Analysis Proposal

Sean Stuntz

January 17, 2018

Bioinformatics Automated Workflow System (BAWS)

BAWS allows users to quickly filter, visualize, and analyze taxonomic marker data.

Why is it Useful?

The app is intended to, as the name describes, automate various data formatting, visualization and statistical procedures for 16s RNA genetic sequence taxonomic markers. In a nutshell, the app will generate data tables from a raw data set, assign taxonomy and create a phyloseq object. The phyloseq object can then be used for subsequent analysis or for sharing information between researchers. The intent is to take raw, genetic sequence data and format it or analyze it until it is able to be shared with the microbiome research community.

The app will enable automation for the following processes:

- Referencing 16s RNA gene sequence raw data file, generate data table with user specified number of forward and reverse passes
- Cluster sequences into Operational Taxonomic Units meeting a certain dissimilarity threshold
- Construct sequence table and remove chimeras
- Assign taxonomic markers
- Construct a phylogenic tree
- Build Phyloseq object by consolidating information above

There is nothing novel being performed in the analysis performed within the app. All it is intended to offer is automation of these processes in order save microbiome researchers time when performing routine formatting and analyses. In fact, the app builds off of R packages commonly used in this community including phangorn, ggplot2, dada2, DESeq2, and structSSI. To benefit from this app, users need to have an understanding of genetic sequencing analysis techniques and basic statistics. Users can find the app at my GitHub repository located here.

Feature Table

Description	Priority	Status	Value Provided	Inputs	Outputs	Use	Sufficient Time	Current/Future
Generate data tables from raw data files	1	Not Started	Saves user time with automation based on inputs.	16s RNA gene sequences	Formatted data table	Direct input into priority 2. Cleans up the data and allows for subsequent analysis.	Yes	Current
Cluster sequences into OTUs meeting a certain dissimilarity threshold	2	Not Started	Automation will allow faster data handling and visualization.	Formatted raw data file	Formatted data table	Direct input into priority 3. Builds OTUs which are eventually used to identify taxonomy.	Yes	Current
Construct sequence table and remove chimeras	3	Not Started	Assigns genetic sequences and categoricals based on OTU input	Formatted OTU data file	Formatted data table	Direct input into priority 4.	Yes	Current
Assign taxonomy	4	Not Started	Categrorizes genetic sequences	Sequenced data table	Formatted data table	Direct input in priority 5. Taxonomy categorization is used to identify patterns between genetic sequences.	Yes	Current
Construct a phylogenic tree	5	Not Started	Allows users to visulaize differences in genetic sequences	Formatted taxonomic data file	Phylogenic tree	Direct input into priority 6. Visualization of taxonomic data.	Yes	Current
Build Phyloseq object	6	Not Started	Standardized output for easier communicating, reproducibility, and further analysis	Phylogenic tree and other analyzed formats	Formatted phyloseq object	End product used for subsequent statistics. Enables reproducibility.	Yes	Current