The reconstruction of doubled genomes

Nadia El-Mabrouk

University of Montreal,
Canada

Genome rearrangement: Compare gene orders between genomes.

Genomes = sequences of signed genes (or blocs).

One copy of each gene

b -a d -e -c f

BUT: Usually, many copies of each gene

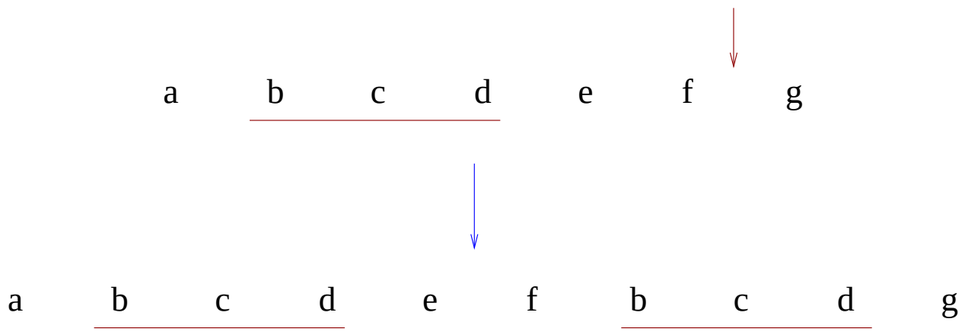
b -a d a -e -c e f d a

Sankoff 1999 ; *Marron, Swenson, Moret 2003*

Problematics: Find the ancestor of a genome with multiple gene copies.

Multigene families due to:

- Single gene duplication;
- Duplication transposition of chromosomal segments;



- Genome-wide doubling events.

Plan

1. Introduction on genome duplication;
2. The *Hannenhalli and Pevzner* theory;
3. Genome halving

   N. El-Mabrouk, D. Bryant, D. Sankoff, RECOMB, 1999

   N. El-Mabrouk and D. Sankoff, SIAM, J. Comp., 2003

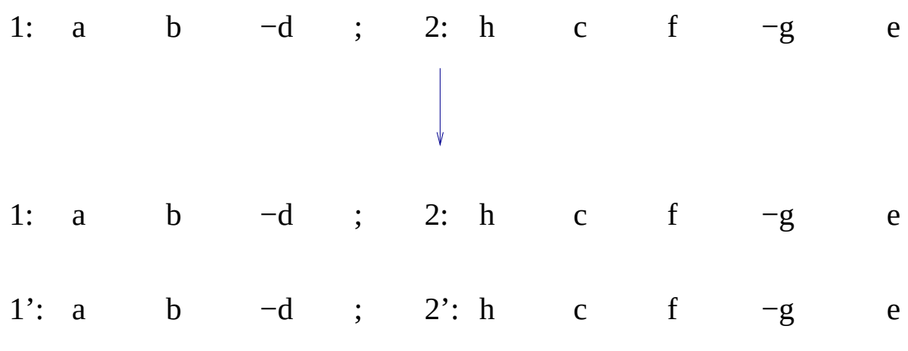
4. Applications to real genomes (yeast and mitochondria);
5. Duplication transposition of chromosomal segments

   N. El-Mabrouk, J. Comp. Sys. Sci., 2002

6. Conclusion.

I. Introduction on genome duplication

Genome duplication or polyploidy:



Source of rapid evolutionary progress.

Evidence across the eukaryote spectrum, two duplications for the vertebrate genome. Particularly prevalent in plants (rice, oats, corn, wheat, soybeans, Arabodopsis ···)

Wolfe, Shields 1997: Traces of duplication in Saccharomyces cerevisiae. 55 duplicated regions representing 50% of the genome.

$\longrightarrow$ From 8 to 16 chromosomes

I : $+2 \bullet -1$
II : $+4 \bullet -3 -7 +\mathbf{8} -5 +6$
III : $+9 \bullet -10 -11$
IV : $+20 +12 +12 +54 +15 +21 \bullet -3 -13 -16 +17 -24 -22 -14$
$-23 -19 +18 -9$
V : $+28 \bullet -25 -27 -4 -26 -13$
VI : $+55 \bullet -36$
VII : $+36 +25 +26 +32 +\mathbf{6} -\mathbf{33} +5 \bullet -30 -34 -31 -29$
VIII : $+35 \bullet -14 -37 -29 -1$
IX : $+38 +39 +27 \bullet$
X : $+10 +40 +41 \bullet -28 -42$
XI : $+42 +40 +43 +35 \bullet -41 -52 -38$
XII : $+53 \bullet -53 -31 -55 -16 -\mathbf{18} -\mathbf{17} -45 -30 -15 -44$
XIII : $+46 +44 +19 \bullet -43 -54 -48 -47 -46$
XIV : $+49 +20 +37 +50 +39 \bullet -11$
XV : $+49 +21 \bullet -22 -52 -50 -23 -45 -51 -47 -2$
XVI : $+48 +32 +33 +51 +8 +24 \bullet -7 -34$

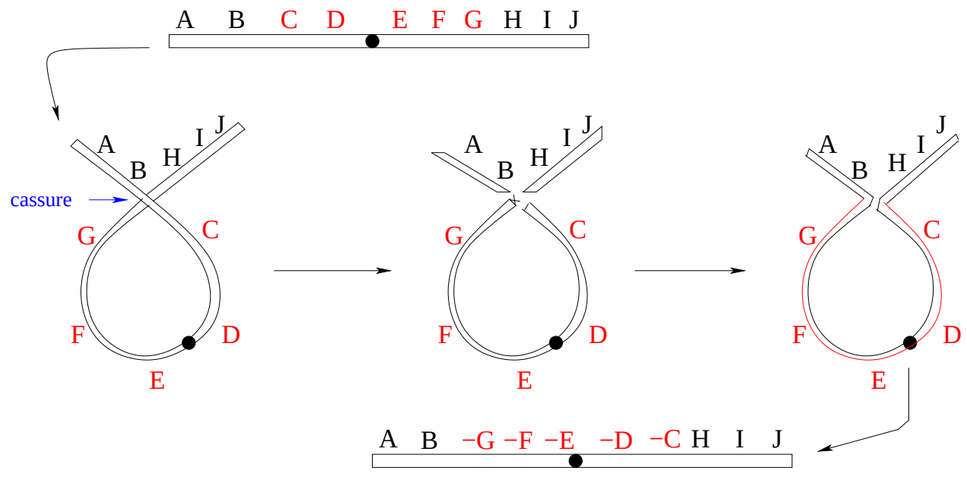
Originally, duplicated genome → two identical copies of each chromosome.

After rearrangements, duplicated chromosomal segments scattered among the genome.

Present-day genome: Signed gene sequences, two copies of each gene.

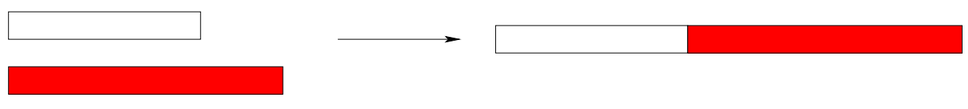
Problem: Reconstruct original gene order at time of duplication.
Minimal number of reversal and/or reciprocal translocations.

Inversion (reversal):

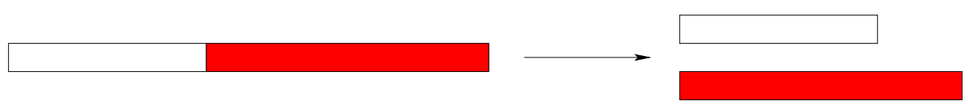


Reciprocal translocation:

Fusion:



Fission:



Problem: Minimum number of inversion and/or translocation, fusion, fission transforming a rearranged duplicated genome G into a perfect duplicated genome H.

Multi-chromosomal case: H has an even number of chromosomes. Not necessarily the case for G.

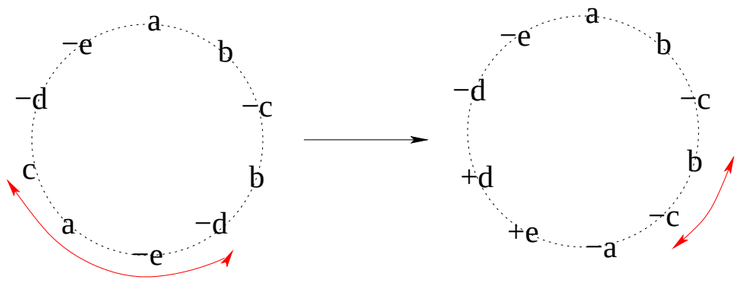
Rearranged duplicated genome:

$$\begin{array}{ll} 1\colon +a\ +b\ -c\ +b\ -d; & 3\colon -e\ +g\ -f\ -d; \\ 2\colon -c\ -a\ +f; & 4\colon +h\ +e\ -g\ +h. \end{array}$$

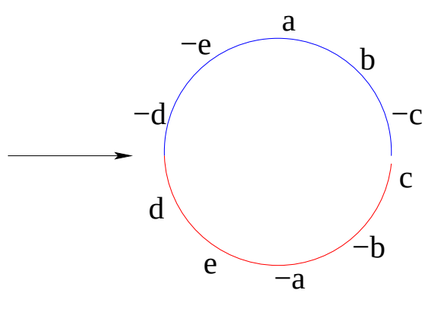
Duplicated genome:

$$\begin{array}{ll} 1\colon +a\ +b\ -d; & 3\colon +h\ +c\ +f\ -g\ +e; \\ 2\colon +a\ +b\ -d; & 4\colon +h\ +c\ +f\ -g\ +e. \end{array}$$

The circular case:



Rearranged genome



Ancestral duplicated genome

Genome rearrangement: Minimum number of rearrangements to transform one signed genome into another.

First polynomial algorithm by *Hannenhalli and Pevzner (1995)*, for:

- reversals only;
- translocations only;
- reversals and translocations.

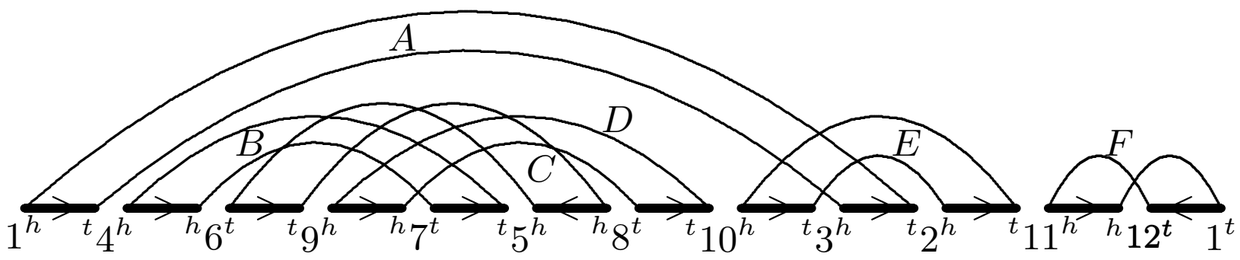
Approach for genome duplication: Find an ancestral genome of G minimizing the HP formula.

II. The Hannenhalli and Pevzner theory

Algorithm based on a breakpoint graph.

G_1: $+1\ +4\ -6\ +9\ -7\ +5\ -8\ +10\ +3\ +2\ +11\ +12$

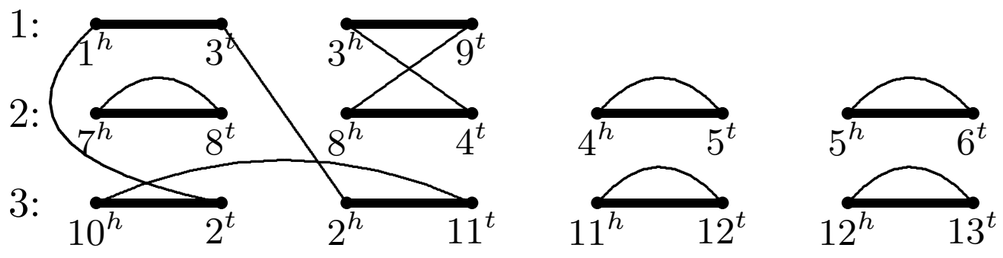
G_2: $+1\ +2\ +3\ +4\ \cdots\ +12$



Multichromosomal case:

G_1: **I**: 1 3 9 **II** : 7 8 4 5 6; **III** : 10 2 11 12 13.

G_2: **I**: 1 2 3 4 5 6 **II** : 7 8 9; **III** : 10 11 12 13.



When $G_1 = G_2$, the number of cycles is maximized

$\longrightarrow$ Perform reversals increasing the number of cycles.

Good component: Can be solved by "good" reversals;

Bad component: Requires "bad" reversals to be solved.

Minimal number $RO(G_1, G_2)$ of rearrangement operations transforming G_1 into G_2:

$$\mathbf{HP}: \quad \mathbf{RO(G_1, G_2) = b - c + m + f}$$

- b: Number of black edges;
- c: Number of cycles;
- m: Number of bad components;
- f: Correction of 0, 1 or 2.

c is the dominant parameter in HP.

III. Genome halving

Partial graph for G:

1: $\overline{O_{11}\quad a_1^t}$ $\overline{a_1^h\quad b_1^t}$ $\overline{b_1^h\quad c_1^h}$ $\overline{c_1^t\quad b_2^t}$ $\overline{b_2^h\quad d_1^h}$ $\overline{d_1^t\quad O_{12}}$

2: $\overline{O_{21}\quad c_2^h}$ $\overline{c_2^t\quad a_2^h}$ $\overline{a_2^t\quad f_1^t}$ $\overline{f_1^h\quad O_{22}}$

3: $\overline{O_{31}\quad e_1^h}$ $\overline{e_1^t\quad g_1^t}$ $\overline{g_1^h\quad f_2^h}$ $\overline{f_2^t\quad d_2^h}$ $\overline{d_2^t\quad O_{32}}$

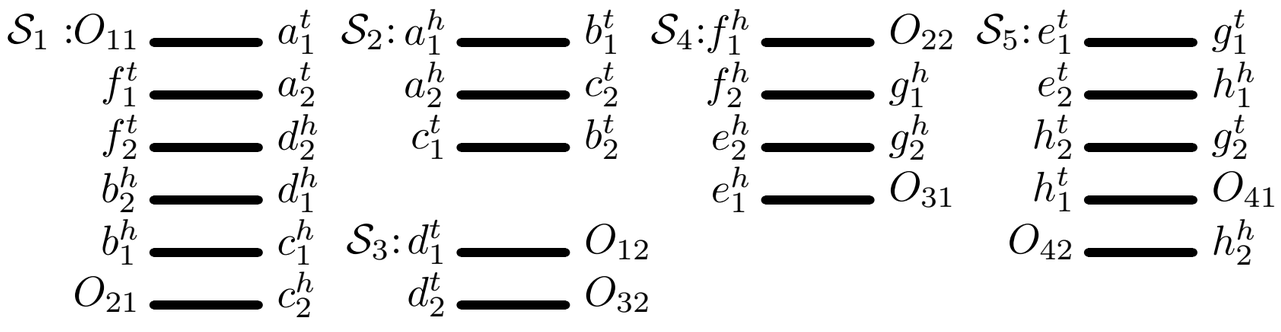
4: $\overline{O_{41}\quad h_1^t}$ $\overline{h_1^h\quad e_2^t}$ $\overline{e_2^h\quad g_2^h}$ $\overline{g_2^t\quad h_2^t}$ $\overline{h_2^h\quad O_{42}}$

Set of valid gray edges: Represent a duplicated genome.

Problem: Find a set of valid gray edges minimizing formula **HP**.

Decomposition into subgraphs

Natural graphs:



Natural graphs of even size are completable.

Amalgamate natural graphs into completable supernatural graphs.

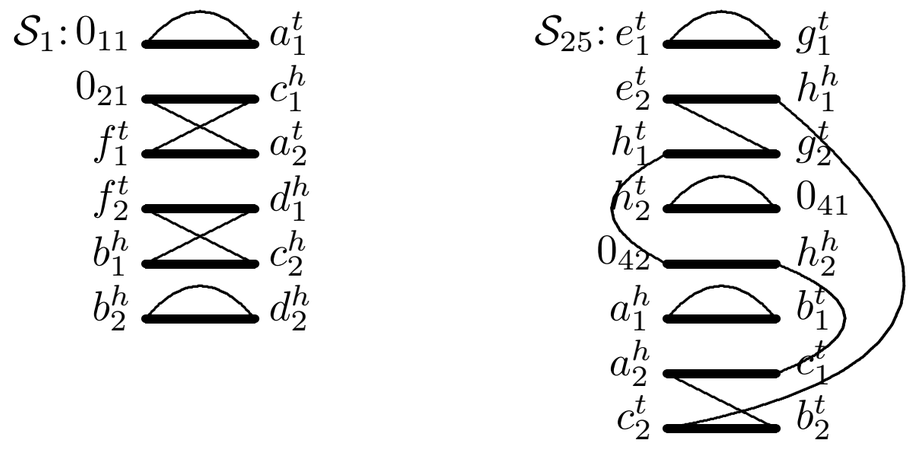
Example: Amalgamate $\mathcal{S}_2$ and $\mathcal{S}_5$

$\longrightarrow$ $\mathcal{S}_1, \mathcal{S}_{25}, \mathcal{S}_3, \mathcal{S}_4$ are supernatural graphs.

Upper bound on the number of cycles

$\mathcal{G}_e$ a supernatural graph of n edges, $\mathcal{G}_e(\Gamma_e)$ a completed graph, and c_e its number of cycles.

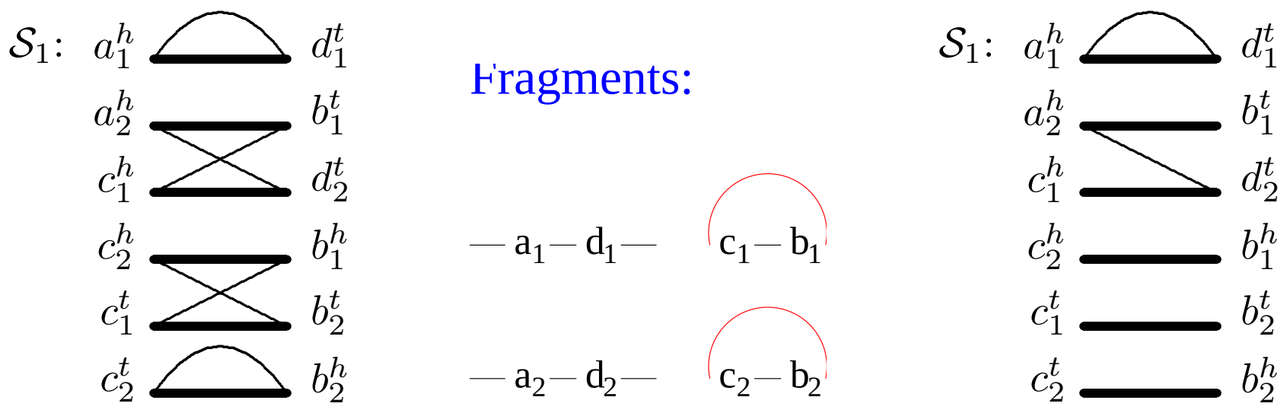
- If $\mathcal{G}_e$ is not amalgamated, $c_e \leq \frac{n}{2} + 1$;
- Otherwise, $c_e \leq \frac{n}{2}$.



Maximizing the number of cycles - Multichromosomal case

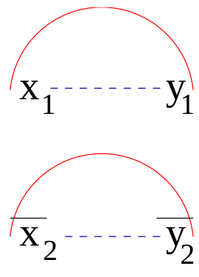
Complete each supernatural graph separately.

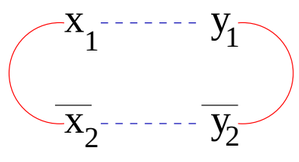
Avoid to create circular fragments.



Maximizing the number of cycles 2

Gray edges creating circular fragments:





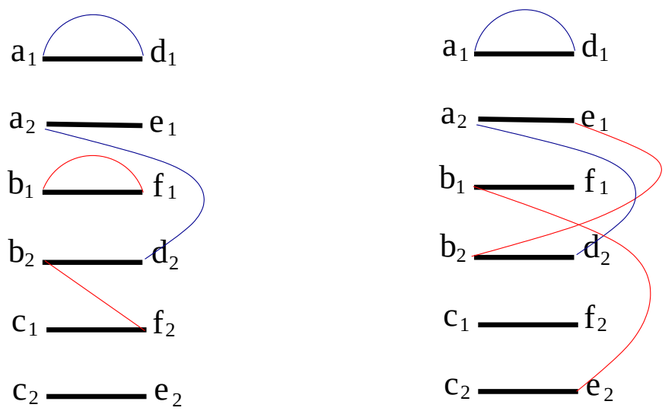
A pair of gray edges not creating circular fragments and not creating a bad subgraph is possible.

Maximizing the number of cycles 3

Algorithm dedouble:

2−edges graph:

n−edges graph:



Linear time algorithm constructing a maximal completed graph containing c cycles:

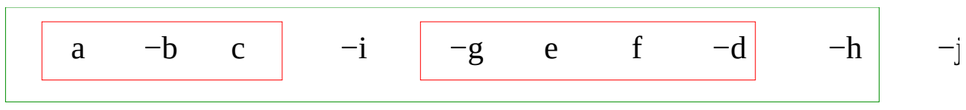
$$c = n/2 + \gamma$$

- n: Number of black edges;
- γ: Number of natural graphs (not amalgamated).

Bad components 1

"Bad components" related to subpermutations (SP) in the HP theory (conserved intervals)

minSP: SP not included in any SP.

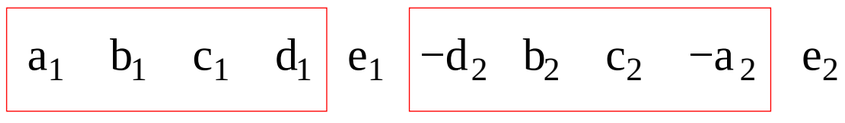


Bijection between SPs and components.

- Rearrangement by translocations: Bad components = minSPs;
- Rearrangement by reversals and/or translocations: Bad components ⊂ minSPs.

Bad components 2

Local SPs:



Lemma: In a maximal completed graph, if ∃ minSP that is not a local SP, then correction to eliminate the minSP.

Corollary: If G does not contain a local SP, then duplicated genome H produced by the algorithm is such that $RO(G, H)$ is minimal.

Bad components 3

General case:

$$RO(G) = \frac{b}{2} - \gamma(G) + m(G) + \phi(G)$$

- b: nb of black edges.
- $\gamma(G)$: nb of natural graphs;
- $m(G)$: nb of "bad" local SPs;
- $\phi(G)$: correction depending on local SPs.

Multichromosomal case: Exact algorithm producing an ancestral genome such that $RO(G, H) = RO(G)$;

Circular case: Uncertainty of up to two reversals.

IV. Application - The yeast genome

Yeast genome: Degenerate tetraploid, duplication 10^8 years ago (*Wolfe and Shields, 1997*). 55 duplicated regions.

I	:	$+2 \bullet -1$
II	:	$+4 \bullet -3 -7 +\mathbf{8} -5 +6$
III	:	$+9 \bullet -10 -11$
IV	:	$+20 +12 +12 +54 +15 +21 \bullet -3 -13 -16 +17 -24 -22 -14$
		$-23 -19 +18 -9$
V	:	$+28 \bullet -25 -27 -4 -26 -13$
VI	:	$+55 \bullet -36$
VII	:	$+36 +25 +26 +32 +\mathbf{6} -\mathbf{33} +5 \bullet -30 -34 -31 -29$
VIII	:	$+35 \bullet -14 -37 -29 -1$
IX	:	$+38 +39 +27 \bullet$
X	:	$+10 +40 +41 \bullet -28 -42$
XI	:	$+42 +40 +43 +35 \bullet -41 -52 -38$
XII	:	$+53 \bullet -53 -31 -55 -16 -\mathbf{18} -\mathbf{17} -45 -30 -15 -44$
XIII	:	$+46 +44 +19 \bullet -43 -54 -48 -47 -46$
XIV	:	$+49 +20 +37 +50 +39 \bullet -11$
XV	:	$+49 +21 \bullet -22 -52 -50 -23 -45 -51 -47 -2$
XVI	:	$+48 +32 +33 +51 +8 +24 \bullet -7 -34$

Sorting by translocations: 45 translocations.

Sorting by inversions+translocations: No local SPs, thus no reversals. Still 45 translocations.

IV. Application - A circular genome

Mitochondrial genome of *Marchantia polymorpha*: Many genes in two or three copies (Oda *et al.*, 1992). Unlikely to be a tetraploid.

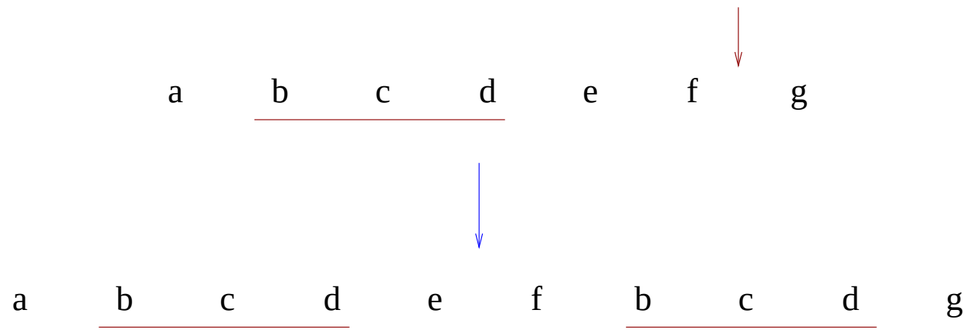
A map with 25 pairs of genes was extracted from the Genbank entry.

$\longrightarrow$ minimum of 25 inversions

Similar to a random distribution $\Longrightarrow$ No trace of duplication.

V. Duplication of chromosomal segments

Duplication of entire regions from one location to another in the genome.



Very recent segment duplication in the human genome (*Eichler et al., 1999*).

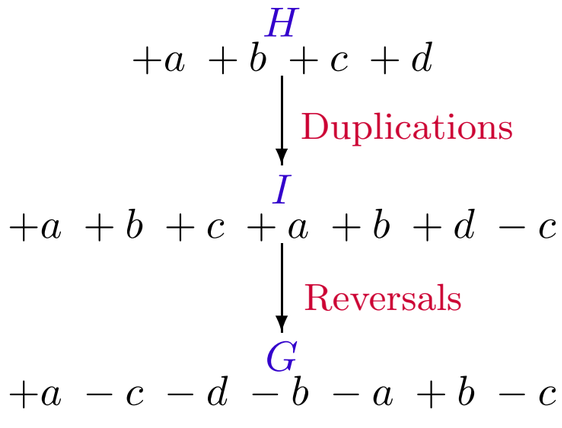
Data: A genome containing many copies of each gene.

Problem: An ancestral genome containing one copy of each gene, minimizing reversals + segment duplication.

$D(G)$ Nb of repeats of G:

$$\underbrace{+a\ -b\ +c}_{S_1}\ +x\ \underbrace{+d-e}_{S_2}\ \underbrace{+e\ -d}_{\overline{S_2}}\ \underbrace{+a\ -b\ +c}_{\overline{S_1}}\ +y$$

Genomes with at most two copies of each gene



A reversal can decrease by at most two the number of repeats of G.

Find I minimizing: $RD(G, I) = D(I) + R(G, I)$.

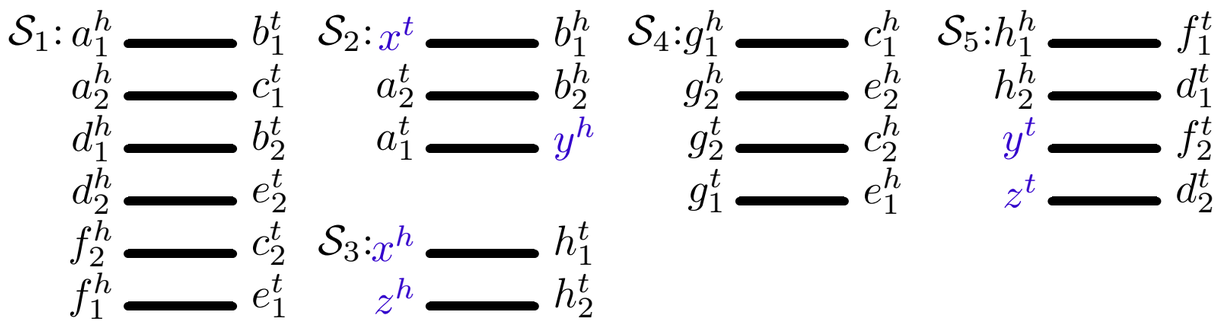
Ignoring "bad components" $\rightarrow$ minimize

$$\Delta(G) = D(I) + b(G) - c(G, I)$$

Genome:

$$a_1\ b_1\ x\ h_1\ f_1\ e_1\ g_1\ -c_1\ -a_2\ -b_2\ -z\ d_2\ e_2\ -g_2\ -c_2\ -f_2\ y$$

Natural graphs:

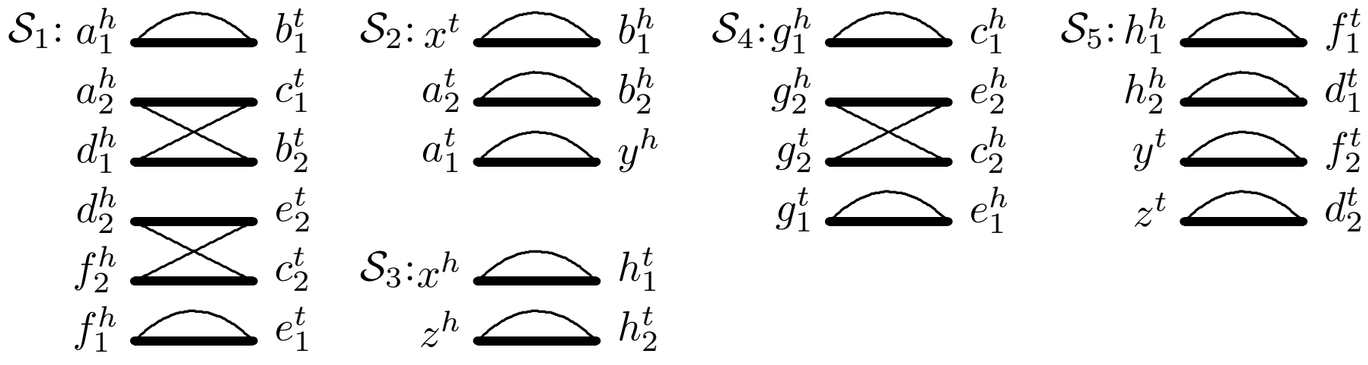


$\mathcal{E}$: Natural graphs of even size with only duplicated genes.

$$\Delta(G) \geq D(G) - |\mathcal{E}|$$

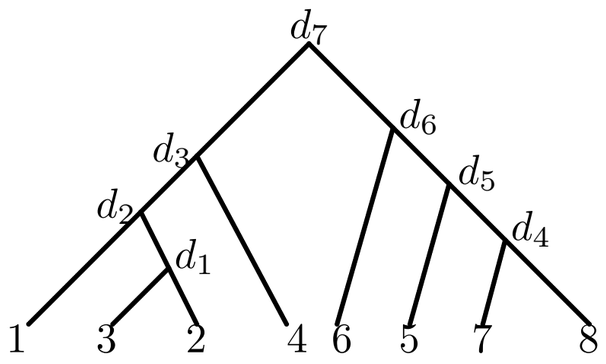
Algorithm

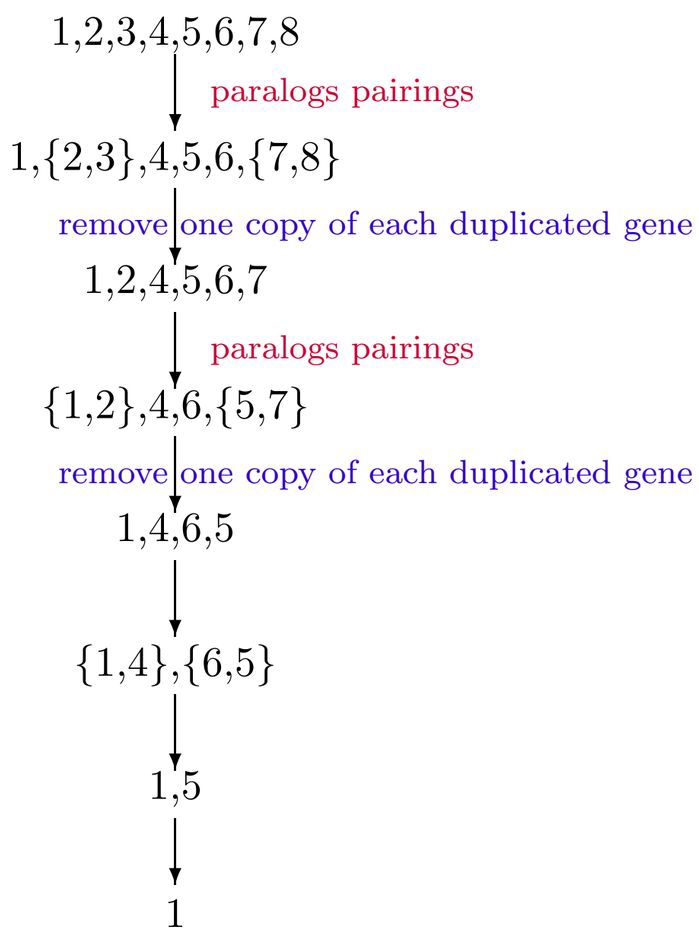
- For graphs not in $\mathcal{E}$, gray edges = black edges;
- For graphs in $\mathcal{E}$, similar to genome duplication.



BUT: Possibly more than one circular fragment. Then, a correction is required.

Approximation algorithm with tight bounds in $O(|\mathcal{E}|n)$.


d_7
d_6
d_3
d_5
d_2
d_4
d_1
1
3
2
4
6
5
7
8


1,2,3,4,5,6,7,8
paralogs pairings
1,{2,3},4,5,6,{7,8}
remove one copy of each duplicated gene
1,2,4,5,6,7
paralogs pairings
{1,2},4,6,{5,7}
remove one copy of each duplicated gene
1,4,6,5
{1,4},{6,5}
1,5
1

VI. Conclusion

First bioinformatics tools to reconstruct the evolutionary history of a single genome.

Genome duplication: A linear-time exact algorithm for reconstructing a pre-doubling ancestral genome in case of reversals, translocations and reversals+translocations.

Segment duplication: A polynomial approximation algorithm with bounds, for reversals.

Extention: Consider the centromere. Some translocations not allowed.