Nathaniel Schleif
Project Proposal
Identification of Indirect Regulators in Gene Regulatory Networks

Introduction: Gene regulatory inference is a useful tool for finding tentative connections that exist between genes. However, one of the weaknesses of this approach is that the connections made between genes are not necessarily direct; just because a gene is shown to have a regulatory influence on another gene, there is no guarantee that there is not intermediary proteins that actually do the work of transmitting the signal. Finding these intermediaries is important because it fleshes out the actual regulatory circuit, important as both a basic research pursuit and in an applied focus of manipulating these circuits. I aim to address this through leveraging GO annotations, DAP-seq data, and data on protein-protein interactions (PPI).

Resources: For this project I do not estimate needing more computing resources than my laptop. I have more powerful computers available to me at home if that proves to be insufficient. As for data resources, I will need quite a few: Firstly, I need an expression dataset; I plan to get this from the CURSE website[1] which pulls data from the Short Read Archive. Specifically I will be choosing a time-series dataset from Arabidopsis. I also need GO annotations which I will get from the TAIR website[2]. PPI data will be gotten from STRING[3] which collects and infers on to other species from known interactions. Finally, protein-DNA interactions will be taken from a large-scale DAP-seq study in Arabidopsis[4].

Methods: My approach will be multi-fold with a few different "levels" that I could bring the project to, depending on available time. Firstly, I plan to construct a directed gene-regulatory network using GENIE3[5] which will be the backbone of the study. I will than use the GO annotation set to identify parental nodes that are not annotated to have any DNA binding properties. These are likely to have intermediates in its regulation. Building on this, I plan to use the DAP-seq to strengthen the GO annotations: Parental nodes that do not have any binding to their children are likely to have intermediaries. Furthermore, the DAP-seq can be used to find potential transcription factors that do actually act on the children nodes and would be the potential intermediary targets of the parent gene. Finally, the PPI information will be used to infer the potential chains of interactions that could connect the parent node to its children. The pathway that has the strongest interactions will be ranked higher than those that consist of weak connections.

Anticipated Results: The principal result will be to identify genes that regulate other genes through intermediaries. I will judge success by being able to recover known genes that influence the expression of other genes through intermediaries. One intermediary in particular that I am interested in is ubiquitin which seems would be a common intermediary that would act post-translationally.

Related Work: Related approaches are ResponseNet and Forest from Omics Integrator which construct interactomes which provides similar information to what I am trying to find.

References:

1. Vaneechoutte, D. and Vandepoele, K. (2018). Curse: building expression atlases and co-expression networks from public RNA-Seq data. Bioinformatics.
2. Lamesch, P., (2011). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Research, 40(D1), pp.D1202-D1210.
3. Szklarczyk, D., (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Research, 45(D1), pp.D362-D368.
4. O'Malley, R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell, 165(5), pp.1280-1292.
5. Huynh-Thu, V., Irrthum, A., Wehenkel, L. and Geurts, P. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE, 5(9), p.e12776.