DOI 10.18699/BGRSSB-2018-064

The use of a horizontally scalable infrastructure in the search for genetic similarity in biodiversity

A. Tskhai[1, 2]*, S. Murzintsev[3]
[1] *Institute for Water and Environmental Sciences SB RAS, Barnaul, Russia*
[2] *Altai State Technical University named after I.I. Polzunov, Barnaul, Russia*
[3] *Altai State University, Barnaul, Russia*
** e-mail: taa1956@mail.ru*

Key words: similarity of genomes, large data, nonrelational databases, search algorithms for repetitions

Motivation and Aim: The exploration of the structural-functional organization of biodiversity continues to be a main direction, developing at the intersection of biology and information technologies [1].
Methods and Algorithms: In this scientific research the problem of rapid detection of genetic similarity is considered in the analysis of databases (DB) of genomes of individuals from ecosystems of different levels. The distributed non-relational DB MongoDB and the Winnowing data processing algorithm are used as the basis for creating the information system. Using a non-relational database to identify genetic similarity, a variant of representing the prints of the structural variations of the genomes in the form of "key-value" was proposed, a program implementation of the developed model was carried out, and computational experiments were carried out, which confirmed the possibility of using the proposed method of genetic similarity search, for example, in a personified analysis of deviations in the gene level.
Results: The results obtained in the scientific research allow to find confirmation of the hypothesis of the applicability of distributed non-relational databases for comparison and searching for deviations in the development of living beings on the basis of a personified analysis of their genomes. The database of the Kyoto encyclopedia of genes and genes KEGG GENOME (Sequenced genomes of various living organisms) was selected as a source of elements of the genome database of organisms, including decoded representations of about five thousand living beings (http://www.genome.jp/kegg/catalog/org_list.html). The genome size of one living creature reaches 1 GB in compressed form. The total volume of genomes of living beings in this information resource, therefore, is approximately five TB. Due to the large amount of data that occurs during the processing of the original information, the transition from relational databases to non-relational databases has been carried out, both to a more flexible and scalable database.
Conclusion: The development of scientific research is considered from several sides. The first direction is the solution of the problem of determining the moment at which it is necessary to add a node to the cluster of computers with increasing the number of deviations considered and increasing the number of genomes in the database of organisms. The second is the practical filling and further formation of the database with as many real genomes as possible. The application of the results obtained in interdisciplinary studies would allow us to speak about the development of the proposed direction of research of genomic disorders. This direction is focused on obtaining an assessment of the probability of genetic abnormalities at the stage of recognition of the potentially unfavorable development of the situation. In the case of gaining access to the production databases of genomes of humans, animals and plants and conducting joint research with genetic specialists, all of the above looks real.

References

1. Noisy V.C., Shokin Yu.I., Kolchanov N.A., Fedotov A.M. (2006) Biodiversity and ecosystem dynamics: information technology and modeling. Novosibirsk: SB RAS Publ. House, 2006.