A pipeline for functional and visual analytics of microbial genetic networks

Leandro Corrêa[1], Ronnie Alves[1,2], Fabiana Goés[1], Cristian Chaparro[2] and Lucinéia Thom[3]

[1] PPGCC - Federal University of Pará, Belém, Brazil
hscleandro@gmail.com
[2] Vale Institute of Technology, Belém, Brazil
ronnie.alves@itv.org, cristian.chaparro@itv.org
[3] PPGC - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
lucineia@inf.ufrgs.br

Abstract. Microorganisms abound everywhere. Though we know they play key roles in several ecosystems, too little is known about how these complex communities work. To act as a community they must interact with each other in order to achieve such *community stability* in which proper functions could help to adapt and survive to unbearable conditions. Thus, to effectively understand microbial genetic networks it is necessary to explore them by means of systems biology. An important challenge in systems biology is to determine the structures and mechanisms by which these complex networks control cell processes. In this paper, we present the FUNN-MG pipeline for functional and visual analytics of microbial genetic networks allowing to uncover strong interactions inside microbial communities.

Keywords: systems biology, gene and pathway enrichment analysis, graph representation, graph visualization, metagenomics

1 Introduction

Microorganisms abound in every part of the biosphere including soil, hot springs, on the ocean floor, high in the atmosphere, deep inside rocks within the Earth's crust and in human tissues. They are extremely adaptable to conditions where no one else could be able to survive.

Their adaptability is mainly due to the fact that they live in complex communities. Interactions inside the microbial networks plays essential functions for the maintenance and survival of the community. Unfortunately, too little is known about microbial interactions.

With the recent advent of High-Throughput Sequencing (HTS) technologies, metagenomic [1] sequencing approaches have been applied to investigate characterizations of diverse microbial communities, including target sequencing of the phylogenetic marker gene encoding 16S rRNA and whole-metagenome shotgun

[1] Metagenomics is a discipline that enables the study of the (meta)genomes of uncultured microorganisms [5].

sequencing [1]. Additionally, the rapid development of numerous computational tools and methodologies have been explored for effective interpretation and visualization of taxonomic and metabolic profiling of complex microbial communities. Putting into perspective applications in several domains such as agriculture [2], medicine [3] and biomineralization [4].

Despite the large advance in computational technologies for metagenomics analysis there is still a lack of proper tools to highlight the key interactions in microbial communities, and consequently the genes associated to essential metabolic pathways [5]. This task is usually referred as functional analysis of microbial genetic networks and most of the available pipelines deal with a list of microbial genes rather than interactions. Thus, the genomics highlight the "static" view of the genes available in a metagenome, but the interaction as well as the function that will be performed must be evaluated by an enrichment analysis over a proper database of metabolic pathways such as KEEG.

Metagenomics data analysis poses challenges that could be handled by the utilization of Machine Learning (ML) techniques. In fact, ML has been applied succesfully in several genomics problems. In the context of functional analysis it can provide new ways to explore graphs by using robust statistics, dealing with uncertainty in the data and boosting the search for "hot spots" in large microbial genetic networks.

In this work we propose a computational pipeline to evaluate functional enrichment of microbial genetic networks. A weighted graph is built with its basis on the genes and pathways properly induced from the relative abundance of the metabolic pathways enriched by the associated metagenomic data. In addition, non-supervised ML is applied to enumerate network components (clusters) of microbial genes presenting strong evidence of both interaction and functional enrichment.

The main contribution of the proposed strategy are:

- A functional enrichment analysis which takes into account microbial gene interactions;
- A new visual analytics system to explore interactively the enriched metabolic pathways in microbial genetic networks;
- the FUNN-MG *R* pipeline for the identification of network components (clusters) having strong functional enrichment in microbial communities.

2 Metagenomic pathway-centric network analysis

Metagenomic data analysis is a complex analytical tasks in both biological and computational senses. In sequence-based metagenomics, researchers focus on finding the entire genetic sequence, the pattern of the four different nucleotide bases (A, C, G, and T) in the DNA strands found in a sample. The sequence can then be analyzed in many different ways. For instance, researchers can use the sequence to analyze the genome of the community as a whole, which can offer insights about population ecology, evolution and functioning. In this work, we propose the FUNN-MG *pipeline* (Figure 1) which provides a functional and visual analytic system for the identification and exploration of the *key* functions of a microbial community.

The *pipeline* has four main tasks (the rounded rectangles in Figure 1) that must be executed sequentially: i) identification of the metabolic pathways, ii) evaluation of the enriched pathways, iii) detection of strong components (clusters) and iv) visualization of the microbial gene-pathway network. The first three steps are related to the ML part of the strategy while the remaining step deals with the visual analytics of the graph patterns extracted in the previous steps. Next section we discuss each one of these steps, leaving one particular section to the visualization strategy.

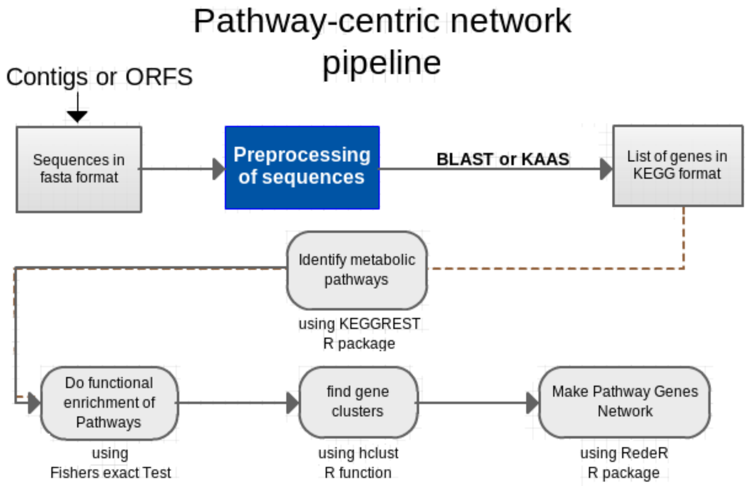
Pathway-centric network pipeline
Contigs or ORFS
Sequences in fasta format
Preprocessing of sequences
BLAST or KAAS
List of genes in KEGG format
Identify metabolic pathways
using KEGGREST R package
Do functional enrichment of Pathways
using Fishers exact Test
find gene clusters
using hclust R function
Make Pathway Genes Network
using RedeR R package

Fig. 1. Metagenomic pathway-centric network pipeline for functional and visual analytics of microbial communities.

3 Materials and Methods

3.1 The metagenomic experimental data

The metagenomic data selected for our experimental study is the Acid Mine Drainage (AMD) biofilm [6], freely available at the site of NCBI [2]. This biofilm sequencing project was designed to explore the distribution and diversity of metabolic pathways in acidophilic biofilms. Acidophilic biofilms are self-sustaining communities that grow in the deep subsurface and receive no significant inputs of fixed carbon or nitrogen from external sources. While some AMD is caused by the oxidization of rocks rich in sulfide minerals, this is a very slow process

[2] http://www.ncbi.nlm.nih.gov/books/NBK6860/

and most AMD is due directly to microbial activity. The AMD metagenome was assembled into 2425 contigs distributed along five main species (see Table 1).

More information regarding the AMD study as well as environmental sequences, metadata and analysis can be obtained at [7].

Species name	Number of contigs
Ferroplasma acidarmanus Type I	412
Ferroplasma sp. Type II	118
Leptospirillum sp. Group II 5-way CG	79
Leptospirillum sp. Group III	959
Thermoplasmatales archaeon Gpl	857

Table 1. The distribution of assembled contigs per species in the AMD metagenome.

3.2 Preprocessing of the metagenomic sequences

We have used the KAAS tool [8] for the identification of 477 microbial genes. This identification was based on the nucleotide percent homology of the groups of orthologous genes [3] found in the KEGG database [9].

The search for microbial genes was carried out in several steps. First, the metagenomic data was split into several groups accordingly to (Table 1), followed by a validation stage of each group within the corresponding species in the KEGG database [7]. KAAS tool was employed sequentially in four steps (Table 2) to obtain the final set of 477 genes:

- Step 1, **finding groups of orthologous genes**: for each specie in the AMD sample we search all its orthologous genes in the KEGG database. For example, the AMD species *Ferroplasma acidarmanus Type I and Type II* are named in KEGG as *Ferroplasma acidarmanus*. So, we use the 530 *contigs* of the associated AMD species as a reference into the KASS tool, retrieving 290 orthologous genes;
- Step 2, **identifying associated species in KEGG**: it basically filters out orthologous genes that are not associated to the reference species. Taking the previous example in Step 1 only 226 genes were kept for the *Ferroplasma acidarmanus* species;
- Step 3, **getting functional annotation in KEGG**: it retrieves the genes associated to pathways in KEGG by using the gene list obtained in Step 2. For instance, 149 genes were retrieved for the *Ferroplasma acidarmanus* specie;
- Step 4, **eliminating duplicated genes**: since pathways are usually associated to one or more genes we deduplicate these genes found in Step 3. So, for the *Ferroplasma acidarmanus* specie we obtained 119 genes.

[3] Orthologous genes are genes in different species that originated by vertical descent from a single gene of the last common ancestor (Homology section on Wikepedia)

All the steps above were executed for all reference species in the AMD sample, taking into account its associated target species in the KEGG database. In (Figure 2) we present this association as well as the distribution of the genes found in the related metagenome.

Id	Species identified	Step 1	Step 2	Step 3	Step 4
fac	*Ferroplasma acidarmanus*	290	226	149	119
lfc	*Leptospirillum ferrooxidans*	450	351	327	217
lfi	*Leptospirillum ferriphilum*	44	33	25	23
tac	*Thermoplasma acidophilum*	26	26	8	8
tar	*Thermoplasmatales archaeon*	412	192	125	107
tvo	*Thermoplasma volcanium*	11	11	3	3
	Genes	1233	839	547	**477**

Table 2. The total number of genes found on each preprocessing step.

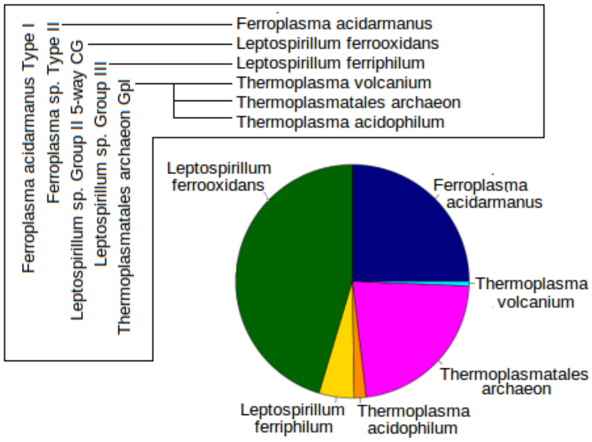

Fig. 2. The dendrogram on the top highlights the association between species in the AMD metagenome and its target species in the KEGG database. In the bottom, a pie chart of the distribution of the 477 genes identified.

3.3 Identifying metabolic pathways

The "*KEGGREST*" *R* package [10] was applied using as reference the list of 477 genes identified, highlighting 95 pathways for the AMD metagenome. Though at this step we cannot assume any strong evidence of functional enrichment regarding to the genes identified.

3.4 Functional enrichment analysis

We devised a functional enrichment strategy based on [11], in which contigency tables are properly set to further apply Fisher's exact test for statistical significance of the enriched metabolic pathways. Fisher's exact test[4] is one of a class of exact tests, so called because the significance of the deviation from a null hypothesis (e.g.: P-value) can be calculated exactly, rather than relying on an approximation that becomes exact in the limit as the sample size grows to infinity, as with many statistical tests.

The main challenge in evaluating the enrichment of a metabolic pathways is the calculation of the probability of finding species covered on each pathway across samples, given that, eventually, only a selected group of species will have an associated pathway. This is also due to the fact that species play distinct roles in the microbial community. As an example, the metabolic pathway *Glutathione metabolism* is annotated for five out of six species identified in the samples (Table 2): *Ferroplasma acidarmanus*, *Leptospirillum ferrooxidans*, *Leptospirillum ferriphilum*, *Thermoplasma acidophilum* e *Thermoplasma volcanium*. So, KEGGREST will only take into account these five species for the enrichment score (Fisher's exact test).

	Gene associated with a pathway	Gene not associated with a pathway	Total gene
Sample	**a** **(6)**	**b** **(364)**	a+b (370)
Population	**c** **(15)**	**d** **(2768)**	c+d (2783)
Total in KEGG	a+c (21)	b+d (3132)	n (3153)

Table 3. The contigency table of the *Glutathione metabolism* pathway which is required for the calculation of the enrichment score.

In Table 3 we present the contigency table required to calculate the enrichment of the *Glutathione metabolism* pathway with respect to the microbial genes found in the samples and its corresponding annotations in KEGG. Having this table, we use the *phyper* function in the "*stats*" *R* package for the enrichment score, followed by a test of significance using the *"Firsher's exact test for count data"* *R* package. Finally, we obtained an enrichment score of 0.0077 (p-value = 0.0292) for the the *Glutathionemetabolism* pathway.

After completing the functional analysis for the 95 metabolic pathways, we obtained a list with only 11 enriched pathways (see Table 4) (p-value $\leq$ 0.05) corresponding to 329 genes. Furthermore, we explore functional modules presenting strong gene interactions by the utilization of a bipartite graph structure $MGP = (G, P, E)$. We called this bipartite graph *Microbial Gene Pathway* (Figure 3. a). MGP vertices are divided into two disjoint sets (G)enes and

[4] http://en.wikipedia.org/wiki/Fishers_exact_test

function	Enrichment	p.value
Purine metabolism	0.033	0.04
Geraniol degradation	6.95e-05	0.01
Cyanoamino acid metabolism	0.008	0.05
Glutathione metabolism	0.007	0.02
Porphyrin and chlorophyll metabolism	0.023	0.03
Metabolic pathways	0.0002	0.0003
Microbial metabolism in diverse environments	0.042	0.05
Carbon metabolism	0.039	0.05
Biosynthesis of amino acids	0.017	0.02
RNA degradation	0.01	0.03
Nucleotide excision repair	0.01	0.03

Table 4. The eleven most significant enriched pathways.

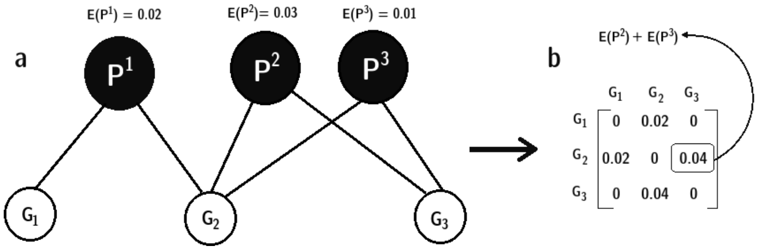

Fig. 3. a) The *MGP* bipartite graph with (G)enes and (P)athways . b) the associated community matrix with the gene-to-gene interaction augmented with the enrichment score.

(P)athways, such that every edge (E) connects a vertex in (G) to one in (P). The enrichment score is annotated in the vertice (P).

3.5 Finding gene clusters

Several groups of genes interact in microbial communities, and some of these interaction are stronger than others. In addition, these interactions usually correlate to the environment in which they are living. We called these strong gene interactions *community patterns*, and potentially they may play a key role in the stability of the microbial genetic network. We have a hypotheses that any perturbation in such patterns could impact directly in the maintenance of the network. We propose a *structural graph clustering* strategy which takes into account a bipartite graph (*MGP*).

The structural graph clustering uses a community matrix (Figure 3.b) based on the genes and its enriched pathways represented in *MGP*. The community matrix observes three main aspects regarding gene-to-gene interactions:

- The existence of one or more metabolic pathways shared by the genes;
- The amount of metabolic pathways in which genes play;

– The enrichment score associated to each metabolic pathway.

The *MGP* bi-partite graph is an interesting computational structure for both the application of ML techniques and interactive visualization of the microbial genetic network [12]. The *community patterns* are obtained directly through the utilization of a hierarchical clustering (*hclust()* R function) technique over the community matrix. The hierarchical clustering solution (Figure 4) requires an euclidean distance matrix that can be built directly through the community matrix. From a biological perspective, the identification of these strong interactions allows for a better understanding of the mechanisms by which these complex networks control cell processes, making it possible to interfere in such processes [13].

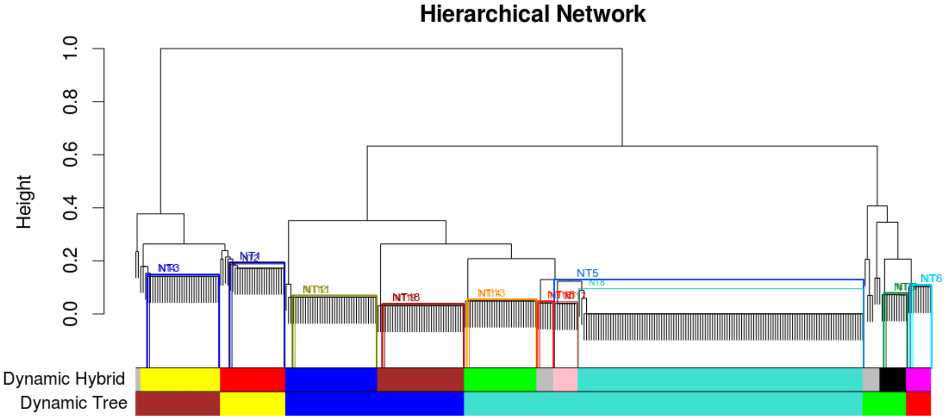

Fig. 4. The hierarchical clustering solution of the community matrix. Two clustering solutions are calculated i) Dynamic Hybrid and ii) Dynamic Tree. Each solution takes into account a distinct cut scheme to form clusters (colored).

The branches of the hierarchical clustering dendrogram correspond to *community patterns* and can be identified using one of a number of available branch cutting methods, for example the constant-height cut or two Dynamic Branch Cut methods. One drawback of hierarchical clustering is that it can be difficult to determine how many (if any) clusters are present in the data set. We employed the *Dynamic Tree Cut R* package to obtain robust clusters [14]. Although the height and shape parameters of the Dynamic Tree Cut method provides improved exibility for branch cutting and module detection, it remains an open research question how to choose optimal cutting parameters or how to estimate the number of clusters in the data set. Two cutting strategies were explored with the *Dynamic Tree Cut*:

– Dynamic tree: the algorithm implements an adaptive, iterative process of cluster decomposition and combination and stops when the number of clusters becomes stable. To avoid over-splitting, very small clusters are joined to their neighboring major clusters;

– Dynamic hybrid: the algorithm can be considered a hybrid of hierarchical clustering and modified Partitioning Around Medoids (PAM), since it involves assigning objects to their closest medoids.

Given that we were looking for compact clusters we decided to use the cutting result obtained with the Dynamic hybrid approach. Thus, 9 clusters and 10 nested subclusters were enumerated. All clusters have the prefix “NT" followed by a sequential number (Table 5). The nested subclusters were calculated with the guide of the *RedeR R* package, and it offered an interesting alternative for the interactive visualization of the microbial genetic networks.

In summary, 308 genes were clustered, corresponding to 96.61% of the enriched pathways related to AMD biofilm. These clusters enclose on average 30 genes, having 6 genes in the most compact cluster and 128 in the largest one. Next, we explore the visual analytic systems over the *MGP* bipartite graph allowing free manipulation of the community patterns as well as the exploration of key hub genes and pathways inside this microbial network.

4 Results and discussion

4.1 Visual analytics system

Given the linked information associated with the concept of microbial communities, it is strongly advised to explore it by graph visualization [15]. The *MGP* bipartite graph fits properly the graph structure required for visualization by the *RedeR R* package. This network visualization system allows several interactive and graph functions such as: zoom, pan, neighborhood highlighting, search, flows, labeling, addition and deletion of graph components.

The structural visualization of the enriched *Microbial Gene Pathway* is presented in Figure 5. The visualization model allows the identification of genes across species and pathways, depicted in distinct colors. It is also possible to explore the degree of connectivity by inspecting the size of the vertices; key players are identified by neighborhood highlighting while clicking on a particular node in the graph network. Such interactive experience allows one to explore resilient aspects of the enriched microbial gene pathway.

The community patterns are explored through the visualization of the graph components associated with the clusters and subclusters (Figure 6). Furthermore, it is also possible to inspect particular spots as well as identify either hub genes, modules or pathways within the network. As an example, the modules are explored as (nested) clusters detected by the proposed pipeline. The Subgroup row in Table 5 identifies these nested clusters. Thus, if one looks to the Group “NT2” we observe a total of 22 genes distributed along the six species (The headers previously described above). NT1 is an example of nested cluster having 1 gene plus 22 genes from NT2, summing up to a total of 23 genes. The symbol “–” shows the there is no nested cluster for that Group.

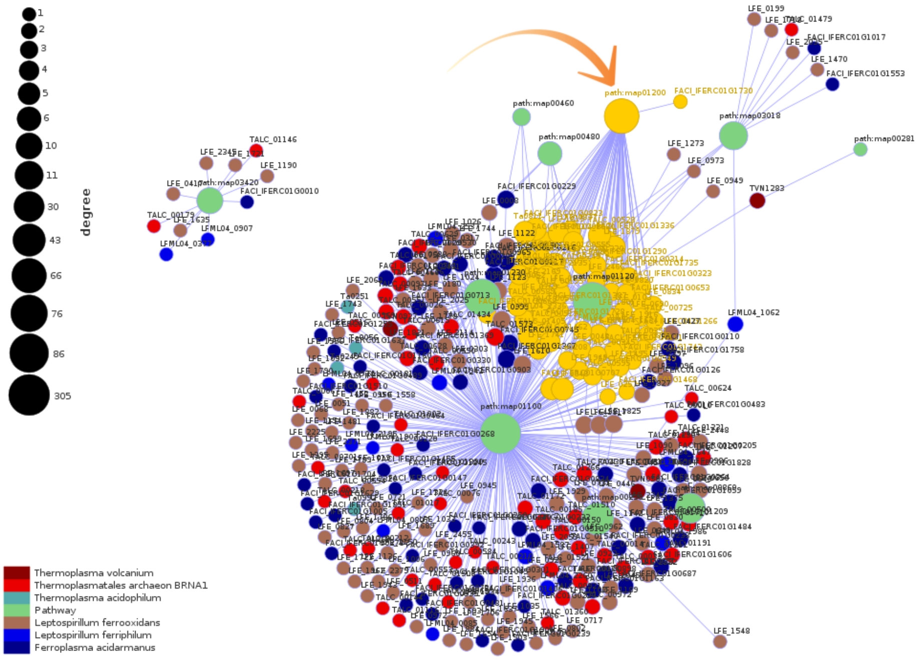

Fig. 5. The enriched Microbial Gene Pathway Network. At the bottom left the legend of the species and associated pathways are represented. Nodes (circles) are related to either species genes or pathways. At the upper left the degree connectivity scale of all nodes. The nodes in highlighting (yellow) are all genes associated to the *Carbon metabolism* pathway (direct orange arrow).

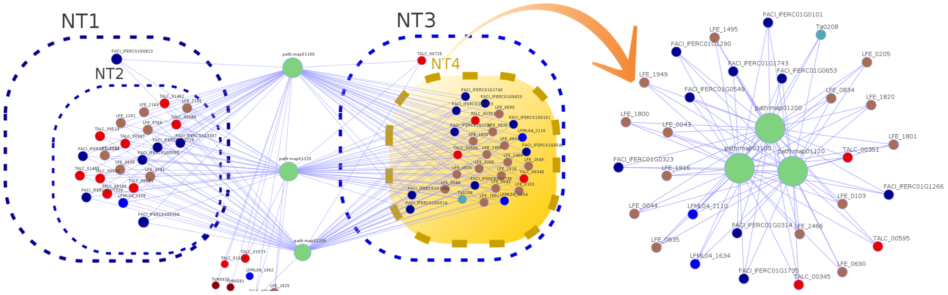

Fig. 6. The representation of the community patterns as clusters (NT1, NT3) and sub-clusters (NT2, NT4). At the right the expanded subnetwork corresponding to elements clustered in NT4.

Group name	NT1	NT2	NT3	NT4	NT5	NT6	NT7	NT8	NT9	NT10
Subgroup	NT2	—	NT4	—	NT17; NT8	NT7	—	—	NT10	—
fac	1	6	0	9	0	0	5	29	0	3
lfc	0	7	0	14	0	1	3	56	1	5
lfi	0	1	0	2	0	0	0	5	0	0
tac	0	0	0	1	0	0	0	4	0	0
tar	0	8	1	3	0	0	0	24	0	0
tvo	0	0	0	0	0	0	0	0	0	1
Total of genes	23	22	30	29	128	9	8	118	10	9

Group name	NT11	NT12	NT13	NT14	NT15	NT16	NT17	NT18	NT19	Total
Subgroup	NT12	—	NT14	—	NT16	—	—	NT19	—	10
fac	0	9	0	7	0	2	1	0	8	80
lfc	0	13	0	12	0	4	5	0	15	136
lfi	0	2	0	2	0	0	2	0	2	16
tac	0	0	0	0	0	0	0	0	0	5
tar	1	10	1	7	1	0	5	1	8	70
tvo	0	0	0	0	0	0	0	0	0	1
Total of genes	35	34	29	28	7	6	10	34	33	—

Table 5. Number of species genes associated to each cluster (Group) and subcluster (Subgroup) calculated by the Dynamic Hybrid cutting strategy.

As an illustration of the visualization, the nested cluster “NT3” having 30 genes is depicted in the middle of Figure 6. As it can be observed the nested cluster “NT3” has 1 gene plus the 29 genes (from “NT4”). The most abundant specie is the “lfc” (colored in brown). Finally, it is presented the eleven enriched pathways (colored in green) connecting all the enumerated nested clusters.

5 Conclusions

The enrichment analysis of microbial genetic networks poses an interesting computational challenge. It is not practical to enumerate all gene-to-gene interaction of a microbial community, so the pathway-centric analysis sound a promising strategy to smooth this combinatorial problem. This strategy has it basis on non-supervised machine learning over a bipartite graph properly built to evaluate the enriched microbial gene pathways.

Interactive visualization of the resulting microbial gene pathway networks allows for the exploration of network metrics enhancing the enrichment analysis. Once all the topological network aspects are understood for a particular metagenome, we envisage the possibility of using such profiles for metagenome comparison as well as classification of unknown microbial genetic network.

Author’s contributions

LC and RA performed the analysis and developped the pipeline. RA and CC supervised the study. LC, RA, CC and LT wrote the manuscript.

Acknowledgements

This work is partially supported by the Brazilian National Research Council (CNPq – *Universal calls*) under the BIOFLOWS project [475620/2012-7].

References

1. Hugenholtz, P., Tyson, G.W.: Microbiology: Metagenomics. Nature **455**(7212) (September 2008) 481–483
2. Fierer, N., Leff, J.W., Adams, B.J., Nielsen, U.N., Bates, S.T., Lauber, C.L., Owens, S., Gilbert, J.A., Wall, D.H., Caporaso, J.G.: Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proceedings of the National Academy of Sciences **109**(52) (December 2012) 21390–21395
3. Bäckhed, F., Ley, R.E., Sonnenburg, J.L., Peterson, D.A., Gordon, J.I.: Host-Bacterial Mutualism in the Human Intestine. Science **307**(5717) (March 2005) 1915–1920
4. Johnston, C.W., Wyatt, M.A., Li, X., Ibrahim, A., Shuster, J., Southam, G., Magarvey, N.A.: Gold biomineralization by a metallophore from a gold-associated microbe. Nat Chem Biol **advance online publication** (February 2013)
5. Wooley, J.C., Godzik, A., Friedberg, I.: A Primer on Metagenomics. PLoS Comput Biol **6**(2) (February 2010) e1000667+
6. NCBI: Metagenomics: Sequences from the environment [internet]. Sequences from the Environment, Tyson (2013)
7. Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., Banfield, J.F.: Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428**(6978) (March 2004) 37–43
8. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M.: KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic acids research **35**(Web Server issue) (July 2007) W182–W185
9. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M.: KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research **40**(Database issue) (January 2012) D109–D114
10. Tenenbaum, D.: KEGGREST: Client-side REST access to KEGG. R package version 1.0.1.
11. Sreenivasaiah, P.K.K., Rani, S., Cayetano, J., Arul, N., Kim, D.H.o..H.: IPAVS: Integrated Pathway Resources, Analysis and Visualization System. Nucleic acids research **40**(Database issue) (January 2012) D803–D808
12. Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabási, A.L.: The human disease network. Proceedings of the National Academy of Sciences **104**(21) (May 2007) 8685–8690
13. A.L. LEHNINGER, N., D.L: Principios da bioquímica. 5 edn. Volume 1. (2005)
14. Langfelder, P., Zhang, B., Horvath, S.: Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. Bioinformatics **24**(5) (2008) 719–720
15. Herman, I., Melancon, G., Marshall, M.S.: Graph visualization and navigation in information visualization: A survey. Visualization and Computer Graphics, IEEE Transactions on **6**(1) (January 2000) 24–43