**RESULTS FROM PRIOR NSF SUPPORT:** Because one outcome of this proposal will be a computational platform for translational gene discovery, we briefly describe our completed NSF Grant DBI-0445666, “Conceptual Data Integration for the Virtual Plant.” The VirtualPlant software platform (www.virtualplant.org) [Katari 2010] 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 within single species. We have demonstrated the use of tools embodied in the VirtualPlant system to generate hypotheses that were subsequently experimentally validated [Gifford 2008; Gutierrez 2007 JExpBot; Gutierrez 2007 Genome Biol; Nero 2009;Thum 2008; Wang 2004;Gutierrez 2008 PNAS].

***Our NSF VirtualPlant grant had four goals***: **Integration**, **Visualization**, **Synthesis**, and **Prediction**.

**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 [Katari 2010; Gutierrez 2007 Genome Biol]. The Arabidopsis multinetwork in VirtualPlant has 16,562 nodes (of which 13,960 are genes) and 97,423 interactions [Katari et al 2010]. This Arabidopsis 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”[Gutierrez 2008 PNAS]. 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) [Gutierrez 2008 PNAS]. In this example, 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 [Gifford 2008; Gutierrez 2007 Genome Biol; Nero 2009; Thum 2008]. A complementary network tool is GeneMania [Wade-Farley 2010] which generates a hypothesis for gene function based on interactions with other genes and their attributes. More recently, AraNet reports a genome-scale functional network for Arabidopsis – which, like VirtualPlant multinetwork, combines data from multiple sources about gene and protein interactions [Lee 2010 Nature Biotech “Rational Association of Genes….]. For a recent review of various plant multinetwork approaches, see [Moreno-Risueno 2009] and [Lee 2010 Nature Biotech].

**Aims 2 & 3. Synthesis and Visualization: *VirtualPlant’s primary analysis tools and functions.*** In addition to the multinetwork, the VirtualPlant platform ([www.virtualplant.org](http://www.virtualplant.org)) houses tools for data analysis, integration and visualization. Here are some examples.

**BioMaps**: BioMaps takes one or more sets of genes and determines which functional terms (GO [Ashburner 2000] or MIPS [Mewes 2004] ) 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 or as a directed acyclic graph [Gutierrez 2007] [Katari 2010].

**Sungear**: Sungear enables a visually interactive and biologist-driven exploration of experiments/lists, all of their disjoint intersections, and their related ontological terms. The position and size of a circle is proportional to the number of genes in the intersection of those lists (see [Poultney 2007]). Biologists find Sungear to be an extremely powerful and interactive tool for analyzing the interrelationships between sets of genes [Gutierrez 2007, J Exp Bot].

**NetMatch:** NetMatch, a Cytoscape plug-in, finds all instances of a query graph (e.g. a network motif) in a larger graph [Ferro 2007]. New versions compute the statistical significance of the motifs (e.g. Transcription factor motifs) found in a network.

**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 [Krouk 2010 Genome Biol; Mirowski 2009]. Our second goal was to extend VirtualPlant to other single species datasets such as Rice (see www.virtualplant.org).

**VirtualPlant Database**: The VirtualPlant database contains some of the most commonly used data types including metabolic pathways from KEGG [Kanehisa 2004] and ARACYC [Mueller 2003], protein-protein interactions from BIND [Bader 2002] and Interolog databases for Arabidopsis [Geisler-Lee 2007], and GeneOntology and annotations from TAIR. The multinetwork database also contains processed data Microarray experiments obtained from NASC [Craigon 2004].

**Software and Data Availability**: VirtualPlant is accessible via the website www.virtualplant.org. Registered users (currently > 700) store their data sets and use many tools to analyze their genomic data such as microarray experiments. The website is available for free when used for non-for-profit purposes.

**VirtualPlant and User Community:** The VirtualPlant user community consists of >700 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).

**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 Cytoscapeplugin for searching biological networks” ***Bioinformatics***, 2007 23(7):910-912

**Applications of VirtualPlant: Hypothesis Generation and Testing**

Lee E, Katari M, Kolokotronis S, Cibrian A, Stamatakis A, Ott M, Little D, Stevenson D,

McCombie WR, Chiu J, Martienssen R, Brenner E, Coruzzi G, DeSalle R (2011) “High

resolution phylogeny of the seed plants: A functional phylogenomic view.” ***PLoS***

***Genetics*** Dec;7(12):e1002411. Epub 2011 Dec 15.

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

Vidal EA, Araus V, Lu C, Parry G, Green PJ, Coruzzi GM, Gutiérrez RA (2010). Nitrate-

responsive miR393/AFB3 regulatory module controls root system architecture in

Arabidopsis thaliana. ***Proc Natl Acad Sci U S A.*** 107(9):4477-82

Das M, Reichman JR, Haberer G, Welzl G, Aceituno FF, Mader MT, Watrud LS, Pfleeger TG,

Gutiérrez RA, Schäffner AR, Olszyk DM. (2010) A composite transcriptional signature

differentiates responses towards closely related herbicides in Arabidopsis thaliana and

Brassica napus. ***Plant Mol Biol***. 2010 Mar;72(4-5):545-56.

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.

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)*

**Computational Publications**

Di Natale R, Ferro A, Giugno R, Mongiovi M, Pulvirenti A and Shasha D (2010) "SING:

Subgraph search In Non-homogeneous Graphs" ***BMC Bioinformatics***,

11:96doi:10.1186/1471-2105-11-96

Zhang X, D. Shasha, Y. Song and J. T. L. Wang (2010) “Fast Elastic Peak Detection for Mass

Spectrometry Data Mining,” *IEEE Transactions on Knowledge and Data Engineering*.

Issue 99. November 29, 2010, doi: 10.1109/TKDE.2010.238

**Plant Systems Biology: Reviews, Books and Outreach**

Kraiser T, Gras DE, Gutiérrez A, González G, Gutiérrez R (2011). From molecular to the

ecosystem level: A holistic view of N-acquisition in plants. ***J Exp Bot***. 62, 1455-66.

Krouk G, Crawford NM, Coruzzi GM, Tsay YF (2010) “[Nitrate signaling: adaptation to fluctuating environments](http://www.ncbi.nlm.nih.gov/pubmed/20093067?itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum&ordinalpos=2)” ***Current Opinion in Plant Biol***. 13(3):266-73

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

C. R. McClung and Gutiérrez RA (2010). Network news: prime time for systems biology of the

plant circadian clock. ***Curr Op Genet and Dev***. 20, 588-598.

Vidal EA, Tamayo KP, and Gutiérrez RA (2010). Gene networks for nitrogen sensing, signaling,

and response in Arabidopsis thaliana. Wiley Interdisciplinary Reviews: ***Systems Biology***

***and Medicine***. 2, 683-693. (doi:10.1002/wsbm.87)

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 VirtualPlant software platform has trained undergraduates (UG), master’s (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, Computer Science), 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, NYU Courant), Aris Tsirigos (PhD, NYU Courant), Saurabh Kumar (PhD, 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). **High School Students**: Angela Fan (Stuyvesant HS) – Siemans Semi-Finalist, Intel Finalist 2012; Jenny Kim (Chapin HS).

**RELEVANCE OF THE PRESENT PROPOSALS TO THE STATED GOALS OF THE PGRP**

We propose first to develop approaches, tools, and pipelines that exploit the large amount of data on well-studied plant species to infer networks on new and emerging species. Second, we propose a cross-species network approach to predict the functionality of genes, then test them on model species and on crop species. Thus, this work will enhance translational research and enable the prediction of network states under untested conditions. Our project addresses the PGRP’s goals in the following ways:

1. *Advance plant systems biology*: Utilize the large amount of data on well-studied plant species to infer networks on new and emerging species (e.g. crop species).

2. *Translate basic discovery to field*: Derive gene networks that will contribute to gene discovery in crops like Maize (Aims 2).

3. *Develop coordinated solutions to data access, analysis and synthesis:* Develop and deploy a platform that will enable researchers to synthesize knowledge across species of interest to identify network modules for hypothesis testing (Aim 3).

4. *Enhance education, training and outreach*: Collaborative training in Plant Systems Biology from biologists and computer scientists. Outreach to science museums.

5. *Broaden societal impacts of Systems Biology*: Enable *in silico* predictions for modifying traits of

agronomic and/or environmental value.

**RESEARCH PLAN**

**Aim 1: Development of InferNET: Inferring networks in data-poor species.**

***Rationale***: In this Aim, we propose to develop a network inference approach, ***InferNET***,that will ***infer*** regulatory networks in a data-poor target species based on gene networks from several data-rich species. The *InferNET* approach is inspired by a variant of the *Robin Hood philosophy* - “Learning from the rich and giving to the poor.” Inferred networks in the data-poor target (e.g. crop) species may be used to derive hypotheses and identify potentially important genes for validation testing. Most of the discussion below concerns co-expression networks, because, with the advent of Next-Gen sequencing, such data is now readily available and/or obtainable for many species. However, the methods we develop apply to other types of network edges (e.g. protein interaction for the Rice protein interaction project [web site (Vidal and Ecker)]), and we will apply them as data becomes available.

***Novelty***: As its name suggests, ***InferNET*** uses data-rich species to ***infer*** regulatory networks in data-poor species. By contrast, existing tools for comparing plant gene networks create networks based only on experimental data and then compares them post-hoc (e.g. CoP [Ogata 2010], StarNet [Jupiter 2009], ATTED-II [Obayashi 2011, and PlaNet [Mutwil 2011]). Also, with the exception of PlaNet, such tools compare only two species at a time. InferNET, like PlaNet, uses the networks from multiple species simultaneously.

***Community Need:*** As the number of available genome sequences increases, it will be common to find a newly sequenced or poorly studied target species “*t*” that is phylogenomically similar according to patristic distance based on maximum parsimony [Fourment and Gibbs /BMC Evolutionary Biology/ 2006, \*6\*:1] (see phylogenetic tree in Fig X.) to those few “data-rich” species having a substantial body of experiments. For targets such as emerging crops and new “boutique” crop species, inferring networks will be particularly valuable.

**IMPLEMENTATION AND TESTING OF *InferNET*:**

**Species:** *InferNET* will mine the current 21 fully sequenced species (Fig. X) and we will add additional species to the *InferNET* analysis pipeline as their sequences become available. This would include species with fully sequenced and annotated genomes, and can also include species with fragmented “gene space” assemblies as are likely to be produced by Next-Generation sequencing technologies.

**Defining data-rich species**: To determine whether a species is indeed “data-rich,” we will use a technique analogous to Statistical Power Analysis [Hill, T. & Lewicki, P. (2007). STATISTICS: Methods and Applications. StatSoft, Tulsa, OK]. Mechanically, this consists of computing the p-values of large positive (r value >= 0.5) and large negative (r value <= -0.5) correlations within some species for the experiments already done on that species. If a large portion (say 70%) of those have p-values below 0.05, then the species is data-rich. Admittedly, these thresholds are somewhat arbitrary, but they divide the 21 species reasonably. For now, thespecies Arabidopsis, Poplar, Medicago, Soybean, Rice, and Maize of Fig. X are designated as data-rich.

**Correlation networks**: The basic co-expression metric we use to identify correlation networks will be Pearson correlation because it has been shown to be particularly useful in inferring functionality in current cross-species network studies [Mutwil 2011] [Usadel 2009], [[Klie et al., 2010](http://www.plantcell.org/content/23/3/895.full#ref-28)]. However, in the course of this study, we will also test our methods of analysis using other metrics including mutual information [Margolin 2006], Mutual Rank [Obayashi et al], and Spearman correlation [Hill 2007]. Our approach will be to train the *InferNET* algorithm using two or more data-rich source species (s1, s2, …), and then to apply the trained model to data-poor target species *t*. We validate the success of each in making valid predictions, using experimental data from data-rich species.

**Learning the Rules**: The *InferNET* training itself will be done as follows: Take several data-rich source species *s1, s2, …, sk,* and *learn* the parameters of a regression model that predicts co-expression edges on one of those species. We then use that model to predict edges in the data-poor target species *t*.

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

**orthotab: target species| target gene | source species | source gene | orthology val1 | orthology val2 …**: gives the gene-to-gene orthology value, according to several different orthology measures, for example: Reciprocal best BLAST hits [Altschul 1997 Nuc Acid Resh], BLAST hits above a threshold, OrthologID [Chiu 2006 Bioinformatics], OrthoMCL [Li 2003 Genome Research], and Inparanoid [O’Brien 2005 Nuc. Acid Resh]. Our preliminary work used reciprocal best BLAST hits. Part of the machine-learning research will be used to determine which orthology method or methods will work best.

**edgetab: species | gene1 | gene2 | edgetype | strength | p-value**: gives the strength and the p-value of a given experimentally supported edge (e.g. by data including expression correlation). In our preliminary studies, we examine gene expression correlations that hold over all conditions. Edge relationships present only under certain conditions (e.g. drought conditions) or in certain tissues (“Gene Spaces”) can be retrieved using a focused set of experiments, as done in Aim 2. The machine learning stays the same, but the data can change. We will consider Microarray and RNA-seq data interchangeable under the proper normalization protocols [Bullard et al 2010].

**speciestab (species1 | species2 | species similarity measure1 | species similarity measure2)**: measures sequence similarity of species according to several criteria (e.g. distance based, for example, average percent identity of protein sequences, or through parsimony). Which similarity measure will work the best can be determined in the course of learning the coefficients of our Species Combining Rule.

**Machine Learning**: Now, to predict an edge between *g1* and *g2* in a data-poor target species *t*, we will combine evidence from edges in one or more data-rich source species s1, s2, …, as well as evidence from any experiments conducted in the data-poor target species *t* itself. The basic machine learning method will be Linear Regression and Regression Trees, with a penalty for complexity. For the sake of performance and robustness to noise, we will use one of the following three machine learning approaches:

1. **Random Forests [**Breiman 2001 Machine learning, Huynh-Thu 2010 PloS On**e]** 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** [Bottou 2010] Bottou demonstrated that a stochastic gradient descent solver for a variety of learning problems (including support vector machine optimization) is able to scale to extremely large datasets, while converging to the predictive performance of traditional optimization algorithms.

3. **Large-Scale L-Regularized Learning [Shalev-Shwartz 2009]** 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.

**Validation testing of InferNET predictions**: The net effect of these machine-learning analyses will be to find the weighting of different factors (e.g. that correlation of source edges is more important than gene sequence orthology), that will lead us to estimate the correlation between two genes in some target species *t*. To determine which machine learning method is best, we will test them first on the data-rich species in “*hide-the-answer*” experiments. That is, we compare the predicted results (e.g. inferred edges in the target species t) that use no expression experiments (*hide-the-answer*) from the target species with the results from the experiments in the target species (see Table X). This analysis gives us precision and recall data to quantify the success of our prediction methods.

**Pilot study of *InferNET***: In our pilot study, we tried to infer Pearson correlation edges in a “target” species, Soy, knowing correlation edges in a “source” species, Arabidopsis, trained using another data-rich species, Medicago, and the gene-by-gene orthology between genes in Arabidopsis and both Medicago and Soy (Fig. X). We selected these three species as an initial proof of concept because (i) there is ample and reliable Affymetrix data for each, enabling us to validate our predictions, and (ii) Medicago and Soybean – both legumes – are quite closely related. The equation for network inference is trained using Arabidopsis and Medicago under an L-Regularized learning algorithm **[Shalev-Shwartz 2009]**.Once we “learn” the rules for network structure using Arabidopsis and Medicago data, we applied this learned equation to infer edges in Soy. To test whether this approach worked, we evaluated the predictions using actual expression data from Soy.

Our preliminary studies show that our “InferNET” learning approach is superior to the Interolog approach which considers only BLAST scores to infer edges [Yu (2004) Genome Research,Annotation Transfer Between Genomes: Protein–Protein Interologs and Protein–DNA Regulogs ]. Under the Interolog approach, if (i) the co-expression edge between *g1* and *g2* in Arabidopsis has a certain correlation value *r,* (ii) *g1’* in the target (Soy) is the reciprocal top BLAST hit for *g1*, and (iii) *g2’* is the reciprocal top BLAST hit for *g2*, then the approach infers a correlation of *r* between *g1’* and *g2’*. As we show in Table X below, *InferNET* has better recall (88% vs. 81%) and precision (77% vs. 69%) than the Interolog approach, even though Interolog by itself is quite informative.

***For our proof of concept study, the regression model had the following form:***

Estimated correlation in **target species *t*** (ECT) = a1\*Mean of Orthology values (MOv) + a2\*correlation of source pair (Cs) + a3\*p-value of correlation of source pair (Ps) (Fig. X).

This form of the regression model equation was chosen based on our expectation that the strength of correlation in the target species will depend on some statistic on the orthology assignments (a1\*MOv) and the strength and confidence in the correlation of expression in source species (a2\*Cs and a3\*Ps). For this study, mean of orthologous values is calculated as follows: if g1 and g2 are the source pair, and g1' and g2’ are the potential new target pair, and g1 and g1’ are reciprocally best BLAST hits (as are g2 and g2’), then we take the Mean of the Orthology values (MOv). We chose the simple linear form of this equation for ease of understanding. The learned model also suggests why the Interolog approach [Yu 2004] works as well as it does. Whereas the MOv value and the correlation values both have absolute values between 0.5 and 1, the coefficient for correlation is 40 times greater than the correlation for orthology (1.2 vs. 0.03), implying that the specific value of orthology is unimportant for reciprocal BLAST hits. Often, the correlation of the edge in the source species by itself predicts the correlation in the target.

Since there are a different number of experiments for each species 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% of all measured correlations, and “highly negatively correlated,” if their correlation is in the bottom 5% (Table X). Thus, our machine-learning algorithm starts from the 5% most positively and negatively correlated pairs in Arabidopsis and infers positive or negative correlations about edges in the target (e.g. Soy) for reciprocal top BLAST hits of those elite pairs.

**[Becca: the table you are doing in power point should replace this one. Do keep the caption however.] We are still working on the exact numbers but the table should look something like this**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predicting networks in Soy (Glycine Max)** | | | | |
| **Method** | **Positive Recall** | **Positive Precision** | **Negative Recall** | **Negative Precision** |
| **InferNET** | **88% (xxx/yyy)** | **79% (xxx/yyy)** | **73% (xxx/yyy)** | **83% (xxx/yyy)** |
| **Interolog** | **81% (xxx/yyy)** | **69% (xxx/yyy)** | **65% (xxx/yyy)** | **78% (xxx/yyy)** |
|  |  |  |  |  |
| **Predicting networks in Medicago** | | | | |
| **Method** | **Positive Recall** | **Positive Precision** | **Negative Recall** | **Negative Precision** |
| **InferNET** | **xx% (xxx/yyy)** | **xx% (xxx/yyy)** | **xx% (xxx/yyy)** | **xx% (xxx/yyy)** |
| **Interolog** | **xx% (xxx/yyy)** | **xx% (xxx/yyy)** | **xx% (xxx/yyy)** | **xx% (xxx/yyy)** |

**Table X Caption**: Positive recall is the number of gene pairs in the target species correctly predicted to be positively correlated divided by the number of gene pairs that are positively correlated. Positive precision is the number of gene pairs correctly predicted to be positively correlated divided by the total number predicted to be positively correlated. Similar for negative correlation. The inferred correlation coefficients are XXXXXXXXX [TO BE FILLED IN].

**Limitations of the Proof-of-Concept Model and Planned Improvements of InferNET:**

**Orthology assignments**: In our future work, instead of using reciprocal top BLAST hits when inferring the correlation between some target pair g1’ and g2’, we will consider all gene pairs g11, g21; g12, g22 such that each g1i is above a similarity threshold GENESIM to g1’, and g2i is above the same similarity threshold GENESIM to g2’. This will imply that many gene pairs may be relevant to the prediction of a given target pair g1’ and g2’. This, in turn, implies the need for some form of aggregation over the correlation for potentially relevant gene pairs. We will include terms for mean (weighted by gene orthology), median, max, and min as the most representative aggregates. Each of the three machine learning mechanisms we will test will determine the weights for each term. We will also determine, based on cross-validation, the best gene orthology threshold, GENESIM.

**Incorporation of target species data**: In further development of *InferNET*, we will incorporate the limited expression data that is already available for the target species into the learning equation. The net result will be for the edge g1’ between g2’, a term for an experimentally derived correlation and a term for the experimentally derived p-value.

**Use of additional species in training**: Further, in future development and testing of *InferNET*, we will be using more than two species for training. For example, based on available expression datasets we might train on Arabidopsis using data from two data-rich legume species (Soy and Medicago) and then apply the learned model on Cucumis (a data-poor species), or we would train on Rice using Maize and Sorghum as data-rich species and apply the model to Brachypodium, Setaria etc. In general, we might learn a model using *s1*, *s2*, *s3*, and *s4* and train on *s5*, then apply that model to a target species *t*. We will first create a model for each source-train species independently (e.g. from Arabidopsis to Glycine and then from Poplar to Medicago). Then we will form a “species combining rule” consisting of a learned joint ranking of the several regression models weighted by phylogenomic similarity. The weights on the conclusions from each species will be learned using one of the three machine learning methods above.

**Expected outcomes of Aim 1 and future directions:** 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 “target” species by inference from one or more data-rich “source” species. The success of the preliminary results suggest reason for optimism. We will apply the same techniques to other edge types (e.g. protein-protein interaction).

**Aim 2A:** **Trait-to-gene network module discovery pipeline**

***Rationale***: We propose to develop and implement methods to build gene networks from co-expression data on multiple crop species to identify genes that are potentially central to a particular trait of interest. We will then test candidate genes by over-expression, knock-outs, and knock-ins in Arabidopsis.

***Novelty***: The trait-to-network module approach follows the spirit of AraNet [Lee 2010] and PlaNet [Mutwill 2011] in that multiple species are used to identify functionality in gene networks. What is novel in our proposed approach is: (i) we use expression data from trait-relevant experiments on multiple crop species to “weight” edges in the network and identify sets of genes associated with a trait, (ii) we identify orthologous genes that are relevant to a trait, some of which may be missing in Arabidopsis, and (iii) our experimental strategy is based on a medium throughput validation testing in an inducible expression system in Arabidopsis.

**Significance: Agronomic traits and phylogenomic context**: Since the dawn of agriculture, farmers and scientists have improved crops by selection and breeding. Now that many crop and non-crop species have been sequenced, these species can be related through phylogeny Fig. X. Among the 21 sequenced species in the tree, the crop species lie at phylogenetic distances of a hundred million years or more from a common ancestor [Chaw et.al. 2004]. Nevertheless, there is evidence for conservation of specific gene functions across these species, sometimes at large phylogenetic distances [Irish and Yamamoto 1995]. As such, complex agronomic traits, such as seed development, seed composition, root architecture, flowering time etc., likely result from medium-sized conserved networks of genes rather than single genes [Espinosa-soto et al., 2006]. Arabidopsis, through its amenability to mutational and transformational studies, has provided the vast majority of knowledge about these traits, e.g. flowering time [Espinosa-soto et al., The Plant Cell Nov 2004], seed development [To et al., The Plant Cell July 2006] and root architecture [Péret et. Al., Trends in Plant Science July 2009]. Because Arabidopsis has limited tolerance to extreme conditions, data about environment-specific responses of genes is easier to obtain in other species [Li et al. 2011, Tuteja et al. 2010]. Examples include drought resistance [Shen Y, Venu RC, Nobuta K, Wu X et al. 2011] and early seed development [http://www.ncbi.nlm.nih.gov/geo/ : GSE29163]. In addition, expression atlases exist for each of several crop species [Severin et al. BMC Plant Biology2010].

**Gene Discovery**: The combined computational-experimental approach described below capitalizes on mining genomic data from crop species to inform the identification of gene network modules in Arabidopsis associated with traits of agronomic interest. This will lead to novel gene discoveries that can be tested using Arabidopsis as a model. For example, in some cases, this comparative network analysis will identify Arabidopsis orthologs of crop genes whose significance in the trait could not have been identified based solely on Arabidopsis data. In other cases, it will identify nodes of networks that are “missing” in Arabidopsis but present in crops. These crop genes could be “knocked-in” to Arabidopsis using a protoplast transient assay system we describe in Aim 2B. Here is the method:

**Step 1**: **Identify** **trait-related expression datasets**: For each trait, collect one set of experiments [NCBI GEO] that are relevant to the trait (e.g. during seed development). Collect a second set of expression data under unrelated conditions. Genes that do not show variation, as determined by a minimum variance cut-off across all experiments, are “housekeeping” genes, and are removed from further analysis. Further filters are applied to remove poor quality data (e.g. low counts). After filtering, raw counts are normalized using a full-quantile normalization method [Bullard et.al. BMC Bioinformatics 2010].

**Step 2**: **Compute gene correlations**: Build gene correlation networks separately in each of species S1…Sn. We will use methods for correlation proven useful in other plant network tools including Pearson correlation coefficient [Usadel et. al, 2009], as well as cosine coefficient (CoP) [Ogata et. al., Bioinformatics 2010] and Mutual rank, [ATTEDII, PlaNet] [Obayashi et.al., NAR 2008; Mutwil et.al., Plant cell, 2011] and test other correlation methods discussed in Aim 1. A gene correlation network (CNi) for species Si consists of edges {g1, g2}, provided the absolute value of the correlation between these two genes is at least 0.5 with p-value <=0.05

**Step 3. Consensus through “weighting” of nodes and edges:** Each gene g, from a specific species network that passes an orthology cutoff (Stringent BLAST e-value cutoff) (Kranthi- where is this cut off determined?) with respect to some Arabidopsis gene g’, will give g’ a “weighted” value. Thus g’ may receive “weight” values from several genes in each network and from multiple species networks. Further, the gene g in one network, may add to the “weight” for nodes or edges of several gene pairs in Arabidopsis as shown in Fig X. If g is not orthologous enough to any gene in Arabidopsis but is orthologous enough to genes in other crop species, then we consider that gene to be a candidate for a “knock-in” experiment in Arabidopsis. If there is an edge between g1 and g2 in a species-specific network, and g1 exceeds the orthology cutoff to g1’ in Arabidopsis (as does g2 with respect to g2’), then add a “weight” to the edge between g1’ and g2’. Experimentally validated edges from Arabidopsis provide additional weights to the network edges and nodes. Edges representing predicted Cis-binding sites for Arabidopsis transcription factors are added to provide confidence and direction to the regulation network. The final network including nodes and edges, with assigned weights, forms the Treturn network within Arabidopsis

**Step 4**: **Identify conserved network modules**: The reciprocal of the weights of the edges form a measure of distance, thus assigning low distance to genes that have often been associated together. Next, we perform k-means clustering or possibly affinity propagation clustering [Frey and Dueck, Science 2007] based on this distance measure. Candidate clusters may be ranked based on the median “weight” count for the nodes, the mean thickness of edges, or edge density.

**Step 5: Mutant studies in Arabidopsis**: Identify genes central to each network module and prioritize genes that are poorly characterized in (or even absent from) Arabidopsis. Mutagenize those genes by creating knock-out mutants, “knock-ins” (of the missing gene), or over-expression lines. This is explained further in Aim 2B.

**Dealing with potential Orthology and Paralogy issues**: By collecting weights supporting nodes and edges from multiple species, the “weighted” network includes all paralogs across species. Because the propagation of edges between all pairs of orthologs may inflate the weight assigned to some pairs of paralogs, we will explore methods to prioritize genes for phenotypic assays by ranking the genes within a network module based on the average number of paralogs and/or gene family members across the species.

**Proof-of Principle Preliminary Analysis:** To test this approach, we used seed development as the trait of interest [Baud and Lepiniec, , 2010]. In Arabidopsis, LEC1, LEC2, FUS3, ABI3 and ABI5 are known master regulators of seed development and likely exert important influence on early stages of seed nutrient accumulation [Santos-Mendoza et. al.2010. The seed trait has been studied in multiple species, and ample mutant phenotype information is available for *in silico* validation of our network predictions [Meinke D et. al., 2008].

**Construction of gene correlation networks**: We have used deep transcriptome data sets from early seed tissue samples of Soybean and Maize to perform this preliminary analysis [NCBI GEO] [REF]. Following the specifications of the first two steps above, we found CNsoybean and CNmaize. Then, we assigned orthologs to Arabidopsis and each other (Kranthi- what do you mean by “each other”? Soy and Maize? Unclear). For this preliminary work, orthology was assigned based on best reciprocal BLAST matches. More sophisticated approaches to orthology assignment will be used in the final work, as discussed above in Aim 1. This unified network contains XX nodes (YY genes) and ZZ edges. The distance between nodes, computed as the reciprocal of edge weight [REFERENCE], was used to cluster the nodes. This distance matrix was subjected to k-means clustering to identify conserved clusters of co-expression [REFERENCE].

REST OF THE PRELIMINARY RESULTS WILL DEPEND ON THE RESULTS OBTAINED AT THIS STAGE. WE ANTICIPATE THAT LEC1, LEC2, FUS3, ABI3 AND ABI5 GENES WILL BE REDISCOVERED IN ADDITION TO OTHER GENES THAT INTERACT WITH THEM.

IF ANY UNANNOTATED GENES APPEAR IN THE NETWORK WE CAN FOCUS ON THOSE FOR CAREFUL ANNOTATION TO IDENTIFY POSSIBLE ROLES. ANY GENES MISSING INARABIDOPSIS ARE OBVIOUS CANDIDATES FOR KNOCK-IN STUDIES. SUCH “MISSING” GENES WILL BE RANKED BY AVERAGE NUMBER OF PARALOGS ACROSS SPECIES.

**Aim 2B: Experimental Validation Strategy**. In Aim 2A, we develop and test the method for exploiting data associated with traits in crop species to inform network studies in Arabidopsis, which ultimately will aid in translational studies back to crops. As proof-of-principle, we tested this approach for a trait, “seed development,” for which there were ample mutant data to validate the genes uncovered in our networks. In Aim 2B, we test in Arabidopsis genes that are important for seed development in crops. We will use a medium throughput dexamethasone inducible transient assay system to assay transcription factors and identify their targets in vivo [Sablowski and Meyerowits Cell 1998]. For TFs that pass initial validation in this transient protoplast system, we will proceed to stable transformants (e.g. T-DNA, overexpression, or “knock in” for cases where the gene is missing in Arabidopsis).

**Experimental method**: We have developed an approach (named “Network Walking”) aimed at analyzing the network targets of a transcription factor in *Arabidopsis* FACS sorted protoplasts. This rapid approach identifies transcription factor targets in less than a week of experimentation following methods developed by Bargmann and Birnbaum [Bargmann BO, Birnbaum KD (2009)]. Using Gateway™technology, we have engineered a vector for which any TF can be fused with a GR (the glucocorticoid receptor) tag. This 35S-TF-GR chimera allows one to (i) overproduce the studied TF in the protoplasts, and to (ii) control its entrance into the nucleus using dexamethasone (DEX) treatment [Lloyd *et al*, 1994.][Sablowski and Meyerowitz, 1998] [Bargmann BO, Birnbaum KD (2009)][Bargmann BO, Birnbaum KD (2010)]. Recently, we successfully used this “network-walking” approach (FACS assisted protoplast selection + DEX fusion) to study network targets of the well-studied TF, ABI3 [Bargmann et al 2012, In Preparation].

Sentinel genes predicted to be targets of the TF (based on the “weighted network”) in seed development will be assayed by Q-PCR to ascertain any affect this gene might have on their steady state expression levels. A transcription factor that significantly changes the expression level of one or more sentinel genes is assumed to be involved in the regulation (direct or indirect) of that gene. Positive results will be followed up with (i) transcriptome responses in the protoplast system, and (ii) *in planta* experiments in Arabidopsis (e.g. T-DNA mutants, overexpression) as well as “knock ins” for genes not present in Arabidopsis. The approach of identifying regulatory hubs was highly successful in identifying master regulatory genes in N-usage in Arabidopsis [Gutierrez et al 2008] Similarly, this “weighted network” approach that is fueled by gene correlation networks in crops will enable us to target genes for study in Arabidopsis with high potential for translational impact in crops.

**Aim 3: X-Net Builder: A Platform for Cross Species Network building and inference.**

***Rationale*:** We propose to build the X-Net Builder, an intuitive web interface that will give biologists access to all the data, tools, and analysis pipelines required to build gene networks based on experimental and/or inferred data. The end user can build both (i) species-specific networks consisting of multiple edge types (multinetworks, for short) and (ii) cross-species weighted networks, where the weights of edges are determined by the amount of support an edge has. Users can create these plant networks using the tools developed in Aims 1 and 2, and query them using the interface described below. This will lead the experimental biologists to identify candidate networks of genes, which they can experimentally validate.

***Novelty*:** Other web-based tools that allow researchers to query and browse plant gene networks made from data-rich species, such as PlaNet (Mutwil 2011) and ATTED-II (Obayashi 2011), offer large pre-calculated networks that may get updated periodically. By comparison, X-Net allows biologists to (i) create predicted networks for data-poor species, (ii) create networks based on subsets of experiments, and (iii) to create multinetworks, and/or weighted networks, using data from multiple species. Because these networks are created “on-the-fly,” X-Net gives researchers the ability to not only select which datasets to use, but also to select parameters such as orthology method and thresholds for multispecies networks.

**The X-Net Platform**: There are two main network analysis functionalities we propose to create in X-Net: (1) the ability to create a species-specific multinetwork for any given species, and (2) the ability to create a multispecies weighted network.

1. **Species-specific multinetwork**: A species-specific multinetwork is simply the union of all different types of interactions. The interface for a species-specific interface would allow the researcher to choose edge types, thresholds (e.g. correlation above 0.6), and sources of data. The species-specific network might come from experimental data or from inference based on InferNET (Aim 1), or from Interolog. For example, (see Fig. 6) a biologist working on Glycine max who wants to use protein-protein information from Arabidopsis would simply:

* + 1. Choose Arabidopsis as the source
    2. Choose Glycine max as the target
    3. Choose an orthology definition and threshold
    4. Click on the “Run Interolog” button
    5. Receive a link to the created network.

2. **Multispecies weighted network:** The multispecies weighted network will allow researchers to combine networks from any number of species into one multi-species network where the edges and nodes have confidence values based on weights determined by the support from multiple species (in the style of Aim 2) (See Fig. 6). To provide this feature to the community, we will create a “***Network Cart***” in VirtualPlant ([www.virtualplant.org](http://www.virtualplant.org)) (Katari et al 2010) that allows plant biologists to store, manage, and refine the networks they create using X-Net. Because the VirtualPlant user community of biologists finds the existing “Gene Cart” feature both intuitive and powerful, we believe that they will be able to perform sophisticated queries with their “Network Carts” as well. This querying feature enables researchers to refine their network analysis and predictions over iterative rounds of data analysis.

**PLAN TO INTEGRATE RESEARCH AND EDUCATION**:

**Cross-training of Biologists and Computer Scientist in Systems Biology**. We have and will continue to implement mechanisms to bridge the gap between computer scientists and biologist. Each year, Dr. Katari (a computer scientist with a PhD in Genetics) leads the **R-boot Camp** (weekly meetings during one semester), to train the biologists in using “R” to analyze genomic data. This trains biologists at all levels in the workings of “R”. Recent “students” have included faculty on sabbatical, Mary Lou Guerinot and Rob McClung of Dartmouth. As a complement, computer scientists from Courant (and visiting computer scientists from the business world) are taught biology through a Molecular & Cell Biology Class (taught by Dr. Coruzzi) and during the weekly joint lab meetings between the Coruzzi Lab (NYU Biology) and NYU Courant (Shasha and Tranchina). This is in addition to courses taught at NYU’s Center for Genomics & Systems Biology: G23.1128 Systems Biology; G23.1130 Applied Genomics & Network Modeling; G23.1127 Bioinformatics & Genomes. Graduate students are co-advised by a Biology and Computer Science faculty. In the last year, we have trained two PhD students, two interns, and two MS students from Courant 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. Computational students will be involved in constructing the pipeline and making it perform through the use of parallelization. Such students will also help develop and optimize machine-learning algorithms for network inference.

**NYU-Stuyvesant High School Intern Intel Program:** The PI of this project serves as a faculty liaison for Intel High School students at NYU’s Center for Genomics & Systems Biology. This program, initiated by the PI, Gloria Coruzzi, involves an annual workshop at NYU’s Center for Genomics and Systems Biology (NYU-CGSB) which hosts 40+ High School Students from NYC Stuyvesant HS, a premier NYC public school specializing in math and science. Students are exposed to presentations by NYU genome faculty. As a result of this activity, this year NYU-CGSB faculty hosted *four* Intel Semi Finalists (out of 300 nation-wide) and *two Intel finalists* (out of 3 from NY and 40 finalists nation-wide). One Intel finalist, **Angela Fan** (Stuyvesant HS), performed her project in the plant genomics and systems biology laboratory of **Gloria Coruzzi**, where she studied the molecular basis for root nutrient foraging. Angela applied a morphometric approach to quantifying the developmental plasticity space of different ecotypes of the model plant species *Arabidopsis thaliana* in laboratory and natural environments. In addition, Angela was named a Siemans Semifinalist. Angela will be a freshman at Harvard in the Fall of 2012. The Coruzzi Lab will also host 2 new Intel Students from Stuyvesant (one junior and one sophomore) starting Summer 2012 who will work on this NSF Plant Genome project. As all students at Stuyvesant learn computer science beginning in sophomore year, this project is perfect training in the application of computer science to a biological problem.

**Public Outreach:** Dr. Dennis Shasha has been a consultant for the New York Hall of Science for the last several years where he helps with the design of computationally and biologically inspired exhibits involving flows, mixtures, and probability. His recent general science book *Natural Computing: DNA, Quantum Bits, and the Future of Smart Machines* discusses the strong influence of biological thinking in future of technology and vice versa. The PIs are periodically consulted by film students on treatments having to do with biological and computational themes.

**Plant to Integrate Diversity**: We are committed to training scientists at the graduate and postdoctoral levels across diversity.  Students trained as part of the parent NSF grant include Hispanic and African-American scientists. Damion Nero, an African-American PhD student, has written programs contributing to the Virtual Plant project. Roberto Jimenez (Systems Admin) associated with this project is of Hispanic origin, as is our collaborator Rodrigo Gutierrez. Unusual for a computational grant, we have numerous female scientists associated with this project: Coruzzi (PI); Rebecca Davidson (Programmer); Varuni Prabhakar (UG Programmer); Ana Arroja (MS); Ranjita Iyer (MS Courant).