Genome assembly algorithms for genetic disorders diagnosis

The order of nucleotides in DNA as the basis of genetic information is essential for understanding the biological activity of every living cell. The process of determining this order within a DNA molecule is called DNA sequencing. Currently we can sequence only fragments of DNA. These fragments are called reads and their characteristics depend on the technology used. Process of reconstructing genomes from such reads is called genome assembly. Features of the reads (length and accuracy) as well as features of the organisms' genome (how repetitive the genome is) have a great impact on the algorithmic complexity of the genome assembly problem.

Now there are two main types of sequencing instruments: short-read sequencers that generate reads of length from 50 to 500 base pairs with high accuracy, and long-read sequencers generating reads from one to several hundreds of thousands base pairs, but much more error-prone. Short-read sequencing methods appeared earlier and have been successfully used in many applications. Limitations of short-read sequencing, especially concerning restricted ability to mapping repetitive elements and spanning structural variants, have left substantial fraction of human genome inaccessible for the biological or medical analysis. Long-read sequencing technologies where introduced to meet the challenges of the assembly of complex and repetitive regions.

Nevertheless, even state-of-the-art biotechnologies cannot unassisted answer any biological or medical question without adequate bioinformatic pipelines and tools tailored for the interpretation of the data they produce. Long-read sequencing is now primary used in research project, but have a great potential for clinical application, especially in conjunction with the expert knowledge about medically relevant loci.

In this project we want to focus on two problems that can be solved using the long-read sequencing data with important clinical applications:

- Finding architecture of complex structural variants (ie. translocations, insertions, duplications, deletions and inversion), possibly involving more than two chromosomes;
- Resolving complex, repetitive regions of human genome including so-called segmental duplication.

We plan to implement algorithms for solving these problems, and then provide tools suitable for clinical interpretation of the obtained results.

In this project we want to create algorithms that leveraging on existing tools for the detection of breakpoints from long-read sequencing data will create a graph representing the structure of chromosomes after rearrangements. Using existing databases of repetitive elements and structural variations common in population or known to cause diseases we will evaluate the pathogenecity of detected structural variations.

Second aim of this project is to create algorithms for resolving regions containing segmental duplication (highly self-similar fragments of DNA that that appear in human genome in more than one copy). We want to approach the problem of finding the most probable scenario of genomic rearrangement mediated by segmental duplications in such regions from a set of scenarios obtained from the literature. Our algorithms we will be based on graph-theory and Bayesian inference approach.

We expect that derived methods and tools, thanks to the use of analytical and computational procedures, will provide the next step in diagnostics of patients with complex chromosomal rearrangements. Project will also allow to systematically explore genetic variations in regions enriched for segmental duplication, often overlooked in disease-association studies but frequently causing genetic disorders. In the long-term project will definitely improve our understanding of organization, diversity, and impact of repetitive regions and structural variations on phenotype, evolution, and disease. It seems that the greatest value of the proposed research, will be the development of the novel bioinformatic methods for studying genetic variations and the diagnostic potential of the proposed tools.