ABSTRACT

Master's thesis: 90 pp., 25 fig., 2 tables, 1 app., 72 sources.

The relevance. This work is devoted to the research of the primary structure of DNA. The problems associated with the research of the primary structure, primarily related to the problems of recognition of the protein-coding regions in the sequenced nucleotide sequences of DNA.

From the math perspective, the given problem is a classification problem, which could be solved by a dozen of different methods. Most of such methods are based on machine learning. The main problem here is that the existing volume of data can't be processed by classical methods of machine learning, especially considering the computation speed, which could be increased in terms of single machine nowadays. Despite the fact that the distributed computing has opened new ways of solving problems that require large computing power, distributed machine learning methods are still relatively new and almost unknown. Ability to use the whole power of distributed computing and computing clusters with machine learning will solve such kind of problems and will give the ability to process large volumes of data much faster.

Connection of paper with scientific programs, themes, topics. Master's thesis is done with regard to plan of department of inductive modeling and control methods of V.M. Glushkov Institue of Cybernetics of NAS of Ukraine in scope of research topic «Development of methods for modeling biological sequences and processes based on the active particles apparatus» (code VF.235.14, state registration number 0114U000359, 2014-2018).

Purpose and objectives of the study. The goal is to increase speed of genome processing through distributed computation.

To achieve this goal it is necessary to solve the following problems:

- make inspection of the known methods of recognition of introns and exons in DNA;
- develop a formal statement of the problem of distributed machine learning to recognize introns and exons in DNA;
- develop algorithms which supposed to solve described problem;

– implement algorithms in the form of software;
– benchmark the efficiency of the developed algorithms through computational experiment and make a conclusion.

The object of the study is a process of the genome decryption.

The subject of the research is the task of recognizing exons and introns in DNA.

Methods are based on machine learning algorithms and distributed computing.

Scientific novelty of the results – introduced distributed machine learning methods which are based on the naïve Bayesian classifier and binary logistic regression and adjusted to solve the problem of recognition of introns and exons in DNA.

Publications. Materials are published in the scope of the iScience XXV international scientific conference «Recent challenges of modern science» [71], international scientific conference «Innovative development of science of the new millennium» [72] and conference «IOT-2017».

MACHINE LEARNING, PATTERN RECOGNITION, CLASSIFICATION, EXON, INTRON, DNA, NAIVE BAYES CLASSIFIER, BINARY LOGISTIC REGRESSION, APACHE SPARK.