

Master's degree in computer science

Existing Datasets Analysis And Report

Methodology

I was asked to perform a research on all the public benchmarks with OTU (operational taxonomic unit), OUT are used to classify groups of closely related individuals, with phenotypes (individuals observable traits that is to say height, eye color, blood type, etc.

For this several tools were used which I’ll address in each of their specific sections and I’ll display the datasets accordingly as well as a brief summary of the outcome of said research.

Kaggle

As a well-known website were it can be found datasets, Kaggle was used first. Keywords such as “OTU”, “OTU DATA”, “PHENOTYPE”, “OTU BENCHMARK” were employed and the following results were obtained:

* OTU:
  + <https://www.kaggle.com/search?q=OTU> Results were obtained but they all pertrains to a dataset regarding Ovarian Tumor (that uses a dataset containing images) and thus doesn’t have relevance for this point
* OTU BENCHMARK:
  + <https://www.kaggle.com/code/eteims/a-dictionary-is-all-i-needed> this dataset was found but this dataset has nothing to do with OTU data and thus should not be taken into consideration
* Operational Taxonomic Unit
  + <https://www.kaggle.com/code/hpbarr/covid-19-literature-report-genet-origin-evol> dataset regarding covid but doesn’t have the wanted data…
* OTU Data
  + <https://www.kaggle.com/code/antaresnyc/prediction-with-random-forest-and-xgboost> explores the explores the relationship between gut microbiota composition and metabolic activity in individuals with ASD. The dataset can be found at <https://www.kaggle.com/datasets/antaresnyc/human-gut-microbiome-with-asd?utm_source=chatgpt.com> and has OTU data, Taxonomic classification and sample metadata
* Phenotype
  + Nothing relevant to the topic.

Github

Via searching “OTU DATA” on github I got a few github repositories that I’ll be further discussing just ahead.

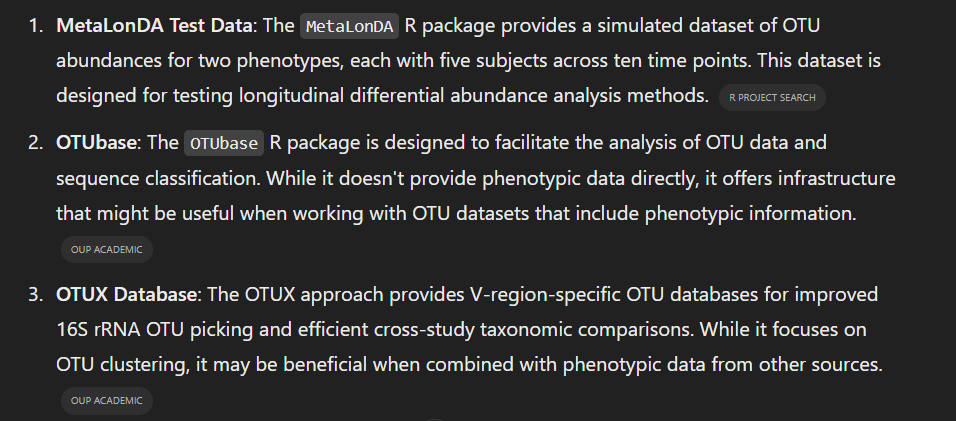
<https://github.com/tbj128/mian> is an open-source web platform designed for the visualization, analysis, and feature selection of 16S rRNA Operational Taxonomic Unit (OTU) tables.

<https://github.com/divya031090/taxoNN_OTU> contains an ensemble learning method that trains CNNs on OTU data stratified by phyla. This deep learning approach organizes OTU data for CNN modeling based on similarities at the phylum level of the taxonomy tree.

ChatGPT

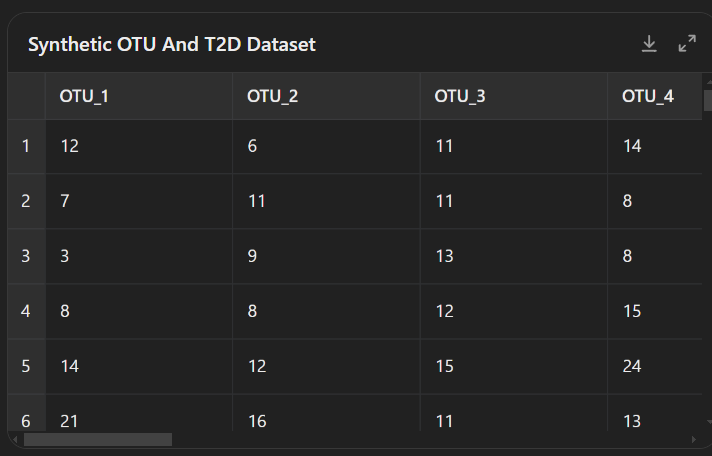
I asked ChatGPT for public benchmark OTU datasets with phenotypes and this was the response:

* <https://search.r-project.org/CRAN/refmans/MetaLonDA/html/00Index.html>
* <https://academic.oup.com/bioinformatics/article/27/12/1700/256371?utm_source=chatgpt.com&login=false> After clicking this link to try to see the package, I got the Page Not Found Error
* <https://academic.oup.com/dnaresearch/article/26/2/147/5280828?utm_source=chatgpt.com&login=false>

****

Another Approach

Seeing as data wasn’t accessible I asked ChatGPT to actually create, synthetically, a dataset with 100 entries (for testing) with OTU data and phenotype related to a certain disease, in this case T2D (Type 2 Diabetes). Surprisingly I got an output…



Results

As a summary I must say there is not much data, if anything at all about this topic. The existing data is either deprecated, inaccessible or just restricted by one or other reasons.

As for the alternative response, I am not sure yet what results the data would give me but no matter what, synthetic data can never be trusted for real world applications and thus can only be a base mark or a starting line for a real approach with real data.