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:

\[
\begin{align*}
1: & \ a \ b \ \sim d \ ; \ 2: \ h \ c \ f \ \sim g \ e \\
I: & \ a \ b \ \sim d \ ; \ 2: \ h \ c \ f \ \sim g \ e \\
1': & \ a \ b \ \sim d \ ; \ 2': \ h \ c \ f \ \sim g \ e 
\end{align*}
\]

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.

→ From 8 to 16 chromosomes

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Duplicated Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>2 • −1</td>
</tr>
<tr>
<td>II</td>
<td>4 • −3 7 +8 5 +6</td>
</tr>
<tr>
<td>III</td>
<td>9 • −10 −11</td>
</tr>
<tr>
<td>IV</td>
<td>20 +12 12 +54 +15 +21 • −3 −13 −16 +17 −24 −22 −14 −23 −19 +18 −9</td>
</tr>
<tr>
<td>V</td>
<td>28 • −25 −27 4 −26 −13</td>
</tr>
<tr>
<td>VI</td>
<td>55 • −36</td>
</tr>
<tr>
<td>VII</td>
<td>36 +25 26 +32 +6 −33 +5 • −30 −34 −31 −29</td>
</tr>
<tr>
<td>VIII</td>
<td>35 • −14 −37 −29 −1</td>
</tr>
<tr>
<td>IX</td>
<td>38 +39 +27 •</td>
</tr>
<tr>
<td>X</td>
<td>10 +40 +41 • −28 −42</td>
</tr>
<tr>
<td>XI</td>
<td>42 +40 +43 +35 • −41 −52 −38</td>
</tr>
<tr>
<td>XII</td>
<td>53 • −53 −31 −55 −16 −18 −17 −45 −30 −15 −44</td>
</tr>
<tr>
<td>XIII</td>
<td>46 +44 +19 • −43 −54 −48 −47 −46</td>
</tr>
<tr>
<td>XIV</td>
<td>49 +20 +37 +50 +30 • −11</td>
</tr>
<tr>
<td>XV</td>
<td>49 +21 • −22 −52 −50 −23 −45 −51 −47 −2</td>
</tr>
<tr>
<td>XVI</td>
<td>48 +32 +33 +51 +8 +24 • −7 −34</td>
</tr>
</tbody>
</table>
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:
1: $+a + b - c + b - d$; 3: $-e + g - f - d$;
2: $-c - a + f$; 4: $+h + e - g + h$.

Duplicated genome:
1: $+a + b - d$; 3: $+h + c + f - g + e$;
2: $+a + b - d$; 4: $+h + c + f - g + e$. 
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.

$\rightarrow$ 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 $G_2$:

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

- $b$: 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: \( O_{11} \overset{\text{a}_1}{\longrightarrow} \overset{\text{b}_1}{\longrightarrow} \overset{\text{c}_1}{\longrightarrow} \overset{\text{b}_2}{\longrightarrow} \overset{\text{d}_1}{\longrightarrow} \overset{\text{O}_{12}}{\longrightarrow} \)

2: \( O_{21} \overset{\text{e}_{2}}{\longrightarrow} \overset{\text{f}_1}{\longrightarrow} \overset{\text{O}_{22}}{\longrightarrow} \)

3: \( O_{31} \overset{\text{g}_1}{\longrightarrow} \overset{\text{h}_1}{\longrightarrow} \overset{\text{i}_1}{\longrightarrow} \overset{\text{O}_{32}}{\longrightarrow} \)

4: \( O_{41} \overset{\text{j}_1}{\longrightarrow} \overset{\text{k}_1}{\longrightarrow} \overset{\text{l}_1}{\longrightarrow} \overset{\text{O}_{42}}{\longrightarrow} \)

Set of valid gray edges: Represent a duplicated genome.

**Problem:** Find a set of valid gray edges minimizing formula \( \text{HP} \).
Decomposition into subgraphs

Natural graphs:

$S_1: O_{11} \rightarrow a_1^1 \quad S_2: a_1^h \rightarrow b_1^f \quad S_4: f_1^h \rightarrow O_{22} \quad S_5: e_1^l \rightarrow g_1^l \quad O_{12} \quad O_{32}$

$S_2: a_2^f \rightarrow a_2^h \rightarrow e_2^l \rightarrow f_2^h \rightarrow g_2^h \rightarrow e_2^l \rightarrow h_2^2 \rightarrow g_2^l \rightarrow b_2^2 \rightarrow b_2^1 \rightarrow c_1^1 \rightarrow c_1^h \rightarrow S_3: d_1^h \rightarrow O_{12} \quad O_{42} \rightarrow h_2^b$

$S_3: d_2^h \rightarrow d_2^b \rightarrow O_{32}$

Natural graphs of even size are completable.

Amalgamate natural graphs into completable supernatural graphs.

Example: Amalgamate $S_2$ and $S_5$

$\rightarrow S_1, S_{25}, S_3, S_4$ are supernatural graphs.
Upper bound on the number of cycles

$G_e$ a supernatural graph of $n$ edges, $G_e(\Gamma_e)$ a completed graph, and $c_e$ its number of cycles.

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

\[ S_1: 0_{11} \leftrightarrow a_1^1 \quad \quad S_{25}: e_1^1 \leftrightarrow g_1^1 \]

\[ 0_{21} \leftrightarrow a_1^h \quad e_2 \leftrightarrow h_1^b \]
\[ f_1^1 \leftrightarrow a_2 \quad h_1^b \leftrightarrow g_2 \]
\[ f_2^1 \leftrightarrow a_1^h \quad h_1 \leftrightarrow g_2 \]
\[ b_1^h \leftrightarrow c_1^h \quad a_1^1 \leftrightarrow b_1 \]
\[ b_2^1 \leftrightarrow d_1^2 \quad a_2 \leftrightarrow b_1 \]
\[ b_2 \leftrightarrow d_2^2 \]
Maximizing the number of cycles - Multichromosomal case

Complete each supernatural graph separately.

Avoid to create circular fragments.

Bad graph:

\[ S_1: a_1^h \rightarrow d_1^h \]
\[ a_2^h \rightarrow b_1^h \]
\[ c_1^h \rightarrow d_2^h \]
\[ c_2^h \rightarrow b_1^h \]
\[ c_1^l \rightarrow b_2^l \]
\[ c_2^l \rightarrow b_2^l \]

\[ a_1^l \rightarrow d_1^l \]
\[ a_2^l \rightarrow b_1^l \]
\[ c_1^l \rightarrow d_2^l \]
\[ c_2^l \rightarrow b_1^l \]
\[ c_1^l \rightarrow b_2^l \]
\[ c_2^l \rightarrow b_2^l \]
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

Algorithm dedouble:

2-edges graph:

\[ \begin{align*}
&\quad a_1 \longrightarrow b_1 \\
&\quad a_2 \longrightarrow b_2
\end{align*} \]

\[ \begin{align*}
&\quad a_1 \longrightarrow a_2 \\
&\quad b_1 \longrightarrow b_2
\end{align*} \]

n-edges graph:

\[ \begin{align*}
&\quad a_1 \longrightarrow d_1 \\
&\quad a_2 \longrightarrow e_1 \\
&\quad b_1 \longrightarrow f_1 \\
&\quad b_2 \longrightarrow d_2 \\
&\quad c_1 \longrightarrow f_2 \\
&\quad c_2 \longrightarrow e_2
\end{align*} \]
Linear time algorithm constructing a maximal completed graph containing $c$ cycles:

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

- $n$: Number of black edges;
- $\gamma$: Number of natural graphs (not amalgamated).
Bad components 1

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

\textbf{minSP}: SP not included in any SP.

\begin{center}
\begin{tabular}{ccccccc}
a & b & c & -i & -g & e & f & -h & -j \\
\end{tabular}
\end{center}

Bijection between \textit{SPs} and \textit{components}.

- Rearrangement by \textit{translocations}: Bad components = minSPs;
- Rearrangement by \textit{reversals and/or translocations}: Bad components $\subset$ minSPs.
Bad components 2

Local SPs:

\[
\begin{array}{cccccccc}
 a_1 & b_1 & c_1 & d_1 & c_1 & -d_2 & b_2 & c_2 & -a_2 & e_2 \\
\end{array}
\]

**Lemma:** In a maximal completed graph, if \( \exists \) 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.

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.

→ minimum of 25 inversions

Similar to a random distribution ⇒ No trace of duplication.
V. Duplication of chromosomal segments

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

```
  a b c d e f g
```

```
  a b c d e f b c d g
```

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.

\[ D(G) \text{ Nb of repeats of } G: \]

\[ +a - b + c + x + d - c + e - d + a - b + c + y \]

\[ S_1 \quad S_2 \quad S_2 \quad S_1 \]
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” → 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:

\[ S_1: a^h_1 \quad b^l_1 \quad S_2: x^l \quad b^h_1 \quad S_4: g^h_1 \quad c^h_1 \quad S_5: h^h_1 \quad f^l_1 \]
\[ a_2^h \quad c_1^l \quad d_2^h \quad e_1^l \quad b_2^h \quad y^h_1 \quad g_2^h \quad e_2^h \quad h_2^h \quad d_1^l \]
\[ d_1^h \quad e_2^l \quad f_2^h \quad f_1^l \quad c_2 \quad z^h \quad h_1^l \quad e_1^h \quad z^l \quad d_1^l \]

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

\[ \Delta(G) \geq D(G) - |\mathcal{E}| \]
Algorithm

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

$S_1$: $a_1^h \rightarrow b_1^h$  $S_2$: $x^t \rightarrow b_1^h$  $S_4$: $g_1^h \rightarrow c_1^h$  $S_5$: $h_1^h \rightarrow f_1^h$

$S_1$: $a_2^h \rightarrow c_1^h$  $a_2^h \rightarrow b_2^h$  $g_2^h \rightarrow e_2^h$  $h_2^h \rightarrow d_1^h$

$S_2$: $a_1^h \rightarrow y^h$  $g_2^h \rightarrow e_2^h$  $y^t \rightarrow f_2^h$

$S_3$: $f_2^h \rightarrow c_2^h$  $h_1^t \rightarrow h_2^t$

$S_4$: $g_1^h \rightarrow e_1^h$  $h_1^t \rightarrow z^h$  $d_2^h$

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

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

32
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.

**Extension:** Consider the centromere. Some translocations not allowed.