**NEIGHBORLY NETWORK INFERENCE**

**PROJECT DESCRIPTION *(Dennis has changed this)***

**RELEVANCE AND JUSTIFICATION TO THE STATED GOALS OF THE ABI/INNOVATIVE**

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

Our vision is to construct species-specific networks of ever-better quality on sets of related species. For each species s, we will use the experiments on s, but also the experiments on phylogenetically neighboring species s1, s2, ... The approaches we will use to infer network edges include intra-species techniques (cis-element analysis, biclustering, time-series analysis, knockout analysis (where feasible)) and inter-species techniques (orthology of genes and, when available, cis-binding sites). In the complete vision, every experiment on species s will add edges (or increase the confidence in intergenic-edges) to s as well as to neighboring species. Moreover, the machine-learning model that gives rise to these predictions will also give rise to a framework for suggesting how to direct experimental effort on species s for the inference of new edges or increasing the confidence in existing edges on s and its neighbors. With the advent of multi-species biclustering [Genome Biol. 2010;11(9):R96. Epub 2010 Sep 29. Multi-species integrative biclustering. [Waltman P](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Waltman%20P%22%5BAuthor%5D), [Kacmarczyk T](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Kacmarczyk%20T%22%5BAuthor%5D), [Bate AR](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Bate%20AR%22%5BAuthor%5D), [Kearns DB](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Kearns%20DB%22%5BAuthor%5D), [Reiss DJ](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Reiss%20DJ%22%5BAuthor%5D), [Eichenberger P](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Eichenberger%20P%22%5BAuthor%5D), [Bonneau R](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Bonneau%20R%22%5BAuthor%5D)], co-regulated modules as well as edges will be found.

This project proposes to use, as a test case, 20 recently sequenced plant species *<list them and refer to figure>*. Owing to new next-gen sequencing techniques, genomic data such as transcriptomic has been generated, but most have been barely explored at the level of gene interaction networks. We will use the genome sequences and transcriptome data to collect and develop software that infers gene-interaction edges using inter-species and intra-species techniques. This is timely, because many additional plant species will soon be sequenced and then expression data will be available. Therefore, it is of interest to combine information about a species with information about neighboring species. The approach and pipeline tools we develop will be deployed using a gaggle-based *<ref>* interface, so biological tools can have easy access to it. Our project addresses several of the ABI goals:

1. *New algorithms for network inference:* Neighborly Network Inference methods for expression (Aim 1) and the generalization to other kinds of edges (Aim 2)

2. *Heterogeneous data*: Use of homology, expression, metabolic, and protein-protein networks (Aim 2).

3. *Tools for biological work-flows:* Helping biologists determine the next experiment to do (aim 3).

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

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

agronomic and/or environmental value.

**RESULTS FROM PRIOR NSF SUPPORT *(Dennis has not touched this)*** This proposal leverages on the accomplishments of the previous parent NSF grant, “Conceptual Data Integration for the Virtual Plant” (DBI-0445666). The VirtualPlant software platform (www.virtualplant.org) {Katari, 2010 #29} integrates genome-wide data concerning the known and predicted relationships among genes, proteins and molecules, as well as genome-scale experimental measurements. VirtualPlant also provides tools that render multivariate information into integrated visual displays (e.g. networks) to highlight biological implications. We have demonstrated the use of tools embodied in the VirtualPlant system to generate hypotheses that were subsequently experimentally validated {Gifford, 2008 #33;Thum, 2008 #35;Nero, 2009 #36;Wang, 2004 #37;Gutierrez, 2007 #38;Gutierrez, 2007 #39}.

**Fig. 1. The VirtualPlant Multinetwork.** The Arabidopsis multinetwork contains genes represented as nodes (A) that are connected by edges of many types (B) including metabolic, protein-DNA, protein-protein, microRNA-RNA, and edges derived from text mining [1]. (C) shows a network neighborhood resulting from querying this multinetwork with microarray data, uncovering a regulatory hub (CCA1) involved in nitrogen signaling [11].

Our NSF-ABI VirtualPlant grant had four goals: integration, visualization, synthesis, and prediction which we have accomplished, as outlined below.

**Aim 1.** **Integration**: *The Arabidopsis Multinetwork: A systems biology tool for hypothesis generation***.** Our VirtualPlant project included assembling the first multinetwork for Arabidopsis, a first step towards a molecular wiring diagram of the plant cell {Katari, 2010 #29;Gutierrez, 2007 #39}. 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 nitrogen-regulated genes resulted in a sub-network of 369 genes connected by one (or more) “expression correlation edges” {Gutierrez, 2008 #34}. 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 #34}. 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 {Gutierrez, 2008 #34}, 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 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 #33;Thum, 2008 #35;Nero, 2009 #36}. A complementary tool is GeneMania {Warde-Farley, 2010 #2}, which generates a hypothesis for gene function based on interactions with other genes and their attributes. For a recent review of various plant multinetwork approaches, see {Moreno-Risueno, 2009 #40}.

**Table I**. **Quantitative Information about the Edge Types of the Arabidopsis Multinetwork**. The multinetwork is described further in [1]. ALL REF NUMBERS ON THIS TABLE WILL CHAGE

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

**BioMaps**: BioMaps takes one or more sets of genes and determines which functional terms (GO {Ashburner, 2000 #43} or MIPS{Mewes, 2004 #44}) 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 {Katari, 2010 #29}.

**Sungear**: Sungear is a visually interactive and biologist-driven exploration of comparisons of the results of many experiments on a genomic scale. Sungear can represent an arbitrary number of experiments/lists, all of their disjoint intersections, and their related ontological terms. The position of a circle and arrows emanating from it indicate the input lists of which it is a subset. The size of a circle is proportional to the number of genes in the intersection of those lists (see {Poultney, 2007 #41}). Many biologists find Sungear to be an extremely powerful and interactive tool for analyzing the interrelationships between sets of genes {Gutierrez, 2007 #38}.

**NetMatch:** NetMatch, a Cytoscape plug-in, finds all instances of a query graph (e.g. a network motif) in a larger graph {Ferro, 2007 #42}. New versions compute the statistical significance of the motifs (e.g. Transcription factor motifs) found in a network.

Up and coming tools for VirtualPlant include **GeneSect** whose purpose it is to take a set of collections of genes and to determine whether any pairwise intersections among those collections are either surprisingly large (against a variety of backgrounds) or surprisingly small. Another new tool under development is a cluster management framework **ClusterBoss** to run some expensive tasks such as correlation and network inference in parallel, which relate directly to Aim 2 of the current application.? IS THIS STILL CORRECT?

**Aim 4. 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 learn regulatory networks {Krouk, 2010 #45;Mirowski, 2009 #47}.

This approach is more fully described in the Research Plan (Aim 2) (DOES THIS ALSO APPLY TO THE NEW GRANT Aim 2????) because it relates to this new NSF ABI proposal. Our second goal was to extend VirtualPlant to other species, such as Rice, which we have done (Fig. 2). NUMBER OF REF IN FIGURE WILL CHANGE>>>>

**Fig. 2. The VirtualPlant Arabidopsis and Rice Home Pages.** TheVirtualPlant software platform ([www.virtualplant.org](http://www.virtualplant.org)) is designed to support multiple species [1]. Shown are the two home pages for Arabidopsis and Rice. Each supports a common set of tools but is implemented on top of a separate database. An analysis within a species will not be slowed down by the addition of another species.

**Virtual Plant and User Community:**

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

**VirtualPlant DB**: The VirtualPlant database contains some of the most commonly used data types including metabolic pathways from KEGG and ARACYC, protein-protein interactions from BIND and Interolog databases, and GeneOntology and Gene annotations from TAIR (see Table I for a complete listing of data sources). The database also contains processed data obtained by analyzing publicly available Microarray experiments obtained from NASC {Craigon, 2004 #48}.

**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.

WE MAY WANT TO REVISE THIS TO REFER TO NEW REVIEW OF GRANT IN ABI AND SHORTEN OR OMIT THIS CRITIQUE OF THE PREVIOUS SUBMISSION TO PLANT GENOME)

**Plant Genome Application IOS-1025989: TRMS “Cross species network inference: From models to crops” (January 26, 2010):** In a prior NSF Plant Genome application, we proposed to build tools to infer networks in newly sequenced or under-analyzed species and to generate experimental data for them. The tools proposed in this current NSF ABI grant application constitute the computational portion only of the proposal that was previously submitted to NSF Plant Genome, which was ranked highly meritorious, but not funded. A*ll six reviewers of the previous Plant Genome application noted that the Cross Species Network Inference (NNI) tools were important, timely and would be of benefit to the entire plant community.* Below are excerpts of reviewer comments related to this point.

**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. This was seen as a strong proposal from an excellent interdisciplinary team of researchers.”

**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. Tool development efforts are well integrated in cyberinfrastucture, including iPlant and Galaxy”.

**Review 2**: “The NNI 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. … 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 by the reviewers 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 in Aim 2 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. We therefore use a method to refine our choice of experiments in Aim 1. Yet another criticism suggested that our techniques for obtaining orthology should be compared with those of InParanoid and OrthoMCL.

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

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

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

Nero D, Kelfer J, Katari MS, Tranchina D, Coruzzi G (2009) “In silico Evaluation of Predicted Regulatory Interactions in Arabidopsis thaliana”. **BMC Bioinformatics**. Dec 21;10(1):435.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**RESEARCH DESIGN**

**Aim 1: Development of the NNI model on Expression data**

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

Our neighboring species strategy for inferring edges between genes in species s starts with pair-wise gene expression correlation data on other species. From that data, we will train a machine-learning algorithm to determine whether there will be correlation between two genes in species s. (It is a separate question when this correlation signifies causality. If a g1 and g2 correlate, g1 is a transcription factor, g2 is not, and g2 has a DNA-binding site that the protein associated with g1 can bind to, then this is some evidence for causality. But the best test is a time-series experiment and analysis*<refs>*, followed by a knockout or overexpression experiment. We plan to continue to pursue such experimental analysis, but that is not the specific topic of this innovation proposal.)

Because the relevant data may come from many species and ultimately from many different kinds of data (e.g. expression, metabolic, protein-protein) as we will see in Aim 3, the input data will come in the form of three tables.

**orthotab: target species| target gene | other species | other gene | orthology val1 | orthology val2 …**: This table gives the gene to gene orthology value, according to several different orthology measures <*refs*>.

**edgetab: species | gene1 | gene2 | edgetype | strength | confidence width | number of different experimental conditions**: This table gives the strength and the length of the confidence interval (the “width”) of a given experimentally determined edge. We consider only experimentally determined edges as an input to this inference algorithm to avoid circular inferences. Note that certain edge relationships may be present only in certain conditions (e.g. drought conditions for plants). In that case, we might want to complicate this table representation to include condition-specific edge strengths. However, we prefer not to do this because that would cause our machine learning method to over-fit the data. Instead, we note simply how many different conditions have been tried that support this edge? (e.g. X/Y support the edge???).

**speciestab species1 | species2 | species similarity measure1 | species similarity measure2**: This table measures sequence similarity according to one of a number of criteria (e.g. distance based or through parsimony).

Now, to predict an edge between g1 and g2 in s, we will combine evidence from edges in other species, as well as evidence from experiments in species s itself. The basic method will be regression and regression trees with a penalty for complexity. For the sake of performance and robustness to noise, we will use some mixture of the following three approaches:

1. Random Forests [Breiman L (2001) Random forests. Machine Learning 45: 5–32]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 L. (2010) Large-scale Machine Learning with Stochastic Gradient Descent. Proceedings of the 19th international Confierence on Computational Statistics. Springer pp. 177-187] Bottou demonstrated that a stochastic gradient descent solver for a variety of learning problems(including support vector machine optimization) is able to scale with extremely large datasets while converging to the predictive performance of traditional optimization algorithms.

3. Large-Scale L-Regularized Learning [Shai Shalev-Shwartz and Ambuj Tewari (2009), Stochastic Methods for L1 regularized loss minimization. Proceedings of the ICML 2009, pages 929-936 http://doi.acm.org/10.1145/1553374.1553493] Stochastic coordinate descent can be used to learn sparse regression models, with small training times even for data sets where both the dimensionality and the number of training points is large.

The net effect of this analysis will be to find the weighting of different factors that will lead us to conclude that two genes in some species are correlated. Then using regression methods <*ref*> and the function of genes (e.g. there can be a causality edge from a transcription factor to a gene but not vice versa), we will tease out causality.

There are regression based approaches in which we assume that the value of a given gene is a function of a small number of other genes val(g) = a1\*val(g1) + a2\*val(g2) + …. We can add cross-terms and then use techniques like ANOVA model simplification to eliminate unnecessary terms. In the DREAM 4 competition, the best method find was a Random Forest approach, so we plan to explore and extend that method as well as a graphical model method.

Determining causality: Irrthum, A., & Wehenkel, L. (2010). PLoS ONE: Inferring Regulatory Networks from Expression Data Using Tree-Based Methods

In a random forest technique [Breiman L (2001) Random forests. Machine Learning 45: 5–32], a regression tree is built for each gene g. A regression tree is similar to a decision tree and determines which other genes could influence g and how they might do that.

Another method applies the Graphical Lasso [Menéndez, P., Kourmpetis, Y. A. I., ter Braak, C. J. F., & van Eeuwijk, F. A. (2010). Gene Regulatory Networks from Multifactorial Perturbations Using Graphical Lasso: Application to the DREAM4 Challenge. PloS one, 5(12), e14147] in order to estimate the sparse inverse covariance matrix of the matrix of gene expression values. The method works as follows: from the covariance matrix of the normalized gene expression values, Graphical Lasso is used to estimate the relationship among pairs of genes. Graphical Lasso uses L­­1 minimization and a penalty parameter to enforce sparsity (so not too many genes are related). The larger the penalty parameter is, the sparser the inverse covariance matrix will be. Because covariance matrices are symmetrical, they are undirected, so apply most usefully to interaction networks (e.g. protein-protein networks) rather than source-target networks (e.g., expression or metabolic).

For a species in which ample genetic perturbation and/or time-series data is available, a workflow might start by using steady state expression data to cluster similar genes together <*refs*> and possibly try to infer causality. The next step of the workflow is to use genetic perturbation data to find some relationships between source and target genes <*refs*> . The idea is simple: if the wild-type species has a certain value for gene g’ and the knockout for some other gene g changes the value of g’, then g must affect g’ (directly or indirectly?). Finally, time-series data permits the teasing out of causality information <*refs*> through a variety of ordinary differential equation approaches. Without going into detail (because we plan to use best-of-breed approaches that we and others have developed), here is an overall workflow: <*workflow*>

CASE STUDY GOES HERE

**Expected Outcomes of Aim 1.**  Our goal is to construct a machine-learning model that can predict, with high precision, the correlation of edges between genes? with a reliable confidence interval, leading to some estimate of causality. We may find that we have a trade-off between precision and recall. That is, we may find that the model gives narrow confidence bounds for some edges and meaninglessly wide ones for other edges. This may be acceptable if the narrow bounds are for those edges having high (positive or negative) correlations.

**Aim 2: Proof-of-principle verification of Neighborly Network Inference (NNI) on heterogeneous data**

***Rationale***. In production, Neighborly Network Inference (NNI) will be used to infer gene networks in nearby species based on many data sources. The purpose of Aim 2 is to verify the NNI methodology by inferring heterogeneous networks in species for which networks have been experimentally determined, and then evaluating the accuracy with which the inferred network predicts the experimental network. Here, we have chosen Arabidopsis and Rice, because they each have the most complete genomic data set to test our methods. Despite the fact that these species are phylogenetically far apart, the high accuracy obtained suggests that the approach could have wide applicability. We discuss 1) our preliminary analysis and results, 2) the overall objectives of this aim, and 3) its expected outcomes.



***Preliminary results***. Our preliminary results demonstrate NNI's ability to infer gene networks from Arabidopsis to Rice with impressive accuracy, as shown in Table II.

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

For the data in Table II NNI was used to *infer a* Rice network that was then compared to the *known* validated data for Rice including metabolic data from KEGG {Kanehisa, 2004 #49}, and protein-protein interaction data from BIND plus other experimentally determined protein interactions {Ding, 2009 #4;Popescu, 2009 #3;Rohila, 2009 #5;de Folter, 2005 #51;Popescu, 2007 #52}. We used validated gene interaction datasets for Arabidopsis and Rice to develop and validate a methodology for inferring networks for other species, and later apply the NNI pipeline to under-analyzed species that have little interaction information (Aim 3). Our approach builds on inference approaches based on expression and homology {Gholami, 2010 #7;Mutwil, 2008 #19;Yu, 2004 #15}, and also based on integration of several different types of associations {Geisler-Lee, 2007 #20;Mutwil, 2008 #19}.

**Below are the steps used in the NNI approach:**

**Step 1.** **Obtain a reference validated Arabidopsis interaction network based on experimentally supported data**. For our validated Arabidopsis networks, we have assembled metabolic interactions (KEGG; 19,688 interactions) {Kanehisa, 2004 #49}, protein-protein interaction data from BIND (949 interactions) {Bader, 2002 #50}, protein-chip interaction data for MADS box (272 interactions) {de Folter, 2005 #51}, and protein chip interactions for Calmodulin (755 interactions) {Popescu, 2007 #52} and the Plant Interactome project (11,374 interactions) (http://signal.salk.edu/interactome.html). (MANNY HAS ADDED NEW PROTEIN INTERACTION DATA AND PROTEIN DNA INTERACTION DATA (e.g. from Siboean Brady) THIS SHOULD BE UPDATED). Many of the metabolic pathways in the KEGG and AraCyc databases are based on computational predictions, while 25% are validated experimentally in the literature {Masoudi-Nejad, 2007 #55;Zhang, 2005 #56}.

**Step 2. Identify Rice homologs of Arabidopsis interaction pairs.** Connect every gene in the Arabidopsis interaction network with its Rice homologs. This technique can employ various homology methods, including either distance or parsimony based. In our preliminary analysis (Table II), we obtained homologs via two commonly used methods, InParanoid {O'Brien, 2005 #24} and OrthoMCL {Li, 2003 #26}. We also experimented with distance-based homology, selecting homologs with BLAST matches stronger than E-value of E-20 to capture one-to-many homology relationships {Tatusov, 1997 #57}, which captures the gene duplication events prevalent in plant genomes {Zhang, 2003 #58}.

**Step 3. Build a Rice correlation network based on publicly available Rice microarray expression experiments.** We downloaded all 48 Rice gene expression experiments on the Affymetrix GPL2025 platform from GEO {Barrett, 2007 #54}. With the aim of finding experiments that both repress and induce the genes of interest (the Rice genes homologous to the genes in the Arabidopsis network), we selected the experiments with the highest variability of expression level across samples for these genes. These were experiments in which at least half the individual gene Z-scores across the samples exceeded 0.5. This selected 8 experiments with a total of 169 samples. We then computed the Pearson correlation of all pairs of the genes of interest. We retain correlation edges between gene pairs whose expression vectors were significantly correlated (p-value <0.05, meaning less than a 5% chance of a non-zero correlation by chance) and absolute value of correlation > 0.5 or >0.7 (Table II).

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

**Step 5. Obtain a reference validated Rice network that contains edges representing known interactions.** Our initial Rice validated network was constructed from 10,976 metabolic interactions and 334 protein-protein interactions for Rice from KEGG {Kanehisa, 2004 #49} and BIND {Bader, 2002 #50}.

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

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

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

To determine whether our general homology plus expression correlation technique would work for other kinds of edges, we tried to infer Rice protein-protein edges from Arabidopsis protein-protein edges and expression data. Unfortunately, there are only 11,241 *non-redundant* validated protein-protein edges in Arabidopsis and only 344 in Rice {Bader, 2002 #50}, so many of our predictions did not fall among those 344, but may one day be validated. Surprisingly, simple homology techniques (reciprocal top Blast hits and InParanoid with homologs of paralogs) each obtained a quite high precision of about 50% and recall of between 4% and 8%. In those techniques, an edge between rice genes r1 and r2 would be inferred when r1 was homologous to a1, r2 to a2, and a1 and a2 formed a validated protein-protein edge in Arabidopsis. Expression data (either on all experiments or just those in which the expression value of potential homologs varied the most) sometimes improved precision but at a severe loss in recall.

These very preliminary results suggest that machine-learning techniques are needed to determine the proper weights of different forms of evidence.

**Step 7. Expand validated and network inference into a “multinetwork” containing multiple edge types**. We will use techniques analogous to Steps 1-6 to infer networks based on other edge types. For example, we will add regulatory interactions including protein-DNA (AGRIS: 343 interactions) {Davuluri, 2003 #59} and miRNA:RNA interactions {Griffiths-Jones, 2006 #60;Gustafson, 2005 #61;Lu, 2005 #62}. Expanding the validated networks to include these datasets will enable us to create an inferred multinetwork that includes: protein-protein, Protein:DNA, miRNA-RNA and Metabolic edges.

**Role of machine-learning.** As one would expect, the choice of data sources, expression experiment selection methods and homology algorithms and parameters greatly influence the accuracy of the inferred Rice networks. That is why the machine learning techniques outlined in aim 1 will be used. The experiments used for gene expression correlation will include many different developmental stages, different organs, and different biotic and abiotic treatments such as the ones recently released for Rice on GEO NCBI {Wang, 2009 #64}.

**Objectives of Aim 2**. Through this work we will evaluate the accuracy of NNI on additional species pairs and data sets. These will include:

1. Use machine-learning techniques to improve the selection of parameters for each form of information (edge type, similarity of species, etc.). In some cases, the target edge type may not constitute a good target. For example, our preliminary results (not shown) indicate that Kinase networks {Ding, 2009 #4;Popescu, 2009 #3;Rohila, 2009 #5} cannot be accurately – recall and precision each top out at a few percent – inferred between Arabidopsis to Rice. One reason may be that TF-target edges – which constitute the majority of edges in Kinase networks – evolve too rapidly to be conserved at the Arabidopsis to Rice phylogenetic distance.

1. As data become available, on an ongoing basis, evaluate the accuracy of NNI for other species pairs and data sets. For example, NCBI now contains 147 experiments on Zea mays, and 37 for Medicago truncatula, and large-scale Arabidopsis and Rice protein interaction datasets are being created and will be made available (Joe Ecker – NSF Plant Interactome Project, see personal communication (letter of support?). In general, we expect that gene network inference will perform better between species that are phylogenetically closer. For example, we predict that inference between Zea mays and Rice will perform better than inference between Zea mays and Arabidopsis because the former are both monocots. We will expand our understanding of the phylogenetic context of gene interactions in Aim 4.

**Expected Outcomes of Aim 2**. We will refine our NNI approach as we experiment with additional data sets, and refine the homology methods and expression data incorporation. We expect that inference statistics will improve as sequence and experimental data become available for species pairs that are phylogenetically closer than Rice and Arabidopsis. More data will enable better inference. This Aim provides a testing ground and validation for the NNI simulation approach that we will automate in Aim 3.

DENNIS- SHOULD PHYOLGENETIC FRAMWORK FOR NETWORK INFERENCE GO HERE? AND LEAVE MOST SPECULATIVE AIM LAST?

**Aim 3: Predicting experimental “Pay-off”: Framework to Determine the Next Best Experiment to Perform (Move to last aim? After phylogenetic framework for network inferences?)**

***Rationale*:** In this aim (our most speculative), we propose a framework to estimate the information we might learn from a new set of experiment instances on a given species in order to determine which will be most useful. The goal is to minimize experimental time and expense.

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

To start, we will evaluate the “*payoff*” of a set of experiment instances *already done* as follows: compare our state of knowledge *before* doing them with our state of knowledge *afterwards*. To measure the difference, consider the edges after the experiments as “better” than those before. The payoff, is the number of edges that have improved, i.e. how many false-positives have been corrected, how many false-negatives have been corrected, and how many borderline cases have been resolved. For example, if we are interested in determining which pairs of genes have a correlation threshold above 0.7 (in absolute value) with a 95% confidence width below 0.2, then we can use either parametric or non-parametric statistics <*refs*> to determine, for each gene pair, whether that pair achieves the threshold (a positive), doesn’t (a negative), or might (e.g. the mean is above 0.7 but the confidence width is too wide). The question is which experimental strategy will resolve as many edges as possible. Because our approach is an inter-species approach, we will look at our machine-learning model to determine the numbers of edges in other species that can also be resolved as well.

Suppose we are given a “budget” of n experimental instances, where each instance is a single assay (e.g. a single microarray or chip-chip assay). We will use the above method to determine which mix of replicates under existing conditions, replicates under c new conditions with r replicates each (where n >= rc), or some number of time-series experiments where there are r replicates during each time-point. The computational method will not determine which conditions to try (that requires biological insight), just how many new conditions would probably lead to the most learning.

For a certain little-studied species s, this “take-away and simulate” strategy may not work, because there may not be enough experiments to take away in that species. For that reason, we might use a different species s’ that is more studied and is statistically similar to this one. Statistical similarity will be measured as follows: take from s’ a subset of its experiments that reflects the diversity of the experiments done on s. For example, if three conditions have been tried on s having 2, 3, and 4 replicates respectively, then find the subset of experiments on s’ having three conditions with 2, 3, and 4 replicates. Next, find the number of edges known to be above threshold in s’, the number known to be below threshold, and the number in between. If those numbers are similar for s’ and s, then try computational experiments on s’ in which we add in our budget of n experimental instances using different combinations of conditions and replicates and see which gives the best payoff. Whichever combination is best then becomes the strategy we will use for s.

In many ways, this work falls in the pool-based sampling subcategory of the active learning framework [Active Learning Literature Survey by Burr Settles, Computer Sciences Technical Report 1648, University of Wisconsin–Madison, January 26, 2010]. In active learning, the learning algorithms “asks questions” to try to optimize the amount of information gained. An example in biology was done by King et al. [ R.D. King, K.E. Whelan, F.M. Jones, P.G. Reiser, C.H. Bryant, S.H. Muggleton, D.B. Kell, and S.G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature, 427(6971):247–52, 2004. R.D. King, J. Rowland, S.G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L.N. Soldatova, A. Sparkes, K.E. Whelan, and A. Clare. The automation of science. Science, 324(5923):85–89, 2009.] to discover metabolic pathways. The idea is that the active learner chooses a mutant and growth medium and sees whether the mutant survives and chooses the most useful one for the purpose. Pool-based sampling is the idea that there exists a large pool of potential experiments to be performed and one must choose among them. The most common approach is “uncertainty sampling” in which one performs experiments on data that one is least certain about (in information theoretic terms, the ones with maximum entropy). (Lewis and Gale, 1994) (Settles and Craven, 2008). Another approach is called Expected Model Change in which we try to learn the experimental instances that would improve our current model as much as possible if we knew the outcome Settles et al. [B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In Advances in Neural Information Processing Systems (NIPS), volume 20, pages 1289–1296. MIT Press, 2008.] Our approach attempts to follow the expected model change approach.

**Preliminary results**: *Manny: I need you for this* 1) The following table shows the number of experiments and their conditions on the species of interest to us. We also note how many expression correlation edges have an absolute value as great as 0.7 and a 95% confidence width of 0.2 or less. Given a budget of 40 experimental instances, we use our method to calculate the payoff in each case. For the purposes of this preliminary work, we do the analysis on each species independently of others.

**Objectives of Aim 3.** Our objective is to provide a tool for experimentalists to suggest which group of experiments to try next on some species s. If the experimentalist wants to learn about a whole group of related species, then our method will use the Neighborly Network framework to estimate the payoff for other species as well as for s itself.

**Expected Outcomes of Aim 3.** Neighborly Network Inference will both infer edges and suggest experimental strategies. The two goals work nicely together because inference is needed to calculate the payoff of an experiment.

AIM 4: GENERATION OF A PHYLOGENIC GENE INTERACTION FRAMEWORK

DENNIS- BEFORE I TRY TO WRITE THIS, DO YOU WANT TO DO A DRAFT?

**TIMELINE:**

**Year 1:** Aim 1. Implement Neighborly Network Inference using a variety of machine learning methods, starting with linear regression and extending to various flavors of stochastic gradient descent. Cross-validate on the *<how many>* expression experiments from our 20 species. Try the same approach among other eucaryotes. Aim 2. Gather and normalize the data for validated protein-protein and metabolic interaction networks for plant species.

**Years 2-3:** Aim 2. Extend the Neighborly Network Inference to other species and data types. Aim 3. Build the framework for determining the best new experiments on cross-validated data. Deploy the first version of the NNI analysis to collaborators (R. Gutierrez, Chile) and a growing community of beta testers.

**PLAN TO INTEGRATE RESEARCH AND EDUCATION**. *Dennis has hardly touched this*

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

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

**Training Postdocs as educators**. In this project, Post-Docs are paired up with graduate students, undergraduate students, and technicians in the laboratory to practice mentoring skills in a research context. At NYU, post-docs are also afforded the opportunity to teach and are mentored by faculty advisors. Post-Docs also receive counseling from their co-mentors and practice presentation skills during regular group-lab meetings, through a Post-Doc seminar series, and at annual poster sessions at NYU.

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

**SHARING OF RESULTS**: The informatic analysis pipelines for Cross Species Network Inference (NNI), discussed in Aim 3 will be made available to the community free of charge, deployed on a website (www.crossspecies.org) linked to several additional platforms, first to VirtualPlant website (www.virtualplant.org), and second to *iPlant* (see S. Goff letter), and third as a webservice. **Publications:** The results of our analysis of the data we generate will be made available through peer- reviewed literature as it is the most appropriate way to make this information available.

**MANAGEMENT PLAN**: To coordinate and facilitate interactions between individuals, Dennis Shasha (NYU Computer Science) will serve as the overall Project Manager and Gloria Coruzzi (NYU Biology) will serve as a biological advisor and conduit to a working lab and the wider plant community. The role of the Project Manager is to oversee the daily operations of the project and ensure that the needs and concerns of the participants are addressed on a day-to-day basis between the participants involved. The project manager will also facilitate communication between PIs, post-docs, graduate students and laboratory technicians by scheduling weekly meetings of all participants to manage immediate issues regarding research needs. We will also schedule day-long meetings twice a semester with our collaborator (Rodrigo Gutierrez, Chile), to do evaluation of work status and long term planning.

**Bioinformatics manager: Dr.** **Manpreet Katari** (NYU Biology) will be in charge of the bioinformatics data. To enable efficient information exchange of raw and processed data, a file server has been set up at the NYU to store and distribute data and its analysis among users at NYU Biology and NYU Courant. Dr. Katari will maintain the web server, database server, and the multinetwork database.

**Software development manager: Dr. Arthur Goldberg** (NYU Courant) will manage the development of new software analysis tools and pipelines to enable Cross Species Network Inference (NNI) which will support the different species and inference, and also new pipelines for cross species analysis, especially as they relate to crop species in coordination with the PI, the programmer Rebecca Davidson, and a computer science doctoral student.

**Website:** We have set up a web site to house the development of Cross Species Network Inference tools and pipelines, which is accessible at: [www.CrossSpecies.org](http://www.CrossSpecies.org)

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

**Role of senior participants:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Institution** | **Role** | **Aim** |
| ***Dennis Shasha***PI | NYU Courant | Project Leader: Computational | Oversee Aims 1, 2, 3 |
| ***Gloria Coruzzi,***Co-PI | NYU Biology | Co-leader: Biological | Oversee Aims 2 & 3 |
| ***Manpreet Katari***Co-PI | NYU Biology | Bioinformatics Manager | Aims 1, 3 |
| ***Arthur Goldberg***Senior Personnel | NYU Courant | Software developer | Aims 1, 3  |
| ***Rodrigo Gutierrez***Consultant | U Catolica,Chile | Assembling validated networks for targets | Aim 1 |

**COORDINATION WITH OUTSIDE GROUPS**

**Please see attached letters of collaboration:**

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

**iPlant (see letter from iPlant Project Director, Steve Goff)** We will coordinate with iPlant to make our Cross species network inference platform (NNI) modular, independent and accessible with and compatible with iPlant, and accessible using other annotation analysis platforms such as Galaxy and Taverna. We will also make our currently developed VirtualPlant tools accessible to iPlant, as per letter by (S. Goff).