PROJECT SUMMARY "ABI Innovation: Neighborly Network Inference"

1. Senior personnel

PI: Dennis Shasha (NYU Courant Institute of Mathematical Sciences)
Co-PIs: Gloria Coruzzi & Manpreet Katari (NYU Biology, Center for Genomics & Systems Biology)
Collaborator: Rodrigo Gutierrez (Pontificia Universidad Catolica de Chile)

2. Intellectual merit of the proposed activity

Species are being sequenced at a vastly increasing rate. When embarking on the study of a newly sequenced species, researchers would benefit from tools that infer gene interaction networks from experiments on phylogenetically neighboring species. We propose to develop such "***Neighborly Network Inference*** (NNI)" tools. Our vision is to construct species-specific interaction networks of many kinds (transcription factor-binding networks, protein-protein, metabolic, miRNA-RNA, etc.) for many species *synergistically*. In this vision for NNI, every experiment in species *s*, will contribute to inferences on *s*, and all related species. The experimental bases of our inference will vary from steady-state wild-type experiments, to time-series experiments, to mutant experiments, all on 21 plant species whose genomes have been fully-sequenced, and are of great practical and research importance (e.g. Arabidopsis, rice, soy). This project will leverage the facilities of the current VirtualPlant software platform (www.virtualplant.org) developed under an NSF Grant (DBI-0445666), that includes Arabidopsis multinetwork data, analysis, integration and manipulation tools [1]. In this grant, we will develop tools that will infer gene interaction networks in a target species, based on measured results in a fully-sequenced source species, and also an approach to select which experiments are likely to be most helpful. While NNI is described with respect to plants, the framework and basic algorithms may be extended to any under-analyzed species. This work will achieve one of the main goals of Systems Biology – predicting network states under untested conditions. ***We divide the work into three aims*:**

Aim 1. Develop the Neighborly Network Inference (NNI) Model on Expression Data. Omic-scale expression correlation is the basis for clustering, transcription factor-target inference, and many other goals. This aim explains a machine-learning framework to infer correlation in a little-studied target species, based on experiments in a well-studied and fully-sequenced source species.

Aim 2. Proof-of-principle verification of Neighborly Network Inference (NNI) on heterogeneous data. This aim extends the previous model to infer non-expression edges even among distant species. Our preliminary results show high precision inference of metabolic and protein:protein networks in Rice, inferred from Arabidopsis as a reference species. The approach uses a fixed set of parameters. After showing the promise of this approach, the aim explains how to choose the parameters in a principled way using machine-learning.

Aim 3. Predicting experimental "*Pay-off*": A Framework to Determine the Next Best Experiments to Perform. This aim proposes a tool to help experimentalists determine how they should spend their assay resources. The basic idea is to define a notion of the "*pay-off*" of a set of experiments. Then the tool tries removing existing experiments, to see which already done experiments give the highest payoffs until now. The experiments giving highest *payoff*, then guide the selection of future experiments.

Justification for ABI Goals: The approach and tools we develop will be deployed using a gaggle-based [2] interface, so biological tools can have easy access to it. Our project addresses several ABI goals:

1. *New algorithms for network inference:* Neighborly Network Inference methods for expression (Aim 1) and the generalization to other kinds of edges (Aim 2)
2. *Heterogeneous data*: Use of homology, expression, metabolic and protein-protein networks (Aim 2).
3. *Tools for biological work-flows:* Helping biologists determine the next experiment to do (Aim 3).

3. Broader impacts of the proposed research This project is the result of a long-standing and highly successful collaboration between biologists at NYU and elsewhere, and computer scientists at NYU's Courant Institute of Mathematical Sciences. The Systems Biology tools resulting from this project will empower biologists to use genomic data to predict a spectrum of gene networks in biology with broad applications to agriculture, the environment, and health. In addition to scientific results, this collaboration extends to joint training of biologists and computer scientists in the field of Systems Biology.

PROJECT DESCRIPTION

Motivation: Suppose a scientific community is working on a set of related species that have been fully sequenced. Experimenters around the world are doing experiments on individual species under various conditions, at different times. Some species enjoy more experimental attention than others. Whereas an individual scientist may be interested in one or a few species, the community as a whole is interested in increasing knowledge about all the related species as efficiently as possible.

Our vision for "Neighborly Network Inference" (NNI) is to construct species-specific gene interaction networks of ever-better quality, using sets of related species. For each species s, we will use the experiments on s, but also the experiments on phylogenetically neighboring species s1, s2, ... The approaches we will use to infer network edges include intra-species techniques (cis-element analysis, time-series analysis, knockout analysis (where feasible)), and inter-species techniques (orthology of genes and species). In the complete vision, every experiment on species s will add edges (or increase the confidence in intergenic-edges) to s, as well as to neighboring species. In addition, we propose a methodology for suggesting how to direct experimental effort on species s to maximize the information gained from a given experimental budget (measured for example as number of deep-seq runs). With the advent of multi-species clustering [3], co-regulated modules, as well as edges, will be found.

As a test case, and because of their intrinsic importance, the NNI project proposes to use 21 recently sequenced plant species: *Glycine max*, *Medicago truncatula*, *Cucumis sativus*, *Prunus persica*, *Populus trichorpa*, *Manihot esculenta*, *Ricinus communis*, *Citrus sinensis*, *Arabidopsis thaliana*, *Carica papaya*, *Eucalyptus grandis*, *Vitis vinifera*, *Mimulus guttatus*, *Aquilegia coerulea*, *Sorghum bicolor*, *Zea mays*, *Setaria italica*, *Oryza sativa*, *Brachypodium distachyon*, *Physcometrille patens*, *Selaginella moellendorfii* (Fig. 1). To construct the phylogenomic tree shown in Fig. 1, we used OrthologID [4] to process 731,093 amino acid sequences from 21 completely sequenced land plant genomes, including two outgroups (Selaginella and Physcomitrella). OrthologID produced 45,156 sets of orthologs from 10,054 multi-species gene families, and assembled a simultaneous analysis (SA) matrix with 21,271 partitions. At least 5 taxa are present in each partition in this matrix, which has 12.9 million characters. The total evidence (TE) tree is the most parsimonious tree generated from the SA matrix using a combination of drifting, ratchet, and fusion in TNT [5]. For a more detailed description of our method for constructing phylogenomic trees see [6-8].

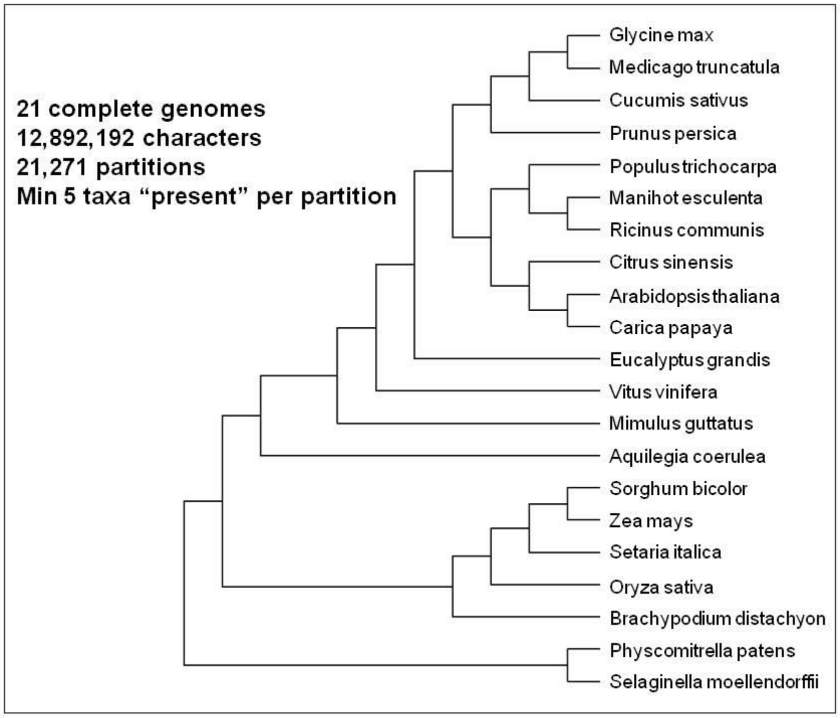


Fig. 1: Phylogenomic tree of 21 plant species with complete genome sequences. The total evidence tree shown here, was creating using OrthologID [4]. It is the most parsimonious tree generated from simultaneous analysis matrix, using combination of drifting, rachet, and fusion in TNT [5], as described in the text and in [6,7].

Next-gen sequencing techniques and microarray platforms have been used to generate transcriptome data for some of these 21 species, but only a few of these fully-sequenced species have gene interaction networks. The software tools we will collect and build in this NNI project will generate gene-interaction edges using inter-species and intra-species techniques. This is timely, because many additional plant species will soon be sequenced and then expression data will be available.

RESULTS FROM PRIOR NSF SUPPORT: This proposal leverages on the accomplishments of the completed parent NSF Grant DBI-0445666, “Conceptual Data Integration for the Virtual Plant”. The VirtualPlant software platform (www.virtualplant.org) [1] developed in that grant, integrates genome-wide data concerning the known and predicted relationships among genes, proteins and molecules, as well as genome-scale experimental measurements. VirtualPlant also provides tools that render multivariate information into integrated visual displays (e.g. networks) to highlight biological implications. We have demonstrated the use of tools embodied in the VirtualPlant system to generate hypotheses that were subsequently experimentally validated [9-15].

Our NSF VirtualPlant grant had four goals: **Integration**, **Visualization**, **Synthesis**, and **Prediction**, which we have accomplished, as outlined below.

Aim 1. Integration: ***The Arabidopsis Multinetwork**: **A systems biology tool for hypothesis generation*****.** Our VirtualPlant project included assembling the first multinetwork for Arabidopsis, a first step towards a molecular wiring diagram of the plant cell [1,11]. The Arabidopsis multinetwork in VirtualPlant has 16,562 nodes (of which 13,960 are genes) and 97,423 interactions (Fig. 2B, & Table 1). The multinetwork enables researchers to interpret transcriptome data in the context of all known sources of interaction including protein, DNA, RNA, etc. In one example, a query against the Arabidopsis multinetwork with 834 nitrogen-regulated genes resulted in a sub-network of 369 genes connected by one (or more) “expression correlation edges” [15]. At the top of the resulting list of network TF “hubs” (with 47 connections to targets in the N-regulatory network) was the central clock control gene CCA1, a Myb family transcription factor (TF) [15] . Exploration of the network “neighborhood” surrounding this CCA1 TF hub revealed connections to target genes in N-assimilation (Fig. 2C). Using Arabidopsis lines that over-express 35S::CCA1 and by Chromatin-IP, we validated these predicted TF→Target interactions and showed, using phase response curves, that distinct N-metabolites can advance or delay the circadian phase of CCA1 mRNA expression [15]. Thus, we derived and validated the novel hypothesis that nitrogen-regulation of CCA1 mRNA expression sets the circadian clock. Other examples of networks derived and validated using the VirtualPlant multinetwork are reported in [9,11,12,13].

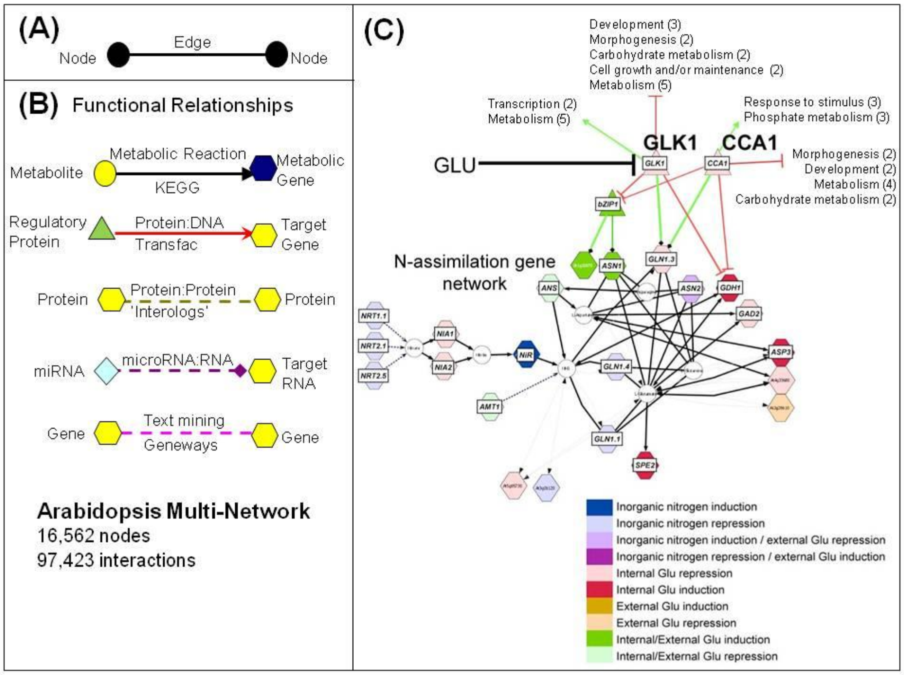
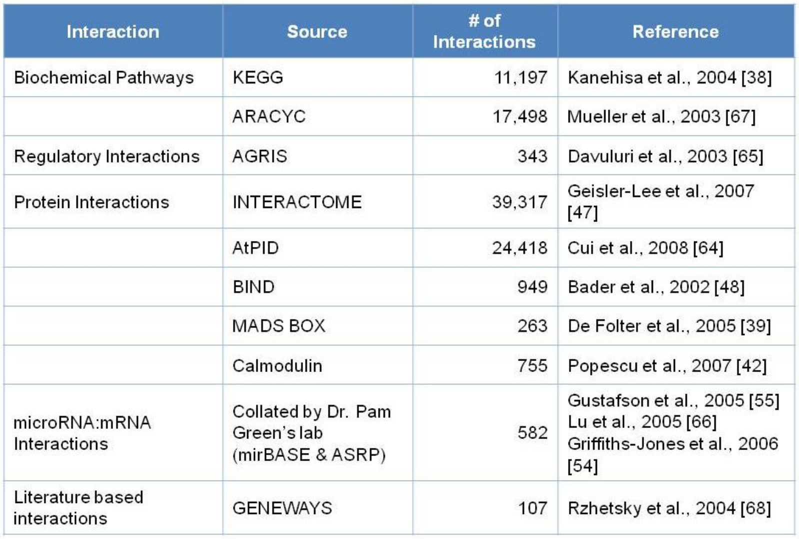


Fig. 2: The VirtualPlant Multinetwork. The Arabidopsis multinetwork contains genes represented as nodes (A) that are connected by edges of many types (B) including metabolic, protein-DNA, protein-protein, microRNA-RNA, and edges derived from text mining [1]. (C) shows a network neighborhood resulting from querying this multinetwork with microarray data, uncovering a regulatory hub (CCA1) involved in nitrogen signaling [15].

A complementary network tool is GeneMania [16], which generates a hypothesis for gene function based on interactions with other genes and their attributes. For a recent review of various plant multinetwork approaches, see [17].



Interaction	Source	# of Interactions	Reference
Biochemical Pathways	KEGG	11,197	Kanehisa et al., 2004 [38]
	ARACYC	17,498	Mueller et al., 2003 [67]
Regulatory Interactions	AGRIS	343	Davuluri et al., 2003 [65]
Protein Interactions	INTERACTOME	39,317	Geisler-Lee et al., 2007 [47]
	AtPID	24,418	Cui et al., 2008 [64]
	BIND	949	Bader et al., 2002 [48]
	MADS BOX	263	De Folter et al., 2005 [39]
	Calmodulin	755	Popescu et al., 2007 [42]
microRNA:mRNA Interactions	Collated by Dr. Pam Green's lab (mirBASE & ASRP)	582	Gustafson et al., 2005 [55] Lu et al., 2005 [66] Griffiths-Jones et al., 2006 [54]
Literature based interactions	GENEWAYS	107	Rzhetsky et al., 2004 [68]

Table 1: Quantitative Information about the Edge Types of the Arabidopsis Multinetwork. The multinetwork contains interaction data from several different sources. Here, we list some of the main sources of interactions. For a detailed description see [1].

Aims 2 & 3. Synthesis and Visualization: ***VirtualPlant's primary analysis tools and functions.*** In addition to the multinetwork, the VirtualPlant platform houses other tools for data analysis, integration and visualization. Below is a list of three exemplary tools deployed through VirtualPlant.

BioMaps: BioMaps takes one or more sets of genes and determines which functional terms (GO [18] or MIPS [19]) are statistically over-represented in each set, with respect to a background population (e.g. Arabidopsis genome). The output is presented in either a tabular format that can be downloaded to Microsoft Excel or a graphical representation based on the GO directed acyclic graph [1].

Sungear: Sungear is a visually interactive and biologist-driven exploration of comparisons of the results of many experiments on a genomic scale. Sungear can represent an arbitrary number of experiments/lists, all of their disjoint intersections, and their related ontological terms. The position of a circle and arrows emanating from it, indicate the input lists of which it is a subset. The size of a circle is proportional to the number of genes in the intersection of those lists (see [20]). Many biologists find Sungear to be an extremely powerful and interactive tool for analyzing the interrelationships between sets of genes [10].

NetMatch: NetMatch, a Cytoscape plug-in, finds all instances of a query graph (e.g. a network motif) in a larger graph [21]. New versions compute the statistical significance of the motifs (e.g. Transcription factor motifs) found in a network.

Up and coming tools for VirtualPlant include **GeneSect,** whose purpose it is to take a set of collections of genes and to determine whether any pair-wise intersections among those collections are either surprisingly large (against a variety of backgrounds), or surprisingly small. Another new tool under development is a cluster management framework **ClusterBoss,** to run some expensive tasks such as correlation and network inference in parallel, which relates directly to the aims of the current NSF NNI proposal.

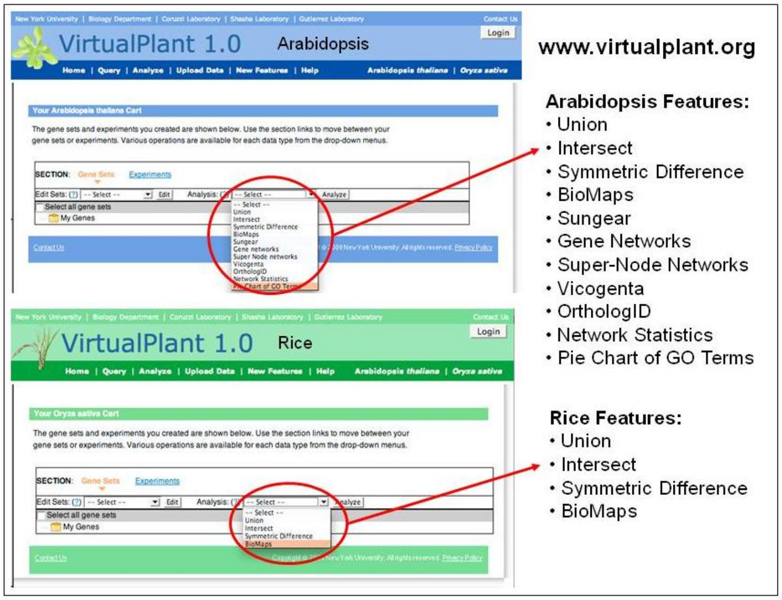
Aim 4. ***Predictions: Extensions into time and species*****.** We have approached dynamic network modeling by applying a machine learning method called "State Space" analysis to time-series data in Arabidopsis to learn regulatory networks [22,23]. Our second goal, was to extend VirtualPlant to other species, such as Rice, which we have done as shown in Fig. 3.

Virtual Plant and User Community:

The VirtualPlant user community currently consists of 635 registered academic and commercial users from 36 countries. Among the 347 registered US users, 181 are from academia and 166 are from companies. Examples of commercial users include: Monsanto, Pioneer, Ceres, Syngenta and Unilever. Other countries that also have many users include: UK (78), Australia (27), Germany (24), Chile (22), France (15), Italy (11), Spain (10), Canada (9), Japan (8), Korea (8). In addition, many anonymous unregistered users use VirtualPlant, but cannot store their datasets for later iterative analysis.

VirtualPlant DB: The VirtualPlant database contains some of the most commonly used data types including metabolic pathways from KEGG and ARACYC, protein-protein interactions from BIND and Interolog databases, and GeneOntology and Gene annotations from TAIR (see Table I for a complete listing of data sources). The database also contains processed data obtained by analyzing publicly available Microarray experiments obtained from NASC [24].

Software and Data Availability: VirtualPlant is accessible via the website www.virtualplant.org. Registered users (currently > 630) store their data sets and use many tools to analyze their genomic data such as microarray experiments. The website does not require a password and is available for free when used for non-for-profit purposes.


The gene sets and experiments you created are shown below. Use the section links to move between your gene sets or experiments. Various operations are available for each data type from the drop-down menus.

The gene sets and experiments you created are shown below. Use the section links to move between your gene sets or experiments. Various operations are available for each data type from the drop-down menus.

New York University | Biology Department | Coruzzi Laboratory | Shasha Laboratory | Gutierrez Laboratory
Contact Us
Login
VirtualPlant 1.0
Arabidopsis
Home | Query | Analyze | Upload Data | New Features | Help
Arabidopsis thaliana | Oryza sativa
Your Arabidopsis thaliana Cart
SECTION: Gene Sets Experiments
Edit Sets: (?) -- Select -- Edit
Analysis: (?) -- Select -- Analyze
Select all gene sets
My Genes
-- Select --
Union
Intersect
Symmetric Difference
BioMaps
Sungear
Gene networks
Super Node networks
Vicogenta
OrthologID
Network Statistics
Pie Chart of GO Terms
Contact Us
© 2008 New York University. All rights reserved. Privacy Policy
www.virtualplant.org
Arabidopsis Features:
• Union
• Intersect
• Symmetric Difference
• BioMaps
• Sungear
• Gene Networks
• Super-Node Networks
• Vicogenta
• OrthologID
• Network Statistics
• Pie Chart of GO Terms
New York University | Biology Department | Coruzzi Laboratory | Shasha Laboratory | Gutierrez Laboratory
Contact Us
Login
VirtualPlant 1.0
Rice
Home | Query | Analyze | Upload Data | New Features | Help
Arabidopsis thaliana | Oryza sativa
Your Oryza sativa Cart
SECTION: Gene Sets Experiments
Edit Sets: (?) -- Select -- Edit
Analysis: (?) -- Select -- Analyze
Select all gene sets
My Genes
-- Select --
Union
Intersect
Symmetric Difference
BioMaps
Contact Us
Copyright © New York University. All rights reserved. Privacy Policy
Rice Features:
• Union
• Intersect
• Symmetric Difference
• BioMaps

Fig. 3: The VirtualPlant Arabidopsis and Rice Home Pages. The VirtualPlant software platform (www.virtualplant.org) is designed to support multiple species [1]. Shown are the two home pages for Arabidopsis and Rice. Each supports a common set of tools, but is implemented on top of a separate database. An analysis within a species will not be slowed down by the addition of another species.

NOTE IN REVISION: This proposal is a revision of a previous application to NSF Plant Genome (IOS-1025989: TRMS "Cross species network inference: From models to crops" (January 2010), and later to NSF ABI Innovation (ABI-1062434): "Cross species network inference" (July 2010). Both panel reviews endorsed the novelty and importance of creating tools to enable network inference across species. In the NSF Plant Genome panel, the reviewers noted: "Shasha et. al. propose to develop, validate and deploy an analysis pipeline for comparative inference of gene function and interaction based on similarities in NT sequence, regulatory regions and transcription patterns. Such a tool is *sorely needed* with the growing number of genome and trancriptome sequences coming available for the emerging model and non-model species. … As such, the proposed development of a web-based Cross-species network inference database and analysis tool would be a major contribution."

In the more recent ABI submission, the reviewers appreciated the "novelty and importance of the proposed work and its broader impacts". "The proposed resource will contribute to increased understanding of biochemical, regulatory, RNA and protein-protein interactions in plant species, with broad applications to agricultural and environmental issues. The biologist-friendly tools will benefit the plant biology research community." However, because that ABI submission included a mixture of innovation and development activities, the PIs were encouraged by the PO to select one aspect for a resubmission. In this revised ABI Innovation Application, we focus on *innovation*, related to the creation of cross-species network inference tools using an approach called "Neighborly Network Inference", which exploits orthology, expression and other data, as well as phylogenetic position in making inference about network associations across species.

PUBLICATIONS: Peer reviewed journal articles, chapters and books:

VirtualPlant: Tool development for Plant Systems Biology

Katari MS, Nowicki S, Aceituno F, Nero D, Kelfer J, Thompson L, Cabello J, Davidson R, Goldberg A, Shasha D, Coruzzi G, Gutierrez R (2010) "VirtualPlant: A software platform to support Systems Biology research". ***Plant Physiol***. Feb; 152:500-15

Nero D, Kelfer J, Katari MS, Tranchina D, Coruzzi G (2009) "*In silico* evaluation of predicted regulatory interactions in Arabidopsis thaliana". ***BMC Bioinformatics***. Dec 21;10(1):435

Poultney C, Gutierrez R, Katari MS, Gifford M, Paley W, Coruzzi G and Shasha D (2007) "Sungear: Interactive visualization, exploration & functional analysis of genomic datasets". ***Bioinformatics***, 23:259-61

Ferro A, Giugno R, Pigola G, Pulvirenti A, Skripin D, Bader G, Shasha D, "NetMatch: a Cytoscape plugin for searching biological networks" ***Bioinformatics***, 2007 23(7):910-912

Applications of VirtualPlant: Hypothesis Generation and Testing

Krouk, G, Mirowski, P, LeCun, Y, Shasha, D and Coruzzi, G. (2010) Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate. ***Genome Biology*** 11 (12), R123

Krouk, G, Crawford, NM, Coruzzi, GM and Tsay, YF. (2010) Nitrate signaling: adaptation to fluctuating environments. ***Curr Opin Plant Biol*** 13 (3), 266-273

Krouk G, Tranchina D, Lejay L, Cruikshank A, Shasha D, Coruzzi G and Gutierrez R (2009) "A systems approach uncovers restrictions for signal interactions regulating genome-wide responses to nutritional cues in Arabidopsis." ***PloS Comp Biol***. Mar;5(3):e1000326. *(Highly Accessed).*

Gutierrez R, Stokes T, Thum K, Xu X, Obertello M, Katari M, Tanurdzic M, Dean A, Nero D, McClung R and Coruzzi G (2008) "Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1" ***Proc. Natl Acad Sci USA*** 105, 4939-4944. *(Faculty of 1000 recommended: Factor 3)*

Gutierrez R, Gifford M, Poultney C, Wang R, Shasha D, Coruzzi G, Crawford N (2007) "Insights into the genomic nitrate response using genetics and the Sungear Software System" ***Journal of Experimental Botany*** doi: 10.1093/jxb/erm079

Gutierrez R, Lejay L, Chiaromonte F, Shasha D, Coruzzi G (2007) "Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive biomodules in Arabidopsis" ***Genome Biology***, 8: R7. *Faculty 1000 (Must Read: Factor 6)*

Plant Systems Biology: Reviews, Books and Outreach

Ruffel S, Krouk G, Coruzzi G (2010). "A Systems View of Responses to Nutritional Cues in Arabidopsis: Towards a Paradigm Shift for Predictive Network Modeling". ***Plant Physiol***. Feb; 152;445-52

Gutierrez R, Coruzzi G., Eds (2009) Book: "Plant Systems Biology", ***Annual Plant Reviews***; Blackwell Publishing: Oxford, UK, 2009, Vol. 35. 360 pages.

Coruzzi GM, Burga A, Katari MS, and Gutierrez RA (2009) "Systems Biology: Principles and Applications in Plant Research". In "Plant Systems Biology", ***Annual Plant Reviews***; Blackwell Publishing: Oxford, UK, 2009, Vol. 35. Pgs 3-31. *Book Chapter.*

Gifford M, Gutierrez R, and Coruzzi G (2006) "Modeling the Virtual Plant: A Systems Approach to Nitrogen-Regulatory Gene Networks". Essay 12.2 Chapter 12. Assimilation of mineral nutrients; In ***A Companion to Plant Physiology,*** Fourth Edition, Lincoln Taiz and Eduardo Zeiger, http://4e.plantphys.net/article.php?ch=12&id=352

Gutierrez R, Shasha D and Coruzzi G. (2005) "Systems Biology for the Virtual Plant". ***Plant Physiol.*** Vol 138, pp 550-554.

Education & Training: The development of the Systems Biology tools and the Virtual Plant software platform has trained undergraduates (UG), MS and PhD students in Systems Biology. Students trained include **Undergraduates**: Steve Nowicki (NYU CAS), Varuni Prabhakar (Barnard College), Rebecca Davidson (BS Computer Science); **Masters Students**: Ana F. Arroja (MS student, NYU Courant), Ranjita Iyer (MS Computer Science), Jonathan Kelfer (MS Computer Science), Jesse Lingeman (MS Computer Science), Lee Parnell (MS Computer Science), Jarod Wang, (MS Computer Science); **PhD Students**: Chris Poultney (PhD student, NYU Courant), Aris Tsirigos (PhD student, NYU Courant), Saurabh Kumar (PhD student, NYU Courant). These students have gone on to PhD programs (Prabhakar and Parnell), post-docs (Poultney and Tsirigos) as well as to industry (Kelfer, Wang Medidata Solutions).

RESEARCH DESIGN

Aim 1: Development of the Neighborly Network Inference (NNI) model on Expression data

Rationale. With the advent of Next-gen sequencing technologies, it will be increasingly common to find a newly sequenced species *s*, that is phylogenetically similar to other species on which there are already available experiments. Because many of those experiments will be genome-wide assays such as transcriptome expression measurements, we start with the network inference of positive and negative expression correlation for a hypothetical newly sequenced species *s*.

Our neighboring species strategy for inferring edges between genes in species *s*, starts with pair-wise gene expression correlation data on other species. From that data, we will train a machine-learning algorithm to determine whether there will be correlation between two genes in species *s*. In addition to the simple Pearson correlation we use in this preliminary work, we will use related techniques such as mutual information [25], and Spearman correlation. (Note: It is a separate question to determine whether correlation signifies causality. If genes g1 and g2 correlate, g1 is a transcription factor, g2 is not, and g2 has a transcription factor-binding site that the protein associated with g1 can bind to, then this is some evidence for causality. The best test is time-series experiment and analysis [22,26-29], followed by a knock-out or over-expression experiment. As that data becomes available, we will use it as part of our network inference project.).

The input for our algorithm will be in the three formats described below.

orthotab: target species| target gene | other species | other gene | orthology val1 | orthology val2 …: gives the gene-to-gene orthology value, according to several different orthology measures for example: reciprocal best blast [30] hits, OrthologID [4], OrthoMCL [31], and Inparanoid [32].

edgetab: species | gene1 | gene2 | edgetype | strength | p-value | number of different experimental conditions: gives the strength and the p-value (the probability it could arise by chance – we evaluate this using a non-parametric re-sampling approach) of a given experimentally determined edge. We consider only experimentally determined edges as an input to this inference algorithm to avoid circular inferences. Note that certain edge relationships may be present only in certain conditions (e.g. drought conditions for plants). In that case, the tools we propose could be used just for the conditions of interest. In our preliminary work, we find correlations that generally hold over all conditions.

species1 | species2 | species similarity measure1 | species similarity measure2: measures sequence similarity according to one of a number of criteria (e.g. distance based, for example average percent identity of protein sequences, or through parsimony).

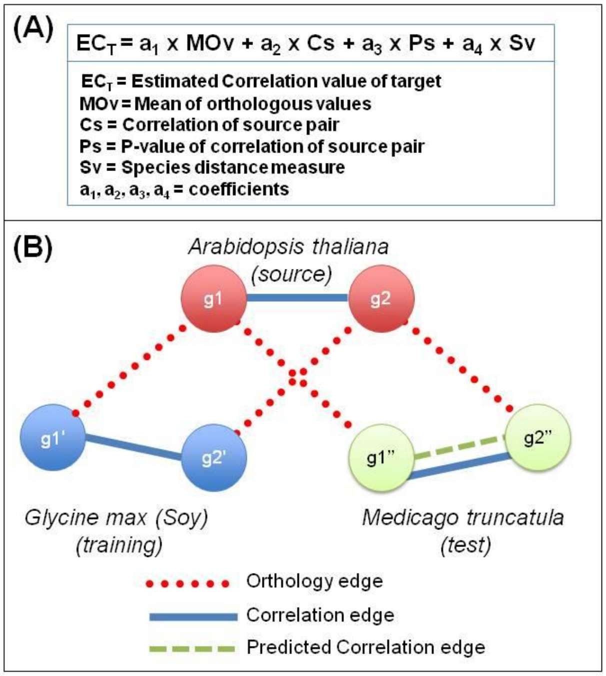
Now, to predict an edge between *g1* and *g2* in species *s*, we will combine evidence from edges in other species, as well as evidence from experiments in species *s* itself. The basic method will be regression and regression trees, with a penalty for complexity.

For the sake of performance and robustness to noise, ***we will use some mixture of the following three approaches***:

1. **Random Forests** [33,34] Random forests are ensembles of decision trees which are constructed from random subsets of the data. They're fast to train, easy to parallelize, and perform extremely well.
2. **Large-Scale SVM Regression** [35] Bottou demonstrated that a stochastic gradient descent solver for a variety of learning problems (including support vector machine optimization), is able to scale with extremely large datasets, while converging to the predictive performance of traditional optimization algorithms.
3. **Large-Scale L-Regularized Learning** [36] Stochastic coordinate descent (a method related to stochastic gradient descent, but with a slightly different update rule), can be used to learn sparse regression models, with small training-times, even for data sets where both the dimensionality and the number of training-points is large.

The net effect of this analysis, will be to find the weighting of different factors that will lead us to conclude that two genes in some species are correlated. Then, using available Arabidopsis time-series data [22], and other datasets that are currently being generated in our lab and others, we will combine correlation with time-series [22,26-29] and perturbation approaches using Graphical Lasso [37] to form causal networks.

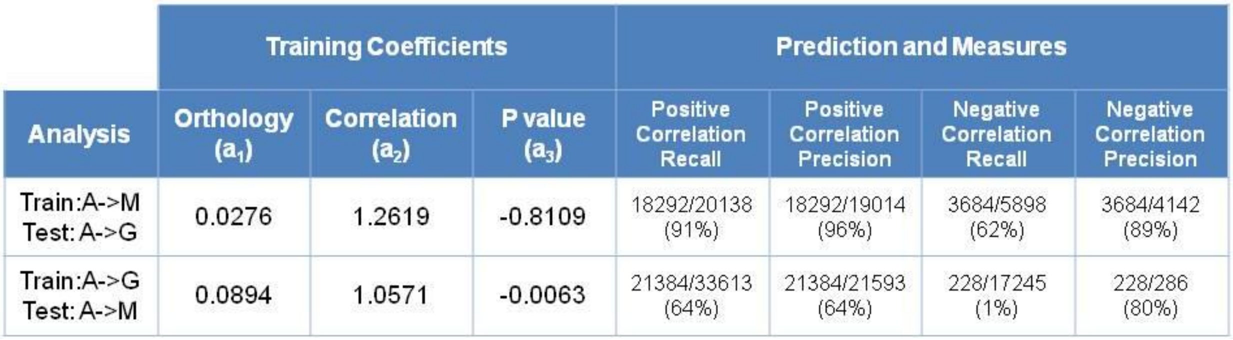
Preliminary Results. In our initial case study, we consider steady-state data on three species Arabidopsis (A), Medicago (M), and Soy (G) (*Glycine max*) Fig. 4 & Table 2. We selected these three species as an initial test case because (i) there is ample and reliable Affymetrix data for each, and (ii) Medicago and Soybean -- both legumes -- are quite closely related (more so than Arabidopsis and Rice, as we discuss in the preliminary work for Aim 2). We tested the ability to infer Pearson correlation edges in a "target" species, knowing only correlation edges in a "source" species, and the gene-by-gene orthology between genes in the source species, and genes in the target species (Fig. 4). For this study, we analyze only those genes that are conserved across all three species - Arabidopsis, Medicago and Soybean. Since there are a different number of experiments for each, and experiments from different sources, the distribution of correlation values can vary. So, we define two genes as "highly positively correlated", if their correlation is in the top 5%, and "highly negatively correlated", if their correlation is in the bottom 5%, and "in between" otherwise. Thus, our machine-learning algorithm predicts which of these three categories (positive, between, or negative) an edge in the target species is in. To assess the quality of the predictions, we compare the predicted results (that use no expression experiments in the target species), with the results from the experiments in the target species.


(A)
$EC_T = a_1 \times MOv + a_2 \times Cs + a_3 \times Ps + a_4 \times Sv$
EC_T = Estimated Correlation value of target
MOv = Mean of orthologous values
Cs = Correlation of source pair
Ps = P-value of correlation of source pair
Sv = Species distance measure
a_1, a_2, a_3, a_4 = coefficients
(B)
Arabidopsis thaliana
(source)
g1
g2
g1'
g2'
g1"
g2"
Glycine max (Soy)
(training)
Medicago truncatula
(test)
Orthology edge
Correlation edge
Predicted Correlation edge

Neighborly Network Inference Model. Panel A, describes the equation used on the training data to determine the coefficients (a1, a2, a3..), which are then used for predicting the correlation edges in **Panel B**. Panel B shows an example where the model is trained (e.g. coefficients are determined) using correlation data in Arabidopsis (A) and Soy (G, Glycine max) as well as orthology data between A and G. Then, the model is used to predict correlated edges in M (Medicago) (a neighbor species of G), given the coefficients determined in training, and orthology between genes in A and M and correlations in A. The results of those predictions are then compared to experimentally determined correlation edges in M for validation. When training on several pairs of species, then coefficient a4 species distance measure will be used in training and predictions.

Because stochastic gradient descent is a machine learning technique, we train a linear equation of the form estimated correlation in target = a1*mean of orthologous values + a2*correlation of source pair + a3*p-value of correlation of source pair, and + a4*species distance measure (Fig. 4). Here, mean of orthologous values is calculated as follows: if g1 and g2 are the source pair, and g1' and g2’ are the potential target pair, and g1 and g1’ are reverse top blast hit matches (as are g2 and g2’), then we take the mean of the orthology values, in this case percent identity, between g1 and g1', and between g2 and g2'.

However, this equation is too simple because (i) it ignores experiments already done in the target species, and (ii) many gene pairs (besides reverse top blast hits) in the source species, may be relevant to the target pair g1 and g2, for example paralogs. For point (i), when some expression data about the target species is available, we will add a term of the form c*prelimcorrelation(g1,g2), that takes into account the correlation between g1 and g2, based on the experiments performed so far, though we did not do that here. For point (ii), we may require some form of aggregation over the gene pairs of the source species that are orthologous above a threshold to g1 and g2. (Note: That is unnecessary in this preliminary study, where we focus on reverse top blast hits.) When using a threshold, cross-validation on a training set, would set the level of the threshold. Finally, once we have data on many pairs of species, we will include a term that measures the similarity of species.



	Training Coefficients			Prediction and Measures			
Analysis	Orthology (a_1)	Correlation (a_2)	P value (a_3)	Positive Correlation Recall	Positive Correlation Precision	Negative Correlation Recall	Negative Correlation Precision
Train:A->M Test: A->G	0.0276	1.2619	-0.8109	18292/20138 (91%)	18292/19014 (96%)	3684/5898 (62%)	3684/4142 (89%)
Train:A->G Test: A->M	0.0894	1.0571	-0.0063	21384/33613 (64%)	21384/21593 (64%)	228/17245 (1%)	228/286 (80%)

Table 2: Neighborly Network Inference between Arabidopsis (A), Medicago (M), and Soy (G, *Glycine max*). The table is separated into two parts – (Left) Coefficients obtained from training and (RIGHT) The precision and recall of the correlation predictions. The analysis was performed reciprocally, using A→ M for training, and then predicting G, or using A→ G as training, and M for test. Recall is less for negative correlation values because the training set is smaller.

We have assigned coefficients to the linear equation using Arabidopsis (A) as source species, and Soy (G, *Glycine max*) as the target. Then, we use those coefficients to infer edges in Medicago (M), based on edges in Arabidopsis. Then, we will do another test in which Soy and Medicago reverse roles. Examples of these results are summarized in Fig. 4 and Table 2.

When we train using Arabidopsis (A) and Medicago (M) data, we get values a1 = 0.0276, a2 = 1.2619, a3 = -0.8109. We then test this using Arabidopsis and Soy (G), to get 18,292 predicted highly positive correlations, 3,684 predicted highly negative correlations. This gives us a recall of 0.91, for highly positive correlations, with a precision of 0.96, and for highly negative correlations, we get a recall of 0.62, and precision of 0.89 (Table 2).

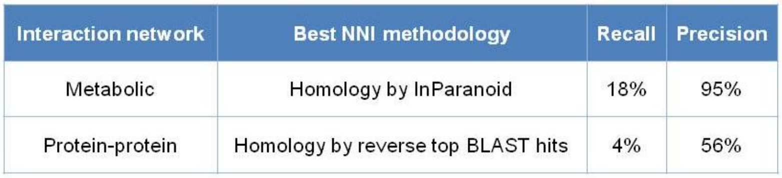
When we train using Arabidopsis (A) and Soy (G) data, we get values a1 = 0.0894, a2 =1.0571, a3 =-0.0063. We then test this using Arabidopsis (A) and Medicago (M), to get 21,384 predicted highly positive correlations, and 228 predicted highly negative correlations. This gives us a recall of 0.99 for highly positive correlations, with a precision of 0.98, and recall of 0.01 and precision of 0.8, for highly negative correlations. Recall is less for negative correlation values because the training set is smaller (Table 2).

In this preliminary test, we only used one pair of species to train. As we develop this aim, we will train on several pairs of species, in which case coefficient a4*species distance measure will be used in both training and predictions. Note also that this preliminary experiment makes predictions only about pairs in the target species whose members are highly orthologous to some pair in the source species. Our recall numbers would be much lower if we were measuring our success against identifying ALL correlation edges in the target species. Orthology helps and may identify some of the most important edges, but this technique complements rather than replaces in-species experimentation.

Expected Outcomes of Aim 1. Our goal in this Aim, is to construct a machine-learning model that can predict, with high recall and precision, the expression correlation of edges between genes in a little-studied species, by inference from a well-studied species. As more data about the species becomes available, we then apply the rest of our workflow to find a refined causal network.

Aim 2: Proof-of-principle verification of Neighborly Network Inference (NNI) on heterogeneous data. ***Rationale***. In production, Neighborly Network Inference (NNI) will be used to infer edges between genes of many types. In Aim 1, we concerned ourselves only with expression correlation. The purpose of Aim 2, is to apply the NNI methodology to other kinds of networks (e.g. metabolic and protein-protein). As a test case, we infer such networks in species for which networks have been experimentally determined, and then evaluate the accuracy with which the inferred network predicts the experimental network. Here, we have chosen Arabidopsis and Rice, because they each have the most complete genomic data set to test our methods. Despite the fact that these species are phylogenetically far apart (Fig. 1), the high accuracy obtained in our preliminary studies, suggests that the approach could have wide applicability. We discuss 1) our preliminary analysis and results, 2) the overall objectives of this aim, and 3) its expected outcomes.

Preliminary results. Our preliminary results demonstrate NNI's ability to infer gene networks from Arabidopsis to Rice with impressive accuracy, as shown in Table 3. For the data in Table 3, NNI was used to *infer a* Rice network that was then compared to the *known* validated data for Rice, including metabolic data from KEGG [38], and protein-protein interaction data from BIND plus other experimentally determined protein interactions [39-43]. Our approach builds on inference approaches based on expression and homology [44-46], and also based on integration of several different types of associations [45,47].



Interaction network	Best NNI methodology	Recall	Precision
Metabolic	Homology by InParanoid	18%	95%
Protein-protein	Homology by reverse top BLAST hits	4%	56%

Table 3: Validation of Network Inference using Arabidopsis (Reference) and Rice (Target). Inferring metabolic or protein interaction relationships in Rice based on homology alone (to Arabidopsis) data (using Rice expression data), yields high precision relative to the validated network (of Rice).

Below are the steps used in the NNI approach:

Step 1. Obtain a reference validated Arabidopsis interaction network based on experimentally supported data. For our validated Arabidopsis networks, we assembled metabolic interactions (KEGG; 19,688 interactions) [38], protein-protein interaction data from BIND (949 interactions) [48], protein-chip interaction data for MADS box (272 interactions) [39] and protein chip interactions for Calmodulin (755 interactions) [42], and the Plant Interactome project (11,374 interactions) (http://signal.salk.edu/interactome.html). Many of the metabolic pathways in the KEGG and AraCyc databases are based on computational predictions, while 25% are validated experimentally in the literature [49,50].

Step 2. Identify Rice homologs of Arabidopsis interaction pairs. Connect every gene in the Arabidopsis interaction network with its Rice homologs. This technique can employ various homology methods, including either distance or parsimony based. In our preliminary analysis (Table 3), we obtained homologs via two commonly used methods, InParanoid [32] and OrthoMCL [31]. We also experimented with distance-based homology, selecting homologs with BLAST matches stronger than E-value of E-20 to capture one-to-many homology relationships [51], which captures the gene duplication events prevalent in plant genomes [52].

Step 3. Build a Rice correlation network based on publicly available Rice microarray expression experiments. We downloaded all 48 Rice gene expression experiments on the Affymetrix GPL2025 platform from GEO [53]. With the aim of finding experiments that both repress and induce the genes of interest (the Rice genes homologous to the genes in the Arabidopsis network), we selected the experiments with the highest variability of expression level across assays for these genes. These were experiments in which at least half the individual gene Z-scores across the assays exceeded 0.5. This selected 8 experiments, with a total of 169 assays (e.g. cel files). We then computed the Pearson correlation of all pairs of the genes of interest. We retain correlation edges between gene pairs whose expression vectors were significantly correlated (p-value <0.05, meaning less than a 5% chance of a non-zero correlation by chance), and absolute value of correlation > 0.5 or >0.7 (Table 3).

Step 4. Build an *inferred* Rice network. Initially, we infer a Rice network that contains the edges that connect homologs to the network in Arabidopsis. We then refine the inferred Rice network, by retaining only edges that *both* connect homologs to the network in Arabidopsis, *and* connect genes whose expression values in the Arabidopsis experiments selected in Step 3, correlate more strongly than 0.5 or 0.7. Conceptually, homology suggests a set of possible network edges in the target species, and strong correlation of expression levels refines the set. Notice that we are *using* expression to *infer* other relationships (metabolic and protein-protein). This network is called the *inferred* Rice network.

Step 5. Obtain a reference validated Rice network that contains edges representing known interactions. Our initial Rice validated network was constructed from 10,976 metabolic interactions and 334 protein-protein interactions for Rice from KEGG [38] and BIND [48].

Step 6. Evaluate *Inferred* Rice Network. This step computes the similarity and p-value (significance) between the *inferred* and validated Rice networks, by using a network intersection tool called ***NetSect*** which is described below. We evaluated the quality of each subset of edge types in the inferred network.

NetSect*. Evaluating the Accuracy of the *Inferred* Network**. Given networks *N* ("inferred") and *M* (validated), with edges *E(N)* and *E(M)* respectively, one can measure their similarity by computing *size(intersection(E(N), E(M))) / size(union(E(N), E(M))),* which equals *1* when *E(N)* and *E(M)* are identical and zero when they are disjoint. We will also compute the recall and precision of the *inferred* network's ability to predict edges in the reference network. To compute a p-value for the *inferred* network's reconstruction of the reference network, ***NetSect computes the similarity of the inferred and validated networks and then computes a p-value by comparing the sample similarity with the similarity of a collection of random networks having the same topology (i.e. isomorphic) as the inferred network, with vertices drawn from the entire genome. This use of randomness corresponds to the null hypothesis that the inferred network is no better than a random choice of edges.

Analysis of preliminary results. Two main conclusions arise from our preliminary analysis of Neighborly Network Inference (Steps 1-6, above) for Rice networks inferred from Arabidopsis data, and validated using rice data, as shown in Table 3. *First*, homology alone does an excellent job of inferring networks, even for distantly related species. For metabolic edges, of the 2,165 edges in the Rice metabolic network inferred via homologs from InParanoid, 94.8% or 2,053 are validated in the Rice KEGG metabolic interactions, while the inferred network's recall is 17.8%. The precision of the metabolic network prediction is so high, that we hypothesize many of the predicted protein interaction edges that haven't yet been detected experimentally. *Second*, restricting inferred edges to gene pairs with highly correlated expression data, enhances the inference's precision, but invariably dramatically worsens its recall. For example, intersecting with edges between genes with |correlation| > 0.5, reduces the recall to 0.6% for metabolic edges (not shown).

To determine whether our general homology plus expression correlation technique would work for other kinds of edges, we tried to infer Rice protein-protein edges from Arabidopsis protein-protein edges and expression data. Unfortunately, there are only 11,241 *non-redundant* validated protein-protein edges in Arabidopsis and only 344 in Rice [48], so many of our predictions did not fall among those 344, but may one day be validated. Surprisingly, simple homology techniques (reciprocal top Blast hits and InParanoid with homologs of paralogs) each obtained a quite high precision of about 50%, and recall of between 4% and 8%. In those techniques, an edge between rice genes r1 and r2, would be inferred when r1 was homologous to a1, r2 to a2, and a1 and a2 formed a validated protein-protein edge in Arabidopsis. Expression data (either on all experiments or just those in which the expression value of potential homologs varied the most) sometimes improved precision, but at a severe loss in recall.

These very preliminary results suggest that machine-learning techniques, like the stochastic gradient descent method used in Aim 1, can help determine the proper weights of different forms of evidence.

Step 7. Expand validated and network inference into a "multinetwork" containing multiple edge types. We will use techniques analogous to Steps 1-6 to infer networks based on other edge types. For example, we will add miRNA:RNA interactions [54-56]. Expanding the validated networks to include these datasets will enable us to create an inferred multinetwork that includes: protein-protein, Protein:DNA, miRNA-RNA and Metabolic edges.

Role of machine-learning. As one would expect, the choice of data sources, expression experiment selection methods and homology algorithms and parameters, greatly influence the accuracy of the inferred Rice networks. That is why the machine learning techniques outlined in Aim 1 will be used. The experiments used for gene expression correlation, will include many different developmental stages, different organs, and different biotic and abiotic treatments, such as the ones recently released for Rice on GEO NCBI [57].

Expected outcomes and objectives of Aim 2. Through this work, we will evaluate the accuracy of NNI on additional species pairs and data sets. These will include:

1. **Tailor the selection of parameters** and data sources to each form of information (edge type, similarity of species, etc.) For example, our preliminary results (not shown) indicate that Kinase networks [40,41,43] cannot be accurately inferred between Arabidopsis to Rice (e.g. recall and precision each top out at a few percent). One reason for this may be that Transcription Factors – which constitute the majority of targets in Kinase networks – evolve too rapidly to be conserved at the Arabidopsis to Rice phylogenetic distance.
2. As **new data** become available on an ongoing basis, we will evaluate the accuracy of NNI for other species pairs and data sets. For example, NCBI now contains 387 Affymetrix experiments on *Zea mays*, and 145 for *Medicago truncatula*, and large-scale Arabidopsis and Rice protein interaction datasets are being created and will be made available (NSF Plant Interactome Project, http://signal.salk.edu/interactome.html). For some edge types, we expect that gene network inference will perform better between species that are phylogenetically closer. For example, we predict that inference between *Zea mays* and Rice, will perform better than inference between *Zea mays* and Arabidopsis, because the former are both monocots.
3. We will **stress-test** our algorithms for their sequential and parallel performance, by generating large-scale artificial data sets. The results of this test will be to choose and design machine-learning algorithms that scale better while giving as promising results.

Aim 3: Predicting experimental "Pay-off": Framework to Determine the Next Best Experiment to Perform. ***Rationale***: In Aim 3, we propose a framework to estimate the information we might learn from a new set of experiment assays (hereafter simply "assay") on a given species, in order to determine which types of new experiments will be most useful. The goal is to minimize experimental time and expense, with respect to creating a network. We understand the importance of replicates with respect to identifying differentially expressed genes, however, replicates may not be as informative as new experiments for calculating condition-independent correlation between genes.

To establish intuition for this aim, suppose that some species has many replicates/assays in some experimental conditions already, but many important conditions remain experimentally unexplored. It will probably be less useful to perform yet more assays on the already studied conditions, rather than to study new ones. On the other hand, if some important experimental conditions are particularly vulnerable to noise, then it may be useful to repeat assays in those conditions. The question is: how do we anticipate which experimental strategy will be most useful?

To start, we will evaluate the "*payoff*" of performing a set of assays as follows: compare our state of knowledge *before* doing them, with our state of knowledge *afterwards*. To measure the difference, consider the edges *after* the assays to be closer to the truth. The *payoff* is the number of edges that have improved, i.e. how many false-positives have been corrected, how many false-negatives have been corrected, and how many borderline cases have been resolved. For example, suppose we are interested in determining which pairs of genes have a correlation threshold above 0.7 (in absolute value), with a p-value below 0.2. Then, we determine for each gene pair, both before and after the set of experiments, whether that pair achieves the threshold (a positive), doesn't (a negative), or might (e.g. the mean is above 0.7 but the p-value is too large). Pairs that have changed categories contribute to the *payoff*.

Suppose we are given a "budget" of n assays, e.g. a single microarray or chip-chip assay. We will use the above *payoff* measure, to determine which mix of replicates under existing conditions, replicates under c new conditions, with r replicates each (where n = rc), or some number of time-series experiments, where there are r replicates during each time-point. The computational method will not determine which conditions to try (that requires biological insight), just how many new conditions would probably lead to the most learning. For example, suppose that removing replicate assays from the database of assays for some species s, leads to almost the same correlation predictions as including those replicates (i.e. the payoff from including those replicates is low). Suppose, further, that removing conditions changes correlation predictions a lot. Then, our next experiments should explore more conditions.

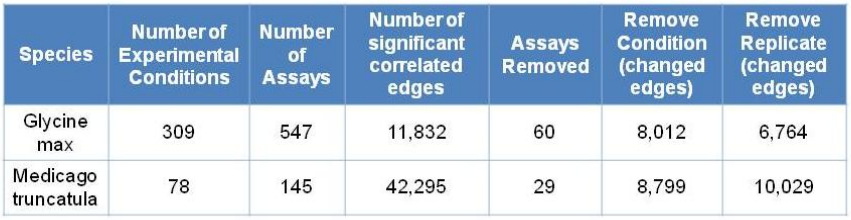
For a certain little-studied species s, this "take-away-and-simulate" strategy may not work, because there may not be enough assays to take away in that species. For that reason, we might use a different species s', that is more studied and is statistically similar to this one. Statistical similarity will be measured as follows: take from s' a subset of its experiments that reflects the diversity of the experiments done on s. For example, if three conditions have been tried on s, having 2, 3, and 4 replicates respectively, then find the subset of experiments on s' having three conditions with 2, 3, and 4 replicates. Next, using only those three conditions having those replicates, find the number of edges calculated to be above threshold in s', the number calculated to be below threshold, and the number in between. If those numbers are similar for s' and s, then try computational experiments on s', in which we use the take-away-and-simulate strategy for n assays on s' to determine the best strategy for s.

In many ways, this work falls in the pool-based sampling subcategory of the active learning framework [58]. In active learning, the learning algorithm "asks questions" to try to optimize the amount of information gained. An example in biology was done by King et al. [59,60] to discover metabolic pathways. The idea is that the active learner chooses a mutant and growth medium, and sees whether the mutant survives, and chooses the most useful growth medium for the purpose. Pool-based sampling, is the idea that there exists a large pool of potential experiments to be performed, and one must choose among them. The most common approach is "uncertainty sampling", in which one performs experiments on data that one is least certain about (in information theoretic terms, the ones with maximum entropy) [61,62]. Another approach is called Expected Model Change, in which we try to learn the assays that would improve our current model as much as possible, if we knew the outcome [63]. Our approach attempts to follow the Expected Model Change approach.

Preliminary results: 1) Table 4, shows the number of experimental conditions, and total number of assays on the species of interest to us. In this case, Soy (*Glycine max*) and Medicago. We also note how many expression correlation edges have an absolute value as great as 0.7, and a p-value of 0.2 or less. Given a budget of 60 assays, we use our method to calculate the *payoff* in each case. For the purposes of this preliminary work, we do the analysis on each species independently of others. In our research for the proposal, we will use orthology-inferred edges as well.

In our test case, we "took away" 60 assays from Soy (*Glycine max)* using two strategies: (i) Remove conditions having the smallest number of replicates until the total number of assays removed equals 60, *OR* (ii) Take away assays from conditions having the most replicates first, until the total number of assays removed equals 60. We did the same for Medicago, but only removed 20% of the assays (29).

Our preliminary experiments (Table 4) showed that for Soy (*Glycine max)* more conditions gave a bigger *payoff*, than more replicates. Specifically, 6,764 gene-pair correlations changed categories (e.g. from negative to between or from between to positive) when replicates were removed, and 8,012 gene pairs changed categories when experimental conditions were removed. Surprisingly, Medicago gave the opposite results. For Medicago, more genes changed categories where replicates were removed (10,029), compared to when conditions were removed (8,799). There are many reasons why different strategies would be better for different species or experimental datasets, but it is important note that our algorithm can capture such discrepancies.



Species	Number of Experimental Conditions	Number of Assays	Number of significant correlated edges	Assays Removed	Remove Condition (changed edges)	Remove Replicate (changed edges)
Glycine max	309	547	11,832	60	8,012	6,764
Medicago truncatula	78	145	42,295	29	8,799	10,029

Table 4: Experimental "Pay-off" prediction. Affymetrix data for Medicago or Soy (*Glycine max*) was downloaded from GEO [53] and normalized. Assays (individual hybridizations) were removed in two strategies: entire experimental conditions or individual replicates across many conditions. A significant correlation is defined as >=0.7 or <=-0.7 and p-value <=0.2. Results suggest that for Medicago it is better to add replicates and for Soy it is better to add new conditions.

Expected Outcomes and Objectives of Aim 3. Our objective is to provide a tool for experimentalists to suggest which group of assays to try next on some species *s*, in order to learn as much as possible. If the experimentalist wants to learn about a whole group of related species, then our method will use the Neighborly Network Inference (NNI) framework to estimate the experimental *payoff* for other species, as well as for s itself. While our preliminary work has been concerned with expression correlation, inference of other kinds of edges (e.g. metabolic, protein-protein as discussed in Aim 2) can use the same technique. Neighborly Network Inference will both infer edges Aims 1 and 2, and also suggest experimental strategies in Aim 3. These two goals work nicely together because inference is needed to calculate the *payoff* of an experiment.

VISUALIZING AND INTEGRATING OUTCOMES OF NNI: When we succeed, Neighborly Network Inference will provide a collection of computational tools to help infer pair-wise relationships (correlation, protein-protein relationships etc.) among plant genes, though similar techniques could be used for other biological entities for which orthology is important. Our goal is to help biologists do their job efficiently and economically. To help them gain insight, we will integrate results across all 21 species (or any phylogenetically related group) using the following simple visualization.

Visualizing the phylogenetic placement of correlated gene pairs. Suppose that *g1* and *g2* are highly correlated across many species in a clade c, but none outside c. In this instance, we will "decorate" the phylogenetic tree (Fig. 1), at the basal node of *c* with *g1* and *g2*, and record the source and number of species having a high correlation value between *g1* and *g2* within *c*. This will permit queries of the form: which gene pairs are highly correlated at some clade? For which clades does this gene pair show high correlation for at least a fraction f, of the species in that clade? (refer to Fig. 1). This visualization will enable a cumulative analysis of gene pair correlations across many species. This will reveal which gene pairs are highly conserved across deep nodes of the phylogenetic tree, and which gene pair correlations are derived. Such information will be useful for practical purposes in inspiring experimental studies from our findings.

TIMELINE: Year 1: Aim 1. Implement Neighborly Network Inference using a variety of machine learning methods, starting with linear regression and extending to various flavors of stochastic gradient descent. Verify the algorithms on simulated data (where we know the ground truth). Cross-validate on the expression experiments from our 21 species. Try the same approach among other eukaryotes. Aim 2. Gather and normalize the data for validated protein-protein and metabolic interaction networks for plant species. **Years 2-3:** Aim 2. Extend the Neighborly Network Inference to other species and data types. Aim 3. Build the framework for determining the best new experiments on cross-validated data. Deploy the first version of the NNI analysis to collaborators (R. Gutierrez, Chile), and other beta-testers.

Future directions: In this NSF ABI innovation grant, in Years 1-3, we will be developing and testing methods for gene network inference across species. As we validate our NNI approaches, we would like to make them available to the plant community to empower network studies across species and generate testable hypotheses for interactions of genes whose functions are conserved across species. Therefore, in the future, we envision developing an NNI analysis pipeline and interface using our software platform VirtualPlant (www.virtualplant.org), which could potentially be implemented as an extension of this NSF ABI *innovation* grant in years 4 and 5, under NSF ABI *development* funding.

PLAN TO INTEGRATE RESEARCH AND EDUCATION:

Cross training of Biologists and Computer Scientist in Systems Biology. The development of Systems Biology tools in this project has and will involve biologists teaching computer scientists about topics like genetics, experimental genomics, and the computational challenges of analyzing genomic data. We do this informally at our weekly joint lab meetings at which graduate students and post docs from NYU Biology and NYU Courant each present their work to the group. This project involves a resident full-time senior programmer (Arthur Goldberg) and part-time systems administrator (Roberto Jimenez) working within a Biology lab, interacting closely with wet-bench biologists. The PI computer scientists (Shasha and Katari), are also involved in training and engaging computer scientist students at all levels in the emerging field of Systems Biology. In the last year, they have trained two PhD students, two interns and two MS students from Courant working in this environment. For a complete listing of students trained in the past 4.5 years, see Education and Training section in Results from Prior support.

Workshops and Classroom Training in Genomics and Systems Biology: We also provide formal training in the form of workshops and classes to enable Systems Biology. Examples of this include a weekly software workshop in "R", which aims to teach biologists how to analyze their own genomic data. A workshop on Virtual Plant has been taught two times, once by Jonathan Kelfer, a MS student working on the project and most recently by Manpreet Katari, co-PI. Students have included several faculty on sabbatical at NYU including most recently: Mary Lou Guerinot and Rob McClung of Dartmouth. Students will be exposed to Genomics and Systems Biology also through a series of formal courses offered by faculty at NYU's Center for Genomics and Systems Biology including: G23.1128 Systems Biology; G23.1130 Applied Genomics: Introduction to Bioinformatics & Network Modeling; G23.1127 Bioinformatics & Genomes. PhD students have and will continue to present their work in the weekly PhD seminar series hosted by the Biology Department. Computational students will be involved in constructing the pipeline and making it perform through the use of parallelization. Such students will also help to develop and test optimization and machine learning algorithms for network inference.

PLAN TO INTEGRATE DIVERSITY: We are committed to training scientists at the graduate and postdoctoral levels who can do independent research that cuts across fields and expertise in genomics. Our research team is also committed to diversity. Researchers in our current and previous NSF grants included Hispanic and African-American scientists. We will continue to actively seek out and recruit scientists from under-represented minorities to participate in our research in our continuing commitment to increase diversity in our research program. Five female scientists are associated with this project: Coruzzi (co-PI); Rebecca Davidson (Programmer); Varuni Prabhakar (UG Programmer); Ana Arroja (MS); Ranjita Iyer (MS Courant). Damion Nero a minority, recently graduated PhD student, has written programs contributing to the Virtual Plant project. Roberto Jimenez (Systems Admin) associated with this project is of Hispanic origin.

SHARING OF RESULTS:

Publications: The results of our analysis of the data we generate will be made available through peer-reviewed literature as it is the most appropriate way to make this information available.

MANAGEMENT PLAN: To coordinate and facilitate interactions between individuals, Dennis Shasha, the PI (NYU Computer Science) will also serve as the overall Project Manager. Gloria Coruzzi (NYU Biology) will serve as a biological advisor and conduit to a working lab and the wider plant community. The role of the Project Manager is to oversee the daily operations of the project and ensure that the needs and concerns of the participants are addressed on a day-to-day basis between the participants involved. We will also schedule day-long meetings twice a semester with our collaborator (Rodrigo Gutierrez, Chile), to do evaluation of work status and long-term planning.

Bioinformatics manager: Dr. Manpreet Katari (NYU Biology) will be in charge of the bioinformatics data. To enable efficient information exchange of raw and processed data, a file server has been set up at the NYU to store and distribute data and its analysis among users at NYU Biology and NYU Courant. This will be maintained by **Dr. Roberto Jimenez**, the Systems Administrator for this project, who will also maintain the web server, database server, and update the multinetwork databases.

Senior Programmer: Dr. Arthur Goldberg (NYU Courant, current affiliation-Memorial Sloan Kettering) will manage the development of new software analysis tools and pipelines to enable Neighborly Network Inference (NNI) which will support the different species and inference, and also new pipelines for cross species analysis, especially as they relate to crop species in coordination with the PI, and a computer science doctoral student.

Principal Investigators: Shasha and Coruzzi will each supervise personnel, organization, intellectual developments and contributions.

Role of Participants:

Name	**Institution**	**Role**
Dennis Shasha-PI	NYU Courant	Project Leader: Computational
Gloria Coruzzi-Co-PI	NYU Biology	Co-leader: Biological
Manpreet Katari-Co-PI	NYU Biology	Bioinformatics Manager
Arthur Goldberg-Senior Programmer	NYU Courant	Programmer
Rodrigo Gutierrez-Consultant	UCatolica Chile	Assembling validated networks for target species

COORDINATION WITH OUTSIDE GROUPS:

Please see attached letter of collaboration:

Rodrigo Gutierrez (U Catolica, Chile) Dr. Gutierrez, the creator of the Arabidopsis multinetwork (Gutierrez et al 2007) will assist in the assembly of multi-networks for crop species in the list of 21 species including Vitis (Grape), Corn, and Medicago.

REFERENCES CITED

1. Katari, MS, Nowicki, SD, Aceituno, FF, Nero, D, Kelfer, J, Thompson, LP, Cabello, JM, Davidson, RS, Goldberg, AP, Shasha, DE, Coruzzi, GM, and Gutierrez, RA, *VirtualPlant: a software platform to support systems biology research.* Plant Physiol, 2010. **152**(2): p. 500-515.
2. Shannon, PT, Reiss, DJ, Bonneau, R, and Baliga, NS, *The Gaggle: an open-source software system for integrating bioinformatics software and data sources.* BMC Bioinformatics, 2006. **7**: p. 176.
3. Waltman, P, Kacmarczyk, T, Bate, AR, Kearns, DB, Reiss, DJ, Eichenberger, P, and Bonneau, R, *Multi-species integrative biclustering.* Genome Biology, 2010. **11**(9): p. R96.
4. Chiu, JC, Lee, EK, Egan, MG, Sarkar, IN, Coruzzi, GM, and DeSalle, R, *OrthologID: automation of genome-scale ortholog identification within a parsimony framework.* Bioinformatics, 2006. **22**(6): p. 699-707.
5. Goloboff, PA, Farris, JS, and Nixon, KC, *TNT, a free program for phylogenetic analysis.* Cladistics, 2008. **24**(5): p. 774-786.
6. Cibrián-Jaramillo, A, De la Torre-Bárcena, J, Lee, E, Katari, M, Little, D, Stevenson, D, Martienssen, R, Coruzzi, G, and DeSalle, R, *Using Phylogenomic Patterns and Gene Ontology to Identify Proteins of Importance in Plant Evolution.* Genome Biol and Evol, 2010. **2**(0): p. 225.
7. de la Torre-Barcena, JE, Kolokotronis, SO, Lee, EK, Stevenson, DW, Brenner, ED, Katari, MS, Coruzzi, GM, and DeSalle, R, *The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data.* PLoS One, 2009. **4**(6): p. e5764.
8. Lee EK, C-JA, Kolokotronis SO, Katari MS, Stamatakis A, Ott M, Chiu JC, Little DP, Stevenson DW, Brenner ED, McCombie RW, Martienssen RA, Coruzzi GM, DeSalle R, *Functional Phylogenomics of the Seed Plants.* PloS Genetics, 2011. **Conditionally accepted**.
9. Gifford, ML, Dean, A, Gutierrez, RA, Coruzzi, GM, and Birnbaum, KD, *Cell-specific nitrogen responses mediate developmental plasticity.* Proc Natl Acad Sci (USA), 2008. **105**(2): p. 803-808.
10. Gutierrez, RA, Gifford, ML, Poultney, C, Wang, R, Shasha, DE, Coruzzi, GM, and Crawford, NM, *Insights into the genomic nitrate response using genetics and the Sungear Software System.* J Exp Bot, 2007. **58**(9): p. 2359-2367.
11. Gutierrez, RA, Lejay, LV, Dean, A, Chiaromonte, F, Shasha, DE, and Coruzzi, GM, *Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis.* Genome Biol, 2007. **8**(1): p. R7.
12. Nero, D, Katari, MS, Kelfer, J, Tranchina, D, and Coruzzi, GM, *In silico evaluation of predicted regulatory interactions in Arabidopsis thaliana.* BMC Bioinformatics, 2009. **10**: p. 435.
13. Thum, KE, Shin, MJ, Gutierrez, RA, Mukherjee, I, Katari, MS, Nero, D, Shasha, D, and Coruzzi, GM, *An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in Arabidopsis.* BMC Syst Biol, 2008. **2**: p. 31.
14. Wang, R, Tischner, R, Gutierrez, RA, Hoffman, M, Xing, X, Chen, M, Coruzzi, G, and Crawford, NM, *Genomic analysis of the nitrate response using a nitrate reductase-null mutant of Arabidopsis.* Plant Physiol, 2004. **136**(1): p. 2512-2522.
15. Gutierrez, RA, Stokes, TL, Thum, K, Xu, X, Obertello, M, Katari, MS, Tanurdzic, M, Dean, A, Nero, DC, McClung, CR, and Coruzzi, GM, *Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1.* Proc Natl Acad Sci (USA), 2008. **105**(12): p. 4939-4944.
16. Warde-Farley, D, Donaldson, SL, Comes, O, Zuberi, K, Badrawi, R, Chao, P, Franz, M, Grouios, C, Kazi, F, Lopes, CT, Maitland, A, Mostafavi, S, Montojo, J, Shao, Q, Wright, G, Bader, GD, and Morris, Q, *The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.* Nucleic Acids Res, 2010. **38 Suppl**: p. W214-220.
17. Moreno-Risueno, MA, Busch, W, and Benfey, PN, *Omics meet networks-using systems approaches to infer regulatory networks in plants.* Curr Opin Plant Biol, 2009.

18. Ashburner, M, Ball, CA, Blake, JA, Botstein, D, Butler, H, Cherry, JM, Davis, AP, Dolinski, K, Dwight, SS, Eppig, JT, Harris, MA, Hill, DP, Issel-Tarver, L, Kasarskis, A, Lewis, S, Matese, JC, Richardson, JE, Ringwald, M, Rubin, GM, and Sherlock, G, *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nature Genetics, 2000. **25**(1): p. 25-29.
19. Mewes, HW, Amid, C, Arnold, R, Frishman, D, Guldener, U, Mannhaupt, G, Munsterkotter, M, Pagel, P, Strack, N, Stumpflen, V, Warfsmann, J, and Ruepp, A, *MIPS: analysis and annotation of proteins from whole genomes.* Nucleic Acids Res, 2004. **32**(Database issue): p. D41-44.
20. Poultney, CS, Gutierrez, RA, Katari, MS, Gifford, ML, Paley, WB, Coruzzi, GM, and Shasha, DE, *Sungear: interactive visualization and functional analysis of genomic datasets.* Bioinformatics, 2007. **23**(2): p. 259-261.
21. Ferro, A, Giugno, R, Pigola, G, Pulvirenti, A, Skripin, D, Bader, GD, and Shasha, D, *NetMatch: a Cytoscape plugin for searching biological networks.* Bioinformatics, 2007. **23**(7): p. 910-912.
22. Krouk, G, Mirowski, P, LeCun, Y, Shasha, DE, and Coruzzi, GM, *Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate.* Genome Biology, 2010. **11**(12): p. R123.
23. Mirowski, P, Madhavan, D, Lecun, Y, and Kuzniecky, R, *Classification of patterns of EEG synchronization for seizure prediction.* Clin Neurophysiol, 2009. **120**(11): p. 1927-1940.
24. Craigon, DJ, James, N, Okyere, J, Higgins, J, Jotham, J, and May, S, *NASCArrays: a repository for microarray data generated by NASC's transcriptomics service.* Nucleic Acids Res, 2004. **32**(Database issue): p. D575-577.
25. Margolin, AA, Nemenman, I, Basso, K, Wiggins, C, Stolovitzky, G, Dalla Favera, R, and Califano, A, *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC bioinformatics, 2006. **7 Suppl 1**: p. S7.
26. Greenfield, A, Madar, A, Ostrer, H, and Bonneau, R, *DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models.* PLoS One, 2010. **5**(10): p. e13397.
27. Kim, S, Kim, J, and Cho, KH, *Inferring gene regulatory networks from temporal expression profiles under time-delay and noise.* Comput Biol Chem, 2007. **31**(4): p. 239-245.
28. Madar, A, Greenfield, A, Vanden-Eijnden, E, and Bonneau, R, *DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator.* PLoS One, 2010. **5**(3): p. e9803.
29. Yu, J, Smith, VA, Wang, PP, Hartemink, AJ, and Jarvis, ED, *Advances to Bayesian network inference for generating causal networks from observational biological data.* Bioinformatics, 2004. **20**(18): p. 3594-3603.
30. Altschul, SF, Madden, TL, Schaffer, AA, Zhang, J, Zhang, Z, Miller, W, and Lipman, DJ, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucl.Acids Res., 1997. **25**(17): p. 3389-3402.
31. Li, L, Stoeckert, CJ, Jr., and Roos, DS, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* Genome Res, 2003. **13**(9): p. 2178-2189.
32. O'Brien, KP, Remm, M, and Sonnhammer, EL, *Inparanoid: a comprehensive database of eukaryotic orthologs.* Nucleic Acids Res, 2005. **33**(Database issue): p. D476-480.
33. Breiman, L, *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
34. Huynh-Thu, VA, Irrthum, A, Wehenkel, L, and Geurts, P, *Inferring regulatory networks from expression data using tree-based methods.* PLoS One, 2010. **5**(9).
35. Bottou, L. *Large-scale machine learning with stochastic gradient descent*. in *Proceedings of the 19th International Conference on Computational Statistics*. 2010: Springer:Paris.
36. Shalev-Shwartz, S and Tewari, A. *Stochastic methods for l 1 regularized loss minimization*. in *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009: ACM:Montreal, Quebec, Canada.
37. Menendez, P, Kourmpetis, YA, ter Braak, CJ, and van Eeuwijk, FA, *Gene regulatory networks from multifactorial perturbations using Graphical Lasso: application to the DREAM4 challenge.* PLoS One, 2010. **5**(12): p. e14147.

38. Kanehisa, M, Goto, S, Kawashima, S, Okuno, Y, and Hattori, M, *The KEGG resource for deciphering the genome.* Nucleic Acids Res, 2004. **32 Database issue**: p. D277-280.

39. de Folter, S, Immink, RG, Kieffer, M, Parenicova, L, Henz, SR, Weigel, D, Busscher, M, Kooiker, M, Colombo, L, Kater, MM, Davies, B, and Angenent, GC, *Comprehensive interaction map of the Arabidopsis MADS Box transcription factors.* Plant Cell, 2005. **17**(5): p. 1424-1433.

40. Ding, X, Richter, T, Chen, M, Fujii, H, Seo, YS, Xie, M, Zheng, X, Kanrar, S, Stevenson, RA, Dardick, C, Li, Y, Jiang, H, Zhang, Y, Yu, F, Bartley, LE, Chern, M, Bart, R, Chen, X, Zhu, L, Farmerie, WG, Gribskov, M, Zhu, JK, Fromm, ME, Ronald, PC, and Song, WY, *A rice kinase-protein interaction map.* Plant physiology, 2009. **149**(3): p. 1478-1492.

41. Popescu, SC, Popescu, GV, Bachan, S, Zhang, Z, Gerstein, M, Snyder, M, and Dinesh-Kumar, SP, *MAPK target networks in Arabidopsis thaliana revealed using functional protein microarrays.* Genes Dev, 2009. **23**(1): p. 80-92.

42. Popescu, SC, Popescu, GV, Bachan, S, Zhang, Z, Seay, M, Gerstein, M, Snyder, M, and Dinesh-Kumar, SP, *Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays.* Proc Natl Acad Sci U S A, 2007. **104**(11): p. 4730-4735.

43. Rohila, JS, Chen, M, Chen, S, Chen, J, Cerny, RL, Dardick, C, Canlas, P, Fujii, H, Gribskov, M, Kanrar, S, Knoflicek, L, Stevenson, B, Xie, M, Xu, X, Zheng, X, Zhu, JK, Ronald, P, and Fromm, ME, *Protein-protein interactions of tandem affinity purified protein kinases from rice.* PLoS One, 2009. **4**(8): p. e6685.

44. Gholami, AM and Fellenberg, K, *Cross-species common regulatory network inference without requirement for prior gene affiliation.* Bioinformatics, 2010. **26**(8): p. 1082-1090.

45. Mutwil, M, Obro, J, Willats, WG, and Persson, S, *GeneCAT--novel webtools that combine BLAST and co-expression analyses.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W320-326.

46. Yu, H, Luscombe, NM, Lu, HX, Zhu, X, Xia, Y, Han, JD, Bertin, N, Chung, S, Vidal, M, and Gerstein, M, *Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.* Genome Res, 2004. **14**(6): p. 1107-1118.

47. Geisler-Lee, J, O'Toole, N, Ammar, R, Provart, NJ, Millar, AH, and Geisler, M, *A predicted interactome for Arabidopsis.* Plant Physiol, 2007. **145**(2): p. 317-329.

48. Bader, G, Betel, D, and Hogue, C, *BIND: the Biomolecular Interaction Network Database.* Nucleic Acids Res., 2002. **31**: p. 248.

49. Masoudi-Nejad, A, Goto, S, Endo, TR, and Kanehisa, M, *KEGG bioinformatics resource for plant genomics research.* Methods Mol Biol, 2007. **406**: p. 437-458.

50. Zhang, P, Foerster, H, Tissier, CP, Mueller, L, and Paley, S, *MetaCyc and AraCyc. Metabolic pathway databases for plant research.* Plant Physiol., 2005. **138**: p. 27.

51. Tatusov, RL, Galperin, MY, Natale, DA, and Koonin, EV, *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic Acids Res, 2000. **28**(1): p. 33-36.

52. Zhang, J, *Evolution by gene duplication: an update.* TRENDS in Ecology and Evolution, 2003. **18**(6): p. 292-298.

53. Barrett, T, Troup, DB, Wilhite, SE, Ledoux, P, Rudnev, D, Evangelista, C, Kim, IF, Soboleva, A, Tomashevsky, M, and Edgar, R, *NCBI GEO: mining tens of millions of expression profiles--database and tools update.* Nucleic Acids Res, 2007. **35**(Database issue): p. D760-765.

54. Griffiths-Jones, S, Grocock, RJ, van Dongen, S, Bateman, A, and Enright, AJ, *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Res, 2006. **34**(Database issue): p. D140-144.

55. Gustafson, AM, Allen, E, Givan, S, Smith, D, Carrington, JC, and Kasschau, KD, *ASRP: the Arabidopsis Small RNA Project Database.* Nucleic Acids Res., 2005. **33**: p. D637.

56. Lu, S, Sun, YH, Shi, R, Clark, C, Li, L, and Chiang, VL, *Novel and mechanical stress-responsive microRNAs in Populus trichocarpa that are absent from Arabidopsis.* Plant Cell, 2005. **17**: p. 2186.

57. Wang, L, Xie, W, Chen, Y, Tang, W, Yang, J, Ye, R, Liu, L, Lin, Y, Xu, C, Xiao, J, and Zhang, Q, *A dynamic gene expression atlas covering the entire life cycle of rice.* Plant J, 2009.
58. Settles, B, *Active learning literature survey*. Computer Sciences Technical Report, 2010. **1648**.
59. King, RD, Rowland, J, Oliver, SG, Young, M, Aubrey, W, Byrne, E, Liakata, M, Markham, M, Pir, P, Soldatova, LN, Sparkes, A, Whelan, KE, and Clare, A, *The automation of science.* Science, 2009. **324**(5923): p. 85-89.
60. King, RD, Whelan, KE, Jones, FM, Reiser, PG, Bryant, CH, Muggleton, SH, Kell, DB, and Oliver, SG, *Functional genomic hypothesis generation and experimentation by a robot scientist.* Nature, 2004. **427**(6971): p. 247-252.
61. Lewis, D and Gale, W, *Training text classifiers by uncertainty sampling.* Research and Development in Information Retrival, 1994.
62. Settles, B and Craven, M. *An analysis of active learning strategies for sequence labeling tasks*. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2008: ACL Press.
63. Settles, B, Craven, M, and Ray, S. *Multiple-instance active learning*. in *Advances in Neural Information Processing Systems (NIPS)*. 2008: MIT Press.
64. Cui, J, Li, P, Li, G, Xu, F, Zhao, C, Li, Y, Yang, Z, Wang, G, Yu, Q, and Shi, T, *AtPID: Arabidopsis thaliana protein interactome database--an integrative platform for plant systems biology.* Nucleic Acids Res, 2008. **36**(Database issue): p. D999-1008.
65. Davuluri, R, Sun, H, Palaniswamy, S, Matthews, N, Molina, C, Kurtz, M, and Grotewold, E, *AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors.* BMC Bioinformatics, 2003. **4**(1): p. 25.
66. Lu, C, Tej, SS, Luo, S, Haudenschild, CD, Meyers, BC, and Green, PJ, *Elucidation of the small RNA component of the transcriptome.* Science, 2005. **309**: p. 1525.
67. Mueller, LA, Zhang, P, and Rhee, SY, *AraCyc: a biochemical pathway database for Arabidopsis.* Plant Physiol, 2003. **132**(2): p. 453-460.
68. Rzhetsky, A, Iossifov, I, Koike, T, Krauthammer, M, Kra, P, Morris, M, Yu, H, Duboue, PA, Weng, W, Wilbur, WJ, Hatzivassiloglou, V, and Friedman, C, *GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data.* J Biomed Inform, 2004. **37**(1): p. 43-53.