Algorithms for Genome Rearrangements

Pedro Feijão

Lecture 3 – Sorting by Signed Reversals

Summer 2015

pfeijao@cebitec.uni-bielefeld.de

Definitions

- A **signed permutation** is a permutation on the set $\{0, 1, \ldots, n\}$ in which every element has a *sign*. To simplify, permutations will always start with 0 and end with n. *For example*:

$$\pi_1 = (0 \quad -2 \quad -1 \quad 4 \quad 3 \quad 5 \quad -8 \quad 6 \quad 7 \quad 9)$$

- A **point** $p \cdot q$ is a pair of consecutive elements in the permutation. In the above example, $0 \cdot -2$ and $-2 \cdot -1$ are the first two points of π_1.
- When a point is in the form $i \cdot (i+1)$ or $-(i+1) \cdot -i$ it is called an **(conserved) adjacency**. Otherwise, it is a **breakpoint**.

Breakpoints

$$\pi_1 = (0 \quad -2 \quad -1 \quad 4 \quad 3 \quad 5 \quad -8 \quad 6 \quad 7 \quad 9)$$

- In this permutation, there are *two* adjacencies, $-2 \cdot -1$ and $6 \cdot 7$, and *seven* breakpoints.
- The **Breakpoint Distance** is the number of breakpoints in a permutation, that is, distance from the **identity**:

$$\mathrm{Id} = (0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9)$$

- It is one the simplest measure of dissimilarity for genome rearrangements. *Notation*: $d_{\mathsf{BP}}(\pi_1) = 7$.

For instance, the permutation

$$\pi_2 = (0 \quad -4 \quad -3 \quad -2 \quad -1 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9)$$

has 2 breakpoints, which means that π_2 is *closer* to the identity than π_1.

Reversals

- An **reversal** of a permutation interval reverts the *order* and *sign* of all elements of the interval.

$$\pi_1 = (0 \quad -2 \quad \underline{-1 \quad 4 \quad 3 \quad 5} \quad -8 \quad 6 \quad 7 \quad 9)$$

$$\pi_1' = (0 \quad -2 \quad -5 \quad -3 \quad -4 \quad 1 \quad -8 \quad 6 \quad 7 \quad 9)$$

- The **reversal distance** is the minimum number of reversals needed to transform one permutation into another (usually the other permutation is the identity). Notation: $d_R(\pi_1)$.
- Finding such a scenario of reversals is called **sorting by reversals**.
  - *Distance* vs. *Sorting*

BP vs. Reversals

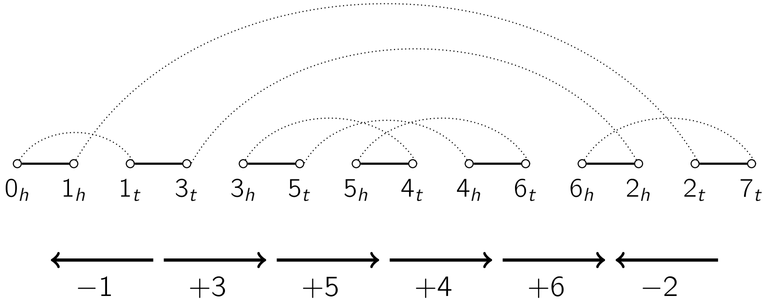
- A reversal changes the number of breakpoints by at most 2.
- This gives a simple *lower bound* for the reversal distance:

$$d_R(\pi_1) \geq \frac{d_{\mathrm{BP}}(\pi_1)}{2}$$

- Using BP for lower bound is an useful *first approach* in many models.

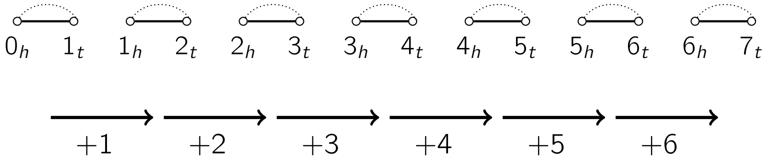
Breakpoint Graph - Genomes as Graphs

- The BP graph of a is a very useful structure for studying rearrangement problems. Notation $BP(\pi)$.
- **Vertices** are the gene extremities (tail and head).
- **Black edges** between consecutive gene extremities (reality edges).
- **Grey edges** between consecutive gene extremities of the identity (desire edges).



Breakpoint Graph

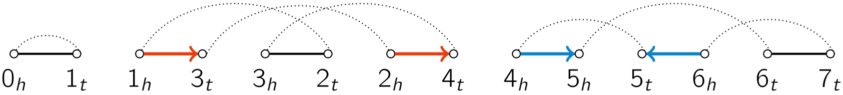
- When the input genome is the identity, the BP graph is composed of n **trivial cycles**.



- Sorting is equivalent to **increasing the cycles of the BP graph**.
- What happens in the BP graph when a reversal is applied?

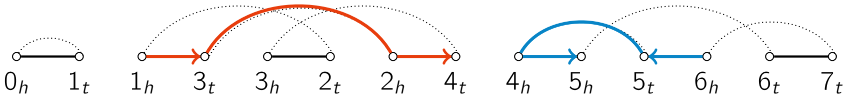
BP Graph Elements

- Two black edges in they same cycle are **convergent** if, when traversing the cycle both edges induce the *same direction*. Otherwise, they are **divergent**.



BP Graph Elements

- A grey edge is **oriented** if its two incident black edges are *divergent*, otherwise the edge is **unoriented**.



- Equivalently, a grey edge is **oriented** if it “contains” an odd number of vertices, and **unoriented** otherwise (even number of vertices).

BP Graph Elements

- A cycle is **oriented** if it contains *at least one* oriented edge. Otherwise, it is **unoriented**.

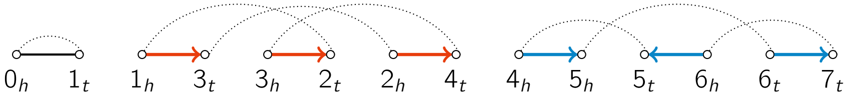
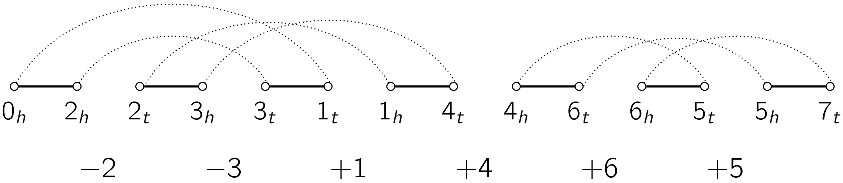


Figure: Example of unoriented and oriented cycles.

BP Graph Components

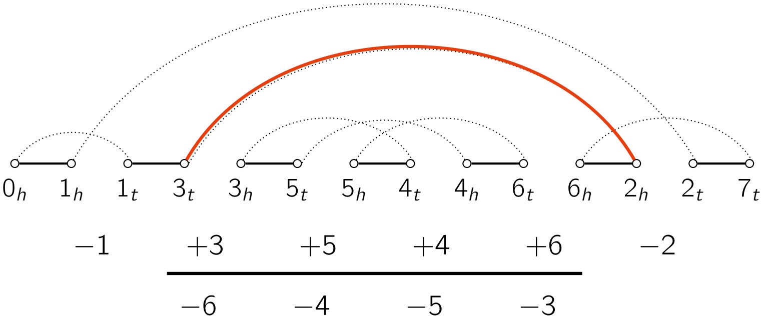
- Two cycles are **connected** if they have overlapping edges.
- A **component** is a subset of connected cycles.



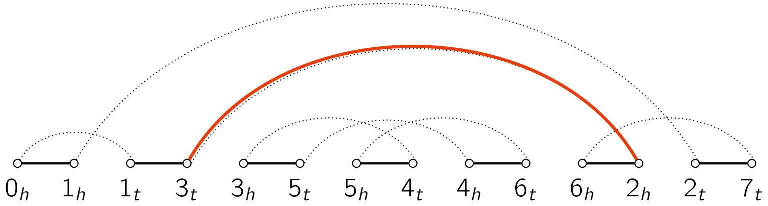
- An **oriented component** has at least one oriented cycle, otherwise it is a **unoriented component**.

Inducing Reversals

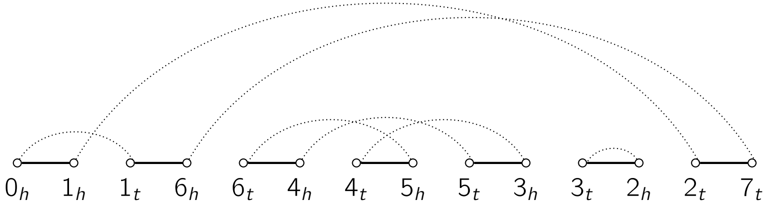
- A reversal **induced** by a grey edge (equivalenty, by two black edges) reverses the elements that are *completely* contained in the edge.



0_h 1_h 1_t 3_t 3_h 5_t 5_h 4_t 4_h 6_t 6_h 2_h 2_t 7_t
-1 $+3$ $+5$ $+4$ $+6$ -2
-6 -4 -5 -3



$$-1 \quad \underline{+3 \quad +5 \quad +4 \quad +6} \quad -2$$



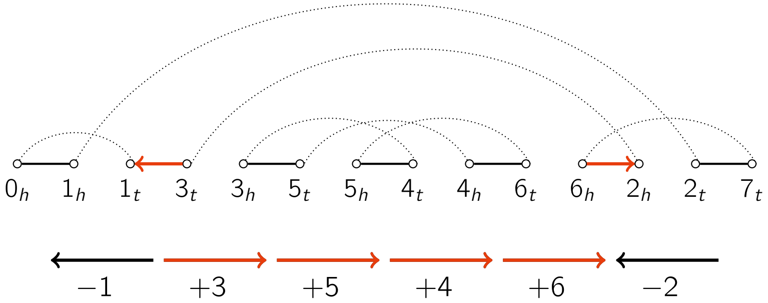
$$-1 \quad -6 \quad -4 \quad -5 \quad -3 \quad -2$$

Reversals and effect on cycles

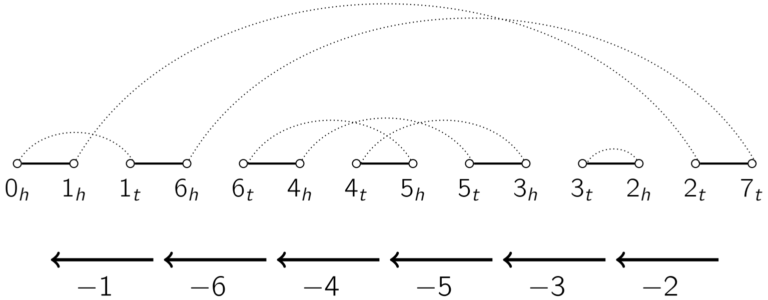
1. Black Edges are on the **same cycle**:
   - **Type I**: Divergent edges: breaks the cycle. $\Delta C = +1$.
   - **Type II**: Convergent edges: $\Delta C = 0$, may change cycle orientation.
2. Black Edges on **different cycles**:
   - **Type III**: Merges the two cycles. $\Delta C = -1$.

So far, we only used **Type I** operations, to sort oriented components.

Type I - Same Cycle, divergent

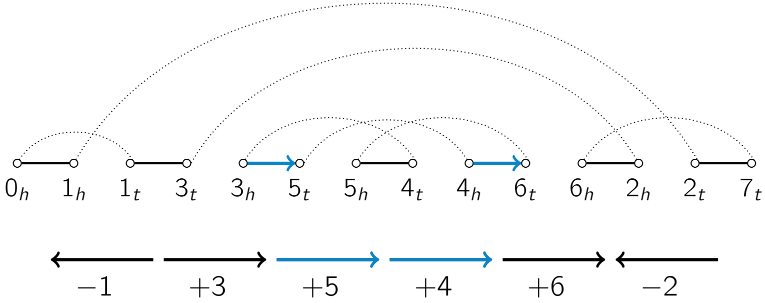


Type I - Same Cycle, divergent

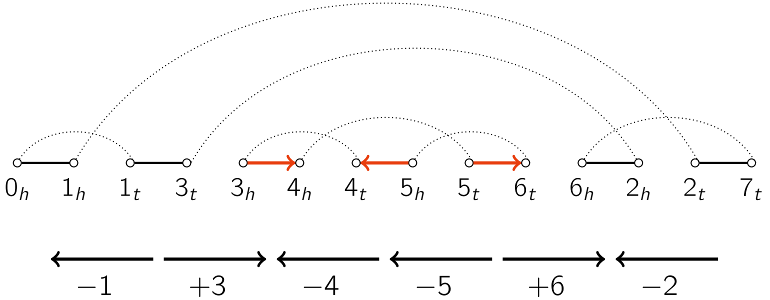


This reversal increases the number of cycles by one, $\Delta C = +1$.

Type II - Same Cycle, convergent

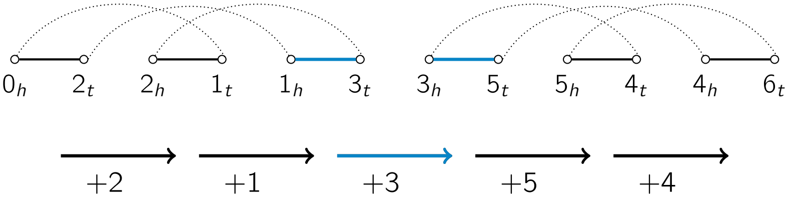


Type II - Same Cycle, convergent

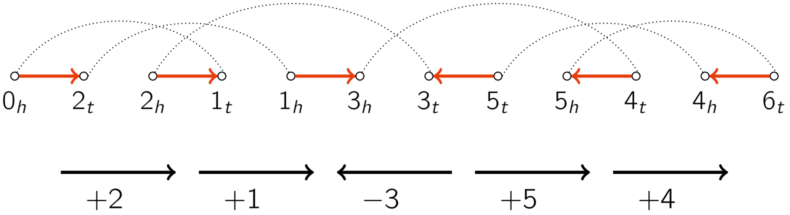


Does not change number of cycles ($\Delta C = 0$), but the cycle is **oriented**.

Type III - Different Cycles



Type III - Different Cycles



Merges the two cycles, decreasing the number of cycles by one ($\Delta C = -1$), but the new cycle is **oriented**.

Breakpoint Graph - Lower Bound

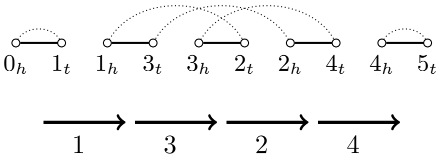
- A reversal changes the number of cycles of the BP graph at most by 1.
- Then, we have a **lower bound** for the reversal distance:

  $$d_R(\pi) \geq N - C$$

  where C is the *number of cycles* in the BP graph of π.
- This bound is usually **tight**, that is, most of the times it is exactly the reversal distance.
- When is this bound not *exactly* the distance?
  - When it is not possible to increase the cycles of BP with a reversal.
  - That occurs in the presence of **unoriented components**.

Unoriented components

- In the example below, there is no reversal that increases the number of cycles.

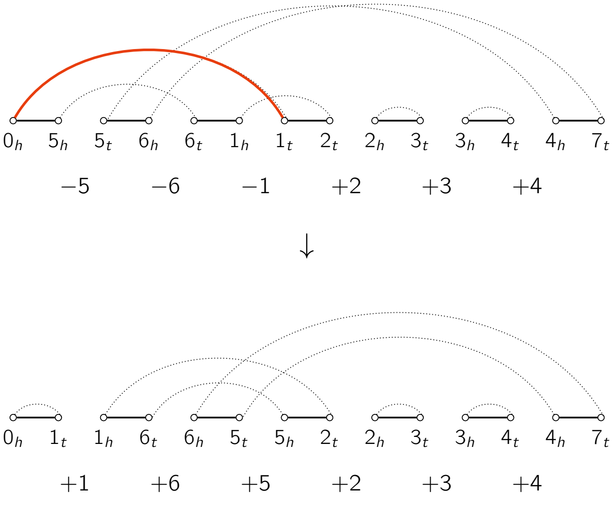


- The lower bound is $N - C = 5 - 3 = 2$, but the real distance is 3, because one extra reversal is needed to *orient* the unoriented cycle in the BP graph.
- Let's first consider the *good* cases, without unoriented components.

Sorting oriented components

- If there are only oriented components, there is always a reversal that increases the number of cycles.
- The problem is, after such a reversal, it is possible the some components become **unoriented**.

Bad reversal - Example



- Increased number of cycles but created a bad component!

Finding "good" reversals

- Is it possible to find a reversal that increases the number of cycles **AND** also does not create an unoriented component? **YES!**

Sorting oriented components

Theorem (Hannenhalli-Pevzer, 95)

If the graph $BP(\pi)$ has only **oriented components***, then*

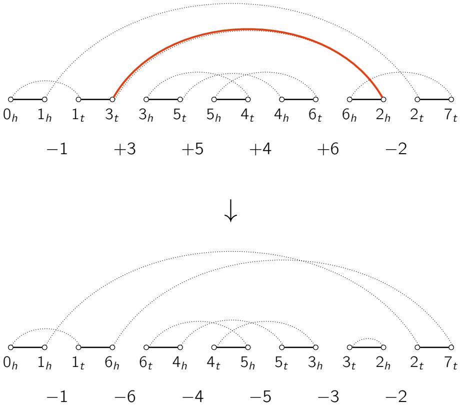
$$d_R(\pi) = N - C$$

where N is the number of elements of π and C is the number of cycles of $BP(\pi)$.

- This means that there is always at least one “good” reversal, that increases the number of cycles of $BP(\pi)$ and *does not create any unoriented component*.
- These are called **safe reversals**. How can we find them?

Safe reversals - Definitions

- The **score** of a reversal is the number of *oriented edges* in the BP graph, *after* the application of the reversal.


0_h 1_h 1_t 3_t 3_h 5_t 5_h 4_t 4_h 6_t 6_h 2_h 2_t 7_t
-1 $+3$ $+5$ $+4$ $+6$ -2
↓
0_h 1_h 1_t 6_h 6_t 4_h 4_t 5_h 5_t 3_h 3_t 2_h 2_t 7_t
-1 -6 -4 -5 -3 -2

The score of this reversal is **two**.

Safe reversals

- **Safe reversals** are reversals that increase the number of cycles of the BP graph by one and do not create new unoriented components.
- Can we always find safe reversals? Yes:

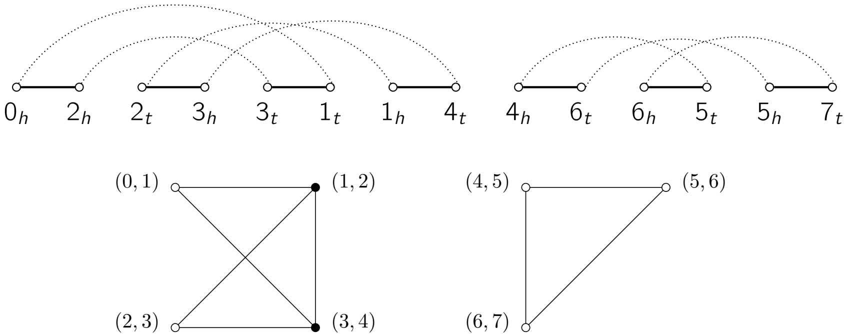
Theorem (Bergeron, 2001)

Among all possible oriented reversals, a reversal of maximal score is always safe.

- **Algorithm**: Apply maximal score reversals until all components are sorted.

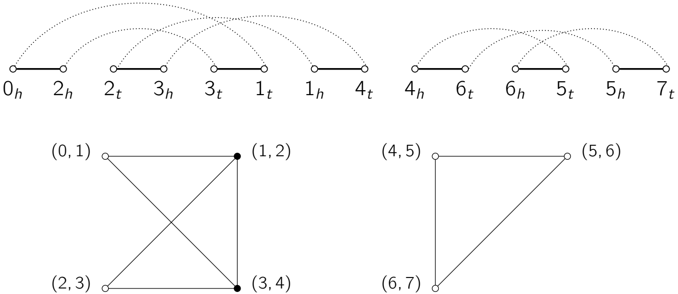
Finding safe reversals with the Overlap Graph

- The **overlap graph** $O(\pi)$ is a graph where:
  - Vertices are the grey edges of $BP(\pi)$. If the edge is oriented, the vertex is black, otherwise is white.
  - When two grey edges overlap, there is an edge between the corresponding vertices.



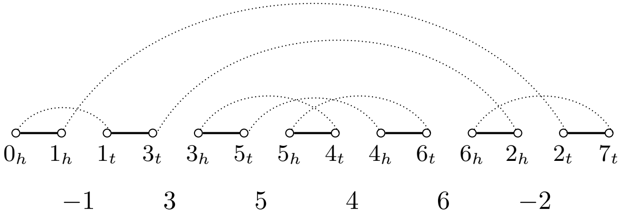
BP Graph vs Overlap Graph

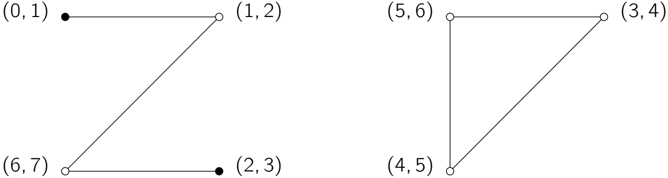
BP Graph	Overlap Graph
Component	Connected component
Oriented edge	Black vertex, *odd degree*
Unoriented edge	White vertex, *even degree*
Oriented component	Component with at least 1 black vertex
Unoriented component	Component with only white vertices



Another Example

$$\pi = [\,-1\ \ 3\ \ 5\ \ 4\ \ 6\ \ -2\,]$$

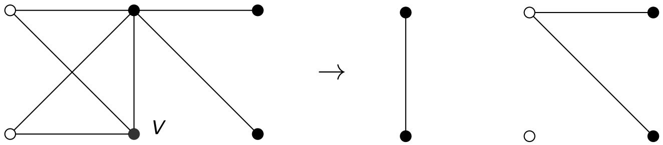




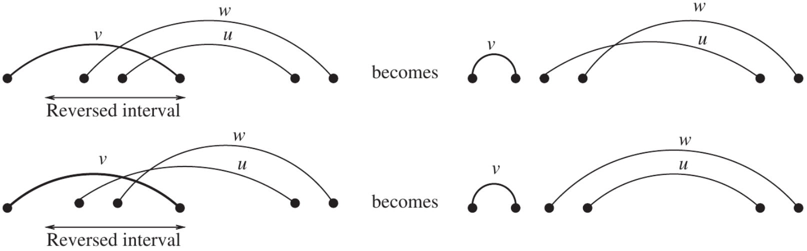
Effect of Reversal in the Overlap Graph

- A reversal *induced by a vertex v* is the reversal that is induced by the corresponding grey edge in the breakpoint graph.
- What happens in $O(\pi)$ after applying an oriented reversal in a vertex v?

1. The subgraph induced by v and its neighbours is **complemented**.



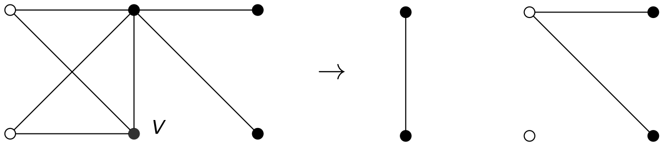
Why?



A. Bergeron / Discrete Applied Mathematics 146 (2005) 134–145

Effect of Reversal in the Overlap Graph

2. All neighbours of *v* have their orientation inverted.



Why?

Reversal Score with $O(\pi)$

We know how the overlap graph changes with a reversal, then it is possible to find an equation for the reversal score of any vertex v:

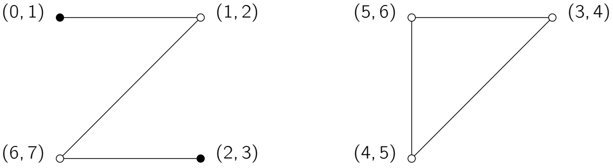
Definition (Reversal score)

The score of a reversal induced by a vertex v in the overlap graph is given by

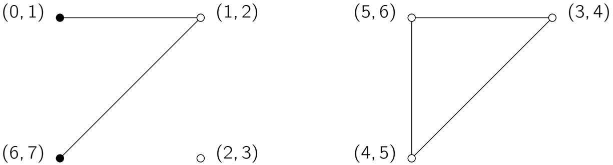
$$s(v) = T + U - O - 1$$

where T is the number of oriented vertices in the graph, U and O are the number of unoriented and oriented vertices adjacent to v, respectively.

Reversal Score - example



For $v = (2, 3)$, we have $T = 2$, $U = 1$, $O = 0$. Therefore $s(v) = T + U - O - 1 = 2$.
After applying the reversal, we have the following graph:



and we see that the score (number of oriented vertices) is indeed 2.

Sorting Example

$$\pi = (0 \quad 3 \quad 1 \quad 6 \quad 5 \quad -2 \quad 4 \quad 7)$$