Technical Faculty
"Mihajlo Pupin" Zrenjanin
University of Novi Sad
Zrenjanin, SERBIA
http://www.tfzr.uns.ac.rs/

Faculty of Information and
Communication Technologies
University of St. Clement Ohridski
Bitola, MACEDONIA
http://fikt.edu.mk/

International Conference on
Applied Internet and Information Technologies
ICAIIT 2017

PROCEEDINGS

Zrenjanin
October 5-6, 2017

Organizer:
Technical Faculty ”Mihajlo Pupin”,
University of Novi Sad,
Zrenjanin, Republic of Serbia

Faculty of Information and Communication Technologies
University of St. Clement Ohridski
Bitola, Macedonia

Publisher:
University of Novi Sad, Technical Faculty ”Mihajlo Pupin”
Djure Djakovica bb, Zrenjanin, Republic of Serbia

For publisher:
Dragica Radosav, Ph. D, Full Professor,
Dean of the Technical Faculty ”Mihajlo Pupin”

Technical preparation and design:
Vladimir Brtka, Dejan Lacmanović, Predrag Pecev, Željko Stojanov

Cover design:
Predrag Pecev

CIP - Каталогизација у публикацији
Библиотека Матице српске, Нови Сад

004(082)

INTERNATIONAL Conference on Applied Internet and Information Technologies (7 ; 2017 ; Zrenjanin)
Proceedings [Elektronski izvor] / [7th] International Conference on Applied Internet and Information Technologies ICAIIT 2017, Zrenjanin, October 5-6, 2017 ; [organizer] Technical Faculty "Mihajlo Pupin", Zrenjanin. - Zrenjanin : Tehnički fakultet "Mihajlo Pupin", 2017. - 1 elektronski optički disk (CD-ROM) : tekst ; 12 cm

Nasl. sa naslovnog ekrana. - Str. [VI]: Introduction / Vladimir Brtka. - Bibliografija uz svaki rad.

ISBN 978-86-7672-304-1
1. Technical faculty "Mihajlo Pupin" (Zrenjanin)
а) Информационе технологије – Зборници

COBISS.SR-ID 317671687

An Environment for Metagenomic Analysis

Evgeny Cherkashin[1,2,4], Alexey Shigarov[1,2], Fedor Malkov[1,2],
Kristina Pascal[2], and Alexey Morozov[3]

[1] V.M.Matrosov's Institute of System Dynamics and Control Theory of SB RAS,
Lermontov str. 134, Irkutsk, 664033, Russia
{shig,eugeneai}@icc.ru
[2] Irkutsk scientific center of SB RAS, Lermontov str. 134,
Irkutsk, 664033, Russia
[3] Limnological Institute of SB RAS, Ulan-Bator str. 3, Irkutsk, 664033, Russia
[4] National research Irkutsk state technical university, Lermontov str. 83,
Irkutsk, 664033, Russia

Abstract. Metagenomic analysis allows describing microbial community with a previously unavailable precision, but requires considerable computing power for solving bioinformatics problems and participation of domain specialists at the stage of the result interpretation. This complicates the implementation of the analysis in a broad biological practice. The development of a domain user-friendly software environment for storage and analysis of metagenomic data has been started. The usage of a dataflow programming system for representation of metagenomic analysis and the schema for a SQL database for storage of the metadata are considered as units of the environment.

Keywords: bioinformatics, metagenomic analysis, Big Data.

1. Introduction

In the last decade, thanks to the invention of next-generation sequencing (NGS) methods and their introduction in practice of research of biological systems a field of research of molecular genetics, namely metagenomics, has been arisen. Its basic principle is that the object under investigation is not a separate microscopic organism, but their communities (microbiomes). The sampled probes stand out with a total DNA sequencing data over the whole set of genes of all microorganisms in the probe. That is, the studied object is the microbiome as a whole, not only those organisms which can be cultivated in laboratory conditions or identified with microscopic or microbiological methods.

Metagenomics allows us to describe a significant number of new groups on all taxonomic levels, broadening the field of view of the world science. A characteristic example is the recently discovered group CPR (*candidate phyla radiation*). No CPR is isolated in a culture at the moment. According to genomic data, its representatives differ in the set of ribosomal proteins, the absence of certain key metabolic pathways and the presence of self-splicing introns in genes 16S rRNA [1]. Phylogenetic analysis indicates that this group is a sister to all other bacteria, and the level of divergence is not inferior to bacteria, not to mention the eukaryotes [2].

There are two main types of metagenomic studies. The first one, which is simpler, is called *analysis of the amplicons*. In this case a specific taxonomic marker is amplified and sequenced. The marker is universal for the studied species. Usually, the sequence of the small subunit of ribosomal RNA is used as the marker, as this gene is widely used in phylogenetics. The gene is available in numerous reference sequences. For example, the release 128 of widely used in amplicon analysis SILVA database [3] contains 645 151 unique rRNA. The reads obtained from the DNA sequences extracted from the sample under investigation are compared to the sequences in databases, attributing them to a particular taxon of a taxonomic level, obtaining information about the diversity of the microbiome in the studied environment.

The second approach is known as *metogenomic shotgun method*. It is based on sequencing the whole DNA sample instead of the specific locus. With sufficient coverage, this approach allows describing the taxonomic composition of the community, as well as the genes of functional or structural proteins presented in the representatives of the community, including viral ones [4]. On the basis of metagenomic data, metabolic interactions in individual microbiomes can be determined using the databases ePGDBs (environmental pathway/genome databases) [5]. In several works, full genomes of individual species were isolated from metagenomic dataset reads [6].

In recent years the amplicon analysis was applied in microbiome studies for different environments of lake Baikal. The researchers of the Limnological Institute of SB RAS described the under-ice bacterial communities associated with blooms of diatoms [7] and bacteria in photic layer during spring [8]. Bacteria inhabiting the Baikal sponges were studied as well [9]. Finally, the bacterial communities of bottom sediments in the areas of hydrocarbon yields [10,11] were investigated.

In order to carry out the metagenomic studies the significant computational resources and bioinformatics skills are required for data processing and interpretation. The software used for analysis of amplicons includes various library modules of sequence processing, for example, Mothur [12], USearch [13], statistical packages and development environments of data mining algorithms, e.g., R (**`https://www.r-project.org`**). In order to carry out the studies of metagenomic data, the specialists are required to be able scripting the command shell of an operating system (Linux, Windows), running programs in a distributed computing environment and cluster computing systems, and programming with general-purpose languages, usually R or Python.

Another important problem is the organization of a centralized data storage and providing the efficient regulated access to the data for the users. At the moment the staff of LIN SB RAS conducted numerous amplicon research of the different ecotopes of lake Baikal, the data were collected for several years. There are no strict rules of the storage policy of input, intermediate data and the obtained results. Comparison and integration of data from different studies is also complicated due to its heterogeneity, resulting from the use of various software. The implementation of a system for storing input data, metadata, and results of metagenomic studies in a unified form will simplify the integration of results from different studies and the comparative analysis.

The goal of this study is a software environment development for supporting the processes of new-generation sequencing with organizational, informational and computational resources.

2. The domain analysis

Domain analysis showed that the problems solved in the bioinformational part of metagenomic analysis, together with NGS itself, are well represented within the paradigm of Big Data. At the moment, the scientific community developed data formats for representation and storage of metagenomic information, algorithms and software modules including distributed and parallel implementations on cluster computing systems providing different stages of data analysis.

The solution of the problems within the Big Data paradigm requires the biologist to have software development skills to be a professional programmer in bioinformatics. In order to carry on the analysis of each probe, biologist is to construct and execute a separate program script or perform stage-by-stage execution manually to control each step's results quality. This approach significantly slows down the process of obtaining the final result.

The proposed organization of studies is based on the creation of an information-computational environment that allows one to design and execute scenarios, giving the input data in various formats from various sources, e.g., files, databases, servers of metagenomic information. The environment must also support a cloud storage for intermediate data and the obtained results. A collaborative project of LIN SB RAS and ISDCT SB RAS is devoted to the construction of the environment for the research support. The following problems are to be solved within the project.

1. The subject area and its functional modeling. The classes of functions (problems) are being recognized and presented in the form of software modules. Modules form scripts of problem solving, network graphs of modules connected by data transmission.
2. Metadata descriptions of the modules and structures of input and output data. At this stage, it is necessary to deal with the problem of integration with external information and computational resources. In this case, the standards and standard means of data modeling like ontologies are of critical usage.
3. Decomposition of the input/output data formats and implementation of subsystems of their transformation, accumulation, storage and effective (according to the criteria of time and computational complexity) regulated access.
4. Construction of virtual executional environments and software interfaces for modules, whose source code is inaccessible due to the lack of the source code or licensing restrictions.
5. Development a customized user interface for high level control of the scenario executions. At this stage, a visual programming with the user interface for script development and execution is required to provide flexibility for managing computational processes by domain specialists.
6. Development of subsystems of visualization and interpretation of obtained results, including the modules for interpretation of the process of metagenomic analysis.

3. Dataflow representation of the domain

A popular approach to the representation of the computational process is dataflow programming [14]. The data flow programs are constructed as a combination of the executable modules. The modules receive input data, process it, and produce output. The approach is being developed since the 1970-ies.

An example of usage of the script construction system under development is shown in Fig. 1. The figure shows an example of an initial stage of a computing process of analysis of the amplicons.

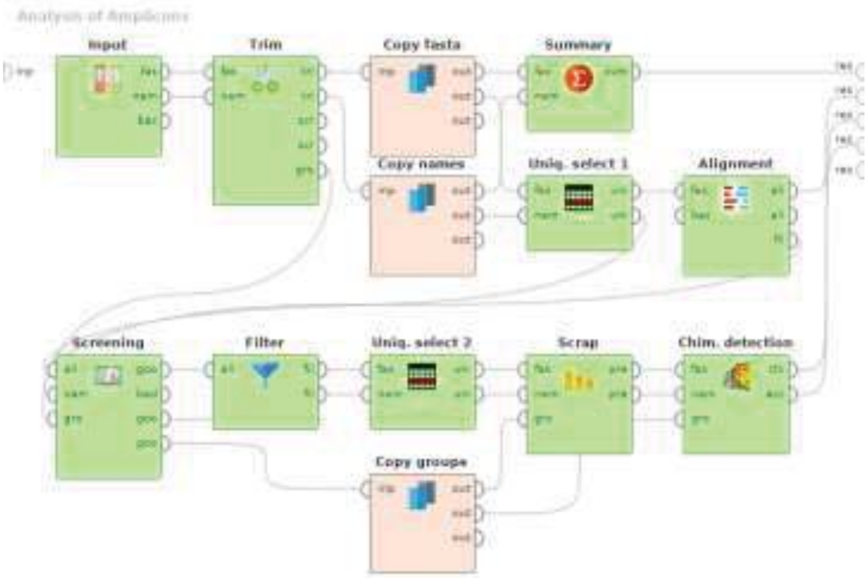

Fig. 1. The initial stage of the metagenomic analysis represented as a data flow.

The presented script was constructed by means of the software package Rapidminer Studio (https://rapidminer.com/) supplemented by our extension module for description of the amplicon analysis stages. The scenario includes the following operations:

- definition of a research project as a set of input files containing the sequencing data in a directory (module “Input”);
- trimming reads (module “Trim”);
- the module “Summary” is used for visual quality analysis of the results of the previous steps;
- the reduction of the volume of input by the removal of insignificant information, for example, overlapping sequences (modules, “Uniq. select …”);
- alignment of sequences to the reference database (module “Alignment”);
- filtering sequences according to specified criteria (module “Screening”);
- removing alignment columns based on specified criteria, for example, empty columns (module “Filter”);
- removal of sequences containing sequencing error (the module “Scrap”);
- detection of chimeras (module “Chim. detection), etc.

The diagram shows service modules of RapidMiner Studio, which are necessary to distribute the same type of information between modules (e.g., “Copy groups”). The necessity of introduction of such modules is a feature of Rapidminer Studio; it supposes that in a general case the modules make changes in the data under processing without copying it.

Each module receives file names as input and creates new file set as the result. The operation of the module depends on the parameters specified by the user via user interface of each module. The results of the script are sent to output ports and displayed by the Rapidminer Studio visualization subsystem in a convenient form to the user. The system supports presentation of a scenario as a new block with its input and output ports, as well as a cloud storage and execution of scripts, creating distributed computing environment. Rich feature set of Rapidminer Studio and various services provided by its developers were the main reason for choosing this system as a development environment for informational-computation resources of the project.

4. The database supporting metagenomic analyses

Assessment of world experience of organization of scientific research in the field of Data Science showed that the use of cloud technologies is a necessary basis for the interaction organization of the researchers. A specialized data storage should be a unit of the environment to ensure effective user access and computing processes to the data of research.

Database for microbiome-based metagenomic analysis data (Fig. 2.) provides the storage facility on all the stages of the microbiome studies from the probe sampling to the publication of the scientific meaningful results. The scheme in the Fig. 2 represents database structure as an ER-diagram. The scheme contains data about sampling, analysis of physicochemical and biological parameters of the probes, the sequencing results, the applied equipment and software, taxonomic databases, methods of the analysis of the collected material, publications of the obtained results and the participated researchers. It also allows us to store the processing scripts of analysis of metagenomic data, including software tools, commands, and configuration files. The latter function allows one to save the state of the computational process and restart it from the specified point.

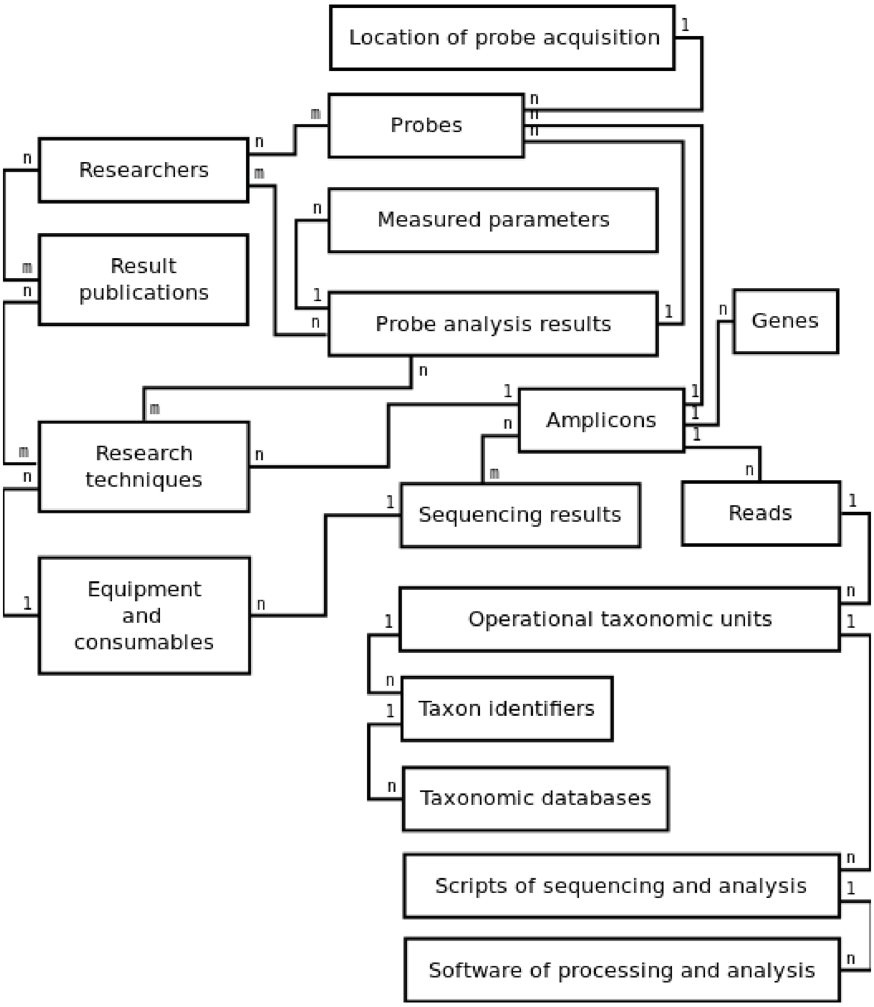

Fig. 2. A general schema representation of the database for microbiome research based on metagenomic analysis

The model is implemented by means of Django framework (`https://www.djangoproject.com`). The framework supports automatic definition of rational table structures representing many-to-many relations, and generation a customizable interface for the administrative panel, allowing testing the developed model. These same tools are used for the implementation of the project web site.

Cloud-based storage and dedicated storage of the metagenomic data will allow one to create online services for joint processing of sequencing data from different studies by specialized software and to publish information in the Internet. In order to achieve this goal, the following must be carried out:

- a software interface implementation for data access;
- filling in the database with information collected and processed as a result of studies of the microbiome of lake Baikal in 2009-2015;
- realization of the scenario design and execution to support the metagenomic data analysis in a distributed computing environment.

5. Conclusion

Modern problems of development of a distributed software environment for the implementation of organizational, informational and computational resources for scientific microbiological studies based on metagenome analysis are presented in the article. A generalized domain model of system-level is conducted, as well as the requirements are stated to the development environment and problems to be solved. A computational model of the process of analysis of the amplicons is being constructed and implemented. Aspect of informational supply of the computational process is represented by realization of the problem of cloud storage for computing processes (scenarios of metagenomic data processing), as well as by construction of a database for storing input and intermediate data, results of the scenarios execution. The database is used as a basis of an information portal construction for processing metagenomic data and presenting the results of scientific community.

6. Acknowledgments

The results obtained with the support of the following projects:

- Irkutsk scientific center of SB RAS No 4.1.2;

The results obtained with the use of the network infrastructure of Telecommunication center of collective use "Integrated information-computational network of Irkutsk scientific-educational complex" (http://net.icc.ru).

References

1. Brown, C. T., Hug, L. A., Thomas, C. B. *et al*. Unusual biology across the group comprising more than 15% of domain Bacteria. Nature. — 2015. — Vol. 523. — P. 208-211.1.
2. Hug, L. A., Brett, J. B, Anantharaman K. *et al*. A new view of the Tree of Life. Nature Microbiology. — 2016. — Vol. 1. — P. 16048.
3. Quast, C., Pruesse, E., Yilmaz, P. *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tool. Nucleic Acids Research. — 2013. — Vol. 41. — P. 590-596.
4. Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A. *et al*. Uncovering earth`s virome. Nature. – 2016. – Vol. 536. – P. 425–430.
5. Hanson, N.W., Konwar, K.M., Wu, S.J., Hallam, S.J. Introduction to the analysis of environmental sequence information using metapathways. Comp. Meth. Next Gener. Sequenc. Data Analysis. – 2016. – P. 25–56.
6. Iverson, V., Morris, P. M., Frazar, C. D. Untangling genomes from metagenomes: Revealing an uncultured class of marine Euryarchaeota. Science. — 2012. — Vol. 335. — P. 587-590.
7. Bashenkhaeva, M. V., Zakharova, Y. R. Petrova, D. P. Sub-Ice microalgal and bacterial communitie in freshwater Lake Baikal, Russia. Environmental Microbiology. — Vol. 70, No. 3. — P. 751-765.
8. Mikhailov, S, Zakharov, Y. R., Galachyants, Y. P. *et al*. On the uniformity in taxonomic composition of bacterial communities in the photic layer of the three basins of lake Baikal, which differ in composition and abundance of spring phytoplankton. Reports of Academy of Sciences. — 2015. — Vol. 465, No. 5. — P. 620-626.

9. Gladkikh, A. C., Kalyuzhnaya, O. V., Belykh, O. I., Ahn, T. S., Parfenova, V. V. Analysis of the bacterial community of two endemic sponges from lake Baikal. Microbiology. — 2014. — Vol. 83, No. 6. — P. 682-693.
10. Zemskaya, T. I, Lomakina, A. V., Mamaeva, E. V. *et al*. Bacterial communities in sediments of Lake Baikal from areas with oil and gas discharge. Aquatic Microbial Ecology. — 2015. — Vol. 75. — P. 95-109.
11. Bukin, S. V., Pavlova, O. N., Manakov, A. Y. *et al*. The ability of microbial community of Lake Baikal bottom sediments associated with gas discharge to carry out the transformation of organic matter under thermobaric conditions. Frontiers in microbiology. — 2016. — Vol. 7. — P. 690.
12. Schloss, P. D. et al. Introducing Mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology. — 2009. — Vol. 75 (No. 23). — P. 7537-7541.
13. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. — 2010. — Vol. 26, No. 19. — P. 2460-2461.
14. Johnston, W.M., Hanna, J.R.P., Millar, R.J. Advances in Dataflow Programming Languages. ACM Computing Surveys. – 2004. – Vol. 36. – P. 1–34.