Topics of today:

Canonical inversion distance and sorting:

1. Relational / Breakpoint diagram
2. Split / Neutral / Joining inversions
3. Good / bad components
4. Hurdles and fortress

Canonical inversion model - circular chromosomes

(Unichromosomal genomes $\equiv$ chromosomes)

Given two canonical circular chromosomes $\mathbb{A}$ and $\mathbb{B}$,...

Canonical Inversion Distance Problem: Compute the minimum number of inversions required to transform $\mathbb{A}$ into $\mathbb{B}$.

Denote by $\mathrm{d}_{\mathrm{INV}}(\mathbb{A}, \mathbb{B})$ the inversion distance of $\mathbb{A}$ and $\mathbb{B}$.

Canonical Inversion Sorting Problem: Find a sequence of $\mathrm{d}_{\mathrm{INV}}(\mathbb{A}, \mathbb{B})$ inversions that transform $\mathbb{A}$ into $\mathbb{B}$.

Relational diagram of canonical circular chromosomes

Given canonical circular chromosomes $\mathbb{A}$ and $\mathbb{B}$, their **relational diagram** $RD(\mathbb{A}, \mathbb{B}) = (V, E)$ is described as follows:

1. $V = V(\xi(\mathbb{A})) \cup V(\xi(\mathbb{B}))$: there is a vertex for each extremity of each gene in $\mathbb{A}$

   and a vertex for each extremity of each gene in $\mathbb{B}$

   The vertices corresponding to $\xi(\mathbb{A})$ are drawn in an upper line,
   while the vertices corresponding to $\xi(\mathbb{B})$ are drawn in a lower line.

   In each line, the vertices must follow the same (circular) order of the corresponding extremities in the respective chromosome, according to one of the two reading directions.

   Each vertex v has a label $\ell(v)$, that corresponds to the extremity it represents.

2. $E = E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B}) \cup E_\xi$, where:

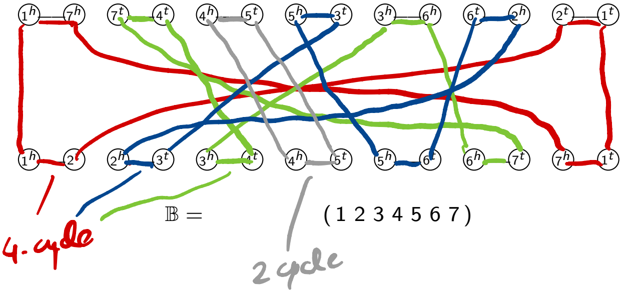
   - **Adjacency edges:** $\begin{cases} E_\alpha(\mathbb{A}) = \{uv : u, v \in V(\xi(\mathbb{A})) \text{ and } \ell(u)\ell(v) \in \alpha(\mathbb{A})\} \\ E_\alpha(\mathbb{B}) = \{uv : u, v \in V(\xi(\mathbb{B})) \text{ and } \ell(u)\ell(v) \in \alpha(\mathbb{B})\} \end{cases}$
   - **Extremity edges:** $E_\xi = \{uv : u \in V(\xi(\mathbb{A})) \text{ and } v \in V(\xi(\mathbb{B})) \text{ and } \ell(u) = \ell(v)\}$

Note that:

- Let $n = |\mathcal{G}_\star|$. The number of edges in $E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B})$ is $2n$ (n adjacency edges per chromosome).

Relational diagram of canonical circular chromosomes

$\mathbb{A} = \quad (1\ \bar{7}\ 4\ 5\ 3\ \bar{6}\ \bar{2})$



1^h 7^h 7^t 4^t 4^h 5^t 5^h 3^t 3^h 6^h 6^t 2^h 2^t 1^t
1^h 2^t 2^h 3^t 3^h 4^t 4^h 5^t 5^h 6^t 6^h 7^t 7^h 1^t
4-cycle
2 cycle

$\mathbb{B} = \quad (1\ 2\ 3\ 4\ 5\ 6\ 7)$

$n = |\mathcal{G}_\star| = 7$

Every vertex has degree two:

$RD(\mathbb{A}, \mathbb{B})$ is a collection of (even) cycles (alternating edes in E_ξ and in $E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B})$)

cycle with k edges in $E_\alpha(\mathbb{A}) \cup E_\alpha(\mathbb{B})$: k-cycle

$\mathcal{C}$ = set of cycles in $RD(\mathbb{A}, \mathbb{B})$

If $\mathbb{A} = \mathbb{B}$,
$RG(\mathbb{A}, \mathbb{B})$ has only 2-cycles:

$$2n = 2|\mathcal{C}| \quad \Rightarrow \quad n = |\mathcal{C}|$$

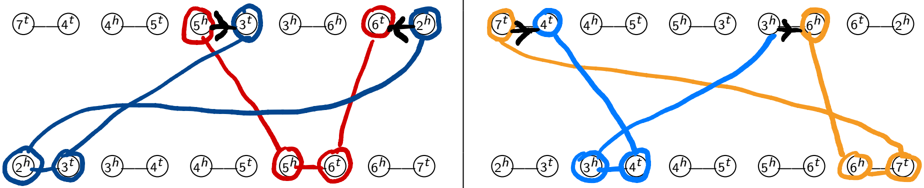
Otherwise, if $\mathbb{A} \neq \mathbb{B}$:

$$n > |\mathcal{C}|$$

Types of inversion and lower bound for the inversion distance

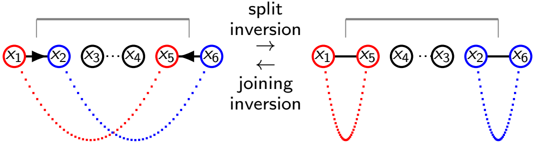
Assign one (arbitrary) direction to each cycle of $RD(\mathbb{A}, \mathbb{B})$

$\mathbb{A} = \quad (1\ \bar{7}\ 4\ 5\ 3\ \bar{6}\ \bar{2})$ $\qquad\qquad$ $\mathbb{A} = \quad (1\ \bar{7}\ 4\ 5\ 3\ \bar{6}\ \bar{2})$

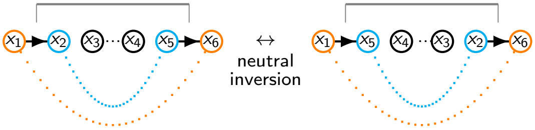


7^t 4^t 4^h 5^t 5^h 3^t 3^h 6^h 6^t 2^h
2^h 3^t 3^h 4^t 4^h 5^t 5^h 6^t 6^h 7^t
7^t 4^t 4^h 5^t 5^h 3^t 3^h 6^h 6^t 2^h
2^h 3^t 3^h 4^t 4^h 5^t 5^h 6^t 6^h 7^t

$\mathbb{B} = \quad (1\ 2\ 3\ 4\ 5\ 6\ 7)$ $\qquad\qquad$ $\mathbb{B} = \quad (1\ 2\ 3\ 4\ 5\ 6\ 7)$



x_1 x_2 x_3 ⋯ x_4 x_5 x_6
split inversion →
← joining inversion
x_1 x_5 x_4 ⋯ x_3 x_2 x_6



x_1 x_2 x_3 ⋯ x_4 x_5 x_6
↔
neutral inversion
x_1 x_5 x_4 ⋯ x_3 x_2 x_6

Lower bound for the inversion distance: $\quad d_{\text{INV}}(\mathbb{A}, \mathbb{B}) \geq n - |\mathcal{C}|$

Types of cycles

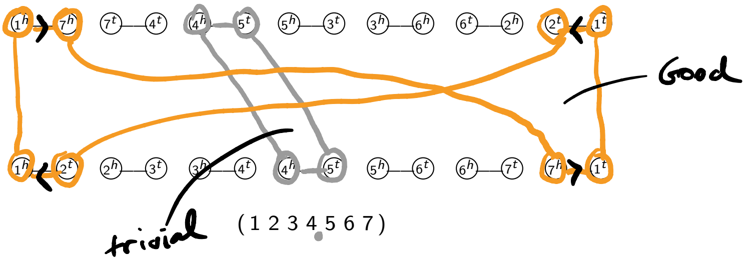
Trivial cycle: one adjacency in each chromosome

2-cycle (sorted)

Good cycle: $\begin{cases} \text{at least one pair of adjacencies with opposite directions in chromosome } \mathbb{A} \\ \text{at least one pair of adjacencies with opposite directions in chromosome } \mathbb{B} \end{cases}$

Can be split into two cycles by applying an inversion in $\mathbb{A}$ or in $\mathbb{B}$

$(1\ \bar{7}\ 4\ 5\ 3\ \bar{6}\ \bar{2})$


1^h 7^h 7^t 4^t 4^h 5^t 5^h 3^t 3^h 6^h 6^t 2^h 2^t 1^t
1^h 2^t 2^h 3^t 3^h 4^t 4^h 5^t 5^h 6^t 6^h 7^t 7^h 1^t
Good
trivial

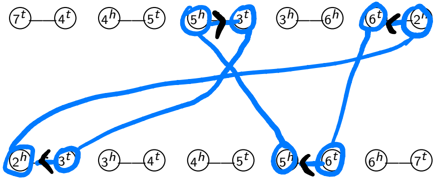
$(1\ 2\ 3\ 4\ 5\ 6\ 7)$

Types of cycles

Semi-good cycle: $\begin{cases} \text{at least one pair of adjacencies with opposite directions in one of the two chromosomes} \\ \text{all adjacencies have the same direction in the other chromosome} \end{cases}$

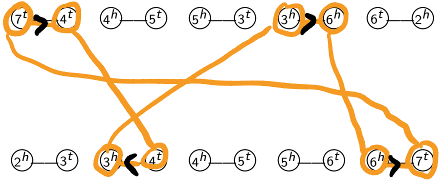
Can be split into two cycles by applying an inversion only in $\mathbb{A}$ or only in $\mathbb{B}$

$(\,1\ \bar{7}\ 4\ 5\ 3\ \bar{6}\ \bar{2}\,)$



$(\,1\ 2\ 3\ 4\ 5\ 6\ 7\,)$

$(\,1\ \bar{7}\ 4\ 5\ 3\ \bar{6}\ \bar{2}\,)$



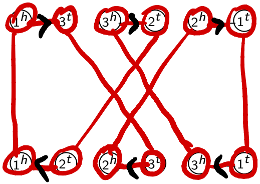
$(\,1\ 2\ 3\ 4\ 5\ 6\ 7\,)$

Types of cycles

Bad cycle: $\begin{cases}\text{all adjacencies in chromosome } \mathbb{A} \text{ have the same direction}\\ \text{all adjacencies in chromosome } \mathbb{B} \text{ have the same direction}\end{cases}$

Cannot be split into two cycles

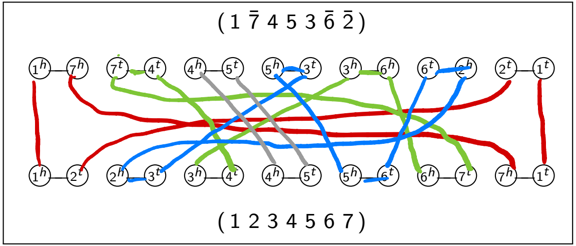
(1 3 2)



(1 2 3)

Relational diagram $\cong$ Breakpoint diagram

Relational diagram:



$(1\ \bar{7}\ 4\ 5\ 3\ \bar{6}\ \bar{2})$
1^h 7^h 7^t 4^t 4^h 5^t 5^h 3^t 3^h 6^h 6^t 2^h 2^t 1^t
1^h 2^t 2^h 3^t 3^h 4^t 4^h 5^t 5^h 6^t 6^h 7^t 7^h 1^t
$(1\ 2\ 3\ 4\ 5\ 6\ 7)$

Looking either at the top line or
at the bottom line of the diagram:

Two interleaving cycles: $c \ldots c' \ldots c \ldots c'$

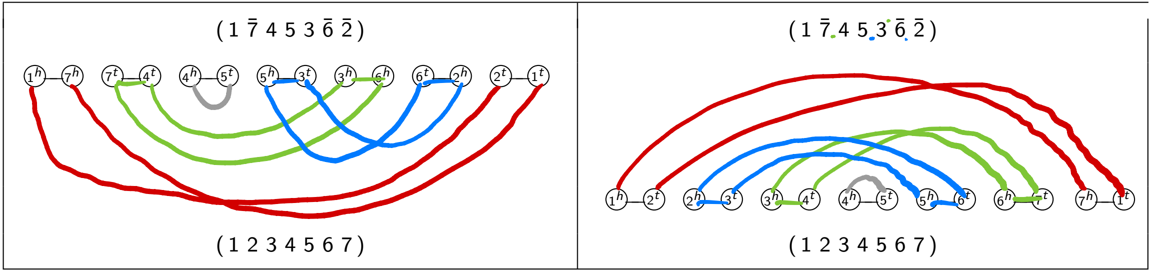
Interleaving sequence of cycles: $c_1, c_2, \ldots, c_k$ such that c_i and c_{i+1} are interleaving for all $1 \le i \le k-1$

Interleaving component or simply **component** K:

$$\begin{cases} \text{for each pair of cycles } c, c' \in K \text{ there is an interleaving sequence from } c \text{ to } c' \\ K \text{ is maximal} \end{cases}$$

c' c c' c c'' c c''

Breakpoint diagrams:

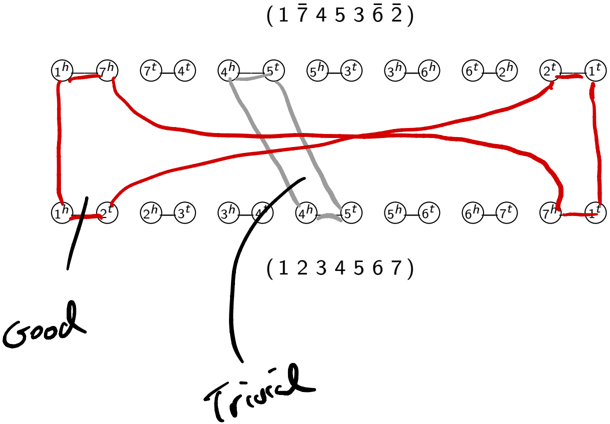


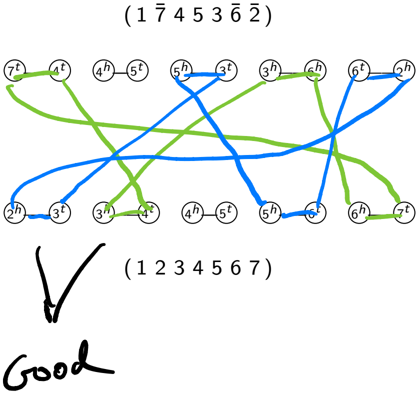
$(1\ \bar{7}\ 4\ 5\ 3\ \bar{6}\ \bar{2})$
1^h 7^h 7^t 4^t 4^h 5^t 5^h 3^t 3^h 6^h 6^t 2^h 2^t 1^t
$(1\ 2\ 3\ 4\ 5\ 6\ 7)$
$(1\ \bar{7}\ 4\ 5\ 3\ \bar{6}\ \bar{2})$
1^h 2^t 2^h 3^t 3^h 4^t 4^h 5^t 5^h 6^t 6^h 7^t 7^h 1^t
$(1\ 2\ 3\ 4\ 5\ 6\ 7)$

Types of (interleaving) components

Trivial component: only one trivial 2-cycle

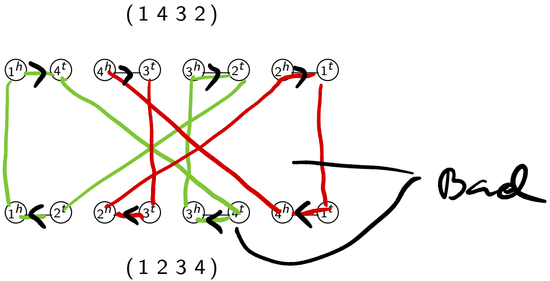
Good component: at least one good or semi-good cycle


(1 $\bar{7}$ 4 5 3 $\bar{6}$ $\bar{2}$)
1^h 7^h 7^t 4^t 4^h 5^t 5^h 3^t 3^h 6^h 6^t 2^h 2^t 1^t
1^h 2^t 2^h 3^t 3^h 4^t 4^h 5^t 5^h 6^t 6^h 7^t 7^h 1^t
(1 2 3 4 5 6 7)
Good
Trivial


(1 $\bar{7}$ 4 5 3 $\bar{6}$ $\bar{2}$)
7^t 4^t 4^h 5^t 5^h 3^t 3^h 6^h 6^t 2^h
2^h 3^t 3^h 4^t 4^h 5^t 5^h 6^t 6^h 7^t
(1 2 3 4 5 6 7)
Good

Types of (interleaving) components

Bad component: only bad cycles



Quiz 1

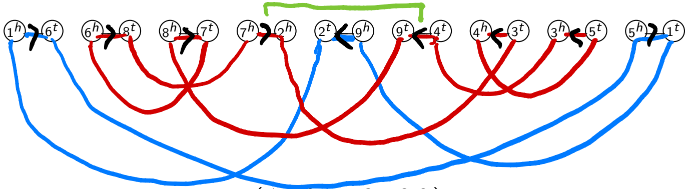
1 Which of the following statements about the relational diagram are true?

- ~~A~~ A cycle can always be split into two cycles with an inversion.
- ~~B~~ A joining inversion cannot be optimal.
- ~~C~~ A split inversion is always optimal.
- [✓ D] It is always possible to split a good or a semi-good cycle into two.
- [✓ E] A bad cycle cannot be split by an inversion.

Unsafe inversions

A split inversion applied to a good cycle of a good component can create bad components

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

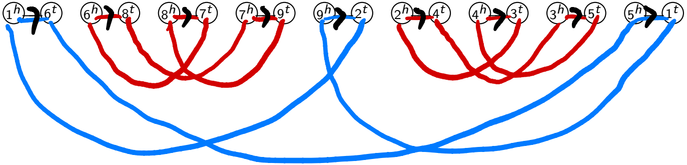


1^h 6^t 6^h 8^t 8^h 7^t 7^h 2^h 2^t 9^h 9^t 4^t 4^h 3^t 3^h 5^t 5^h 1^t

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

– One good component

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



1^h 6^t 6^h 8^t 8^h 7^t 7^h 9^t 9^h 2^t 2^h 4^t 4^h 3^t 3^h 5^t 5^h 1^t

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

– 3 bad components

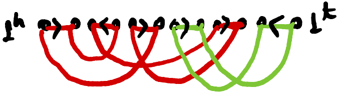
Sorting a good component - finding safe split inversions

Target adjacency: $\begin{cases} \text{good} \\ \text{bad} \end{cases}$

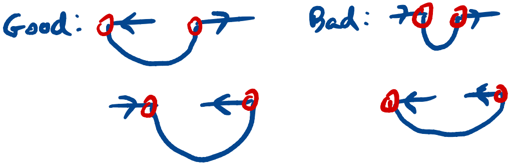
Overlapping target adjacencies

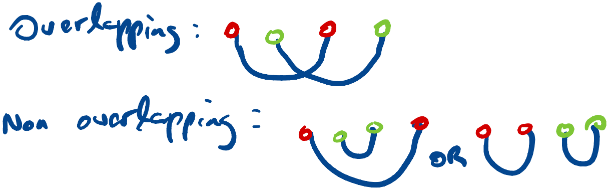
Overlap graph of a good component

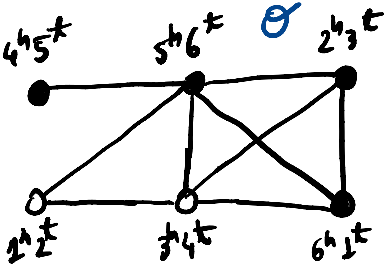
(1 5 $\bar{4}$ 2 $\bar{6}$ $\bar{3}$)



(1 2 3 4 5 6)

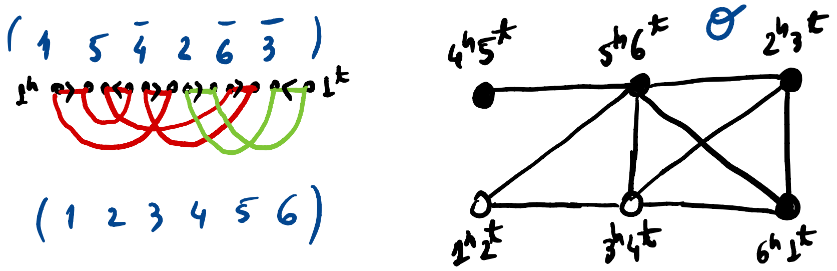






good adjacency
⇓
black vertex

bad adjacency
⇓
white vertex

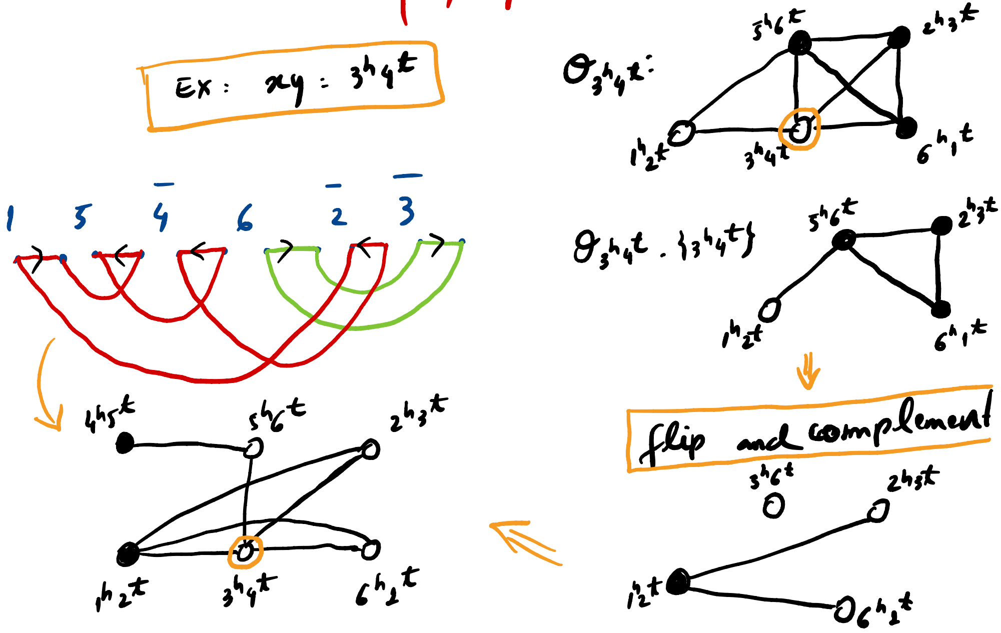


θ_{xy} : subgraph of θ composed of xy and its adjacent vertices

Effect of an inversion on the overlap graph θ :

if xy is bad: { 1) flip the types of the vertices of $\theta_{xy} - \{xy\}$
2) complement the edges of $\theta_{xy} - \{xy\}$

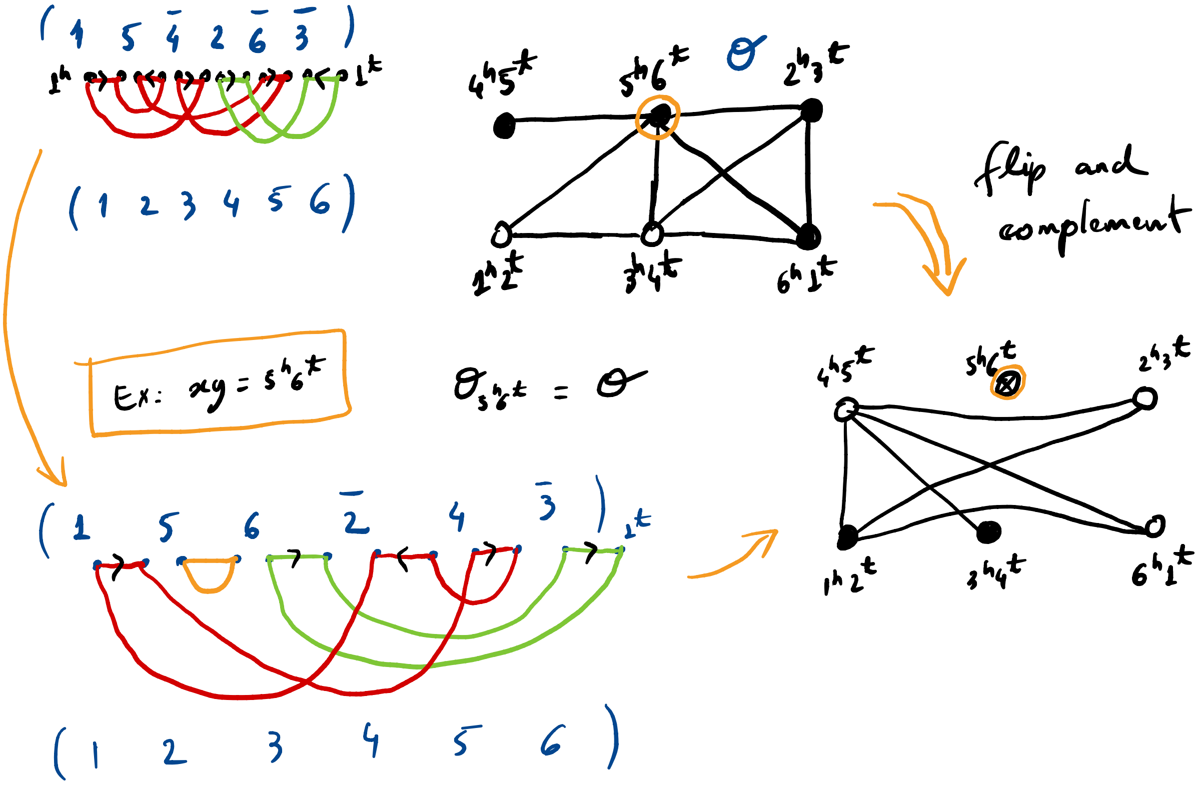
Ex : $xy = 3^h4^t$



if xy is good { 1) flip the types of all vertices of G_{xy}
2) complement the edges of G_{xy}

Ex: $xy = 5^h6^t$

$\theta_{5^h6^t} = \theta$



Sorting a good component - finding safe split inversions

G: # of good adjacencies in RD(A,B)
g(xy): # of good adjacencies overlapping xy in RD(A,B)
b(xy): # of bad adjacencies overlapping xy in RD(A,B)

score(xy): # good adjacencies in the diagram <u>after</u> fixing xy

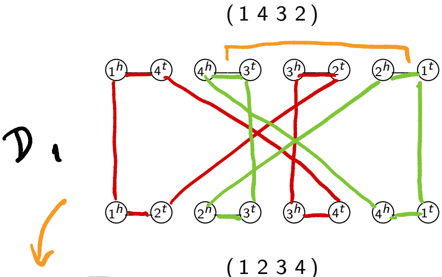
$$\boxed{score(xy) = G + b(xy) - g(xy) - 1}$$

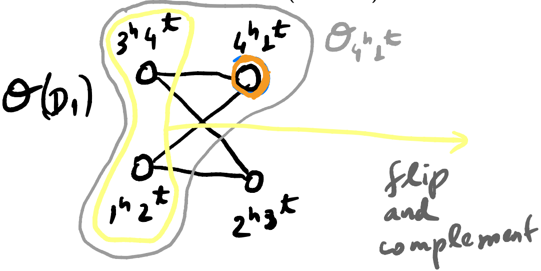
An inversion that fixes a good target adjacency with maximal score is <u>SAFE</u>. (does not create new bad components)

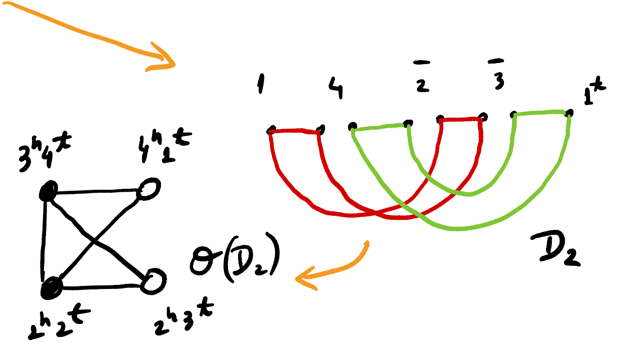
Sorting a bad component with a neutral inversion

Overlap graph of a bad component

Any neutral inversion applied to a bad adjacency of a bad component K turns K into a good component


(1 4 3 2)
1^h 4^t 4^h 3^t 3^h 2^t 2^h 1^t
D_1
1^h 2^t 2^h 3^t 3^h 4^t 4^h 1^t
(1 2 3 4)



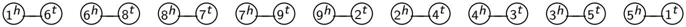


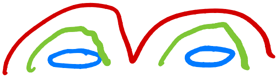
Sorting bad components - hurdles

K_1, K_2 and K_3 are three distinct components in $RD(\mathbb{A}, \mathbb{B})$ so that $K_3 \ldots K_1 \ldots K_1 \ldots K_3 \ldots K_2 \ldots K_2$

$\Rightarrow$ K_3 **separates** K_1 and K_2

(1 6 8 7 9 2 4 3 5)





(1 2 3 4 5 6 7 8 9)

By joining with an inversion two cycles c_1 and c_2, that belong to two distinct components K_1 and K_2 respectively, we merge not only the components K_1 and K_2, but also all components that separate K_1 and K_2 , into a single **good** component K.

Sorting bad components - simple hurdles and super hurdles

h : number of hurdles in $RD(\mathbb{A}, \mathbb{B})$

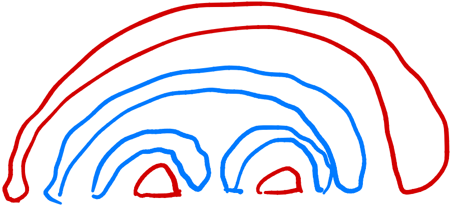
hurdle : bad component that does not separate 2 bad components

super hurdle K : fixing K by a neutral inversion creates a new hurdle

Sorting bad components - simple hurdles and super hurdles

Sorting bad components - fortress

$$f : \begin{cases} 0 & RD(\mathbb{A}, \mathbb{B}) \text{ is not a fortress} \\ 1 & RD(\mathbb{A}, \mathbb{B}) \text{ is a fortress} \end{cases}$$



fortress: A) odd # of hurdles

B) all hurdles are super hurdles

Canonical inversion distance of circular chromosomes

$$\mathsf{d}_{\mathrm{INV}}(\mathbb{A}, \mathbb{B}) = \underline{n - |\mathcal{C}| + h + f}$$

Quiz 2

1 Which of the following statements about the inversion model are true?

- A The inversion distance depends only on the number of cycles in the relational diagram.
- B Every bad component in the diagram is a hurdle.
- C A good component can always be sorted with (safe) split inversions.
- D A super hurdle can be optimally sorted with a neutral inversion.
- E A diagram with an even number of bad components can be a fortress.

References

Transforming Cabbage into Turnip: Polynomial Algorithm for Sorting Signed Permutations by Reversals

(Sridhar Hannenhalli and Pavel A. Pevzner)

Journal of the ACM, Vol. 46, No. 1, pages. 1–27 (1999)

The Inversion Distance Problem

(Anne Bergeron, Julia Mixtacki and Jens Stoye)

In: Mathematics of Evolution and Phylogeny. Gascuel O (Ed); (2005)