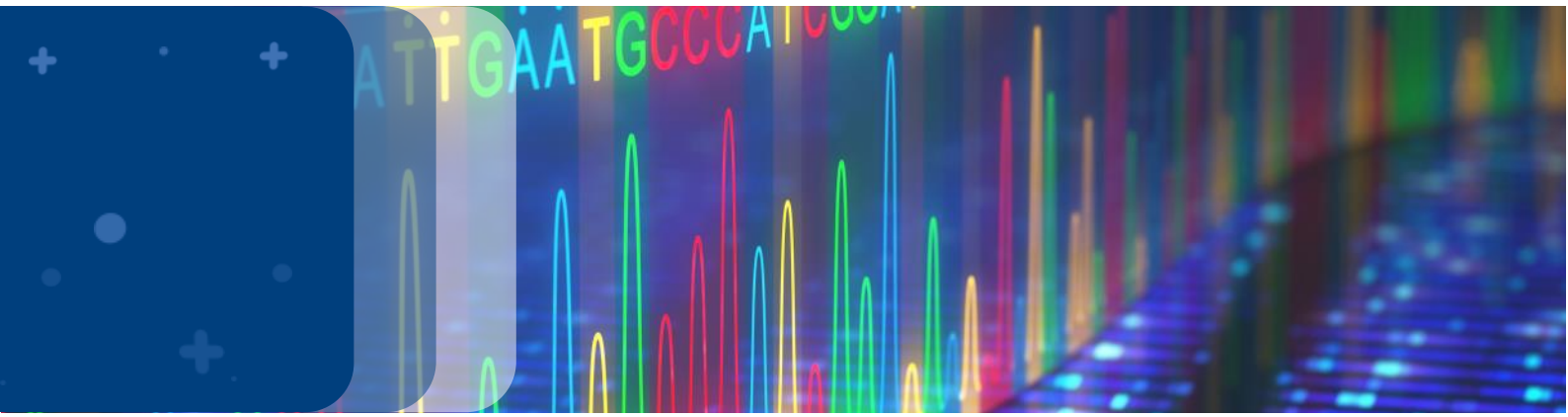
supper&supper



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

Project Goal

Human gut microbiome is an complex and dynamic union of microorganisms which inhabits the gastrointestinal tract. It contributes to several aspects of host physiology, including metabolism, maturation of the immune system, behavior and even the neurology. By observing the changes of the gut microbiota composition can provide insight of host-microbiome interaction and may suggest new options for therapeutic intervention.

Project goal was to develop a pipeline to cluster the microbiome sequences from the colon based on the order of nucleotides and to subsequently match the formed clusters with known sequences by the blast algorithm. According to the hypothesis, the homogeneous sequences can be gathered into the same cluster which might correspond to the same taxonomy. The representative sequences from the cluster were then matched to the known bacteria genus or species in the NCBI database. The results of the project can be further used for the clinical treatment.

Provided Data

The samples were collected from colon during colonoscopy. The colonoscope was inserted and pushed forward from the rectum through the entire colon to the caecum. The air injection was employed on the way back and the intestine was unfold so that the entire mucosa can be scanned. RNA was isolated from colon and the 16S rRNA sequencing was applied to identify the bacteria. The raw sequenced data (.FASTQ) files were provided.

Challenges

First, the mixture of the human genome and bacteria genome increases the difficulty of the microbiota species identification. There was no prior information provided for clustering, so the unsupervised method should be used. Second, the technical error from the machine might occur during sequencing. It might generate useless information (nucleotides labeled as “N”) or even false information of bacteria genome. Third, by the limitation of the technique, the samples can not all run at the same time. Thus, the batch effect should be normalized.

Localization of the sampling

Applied Methods

Self Organizing Map (SOM)

SOM is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional discretized representation of the input. SOM can adjust the batch effect and controlling the false positive error from technical error. All sequences need to be cut up to 250 nucleotides from the primer start for the training. 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 batch effect 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.

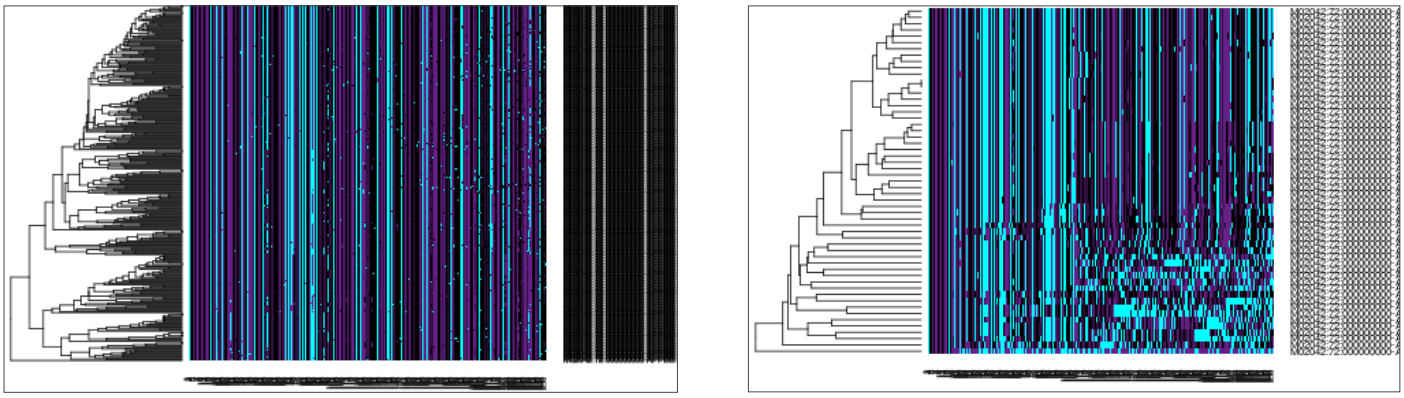
Thus, after two SOM runs each sequence was in an appropriate cluster with average cluster homogeneity above 90%.

Basic Local Alignment Search Tool (BLAST)

The BLAST algorithm was used to identify which known microorganism sequences are presented in the detected samples. It can separate the human genome and bacteria genome. 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 allow us to have deeper understanding of the microbiota composition within the patients. The results can be further used for the investigation of host-microbiota interaction, drug effect and environmental influence.

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