Taxator-Tk Benchmarking and Ensemble Classifier

Nicole Buleza, Anna Lu, Keyur Shah
ECES T480/T680: Statistical Analysis of Genomics
https://github.com/ahl54/ECEST480-680

March 24, 2017

Contents

Abstract

In metagenomics and in medicine, false positives are great concerns as they can affect medical diagnostics and infectious agent tracking. Often, these overestimations are attempted to be addressed on a tool-by-tool basis and by conservatively, or not at all, taking into account the potential bias of databases and methods on performance. One source for potential bias is the type of gram stain of bacteria. Gram-positive bacteria have thick, peptidoglycan cell walls and are stained purple. Gram-negative bacteria have thinner cell walls and are counter-stained pink. It was an observation from preliminary data that certain tools performed better on type of gram stain. In this paper, we evaluate if gram stain type correlates to tools that use k-mer based methods in their sequence assembly and then gauge if a bias exists. Accordingly, we will build an ensemble classifier to reduce potential bias from tools that use k-mer length in sequence assembly. We will, in addition, attempt to do a benchmark comparison on shotgun-sequencing data for taxonomic labeling with the tool, Taxator-Tk.

I. Overview of Project and Goals

This project was inspired by Dr. Rosen's paper and had two distinct parts. As such, this report has been broken into Part 1 and Part 2 accordingly. The primary objective for Part 1 was to benchmark the data sets using PhyloPythiaS+ and Taxator-Tk. The primary goal for Part 2 was to build an ensemble classifier to test its performance on the data. As for writing this report, several issues arose that severely hindered the aforementioned goals.

PhyloPythiaS+ Evaluation Discontinued

We discontinued evaluating PhyloPythiaS+ because the tool required updates to be downloaded from their servers. However, the server host returns a '`403 Forbidden`' error when running the following command for a latest software update: `-update s`. The same error is returned also for the following command for fetching reference data: `-update t`. One failed attempt made to resolve this problem was to spoof the web request to use various different browser headers: Safari, Mozilla, Chrome, Opera, and Linux based browsers. Ultimately PhyloPythiaS+ was not installed due to these unresolved errors, but exists at

```
/mnt/HA/groups/rosenclassGrp/benchmark_project/PhyloPythiaS+
```

An abbreviated version of our progress can be seen in Figure 1. The obstacles encountered will be expanded upon and discussed in further detail for each part of the project in sections II and III.

Week	Milestones
5	Installation of PhylopythiaS+ and Taxator-Tk on Proteus
6	PhylopythiaS+ server forbidden error 403, continue installation of Taxator-Tk on Proteus
7	Taxator-Tk pipeline issues 1-4 resolved. Troubleshoot alignment format error 5. Unable to install PhylopythiaS+. Communicate issues with Dr. Rosen
8	Discontinued PhylopythiaS+ benchmarking. Continued to work pipelining Taxator-Tk results. Received output data and began working on ensemble classifier.
9	Continued to work on outputting Taxator-Tk results. A data loading and parser script was written in MATLAB for the ensemble classifier.

10	Resolved % Taxator-Tk issues. Email correspondence with creator of tool and Dr. Rosen. Generated graphs from TRUTH data. Presentation.
11	Developed distance matrix heatmap for relative abundances. Documentation of project methods and analysis.

Figure 1. Project timeline

II. Part 1: Benchmarking Taxator-Tk

About the Benchmarking Data:

The IMMSA dataset contains 81 number of gz compressed fasta files (McIntyre, 2017). The data exists on Proteus at the following path:

```
/mnt/HA/groups/rosenclassGrp/benchmark_project/data/nist-immsa/IMMSA/
```

i. Obstacles Encountered

Taxator-tk Installation

Taxator-tk is currently installed on Proteus at the following path:

```
/mnt/HA/groups/rosenclassGrp/benchmark_project/taxator-tk
```

Installation began week 4 and required several dependency resolutions on Proteus. The first dependency required a module called 'Boost' which administrator David upon his return installed on Proteus for us on February 7 in week 5. The second dependency required a g++ library call from cmake installer overriding the default system gcc compiler. The solution to this problem was resolved by trial and error. Initial attempts to call make, `./build.sh`, and cmake each from fresh installations failed.

Next we attempted to export path variables specifically for gcc and g++ such that CMake could call the Proteus g++ compiler instead of the default Linux system gcc path. Note that this partial solution only works on a fresh clone of Taxator-tk from source. Path variable exports created after calling `./build.sh` generates a CMakeCache.txt that again provides the unwanted compiler. The order of steps are crucial to a successful build:

```
module load boost/openmpi/gcc/64/1.57.0

export CC=$( which gcc )

export CXX=$( which g++ )

git clone https://github.com/fungs/taxator-tk.git
```

Taxator-tk Pre-alignment

Taxator-tk requires a pre-alignment step before classification. Taxator-tk recommends three alignment options: LAST2MAF, BLASTXML2, Native BLAST. We chose to align using BLASTn alignment without e-value filtering at 1e-20 used by the authors (pp. 16) because the

USAGE.md Taxator-tk warns that e-value filtering negatively affects the binning results. Script to align and run a data file through Taxator-tk exists at the following path:

```
/mnt/HA/groups/rosenclassGrp/benchmark_project/scripts/taxator_master_job.sh
```

ii. State of Project

Current Implementation

Taxator-tk is installed and scripts written for alignment and binning workflow are written. Currently pre-alignment and binning scripts are being tested and revised on Proteus for one sample output.

Results

No bins were generated from the sample dataset provided: b9.pass.2d.fasta
Expected binning output format from Taxator-tk has the following columns:
@@SequenceID TaxID _TaxatorTK_Support _TaxatorTK_Length

iii. Discussion

Analysis of Results/Interpretations

Taxator-tk is not ready for production use since the alignment file input requires manual formatting and the author provided mapping.tax from the refpack throws a bad taxon mapping error. Error reports and attempts to resolve the mapping.tax are detailed at https://github.com/ahl54/ECEST480-680/issues.
Taxator-tk issues, including the bugs described above, are also documented at https://github.com/fungs/taxator-tk/issues.

III. Part 2: Ensemble Classifier

Goals

The goal of the project is to explore the data and answer the question: Does any relative abundance per taxa bias exist in an ensemble of tools? A subgoal of the project is to classify and compare data on bacterial feature as either gram positive or gram negative for comparison of any weighted bias on that feature.

About the Data

A data loading and parser script was written in MATLAB for the ensemble classifier. Identified were 136 ground truth (TRUTH) files, which were divided into four subtaxa levels {40 full, 40 genus, 40 species, 16 subspecies} as a structure. CosmosID datasets are excluded from following analysis. Each file corresponds to a single tool or ground truth. Every file contains five header columns with example values shown as follows:

TaxID	Number_of_reads	Relative_abundance	rank	Name
28216	50000.00000	0.04167	class	Betaproteobacteria

i. Obstacles Encountered

File parsing

File parsing was a majority of the coding effort due to mixed data types: float for relative abundance, string for taxa and file names. Loading the data requires wildcard expression filtering (TRUTH vs non-truth) on file names. Since MATLAB does not have string exclusionary functions like regular expression, non-truth files were filtered to remove any truth files after aggregating all files. Data within files were stored within structures to resolve mixed data types.

We manually entered headers to all the output files. This was done because, MATLAB saves the first row of the file as structure headers, and we were losing information as a result of it.

Identifying gram positive and gram negative bacteria

NCBI taxonomy database provides information about whether a bacteria is gram positive or gram negative. However, the classification of gram positive and gram negative values were manually entered and often inconsistently located in paragraph descriptions. Therefore gram positive or negative classification could not be programmatically extracted. We manually collected the data for a total of 433 unique taxa as a lookup table for comparing gram positive versus gram negative bacteria distinctly.

Distance metric for relative abundance

Relative abundance is a continuous value in units of a percentage. In order to compare relative abundances between tools, relative abundance values must be normalized. There was some uncertainty in determining the appropriate distance metric for normalizing relative abundance. We selected the L1 norm, Euclidean distance by Dr. Rosen's recommendation, however, it is notable that the Euclidean distance may be biased in favor of high

Visualizing meaningful data

Visualizing the data is difficult due three dimensionality: relative abundance, taxa, and tool. Initially we displayed the data as 2D histograms with one histogram per tool, plotting organisms (taxa) versus relative abundance. This generated a lot of histograms and wasn't clearly comparable between tools. A distance matrix of relative abundance as a distance, with 2D heatmap taxa of truth vs taxa of tools would resolve the dimensionality representation. Additionally the distance matrix heatmap shows a color scale for identifying specific taxa of interest above a quantitative distance threshold.

ii. Project Implementation

Methods

Implementation consists of four core scripts: load_truth.m, load_tools.m, parse_ra.m, and relative_abundance.m. Load_truth.m and load_tools.m extract the truth and non-truth files from the output directory and returns a structure subdivided into full, genus, species, and subspecies levels for both truth and non-truth files respectively. Parse_ra.m extracts two columns from each file: the relative abundances in the third column and the organism from the last column. Load_data.m extracts data from directories that have been organized separately into truth and non-truth folders.

Relative_abundance.m contains the ensemble classifier. The ensemble classifier is an intersection of relative abundance for each taxa between ground truth and tools. We chose the same distance metric used by the authors, L1-norm the Euclidean distance as our normalization metric on relative abundances. The distance matrix is visually represented as a heatmap for color identification of most similar taxa by relative abundance within a tool.

iii. Results and Interpretation

In Figure 2, our group graphically represents the TRUTH values for BMI Genus level reads. The TRUTH data set is the established ground "truth" values for each taxa. In Figure 3, we see that only 3 tools matched the ground truth's relative abundance value. The tools that do match truth (Metaphlan, BlastMeganFiltered, and BlastMeganFilteredLiberal) use mechanisms other than discriminant k-mer based algorithms. The tool KrakenFiltered, however, is an exact k-mer based algorithm and substantially overestimates the relative abundance compared to the truth value. This make sense if gram negative bacteria favor k-mer based algorithms as Pervotella is a gram negative genus. Conversely, as seen in Figure 4, KrakenFiltered returns zero or negligible relative abundance for Peptostreptococcus, a gram positive genus. The tool ClarkDefault, which is a discriminant k-mer based tool, showed zero relative abundance with the gram negative bacteria and 0.03 relative abundance with the gram positive bacteria. ClarkDefault results were the opposite of what was expected if gram stain bias does exist.

In the heatmap in Figure 5, the data shows minimal variance between relative abundances across various tools. This suggests bias does not exist or does not significantly impact each tool's computation for relative abundance. If gram stain or some other confounding variable was significantly affecting the ten tools' computation process, we would expect to see more variance in color scale. The brighter colors represent disimilarity (larger distances) between the tools for each taxa compared to truth values. Some tools are dissimilar to the truth, depicted by red or white lines, however, there is no pattern or gradient suggesting specific tools are more dissimilar than other tools based on relative abundance.

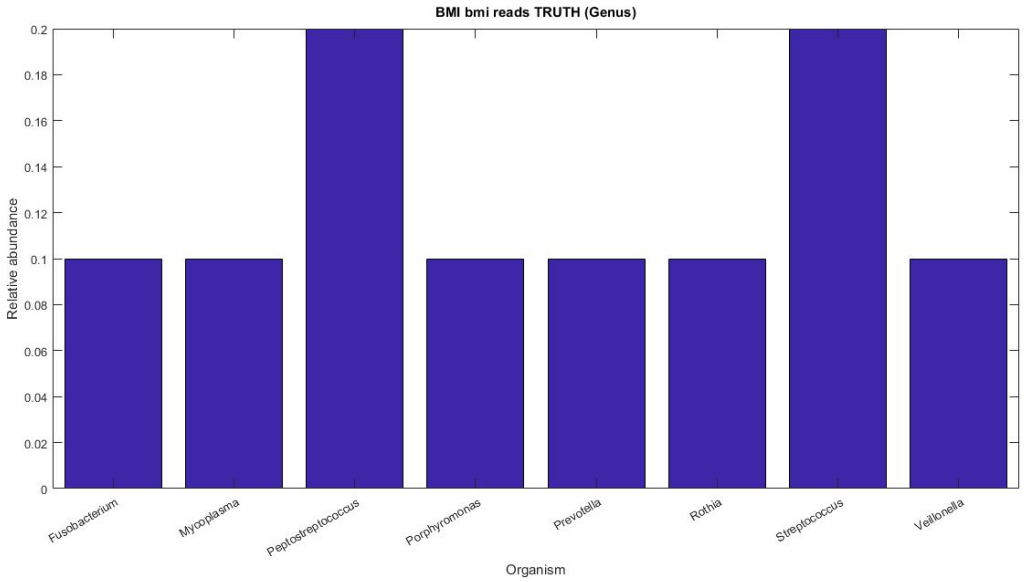

Figure 2. Relative abundance vs Organism (taxa) histogram for BMI bmi reads TRUTH (genus) method.

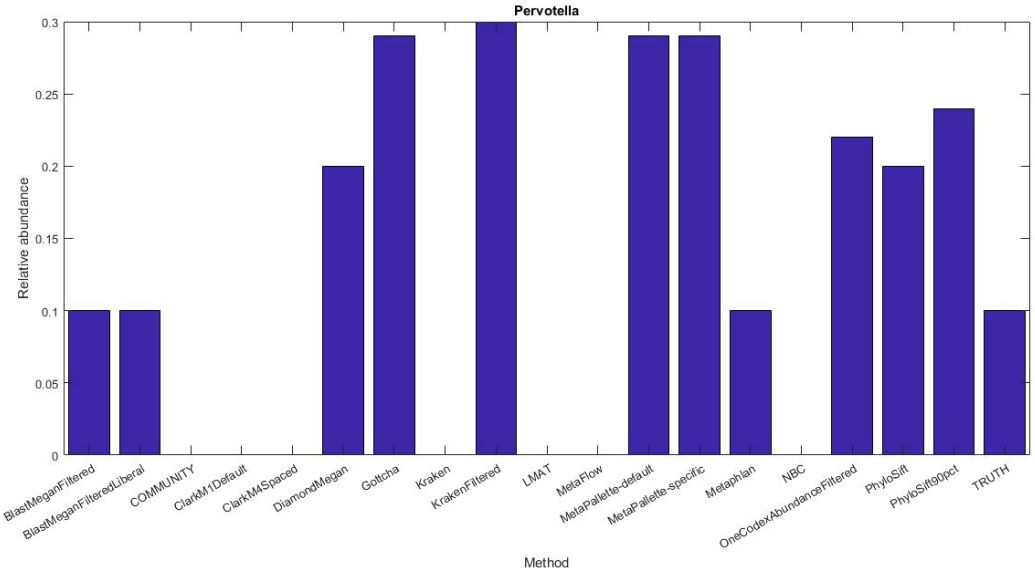

Figure 3. Relative abundance vs Method (tool) histogram for taxa Pervotella. Eight tools show potential bias for higher relative abundance for Pervotella, a gram negative bacteria, compared to the TRUTH relative abundance. Three tools matched the ground truth and seven tools did not detect Pervotella from the same data.

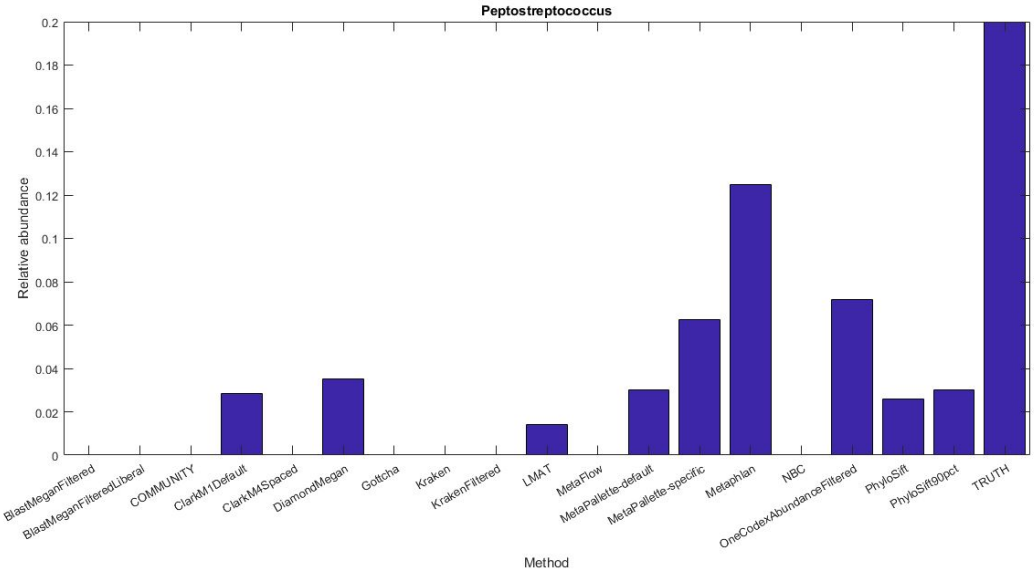

Figure 4. Relative abundance vs Method (tool) histogram for taxa Peptostreptococcus. Nine tools show potential bias for lower relative abundance for Peptostreptococcus, a gram positive bacteria, compared to the TRUTH relative abundance. None of the tools matched the ground truth and nine tools did not detect Peptostreptococcus from the same data.

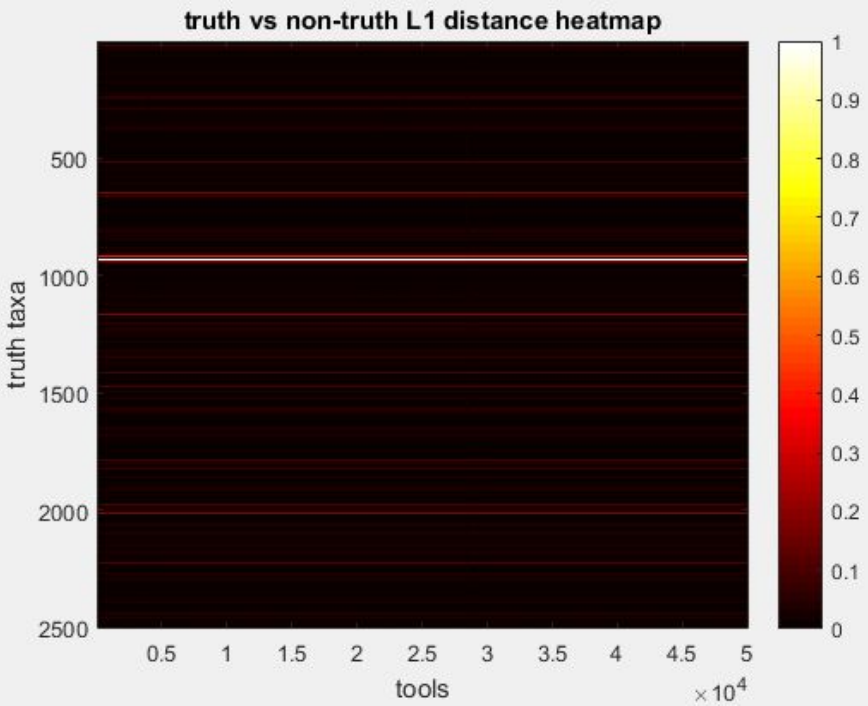

Figure 5. A relative abundance distance heatmap of truth versus tool per taxa are depicted. A majority of similar (zero/low, dark) distances exist between truth and tool per taxa, meaning most tools identify a particular taxa. Brighter, red lines indicate greater distances between tool vs truth, meaning variable relative abundances were measured by various tools for a some specific taxa. A white line at 1.0 indicates complete dissimilarity between tool and truth value for relative abundance.

IV. References

J. Dröge, I. Gregor, and A. C. McHardy, *Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods.* Bioinformatics 2015 31: 817-824. doi:10.1093/bioinformatics/btu745

A. B. R. McIntyre, R. Ounit, E. Afshinnekoo, R. Prill, G. L. Rosen, E. Hénaff, N. Alexander, S. Ahsanuddin, S. Tighe, N. A. Hasan, P. Subramanian, K. Moffat, C. Heberling, M. Dadlani, S. Minot, S. Levy, N. Greenfield, R. R. Colwell, C. E. Mason. *Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers.* 2017 pre-print in review