Final Project Proposal

Yashwanth Manne

Background, goals, & significance

- What is the problem you are hoping to address?

I'm planning on linking ATP-binding cassette (ABC) transporters and lipid transfer proteins (LTPs) in arabidopsis plants. ABC transporters are typically used in transporting small hydrophobic molecules across membranes for defense, cuticle formation, or signaling. These transporters often work in parallel with LTPs that "pick up" the compound from the ABC transporters. Through this project, I hope to understand possible genes involved in the transport of small hydrophobic particles and the LTPs' and ABC transporters' role in this.

- What is the current approach to solving this problem & what are its limitations?

Currently, in order to link two genes, co-expression data is used with a Pearson coefficient as the main summary statistic. While this isn't a problem in most cases, in the rare case that the data has a strong outlier or a characteristic, non-linear shape, the approach falls short. Furthermore, the results of co-expression studies only suggest correlations between any two genes/proteins. In reality, these may be spurious correlations. As a result, it is important to compare this with a corresponding protein-protein interaction network to eliminate spurious relations.

- What will you do & why is it likely to succeed?

I aim to gather existing co-expression data from ATTED and cross-reference relations with high Pearson correlation scores to known protein-protein interactions from BioGRID and then rank the top 50 relations for future study. This approach is likely to succeed as the cross-referencing allows for the removal of spurious relations and more clues in what to look for.

- If successful, what is the broader impact?

Understanding the transport of small hydrophobic molecules in plants is important for basic research as it provides a branching point for future research to further shape our understanding of the cellular mechanisms of plants. Additionally, a base understanding of small hydrophobic molecule transport can aid in the production of high-value plant products as a lot of these compounds have bioactive properties. Specifically, my project will provide key genes that should either be looked into further via pulldown assays or knockout studies. If my project results in no pairing between LTPs and ABC

transporters, it can be a possible clue to rework the current theory of how small hydrophobic molecules are transported in plants.

Datasets

- What datasets will you use?

I will use the ATTED-II co-expression data for arabidopsis as well as the mapped protein-protein interactions from BioGRID.

- Where are these datasets from (databases, publications, etc.)?

Biological General Repository for Interaction Datasets (BioGRID), which is a curated database of genetic, protein, and chemical interactions as well as post-translational modifications. ATTED-II is a plant coexpression database created from numerous DNA microarray analysis studies.

- What exactly do these datasets contain and in what format?

The “BIOGRID-ORGANISM-3.5.181.tab2.zip” file that I downloaded contains the protein-protein interactions of all organisms but I will specifically be using the file for Arabidopsis: “BIOGRID-ORGANISM-Arabidopsis_thaliana_Columbia-3.5.181.tab2”. This file is formatted as a Tab 2.0 Delimited Text file and contains all interaction and associated annotation data.

The ATTED-II data seems to be a rma.mrgeo.d file. Specifically, the folder that I downloaded, “Ath-mB.v17-08.G20819-S16033.rma_combat.mrgeo.d,” contains multiple files, with each being a coexpressed gene list for that specific query gene. Each file has the gene ID, a mutual rank of the gene (MR), and the Pearson’s correlation coefficient (PCC) of each gene to a given gene. The creators of the data recommended that MR be used rather than the PCC for data analysis.

- Computational methods/approach

Because of the fact that the data is already cleaned up in that the MR and PCC values are already given, I think I should be able to simple python packages for all my computational needs. If needed, I will need to talk to you, Arjun, to get software that I should use instead of the tools I learned in CMSE 201.

- What are the analytical/computational methods you are planning to use?

Because of the fact that the data is already given to me with MR and PCC values, I simply plan to go through an enrichment test using the MR values of the ATTED and the connected nodes of a specific distance away on the protein-protein map. In order to

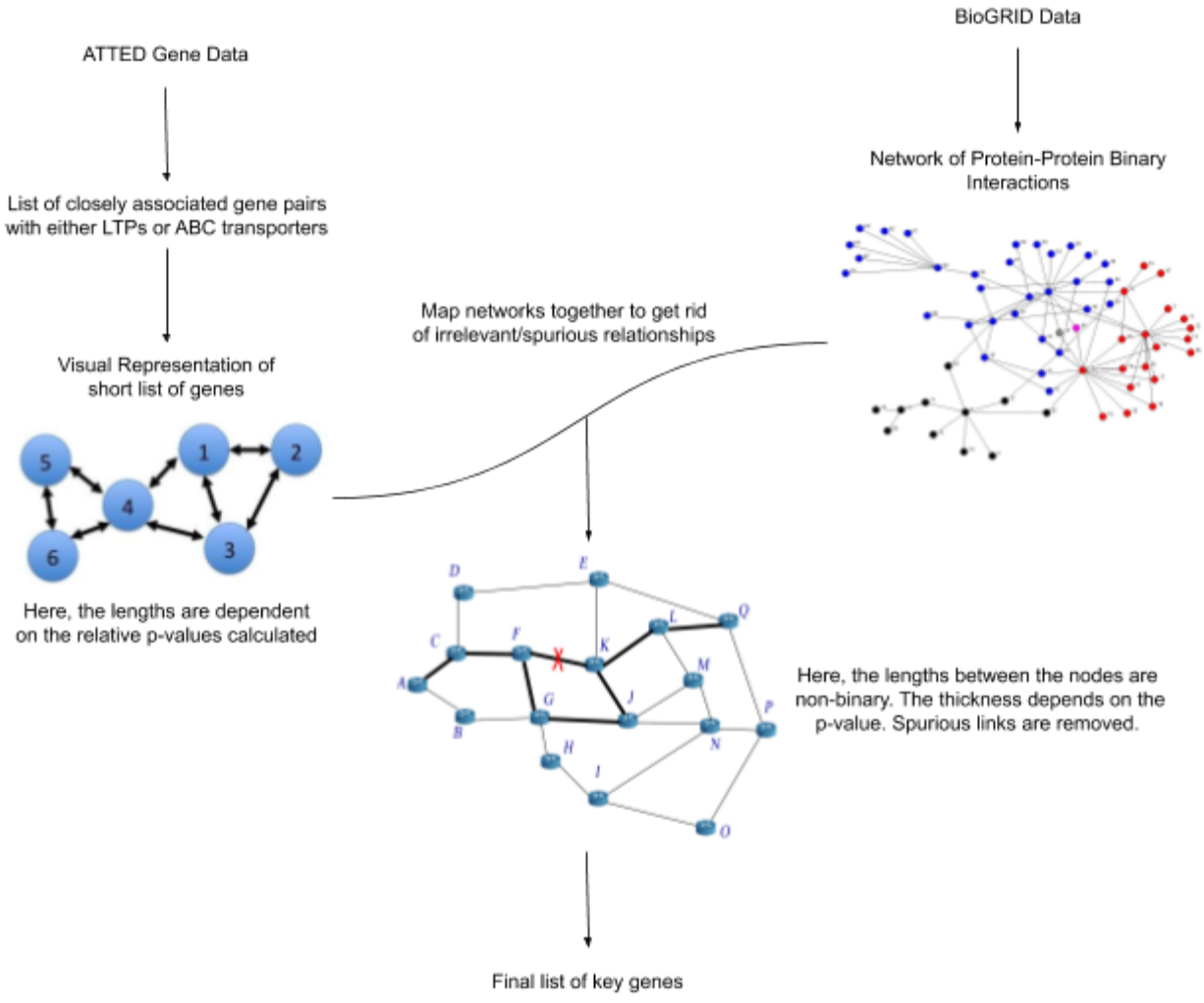
simplify, I might need to combine similar genes to make modules instead of treating each gene as a single node to save computing power.

- What are the specific software you will use?

Specifically, I will be using Numpy, Networkx, and Matplotlib libraries to work with the BIoGRID data and make a non-boolean network of the important genes.

I'm not sure how to approach the ATTED dataset. I will likely use the Networkx library for the data-mapping here as well.

- Include a detailed flowchart of your approach here.



Evaluation plan

- How will you evaluate the results that you get?

    - Think in terms of how to test if a) your approach is working correctly without errors and b) your results make quantitative/biological sense and are meaningful.

I plan to evaluate my approach to see if my code is working properly by trying to find existing papers using the same database and used a similar approach to identify a list of genes. Then, I could try inputting the same data as in the paper to see if I get the same output that the paper mentioned.

In trying to make sense of the biological sense of the list of genes that my program outputs, I would consult with Dr. Benning to check if the genes mentioned/proteins of the genes were mentioned in possible transport mechanisms by any existing papers or any if any genes similar to those identified have been characterized in a similar roles in other organisms. I think the results will be tested in a biological sense inherently in the program by cross-referencing it with the protein-protein map that has already been characterized.

Potential challenges & alternative approaches

- What are some assumptions you are making that could turn out to be not valid?

I think the major assumption I am making is that there is a genetic component in the transport of the small hydrophobic particles. I don't expect this assumption to fall apart based on the papers that I have read so far. An assumption that could turn out to be not valid is that there is only a linear-like relation between any two genes. That is, there are no instances where the association between any two given genes have a non-linear relation or a clustering-type relation. Another problem that may come up would be that the genes have a direct association with proteins in the protein-protein interaction map.

- What are some potential limitations of your dataset or approach that might prevent you from achieving your aforementioned goals?

According to Sipko van Dam, et. al, a limitation of co-expression analysis on the splice variant level is the introduction of biases because it is difficult to determine which splice variant is expressed if multiple splice variants share the same expressed exon.

- What will you do as alternatives if you hit those limitations?

To be completely honest, I'm not sure that I have enough experience to go around these errors. At this point, I would have to go to you, Arjun, and ask if I could do any troubleshooting.

Specific milestones

- What is the list of specific results/outcomes you will work towards?

1. Download the datasets and figure out how to load the data in a programmable manner for Python.
2. Create a few preliminary charts using the Pandas library in Python
3. Create a network map of genes using the ATTED data.
4. Map co-expression data onto the protein-protein network to convert the binary edges to floats, resulting in a network with lengths of varying probabilities (0-1).
5. Use the updated network for some sort of enrichment analysis from the original ATTED data.
6. Find a list of 50 gene-pairs to study in the future.

References:

Databases:

Takeshi Obayashi, Yuichi Aoki, Shu Tadaka, Yuki Kagaya, Kengo Kinoshita, ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of the Statistical Property of the Mutual Rank Index, *Plant and Cell Physiology*, Volume 59, Issue 1, January 2018, Page e3, https://doi.org/10.1093/pcp/pcx191

Obayashi T., Hayashi S., Saeki M., Ohta H. & Kinoshita K. (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Research* 37,D987–D991.

Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, Mike Tyers, BioGRID: a general repository for interaction datasets, *Nucleic Acids Research*, Volume 34, Issue suppl_1, 01 January 2006, Pages D535–D539, https://doi.org/10.1093/nar/gkj109

Other:

Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, João Pedro de Magalhães, Gene co-expression analysis for functional classification and gene–disease predictions, *Briefings in Bioinformatics*, Volume 19, Issue 4, July 2018, Pages 575–592, https://doi.org/10.1093/bib/bbw139

Björn Usadel, Takeshi Obayashi, Marek Mutwil, Federico M. Giorgi, George W. Bassel, Mimi Tanimoto, Amanda Chow, Dirk Steinhauser, Staffan Persson, Nicholas J. Provart, Co-expression tools for plant biology: opportunities for hypothesis generation and caveats, *Plant, Cell and Environment*, Volume 32, Issue 12, December 2009, Pages 1633–1651, https://doi.org/10.1111/j.1365-3040.2009.02040.x

Serin Elise A. R., Nijveen Harm, Hilhorst Henk W. M., Ligterink Wilco, Learning from Co-expression Networks: Possibilities and Challenges, *Frontiers in Plant Science,* Volume 7, 2016, Pages 444, https://doi.org/10.3389/fpls.2016.00444

Useful Links

http://atted.jp/data/locus/818558.shtml