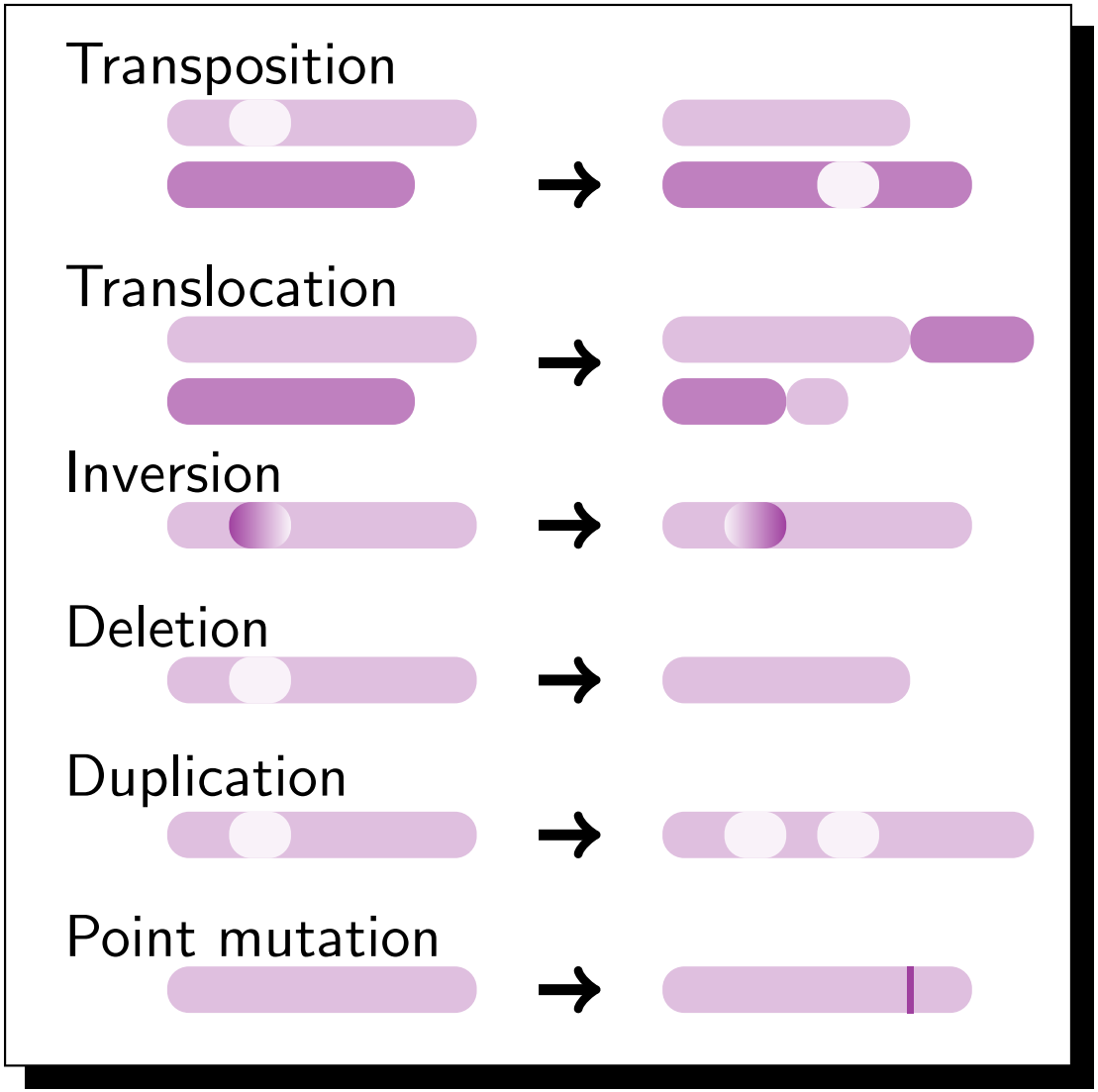
# Gene Family Assignment-Free Comparative Genomics

Daniel Doerr  Annelyse Thévenin  Jens Stoye

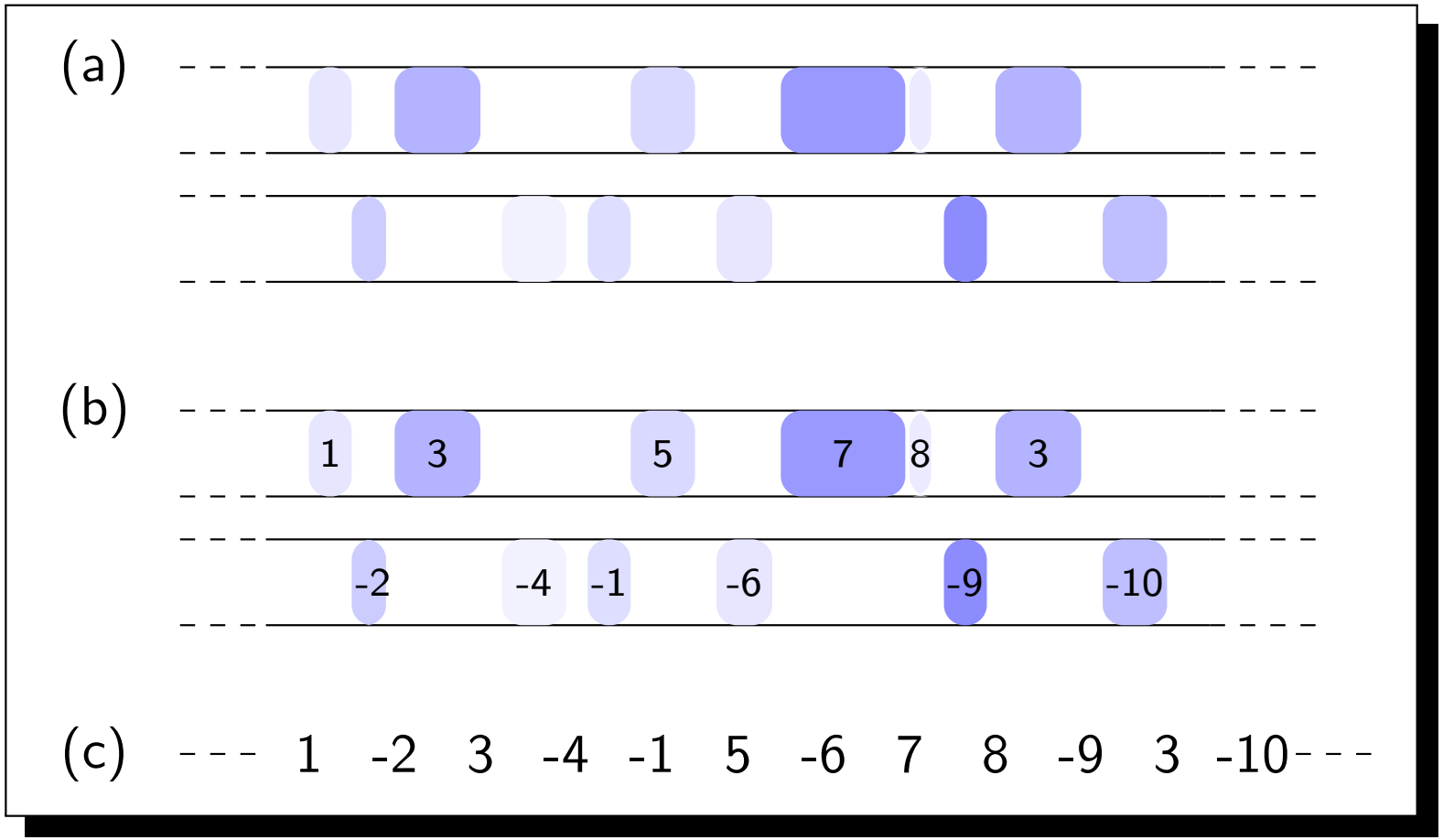
Genome Informatics, Faculty of Technology and
Institute for Bioinformatics, Center for Biotechnology (CeBiTec), Bielefeld University, Germany


# Similarities and differences between species

# Mutations

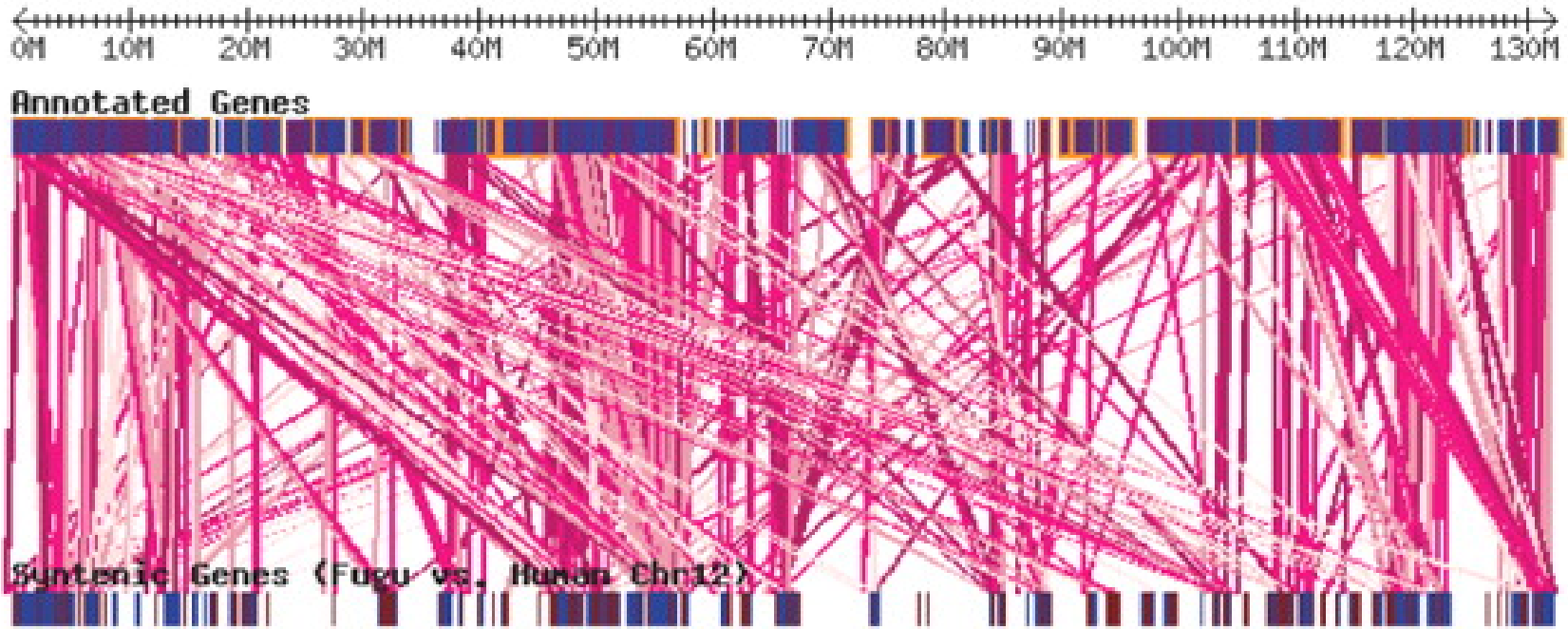

Transposition
Translocation
Inversion
Deletion
Duplication
Point mutation

# Representation of a genome

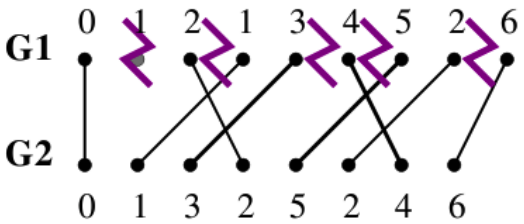


# Comparative genomics

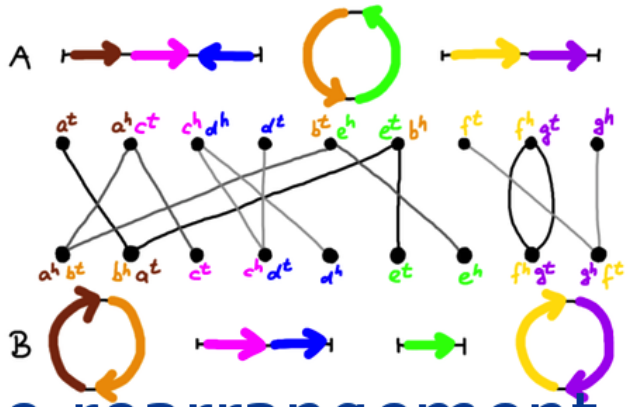
The order of genes is relevant.



# Applications in Comparative Genomics



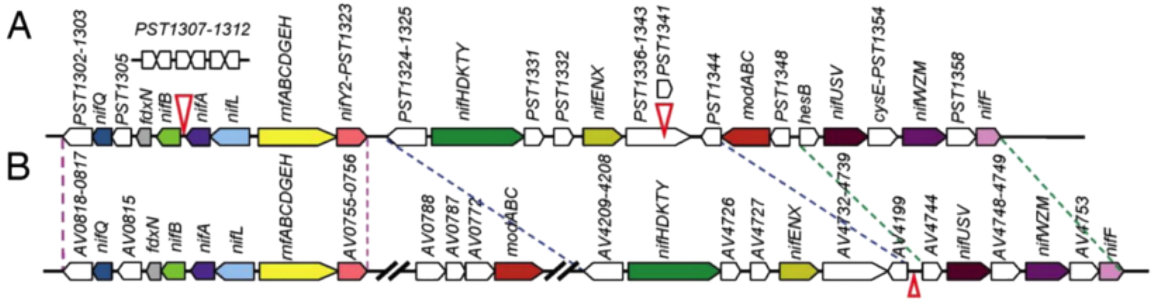

Gene order




Gene rearrangement

## Comparative Genomics

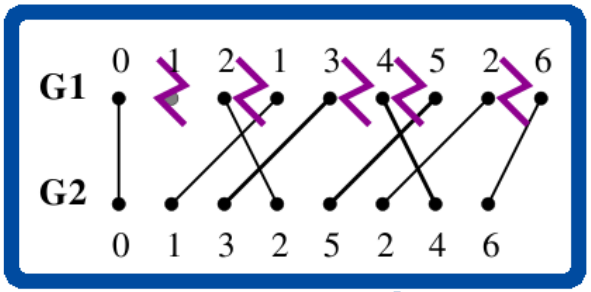
Gene cluster



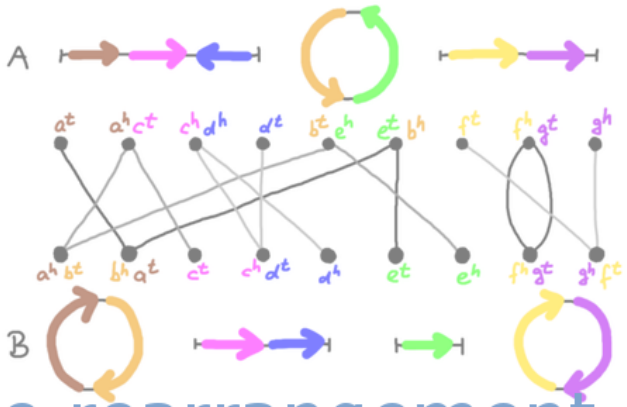

... and

many more

# Comparing gene order of two genomes




Gene order




Gene rearrangement

## Comparative Genomics

Gene cluster


... and many more

# Model 1: No duplication

## Two genomes without duplication

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $G_1$ | $+0$ | $+1$ | $+2$ | $+3$ | $+4$ | $+5$ | $+6$ | $+7$ | $+8$ | $+9$ |
| $G_2$ | $+0$ | $+7$ | $+3$ | $-5$ | $-4$ | $+6$ | $+1$ | $+2$ | $-8$ | $+9$ |

A chromosome is a signed permutation.

# Number of adjacencies

## A measure of similarity

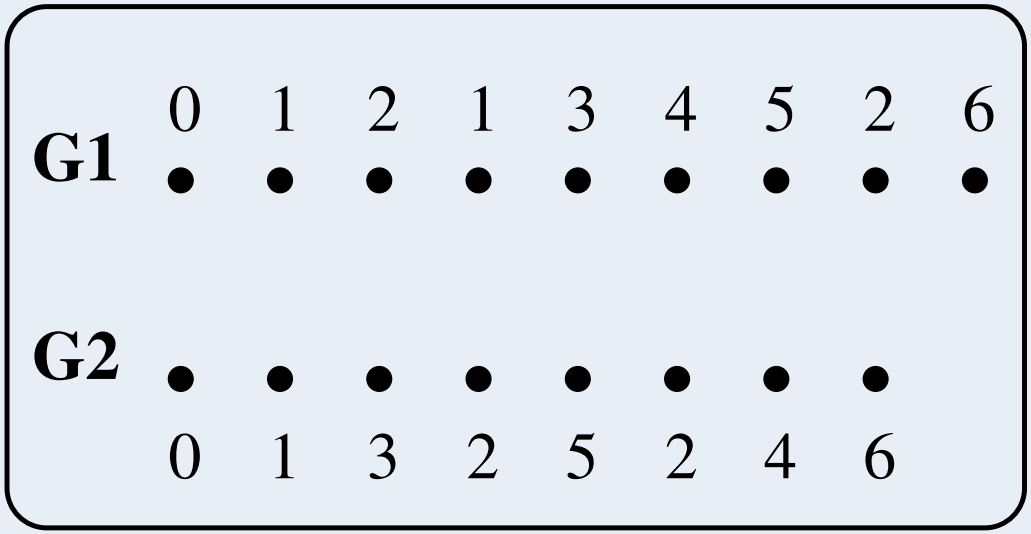
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $G_1$ | $+0$ | $+1$ | $+2$ | $+3$ | $+4$ | $+5$ | $+6$ | $+7$ | $+8$ | $+9$ |
| $G_2$ | $+0$ | $+7$ | $+3$ | $-5$ | $-4$ | $+6$ | $+1$ | $+2$ | $-8$ | $+9$ |

$\Rightarrow$ 2 adjacencies between $G_1$ and $G_2$.

# Model 2: With duplications

## Two genomes with duplications

A chromosome is a sequence over a set of signed characters.

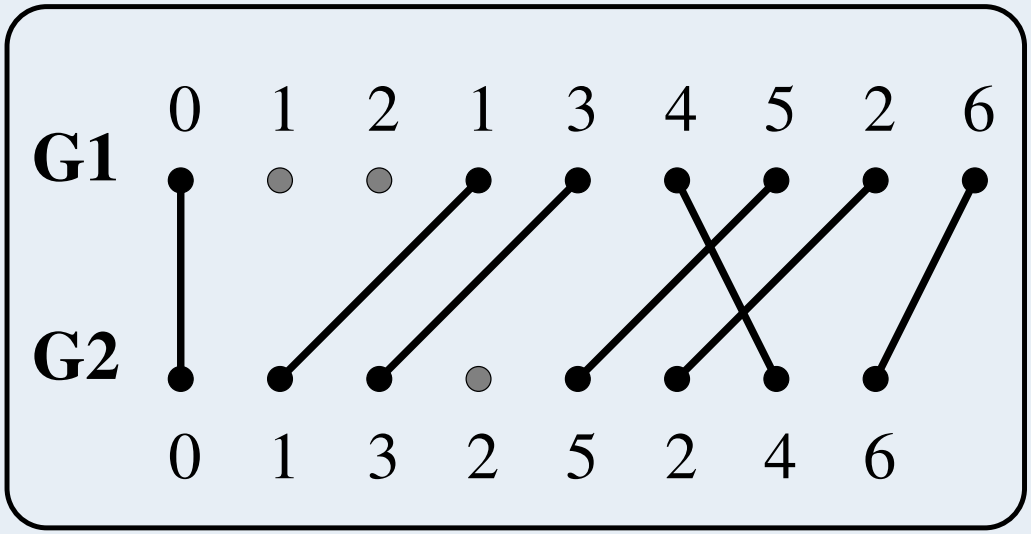



We need to find a matching between both genomes.

# Model 2: With duplications

## Two genomes with duplications

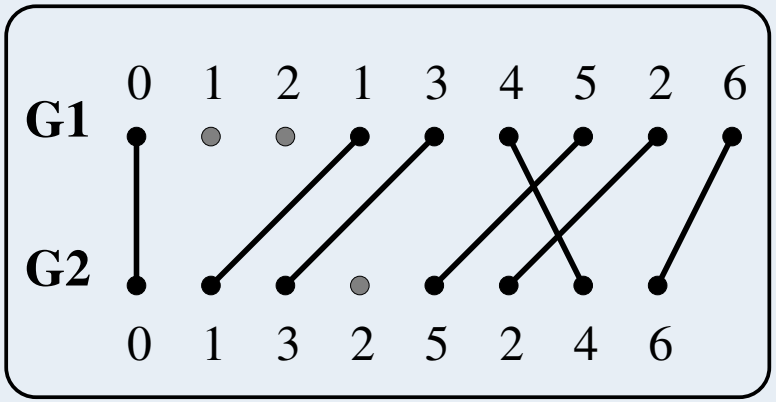
A chromosome is a sequence over a set of signed characters.

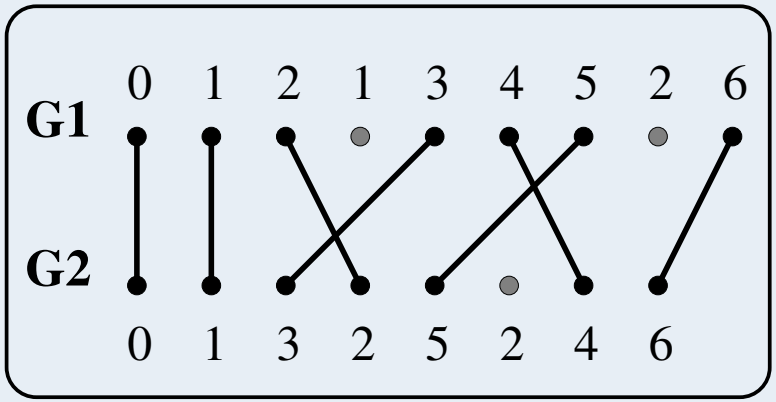



We need to find a matching between both genomes.

# Number of adjacencies varies in matchings

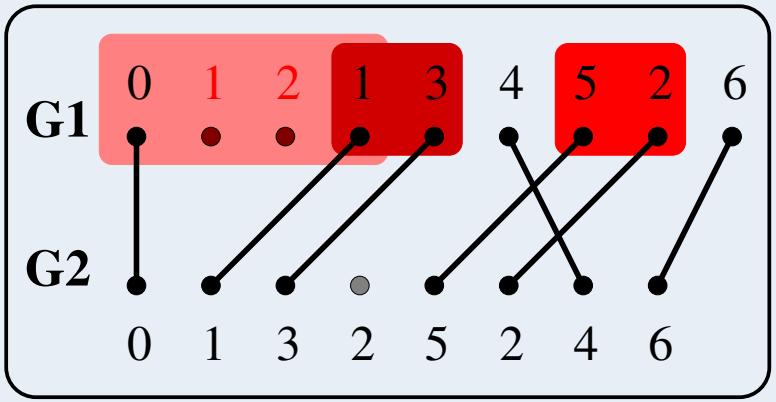
Example: Number of adjacencies in two different matchings:



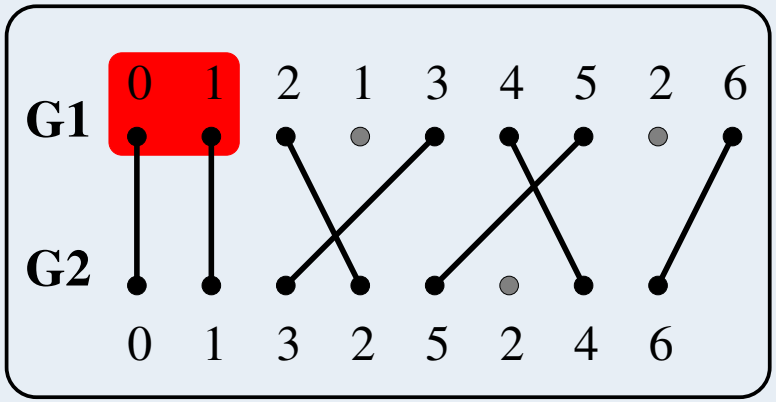



# Number of adjacencies varies in matchings

Example: Number of adjacencies in two different matchings:




*3 adjacencies*




*1 adjacency*

# Previous results

Maximizing the number of adjacencies between two genomes with duplication is a **NP-hard** problem.

There exists **exact** (and realistic) **programs and heuristics** to resolve it. [Angibaud *et al.*, 2008].

# Gene family consequences

## Pros

+ Subsequent analyses produce **reasonable results**.
+ Facilitates **simple** but **powerful** datastructure.
+ Gene family information: Many **databases** and tools available.

## Cons

– Wrong gene family assignments produce **incorrect results** in subsequent analyses.
– Datastructure: Strong and weak homology assumptions are **indifferent**.
– Gene family concept **not applicable** for all biological phenomena.

# Gene family consequences

## Pros

+ Subsequent analyses produce **reasonable results**.
+ Facilitates **simple** but **powerful** datastructure.
+ Gene family information: Many **databases** and tools available.

## Cons

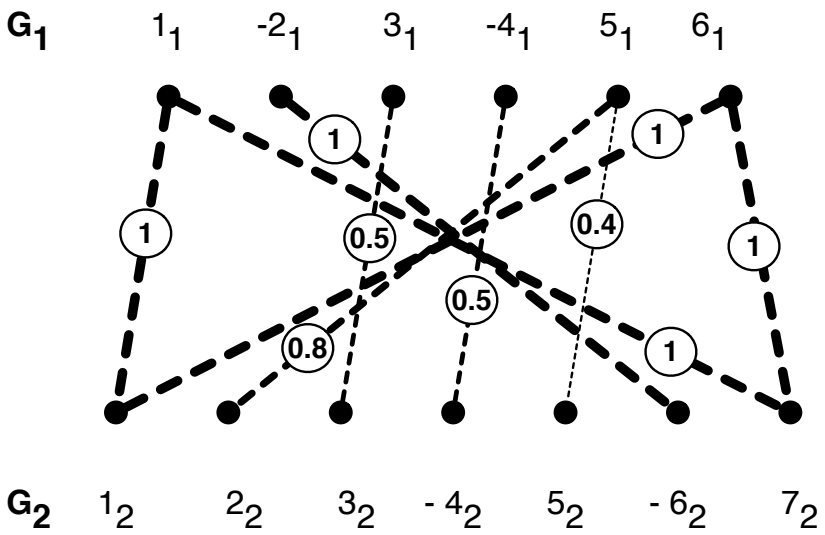
– Wrong gene family assignments produce **incorrect results** in subsequent analyses.
– Datastructure: Strong and weak homology assumptions are **indifferent**.
– Gene family concept **not applicable** for all biological phenomena.

# New model

## Gene family assignment-free

# Gene family assignment free

Normalized similarity measure: $\sigma : G_1 \times G_2 \rightarrow [0, 1]$




Datastructure is an ordered weighted bipartite graph.

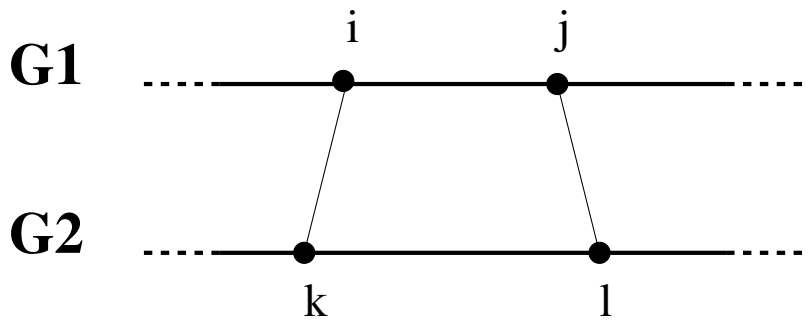
# Conserved Adjacencies

## Scoring scheme for adjacencies

**Adjacencies** in a matching $\mathcal{M}$ are scored according to the measure $\sigma$ of the corresponding edges as follows:

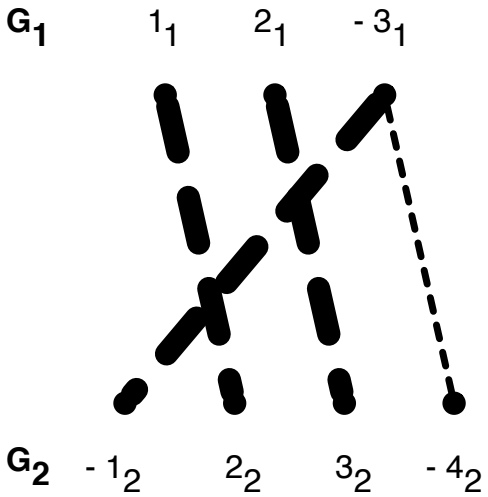
$$s(i,j,k,\ell) = \begin{cases} \sigma(G_1[i], G_2[k]) \cdot \sigma(G_1[j], G_2[\ell]) & \text{if adjacent } * \\ 0 & \text{otherwise} \end{cases}$$

*adjacent: $(G_1[i], G_1[j])$ and $(G_2[k], G_2[\ell])$ are saturated and consecutive (taking sign into account)



# Adjacencies or edges?

Quantifying the **quality of a matching** $\mathcal{M}$: Adjacencies vs edges.




$$adj(\mathcal{M}) = \sum_{\substack{0 \le i < j \le |G_1|, \\ 0 \le k, \ell \le |G_2|}} s(i,j,k,\ell)$$

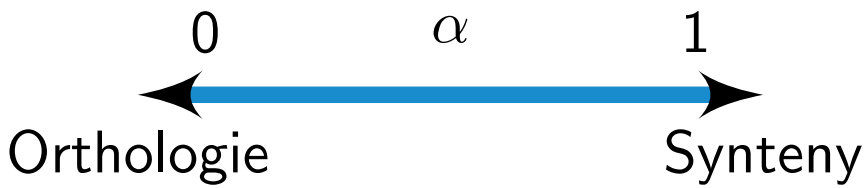
$$edg(\mathcal{M}) = \sum_{e \in \mathcal{M}} \sigma(e)^2$$

# Studied problem

## Family-free(FF)-Adjacencies problem

Given two genomes $G_1$ and $G_2$ and some $\alpha \in ]0, 1]$, find an intermediate matching $\mathcal{M}$ such that at least one edge per connected component is covered and the following formula is maximized:

$$\mathcal{F}_\alpha(\mathcal{M}) = \alpha \cdot adj(\mathcal{M}) + (1 - \alpha) \cdot edg(\mathcal{M}) .$$



# Our strategy

**Goal:** Family-free comparative genome analysis.

**Strategy:** Resolve a particular case: For a given pair of genomes $G_1$ and $G_2$, find optimal solution for FF-adjacencies problem (**NP-hard** problem).

**Method:** Exact algorithm and heuristic.

# Algorithms

# Algorithms

**FFAdj-Int** Exact algorithm, implemented as pseudo-boolean program, based on previous work

[Angibaud *et al.*, 2008]

**FFAdj-MCS** Heuristic, based on LCS - Longest Common Substring [Marron *et al.*, 2004]

# Evaluation of our methods

# Experimentation

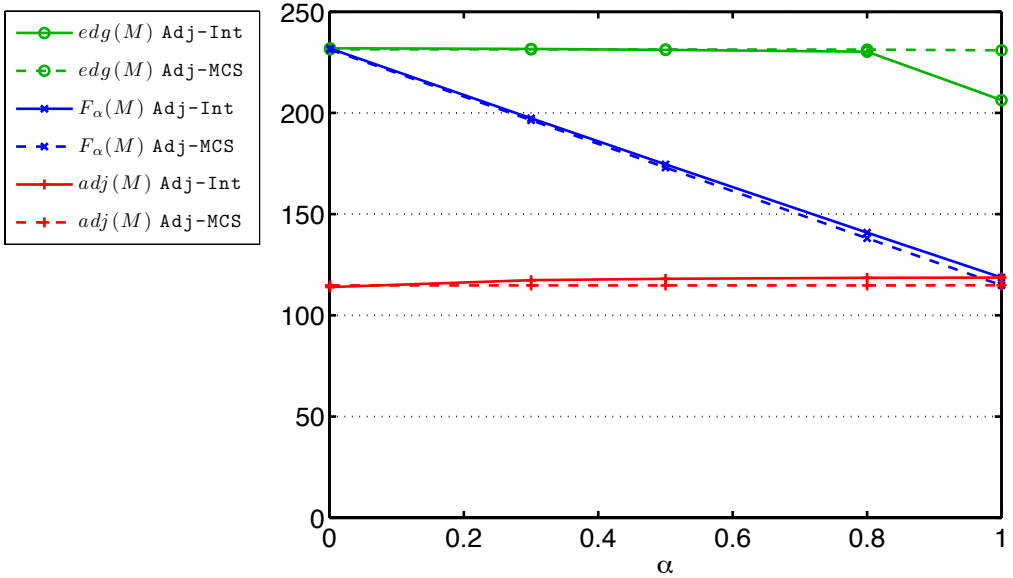
## Dataset

- 12 $\gamma$-proteobacteria complete genomes,
- Size: Between 564 and 5571 genes,
- Already used in [Angibaud *et al.*, 2008],
- 7.6% of duplicated genes.

- The parameter $\alpha$ is in $\{0.001, 0.3, 0.5, 0.8, 1\}$.
- Pairwise normalized similarities $\sigma$ were obtained using the Relative Reciprocal BLAST Score (RRBS)
- The solver used is `CPLEX` `http://www.ilog.com/products/cplex`.

# Evaluation of our algorithms

- For **40 out of 66 pairs of genomes** we could solve the pseudo-boolean program for all values of $\alpha$.
- The heuristic `FFAdj-MCS` deviates in the objective by **less than 3%** (between 0.2% for $\alpha = 0.001$ and 2.9% for $\alpha = 1$).



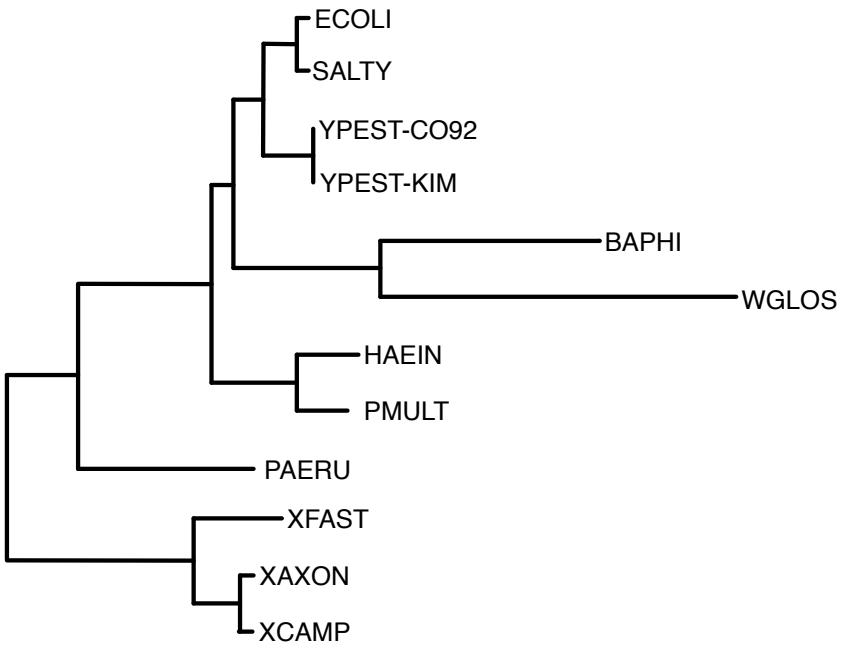
# Comparison with previous results

## With or without gene family assign

Although the number of adjacencies is artificially increased in the gene family assignment study (the genes are unsigns), we observed **the same number of adjacencies** relative to the size of the matching (which increase) in the results of `FFAdj-Int` (for $\alpha = 1$).

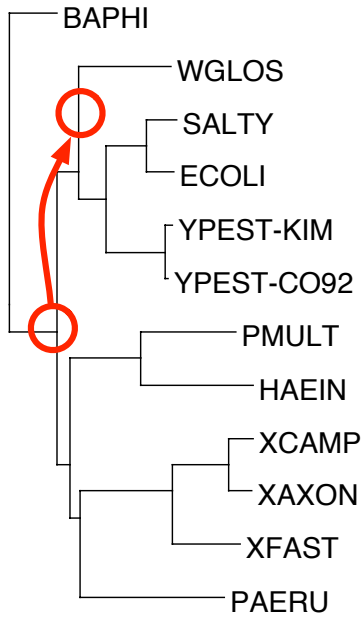
# Reconstructed trees

True phylogeny [Lerat, 2003]




Reconstructed tree

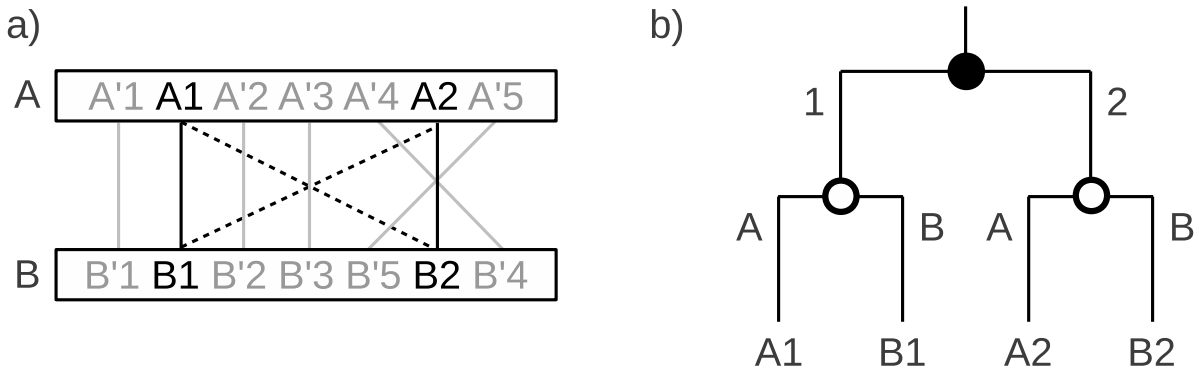
(Robinson-Foulds distance: 2)




This branch is known to be particularly hard to reconstruct since the two organisms diverged far from each other.

# Orthology detection

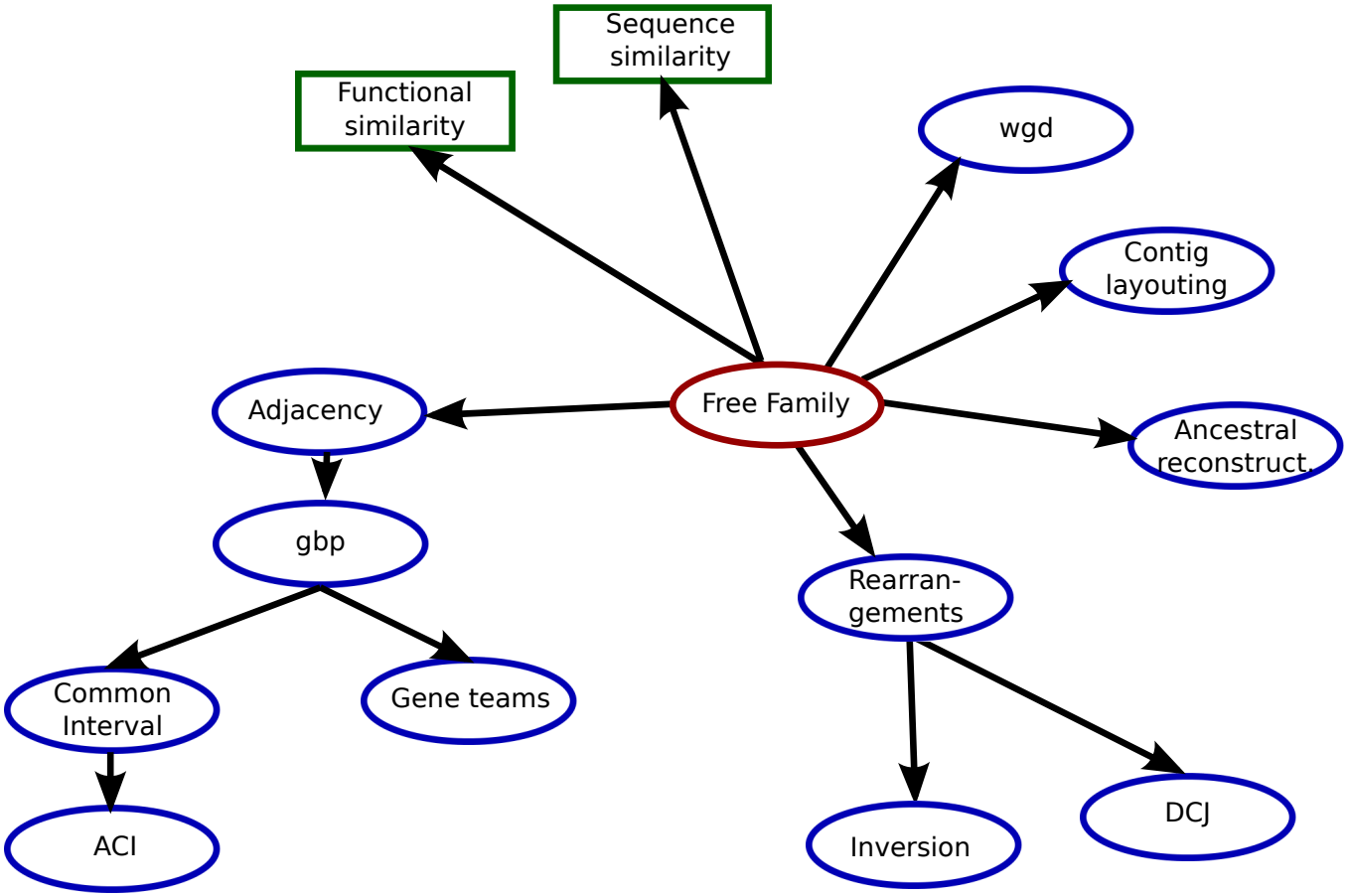
In collaboration with Marcus Lechner, Maribel Hernandez-Rosales, Nicolas Wieseke, Sonja J. Prohaska and Peter Stadler, we improve the **detection of orthology** by combining clustering and synteny.



# Other free-family projects

# Apply the new model to different problems

In collaboration with Marília Braga, Cédric Chauve, Katharina Jahn and Roland Wittler.



# Acknowledgments

Unterstützt von / Supported by

Alexander von Humboldt
Stiftung/Foundation