Genome sequencing has an unprecedented impact on our understanding of the species living on Earth. Since 2007, when the sequencers by Illumina started to dominate the market, the costs of sequencing decreased rapidly, e.g., from 10M USD to 1K USD for a human genome. This allowed many large-scale sequencing projects to launch, e.g., The 1000 Genomes Project, The Genomics England 100,000 Genomes Project, The French Plan for Genomic Medicine 2025, Earth BioGenome Project, The Human Cell Atlas, The 100K Pathogen Genome Project. The recent estimates suggest that in the 2025, the global acquisition rate of sequencing will be close to 1zetta-bytes/year, while 2–40EB/year will be stored for a long term.

Nowadays, the majority of the sequenced data come from the second generation of instruments produced by Illumina. This technology has a lot of assets, including high throughput and very good quality. Unfortunately, it has also some drawbacks. The most important one is that the fragments of genomes (reads) that can be decoded are very short, up to 250 base pairs, which can be compared with 3 billion base pairs in a human genome. Therefore, to determine a sequenced genome it is necessary to solve a puzzle. The reads come from randomly picked genome regions and the total length of the reads is usually tens time more than the genome length. Thus, there are significant overlaps that can be found. Of course the problem can be the volume of data, e.g., a billion of reads is quite typical for a single experiment. Moreover, some parts of genomes are highly repetitive and the repetitions are much longer than the read length. Thus, this technology does not allow to reconstruct the complete genome. Currently about 5% of human genome is still unknown.

The Illumina's instrument limitations has given the opportunity of the 3rd generation of sequencers to emerge. The most prominent examples are Single Molecule Real Time (SMRT) technologies by Oxford Nanopore Technology (ONT) and Pacific Biosciences (PacBio). The technical details of the ONT and PacBio instruments are different, but both offer much longer reads, even up to 1 million base pairs. Unfortunately, their serious drawback is the base error rate ~10%, which can be compared with 0.1% of Illumina instruments. Nevertheless, a remarkable symptom that in the near future we should experience the change in how the data are sequenced is the recent acquisition of Pacific Biosciences by the market leader Illumina.

The high error rate and the much longer reads require to redesign the existing software, or create completely new tools, for reads analysis and compression. Several successful attempts into this direction has been done. Nevertheless, still the existing tools are far from being perfect. Therefore, the main aim of this project is to design new algorithms and data structures, as well as prepare prototype implementations for several problems that can be encountered when we deal with the third generation sequencing data. Some of the tools will be related with the compression of sequenced reads and the results of some analyses. Some other will aim at finding the overlaps among the long and noisy reads. One of our goals is also to prepare a tool that can be used in the almost real-time diagnosis in the medicine.