Inferring genomic duplication events.

History of evolution could be represented by a species tree. Such tree in structure and purpose resemble a genealogical tree, which is created only for male members of the family with the same surname. In order to obtain structure of a species tree one can use comparison of genomes of selected species (A,C,G,T sequence alignment).

Generally speaking, the gene is a fragment of the genome sequence, which is encoding a protein. Analogically, we can divide genes into families and for every family create gene tree by sequence comparison. A distance between two elements of a tree can be measured by the total number of edges between them.

Hypothetically it may happen that in a species tree for a human, a monkey and a pig, the distance between the human genome and the genome of a monkey is smaller than distance between the genomes of a human and a pig. On the other hand, we may assume that the human gene is closer to the pig gene, rather than to the monkey gene, for instance when we consider the gene tree of hemoglobins. The process known as reconciliation is used to explain this inconsistency.

The phenomenon of gene duplication is related to the reconciliation. In general, we want to find the minimal number of gene duplications that explains inconsistency between species tree and gene trees.

Given a set of gene trees and a species tree, the Episode Clustering (EC) problem is to find the minimal number of locations on the species tree. Such locations, i.e., nodes of the species tree, are called *duplication episodes*. A single duplication episode is usually related to multiple duplication events that induce the set of subtrees in gene trees. The height of the tallest subtree determines the number of events in a selected episode. Minimal Episode (ME) problem is to group duplications into minimal number of events. Grouping gene duplication onto minimum number of locations do not always give a minimum number of events and vice versa.

For definiteness we consider binary trees. A rooted tree is a tree with selected node called root. For a species tree this is a node with no ancestors, which is representing the ancestor of every species in that tree. When we study relations between genes it is more natural to represent them as an unrooted tree (without the highlighted item from whom all other descent). Unrooted gene family trees are frequently inferred by phylogenetic methods. EC and ME problems have linear time solutions for the version, when gene trees are rooted. We proposed a solution for unrooted EC problem. However, the complexity of unrooted EC problem and unrooted ME problem are still open problems without solution. The aim of our project is to study them and find the solutions.

In our preliminary work we focused on EC problem for unrooted trees. We shown an efficient fixed parameter tractable algorithm that reduces this problem into the episode clustering problems defined for rooted trees. Article describing results of that work was recently accepted to BMC Genomics. We used the theory of unrooted reconciliation. That preliminary work could be treated as the proof of concept for the work in this project.

Our goal is to provide a solution to an open problem, which is unrooted ME problem. Our strategy is to study the complexity class of the unrooted EC and unrooted ME problems. Our goal is to implement invented solutions. As a result of this project we will provide novel tools that not only could improve known results but also will be more applicable than existing ones by dealing with problems that existing tools cannot solve. Our tools due to their independence of the fact if the input gene trees are rooted or unrooted could be easily incorporated into pipelines made from existing applications. Moreover, our comparative study of the implementation of our FPT-algorithm shown that we can improve known results on genomic duplication inference from real datasets for unrooted EC problem.

Our aim is also to perform experiments on real biological datasets. We believe that they could lead to novel results in the area of multiple gene duplication and whole genome duplication.

Summing up, our research will enrich the field of comparative genomics. To mention some practical usage, evolutionary trees were used to study the dynamic range of patients' cancer progressions, in order to tailor corresponding treatments. Species trees were used to develop pesticides and to predict outbreaks of infectious diseases.