Reconstruction of ancestral gene orders

Pedro Feijao

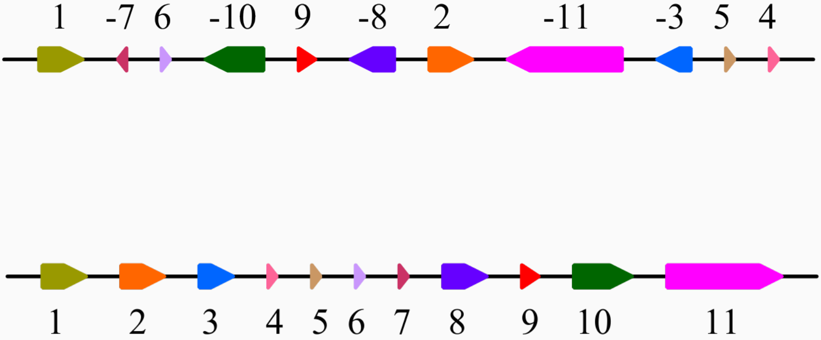
Banff, February 17, 2017

SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

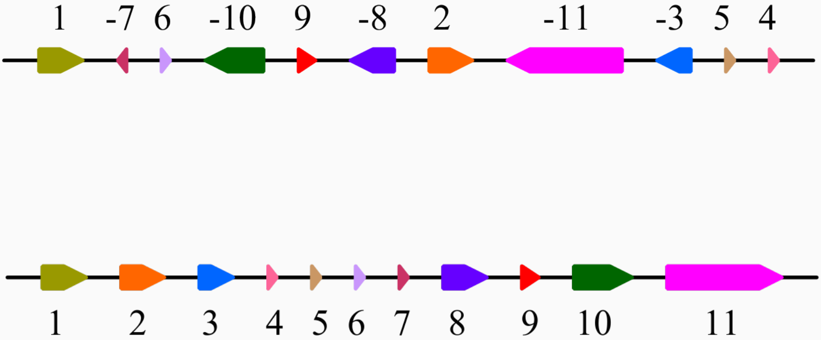
Universität Bielefeld
Technische Fakultät
Genominformatik

Introduction

Genomes



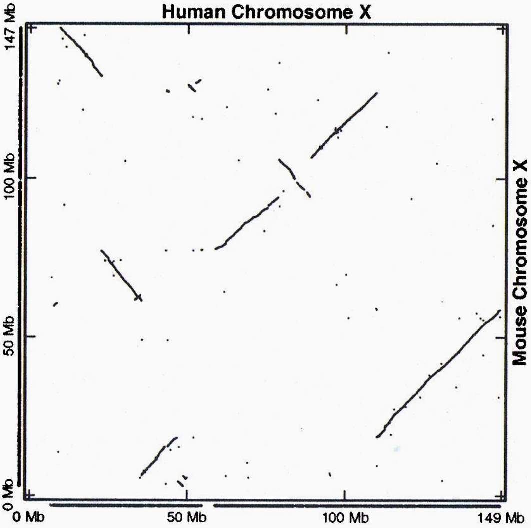
Mouse X-Chromosome

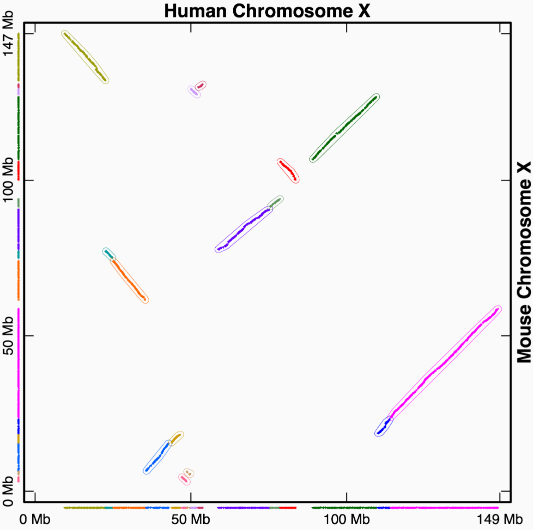


Human X-Chromosome

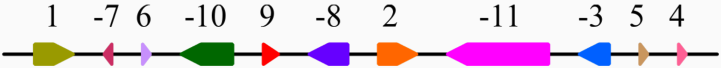
(Pevzner and Tesler, 2003)

Genomes

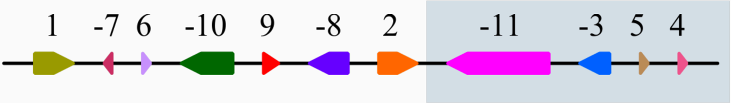

Human Chromosome X
Mouse Chromosome X
0 Mb
50 Mb
100 Mb
149 Mb
0 Mb
50 Mb
100 Mb
147 Mb


Human Chromosome X
Mouse Chromosome X
0 Mb
50 Mb
100 Mb
149 Mb
0 Mb
50 Mb
100 Mb
147 Mb

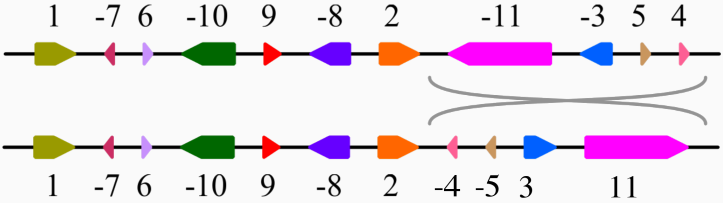
Reversal



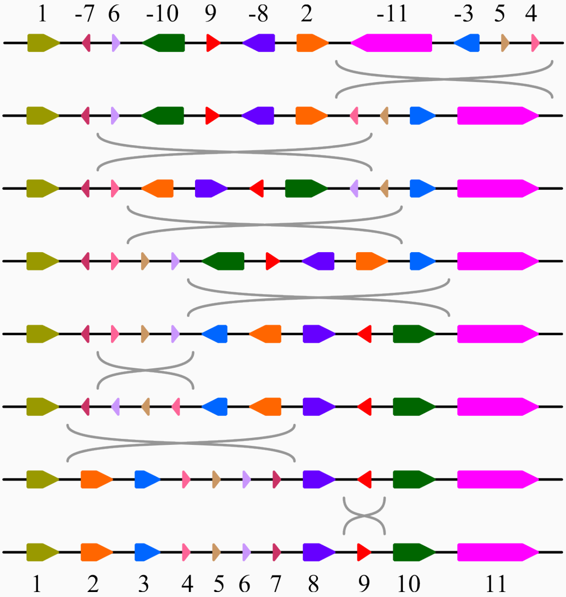
Reversal



Reversal



Rearrangement Scenario



Genome Rearrangement Problems

- **Distance**: Minimum # of rearrangements from A to B?

Genome Rearrangement Problems

- **Distance**: Minimum # of rearrangements from A to B?
- **Scenario**: Which rearrangements? (also called *Sorting*)

Genome Rearrangement Problems

- **Distance**: Minimum # of rearrangements from A to B?
- **Scenario**: Which rearrangements? (also called *Sorting*)
- **Phylogeny**: How did the genomes evolve?

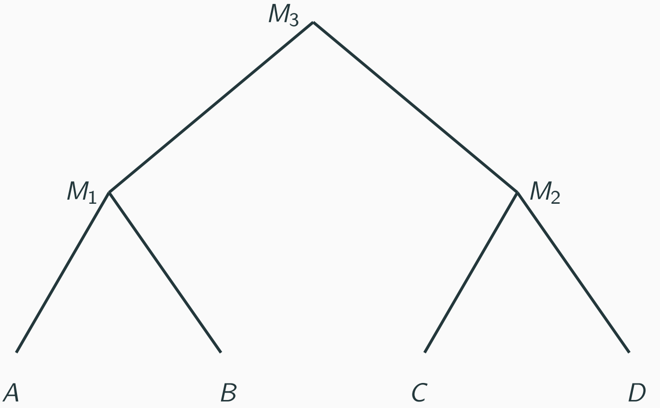
Genome Rearrangement Problems

- **Distance**: Minimum # of rearrangements from A to B?
- **Scenario**: Which rearrangements? (also called *Sorting*)
- **Phylogeny**: How did the genomes evolve?
- **Ancestral Reconstruction**: How do the ancestors look like?

translocations
algorithm
adjacencies
breakpoint
segments
model
probabilistic
reciprocal
computing
sorting
DCJ
genes
approximation
double
maximum
ancestral
problem
reconciled
graphs
exact
problems
InDels
median
linear
join
loss
SCJ
contiguous
duplication
genome
complexity
signed
algorithms
segmental
comparing
cut
orders
using
evolution
gene
distance
duplicate
genomic
programming
DCJ-indel
multichromosomal
parsimony
reversals
reconstruction
phylogeny
duplication

Ancestral Reconstruction

Ancestral Reconstruction



Input: Tree and genomes A, B, C, D

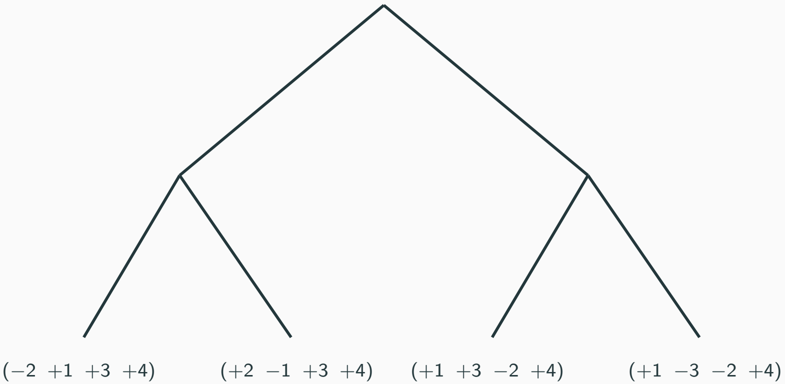
Ouput: Ancestral genomes (M_1, M_2, M_3)

Ancestral Reconstruction

- Distance-based Methods

- Homology-based Methods

Distance-based Methods



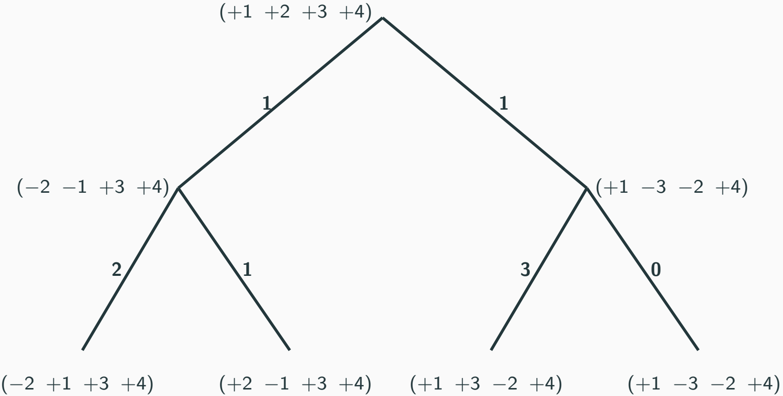
$(-2\ \ +1\ \ +3\ \ +4)$

$(+2\ \ -1\ \ +3\ \ +4)$

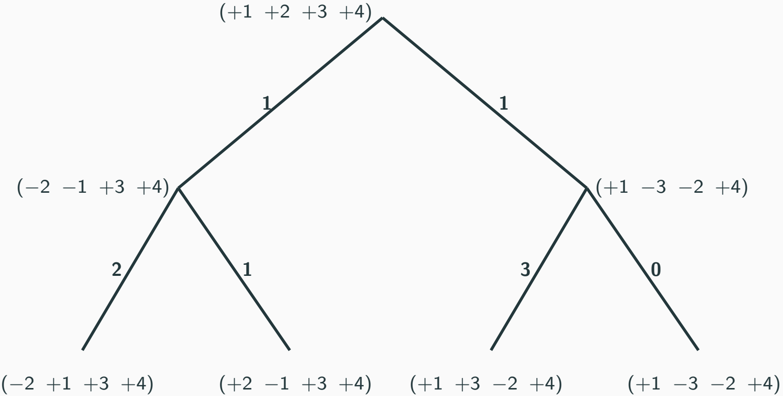
$(+1\ \ +3\ \ -2\ \ +4)$

$(+1\ \ -3\ \ -2\ \ +4)$

Distance-based Methods

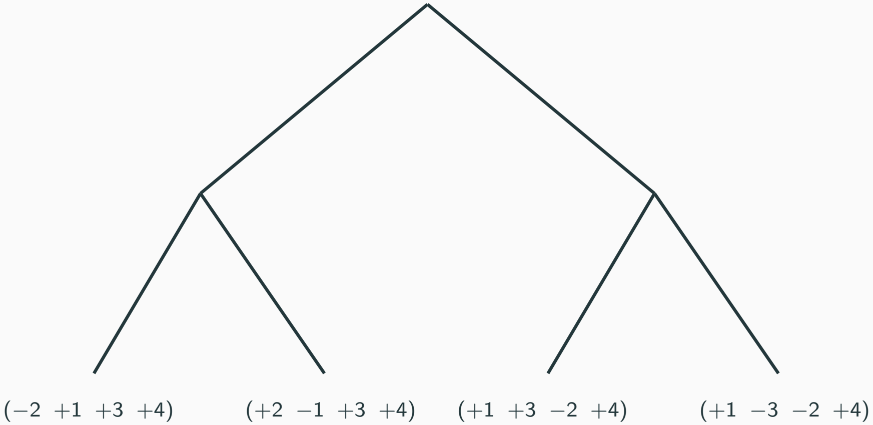


Distance-based Methods



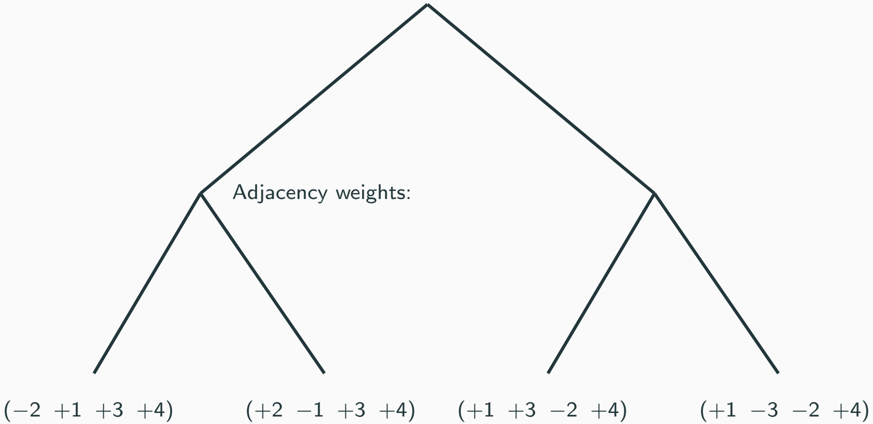
Find ancestral genomes that **minimize events** on the tree
→ **Small Parsimony Problem**

Homology-based Methods



$(-2\ \ +1\ \ +3\ \ +4)$ $(+2\ \ -1\ \ +3\ \ +4)$ $(+1\ \ +3\ \ -2\ \ +4)$ $(+1\ \ -3\ \ -2\ \ +4)$

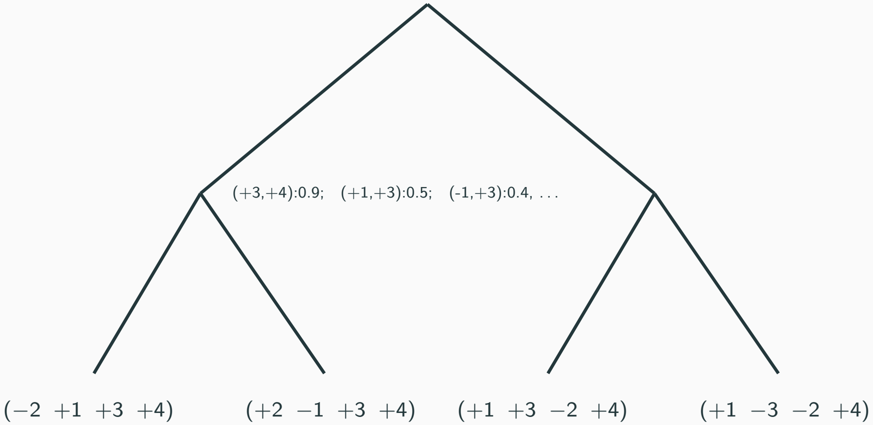
Homology-based Methods



Adjacency weights:

$(-2\ \ +1\ \ +3\ \ +4)$ $(+2\ \ -1\ \ +3\ \ +4)$ $(+1\ \ +3\ \ -2\ \ +4)$ $(+1\ \ -3\ \ -2\ \ +4)$

Homology-based Methods



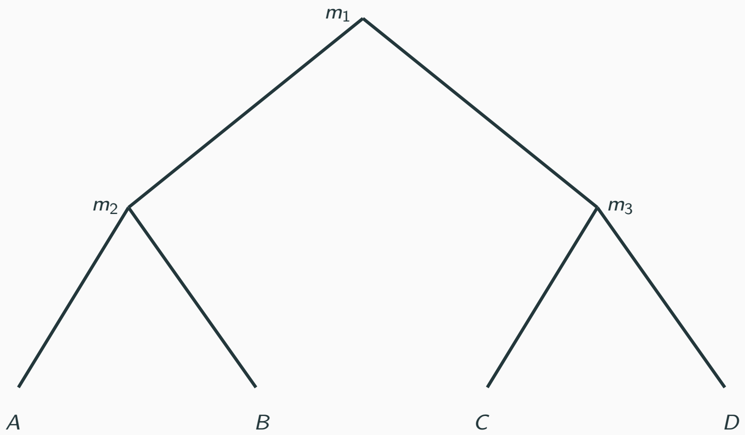
(+3,+4):0.9; (+1,+3):0.5; (-1,+3):0.4, . . .

$(-2\ \ +1\ \ +3\ \ +4)$ $(+2\ \ -1\ \ +3\ \ +4)$ $(+1\ \ +3\ \ -2\ \ +4)$ $(+1\ \ -3\ \ -2\ \ +4)$

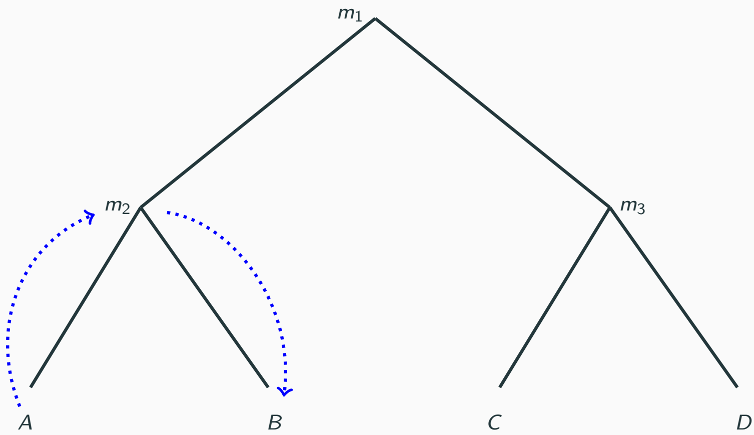
Ancestral Reconstruction methods

- Distance-based methods:
  - Assume a rearrangement model
  - Minimize branch lengths
- Homology-based methods:
  - Find conserved structures
  - Maximize some weight/probability function

Intermediate Genomes



Intermediate Genomes



Problem Definition

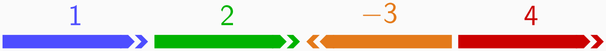
Ancestral Reconstruction where internal nodes

are

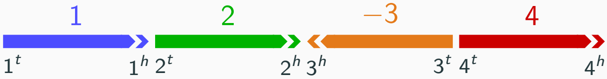
Intermediate Genomes of its children.

Definitions

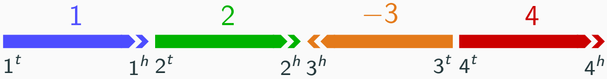
Genome model



Genome model

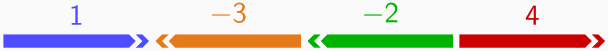


Genome model

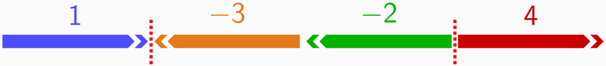


$$A = \{\circ 1^t, 1^h 2^t, 2^h 3^h, 3^t 4^t, 4^h \circ\}$$

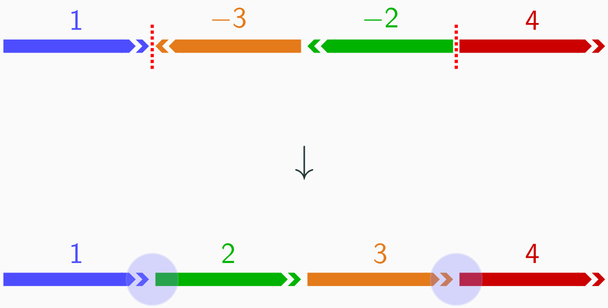
The Double-Cut-and-Join (DCJ) operation



The Double-Cut-and-Join (DCJ) operation



The Double-Cut-and-Join (DCJ) operation



Breakpoint Graph $BP(A, B)$

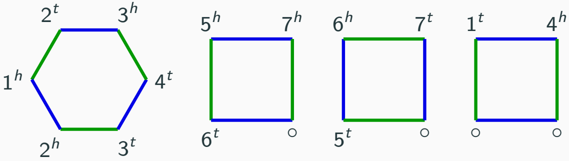
$$A = \{\circ 1^t, 1^h 2^t, 2^h 3^t, 3^h 4^t, 4^h \circ, \circ 5^t, 5^h 6^t, 6^h 7^t, 7^h \circ\}$$

$$B = \{1^h 2^h, 2^t 3^h, 3^t 4^t, 4^h 1^t, \circ 6^t, 6^h 5^t, 5^h 7^h, 7^t \circ\}$$

Breakpoint Graph $BP(A, B)$

$$A = \{\circ 1^t, 1^h 2^t, 2^h 3^t, 3^h 4^t, 4^h \circ, \circ 5^t, 5^h 6^t, 6^h 7^t, 7^h \circ\}$$

$$B = \{1^h 2^h, 2^t 3^h, 3^t 4^t, 4^h 1^t, \circ 6^t, 6^h 5^t, 5^h 7^h, 7^t \circ\}$$



A-edges are drawn in green, and *B*-edges in blue.

DCJ Distance

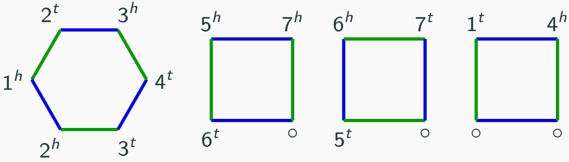
$$d_{\mathrm{DCJ}}(A, B) = N - C$$

where N is the number of genes and C is the number of cycles in $BP(A, B)$.

(Bergeron et al, 2006)

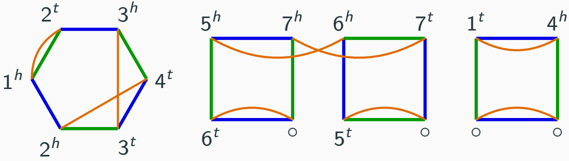
Intermediate Genomes

Genomes are **matchings** in the BP graph:



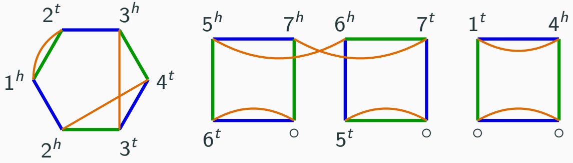
Intermediate Genomes

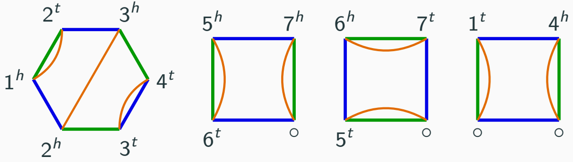
Genomes are **matchings** in the BP graph:



Intermediate Genomes

Genomes are **matchings** in the BP graph:





Intermediate Genomes

Edges are non-crossing chords in the cycles of $BP(A, B)$

$$\Rightarrow$$

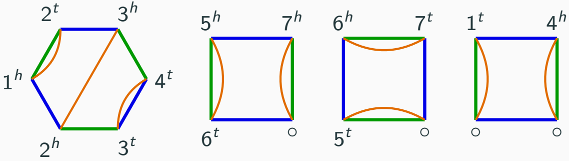
Intermediate Genome of A and B

Intermediate Genomes

Edges are non-crossing chords in the cycles of $BP(A, B)$

$\Rightarrow$

Intermediate Genome of A and B



Intermediate Genomes

- Very easy to detect (linear time)

Intermediate Genomes

- Very easy to detect (linear time)

- Reduces the search space. In the example:

Intermediate Genomes

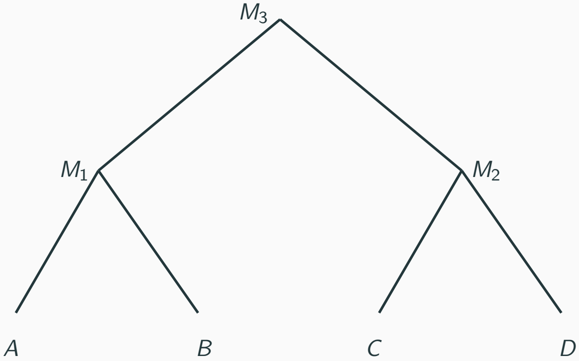
- Very easy to detect (linear time)
- Reduces the search space. In the example:
    - 34,459,425 possible genomes

Intermediate Genomes

- Very easy to detect (linear time)
- Reduces the search space. In the example:
    - 34,459,425 possible genomes
    - Only 40 intermediate genomes between A and B.

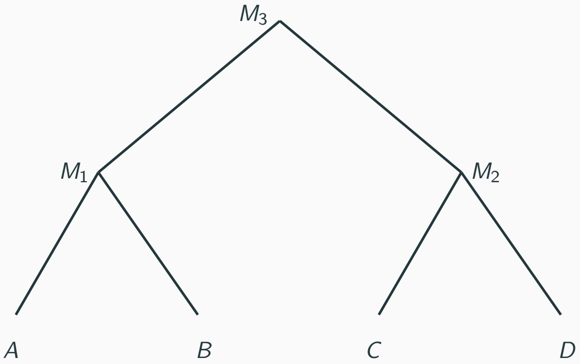
Methods

Ancestral Reconstruction with IG



- Small parsimony with the restriction that internal nodes are IG's of the children.

Ancestral Reconstruction with IG



- Small parsimony with the restriction that internal nodes are IG's of the children.

- Still NP-hard

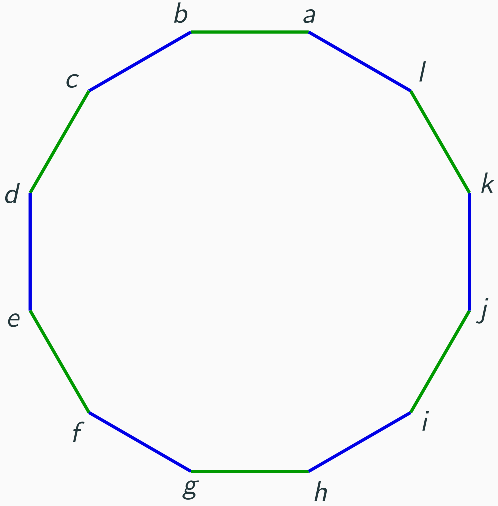
Maximum Weight Intermediate Genome

- Given adjacency weights, can we find an IG with maximum weight?
- **Maximum Weight Independent Set**: Polynomial Time

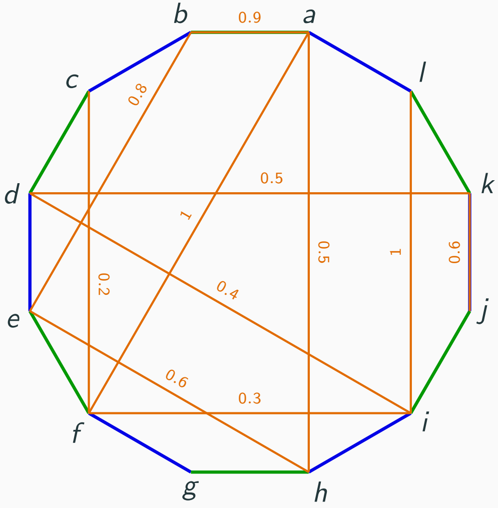
Finding Adjacency Weights

- DeClone (Chauve et al., 2015)
- New proposed algorithm based on InferCARs (Ma et al., 2006).

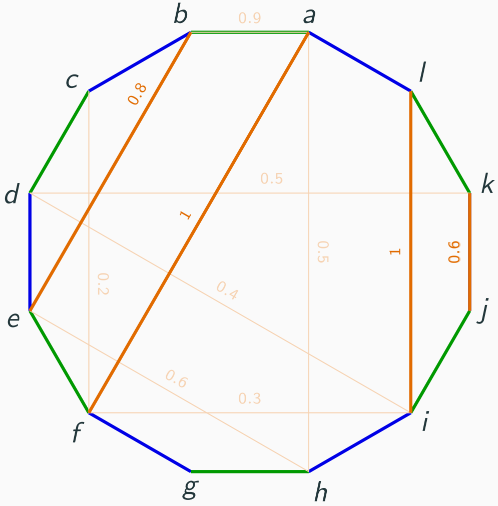
Example



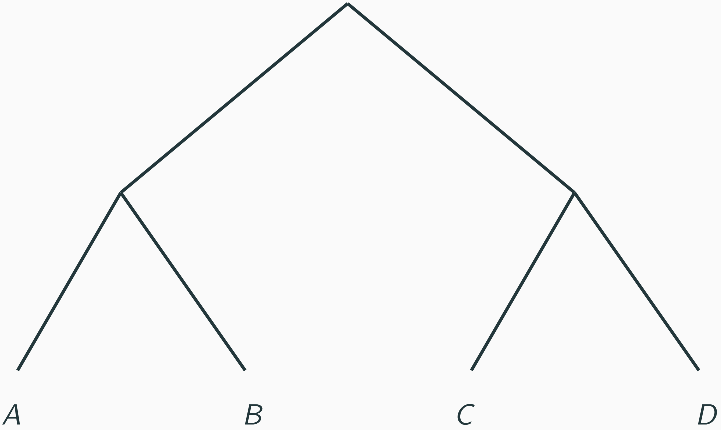
Example



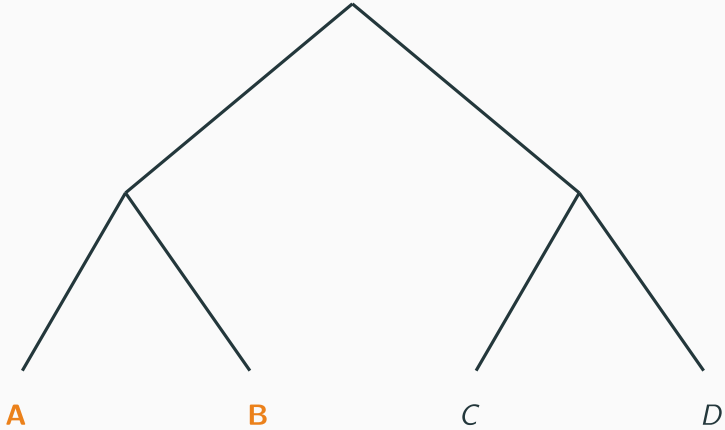
Example



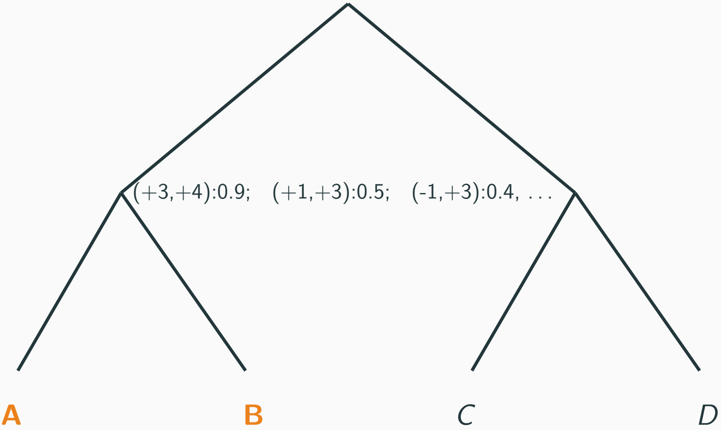
Algorithm Overview



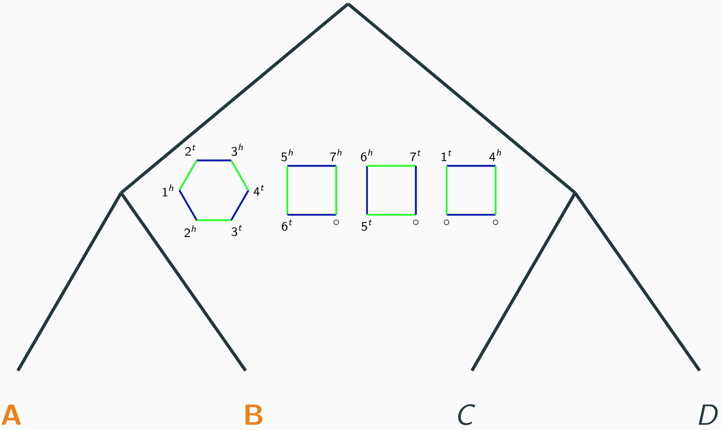
Algorithm Overview



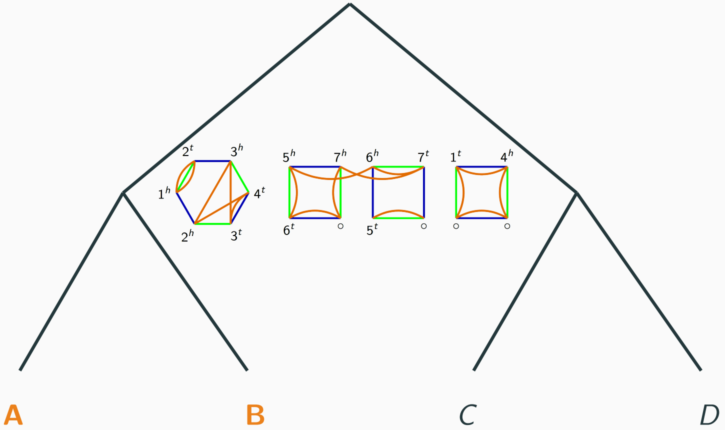
Algorithm Overview



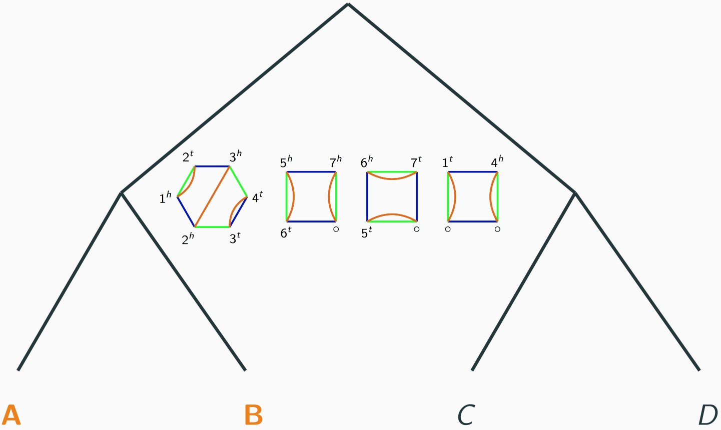
Algorithm Overview



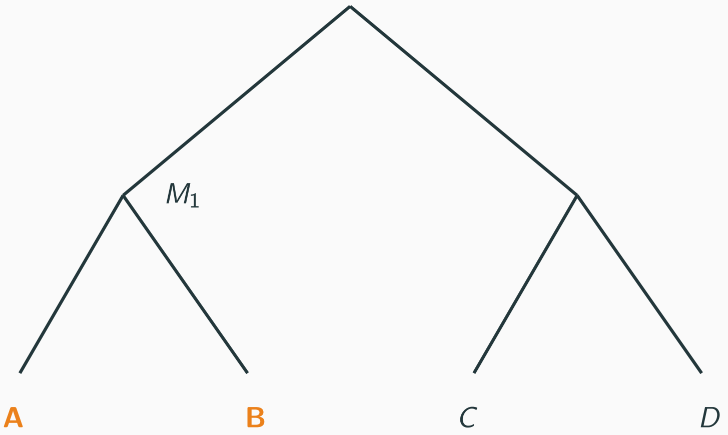
Algorithm Overview



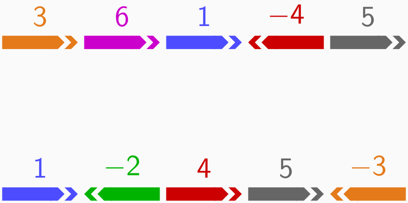
Algorithm Overview



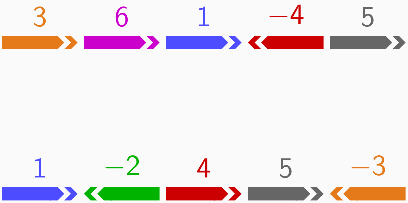
Algorithm Overview



Genomes with unique genes

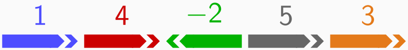


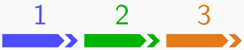
Genomes with unique genes



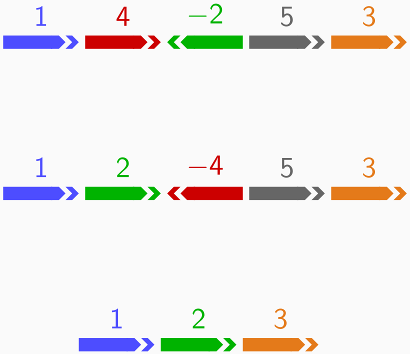
DCJ InDel Model (Braga et al., 2010; Compeau, 2012)

DCJ InDel example

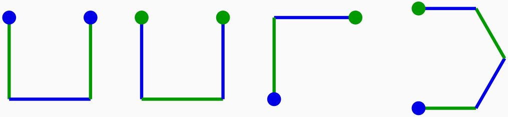




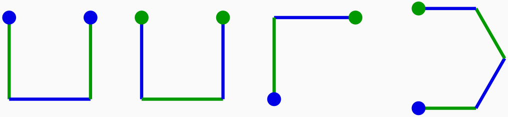
DCJ InDel example



Breakpoint graph with Unique genes

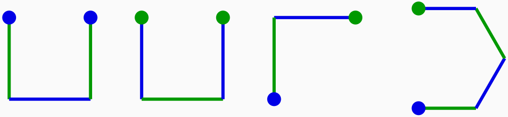


Breakpoint graph with Unique genes



New components: AA-, BB-, AB- , A-, and B-paths.

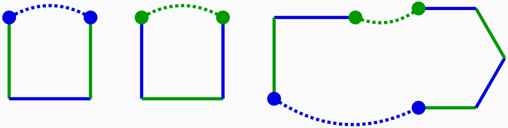
Breakpoint graph with Unique genes



New components: AA-, BB-, AB- , A-, and B-paths.

Find an **optimal completion**

Breakpoint graph with Unique genes

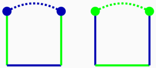


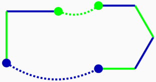
New components: AA-, BB-, AB- , A-, and B-paths.

Find an **optimal completion**

Optimal completion

- AA- and BB- components are closed





- AB- are paired

- A- and B- paths are paired, with opposing parity.

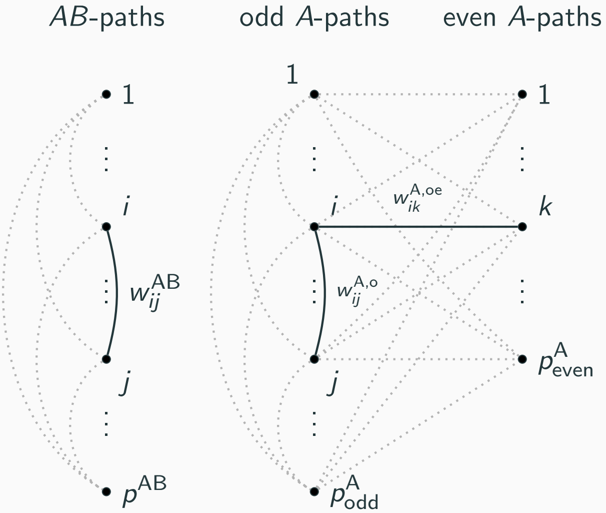
- Sometimes A-, B- and AB- paths are joined in a **triplet**.

Maximum Weight Intermediate Genome

- How to find a completion with maximum weight?

- Calculate all possible pairings and solve a *Maximum Weight Matching*

Maximum Weight Matching



MWM is not always optimal

- Sometimes A-, AB-, B- triplets are possible.
- Triple matching is usually NP-hard, but it is still open in this case.

Results

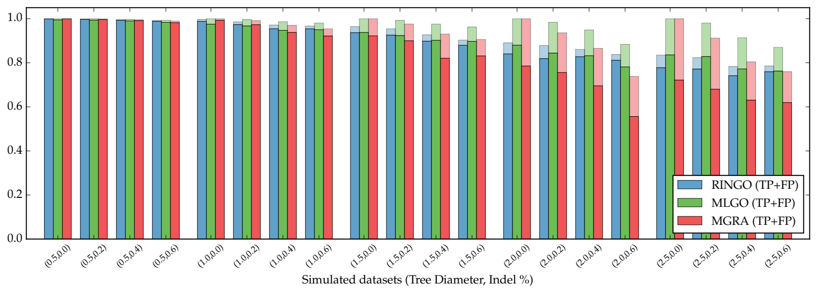
Methods

- RINGO - ancestral **R**econstruction with **IN**termediate **G**en**O**mes (Feijao and Araujo, 2016)

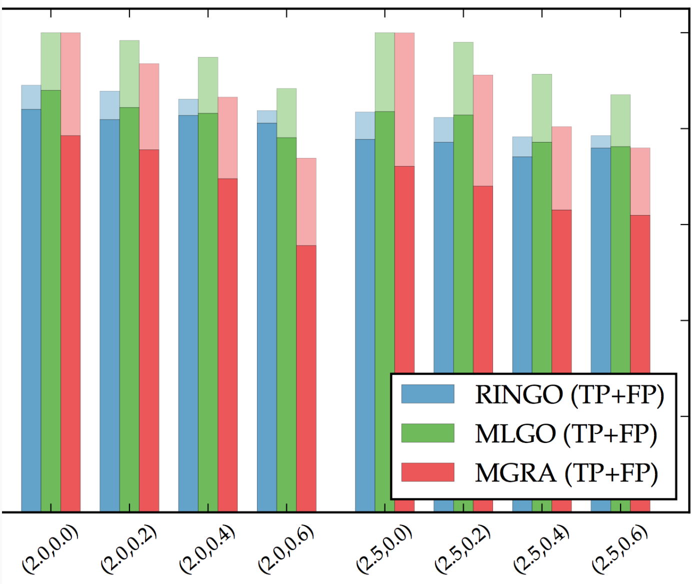
- MGRA2 (Avdeyev et al., 2016)

- MLGO (Hu et al., 2014)

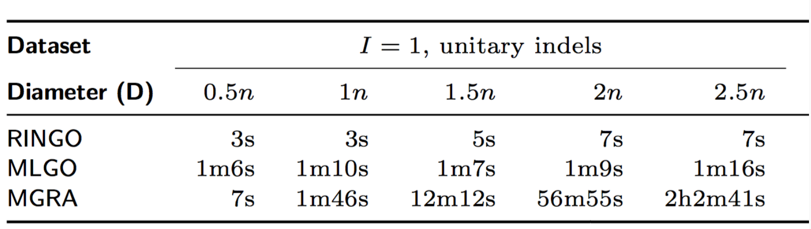
Quality Results



Quality Results



Running Time



Dataset	$I = 1$, unitary indels				
Diameter (D)	$0.5n$	$1n$	$1.5n$	$2n$	$2.5n$
RINGO	3s	3s	5s	7s	7s
MLGO	1m6s	1m10s	1m7s	1m9s	1m16s
MGRA	7s	1m46s	12m12s	56m55s	2h2m41s

Current Challenges

- Duplicated genes
- Statistical models

Acknowledgments

- Elói Araújo (UFMS, Brazil)

UFMS

- Jens Stoye (Bielefeld University, Germany)

Universität Bielefeld

Technische Fakultät
Genominformatik

Thanks!