**MOTIVATION AND NOVELTY.** This grant aims to enable plant biologists to infer and exploit gene networks across a wide variety of plant genomes for agronomic benefit. The *novel features* of our approach to cross-species networks include: The development of “*InferNet*” a machine-learning approach exploiting data-rich species to *infer* networks in data-poor species (Aim 1); To *learn* a “weighted” network in a data-rich model species using data from “target” (e.g. crops) species, and test candidate genes (e.g. network hubs) in a rapid “Network walking” assay, with cross-validation of promising candiates in crops (Aim 2); The development of an “X-Net” pipeline, to enable *on-the-fly* construction of gene networks - including multinetworks (e.g multiple edge-types)- for (i) *any single species* using inferred and/or real interaction data and (ii) networks *across* multiple plant species. “X-Net” will empower biologists to generate and evaluate the significance of network hubs and modules across plant species, to enhance gene discovery and translational research.

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

1. *Advance Plant Systems Biology*: Utilize the large amount of data on well-studied data-rich species

to support and infer gene networks on new and emerging (crop) species (Aim 1).

2. *Translate basic discovery to field*: Exploit crop data to derive “weighted” gene networks in data-rich

models, with pilot validation studies from models-to-crops (Aim 2).

3. *Develop coordinated solutions to data access, analysis and synthesis:* Develop and deploy the “X-Net”

software platform, to enable plant biologists to synthesize knowledge across species to identify

network hubs and modules for hypothesis derivation and testing (Aim 3).

4. *Enhance education, training and outreach*: Collaborative training in Plant Systems Biology across

 biologists and computer scientists, with Outreach to High School students and Science Museums.

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

agronomic and/or environmental value.

**BACKGROUND AND SIGNIFICANCE.**

***Success*: Enabling gene correlation/interaction networks in Arabidopsis and models**.Studies have shown that functionally related genes tend to be transcriptionally coordinated (i.e., co-expressed) [[Stuart et al., 2003](http://www.plantcell.org/content/23/3/895.full#ref-56) Science; [Persson et al., 2005](http://www.plantcell.org/content/23/3/895.full#ref-49) PNAS]. Using “guilt-by-association” approaches, such co-expression network analyses have proved valuable for rapid inference of gene function, subcellular localization, and pathway discovery [[Wei et al., 2006](http://www.plantcell.org/content/23/3/895.full#ref-64) Plant Physiol; [Yonekura-Sakakibara et al., 2008](http://www.plantcell.org/content/23/3/895.full#ref-65) Plant Cell; San Clemente et al., 2009 ; [Usadel et al., 2009](http://www.plantcell.org/content/23/3/895.full#ref-63) Plant Cell Environment; [Klie et al., 2010](http://www.plantcell.org/content/23/3/895.full#ref-28) J. Computational Biology]. Network tools developed to integrate co-expression data with other sources of gene interaction data (e.g. protein-DNA, protein-protein) have been deployed for models like Arabidopsis. For example, “AraNet” – a probabilistic network tool [Lee et al 2010-Nature Biotech]- has been successfully used to identify genes associated with traits. Likewise, the Arabidopsis Multinetwork [Gutierrez et al 2007 Genome Biology, Gutierrez et al 2008 PNAS, and Katari et al 2010] - developed by the PIs under previous NSF funding - has been used to derive and validate biological hypotheses for hubs and biomodules involved in nutrient sensing [Gutierrez et al Genome Biol 2007, Gutierrez PNAS 2008] [Nero 2009] [Gifford et al 2008].

***Challenge*: Transfering “network knowledge” from models-to-crops**. In comparing gene networks across species, several studies have explored the hypothesis that the best ortholog for a gene must have a similar “network neighborhood”, as well as orthology. Such analysis platforms include **PlaNet** [Mutwil 2011 Plant Cell], **StarNet** [Jupiter 2009 BMC Bioinformatics], **CoP** [Ogata 2010 Bioinformatics], and **ATTED-II** [Obayashi 2011 Plant and Cell Physiology]. All of these platforms build co-expression networks first within each species- and then compare them either pairwise (Starnet, CoP, ATTEDII) or amongst multiple species (PlaNet). Despite their success, one of the limitations of all of these approaches is that they *assume* the existence of enough data in the target (e.g. crop) species, to construct reliable co-expression (and potentially other) networks under conditions of interest.

***Solution*: “X-Net: A machine learning approach to plant genome networks”.**  With the advent of Next-Gen technology, plant genome exploration is not limited by DNA or RNA analysis. However, because most of the **newly sequenced species** will be “*data poor*” (compared to the models), we propose a ***novel approach*** that takes advantage of “*data-rich”* species to *learn* and *infer networks* in data-poor species, to augment experimental evidence. This approach is inspired by the Robin Hood philosophy of “learning from the rich and giving to the poor”. Unlike the existing plant network methods above, the *InferNet* method aims to *learn* and *infer* networks for any species of interest. A complementary **novel feature** of our cross-species network learning approach, is to exploit trait-associated expression data from crop species, to learn “weighted” networks to enhance model-to-crop predictions. Finally, our “**X-Net”** pipeline which embodies these approaches in a user-friendly format will enable plant biologists to construct networks and multinetworks (i.e. networks with multiple edge types) “*on the fly*” for any species and to create weighted networks between species. This interplay among species networks will enhance transfer of knowledge between diverse species, models and crops.

-------------------------------------------------------------------------------------------------------------------------------

**RESULTS FROM PRIOR NSF SUPPORT:** One outcome of our proposal will be a computational platform to create networks for translational gene discovery, so we briefly describe our success in a completed NSF Grant DBI-0445666, “Conceptual Data Integration for the Virtual Plant.” The VirtualPlant software platform (www.virtualplant.org) integrates genome-wide data concerning the known and predicted relationships among genes, proteins, and molecules, as well as genome-scale experimental measurements [Katari 2010]. VirtualPlant also provides tools that render multivariate information into integrated visual displays to highlight biological implications within a single species (e.g. Arabidopsis or Rice). 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***: 1. Integration, 2 & 3. Visualization & Synthesis, and 4. Prediction. Our accomplishments in each are highlighted below.

**Aim 1.** **Integration**: ***The Arabidopsis Multinetwork****:* ***A systems biology tool for hypothesis generation*.** The VirtualPlant project included assembling the first Arabidopsis Multinetwork, a first step towards a molecular wiring diagram of the plant cell [Katari 2010 Plant Physiol] [Gutierrez 2007 Genome Biol]. This multinetwork which has 16,562 nodes and 97,423 interactions, enables researchers to interpret transcriptome data in the context of all known sources of interaction including protein, DNA, RNA, etc. [Katari et al 2010] (Fig. 1B). In one example, a query against the multinetwork with 834 nitrogen-regulated genes, resulted in a list of network TF “hubs” (with 47 predicted regulatory connections to targets in the N-regulatory network) including the central clock control gene CCA1 (see Fig. 1B) [Gutierrez 2008 PNAS]. This subnetwork enabled us to derive and validate the novel hypothesis that N-regulation resets the circadian clock. Other examples of networks derived and validated using this multinetwork are reported in [Gifford 2008; Gutierrez 2007 Genome Biol; Nero 2009; Thum 2008]. A complementary network tool is GeneMania [Wade-Farley 2010] [Moreno-Risueno 2009] which generates a hypothesis for gene function based on interactions with other genes and their attributes. Another effort, AraNet reports a genome-scale functional network for Arabidopsis – which combines data from multiple sources about gene and protein interactions, resulting in a far larger network [Lee 2010 Nature Biotech “Rational Association of Genes….].

 **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. Below are a few (of many) 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 (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], including statistical significance.

 **Aim 4. *Predictions: Extensions into time and species*.** We have accomplished 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 accomplished goal was to extend VirtualPlant to other single species datasets such as Rice (see www.virtualplant.org).

**VirtualPlant Database & User Community**: **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]. The **VirtualPlant User Community** consists of >700 *registered* academic and commercial users from 36 countries, as well as unregistered users. 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: VirtualPlant: Tool development for Plant Systems Biology**

Katari MS, …., 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 & analysis of genomic datasets”. ***Bioinf***, 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**

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. ***PNAS.*** 107(9):4477-82

Krouk G, .. Shasha D, Coruzzi G & Gutierrez R (2009) “Systems approach uncovers restrictions for signal interactions regulating genome-wide responses .” ***PloS Comp Biol***. Mar;5(3):e1000326.

Gutierrez R, ..., Nero D, McClung R and Coruzzi G (2008) "Systems approach identifies an organic nitrogen-responsive gene network regulated by the master clock control gene CCA1" ***PNAS*** 105, 4939-4944. *(Faculty of 1000: Factor 3)*

Gutierrez R, …Shasha D, Coruzzi G, Crawford N (2007) "Insights into the genomic nitrate response using genetics and the Sungear Software System" ***J Exp Bot*** doi: 10.1093/jxb/erm079

Gutierrez R, …, Shasha D, Coruzzi G (2007) "Qualitative network models & genome-wide expression data define C/N-responsive biomodules " ***Genome Bio***l 8: R7. *Faculty 1000 (Must Read: Fact 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 Bioinf***, 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 Spec Data

 Mining,” ***IEEE Transactions*** *on Knowledge & Data Engineering*. Issue 99. November 29, 2010,

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

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

Coruzzi GM, Burga A, Katari MS, & Gutierrez RA (2009) “Systems Biology: Principles & Applications in Plant Research”. “Plant Systems Biology”, ***Annual Plant Reviews***; Blackwell, , UK, 2009, Vol. 35. Pgs 3-31.*.*

Gifford M, Gutierrez R, and Coruzzi G (2006) "Modeling the Virtual Plant ". Essay 12.2 Chapter 12.; In ***A Companion to Plant Physiology,*** , http://4e.plantphys.net/article.php?ch=12&id=352

**Education & Training**: Undergraduates (UG), master’s (MS) and PhD students have learned Systems Biology. **Undergraduates**: Steve Nowicki (NYU, CAS), Varuni Prabhakar (Barnard College), Rebecca Davidson (BS, Computer Science); **Masters Students (Computer Science)**: Ana F. Arroja , Ranjita Iyer, Jonathan Kelfer, Jesse Lingeman, Lee Parnell, Jarod Wang,; **PhD Students (NYU Courant)**: Chris Poultney, Aris Tsirigos, Saurabh Kumar; Damion Nero (NYU Biology). These students have gone on to PhD programs (Prabhakar and Parnell), post-docs (Poultney and Tsirigos) and to industry (Kelfer, Wang -Medidata Solutions; Damion Nero, Statistician Programmer, FOJP Service Corp). **High School Students**: Angela Fan (Stuyvesant HS) – Siemans Semi-Finalist 2011, Intel Finalist 2012; Jenny Kim (Chapin HS).

-------------------------------------------------------------------------------------------------------------------------------

**RESEARCH PLAN**

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

***Rationale and Novelty***: Existing tools for comparing plant gene networks are based on existing experimental data for each species and then compare them post-hoc (e.g. CoP [Ogata 2010], StarNet [Jupiter 2009], ATTED-II [Obayashi 2011, and PlaNet [Mutwil 2011]). By contrast, ***InferNET*** will ***infer*** regulatory networks in a data-poor target species, based on gene networks *learned* from several data-rich species. 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. The discussion below concerns co-expression networks, because Next-Gen sequencing makes such data readily obtainable for any species. However, the methods we develop can apply to other types of network edges, including protein interactome, as the data becomes available.

 ***Community Need:*** As the number of available plant genome sequences rapidly increases due to advances in Next-Gen technology, it will be common to find a newly sequenced or poorly studied target species “*t*” that is phylogenomically similar to those few “data-rich” species having a substantial body of experiments (see Fig. 2). For new targets such as emerging crops and “boutique” crop species, inferring networks will be particularly valuable.

**Implementation and Validation Testing of *InferNET*:**

**Species Phylogeny and Expression Data:** To mine expression data within a phylogenetic framework, we constructed a parsimony-based phylogenomic tree for 21 fully sequenced species using Ortholog ID [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] (Fig. 2). This simultaneous analysis (SA) matrix with 21,271 partitions (genes) has at least 5 taxa are present in each gene partition in this matrix (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 [Goloboff, PA, Farris, JS, and Nixon, KC, *TNT, a free program for phylogenetic analysis.* Cladistics, 2008. **24**(5): p. 774-786.5] For a detailed description of our method for constructing phylogenomic trees see [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]. The phylogenetic distance between species can be calculated according to patristic distance, based on maximum parsimony [Fourment and Gibbs /BMC Evolutionary Biology/ 2006, \*6\*:1] Available expression data for each species is represented as a pie, whose overall size indicates the relative amount of expression data, while pie sections represent data sources: Microarray Data (Blue) and Next-Gen RNA-Seq Data (Red) (Fig. 2). We will add additional species to the *InferNET* analysis pipeline, as their sequences and expression data become available. This will include species with fully-sequenced and annotated genomes, as well as species with fragmented “gene space” assemblies that 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, the “data rich” species are Arabidopsis, Poplar, Medicago, Soybean, Rice, and Maize (see Fig. 2). Such data-rich species will be used to train the *InferNet* model in a source species (e.g. Arabidopsis), to predict models in data-poor species (see Fig. 3)

 **Defining 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*.

 **Learning the InferNet 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 accurately predicts co-expression edges on one of those data-rich species. We then use that model to *predict* edges in the data-poor target species *t* (see Figs. 3 & 4).

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

**(1) 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(s) will work best.

 **(2) 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 separately, while testing their interchangeablity under the proper normalization protocols [Bullard et al 2010].

 **(3)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 Approaches**: 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. To achieve high performance and good 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*: “Hide-the-Answer”.*** 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. This analysis gives us both precision and recall data to quantify the success of our prediction methods (see Table I).

**A 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. 4). 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 phylogenetically. 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 for networks in Soy, using actual expression data from Soy (see results in Table I). 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 I 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 InferNet 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).

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). The learned model also suggests why the Interolog approach [Yu 2004] works as well as it does. Whereas the MOv value (percent similarity) 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 less important for reciprocal BLAST hits. Often, the correlation of the edge in the source species by itself predicts the correlation in the target.

**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. For example, 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 (e.g. based on patristic distance based on maximum parsimony [Fourment and Gibbs /BMC Evolutionary Biology/ 2006, \*6\*:1]). The weights on the conclusions from each species will be learned using one of the three machine learning methods above.

**Expected outcomes and future directions:** Our goal in Aim 1 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 suggests both reason for optimism and room for improvement. We will test results using other methods for correlation and orthology, and also apply the same techniques to other edge types (e.g. protein-protein interaction), as the data become available for the source species Arabidopsis [Arabidopsis Interactome Mapping Consortium (2011) Science 29 July 2011: Vol. 333 no. 6042 pp. 601-607 **Evidence for Network Evolution in an Arabidopsis Interactome Map]**, and Rice [ 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, Song WY. 2009. A rice kinase-protein interaction map. Plant Physiol. 149(3):1478-92. ] [Rohila JS, Chen M, Chen S, Chen J, Cerny R, Dardick C, Canlas P, Xu X, Gribskov M, Kanrar S, Zhu J-K, Ronald P and Fromm ME. 2006. Protein-protein interactions of tandem affinity purification-tagged protein kinases in rice. The Plant Journal. 46, 1-13.] [Rohila JS, Chen M, Chen S, Chen J, Cerny R, Dardick C, Canlas P, Fuji H, Gribskov M, Kanrar S, Knoflicek L. Stevenson B, Xie M, Xu X,Zheng X, Zhu J-K, Ronald P and Fromm ME. 2008. Protein-Protein Interactions of TAP-Tagged Protein Kinases in Rice. (2012) Under Revision for Molecular and Cellular Biology. ]

**Aim 2: A trait-to-gene “weighted” network discovery pipeline: Learning (2A) and Validation (2B) *Rationale***: We propose to develop methods to build “weighted” gene networks from co-expression data on crop species, to identify genes that are potentially central to a particular trait of interest. We will then test selected candidate regulatory genes (e.g. TF network hubs) initially in a rapid transient assay system (e.g. “*Network Walking*”), to validate predicted network targets. Based on transient analysis, selected genes will be subject to over-expression, knock-outs, or knock-ins in Arabidopsis. Candidates with phenotypes in Arabidopsis will be tested in Maize in transient cell-based assays (Coruzzi, NYU), *in planta* (Martienssen, CSHL), and in diverse field-conditions (Moose, University of Illinois) to validate translatability of our “weighted” network approach (see Fig. 5 for overall design) ***Novelty***: Our trait-to-gene “weighted” network learning approach, follows the spirit of AraNet [Lee 2010] and PlaNet [Mutwill 2011], in that multiple species are used to identify functionality in gene networks. The novelty in our approach is: (i) we use expression data from trait-relevant experiments on 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, (iii) our experimental validation strategy is based on a medium-throughput testing of predicted networks in an inducible expression system called “***network walking***”, and (iv) a follow-up validation of selected genes *in planta* first in Arabidopsis, and later in Maize, to assess translatability of the network prediction and validation methods from model to crop. ***Significance*: Agronomic traits and phylogenomic context**. Since the dawn of agriculture, farmers and scientists have improved crops by selection and breeding. Among the 21 sequenced plant species in the phylogenomic tree of Fig. 2, the crop species lie at phylogenetic distances of a hundred million years or more from a common ancestor [Chaw et.al. 2004]. Nevertheless, specific gene functions are conserved 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., The Plant Cell Nov 2004, To et al., The Plant Cell July 2006]. Arabidopsis 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, however, 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]. Additionally, involvement of specific genes in agronomic traits is perhaps better investigated in those species. Hence, expression atlases [Severin et al. BMC Plant Biology2010] and numerous individual expression assays exist for each of several crop species, which we will mine in this approach. ***Gene Discovery***: Our combined computational-experimental approach mines transcriptomic data from crop species to inform the identification of gene network modules associated with traits of agronomic interest. This will lead to novel gene discoveries that can be first tested in Arabidopsis and then validated in crops. In some cases, this comparative network analysis will identify Arabidopsis orthologs of crop genes, whose significance to the trait would not be evident solely from Arabidopsis data. In other cases, it will identify nodes of networks that are “missing” in Arabidopsis, but present in the crops. These crop genes could be “knocked-in” to Arabidopsis, first using the protoplast transient assay system we describe in Aim 2B. Promising candidates will be validated *in planta* first in Arabidopsis, and later in maize, as proof-of-principle. Aim 2A is the computational method:

**Aim 2A: A trait-to-gene “weighted” network discovery pipeline**

**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. gene regulation during seed development). Collect a second set of expression data under unrelated conditions. Transcriptomic data from Next-Gen sequencing or microarrays can be used as available (see Fig. 2). Genes that do not show variation, as determined by a minimum variance cut-off across all experiments are designated “housekeeping” genes, and are removed from further analysis. For Next-Gen data, further filters are applied to remove poor quality data (e.g. low counts). After filtering, raw counts are normalized using a full-quantile 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., Plant, cell and environment, 2009], cosine coefficient (CoP) [Ogata et. al., Bioinformatics 2010], Mutual rank [ATTEDII, PlaNet] [Obayashi et.al., NAR 2008; Mutwil et.al., Plant cell, 2011] and other methods discussed in Aim 1. A gene correlation network (CNi) for species Si consists of edges {g1, g2} such that the absolute value of the correlation between these two genes is at least 0.7 with p-value <=0.05

**Step 3. Consensus through “weighting” of nodes and edges:** First form correlation networks in each species of interest (e.g. each crop). A gene g in such a network that passes an orthology cutoff (e.g. stringent BLAST e-value cutoff) with respect to some Arabidopsis gene g’, will give g’ a vote which we will call a “weight” from now on (because we will eventually give different “weights” to different species depending on phylogenomic distance and numbers of paralogs). Thus g’ may receive “weights” from several genes in each network and from multiple species networks as in Fig 5. If g does not pass the orthology cutoff with respect 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 [Gutierrez et al, 2008, Nero et al 2009] are added to provide confidence and direction to the regulation network. The final network including nodes, edges and 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 clustering using 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: Experimental studies in Arabidopsis**: Identify genes central to each network module for testing in vivo (e.g. transient “Network walking” in protoplasts and *in planta*) (see Aim 2B). Prioritize genes that are poorly characterized in (or even absent from) Arabidopsis.

**Step 6: Experimental studies in Maize:** Promising candidate genes from Step 5, will be tested in Maize in transient assays and *in planta* (see below). Maize mutants will be tested in greenhouse and field conditions, to ascertain the effect of the gene on the trait of interest.

**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 “weighted network” Analysis:** To test our “weighted network” approach for trait-to-network node discovery, we use “seed development” as the trait of interest [Baud and Lepiniec, Progress in lipid research, 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., Trends in plant sciences 2008]. In addition, discoveries related to seed networks may have obvious economic value.

 **Construction of gene correlation network (CN)**: We have used deep transcriptome data sets from early seed tissue samples of Soybean and Maize to perform this preliminary analysis [NCBI GEO]. Following the specifications of the first two steps above, we found CNsoybean and CNmaize. Then, we assigned orthologs to Arabidopsis, Soybean and Maize. 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.

KRANTHI NEEDS TO COMPLETE THIS SECTION!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

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: Transient “*Network Walking*” and *In Planta***. Aim 2A finds “weighted” networks to identify candidate genes for functional studies in Arabidopsis, which ultimately will aid in translational studies back to crop (see Fig. 5). As an *in silico* proof-of-principle, we tested seed development as a trait, for which there is ample mutant data with which to validate the genes uncovered in our networks. In the course of this grant, we will expand this approach to identify genes associated with nitrogen-use trait, using N-responsive transcriptome data from Maize (S. Moose, unpublished) to inform weighted networks in the model (Arabidopsis). To validate our TF🡪target predictions based on the “weighted” networks from Aim 2A, we will use a medium-throughput dexamethasone inducible transient assay system, in which transcription factors and their targets can be assayed *in vivo* [Sablowski and Meyerowitz Cell 1998](see details below). For TFs that pass initial validation in the transient system, we will proceed to stable transformants (e.g. T-DNA, overexpression, or “knock in” for cases where the gene is missing in Arabidopsis), and perform tests for phenotypic effects in Maize for selected candidate genes.

**“Network Walking”: A rapid approach to validating network predictions.** In “*Network Walking*”, TFs are transiently expressed in FACS sorted protoplasts, and activation of predicted target genes is validated by RNA analysis (Q-PCR and/or transcriptome). This approach identifies transcription factor targets in less than a week of experimentation, following methods developed by Bargmann and Birnbaum [Bargmann BO, Birnbaum KD (2009) Positive fluorescent selection permits precise, rapid, and in-depth overexpression analysis in plant protoplasts. *Plant Physiol* **149:** 1231-1239.][Bargmann BO, Birnbaum KD (2010) Fluorescence activated cell sorting of plant protoplasts. *J Vis Exp*.] Using Gateway™ technology, we have engineered a vector with a GFP marker, for which any TF can be fused with a GR (the glucocorticoid receptor) tag, and successful transformants are isolated by FACS cell-sorting. This 35S-TF-GR chimera allows one to i) overproduce the studied TF in the protoplasts, and to ii) control the TF entrance into the nucleus using a dexamethasone (DEX) treatment [Lloyd *et al*, 1994 Lloyd AM, Schena M, Walbot V, Davis RW (1994) Epidermal cell fate determination in Arabidopsis: patterns defined by a steroid-inducible regulator. *Science* **266:** 436-439.][Sablowski and Meyerowitz, 1998 Sablowski RW, Meyerowitz EM (1998) A homolog of NO APICAL MERISTEM is an immediate target of the floral homeotic genes APETALA3/PISTILLATA. *Cell* **92:** 93-103.] [Bargmann BO, Birnbaum KD (2009) Positive fluorescent selection permits precise, rapid, and in-depth overexpression analysis in plant protoplasts. *Plant Physiol* **149:** 1231-1239.][Bargmann BO, Birnbaum KD (2010) Fluorescence activated cell sorting of plant protoplasts. *J Vis Exp*.]. We have successfully validated this “Network-walking” method to identify network targets of the well-studied TF, ABI3 [Bargmann et al 2012, In Preparation] in Arabidopsis.As part of this grant, we will adapt the “Network walking” approach to Maize protoplasts. If successful, this will enable rapid cross-validation of our network predictions between Arabidopsis and Maize. This transient assay system should be readily adaptable to Maize protoplasts, based on studies from the Sheen lab (MGH) in which both Arabidopsis and Maize protoplasts are used in transient expression of signal transduction components (see [Sheen (2001) “Signal Transduction in Maize and Arabidopsis mesophyll protoplasts. Plant Physiol. Vol 127; 1466-1475].

**Prioritization of Genes and *In Planta* Studies**: Following transient expression studies in protoplasts, sentinel genes predicted to be targets of the TF (based on the “weighted network”) will first be assayed by Q-PCR for validation. A transcription factor that significantly changes the expression level of one or more target 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, and knock-ins). We will first test transgenics/mutants for (i) molecular phenotypes (e.g. changes in predicted target genes of the TF). Ones that show an altered molecular response will be tested for (ii) physical phenotypes (e.g. seed development). Genes validated to affect a trait of interest in Arabidopsis, will be translated back to one or more orthologs in Maize. As outlined above, we will also attempt to adapt the transient “network walking” assay system to Maize protoplast, to enhance our translation testing of candidate network hubs. Two maize mutants for each of the ten most promising genes TF hubs affecting networks associated with seed development in Arabidopsis will be identified from the maize transposon library (in collaboration with Rob Martienssen, CSHL) [May BP, Liu H, Vollbrecht E, Senior L, Rabinowicz PD, Roh D, Pan X, Stein L, Freeling M, Alexander D, Martienssen R. Proc Natl Acad Sci U S A. 2003 Sep 30;100(20):11541-6]. Each Maize mutant will be assayed for changes in seed development by observing ear phenotypes and/or seed developmental defects (CSHL). Where needed, double mutants for candidate genes will be generated by crossing existing mutant lines.

***In the N-use “trait” study***, we will test TF hubs associated with “weighted” networks formed using N-treatment transcriptome data from Maize (in collaboration with Stephen Moose, U. Illinois) and Arabidopsis (Coruzzi, NYU). We know from a preliminary comparison that Maize and Arabidopsis share N-regulation of key target genes in the N-assimilation pathway; N-induction of nitrate and nitrite reductase, N-repression of L-asparaginase, dark-induction of asparagine synthetase 1. Thus, the “weighted networks” will enable us to predict master regulatory hubs involved in control of the N-regulatory network, as we have done previously in Arabidopsis [e.g. see Gutierrez et al 2008 PNAS] [Nero D, Krouk G, Tranchina D, Coruzzi GM (2009) “[A system biology approach highlights a hormonal enhancer effect on regulation of genes in a nitrate responsive "biomodule".](http://www.ncbi.nlm.nih.gov/pubmed/19500399?itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum&ordinalpos=4) ***BMC Syst Biol***., 3:59.], but now with relevance to Maize. We will attempt cross validation of TFs first in the protoplast system using Arabidopsis and Maize protoplasts. For TFs confirmed in the protoplast system, we will test selected ones in Arabidopsis to prioritize Maize testing. Maize mutants will be grown in field conditions under diverse nitrogen regimes shown to affect expression of N-assimilation genes, to assay phenotypic changes in N-use efficiency [Seebauer, J., **Moose, S.P.**, Fabbri, B., Crossland, L. and Below, F.E. (2004)  Amino acid metabolism in young maize earshoots: implications for assimilate movement and nitrogen signaling.  *Plant Physiol.* 136: 4326-4334]. For analysis of maize mutants, confirmed homozygous mutant plants and wild-type sibling controls will be grown under N-limiting conditions in either a hydroponic system and/or in nitrogen-responsive field plots (U. Illinois).  In addition to any obvious visible phenotypes associated with N-deficiency such as yellowing of leaves (quantified by SPAD chlorophyll meter), faster leaf senescence, or reduced growth, we will also measure biomass accumulation and free amino acid profiles of leaves and developing earshoots.  Corn is a hybrid crop and so agronomic N-utilization is typically defined as grain yield per unit of N supply, but grain yield is also influenced by heterosis and rate of shoot maturation (typically observed as flowering time but includes how fast seeds reach physiological maturity).  We have found that measuring ear-shoot amino acids is a more robust indicator of genetic variation in N-utilization that is largely independent of these other confounding factors.  It would also be possible to assay expression of marker genes for N-metabolism, particularly if a change in amino acid profiles is observed.

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

***Rationale*:** X-Net Builder will be 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 enable experimental biologists to identify 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 are updated periodically. By comparison, **X-Net** will allow biologists to (i) create predicted networks for data-poor species, (ii) create networks based on subsets of experiments, and (iii) 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 biologists to enable Systems Biology. Each year, Dr. Katari (a computer scientist with a PhD in Genetics) leads the “**R-Boot Camp”** (weekly sessions held 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, G. Krouk (CNRS, Montpellier), and others. As a complement, computer scientists from Courant (and visiting computer scientists from the business world (e.g. John Sabini) 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 (Dennis Shasha and Dan Tranchina). This is in addition to courses taught at NYU’s Center for Genomics & Systems Biology (e.g. 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 the Education and Training section in Results from Prior support. Computational students will help develop the learning pipeline and making it perform through the use of parallelization.

 **High School Intern Program in Systems Biology:** The PI of this project serves as a faculty liaison for 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. As a result of this activity, this year NYU- faculty from the Genome Center 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), who performed her project in the plant systems biology laboratory of **Gloria Coruzzi**, was a Siemans Semifinalist and Intel Finalist. We will host 2 new Intel Students from Stuyvesant and a returning student from Chapin HS in 2012. As Stuyvesant students 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 scientific advisor 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 from NYU Tisch School 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 include an African-American Damion Nero, who has written programs contributing to the Virtual Plant project and Roberto Jimenez (Systems Admin) of Hispanic origin. Unusual for a computational grant, in addition to the PI, we have 4 female computer scientists associated with this project: Rebecca Davidson (Programmer); Varuni Prabhakar (UG Programmer); Ana Arroja (MS); Ranjita Iyer (MS Courant).