**RESULTS FROM PRIOR NSF SUPPORT** This proposal leverages on the accomplishments of the previous parent NSF ABI grant, “Conceptual Data Integration for the Virtual Plant” (DBI-0445666) to now develop new network inference tools to make predictions for network states within a species (e.g. over time) and across species (Cross-species network inference).

Under the previous cycle of funding, we developed a software platform called VirtualPlant a plaform that deploys new informatics, visualization, math and statistic tools to enable dynamic modeling and visualization of molecular networks operating within cells. We integrated these tools in a common user-friendly platform ([www.virtualplant.org](http://www.virtualplant.org)) that has enabled researchers (>500 current users) - including biologists with no computer training- to understand how internal and external perturbations affect processes, pathways & gene regulatory networks controlling plant growth and development. The goal is to (1) explain mechanistically how molecular network changes evoke responses and (2) to predict molecular network states under untested conditions or after modifications (e.g. gene mutations). Predictive models are useful for developing strategies that intervene in molecular networks for biotechnological, therapeutic or diagnostic purposes.

We accomplished the four major goals set out in our original VP proposal: **(1) *Integration***: to integrate known relationships among genes, proteins and molecules (extracted from public databases, the published literature and generated with predictive algorithms) with experimental measurements (e.g. mRNA measurements). **(2) *Visualization***: to develop novel visualization techniques that render the multivariate information in visual formats that facilitate extraction of biological concepts. **(3) *Synthesis***: to integrate mathematical and statistical methods with the novel visualization techniques to help summarize the data and derive hypothesis for testing. **(4) *Predictions***: to use machine-learning methods to make predictions of the molecular network state under untested conditions, a goal of systems biology. Below are the details of our accomplishments in these four areas.

**DETAILED PROGRESS REPORT:**

**Virtual Plant: A Software Platform for Data Integration, Analysis and Visualization.** The VirtualPlant software platform (**www.virtualplant.org**) [1] 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 visual displays (e.g. networks) to highlight biological implications. Importantly, VirtualPlant's software architecture and data model have been designed and created in a generic, species-independent manner to facilitate the addition of new organisms and tools. We have demonstrated the use of tools embodied in the VirtualPlant system to generate hypotheses that were subsequently experimentally validated [7, 9-13].

**1.** **Integration**: *The Arabidopsis Multinetwork: A systems biology tool for hypothesis generation***.** Our VirtualPlant work included assembling the first multinetwork for Arabidopsis, a first step towards a molecular wiring diagram of the plant cell [1, 5]. For a very recent review of this and other plant regulatory networks see [6]. (Manny- We need to say something about why our approach to the multinetwork is more suited to our goals than other available tools). It would be particulary good if we could say we are complementary to other efforts. The Arabidopsis multinetwork in VirtualPlant has 16,562 nodes (of which 13,960 are genes) and 97,423 interactions (Fig. 1B, Table I). The multinetwork enables researchers to interpret transcriptome data in the context of all known sources of interaction including protein, DNA, RNA, etc. In one example, a query against the Arabidopsis multinetwork with 834 N-regulated genes resulted in a sub-network of 369 genes connected by one (or more) “edges” [7]. In that example, predictions of TF🡪Target connections were based on significant correlation or anti-correlation (>0.7 or <-0.7 with p-value < 0.01) and statistical over-representation of cis-regulatory elements (CRE) compared to the entire genome [7, 8]. The resulting network analysis identified potential “master” regulators of this N-responsive sub-network. At the top of the list of network TF “hubs” (with 47 connections to targets in the N-regulatory network) is the central clock control gene CCA1, a Myb family transcription factor (TF) [7]. Exploration of the network “neighborhood” surrounding this CCA1 TF hub revealed connections to target genes in N-assimilation (Fig. 1C). Using Arabidopsis lines that over-express 35S::CCA1 and by Chromatin-IP [7], we showed, using phase response curves, that distinct N-metabolites can advance or delay the circadian phase of CCA1 expression. Thus, we derived and validated the novel hypothesis that N-regulation of CCA1 mRNA expression sets the circadian clock. This is one example showing how the Arabidopsis multinetwork and associated software tools in VirtualPlant enabled Systems Biology approaches to derive and test new biological hypotheses. Other examples of networks derived and validated using the multinetwork are reported in [7, 9, 10, 13].

**2 & 3. Synthesis and Visualization: VirtualPlant’s primary analysis tools and functions:** Below is a list of some of the current Systems Biology tools deployed through VP.

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

**Sungear**: Sungear is a visually interactive and biologist-driven exploration of standard questions on many experiments on a genomic scale. Sungear can represent an arbitrary number of experiments/lists, all of their disjoint intersections, and their related ontological terms. The position of a circle and arrows emanating from it indicate the input lists of which it is a subset. The size of a circle is proportional to the number of genes in the intersection of those lists (see [17]). Many biologists find Sungear to be an extremely powerful and interactive tool for analyzing the interrelationships between sets of genes [57].

**Gene Networks**: Gene Network analysis allows users to query our Gene Network data housed in the multinetwork and displays the results in a graph using Cytoscape, an open source project for which we have built a plug-in. The tool allows users to filter interactions before displaying the graph [12].

**Supernode Networks**: The Supernode Network helps summarize the results of a Gene Network analysis. Individual genes in the gene network are grouped (Supernode) on the basis of their functional annotations and they are associated with other Supernodes with edges determined from the Gene Network data. A Supernode’s size is determined by the number of genes that it contains.

**NetMatch:** NetMatch, a cytoscape plug-in, finds all instances of a query graph (e.g. a network motif) in a larger graph [18]. New versions compute the statistical significance of the motifs found.

Up and coming tools include GeneSect whose purpose it is to take a set of collections of genes and to determine whether any pairwise intersections among those collections are either surprisingly large (against a variety of backgrounds) or surprisingly small. Another tool is a powerful cluster architecture to run some expensive tasks such as correlation and network inference in parallel.

**4. Predictions:** Following an initial three years of funding of the NSF VirtualPlant Grant (DBI-0445666), we received a two-year “creativity extension” (2008-2010) to accomplish two additional goals. The first was to develop dynamic network modeling tools for Arabidopsis. We have approached the dynamic network modeling by applying a machine learning method called “State Space” analysis to time-series data in Arabidopsis learn regulatory networks [19, 56]. This approach is more fully described in the Research Plan (Aim 2) as it relates directly to the new Cross Species Network Inference (CSNI) approach, we will develop in this new NSF ABI proposal. Our second goal was to extend VP to other species and we therefore also developed a VirtualPlant version of Rice (Fig. 2).

**Virtual Plant User Features and community:**

**User-specific “GeneCart” Function**: A key challenge to analyzing genomic data is the complex analysis workflow required by currently available software. VirtualPlant integrates multiple tools into a single platform that standardizes the representation of their inputs and outputs so that the output of almost any analysis can be stored in a user’s “GeneCart” and later input to any VirtualPlant analysis tool. This unique feature facilitates iterative cycles of data analysis and experimentation which are key to a Systems Biology approach [15, 16]. Three working examples described in [1] illustrate the use of VirtualPlant to perform iterative data analyses that build and refine testable biological hypotheses.

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

**VirtualPlant DB**: The VirtualPlant database contains some of the most commonly used data types including metabolic pathways from KEGG and ARACYC, protein-protein interactions from BIND and Interolog databases, and GeneOntology and Gene annotations from TAIR (see Table I for a complete listing of data sources). The database also contains processed data obtained by analyzing publicly available Microarray experiments obtained from NASC [14].

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

**Plant Genome Application XXXXXX: TRMS “Cross species network inference: From models to crops” (DATE HERE):** The goals of our current application project two new areas of intense interest and need to develop tools and pipelines to enable the plant genome community to deploy new approaches to data generation and analysis within and across species: Network dynamics (e.g. within a species over time) and network evolution via cross species network inference (CSNI). These two aspects were aims of a proposal previously submitted to NSF Plant Genome, ranked highly meritorious. Below are excerpts of reviewer comments related to a need for a cross species network inference tool. *All six reviewers noted that the cross species network inference tool is a timely and would be of benefit to the entire plant community.*

**Overall Panel Review**: “The effort to make network inference applicable across plant species is important and timely. There was no doubt the proposed methods would be effective. There is excellent potential for tools from this project to be widely applied.”

**Review 1**: “This project proposes to leverage the VP platform to create a pipeline of tools for cross species network inference in plants. This is a highly relevant effort that will benefit many ongoing hypothesis driven projects that lack the tools or capability to include network analysis. The large effort in implementation is well justified as this will be a major resource and wide usability will depend on stability, power and ease of use. I think there will be a lot of “bang for the buck” including novel scientific insights. This is true interdisciplinary integration of biology and computer science with research and tools that can broadly implact plant genome research. Tool development efforts are well integrated in cyberinfrastucture, including iPlant and Galaxy”.

**Review 2**: “The CSNI tool would likely be used by the wider plant biology community.”

**Review 3**: “With the emerging genome sequences and functional genomics datasets now available for other plant species, the time has now come to apply the gene network construction and analysis functions within the VP to crop plants.”

**Review 4**: “A resource will be created for the entire scientific community (the cross species network inference pipeline) which will be freely available on the web. This work will…develop a tools that will advance research in many areas of plant biology.”

**Review 5**: The proposed science is of high quality and internationally competitive. The application area is of the highest importance.

**Review 6**: Shasha et al propose to develop, validate and deploy an analysis pipeline for comparative inference of gene function and interaction based on similarities in NT sequence, regulatory regions and transcription patterns. Such a tool is sorely needed with the growing number of genome and trancriptome sequences coming available for the emerging model and non-model species. In summary, the ultimate objective of the proposes research to develop a user friendly platform for Cross Species network inference is extremely important. Advancement of this objective could evolve syngergistically with the work of the iPlant genotype-to-phenotype (pPG2P) and the plant tree of life (iPToL) grand challenge initiatives. As such the proposed development of a web based Cross species network inference database and analysis tool would be a major contribution.

There were some criticisms as well: one pointed out that certain network edges should enjoy more confidence than others. The reviewer suggests that we reflect confidence in weights. Our time series machine learning approach will do that. Another reviewer pointed out that using correlation across all experiments may work less well than choosing experiments carefully depending on the genes of interest. Yet another criticism suggested that our techniques for obtaining orthology should be compared with those of InParanoid and OrthoMCL. Finally there was the criticism that we don’t take sufficient advantage of the semantic web. We will address these criticisms as they come up in the rest of this proposal.

**PUBLICATIONS: Peer reviewed journal articles, chapters and books:**

**VirtualPlant: Tool development for Plant Systems Biology**

Katari M, Nowicki S, Aceituno F, Nero D, Kelfer J, Thompson L, Cabello J, Davidson R, Goldberg A, Shasha D, Coruzzi G, Gutierrez R (2009) “VirtualPlant: A software platform to support Systems Biology research”. **Plant Physiol**. Dec 9 *(Epub ahead of print) (SUZAN UPDATE THIS).*

Nero D, Kelfer J, Katari M, 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 M, Gifford M, Paley W, Coruzzi G and Shasha D (2007) “Sungear: Interactive visualization, exploration & functional analysis of genomic datasets”. **Bioinformatics**, 23:259-61.

Ferro A, Giugno R, Pigola G, Pulvirenti A, Skripin D, Bader G, Shasha D, “NetMatch: a Cytoscape Plugin for Searching Biological Networks” **Bioinformatics**, 2007 23(7):910-912.

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

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

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

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

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

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

Ruffel S, Krouk G, Coruzzi G (2009). "A Systems View of Responses to Nutritional Cues in Arabidopsis: Towards a Paradigm Shift for Predictive Network Modeling”. **Plant Physiol**. Nov 25 *(epub ahead of print)*

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

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

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

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

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