Reconstruction of rearrangement history and ancestral genomes in yeast

Guénola Drillon, Alessandra Carbone and Gilles Fischer
guenoladrillon@gmail.com, alessandra.carbone@lip6.fr and gilles.fischer@upmc.fr

Laboratoire Génomique des Microorganismes - CNRS - Université Pierre et Marie Curie

Introduction

The reconstruction of ancestral genomes for groups of species and the precise identification of associated chromosomal rearrangements are basic questions in evolution.

By comparing chromosomes in different species, blocks of consecutive genes, called **synteny blocks**, inherited from their last common ancestor, can be identified. **Breakpoints**, that are regions between synteny blocks, result from rearrangements such as : **inversions** of a DNA segment within a chromosome, **reciprocal translocations** between two chromosomal arms, **duplications**, **insertions** and **deletions** of DNA segments.

Many models have been already proposed [1,2]. All models are based on the principle of **parsimony**: the differences between two genomes have to be explained by a minimum number of rearrangements. The main challenge is to precisely model the information contained in genomes to avoid loosing crucial evidences, allowing to trace back as many as possible rearrangements that occurred during evolution.

Our algorithm is designed in order to:

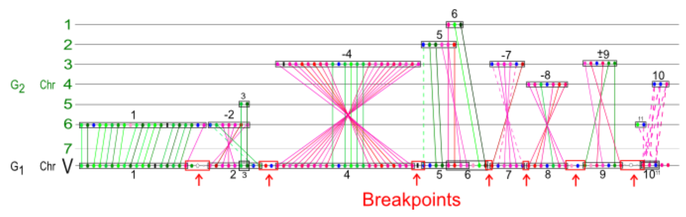
1. identify all the rearrangements that occurred in the different lineages and their resulting breakpoint regions
2. reconstruct ancestral genomes with the maximum number of genes and the minimum number of chromosomes.

Algorithm

Input: Protein sequences from n species; Phylogenetic tree of the n species

A Construction of Synteny Blocks by pairwise comparisons of species

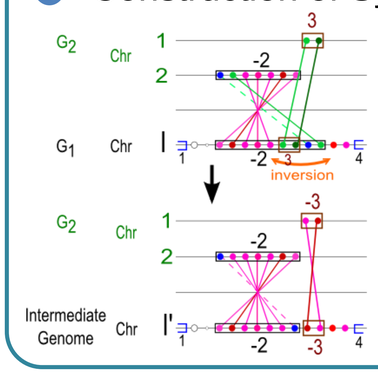
1- Identification of homologues by Bi-Directional Best Hits (BDBH)
2- Construction of blocks with a Δ=5 : two consecutive pairs of homologues cannot be separated by more than 5 genes failing the BDBH condition
3- Refinement of blocks: we add homologous genes having at least 30% of similarity over at least 50% of their length
4- Fusion of neighboring blocks in both genomes
5- Definition of the sign for blocks



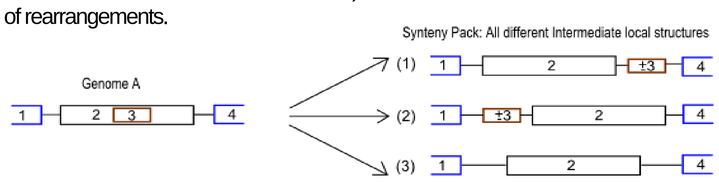
Comparison between chromosomes of species G1 and G2 :
We observe several combinatorial arrangements between genes within synteny blocks: perfect order conservation (block 1), inclusion (3 in 2 in G1), microrearrangement (4), overlapping (5 and 6 in G1), unsigned blocks (9 in G2), duplication (10 and 11 in G2).

Tolerating small rearrangements inside blocks in order to keep all the synteny relationships

B Construction of Synteny Packs to resolve ambiguities in synteny blocks



Problem: some rearrangements can be hidden by small super-imposed inversions
Idea: to pass by an intermediate genome devoid of additional inversions
Solution: test locally all possible intermediate genomes (the notion of synteny pack describes all intermediate local structures) to be able to reconstruct the actual succession of rearrangements.

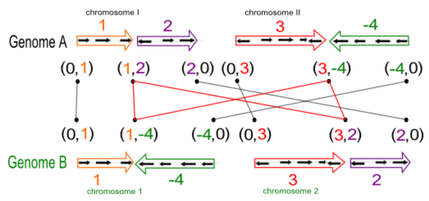


Unraveling super-imposed rearrangements

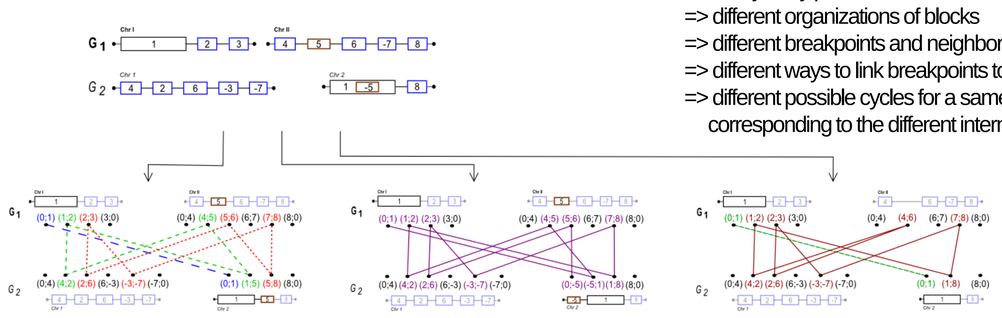
C Identification of Rearrangements by adjacency graph [3]

1- Cycles and Paths between synteny blocks

Let two genomes, A and B, composed of two chromosomes and four synteny blocks (⟹) each, we define the adjacency graph linking the blocks extremities (*i.e.* the breakpoints).
A translocation (as an inversion) forms a small cycle of length four linking the four breakpoints involved in the rearrangement (red cycle).
A path represents rearrangements involving telomeres.



2- Cycles and Paths between synteny packs



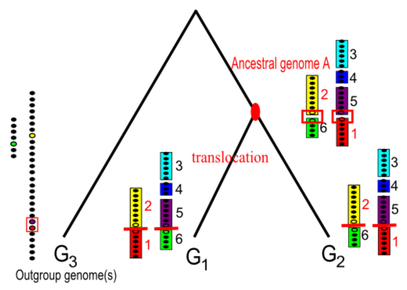
For each synteny pack, one has:
=> different organizations of blocks
=> different breakpoints and neighbors for each block
=> different ways to link breakpoints together
=> different possible cycles for a same breakpoint corresponding to the different intermediate genomes

Among the cycles passing by a same breakpoint, only one has a biological meaning and relevant characteristics for the reconstruction of the history of rearrangements. The principle of parsimony drives us to choose the most parsimonious cycle.

In general, cycles are defined by translocations and/or inversions. A cycle of length 2n corresponds to n-1 inversions and/or translocations. Shorter is the cycle, smaller is the number of rearrangements. Parsimony implies the choice of the shorter cycles as the most probable for the evolutionary process. For the case illustrated here, the first intermediate local structure implies three small cycles (of length 1, 4, 6), whereas the second or third ones implies at least one long cycle (of length 11 or 8).

Parsimony allows to select the most probable rearrangements among different possibilities

D Ancestral genome reconstruction by comparison to the outgroups



1- At the synteny block level

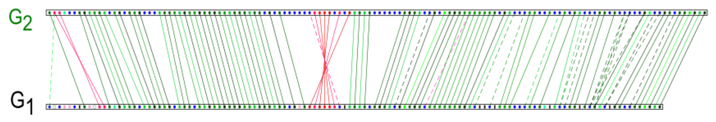
A score is attributed to each breakpoint in G1, corresponding to the proximity of the homologues (in G3) of its neighboring genes. The same is done with breakpoints in G2.

Depending of the different scores obtained for each linked breakpoints, their associated rearrangement are located on the branch where the rearrangement took place and the local ancestral synteny is deduced.

Multiple pairwise comparisons allow us to recover information only shared by few genomes

2- At the gene level

Each block of the ancestral genome has two versions coming from G1 and from G2, which differ by a number of microrearrangements: deletions, duplications, insertions and inversions.



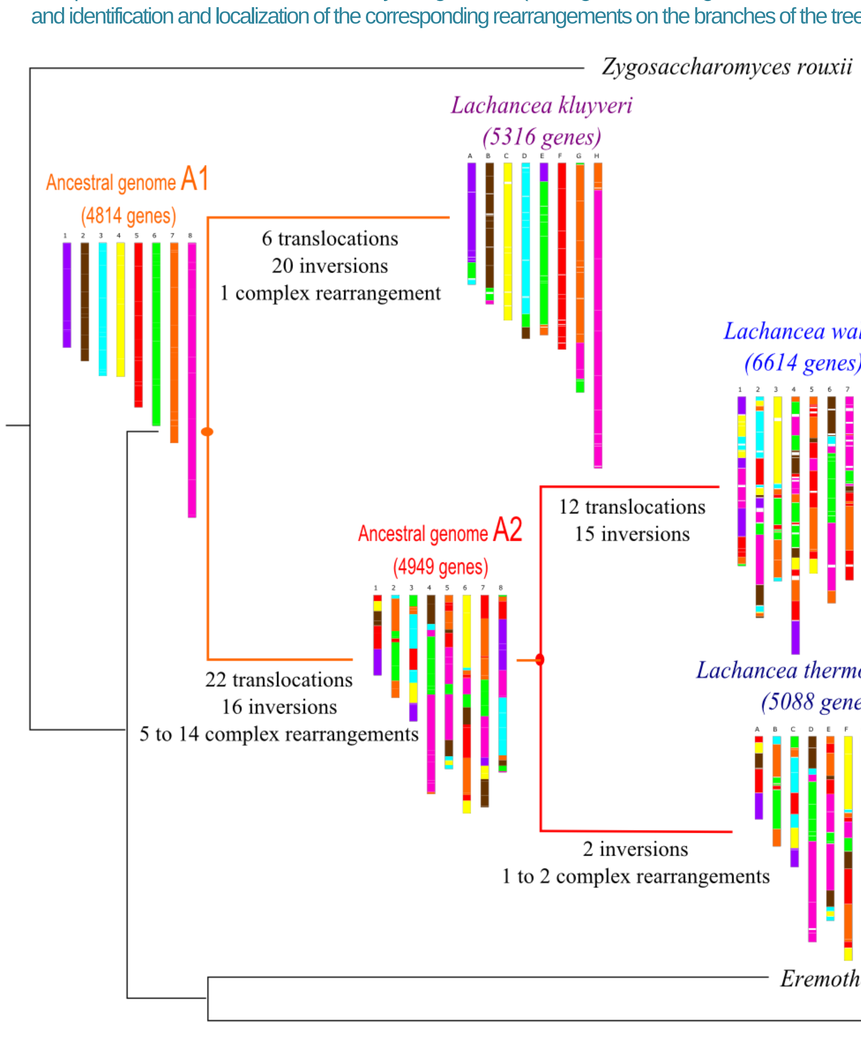
To determine the ancestral gene content and order, there are 2 different steps:
1 - Removing false positive homologies
2 - Detection of small duplications, inversions, insertions and deletions by comparison with the outgroups
In the rare case where the outgroups are not helpful, we choose the order observed in the less rearranged genome.

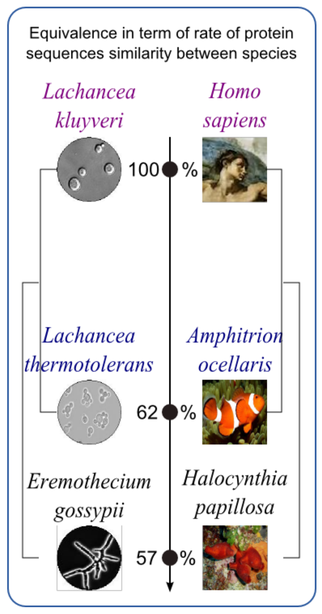
Inference of ancestral gene content and gene order together with a localization of small duplications, deletions, insertions and inversions along the phylogenetic tree

Output: Ancestral Genome; List of Rearrangements associated to branches; Associated Breakpoints and re-use estimations

Results

Complete reconstruction of two ancestral yeast genomes (sharing 90% of their genes with extant species) and identification and localization of the corresponding rearrangements on the branches of the tree





Conclusion

Based on
1. a precise definition of synteny blocks, using a parameterized proximity of genes within the blocks and two different measures of homology between genes,
2. an algorithm combining the notion of linked breakpoints [1] (adapted to deal with the new notion of synteny pack) and the local adjacency comparison [2] (adapted to exploit the strength of pairwise comparison of several genomes),

we are able to:
1. reconstruct several complete ancestral genomes with a realistic number of chromosomes
2. retrace the history of chromosomal rearrangements and precisely define the breakpoint regions.

These reconstructions are the first step to study genome structures, re-use breakpoints, mechanisms of rearrangement and more generally genome evolution.

[1] M. A. Alekseyev and P. A. Pevzner. Breakpoint graphs and ancestral genome reconstructions. Genome Research, 19(5):943-957, 2009
[2] J. Ma, L. Zhang, B. B. Suh, B. J. Raney, R. C. Burhans, W. J. Kent, M. Blanchette, D. Haussler, and W. Miller. Reconstructing contiguous regions of an ancestral genome. Genome Research, 16(12):1557-1565, 2006
[3] A. Bergeron, J. Mixtacki, and J. Stoye. On computing the breakpoint reuse rate in rearrangement scenarios. 2008