MGRA2 performance improvement. Development of the median solver adapted to indel-events.

Student: Artem Kupchinskiy, SPbAU

Instructors: Pavel Avdeev, SPbAU,
Max Alexeyev, George Washington University

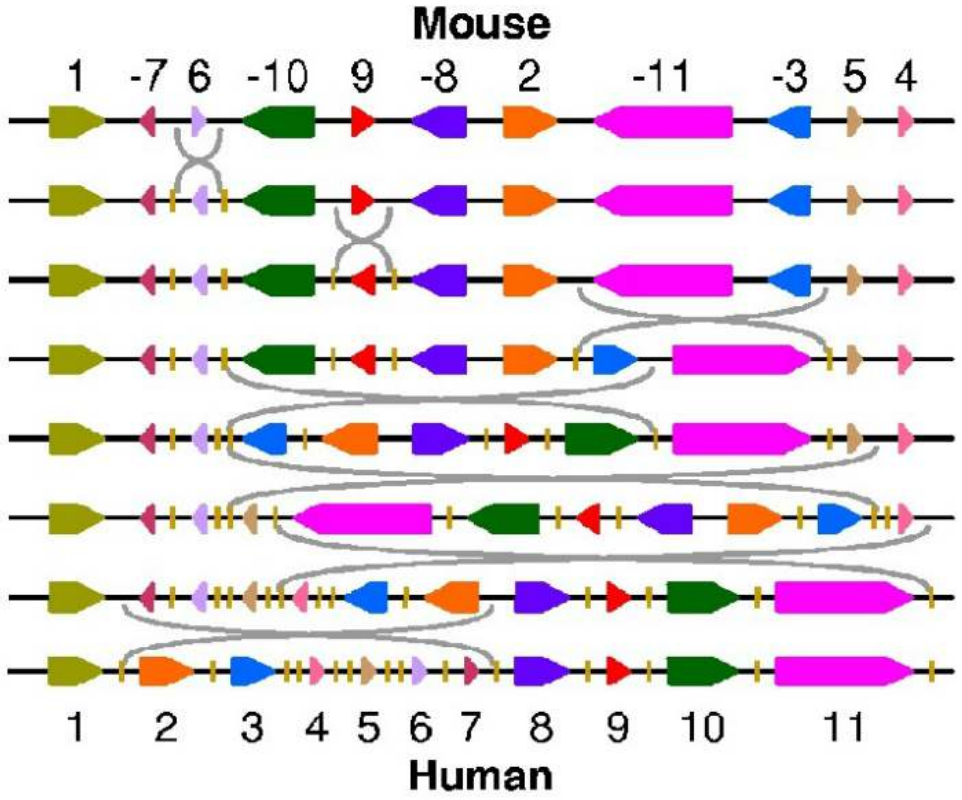
Brief excursion in rearrangement research history

1937 - Dobzhanskiy, Stuart show rearrangement structure on a group of Drosophila obscura species

1987 - Sankoff translated the phylogeny problem on inversion to a computational language. He introduced two key concepts:
The edit distance - the shortest sequence of rearrangments transforming one input genome to another.
The median - given three genomes, construct a fourth genome that minimizes the sum of its pairwise distances to the other three.

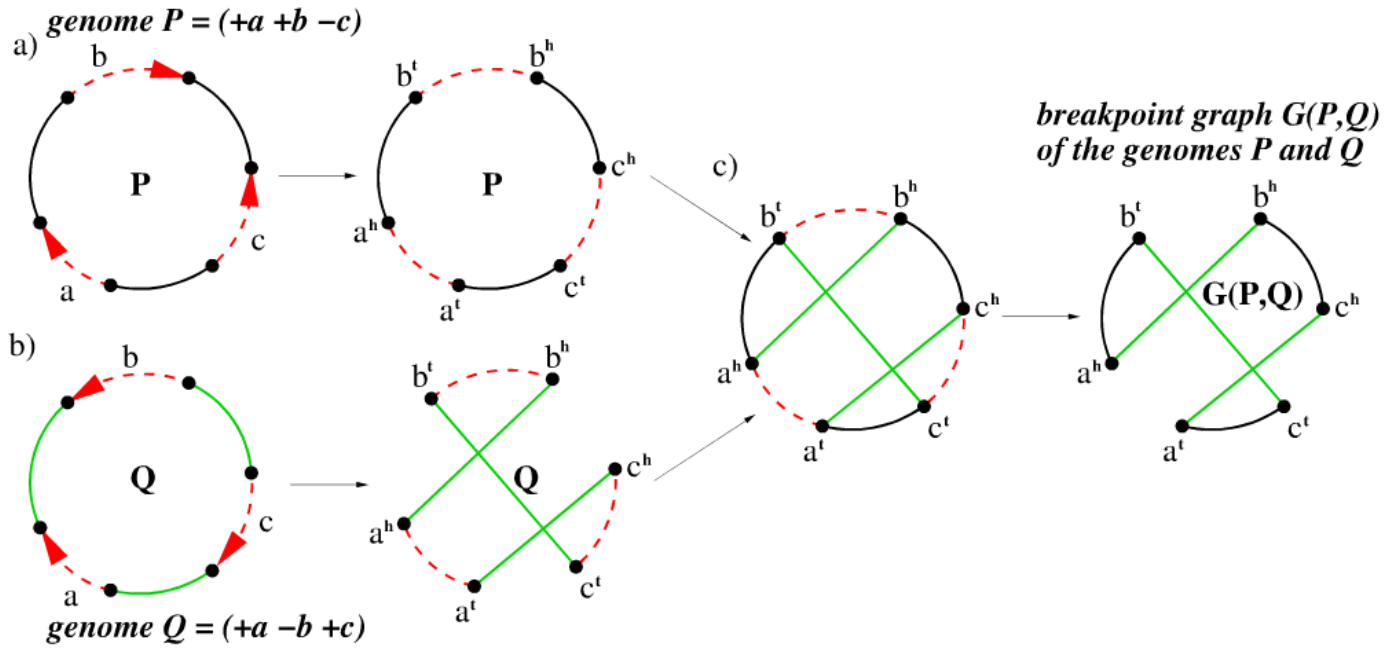
Rearrangement model

Every genome consists of many oriented synteny blocks



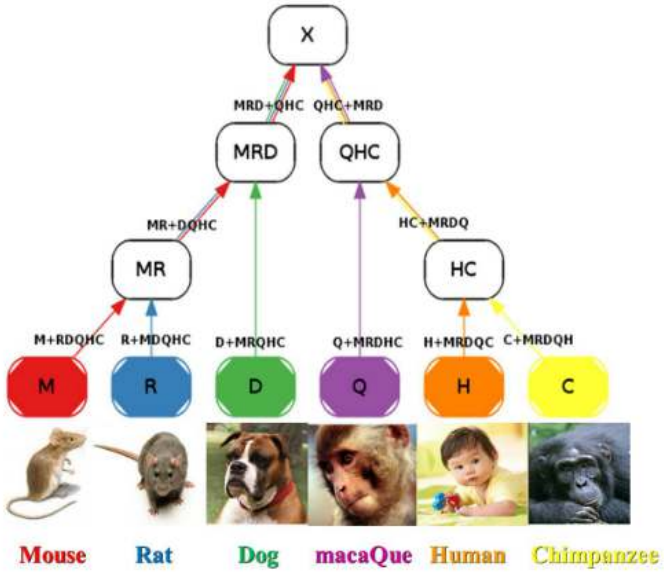
Rearrangement model

Every rearrangement event is considered as a two-break in break-point graph



Rearrangement model

The features of decent evolutionary path:
1. The minimum number of 2-breaks
2. Consistency with a given phylogenic tree



Tools: GASTS vs MGRA

GASTS solves the problem iteratively. It updates inner nodes by solving the median problem until the score stops increasing.

MGRA2 uses for flexible approach. It consider a generalizetion of a break-point graph - multiple break-point graph. It allows to update the whole tree in the same time

MGRA2 shows better scores except
3 genomes - case.

My goal: to close this gap

Couple words about complexity and heuristics for the median problem

The variants number for checking is exponential.
Branch and bound approach can be useful

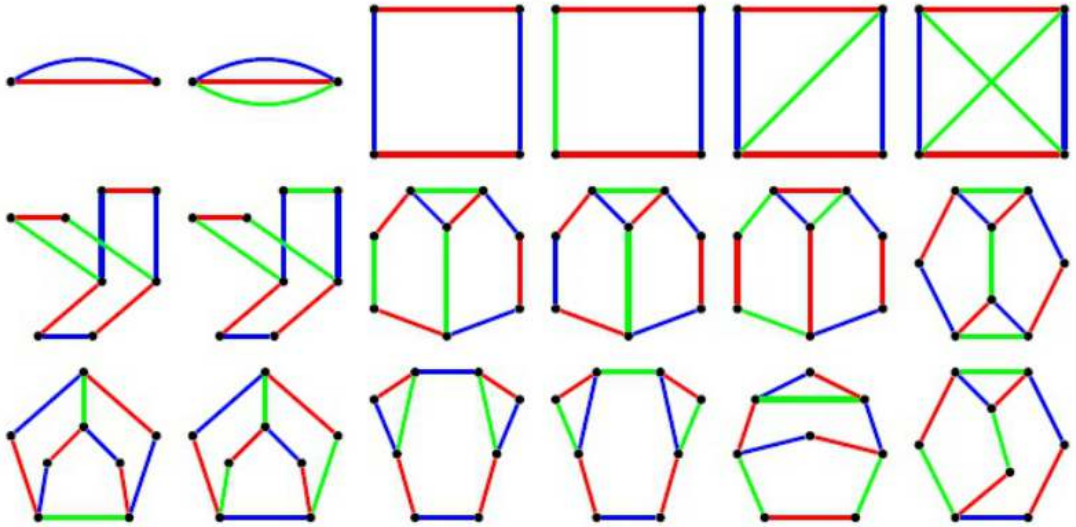


Fig. 7. Simple adequate subgraphs of size 1, 2 and 4 for MBGs on three genomes. See reference [11] for how they were identified.

Current results and further goals

1.Several adequate graph patterns were carried from GASTS to MGRA.
2. The number of such patterns was estimated on a simulated dataset. As expected, it is a small one.

1. To finish developing median-solver in MGRA using heuristics from GASTS or, probably, more modern estimates.
2. To improve current MGRA-heuristics. There are some patterns where it seems promising(4-Cycles)