Growth of tumors and populations of cancer cells stems from accumulation of mutation changes acquired along the period of many years preceding diagnosis of cancer. Mutated cells are growing in uncontrollable manner, which leads to development of cancer diseases.

Since many years it is known that an important factor, which characterizes cancer cells is the genetic diversity of the population of cancer cells, called clonal structure. The name “clonal structure” stems from the fact that each of the genetically mutated cells can initiate the development and growth of the whole sub-population of descendant cancer cells, which is called a clone. Genetic (clonal) diversity of cancer cells results in the fact that as a population they gain more malignant character and become resistant to therapies.

Clonal theory of cancer development was formulated already in seventies of the last century by P. Nowell. In recent years, thanks to the new techniques of DNA sequencing, clonal theory was fully confirmed. A lot of research studies is now oriented towards precise characterization of clonal structures of cancers and using detected/estimated clonal structures in cancer diagnosis and therapy. These researches include a lot of mathematical modeling and the use of information technologies, which stems from the character of DNA sequencing methodology and data.

Cancer tissue is an evolving population of cancer cells with diversity increasing as the tumor advances in development. Gathering insights to tumor structure and development by analysis of DNA sequencing cancer genomics data and by developing mathematical modeling approaches, is of great interest and significant practical importance. However it poses many challenges for researchers. Tumor sequencing data available in the cancer genomics TCGA database (The Cancer Genomics Atlas Research Network, 2008) concern huge number of cell samples. Cell count in cancer tissues exceeds billions, and biopsies include upwards of millions of cells. Developing algorithms for analyses of cancer genomics sequencing data requires addressing not only the problem of large sample size but also numerous additional issues specific to that type of data. Typical sequencing cancer genomics data include reads obtained from a mixture rather than from separate cancer cells, which calls for the development of specialized approaches combining large sample modeling with statistical methodologies for describing mutations discovery processes. Mutational events of diverse types, such as chromosomal duplications, the loss of heterozygosity and rearrangements interfere with the point mutation processes. Mutations seen in the cancer sequencing samples are classified as either driver or passenger, which is related with their roles in the selection mechanisms in the carcinogenesis process. Driver mutations are defining new clones creating cancer cell population subdivisions, leading to the need for further model refinement.

In the planned project a research will be pursued focused on development of algorithms for estimating clonal structures of cancer cell populations and on estimation of their growth rates. The research will be conducted in cooperation between two scientific groups, bioinformaticians and molecular biologists. The designed algorithms will be verified on the basis of data from cancer genomic sequences available in the Internet. Also, experiments concerning sequencing of DNA of tissues of thyroid cancer will be performed. Clinical database of thyroid cancer tissues was created by clinicians within the framework of earlier scientific projects. It will be partly made available for the research planned in the project.

We have already downloaded substantial volumes of data from the TCGA database (over 5TB) and we have pursued preliminary researches concerning developing simple models of clonal evolution. We studied correlations between parameters of our model with the data on survival of 50 glioblastoma multiform patients obtained from the TCGA database. Much more exhaustive and detailed approaches are scheduled in the planned project.