DOI 10.18699/BGRSSB-2018-048

Graph-oriented database for analysis of prokaryotic communities -*omics* data

A. Ryasik[1]*, T. Ermak[1, 2], M. Orlov[1], E. Zykova[1], A. Sorokin[1]
[1] *Institute of Cell Biophysics RAS, Pushchino, Russia*
[2] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
* *e-mail: ryasik.aa@gmail.com*

Key words: metabolic reconstruction, metabolic model assembly, -*omics* data integration, metagenomics

Motivation and Aim: We have constructed a graph database containing information about proteome, metabolome, transcription and translation regulation in various prokaryotes, that is suitable as for analysis of physiology of microbial communities, as for analysis of strain and isolates differences within species. All the data were taken from well-known external databases (Genbank, Uniprot, RAST, etc.). Advantages of data storage in graph structures allowed us to construct metabolic models of the organisms and strains based on the similarity with reference genomes. The aim of the research is the automatic assembly of the metabolic network flux models for bacteria by their protein similarity to reference genome, preparation of metabolic networks for the communities and comparative analysis of these models that can be applied for the distinction of the strains and prediction of physiological abilities of the communities.

Methods and Algorithms: We recently developed the BioGraph database, graph-oriented storage for information about prokaryotic organisms. We collected various -*omics* data (genomic, proteomic, taxonomic) and integrated it by graph representation with predefined types of nodes and edges. This approach was used to create strict rules for integration of -*omics* data of various origin. The implemented network of structural and semantic similarity relationships between proteins, genes, organisms, and reactions enables constructing models for close strains and species using the reference model as a template. Furthermore, such approach makes it possible to extract the existing model from the database as an SBML file. The developed tool was applied for analysis of different bacterial societies such as human microbiome. Also, we build a system of asynchronous queries based on actors for complicated comparison between the large number of nodes with multiple conditions, which could not be completed with online queries. Also, asynchronous queries allow fetching the data from remote sources such as UniProt to keep the BioGraph up to date.

Results: We developed a graph database that contains different -*omics* data; developed tools that allow to assemble metabolic flux models in automatic mode from different -*omics* data such as genomic annotation or proteins and genes similarity; tools for complex graph queries and database updates.

Availability: Source code of the software is available on Github: https://github.com/arc7an/scalaBiomeDB