ECES T480/680 Statistical Analysis of Genomics

Nicole Buleza, Anna Lu, Keyur Shah

March 05, 2017

Project Progress Report

Current State

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.

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.

Goals Achieved and Preliminary Results

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/
```

PhyloPythiaS+ Evaluation Discontinued

We have discontinued evaluating PhyloPythiaS+ because the tool requires 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+
```

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 megablast with e-value filtering at 1e-20 because the pre-print authors applied the same method for a BLAST tabular output format (pp. 16). Script to recursively align each of the fasta.gz data files exists at the following path:

```
/mnt/HA/groups/rosenclassGrp/benchmark_project/scripts
```

Ensemble Classifier

CosmosID datasets are excluded from following analysis. A TRUTH data loading and parsing script called load.m is available. Pseudocode for the distance metric evaluation is described in future goals below.

Future Goals

We still need to determine if taxator-tk output is appropriate for Dr. Rosen's colleague's benchmarking evaluation scripts. Default Taxator-tk output from at minimum, one fasta sample, will be provided to Dr. Rosen by March 11 of Week 9 for benchmarking script compatibility.

An ensemble classifier, yet to be implemented, must answer the question, "Which tool performs better on relative abundance per taxa level?" Data exploration of relative abundance on a per taxa level will be developed using two approaches. First, a Mahalanobis distance metric (using MATLAB function pdist) is proposed as an alternative to L1 Manhattan distance, justified by improved accuracy as a distance metric in kNN classification (Weinberger 2009).

In order to address undetected taxa for some tools, a UNION set operation may be taken of all taxa detected tools first to reduce the number of iterations through taxa and tools. The pseudocode to generate a distance matrix for all detected taxa and tools based on their relative abundance (RA) is as follows:

```
for each taxa:
    for each tool:
                distance matrix include:
                Mahalanobis_distance ( this tool RA, other tools' RA )
```

Timeline

Project was assigned. Dr. Rosen sent us the data for the benchmarking project at the end of week 3. We began trying to install the two tools for benchmarking in Week 4. Taxator-tk required several dependencies that needed to be installed by David in order for us to work with them. During week 4, Anna met with Dr. Rosen to discuss the existing work and expectations of the benchmarking project and ensemble classifier. David finished all the installs, and we were able to start benchmarking our data in week 5. We attempted to install PhylopythiaS+ week 6; however, encountered aforementioned errors. Progress was mostly postponed week 6 as Keyur and Anna prepared for their tutorial in week 7. By the end of week 7, we had benchmarked all Taxator-tk data, but were still unable to work around the error given by PhylopythiaS+. On Tuesday of week 8, we met with Dr. Rosen before class to discuss problems with installing PhylopythiaS+. Dr. Rosen instructed us to discontinue the benchmark for PhylopythiaS+ and to focus on the ensemble classifier with Taxator-tk. Also, Dr. Rosen provided '*OUTPUT*' data for the ensemble classifier and introduced us to the other group working on the benchmarking project. On Friday of week 8, Anna and Moon met with Dr. Rosen to discuss the results from their group's output and classifier approaches.

For the remaining two weeks, we will be providing Dr. Rosen with our Taxator-tk outputs by March 11 of week 9 and will continue working on the ensemble classifier throughout week 10 and 11. The final project results and a reporting pdf of statistical analysis and interpretation will be packaged and emailed to Dr. Rosen by March 24.

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

Q. Weinberger and L. K. Saul., *Distance Metric Learning for Large Margin Nearest Neighbor Classification.* Journal of Machine Learning Research 10 (2009) 207-244