In the recent years, the progress in bioinformatics is dramatic. In particular it is possible to sequence individual genomes quite cheaply. These techniques become accessible even for diagnostic purposes. As a consequence the promises of personalized medicine turn into reality.

There are also some drawbacks of this situation. One of them is the huge size of the acquired data. For example, in a typical genome sequencing of a single human, the raw data consume about 300 GB. When we take into account also the data from following stages of analyzes the total size exceeds 1 TB. It is not a concern for a single person. We should, however, remember that in the personalized medicine the genome sequencing will be a routine task for many patients. This would lead to huge challenges for laboratories that would store gigantic amounts of data. To imagine the scale of the problem, it suffice to mention one of the recently published estimations, according to which in 2025 data from genome sequencing will be acquired at about 1 zettabyte (10^{21}) per year, of which about 2–40 exabytes (10^{18}) will require long-term deposition. These values are enormous, but we should also mention that the contemporary realistic cost of one year storage and 15 transfers of 1 TB of data is about 1 thousand USD.

A classic attempt to reduce the scale of such problems is data compression. Thanks to the reduction of amount of space occupied, data can be faster (and cheaper) transferred and stored, which is of large practical importance. The compression ratios that can be achieved strongly depends on the data and on whether a lossy compression (controlled removal of some parts of information) is an option.

In our project we focus on results of one of the last stages of genomic data processing pipelines, i.e., complete genomic sequences of individuals and descriptions of differences (variants) between individuals and so-called reference genome (some approximation of genomes of all individuals of a given species). Nowadays, taking into account such data, researchers try to predict the risk of genetic diseases.

The main scientific goals of the project are:

- development of an algorithm for compression of collections of genomic sequences of individuals of the same species,
- development of compressed data structure allowing highly efficient storage of description of variants present in a huge collection of individuals; the fast access to the compressed data will be also important,
- development of an universal compression algorithm of VCF (Variant Call Format) files for storage of results of genomic analyzes, especially huge collections of individuals,
- development of compression algorithm for gVCF files containing detailed results of analyzes of a single individual.

The common factor of all the mentioned goals is a development of novel compression algorithms or compressed data structures for various types of genomic data. In consequence, we should develop tools significantly reducing costs of storage, transfer, and analyzes of genomic data. We believe that due to our works genome analyzes will be available for broader range of researchers, especially for those that have no (or limited) access to expensive IT infrastructure. Often the analyzes could be made using a small workstations, or even laptops.