Algorithms for Genome Rearrangements

Pedro Feijão

Lecture 2 – Sorting by Reversals II

Summer 2015

pfeijao@cebitec.uni-bielefeld.de

Sorting by Unsigned Reversals

Last Lecture:

- First algorithm: Bring each element in position. $\dfrac{n-1}{2}$ approximation, not good.

- Better idea: fix *breakpoints*, preserving adjacencies. We found a $4-$ and $2-$approximation.

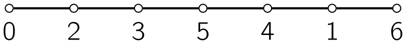
Is there a way to improve this?

Permutations as Graphs

The **graph** G_π of a permutation π is a graph where:

- The vertices are the elements of π (adding 0 and $n+1$).
- Consecutive elements are connected by an edge.

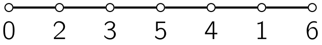
For example, for $\pi = (\ 2\ \ 3\ \ 5\ \ 4\ \ 1\ )$, we have G_π:

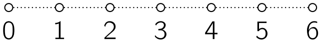


Comparing Permutation Graphs

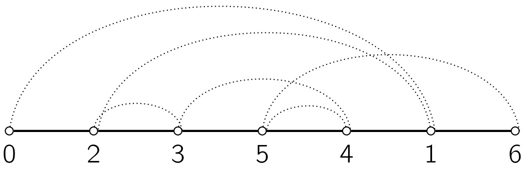
Given two permutations, we can combine their graphs, *reusing the same vertices*, with different edge colors.

For instance, for $\pi = (\ 2\ \ 3\ \ 5\ \ 4\ \ 1\ )$ and $\sigma = (\ 1\ \ 2\ \ 3\ \ 4\ \ 5\ )$, the graphs





can be build together like this:



Breakpoint Graph

- The **Breakpoint Graph** (Bafna and Pevzner, 1996) of a permutation π, denoted by $\mathrm{BP}(\pi)$, is built by combining the graph G_π with the graph of the *identity permutation*.

Breakpoint Graph Definition

The breakpoint graph $\mathrm{BP}(\pi)$ of a permutation π is a graph $G = (V, E_b \cup E_g)$ where

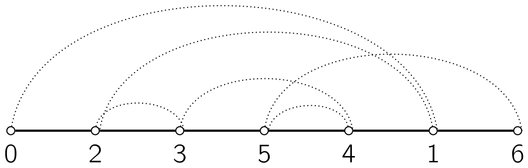
- **Vertices**: $V = \{0, 1, \ldots, n, n+1\}$.
- **Black edges**: $E_b = \{(\pi_i, \pi_{i+1}) : i = 0, \ldots, n\}$
- **Gray edges**: $E_g = \{(i, i+1) : i = 0, \ldots, n\}$

Properties of the BP graph

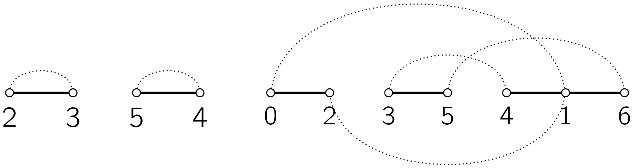
- All vertices are adjacent to the same number of gray and black edges.
  - Vertices 0 and $n + 1$: 1 gray and 1 black edge.
  - All other vertices: 2 gray and 2 black edges.
- This means that BP can be decomposed into *alternating cycles*.

An **alternating cycle** is a cycle in which the edges are alternating between two colors.

Breakpoint graph decomposition – BGD

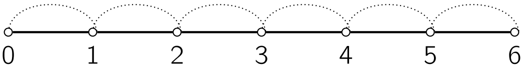


Can be decomposed like this:



BGD of the Identity permutation

For instance, take $i = (\ 1\ \ 2\ \ 3\ \ 4\ \ 5\ )$. The BP graph is

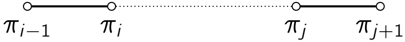


- In this case the BGD has $n + 1$ cycles, which is the maximum.
- The reversal problem becomes: *increase the number of cycles until we reach* $n + 1$.

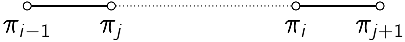
Now let's try to find what is the effect of a reversal in a genome graph, and also in a BGD cycle.

Effect of a Reversal in the Genome Graph

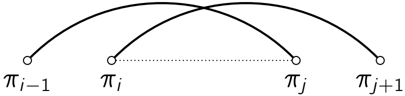
Consider a genome graph G_π, and the block from position i to j.



After the reversal $\rho(i, j)$, we have:



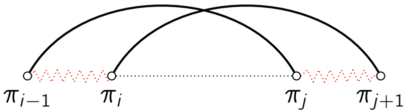
Which is the same as doing:



Effect of a Reversal in the Genome Graph

Applying the reversal $\rho(i, j)$ in a genome graph G_π is the same as:

- Removing the edges (π_{i-1}, π_i) and (π_j, π_{j+1})
- Adding the edges (π_{i-1}, π_j) and (π_i, π_{j+1})



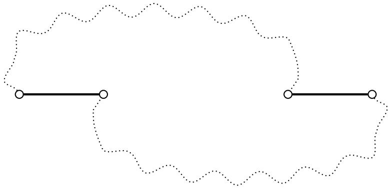
- The size of the reversal does not change this.
- Any two edges of a graph G_π **define a reversal**.

Effect of a Reversal in BGD Cycles

The effect of a reversal in a decomposition cycle depends on the **reversal edges**. There are two cases:

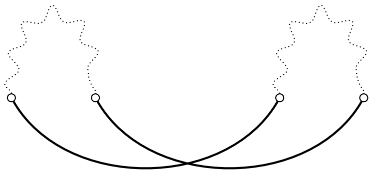
- The reversal edges are in two different cycles

- The reversal edges are in the same cycle



Edges in Different Cycle

After applying the reversal:

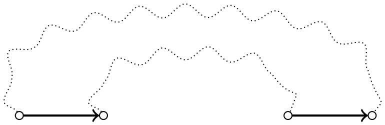


The cycles are **merged**, and the total number of cycles decreases by 1.

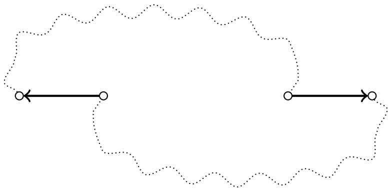
Edges in Same Cycle

When both edges are in the same cycle, there are two possible cases:

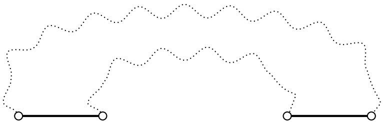
- **Directed edges**: both edges point to same direction when traversing the cycle.



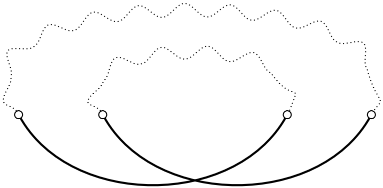
- **Undirected edges**: edges point to different direction when traversing the cycle.



Directed Case

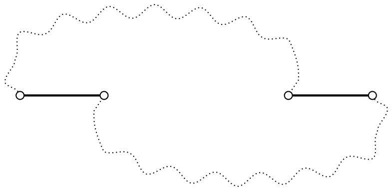


After applying the reversal:

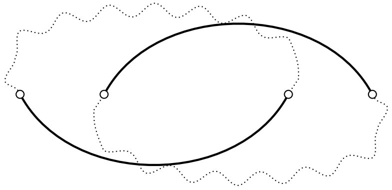


Still one cycle; the number of cycles does not change.

Undirected Case



After applying the reversal:



The cycle is **split** in two cycles. Therefore, the total number of cycles *increases* by 1.

Lower Bound with BP graph

As we saw, given a breakpoint graph decomposition, each reversal can only change the number of cycles by $+1$, 0, or -1.

Then, for a given $\mathrm{BP}(\pi)$ and a decomposition with maximum number of cycles $c(\pi)$, we have that

$$d(\pi) \geq n + 1 - c(\pi)$$

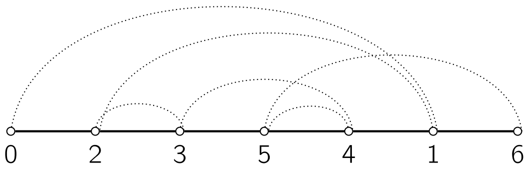
Proof: Since we can increase $c(\pi)$ by at most one, to sort π (sorting is the same as increasing $c(\pi)$ to $n+1$) we need at least $n + 1 - c(\pi)$ operations.

Lower Bound example

For the permutation $\pi = (\ 2\ \ 3\ \ 5\ \ 4\ \ 1\ )$, what is the breakpoint bound?

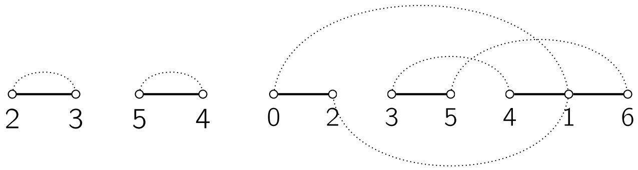
$$\pi = (\ 0 \mid 2\ \ 3 \mid 5\ \ 4 \mid 1 \mid 6\ ) \quad \Rightarrow \quad b(\pi) = 4 \quad \Rightarrow \quad d(\pi) \geq 2$$

And the cycle bound? The BP graph is:



Lower Bound example

One max cycle decomposition is:



Which means that

$$d(\pi) \geq n + 1 - c(\pi) = 6 - 3 = 3$$

Can we sort it in the number of reversals of the cycle bound?

Lower Bound Tightness

- The cycle bound is very tight, meaning that for most permutations π, we have that $d(\pi) = n + 1 - c(\pi)$.
- When is the bound not tight?
- When all the maximum cycle decompositions have some *directed* cycles, in a way that would force a reversal that does not increase the number of cycles.

Review of Unsigned Reversal

- BP(π) cycle decomposition gives better results than the breakpoint approach. But, BGD is not unique.
- It is also not easy: finding a max-cycle BGD is NP-hard.
- Directed cycles may make the distance $d(\pi) > n + 1 - c(\pi)$

Next episode: Signed Reversals

- Similar ideas will be used: BP graph decomposition.
- Good thing: BGD *is* unique.
- The difficult part is also related to directed cycles, but can it be treated polynomially.