The dramatic progress in computational biology in the last few years, and in particular development of cheaper and more available next-generation sequencing methods, is accompanied with the rapid growth of data, which need to be stored, transferred, analyzed, seriously challenging the existing hardware infrastructures. This phenomenon is sometimes called "data deluge", and indeed this is not an overstatement.

In the years from 2003 to 2015 the real cost of sequencing of a single human genome dropped from about 30 million dollars to about 4 thousand dollars. Moreover, recently new sequencers like Illumina TenX were launched. They are expensive but offer human genome sequencing for as little as 1 thousand dollars. Other sequencers of the newest generation (IonTorrent, Oxford Nanopore) produce long reads (much longer than obtained from Illumina machines). They error rate of these reads is high, nevertheless, they length can compensate the low quality. All these leads to the conclusion that the personalized medicine seems to be just behind the corner as we have a fast and relatively cheap way of recovering the genome.

It is worth noting that the world-largest genome sequencing center, Beijing Genomics Institure (BGI) in China uses 230 sequencing instruments, and most of them are the newest Illumina HiSeq 2000/2500. The theoretical throughput of the institute is about 1.5 Pbp per year, which results in at least 10 PB of space. The storage sizes in the largest sequencing institutes are on the order of a few tens of few hundreds of PBs. It is hard to give precise costs of storage in such institutes as BGI, but some good estimation could be the costs of storage and transfer of one of the largest data centers and cloud computing centers, which is Amazon EC2. The cost of one year storage of 1 TB of data together with the cost of 15 downloads of these data is more than 1 thousand dollars. it is still about 4 times less than the total cost of sequencing (including the IT costs), but the ratio changes - sequencing costs fall rapidly, while IT costs diminish much more moderately. Thus, many institutes decide to remove archive data, when only 2-year-old files can be considered old and thus removed.

The traditional means of handling huge data are compression techniques. Those methods allow for not only space reduction and data circulation (e.g., among research institutes) speed-up, but also, in many cases, for data analysis acceleration directly in the compressed form. An important motivation for compression in bioinformatics is the possibility to process more data at a time on a particular machine, with a limited amount of RAM. Note this is crucial e.g. for tasks like de novo mammalian genome assembly, where the auxiliary structures (e.g., string graph, de Brujin graph) are many times greater than the corresponding genome!

The main goals of the project are:

- development of a memory-efficient and fast data structure representing results of genomes of a collection of individuals of the same species,
- development of new algorithms for analyzing data from genome sequencing, making use of the knowledge of more than a single reference genome, i.e. genomic sequences of a collection of individuals will be used to improve the results,
- development of new algorithms for alignment of genomic sequneces; moreover, the new compression algorihtm will be invented for the results of such alignments,
- improvement of the invented fast and memory effifient algorithms for k-mer counting; moreover, the obtained efficient k-mer counter will be used to solve some problems from the genome analysis field, in which the frequencies of k-mers are crucial.

The common factor of the project goals is data compression. Thanks to it, it will be possible to reduce the requirements of necessary disk space. What is most important, a compact representation of data in RAM (or temporary disk), will allow to run the developed algorithms, opposite to the existing algorithms, at relatively cheap workstations. In the longer perspective, this will allow to reduce the costs of processing in bioinformatic centers. Moreover, individual researchers will be able to solve some of the bioinformatic tasks on their own PCs, rather than on expensive servers.