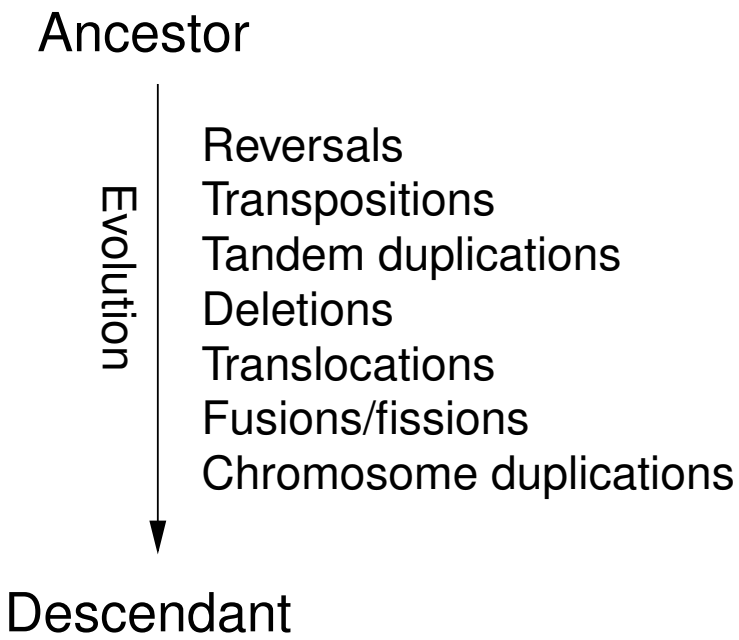
ulm university universität ulm

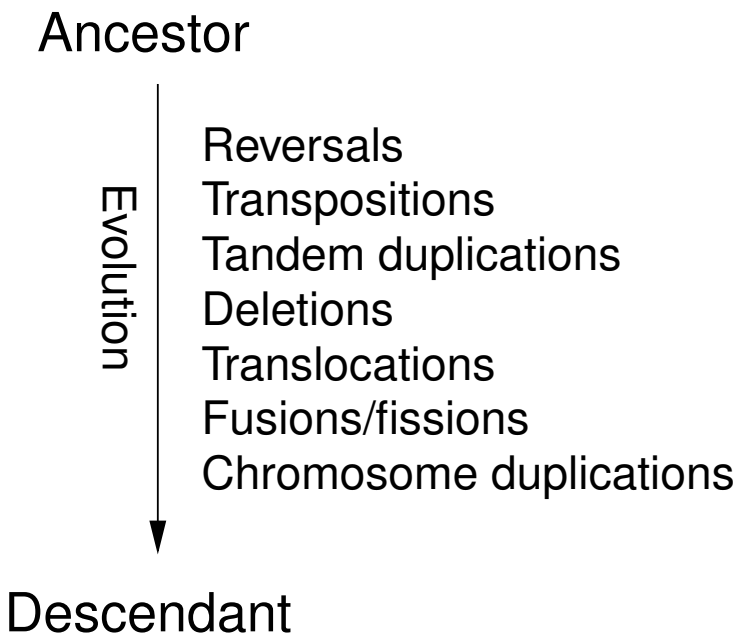
Genome rearrangements with duplications

Martin Bader | January 21, 2010

Genome Rearrangement Problems



Genome Rearrangement Problems

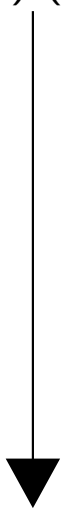


- Further restriction: chromosomes in ancestor are disjoint or identical

Example

Transform $\rho = (\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ into $\pi = (\overleftarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overleftarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

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



$(\overleftarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overleftarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

Example

Transform $\rho = (\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ into $\pi = (\overleftarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overrightarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ tandem duplication

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

Example

Transform $\rho = (\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ into $\pi = (\overleftarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overrightarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ tandem duplication

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ translocation

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{2}\ \overleftarrow{7}\ \overleftarrow{6}\ \overleftarrow{5})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

Example

Transform $\rho = (\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ into $\pi = (\overleftarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overrightarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ tandem duplication

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ translocation

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{2}\ \overleftarrow{7}\ \overleftarrow{6}\ \overleftarrow{5})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$ deletion

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{2}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

Example

Transform $\rho = (\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ into $\pi = (\overleftarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overleftarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ tandem duplication

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ translocation

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{2}\ \overleftarrow{7}\ \overleftarrow{6}\ \overleftarrow{5})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$ deletion

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{2}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$ reversal

$(\overrightarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overleftarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

Example

Transform $\rho = (\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ into $\pi = (\overleftarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overleftarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ tandem duplication

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$ translocation

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{2}\ \overleftarrow{7}\ \overleftarrow{6}\ \overleftarrow{5})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$ deletion

$(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{2}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$ reversal

$(\overrightarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overleftarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$ reversal

$(\overleftarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overleftarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

Algorithm: outline

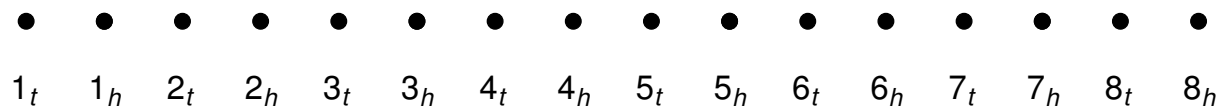
- Extension of algorithm for unichromosomal genomes (APBC 2009)
- Define a lower bound on $d(\rho, \pi)$ based on the breakpoint graph
- Start with π, sort backwards to ρ $\Rightarrow$ apply inverse operations
- Find operations on π that decrement the lower bound
- Apply the “best” of them (Greedy algorithm)
- If algorithm gets stuck, use fall-back algorithm

The breakpoint graph

- Invented by Bafna and Pevzner for genomes without duplicates

The breakpoint graph

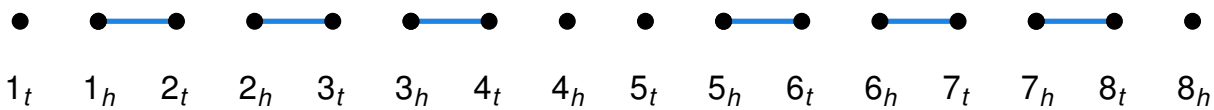
- Invented by Bafna and Pevzner for genomes without duplicates
- Write extremities of all genes on a straight line



Example: $\rho = (\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$, $\pi = (\overleftarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overleftarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

The breakpoint graph

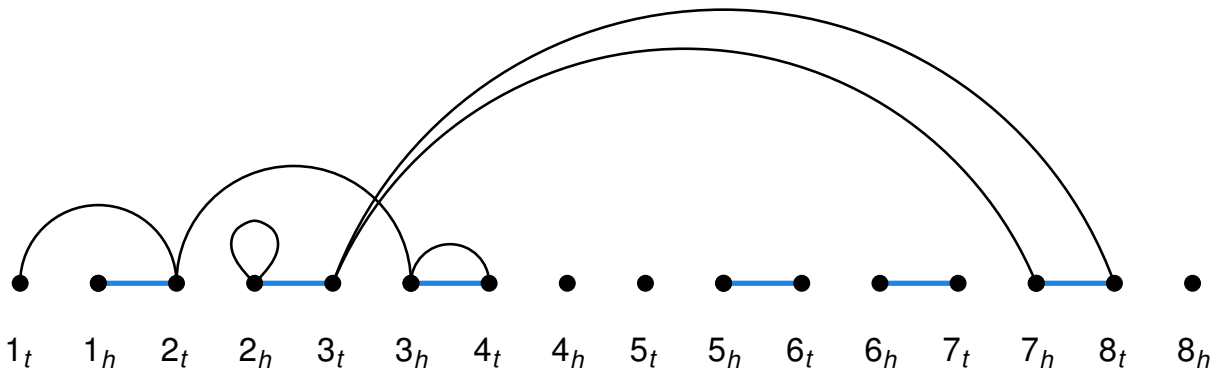
- Invented by Bafna and Pevzner for genomes without duplicates
- Write extremities of all genes on a straight line
- Add blue edges according to adjacencies in ρ



Example: $\rho = (\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8})$, $\pi = (\overleftarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overleftarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

The breakpoint graph

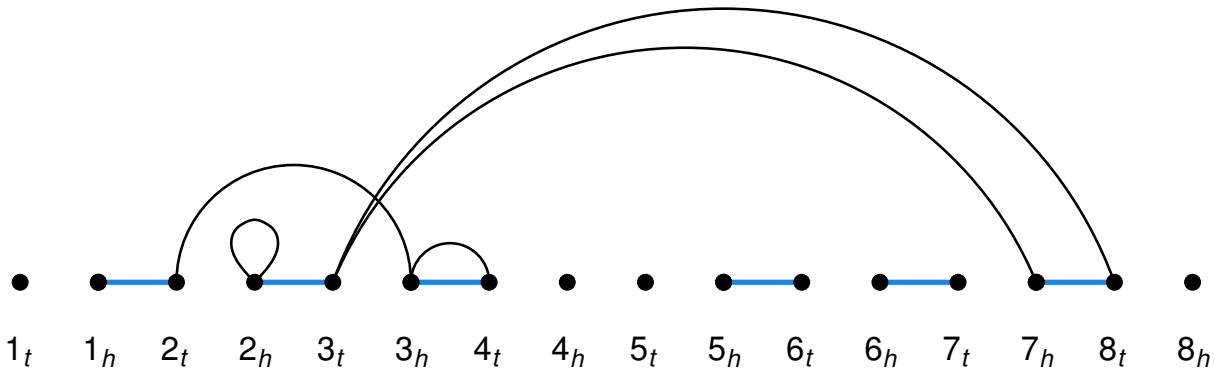
- Invented by Bafna and Pevzner for genomes without duplicates
- Write extremities of all genes on a straight line
- Add blue edges according to adjacencies in ρ
- Add black edges according to adjacencies in π



Example: $\rho = (\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8}),\ \ \pi = (\overleftarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overleftarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

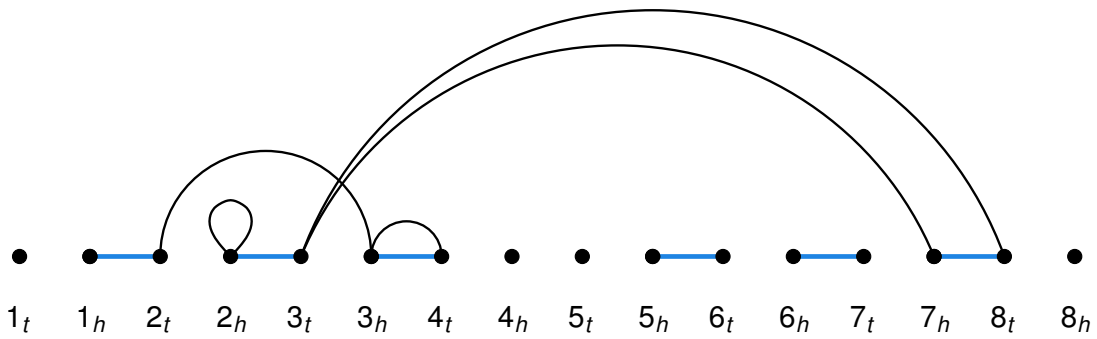
The breakpoint graph

- Invented by Bafna and Pevzner for genomes without duplicates
- Write extremities of all genes on a straight line
- Add blue edges according to adjacencies in ρ
- Add black edges according to adjacencies in π
- If x is a telomere in ρ, remove all edges incident to x



Example: $\rho = (\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4})\,(\overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8}),\ \ \pi = (\overleftarrow{1}\ \overrightarrow{2}\ \overleftarrow{2}\ \overleftarrow{3}\ \overleftarrow{7})\,(\overleftarrow{4}\ \overleftarrow{3}\ \overrightarrow{8})$

The breakpoint graph



- Component: connected component (graph theory)
- Loop: black edge (v, v)
- For unichromosomal genomes, the lower bound depends on the number of components and loops

Handling incorrect telomeres

- Incorrect telomeres:

$$T(\rho,\pi) = \sum_{x_{t/h} | t(\rho,x_{t/h})>0} \max\{t(\rho,x_{t/h}) - t(\pi,x_{t/h}),0\} \quad + \sum_{x_{t/h} | t(\rho,x_{t/h})=0} t(\pi,x_{t/h})$$

Example: $\rho = (\overrightarrow{1}\ \overrightarrow{2}\ )\ (\ \overrightarrow{1}\ \overrightarrow{2}\ )\ (\overrightarrow{3}\ \overrightarrow{4}),\ \ \pi = (\overrightarrow{1}\ \overrightarrow{2}\ \overleftarrow{3}\ )\ (\overrightarrow{3}\ \overrightarrow{4}\ )\ (\ \overrightarrow{4}\ )$

Handling incorrect telomeres

- Incorrect telomeres:

$$T(\rho,\pi) = \sum_{x_{t/h} | t(\rho,x_{t/h})>0} \max\{t(\rho,x_{t/h}) - t(\pi,x_{t/h}),0\} \quad + \sum_{x_{t/h} | t(\rho,x_{t/h})=0} t(\pi,x_{t/h})$$

Example: $\rho = (\overrightarrow{1}\ \overrightarrow{2}\ )\ (\ \overrightarrow{1}\ \overrightarrow{2}\ )\ (\overrightarrow{3}\ \overrightarrow{4}),\ \ \pi = (\overrightarrow{1}\ \overrightarrow{2}\ \overleftarrow{3}\ )\ (\overrightarrow{3}\ \overrightarrow{4}\ )\ (\ \overrightarrow{4}\ )$

Translocation: $\rho = (\overrightarrow{1}\ \overrightarrow{2}\ )\ (\ \overrightarrow{1}\ \overrightarrow{2}\ )\ (\overrightarrow{3}\ \overrightarrow{4}),\ \ \pi = (\overrightarrow{1}\ \overrightarrow{2}\ )\ (\overrightarrow{3}\ \overrightarrow{4}\ )\ (\overrightarrow{3}\ \overrightarrow{4}\ )$

Handling incorrect telomeres

- Incorrect telomeres:

$$T(\rho,\pi) = \sum_{x_{t/h} | t(\rho,x_{t/h})>0} \max\{t(\rho,x_{t/h}) - t(\pi,x_{t/h}),0\} \quad + \sum_{x_{t/h} | t(\rho,x_{t/h})=0} t(\pi,x_{t/h})$$

- Each operation can decrease $T(\rho,\pi)$ by at most 2.
- If an operation decreases $T(\rho,\pi)$, neither a component can be split nor a loop can be removed

Example: $\rho = (\overrightarrow{1}\ \overrightarrow{2})\ (\overrightarrow{1}\ \overrightarrow{2})\ (\overrightarrow{3}\ \overrightarrow{4}),\ \ \pi = (\overrightarrow{1}\ \overrightarrow{2}\ \overleftarrow{3})\ (\overrightarrow{3}\ \overrightarrow{4})\ (\overrightarrow{4})$

Translocation: $\rho = (\overrightarrow{1}\ \overrightarrow{2})\ (\overrightarrow{1}\ \overrightarrow{2})\ (\overrightarrow{3}\ \overrightarrow{4}),\ \ \pi = (\overrightarrow{1}\ \overrightarrow{2})\ (\overrightarrow{3}\ \overrightarrow{4})\ (\overrightarrow{3}\ \overrightarrow{4})$

A lower bound

- Handle telomeres separately from components and loops
- Handling components and loops works like in the unichromosomal case
- Number of components is maximized when sorted ($=: c(\rho)$)
- Number of loops and incorrect telomeres is 0 when sorted

$$d(\rho,\pi) \geq lb(\rho,\pi) = c(\rho) - C(\rho,\pi) + \sum_{\text{Components } C_i} \left\lceil \frac{L_i(\rho,\pi)}{2} \right\rceil + \left\lceil \frac{T(\rho,\pi)}{2} \right\rceil$$

The algorithm

- Define disruption measure $\tau(\rho, \pi)$ based on the number of breakpoints and wrong number of occurrences of elements

while $\pi \neq \rho$ **do**
  find all operations on π that decrease $lb(\rho, \pi)$
  find some additional promising operations
  if operation found **then**
    apply an operation that maximizes $(\Delta lb, \Delta\tau)$
  else
    use fall-back algorithm
  end if
end while

The fall-back algorithm

- Main algorithm can get stuck
- This will never happen when each element has the same number of occurrences in ρ and π
- Add segments of consecutive elements to π until elements belonging to the same chromosome in ρ have the same multiplicity in π
- Copy/delete chromosomes in ρ until all elements have the same multiplicity in ρ and π
- Sort with main algorithm, but avoid duplications and deletions

The fallback algorithm: example

$$\rho = (\overrightarrow{1}\ \overrightarrow{2})\,(\overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5})$$

$$\pi = (\overrightarrow{1}\ \overrightarrow{2})\,(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3})\,(\overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5})$$

The fallback algorithm: example

$$\rho = (\overrightarrow{1}\ \overrightarrow{2})\,(\overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5})$$

$$(\overrightarrow{1}\ \overrightarrow{2})\,(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5})\,(\overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5})$$

deletion

$$\pi = (\overrightarrow{1}\ \overrightarrow{2})\,(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3})\,(\overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5})$$

The fallback algorithm: example

$$\rho = (\overrightarrow{1}\ \overrightarrow{2})\,(\overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5})$$

$2\times$ chromosome duplication

$$(\overrightarrow{1}\ \overrightarrow{2})\,(\overrightarrow{1}\ \overrightarrow{2})\,(\overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5})\,(\overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5})$$

$$(\overrightarrow{1}\ \overrightarrow{2})\,(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5})\,(\overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5})$$

deletion

$$\pi = (\overrightarrow{1}\ \overrightarrow{2})\,(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3})\,(\overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5})$$

The fallback algorithm: example

$$\rho = (\vec{1}\ \vec{2})\,(\vec{3}\ \vec{4}\ \vec{5})$$

$2\times$ chromosome duplication

$$(\vec{1}\ \vec{2})\,(\vec{1}\ \vec{2})\,(\vec{3}\ \vec{4}\ \vec{5})\,(\vec{3}\ \vec{4}\ \vec{5})$$

fusion

$$(\vec{1}\ \vec{2})\,(\vec{1}\ \vec{2}\ \vec{3}\ \vec{4}\ \vec{5})\,(\vec{3}\ \vec{4}\ \vec{5})$$

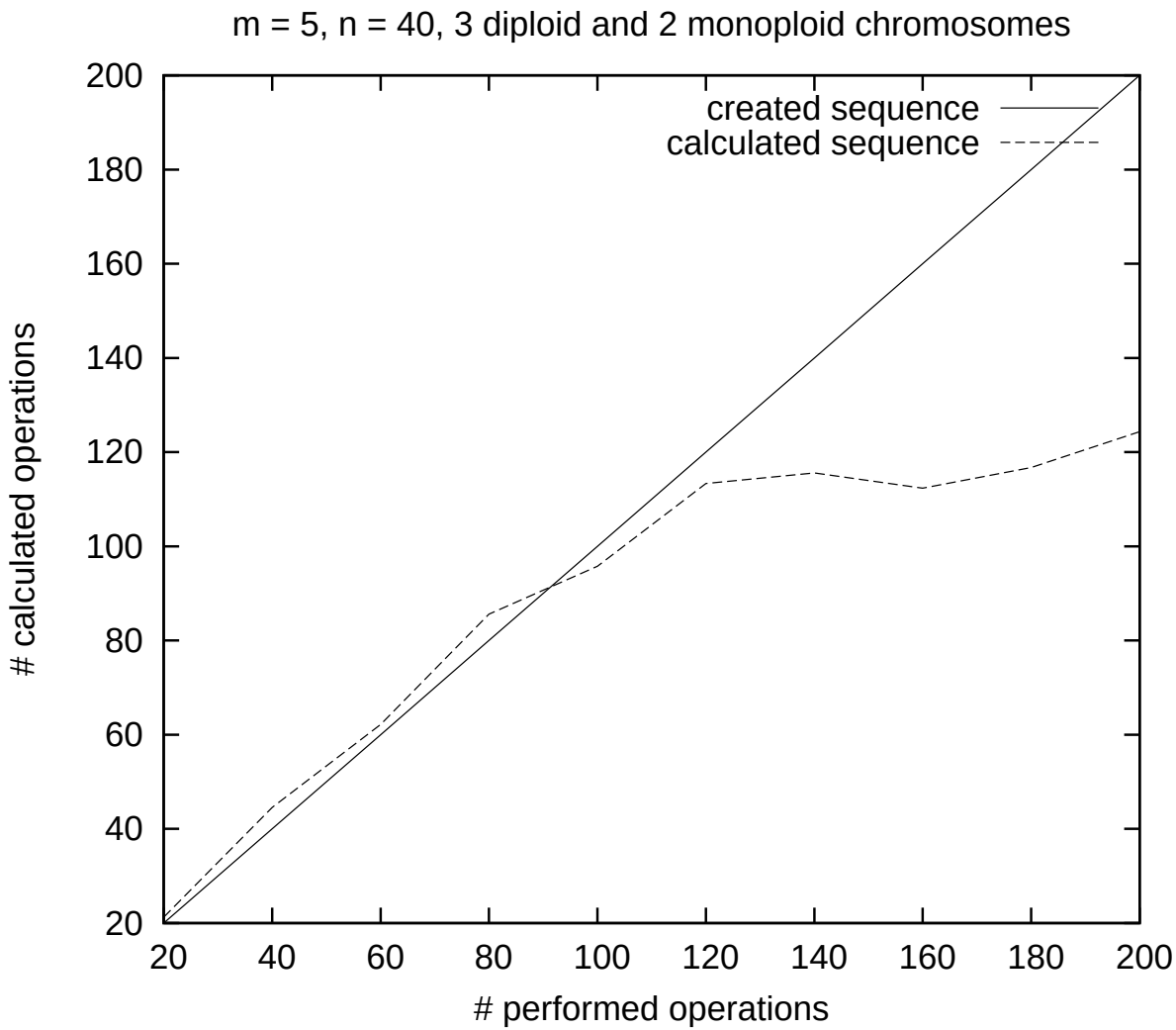
deletion

$$\pi = (\vec{1}\ \vec{2})\,(\vec{1}\ \vec{2}\ \vec{3})\,(\vec{3}\ \vec{4}\ \vec{5})$$

Experimental results

- Genome ρ: m disjoint chromosomes of n elements, each chromosome has 1 or 2 copies
- Apply $\alpha \cdot n \cdot m$ random operations ($\alpha \in [0,1]$) to obtain genome π
- Use algorithm to find sorting scenario between ρ and π
- Compare # applied operations to # calculated operations

Experimental results



Conclusion and Future work

- Algorithm works well for closely related genomes
- Possible improvements:
  - Tighter lower bound
  - Finding an upper bound
  - Improving the heuristics
  - Post-processing of the sequences
  - More realistic weighting of the operations

Thanks for your attention!

Any questions?