ASAR Database: An R Tool for Visual Analysis and Storage of Metagenomes

Askarbek Orakov[1,2], Nazgul Sakenova[1,2], Igor Goryanin[1,4,5] and Anatoly Sorokin[3,6]
[1]*Okinawa Institute of Science and Technology, Onna-son, Japan*
[2]*School of Science and Technology, Nazarbayev University, Astana, Kazakhstan*
[3]*Institute of Cell Biophysics RAS, Pushchino, Russia*
[4]*University of Edinburgh, School of Informatics, Edinburgh, U.K.*
[5]*Tianjin Institute of Industrial Biotechnology, Biodesign Centre, Tianjin, China*
[6]*Moscow Institute of Physics and Technology, Dolgoprudny, Russia*

Keywords: Metagenomics, Interactive Data Analysis, Functional Analysis, KEGG Pathways, Taxonomic Analysis.

Abstract: The functional and taxonomic analysis is the critical step in understanding the interspecies interaction within the microbial communities. Currently, these types of analysis are run independently, which makes interpretation of the results hard and error-prone. Here we present ASAR (Advanced metagenomic Sequence Analysis in R) Database, the interactive tool and the databases for storage and exploratory analysis of the metagenomic sequencing data along three dimensions: taxonomy, function, and environmental conditions.

1 INTRODUCTION

It is known that 99% of prokaryotic species are not culturable (Schloss and Handelsman, 2005) either at all or culture conditions are not known. In that circumstances, the metagenomic analysis becomes an essential experimental technique for our understanding of composition and functional properties of microbial communities. In addition to that, decreasing the cost of sequencing and increasing throughput of sequencing machinery cause rapid growth in availability of the metagenomic data, which makes the development of tools for functional, taxonomic and metabolic analyses of metagenomes extremely important (Hugenholtz and Tyson, 2008; Lindgreen et al., 2016). Recently the whole genome sequencing (WGS) become more and more popular in comparison with 16S, Ribosomal Intergenic Spacer Analysis (RISA), which compares the sizes of the intergenic region between the 16 S rRNA (rrs) and 23 S rDNA (rrl) genes, and other amplicon sequencing techniques as it not only provides information about the taxonomical composition of the biome but highlight its functional abilities via mapping DNA reads on to protein function database. However, even most promising current metagenomic analysis tools usually provide either only taxonomic (Menzel et al., 2016) or just functional (Westbrook et al., 2017) analysis. Some tools implement both types of analysis but independently (Keegan et al., 2016). That renders data analysis incomplete and leaves a lot of information contained in the metagenomic datasets undiscovered. Recently we have developed the ASAR (Advanced metagenomic Sequence Analysis in R) application (Orakov et al., 2017) to fill that gap.

Simultaneous analysis of taxonomic and functional annotations at the reading level could help answer many important questions, such as, which taxonomic group in a sample is the main contributor to a particular function or metabolic pathway. Moreover, ability to analyze changes in microbiomes in the context of the metabolic network is the critical requirements for understanding biochemical processes in the community and the presence of competition or symbiosis between species. Discovering the most important metabolic pathways would also considerably improve the understanding of microbial community evolution. The core advantage of ASAR is the ability to perform taxonomic and functional analyses simultaneously, by interactive subsetting and aggregating abundance data at various levels of taxonomical and functional hierarchy. It is designed to let researchers drill down towards the most meaningful view of their data in a convenient way. It is also possible to perform the compa-

Orakov, A., Sakenova, N., Goryanin, I. and Sorokin, A.
ASAR Database: An R Tool for Visual Analysis and Storage of Metagenomes.
DOI: 10.5220/0006722801960200
In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 3: BIOINFORMATICS*, pages 196-200
ISBN: 978-989-758-280-6

rative analysis of the KEGG metabolic pathways (Kanehisa et al., 2016), by exploring the pathways enrichment and visualizing the pathways themselves.

Original ASAR application was designed to deal with data in memory. It is not uncommon in metagenomics to have the repetitive collection of samples as a time series. In this case, application sometimes has to deal with datasets of hundreds of samples, which do not fit into the memory of regular workstation. To handle large datasets, we augmented the ASAR application with the database to store raw data and perform the aggregation and selection.

2 METHODS

The application was written in R programming language (R Core Team, 2017) and Shiny platform (Chang et al., 2017) was used to make it web-based and user-interactive. Thanks to R and Shiny, the application can both be used locally at machines with installed R and as web-service. MonetDB (https://www.monetdb.org/) was used as DBMS and MonetDB.R (Muehleisen et al., 2017) package was used for connection between R application and DBMS. The application requires following R packages: dplyr (Wickham et al., 2017) and data.table (Dowle and Srinivasan, 2017) for efficient data manipulation; ggplot2 (Wickham, 2016), gplots (Warnes et al., 2016), RColorBrewer (Neuwirth, 2014) and d3heatmap (Cheng and Galili, 2016) for visualization; pathview (Luo et al., 2013) and png (Urbanek, 2013) for exporting the results.

Two datasets were used for development of the application. The small dataset contains 11 metagenomes (total size 45 GB) from swine waste microbial fuel cell (MFC) performance analysis project (Khilyas et al., 2017). The moderate dataset consists of 172 metagenomes (total size 195 GB) from longitudinal monitoring of the MFC wastewater treatment of Spent Wash (Dimou et al., 2014). Both datasets were loaded into the database separately. The small dataset was used for the performance comparison with the in-memory application. The moderate dataset does not fit into memory, so it was used for demonstration of the performance of the database version of the app.

3 RESULTS

3.1 The WGS Data

Sequencing data usually comes as a set of short DNA reads, which are mapped to genomic and functional databases for annotation by tools like Kaiju (Menzel et al., 2016), Paladin (Westbrook et al., 2017), and MG-RAST (Keegan et al., 2016). After joining of taxonomical and functional annotation, the data form 2D matrix with species in rows and functions in columns. In that matrix, each cell contains the abundances of reads mapped to the particular function in particular species. Analysis of single metagenome is quite rare, usually, metagenomes obtained at several sets of environmental conditions, time points and perturbations are analyzed. That set of samples forms the third dimension of the dataset.

The analysis of multidimensional datasets is a tricky task; this is one reason why people usually analyze taxonomy and functional data separately: aggregation along functional or taxonomic dimension forms the 2D matrix from the data, which is more straightforward for visualization and interpretation. The similar type of task was solved in business analytics in the middle of 80s by development concept of the data cube (Kimball and Ross, 2011). In our case, the data cube is the 3D array with taxonomy, function, and metagenome as dimensions and read counts as cell content. Elements of two of dimensions form hierarchies: taxonomic and functional. The components of metagenome dimension usually organized into kind of design matrix either explicitly by planning experiment upfront, or implicitly by exploring the spatial and temporal variability of a microbial community under investigation.

We designed ASAR (Orakov et al., 2017) application for interactive analysis of the whole dataset by application aggregation and selection operations dynamically and exploration of the obtained 2D matrices visually. At the moment we are using the annotation files generated by MG-RAST pipeline, but any other annotation pipeline, which assigns annotation at the DNA read level, such as Kaiju, Paladin, QIIME, etc., would give similar results. The MG-RAST was chosen because its annotation is based on the common database and so self-consistent. The procedure of import other types of data is the same, but mismatches caused by use different references during DNA read annotation won't be fixed.

3.1.1 Selection and Aggregation

Interactive application of Selection and Aggregation is the essential steps of dynamic exploration of the 3D data cube. Selection operation allows the user to navigate through hierarchy by selecting element at some higher level of the tree and analyze the subset of the cube underneath part chosen. For example, the user can choose *Deltaproteobacteria* at the class level of taxonomy and restrict consideration to species and

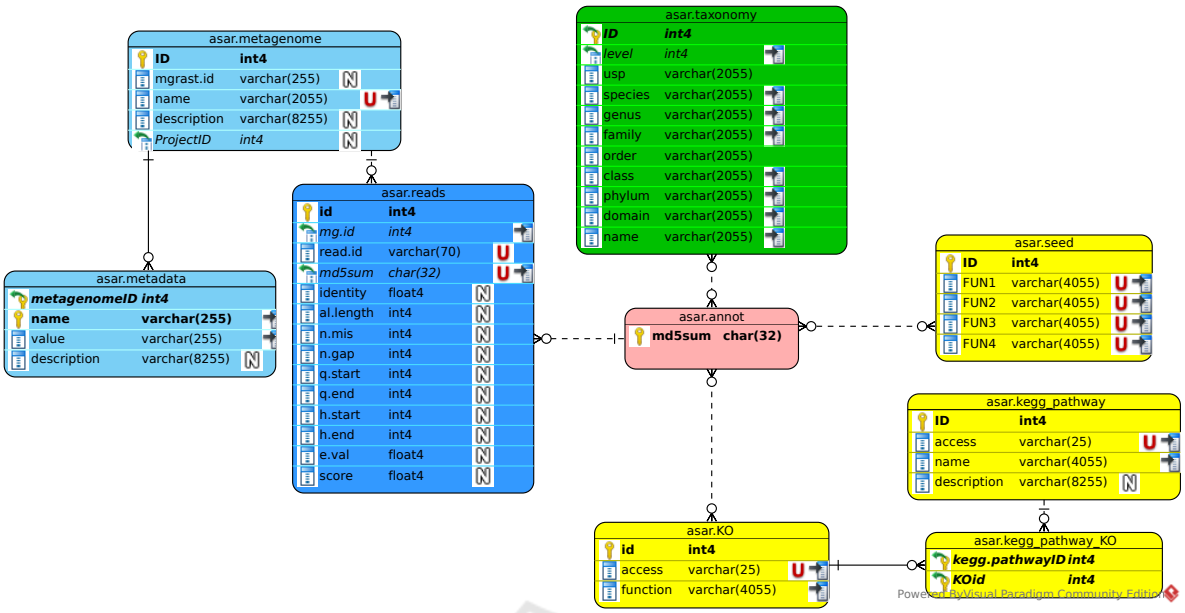

Figure 1: The database structure diagram. The fact table is shown in blue, functional annotation shown in yellow, taxonomic annotation shown in green. 'U' icon indicates the columns with the unique constraint. 'N' icon indicates the nullable columns. Columns included in the primary key are marked by the key symbol.

functions in that class only. Aggregation operation allows the user to summarize the data at some level of the hierarchy, by summing up the content of cells corresponding to the same element at selected Aggregation level. For example, the reliability of data at strain level is usually low, so it is common to *Aggregate* the data up to the genus level.

The application of Selection and Aggregation to the metagenome axis, by choosing the particular field in the metadata and use it as a hierarchy level for selection and aggregation. For example, analysis of microbial fuel cell (MFC) microbiome usually consider anodic biofilm separately from the planktonic community, so we can choose "Part of MFC" field from metadata and Select "Anode" value to study the composition of anodic biofilm only. We can also aggregate all planktonic communities into one matrix and explore their dependence on time or initial community composition.

3.1.2 SEED Annotation Analysis

Shiny Application has five tabs, four of which are heatmaps and last is the KEGG pathway diagram (Figure 2).

First three heatmaps are three different projections of the 3D dataset: function vs. taxonomy (F/T), function vs. metagenome (F/M) and taxonomy vs. metagenome (T/M). So, in the first heatmap, you can see the abundance of each intersection between function and taxon in a single metagenome sample. The next two heatmaps represent traditional functional and taxonomic analysis and allow to compare enrichments of functions or taxons among selected metagenomes. For both functions and taxons user can choose particular level and value to work with and the level at which all data will be aggregated. The taxonomic hierarchy levels are taken from MG-RAST (Keegan et al., 2016) and SEED hierarchy (Overbeek et al., 2013) is used for functions.

3.1.3 KEGG Annotation Analysis

In the fourth heatmap one can compare KEGG pathway (Kanehisa et al., 2016) enrichments in order of the descending value of standard deviation among selected samples in a selected taxon. After one finds the pathway of interest, choosing the pathway name in the last tab will draw its KEGG diagram. In the diagram, every enzyme will correspond to a rectangle where samples are colored according to values of their contribution to the abundance of that enzyme in the community.

3.2 Database Structure

The structure of the database (Figure 1) follows standard Online analytical processing (OLAP) Snowflake pattern (Ponniah, 2010) with *asar.reads* as the fact table. The icons on the diagram follow The *metagenome* and *annot* tables define two main dimensions.

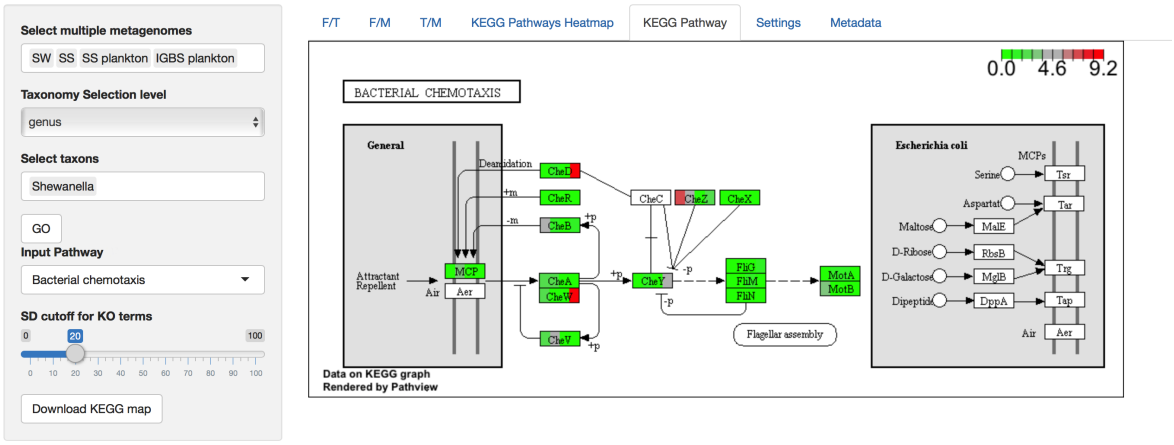

Figure 2: The KEGG diagram visualization. The selection panel is on the left. It is possible to choose metagenomes of interest, taxonomy selection level and value and the pathway to show. The "GO" button prevents unintended drawing of the pathway, which requires access to the KEGG database and is time-consuming.

The former one for non-hierarchical sample dimension, the latter for both hierarchical taxonomic (*taxonomy* table) and functional (*function* table) dimensions. The *ko* table, another connection of the *annot* table, provides KEGG annotation for pathway analysis.

All selection and aggregation operations along taxonomy and function dimensions of the data cube are performed in memory in the same way as in original ASAR application, while for selection and aggregation operations along metagenomic dimension the SQL queries to the database is used. That way of interaction with the database was chosen to reduce the response time of the application, as the manipulation along the functional and taxonomic dimensions are much more often compare to modification of selection and aggregation criteria along samples dimension.

The taxonomic hierarchy levels are taken from MG-RAST (Keegan et al., 2016). It consists of eight levels from domain to strain levels. The read could be assigned to any level of the taxonomy, so "least common ancestor" annotation method could be used. The structure of functional annotation follows the SEED hierarchy (Overbeek et al., 2013) that consists of four levels. Level 1 of the SEED hierarchy corresponds to individual enzyme functions, while major functional groups like "DNA metabolism" or "Virulence" form the level 4 of the tree. Specific kind of annotation is KEGG orthology, which is required for mapping of metagenomic data onto the KEGG pathway.

4 CONCLUSIONS

Our post-annotation analysis and visualization tool uses data integration algorithm to merge taxonomic and functional data annotated at the DNA read level. The resulting 3D dataset with axes of Functions, Taxonomy and Metagenome samples is visualized via three heatmaps of each axis versus two others (F/T, F/M, T/M). Additionally, KEGG pathway enrichment sorting/heatmap and its map visualization are implemented.

We have tested the performance of the database on Intel Core i5, 32 GB RAM workstation with small (11 metagenomes, total size 45 GB) and moderate (172 metagenomes, total size 195 GB) datasets. The average response time was in a range of 10 sec for in-memory data transformation and up to 2 min for DB SQL query. The database upload time was 5 minutes for the small dataset and 10 minutes for the moderate one. The source code of ASAR is free and accessible at GitHub (https://github.com/Askarbekorakov/ASAR).

ACKNOWLEDGEMENTS

Members of OIST BSU UNIT, Irina Khilyas for providing data, OIST NGS Section for sequencing services. Dr. Jeannette Kunz for supporting internship.
This work has been supported by the OIST funding.

REFERENCES

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2017). *shiny: Web Application Framework for R.*

Cheng, J. and Galili, T. (2016). *d3heatmap: Interactive Heat Maps Using 'htmlwidgets' and 'D3.js'.* R package version 0.6.1.1.

Dimou, O., Andresen, J., Feodorovich, V., Goryanin, I., Harper, A., and Simpson, D. (2014). Optimisation of scale-up of microbial fuel cell for sustainable wastewater treatment with positive net energy generation. *New Biotechnology*, 31:S213.

Dowle, M. and Srinivasan, A. (2017). *data.table: Extension of 'data.frame'*. R package version 1.10.4-3.

Hugenholtz, P. and Tyson, G. W. (2008). Microbiology: metagenomics. *Nature*, 455(7212):481–483.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–62.

Keegan, K. P., Glass, E. M., and Meyer, F. (2016). MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. *Methods in Molecular Biology (Clifton, N.J.)*, 1399:207–233.

Khilyas, I. V., Sorokin, A., Kiseleva, L., Simpson, D. J. W., Fedorovich, V., Sharipova, M. R., Kainuma, M., Cohen, M. F., and Goryanin, I. (2017). Comparative Metagenomic Analysis of Electrogenic Microbial Communities in Differentially Inoculated Swine Wastewater-Fed Microbial Fuel Cells. *Scientifica*, 2017(5-6):1–10.

Kimball, R. and Ross, M. (2011). *The Data Warehouse Toolkit*. The Complete Guide to Dimensional Modeling. John Wiley & Sons.

Lindgreen, S., Adair, K. L., and Gardner, P. P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, 6(1):19233.

Luo, Weijun, Brouwer, and Cory (2013). Pathview: an r/bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1831.

Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7:11257.

Muehleisen, H., Damico, A., Raasveldt, M., Lumley, T., and Team, M. D. (2017). *MonetDBLite: In-Process Version of MonetDB for R*. MonetDB. R package version 0.4.1.

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2.

Orakov, A., Sakenova, N., Sorokin, A., and Goryanin, I. (2017). *ASAR: visual analysis of metagenomes in R*. OIST, Okinawa, Japan.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., and Stevens, R. (2013). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, 42(D1):D206–D214.

Ponniah, P. (2010). *Olap in the Datawarehouse*, pages 373–406. John Wiley & Sons, Inc.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schloss, P. D. and Handelsman, J. (2005). Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biology*, 6(8):229.

Urbanek, S. (2013). *png: Read and write PNG images*. R package version 0.1-7.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. (2016). *gplots: Various R Programming Tools for Plotting Data*. R package version 3.0.1.

Westbrook, A., Ramsdell, J., Schuelke, T., Normington, L., Bergeron, R. D., Thomas, W. K., and MacManes, M. D. (2017). PALADIN: protein alignment for functional profiling whole metagenome shotgun data. *Bioinformatics (Oxford, England)*, 33(10):1473–1478.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., Francois, R., Henry, L., and Müller, K. (2017). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.4.