BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2019
http://excel.fit.vutbr.cz

Estimation of Bacterial Functions in Samples Based on 16S rRNA

Bc. Michaela Bieliková*

Abstract

Humans are hosts to an enormous variety of microbes, bacterial, archaeal, fungal, and viral. Unfortunately, science knows only little about them. Since most of the bacteria has not been studied yet, the main question for a given sample is not only which species of bacteria a specific sample contains, but also what can the bacteria in this sample do (i.e. lipid digestion or resistance to antibiotics). This task is called functional profile prediction and it will be the main focus of this paper. In this paper, I introduce methods for functional analysis, describe existing tools and then design a new tool inspired by them, which implements different methods for the prediction. The results of the experiments imply, that the implemented tool is accurate and useful when using the same method for experimental evaluation as existing tools. However, I propose a new approach to evaluation, that concerns only the most specific bacterial functions, where the results differ from the classic one. In the end, I discuss possible implications of this difference.

Keywords: bioinformatics — metagenomics — bacterial functional profile — KO profile — 16S rRNA — PiCRUST

Supplementary Material: Github repository of the created tool

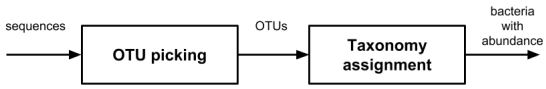
*xbieli06@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Humans are hosts to an enormous variety of microbes. Some of these are invaders that can cause serious diseases, but there is a lot of microbes that are essential to human life. Particularly gut microbiome is crucial for the regular function of the digestion tract. In the last years, it was proven that irregularities in gut microbiome are linked to many conditions ranging from digestion tract diseases like inflammatory bowel disease to antibiotic resistant infections [1]. Unfortunately, because of a big variety of present bacterial species and the impossibility to cultivate most of them in laboratories, the gut microbiome is not well described. Modern approaches in microbiology, specifically high-throughput sequencing and metagenomics, seem to be able to solve these problems and allow us to study microbiome thoroughly and understand how it is connected to human health [1, 2].

Since most of the bacteria present in gut microbiome has not been studied yet, the main question is not which species of bacteria a specific sample contains, as we lack named species for most of the bacteria present in the sample, but instead what can the bacteria do (i.e. lipid digestion or resistance to antibiotics). This task is called functional profile prediction and it will be the main focus of this paper. Functional profile prediction is based on the observation, that bacteria species with similar RNA sequence tend to have sim-

Figure 1. Diagram showing steps of bacterial composition analysis



ilar functions, whereas between species with small RNA similarity the functional profile differs [1].

In this paper, we will introduce existing bioinformatics tools for functional profile prediction, namely PiCRUST, and Tax4Fun. We will discuss the different methods they use for prediction. Then we will design a new tool inspired by them, that will implement the classic methods for functional profile prediction, but should also include a new approach to this problem based on linear regression.

In Chapter 2, I will define theoretical background needed to understand this paper. In Chapter 3, I will discuss the created tool. Chapter 4 contains experimental evaluation of the tool. In the last Chapter, I will talk about the future work and possible extension of the created tool.

2. Theoretical background

In this section, we will define the field this paper relates to - metagenomics. Then, we will discuss the details of the functional composition analysis of a given sample. This section is based mostly on papers from Xochitl C. Morgan [1], Jay-Hyun Jo [3] and Andreas Hiergeist [2], where more detailed information can be found.

2.1 Metagenomics

Metagenomics is a study of genetic material recovered directly from samples. It does not require isolating the DNA of individual species, neither cultivating in laboratories.

There are two main types of analysis often performed in metagenomics. The first one is taxonomic, where the main question is: Which bacteria are present in the given sample? The second one, the main focus of this thesis, is functional: What can the bacteria in this sample do?

The steps of taxonomic analysis can be seen in Figure 1.

2.2 16S rRNA

To minimize the length of the DNA sequence that must be processed to determine the species and functional profile, only a part of genetic information, called marker gene, is used. Marker gene needs to have the following attributes:

- It is present in every organism we want to study
- It is unique for every species
- It is similar for closely related species and different for non-related species

For bacteria, a commonly used marker gene is 16S RNA. It contains conserved regions, that are consistent among all species, and variable regions, that are different. In the taxonomic analysis, we study and compare the variable regions to determine which species we are dealing with.

2.3 Functional profile analysis

In functional analysis, we want to find the different metabolic functions of organisms in the sample, as well as to estimate their abundance — how many organisms in the sample have this function. The process of functional analysis is shown in Figure 2. Functional profiles have the form of KO identifiers with abundance in the corresponding sequences. KO identifiers refer to molecular functions and can be found in the Kegg Orthology database [4]. The functional profiles represent which bacterial functions in what quantity are present in the given sample.

The input of functional analysis can be either raw sequences or already preprocessed and clustered data (OTU table — a matrix that gives the number of reads per sample per OTU). For the purpose of explaining all steps of the analysis, we will suppose we start with raw DNA (or RNA in microbiome research) sequences.

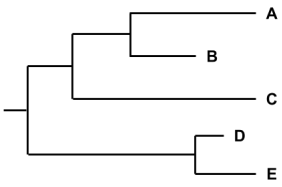
In the data preprocessing, DNA sequences that are very similar (95-98% similarity) are clustered into groups called OTUs. OTU stands for the operational taxonomic unit and is used as a synonym to species since we lack a real, named species corresponding to most of the clusters. Each OTU has an identifier, but these are generic and are not consistent among different samples, so to represent OTUs we use the nucleotide sequences.

The basic principle of functional analysis is to compare representative sequences of OTUs to a reference database that contains the functional profile of previously studied organisms and find the best match. For the genetic material that cannot be paired with a known organism, we can search for the most similar organisms and deduce the functional profile from there.

A big question is, how to find the most similar organisms to a given sentence. Various methods for solving this problem exist. The naive algorithm is based on analyzing the sequences and finding the most similar one - these methods will be further called distance based. More advanced algorithms are based on

Figure 2. Diagram showing steps of functional analysis of a sample

Figure 3. Example of a phylogenetic tree



constructing a phylogenetic tree, that represents evolutionary relationships between the species, of all OTUs and deduces the estimated functional profile from the tree structure.

2.3.1 Functional analysis methods

In this section, I will describe two distinctive groups of methods for functional analysis.

Phylogenetic tree based algorithms

This group of methods is commonly used in bioinformatic tools for functional analysis. It is based on constructing a phylogenetic tree which is a graph that represents evolutionary relations between organisms. Each node of such a tree represents a species. Some of them, specifically the leaves, are living species, while the others may be extinct or only theoretical. The common parent of two nodes is their most probable evolutionary ancestor.

An example of a phylogenetic tree can be seen in Figure 3. This is a tree where the lengths of individual lines between nodes represent the estimated time of evolution. If the line is short, the nodes it connects are very similar, since the time for evolution is short which implies fewer changes in the genome compared to the long lines.

From the phylogenetic tree, we can estimate the evolutionary relationship between different species. Then it is possible to infer a correct combination of known functional profiles for all species for which the functional profile was not found in the reference database.

The inference of unknown functional profiles can be done by finding the nearest nodes with known profiles. We can search for a certain number of known profiles, or limit the search by sequence similarity to the investigated. After we have a set of nodes with known profiles, we compute a consensus profile based on the distance to the investigated node — it can be a simple average, or closer nodes may have a bigger weight than the more distant ones.

Distance based algorithms

The basic idea used in these algorithms is my original work that I introduced in my masters thesis. It is based on analyzing the representative sequences of given OTUs and comparing them to reference sequences with known functional profiles. The resulting functional profile is then inferred from the most similar reference OTUs.

To speed up the search, the similarity between sequences is usually precomputed and stored in a distance matrix. The rows and columns of the distance matrix represent the OTUs and the numbers in the matrix represent the distance of OTUs in the corresponding row and column.

To compute the similarity between sequences, different methods can be used. One of them is to simply count the number of equal characters in their sequence alignments. Others punish the differences according to their evolutionary probability — because of the different chemical nature of the nucleotides in RNA, certain changes in the sequences are more probable than the others. There are various matrices that express the probability of interchange between the nucleotides [5].

2.4 Existing tools

The most used tools for functional analysis are Picrust [6] and Tax4Fun [7]. They both implement the phylogenetic tree approach. The main difference between them is their reference database. Picrust uses Greengenes [8], which is outdated, while Tax4Fun uses Silva [9], which is a newer database that is still frequently updated. To eliminate this disadvantage, the creators of Picrust developed Picrust 2, which is not dependent on reference database [10]. Unfortunately, Picrust 2 is still in beta version.

Picrust is a bioinformatic software package implemented in Python and R, while Tax4Fun is an open-source package for R. Tax4Fun is newer and tries to alter the prediction method of Picrust to make it more accurate. It is also easier to use and faster.

The work-flow of Picrust can be divided into two parts, Gene content inference, and Metagenome inference. In Gene content inference, Picrust takes the reference OTU table from Greengenes database and gene content table from IMG, which is a table containing functional profiles for known genomes. It creates a tree featuring all OTUs from the reference database using ancestral state reconstruction algorithm. For OTUs with an unknown functional profile, an estimated profile is computed, using the position of the given OTU in the phylogenetic tree and the closest OTUs with a

known functional profile. This step is independent on the sample, so it is computed only once. [11]

In Metagenome inference, Picrust takes an user-provided table of OTUs, and using the gene content table from the previous step, predicts metagenomic content of the given sample. The prediction is done by summing up the functional profiles (obtained in the previous step) corresponding to OTUs in the input table while taking into account their abundance. [11]

3. Created tool

In my masters thesis, I have created a new tool for functional profile prediction. It is not dependent on a reference database, as Picrust is. It implements various methods for dealing with OTUs with the unknown functional profile.

The dataflow of the designed tool can be seen in Figure 4. The yellow modules (Input parser, Known profile resolver, Output generator) will be the same for every sample and method for dealing with unknown OTUs, the pink one (Unknown profile resolver) differs and its accuracy is the target of the experiments.

The Input parser loads the data from the input sample. Then the Known profile resolver determines, which OTUs have known functional profiles, and which do not. The unknown ones are then processed by the Unknown profile resolver, which tries to estimate the most probable KO profile composition. Both the known and the estimated profiles are then merged together in the Output generator.

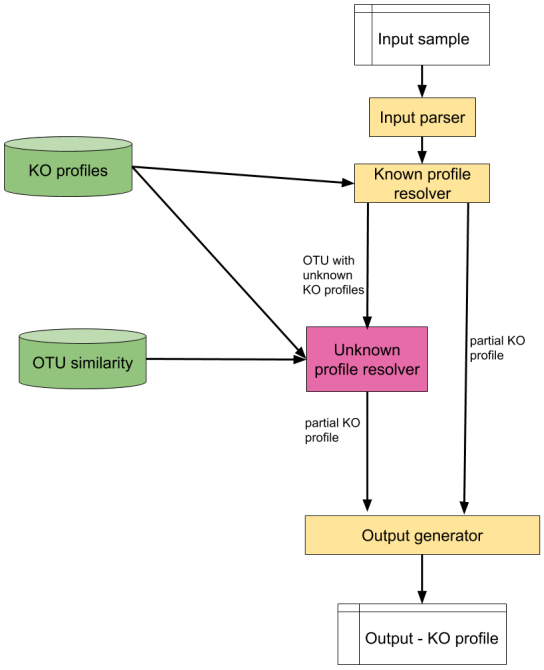
The Unknown profile resolver currently implements three methods for functional profile prediction, two are distance based and one uses a phylogenetic tree. They will be described in detail in the following sections. The results of the experiments will also be given.

The green modules (KO profiles, OTU similarity) represent data sources. They are precomputed and saved in files, but the tool contains code for the precomputation so it is possible to repeat it for other reference databases. KO profiles data source is a table of known species with corresponding KO profiles. OTU similarity is a data source of similarity between OTUs with known and with unknown functional profiles. It can either be a similarity matrix or a phylogenetic tree, or anything else, that somehow represents similarity between OTUs.

4. Experiments

To simulate the real-life situation, where we do not have information about many functional profiles, I have created a version of the reference table with a

Figure 4. Design of the created tool for functional profile prediction based on 16S rRNA data



fraction of rows missing. For 0% missing, the prediction should be 100% accurate, since we have all the data and no estimation is needed.

For each method, the accuracy was tested on 10 artificial samples. To simulate missing functional profiles, a part of reference KO profile table was randomly deleted. The ratio of the deleted table was incrementally increased, from 0% to 90%, to see how much the accuracy drops with more profiles missing. Since the deletion from the reference table was randomized, this step was performed 10 times. To summarize, for each ratio of missing functional profiles, I performed 100 tests.

In the visualization of the results of the experiments I always show the correlation between the expected and the computed functional profile in a boxplot. The y-axis will be the correlation and the x-axis the ratio of the known functional profiles. This way we can see how much the accuracy drops when we do not have enough reference data.

The correlation coefficients are computed as Pearson Product-Moment Correlation, which shows the linear association between two vectors. The values of the coefficients can range from -1 to 1, where 0 means no association between the vectors, values bigger than 0 show positive association and values smaller than 0 show negative association. The proximity to -1 and 1 show the strength of the association. Gen-

erally, values bigger than 0.5 are considered a strong association [12].

4.1 Distance based methods

I have experimented with two methods for functional prediction based on distance. I have used the Greengenes database, for which a global alignment of all sequences is available. Greengenes database was chosen so I can compare my results with Picrust in the future. However, the tool is not dependent on the database, and if another multiple sequence alignment data are used, the prediction will work.

Both methods look for the most similar OTUs with a known functional profile. The similarity is measured by the distance of the multiple sequence alignment. The number of similar sequences that are taken into account is an attribute I tried to experiment with. The best results were achieved when the estimated profile was computed as the average of profiles of 4 most similar sequences.

The difference in the implemented methods is the computation of the distance metric. The first one simply counts the number of similar symbols in the multiple sequence alignment. The second one uses a transition/transversion scoring matrix for nucleotides. It takes into account the chemical attributes of RNA bases and the probability of exchange of the nucleotide pairs.

Surprisingly, better results were achieved with the simpler method, that does not respect the chemical properties of nucleotides. This might be caused by a wrong approach to gaps in the scoring matrix, or by the fact that the distance matrix was computed from the global alignment of all Greengenes database sequences, so the impact of alignment errors might be significant. Therefore, with respect to the limited scope of this paper, this is the only method which will be discussed further. We can see the results of the average-finding method in Figure 5.

4.2 Phylogenetic tree based methods

I have also implemented a method based on a phylogenetic tree. It uses a UPGMA (short for "unweighted pair group method with arithmetic mean", an algorithm for creating a phylogenetic tree based on sequence similarity) method to compute a phylogenetic tree of all the OTUs from the Greengenes database, both the ones with known and unknown KO profiles. Then the unknown OTUs are determined as the average profile of OTUs which are connected to it in the tree.

The tree is computed before the functional prediction. The script for the tree creation is a part of the tool, so the process can be repeated for any reference data.

Figure 5. Evaluation of a distance based method for functional prediction on complete functional profiles. The x-axis represents the ratio of known profiles in the sample, the y-axis represents the correlation between expected and computed result.

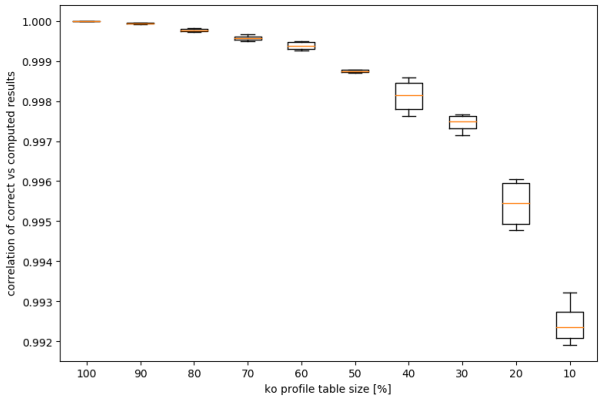
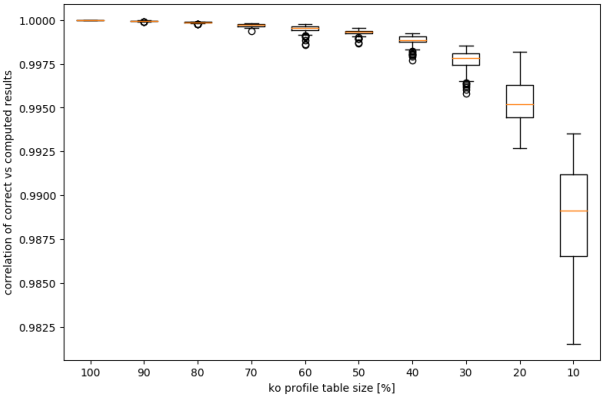


Figure 6. Evaluation of a phylogenetic tree based method for functional prediction on complete functional profiles. The x-axis represents the ratio of known profiles in the sample, the y-axis represents the correlation between expected and computed result.



In UPGMA, each item is paired with the closest item by a given distance matrix. The pair is connected and is assigned a parent node in the resulting tree. In the name of the node, I store names of all its children. That means that the root has names of all the OTUs in the reference database. The closest OTUs to any OTU are the ones with which it was connected.

When looking for the most similar OTU, I search the tree and find, in which point the searched OTU was connected with some other node. The name of the node with which it was connected are the ones that are the most similar, and the estimated profile is inferred from them. The results can be seen in Figure 6.

4.3 Comparison

As we can see, the method based on average is slightly better than the one based phylogenetic tree. However,

Figure 7. Evaluation of a distance based method for functional prediction on the rarest functions. The x-axis represents the ratio of known profiles in the sample, the y-axis represents the correlation between expected and computed result.

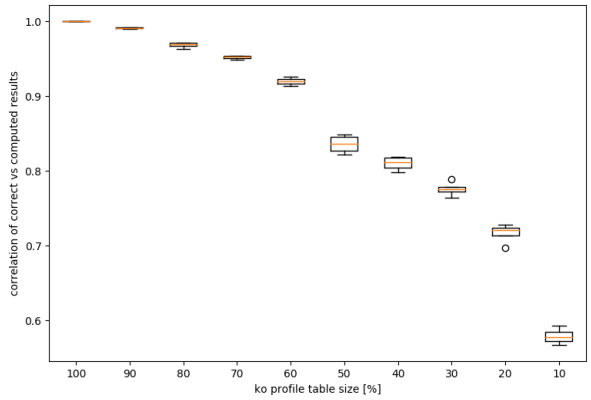
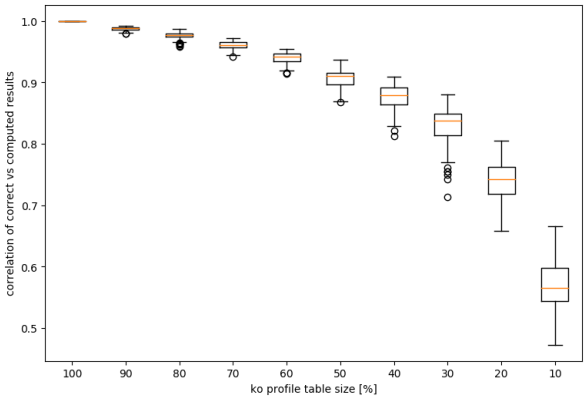


Figure 8. Evaluation of a phylogenetic tree based method for functional prediction on the rarest functions. The x-axis represents the ratio of known profiles in the sample, the y-axis represents the correlation between expected and computed result.



the correlation between expected and computed profiles is surprisingly high, over 99% even when only 10% of the profiles are known. Picrust and Tax4Fun also report a high prediction accuracy [11, 13].

The high correlation can be caused by a number of reasons, the most probable is **common metabolic functions** — each bacteria must have basic functions for translation, transcription, and processing of common metabolites. This is the part of the functional profile, that is the same for every species of bacteria, independent on the sample. A number of KO specific for a certain species of bacteria is much smaller, so it might not have that much of an effect for the correlation.

Since every bacteria has common metabolic functions, it is expected that they will be present in every sample. More interesting are the functions that are exclusive for a certain species. From this reason, I have altered the correlation so that it takes into account only those KOs, that are present in less than 1% of the species in the reference table.

With this approach, the results between the methods are more different, so it is more useful for comparing different algorithms. We can see the new results in Figures 7 and 8. Now, we can see that the correlations drop more significantly, and with only 10% of the profiles known, we are at the 60% match. The difference between the methods is still minimal, which might indicate, that while 16S rRNA is immensely useful as a species identifier, the connection between the 16S sequence and functional profile is not very significant and to predict functional profile more accurately, we need to look at the whole genome. To confirm this hypothesis, more analysis and evaluation have to be performed.

5. Conclusion

In this paper, I have introduced functional profile analysis, which is an important part of metagenomic research. I have discussed the most used methods and created a tool that implements them.

The focus of this paper is the comparison of different methods for functional analysis. I have shown, that the classic approach to experimental evaluation — when we look at the whole functional profile — gives a different result than the evaluation that looks only at the more specific functions.

The significance of the evaluation of the specific functions is that it gives us more detailed information about a sample. The common metabolic functions are a part of every bacteria species, so it is not a surprise to find them in every sample. More rare functions are more informative and harder to predict.

In the future, I would like to compare my tool with Picrust and Tax4Fun. It would be interesting to see how they stand in the tests of only specific KOs since they are widely used in bioinformatics research. I will also add some more methods for functional analysis, including one based on linear regression, to try a more computer-sciences inspired approach.

Acknowledgements

I would like to thank my supervisor Ing. Stanislav Smatana for his help.

References

[1] Xochitl C. Morgan and Curtis Huttenhower. Chapter 12: Human microbiome analysis. In *PLoS Computational Biology*, 2012.

[2] Andreas Hiergeist, Udo Reischl, and André Gessner. Multicenter quality assessment of 16s ribosomal dna-sequencing for microbiome analyses reveals high inter-center variability. *International journal of medical microbiology : IJMM*, 306 5:334–342, 2016.

[3] Jay-Hyun Jo, Elizabeth A. Kennedy, and Heidi Kong. Research techniques made simple: Bacterial 16s ribosomal rna gene sequencing in cutaneous research. *Journal of Investigative Dermatology*, 136:e23–e27, 03 2016.

[4] KEGG Orthology database. [Online: https://www.genome.jp/kegg/ko.html]. [Visited 4.12.2018].

[5] William R. Pearson. An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinformatics*, page Chapter 3:Unit3.1, 2013.

[6] Picrust webpage. [Online: http://picrust.github.io/picrust/]. [Visited 4.12.2018].

[7] Tax4Fun webpage. [Online: http://tax4fun.gobics.de/]. [Visited 4.12.2018].

[8] Greengenes database webpage. [Online: http://greengenes.secondgenome.com/]. [Visited 28.12.2018].

[9] Silva database webpage. [Online: https://www.arb-silva.de/]. [Visited 28.12.2018].

[10] Picrust 2 github repository. [Online: https://github.com/picrust/picrust2]. [Visited 4.12.2018].

[11] Morgan G I Langille, Jesse Zaneveld, J Gregory Caporaso, Daniel Mcdonald, Dan Knights, Joshua A Reyes, Jose C Clemente, Deron Burkepile, Rebecca Vega Thurber, Rob Knight, Robert G Beiko, and Curtis Huttenhower. Predictive functional profiling of microbial communities using 16s rrna marker gene sequences. *Nature biotechnology*, 31, 08 2013.

[12] Pearson Product-Moment Correlation. [Online: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php]. [Visited 1.4.2019].

[13] Kathrin P Aßhauer, Bernd Wemheuer, Rolf Daniel, and Peter Meinicke. Tax4fun: Predicting functional profiles from metagenomic 16s rrna data. *Bioinformatics (Oxford, England)*, 31, 05 2015.