A FRAMEWORK FOR ANALYSIS OF METAGENOMIC SEQUENCING DATA

A. MURAT EREN
Department of Computer Science, University of New Orleans, 2000 Lakeshore Drive,
New Orleans, LA 70148, USA
Email: aeren@uno.edu

MICHAEL J. FERRIS
Departments of Pediatrics and Microbiology Immunology and Parasitology, Louisiana State University Health Sciences Center
New Orleans, LA 70112, USA
Email: mferris@chnola-research.org

CHRISTOPHER M. TAYLOR
Department of Computer Science, University of New Orleans, 2000 Lakeshore Drive,
New Orleans, LA 70148, USA
Email: taylor@cs.uno.edu

The human body is home to a diverse assemblage of microbial species. In fact, the number of microbial cells in each person is an order of magnitude greater than the number of cells that make up the body itself. Changes in the composition and relative abundance of these microbial species are highly associated with intestinal and respiratory disorders and diseases of the skin and mucus membranes. While cultivation-independent methods employing PCR-amplification, cloning and sequence analysis of 16S rRNA or other phylogenetically informative genes have made it possible to assess the composition of microbial species in natural environments, until recently this approach has been too time consuming and expensive for routine use. Advances in high throughput pyrosequencing have largely eliminated these obstacles, reducing cost and increasing sequencing capacity by orders of magnitude. In fact, although numerous arithmetic and statistical measurements are available to assess the composition and diversity of microbial communities, the limiting factor has become applying these analyses to millions of sequences and visualizing the results. We introduce a new, easy-to-use, extensible visualization and analysis software framework that facilitates the manipulation and interpretation of large amounts of metagenomic sequence data. The framework automatically performs an array of standard metagenomic analyses using FASTA files that contain 16S rRNA sequences as input. The framework has been used to reveal differences between the composition of the microbiota in healthy individuals and individuals with diseases such as bacterial vaginosis and necrotizing enterocolitis.

1. Background

Understanding the composition of microbial communities is important since microbes drive global nutrient cycles and there is a significant correlation between human microbial community composition, health and disease [1, 2]. Although they are not visible to the naked eye, microbes are ubiquitous in nature. Microbial cells constitute a large portion of the Earth's biomass [3] and the human body is colonized by bacteria in the gastrointestinal tract, oral cavity, skin, airway passages and urogenital system [4]. The 16S rRNA gene sequence has been widely used to detect bacterial species in natural specimens and to establish phylogenetic relationships among them. All bacteria possess this gene, which has highly conserved regions that are needed to construct phylogenies and are useful targets for PCR amplification and pyrosequencing analyses of microbial communities. The 16S rRNA gene also has hypervariable regions that are diverse enough to identify individual species [5]. Because of the large amount of sequence information associated with PCR amplification and pyrosequencing of 16S rRNA genes from microbial communities, a variety of statistical methods and extensive computational aid is needed for the analysis of the data. The primary goal of our work is to bring the analysis of large amounts of microbial community sequence data within reach of scientists who have only basic computer skills.

2. Framework

There are several computational methods available to process microbial community 16S rRNA gene sequence data in order to understand and compare bacterial populations within them. Most of these were not designed to manipulate large pyrosequencing files. Preparing individual scripts in order to manipulate large sequencing files for each analysis is a difficult solution that requires extensive programming skills and experience to maintain. We present a software framework that overcomes many of these challenges of metagenomic sequencing data analysis and provides researchers with an easy way to analyze and interpret their data.

2.1. *Motivation*

Software packages that are available to researchers to process 16S rRNA gene sequence data can be divided into two groups: those that are hosted on a server and used via web interfaces, and those that are downloaded and run locally. Both approaches have their benefits and their limitations. Online ribosomal sequence analysis applications and pipelines, such as Microbial Community Analysis (MiCA) [6] and the Ribosomal Database Project (RDP) pipeline [7], require researchers to upload their data over the Internet and work using web interfaces that are designed to be easy to operate. However online analyses usually have stringent limitations on the number of sequences that can be analyzed (or number of runs or permutations), primarily due to the fact that scarce resources, such as CPU time, memory size and network bandwidth, must be shared by many researchers in any centralized approach. Another limitation of this approach is that the software cannot be customized and enhanced for specialized analysis since it is running on another group’s server. On the other hand software that can be downloaded and run locally such as MOTHUR [8] and QIIME [9], permits researchers to use their own computational resources without requiring them to upload their data to another server. However, since most of these applications necessitate the use of command line interfaces to perform function calls, the learning curve for these tools is steep and a significant investment of time is required to learn and operate them.

Another aspect of available 16S rRNA analysis software that limits its utility is the “pipeline” approach. Pipeline approaches are a model of computing where a set of applications are connected to each other such that output from one application becomes input to one or more applications in the subsequent stage. A pipeline approach is not an efficient structure for an application that is designed to analyze sequencing data. Applications in a pipeline cannot use previous applications’ resources; these resources may need to be re-allocated or re-computed at every stage of the pipeline. This redundancy is not efficient use of computational resources and negatively impacts overall performance. In addition, the process of file upload, analysis and download, which may be repeated at different stages, is time consuming since the user must wait for output and must often upload results again for the next stage of analysis. Lastly, the preponderance of intermediate results from different stages of the pipeline that the user must manage is a large burden that can easily lead to mistakes due to human error.

Our goal is to design an extensible, easy-to-use software framework that is liberated from these issues as much as possible by offering a hybrid solution. During its development, our framework has been tested and used by microbial community researchers studying the microbiota associated with various diseases such as bacterial vaginosis and necrotizing enterocolitis. Researchers using the framework were empowered to analyze their own samples, test hypotheses, and produce publication quality figures in order to communicate their results.

2.2. *Technical Features*

The framework is developed on the Pardus Linux distribution using the Python programming language and open source scientific computing tools and libraries such as SciPy (http://scipy.org) and matplotlib (http://matplotlib.sourceforge.net/). A reliance on open source development tools and libraries will allow us to easily extend the framework and make it portable to non-Linux-based environments.

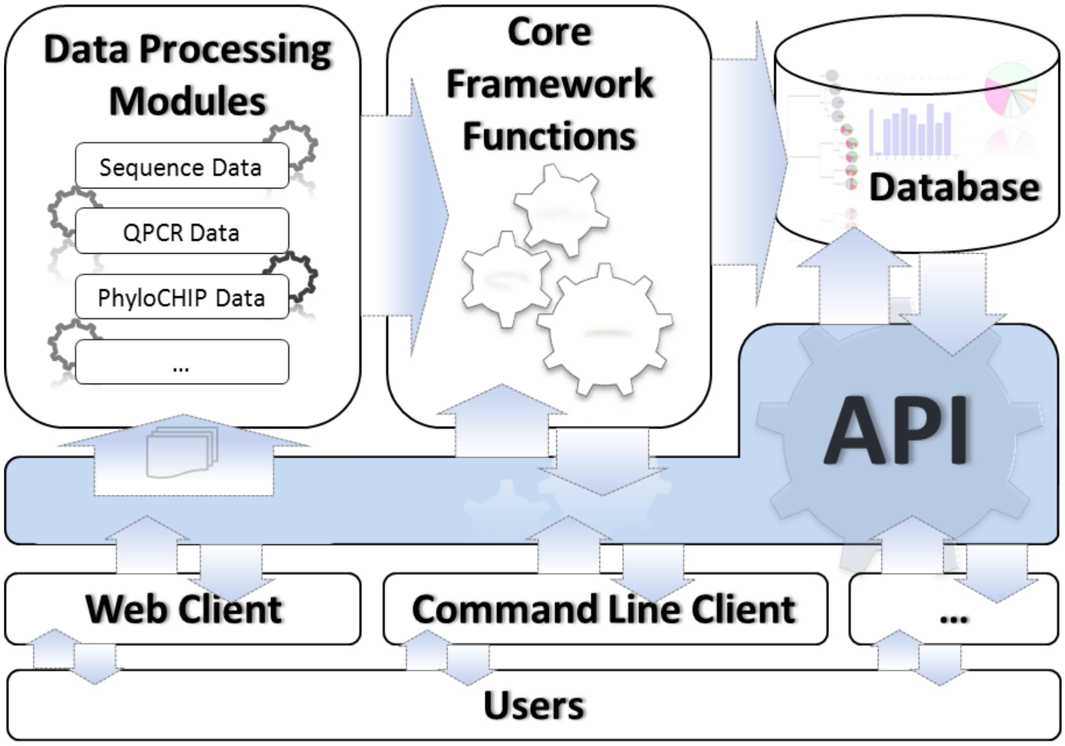

Figure 1: Architectural overview of the framework.

Figure 1 shows an architectural overview of the framework with two major components: A multi-threaded server application that runs in the background performing data processing and core framework functions and interfaces for users to interact with the server.

2.2.1. *Server*

The server performs all manner of computational tasks and figure generation. The multi-threaded design of the server allows it to run multiple analyses concurrently and handle queries simultaneously. The server exposes its functions via an application programming interface (API). This makes it possible for different types of clients to be written and interact with the server seamlessly (Figure 1). This flexibility also allows our framework to be used in both the graphical, user-friendly manner or invoked by scripts for automated analysis of large numbers of data sets.

The server has more than one data processing module, and a set of core functions that is separated from the data. This modularity allows us to extend the server's core functions and analysis capabilities to different types of inputs, such as quantitative PCR data.

2.2.2. *Client*

Any client that can communicate via UNIX domain socket or TCP/IP protocols can query and submit tasks to the server through the API. The default client of the framework is a set of Django (http://www.djangoproject.com) powered web interfaces. The web client allows users to connect to and use the framework via their web browser. Thus, users can interact with the default client of the framework using any operating system and Internet browser they choose.

2.3. ***Implementation Status and Limitations***

The framework is still under development and currently the server analyzes 16S rRNA sequences only using RDP's naïve Bayesian classifier [10] and performs all analyses based on genus level taxonomy assigned by RDP. However, the modular nature of the framework allows us to extend its capabilities easily and we are currently working on implementing other data processing modules for phylogenetic analysis based solely on sequence similarity.

Currently the server is being used in house by several biological researchers for a number of active research projects. The most demanding project that has been analyzed on the framework included 166 samples with more than 2 million sequences. The framework server is installed on a Linux server as we work towards our first stable release. We are working to port the server to other platforms such as Mac OSX in order to distribute it more widely upon release.

It is also important to note that the classification of sequence data (currently performed via the RDP classifier) is independent and orthogonal to the downstream analysis and visualization tools. In fact, any data set that contains names and associated abundance values can be slipped into the framework and processed through the downstream analysis and visualization. As a concrete example of this, we have implemented a facility for quantitative PCR data to be loaded into the framework and analyzed in a similar manner to classified 16S rRNA sequencing data. We intend to extend this facility to microarray data as well.

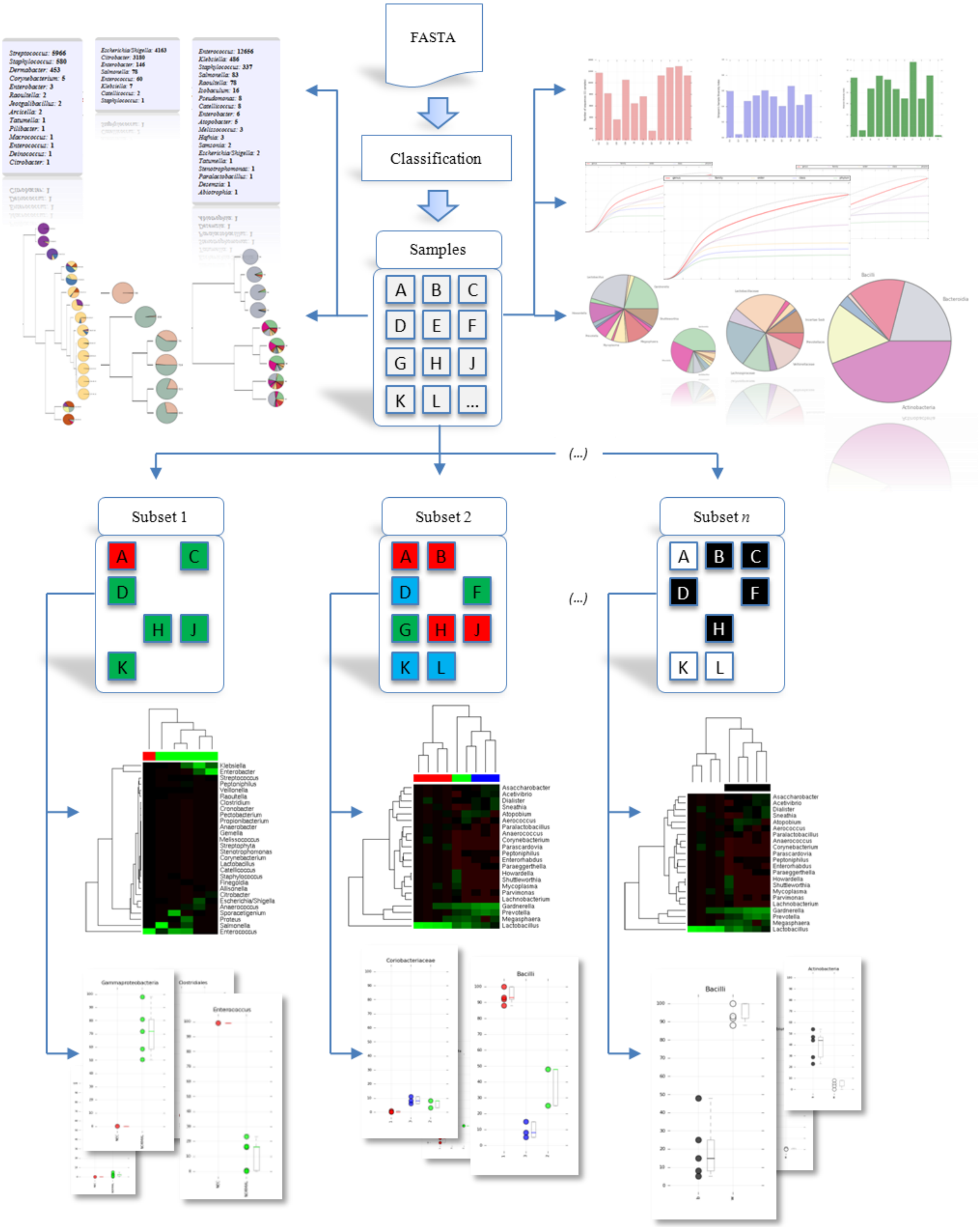

Figure 2: Basic workflow of the framework. Analysis begins with the submission of a FASTA formatted 16S rRNA sequence file.

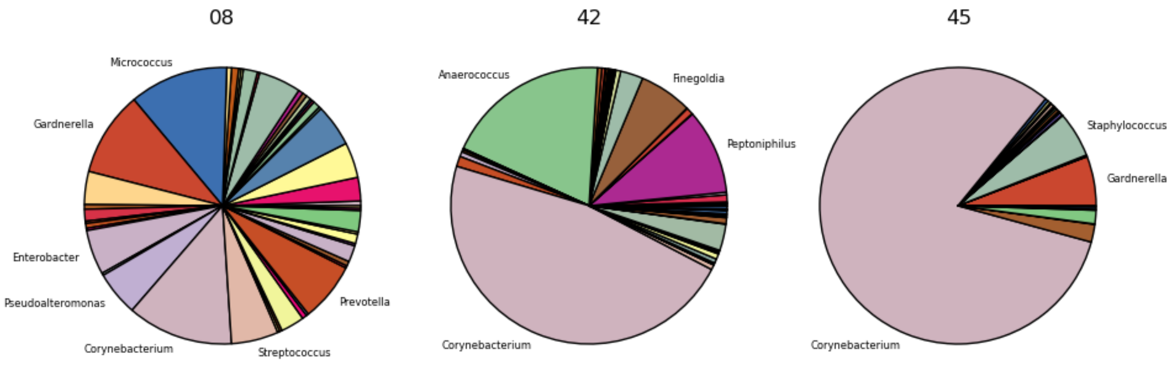

Figure 3: Example pie chart figures show bacterial composition at the genus level of three random samples from a bacterial vaginosis study analyzed using the framework.

3. Workflow

Ease of use and extensibility are key design concerns for the framework. Hence, most of the analysis tasks are performed without requiring any *a priori* knowledge to be provided by the researcher. The basic workflow of the framework is illustrated in Figure 2. Readers are encouraged to visit http://meren.org/framework/ to view an example analysis performed with the framework.

An analysis begins by submitting a FASTA formatted file containing 16S rRNA gene sequences. The file can contain multiple FASTA files originating from multiple environmental or clinical specimens. The framework then employs RDP's naïve Bayesian classifier [10] for rapid assignment of sequences to the taxonomic groups at the phylum, class, order, family and genus levels and the framework proceeds to perform unsupervised preliminary analyses on the samples acquired from the RDP classifier results. These analyses include:

- Calculations of total and percent abundance of bacteria in every sample,
- Bar chart representation of the number of sequences acquired for each sample,
- Bar chart representation of Shannon and Simpson's diversity indices,
- Pie chart representations of samples based on their bacterial compositions at each taxonomic level ranging from phylum to genus (Figure 3),
- Rarefaction curves to illustrate the degree of diversity covered by each sample (Figure 4),
- Hierarchical clustering dendrograms that illustrate how samples clustered based on their bacterial composition at different taxonomic levels (Figure 7).

Once this set of unsupervised alpha-diversity analyses is completed, researchers can assign keys to desired samples and create subsets of samples for further investigation. The user defines subsets by assigning samples to groups, and then assigns a color to each of those groups for visualization. There is no limit on the number of subsets the user may define. The framework automatically ignores samples that are present in the original library if they are not assigned into any groups in a defined subset.

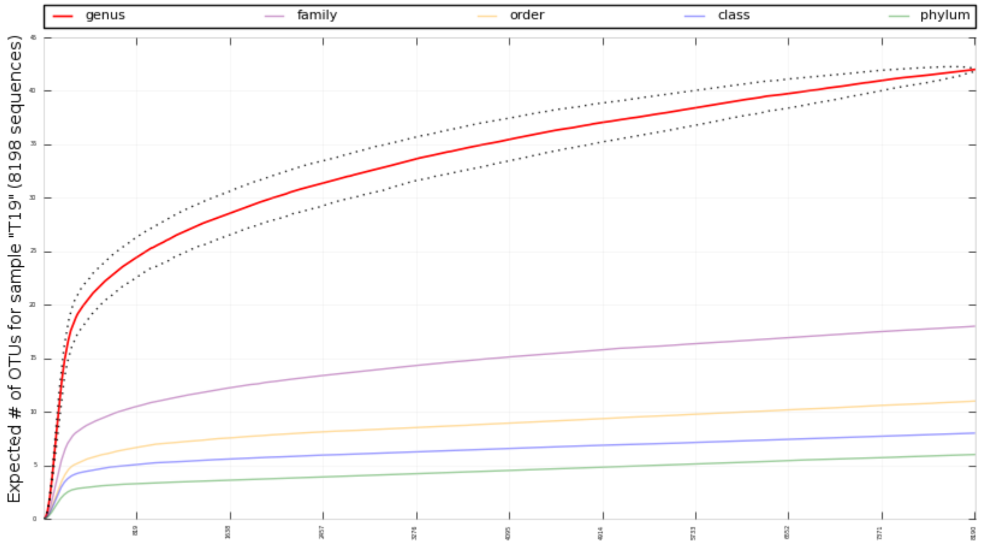

Figure 4: In this example set of rarefaction curves, species richness and expected number of OTUs are shown at different taxonomic levels of a sample that was analyzed using the framework.

When the newly defined subset of samples is submitted for analysis, dot plots of every operational taxonomic unit (OTU) at each taxonomic level ranging from phylum to genus are generated. Box plots are attached alongside the dot plots to illustrate the abundance of each individual OTU across subsets of samples (Figure 5). Complete linkage clustering analysis is performed to assess similarities between microbial communities based on the percent abundance of the taxa they contain. These clustering results are displayed as dendrograms along with heatmaps illustrating the abundance of taxa in each sample (Figure 6). Heatmaps can be refined further to eliminate very low abundance OTUs or to use logarithmic values.

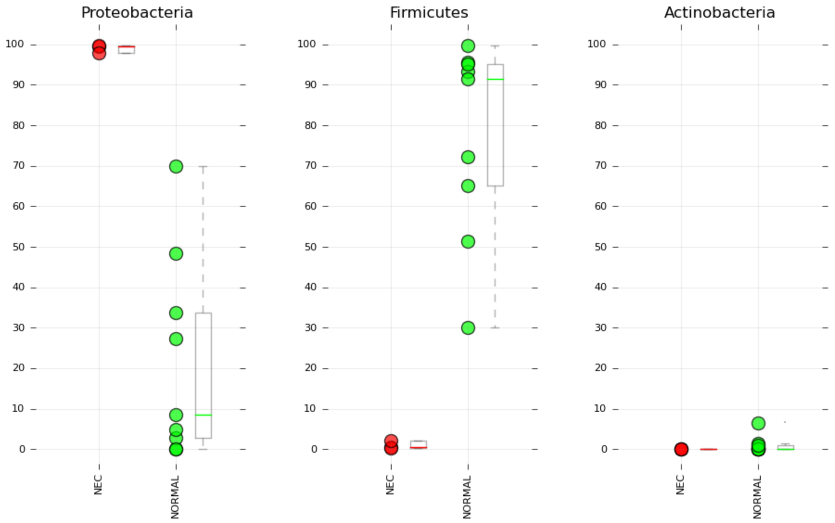

Figure 5: Example dot plot of a subset of samples assigned to two categories, NEC (green) or NORMAL (red) showing differences in the percent abundance of three different OTUs at the phylum level.

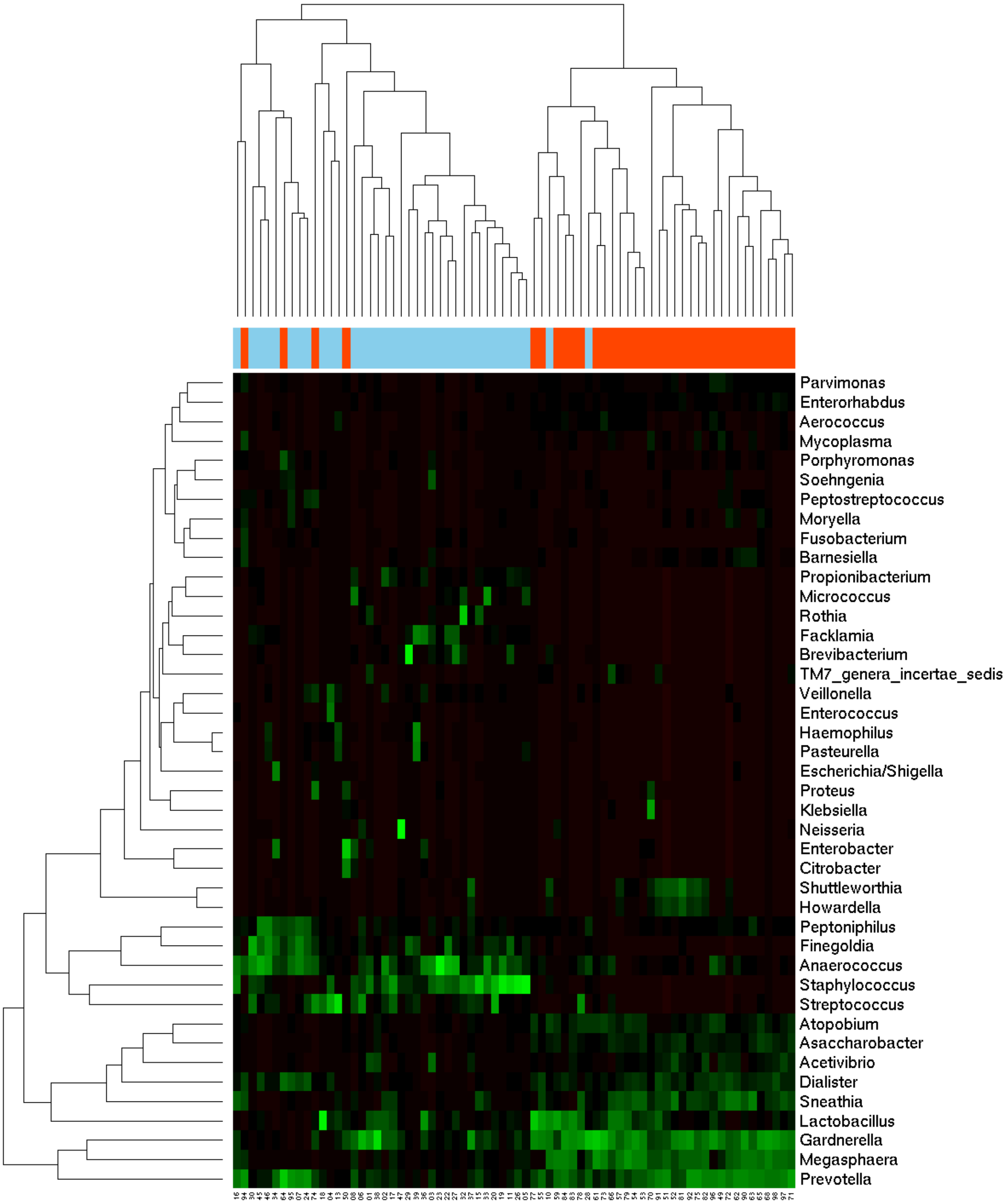

Figure 6: Example heatmap generated by the framework showing how a subset of samples clustered based on their microbial flora at the genus level. Within this particular subset of samples, the cyan color represents penile skin swab samples collected from male patients and the red color represents vaginal swab samples gathered from female patients. The vaginal swab samples largely cluster together on the left of the heatmap, while the penile skin swab samples cluster together on the right side of the heatmap.

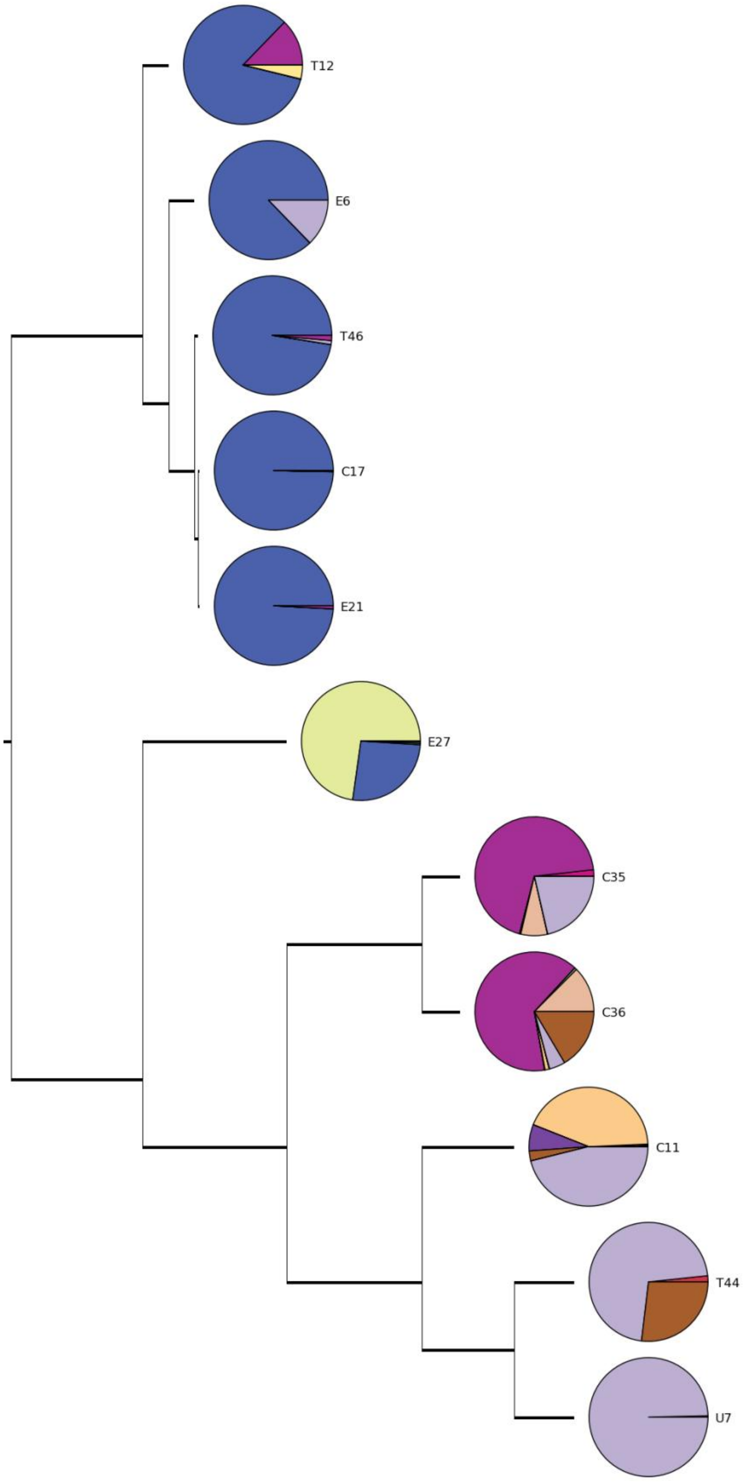

Figure 7: Example dendrogram generated by the framework showing how samples from a necrotizing enterocolitis study were clustered based on their microbial composition at the family level. Smaller versions of the pie chart representations of samples attached to the tree provide additional visual evidence for clustering results.

4. Discussion and Future Work

Many existing tools for metagenomic sequence analysis force biologists to learn application specific details to run various tests on their 16S rRNA sequence data. This burden may cause researchers to eschew new methods and tools for analysis in favor of those that they have already worked to become familiar with. A framework that provides ease of use and seamless integration of new methods as they appear will encourage researchers to try new methods.

We plan to enhance the framework with a variety of additional components in the future. We are currently working to add phylogeny based beta diversity analysis methods, such as UniFrac [11]. It is also worth noting that the longer read lengths being produced by the Illumina Genome Analyzer IIe have made deep sequencing of entire metagenomes feasible. This will allow researchers to go beyond simple classification based on 16S rRNA and on to analysis of complete metagenomes. Our framework provides the infrastructure for further development of features to address assembly, classification and processing of broader metagenomic sequencing data while maintaining the ease of use through web-based client interfaces.

Finally, our framework provides an important separation between classification of metagenomic sequencing data, and analysis and visualization of the classified data. We have currently implemented a front-end that uses the RDP classifier to interpret pyrosequencing reads of 16S rRNA into their taxonomic categories. We intend to enhance the utility of the framework by developing other front-end classifiers that may use the NCBI taxonomy or perform classification based solely on edit distance of sequences to further explore intra-genus and intra-species diversity. We are also working on a facility to utilize the analysis and visualization features of the framework on other data types such as quantitative PCR and microarray data, which can be slipped into the framework past the classification front-end.

5. Conclusion

Although there are a variety of tools currently available for metagenomic sequence analysis, they impose unnatural paradigms or restrictive limitations on biological researchers who may have only rudimentary computer skills. Pipeline approaches force users to select analyses to perform on their data instead of performing a comprehensive analysis by default. They also place a burden on the user for maintenance and routing of intermediate results that can lead to errors. Web based applications have advantages in terms of ease of use, but can be restrictive in the quantity of data they allow to be analyzed and the amount of user interaction required to perform an analysis. None of these approaches, by themselves, provide a viable alternative for microbial community researchers to analyze their data without scaling a significant learning curve.

Our framework provides a scalable, hybrid approach to the problem of metagenomic sequence analysis. Researchers can run the framework on their own computational resources and are not faced with limitations on the quantity of sequences or number of analyses they can perform. They are also able to use familiar web-based interfaces to access the server and do not need to shepherd analyses through a pipeline and manage intermediate results. All of the standard analysis methods are run at the push of a button and the user is presented with an intuitive interface to group samples for further directed analyses. This flexibility and ease of use has allowed microbial community researchers to perform their own analyses and generate publication quality figures to communicate their results with relative ease.

Acknowledgements

This work is supported by funding from the Research Institute for Children in New Orleans.

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.

References

1. D. A. Sandoval, R. J. Steeley, *Science* **328**(5975):179-80 (2010).
2. Y. Wang, J. D. Hoenig, K. J. Malin, S. Qamar, E. O. Petrof, J. Sun, *et al*, *ISME J.*, doi:**10.1038**/ismej.2009.37 (2009).
3. W. B. Withman, D. C. Coleman, W. J. Wiebe, *Proc. Natl. Acad. Sci. USA* **95**, 6578–6583 (1998).
4. NIH HMP Working Group, J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, *et al*, *Genome Res*. **19**(12):2317-23 (2009).
5. G. J. Olsen, C. R. Woese, *FASEB J.*, **7**:113–123 (1993).
6. C. Shyu, T. Soule, S. J. Bent, J. A. Foster, L. J. Forney, *Microb. Ecol.* **53f4**, 562 (2007).
7. J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, *et al*, *Nucleic Acids Res.* **D**, 141 (2009)
8. P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, *et al*, *Appl. Environ. Microbiol.* **75**(23):7537-41 (2009).
9. J. G. Caporaso, J. N. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, *et al*, *Nature Methods* doi:**10.1038** / nmeth.f.303 (2010).
10. Q. Wang, G.M. Garrity, J.M. Tiedje, J.R. Cole, *Appl Environ Microbiol.,* **73**(16):5261-7 (2007).
11. M. Hamady, C. Lozupone, R. Knight, *ISME J.*, doi:**10.1038**/ismej.2009.97 (2009).