CoreNET

A Bioinformatics Tool for Cross Species Network Analysis

Author : Gregory Hamilton

Mentors: Dr. Manpreet Katari and Dr. Gloria Coruzzi

May 3, 2016

A Thesis in fulfillment of the Masters in Biology Degree.

Table of Contents

Acknowledgements

I would like to thank my mentors Dr. Manpreet Katari and Dr. Gloria Coruzzi for their assistance and support during my thesis work. The creation of the Network Tools 2.0 suite of programs and the CoreNET tool would not have been possible without them. Dr. Katari's guidance as both a professor and mentor allowed me to go from knowing very little about Bioinformatics and programming to becoming a full-fledged Bioinformatics Developer in the course of my Master's work at NYU. I'd also like to thank the entire Coruzzi Lab for helping to create an environment where I could succeed and create the tool I set out to make.

Additionally I'd like to thank my friends, family and partner, Nadia, for supporting me through my Master's work. Without their support I would not have been able to finish my Master's thesis nor would I be moving forward to PhD program. I'm truly grateful for all their help.

Summary

Complex biological networks that provide the underpinnings to many of the questions scientists are attempting to address. Due to limitations of the experimental methods for deriving these interactions, there is a large gap in our knowledge base for these networks. To fill that gap a large variety of methods have been developed to predict these network from expression data. Yet, there are relatively few tools for translating these networks from model species to the agronomically important crops.

In service of that need CoreNET, a Python based tool that allows for rapid Cross Species Network Analysis, was developed. CoreNET accepts normalized expression data, homology data and known network data from two species of interest as input and returns conserved cross species network modules for each of the input species. In addition to developing CoreNET, VirtualPlant Networking Tools 2.0 was developed. Network Tools 2.0 is a suite of tools for storing, querying and predicting MultiNetworks. The tools are flexible/efficient and can be executed independently of the VirtualPlant platform. The two major parts of Network Tools 2.0 are the MultiNetwork Storage/Query and the Correlation network prediction tools. The MultiNetwork Storage/Query Tool allows for a user to store SIF format MultiNetworks and quickly Query the stored data. The Correlation Network prediction tool takes normalized expression data and predicts a Pearson pairwise correlation regulatory network.

To demonstrate the power of the CoreNET tool, three case studies were performed. The first one compares Arabidopsis and Rice Nitrogen regulatory networks, an update to previous work done in the Coruzzi Lab (Orbtello et al. 2015). The second compares Arabidopsis and Maize Nitrogen regulatory networks using two previously published studies (Gutierrez et al 2008 and Yang et al 2011). The focus of both studies are to uncover evolutionarily conserved Nitrogen Regulatory Networks. The third case study looks combines the data from the first two studies to find the highly conserved modules all three species.

Introduction

Biology at its most basic level is a study of complex molecular interactions inside a living organism. While it is important to understand the mechanics behind the individual interactions, the field of biology has expanded its gaze to include how these interactions play a combinatorial role in networks that are critical to biologically relevant questions. One aspect of Systems Biology is the study of these complex networks or systems that are critical for understanding everything from how to predict breast cancer progression (Chaung et al. 2006) to how crop species respond to nitrogen (Yang et al. 2011). While gains have been made in developing experimental methods to better elucidate the complex interactions that are the basis of these networks, we are still a long way off from having the full picture of what interactions are actually occurring and how that is resulting in a particular phenotype.

To fill this void, computational biologists have begun developing algorithms to predict these Gene Regulatory Networks (GRNs) or interactomes. Typically, these algorithms utilize experimentally determined interactions and gene expression data, either from RNA-seq or microarrays, to reverse engineer the network. A wide variety of models have been developed due to the importance of solving this critical issue and a yearly conference, Dialogue on Reverse Engineering Assessments and Methods or DREAM for short, is held to assess the state of the currently available algorithms. One of the earliest and still most commonly used correlation methods is Pearson, which can be used to predict a relationship between two genes based on their expression patterns across a series of experiments. While this method is tried and true, it is not without its drawbacks. More complex and robust algorithms have been developed to tackle this problem. These algorithms can be categorized into particular groups, such as Mutual Information, Bayesian, Random Forest, Differential Equation, and a possible combinatorial model utilizing aspects from each of the

previous groups (Marback et al 2012). While we have begun to solve this problem, there are additional hurdles Systems Biologists face.

One such hurdle lies in translating lab results to the field. Often times it can be tough to translate experimental data to something that can be utilized outside of the lab. This is common in all fields of biology, whether it is attempting to translate an Arabidopsis experiments into something that leads to improved crop yields or translating a finding in a genetically modified mouse model for a disease into a possible treatment in humans. With that in mind, a new set of Systems Biology studies have begun to spring up that focus on integrating data from a model species that has been well studied with data from the species of interest. Often these "Cross Species" studies, attempt to generate Gene Regulatory Networks and then cross the generated networks to find conserved network modules utilizing homology data linking the two species. These studies have provided some really important insights, such as; uncovering a conserved Nitrogen Regulatory Network between two evolutionarily distant species, *Arabidopsis thaliana* and *Oryza sativa,* (Obertelloet al.2015) or uncovering conserved master regulators, FOXM1 and CENPF, as drivers for malignant prostate cancer (Aytes et al. 2014). Although these studies are on the rise and are of great importance there is a lack of bioinformatics tools to assist researchers in this sort of study.

With that in mind, we have developed CoreNET, a tool aimed at assisting researching in performing cross species gene regulatory network analysis. The tool aims to make this sort of studies easier and more accessible by performing the predictive Gene Regulatory Network creation for both species and then crossing those two networks to return only the conserved interactions. As a novel addition, the tool can also integrate known network information into the predictive networks, creating more robust and complex networks.

Introduction

In concert with developing CoreNET, an additional set of tools for storing, querying and creating correlation networks were created, Network Tools 2.0. These tools aim to allow biologists to have the ability to local store and query their own custom networks as well as create correlation networks from expression matrices. Both sets of tools, CoreNET and Network Tools 2.0, aim to be lightweight enough that they can be run locally as well as have the ability to scale to larger projects and be run on a HPC cluster. Lastly, both sets of tools will be integrated into the next update to VirtualPlant (Katari et al. 2010).

Aims

1) Develop Virtual Plant Networking Tool 2.0
    a. The goal of developing this tool set was to create a set of tools for storing, querying and predicting MultiNetwork data that can be utilized with and without the VirtualPlant platform.
    b. This tool can allow any researcher with minimal programming skills to store their own custom networks and perform basic MultiNetwork analysis. The need to store networks locally has gone relatively unaddressed throughout the bioinformatics community.
2) Develop a Cross Species Network Analysis Tool
    a. CoreNET aims to be the first automated bioinformatics tool for identifying orthologous networks using experimental data from both, the model and target, species.
    b. It allows researchers to analyze translational datasets for conserved subnetworks in a highly flexible manor.
3) Case Study 1 – Rice vs Arabidopsis Nitrogen (Revisit Obertello et al 2015

a. The purpose of revisiting this paper is twofold. First, it was the inspiration behind the creation of CoreNET so, it provides the perfect data to test the tool. Second, we have performed a new OrthoMCL run that provides better orthology grouping so we will be able to update the conserved OrthoMCL network from the paper.

4) Case Study 2 – Maize vs Arabidopsis Nitrogen Regulatory Network
   a. Utilizing publically available data from two previously published papers (Guterriez et al 2008. and Yang et al. 2011), we will attempt to uncover conserved nitrogen regulatory networks between the two evolutionarily distant species.
5) Case Study 3 – Arabidopsis x Maize x Rice
   a. Using the Arabidopsis and Rice data from case study one and the Maize data from case study two we will perform a three species analysis using CoreNET. The goal is to uncover highly conserved network modules for all three species by crossing them with one another and then intersecting the conserved networks.

VirtualPlant Network Tools 2.0

The new Networking tools for virtual plant were developed fully in Python 2.7 (Python Software Foundation) and are able to run independent of the Virtual Plant platform. All of the parts of the new tool are species agnostic, meaning they can be used for any species (including those outside of the Virtual Plant Platform. The new tool set has two components a MultiNetwork module and a Correlation Network Prediction module. For a detailed documentation of Network Tools 2.0 refer to the supplementary material.

The MultiNetwork module contains two key elements, MultiNetwork data storage and MultiNetwork query. Utilizing a SQLite database file structure, the MultiNetwork data storage program accepts standard SIF files, a basic flat file format for storing network information, for input and stores the interactions in the SQLite database. To enhance the power of the tool we require all interactions to be uploaded in a standard format that categorizes the interaction types. We currently have five interactions types protein:protein, metabolic, regulatory, miRNA, and modify. All interaction types except, protein:protein and correlation edges, are considered directional. The database has a simple schema (Fig. 1) that allows for rapid database creation and querying. Additionally, it allows the database to highly portable and lightweight.

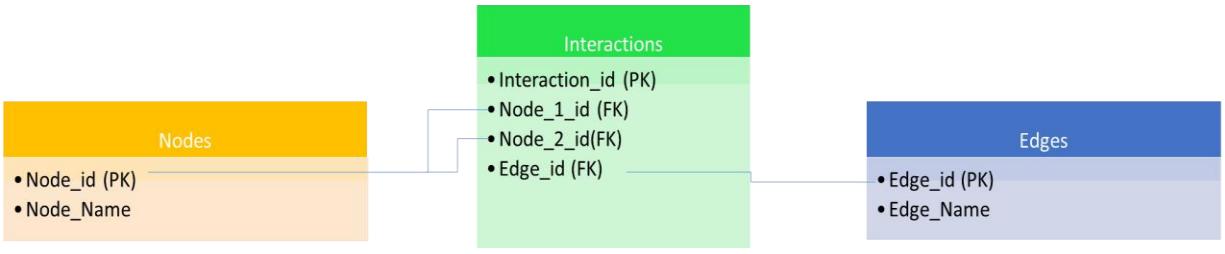

Figure 1The SQLite Database Schema for MultiNetwork storage

The database is created with the SQLalchemy and SQLite3 Python packages (Bayer M SQLalchemy). SQLalchemy allowed for intuitive mapping of database objects to python objects and SQLite3, a built in python package, allows for direct interaction with SQLite databases. The MultiNetwork data storage tool can be utilized outside of Virtual Plant to create custom MultiNetwork databases for any species of interest. The MutliNetwork Storage program is made up of two scripts. The first Initialize_Network_DB.py initializes the network database and Insert_Netowrk_Data.py Inserts SIF file data into the database. The insert provides The MultiNetwork Query program accepts a list of genes, 1 or 0 (for 1 or 0 hop interactions) and an optional edge list. Then queries your database for either zero hop, interactions where both nodes are contained in your input list, or one hop interactions, only one of the nodes needs to be contained within your input list and will filter the interactions based on the optional edge list you provide. The output is in the form of a SIF file. A single script performs the Network query, Network_Query.py. All parts of the MultiNetwork tool are currently limited to running in command line, an example of how to run these programs can be found in figure 2.

```
PS C:\Users\Greg Hamilton> python .\Initialize_Network_DB.py -d Ex.sqlite
PS C:\Users\Greg Hamilton> python .\Insert_Network_Data.py -d Ex.sqlite -i .\newKegg.sif
.\newKegg.sif
Edges Added : 4
Nodes Added :  2610
Interactions Added :  11197
PS C:\Users\Greg Hamilton> python .\Network_Query.py -d Ex.sqlite -g .\1000genes.txt -o Test
PS C:\Users\Greg Hamilton>
```

Figure 2- Command Line example of running the MultiNetwork Programs

Methods

The Correlation Network Prediction tool accepts normalized expression data and returns a correlation network in the form of a SIF file. Ideally, the expression data provided will only contain the differentially expressed genes but if it is the full expression matrix you can provide a gene list to filter the expression data for only the differentially expressed genes. Additionally, if you do not provide a gene list the full expression set will be run through the correlation algorithm. Correlation Network Prediction performs Pearson Pairwise correlation and only returns the correlations that meet a user chosen p-value cutoff, the tool defaults to a cutoff of < 0.05. Pearson Correlation is done utilizing a custom function built with the help of Akshay Jain, a former Computer Science Masters Student who developed the core of Virtual Plant 2.0. The tool also utilizes two python packages, NumPy (Van der Walt, S. et al. 2011) and SciPy (Jones, E. et al 2016). NumPy allows for the expression data to be stored in a custom array object, allowing for rapid iteration through the object when performing correlation. SciPy is utilized to generate a P-value corresponding to each pair-wise correlation. Once the correlation network is generated the user can choose to query their MultiNetwork Database for zero hop interactions based on the genes that are the nodes in the correlation network to create a combined network. Alternatively, the user can choose to intersect the correlation edges with predicated regulatory edges based on TF Binding motifs; such as, those from the AGRIS database. This will result in a regCorr or regNegCorr edge type in the output SIF file. The Correlation Network Prediction program is a single script, Network_Correlation.py, which can only be run in command line outside of VirtualPlant 2.0 platform. Lastly, the tool was built to allow parallel computing for increased speed but this functionality is still in active development.

CoreNET

The CoreNET tool was also built in Python 2.7 and can run independent of the Virtual Plant platform. CoreNET was built in a modular fashion to allow for rapid improvement/iteration and so the individual parts can be utilized as their own tools. CoreNET has four fundamental modules, Network Tools, Orthology Tools, Network Prediction, and Network Cross. For input CoreNET accepts normalized expression data, known MultiNetwork data and Orthology mapping data between the two species. The workflow can be viewed in figure 3.

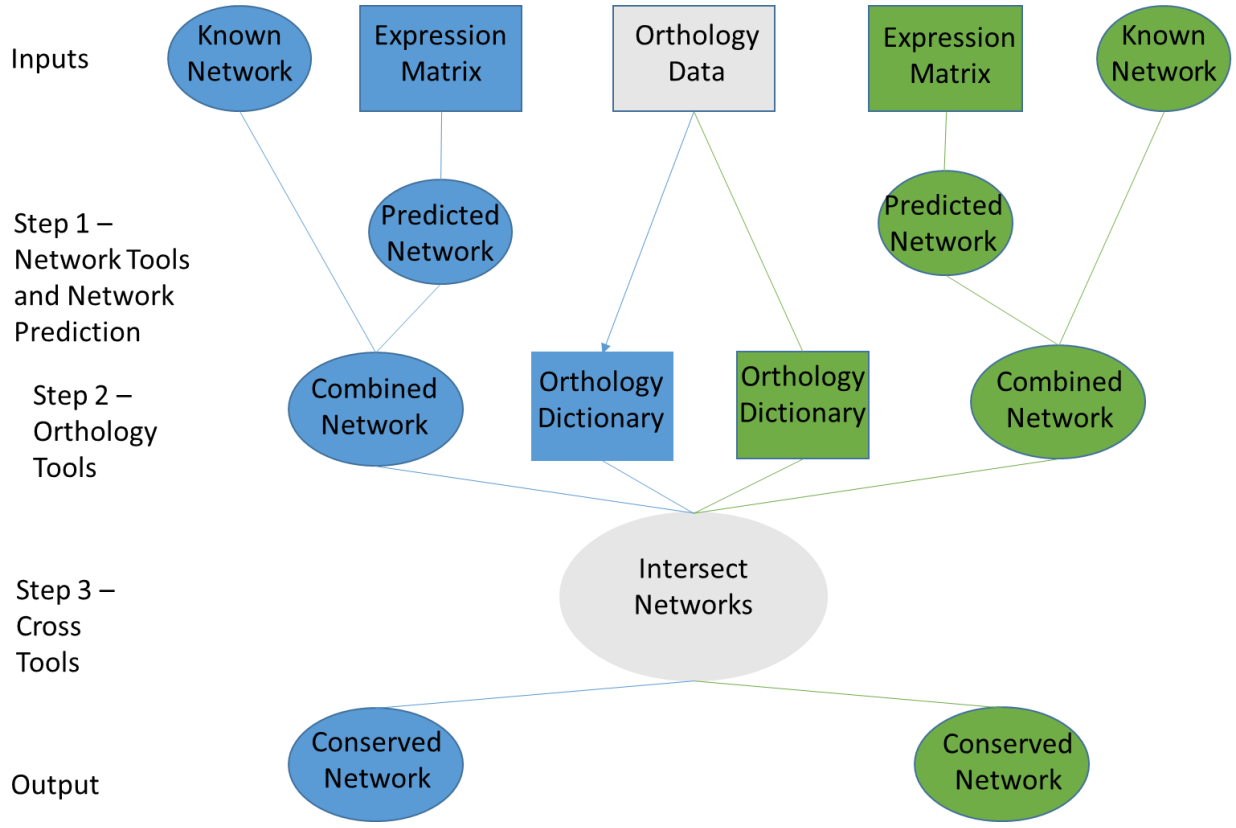

Figure 3 - CoreNET workflow

Network Prediction currently contains only one prediction type, Pearson Correlation, but there are plans to add additional prediction methods with a future update. This utilizes many of

the same functions/scripts as the Correlation Network Prediction module of Virtual Plant Network Tools 2.0.

Network Tools are a series of functions for working with network data. This includes functions for pulling zero hop interaction network either from a SQLite database and/or known network SIF files. This is done by first pulling a gene list from the predicted network nd then filtering either a SIF file or querying the database for the zero hop interactions based around the gene list. It also has functions for creating a union of known and correlation/predicted networks for a combined network. Lastly, it has a function for outputting the custom dictionary style network object to a SIF file.

Orthology Tools contains functions for working with orthology data in the form of either an OrthoMCL output or a SIF file style (an Example of this style can be found in the supplementary material, OrthologyExample.SIF). The tool pulls orthology data from those two forms utilizing the gene list, created by Network Tools. It can then create a union of the orthology data from the two styles of input in the form of a single orthology dictionary for each species. This is done because key look ups will allow the algorithm the crosses the networks to lookup the orthologs at a faster rate.

The CoreNET tool takes the combined networks from the two species and crosses them utilizing the orthology data. It first filters both combined networks down to only contain interactions that have genes whose orthologous pair also has interactions in the other species' combined network. It then intersects the two networks using the orthology dictionary generated

by orthology tools. Regulatory and Modification edges are treated as having direction while Correlation/Predicted, protein:protein, and Metabolic edges are treated as directionless. While many metabolic edges have directionality, due to the inconsistency of metabolic input data we treat these edges as directionless to ensure we do not miss any in the intersect. The interactions in species one are filtered based on their types. Directionless edges then have their nodes replaced with orthologs from species two. Then checked against the combined network of species two with the directionality maintained and then again with the directionality reversed. If the interaction is present and of the same type in species two, it is added to the conserved network. For the edges that have direction, we only check for interactions in species 2 that have the same direction. Additionally for correlation edges we only keep the correlations that are in the same direction, for example, keep a negative correlation edge if it is a negative correlation edge in species two's combined network.

Depending on what the user would like they can choose to output a conserved network for one or both of the species. The conserved network(s) are then output as SIF files using a function from Network Tools. This tool, like the Network Tools 2.0 suite of programs, can currently only be run in command line. For a detailed documentation of CoreNET refer to the supplementary material (CoreNETManual.pdf).

Tool Benchmarking

All tools were benchmarked for time and memory usage in order to determine the limits and viability of the tools outside of the VirtualPlant Platform. Memory usage was assessed utilizing the memory_profiler Package, which is freely available on GitHub (Pedregosa F. et al.).

The memory benchmarks had to be done separately from the timing benchmarks as memory profiling slows down the computation to ensure proper memory measurements. The benchmarks were done and a 2014 Dell XPS 13 laptop with 8GB of ram, only 6.5 GB of the ram was available for use, and an Intel Core i7 CPU @ 2.00 GHz, running Windows 10. All benchmarking was performed on this laptop to demonstrate the lightweight and flexible nature of the tools.

For Network Tools 2.0, the MultiNetwork Data Storage module was benchmarked using the Arabidopsis MutliNetwork data with the addition of the predicted AGRIS database regulatory interactions, 9,000,000+ additional interactions. The Query Module was tested using randomly generated gene lists that ranged from 1,000 genes to 15,000 genes, in increments of 1,000. Additionally, a benchmark was performed with the full TAIR10 Arabidopsis Genome which consists of 41,671 genes. All of this was done with and without protein:protein edge filtering and for 0 and 1 hop interactions, for a total of 4 different time/memory benchmarks for each gene list. All of the gene lists used are provided in the supplemental materials.

The Network Correlation prediction tool was benchmarked using the unpublished Expression data from the Coruzzi Lab. This data consists of normalized expression values across 24 experiments for 15,528 differentially expressed genes. The tool was then run using randomly generated gene lists (pulling from the 15,528 genes) ranging from 1,000 to 13,000 in increments of 1,000. The default p-value cutoff was used, < 0.05. Additionally, the benchmarking was re-run with a protein:protein edge filter.

CoreNET was benchmarked while performing the first two Case Studies. With benchmark data obtained for each of the steps in the tool: Creating the combined Networks, Pulling the Orthology data and Crossing the networks. Additionally, to obtain benchmark data the tool was run with 4 variations of the maize expression data from case study two. Allowing us to capture six data points for each of the metrics. The expression matrices ranged in size from 142 genes with four experiments to 5067 genes with 72 experiments.

Homology Data

The homology data used to cross the networks was generated by Dr. Karnthi Varala. An OrthoMCL v5 (Fischer, S. et al. 2011) run was performed utilizing the following species, *Arabidopsis thailiana , Zea mays, Glycine max, Manihot esculenta, Oryza sativa ,Bracypodium distachyon, Sorghum bicolor, Physcomitrella patens, and Medicago sativa* . The data for required for the OrthoMCL run was obtained through the Phytozome v11 database (Goodstein DM et al 2012). OrthoMCL was utilized due to its ability to differentiate Orthologs and Paralogs, as well as, its ability to be run with multiple species at once. Additionally, OrthoMCL has proven to have a high specificity for identifying Orthologs meaning the result will have a lower number of false positives than other methods, such as Reverse Blast Hit analysis. The major short coming of orthoMCL is that it may miss possible orthologs due to its strict cut off. This run of OrthoMCL will also allow for all Case Studies to be performed with the same homology data. This is critical for the final case study as it will allow us to directly compare the conserved networks to one another.

Arabidopsis MultiNetwork

For the known network information, I utilized the MultiNetwork created for VirtualPlant (Katari M.S. et al 2010). This network consists of 47,000+ metabolic interactions primarily from KEGG(Kaneshisa et al., 2004) and AraCyc(Mueller et al. 2003), ~ 67,000 protein:protein interactions from AtPID (Cui et al., 2008), Bind, Braun, Calmodulin (Popescu et al., 2007), MADS BOX (de Folter et al., 2005), EckerArray (Arabidopsis Interactome Mapping, 2011), Frommer (Jones et al. 2014), literature based Geneways (Rzhetsky et al., 2004), InteractomeDirect Evidence and, and 11,000+ Regulatory interactions from the AGRIS database (Davuluri et al., 2003). The network has 26,000+ genes represented and 3,000+ metabolites. This robust network was stored in a SQLite Database, Arab.sqlite, using Network Tools 2.0. This database was queried by the CoreNET tool in both case studies to create the Arabidopsis combined networks.

Rice MultiNetwork

The Rice MutliNetwork used was the same as the one in Orbtello et al. 2015. It consisted of metabolic 9,625 interactions from RiceCyc (Zhang et al. 2010). 1,173 experimentally derived protein:protein interactions from the PRIN database (Gu H.B. et al 2011) and the Rice Kinase Database (Dardick, C. et al 2007). Lastly, due to the limited Network information available for Rice, 1,258,608 computationally predicted protein:protein interactions from the PRIN database and the Rice Journal database were included. The Rice MultiNetwork was placed into a SQLite database, RiceMultiNet.sqlite, using NetworkTools 2.0.

Maize MultiNetwork

The Maize MultiNetwork consisted of 1,408,146 interactions. There were 2,762,560 protein:protein interactions pulled from the PPIM database (Zhu, G. et al 2016) , which contained predicted interactions and experimentally determined interactions. All low confidence predicted protein:protein interactions were discarded leaving only the experimentally determined, medium and high confidence predicted interactions which totaled 1,387,035. Metabolic interactions from CornCyc(Zhang et al. 2010), 21112 total interactions, were also included. The Maize MultiNetwork was placed into a SQLite database, MaizeMultiNet.sqlite, using NetworkTools 2.0.

Case Study 1 – Revisiting Obertello et al. 2015

The data sets for the first case study come from a paper published last year, *Obertello et al. 2015.* The data consists of two Microarray experiments, one for *Oryza sativa* and the other for *Arabidopsis Thaliana*. Each species were grown for 12 days in basal MS salts with 0.5 mM ammonium succinate and 3 mM of sucrose at a pH of 5.5. 24 hours prior to treatment the plants were transferred to media containing only basal MS salts. Then the experimental condition plants were treated with 20 mM KNO3 and 20 mM NH4NO3 while the control plants were treated with 20 mM KCl. After treatment the roots and shoots were harvested for Affymetrix Microarray analysis. The raw data can be found at accession number GSE38102 in the Gene Expression Omnibus database (Obertelloet al 2015).

To avoid any deviation from the original paper, the normalized expression data from the paper was used in the case study. The data had undergone RMA normalization and ANOVA

analysis to determine the differentially expressed genes; resulting 451 total differentially expressed genes in *O. sativa* and 1,417 differentially expressed genes in *A. thaliana.* The expression files were than parsed to separate the differentially expressed root and shoot genes for each species. The individual tissue types from each species were run against one another in the CoreNET tool with the new OrthoMCL data and the MultiNetworks mentioned above. A 0.05 p-value cutoff was used for the Correlation Network prediction. The combined network for each species was saved and can be found in the supplementary materials. The resulting networks for the two tissues were combined, creating a single conserved network for each species. The original study's OrthoMCL conserved network was compared utilizing a custom made python script that allows for direct comparison of two SIF files. The script has been included with the supplemental materials. Unique gene lists from the Rice and Arabidopsis conserved networks were pulled then run through VirtualPlant's BioMaps tool (Katari et al. 2010) to check for enriched Biological Process GO-terms, and KEGG pathways. Additionally, this was done to assess if the resulting conserved networks contained relevant genes and interactions.

Case Study 2 – Arabidopsis thaliana x Zea mays

The data for case study two comes from previously published works, the first an Arabidopsis microarray experiment (Guitierrez et al. 2008) and the second is a Maize microarray Experiment (Yang et al. 2011). The Arabidopsis plants were grown with a sixteen hour light cycle for fourteen days in basal MS salts with 0.5% sucrose, 0.8% BactoAgar and 1mM of KNO_3. After the initial fourteen days the plants were treated with 20mM of KNO_3 and 20mM of NH_4NO_3 for the first two hours of their light cycle. Additionally, the plants were treated with either 1mM

MSX, 10mM glutamic acid, 10mM glutamine, or some combination of the three. After treatment finished the whole plants were harvested for microarray analysis.

The Maize plants were grown for ten days in pots with either 2 mM or 20 mM of NH_4NO_3 . After ten days, the plants were supplied with 100mL of nutrient solution every other day. The 2mM plants were harvest after 28 days and the 20mM plants at 21 days. Additionally, a set of "recovery plants", which were grown in 2mM of NH_4NO_3 , were treated with 20mM of NH_4NO_3 for 2, 15 or 26 hours to recover from the low nitrogen stress. There were a total of four Maize lines used and the plants were harvested at varying time points. This made for a total of 72 conditions. The raw microarray data was obtained from NCBI's gene expression omnibus database, Accession number GSE32361.

Both datasets were MAS5 normalized and 2-way ANOVA's were performed with a 0.05 FDR correction. Only genes with the two fold expression change were kept, reducing the size of the Arabidopsis dataset to 815 unique genes. Any genes that mapped to multiple probes had their expression level averaged across the probes. The Maize dataset was analyzed five different ways looking at each of the different factors or some combination of them. For the purpose of benchmarking the tool all versions of the analysis were used except for one which looked for genes with differential expression in the recovery group based on the length of the recovery period. The four analysis used for benchmarking were, Exp1 or the full dataset analysis which analyzed the data for differential expression based on either Nitrogen treatment or harvest time resulting in a list of 9,800 probes, 5067 genes, that were either Nitrogen regulated or Nitrogen regulated and responding to time of day. Exp2 or the line 4 analysis was essentially the same as

Exp1 but removed all but one of the Maize lines, line 4, resulting in a dataset containing 5,057 + differential expressed genes. Exp4, or the low nitrogen vs recovery analysis, looked for differentially expressed genes between the low nitrogen, 2mM, group and the recovery group but only for the 10am harvest to control of any time effect, finding 1009 differential expressed probes which mapped to 627 genes. Lastly, Exp5, or the low vs high nitrogen analysis, looked for differentially expressed genes between the high, 20mM, and low, 2mM, nitrogen groups regardless of the harvest time or Maize line. This analysis uncovered 2554 differentially expressed genes. The initial analysis was performed by a previous master's student of the Corruzi lab, Stuti Srivastava and Dr. Manpreet Katari. For the purpose, of the case study we will focus on the conserved networks derived from running the High vs Low nitrogen dataset and the line 4 analysis Maize dataset through the CoreNET tool. Due to the size of the line 4 only analysis, it was run through CoreNET with a p-value cut off of 0.01 to limit the number of false positives. The unique gene lists were pulled from each of the conserved networks and run through the Biomaps tool on VirtualPlant to check for enriched Biological Process GO-terms, and metabolic pathways. Additionally, the conserved network were analyzed with in an iPython notebook session to generate network statistics.

Case Study 3 – Arabidopsis x Maize x Rice

The Arabidopsis expression data and the Rice expression data from Case Study 1 was used for the final comparative analysis. The Root and Shoot Tissue data were individually crossed with each of the four Maize expression datasets. The Arabidopsis data from case study 1 was used to ensure continuity between the species. Additionally, the experiment focused on only two distinct

treatment growth conditions which it shared with the Rice Expression data, High Nitrogen and Nitrogen deprivation. During this analysis, correlation networks for each of the datasets were generated using Network Tools 2.0 and they can be found in the supplemental materials. The conserved networks generated for Arabidopsis and Rice in Case Study 1 were carried over. A p-value of 0.05 was used for the correlation Network prediction when running CoreNET and the same OrthoMCL file mentioned above was used. The separate tissue networks were combined forming single union networks for each species comparison. Only unique interactions were kept when combining the networks. This resulted in 18 union networks. Network statistical analysis was performed on the 18 union networks. The statistical data can be found in the supplementary material. All of the Maize/Arabidopsis union networks were visualized and had their unique gene's characterized by the BioMaps tool. This was done a follow up to Case Study 2. The Arabidopsis data from Obertello et al. is better suited for comparative analysis with the Maize data.

The Maize/Rice and Maize/Ara union networks were than intersected with the Ara/Rice conserved network for to produce nine intersect networks. The networks were intersected to find the highly conserved interactions. Network statistical analysis and Cytoscape visualization was performed on these intersect networks. The unique gene lists were pulled from each of the networks and analyzed with VirtualPlant's BioMaps tool for enriched Biological Process GO terms, PlantCyc pathways, and KEGG pathways. Intersect Networks and the full analysis can be found in the supplemental materials.

Benchmarks

The MultiNetwork insert and Query Results (fig 4) show that the tool can be utilized on essentially any modern laptop. The timing and peak memory for the MutliNetwork tool shows linear growth.

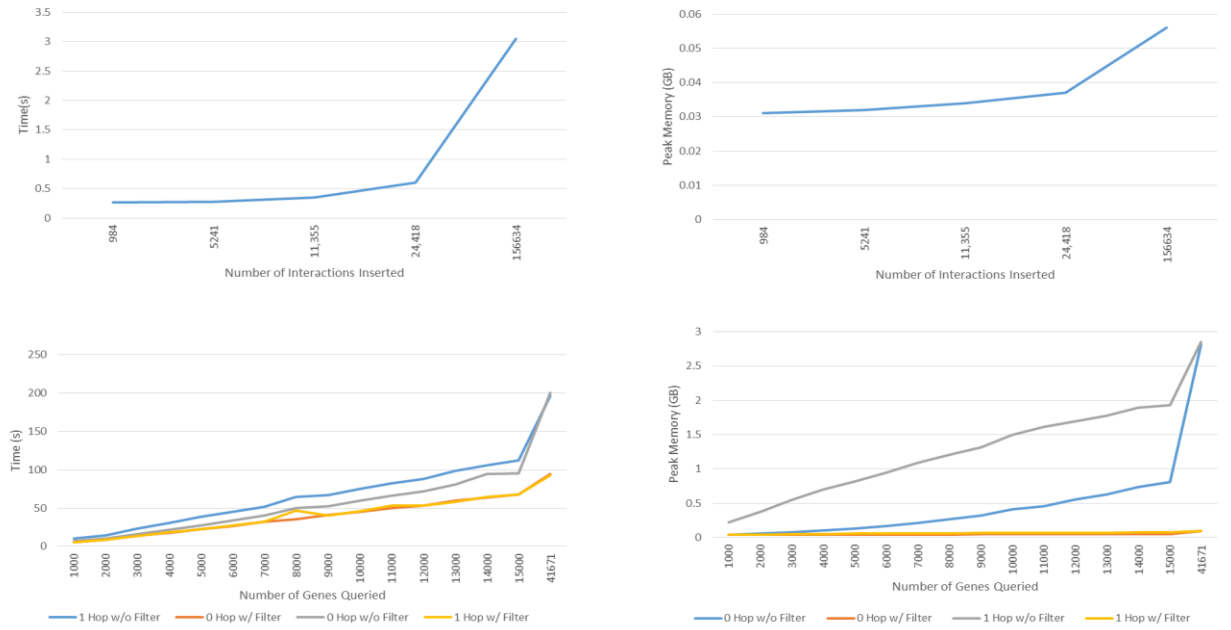

Figure 4 - The Benchmark results for the MultiNetwork Tool - A & B - Timing and Peak Memory for the MultiNetwork Inesrt C & D - Timing and Peak Memory for the MultiNetwork Query

The time and memory result for inserting AgrisPred interactions where left out of the figure because it consisted of significantly more interactions that anything previous, 9,787,820 interactions. This took 357.46 seconds and had a peak memory usage of 2.3 GB, which is still well within the capabilities of a modern laptop. The resulting Network SQLite database was only 410 MB. The memory usage for the MultiNetwork query was significantly lower for the runs with the edge filter, this is because the tool didn't load any interactions that did not match the specified

edges. The time savings for the filtered vs non-filtered runs was not significant because the query itself makes up the largest portion of time used by the tool. The Peak Memory is significantly lower for the zero hop runs because the tool does not load any of the interactions into memory where both nodes, if they are genes, are present in the input gene list. Lastly, the peak memory and timing intersect for the respective zero and one hop runs when querying the full genome because we are pulling all available interactions.

The Network Correlation Benchmark Results (figure 5) indicate that the tool can be utilized without high powered cluster computing up to an expression matrix that contains 12,000 genes. Any expression matrix larger caused the program to terminate as it hit the memory limit of the laptop utilized for the benchmarking.

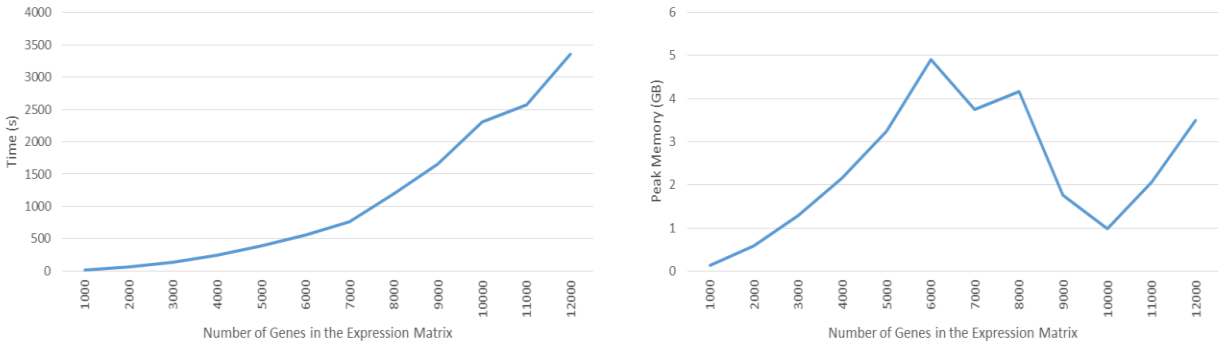

Figure 5 - Network Correlation Tool Benchmarks

Interestingly, after 7,000 genes the peak memory follows an irregular pattern. This can be explained because by the nature of the Python Garbage collector which will only activate if required.

The CoreNET Tool Benchmark results (fig 6) show the tool is lightweight enough to be able to run on most modern laptops. At the upper limit of the test, the tool was able to analysis and cross a Maize dataset containing 5067 genes with data for 72 experiments with an Arabidopsis dataset containing 815

genes with data for 8 experiments in slightly more than six minutes with a peak memory of 2.938 GB, less than the amount of ram in most modern cell phones.

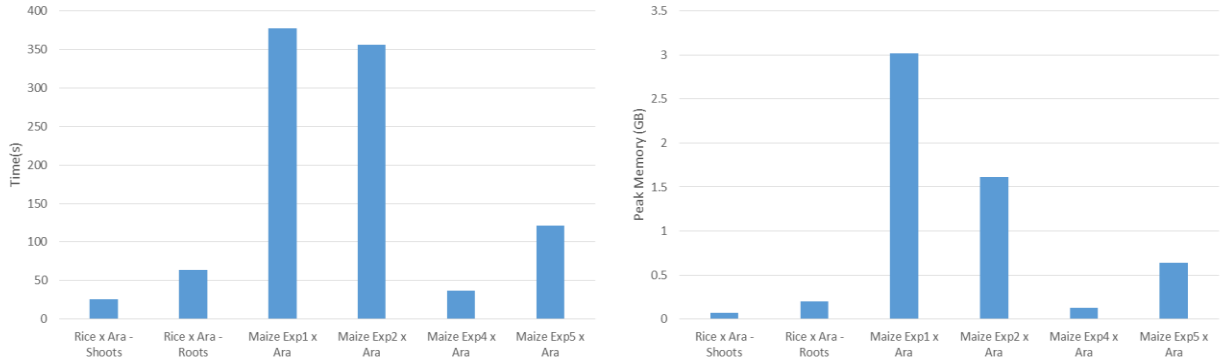

Figure 6 - Benchmark Results for the CoreNET tool

Case Study 1 – Revisiting Obertello et al. 2015

The Arabidopsis conserved network (fig 7) contains 65 unique genes, nine more than the original paper, and 23 metabolites with 573 total interactions, 533 were unique. The difference in unique and total interactions is due to a number of the conserved correlation edges being present in both roots and shoots. The same 23 metabolites were found in the Rice conserved network (fig 7). The rice network contained 58 unique genes, ten more than the Obertello et al. paper. The rice conserved network had a total of 422 total interactions and 399 unique interactions, 16 more than the Obertello et al. paper. The additional genes and interactions

present are a direct result of the better orthology grouping with the more recent OrthoMCL run. While the increase in the number of unique Genes is not surprising, the increase in the number

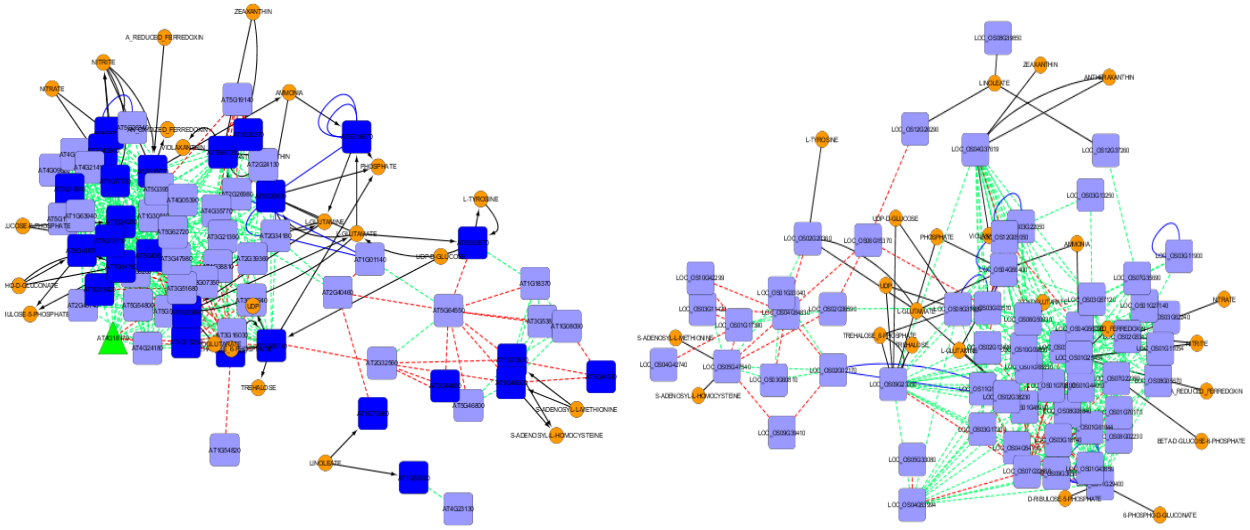

Figure 7- Cytoscape Visualization of the conserved networks with Arabidopsis on the left and Rice on the right

of interactions is. The CoreNET Algorithm has stricter requirements for what is considered a conserved interaction than the original workflow. This is clear when examining the types of interactions conserved, the conserved rice network generated by CoreNET contained ten protein:protein and 37 metabolic interactions while the initial workflow contained twenty protein:protein and 46 metabolic interactions.

A BioMaps (Table 1) analysis of the 65 unique Arabidopsis genes reveled enrichment for a number nitrogen response related GO terms and KEGG Pathways. This result was to be expected as the initial experiments were assessing expression levels in response to Nitrogen treatments. The full results for the Arabidopsis BioMaps analysis can be found in the supplementary materials along with the BioMaps results for the unique rice genes. Additional functional analysis revealed that the WRKY transcription factor was among the list of conserved genes in both species.

GO ID	Term	Observed Frequency	Expected Frequency	p-value
GO:0042126	nitrate metabolic process	4 out of 61 genes, 6.6%	10 out of 24961 genes, 0%	4.70E-06
GO:0042128	nitrate assimilation	4 out of 61 genes, 6.6%	10 out of 24961 genes, 0%	4.70E-06
GO:0071941	nitrogen cycle metabolic process	4 out of 61 genes, 6.6%	11 out of 24961 genes, 0%	4.70E-06
GO:0042221	response to chemical stimulus	18 out of 61 genes, 29.5%	1892 out of 24961 genes, 7.6%	3.22E-05
GO:0050896	response to stimulus	25 out of 61 genes, 41%	3689 out of 24961 genes, 14.8%	4.37E-05
GO:0010035	response to inorganic substance	9 out of 61 genes, 14.8%	507 out of 24961 genes, 2%	0.000231
GO:0010167	response to nitrate	3 out of 61 genes, 4.9%	20 out of 24961 genes, 0.1%	0.00111
GO:0009611	response to wounding	5 out of 61 genes, 8.2%	145 out of 24961 genes, 0.6%	0.00136
GO:0006809	nitric oxide biosynthetic process	2 out of 61 genes, 3.3%	4 out of 24961 genes, 0%	0.00288
GO:0046209	nitric oxide metabolic process	2 out of 61 genes, 3.3%	4 out of 24961 genes, 0%	0.00288
GO:0009051	pentose-phosphate shunt, oxidative branch	2 out of 61 genes, 3.3%	6 out of 24961 genes, 0%	0.00486
KEGG ID	**Term**	**Observed Frequency**	**Expected Frequency**	**p-value**
910	Nitrogen metabolism	6 out of 20 genes, 30%	42 out of 3011 genes, 1.4%	1.09E-05
ENERGY METABOLISM	Energy Metabolism	10 out of 20 genes, 50%	322 out of 3011 genes, 10.7%	0.000177
METABOLISM	Metabolism	20 out of 20 genes, 100%	1741 out of 3011 genes, 57.8%	0.000177
30	Pentose phosphate pathway	5 out of 20 genes, 25%	53 out of 3011 genes, 1.8%	0.000194
480	Glutathione metabolism	4 out of 20 genes, 20%	57 out of 3011 genes, 1.9%	0.00328
CARBOHYDRATE METABOLISM	Carbohydrate Metabolism	9 out of 20 genes, 45%	466 out of 3011 genes, 15.5%	0.00813
METABOLISM OF OTHER AMINO ACIDS	Metabolism of Other Amino Acids	5 out of 20 genes, 25%	139 out of 3011 genes, 4.6%	0.00813

Table 1 – BioMaps results for enriched Biological Processes GO terms and KEGG pathways for the Conserved Arabidopsis Genes

Case Study 2

The Arabidopsis Conserved Network (fig 7) for the Line 4 dataset cross contains 282 unique genes, 8 of which are transcription factors. The Network has a total of 1241 interactions. 925 are correlation based interactions, 61 are protein:protein interactions, and 255 are metabolic interactions. The Maize Conserved Network has a similar breakdown. It contains 301 unique genes and 1893 total interactions. 1,500 of the 1,893 are correlation based interactions while, 92 are protein:protein and 301 are metabolic. The Arabidopsis Conserved Network (fig 8) for the Low vs high nitrogen group maize cross contains 150 unique genes, 5 transcription factors,

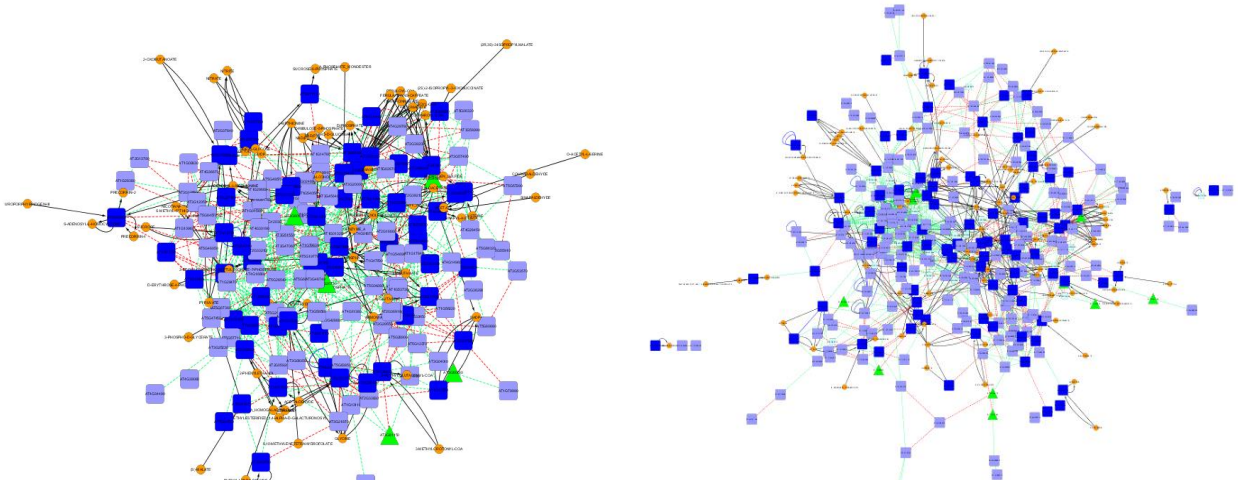

Figure 8 - The Arabidopsis Conserved Networks for the Maize Low vs high nitrogen group Cross (left) and the Maize Line 4 dataset Cross(right)

and a total of 941 interactions. The results of the BioMaps analysis found 133 enriched Biological process GO terms and two enriched KEGG pathways for the ArabidopsisXMazieLine 4 dataset Network. It found 85 enriched Biological Process GO terms and three enriched KEGG pathways for the ArabidopsisXMazieLow vs high nitrogen group Network. The enriched GO terms for both networks contained mostly Nitrogen regulation related GO terms but there were a large number of additional enriched GO terms; suggesting that the data sets may not have been ideal for the comparative analysis as we likely conserved a number interactions that were not related specifically to nitrogen regulation. The entirety of the BioMaps results for case study 2 can be found with the supplementary materials.

Case Study 3

The Arabidopsis/Maize networks (fig 9) varied dramatically in size. The Line 4 dataset cross resulted in a network that consisted of 10,940 interactions, 10,682 of which were correlation based, and 354 unique genes. The BioMaps result found a large number of metabolic

process related GO terms, 2 KEGG pathways, Metabolism and Nitrogen Metabolism were found to be enriched, and the MIPPS functional categories found to be enriched varied but a number of Nitrogen metabolism functions were found to be enriched. Additionally, 21 Transcription factors were found to be conserved. The Low vs high nitrogen group Cross contained 3381 interactions, 3266 of which were correlation based. There were 193 unique genes present in the network, one of which was a transcription factor part of the B3 TF family. The BioMaps Functional analysis of the 193 unique genes, return no enriched KEGG pathways, 17 enriched Biological Process GO terms, and 13 enriched MIPPS categories. The enriched GO terms and MIPS categories were mostly related to nitrogen metabolism and metabolism in general. The final Arabidopsis/Maize union network, the Low nitrogen vs recovery group cross, contained 652 interactions and 63 unique genes with 4 transcription factors. The BioMaps analysis returned

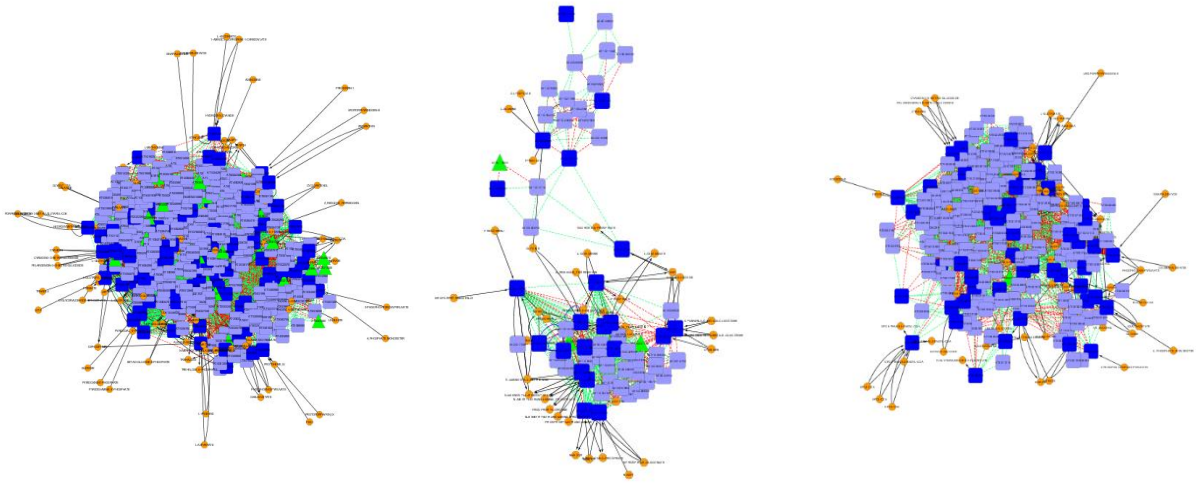

Figure 9 - Arabidopsis/Maize union networks (Line 4 dataset left, Low nitrogen vs recovery group middle, Low vs high nitrogen group right)

12 Biological Process GO terms, 3 KEGG pathways and 17 MIPPS categories. Interestingly, the vast majority of these were directly related to nitrogen regulation/metabolism. Due to this, the

intersect analysis focused primarily on the conserved networks derived from crosses with the Low nitrogen vs recovery group Maize dataset. The full network statistic data and BioMaps results data can be found in the supplementary materials.

The Arabidopsis intersect Network (fig 10), for the Low nitrogen vs recovery group cross, has 118 interactions and 20 unique genes (table 2). There are 87 interactions are correlation, 11 negative and 76 positive, 1 protein:protein interaction and 30 metabolic interactions. The BioMaps (table 3) analysis returned 8 Biological Process GO terms, 4 AraCyc pathways, 2 KEGG pathways, and 17 MIPPS categories. All of the enriched GO terms were related nitrogen regulation/metabolism. The Maize intersect network has 96 interactions and 18 unique genes. The Rice intersect network has 81 interactions with 15 unique genes.

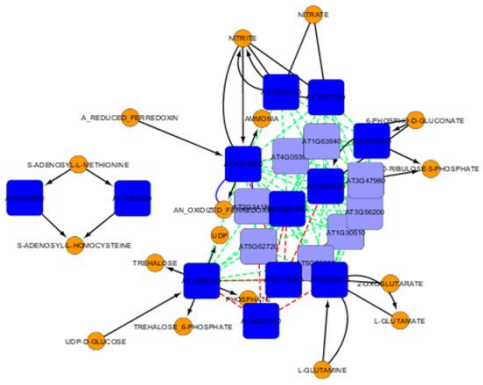

Figure 10- The Arabidopsis Intersect Conserved Network

Genes	Functional Data
At1g60140	ATTPS10, TPS10, TPS10, trehalose phosphate synthase
At1g77760	GNR1, NIA1, NR1, nitrate reductase 1
At4g15280	UGT71B5, UDP-glucosyl transferase 71B5
At5g50200	ATNRT3.1, NRT3.1, WR3, nitrate transmembrane transporters
At5g53460	GLT1, NADH-dependent glutamate synthase 1
At3g47980	Integral membrane HPP family protein
At5g40850	UPM1, urophorphyrin methylase 1
At2g34180	ATWL2, CIPK13, SnRK3.7, WL2, CBL-interacting protein kinase 13
At2g15620	ATHNIR, NIR, NIR1, nitrite reductase 1
At1g30510	ATRFNR2, RFNR2, root FNR 2
At1g64190	6-phosphogluconate dehydrogenase family protein
At4g05390	ATRFNR1, RFNR1, root FNR 1
At5g62720	Integral membrane HPP family protein
At3g56200	Transmembrane amino acid transporter family protein
At1g63940	MDAR6, monodehydroascorbate reductase 6
At1g37130	ATNR2, B29, CHL3, NIA2, NIA2-1, NR, NR2, nitrate reductase 2
At5g41670	6-phosphogluconate dehydrogenase family protein
At3g21760	HYR1, UDP-Glycosyltransferase superfamily protein
At1g48600	AtPMEAMT, PMEAMT, S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
At1g73600	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein

Table 2- Unique Genes in the Conserved Intersect Arabidopsis Genes

Of the 18 unique genes found in the Maize network, only 8 have annotation data which likely explains the very limited BioMaps result (table 3). The Maize network has annotations for all of the unique gene's present and had a similar Biomaps result(table 3) to the Arabidopsis network.

Arabidopsis BioMaps Results				
GO ID	Term	Observed Frequency	Expected Frequency	p-value
GO0042126	nitrate metabolic process	3 out of 20 genes, 15%	10 out of 24961 genes, 0%	6.83E-06
GO0042128	nitrate assimilation	3 out of 20 genes, 15%	10 out of 24961 genes, 0%	6.83E-06
GO0071941	nitrogen cycle metabolic process	3 out of 20 genes, 15%	11 out of 24961 genes, 0%	6.83E-06
GO0006809	nitric oxide biosynthetic process	2 out of 20 genes, 10%	4 out of 24961 genes, 0%	0.000235
GO0046209	nitric oxide metabolic process	2 out of 20 genes, 10%	4 out of 24961 genes, 0%	0.000235
GO0010035	response to inorganic substance	5 out of 20 genes, 25%	507 out of 24961 genes, 2%	0.000916
GO0010167	response to nitrate	2 out of 20 genes, 10%	20 out of 24961 genes, 0.1%	0.00257
GO0044271	cellular nitrogen compound biosynthetic process	4 out of 20 genes, 20%	445 out of 24961 genes, 1.8%	0.0064
KEGG ID	Term	Observed Frequency	Expected Frequency	p-value
910	Nitrogen metabolism	4 out of 9 genes, 44.4%	42 out of 3011 genes, 1.4%	6.18E-05
ENERGY METABOLISM	Energy Metabolism	6 out of 9 genes, 66.7%	322 out of 3011 genes, 10.7%	0.000546
AraCyc ID	Term	Observed Frequency	Expected Frequency	p-value
PWY-5944	zeaxanthin biosynthesis	2 out of 14 genes, 14.3%	4 out of 3216 genes, 0.1%	0.00605
PWY4FS-3	phosphatidylcholine biosynthesis III	2 out of 14 genes, 14.3%	4 out of 3216 genes, 0.1%	0.00605
PWY4FS-4	phosphatidylcholine biosynthesis IV	2 out of 14 genes, 14.3%	4 out of 3216 genes, 0.1%	0.00605
PWY-3385	choline biosynthesis I	2 out of 14 genes, 14.3%	7 out of 3216 genes, 0.2%	0.00864
PWY4FS-2	phosphatidylcholine biosynthesis II	2 out of 14 genes, 14.3%	7 out of 3216 genes, 0.2%	0.00864
Rice BioMaps Results				
GO ID	Term	Observed Frequency	Expected Frequency	p-value
GO:0006091	generation of precursor metabolites and energy	7 out of 13 genes, 53.8%	1530 out of 24036 genes, 6.4%	0.000198
GO:0042126	nitrate metabolic process	2 out of 13 genes, 15.4%	5 out of 24036 genes, 0%	0.000198
GO:0042128	nitrate assimilation	2 out of 13 genes, 15.4%	5 out of 24036 genes, 0%	0.000198
GO:0005991	trehalose metabolic process	2 out of 13 genes, 15.4%	28 out of 24036 genes, 0.1%	0.00204
GO:0005992	trehalose biosynthetic process	2 out of 13 genes, 15.4%	27 out of 24036 genes, 0.1%	0.00204
GO:0046351	disaccharide biosynthetic process	2 out of 13 genes, 15.4%	27 out of 24036 genes, 0.1%	0.00204
GO:0005984	disaccharide metabolic process	2 out of 13 genes, 15.4%	43 out of 24036 genes, 0.2%	0.00395
GO:0006112	energy reserve metabolic process	2 out of 13 genes, 15.4%	57 out of 24036 genes, 0.2%	0.00529
GO:0006118	electron transport	5 out of 13 genes, 38.5%	1292 out of 24036 genes, 5.4%	0.00529
GO:0015980	nergy derivation by oxidation of organic compounc	2 out of 13 genes, 15.4%	83 out of 24036 genes, 0.3%	0.00986
Maize BioMaps Results				
GO ID	Term	Observed Frequency	Expected Frequency	p-value
GO:0008271	ndary active sulfate transmembrane transporter ac	2 out of 16 genes, 12.5%	9 out of 16986 genes, 0.1%	0.00412
GO:0008272	sulfate transport	2 out of 16 genes, 12.5%	11 out of 16986 genes, 0.1%	0.00412
GO:0015116	sulfate transmembrane transporter activity	2 out of 16 genes, 12.5%	9 out of 16986 genes, 0.1%	0.00412
GO:0055114	oxidation-reduction process	8 out of 16 genes, 50%	2008 out of 16986 genes, 11.8%	0.00986
CornCyc ID	Term	Observed Frequency	Expected Frequency	p-value
PWY-5194	siroheme biosynthesis	2 out of 10 genes, 20%	2 out of 3691 genes, 0.1%	0.00216
NITROGEN-DEG	Nitrogen Compounds Metabolism	3 out of 10 genes, 30%	42 out of 3691 genes, 1.1%	0.00522
NITRATE-REDUCTION	Nitrate Reduction	2 out of 10 genes, 20%	13 out of 3691 genes, 0.4%	0.00931
PWY-381	nitrate reduction II (assimilatory)	2 out of 10 genes, 20%	13 out of 3691 genes, 0.4%	0.00931

Table 3- BioMaps result for the Intersect Networks

Network Tools 2.0

The Network Tools 2.0 MultiNetwork programs provide a user friendly and light weight means for storing and querying MutliNetwork data. Based on the benchmark results, researchers will be able to store and access their data locally without having to invest in an expensive server setup. Additionally, utilizing these tools requires little knowledge of programming. One current drawback is that researchers will need to be comfortable running command terminal commands. To help alleviate any potential anxieties surrounding this requirement, a detailed guide can be found in the Network Tools 2.0 documentation. With this obstacle in mind, development of a user friend front is in the development pipeline for a future update to the stand alone version of the tool.

The Correlation Network program provides a highly scalable and relatively lightweight way to perform correlation analysis of normalized expression data. The tool didn't hit any memory issues until running expression matrices that are well beyond what would typically be used of this type of analysis. With that being said, the tool can easily scale too much larger datasets. A modified version of the tool, with the Network Query portion removed, was utilized by Ren Yi (a rotation student in the Coruzzi Lab) to a perform Pearson correlation on a dataset containing 60,000 + genes and 200 + experiments of Maize data. To perform this analysis the tool was executed on NYU's High Performance Computing Cluster. The current major limitation of the tool is the time required to perform the analysis which climbs rapidly has the size of the expression matrix grows. To combat this problem, the tool was designed to be capable of running

in parallel. This will dramatically reduce the time the analysis takes but at the cost of added memory usage. This feature is currently in development and will be added with the next update to the tool. Additionally, this tool suffers from the same limitation as the MultiNetwork Programs. It requires the ability to run the program on the command line. As with the MultiNetwork Program this issue is addressed through providing a detailed guide in the Network Tools 2.0 Manual but a long term solution will be to add a front end user interface in a later update.

CoreNET

CoreNET is a unique and flexible tool that can allow Biologists to rapidly uncover conserved network modules between evolutionarily distant species. Additionally, the benchmark results quite clearly indicate that CoreNET can be run without the need for high powered computing. This will allow CoreNET to be used by researches who otherwise would not have access to such a powerful bioinformatics tool. Currently, CoreNET suffers from two limitations. The first is that it requires command line input to be run but as with the Network Tools 2.0 tools a manual with a step by step guide is provided with the supplementary material to help alleviate that problem. The second limitation is that the tool currently only relies on Pearson Pairwise correlation to create its predictive gene regulatory networks. Correlation while useful can lead to a large number of false positive interactions. To address this limitation, CoreNET was designed modularly so that additional Gene Regulatory Network prediction algorithms can be incorporated. Ideally, the tool will use a combination of a few models as a “community” style approach to predicting GRNs has proven to be the most successful (Marbach et al. 2012).

Inclusion of spearman correlation and/or adding the ability to intersect the correlated edges with regulatory edges predicted through cis binding motifs will likely be part of the next update to the tool.

Case Study Results

The first case study produced the expected result, which was a slightly larger network than the one produced for the original paper (Obertello et al. 2015). The additional interactions were due to the improved orthology mapping provided by the updated orthoMCL run (Fisher et al. 2011). As expected, the same TF, WRKY, was found to be conserved with the updated data as the one in the original study. Lastly, there were some protein:protein edges missing in the update due to the nature of how CoreNET determines conserved interactions. In the initial paper, some protein:protein interactions we kept even if they weren't present in Arabidopsis as only the correlation networks were crossed. Meaning, that any as long as the nodes in the protein:protein interaction were part of a conserved correlation interaction that protein:protein interaction was kept as well.

Case Study two produced some large "conserved networks" but the functional analysis of the unique gene lists indicated that we were "conserving" a significant number of interactions that were not Nitrogen regulated. This was likely due to the Gutierrez et al and Yang et al datasets not being ideal for comparison to one another. This became even clearer when crossing the Yang et al dataset with the Obertello et al. dataset, as the resulting networks contain primarily genes associated with Nitrogen regulation.

The third and final case study produced a number of interesting conserved networks. In particular, the intersect Network from the Ara/Maize Low nitrogen vs recovery group and Ara/Rice networks contains almost entirely genes associated with nitrogen regulation. The highly conserved sub network uncovered by this double cross analysis need to be studied further to assess its validity but could very likely contain interactions that are core to Nitrogen metabolism and Nitrogen response. While, the BioMaps results for the Maize and Rice intersected networks were more limited, they still provided some important insights. The Rice results showed enrichment for a number of Biological process GO terms associated with Nitrogen Metabolism/regulation. Additionally, the Maize results showed enrichment for KEGG pathways associated with Nitrogen metabolism/regulation. The BioMaps enrichment analysis for Maize and Rice were limited due to the limited functional annotation information for both species. The study might have also been hindered by the exclusive use of orthoMCL for homology data. By using this exclusively we may have lost a number of conserved interactions that could have been important to this system. For a follow up study, the analysis should be repeated utilizing additionally homology data; such as, homology mapping generated using the Reverse BLAST method mentioned in the Obertello et al paper. This would broaden our pool of orthologs and allow for us to capture and conserved interaction lost due to the strict cut offs imposed by orthoMCL. For example, we found conserved interactions involving the WRKY TF family in the Arabidopsis/Rice network and the Arabidopsis/Maize network; this TF was lost when creating the intersect network. WRKY transcription factors play a role in responding to biotic and abiotic stresses. The Arabidopsis/Rice network found WRKY28 , AT4G18170, to be conserved. The

Arabidopsis/Maize network found WRKY75, AT5G13080, to be conserved. Both are members of the group II-c, subset of WRKY TFs. All of the annotation information was obtained through VirtualPlant (Katari et al 2010).

Lastly, a large number of networks were generated for all of the case studies due to the relative ease at which CoreNET can be run. While basic statistical analysis was performed on all of these networks, the ones not covered in this manuscript may contain interactions that are relevant to Nitrogen regulation so further analysis should be done.

References

Arabidopsis Interactome Mapping, C. (2011). Evidence for network evolution in an Arabidopsis interactome map. Science *333*, 601-607.

Aytes, A., Mitrofanova, A., Lefebvre, C., Alvarez, M.J., Castillo-Martin, M., Zheng, T., Eastham, J.A. Gopalan, A., Pienta, K.J., Shen, M.M., Califano, A., and Abate-Shen, C. (2014) Cross-species anaylsis of genome-wide regulatory networks identifies a synergist interaction between FOXM1 and CENPF that drives prostate cancer malignancy. Cancer Cell. 25(5). 638-651.

Bayer, M. SQLAlchemy – A Database Toolkit for Python. http://www.sqlalchemy.org/

Chuang, H. , Lee, E., Liu, Y., Lee, D., and Ideker, T. (2006). Network-based classification of breast cacer metastasis. Molecular Systems Biology. 3(140). 1 -10.

Cui, J., Li, P., Li, G., Xu, F., Zhao, C., Li, Y., Yang, Z., Wang, G., Yu, Q., Li, Y.*, et al.* (2008). AtPID: Arabidopsis thaliana protein interactome database--an integrative platform for plant systems biology. Nucleic acids research *36*, D999-1008.

Dardick, C., Chen, J., Richter, T., Ouyang, S. and Ronald, P.. The Rice Kinase Database. A Phylogenomic Database for the Rice Kinome. Plant Physiology, 2007, 143(2):579-586.

Davuluri, R.V., Sun, H., Palaniswamy, S.K., Matthews, N., Molina, C., Kurtz, M., and Grotewold, E. (2003). AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC bioinformatics *4*, 25.

de Folter, S., Immink, R.G., Kieffer, M., Parenicova, L., Henz, S.R., Weigel, D., Busscher, M., Kooiker, M., Colombo, L., Kater, M.M.*, et al.* (2005). Comprehensive interaction map of the Arabidopsis MADS Box transcription factors. The Plant cell *17*, 1424-1433.

Devaiah BN, Karthikeyan AS, Raghothama KG (2007) WRKY75 transcription factor is a modulator of phosphate acquisition and root development in Arabidopsis. Plant Physiology 143: 1789–801

Dharmawardhana P, Ren L, Amarasinghe V, Monaco M, Thomason J, Ravenscroft D, McCouch S, Ware D, Jaiswal P (2013) A genome scale metabolic network for rice and accompanying analysis of tryptophan, auxin and serotonin biosynthesis regulation under biotic stress. Rice 6: 15

Ding X, Richter T, Chen M, Fujii H, Seo YS, Xie M, Zheng X, Kanrar S, Stevenson RA, Dardick C, et al (2009) A rice kinase-protein interaction map. Plant physiology 149: 1478–92

Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., Iodice, J. B., Shanmugam, D., Roos, D. S. and Stoeckert, C. J. Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups. Current Protocols in Bioinformatics. 2011 35:6.12.1–6.12.19.

Gutierrez, R.A., Stokes, T.L., Thum, K., Xu, X., Obertello, M., Katari, M.S., Tanurdzic, M., Dean, A., Nero, D.C., McClung, C.R., et al. (2008). Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1. Proceedings of the National Academy of Sciences of the United States of America 105, 4939-4944.

Geisler-Lee, J., O'Toole, N., Ammar, R., Provart, N.J., Millar, A.H., and Geisler, M. (2007). A predicted interactome for Arabidopsis. Plant physiology *145*, 317-329.

Gu, H.B., Zhu P, Jiao Y, Meng Y, Chen M (2011) PRIN: a predicted rice interactome network. BMC bioinformatics 12: 161

Gustafson, A.M., Allen, E., Givan, S., Smith, D., Carrington, J.C., and Kasschau, K.D. (2005). ASRP: the Arabidopsis Small RNA Project Database. Nucleic acids research *33*, D637-640.

Jones, A.M., Xuan, Y., Xu, M., Wang, R.S. Ho, C.H. Lalonde, S. You, C.H., Sardi, M.I., Parsa, S.A., Smith-Valle, E., Su, T., Frazer, K.A., Pilot, G., Pratelli, R., Grossmann, G., Archarya, B.R., Hu, H.C. Engineer, C. Villers, F., Ju, C., Takeda, K., Su, Z. Dong, Q. Assmann, S.M., Chen, J., Kwak, J.M., Schroeder, J.I., Albert, R. , Rhee, S.Y. and Frommer ,W.B. (2014) Border control—a membrane-linked interactome of Arabidopsis. Science. 344(6185). 711-716.

Jones E, Oliphant E, Peterson P, *et al.* SciPy: Open Source Scientific Tools for Python, 2001-, http://www.scipy.org/

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. Nucleic acids research *32*, D277-280.

Katari, M.S., Nowicki, S.D., Aceituno, F.F., Nero, D., Kelfer, J., Thompson, L.P., Cabello, J.M., Davidson, R.S., Goldberg, A.P., Shasha, D.E.*, et al.* (2010). VirtualPlant: a software platform to support systems biology research. Plant physiology *152*, 500-515.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M.*, et al.* (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic acids research *40*, D1202-1210.

Marbach, D. , Costello, J.C., Kuffner, R., Vega, N.M., Prill, R.J. , Camacho, D.M., Allison, K.R. THE DREAM5 Consortium, Kellis, M., Collins, J.J. and Stolovitzky, G. (2012) Widom of crowds for robust gene network inference. Nature Methods Analysis. 9(8). 796 – 806.

References

Mueller, L.A., Zhang, P., and Rhee, S.Y. (2003). AraCyc: a biochemical pathway database for Arabidopsis. Plant physiology *132*, 453-460.

Obertello M., Shrivastava S., Katari, M. and Coruzzi, G. (2015) Cross-species network analysis uncovers conserved nitrogen-regulated network modules in rice. Plant Physiology

Pedregosa F., and Gervais P. memory_profiler. https://github.com/fabianp/memory_profiler

Popescu, S.C., Popescu, G.V., Bachan, S., Zhang, Z., Seay, M., Gerstein, M., Snyder, M., and Dinesh-Kumar, S.P. (2007). Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays. Proceedings of the National Academy of Sciences of the United States of America *104*, 4730-4735.

Python Software Foundation. Python Language Reference, version 2.7. Available at http://www.python.org

Rohila J, Chen M, Chen S, Chen J, Cerny R, Dardick C, Canlas P, Fujii H, Gribskov M,
Kanrar S, et al (2009) Protein-Protein Interactions of Tandem Affinity Purified Protein
Kinases from Rice. PloS one 4:

Rohila JS, Chen M, Chen S, Chen J, Cerny R, Dardick C, Canlas P, Xu X, Gribskov M,
Kanrar S, et al (2006) Protein-protein interactions of tandem affinity purification-tagged
protein kinases in rice. The Plant journal : for cell and molecular biology 46: 1–13

Shannon,P. et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res., 13, 2498–2504.

Van der Walt, S. Chris Colbert and Gaël Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering, 13, 22-30 (2011), DOI:10.1109/MCSE.2011.37

Xu, G., Fan, X., and Miller, A.J. (2012). Plant nitrogen assimilation and use efficiency. Annual review of plant biology *63*, 153-182.

Yang, X.S., Wu, J., Ziegler, T.E., Yang, X., Zayed, A., Rajani, M.S., Zhou, D., Basra, A.S., Schachtman, D.P., Peng, M.*, et al.* (2011). Gene expression biomarkers provide sensitive indicators of in planta nitrogen status in maize. Plant physiology *157*, 1841-1852. Goodstein DM, Shu S, Howson R, et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*. 2012;40(Database issue):D1178-D1186. doi:10.1093/nar/gkr944.

Zhang, P., Dreher, K., Karthikeyan, A., Chi, A., Pujar, A., Caspi, R., Karp, P., Kirkup, V., Latendresse, M., Lee, C., Mueller, L.A., Muller, R. and Rhee, S.Y. (2010) Creation of a Genome-Wide Metabolic Pathway Database for Populus trichocarpa Using a New Approach for Reconstruction and Curation of Metabolic Pathways for Plants. Plant Physiology 153(4): 1479-1491.

Zhu, G., Wu, A., Xu, X., Xiao, P., Lu, Le., Liu, J., Cao, Y., Luonan, C., Wu, J., and Zhao, X. (2016) PPIM: A Protein-Protein Interaction Database for Maize. Plant Physiology. 170: 618-626