Improved Orthology Assignments for Functional and Evolutionary Genomics

Maribel Hernández Rosales

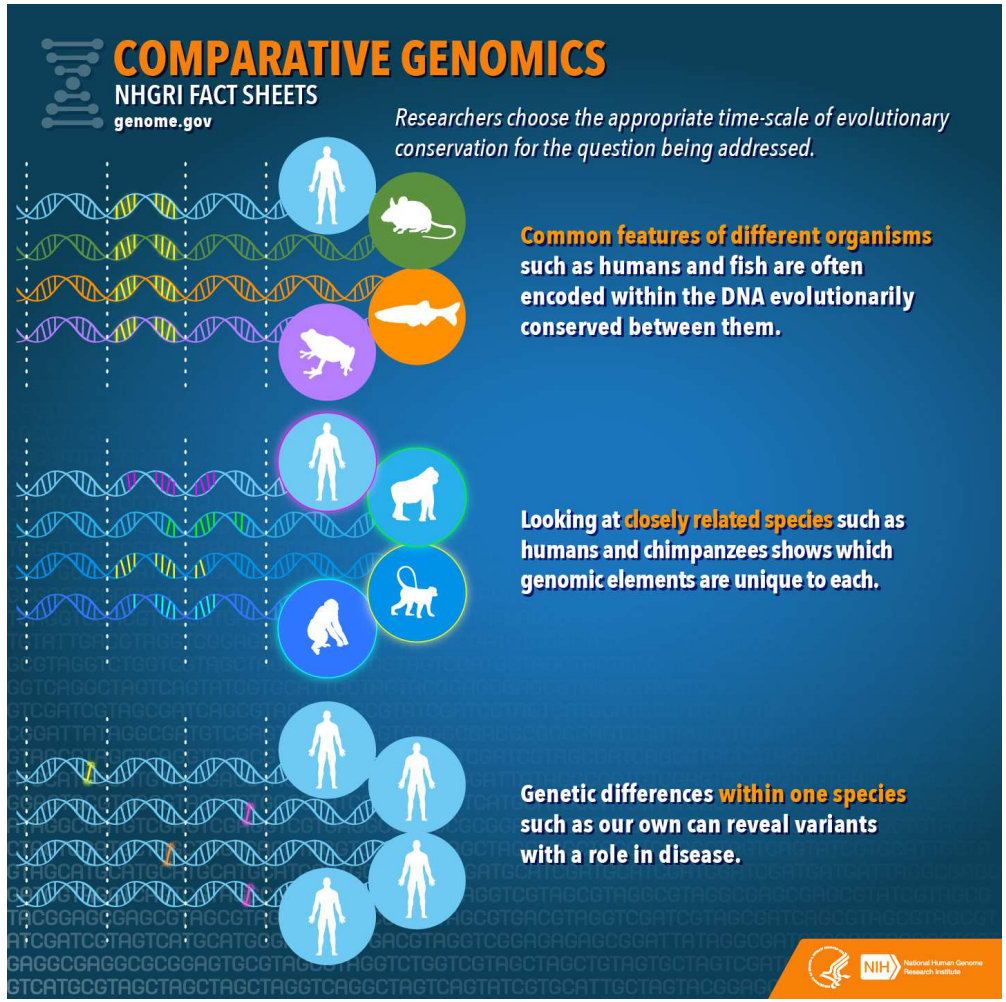
Conacyt - Instituto de Matemáticas, UNAM Juriquilla, México

8. November, 2018
Computational Challenges in High-Throughput Genomics
with Application to Precision Medicine

BIRS-CMO

Why is it important to distinguish orthologs from other type of homlogs?

- Comparative Genomics
- Functional Genomics
- Evolution Reconstruction

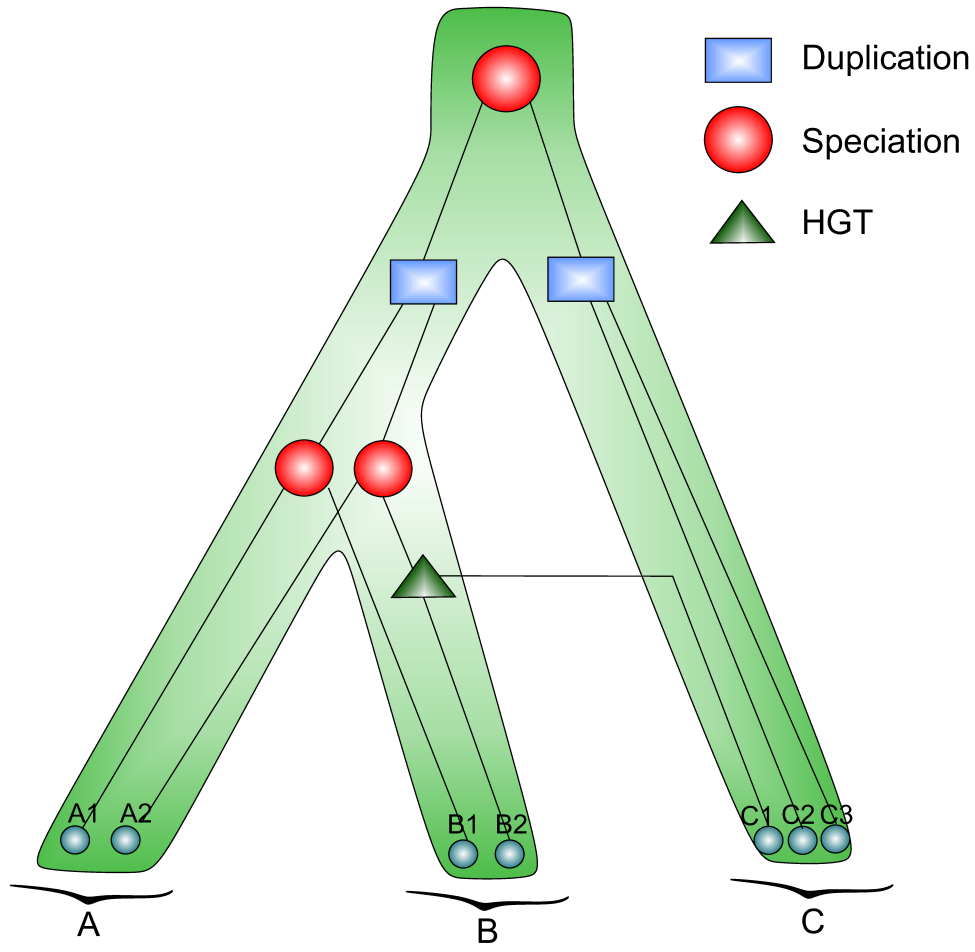


Events of gene evolution

- Speciation
- Gene duplication
- Horizontal gene transfer (HGT)
- Gene loss
- Rearrangements, retrotransposition, ...

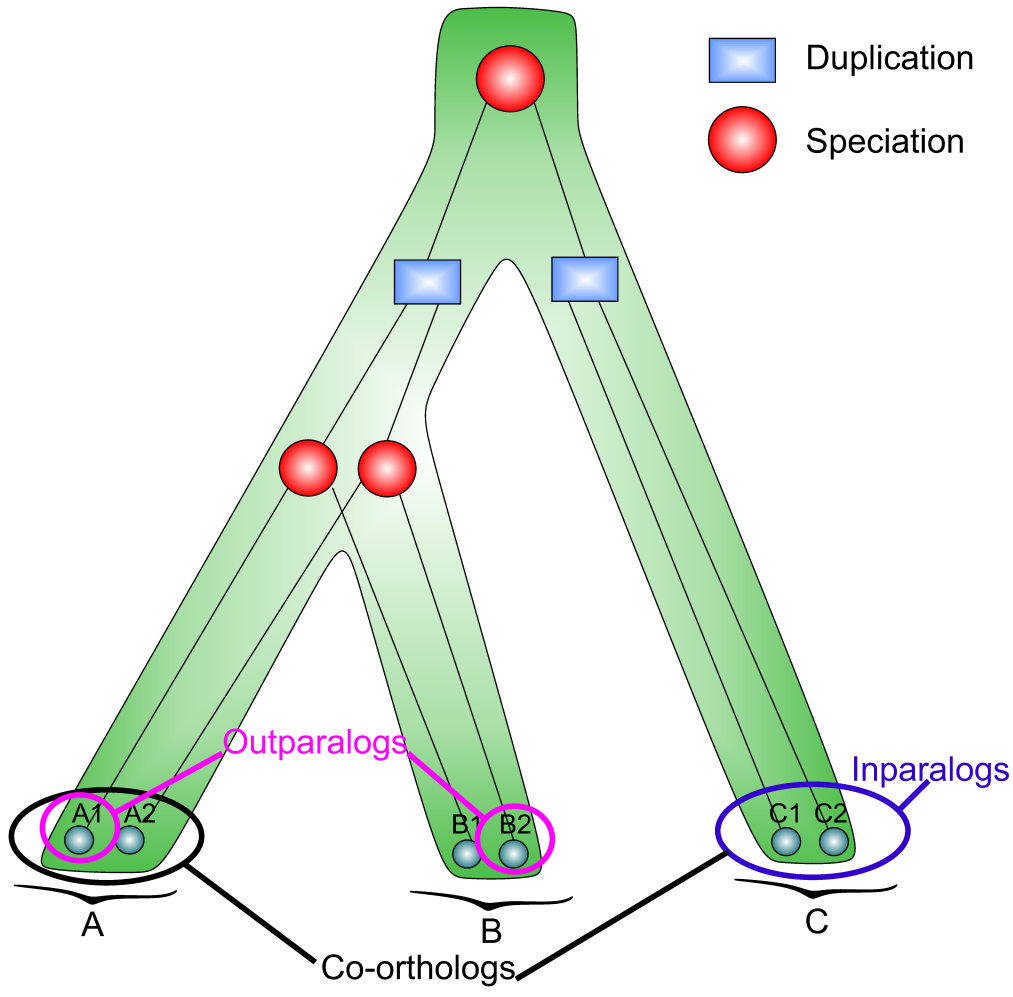
Homology, Orthology, Paralogy & Xenology

- Homologs: genes that descend from a common ancestral gene
- Orthologs: genes related via a speciation event
- Paralogs: genes originated via a duplication event
- Xenologs: genes related via a horizontal gene transfer event



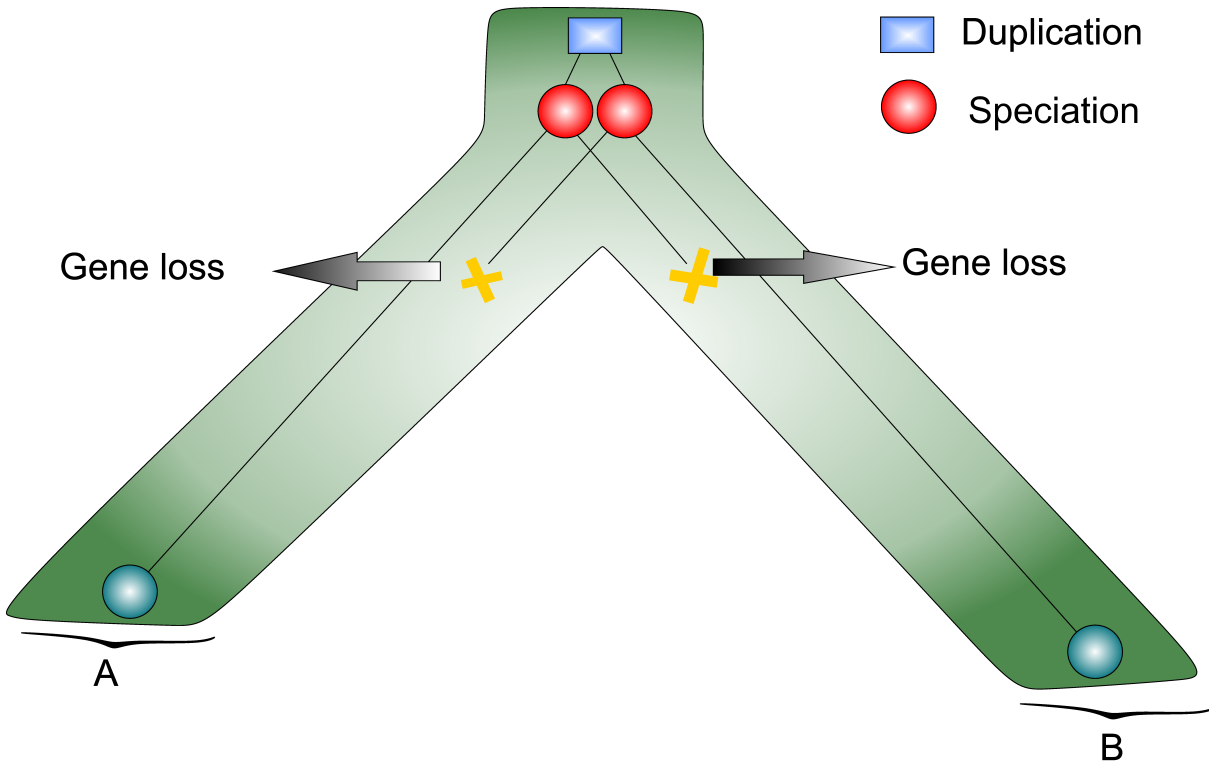
Effects of gene duplication

- In-paralogs: genes arising via duplication in a lineage after a speciation event
- Out-paralogs: genes arising via duplication before a speciation event
- Co-orthologs: a group of genes in one species that is orthologous to a gene or a group of genes in another species due to gene duplications.



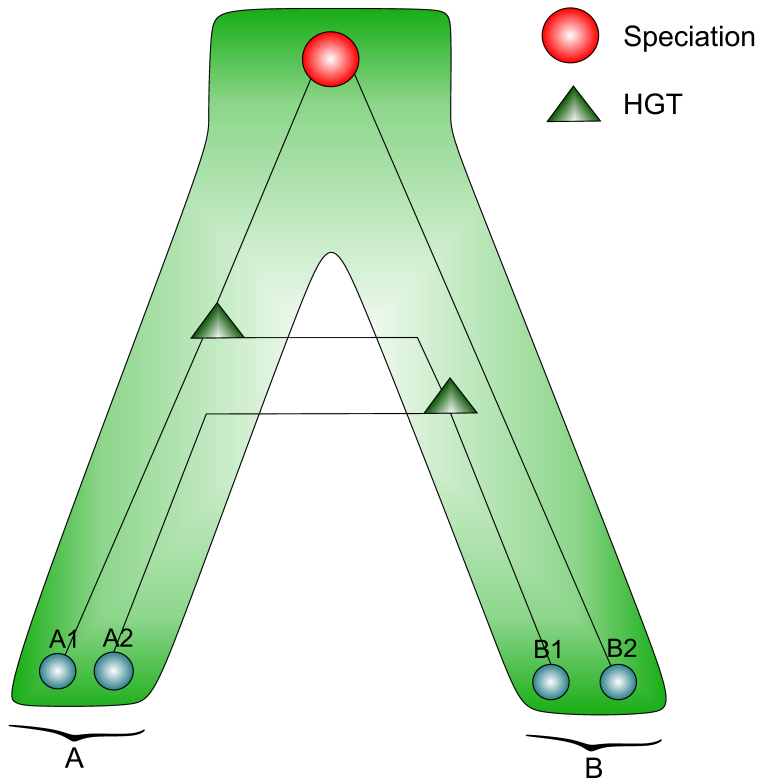
An effect of Gene loss

- Pseudo-orthologs: genes that are actually paralogs but appear to be orthologs due to gene loss.

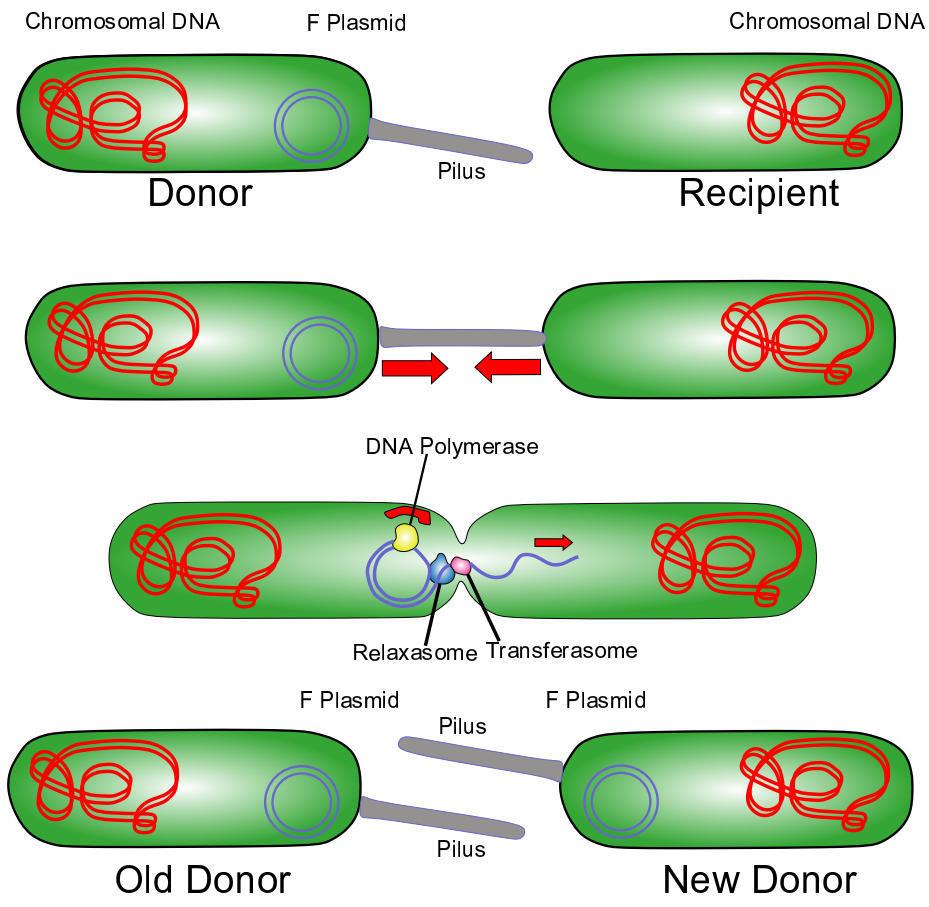


An effect of Horizontal Gene Transfer (HGT)

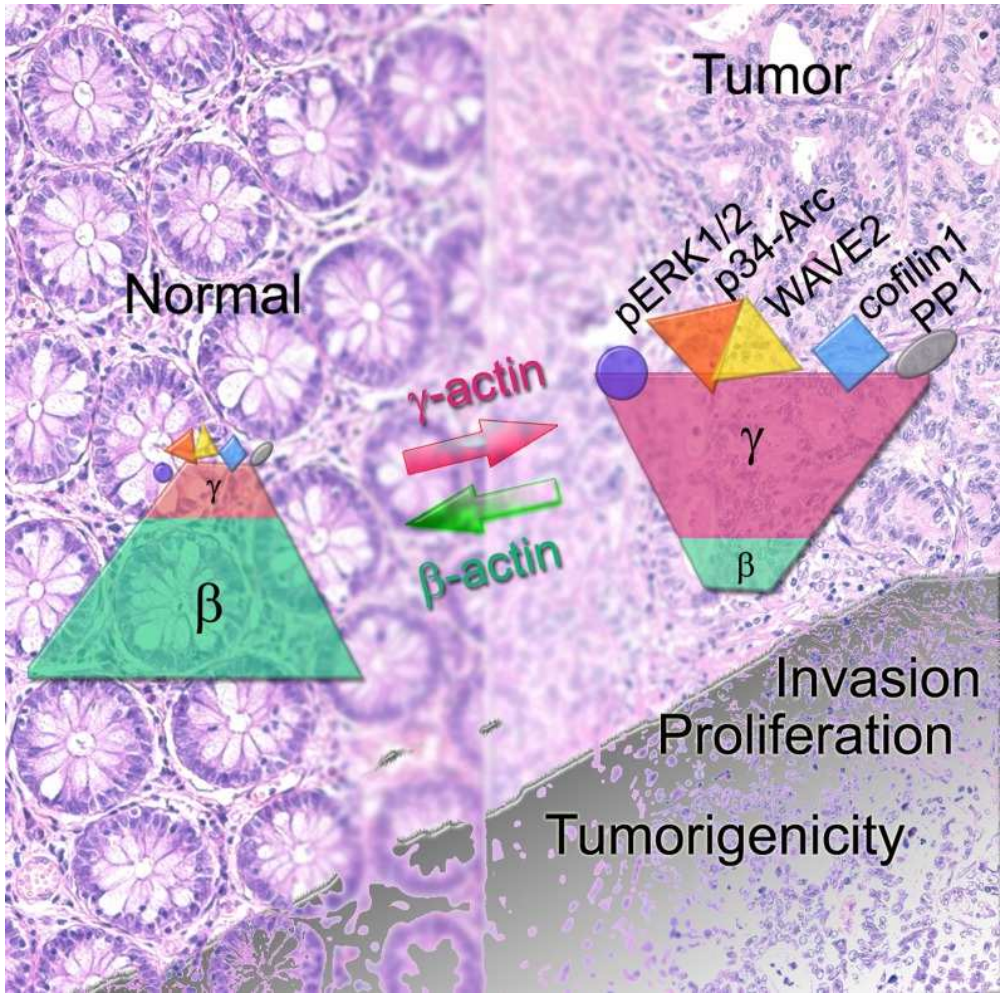
- Pseudo-paralogs: genes that are actually xenologs but appear to be paralogs due to consecutive HGTs.



HGT in antibiotic resistance



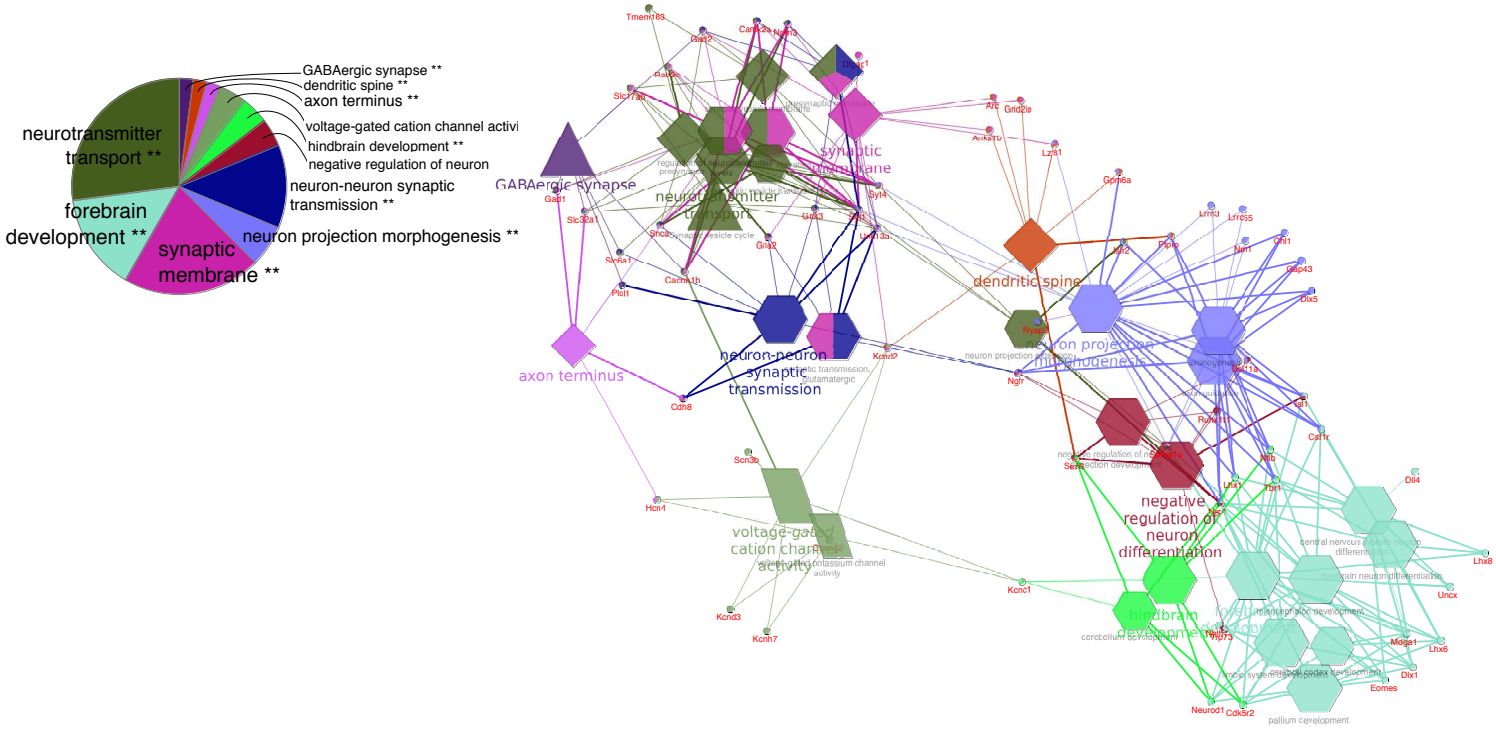
β and γ actin paralogs [1]



[1] Tumor promotion by γ and suppression by β non-muscle actin isoforms, Dugina et al.,Oncotarget., 2015

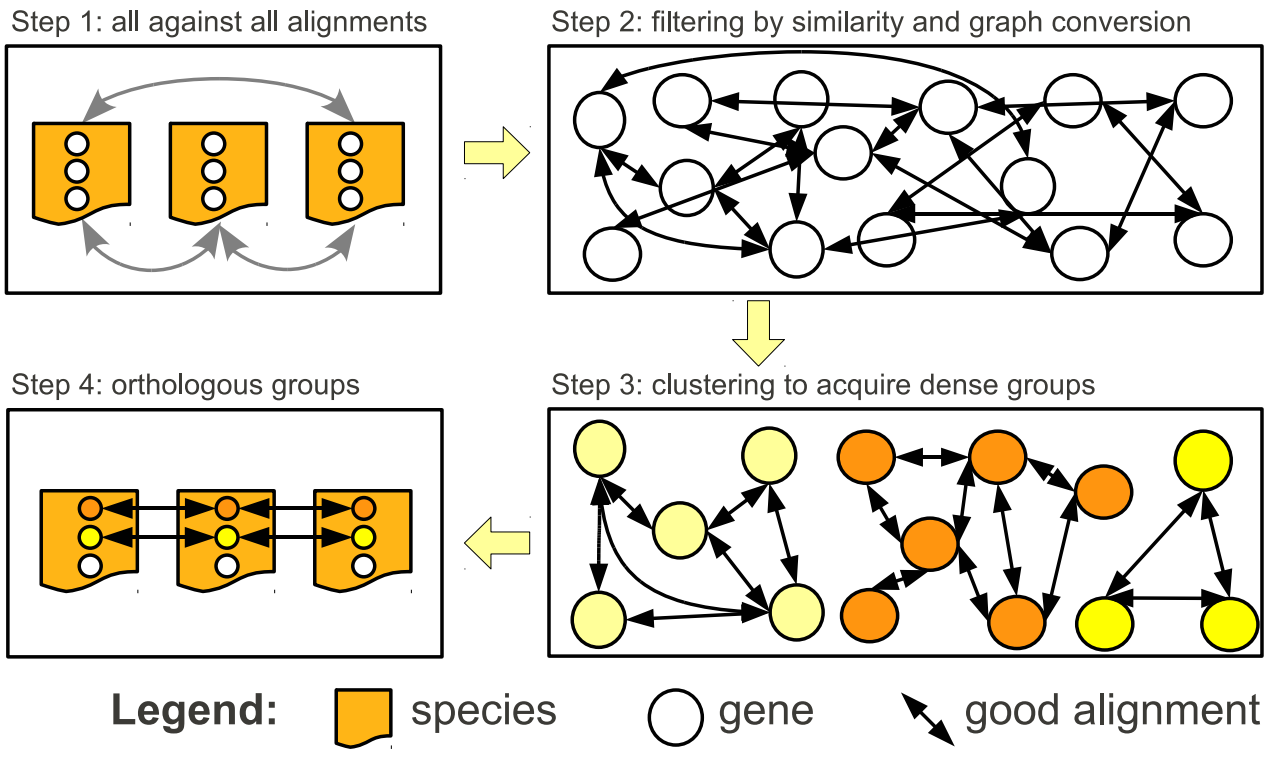
Ortholgs in chicken and mouse brain development

E



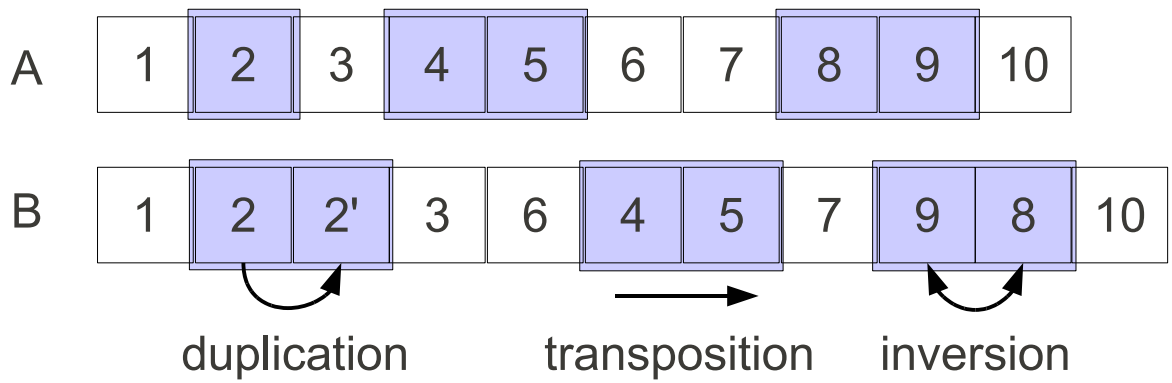
Estimating orthology: `Proteinortho`

- determines orthologous relationships based on
  - Sequence similarity: `BLAST`
  - Cluster density: spectral partitioning



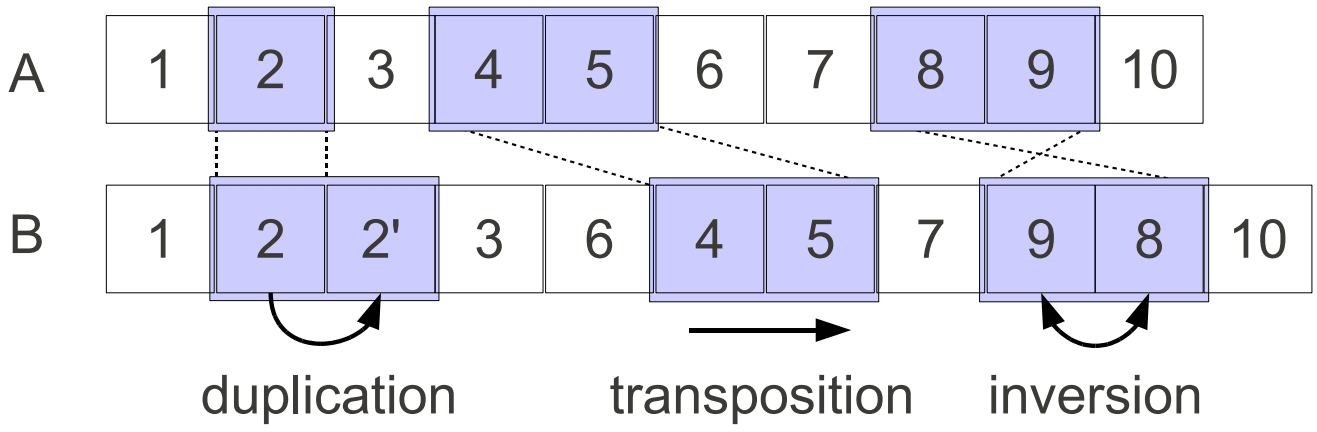
PoFF: Synteny extension [2]

- contextual conservation of gene loci
- passed down from a common ancestor
- indicates functional relationships of the genes
- identifies genomic rearrangements



[2] **Orthology Detection Combining Clustering and Synteny for Very Large Data Sets.**, Marcus Lechner, Maribel Hernandez-Rosales, et al. *PlosONE*, 2014

Synteny extension

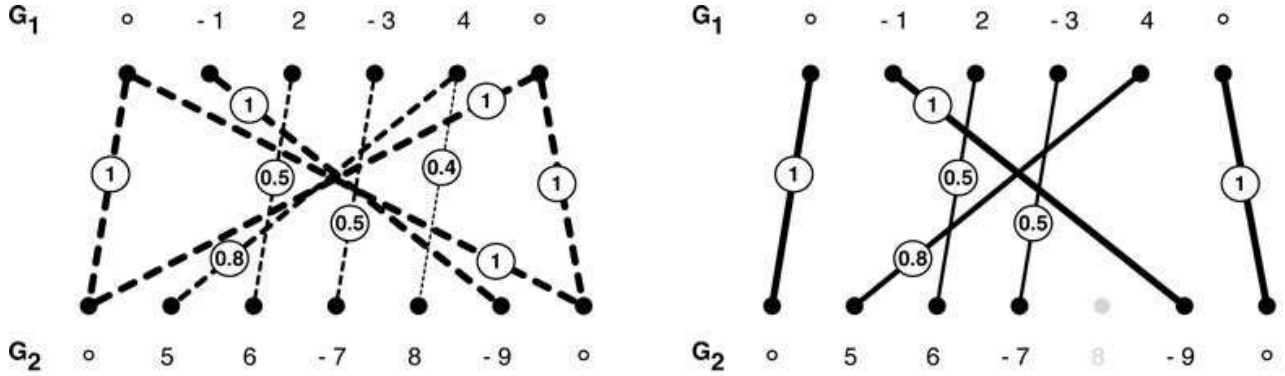


`Family-Free-Adjacencies` heuristic

- maximizes the balance between
  - the total number of edges
  - and the number of adjacencies
- prefers syntenically conserved genes

FF-Adjacencies

Gene family assignment-free comparative genomics



Find a *minimal* matching that maximizes

$$F(M) = adj(M) + edg(M)$$

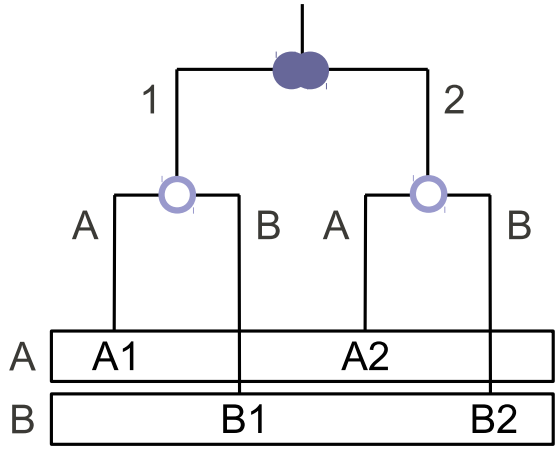
where

$$adj(M) = \sum Conserved\ adjacent\ edges\ in\ G_1\ and\ G_2$$

and

$$edg(M) = \sum Edge\ weights$$

Idea

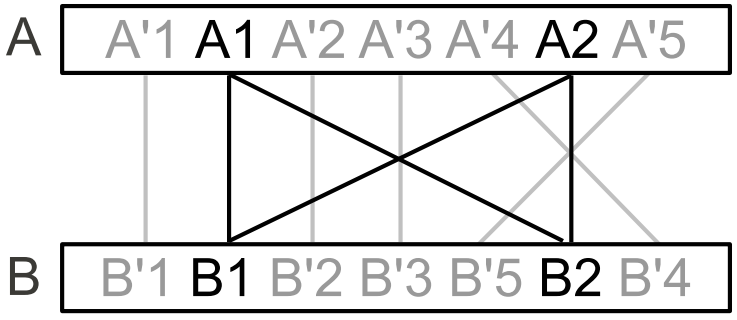


duplication

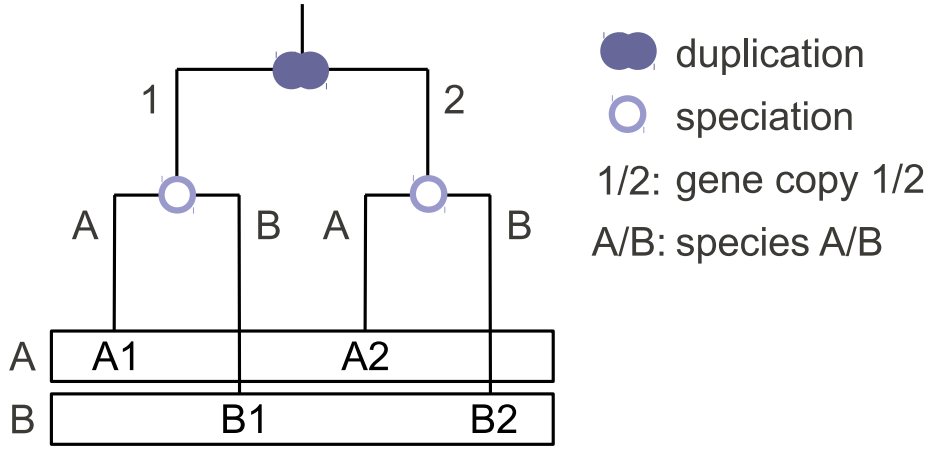
speciation

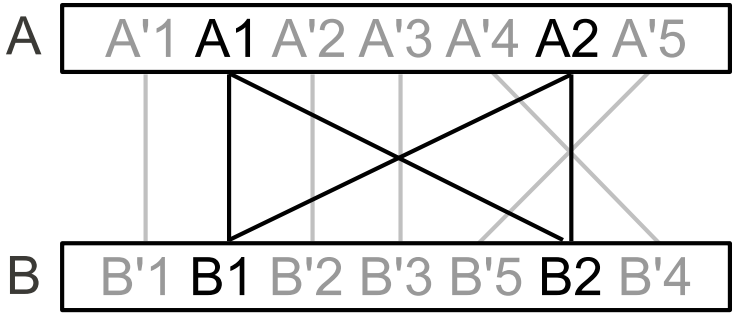
1/2: gene copy 1/2

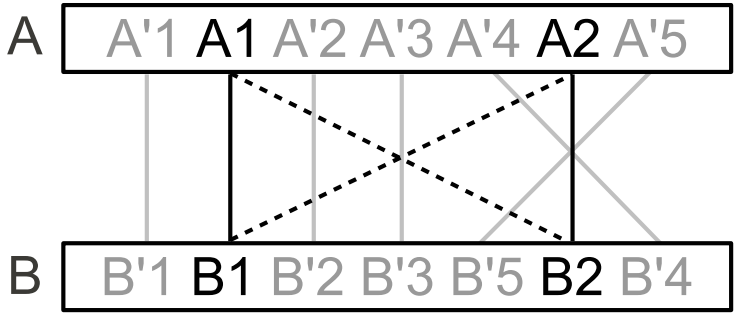
A/B: species A/B



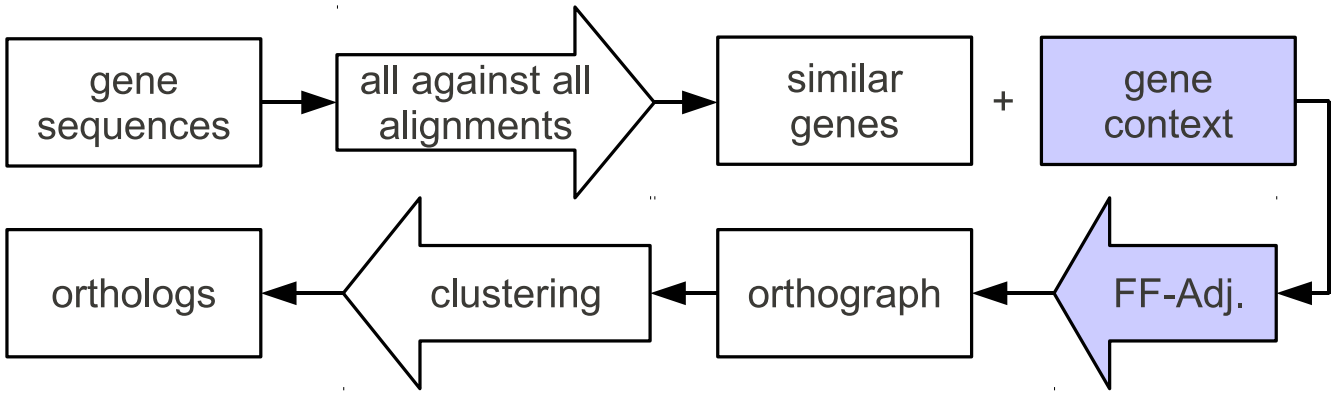
Idea


1
2
A
B
A
B
A
A1
A2
B
B1
B2
duplication
speciation
1/2: gene copy 1/2
A/B: species A/B


A
A'1 A1 A'2 A'3 A'4 A2 A'5
B
B'1 B1 B'2 B'3 B'5 B2 B'4


A
A'1 A1 A'2 A'3 A'4 A2 A'5
B
B'1 B1 B'2 B'3 B'5 B2 B'4

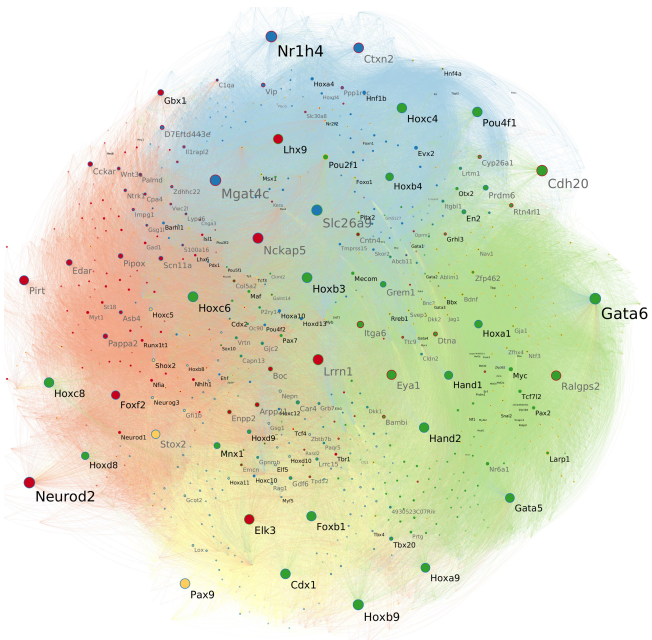
Synteny-enhanced workflow

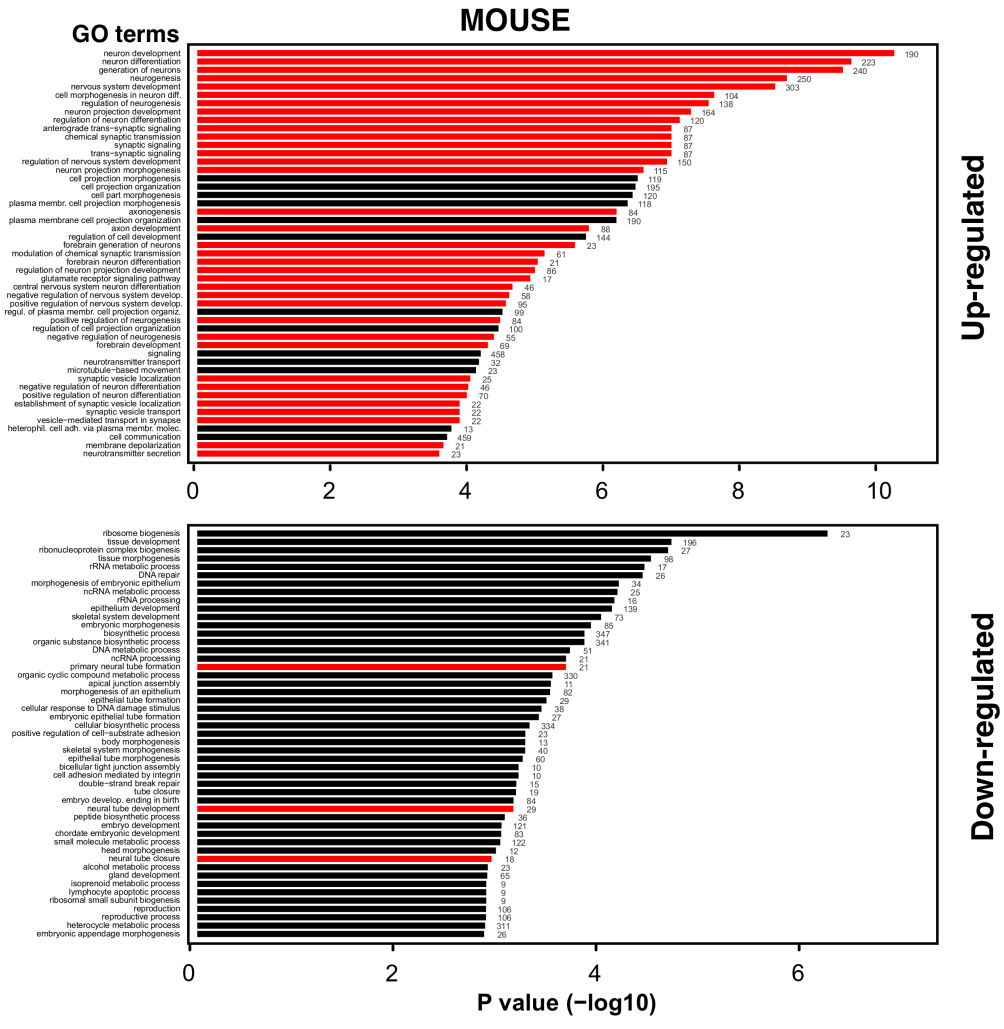


Options

- 1:1 orthologs: `FF-Adjacencies` is run once
- Noise reduction: `FF-Adjacencies` is iterated several times and can detect multiple copies of genomic areas

Application to brain development in mouse and chicken

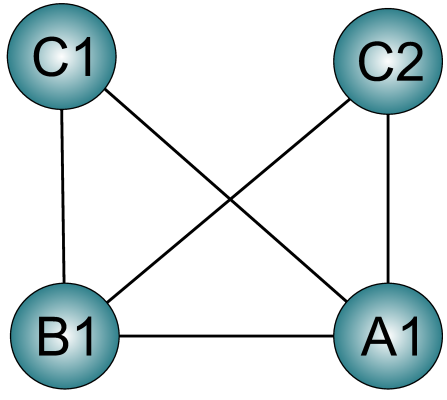


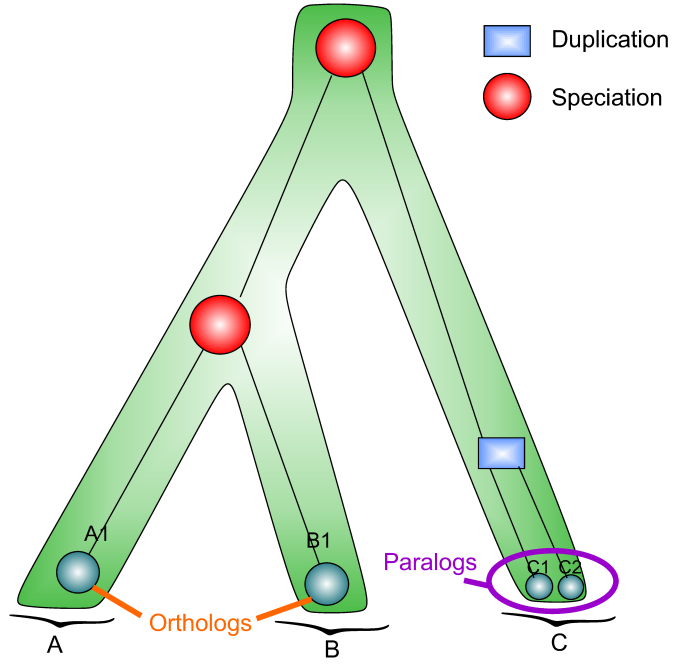


Orthologs and Paralogs

The orthology relation:

$\Theta = \{(A1, B1), (A1, C1), (A2, B1), (A2, C1), (B1, C1)\}$





Estimating orthology directly from the data

1. We know the assignment of genes to species.
2. We can estimate orthology directly from the data (e.g. `ProtheinOrtho`)

Aim: Correct the initial estimate to the "closest" orthology relation that fits the data.

$\Longrightarrow$ We need a characterization of an "ideal" orthology relation.

The question is: For a given orthology relation when does there exist a (gene) tree T with

- $\mathit{Event}(\mathrm{lca}(x,y)) = \bullet = \mathit{speciation}$ for all $(x,y) \in \Theta$ and
- $\mathit{Event}(\mathrm{lca}(x,y)) = \blacksquare = \mathit{duplication}$ for all $(x,y) \notin \Theta$?

Result[3]

If we represent every pair $(x, y) \in \Theta$ as an edge of a graph G, then we have the following result:

- $G_{\bullet} = G_{\ominus}$ **is a Cograph.**

Theorem (Corneil et al, 1981)

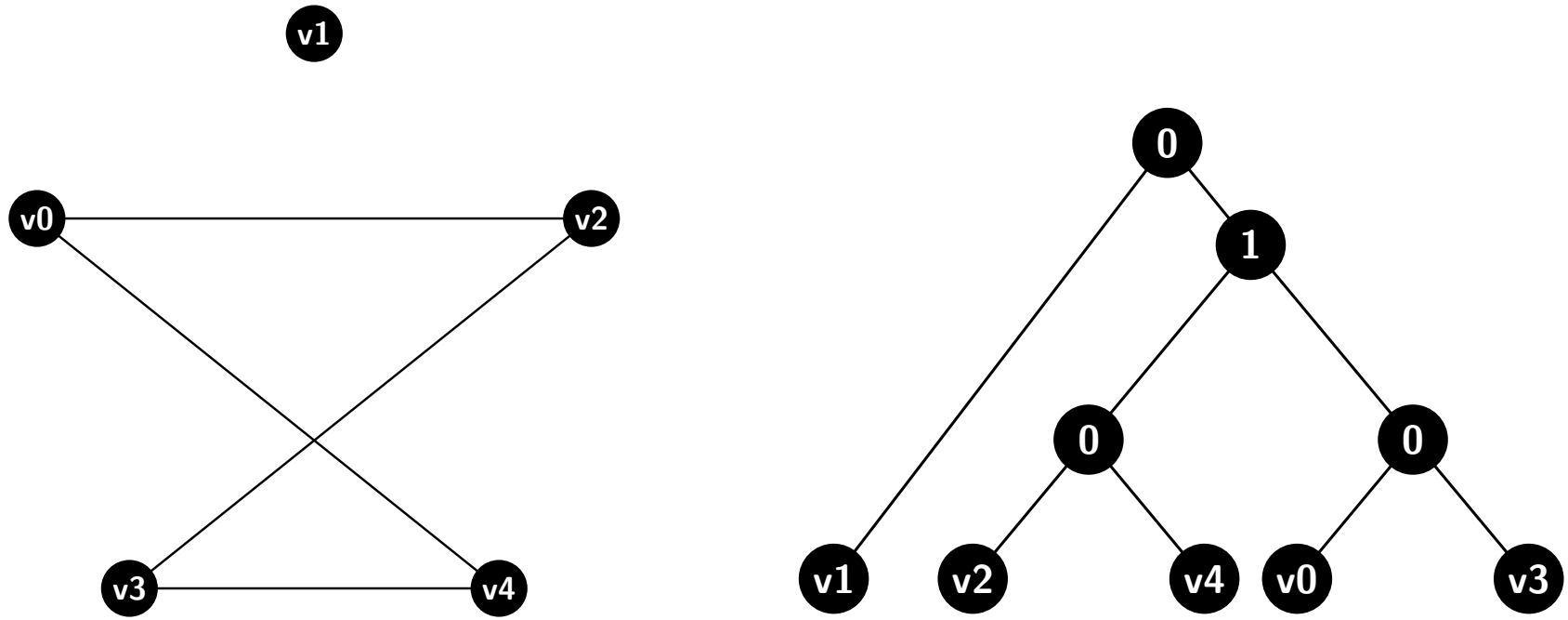
G is a Cograph if and only if G does not contain induced P_4's.

Forbidden P_4:

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

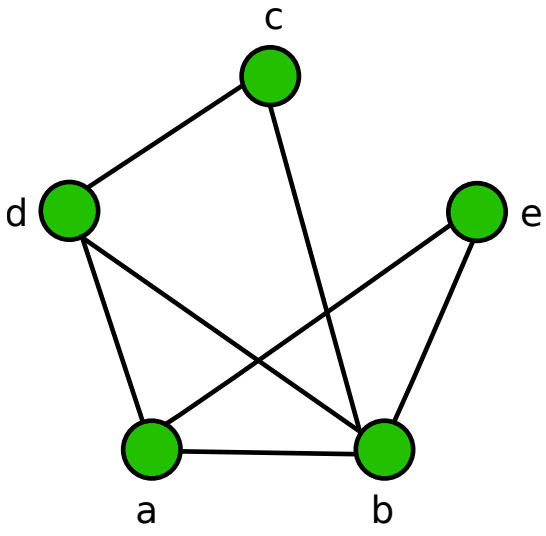
Cotree

A property of a cograph is that it is associated to a unique tree representation: **a Cotree**.

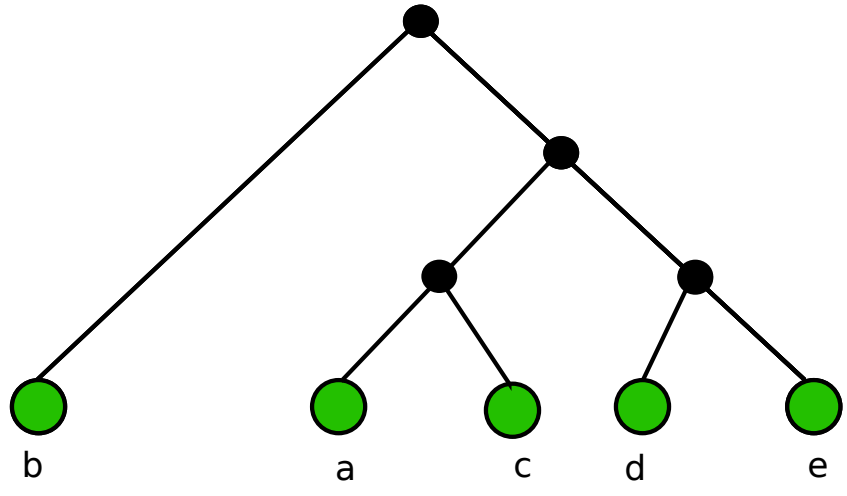


$(x, y) \in E(G)$ if and only if $\mathrm{lca}(x, y) = 1$

Example: a non-valid orthology relation



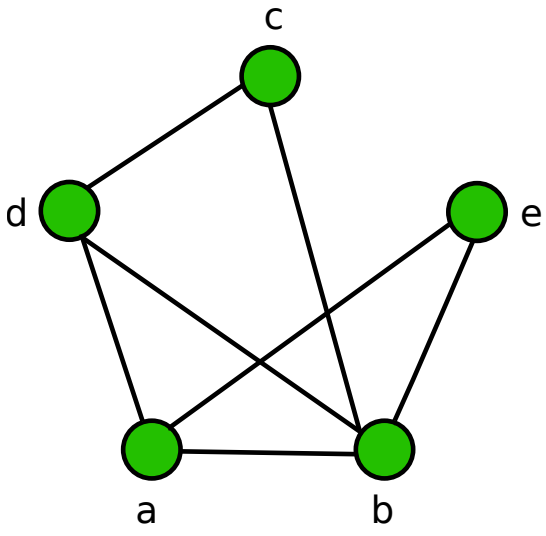
(a)



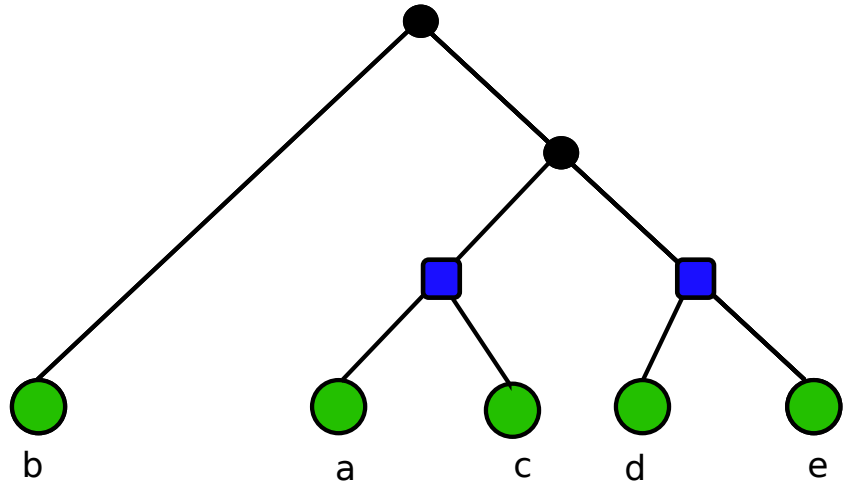
(b)

- (a) a graph representing the orthology relationships
- (b) the predicted gene tree

Example



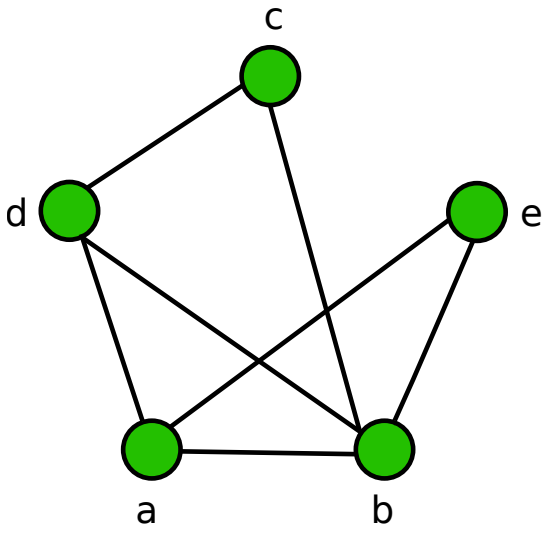
(a)



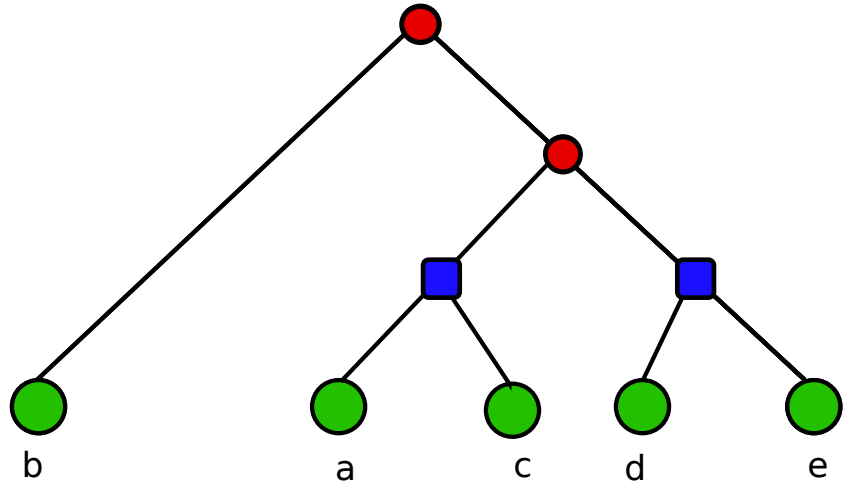
(b)

- (a) a graph representing the orthology relationships
- (b) the predicted gene tree

Example



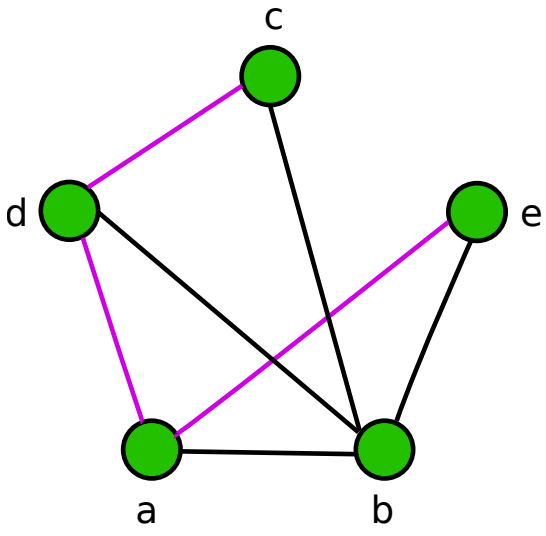
(a)



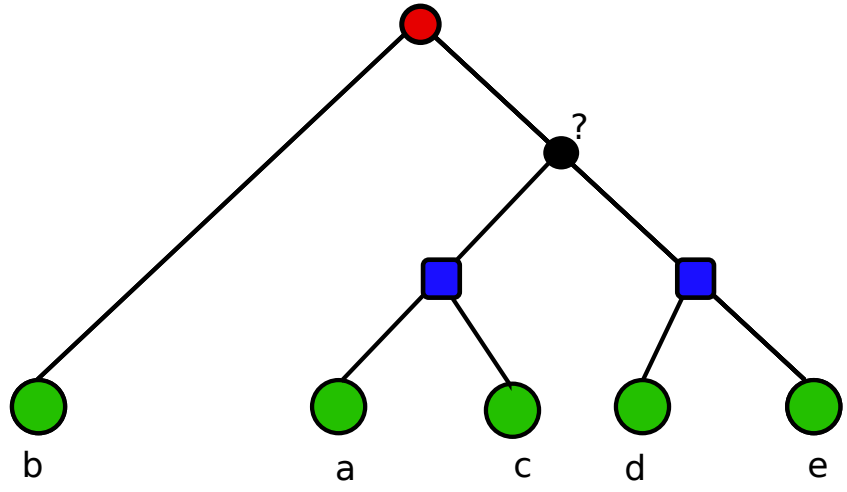
(b)

- (a) a graph representing the orthology relationships
- (b) the predicted gene tree

Example



(a)

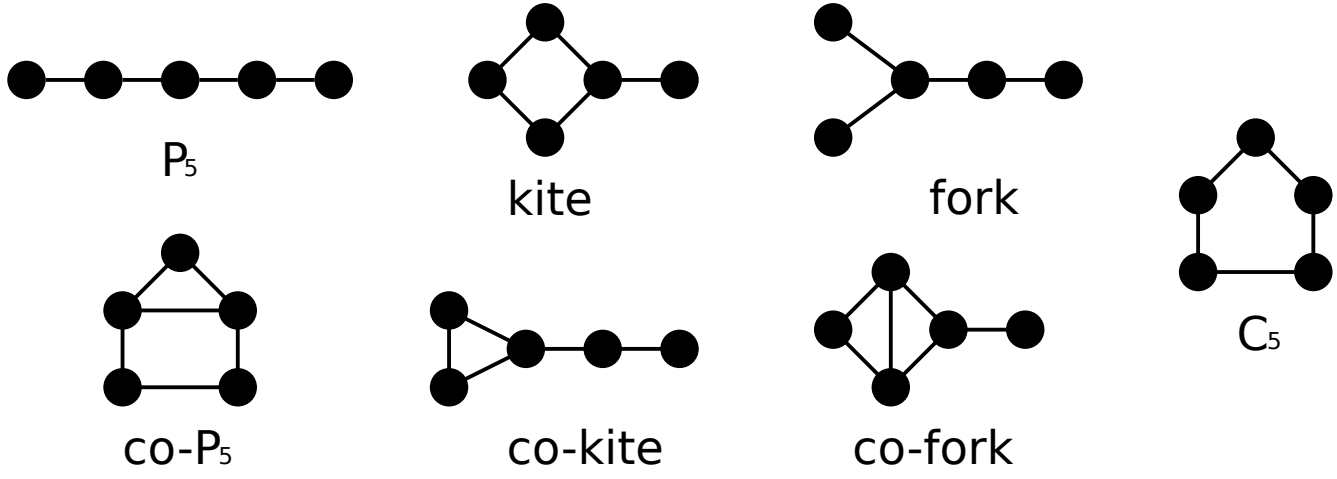


(b)

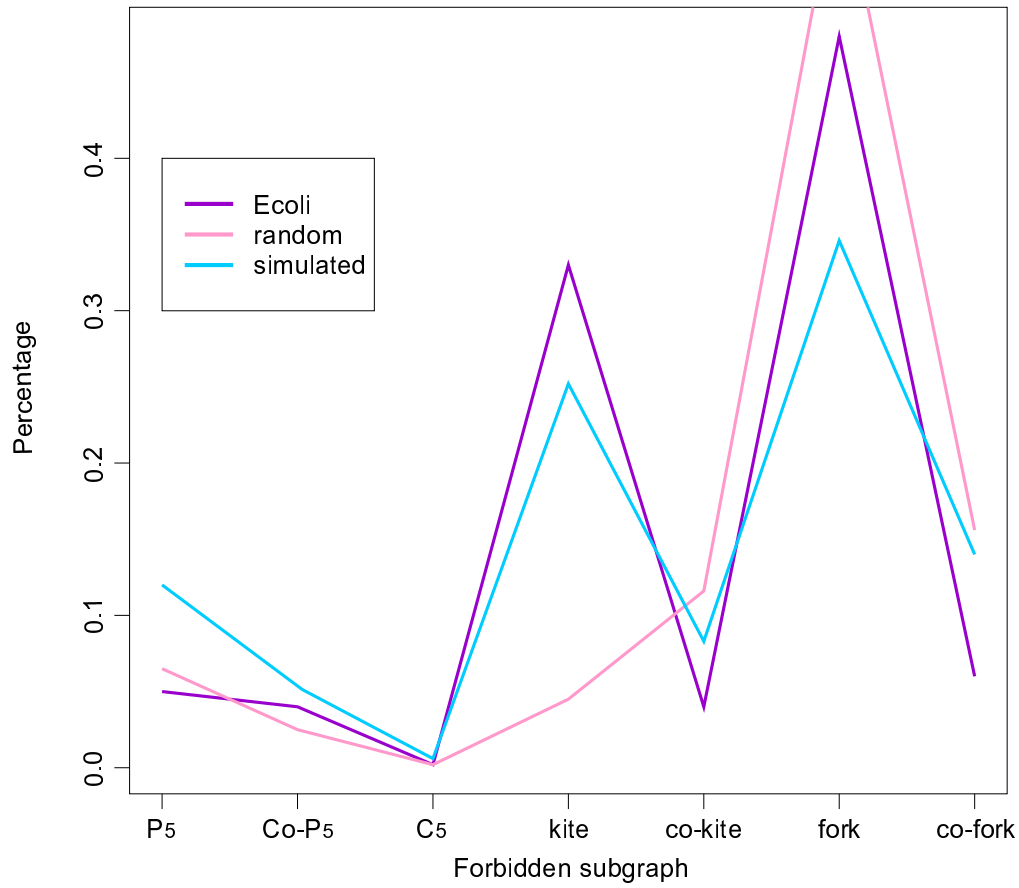
- (a) a graph representing the orthology relationships
- (b) the predicted gene tree

The cograph editing problem (another presentation!

- Consists of a method that converts any given graph to a cograph with a minimum edge modification;
- It has been proved to be NP-complete!
- Objective: present a heuristic approach to solve the cograph editing problem.

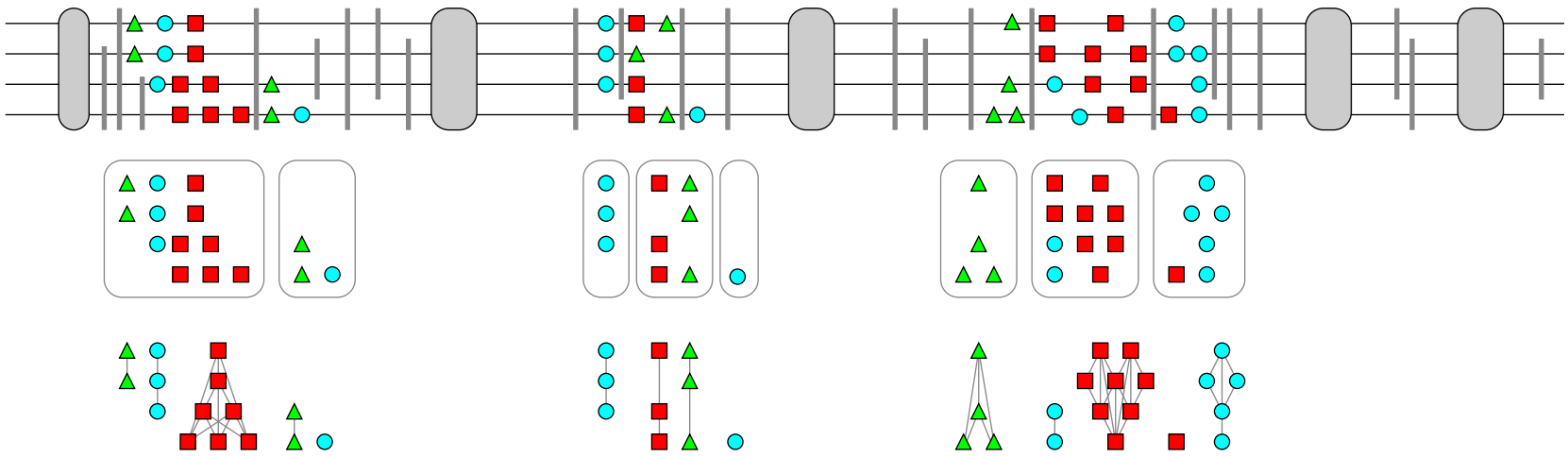


Application to the *E.coli* Pangenome



A “noise preference” gave us evidence of gene fission and fusion.

Application: indirect orthology prediction of tRNAs in Primates [4]

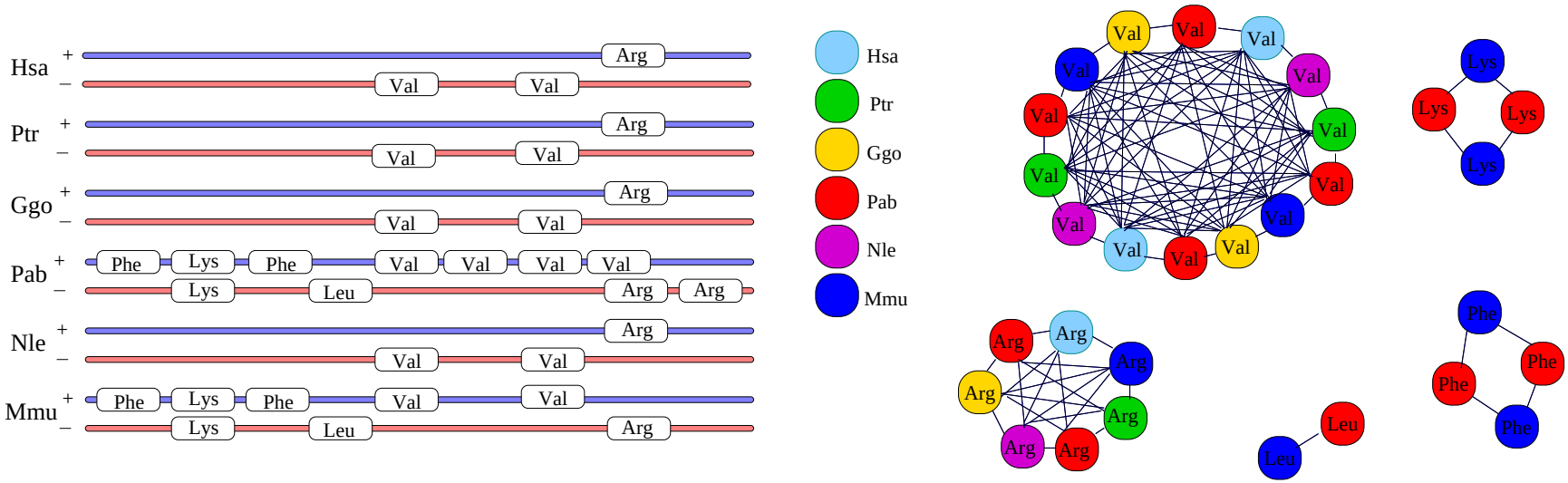


Genomic organization of tRNAs.

---

[4] **Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies.**, Cristian A Velandia-Huerto, Sarah J Berkemer, Anne Hoffmann, Nancy Retzlaff, Liliana C Romero Marroquin, Maribel Hernández Rosales, Peter F Stadler,C lara I. Bermudez-Santana, *BMC Genomics*, 2016.

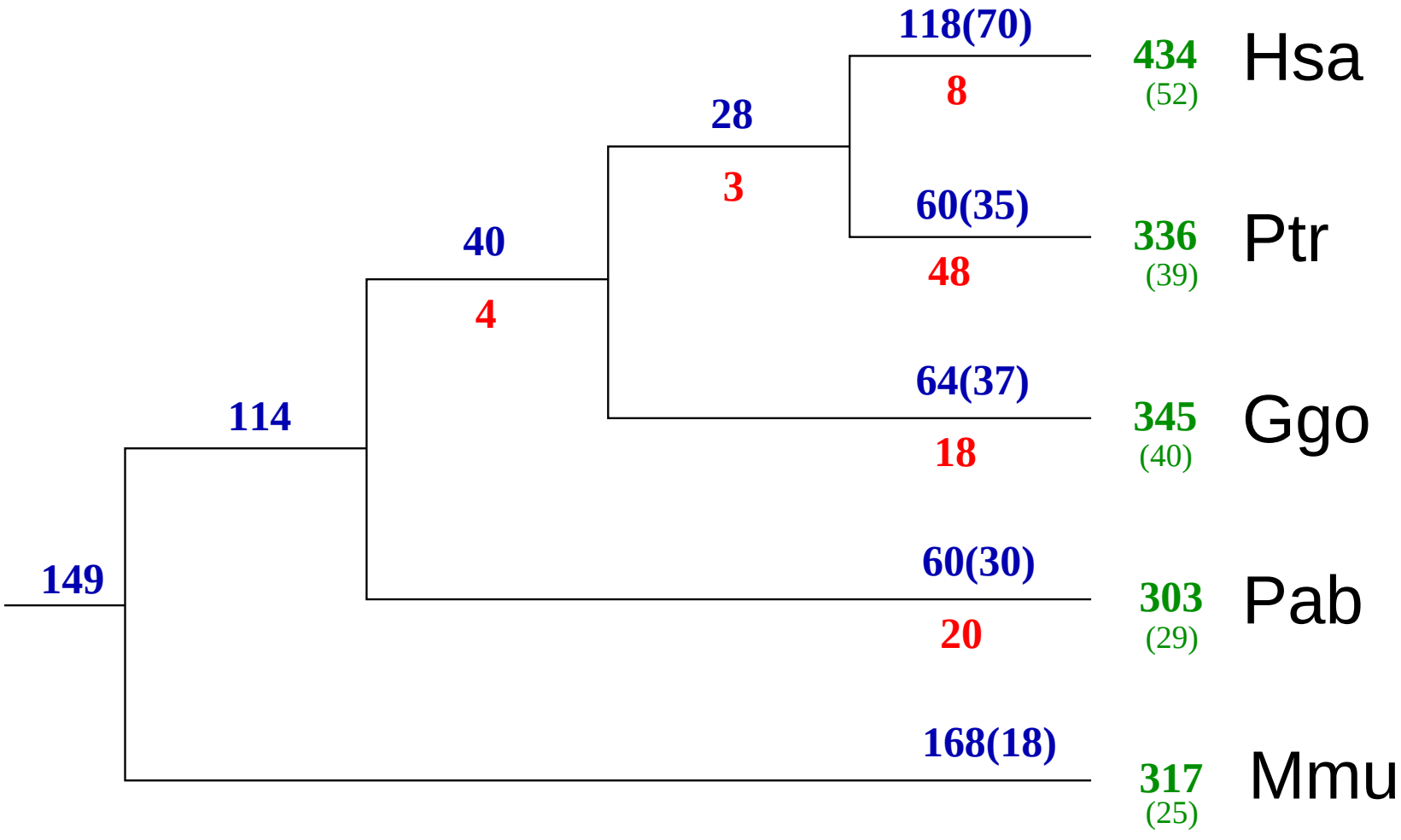
Identification of tRNA orthologs



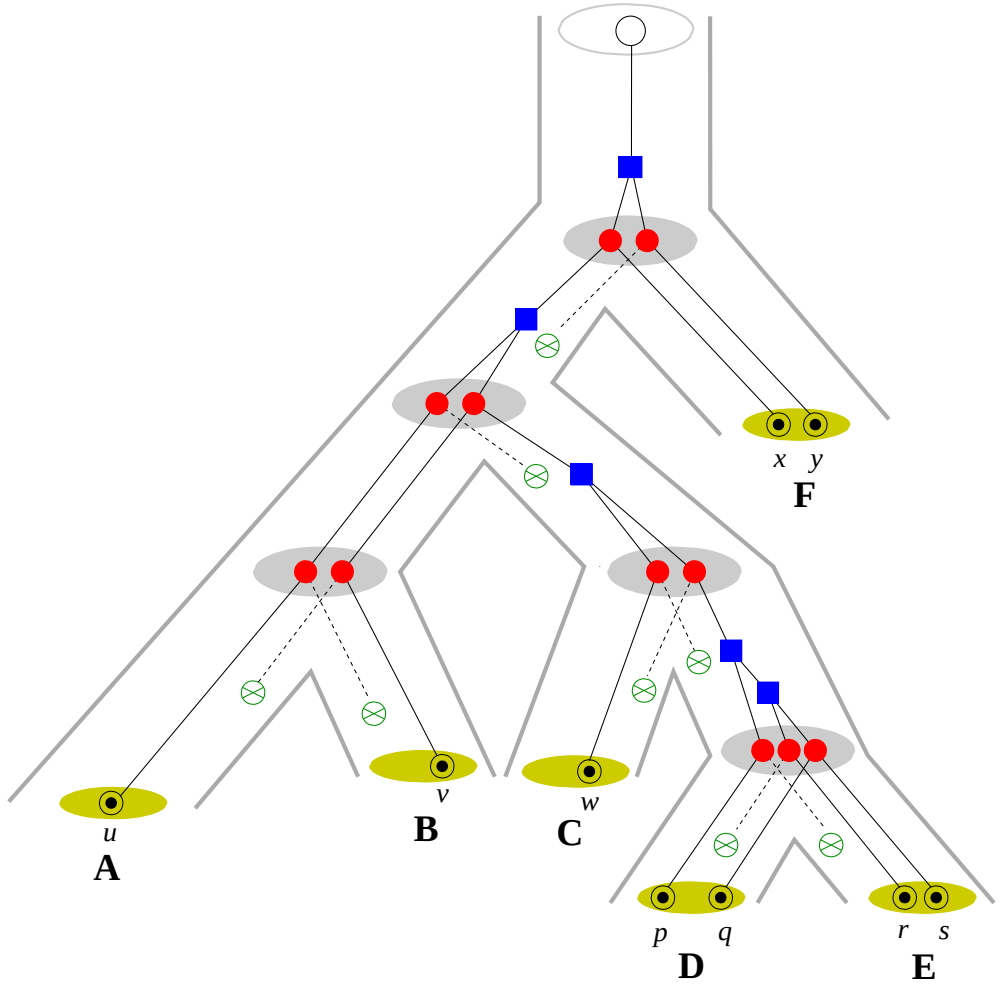
Graphs corresponding to tRNA clusters. Edges indicate that the tRNA sequences are

sufficiently similar to be possible orthologs. Different species are distinguished by colors. The tRNAs isoacceptor classes are indicated by their 1-letter codes: Phe (F), Lys (K), Leu (L), Val (V), Arg (R).

Gain, loss and duplications of tRNAs in primates



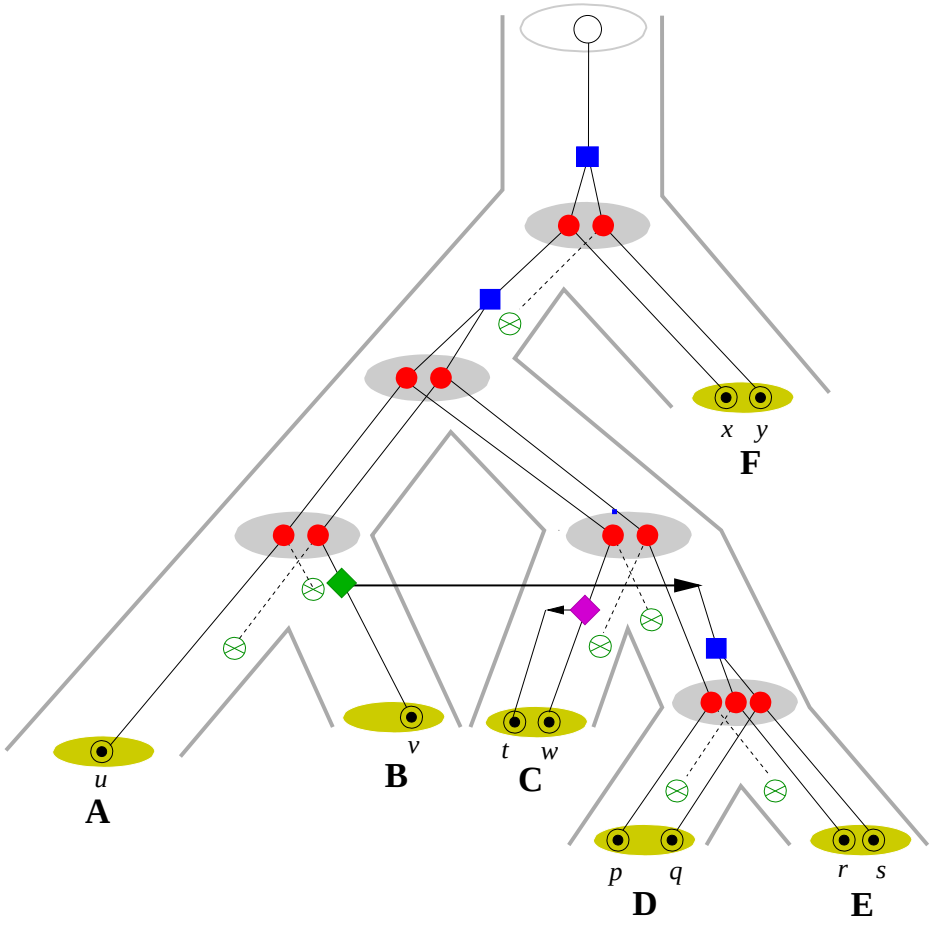
Reconciliation Tree

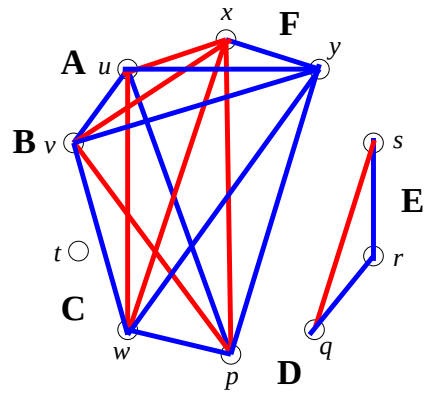


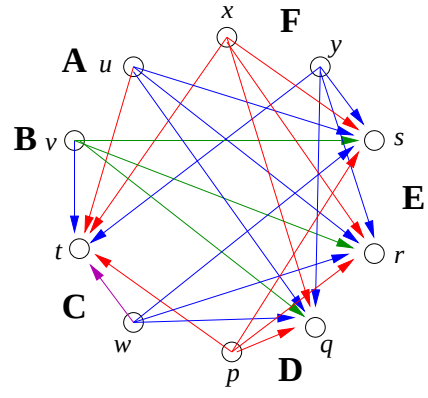
Conclusions

- Accurate annotation of orthologs and paralogs are esencial for functional genomics.
- Details on the evolution of very large gene families are helpful for better annotation.
- Fast and efficient algoritms for large-scale genome data are needed.
- Mathematical characterizations of valid evolutionary relations are important for noise detection in real data.

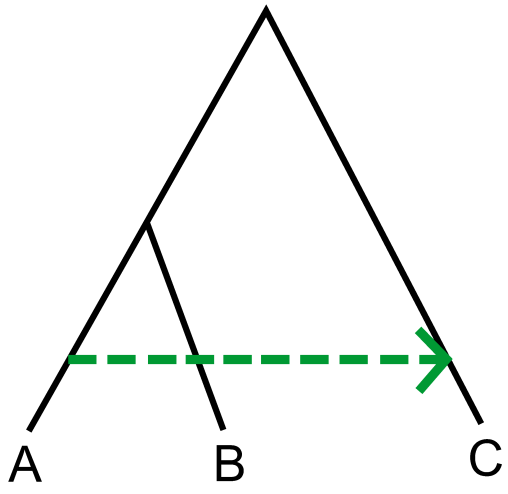
Directed Cographs: work in process

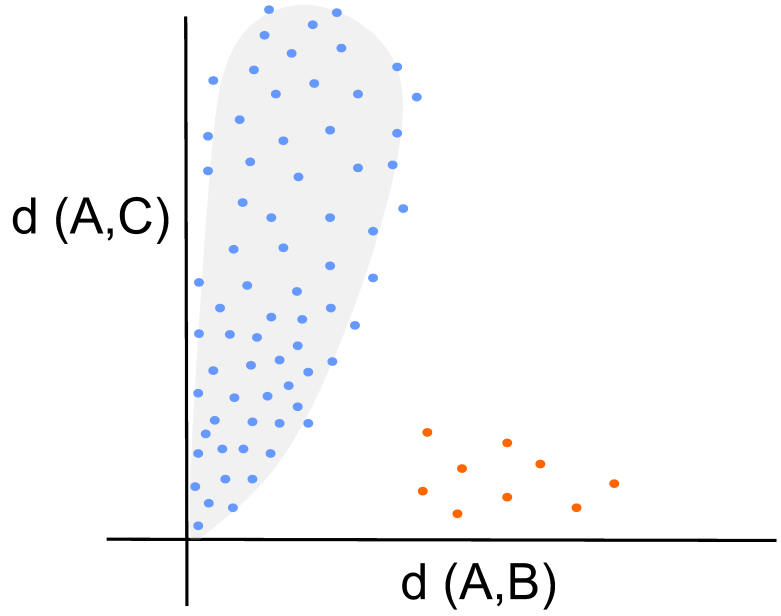






Statistical method for HGT identification





Thanks!

Special thanks to:

- Markus Lechner, University of Marburg, Germany
- Marc Hellmuth, University of Greifswald, Germany
- Nicolas Wieseke, Peter Stadler, University of Leipzig, Germany
- Gabriel Moreno, University of Wilfried Laurier, Canada
- Mia Walter, University of Brasilia, Brazil
- Alfredo Varela, INB, UNAM Juriquilla
- Students of the Bioinformatics Group at UNAM Juriquilla
- DAAD, Germany
- Conacyt, Mexico