Bipartite Graphs for Visualization Analysis of Microbiome Data

Karel Sedlar[1], Petra Videnska[2], Helena Skutkova[1], Ivan Rychlik[3] and Ivo Provaznik[1]

[1]Department of Biomedical Engineering, Brno University of Technology, Brno, Czech Republic. [2]Research Centre for Toxic Compounds in the Environment RECETOX, Masaryk University, Brno, Czech Republic. [3]Veterinary Research Institute, Brno, Czech Republic.

Supplementary Issue: Bioinformatics Methods and Applications for Big Metagenomics Data

ABSTRACT: Visualization analysis plays an important role in metagenomics research. Proper and clear visualization can help researchers get their first insights into data and by selecting different features, also revealing and highlighting hidden relationships and drawing conclusions. To prevent the resulting presentations from becoming chaotic, visualization techniques have to properly tackle the high dimensionality of microbiome data. Although a number of different methods based on dimensionality reduction, correlations, Venn diagrams, and network representations have already been published, there is still room for further improvement, especially in the techniques that allow visual comparison of several environments or developmental stages in one environment. In this article, we represent microbiome data by bipartite graphs, where one partition stands for taxa and the other stands for samples. We demonstrated that community detection is independent of taxonomical level. Moreover, focusing on higher taxonomical levels and the appropriate merging of samples greatly helps improving graph organization and makes our presentations clearer than other graph and network visualizations. Capturing labels in the vertices also brings the possibility of clearly comparing two or more microbial communities by showing their common and unique parts.

KEYWORDS: metagenomics, OTU table, 16S rRNA, bipartite graph, visualization analysis, graph modularity

SUPPLEMENT: Bioinformatics Methods and Applications for Big Metagenomics Data

CITATION: Sedlar et al. Bipartite Graphs for Visualization Analysis of Microbiome Data. *Evolutionary Bioinformatics* 2016:12(S1) 17–23 doi: 10.4137/EBO.S38546.

TYPE: Technical Advance

RECEIVED: February 02, 2016. **RESUBMITTED:** April 17, 2016. **ACCEPTED FOR PUBLICATION:** April 23, 2016.

ACADEMIC EDITOR: Jike Cui, Deputy Editor in Chief

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 3999 words, excluding any confidential comments to the academic editor.

FUNDING: Supported by BUT intern grant agency, interfaculty junior project FEKT/FIT-J-16-3335. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: sedlar@feec.vutbr.cz

Paper subject to independent expert blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

The rapid progress that DNA sequencing techniques have undergone in the last decade has changed the way in which metagenomics research is carried out.[1,2] Sequencing of 16S rRNA gene has become a relatively easy way to study microbial composition and diversity.[3] However, to process the amount of data, associated bioinformatics tools have to handle the resulting high dimensionality that is dependent on the number of operational taxonomic units (OTUs). From the beginning, one of the main goals of metagenomics research has been to describe and compare several samples. Unfortunately, standard metrics for dimensionality reduction techniques do not take into account the phylogenetic information that sequences contained. For this reason, UniFrac[4] metric based on phylogenetic analysis was proposed. Using phylogenic information was a great advantage, because even unidentified sequences, due to a lack of reference database, could be used for the comparison of samples by UniFrac and for visualization by following principal coordinate analysis (PCoA). On the other hand, the method had to tackle the increasing output of sequencing machines. Its array-based implementation, fast UniFrac,[5] provided orders of magnitude improvements, making UniFrac followed by PCoA a standard technique used for visualization in microbiome studies. However, dimensionality reduction causes new axes, which are principal coordinates, that do not correspond to specific OTUs, but rather to abstract OTUs. This can be disadvantageous in cases where common and unique taxa should be captured and compared for different environments because those particular taxa cannot be connected with the samples or environments directly, but can be done only indirectly using PCoA biplot. Moreover, the nature of UniFrac makes it heavily dependent on the appropriate use of databases and phylogenetic methods.

Although the progress in DNA sequencing brings the possibility to get a deep insight into a metagenome by shotgun sequencing, which is the current trend, cheaper amplicon sequencing-based technique remains the standard approach for investigating the diversity of microbial communities.[6,7] Thus, there is still need to develop algorithms for processing amplicon sequencing that can be used for data quantitation by clustering similar sequences together and by assigning taxonomy to the clusters using a reference database.[8] With the growing volume of current databases, it is possible to identify even uncultured microorganisms relatively reliably down to the genus level.[9] A number of various pipelines for processing data and identifying sequences are now available, eg, Megan,[10] mg-RAST,[11] QIIME,[12] etc. An OTU table containing an abundance of OTUs in different samples is presented as the output. Various visualization techniques can be used for the analysis of this table. Among them, one can use direct visualization by a cluster heat map, or a correlation analysis followed by a cluster heat map.[13] Although labels can be added and a range of clustering techniques can be utilized, the rectangular shape of the visualization may hide some small yet important clusters. Venn diagrams also suffer from a similar kind of rigid presentation, which makes them difficult to follow, especially for larger datasets. The abovementioned PCoA can not only be used for dimensionality reduction, but various different techniques can also be applied, for example principal component analysis, nonmetric multidimensional scaling, FastMap, or MetricMap.[14] Last but not least, a range of different network representations can provide visual analysis. Even though these visualizations are able to reveal microbial interactions, are usable for dynamic modeling of bacterial communities, or can process high-dimensional data,[15] there is still a lack of simple workflow that could suitably complement UniFrac PCoA visualization.

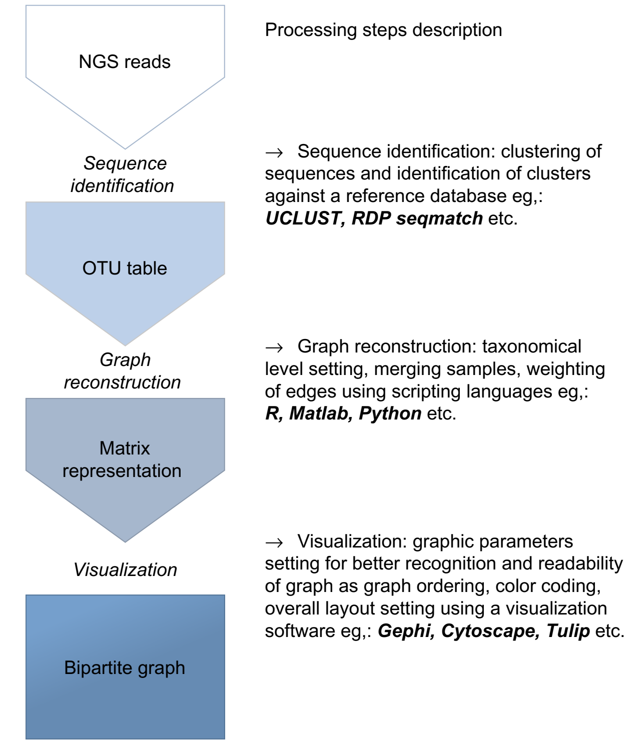
Currently, the increasing output of sequencing platforms is causing such a massive increase in the number of nodes in network representations that the resulting representations are becoming uninformative and are not human readable. The significant reduction in the price for sequencing is having another impact. It is cheaper to acquire new data by sequencing than it is to store and process them.[16] Moreover, this huge amount of sequences contains contaminations, so it is not necessary to use every sequence.[17] Rather, an overall view can be more desirable. In an amplicon-based approach, where the reference database is limited to the target gene, even the BLAST[18] identification for larger datasets is possible within a reasonable time, and a reduction of data can be achieved by focusing on the higher taxonomical levels. Herein, we present the bipartite graph visualization that was inspired by the network analysis implemented in QIIME, which, unfortunately, lacks clear description. However, we show that several steps for OTU table preprocessing are required to meet the current needs and provide a clear and informative visualization. We also demonstrate important graph features to be independent of taxonomical level, making community detection also available for highly reduced data. Based on our previous outline of the approach in comparison of microbial samples,[19,20] we now provide a deeper analysis showing the patterns that can be revealed by giving different weights to partitions, OTUs, and samples. These can provide additional information to UniFrac-PCoA analysis, making this approach an advantageous complement to the standard technique.

Materials and Methods

Test dataset. For the verification and presentation of our approach, we utilized data regarding the microbiota composition in the cecum of egg-laying hens during their whole lives. The three experiments are described and data are published.[21] Shortly, in the first experiment, long-term on-farm development of cecal microbiota was described. The three chickens or hens were taken from the flock, sacrificed at weeks 1, 2, 3, 4, 8, 12, 16, 19, 22, 26, 34, 38, 45, 51, 55, and 60. In the second experiment, short-term development of cecal microbiota was observed in newly hatched chickens. Three chicks were sacrificed on days 4, 7, 10, 13, 16, and 19. The third experiment verified the long-term experiment. Three chickens or hens were taken from the flock, sacrificed at weeks 3, 7, 16, 28, 40, and 52. Cecal contents were collected from all sacrificed birds and were frozen at −20 °C. The isolated DNA served as a template for the preparation of 16S rDNA library, which was sequenced by 454 Junior (Roche).[21] UniFrac analysis followed by PCoA was used for visual presentation and dimensionality reduction. The data were classified into four clusters by Ward's hierarchical clustering using Mahalanobis distance. Comparing the clusters to area flowcharts representing abundance of different taxa together with biological considerations, four main stages of microbiota development represented by four clusters were determined. This made the dataset an appropriate reference for presenting our workflow.

Overall workflow. Sequence reads were clustered at the 97% similarity level using UCLUST[22] and identified with RDP Seqmatch.[23] The identified clusters of sequences represented OTUs and a number of sequences in the cluster stands for abundance of an OTU in particular samples. Besides the abovementioned techniques, other methods for OTU picking are applicable in this step. Matlab 2014a and R[24] were used for OTU table preprocessing and for the construction of bipartite graphs, but any suitable scripting language can be used. The final graphs were visualized with Gephi[25] using the force-directed layout ForceAtlas2.[26] Different stages of microbiota composition were determined by community detection based on modularity optimization[27] using different resolutions according to the graph size.[28] Also this step can be implemented in any suitable graph visualization software. The overall commented workflow is shown in Figure 1.

Data transformation. An OTU table, which is $P\ m \times n$ matrix, was obtained as a result of sequence identification. Each of the m rows represents different OTUs; therefore, the precise number of rows depends on taxonomical level. While the number of rows is more when OTU stands for species, it can be reduced when OTU stands for phylum. Each of the n columns represents different samples; therefore, the precise number of columns depends on a number of samples and can be reduced by sample merging. The numbers in the table show the abundances of particular OTUs in the different samples. Due to dissimilar sequencing depth for each sample (from 292 to 47,657 sequences in the source dataset), an OTU table representing relative abundances was used. We reconstructed six matrices where number of rows was reduced by focusing on higher taxonomical levels, see Table 1. For every matrix



→ Sequence identification: clustering of sequences and identification of clusters against a reference database eg,: ***UCLUST, RDP seqmatch*** etc.

→ Graph reconstruction: taxonomical level setting, merging samples, weighting of edges using scripting languages eg,: ***R, Matlab, Python*** etc.

→ Visualization: graphic parameters setting for better recognition and readability of graph as graph ordering, color coding, overall layout setting using a visualization software eg,: ***Gephi, Cytoscape, Tulip*** etc.

Figure 1. Flowchart describing proposed workflow. Every main step is implementable in several scripting languages/software according to the preferences of users.

from another set of five matrices, the number of columns was reduced by merging the samples. Although in Ref. 18 we used merging the samples according to the environment, here we performed a different strategy based on a fusion of samples from the detected communities. This approach better represents the original pattern and allows combining data from different experiments and the handling of uneven sampling in time. Additional reduction was done in row for five matrices by omitting low-abundant OTUs (Table 2). Due to the non-normal distribution in different samples tested by Shapiro–Wilk test, we used merging of samples based on median in way that the row of selected columns (samples) was replaced by its median value.[18]

The preprocessed OTU table was easy to use for graph reconstruction. The values of m and n represent the sizes of partitions, with m being the number of taxa and n being the number of samples or communities. A Boolean biadjacency B $m \times n$ matrix representing the connections between partitions can be reconstructed in this way:

$$B_{i,j} = \begin{cases} 0, & P_{i,j} \le t \\ 1, & P_{i,j} > t \end{cases} \qquad \text{for } 1 \le i \le m \ \text{ and } 1 \le j \le n. \tag{1}$$

A connection between the ith taxon and the jth sample/community is created when the taxon was detected in this particular sample/community, and its abundance exceeded the threshold t. Usually, the threshold is set as 0. However, weak connections can be omitted by setting the threshold higher. To better reveal the hidden patterns in the microbiota composition, we suggest rate edges according to the relative abundance of the taxon across samples. This approach provides suitable results for showing the common and unique parts of microbiota composition for different samples/communities, because all the taxa are given an equal priority that is independent of the abundance of the taxon. A weighted biadjacency matrix W can be computed as:

$$W_{i,j} = 10 \frac{P_{i,j}}{\max(P_i)} \qquad \text{for } 1 \le i \le m \ \text{ and } 1 \le j \le n, \tag{2}$$

where the weight of the edge ranges from 0 to 10. This value was used not only for the width of an edge in the visualized graph but also for the final layout computation. The adjacency matrix $A_{r,r}$ representing the graph is then reconstructed as a square matrix:

$$A_{r,r} = \begin{pmatrix} 0_{m,m} & W \\ W^T & 0_{n,n} \end{pmatrix}, \tag{3}$$

where $r = m + n$. It represents the number of nodes in the final bipartite graph.

To distinguish the vertices between both partitions for visual presentation, we decided to present vertices from taxa partition as smaller than vertices from the second partition.

Table 1. Summary of parameters describing the reconstructed graphs for reduction of taxa partition.

	WHOLE OTU	GENUS	FAMILY	ORDER	CLASS	PHYLUM
No. of vertices	18,503	280	139	98	73	66
No. of edges	37,356	5,776	2,268	1,155	656	407
Average degree	4.037	41.257	32.633	23.571	17.973	12.333
Graph density	<0.000	0.148	0.236	0.243	0.250	0.190
Average path length	3.713	2.118	2.042	2.053	1.916	1.960
Modularity	0.577	0.281	0.287	0.303	0.288	0.263
No. of communities	4	4	4	4	4	4

Table 2. Summary of parameters describing the reconstructed graphs for reduction by abundance threshold.

THRESHOLD	0	0.005	0.01	0.05	0.1
No. of vertices	64	21	16	12	10
No. of edges	156	33	25	16	10
Average degree	4.875	3.143	3.125	2.667	2
Graph density	0.077	0.157	0.208	0.242	0.222
Average path length	2.105	2.181	2.133	2.242	2.711
Modularity	0.224	0.338	0.340	0.377	0.411
No. of communities	4	4	4	4	4

Any arbitrary distinction, eg, circles vs. squares, can be used. Both partitions are identifiable from the matrix and their presentation is therefore dependent on software used for visualization (Gephi, Cytoscape, etc.).

Results and Discussion

Original data visualization. The entire dataset consisted of 18,451 OTUs detected in 52 cecal microbiota samples obtained during the whole life of egg-laying hens. For the first analysis, we only transformed the absolute values of OTU table into relative counts. We consider that this is a sufficient and suitable way for OTU table preprocessing, because of following edge weighting and data reduction by sample merging. Due to different sequencing depths and substantially different composition of the microbiota between the different stages of the lives of the hens, rarefaction may transform the data in an inappropriate way by removing the differences between samples; however, this was not tested because the results without using rarefaction satisfied the purpose of the proposed visualization. The resulting graph is shown in Figure 2. Many small clusters of the vertices from the taxa partition are observable in the graph. These clusters were divided into four large communities, according to modularity optimization, that are represented by four different colors. This result agrees with data description containing four main clusters. Unfortunately, vertices with a low average degree repulsed the vertices with high degrees. These were the vertices that represented the samples. Although this repulsion did not affect community detection, the resulting comet shape of the graph together with an enormous number of edges makes visualization unclear and prevents capturing the labels of the vertices. Although the content of particular communities could be analyzed by further inspection using additional graph algorithms, the purpose of the presented workflow is to visualize the data to the naked eye without the need for additional steps. Therefore, the data reduction is needed before community detection and graph presentation.

Reduction of taxa partition. In the next step, we tried to provide visualization that would on the one hand preserve the correct distribution of communities, but on the other hand, present the distribution in a much clearer way. Such a presentation can be achieved by data reduction. This reduction should not affect the results of the analysis while improving the overall layout. Thus, we decided to observe the overall influence of reducing the taxa partitions. We reconstructed the graphs gradually from the genus to the phylum levels. The results are summarized in Table 1. During the reduction of vertices from taxa partition, division of vertices from sample partition remained satisfactorily consistent when 89.5% vertices remained in cluster 1, 88.3% in cluster 2, 95.4% in cluster 3, and 100% in cluster 4. These four clusters represented the four biological stages defined in Ref. 21, when cluster 1 contained mostly samples from newly hatched chickens, cluster 2 mostly 2–4-week-old chickens, cluster 3 mostly 8–26-week-old chickens, and cluster 4 the rest of samples from hens. Smaller size of cluster 1 and its overlap with cluster 2 caused its lowest consistency.

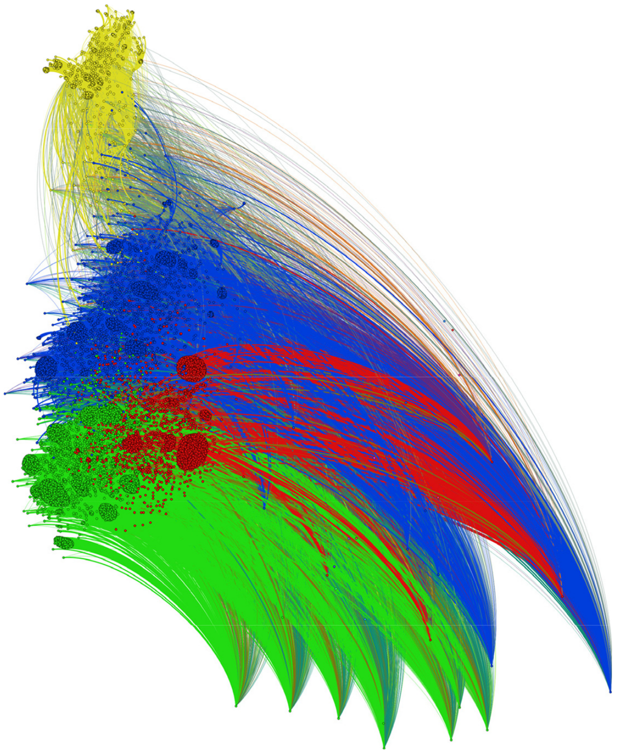


Figure 2. Bipartite graph reconstructed from the entire OTU table with four detected communities.
Notes: Due to different sequencing depth for particular samples, relative abundances of OTUs were used. Communities were detected based on modularity maximization. Vertices (samples and OTUs) within the same community are colored with the same color.

The most noticeable difference was between the whole OTU graph and the genus graph, which is not surprising, because resolution of V3/V4 16S rRNA is sufficient only to the genus level, and the whole OTU graph consists of many unidentified sequences on the species level. The number of vertices was reduced 66 ×, to 280 identified genera. Because the number of edges was reduced only 7 ×, the average degree of vertices and the graph density increased. Thus, the organization of such a graph is clearer. Although modularity decreased in comparison to the whole OTU graph, four main communities were still detected. Other moves to higher taxonomical levels caused additional reduction of vertices and edges, while modularity and detected communities remained similar. Although the reduction of taxa partitions according to higher taxonomical levels improved the graph layout while the results of community detection remained similar, 407 edges are still too many to provide a clear overview of the data. Moreover, in the phylum graph, 52 of 66 vertices represented samples, so community detection is a matter of sample clustering, and such clusters were already revealed by UniFrac-PCoA.[21] On the other hand, distribution of phyla among these four stages of microbiota development would bring interesting additional information, because even in a PCoA biplot, the connection between samples and phyla was not easy to determine. The combination of UniFrac-PCoA, area charts, and hierarchical clustering was needed to reveal this pattern.[21]

Reduction of sample partition. On the contrary, a bipartite graph contains the information about the connections between taxa and samples by definition. In addition, another reduction can be applied to the graph by merging the samples together. By combining the samples according to the cluster they are assigned to, the bipartite graph can provide a very simple yet very powerful visualization, as presented in Figure 3.

The partition representing samples was now replaced by a partition standing for the different stages of microbiota development. The size of the vertices from the partition representing stages was used to distinguish both partitions to the naked eye; it did not play any role in the graph layout computation or in community detection. Because all the samples were related to one organism, a hen, and the purpose was to describe the microbiota composition, this kind of graph denoted the situation in a very clear way. The number of vertices and edges was reduced to 15 and 28, respectively, while other parameters such as modularity (0.255), graph density (0.267), and average path length (1.886) remained more or less the same as in the graphs representing all the samples. Moreover, it revealed the same pattern as did the whole OTU graph in Figure 2, because every stage was included in the different community, but in a way that provided information about the main connections among the stages and their most abundant phyla directly to the naked eye. Thus, it presented the correctness of the division of microbiota development as well as the robustness of the bipartite graph representation against the merging of the samples.

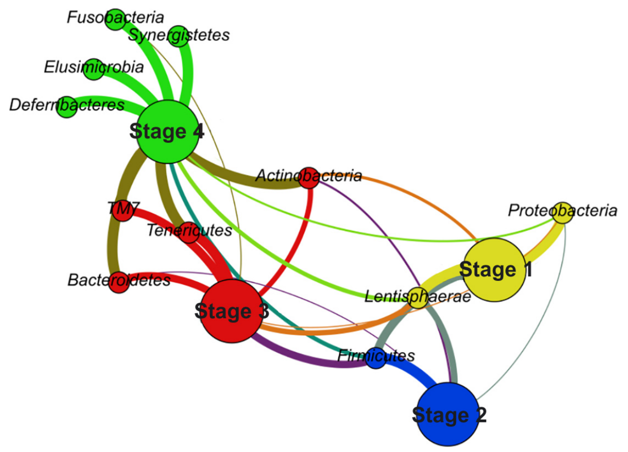


Figure 3. Bipartite graph representing four stages of microbiota development.

Notes: Four communities were detected based on modularity maximization. The color coding of particular communities (yellow, blue, red, and green) corresponds to the color coding used in Figure 2. Other colors represent intercommunity connections. Partitions are distinguished by the size of the vertices.

Reduction by abundance threshold. Eventually, we decided to examine the impact of the threshold t in formula (1) on community detection. Bipartite graphs showing four detected stages and particular bacterial families were used to present the results. While the four selected stages consisted of many different bacterial families, most of them formed less than 1% of the microbiota composition. Thus, we decided on another reduction that was based on focusing only on considerably abundant taxa. The results are presented in Table 2.

A lot of the bacterial families did not reach even 0.5% abundance in particular stages, which led to a massive reduction of the graph, even for very low threshold ($t = 0.005$), from 64 to 21 vertices and from 156 to 33 edges. Further increases of the threshold meant additional graph reduction, while the modularity increased. This is caused by two main facts. First, the average degree of vertices from the partitions representing stages is higher than the average degree in general, and communities are therefore formed around these high-degree vertices. Second, every taxon was given the same priority, according to formula (2), which was determined by its relative abundance across the samples rather than by its relative abundance within samples. Thus, by a reduction of the edges connecting different communities coupled with an increasing threshold, the modularity was also increased. That is also the reason why in every graph from Table 2, no two vertices representing stages were assigned to the same clusters. Our bipartite graph representation was designed primarily for the comparison of microbiota compositions in different stages or environments in a visual manner.[20] Therefore, this feature demonstrated that increasing the threshold did not influence the results highlighting different communities. Although the communities remained coupled with different stages, the increased modularity indicated the changes between and within the communities. These changes should be taken into account when making any conclusions about the diversity of the communities. This kind of visualization primarily shows most abundant, thus most typical taxa for a related community while information about overall taxa richness and their occurrence in samples or communities is lost. Even a low threshold can lead to a massive

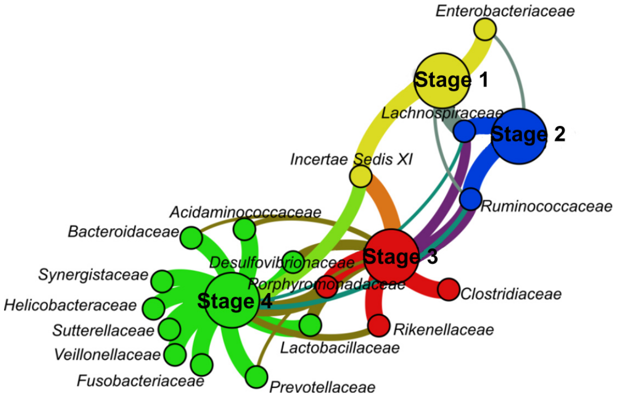


Figure 4. Bipartite graph showing bacterial families exceeding the threshold of 0.5% abundance.

Notes: The color coding is similar to the color coding used in the previous figures. The threshold $t = 0.5\%$ was used to omit any vertices as well as edges that do not meet the requirement for sufficient abundance.

reduction of vertices and especially edges when a lot of low-abundant taxa is present in particular samples. An example is given in Figure 4, a graph showing families with at least 0.5% abundance.

Although the family *Fusobacteriacae* was captured in both stages 4 and 3, the threshold was exceeded only in stage 4. Therefore, the edge connecting *Fusobacteriacae* and stage 4 had maximum weight, because no other connection for *Fusobacteriacae* was made. On the contrary, the edge connecting *Ruminococcaceae* with stage 4 had a lower weight, despite the fact that *Ruminococcaceae* are approximately 10 × more abundant in stage 4 than *Fusobacteriacae*. This is no mistake, nor a disadvantage, because the proposed graph representation was meant to determine the strongest connections among taxa and samples, stages, or environments in a manner in which every taxon has the same priority. However, it should not be utilized for any studies describing overall diversity, especially when a nonzero threshold t is used.

The presented workflow is a suitable complement to the current techniques, eg, UniFrac-PCoA, and can provide advantageous and clear visualization. It is applicable to any kind of microbiota studies, but its main use can be found in descriptive studies where more environments, developmental stages of microbiota composition, or microbial communities are being compared. As it is mainly meant as descriptive approach considering the relation between sample and microbiota composition, no relations among taxa can be revealed.

Conclusion

Amplicon-based metagenomics is a standard approach for investigating the diversity of microbial communities. Although several different techniques can be used for a visualization analysis of microbiome data, there is still room for further improvement, especially for processing data with assigned taxonomy by a reference database. In this article, we proposed a novel workflow for the reconstruction of a bipartite graph that can present the common and unique parts of different samples or communities very clearly. As we demonstrated, the overall layout of the graph can be greatly improved by data reduction coupled with moving to higher taxonomical levels without affecting the result of the analysis in an inappropriate way. It also allows additional analysis of the detected communities by merging the samples without affecting the detected patterns. The results of community detection are also robust against any reduction of data by considering only taxa with abundance higher than a selected threshold, because every taxon is given the same priority during graph construction. Our graph's representation revealed the same pattern as the standard UniFrac-PCoA technique for the reference dataset. However, it provided additional information about the unique and the common parts of the microbiota composition during the different stages. On the other hand, information about the overall diversity was lost, especially when a nonzero threshold was used. Thus, our approach can supply the UniFrac-PCoA analysis with different aspects of the data that would otherwise require combination of several additional techniques.

Acknowledgments

The authors thank the MetaCentrum under the program LM2010005 and the CERIT-SC under the program Centre CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, Reg. no. CZ.1.05/3.2.00/08.0144 for providing computational resources.

Author Contributions

Conceived and designed the experiments: KS, PV. Analyzed the data: KS. Wrote the first draft of the manuscript: KS. Agreed with manuscript results and conclusions: KS, PV, HS, IR, IP. All the authors reviewed and approved the final manuscript.

REFERENCES

1. Land M, Hauser L, Jun S, et al. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics*. 2015;15(2):141–61.
2. Simon C, Daniel R. Metagenomic analyses: past and future trends. *Appl Environ Microbiol*. 2011;77(4):1153–61.
3. Fierer N, Breitbart M, Nulton J, et al. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl Environ Microbiol*. 2007;73(21):7059–66.
4. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*. 2005;71(12):8228–35.
5. Hamady M, Lozupone C, Knight R. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J*. 2009;4(1):17–27.
6. Scholz M, Lo C, Chain P. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr Opin Biotechnol*. 2012;23(1):9–15.
7. Klindworth A, Pruesse E, Schweer T, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res*. 2012;41(1):e1.
8. Zhou J, Wu L, Deng Y, et al. Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J*. 2011;5(8):1303–13.
9. Kim O, Cho Y, Lee K, et al. Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int J Syst Evol Microbiol*. 2012;62(3):716–21.

10. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res*. 2007;17(3):377–86.
11. Meyer F, Paarmann D, D'Souza M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*. 2008;9(1):e386.
12. Caporaso J, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7(5):335–6.
13. Wilkinson L, Friendly M. The history of the cluster heat map. *Am Stat*. 2009;63(2):179–84.
14. Gonzalez A, Knight R. Advancing analytical algorithms and pipelines for billions of microbial sequences. *Curr Opin Biotechnol*. 2012;23(1):64–71.
15. Faust K, Raes J. Microbial interactions: from networks to models. *Nat Rev Microbiol*. 2012;10(8):538–50.
16. Khanipov K, Golovko G, Rojas M, et al. CoCo: an application to store high-throughput sequencing data in compact text and binary file formats. 2015 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Washington, DC: IEEE; 2015:1117–22.
17. Brenner J, Putonti C. HAsh-MaP-ERadicator: filtering non-target sequences from next generation sequencing reads. 2015 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Washington, DC: IEEE; 2015:1100–1.
18. Altschul S, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–10.
19. Videnska P, Rahman M, Faldynova M, et al. Characterization of egg laying hen and broiler fecal microbiota in poultry farms in Croatia, Czech Republic, Hungary and Slovenia. *PLoS One*. 2014;9(10):e110076.
20. Sedlar K, Skutkova H, Videnska P, et al. Bipartite graphs for metagenomic data analysis and visualization. 2015 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Washington, DC: IEEE; 2015:1123–8.
21. Videnska P, Sedlar K, Lukac M, et al. Succession and replacement of bacterial populations in the caecum of egg laying hens over their whole life. *PLoS One*. 2014;9(12):e115142.
22. Edgar R. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26(19):2460–1.
23. Wang Q, Garrity G, Tiedje J, et al. Classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol*. 2007;73(16):5261–7.
24. TEAM R. *Core. R*: *A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing; 2014.
25. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Weblogs and Social Media*. Menlo Park, CA: Association for the Advancement of Artificial Intelligence; 2009.
26. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One*. 2014;9(6):e98679.
27. Blondel V, Guillaume J, Lambiotte R, et al. Fast unfolding of communities in large networks. *J Stat Mech*. 2008;2008(10):10008.
28. Lambiotte R, Delvenne J, Barahona M. Random walks, Markov processes and the multiscale modular organization of complex networks. *IEEE Trans Netw Sci Eng*. 2014;1(2):76–90.