

STATISTICAL ANALYSIS AND ERGODIC PROPERTIES OF GENOMES AND PROTEINS, MARKOV CHAINS BASED LEARNING PROCEDURE

A.Gupal(1), O.Potashko(2)

**1-Institute of Cybernetics of National Academy of Sciences of Ukraine
Kiev, Ukraine alexandra_vagis@yahoo.com**

2-Ukrainian Astrobiology Association, Kiev, Ukraine potashko@mail.ru

Genetic properties of organisms grow out known processes much not enough and mechanisms. Studying with the help of the determined models poorly effectively. We shall believe, that genetic properties are results of stochastic processes. Therefore the information which contains in DNA and proteins, is investigated on the basis of probability models.

The analysis of DNA and proteins represents huge interest as they contain the information which collected and improved by the nature during evolution during billions years. To solve various problems of a statistical property for such objects, it is necessary to construct probability models of the description of sequences of DNA and proteins. We use Markov chains and Beyes networks.

Human Genome. The statistical analysis of human genome is lead. The following law is established: frequencies of letters A and T, and also C and G are equal among themselves for all 24 chromosomes, i.e. the principle complementation is carried out not only for two opposite chains, but also for each chain separately. Chromosomes are indivisible parts of DNA. Concurrence of frequencies nucleotides at all chromosomes is connected to carrying out of process DNA replication. As this process proceeds in the water environment of a cell, that the normal course of replication proceeds successfully if frequencies nucleotides in a chromosome correspond to nucleotides concentration in water solution. The statistical analysis of estimations of transitive probabilities is lead, is shown, that estimations asymptotically are normal, received dispersions and covariatively this limiting distribution.

Calculations have shown, that estimations of transitive probabilities also are rather stable on all chromosomes human genome. The analysis of others genomes higher organisms (mouse and rats) has shown, that at these genomes the same chain of record of the information, as in human genome.

Calculations on human genome have confirmed a conclusion that homogeneous Markov chains in the best way corresponds to the data which contain in genome chromosomes. By virtue of that chromosomes contain the information on many thousand genes, appeared, that frequencies of letters on different sites of a chromosome can differ considerably from each other; and only right at the end of a chromosome the saved up frequency is stabilized at a level of values of the frequency calculated on all length of a chromosome.