Nathaniel Schleif
Project Progress Report
Identification of Indirect Regulators in Gene Regulatory Networks

Proposal Updates: The main modification to my initial proposal is that I no longer will be using time-series data as the basis of my analysis. The reason for this is that it is very data intensive to get good training on the data and not a lot of those sort of data sets exist in plants unfortunately. Consequently I will be changing to just doing "stable state" network inference from RNA-seq data. Specifically I will be using data from a developmental from maize developed by Walley *et al*. (2016), though I am still trying to find a good dataset in Arabidopsis as that would be easier to work with. I can still use the String database for protein-protein interactions and the RNA-seq data has already been processed.

Progress:

- I trialed the dynGENIE3 program as well as the previously mentioned data. Analysis of these datasets resulted in extremely weak associations which led me to explore other options.
- I brought in the new dataset from Walley into R, the required language for GENIE3, the "new" approach I am using to analyze the stable state data.
- I moved to a computer in my lab that is much more powerful. Installing Linux and setting up R / python took awhile but is complete.
- I developed a randomization scheme for creating the null distribution I will use to define "true connections." Unfortunately GENIE3 does not suggest a way to declare substantive connections and so I will be picking an FDR based on how the data distribution looks. I am currently waiting for it to run.

Challenges:

- Changing over to GENIE3 has been a bit of an ongoing challenge. It does not clearly communicate via the command line if it is making progress so I will hopefully know tomorrow if it is actually working. If it does not work on the time scale of a 24 hour period, I will explore moving on to HTCondor.
- If I am sticking with this maize dataset I will need to find a source for transcription-factor binding information which is still an open question. There is a lot less research done on that front in maize and resultantly I might need to lean on orthologs of transcription factors as found in Arabidopsis (from the JASPAR database) and use their known binding activity.
- The biggest problem will be in interpreting the data. I plan to use UniProt Swiss-Prot database to identify genes that specifically have some level of research behind them such that I can infer if my identified connections are accurate.