A Framework for Analysis of Metagenomic Sequencing Data

A. Murat Eren, Michael Ferris, and Christopher Taylor

Department of Computer Science, University of New Orleans, New Orleans, LA 70148 & Research Institute for Children, Children’s Hospital, New Orleans, LA 70118

Abstract

The number of microbial cells in the human body is an order of magnitude greater than the number of human cells that make up the body itself. Changes in the ecology of the human microbiota are highly associated with intestinal and respiratory disorders and diseases of the skin and mucus membranes. Recent advances in high throughput pyrosequencing have made monitoring these changes feasible by targeting microbial genes that are phylogenetically informative, such as the 16S rRNA gene. Although a variety of analysis methods are available to assess the composition and diversity of microbial communities, applying these methods to millions of sequences and visualizing the results is a cumbersome task. Here we introduce a new, easy-to-use, extensible visualization and analysis software framework that facilitates the interpretation of large amounts of metagenomic sequencing data. At the present stage of development, the framework automatically performs an array of standard analyses using FASTA files that contain 16S rRNA sequences as input. It classifies the sequences at each taxonomic level ranging from phylum to genus, displays pie chart representations of the percent abundance of operational taxonomic units (OTUs) at each level, performs rarefaction analysis to assess the adequacy of sampling, and calculates diversity indices. The framework also assesses similarities between the microbial communities based on sequence composition and relative abundance by performing complete linkage clustering analysis. Clustering results are displayed as dendrograms along with heatmaps. The framework has been used to discover associations between changes in the microbiota and diseases such as bacterial vaginosis and necrotizing enterocolitis.

Analysis

Analysis is started by submitting a FASTA formatted 16S rRNA sequence file. Once the analysis is complete, the framework automatically generates a dynamic web page for the study which contains various information about the samples in the submitted library. The image below shows a screenshot of the information page for an example analysis on the 16S rRNA molecule of Escherichia coli.

Sample Maps

Sample maps are subsets of the library originally analyzed. Once a sample map is defined, heatmaps along with dendrograms based on the percent abundance of the OTUs in included samples are created in five taxonomic levels. Further refinements are possible on automatically generated heatmaps to observe how changes affect the clustering of samples. The figure below illustrates the basic workflow with sample maps.

Summary & Acknowledgements

There are numerous software packages for interpretation of metagenomic sequencing data in order to assess the composition and diversity of microbial communities. While some software packages are available for download (such as ARB and MOTHUR), others expect researchers to upload their data to a web server and analyze their data in an iterative manner (such as the RDP Pipeline). Both approaches have their pitfalls.

Downloadable applications do not limit the amount of data that can be analyzed and allow researchers to use their own computational resources, but they lack easy-to-use interfaces and therefore expect researchers to learn and use a command line interface. Web based applications provide somewhat easier to use interfaces but limit the amount of sequence data that can be analyzed due to the fact that they serve many people and physical resources and computational time are scarce.

Our framework allows researchers to utilize their own computational resources to analyze metagenomic sequencing data with an easy-to-use web based interface.

This work was supported by funding from the Research Institute for Children. We thank Marcela Zozaya-Hinchliffe, Johana Norori, Lynn Tran, and Duna Penn for sample sequences used in these studies and helpful suggestions throughout the development of this work.

A. Murat Eren, Michael Ferris, and Christopher Taylor

Department of Computer Science, University of New Orleans, New Orleans, LA 70148 & Research Institute for Children, Children’s Hospital, New Orleans, LA 70118

This framework is implemented and tested on Pardus 2008.2 Canis aureus with Python 2.5.4, Django 1.1.1, SciPy 0.6.0, Matplotlib 1.0.5, RPy 1.2.1 (R 2.9.0). Thanks to all open source software developers for their efforts and generosity.