Semester work -
Learning representations of microbe–metabolite interactions

Lupașcu Marian
Syntactic Modeling of Biological Systems
University of Bucharest

May 31, 2020

Context

Metabolic-microbial relationships are essential for the study of the microbiome. In the paper "Learning representations of microbe-metabolite interactions" published on November 4, 2019 in the scientific journal Nature, a new method is introduced that has the power to analyze the metabolite-microbe relationships. This new method is based on a technology not used so far in the study of metabolized microbial interactions, namely machine learning.

It is proved by 5 experiments: two experiments on cystic pulmonary fibrosis, one on the wetting of the biocrust, in the analysis of the impact of a high fat diet in murine a bacterium responsible for the excess production of a new bile acid is determined and in the analysis of inflammatory bowel disease and the colon identify a bacterium responsible for this disease as it was not initially associated with this disease in the Human Microbiome Project, as this method of analyzing metabolite microbe interactions has higher performance than previous methods (which are purely statistical) to do this thing.

Previous results

Two of the methods used so far are Pearson and Spearman correlations that analyze the interactions between two or more events as independent. Presumption that imposes many restrictions and decreases the degree of generalization of the analysis model. Very briefly, each of these two correlations involves a reduction of events to a combination of independent bilinear problems.

Other established techniques are SparCC and proportionality which are robust and scale invariant for a single data set. But as in this paper we are dealing with two data sets, one for microbes and one for metabolites, and the relationships and interactions between them will be analyzed, it turns out that we are working with multiomic data sets so these methods will not give very results. good thing that is demonstrated in the paper.

SPIEC-EASI is the only method compared to the paper model, as it is the only one capable of handling multiomic data sets. But on the analyzed case, namely the microbial-metabolite relations, this method also does not work very robustly due to the differences of the units of measurement and the differences of measurement at the time when the mass spectrometry was made, for the construction of the set date.

Microbe

A microbe or microorganism is a plant or animal organism of microscopic size. They are of several types, including: bacteria, fungi and viruses (in paper only these three classes of microorganisms were analyzed because only these are pathogenic to humans, for example microscopic algae do not cause any disease or infection to humans, so there is no point in observing any interaction of metabolized algae in humans as there will be no such interaction). In addition, some biologists do not consider viruses as microorganisms because they are not organisms, defined by a living system, capable of reproduction. Attributes that are not specific to any virus, which needs a host for reproduction and is zombie rather than alive. Next we will analyze the particularities of each such microorganism.

Bacteria are prokaryotic microorganisms (genetic information is free through the cytoplasm, it is not covered by a membrane), unicellular, alive, capable of reproducing alone. Bacteria were among the first life forms on Earth and are found in all possible habitats of the planet: water (including in the Marianas pit), soil (even kilometers deep in the soil), air, radioactive waste, etc. They are among the most extensive life forms on the planet: 5×10^30 bacteria divided into $10^7 - 10^9$ different species of bacteria. For example, only in the intestinal microflora of a man of 70 kilograms (30 trillion cells) live 39 trillion bacteria.

Fungi are unicellular (yeast) or multicellular organisms, capable of reproducing on their own. They are found in almost any habitat: on the ground, in the ground, in the water, in the air. Some are parasites, such as mold, scabs or canker. A very small number of such fungi infect animals and humans. In humans, the most common infections are skin infections. In this paper only molds and yeasts that are pathogenic to humans are analyzed. For example, Candida, which is a type of yeast, can cause infections in people with a weakened immune system. The total number of fungal species is $1,5 \times 10^6$.

Viruses are the simplest form of that life, they are not alive, but they still have genetic material (DNA or RNA) and do not consume matter or produce energy. The only way to multiply is by parasitizing, affecting living cells and forcing them to multiply their genetic code. It is estimated at over 10^8 types of viruses.

Metabolite

Metabolites are intermediate or final products of metabolism (the totality of biochemical and energetic transformations and processes in an organism). They have various functions: from protection (antibodies, amylase, lysozyme), fuel (glucose, ATP), catalyst, synthesis (vitamins), interaction with other organisms (pheromones), etc. Metabolites are small molecules produced by an organism in order to stay alive. It is currently estimated at around 114,100 metabolites in humans, of which only 18,608 metabolites are detected and quantified in humans.

Architecture

Machine learning is a subdomain of computer science, in this case, artificial intelligence in which the main objective is to give the computer a chance to learn (to have the ability to generalize some examples it receives). Compared to the usual programming paradigms in which an algorithm is written by a person (algorithms that will later have an input at the input and generate an output), in the artificial intelligence paradigm, one algorithm is generated (is output) by another learning algorithm , which has as input pairs input-output (in the paradigm of supervised learning) - the training step. In the end, the output algorithm - the model, will receive as input the same type of input, other than those from the training step and will generate new outputs - the test step and implementation.

Paper model - mmvec, is an artificial neural network - Multilayer perceptron. The two essential characteristics of a neural network are: architecture - the structure of connections and invention - the mechanisms for adjusting connections. The architecture of the mmvec network is a standard one: an input layer, a hidden (latent) layer and an output layer (the dimensions of these layers vary - they are input arguments of the algorithm).

The layers are interconnected with dense / fully connected (FC) layers. A dense layer is one of the simplest layers of a neural network. It consists of a series of perceptrons in which each of them is connected to all perceptrons in the previous layer and to all perceptrons in the next layer. In the figure below we have a diagram of a perceptron (mathematical neuron) - in which we have n inputs (simulates n dendrites), a bias (another dendrite), a summation function, which gathers these inputs multiplying previously with their activations (activation / weight is a parameter to be learned) and the activation function (which changes the summation result for the next perceptron - usually nonlinear).

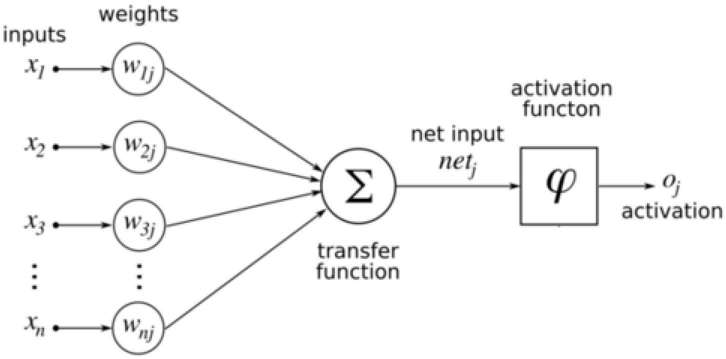

Figure 1: The perceptron

Where:

$$y_k = \varphi\left(\overbrace{\sum_{i=1}^{m} x_i \cdot w_{ki} + b_k}^{v_k}\right)$$

This model is inspired by how word2vec estimates the output vector (a vector the size of several hundred) depending on the input context. word2vec is an artificial neural network of the multilayer perceptron (MLP) type in NLP (Natural Language Processing) whose main purpose is to disambiguate words. We want the proximity of words with a similar meaning (student - university - course are close but far from the word cactus) and the distance of words that have no syntactic connection. There are some major differences to consider. In word2vec, a skipgram was proposed (designed to take into account the sequential nature of the text), which on paper is not necessary because there is no sequential nature of DNA (DNA is continuous and a certain pattern follows - otherwise there may be gaps in the DNA). As a result, skipgrams do not make sense to be used in mmvec architecture.

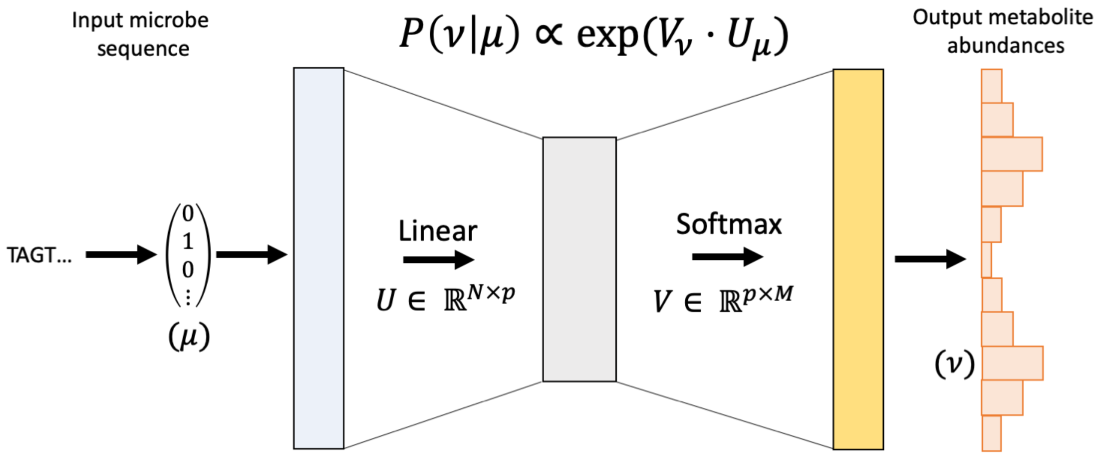
Input microbe sequence
$P(\nu|\mu) \propto \exp(V_\nu \cdot U_\mu)$
Output metabolite abundances
TAGT...
$\begin{pmatrix}0\\1\\0\\\vdots\end{pmatrix}$
(μ)
Linear
$U \in \mathbb{R}^{N\times p}$
Softmax
$V \in \mathbb{R}^{p\times M}$
(ν)

Figure 2: mmvec architecture

The mmvec network input is a piece of DNA sequenced for a particular microbe, say n nucleotides. Then this nucleotide sequence (containing the letters A, T, C, G) is numerically encoded in 4 dimensional vectors, namely: A - (1,0,0,0), T- (0,1,0,0) , C - (0,0,1,0) and G - (0,0,0,1). So the input layer of the mmvec model has the size 4n. Then this layer is connected by a dense layer to the latent layer the size of several tens of times larger than the input start. It is connected through a dense layer to the output layer that has softmax activation. Output is a m dimensional vector of activations / probabilities: high activation means high probability of interaction between the microbe that had the input DNA and the respective metabolite. In this way the metabolized microbial interactions are estimated.

As a learning algorithm was used SGD (Stochastic Gradient Descent) with the function of loss: log loss / cross entropy loss, on a data set of 562 DNA sequences of microbes, 26,966 metabolites (not only from humans) and 400 samples -hate. As filtration: remove microbes that did not appear in at least 10 samples (have very little activation). The training takes place on the GPU (graphics processing) using Tensorflow technology, and lasts somewhere in a few hours.

Experiments

Initially demonstrated the effectiveness of the mmvec model on Pseudomonas aeruginosa (P. aeruginosa) in a simple environment - cystic pulmonary fibrosis. It is then shown that mmvec can resolve the contradictory relationships of metabolized cyanobacteria in a study on the wetting of biocrust in semi-desert soil. Also, the analysis of the microbes generating cystic pulmonary fibrosis confirms the efficiency of mmvec compared to the other methods used for this type of analysis up to mmvec. Finally, we analyze the metabolic changes in the bile and small intestine for murinae fed a high-fat diet (HFD) and identify a new bacterium that is responsible for producing excess bile acid, and implicitly responsible for this disease. Finally, the power of this model is confirmed by identifying a bacterium that was not initially associated with inflammatory bowel disease in the Human Microbiome Project, demonstrating that mmvec can discover the etymology of diseases in complex environments - small intestine and colon.

Cystic pulmonary fibrosis

Cystic pulmonary fibrosis is characterized by chronic inflammation and repeated infections of the lung, often with antibiotic-resistant bacteria (especially P. aeruginosa). It leads to the destruction of lung tissue and lung failure (deformities of the alveoli of the lungs by inflammation of the pulmonary tissue). It is the most common cause of death in this disease.

Previous studies have shown that certain changes in the environment in the lungs can influence this infectious disease. That is, if the oxygen concentration in the lungs decreases and the pH drops below 7, to an acidic one then the community of P. aeruginosa will decrease (without using antibiotics) and will be replaced by one of bacteria that produce ferments (harmless for lungs), which will substantially exceed the community of P. aeruginosa. Due to the simplicity of this model (P. aeruginosa and ferments) this experiment is performed for the first time and has the role of confirming the efficiency of the mmvec network.

Initially, the nucleotide sequence is known for P. aeruginosa and for fermenting bacteria. These DNA sequences of the mmvec network are given as input, after which the network will encode these strings into input vectors larger than the number of nucleotides (factor of 4). In the end, the network will generate the probabilities of belonging to some metabolites. And it is observed that for P. aeruginosa the metabolites of interest are: ammonia and amino acids, because P. aeruginosa releases these metabolites. And for yeast-producing bacteria, the metabolites of interest are sugar and acids. This makes sense because these bacteria consume sugar and produce acidity in the lungs. This can also be seen in the graph below.

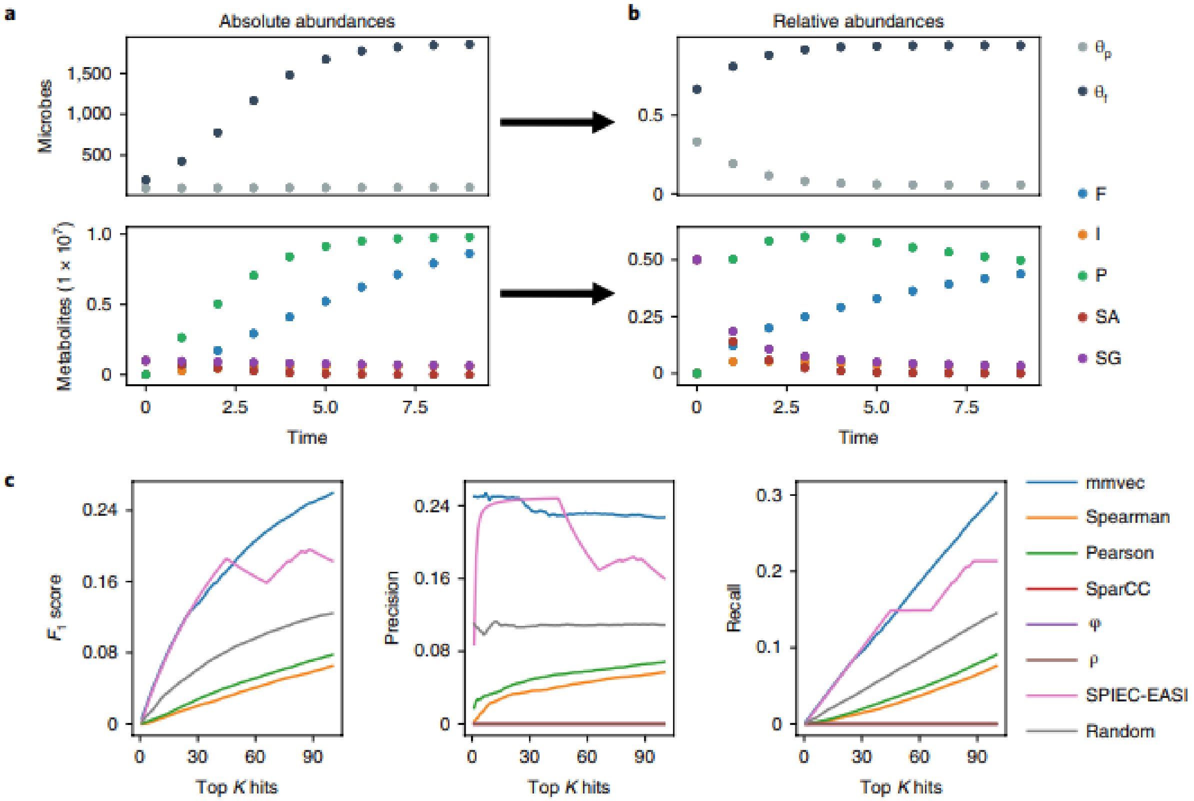

Figure 3: At the top of the graph it can be seen how the abundance of the P. aeruginosa community decreases and the fermentation community increases (in a and b). At the bottom of the graph (a and b) you can see how the metabolic environment is influenced: the amount of ammonia and amino acids decreases because P. aeruginosa no longer produces these metabolites by decreasing the community. And the amount of sugar begins to decrease and the amount of acids begins to increase because yeasts consume sugar and produce acidity (growing community). At the bottom of the graph (c) it can be seen that the performances on the 3 established scores (Precision, Recall and the F1 score - harmonic average Precision Recall) the performance of the mmvec network is superior to the previous methods.

Cystic pulmonary fibrosis II

As I said above pulmonary fibrosis is a disease that occurs mainly by infectious pathways from a number of bacteria. The main bacteria are P.aeruginosa (pathogen) and anaerobic bacteria (Veillonella, Fusobacterium, Prevotella and Streptococcus). These bacteria occupy completely separate niches in cystic pulmonary fibrosis infection: P. aeruginosa loves oxygen and the basic environment of the earth and anaerobic bacteria hate oxygen and love an acidic environment in the lungs (pH below 7).

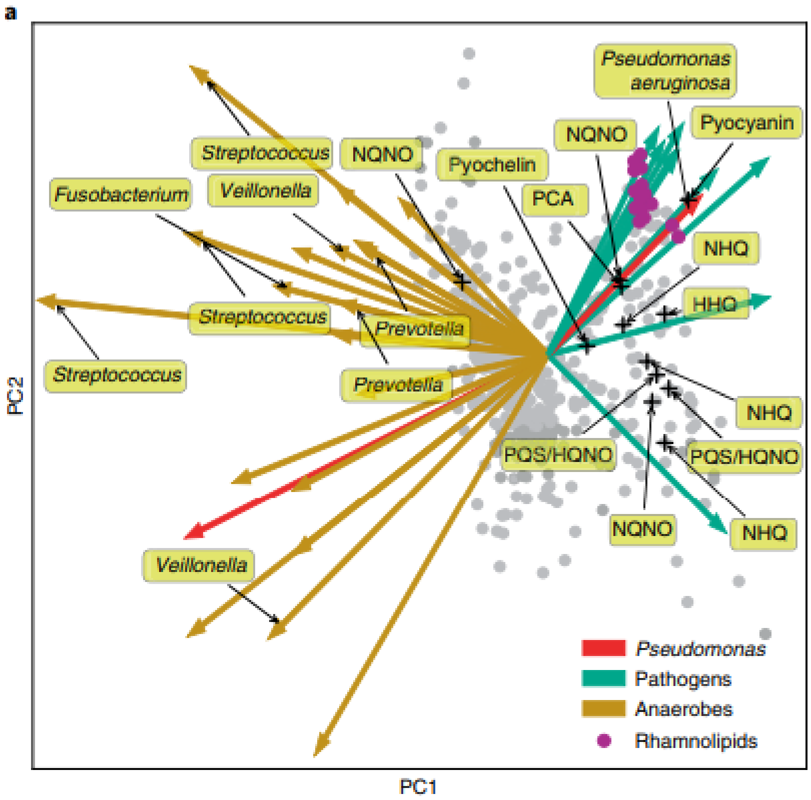

Figure 4: Separation of the pathogen from anaerobic bacteria

Identifying this clear separation between bacteria was a confirmation of the fact that the mmvec network works well in this experiment as well. For this, exactly as in the above experiment, the DNA sequence for P.aeruginosa and for the 4 anaerobic bacteria is passed to the network. Mmvec has achieved a clear separation of bacteria in terms of the metabolites with which they interact. Namely: 4-hydroxy 2-heptyl quinoline, pio cyanine, phenazine-1 carboxylic acid, 2-phenyl 4 hydroxy quinoline, 2-heptyl, 4 hydroxy crinoline and pyrocline (all these molecules are among the green arrows in the figure below) and are strongly correlated with P. aeruginosa (also one of the green arrows in the image - a sign that these metabolites are positively correlated with this pathogen). In addition, there are strong activations for a number of rhamnolipids that are a virulence factor for P. aeruginosa. As for anaerobes, it can be seen that all these bacteria are strongly correlated with each other and also correlated with P. aeruginosa and its specific metabolites. The clear separation can also be seen in the graph below. In addition, this experiment is purely theoretical and has the role of demonstrating the performance of mmvec.

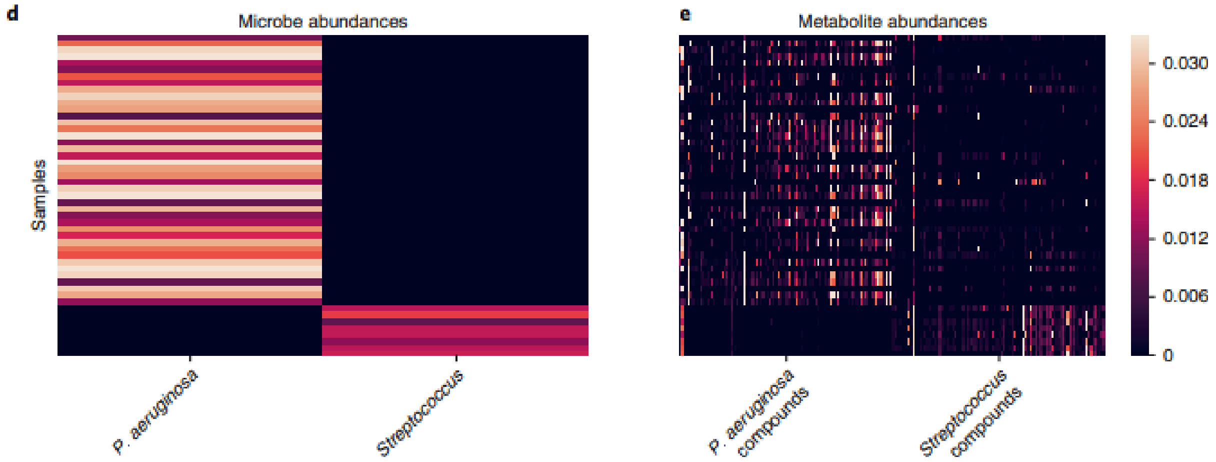

Figure 5: On the left side (d) you can see the abundance of the two classes of microbes analyzed (P. aeruginosa and anaerobes represented by Streptococcus) and on the right side (e) the metabolites with which they interact. What is notable is that the metabolites for P.aeruginosa do not interact with those of Streptococcus but vice versa this rarely happens, and the figure (s) is more of a diagonal rush which means that mmvec has a fairly high accuracy.

Soil biocrust wetting event

Biocrust is a piece of soil in the desert or semi-desert environment. This soil, when it is taken - it is dry, has a large population of Microcoleus Vaginatus and a small population of Bacilli. When this soil is moistened after a period of 50, the opposite effect is observed: a large population of Bacilli and a small population of Microcoleus Vaginatus.

Moreover, in this interaction of bacteria (microbes) are known from previous studies 100% of the relationships of metabolized microbes. So to see the performance of mmvec compared to the other methods, the DNA sequences for Microcoleus Vaginatus and Bacilli are passed to the network. And the network will generate specific metabolites for Microcoleus vaginatus and Bacilli. The results are surprising, namely a performance of 10% above the best estimate, the one made by the Spearman correlation. In the figure below you can see the improvement, in terms of false positive and negative.

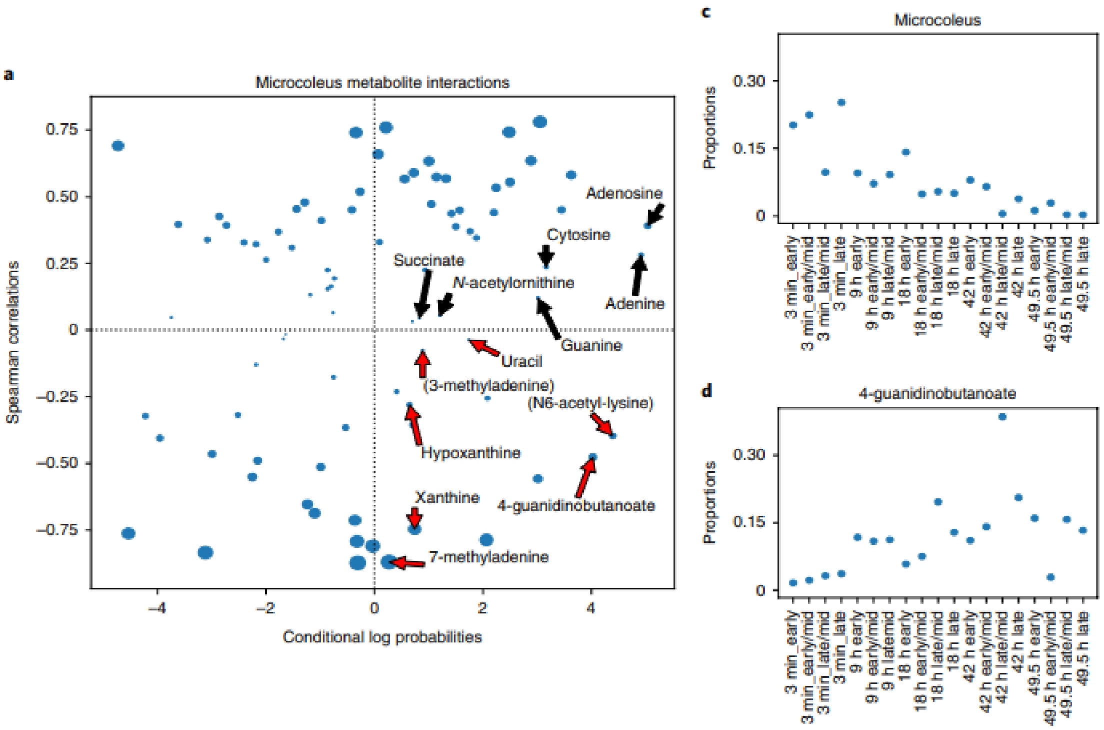

Figure 6: (a) Prediction differences between the Sperman correlation and the mmvec network. The red arrows represent metabolites predicted as being released by the Sperm correlation are actually produced, which is also confirmed by mmvec. The arrows in black, unfortunately, show falsely positioned examples labeled by mmvec. (c) Evolution of populations of Microcoleus Vaginatus initially and during the 50 hours after watering the biocrust - a decrease in the number of such bacteria can be observed. In (d) we have the evolution of the population of Bacilli initially and after 50 times - an increase in the number of such bacteria can be observed

The effects of a high-fat diet in the murinae model

Next we will move to a more complex environment, namely the small intestine of the mouse murinae. A new type of bile acid (phenylalanine cholic amidate) has recently been discovered that is not conjugated to glycine or taurine, as is normally the case with bile acid. Several previous studies have shown that this type of bile acid is microbial and grows with a high-fat diet (HDF). As it comes microbially, those in the paper thought of obtaining the microbes that generate this metabolite. This involves the reverse problem: giving a metabolite, or a sample with several metabolites to obtain the microbes that generate that sample of metabolites. Well, this is extremely difficult so the paper comes with a brute force solution.

The solution involves iterating through all the microbes suspected of agenerating the input sample. For these microbes, the DNA that passes to the mmvec network is known. The network output will be a probability vector for all metabolites (with high activation for probable metabolites and low activation for unlikely ones). Then the distance between these vectors and the reference vector is made (the reference sample, the one extracted from the mouse's intestine). In the end, only vectors that have a distance less than a threshold are selected or receive x such vectors, the vectors corresponding to some microbes. In this way the microbes that generate a certain vector of metabolites (sample) are identified.

In this experiment, with the method stated above, the bacterium Clostridiales is identified is responsible for the excess production of this type of bile acid. Because this bacterium consumes this type of bile acid, therefore the bile in will try to compensate for the acid loss, consumed by the bacterium in excess. This can be seen in the figure below.

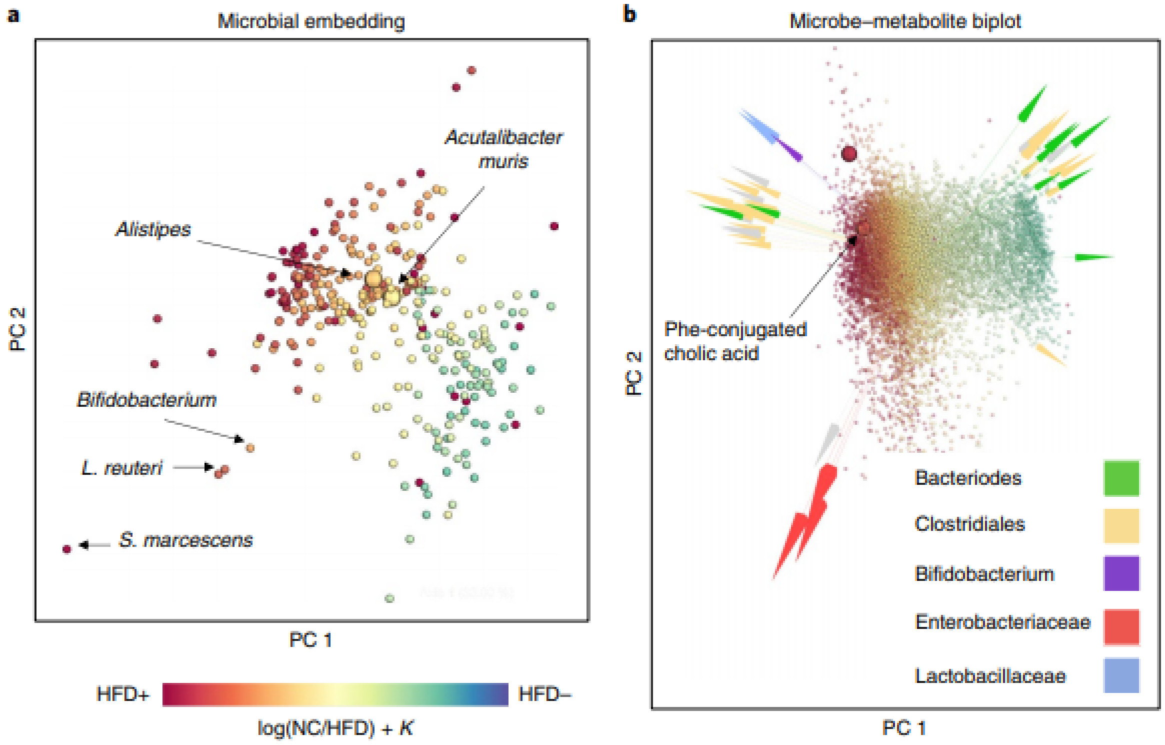

Figure 7: On the right side you can see how the new bile acid is correlated with a high fat diet (it is red) and the bacteria responsible for excess production are: Bacteroides (known to be responsible) and Clostridiales (not known to be responsible until the mmvec)

The effects of a high-fat diet in the murinae model

Inflammatory bowel diseases In the last experiment, and the most complex, the one in which the inflammatory disease of the small intestine and colon in humans is analyzed, it is desired to give a clearer note of the influence of the microbiome in the accusation of this disease. It is known that the microbiome influences this disease, but its role is poorly understood. That is why in the initial Human Microbiom Project, until mmvec, only Klebsiella is associated with this disease. Moreover, it is known that this disease usually occurs in people with excess bile acid, and carnitine has anti-inflammatory effects in this disease.

When analyzing samples from the small intestine of people suffering from this condition and proceed in the same way as in the experiment above, two interesting things are discovered. First of all, it is discovered on a theoretical level that a class of Clostridium is also responsible for this condition. Secondly, it is discovered that the bacterium Propionibacterium freudenreichii, part of the intestinal microflora is an extremely beneficial microbe for this disease. These experiments took place on a theoretical level and were then confirmed in the laboratory as valid. Both bacteria are not included among the bacteria related to this disease in the Human Microbiome Project.

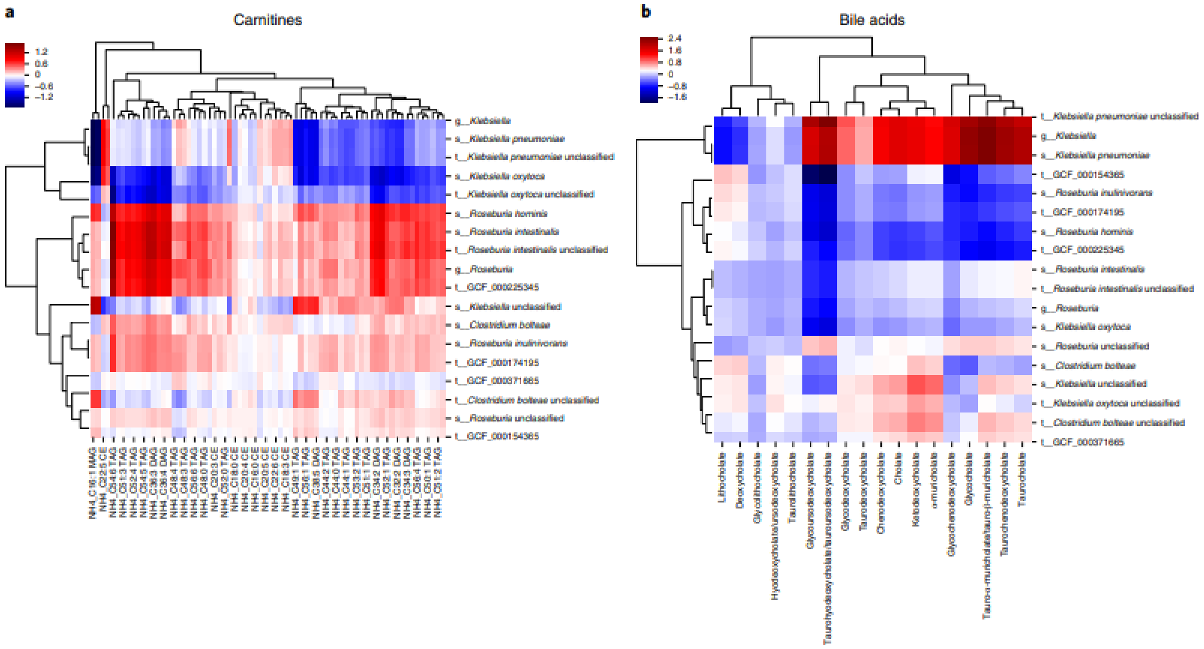

Figure 8: It can be seen that Propionibacterium freudenreichii rich in carnitine (anti-inflammatory effect for IBD) is not responsible for excess bile acid, quite the contrary. And a class of Clostridium (the lower ones) that are low in carnitine and produce excess bile acid inflamed small intestine and colon

Conclusions

We have seen how the current methodology solves essential problems such as the stage of cystic pulmonary fibrosis due to infection, in the first phase only with P. aeruginosa and then with a family of anaerobic bacteria. Then the relationship of the metabolite microbe to the moisture of the biocrust is studied. In all these experiments the performances of the current methodology are validated compared to the existing ones. Finally, there are a number of bacteria responsible for various diseases, such as inflammatory bowel disease and excess production of bile acid, in complex models: mice and humans. In light of these considerations, the current mmvec network still has a number of limitations and constraints. But it is the most powerful tool with which to metabolize microbial relationships at a theoretical level.