18 Genome rearrangement

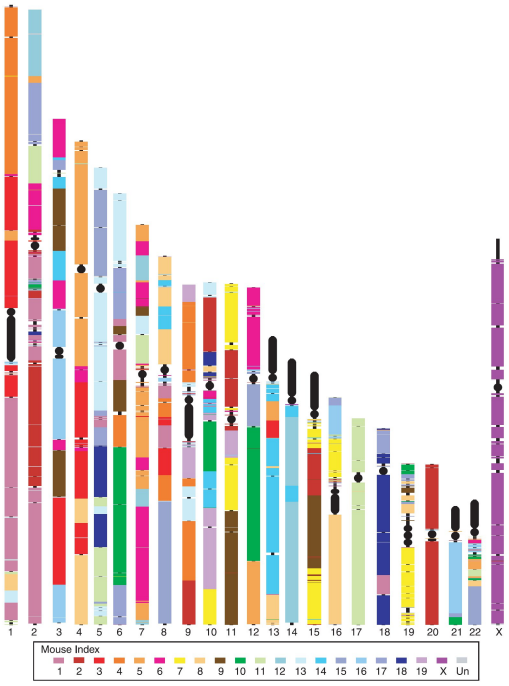
This exposition is based on:

1. Tannier, E., Zheng C., Sankoff, D. (2009) *Multichromosomal median and halving problems under different genomic distances.* BMC Bioinformatics 10: 120.
2. Bergeron, A., Mixtacki, J., Stoye, J. (2006) *A unifying view of genome rearrangements.* In: Proceedings of the 6th International Workshop on Algorithms in Bioinformatics (WABI), LNBI 4175:163–173.
3. Hannenhalli, S., Pevzner, P. (1999) *Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals.* ACM 46(1):1–27.

18.1 Introduction

Genomes may be identical in gene sequence, but differ in gene order. By looking at genome rearrangement, we study the organization and evolution of genomes at the level of blocks (genes), not at single characters.

Genome rearragement is a common mode of molecular evolution especially in bacteria, but also occurs in other organisms. Evolution of genomes by rearrangement can happen at a different rate than single character substitutions, but both modes of evolution should be taken into account to infer ancestral sequences or phylogenetic trees.



Source: Sankoff, D., Eichler, E. (Science 2003)

18.2 Modelling genomes

A *block (gene)* is an oriented segment of DNA.

We model blocks by positive integer numbers b.

Each block has two *extremities*, a *head* b^h representing the start (5′-end) and a *tail* b^t representing the end (3′-end) of the block in forward orientation.

We model the orientation of a block by adding a sign to the block's integer identifier: '+' for forward orientation from head to tail and '−' for the reverse complement from tail to head.

A genome without duplications can be modeled as signed permutation on a set of positive integer numbers, for example $G = (+2, -1, +3, +4)$.

Karyotypes of genomes:

Genomes can consist of multiple chromosomes. Then, a genome is a set of chromosomes and each chromosome is a sequence of signed integer numbers.

A genome with only one chromosome is called *unichromosomal*, a genome with more than one chromosome *multichromosomal*.

Chromosomes can be *linear* or *circular*. Linear chromosomes are bounded by zero length *telomeres*. We denote telomeres by $\bullet$.

Examples:
Linear chromosome: $(\bullet, -3, +1, +2, -5, -4, +6, \bullet) = (\bullet, -6, +4, +5, -2, -1, +3, \bullet)$
Circular chromosome: $(+2, -1, +3, +4) = (+4, +2, -1, +3) = (-3, +1, -2, -4)$

An *adjacency* is an unordered pair of extremities. Ends of linear chromosomes are *telomeric adjacencies*.

Definition 1. Given a set of blocks (integer numbers), a *genome* is a set of adjacencies such that head and tail of each block appears in exactly the same number of adjacencies.

For example, the circular unichromosomal genome $(+2, -1, +3, +4)$ consists of four adjacencies $\{2^t, 1^t\}, \{1^h, 3^h\}, \{3^t, 4^h\}, \{4^t, 2^h\}$, and the linear genome $(\bullet, +2, +3, \bullet), (\bullet, +1, -5, -4, \bullet)$ with two chromosomes consistis of seven adjacencies $\{\bullet, 2^h\}, \{2t, 3^h\}, \{3^t, \bullet\}, \{\bullet, 1^h\}, \{1^t, 5^t\}, \{5^h, 4^t\}, \{4^h, \bullet\}$.

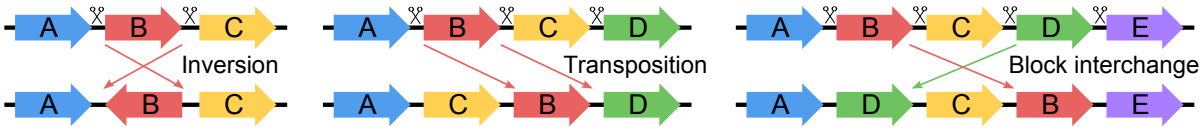
Note that only linear unichromosomal genomes without duplications and without gain/loss can be modeled as signed permutations (each extremity of a block appears exactly once in the set of adjacencies).

18.3 Rearrangement operations

In comparative genomics, the order and/or content of a genome with respect to one or more other genomes is examined.

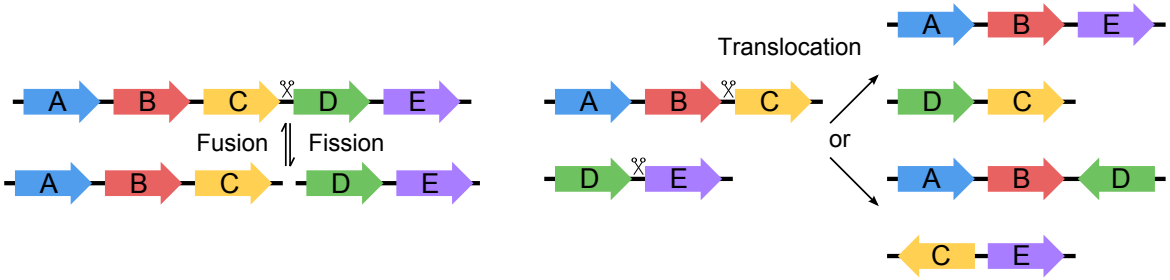
Operations among genomes:

- Rearrangement operations that affect the order and orientation of blocks within a chromosome include *inversion (reversal)*, *transposition*, and *block interchange*.



- The content of a genome can be changed by *gain/loss* and *duplication*. These are operations that affect the copy number of blocks.

- Multichromosomal genomes can also evolve by *translocation* and *fusion/fission*.



In order to measure similarity of genomes in terms of rearrangements several pairwise distances have been introduced.

Here, we examine the

- breakpoint distance,
- reversal distance, and
- double-cut-and-join (DCJ) distance.

Based on such a distance measure a number of computational problems can be solved including the *median problem* and the *halving problem*. The complexity of these problems depends on the choice of distance *and* karyotypic framework.

18.4 Breakpoint distance

The breakpoint distance is a simple distance that does not differentiate between different evolutionary operations, but instead aims to count the minimal number of breakage events that are necessary to transform one genome into another.

Definition 2. A *breakpoint* among two genomes is an adjacendy from one genome that is disrupted in the other.

The breakpoint distance $d_{BP}(G_1, G_2)$ counts the number of breakpoints among two genomes G_1 and G_2.

Note that the same breakpoint can be used by multiple rearrangement operations, i. e. it can be re-used. The breakpoint distance is thus a lower bound to the actual number of breakage events that happened during evolution.

Given two genomes G_1 and G_2 on a set of n blocks, the breakpoint distance can be defined as:

$$d_{BP}(G_1, G_2) := n - a(G_1, G_2) - \frac{t(G_1, G_2)}{2}$$

$a(G_1, G_2)$ .. number of adjacencies from G_1 conserved in G_2

$t(G_1, G_2)$ .. number of telomere adjacencies from G_1 conserved in G_2

This definition of the breakpoint distance is symmetric, i. e. $d_{BP}(G_1, G_2) = d_{BP}(G_2, G_1)$. This is also true for genomes that do not consist of the same set of blocks.

In genomes that consist of the same set of blocks the breakpoint distance equals the number of breakpoints per genome. If there is gain/loss between the genomes, the number of breakpoints can differ from the breakpoint distance, and the breakpoint count may be asymmetric.

Breakpoint graphs:

Given two genomes on a set of n blocks, the *breakpoint graph* contains one vertex per block extremity. Edges in the breakpoint graph correspond to adjacencies within the genomes. Sometimes a second type of edge is introduced to connect the heads and the tails of each block.

Breakpoint graphs are used in a number of different ways. Depending on the application a single telomere vertex or one telomere vertex per extremity with edges according to telomere adjacencies may be added to the graph.

18.5 Reversal distance

The first polynomial time algorithm to compute the reversal distance for linear genomes was introduced by Hannenhalli and Pevzner.

The reversal distance of two genomes is defined as the minimal number of inversions necessary to transform one genome into the other.

"Sorting by reversals"

The reversal distance can be computed in polynomial time by inspecting the breakpoint graph of two genomes for certain substructures; ("hurdles", "fortresses", and cycles).

Given two genomes G_1 and G_2 on a set of n blocks, the reversal distance is

$$d_{REV} = n - c(G_1, G_2) - 1 + h(G_1, G_2) + f(G_1, G_2)$$

$c(G_1, G_2)$ .. number of cycles

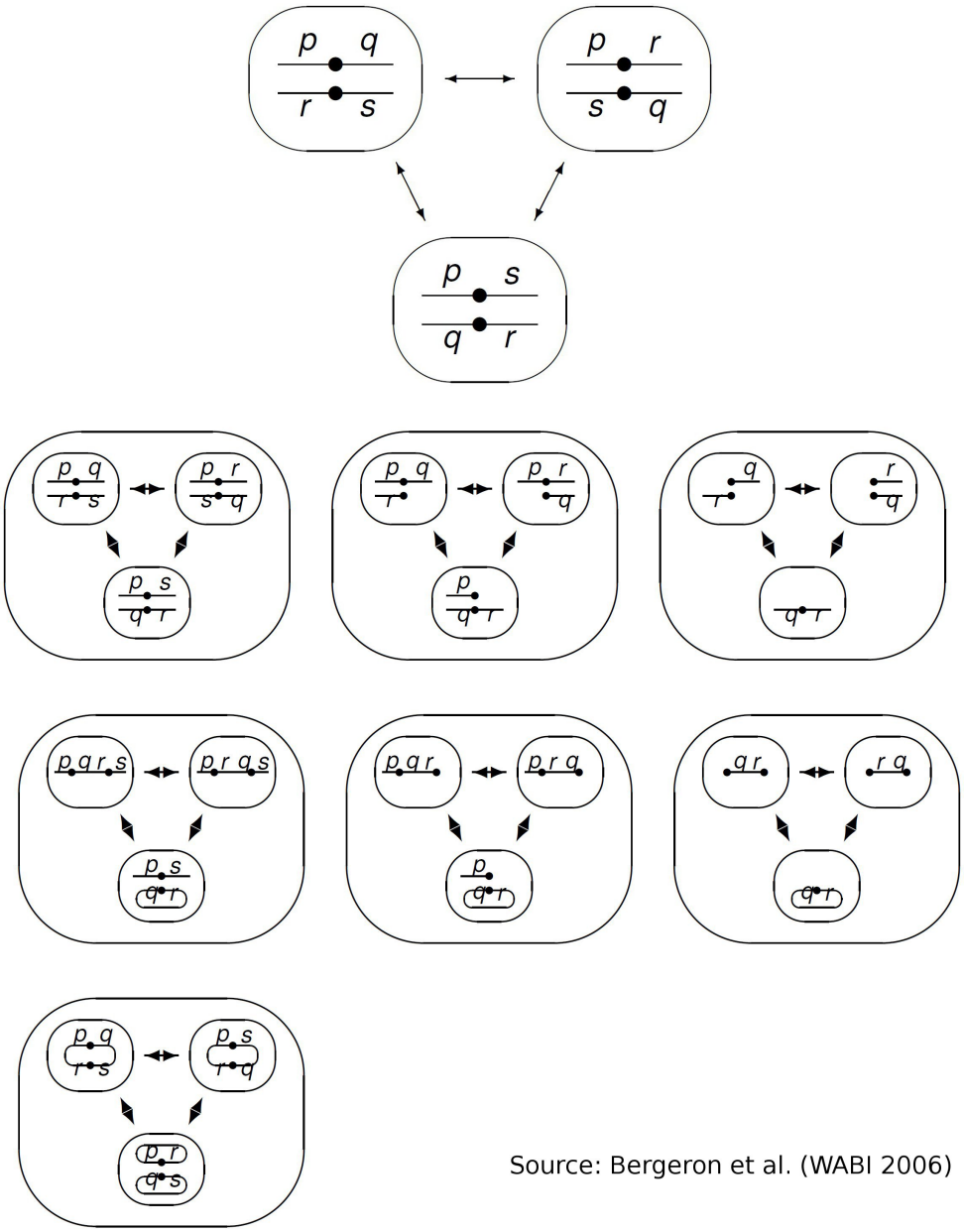
$h(G_1, G_2)$ .. number of hurdles

$f(G_1, G_2) = 1$ if G_1, G_2 have a fortress, otherwise 0

18.6 Double-cut-and-join (DCJ) distance

A very general karyotypic framework together with the DCJ operation has been introduced by Yancopoulos et al. in 2005. They allow multichromosomal genomes and both circular and linear chromosomes. It is defined without gain/loss and without duplications, but extensions have been published for both of these operations.

The DCJ operation acts on two (telomere) adjacencies $\{p, q\}$ and $\{r, s\}$. It cuts the two adjacencies and reconnects (joins) the four loose ends to form two new adjacencies, $\{p, r\}$ and $\{q, s\}$, or $\{p, s\}$ and $\{q, r\}$.



Source: Bergeron et al. (WABI 2006)

Depending on whether the two adjacencies involved in a DCJ operation are telomeric or not, lie on the same chromosome or on different chromosomes, and lie on circular or linear chromosomes, the karyotype of the genome may change.

We may represent a genome by its *genome graph* to better visualize the effect of DCJ operations:

Given a genome consisting of adjacencies and telomere adjacencies, the genome graph contains a vertex for each adjacency and for each telomere adjacency. Two vertices are connected if their adjacencies contain an extremity from the same block.

Given two genomes G_1 and G_2 defined on the same set of blocks. Find a shortest sequence of DCJ operations that transforms G_1 into G_2.

The length of a shortest sequence of DCJ operation is the *DCJ distance* $d_{DCJ}(G_1, G_2)$ of the two genomes G_1 and G_2.

The genome graph represents a single genome. To model two genomes within the DCJ framework, the *adjacency graph* has been introduced:

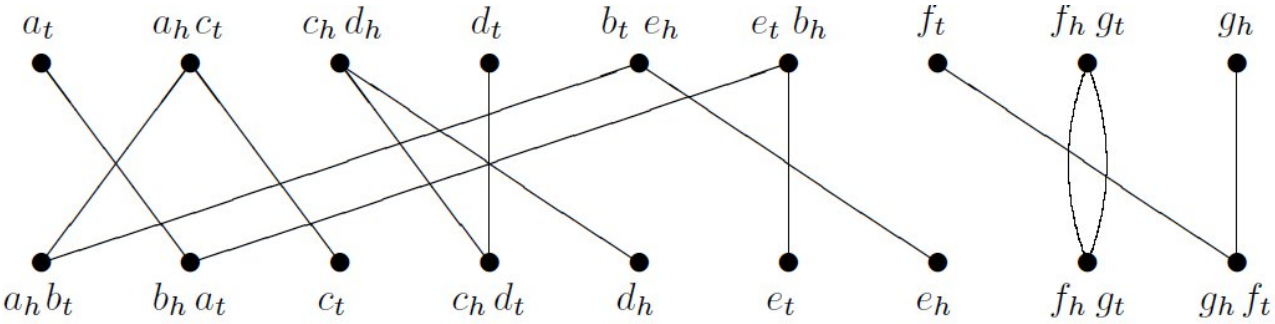
Given two genomes G_1 and G_2, the adjacency graph contains two disjoint sets of vertices, V_1 for G_1 and V_2 for G_2, with vertices for each adjacency from the genomes. Two vertices $v_1 \in V_1$ and $v_2 \in V_2$ are connected by an edge if they share an extremity.

Note that the adjacency is a multigraph; if G_1 and G_2 share an adjacency, the two corresponding vertices are connected by two edges that form a cycle.

Example 3. Adjacency graph

$A = \left\{\{\bullet, a^t\}, \{a^h, c^t\}, \{c^h, d^h\}, \{d^t, \bullet\}, \{b^h, e^t\}, \{e^h, b^t\}, \{\bullet, f^t\}, \{f^h, g^t\}, \{g^h, \bullet\}\right\}$

$B = \left\{\{a^h, b^t\}, \{b^h, a^t\}, \{c^t, \bullet\}, \{c^h, d^t\}, \{\bullet, d^h\}, \{\bullet, e^t\}, \{e^h, \bullet\}, \{f^h, g^t\}, \{g^h, f^t\}\right\}$



The adjacency graph has two interesting properties, the number of cycles and the number of paths with an odd number of edges.

Theorem 4. *Given two genomes G_1 and G_2 defined on the same set of n blocks. The DCJ distance of the two genomes is*

$$d_{DCJ}(G_1, G_2) = n - \left(c(G_1, G_2) + \frac{p(G_1, G_2)}{2}\right)$$

where $c(G_1, G_2)$ is the number of cycles and $p(G_1, G_2)$ the number of odd paths in the adjacency graph of G_1 and G_2.

The following algorithm computes the DCJ distance in $O(n)$ time:

18.7 The median problem

The median problem aims to reconstruct an ancestral genome, the common ancester of two genomes given a third genome as an outgroup.

Given three genomes G_1, G_2, and G_3. A *median genome M* is a genome that minimizes

$$d(G_1, M) + d(G_2, M) + d(G_3, M) \quad .$$

```
(1)  for each adjacency {p, q} ∈ G2 do
(2)      let u be the element of genome G1 that contains p
(3)      let v be the element of genome G1 that contains q
(4)      if u ≠ v
(5)        then replace u and v in G1 by {p, q} and (u \ {p}) ∪ (v \ {q})
(6)      fi
(7)  od
(8)  for each telomere adjacency {p, •} ∈ G2 do
(9)      let u be the element of genome G1 that contains p
(10)     if u is an adjacency
(11)       then replace u in G1 by {p, •} and (u \ {p}) ∪ {•}
(12)     fi
(13) od
```

The median problem can be solved in polynomial time for multichromosomal genomes of circular or mixed karyotype using the breakpoint distance. For all other combinations of distances and karyotypes, the computation is either NP-hard or unknown but expected to be NP-hard (Tannier *et al.*).

A median genome allows us to make assumptions on the evolutionary history of extant genomes. Under the assumption of parsimony, a median genome represents an ancestral genome.

Given a set of extant genomes and a phylogenetic tree topology, the *small phylogeny problem* is to assign ancestral sequences to the internal nodes of the phylogenetic tree.

18.8 The halving problem

The goal of halving is to reconstruct the ancestor of a genome that has undergone whole-genome duplication at the time of the duplication event.

An *all-duplicates genome* is a genome that contains each block exactly twice.

The *double distance* of an all-duplicates genome G_1 and an ordinary genome G_2 is the asymmetric distance $d(G_1, G_2)$ between the all-duplicates genome and the union of two copies of the ordinary genome's set of chromosomes.

Given an all-duplicates genome G_1 and an ordinary genome G_2. The *halving problem* is to find an ordinary genome that minimizes the double distance of G_1 and G_2.

The complexity of the halving problem is not known for some combinations of karyotypes and distances, but can be solved in polynomial time for the known ones.

18.9 Summary

- The arrangement of ancestral genomes can be reconstructed using genomic rearrangement distances, such as the breakpoint distance, the reversal distance, or the double-cut-and-join (DCJ) distance.
- The complexity of problems such as the median and halving problem depends on the choice of genomic distance and on the karyotypic framework.