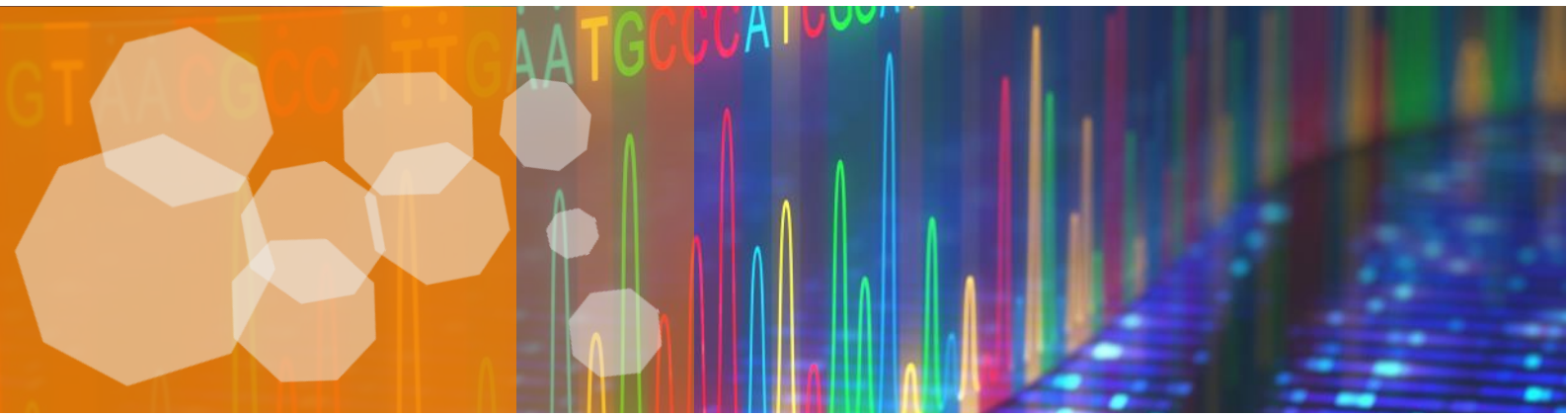


Clustering Approach For The Grouping And Identification Of DNA Sequences From Microbiome Sequences

Project Goal

Although bacteria are generally associated with diseases, there are strains that are crucial for life. The human microbiome includes all microorganisms living inside and on the human body. It plays an important role for the induction and function of the host immune system. Information about changes of the microbiota composition can provide insight to microbial networks and may suggest new options for therapeutic intervention.

Project goal was to develop a pipeline to cluster the colon sequences based on the order of nucleotides and to subsequently match the formed clusters with known sequences by the blastn algorithm. According to the hypothesis that high homogeneity between sequences of the same clusters correspond to the same species, representative sequences were compared with sequences from the NCBI database. The results of the project can be used for the individual therapeutic approach.

Provided Data

The samples collected for microbiome analysis are usually composed of stool and blood samples. Colon samples used for the project were taken during colonoscopy. The colonoscope was inserted and it was pushed forward from the rectum through the entire colon to the caecum. On the way back, using air injection, the intestine was unfold so that the entire mucosa was scanned. The dataset of the colon, obtained from the 16S rRNA sequencing, was analyzed.

Challenges

There was no prior information of contaminations during an experiment. After sequencing step some nucleotides labeled as “N” occurred.

Since not all the clusters were at a high quality after the first run of the algorithm, the second run was necessary.

Localization of the sampling

Applied Methods

Self Organizing Map (SOM)

For the algorithm training all sequences were cut up to 250 nucleotides from the primer start. Taking into consideration that SOM operates only with numerical data, the nucleotides were firstly translated into numbers. The nucleotides A, C, T, G and N were replaced with the numbers 1, 2, 3, 4, 5 respectively. During the training cycle all sequences were distributed 10.000 on the network and the necessary adjustments were carried out. For each cluster a corresponding heatmap was stored. Each nucleotide got a unique color for visualization of cluster homogeneity. Mean of cluster correlation under 0.95 was a crucial value for the second SOM run to identifying the sub clusters of better quality.

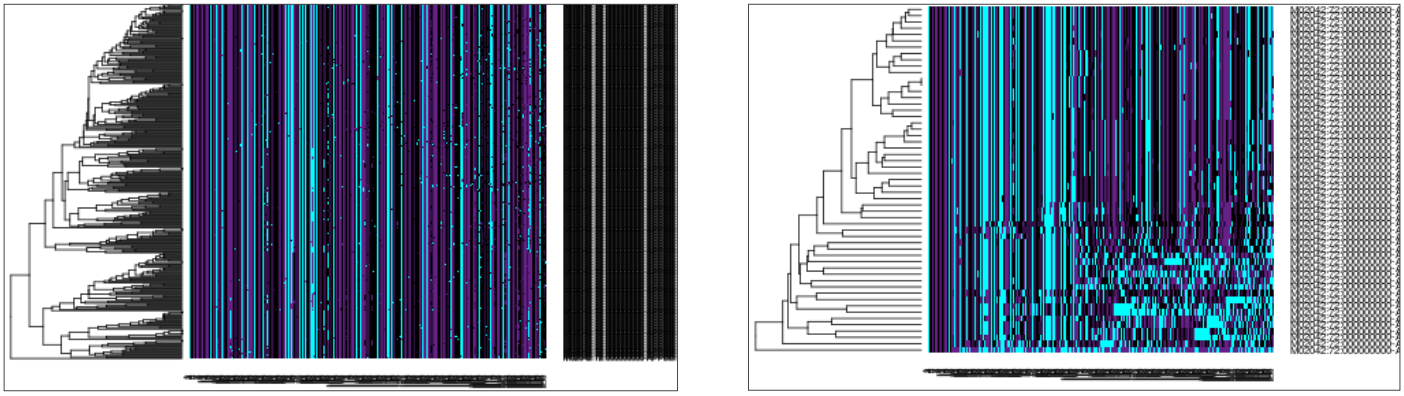
Thus, after two SOM runs each sequence was in an appropriate cluster with average cluster homogeneity above 90%.

Basic Local Alignment Search Tool (BLAST)

The BLAST algorithm was used to define which known microorganism sequences are presented in the detected samples. To minimize the operation time 50 representatives from each of the largest clusters were randomly selected and stored as FASTA file. Finally, for each cluster the most frequent BLAST results were detected.

Project Outcome

The SOMs provided clustering with high homogeneity. Neural networks made it possible to identify similarities in huge amount of DNA sequences based on Euclidean distance. Additionally, according to heatmaps a user can identify which pattern dominates in a respective cluster.



Heatmaps of clusters contained 389 sequences with a mean score of 0.9511 (left) and 57 sequences with a mean of 0.7504053 (right).

According to the BLAST results the identified microorganism sequences of the database allow us to make a conclusion about the content of a person's 16S microbiota from colon.

Cluster	Number of sequences	The most frequent BLAST result
4	9505	Ruminococcus flavefaciens
5	5971	Ruminococcus bromii
6	5566	Bacteroides thetaiotaomicron 8713
7	5195	Bacteroides vulgatus strain JCM 5826
8	4397	Clostridium fusiformis
9	3810	Ruminococcus sp. ID1
10	4782	Coprococcus catus VPI-C6-61