

Data lake platform of the bacterial strain properties for microbiology

Mukhin A.^{1,2*}, Demenkov P.^{1,2}, Kazantsev F.^{1,2,3}, Lakhova T.^{1,2}, Klimenko A.^{1,2,3}, Lashin S.^{1,2,3}

¹ *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

² *Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

³ *Novosibirsk State University, Novosibirsk, Russia*

* mukhin@bionet.nsc.ru

Key words: database, data-lake, microbiology, SQL, java, genomes, annotation

Motivation and Aim: Modern high-throughput sequencing methods lead to generation of large amounts of genomic data. These data then processed with generating additional bunches of data. All of it should be collected in a structured way like a data warehouse. In developing such repository, one must also consider that data may be described either incompletely, or may not be fully annotated, or may be described using non-standard or erroneous formats [1]. Because of these problems, it is impossible to reuse the primary study results in subsequent computational experiments. All of these things must be taken into account in solving the research tasks of finding biotechnologically meaningful properties of microorganisms. The work is aimed at the tasks of the Kurchatov Genomics Center of the ICG SB RAS that collects large amounts of bacterial strains with biotechnological potential and performed DNA sequencing.

Methods and Algorithms: In this work we used Java and Kotlin programming languages, web framework Spring Boot to build the server part of the system. HTML, CSS and JavaScript programming language are used to build client part using VueJS framework. PostgreSQL is used as the database management system. The Data Lake architectural concept was chosen to create the repository. It stores both primary and derived structured data within a single software system. The reuse data problem is solved with taking into account FAIR [2] principles. The task of comprehensive describing the bacterial strain properties is important from both a scientific point of view and an engineering one (creating producers of certain substances or maximizing biomass).

Results: The developed system is named MicrobialDataLake. It includes data pipelines from a series of bash-based scripting tools: bacterial genome assembly and annotation, taxonomic identification of the strain, generation of structural (graph) and mathematical frame models. These results form the informational basis of the MicrobialDataLake system and are located in the raw data repository. To describe the results of analysis and links to the raw data, a JSON schema format has been implemented. The PostgreSQL database management system was also deployed to describe in detail biotechnologically relevant properties of microbial strains (such as the lists of enzymes synthesized by the strain, taxonomic affiliation, gene and protein sequences, etc.). To parse and upload the structured properties of each microorganism, a script was implemented in the Kotlin programming language. A natural tool for working with genomes is the genome browser; the JBrowse2 browser for genome visualization was integrated into the software system. A web service providing access to the data via REST API was also implemented using Java and Spring Boot framework. The Kotlin tools and the Java web service were

developed using common domain classes and database accesses in order to develop a unified platform.

At the moment, the data of 1004 bacterial strains are presented in the system.

Acknowledgements: The study is supported by the Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS (075-15-2019-1662).

References

1. Roche D.G., Kruuk L.E.B., Lanfear R., Binning S.A. Public data archiving in ecology and evolution: how well are we doing? *PLOS Biol.* 2015;13(11):e1002295. doi: 10.1371/journal.pbio.1002295.
2. Wilkinson M.D. et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016;3(1):1-9. doi: 10.1038/sdata.2016.18.