TJHSST Computer Systems Lab Senior Research Project
Statistical Analysis of Mouse Gut Microbiota
2009-2010

Alex Tran

June 10, 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, and can also be applied to any taxonomic group being counted for analysis, however several tools found in the software are exclusively beneficial to the analysis of gut microbiota specifically those found in mice.

Keywords: Computational Genomics, Genomics, Gut Microbiota, Statistics, Microbiota

1 Introduction

1.1 Scope of Study

To use genomics to observe and document the gut microbiotal community in the fecal matter of mice during all stages of analysis. From the sequencing of genetic material to the processing of the microbial community. To create an all encompassing suite to streamline the analysis of any microbiotal and community.

1.2 Expected results

To be able 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. Implemented Gui utilization that will display statistical information on each run, which can be used as visuals in a report, and to be able to do a full run from the ground up given 16sRNA sequencing data that can accurately and efficiently identify individual microbiota within the commmunity. The end result of which will be an all comprehensive applictaion that will be able to take generated data, and give the user a full evaluation of these data.

2 Literature Review

Other research has observed that general imbalances inside the gut microbiome can lead to differences in tendencies towards Obesity in mice(classified as OB/OB)(Turnbaugh, Gordon).It has also observed that within various populations of mice the mirobiome appears to be so diverse that there is little known correlation between an individual and the population of their microbiome, however it was also shown that though the population in the microbiome is diverse there are tendencies based on the diet of an individual mouse for an increased representation of certain strains of bacteria to be present within the microbiotal community (Turnbaugh, et. al.)

Though General imbalances have been observed in the gut microbiome of obese micethere 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)

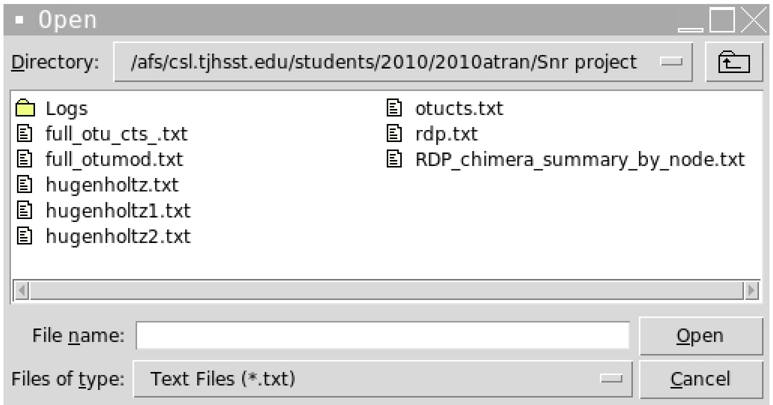
At this point in time the development of computational tools to identify regulatory networks is far behind the tools, which handle sequencing and gene prediction. At this time there are 3 standout methods for identifying microbiotal transcription and regulatory networks in genes. The first method is the prediction of TF (transcription factor) promoters in genes which mark the Transcription Start Site (TSS,) and begins transcription. This is at best an inexact science and we do not have enough knowledge of the regulatory network to utilize this method accurately. The next is the CpG islands and promoter prediction, which uses the location of CpG pairs in DNA as a predicted point of the start of transcription, because CpG pairs are so uncommon (less than 6 percent) of pairs this method may prove to be correct, however this is a trial and error process and inexact to say the least. The final method is the transcript mapping approach, which uses gene identificaiton software to eliminate known genes and combs the remaining data to cross reference known genomes to attempt to identify TF (transcription factors.) (Ping Qiu)

3 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. Programming then moved on to allow for mouse movement to allow for more user friendliness, and to aid between two varying samples.

Following these simple processes had been implemented a system, which maintained the freedom to use any taxanomic tree desired, but also allowed the program to default to the most commonly used tree, and for the user to reopen recently used files by utilizing the cPickle archiving system for python. To do this it became necessary to create a hashmap, which would keep track of the number of times each tree was used, which the program would default to if the user did not specify a taxanomic tree, as well as allow the user to call up the 10 most recently used files. The cPickle archiving system is not dissimilar to the way in which Microsoft Office's archiving system works in the way it uses a shared cache file between various instances of the program that is read when the program is opened, and editted upon closing the program. To test and verify results several premade computational genomics tools that each perform a part of what the finalized project does as one universal tool were used.

The next step was to include tools for the comparison of a number of experiments rather than just the simple viewing of a single microbiotal community. To accomplish this a driver/slave system was implemented that allowed for a simple wrapper program to run and share information across multiple independent files, which could be opened and closed at will. This required the creation of a Listener in Tkinter, which could listen for changes done in one file, which should be mirrored across all files. This is not a built in tool in Python and Tkinter, and the streamlining of this functionality to run memory efficiently. The end result of this allowed for the implementation of a Chi Square test of significance to be performed across a number of files.



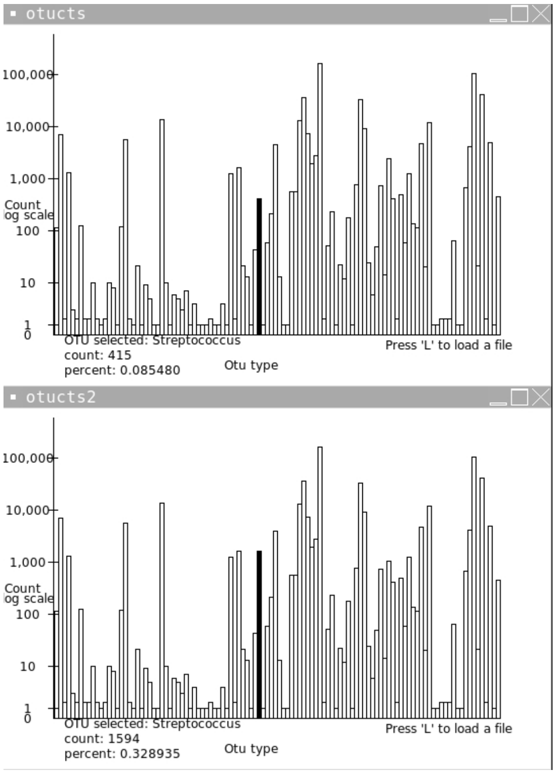
Directory: /afs/csl.tjhsst.edu/students/2010/2010atran/Snr project

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

File name:

Files of type: Text Files (*.txt)

Figure 1: Screenshot of the file browser set to find .txt files



OTU selected: Streptococcus
count: 415
percent: 0.085480

OTU selected: Streptococcus
count: 1594
percent: 0.328935

Figure 2: Screenshot of two loaded files, with a data point selected

4 Quality assessment

Input data and 16sRNA sequences were received from the Gordon Lab at Washington University in St. Louis. The main goal of this project was to provide a user friendly OTU analysis suite, which could be used and interpreted by anyone familiar with genetics, but not necessarily with computer programming or use. To achieve this goal a large portion of development time and assessment was done in the minor debugging stage to present the user with the friendliest interface possible. To input a file the user is presented with the self explanatory file browser pictured in Figure.1 When a user inputs a file they are presented with the screen seen in Figure.2.

5 Results

From the outset the goal for the goal of this project was to be able to present data from a 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.

The results from this study are potentially very helpful to the field of genomics, because tools for analysis of gut microbiotal communities are a field that is in its early stages of development, and at this point most tools which have been implementedd into the final application are only available separately, and are typically very expensive.

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.

6 Discussion

The program in its early stages has proven to be useful in the Gordon Lab for which it was developed, and has also found use outside of the lab using taxanomic trees which were not tested during its development, which is a good sign for the universality of the program, however much could be done to improve upon it in its current state. 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 significance 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 offer 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.

7 Bibliography

He Tao, Li Fei, Gu Jin, Li Ruiqiang, Li Fei, "Computational Identification of 99 Insect MicroRNAs Using Comparative Genomics." Beijing Genomics Institute, Chienese Academy of Sciences, Beijing China, 2008.

Peter J. Turnbaugh and Jeffrey I. Gordon, "The core gut microbiome, energy balance, and obesity.” Center for Genome Sciences, Washington University School of Medicine, St. Louis MO, 2009.

Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight Jeffrey I. Gordon, ”The Human Microbiome Project.” Center for Genome Sciences, Washington University School of Medicine, St. Louis MO, 2009.

Ping Qiu, ”Computational Approaches For Deciphering The Transcriptional Regulatory Network by Promoter Analysis.” Bioinformatics Group and Discovery Technology Department at Schering-Plough Research Institute, Kenilworth NJ, 2003