Reversal distance without hurdles and fortresses

Julia Mixtacki

Faculty of Mathematics
University of Bielefeld, Germany

July 7th, 2004

Joint work with:
Anne Bergeron, LaCIM, Université du Québec à Montréal, Canada
Jens Stoye, Technische Fakultät, Universität Bielefeld, Germany

1. Introduction and biological background
2. Definitions and examples
    - Signed permutations and reversal distance
    - Elementary intervals and cycles
    - Components
3. Computing the reversal distance
    - Sorting oriented components
    - Orienting unoriented components
    - The distance formula
    - Algorithms
4. Summary and outlook

Introduction and biological background

Consider two genomes with the same gene content.

Represent each gene by a signed integer between 0 and n.
The sign represents the orientation of a gene.

$$P = \quad (0 \quad \text{-2} \quad \text{-1} \quad 4 \quad 3 \quad 5 \quad \text{-8} \quad 6 \quad 7 \quad 9)$$

A reversal changes the order and the signs of an interval of genes.

$$\begin{array}{lllllllllll} P = & (0 & \text{-2} & \text{-1} & 4 & 3 & 5 & \text{-8} & 6 & 7 & 9) \\ \cline{4-7} P' = & (0 & \text{-2} & \text{-5} & \text{-3} & \text{-4} & 1 & \text{-8} & 6 & 7 & 9) \end{array}$$

Introduction and biological background

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

$$(0 \quad \underline{-2 \quad -1 \quad 4} \quad 3 \quad 5 \quad -8 \quad 6 \quad 7 \quad 9)$$

$$(0 \quad \underline{-4 \quad 1 \quad 2 \quad 3} \quad 5 \quad -8 \quad 6 \quad 7 \quad 9)$$

$$(0 \quad \underline{-3 \quad -2 \quad -1} \quad 4 \quad 5 \quad -8 \quad 6 \quad 7 \quad 9)$$

$$(0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad \underline{-8 \quad 6 \quad 7} \quad 9)$$

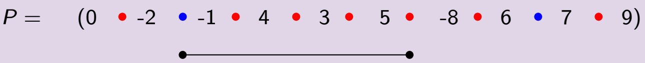
$$(0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad \underline{-7 \quad -6} \quad 8 \quad 9)$$

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

Signed permutations

Definitions

- Signed permutation:



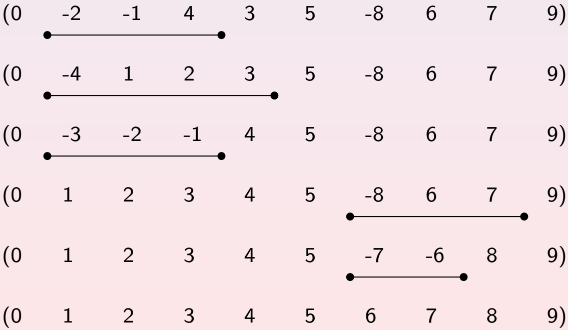
$P = (0 \bullet -2 \bullet -1 \bullet 4 \bullet 3 \bullet 5 \bullet -8 \bullet 6 \bullet 7 \bullet 9)$

- Point: pair of consecutive elements $p \bullet q$
- Adjacency: point of the form $i \bullet i+1$ or $-(i+1) \bullet -i$, otherwise breakpoint
- Interval: defined by its two endpoints

Reversal distance

Definition

Reversal distance $d(P)$: minimum number of reversals needed to transform P into the identity permutation.



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

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

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

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

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

(0 1 2 3 4 5 6 7 8 9)

Theorem (Hannenhalli and Pevzner, 1995)

For a signed permutation P

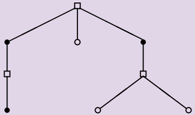
$$d(P) = n - c + h + f$$

where c is the number of cycles, h the number of hurdles, and $f = 1$ if P has a fortress, and $f = 0$ otherwise.

Summary of our results

- If a signed permutation P on the set $\{0, \ldots, n\}$ has c cycles and the associated tree T_P has minimal cost t, then

$$d(P) = n - c + t$$



- Yields a simple linear-time algorithm to compute the reversal distance.

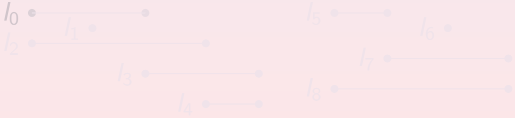
Elementary intervals

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.

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



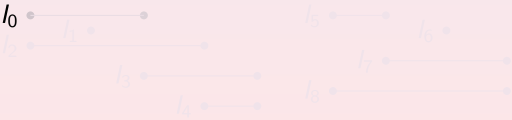
Elementary intervals

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.

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

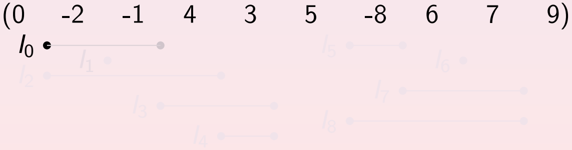


Elementary intervals

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.

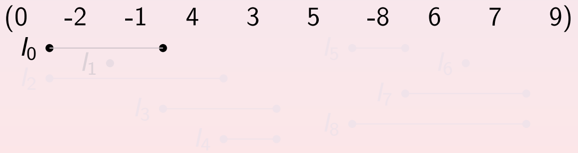


Elementary intervals

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.



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

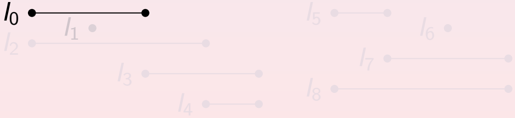
Elementary intervals

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.

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

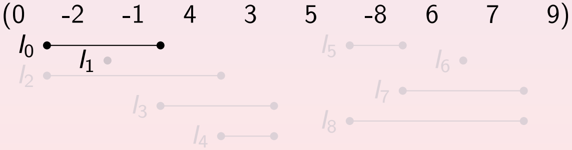


Elementary intervals

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.

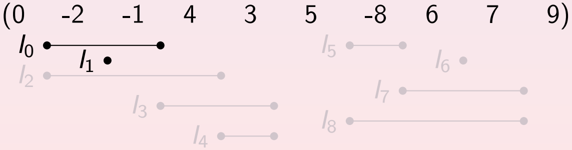


Elementary intervals

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.

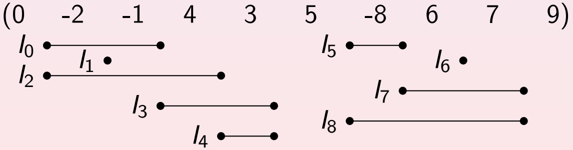


Elementary intervals

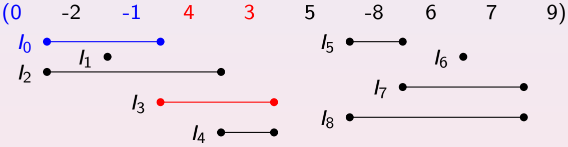
Definition

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.



Elementary intervals



- I_k is oriented if elements k and $k+1$ have different signs, otherwise unoriented

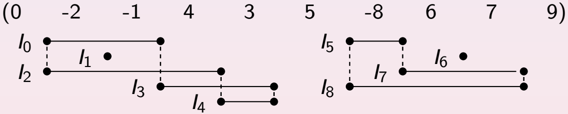
Proposition

Reversing an oriented interval I_k creates either the adjacency $k \bullet (k+1)$ or the adjacency $-(k+1) \bullet -k$.

Cycles

Proposition

Exactly two elementary intervals meet at each breakpoint of a permutation.



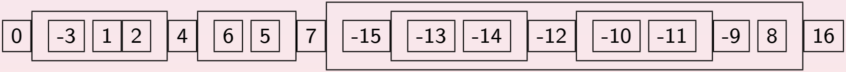
Definition

- Cycle: sequence of points such that two successive points are the endpoints of an elementary interval
- Adjacencies define trivial cycles

Components

Definition

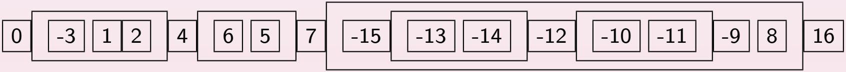
A component is an interval from i to $(i+j)$ or from $-(i+j)$ to $-i$, for some $j>0$, whose set of unsigned elements is $\{i,\ldots,i+j\}$, and that is not the union of two such intervals.



Components

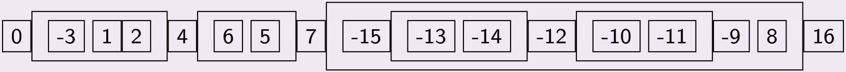
Proposition

Two different components of a permutation are either disjoint, nested with different endpoints, or overlapping on one element.



0 -3 1 2 4 6 5 7 -15 -13 -14 -12 -10 -11 -9 8 16

- Chain: successive linked components
- Maximal chain: cannot be extended to the left or right



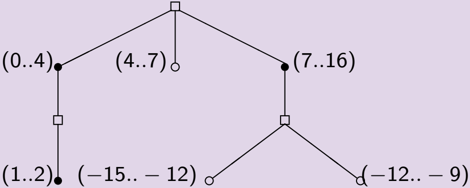
Definition

The tree T_P is defined by the following construction:

1) Each component is a round node.

2) Each maximal chain is a square node whose (ordered) children are round nodes.

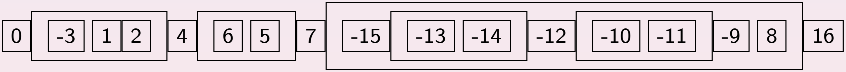
3) A square node is the child of the smallest component that contains this chain.



Components

Definition

A point $p \bullet q$ belongs to the smallest component that contains both p and q.



0 -3 1 2 4 6 5 7 -15 -13 -14 -12 -10 -11 -9 8 16

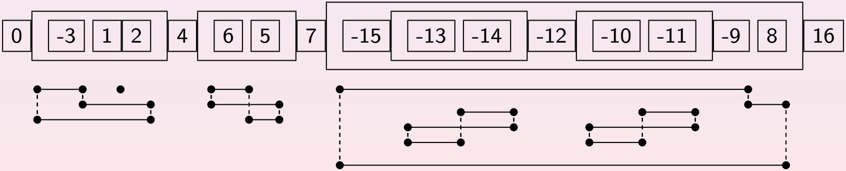
Definition

The sign of a point $p \bullet q$ is positive if both p and q are positive, it is negative if both are negative. A component is unoriented if it has one or more breakpoints and all of them have the same sign. Otherwise the component is oriented.

Components

Proposition

The endpoints of an elementary interval belong to the same component, thus all the points of a cycle belong to the same component.



- An oriented component contains at least two oriented elementary intervals
- All elementary intervals of an unoriented component are unoriented

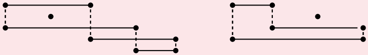
Sorting oriented components

Theorem (Hannenhalli and Pevzner, 1995)

If a permutation P on the set $\{0, \ldots, n\}$ has no unoriented components and c cycles, then

$$d(P) = n - c.$$

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

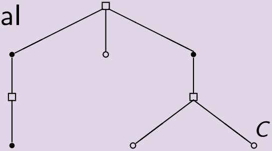


$d(P) = 9 - 4 = 5$

Orienting unoriented components

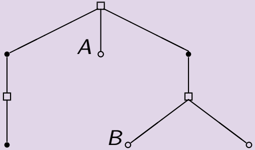
Proposition (Hurdle Cutting)

If a component C is unoriented, the reversal of an elementary interval whose endpoints belong to C, orients C and leaves the number of cycles unchanged.



Proposition (Hurdle Merging)

A reversal that has its two endpoints in different components A and B destroys, or orients, all components on the path from A to B in T_P, without creating new unoriented components.

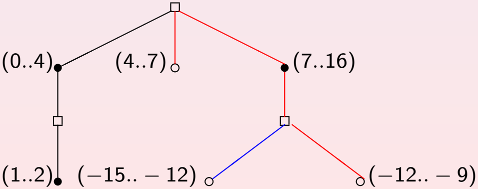


Orienting unoriented components

Definition

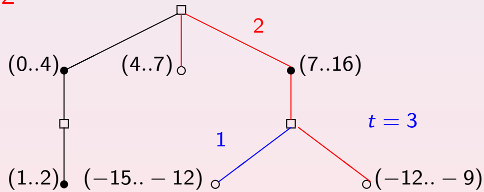
A cover $\mathcal{C}$ of T_P is a collection of paths joining all the unoriented components of P, such that each terminal node of a path belongs to a unique path.

- Each cover of T_P describes a set of reversals that orients all the components of P
- Short path: contains only one component
- Long path: contains two or more unoriented components



Orienting unoriented components

- 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: 2



- t = cost of any optimal cover

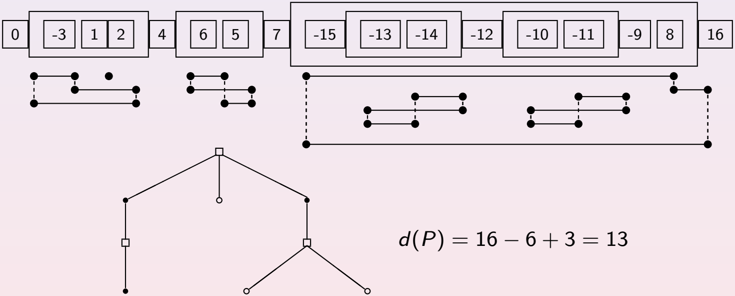
The distance formula

Theorem

If a permutation P on the set $\{0, \ldots, n\}$ has c cycles, and the associated tree T_P has minimal cost t, then

$$d(P) = n - c + t.$$

Algorithms



$$d(P) = 16 - 6 + 3 = 13$$

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 T_P: by a simple pass over the components

Summary and outlook

- Intervals and components are defined directly in the permutation
- Properties of components like inclusion and linkage are represented in a tree
- Simple proof of the reversal distance formula
- Linear-time algorithm to compute the reversal distance
- Next step: application to multi-chromosomal rearrangement problems