Physiology International 107 (2020) 1, 12–17
DOI: 10.1556/2060.2020.00004

AKADÉMIAI KIADÓ

Taxamat: Automated biodiversity data management tool – Implications for microbiome studies

A. VIDA[1,2†], B.L. BODROGI[3], B. BALOGH[1] and P. BAI[1,2,4*]

[1] Department of Medical Chemistry, Faculty of Medicine, University of Debrecen, Debrecen, Hungary

[2] MTA-DE Lendület Laboratory of Cellular Metabolism, University of Debrecen, Debrecen, Hungary

[3] Department of Urology, Borsod-Abaúj-Zemplén County Central and University Teaching Hospital, Miskolc, Hungary

[4] Research Center for Molecular Medicine, Faculty of Medicine, University of Debrecen, Debrecen, Hungary

Received: December 27, 2019 • Accepted: February 25, 2020
Published online: April 14, 2020

ABSTRACT

Working with biodiversity data is a computationally intensive process. Numerous applications and services provide options to deal with sequencing and taxonomy data. Professional statistics software are also available to analyze these type of data. However, in-between the two processes there is a huge need to curate biodiversity sample files. Curation involves creating summed abundance values for chosen taxonomy ranks, excluding certain taxa from analysis, and finally merging and downsampling data files. Very few tools, if any, offer a solution to this problem, thus we present Taxamat, a simple data management application that allows for curation of biodiversity data files before they can be imported to other statistics software. Taxamat is a downloadable application for automated curation of biodiversity data featuring taxonomic classification, taxon filtering, sample merging, and downsampling. Input and output files are compatible with most widely used programs. Taxamat is available on the web at http://www.taxamat.com either as a single executable or as an installable package for Microsoft Windows platforms.

* Corresponding author. Department of Medical Chemistry, University of Debrecen, 4032 Debrecen, Egyetem tér 1, Hungary. Tel.: +36 52 412 345; Fax: +36 52 412 566, E-mail: baip@med.unideb.hu.

†Present address: Soft Flow Hungary LTD., Pécs, Hungary.

KEYWORDS

biodiversity, data management, taxonomy, diversity indices

INTRODUCTION

The widespread availability of next-generation sequencing has led to an incredible growth in the number of biodiversity studies. This boom is most prominent in the analysis of microbiome samples originating mainly from soil, water and the commensal flora of humans and animals.

When it comes to a sample analysis using next-generation sequencing techniques, the workflow goes through the steps of sample collection, DNA isolation, amplification, next-generation sequencing, sequence analysis, construction of a taxon list, curation of the created list, comparing samples and finally creating statistics. There are several online and standalone services and software for sequence analysis and the taxon list construction steps, and similarly many statistics software are available for sample comparison and statistical analysis. However, there are very few if any tools available that provide a simple solution for data curation, such as taxon filtering, sample data merging, and downsampling. These steps are often needed, for example, to exclude host- and food-related sequence data from microbiome samples and to compensate for oversampling when comparing multiple results [13]. Especially researchers with limited possibilities to create their own analysis scripts have difficulties preparing their data for statistical analysis. Taxamat is a simple tool that allows for the management of biodiversity data by automating high-rank taxon filtering, sample file merging and downsampling of oversampled files.

MATERIAL AND METHODS

Input files and data format

Taxamat performs taxonomic classification of an input taxon list, excludes certain main taxa if needed, and sums abundance values of main taxonomy ranks ranging from species up to superkingdom level. In addition, Taxamat also supports automated sample data file merging into one single data table. Merged files (or initial sample files) can be downsampled by random removal of organisms, in order to reduce sampling differences of input samples and to avoid or complement rarefaction [3].

For taxonomic classification, Taxamat requires the user to download database files from the NCBI website (ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/). The required files are packed into the taxdmp.zip archive. From this archive two files are required to be imported into Taxamat: “names.dmp” and “nodes.dmp”. These files are only needed to perform taxonomic hierarchy classification; other functions do not require these additional files.

For sample input file(s), Taxamat requires one or more two-column (hierarchy creation, taxon filtering, and downsampling) or multi-column (downsampling) tab-delimited text file(s) containing diversity data. For merging sample data columns, the tab-delimited text file used as input must not have a header, and the first column should have taxon names while the second column should have positive integers corresponding to abundance values. For downsampling, the two- or multi-column data files must have a header for each column, and the first column

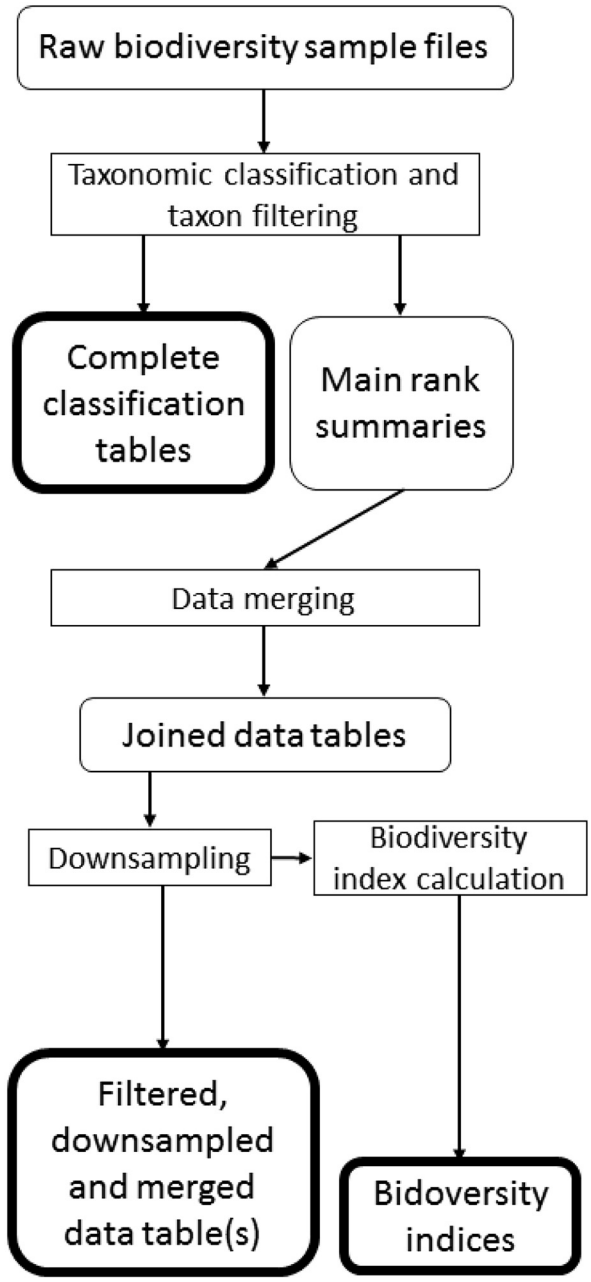

Figure 1. Data processing flow-chart. Flow chart of the data management process. A complete taxonomic categorization and optional filtering is performed on the input data. Main rank summary files can be merged in order to obtain a joint data table. If necessary, merged data can be further processed and downsampled to obtain tables with matching abundance values for further analysis

must contain taxon names while the following columns must have integers corresponding to abundance values for each taxon.

Output files and data format

Taxamat creates several output files. During hierarchy creation, a new tab-delimited file is created for each selected input file containing the following columns: 1. Taxon ID, 2. Query name, 3. Scientific name, 4. Query rank, 5. Species, 6. Genus, 7. Family, 8. Order, 9. Class, 10. Phylum, 11. Kingdom, 12. Superkingdom, 13. Abundance. These output files will have a "_H" suffix attached to indicate the hierarchy creation process. If the appropriate options are selected, then extra files will be created containing the summed value for each taxonomy rank as a two-column tab-delimited text file ("_H_<rank>" suffix). Based on the classification, some main taxa can be filtered out. If taxa are filtered out, then additional files are created with the selected categories removed from the sample. These additional files will have a "_f" suffix applied to indicate filtering. These output files can then be used for data merging, which will output a multi-column tab-delimited text file ("_merged" suffix) with the input file names as headers for each column. This merged file, in turn, can be used for downsampling either to pre-set total abundance values or to the lowest abundance levels ("_downsampled" suffix). Downsampling is an algorithm that iterates through the imported sample file and removes a random organism from the sample at each step until the total number of organisms matches a pre-defined value or the lowest total organism count of a multi-sample table. This step is necessary to compare samples with different sequencing/sampling depth without the need for rarefaction [3].

If the appropriate checkbox is checked, then the most often used biodiversity indices, the Shannon index and the Simpson index (known as Hunter-Gaston index in microbiology) will be calculated as well, and written into a separate file with the "_statistics" suffix attached [10, 11]. The output files are compatible with most widely used statistical software (including Cluster 3.0, an excellent clustering application) and MS Excel [1]. Fig. 1 shows an overview of the analysis process and file formats.

RESULTS AND DISCUSSION

Taxamat, presented in this article provides an easy access option for researchers with limited possibilities to create their own scripts to manage large biodiversity data sets. Taxamat utilizes the up-to date taxonomy database from NCBI while also allows to work with older versions in case of revisiting earlier work. Taxamat guides the user through the steps of a taxonomic classification with the option of filtering for certain taxa. This results in a set of classification tables and summary tables. These intermediary files can subsequently be merged to allow for joint downsampling of the acquired data. This step allows for an unbiased analysis. During this process simple biodiversity indices are calculated and the created files are ready to use with further statistical and biodiversity analysis software.

The taxonomy database files from NCBI are in a tab-delimited text format. As our aim was to develop an easily accessible software, Taxamat uses these files in their native format to avoid the extra burden of database conversion. This, however, results in slow hierarchy creation, as parsing through huge text files is time-consuming. Even at its current state, however, the time requirement

of this process is negligible when compared to other steps of acquiring and evaluating biodiversity data. We still aim to optimize this process for a better user experience in the future.

We advise any users to use the taxon filtering and downsampling options with great care and discretion, as these processes (especially filtering) might alter sample results if the excluded taxa are over-represented in a sample.

For a complete description together with a step-by-step guide and test sample files, the authors invite you to visit http://www.taxamat.com. An example of the analysis of a metagenome sequencing dataset using Taxamat can be found in [13].

The tool we present here has wide range of uses. It can be used in ecology to decipher and characterize complex samples or habitats (e.g. benthos) containing microscopical plants and animals. Another, large field of application is microbiome studies. Pathological changes to the microbiome were described in aging [7, 12], metabolic diseases [4], psychiatric diseases [5] or neoplastic diseases [2, 6, 9]. The general feature of bacterial dysbiosis is the reduction of the complexity of the microbiome [8]. Taxamat offers a comprehensive characterization of microbiome complexity, a unique feature of this software.

DATA AVAILABILITY

The working software can be found at http://www.taxamat.com and the user manual is uploaded as a Supplementary file.

Conflict of interest: The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Authors are grateful to Dr. Karen Uray (Department of Medical Chemistry, University of Debrecen) for improving the use of English in the manuscript.
We were supported by grants from the Hungarian National Research, Development and Innovation Office grants (K123975, GINOP-2.3.2-15-2016-00006), the Momentum fellowship, and the NKM-26/2019 grant of the Hungarian Academy of Sciences and the University of Debrecen and Campus France. The research was financed by the Higher Education Institutional Excellence Programme (NKFIH-1150-6/2019) of the Ministry of Innovation and Technology in Hungary, within the framework of the Biotechnology thematic programme of the University of Debrecen.

REFERENCES

1. de Hoon MJ, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics 2004; 20: 1453–4.
2. Goedert JJ, Jones G, Hua X, Xu X, Yu G, Flores R, et al. Investigation of the association between the fecal microbiota and breast cancer in postmenopausal women: a population-based case-control pilot study. J Natl Cancer Inst 2015; 107: djv147.

3. Gotelli JG, Colwell RK. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol Lett 2001; 4: 379–91.
4. Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc Natl Acad Sci U S A 2012; 109: 594–99.
5. Hsiao EY, McBride SW, Hsien S, Sharon G, Hyde ER, McCue T, et al. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell 2013; 155: 1451–63.
6. Kovács T, Mikó E, Vida A, Sebő É, Toth J, Csonka T, et al. Cadaverine, a metabolite of the microbiome, reduces breast cancer aggressiveness through trace amino acid receptors. Sci Rep 2019; 9: 1300.
7. Lim MY, Song EJ, Kang KS, Nam YD. Age-related compositional and functional changes in micro-pig gut microbiome. Geroscience 2019; 41: 935–44.
8. Maffei VJ, Kim S, Blanchard Et, Luo M, Jazwinski SM, Taylor CM, et al. Biological Aging and the Human Gut Microbiota. J Gerontol A Biol Sci Med Sci 2017; 72: 1474–82.
9. Miko E, Vida A, Kovacs T, Ujlaki G, Trencsenyi G, Marton J, et al. Lithocholic acid, a bacterial metabolite reduces breast cancer cell proliferation and aggressiveness. Biochim Biophys Acta 2018; 1859: 958–74.
10. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948; 27: 379–423 623–656.
11. Simpson EH. Measurement of diversity. Nature 1949; 163: 688.
12. Singh H, Torralba MG, Moncera KJ, DiLello L, Petrini J, Nelson KE, et al. Gastro-intestinal and oral microbiome signatures associated with healthy aging. Geroscience 2019; 41: 907–21.
13. Vida A, Kardos G, Kovacs T, Bodrogi BL, Bai P. Deletion of poly(ADPribose) polymerase-1 changes the composition of the microbiome in the gut. Mol Med Rep 2018; 18: 4335–41.