Functional microbial k-mers: Work in progress

Miroshina Aleksandra

25.07.2023

1 Initial Project

Initial project aimed to establish a link between microbial functions and the k-mers identified in the sample. The core of the algorithm comprises three main steps: functional profiling of samples, k-mer counting, and prediction. Specifically, the prediction step involves utilizing an elastic net model, which effectively learns information about functions and their corresponding k-mers, enabling predictions for the test set. The workflow of the pipeline can be logically divided into three steps. First, the training set is prepared and processed. It undergoes functional profiling using HUMAnN tool, followed by k-mer counting with KMC tool. The resulting output from both tools is then adjusted and used as input for training the elastic net model. In the second part of the program, the test set is prepared and processed similarly to the training set, with functional profiling and k-mer counting. However, the model training step is excluded as it is not required for the test set. The third step involves performing predictions using the information obtained from the second step and the model created during the first step. This predictive step allows for an assessment of how well the model generalizes to new data. By dividing the process into these distinct steps and employing the elastic net model, this pipeline aimed to effectively link microbial functions with the identified k-mers, ultimately providing valuable insights into the functional potential of the microbial community in the sample. In this project, we utilized a diverse dataset consisting of 100 metagenomic paired-end fastq samples for the training set. These samples were acquired from various sources. For the test set we collected 40 samples obtained from patients diagnosed with Inflammatory Bowel Disease (IBD), and 20 samples derived from healthy individuals. All of these samples were obtained from publicly available databases, with the primary sources being MGnify and SRA Explorer.

1.1 Problems

Despite its promising potential, the results of the project did not meet our expectations. Several key challenges emerged during the analysis: Firstly, heterogeneity in the training samples posed a significant hurdle. Secondly, the consideration of an excessive number of k-mers overwhelmed the model's ability to effectively capture relevant functional associations. This led to difficulties in identifying meaningful correlations between k-mers and microbial functions. Thirdly, the vast diversity of functions analyzed also presented a challenge. Moreover, one critical limitation was the lack of taxonomic context in the analysis. By not considering the taxonomic composition of the microbial communities, the model may have missed crucial links between specific microbial taxa and their associated functions.

2 First alternative approach

To enhance the pipeline's performance, we implemented several key improvements as follows:

- reducing heterogeneity in the training data by carefully curating a consistent set of 100 samples (only IBD samples) from Human Microbiome Project 2

- narrow our focus to pathways directly related to human-microbial interactions

- collect relevant gene sequences associated with the chosen pathways, find conserved region across all genes by performing multiple sequence alignment and efficiently split these regions into k-mers

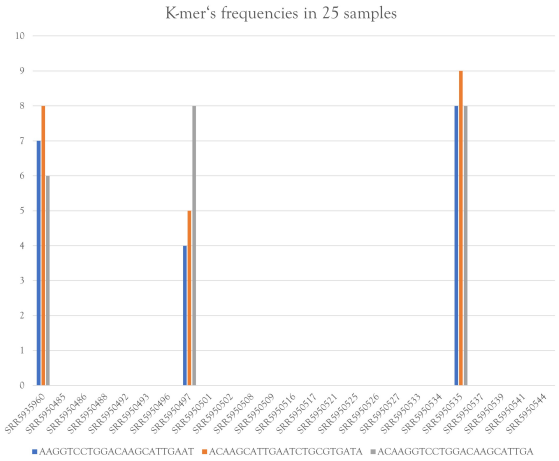

Figure 1: *K-mer distribution across 25 samples for randomly chosen 3 k-mers. X-axis contains sample ids and y-axis contains k-mer's frequencies*

After an extensive literature review, we selected a set of metabolites and their corresponding pathways for detailed examination. The chosen pathways encompassed secondary bile acid biosynthesis, nucleotide, and amino acid biosynthesis, as well as fatty acid biosynthesis. To access the relevant gene sequences essential for our analysis, we leveraged the comprehensive resources available in the KEGG database.

Unfortunately, new results proved to be once again unreliable. The functions identified by the HUMAnN functional profiling tool, which exhibited high prediction scores in our pipeline, did not show any association with the functions mentioned earlier. Furthermore, the k-mers derived from the genetic sequences were noticeably underrepresented in the actual samples obtained from the HMP (see figure 1).

3 Second Approach

In our pursuit of enhancing the results, we adopted a more focused approach by selecting in on specific, highly abundant functions identified by the HUMAnN tool across the 100 samples. The intention behind this was to pinpoint the k-mers that are more likely to be prevalent in the real samples. By narrowing our attention to these specific functions, we aimed to increase the reliability and relevance of our findings.

Consequently, our efforts yielded an improved k-mer representation across the samples (see figure 2). However, the prediction results we unreliable. For instance,despite achieving high prediction scores for certain pathways in our pipeline, we found once again that these pathways showed no connection with amino acid biosynthesis.

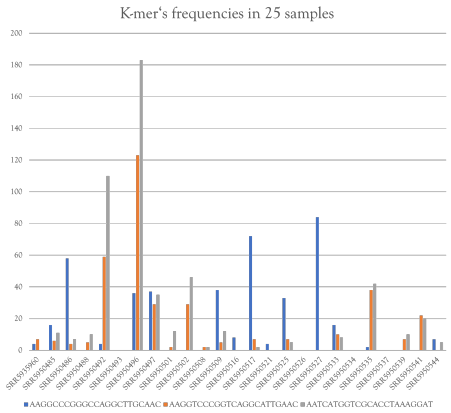

Figure 2: *K-mer distribution across 25 samples for randomly chosen 3 k-mers. X-axis contains sample ids and y-axis contains k-mer's frequencies*

4 Third Approach

In addition to the second approach, we decided to improve our search for meaningful regions across gene sequences by using the expectation maximization algorithm for motif search (STREME). This algorithm improves the runtime of our pipeline, but the tool produces motifs as output that represent the distribution of appropriate sequences which increases the number of possible k-mers and reduces precision, which is important for further applications of our algorithm.

5 Fourth Approach

Previous results also showed lack of conservation across sequences, that is why we have decided to perform conservation analysis from the protein side. We performed multiple sequence alignment of the corresponding protein sequences and searched for the conserved regions. Unfortunately, this approach caused the library size problems as we expected. For instance, for a amino acid sequence $FIPGDGIG$ the corresponding regex representing possible nucleotide sequence would look like this:$(TTT\|TTC)(ATT\|ATC\|ATA)(CCT\|CCC\|CCA\|CCG)(GGT\|GGC\|GGA\|GGG)(GAT\|GAC)(GGT\|GGC\|GGA\|GGG)(ATT\|ATC\|ATA)(GGT\|GGC\|GGA\|GGG)$. From such sequence one could produce 9216 possible nucleotide sequences. The problems arises due to redundancy in the amino acid code and the conservation across amino acid sequences does not necessarily mean conservation on the DNA level.

6 Fifth Approach

In order to make the analysis more precise we made two decisions. Firstly, we incorporated taxonomic information into the approach. Secondly, we explored another highly abundant pathway, specifically, dTDP-L-rhamnose biosynthesis I, which is highly abundant in 99 out of 100 samples.

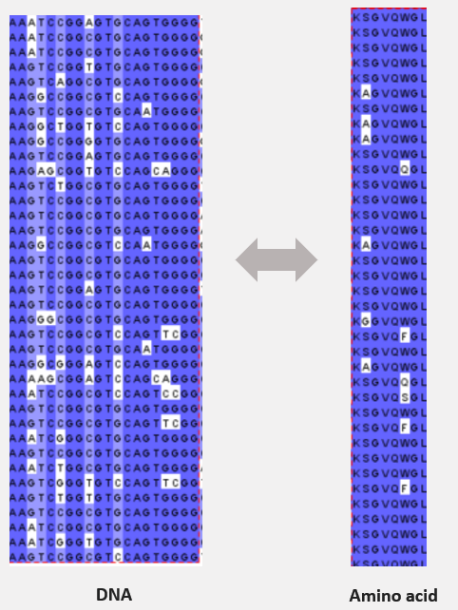

Figure 3: *The figure shows two snapshots of the same region in nucleotide and amino acid sequences. The right side correspond to protein sequence and shows that region is highly conserved, whereas the left side is less conserved, which is caused by redundancy of the amino acid code.*

To accomplish this, we leveraged the KEGG database to gather gene sequences for various enzymes, categorized according to their corresponding genera. To ensure robust conservation analysis, we required a sufficient number of sequences, leading us to select only those genera that possessed at least 15 sequences for the respective genes. Consequently, the number of remaining genera for the target pathway reduced to 29. Unfortunately, the majority of genera exhibited significant dissimilarities in terms of gene sequences, and those that displayed some conserved regions only demonstrated high conservation at the protein level while remaining too variable at the nucleotide level (see figure 3).

7 Current Approach

In our latest approach, we adopted a comprehensive strategy by exploring a wide array of functions within the taxonomic context to identify conservative sequences. Furthermore, we conducted BLAST analyses using these conserved sequences to assess their matches and eliminated those that showed matches with more than one genus. This approach yielded promising results, revealing multiple conservative regions spanning various genera and pathways. To facilitate further analysis, we established a comprehensive database comprising 10 substantial pathways extracted from the KEGG database, encompassing over 500 enzymes. This database includes vital information such as the enzyme name, the genus associated with the conserved k-mers, the number of counts within the corresponding human gut library, and the percentage indicating the degree of matching between the k-mer and the genus. To verify the correlation between the identified k-mers and the genus abundances in the actual samples from the Human Microbiome Project, we conducted a correlation test. The results of the correlation test demonstrated a strong positive correlation between each k-mer and its respective genus. Furthermore, during the abundance analysis, we observed that the k-mers of bacteria prevalent in nearly every human gut were successfully detected by our pipeline in the real samples. Finally, we identified non-overlapping k-mers specific to each genus for every enzyme, resulting in the creation of distinct k-mer pairs for subsequent analysis and experimental investigations. As the next step, we intend to assess the predictive capacity of these k-mers through a series of experiments involving cultured bacteria. These experimental validations will provide valuable insights into the functional relevance of the identified k-mers and their potential applications in real-world scenarios.