Gene and genome duplication

Nadia El-Mabrouk
Université de Montréal
Canada

Plan

1. Genome rearrangement and multigene families
2. Genome duplication
3. Duplication of chromosomal segments
4. Conclusion

Genome rearrangement

Chromosomes evolved by insertion, deletion, movement of genes

Genomic approach: Compare gene orders

Hypothesis: Homologous genes are known

Chromosome ⟹ sequence of signed genes (or blocks)

$$b \quad -a \quad d \quad -e \quad -c \quad f$$

Multigene families

In the human genome ~15% protein genes duplicated

(*Li, Wang, Nekrutenko, 2001*)

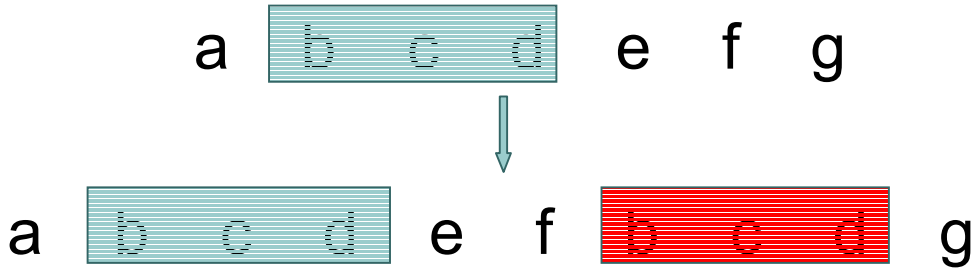
~16% yeast, ~25% Arabidopsis (*Wolfe, 2001*)

Compare sequences of signed genes allowing many copies of each gene

b –a d a –e –c e f d a

Multigene families due to:

- Single gene duplication;
- Segment duplication: Tandem duplication or duplication transposition



- Horizontal gene transfer;
- Genome-wide doubling event

Algorithms and models

- Genome rearrangement with multigene families
  - → Exemplar approach, *Sankoff 1999*
  - → Insertion, deletion, gene duplication,

    Marron,Swensen, Moret 2003
- Reconciliation analysis, projecting gene tree on phylogenetic tree

  Hallett, Lagergren 2000, Page, Cotton 2000;

  Chen, Durand, Farach 2000, Sankoff, El-Mabrouk 2000
- Probabilistic models for the generation of multigene families

Find the ancestor of a genome with multiple gene copies

- Genome duplication

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

- Duplication of chromosomal segments

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

- Genome duplication for unordered chromosomes

  N. El-Mabrouk and D. Sankoff 1998

Plan

1. Genome rearrangement and multigene families
2. Genome duplication
3. Duplication of chromosomal segments
4. Conclusion

Genome doubling

1: a b –d ; 2: h c f –g e

1: a b –d ; 2: h c f –g e

1': a b –d ; 2': h c f –g e

Tetraploid = 4n chromosomes

Evidence across the eukaryote spectrum; Two duplications in early vertebrate evolution (*McLysaght et al. 2002)*

Particularly prevalent in plants (rice, oats, corn, wheat, soybeans, Arabidopsis…)

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

From 8 to 16 chromosomes

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

Originally, duplicated genome =
2 identical copies of each chromosome

After rearrangements, duplicated segments scattered among the genome

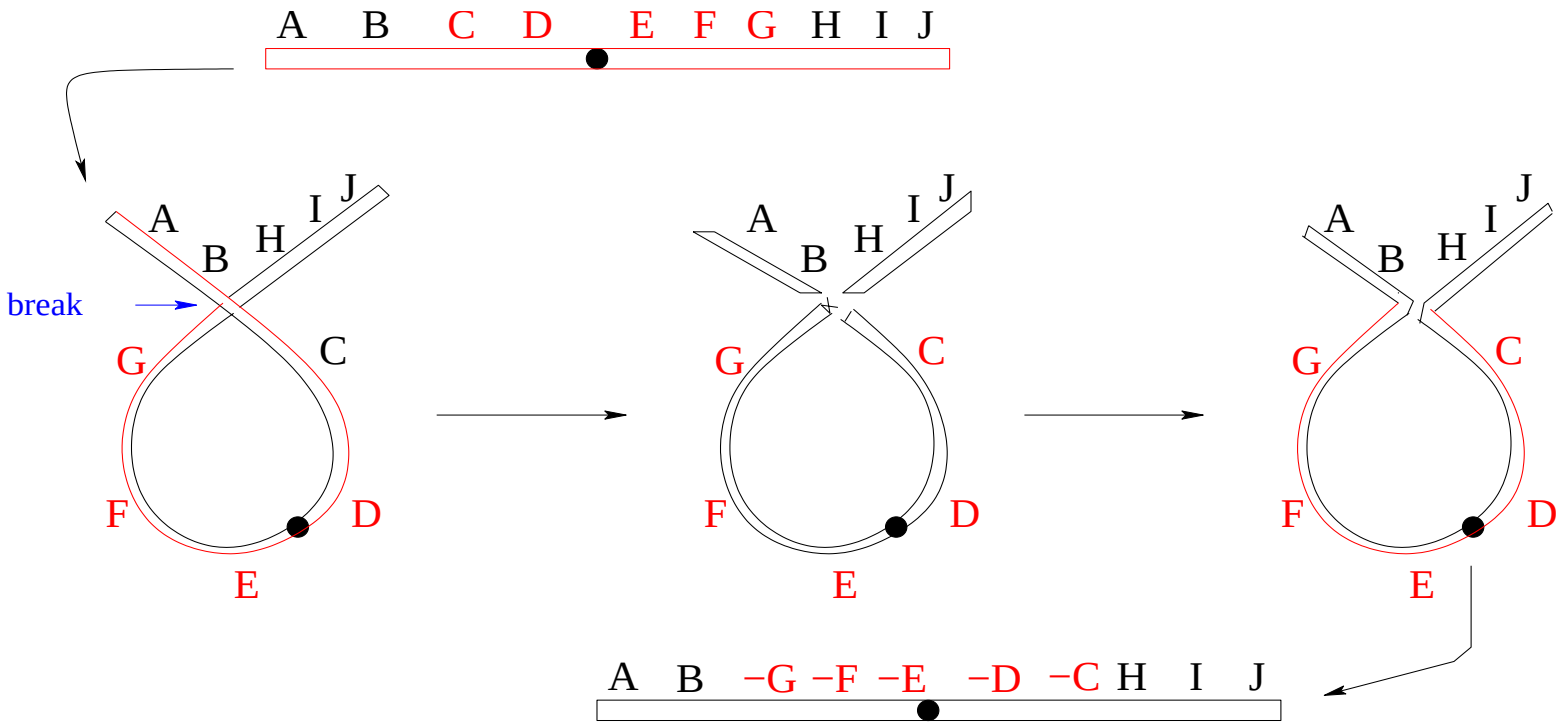
Present–day genome: Signed gene sequences, 2 copies of each gene

Reconstruct original gene order at time of duplication

Minimum number of reversal and/or translocation

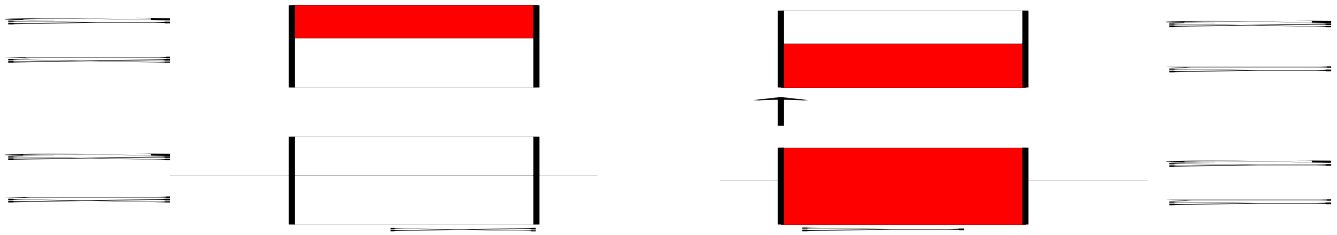

TCS
SSTT634

Inversion


A B C D E F G H I J
break
A B H I J
G C
F D
E
A B H I J
G C
F D
E
A B H I J
G C
F D
E
A B −G −F −E −D −C H I J

Translocation

Reciprocal translocation:



Fusion:

Fission:

Problem:

Rearranged duplicated genome G:

1: +a +b –c +b -d

2: -c -a +f

3: -e +g -f -d

4: +h +e -g +h

Unknown duplicated genome H:

1: +a +b -d

2: +a +b -d

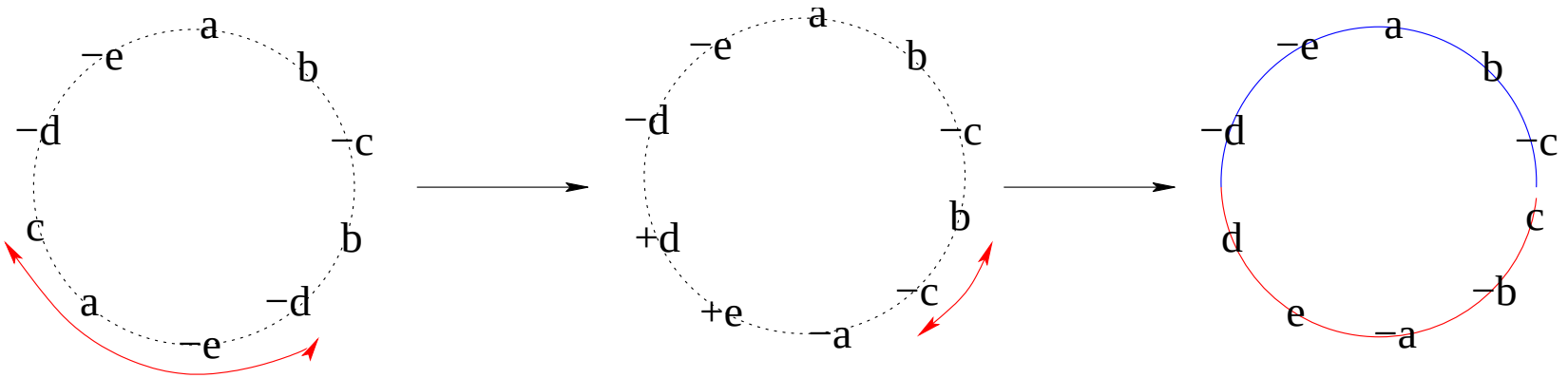
3: +h +c +f -g +e

4: +h +c +f -g +e

Min. num. of inversion and/or translocation transforming G into H

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

The circular case



Rearranged genome

Ancestral duplicated genome

Method

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

First polynomial algorithm by ***Hannenhalli and Pevzner*** for

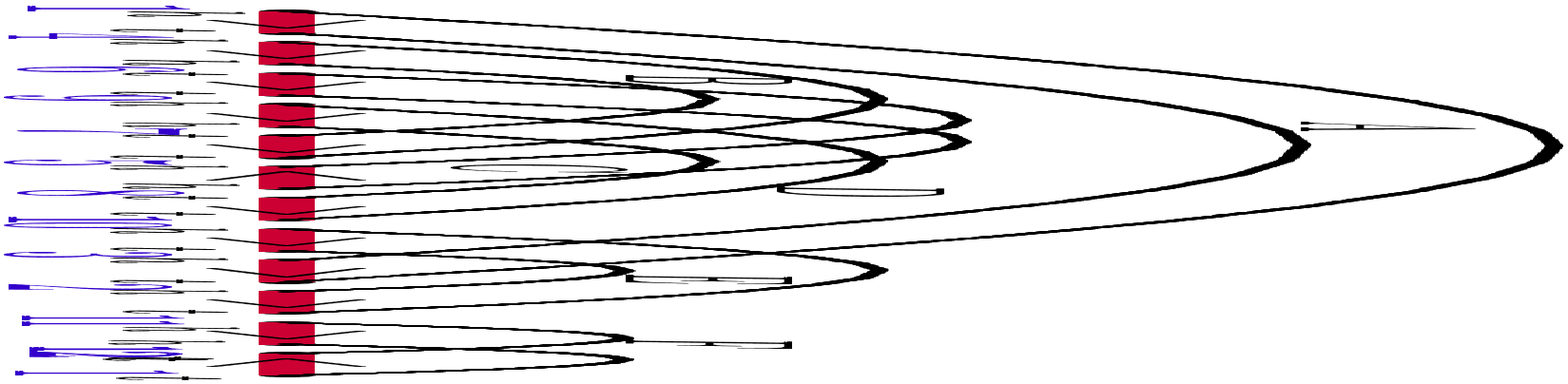
- Reversals only
- Translocations only
- Reversals and translocations

Ancestral duplicated genome of G minimizing the HP formula

The breakpoint graph

G_1 = +1 +4 -6 +9 -7 +5 -8 +10 +3 +2 +11 +12

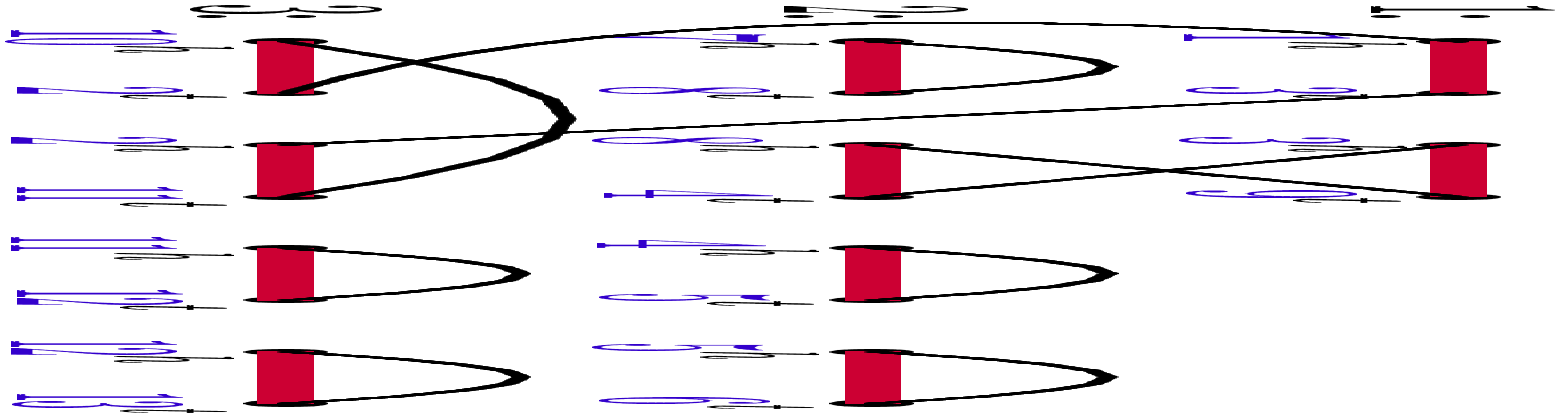
G_2 = +1 +2 +3 +4 +5 +6 +7 +8 +9 +10 +11 +12



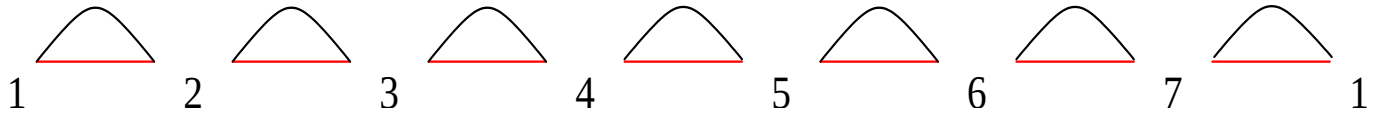
Multichromosomal case

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 , maximum number of cycles



Perform reversals increasing nb of cycles

Good component: Can be solved by good reversals

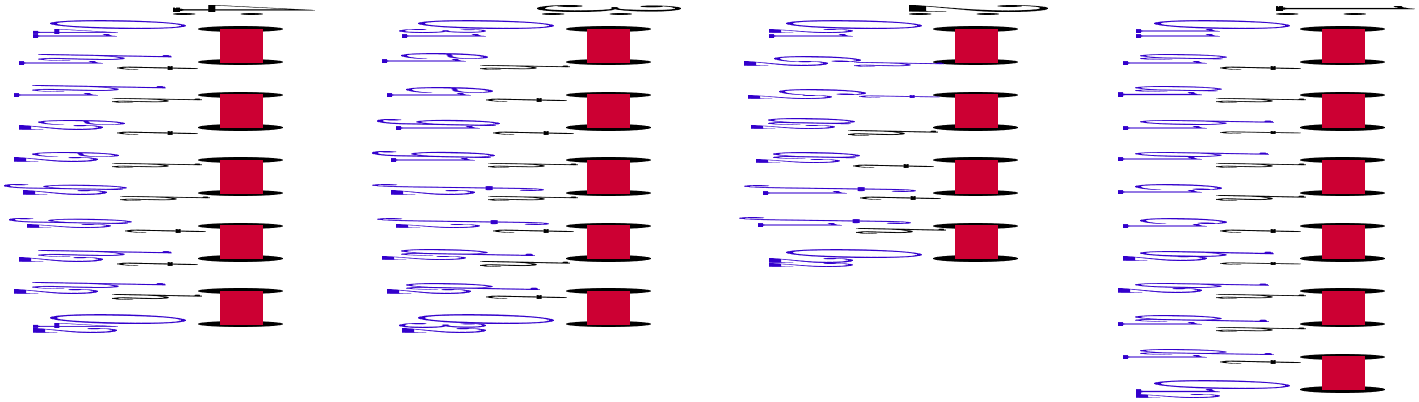
Bad component: Requires bad reversals

HP: $RD(G_1, G_2) = b - c + m + f$

b : red edges; c : cycles; m : bad components; f : Correction of 0, 1 or 2

Genome halving

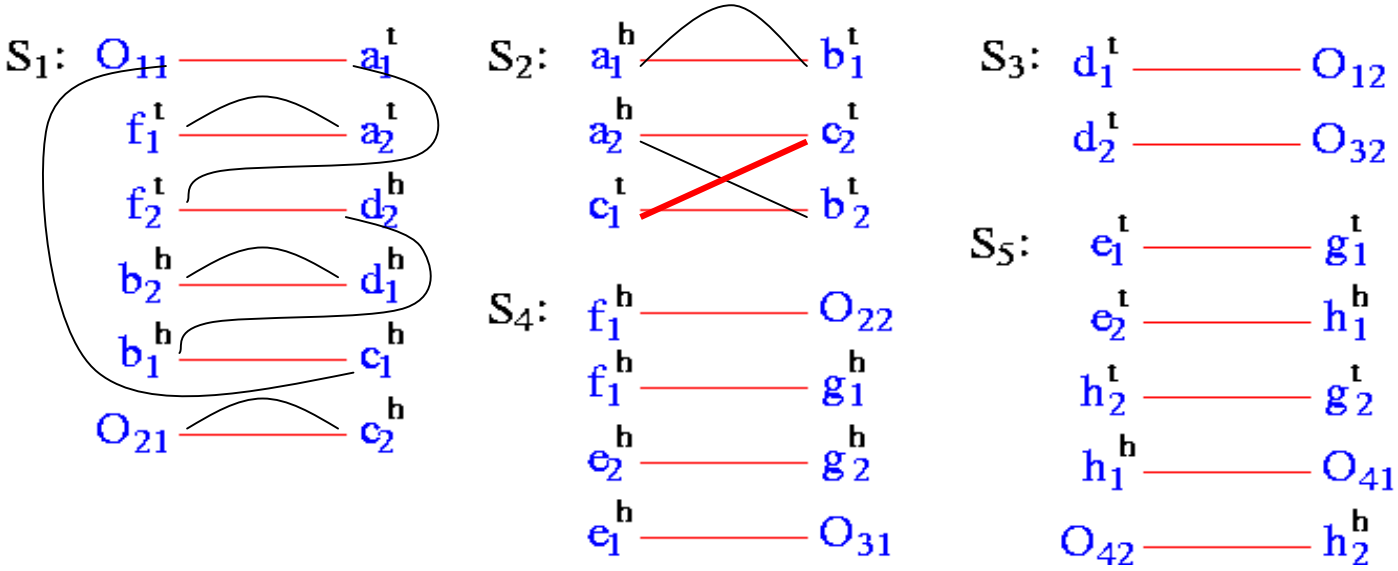
Partial graph for *G:*



Set of valid black edges representing a duplicated genome

Find a set of valid black edges minimizing HP

Decomposition into natural subgraphs



Natural subgraphs of even size are completable

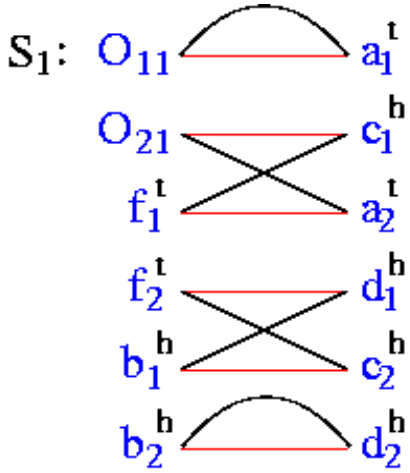
Amalgamate natural graphs into completable supernatural graphs

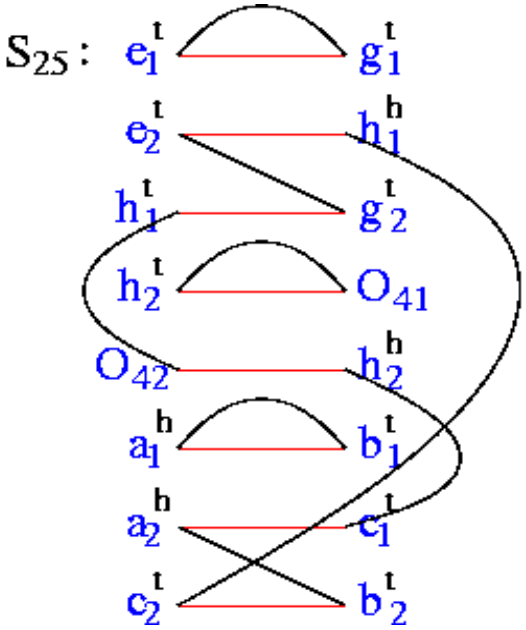
Example: Amalgamate S_2 and S_5 ⟹ S_1, S_{25}, S_3, S_4

Upper bound on the number of cycles

S_e a supernatural graph of n edges, $S_e(\Gamma_e)$ a completed graph with c_e cycles

- If S_e not amalgamated, $c_e \leq n/2 + 1$
- Otherwise, $c_e \leq n/2$

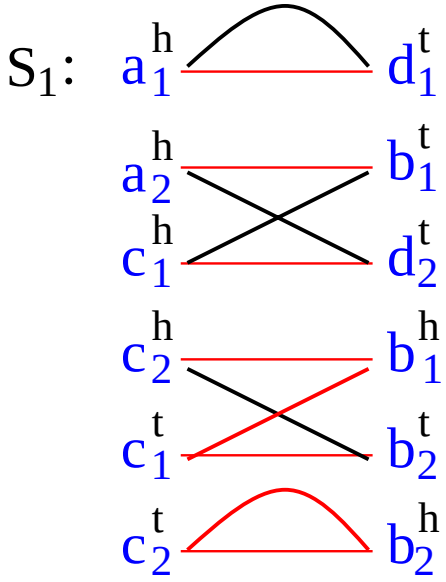


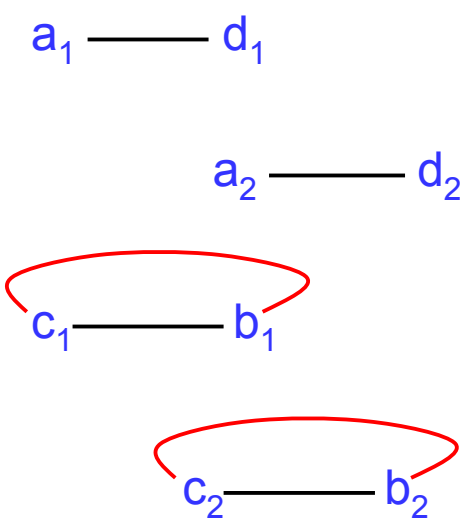


Maximizing cycles – Multichromosomal case

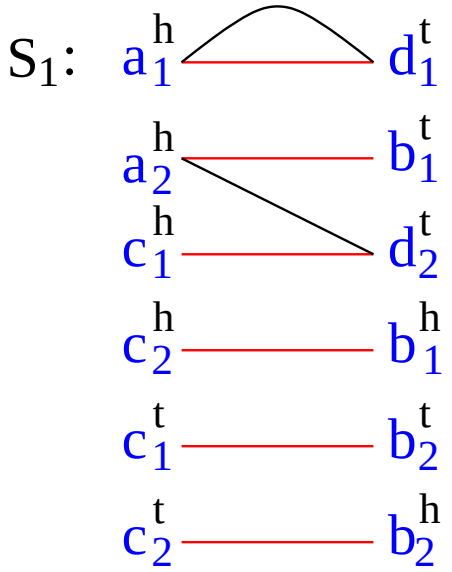
Complete each subgraph separately

Avoid to create circular fragments



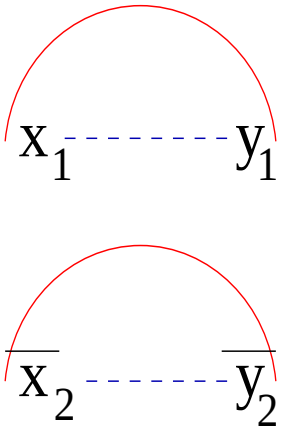


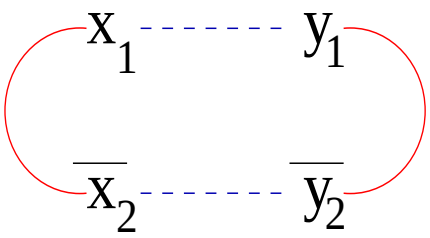
Bad graph:



Maximizing cycles 2

Avoid black edges creating circular fragments:

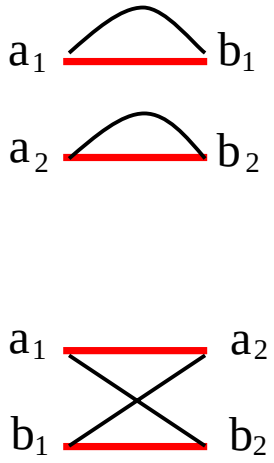




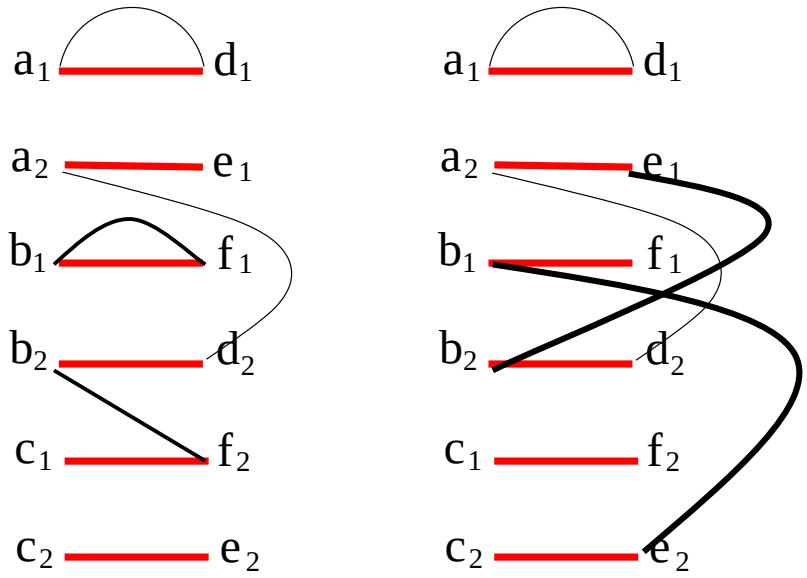
A pair of edges that does not create a circular fragment nor a bad graph is called possible

Algorithm dedouble

2-edges graph:



n-edges graph:



Algorithm dedouble 2

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

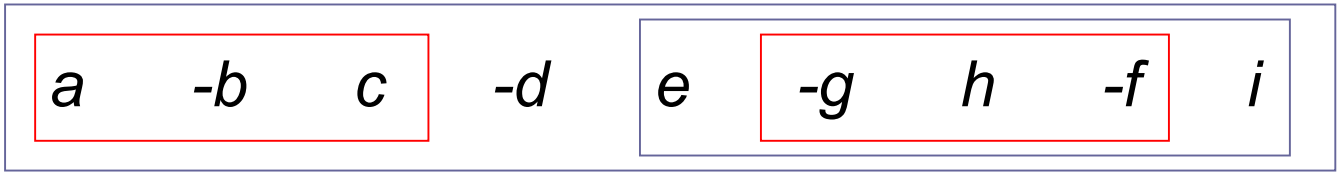
$$c = n/2 + \gamma$$

- n : number of red edges
- γ : number of natural graphs (not amalgamated)

Bad components

Related to subpermutations or conserved intervals

minSP: SP not contained in any SP



- Rearrangement by translocations:

  Bad components = minSPs

- Rearrangement by invertions and/or translocations:

  Bad components subset of minSPs

Bad components 2

Local SPs of *G* :

$$a_1 \quad b_1 \quad c_1 \quad d_1 \quad e_1 \quad -d_2 \quad b_2 \quad c_2 \quad -a_2 \quad e_2$$

Lemma: In a max completed graph, if there is minSP not a local SP, then correction to eliminate the minSP

Corollary: If *G* does not contain local SPs, then duplicated genome H produced by the algorithm is such that:

RO(G,H) minimal

Bad components 3

General case:

$$RO(G) = n/2 - \gamma(G) + m(G) + \phi(G)$$

n : nb of red 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

Circular case: Uncertainty of up to 2 reversals

Application: Yeast genome

Degenerate tetraploid, duplication 10^8 years ago (*Wolfe and Shield, 1997*). 55 duplicated regions

Sorting by translocations: 45 translocations

Sorting by inversions and/or translocations: No local SPs, thus no reversal. Still 45 translocations

A circular genome

Mitochondrial genome of Marchantia polymorpha: many genes in 2 or 3 copies (*Oda et al. 1992*)

Unlikely to be a tetraploid

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

Sorting by reversals: minimum of 25 reversals

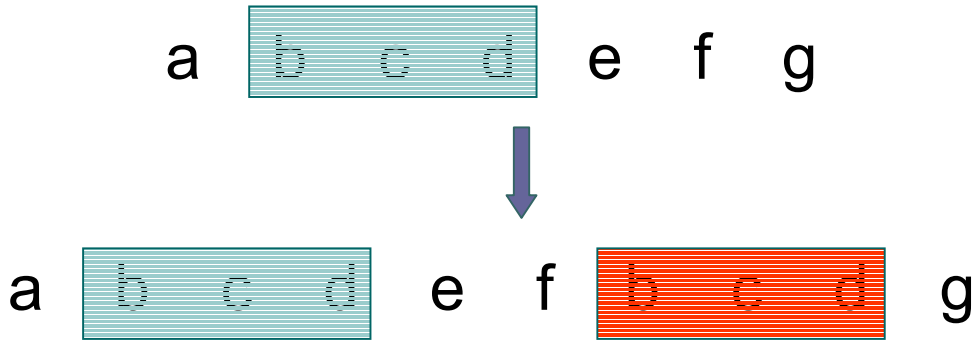
Similar to a random distribution ⟹ No trace of duplication

Plan

1. Genome rearrangement and multigene families
2. Genome duplication
3. Duplication of chromosomal segments
4. Conclusion

Duplication of chromosomal segments

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

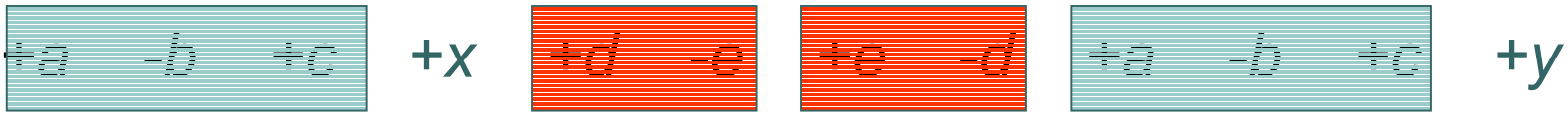


Very recent segment duplication in the human genome (*Eichler et al., 1999*)

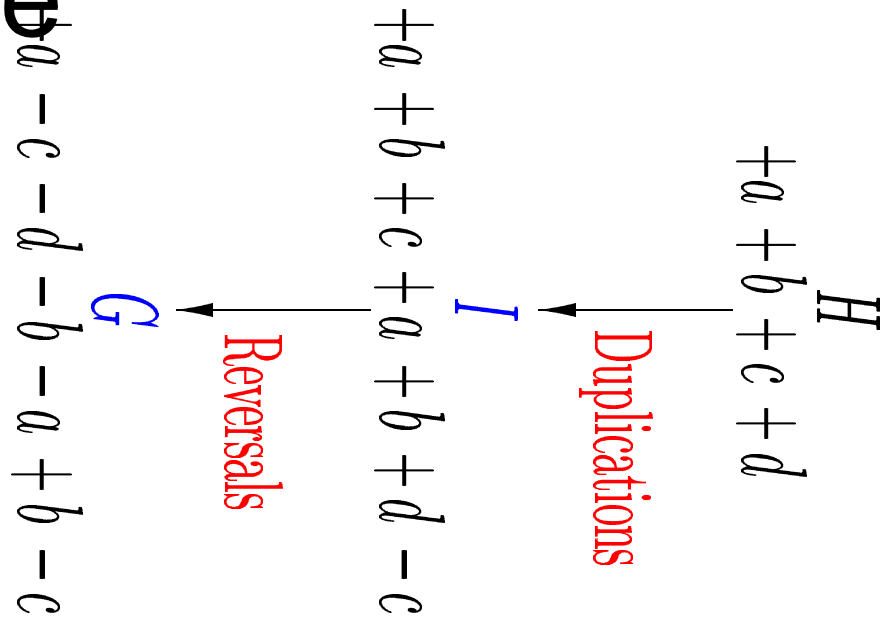
Data: A genome containing many copies of each gene

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

D(G) : Number of repeats of G



At most two copies of each gene



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

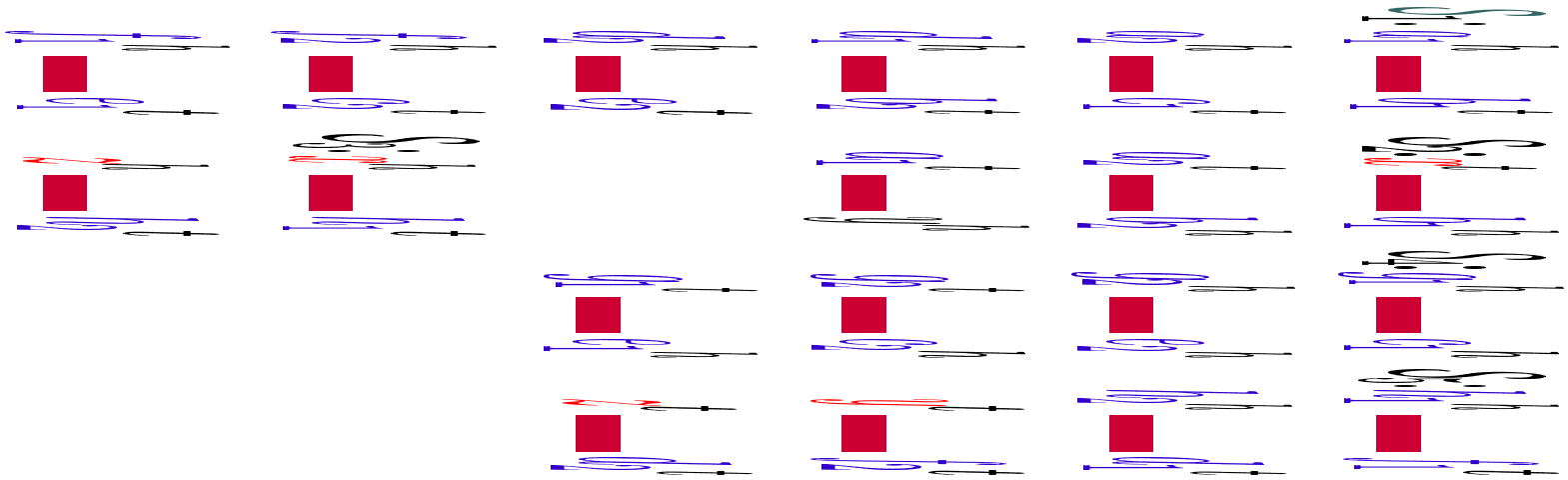
Find I minimizing $RD(G,I) = D(I) + R(G,I)$

Ignoring bad components ⟶ minimize $\Delta(G) = D(I) + n(G) - c(G,I)$

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:

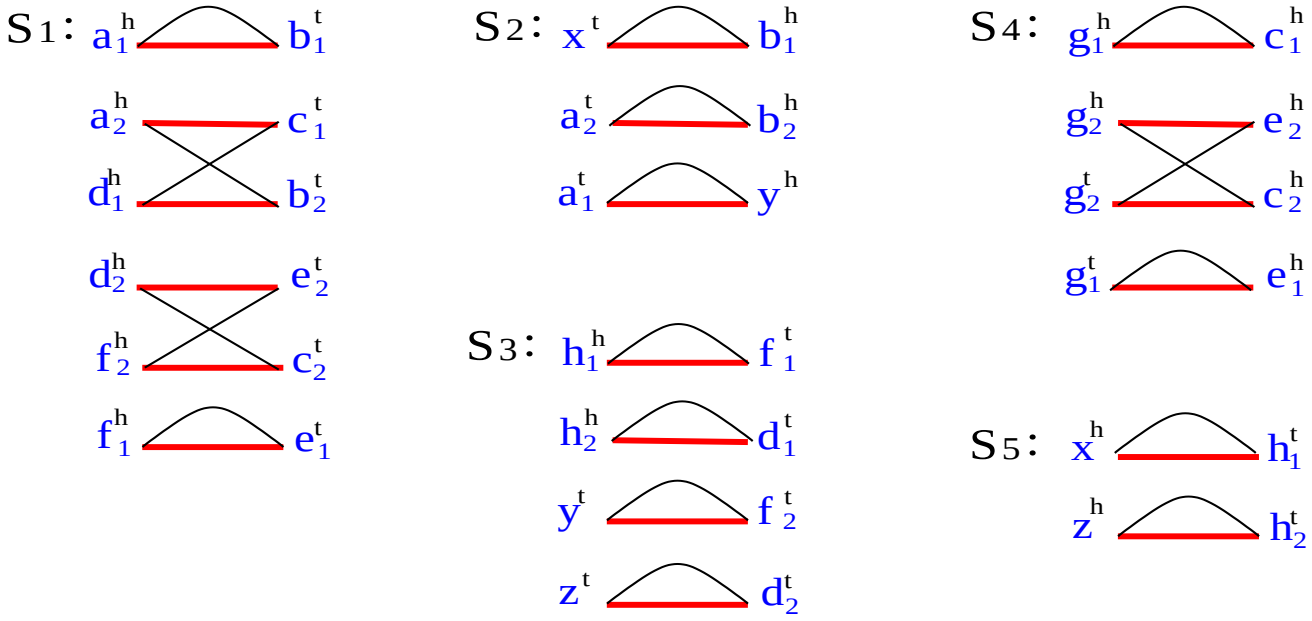


E: Graphs of even size with only duplicated genes

$$\Delta(G) \geq D(G) - |E|$$

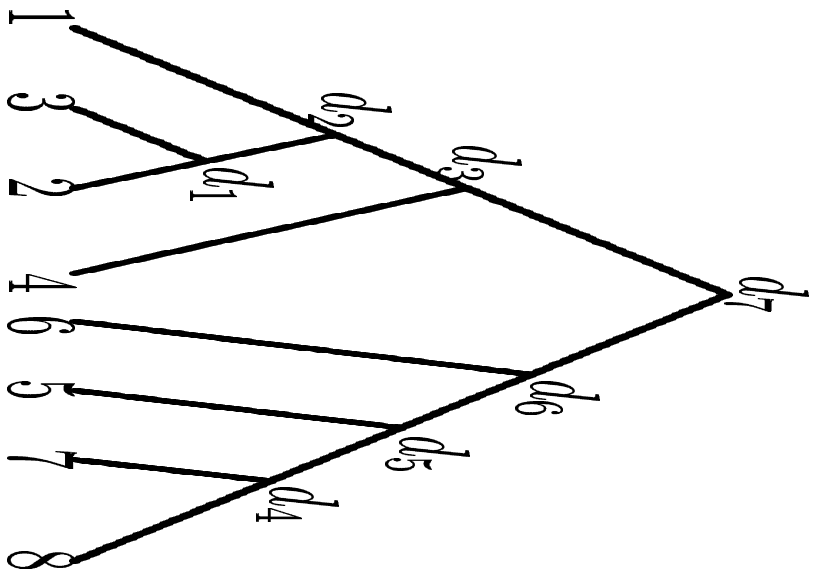
Algorithm

- For graphs not in E, red edges = black edges;
- For graphs in E, similar to genome duplication



BUT: Possibly more than one circular fragment. A correction is required

Approximation algorithm with tight bounds in $O(|E|\, n)$



1, 2, 3, 4, 5, 6, 7, 8

Paralog pairings

1, {2, 3}, 4, 5, 6, {7, 8}

Remove one copy of each duplicate

1, 2, 4, 5, 6, 7

Paralog pairings

{1, 2}, 4, 6, {5, 7}

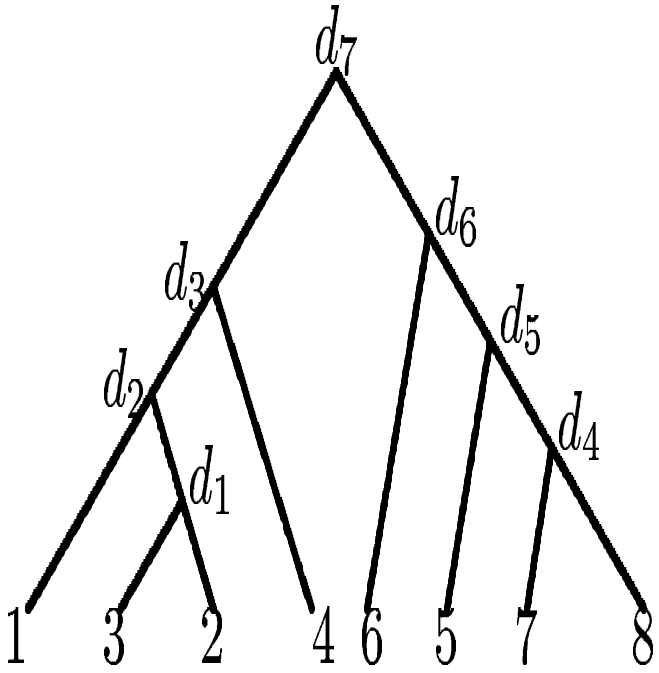
Remove one copy of each duplicate

1, 4, 6, 5

{1, 4}, {6, 5}

1, 6

1



1,2,3,4,5,6,7,8

paralogs pairings

1,{2,3},4,5,6,{7,8}

remove one copy of each d

1,2,4,5,6,7

paralogs pairings

{1,2},4,6,{5,7}

remove one copy of each d

1,4,6,5

{1,4},{6,5}

1,5

1

Conclusion

- First bioinformatics tools to reconstruct the evolutionary history of a single genome
- Genome duplication: A linear-time exact algorithm for reversals and/or translocations
- Segment duplication: A polynomial approximation algorithm with bounds for reversals
- Extension: Consider the centromere. Some translocations not allowed