A unifying view of genome rearrangements

A. Bergeron[1] J. Mixtacki[2] J. Stoye[3,4]

[1] Comparative Genomics Laboratory
Université du Québec à Montréal

[2] International NRW Graduate School
in Bioinformatics and Genome Research
Center for Biotechnology, Universität Bielefeld

[3] AG Genominformatik
Technische Fakultät, Universität Bielefeld

[4] Institute for Bioinformatics
Center for Biotechnology, Universität Bielefeld

PICB Spring School, Shanghai, March 12-16, 2007

1. Genome Rearrangement Problems
   - Biological Background
   - Genomic Distances

2. The DCJ Operation
   - Graphs with Vertices of Degree One or Two
   - The Genome Graph
   - The Adjacency Graph
   - Algorithm for Sorting by DCJ Operations

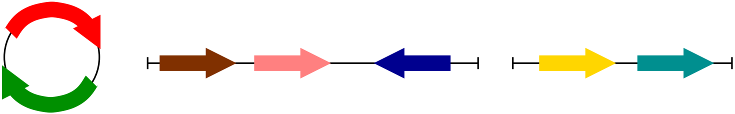
3. Relation to other Genomic Distances
   - The Inversion Distance
   - The Translocation Distance

1 Genome Rearrangement Problems
- Biological Background
- Genomic Distances

2 The DCJ Operation
- Graphs with Vertices of Degree One or Two
- The Genome Graph
- The Adjacency Graph
- Algorithm for Sorting by DCJ Operations

3 Relation to other Genomic Distances
- The Inversion Distance
- The Translocation Distance

Biological Background



- Genome is the entire DNA of a living organism
- Gene is a segment of DNA that is involved e.g. in producing a protein, and its orientation depends on the DNA-strand that it lies on
- Genome consists of chromosomes
- Chromosomes are linear or circular

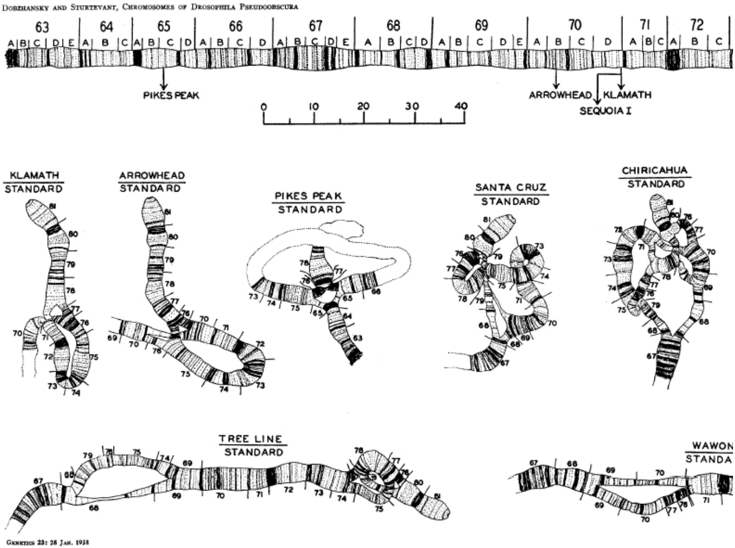


Figure: Dobzhansky & Sturtevant, Genetics (1938)

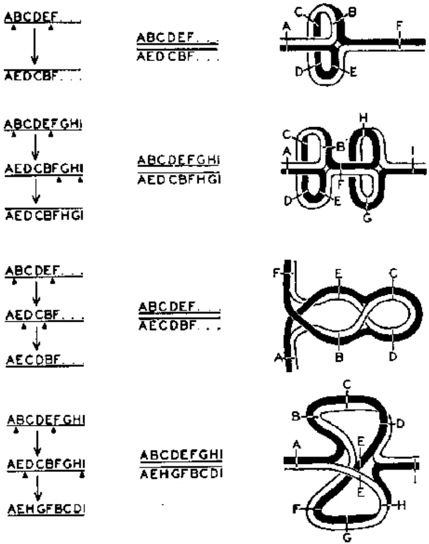


Figure: Dobzhansky & Sturtevant, Genetics (1938)

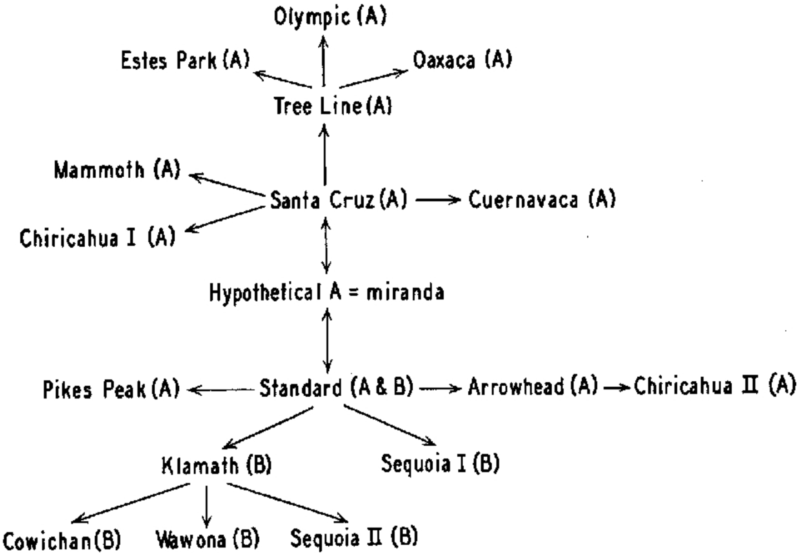


Figure: Dobzhansky & Sturtevant, Genetics (1938)

Conserved synteny between human and mouse:

Exchange of intra- and interchromosomal segments during evolution → Genome rearrangements

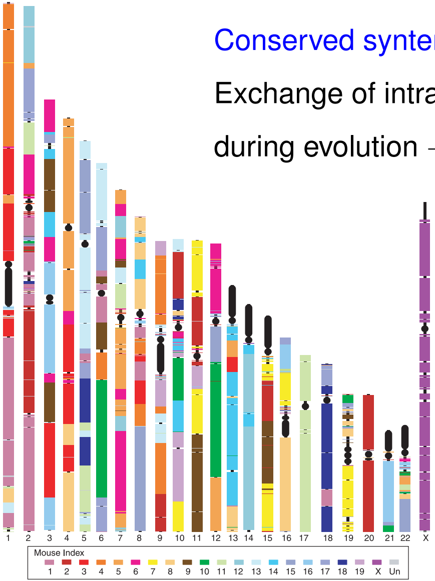
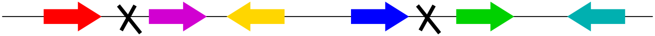


Figure: Eichler & Sankoff, Science (2003)

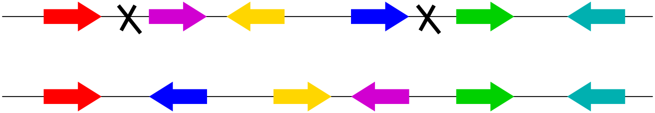
Rearrangement Operations

Inversions reverse the order and the orientation of a segment:



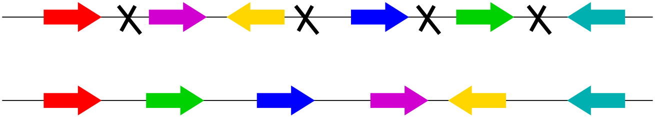
Rearrangement Operations

Inversions reverse the order and the orientation of a segment:

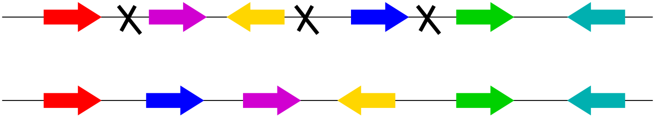


Rearrangement Operations

Block interchanges exchange two segments:

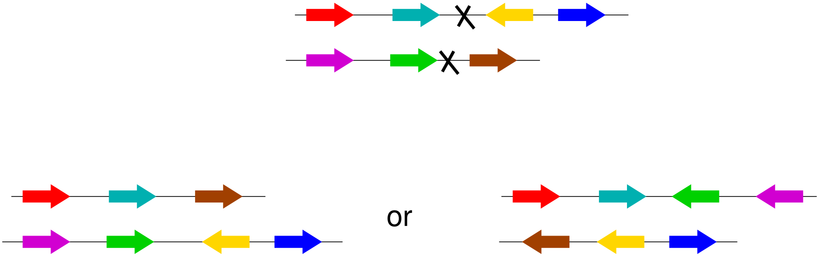


Transpositions are block interchanges whose exchanged segments are adjacent:



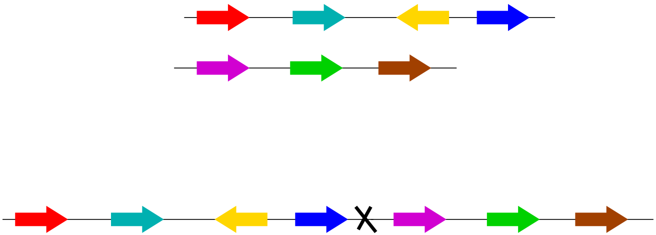
Rearrangement Operations

Translocations exchange two chromosome ends:



Rearrangement Operations

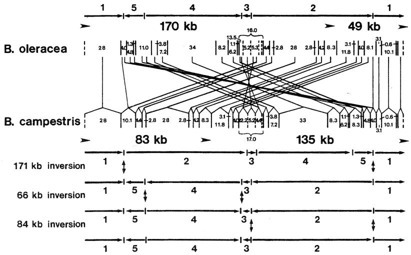
Fusions and fissions are translocations involving or creating empty chromosomes:



Genome Rearrangements

Genome rearrangements change the content andør the order of genes of a genome:

- inversions
- transpositions
- translocations
- fusions and fissions
- ...



(Figure: Palmer & Herbon, 1988)

The number of rearrangements needed to transform one genome into another is a measure for the evolutionary distance between two species

Genomic Distances

Definition

Distance $d(A, B)$: minimum number of operations needed to transform genome A into genome B

1. What kind of genome model?
   - Unichromosomal vs. multichromosomal genomes
   - Linear vs. circular chromosomes
   - Linearly ordered vs. partially ordered chromosomes
   - Duplicates, gene families
2. Which set of operations?
   - Only single operation
   - Weights

Historical Overview

Inversions-only: Sankoff (1992), Bafna & Pevzner (1993), Hannenhalli & Pevzner (1995), Kaplan *et al.* (1999), Bader *et al.* (2001), Bergeron *et al.* (2004)

Translocations-only: Hannenhalli (1996), Bergeron *et al.* (2005)

Inversions, translocations, fusions & fissions: Hannenhalli & Pevzner (1995), Tesler (2002), Ozery-Flato & Shamir (2003)

Block interchanges: Christie (1996)

Transpositions: Bafna & Pevzner (1998), Hartman (2003), Labarre (2005)

Weighted inversions, transpositions & inverted transpositions: Bader & Ohlebusch (2006)

Inversions, translocations, fusions, fissions & block-interchanges: Yancopoulos *et al.* (2005), Bergeron *et al.* (2006)

Historical Overview

Inversions-only: Sankoff (1992), Bafna & Pevzner (1993), Hannenhalli & Pevzner (1995), Kaplan *et al.* (1999), Bader *et al.* (2001), Bergeron *et al.* (2004)

Translocations-only: Hannenhalli (1996), Bergeron *et al.* (2005)

Inversions, translocations, fusions & fissions: Hannenhalli & Pevzner (1995), Tesler (2002), Ozery-Flato & Shamir (2003)

Block interchanges: Christie (1996)

Transpositions: Bafna & Pevzner (1998), Hartman (2003), Labarre (2005)

Weighted inversions, transpositions & inverted transpositions: Bader & Ohlebusch (2006)

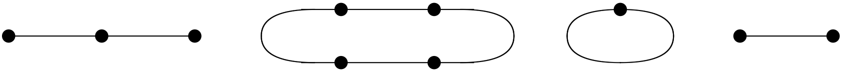
Inversions, translocations, fusions, fissions & block-interchanges: Yancopoulos *et al.* (2005), Bergeron *et al.* (2006)

1. Genome Rearrangement Problems
   - Biological Background
   - Genomic Distances
2. The DCJ Operation
   - Graphs with Vertices of Degree One or Two
   - The Genome Graph
   - The Adjacency Graph
   - Algorithm for Sorting by DCJ Operations
3. Relation to other Genomic Distances
   - The Inversion Distance
   - The Translocation Distance

Let G be a graph where each vertex has degree one or two.

Definitions:

- A vertex of degree one is called external and a vertex of degree two internal
- An internal vertex connecting edges p and q is also denoted by $\{p, q\}$ and an external vertex incident to an edge p by $\{p\}$
- Cycle is a circular component and a path is a linear component
- A cycle or path is even if it has an even number of edges, otherwise it is odd



Definition

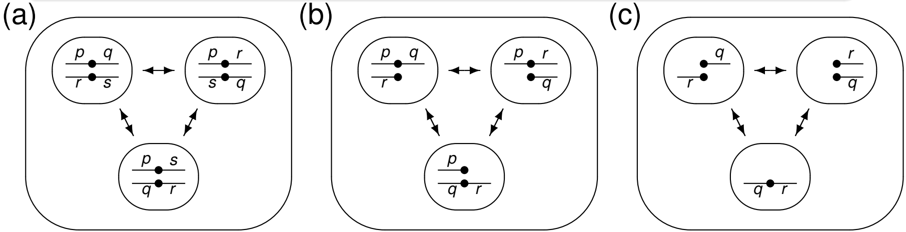
The DCJ operation acts on two vertices u and v of a graph with vertices of degree one or two in one of the following three ways:

(a) If both $u = \{p, q\}$ and $v = \{r, s\}$ are internal vertices, these are replaced by the two vertices $\{p, r\}$ and $\{s, q\}$ or by the two vertices $\{p, s\}$ and $\{q, r\}$.

(b) If $u = \{p, q\}$ is internal and $v = \{r\}$ is external, these are replaced by $\{p, r\}$ and $\{q\}$ or by $\{q, r\}$ and $\{p\}$.

(c) If both $u = \{q\}$ and $v = \{r\}$ are external, these are replaced by $\{q, r\}$.

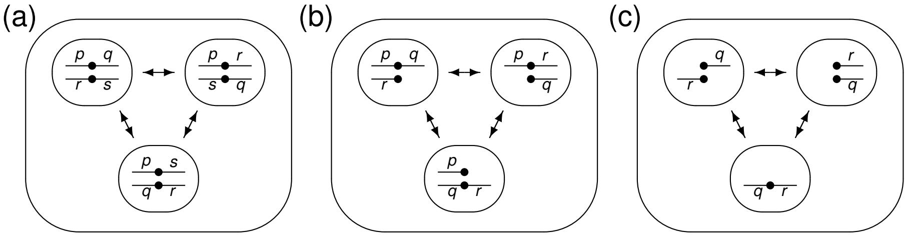
In addition, as an inverse of case (iii), a single internal vertex $\{q, r\}$ can be replaced by two external vertices $\{q\}$ and $\{r\}$.



Global Effects on the Graph

(1) DCJ operation applied on 1 or 2 paths:

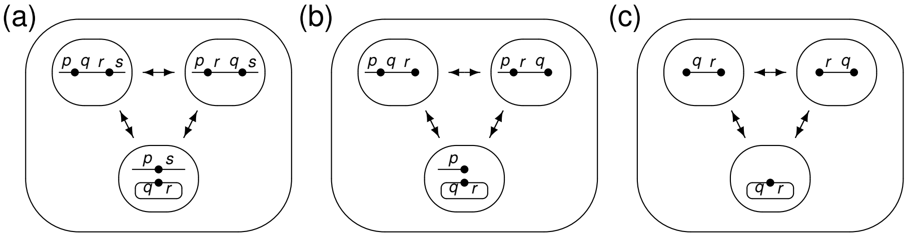
- Path translocation
- Path fusion or path fission



Global Effects on the Graph

(2) DCJ applied on 1 path, or 1 path and 1 cycle:

- Inversions
- Excisions or integrations
- Circularizations or linearizations

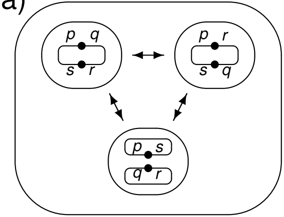


Global Effects on the Graph

(3) DCJ operation applied on 1 or 2 cycles:

- Inversions
- Cycle fusions or cycles fissions

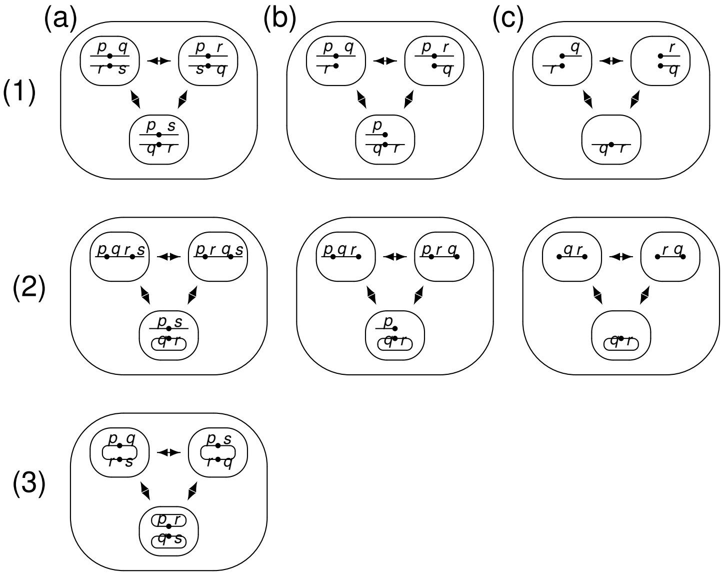
(a)



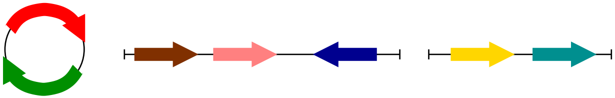
Lemma

The application of a single DCJ operation changes the number of circular or linear components by at most one.

Global Effects on the Graph



Genome Graph



- A gene *a* is an oriented sequence of DNA that starts with a tail a_t and ends with a head a_h
- Head and tail are called the extremities of a gene
- An adjacency of two consecutive genes *a* and *b*, depending on their respective orientation, can be of four different types:

$$\{a_h, b_t\}, \{a_h, b_h\}, \{a_t, b_t\}, \{a_t, b_h\}$$

- An extremity that is not adjacent to any other gene is called a telomere, represented by a singleton set $\{a_h\}$ or $\{a_t\}$

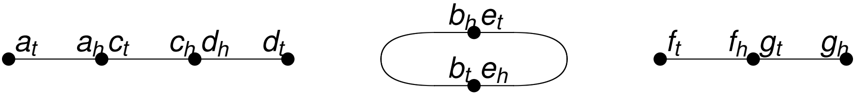
Definition

A genome is a set of adjacencies and telomeres such that the tail or the head of any gene appears in exactly one adjacency or telomere.

$$A = \{\{a_t\}, \{a_h, c_t\}, \{c_h, d_h\}, \{d_t\}, \{b_h, e_t\}, \{e_h, b_t\}, \{f_t\}, \{f_h, g_t\}, \{g_h\}\}$$

Definition

Genome graph: Given a genome, one reconstructs its chromosomes by representing the telomeres and adjacencies as vertices and joining for each gene its tail and its head by an edge.



Observation

The genome graph is a graph with vertices of degree 1 or 2.

$$\begin{aligned} A &= \{\{a_t\},\{a_h,c_t\},\{c_h,d_h\},\{d_t\},\{b_h,e_t\},\{e_h,b_t\},\{f_t\},\{f_h,g_t\},\{g_h\}\} \\ B &= \{\{a_h,b_t\},\{b_h,a_t\},\{c_t\},\{c_h,d_t\},\{d_h\},\{e_t\},\{e_h\},\{f_h,g_t\},\{g_h,f_t\}\} \end{aligned}$$

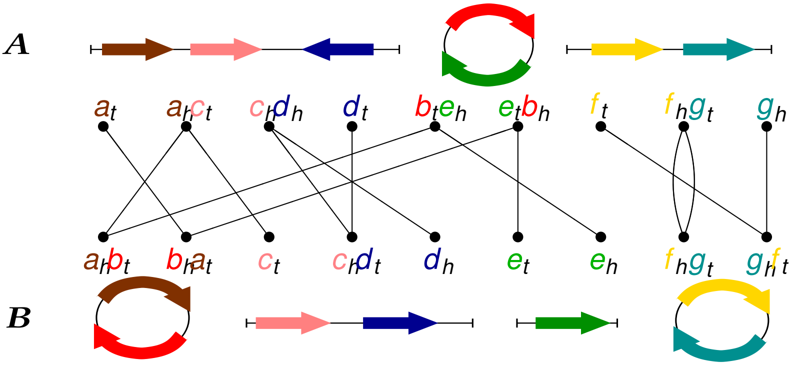
The DCJ Distance Problem

Given two genomes A and B, find a shortest sequence of DCJ operations that transforms A into B. The length of such a sequence is called the DCJ distance between A and B, denoted by $d_{DCJ}(A,B)$.

Adjacency Graph

Definition

The adjacency graph $AG(A, B)$ is a bipartite multi-graph whose set of vertices are the adjacencies and telomeres of A and B. For each $u \in A$ and $v \in B$ there are $|u \cap v|$ edges between u and v.

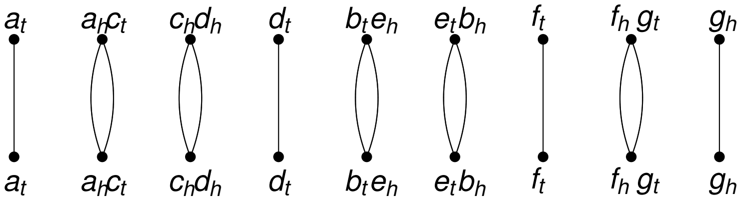


Lemma

Let A and B be two genomes defined on the same set of N genes, then we have

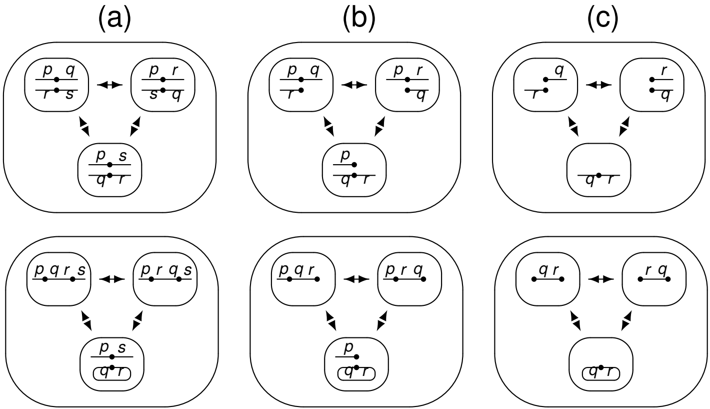
$$A = B \quad \text{if and only if} \quad N = C + I/2$$

where C is the number of cycles and I the number of odd paths in $AG(A, B)$.



Lemma

The application of a single DCJ operation changes the number of odd paths in the adjacency graph by –2, 0, or 2.



Lemma

Let A and B be two genomes defined on the same set of N genes, then we have

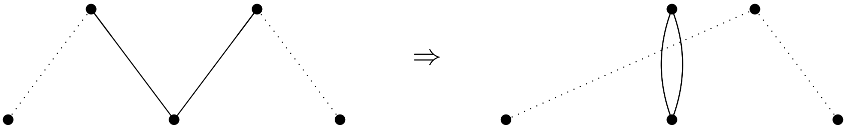
$$d_{DCJ}(A, B) \;\geq\; N - (C + I/2)$$

where C is the number of cycles and I the number of odd paths in $AG(A, B)$.

Sorting by DCJ Operations

1. Generate the adjacencies of *B* that are not yet present in *A*

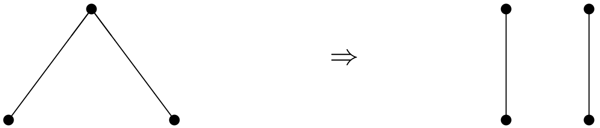
Any pair of edges in the adjacency graph that connect two different vertices of genome *A* with an adjacency $\{p, q\}$ in genome *B* can be transformed by a single DCJ operation into a cycle of length two, plus the remaining structure, reduced by the two edges $\rightarrow$ *C* increases by one!



Sorting by DCJ Operations

2. Generate the telomeres of B that are not yet present in A

All adjacencies of genome B are contained in cycles of length two. There might still be pairs of telomeres of B that form an adjacency in A. These adjacencies can be split into two telomeres, thus creating two odd paths of length one each → I increases by two!



Algorithm for sorting by DCJ operations

```
Let $AG(A, B)$ be the adjacency graph of $A$ and $B$

{Generate the adjacencies of $B$ that are not yet present in $A$}
for each adjacency $\{p, q\}$ in genome $B$ do
    let $u$ be the vertex of $A$ that contains $p$
    let $v$ be the vertex of $A$ that contains $q$
    if $u \neq v$ then
        replace vertices $u$ and $v$ in $A$ by $\{p, q\}$ and $(u \setminus \{p\}) \cup (v \setminus \{q\})$
    end if
end for

{Generate the telomeres of $B$ that are not yet present in $A$}
for each telomere $\{p\}$ in $B$ do
    let $u$ be the vertex of $A$ that contains $p$
    if $u$ is an adjacency then
        replace vertex $u$ in $A$ by $\{p\}$ and $(u \setminus \{p\})$
    end if
end for
```

The DCJ Distance

Theorem (Bergeron, Mixtacki and Stoye 2006)

Let A and B be two genomes defined on the same set of N genes, then we have

$$d_{DCJ}(A, B) \;=\; N - (C + I/2)$$

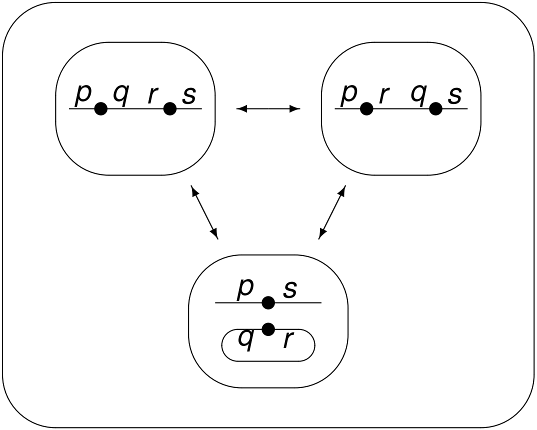
where C is the number of cycles and I the number of odd paths in $AG(A, B)$. An optimal sorting sequence can be found in optimal $O(|A| + |B|)$ time.

1 Genome Rearrangement Problems
- Biological Background
- Genomic Distances

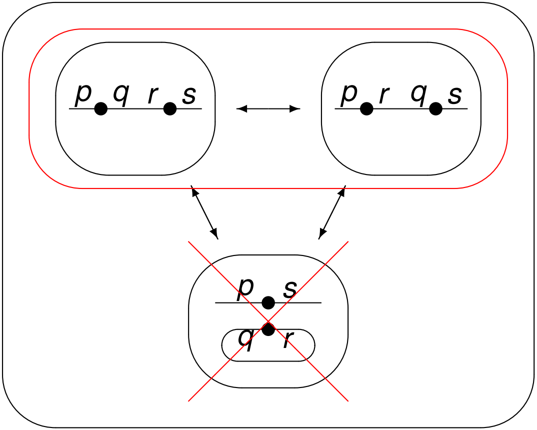
2 The DCJ Operation
- Graphs with Vertices of Degree One or Two
- The Genome Graph
- The Adjacency Graph
- Algorithm for Sorting by DCJ Operations

3 Relation to other Genomic Distances
- The Inversion Distance
- The Translocation Distance

The Inversion Distance Problem



The Inversion Distance Problem

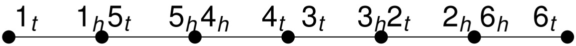


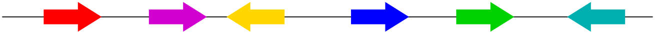
The Inversion Distance Problem

Uni-chromosomal genomes with the same gene content:

- Gene is represented by a signed integer between 1 and N
- Orientation of a gene is represented by the sign

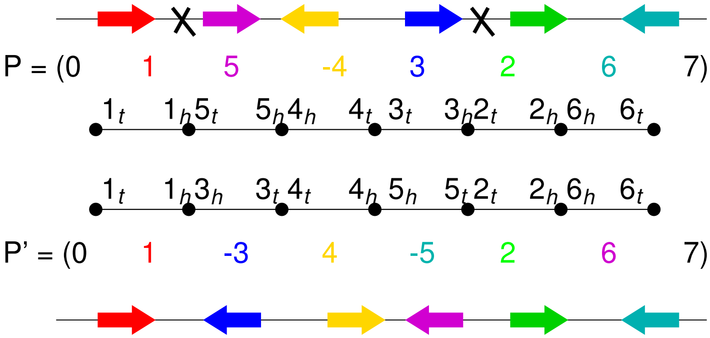
P = (0 1 5 -4 3 2 6 7)





The Inversion Distance Problem

Inversion changes the order and the signs of an interval of genes:



P = (0 1 5 -4 3 2 6 7)

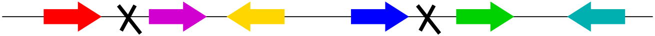
1_t $1_h 5_t$ $5_h 4_h$ $4_t 3_t$ $3_h 2_t$ $2_h 6_h$ 6_t

1_t $1_h 3_h$ $3_t 4_t$ $4_h 5_h$ $5_t 2_t$ $2_h 6_h$ 6_t

P’ = (0 1 -3 4 -5 2 6 7)

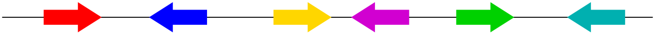
The Inversion Distance Problem

Inversion changes the order and the signs of an interval of genes:



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

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



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

$$\begin{array}{rcccccccc}
P = (0 & 1 & 5 & -4 & 3 & 2 & \underline{-6} & 7) \\
(0 & 1 & \underline{5} & -4 & 3 & 2 & 6 & 7) \\
(0 & 1 & \underline{-5} & \underline{-4} & \underline{3} & \underline{2} & 6 & 7) \\
(0 & 1 & \underline{-2} & -3 & 4 & 5 & 6 & 7) \\
(0 & 1 & 2 & \underline{-3} & 4 & 5 & 6 & 7) \\
Id = (0 & 1 & 2 & 3 & 4 & 5 & 6 & 7)
\end{array}$$

Definition

Inversion distance $d_{Inv}(P)$: minimum number of inversions needed to transform P into the identity permutation

Theorem (Hannenhalli and Pevzner 1995)

For a signed permutation P

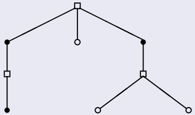
$$d_{Inv}(P) \ = \ N - C - 1 + h + f$$

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 (Bergeron, Mixtacki and Stoye 2004)

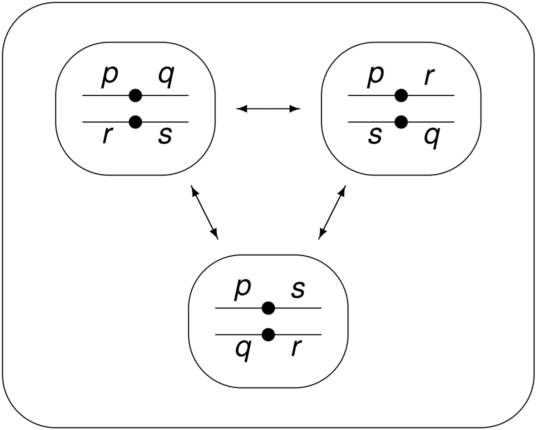
- If a signed permutation P on the set $\{0, \ldots, N-1\}$ has C cycles and the associated tree T_P has minimal cost t, then

$$\begin{aligned} d_{Inv}(P) &= N - C - 1 + t \\ &= d_{DCJ} + t \end{aligned}$$



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

The Translocation Distance Problem



The Translocation Distance Problem

Multi-chromosomal genomes with the same gene content and number of chromosomes:

$$A = \{(4 \quad 3), \quad (1 \quad 2 \quad -7 \quad 5), \quad (6 \quad -8 \quad 9)\}$$

Internal translocation exchanges two non-empty chromosome ends:

$$\begin{aligned} A &= \{(4 \underline{\quad 3}), \quad (1 \quad 2 \quad \underline{-7 \quad 5}), \quad (6 \quad -8 \quad 9)\} \\ A' &= \{(4 \quad -7 \quad 5), \quad (1 \quad 2 \quad 3), \quad (6 \quad -8 \quad 9)\} \end{aligned}$$

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

$$A = \{(4\ \underline{3}), (1\ 2\ \underline{-7\ 5}), (6\ -8\ 9)\}$$

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

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

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

$$B = \{(1\ 2\ 3), (4\ 5), (6\ 7\ 8\ 9)\}$$

Definition

Translocation distance $d(A)$: minimum number of translocations needed to transform A into the identity permutation split in chromosomes sharing the ends of A

Theorem (Hannenhalli 1996)

For a genome *A* with *chr* chromosomes and *N* genes

$$d_{Trans}(A) \ = \ N - C - chr + s + o + 2i$$

where *C* is the number of cycles, *s* the number of minimal subpermutations, $o = 1$ if the number of minimal subpermutations is odd and $o = 0$ otherwise, and $i = 1$ if *P* has an even-isolation and $i = 0$ otherwise.

Summary of our Results (Bergeron, Mixtacki and Stoye 2005)

- Let *A* be a genome with *C* cycles and whose forest F_A has *L* leaves and *T* trees. Then

$$\begin{aligned} d_{Trans}(A) &= N - C - chr + t \\ &= d_{DCJ} + t \end{aligned}$$

where

$$t = \begin{cases} L+2 & \text{if } L \text{ is even and } T = 1 \\ L+1 & \text{if } L \text{ is odd} \\ L & \text{if } L \text{ is even and } T \neq 1. \end{cases}$$

- First correct algorithm for sorting by translocations.

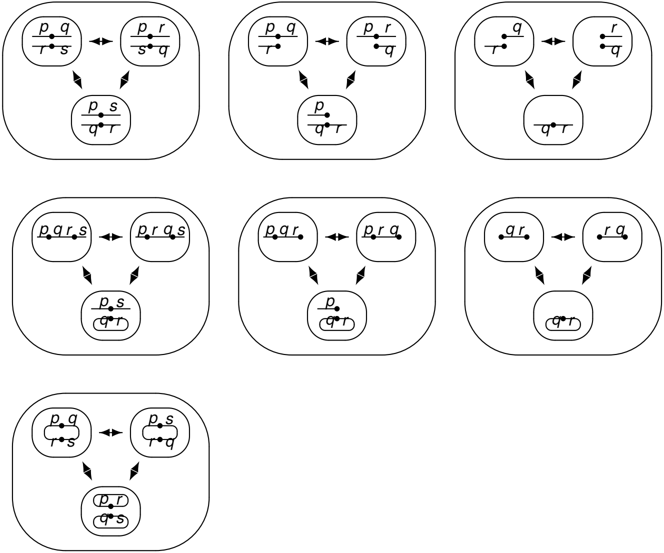
Summary

- Representation of genomes containing linear and circular chromosomes
- All classical operations are modeled by DCJ
- Simple DCJ distance formula
- Linear-time algorithm for sorting by DCJ operations
- Relation to other well-studied models:

$$d(A, B) = d_{DCJ}(A, B) + t$$

where *t* represents the additional cost of not resorting to DCJ operations.

Global Picture on Genome Rearrangement Models



Thanks to:

Julia Mixtacki (Bielefeld)

Anne Bergeron (Montreal)

... and you for your attention!!!

Questions?