HP Distance Via Double Cut and Join Distance

A. Bergeron[1] J. Mixtacki[2] J. Stoye[3]

[1]Université du Québec à Montréal

[2]International NRW Graduate School
in Bioinformatics and Genome Research
Universität Bielefeld

[3]Technische Fakultät
Universität Bielefeld

CPM 2008, Pisa, June 18-20, 2008

1 Introduction
- Biological Background
- The Double Cut and Join (DCJ) Model

2 Computing the General HP Distance
- Components and Oriented Sorting
- Destroying Unoriented Components
- Unoriented Sorting

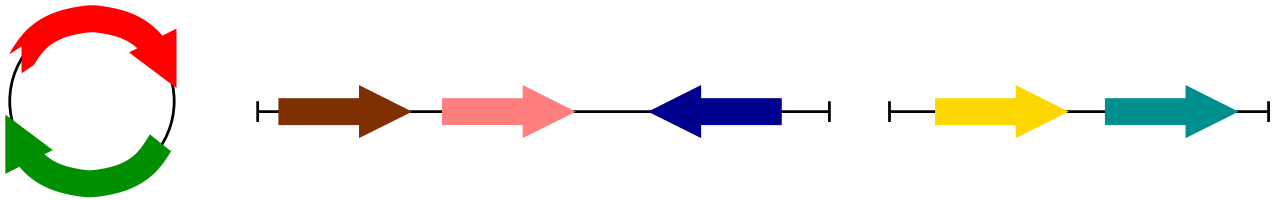
3 Summary

1 Introduction
- Biological Background
- The Double Cut and Join (DCJ) Model

2 Computing the General HP Distance
- Components and Oriented Sorting
- Destroying Unoriented Components
- Unoriented Sorting

3 Summary

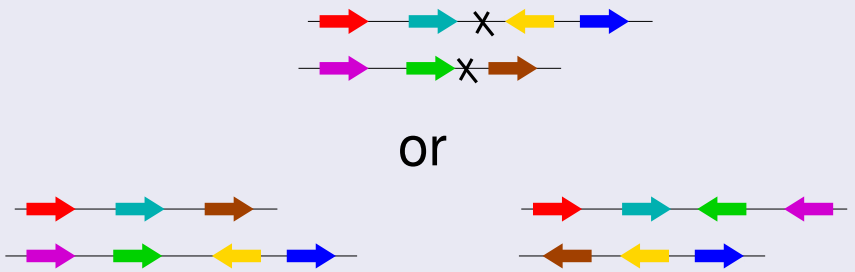
Biological Background



- Genome is the entire DNA of a living organism
- Gene is a segment of DNA that is involved e.g. in producing a protein, and its orientation depends on the DNA strand that it lies on
- Genome consists of chromosomes
- Chromosomes are linear or circular

Operations on *two* chromosomes:

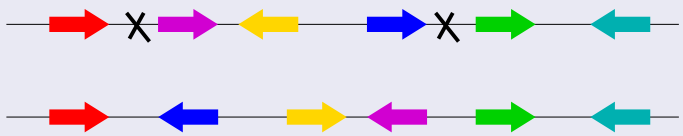
- Translocations exchange two chromosome ends:



- Fusions and fissions are translocations involving or creating empty chromosomes

Operations on *one* chromosome:

- Inversions reverse the order and the orientation of a segment:



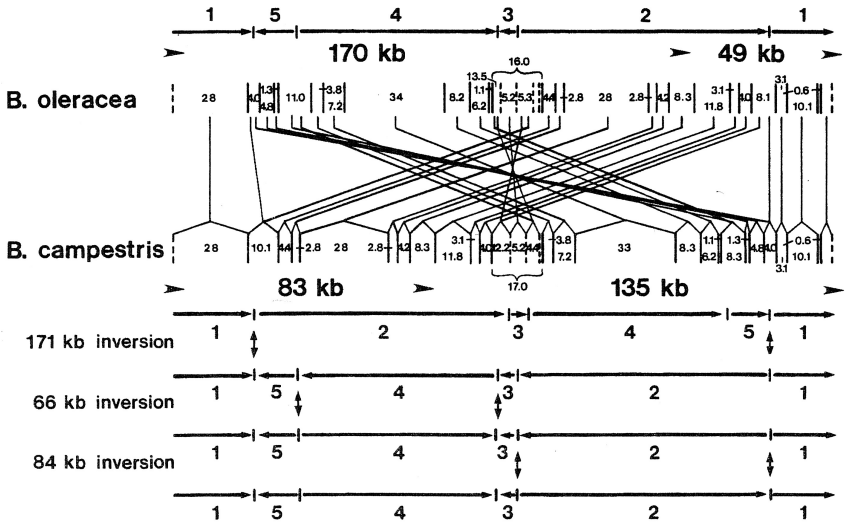
- Block interchanges exchange two segments
- Transpositions are block interchanges whose exchanged segments are adjacent

Genome Rearrangements

Genome rearrangements change the content and/or the order of genes of a genome:

- inversions
- transpositions
- translocations
- fusions and fissions
- ...

(Picture: Palmer & Herbon, 1988)



The number of rearrangements needed to transform one genome into another is a measure for the evolutionary distance between two species

Genomic distance

$d(A, B)$: minimum number of operations needed to transform genome A into genome B

- What kind of genome model?
- Which set of operations?

HP distance d_{HP} (HP 1995, T 2002, O-FS 2003, JN 2007)

- Linear chromosomes
- Translocations, fusions, fissions and inversions

DCJ distance d_{DCJ} (YAF 2005, BMS 2006)

- Linear and circular chromosomes are allowed
- All classical operations are included

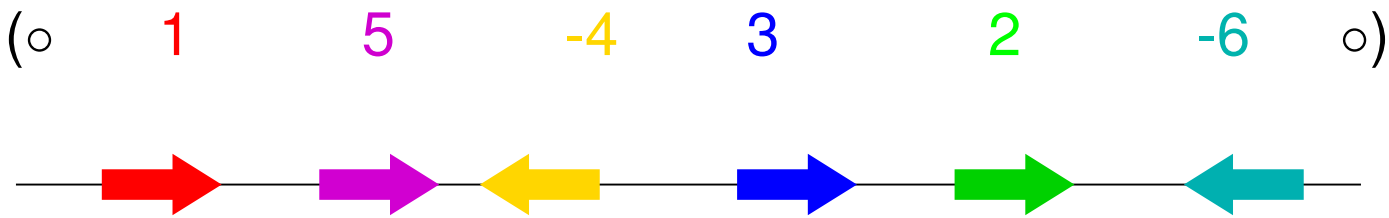
Our Goal:

$$d_{HP} = d_{DCJ} + t$$

The Double Cut and Join (DCJ) Model

Multi-chromosomal, linear genomes with the same N genes:

- Gene is represented by a signed integer between 1 and N
- Chromosome is an ordered sequence of signed genes, flanked by two unsigned telomere markers $\circ = -\circ$



- Interval $(l, \ldots, r)$ is a set of consecutive genes or telomere markers within a chromosome; with extremities $\{l, -r\}$
- Adjacency is an interval of length 2
- Telomere is an adjacency that contains a telomere marker

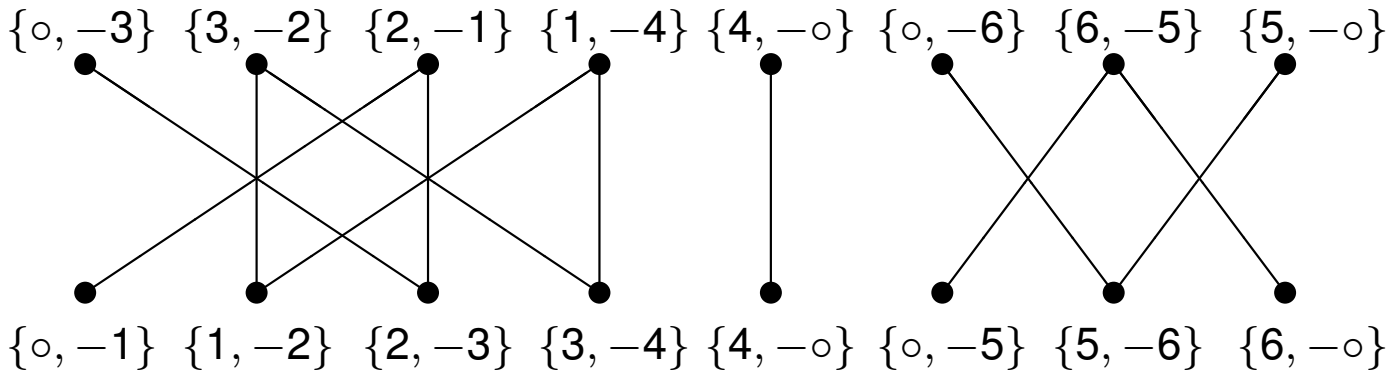
The Double Cut and Join (DCJ) Model

$$A = \{(\circ, 3, 2, 1, 4, \circ), (\circ, 6, 5, \circ)\}$$

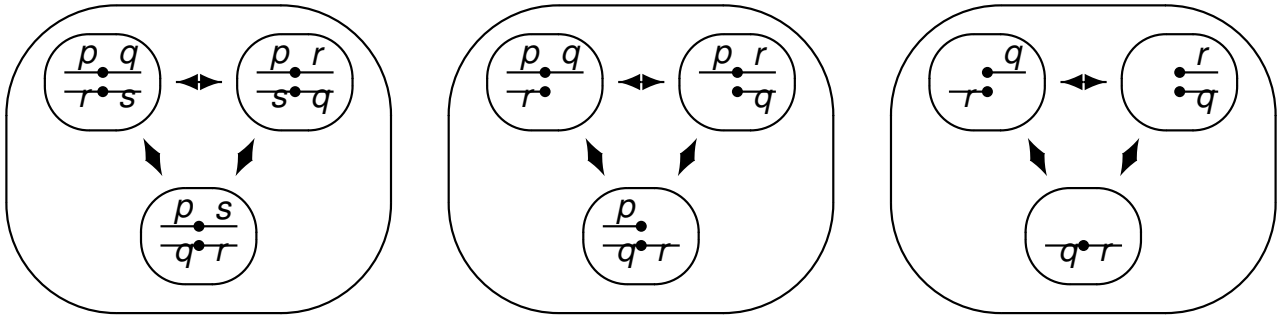
$$B = \{(\circ, 1, 2, 3, 4, \circ), (\circ, 5, 6, \circ)\}$$

Definition 1

Adjacency graph $AG(A, B)$: vertices are the adjacencies of genomes A and B and edges connect either the two adjacencies in which g appears as extremity $+g$, or as $-g$.



DCJ operation acts on two vertices of a graph with vertices of degree one or two in one of the following three ways:



Theorem 1 (BMS 2006)

Let A and B be genomes defined on N genes, then we have

$$d_{DCJ}(A, B) = N - (C + I/2)$$

where $C = \#$ of cycles and $I = \#$ of odd paths in $AG(A, B)$.

DCJ-sorting operation reduces the DCJ distance by 1

Definition 2

A DCJ-sorting operation is oriented if it does not create circular chromosomes.

For linear genomes, oriented operations are necessarily

- inversions,
- translocations,
- fusions, and
- fissions.

Proposition 1

For two linear genomes A and B, we have that

$$d_{DCJ}(A, B) \leq d_{HP}(A, B).$$

1 Introduction
- Biological Background
- The Double Cut and Join (DCJ) Model

2 Computing the General HP Distance
- Components and Oriented Sorting
- Destroying Unoriented Components
- Unoriented Sorting

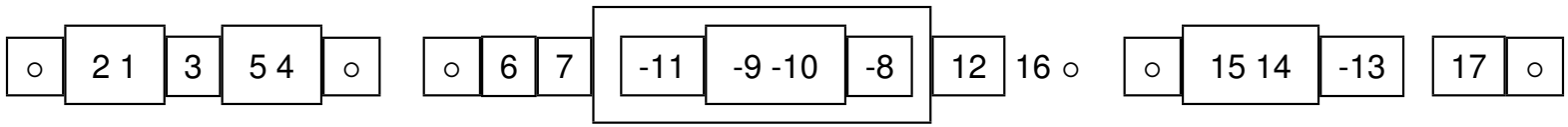
3 Summary

Components

$$A = \{(\circ, 2, 1, 3, 5, 4, \circ), (\circ, 6, 7, -11, -9, -10, -8, 12, 16, \circ), (\circ, 15, 14, -13, 17, \circ)\},$$
$$B = \{(\circ, 1, 2, 3, 4, 5, \circ), (\circ, 6, 7, 8, 9, 10, 11, 12, \circ), (\circ, 13, 14, 15, \circ), (\circ, 16, 17, \circ)\}.$$

Definition 3

An interval $(l, \ldots, r)$ of A is a component if there exists an interval in B:
(a) with the same extremities $\{l, -r\}$,
(b) with the same set of genes, and
(c) that is not the union of two such intervals.



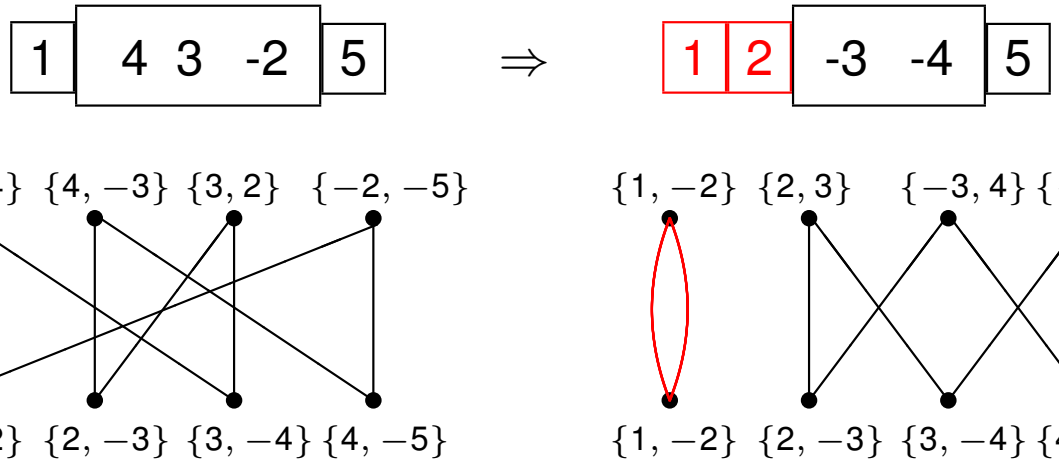
Proposition 2

Components are either disjoint, nested, or overlap on one gene.

Chain: successive linked components

Components and Oriented Sorting

Oriented DCJ-sorting operation:



Lemma 2

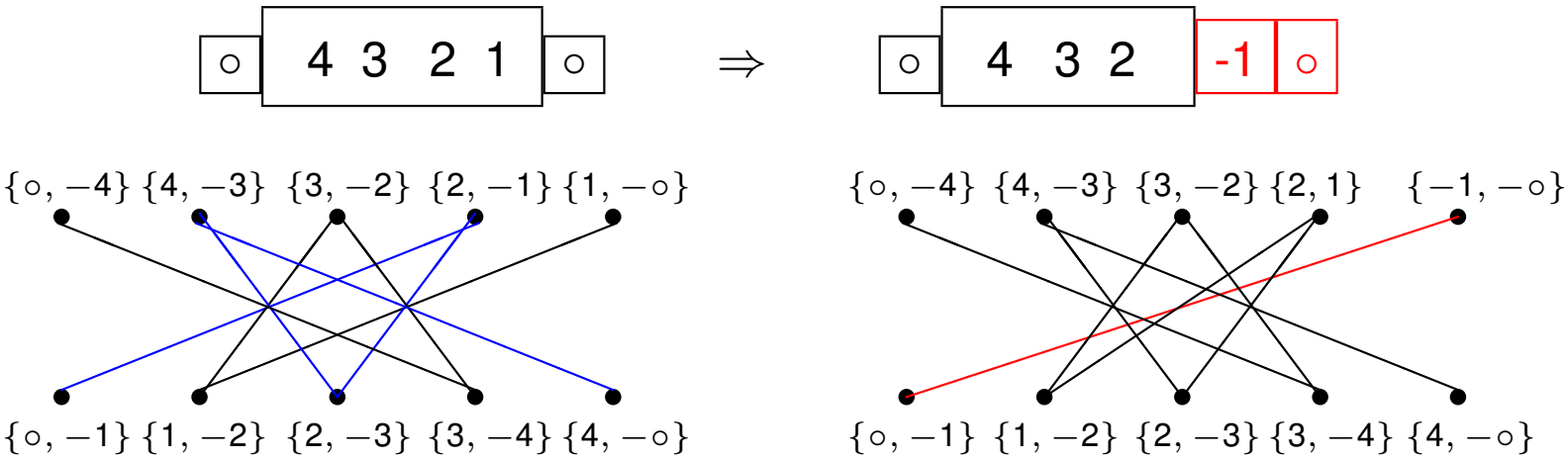
If all elements of a component have the same sign, then no inversion acting on one path/cycle can create a new cycle.

Definition 5

Component is oriented: there exists an oriented DCJ-sorting operation, otherwise it is unoriented.

Components and Oriented Sorting

Oriented DCJ-sorting operation:



Proposition 4

A component is oriented if and only if either its elements have positive and negative signs, or its adjacency graph has two even paths.

Theorem 2

$d_{HP} = d_{DCJ}$ if and only if there are no unoriented components.

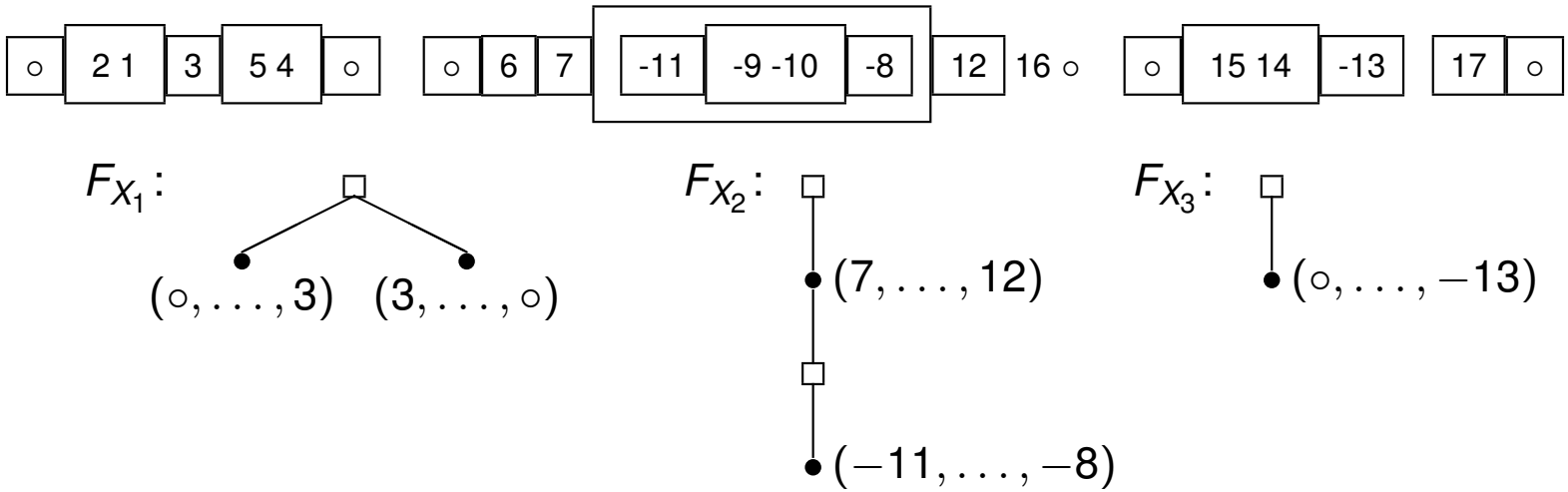
Destroying Unoriented Components

$$A = \{(\circ, 2, 1, 3, 5, 4, \circ), (\circ, 6, 7, -11, -9, -10, -8, 12, 16, \circ), (\circ, 15, 14, -13, 17, \circ)\},$$
$$B = \{(\circ, 1, 2, 3, 4, 5, \circ), (\circ, 6, 7, 8, 9, 10, 11, 12, \circ), (\circ, 13, 14, 15, \circ), (\circ, 16, 17, \circ)\}.$$

Definition 6

The forest F_X of chromosome X is defined by the construction:

1. Each non-trivial component is a round node.
2. Each maximal chain is a square node whose (ordered) children are the round nodes.
3. A square node is the child of the smallest component that contains this chain.

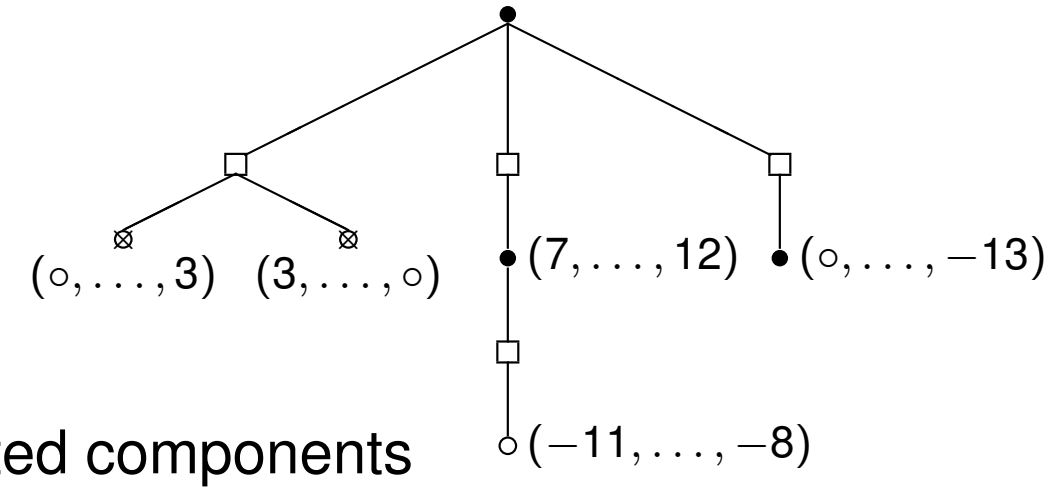


Destroying Unoriented Components

Definition 7

The tree T is given by the construction:

1. The root is a round node.
2. All trees of forests $\{F_{X_1}, F_{X_2}, \ldots, F_{X_K}\}$ of chromosomes $\{X_1, X_2, \ldots, X_K\}$ are children of the root.

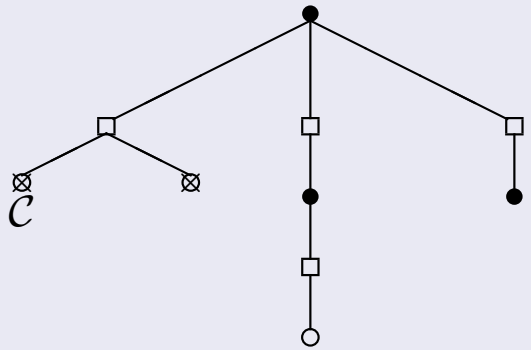


Round nodes are *painted*...
black: the root and all oriented components
white: unoriented components that do not contain telomeres
grey: unoriented components that contain one or two telomeres

Destroying unoriented components

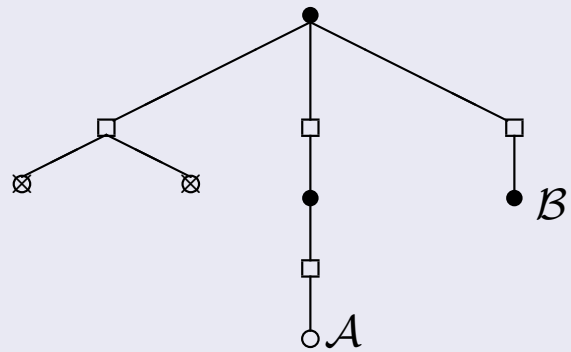
Proposition 7 (Component Cutting)

Unoriented component $\mathcal{C}$:
any inversion within the same cycle/path orients $\mathcal{C}$ and leaves the number of cycles and paths unchanged.



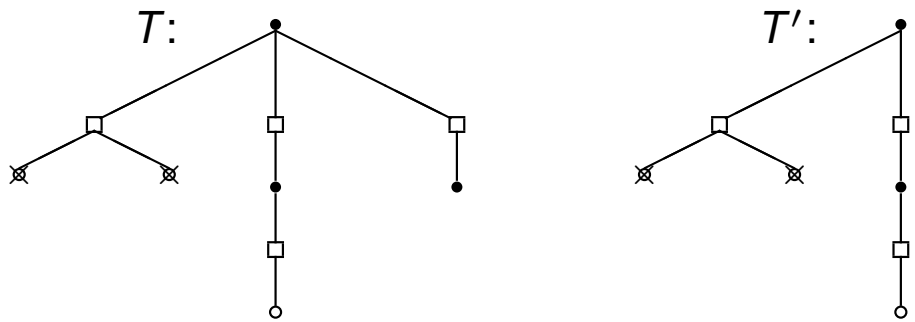
Proposition 8 (Component Merging)

Unoriented components $\mathcal{A}$ and $\mathcal{B}$:
a DCJ operation between $\mathcal{A}$ and $\mathcal{B}$ destroys or orients all components on the path from $\mathcal{A}$ to $\mathcal{B}$, without creating new unoriented components.



Unoriented Sorting

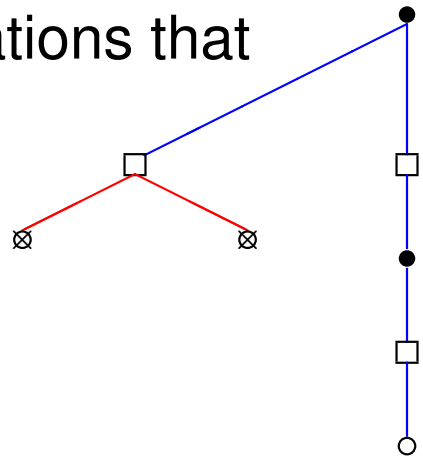
T': smallest subtree of T that contains all the unoriented components



Definition 8

A cover of T' is a collection of paths joining all the unoriented components, such that each terminal node of a path belongs to a unique path.

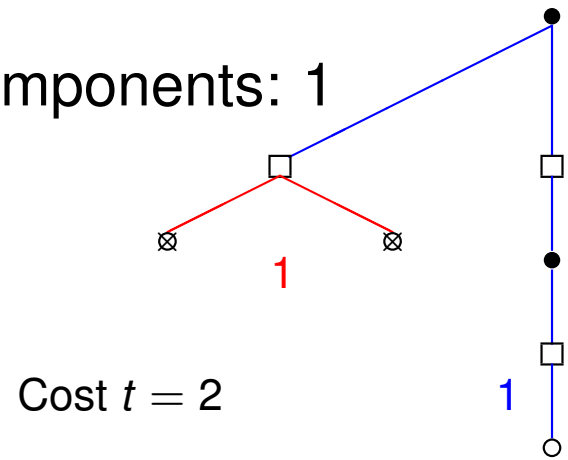
- Each cover of T' describes a set of operations that orients all the unoriented components
- Short path: contains only one white or one grey component
- Long path: contains two or more white or grey components



General HP Distance Formula

Cost of a cover is the sum of the costs of its paths:

1) Cost of a short path: 1
2) Cost of a long path with two grey components: 1
3) Cost of all other long paths: 2



Theorem 3

If t is the cost of an optimal cover of T', then:

$$d_{HP}(A, B) = d_{DCJ}(A, B) + t.$$

Closed formula for t is given in Theorems 4 and 5 (see paper).

1 Introduction
- Biological Background
- The Double Cut and Join (DCJ) Model

2 Computing the General HP Distance
- Components and Oriented Sorting
- Destroying Unoriented Components
- Unoriented Sorting

3 Summary

Summary and outlook

- Relation between the DCJ and the HP genome rearrangement models
- Components are defined directly in the genome
- Properties of components like inclusion and linkage are represented in a tree
- Simple proof of the HP distance formula
- Linear-time algorithm for the HP distance (in my thesis)

Thank you for your attention!