Statistical Analysis of Mouse Gut Microbiota
Alex Tran
Computer Systems Lab 2009-2010

Abstract

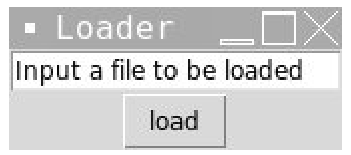
The mouse gut microbiotal community is the population of bacteria, which inhabit the digestive track of a given population of mice. The ability to understand the microbiotal community has implications, which extend beyond mice and into factors of human obesity and the dietary needs of the human body. Computational genomics is an emerging field, which blends both biology and computer science, to analyze genes. This study seeks to utilize several methods emerging within computational genomics to analyze the gut microbiome of mice to observe the effects of varying diets on the gut microbiotal community, as well as create a universal and user friendly tool for researchers to utilize when studying any gut microbiome. The applications of this can extend past studying the gut microbiome, but can also be applied to any taxonomic group being counted for analysis, however several parts of the application are exclusively beneficial to the analysis of gut microbiota specifically those found in mice.

Background

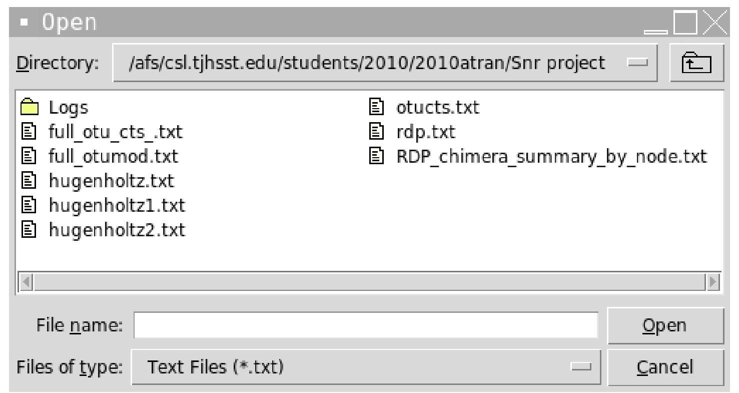
Though General imbalances have been observed in the gut microbiome of obese mice, there is still not enough information or analysis to say pinpoint what specific species of bacteria can lead to the tendency for mice to be obese.(Turnbaugh, Gordon) This lack of information is largely due to the inefficiency and inaccurate nature of current computational 16s sequencing techniques (Turnbaugh et. al.)

To combat these inefficiencies a recent study attempted to use computational genomics to extrapolate microRNA from a "model organism" such as the dorosophilia genus of flies and predict the structure of certain genes in other organisms of similar structure and species. To do this an algorithm was developed that would take microRNA archive it, and use it to predict the structure of certain marker genes that should be present in the following iteration. The hope is that if such an algorithm that can reliably predict the structure of genes given certain restraints could be created it would save a large sum of money on expensive DNA testing, and vastly improve the speed at which research can be done. (He Tao, Li Fei, Gu Jin, Li Ruiqiang, Li Fei)

Too awkward



Becomes



Directory: /afs/csl.tjhsst.edu/students/2010/2010atran/Snr project

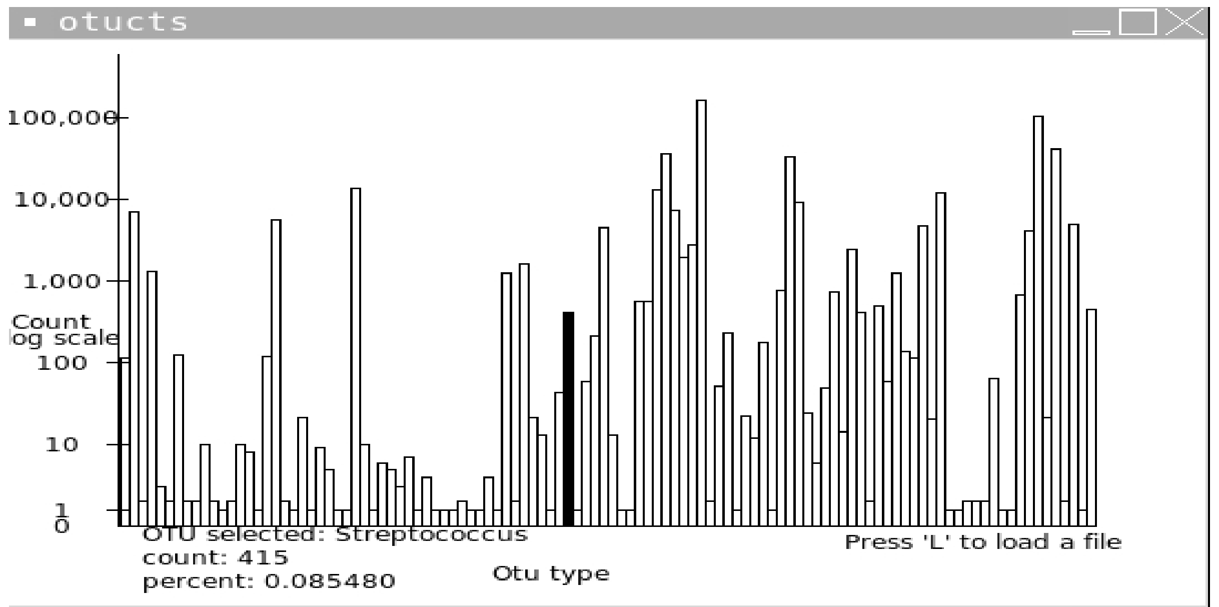
```
Logs                      otucts.txt
full_otu_cts_.txt         rdp.txt
full_otumod.txt           RDP_chimera_summary_by_node.txt
hugenholtz.txt
hugenholtz1.txt
hugenholtz2.txt
```

File name:

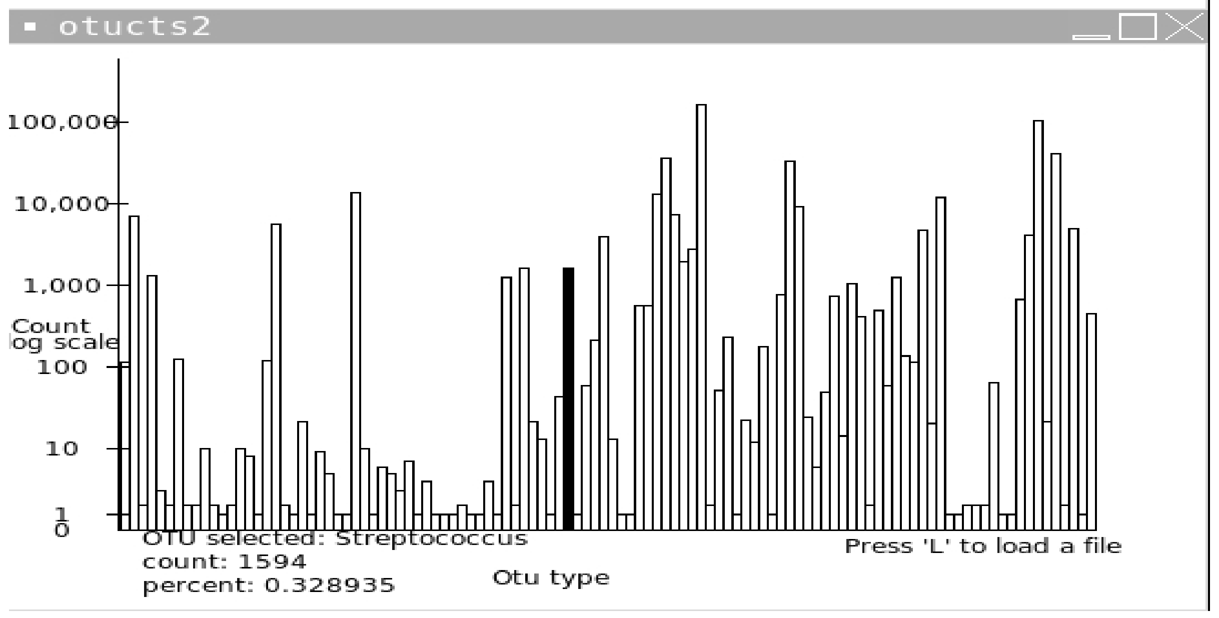
Files of type: Text Files (*.txt)

Development

The program's end result consisted of several parts, which could be created individually in any order. Due to possible problems with time constraints it was been written in reverse depth from the most superficial part to the most abstract. This will allow for user friendly verification as well as a gauranteed presentable result if time had become a problem. It was primarily programmed in Python. Coding began by generating a simple taxonomic tree (originally defaulting to the RDP taxanomic tree,) and filling it with data points. It then moved onto generating a histogram of these data using Tkinter as a base. After this simple histogram was generated specific file loading, and a variety of taxonomic trees were made available to the user to allow the ability of the application to extend past only gut microbiota in mice to any accepted use of taxonomic OTU counting. As development time went on there become a larger focus upon user friendliness resulting in changes to aid with user friendliness as seen above.



```
OTU selected: Streptococcus
count: 415
percent: 0.085480
```



```
OTU selected: Streptococcus
count: 1594
percent: 0.328935
```

Results

From the start the goal of this project was to present data from the microbial community in a way that is user friendly, efficient, and able to represent the data in an accurate way. To achieve this a gui was utilized that allows the display of statistical information on each run in a way that allows the user to more easily gather information from a sample then has ever before been possible. As a result of this analysis of data gathered from the gut microbiome can be done vastly faster than other archaic forms of analysis previously used. It cuts down on the amount of fact checking, and manual calculation that was previously necessary when doing analysis of this nature.

It is also of note that while many of these tools are available in expensive suites most of these require an in depth knowledge of both the field of genetics as well as a firm grasp on scripting in PERL or another comparable programming language, or are programmed to handle a very specific set of data. The goal of this project was to step past these former suites and provide a simple universal tool which could be used across the field or even apply to other fields related only in the need for adherence to a taxanomic tree. The goal was for the program to apply to any situation where counting of a taxanomic group under a given set of criteria is possible, which could vary from the previously stated examining of gut microbiota in mouse fecal matter to the number of a specific species of lemur in Madagascar. The program was therefore refined for gut microbiota, but given the freedom to adapt to many situations.

The program has only served as a functional success largely due to the lack of a professional and user friendly suite similar to the one that was designed, which would likely contain a more complete set of analysis tools. Beyond the use of the chi squared signicance test the tool does not implement many sophisticated statistical analyses, and a future toolset would probably provide more in depth statistical information such as the p-value from such a test, or more options such as the F test statistic. Many features which a similar future program might oer include a more from the ground up approach, which could not be completed due to a lack of access to the raw data by a claim from the company from which the Gordon Lab receives its analysis. Computational genomics is still in its infancy, and the tools which this programs provides, while not complete serve more as a proof of concept that temporarily fill the void where a lack of user friendly software currently exists