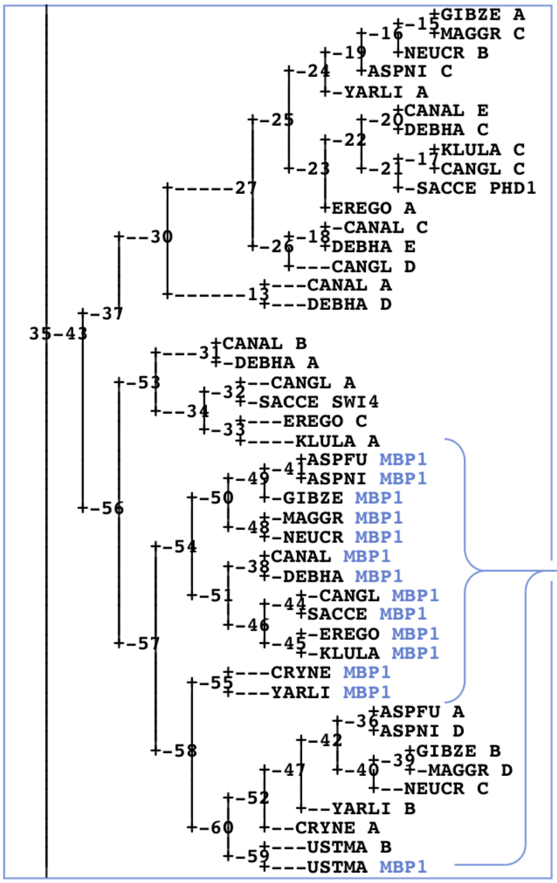
Paralogues:
{2, 4}, {3, 5, 6}

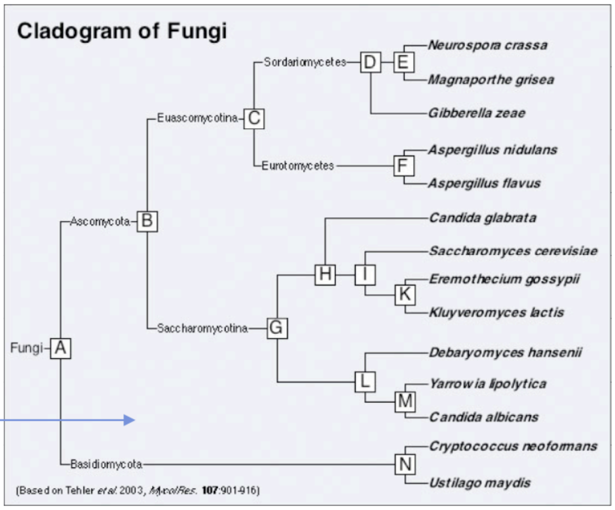


gene tree (mixed)

A selection of sequences (e.g. the result of a PSI-BLAST search contains orthologues and paralogues in the same tree. In order to interpret such a tree, we MUST distinguish between branch points that represent *duplication events* and branch points that represent *speciation events*.

gene tree (mixed)

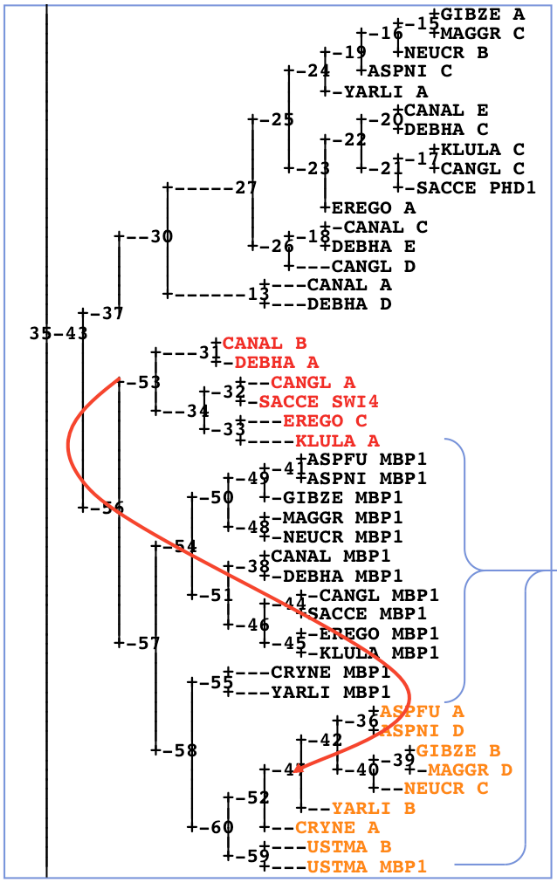




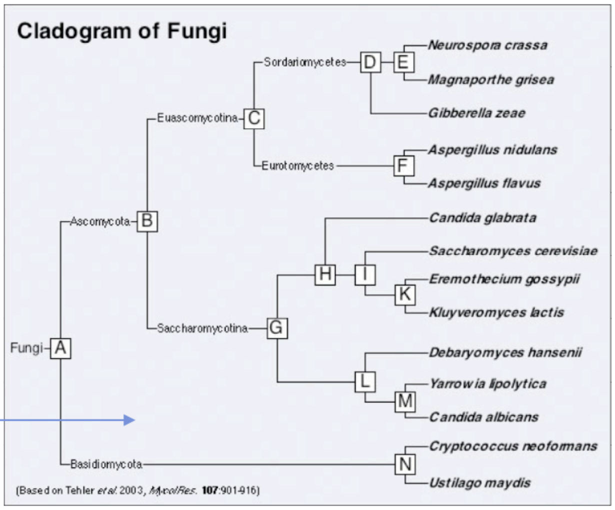
(Based on Tehler *et al.* 2003, *Mycol Res.* **107**:901-916)

Mbp1 orthologues pattern (mostly) according to the species tree in this example.

gene tree (mixed)



```
                              +-15+GIBZE A
                           +-16   +MAGGR C
                        +-19  +NEUCR B
                     +-24  +ASPNI C
                        +-YARLI A
                              +CANAL E
                  +-25     +-20+DEBHA C
                        +-22
                              +-17+KLULA C
                     +-23  +-21   +CANGL C
           +-----27               +-SACCE PHD1
                        +EREGO A
                     +-18+-CANAL C
       +--30      +-26   +DEBHA E
                     +---CANGL D
                       +---CANAL A
           +-------13
                       +---DEBHA D
    +-37
35-43
          +---31+CANAL B
                +-DEBHA A
       +-53       +--CANGL A
              +-32
                  +-SACCE SWI4
          +--34
                  +---EREGO C
              +-33
                  +----KLULA A
                         +ASPFU MBP1
                     +-41
                +-49     +ASPNI MBP1
             +-50    +-GIBZE MBP1
    +-56
                     +-MAGGR MBP1
                +-48
                     +-NEUCR MBP1
          +-54
                     +CANAL MBP1
                +-38
                     +-DEBHA MBP1
             +-51    +-44+-CANGL MBP1
                         +SACCE MBP1
                +-46
                         +-EREGO MBP1
       +-57          +-45
                         +-KLULA MBP1
                +---CRYNE MBP1
             +-55
                +---YARLI MBP1
                                +ASPFU A
                         +-36
                                +ASPNI D
                      +-42
          +-58                     +GIBZE B
                               +-39
                   +-47     +-40   +-MAGGR D
                               +--NEUCR C
                +-52
                      +--YARLI B
             +-60
                   +--CRYNE A
                   +---USTMA B
                +-59
                   +---USTMA MBP1
```



Cladogram of Fungi

(Based on Tehler *et al.* 2003, *Mycol Res.* **107**:901-916)

Violations of the species tree may suggest to re-evaluate the annotations as well as the topology of the gene tree.

Sources Of Error

Errors arise from:

Sampling (indels!)

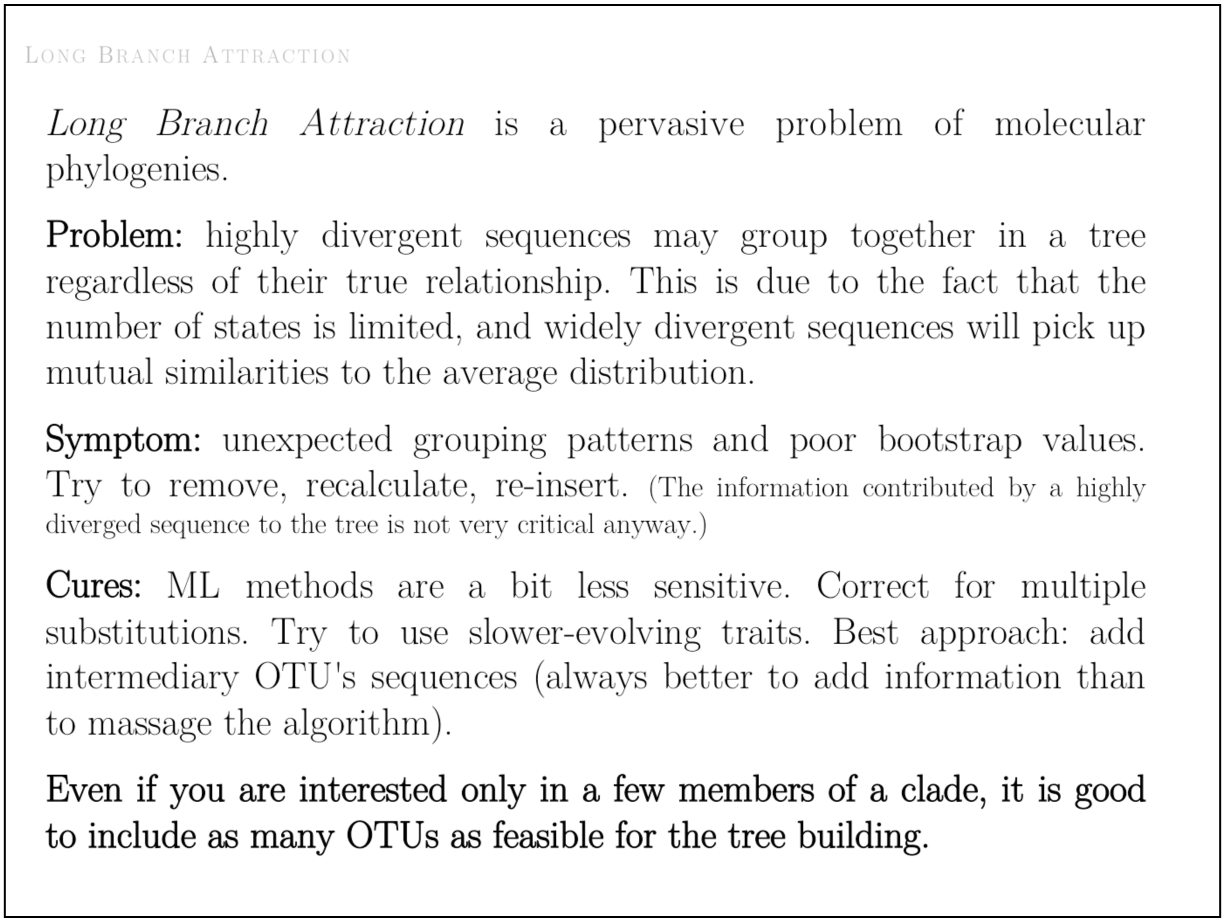
HGT

Methodology

Long branch attraction

...

Probably not from convergence though.



Long Branch Attraction

Long Branch Attraction is a pervasive problem of molecular phylogenies.

Problem: highly divergent sequences may group together in a tree regardless of their true relationship. This is due to the fact that the number of states is limited, and widely divergent sequences will pick up mutual similarities to the average distribution.

Symptom: unexpected grouping patterns and poor bootstrap values. Try to remove, recalculate, re-insert. (The information contributed by a highly diverged sequence to the tree is not very critical anyway.)

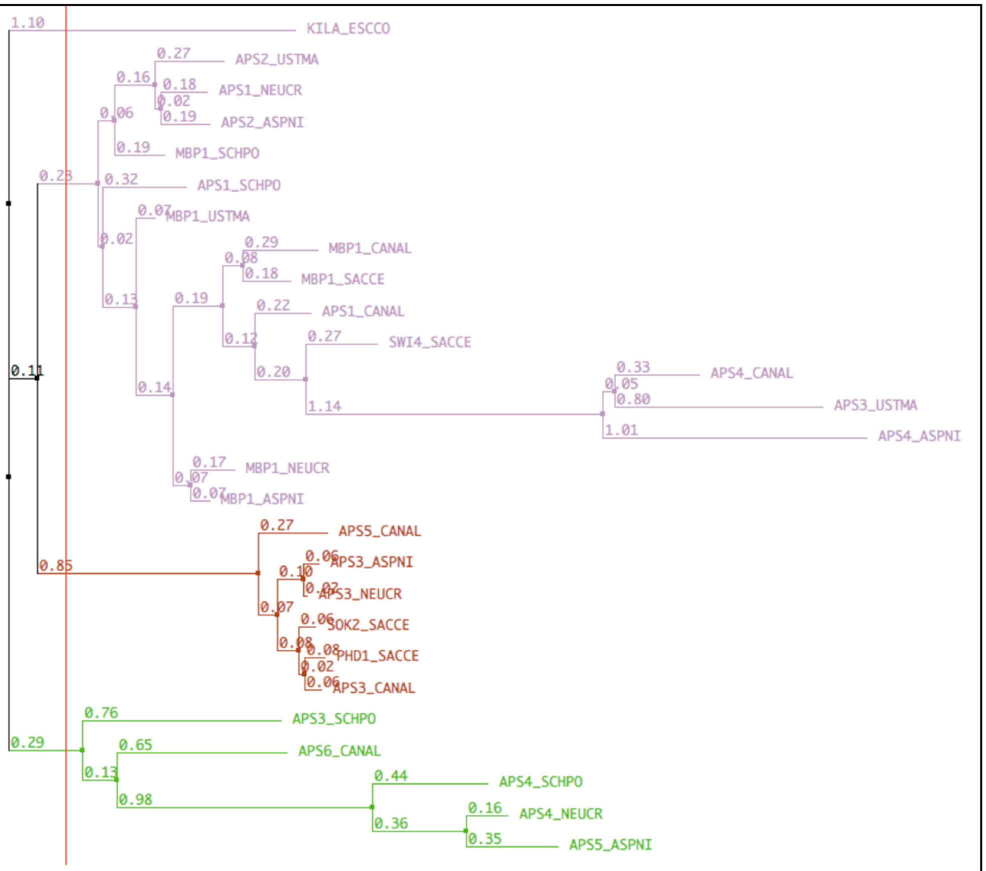
Cures: ML methods are a bit less sensitive. Correct for multiple substitutions. Try to use slower-evolving traits. Best approach: add intermediary OTU's sequences (always better to add information than to massage the algorithm).

Even if you are interested only in a few members of a clade, it is good to include as many OTUs as feasible for the tree building.

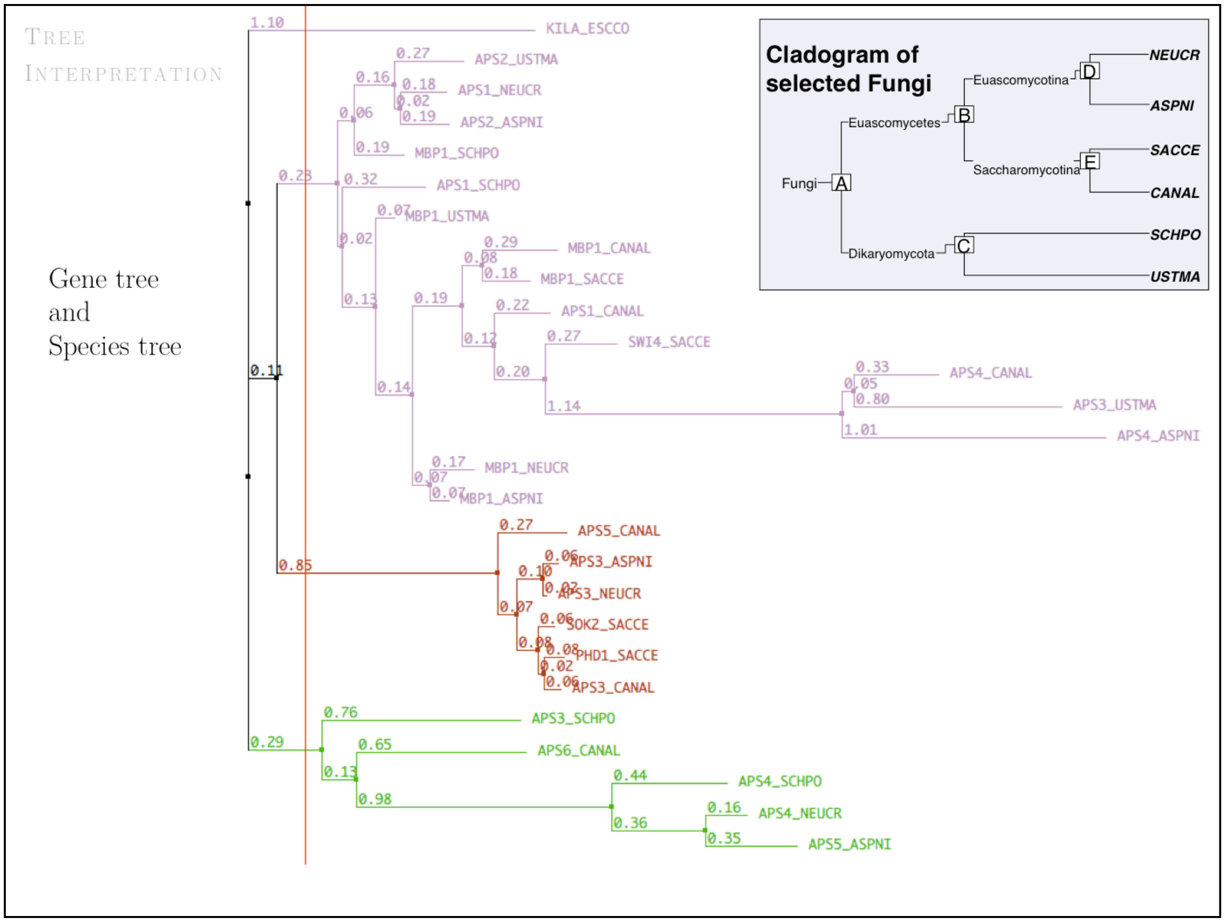
See also: http://en.wikipedia.org/wiki/Long_branch_attraction (this article deserves to be rewritten though, sounds a bit like a high-school project).

Recent paper: Kück P, Mayer C, Wägele J-W, Misof B (2012) Long Branch Effects Distort Maximum Likelihood Phylogenies in Simulations Despite Selection of the Correct Model. PLoS ONE 7(5): e36593. doi:10.1371/journal.pone.0036593.

Interpreting
a mixed
gene tree



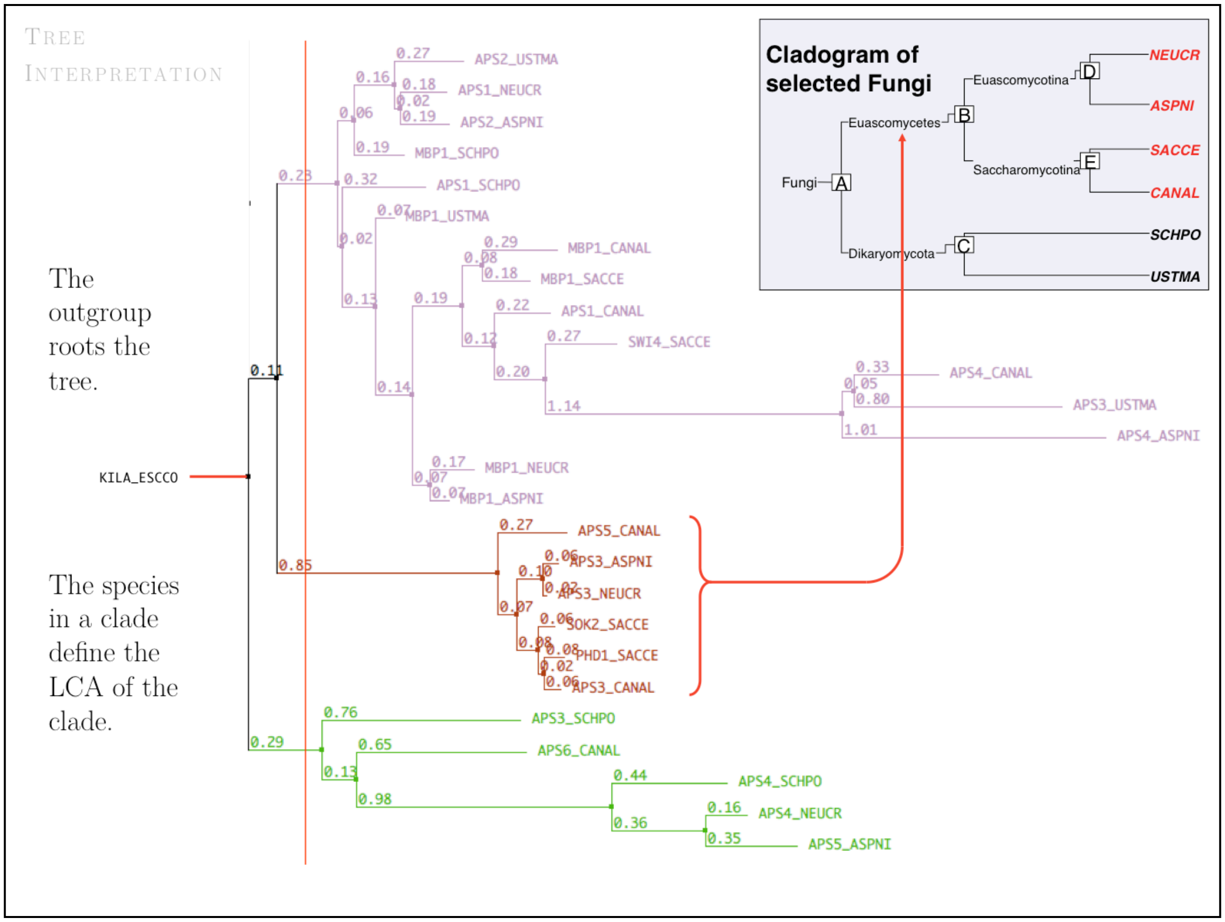
A worked example from course data, displayed in the Jalview tree window.



Tree Interpretation

Gene tree and Species tree

A worked example from course data, displayed in the Jalview tree window.

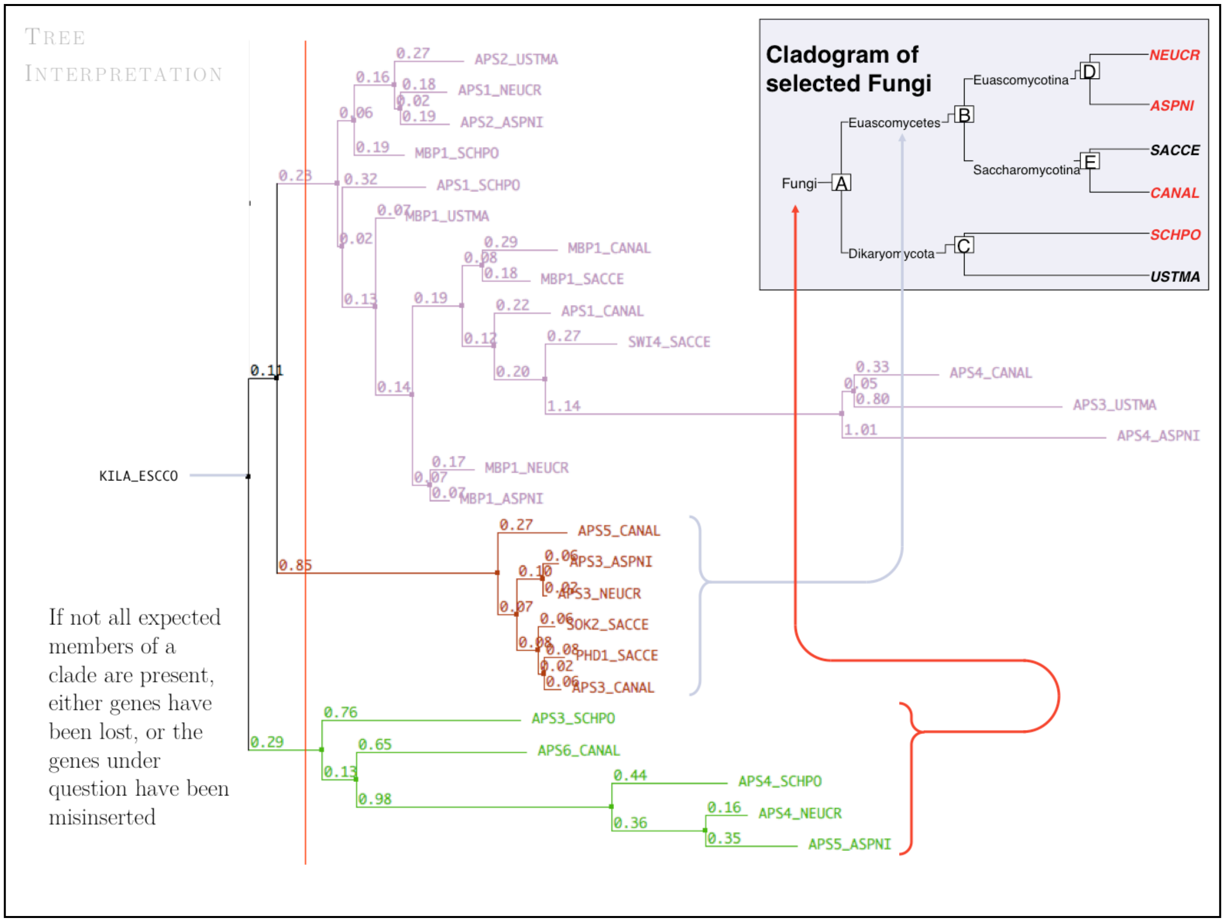


Tree Interpretation

The outgroup roots the tree.

The species in a clade define the LCA of the clade.

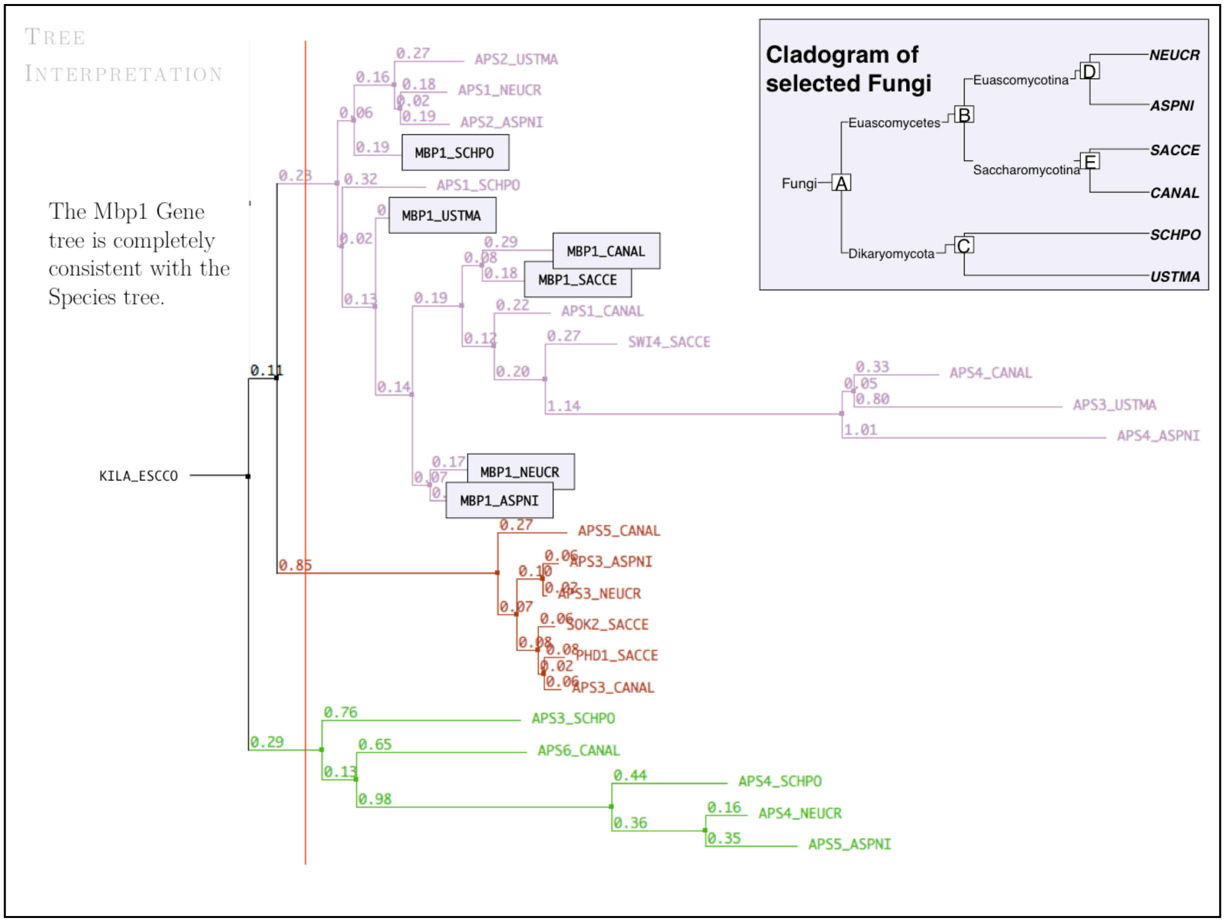
A worked example from course data, displayed in the Jalview tree window.



Tree Interpretation

If not all expected members of a clade are present, either genes have been lost, or the genes under question have been misinserted

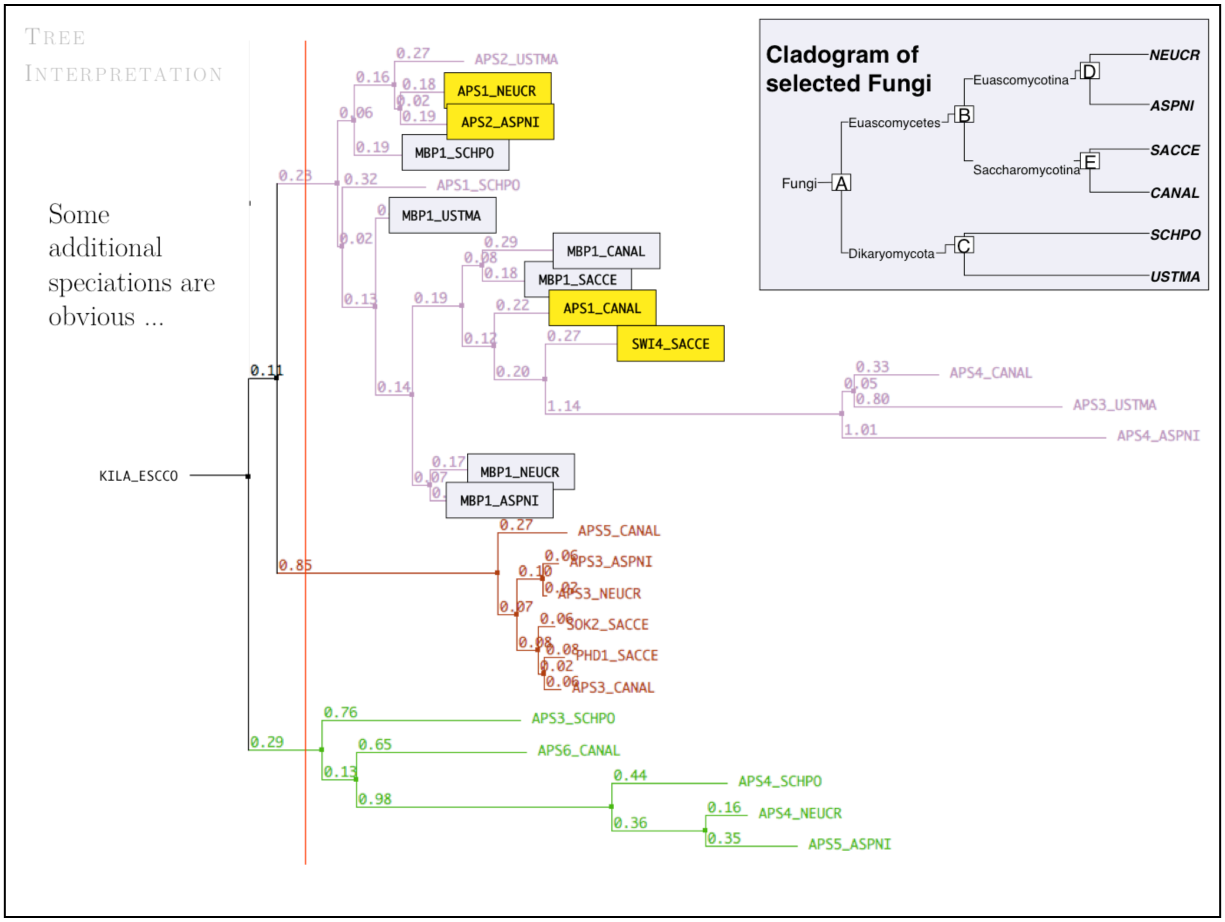
A worked example from course data, displayed in the Jalview tree window.



TREE INTERPRETATION

The Mbp1 Gene tree is completely consistent with the Species tree.

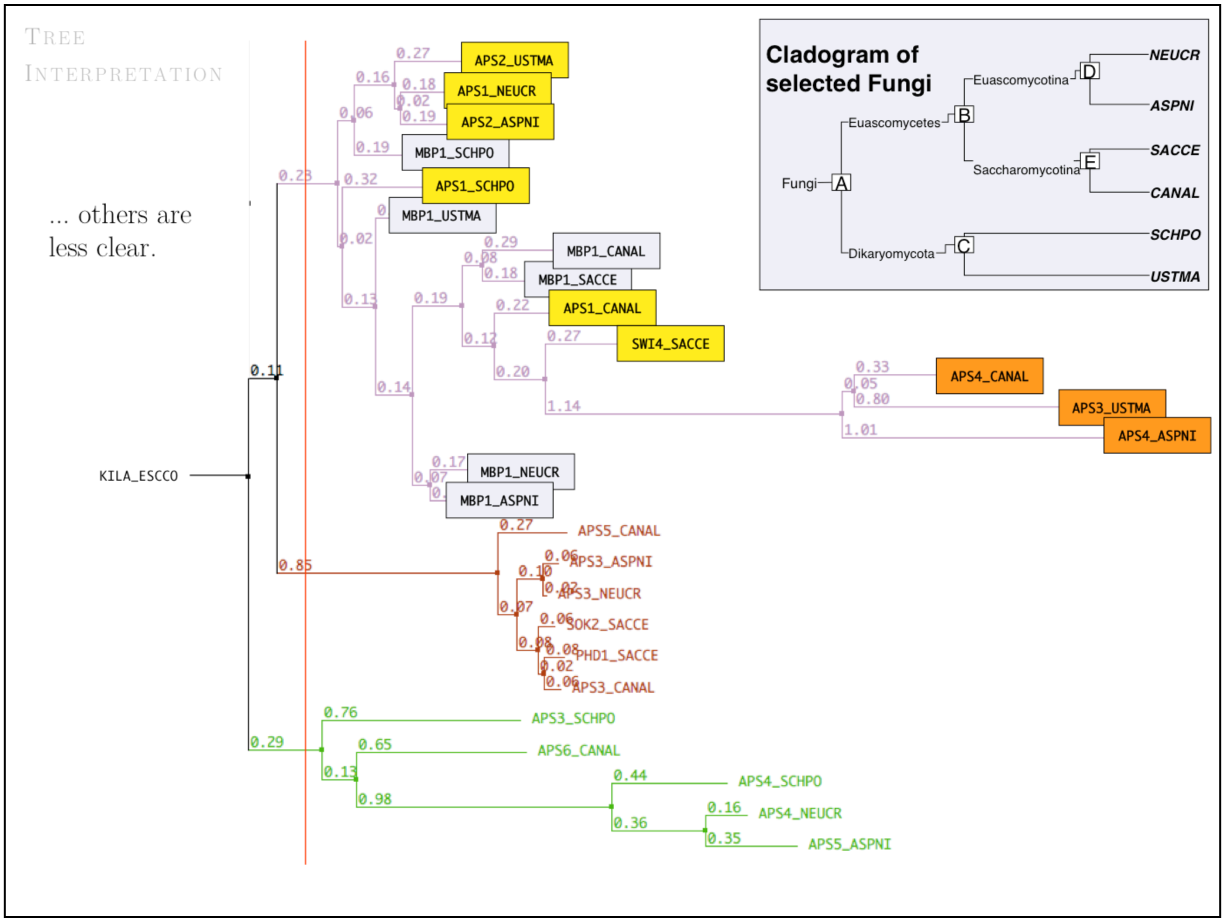
A worked example from course data, displayed in the Jalview tree window.



Tree Interpretation

Some additional speciations are obvious ...

A worked example from course data, displayed in the Jalview tree window.



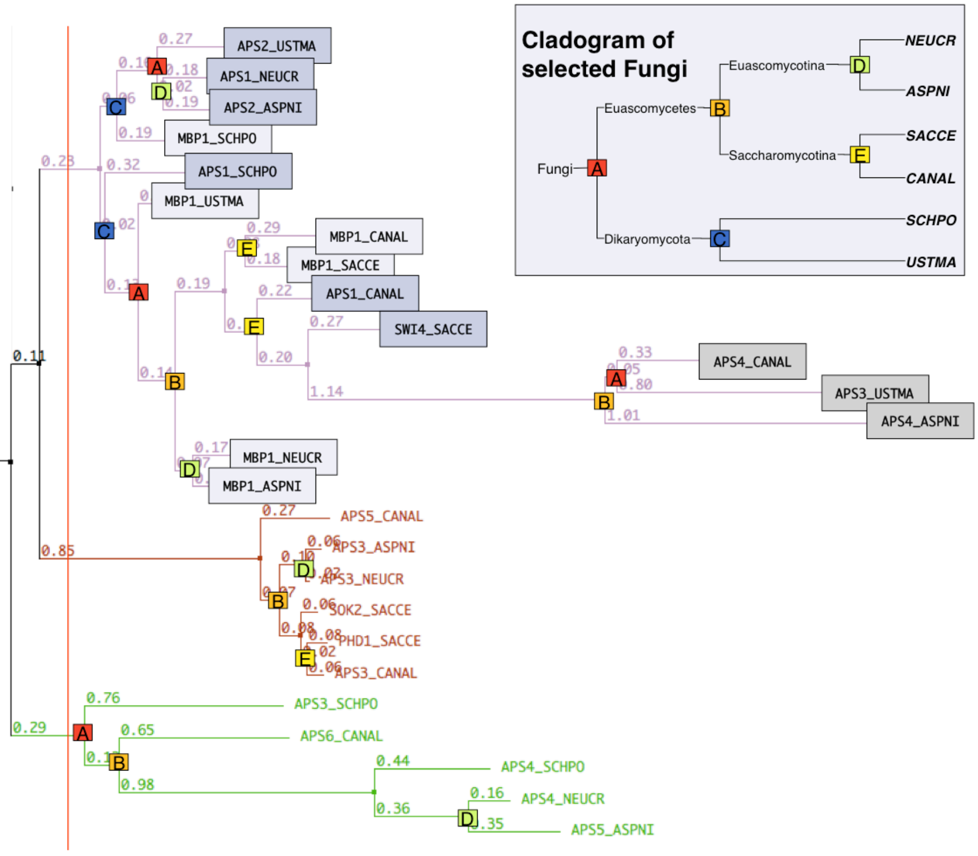
TREE INTERPRETATION

... others are less clear.

A worked example from course data, displayed in the Jalview tree window.

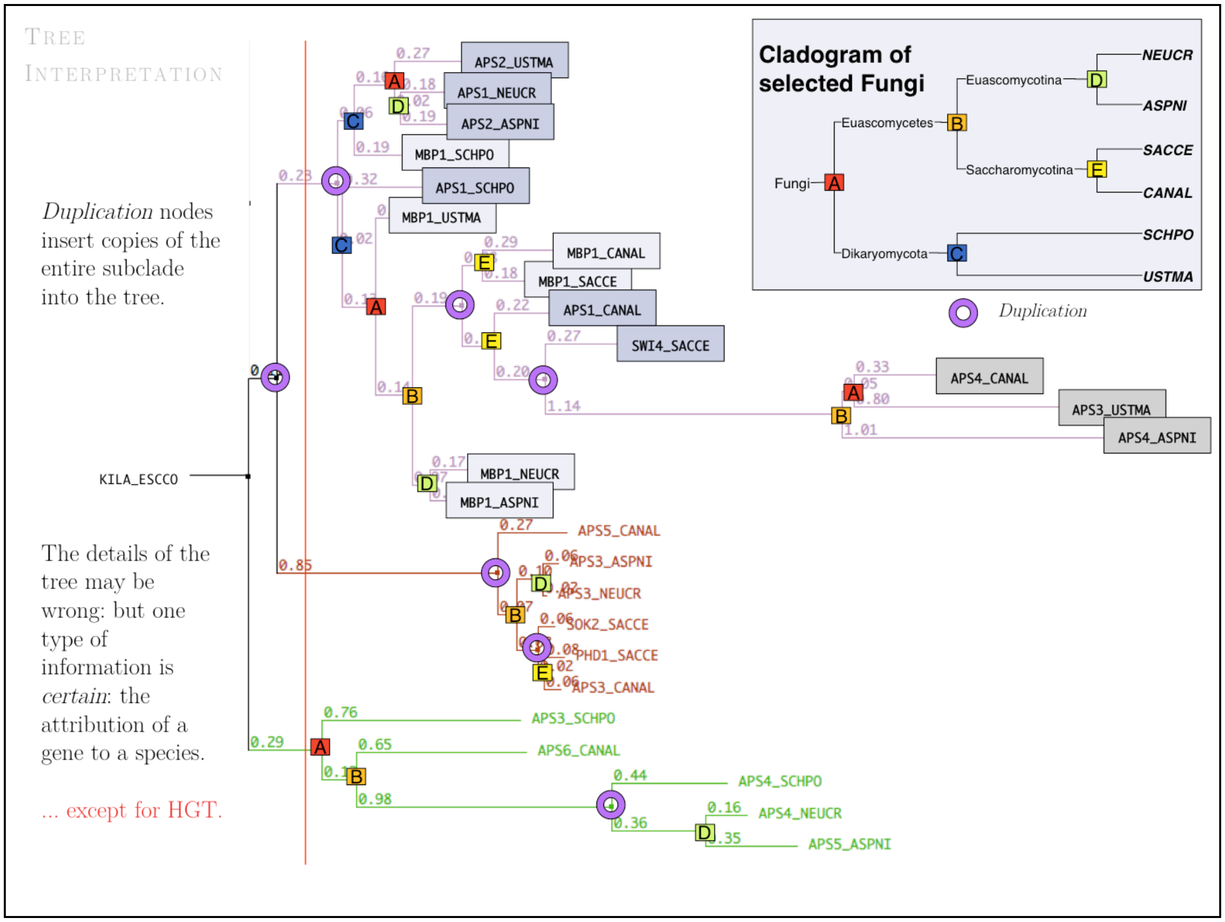
Tree Interpretation

Nodes can either correspond to *duplications* or *speciations*.



Speciation nodes are assigned according to the species that are the OTUs for a clade. All species must be accounted for.

A worked example from course data, displayed in the Jalview tree window.



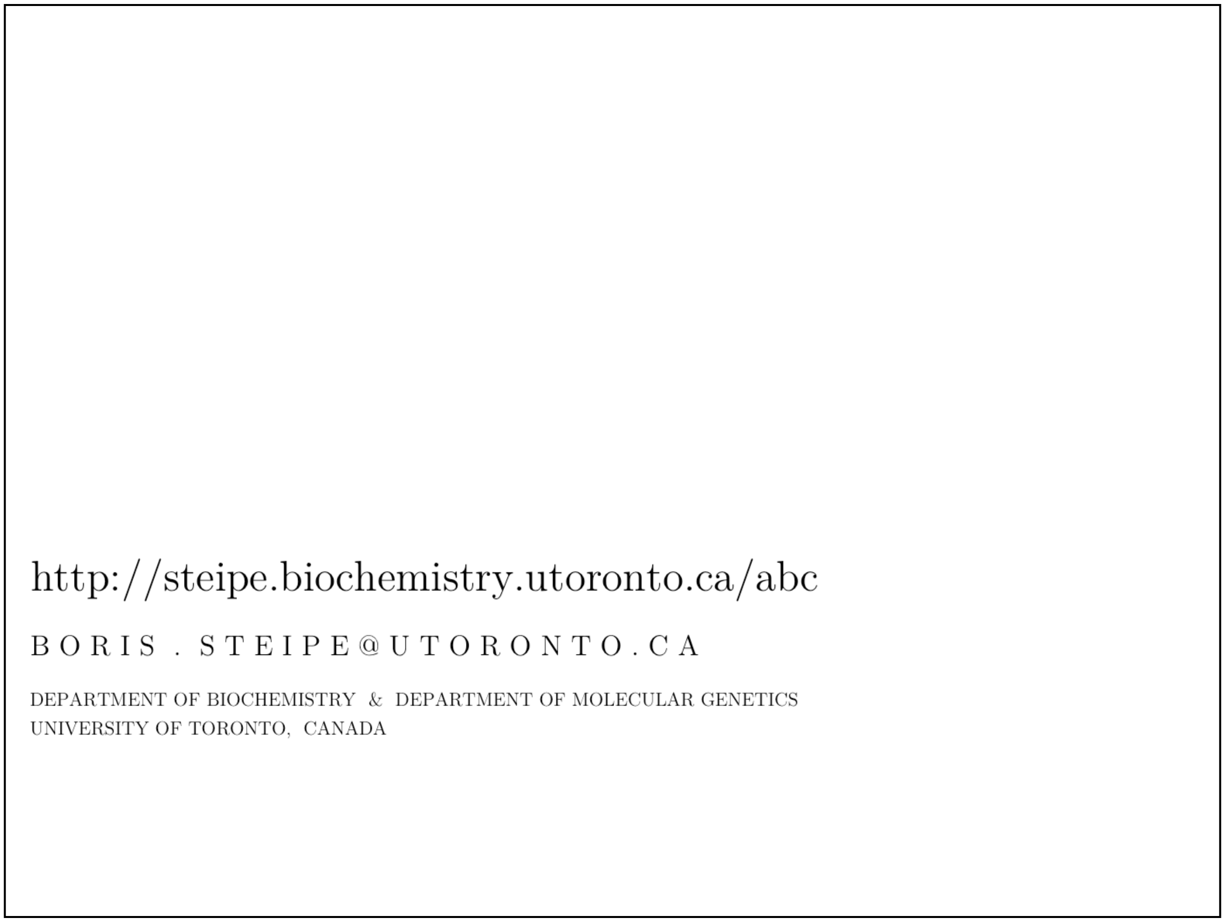
Tree Interpretation

Duplication nodes insert copies of the entire subclade into the tree.

The details of the tree may be wrong: but one type of information is *certain*: the attribution of a gene to a species.

... except for HGT.

A worked example from course data, displayed in the Jalview tree window.



http://steipe.biochemistry.utoronto.ca/abc

BORIS.STEIPE@UTORONTO.CA

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA