On Sorting by Translocations

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

[1]LaCIM
Université du Québec à Montréal

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

[3]Technische Fakultät
Universität Bielefeld

RECOMB 2005, Cambridge, MA, May 14–18, 2005

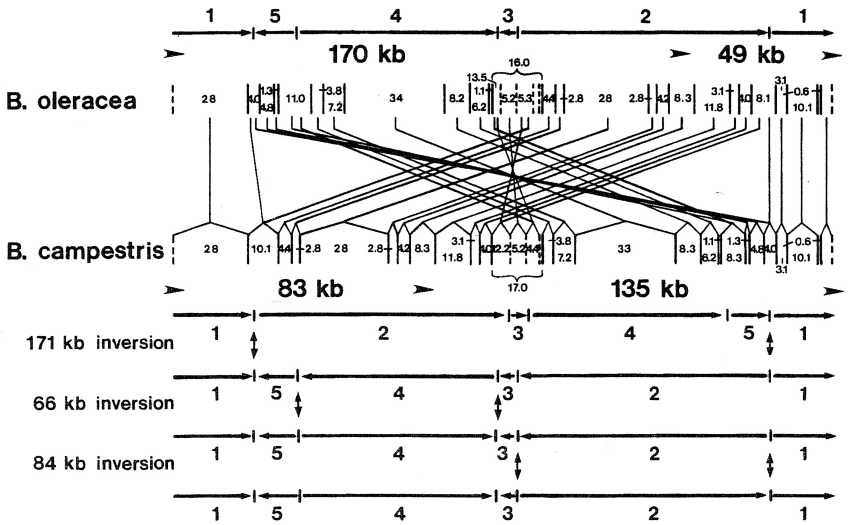
Outline

Outline

Biological background

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

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



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

Biological background

Genomes with the same gene content and number of chromosomes:

- A gene is represented by a signed integer
- A chromosome is a sequence of genes and does not have an orientation, i.e. (6 -8 9) = (-9 8 -6)

$$A_1 = \{(4\ \ 3),\ \ (1\ \ 2\ \ -7\ \ 5),\ \ (6\ \ -8\ \ 9)\}$$

An internal translocation exchanges two *non-empty* chromosome ends:

$$A_1 = \{(4\ \ \underline{3}),\ \ (1\ \ 2\ \ \underline{-7\ \ 5}),\ \ (6\ \ -8\ \ 9)\}$$

$$A_1' = \{(4\ \ -7\ \ 5),\ \ (1\ \ 2\ \ 3),\ \ (6\ \ -8\ \ 9)\}$$

Translocation distance problem

Problem: How many internal translocations do we need to transform one genome into the other?

$A_1 = \{(4\ \ 3), (1\ \ 2\ \ -7\ \ 5), (6\ \ -8\ \ \ 9)\}$

$\{(4\ \ -7\ \ \ 5), (1\ \ 2\ \ \ \ 3), (-9\ \ \ 8\ \ -6)\}$

$\{(4\ \ -7\ \ -6), (1\ \ 2\ \ \ \ 3), (-5\ \ -8\ \ \ 9)\}$

$\{(-9\ \ \ -4), (1\ \ \ 2\ \ \ 3), (-5\ \ -8\ \ -7\ \ -6)\}$

$B_1 = \{(\ 1\ \ \ 2\ \ 3), (\ 4\ \ \ 5), (6\ \ 7\ \ \ 8\ \ \ 9)\}$

Definition

Translocation distance $d(A)$: minimum number of translocations needed to transform A into the identity permutation split in chromosomes sharing the ends of A

Theorem (Hannenhalli, 1996)

For a genome A with N chromosomes and n genes

$$d(A) = n - N - c + s + o + 2i$$

where c is the number of cycles, s the number of minimal subpermutations, $o = 1$ if the number of minimal subpermutations is odd and $o = 0$ otherwise, and $i = 1$ if A has an even-isolation and $i = 0$ otherwise.

Summary of our results

- Let A be a genome with c cycles and whose forest F_A has L leaves and T trees. Then

$$d(A) = n - N - c + t$$

where

$$t = \begin{cases} L + 2 & \text{if } L \text{ is even and } T = 1 \\ L + 1 & \text{if } L \text{ is odd} \\ L & \text{if } L \text{ is even and } T \neq 1. \end{cases}$$

- First correct algorithm for sorting by translocations.

The translocation distance problem

Definition

Concatenation:

- Glue the chromosomes in any order
- Add the elements 0 and $n+1$

$$A_1 = \{(4\ \ 3),\ (1\ \ 2\ \ -7\ \ 5),\ (6\ \ -8\ \ 9)\}$$

$$P_{A_1} = (0\ \ 4\ \ 3\ \ 1\ \ 2\ \ -7\ \ 5\ \ 6\ \ -8\ \ 9\ \ 10)$$

Definition

Inversion:

$$P_{A_1} = (0\ \ 4\ \ 3\ \ 1\ \ 2\ \ -7\ \ 5\ \ 6\ \ -8\ \ 9\ \ 10)$$

$$P'_{A_1} = (0\ \ 4\ \ 3\ \ 1\ \ -5\ \ 7\ \ -2\ \ 6\ \ -8\ \ 9\ \ 10)$$

The translocation distance problem

Sorting by translocations:

$A_1 = \{(4\ 3),\ (1\ 2\ -7\ 5),\ (6\ -8\ 9)\}$

$\{(4\ -7\ 5),\ (1\ 2\ 3),\ (-9\ 8\ -6)\}$

$\{(4\ -7\ -6),\ (1\ 2\ 3),\ (-5\ -8\ 9)\}$

$\{(-9\ -4),\ (1\ 2\ 3),\ (-5\ -8\ -7\ -6)\}$

$B_1 = \{(1\ 2\ 3),\ (4\ 5),\ (6\ 7\ 8\ 9)\}$

Sorting by inversions:

$P_{A_1} = (0\ 4\ 3\ 1\ 2\ -7\ 5\ 6\ -8\ 9\ 10)$

$(0\ 4\ 3\ -5\ 7\ -2\ -1\ 6\ -8\ 9\ 10)$

$(0\ 4\ -7\ 5\ -3\ -2\ -1\ 6\ -8\ 9\ 10)$

$(0\ 4\ -7\ -6\ 1\ 2\ 3\ -5\ -8\ 9\ 10)$

$(0\ 4\ -7\ -6\ 1\ 2\ 3\ -9\ 8\ 5\ 10)$

$(0\ 4\ 9\ -3\ -2\ -1\ 6\ 7\ 8\ 5\ 10)$

$(0\ -9\ -4\ -3\ -2\ -1\ 6\ 7\ 8\ 5\ 10)$

$(0\ -9\ -8\ -7\ -6\ 1\ 2\ 3\ 4\ 5\ 10)$

$(0\ -5\ -4\ -3\ -2\ -1\ 6\ 7\ 8\ 9\ 10)$

$Id = (0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10)$

Outline

Elementary intervals

Definitions

- Signed permutation:

$$P_{A_1} = (0 \;\; 4 \;\; 3 \;\; 1 \;\; 2 \;\; -7 \;\; 5 \;\; 6 \;\; -8 \;\; 9 \;\; 10)$$

- Point: pair of consecutive elements $p \cdot q$
- Adjacency: point of the form $i \cdot i+1$ or $-(i+1) \cdot -i$, otherwise breakpoint

Definition

Black points are points inside chromosomes, all others points are white points

$$A_1 = \{(4 \;\; 3), \;\; (1 \;\; 2 \;\; -7 \;\; 5), \;\; (6 \;\; -8 \;\; 9)\}$$

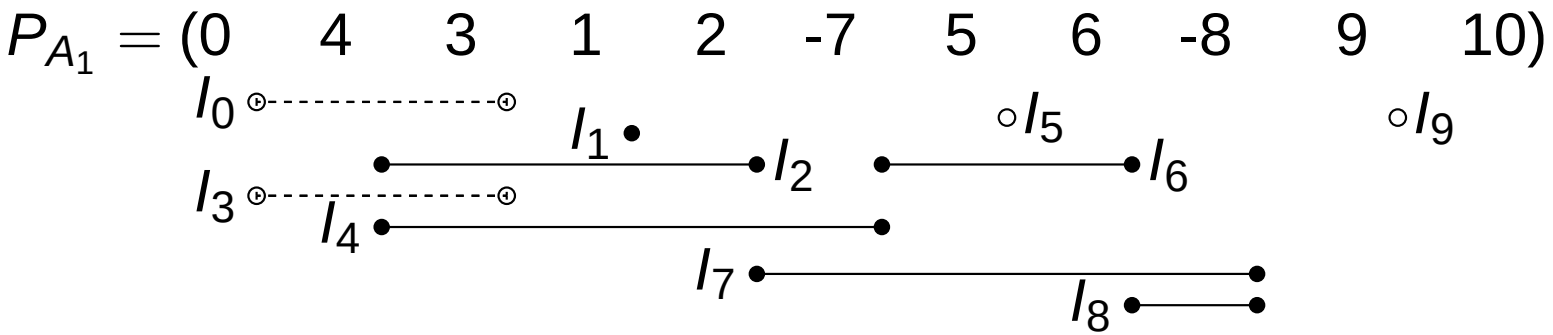
$$P_{A_1} = (0 \;\; 4 \;\; 3 \;\; 1 \;\; 2 \;\; -7 \;\; 5 \;\; 6 \;\; -8 \;\; 9 \;\; 10)$$

Elementary intervals

Definition

The elementary interval I_k is the interval whose endpoints are:

1) the right point of k, if k is positive, otherwise its left point
2) the left point of $k+1$, if $k+1$ is positive, otherwise its right point.

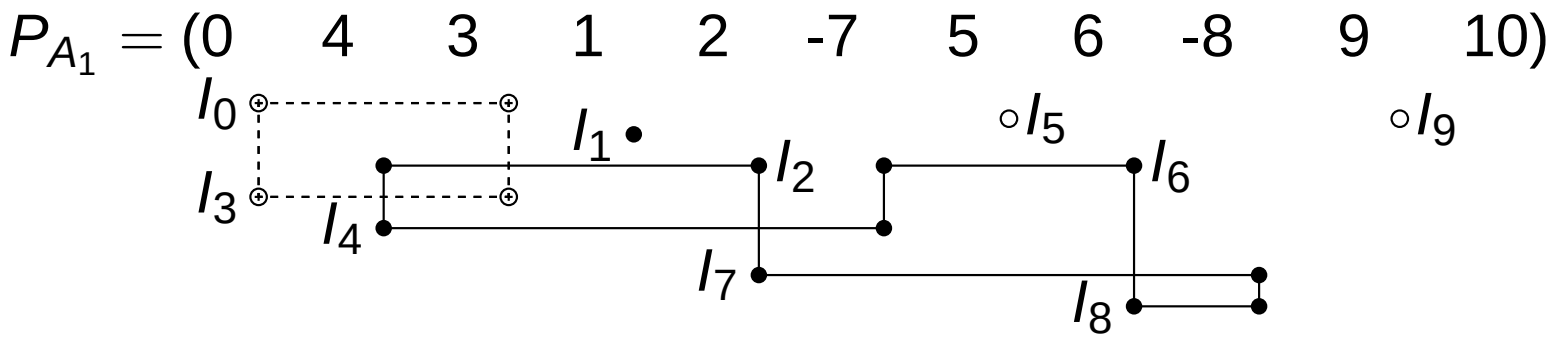


Observation: Exactly two elementary intervals of the same color meet at each breakpoint.

Cycles

Definition

- Black (white) cycle: sequence of black (white) points such that two successive points are the endpoints of an elementary interval
- Adjacencies define trivial cycles



Observation: The number of black cycles equals $n - N$, if and only if genome A is sorted.

Cycles

Lemma 1 (Kececioglu and Ravi, 1995)

A translocation in genome A modifies the number of black cycles of P_A by 1, 0 or -1.

A translocation is called ...

- proper, if it increases the number of black cycles by 1
- improper, if it leaves the number of black cycles unchanged
- bad, if it decreases the number of black cycles by 1.

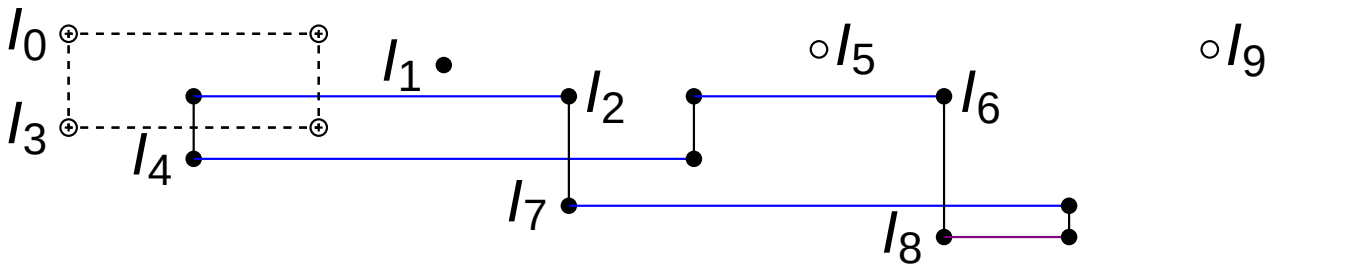
Cycles

Definition

Interchromosomal elementary interval: whose endpoints belong to different chromosomes, otherwise intrachromosomal

$$A_1 = \{(4\ \ 3),\ (1\ \ 2\ \ -7\ \ 5),\ (6\ \ -8\ \ 9)\}$$

$$P_{A_1} = (0\ \ 4\ \ 3\ \ 1\ \ 2\ \ -7\ \ 5\ \ 6\ \ -8\ \ 9\ \ 10)$$



Lemma 2

For each interchromosomal elementary interval in P_A, there exists a proper translocation in the genome A.

Components

Definitions

- Component: an interval from i to $i+j$ or from $-(i+j)$ to $-i$, where $j > 0$, whose set of elements is $\{i, \ldots, i+j\}$, and that is not the union of smaller such intervals
- Intrachromosomal component: its elements belong to the same chromosome

$A_2 =$ {(1 -2 3 8 4 -5 6), (7 9 -10 11 -12 13 14 -15 16)}

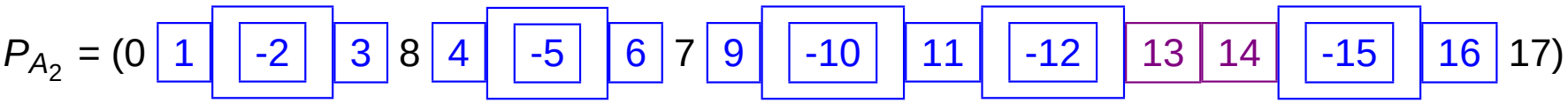
Intrachromosomal components:

P_{A_2} = (0 1 [-2] 3 8 4 [-5] 6 7 9 [-10] 11 [-12] 13 14 [-15] 16 17)

Components

An intrachromosomal component is called ...

- minimal, if it does not contain any other intrachromosomal component
- trivial, if it is an adjacency, otherwise non-trivial



Chain: successive linked intrachromosomal components

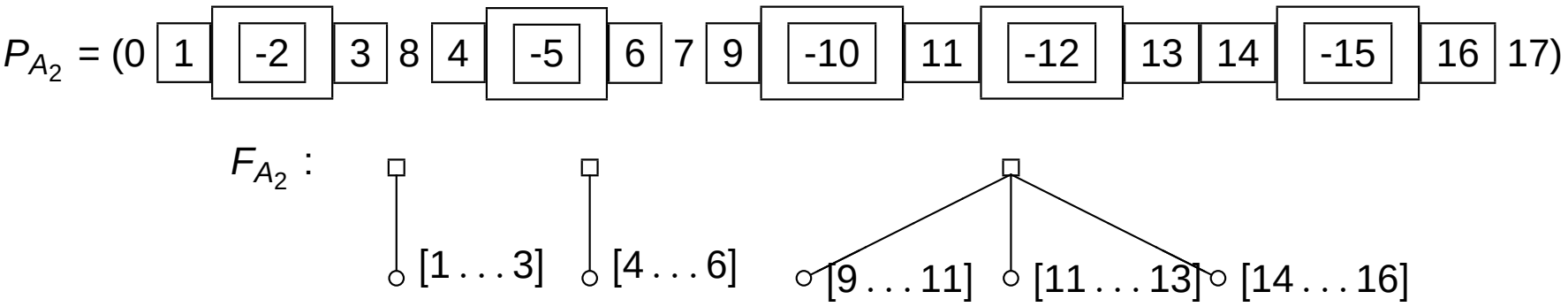
Maximal chain: cannot be extended to the left or right

Definitions

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

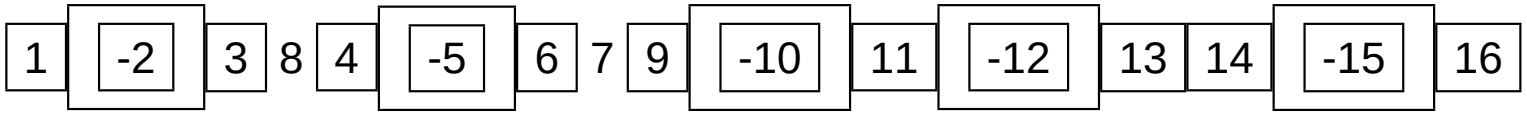
1) Each non-trivial intrachromosomal component is a round node
2) Each maximal chain is a square node whose (ordered) children are the round nodes
3) A square node is the child of the smallest intrachromosomal component that contains this chain

The forest F_A of a genome A is the set of forests $\{F_{X_1}, \ldots, F_{X_N}\}$.

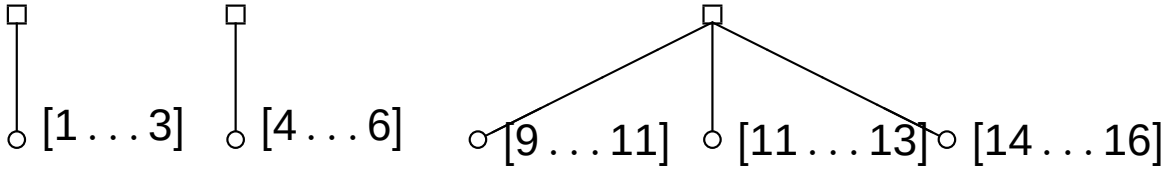


Components

$A_2 =$ {(1 -2 3 8 4 -5 6), (7 9 -10 11 -12 13 14 -15 16)}

P_{A_2} = (0  17)

F_{A_2} :



How can we destroy intrachromosomal components?

- Apply a translocation with one endpoint in the component, and one endpoint in another chromosome
- Such translocations are always bad
- A translocation destroying component C destroys all components that contain C
- A translocation can destroy at most 2 minimal components (leaves)

Let F_A be a forest with L leaves and T trees.

If L even and $T > 1$, destroy the forest optimally:

- Separate the forest (Lemma 3)
- Destroy two leaves (Lemma 4)

Lemma 3

If $T > 1$ and all trees belong to the same chromosome, then the trees can be separated by proper translocations without modifying F_A.

Lemma 4

If L is even and $T > 1$, then there always exists a sequence of proper translocations, followed by a bad translocation, such that the resulting genome A' has $L' = L - 2$ leaves and $T' \neq 1$ trees.

Outline

The distance formula

Theorem 1

Let A be a genome with c black cycles and F_A be the forest associated to A. Then

$$d(A) = n - N - c + t$$

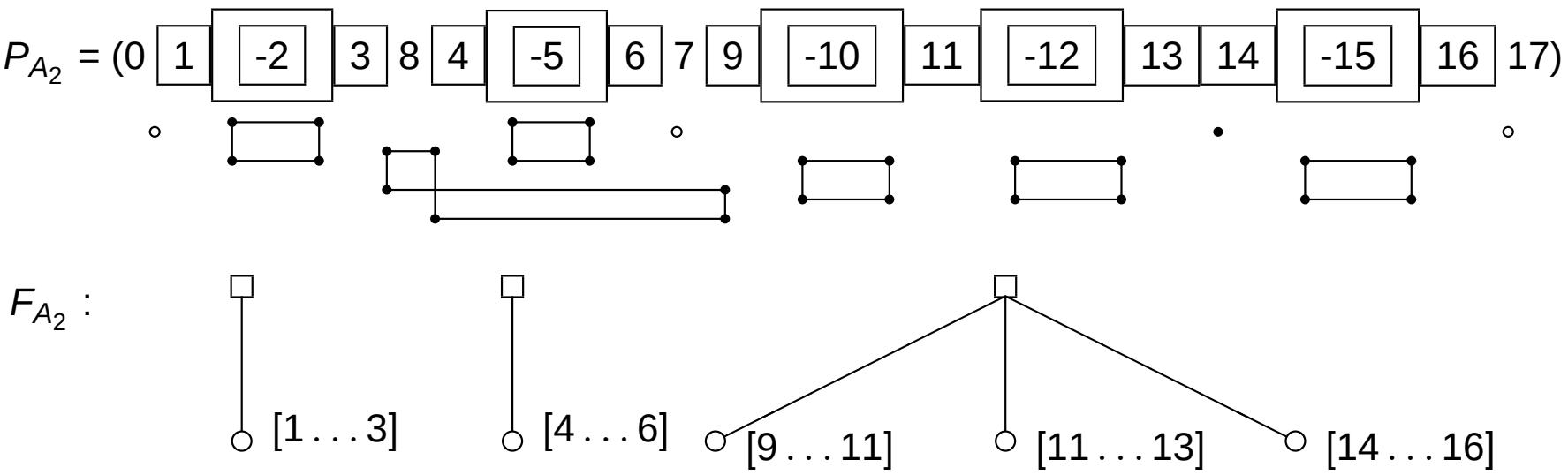
where

$$t = \begin{cases} L + 2 & \text{if } L \text{ is even and } T = 1 \\ L + 1 & \text{if } L \text{ is odd} \\ L & \text{if } L \text{ is even and } T \neq 1. \end{cases}$$

Algorithms

Algorithm for computing the translocation distance

- Cycle identification: by a left-to-right scan of the permutation
- Component identification: by a linear-time algorithm (Bergeron, Heber and Stoye, 2002)
- Construction of F_A: by a simple pass over the components (Bergeron, Mixtacki and Stoye, 2004)



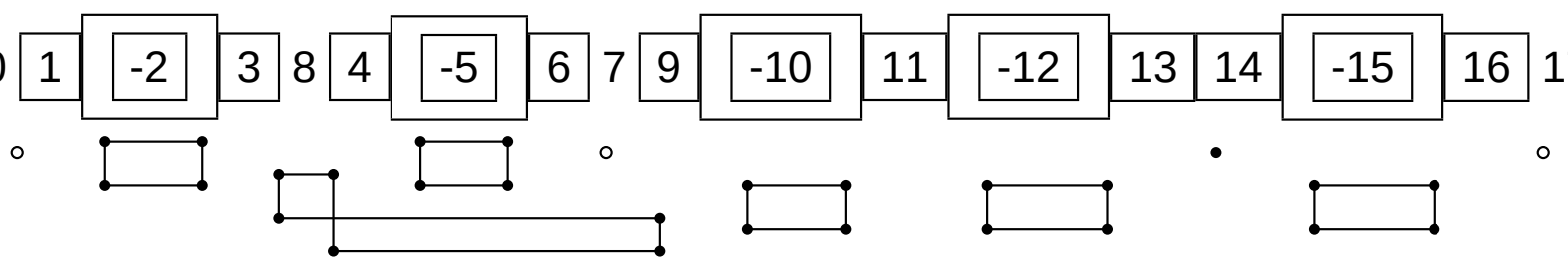
Algorithms

Theorem 2

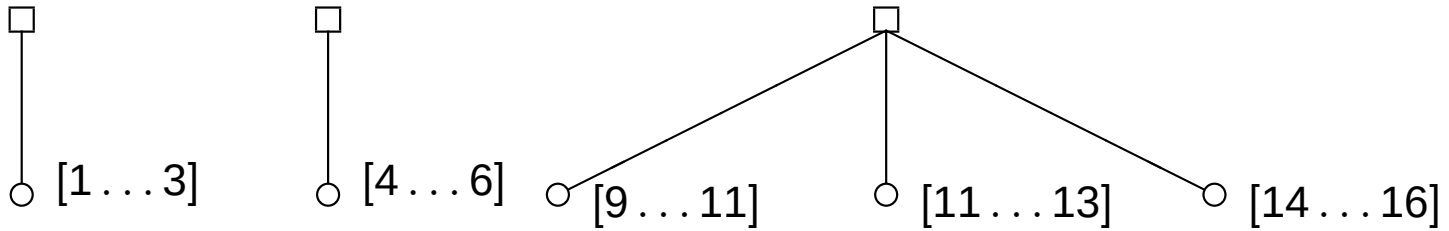
The translocation distance can be computed in linear time.

$A_2 =$ {(1 -2 3 8 4 -5 6), (7 9 -10 11 -12 13 14 -15 16)}

P_{A_2} = (0 1 -2 3 8 4 -5 6 7 9 -10 11 -12 13 14 -15 16 17)



F_{A_2} :



$$d(A_2) = n - N - c + t = 16 - 2 - 7 + 6 = 13$$

Algorithms

Algorithm for sorting by translocations

1: L is the number of leaves, and T the number of trees in the forest F_A associated to the genome A
2: **if** L is even **and** $T = 1$ **then**
3: destroy one leaf such that $L' = L - 1$
4: **end if**
5: **if** L is odd **then**
6: perform a bad translocation such that $T' = 0$, or $T' > 1$ and $L' = L - 1$
7: **end if**
8: **while** A is not sorted **do**
9: **if** there exist intrachromosomal components on different chromosomes **then**
10: perform a bad translocation such that $T' = 0$, or $T' > 1$ and L' is even
11: **else**
12: perform a proper translocation such that T and L remain unchanged
13: **end if**
14: **end while**

Summary

- Modified concepts from the uni-chromosomal case are applied to multi-chromosomal genomes
- New formula for the translocation distance
- Linear-time algorithm to compute the translocation distance
- First correct sorting by translocation algorithm