PFB2017 Project Proposal

Nathalia GG

**Overall Goal:**

The goal of my proposal is to annotate a protein coding gene set from any pathogenic organism, and to build a basic gene ontology organization in html/website format to easily share the annotation information with collaborators.

**Justification:**

I am proposing a simplified gene ontology to gather functional annotation of predicted proteomes and/or RNAseq. I would like to develop my programming knowledge to skills to develop such gene ontology so my advisor and other lab members can access genetic information of the organisms we work with more easily.

My current problem is how I search the most informative set of genes in my data sets. My research focuses on plant-pathogen interactions, and my research goal is to identify candidate genes of a pathogenic fungus that are involved in modulating a plant host. I use non-model organisms, so part of my research also involved sequencing the genomes of 11 fungal isolates and 2 plant cultivars. We collaborate with a bioinformatician who assembled, predicted protein coding genes and annotated these fungal and plant genomes. As much as I appreciate his collaboration, the annotation generated was against standard databases such as SwissProt, which isn’t helping me narrow down options for candidate genes involved in plant disease development.

Recently, I learned about PHI-base (<http://www.phi-base.org)>, a protein database that contains information on genes involved in host-pathogen interactions. This database includes curated biological information of genes of many pathogenic organisms from animal to plant systems that have been experimentally validated. There are many genes of fungal plant pathogens that are included in this database. Therefore, I believe a more tailored functional annotation pipeline of fungal plant pathogens against PHI-base would be more adequate for my research goals.

After today’s class with Dr. Eilbeck, I am also very interested in annotating predicted proteins with GO terms. I believe this annotation will also be useful because the organisms I work with are non-models and newly sequences.

Other useful annotations for plant pathogens are: HMMER search against db-CAN (CAZymes) database (<http://csbl.bmb.uga.edu/dbCAN/)>, and BLASTp search against MEROPS (proteases) (<https://www.ebi.ac.uk/merops/)> database. There also programs that predict signal peptides (SignalP - <http://www.cbs.dtu.dk/services/SignalP/> ), subcellular localization (TargetP - <http://www.cbs.dtu.dk/services/TargetP/> ) that would be informative as well.

In the future, I would like to build such gene ontology with my RNAseq data, which is incomplete at this point (I have to finish a few experiments). My vision is to build a web interface, so my advisor or other people could easily access gene fasta sequences and protein sequences, as well as their functional annotation. Perhaps, I should consider teaching them some programming instead!

**Proposed outline:**

1) Download the predicted proteome (= proteome) of *Magnaporthe oryzae*, or another well studied fungal plant pathogen, that is publicly available in the JGI.

2) Filter the data in Unix:

- remove sequences smaller than 100 aa,

- sort sequences according to size (largest to smallest) rename all fasta protein sequences,

- rename fasta headers to include pathogen abbreviated name (i.e. >Moryzae\_g1, >Moryzae\_g2, etc.).

3) Annotate the predicted proteome against those databases using BLASTp and custom parameters.

- use Python to create a command-line script to call the program

4) Scan predicted proteome against SignalP, and TargetP

5) Use Python scripts to incorporate the results of those analysis into a dictionary data structure.

- build a data structure (dictionary of dictionary) with fasta headers (=gene name) as keys

6) Build a simple gene ontology system to organize the data into a few functions, such as pathogenicity, increased virulence, signal peptide, etc.

- python3 or sqlite3?

7) Mission accomplished? Or is there a way to create a SIMPLE web interface?