**PROJECT SUMMARY** “**ABI Innovation: Cross Species Network Inference**”

**1. Senior personnel**

 **PI:** Dennis Shasha (NYU Courant Institute of Mathematical Sciences)

 **Co-PIs:** Gloria Coruzzi & Manpreet Katari (NYU Biology, Center for Genomics & Systems Biology)

 **Senior Personnel:** Arthur Goldberg (NYU Courant Institute of Mathematical Sciences)

 **Collaborators:** Rodrigo Gutierrez, Catolica Universita de Chile.

**2. Intellectual merit of the proposed activity** We propose to develop a Cross Species Network Inference (CSNI) platform that will enable plant biologists to computationally infer difficult-to-obtain relationships in gene or product networks in less-studied species which we call Targets. CSNI will infer from edges in validated networks in well-studied species, which we call References, such as Arabidopsis, by integrating homology and easily collected data in the Target, such as gene chip expression measurements and RNA-Seq data. Inferred networks in Targets will provide testable hypotheses about their gene and product interactions, as well as suggestions for future experiments, especially time-series experiments and genetic modifications. This project will leverage the facilities of the current VirtualPlant software platform ([www.virtualplant.org](http://www.virtualplant.org)) developed under an NSF Arabidopsis 2010 Grant (DBI-0445666) including Arabidopsis multinetwork data, analysis, integration and manipulation tools [1]. As output, we will provide a pipeline of tools for Cross Species Network Inference (CSNI) to the community via a website ([www.CrossSpecies.org](http://www.CrossSpecies.org)) and the NSF*iPlant*Project (see letter). In addition, our CSNI framework will build on the infrastructure of a generic bioinformatic analysis platform engine such as Taverna [2], Kepler [3], or Galaxy [4]. As a proof-of-principle, we will apply this Cross-Species Network Inference framework to predict metabolic and protein-protein interaction edges in Rice where existing experimental data enables validation. We will then extend CSNI to other genomes, for which experimental data supporting network edges is not yet available. While CSNI is here described with respect to plants, the framework and basic algorithms extend to any under-analyzed species. This work will achieve one of the main goals of Systems Biology – predicting network states under untested conditions.

**We divide the work into three aims:**

**Aim 1.** **Proof-of-principle verification of CSNI, with preliminary results from Arabidopsis to Rice.** We verify our CSNI methods by inferring networks in Target species for which validated networks are known, and evaluating the accuracy of the inferred networks. We'll extend and perform this verification for multiple pairs of Reference and Target species, and multiple types of networks as they data become available. Our preliminary results show high precision in metabolic and protein:protein networks in Rice inferred from Arabidopsis.

**Aim 2. Inference of Regulatory Networks Using Time Series.** As time series constitute a particularly informative method for inferring networks, we show the power of a model resulting from a closely spaced time series transcriptome of a nitrogen-treatment experiment in Arabidopsis *even in the absence of other data*. We describe our “State-Space Analysis” machine learning method and validate the resulting network based on its predictive accuracy on out-of-sample data.

**Aim 3. Develop a Bioinformatic Pipeline for Cross-Species Inference (CSNI).** Provide a biologist-friendly CSNI software platform (www.CrossSpecies.org) that will infer networks in a Target species, given experimentally validated networks in a Reference species, homology information and experimental data in the target species.

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

**PROJECT DESCRIPTION**

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

Our proposal to ABI is novel in its framework for Cross-Species Network Inference (CSNI) and in its algorithmic ideas for network inference from time series. Its draws its impact from the fact that many plant species will soon be sequenced and RNA-seq data obtained and so it is of interest (i) to apply information gathered from other species and (ii) to design informative experiments and analyze them through time series. The approach and pipeline tools we develop will be deployed on a biologist-friendly Web interface that plant scientists can use to infer regulatory networks for any trait of interest. Our project addresses several of the ABI goals:

1. *New algorithms for network inference:* Cross-Species Network Inference methods (Aim 1) and Inference from time series data using State-Space modeling (Aim 2)

2. *Heterogeneous data*: Use of homology and expression to infer metabolic and protein-protein networks (Aim 1).

3. *Tools for biological work-flows:* Cross Species Network Inference (CSNI) Pipeline (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** 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) [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 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 [5-10]. The VirtualPlant grant had four goals: integration, visualization, synthesis, and prediction which we have accomplished, as outlined below.

**Aim 1.** **Integration**: *The Arabidopsis Multinetwork: A systems biology tool for hypothesis generation***.** Our VirtualPlant project included assembling the first multinetwork for Arabidopsis, a first step towards a molecular wiring diagram of the plant cell [1, 10]. 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” [11]. 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) [11]. 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 [11], 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 [5-7]. A complementary tool is GeneMania [12], 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 [13].

**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 [14] or MIPS[15]) 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 [1].

**Sungear**: Sungear is a visually interactive and biologist-driven exploration of comparisons of the results of many experiments on a genomic scale. Sungear can represent an arbitrary number of experiments/lists, all of their disjoint intersections, and their related ontological terms. The position of a circle and arrows emanating from it indicate the input lists of which it is a subset. The size of a circle is proportional to the number of genes in the intersection of those lists (see [16]). Many biologists find Sungear to be an extremely powerful and interactive tool for analyzing the interrelationships between sets of genes [9].

**NetMatch:** NetMatch, a Cytoscape plug-in, finds all instances of a query graph (e.g. a network motif) in a larger graph [17]. 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.

**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 [18, 19]. This approach is more fully described in the Research Plan (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).

**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 [20].

**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 IOS-1025989: TRMS “Cross species network inference: From models to crops” (January 26, 2010):** This prior NSF Plant Genome application proposed to build tools to infer networks in newly sequenced or under-analyzed species and to generate experimental data for several. The tools proposed 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 (CSNI) 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 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.”

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

***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: Proof-of-principle verification of CSNI, with preliminary results from Arabidopsis to Rice.**

***Rationale***. In production, Cross Species Network Inference (CSNI) will be used to infer gene networks in species that lack difficult-to-obtain data, such as experimentally determined protein-protein or metabolic networks. The purpose of Aim 1 is to verify the CSNI methodology by inferring networks in species for which networks have been experimentally determined, and then evaluating the accuracy with which the inferred network predicts the experimental network. 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 CSNI's ability to infer gene networks from Arabidopsis to Rice with impressive accuracy, as shown in Table II

TABLE II: Current Performance measures of Cross species network inference (CSNI)

|  |  |  |  |
| --- | --- | --- | --- |
| Interaction network | Best current CSNI methodology  | Inferred networkRecall | Inferred networkPrecision |
| Protein-protein | Homology by reverse top BLAST hits | 4% | 56% |
| Metabolic | Homology by InParanoid | 18% | 95% |

For the data in Table II CSNI was used to *infer a* Rice network that was then compared to the *known* validated data for Rice including metabolic data from KEGG [21], and protein:protein interaction data from BIND plus other experimentally determined protein interactions [22-26]. 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 CSNI pipeline to under-analyzed species that have little interaction information (Aim 3). See Fig. 3 for overall plan. Our approach builds on inference approaches based on expression and homology [27-29], and also based on integration of several different types of associations [28, 30]. **Below are the steps used in the CSNI 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) [21], protein:protein interaction data from BIND (949 interactions) [31], protein-chip interaction data for MADS box (272 interactions) [25], and protein chip interactions for Calmodulin (755 interactions) [26] and the Plant Interactome project (11,374 interactions) (http://signal.salk.edu/interactome.html). Many of the metabolic pathways in the KEGG and AraCyc databases are based on computational predictions, while 25% are validated experimentally in the literature [32, 33].

**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 [34] and OrthoMCL [35]. 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 [36], which captures the gene duplication events prevalent in plant genomes [37].

**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 [38]. 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 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 [21] and BIND [31].

**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.** We suggest two main conclusions from our preliminary analysis of Cross Species Network Inference (Steps 1-6 above) shown in Table II. *First*, homology alone does an excellent job of inferred 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%. *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%.

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 validated protein-protein edges in Arabidopsis and only 344 in Rice [31], so many of our predictions that did not fall among those 344 will not count as true positives in our analysis 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 other analytical techniques for expression data -biclustering, rank-based correlation, or logistic regression techniques (see Aim 2) may be needed. Additionally, the precision of the metabolic network prediction is so high that we hypothesize that many of the predicted protein interaction edges may represent true interactions that are currently missing from the Rice protein interaction database.

**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) [39] and miRNA:RNA interactions [40-42]. 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.

**Parameter optimization.** 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. To simplify the selection of these parameters for biologists, we will systematically explore the space of these inputs, with the objective of maximizing the accuracy of our network inference predictions. A well-known technique for finding globally optimal parameters is *simulated annealing*, a probabilistic heuristic for finding global minima in large search spaces [43]. Ideally, 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 [44].

**Objectives of Aim 1**. Through this work we will evaluate the accuracy of CSNI on additional species pairs and data sets. These will include:

1. Reverse the inference direction of the preliminary results, and verify the inference of gene networks from Rice to Arabidopsis. This offers the advantage that the networks in Arabidopsis are more complete than those in Rice, so we expect fewer false negatives.
2. As data become available, evaluate the accuracy of CSNI 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 – personal communication). 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.

Also, we will evaluate the effectiveness of inference as a function of the type of gene relationship. For example, our preliminary results (not shown) indicate that Kinase networks [22-24] cannot be accurately – recall and precision each top 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.

**Expected Outcomes of Aim 1**. We will gain greater confidence with our verification of our CSNI mechanisms as we experiment with additional data sets, and refine the homology methods and expression data incorporation. We anticipate that automating the exploration of the input space may find optimal approaches with significantly better statistical outcomes. Also, we expect that inference statistics will improve as genome sequences and expression data become available for species pairs that are phylogenetically closer than Rice and Arabidopsis. This Aim provides a testing ground and validation for the CSNI pipeline approach that we will automate in Aim 3.

**Aim 2: Inference of regulatory networks: Develop a time series expression methodology using multiple data sources to infer regulatory networks.**

***Rationale:*** A principal aim of the biological enterprise is to learn which genes affect which functionalities and how. Whereas it is possible to create models and draw networks based on correlation and other data-driven relationships, the ultimate test of any model is the ability to predict the values of assays (e.g. expression levels) on conditions that were not used in training. If the model is in the form of a causal network, it can be used to suggest gene modifications that may optimize the performance of a species to some useful end. Because causality moves forward in time, time series experiments are a particularly promising source of network structure. In this aim, we describe our methodology in some detail, explain prior results, and explain how we integrate the methodology with other existing information (e.g. validated transcription factor binding sites, structurally based contact-binding sites, and expression data). The idea is to use this methodology for time series experiments in newly sequenced/under-analyzed species. The methodology runs on a standard parallel cluster, so our cluster management software (ClusterBoss, mentioned above) can be used for any site that downloads Virtual Plant.

**Predicting Arabidopsis regulatory networks using time series data and “State Space” analysis a machine learning approach.** The experimental approach of our Arabidopsis time-series [18] was to monitor transcriptome responses to nitrate treatment at 0, 3, 6, 9, 12, 15 and 20 min, using Affymetrix ATH1 chips. In order to build a regulatory network that could predict TF-target interactions, we used a machine learning method, “State-Space” modeling to generate predictions for regulatory networks [19]. The State-Space model synthesizes Bayesian and Markovian approaches (in which each gene’s expression value at a time *t* is assumed to depend directly only on the state of potentially all the genes at the previous time point and indirectly on values from previous time points) [19, 45].

In the “State Space” model depicted in Fig. 5B, each node represents the values of all gene expression at a particular time point. Typical values of all gene expression are depicted as a heat map in Fig. 4. The goal of this approach is to *learn* the function that determines the change in expression of a target gene as a linear (or if needed non-linear) combination of the expression of a relatively small number of transcription factors (typically up to three or four) (Fig. 4). As applied to our problem, the set of all genes at time t is modeled by a “latent” (i.e. hidden) variable (denoted Z(*t*)) from which noisy and sometimes missing observations Y(*t*) are made. Latent variables are represented by large red circles, and observed variables by large black circles in Fig. 5B. The relationship between latent and observed variables is the identity function *h* with added Gaussian noise (represented by a black square in Fig. 5B). An unknown *function f* (represented by a red square in Fig. 5B) relates the values of latent variables Z(t) and Z(t+1) (for all *t)* corresponding to consecutive time measurements as a Markov chain. The *dynamical function f* factors in both transcription factors and their target genes (e.g. other TFs or target genes), as shown in Fig. 4. Learning the *function f* corresponds to finding parameters of *f* that minimize the prediction error while penalizing functions that are excessively complex (i.e. require many transcription factors to determine the change in expression of a target).

An iterative procedure tries to learn the dynamical relationship between latent gene expression variables **z**(*t*) while maintaining the latent variables **z**(*t*) as close as possible to the observed Affymetrix measures **y**(*t*). The algorithm consists in a) minimizing the sum of quadratic errors of the dynamical and the observation models with respect to the latent variables **Z** by using gradient descent on the latent variables [19] (this is the inference step); and in b) minimizing the sum of quadratic errors of the dynamical model using conjugate gradient, LARS [46] or Elastic Nets [47] optimization on the parameters of **F** (this is the learning step). During the learning step, sparse gene regulation networks are obtained by penalizing dense solutions using L1-norm regularization, which amounts to adding a *λ*-weighted penalty to the dynamical error term, as in the LASSO initially described by [48].

To test the ability of the “State Space” approach to generate a *predictive* regulatory network, we built a regulatory network using the Arabidopsis time-series data up to 15 minutes (training set: 0, 3, 6, 9, 12, 15 min) and used the resulting network to *predict* the direction of gene change (up regulation or down regulation) from 15 min to 20 min (Fig. 5). Our State Space predictions of gene regulation were correct for 74% of the genes in a small network of 76 genes (Fig. 5B). As a basis for comparison, the *"naive trend forecast"* that predicted the direction of change from 15 to 20 min to be in the same direction as the movement from 12 to 15 min, was correct for only 52% of the genes, just slightly better than random (Fig. 5C), p-value < 0.006. This “State Space” model can also be used to predict the “most influential TFs” in the network (e.g. the one that is predicted to influences the most genes in the network), and to generate a time-dependent regulatory network model for the control of N-assimilatory pathway genes.

10

When compared with other network inference approaches [49-52] our state-space method showed a slight improvement in accuracy and had a better signal to noise ratio when compared on the same data. Further, the method reduces the importance of initial parameters by using random starting points and bootstrapping, thus offering a principled way to deal with uncertainty and avoid over-fitting in microarray measurements. Further, our method easily allows the addition of “hints” in the form of known transcription factor-binding relationships. Finally, the state space method combined with biclustering generalizes to larger networks. In a network of 550 genes including the original 76, we ran the biclustering algorithm CMonkey[53] using default settings on the non-TF genes. (The biclustering algorithm makes use of metabolic interactions, the Araridopsis Prolinks file, as well as gene expression.) This resulted in a reduced network consisting of 67 TF genes and 63 biclusters among the 483 non-TF genes. On this network of 67 + 63 = 130 supernodes, the state space method predicted the direction (expression up or down) accurately for 67.7% of all supernodes in the 15-20 minute time point compared with 51.9% for the trend forecast prediction.

Here is a summary of the approach for network inference:

1. Find sentinel genes that respond to the given treatment.
2. Test them at a variety of times
3. Decide on the times of global expression experiments by selecting times in which the sentinel genes vary the most.
4. Do the experiments at these time points
5. Analyze transcriptome data using our State space machine learning algorithm
6. Try a leave-out-one test to test whether resulting network gives good enough predictions to be useful.

**Expected outcomes of Aim 2:** The results of this aim will generate a robust package that can be used for network inference across species and for a variety of time series experiments. The goal will be to help the design of time series and to construct dynamic regulatory networks using high performance parallel cluster architecture. From this, we can identify a core set of regulatory networks.

**Aim 3: A Bioinformatic Pipeline for Cross-Species Network Inference (CSNI).**

***Rationale*:** In this aim, we will build a publicly available, production quality, Cross-Species Network Inference (CSNI) pipeline that will provide the plant scientist community (especially those with no informatics training) with a biologist-friendly tool for inferring gene networks in newly sequence/unanalyzed species. CSNI employs data about two species, 1) the under-analyzed species – which we call the *Target* species, and 2) a species that has been deeply studied, which we call the *Reference* species. The basic idea of CSNI is that the larger data set from the reference species will be mapped by homology into the target species, and combined with data about the target species to infer a network for the target species, as described in Aim 1.

***Figure 3A illustrates the CSNI pipeline***. A plant scientist who wants to infer a gene network for a target species will set the free parameters that determine the homology and inference methods of CSNI. These include i) the pair of *Reference* and *Target* species chosen, ii) the data sets selected from these species, iii) the homology mechanism and its parameters (such as BLAST E-value thresholds, COG [36], InParanoid [34], or OrthoMCL [35] (Li et al 2003) if distance-based homology is desired or parsimony methods [54]), and iv) the *Inference* rules which combine these data into the target species' inferred network.

In the *inference step*, the biologist chooses a validated network (e.g. metabolic, protein-protein, etc.) in the *Reference* species. Next, the biologist chooses some parameter settings or allows an optimization technique such as simulated annealing to set those parameters. For example, a combination rule might infer a regulatory edge in the target species if the edge’s genes were connected by an expression edge with correlation > 0.7 and the edge had homologous genes connected by a regulatory edge in the reference species’ validated network. Given all this, CSNI infers the target network.

We plan to deploy CSNI ([www.CrossSpecies.org](http://www.CcrossSspecies.org)) on several platforms, first on our VirtualPlant website (www.virtualplant.org) and second on *iPlant* (see S. Goff letter) where it will join the sophisticated tools currently in development. CSNI will use one of the widely-used bioinformatic workflow engines: Taverna [2], Kepler [3] or Galaxy [4] because workflow engines provide increasingly popular platforms for developing computational genetic analyses, and provide generic support for reproducible bioinformatic analyses.

**Expected outcomes of Aim 3.** The CSNI pipeline analysis constructed in Aim 3, and made available to the community as a biologist-friendly interface, will empower plant biologists to use network approaches to derive testable hypothesis for gene functions in crop species for which limited genomic information is available. Identifying networks conserved between reference and crop species will also enable researchers to focus their translational studies from models to crops.

**TIMELINE:**

**Year 1:** Aim 1. Extend cross species network inference using validated protein:protein and metabolic interaction networks for Rice and Arabidopsis to other homology methods. Extend network inference analysis beyond protein-protein interaction to validated regulatory (AGRIS) edges as well as miRNA-RNA edges. Aim 2: Analyze time series experiments in various plant species to validate our network inference approach. Aim 3: Assemble validated networks in the 3-5 target crop species beginning with Medicago, Corn, Grape. Select bioinformatic workflow platform on which we will deploy.

**Years 2-3:** Aim 2. Make the State Space analysis platform available to the community, including facilities to suggest needed experiments in under-analyzed species. Aim 3. Deploy the first version of the CSNI analysis pipeline for cross species network inference to collaborators (R. Gutierrez, Chile) and other beta testers, including facilities to do parameter optimization using heuristic techniques like simulated annealing and genetic algorithms.

**Years 4-5:** Apply the computational pipeline to infer networks in several crop species for example Corn and Grape. Deploy the full computational CSNI pipeline for cross-species network inference to plant community via CSNI (www.CrossSpecies.org) linked to VirtualPlant, iPlant and a selected workflow platform (e.g. Galaxy). Make all software available as a Webservice.

**PLAN TO INTEGRATE RESEARCH AND EDUCATION**.

 **Cross training of Biologists and Computer Scientist in Systems Biology**. The development of Systems Biology tools in this project has and will involve biologists teaching computer scientists about topics like genetics, experimental genomics, and the computational challenges of analyzing genomic data. We do this informally at our weekly joint lab meetings at which graduate students and post docs from NYU Biology and NYU Courant each present their work to the group. This project involves a team of three resident full time computer scientists working within a biology lab, interacting closely with wet bench biologists. The senior computer scientists (Shasha, Katari and Goldberg) 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 VirtualPLanthas 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 evolutionary 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 (CSNI), 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 (CSNI) 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

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

**Role of senior participants and timeline:**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 (CSNI) 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).

**REFERENCES CITED**

1. Katari, M.S., et al., *VirtualPlant: a software platform to support systems biology research.* Plant Physiol, 2010. **152**(2): p. 500-15.

2. Oinn, T., et al., *Taverna: a tool for the composition and enactment of bioinformatics workflows.* Bioinformatics, 2004. **20**(17): p. 3045-54.

3. Altintas, I., et al., *Kepler: an extensible system for design and execution of scientific workflows.* Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004: p. 423--424.

4. Blankenberg, D., et al., *Galaxy: a web-based genome analysis tool for experimentalists.* Curr Protoc Mol Biol, 2010. **Chapter 19**: p. Unit 19 10 1-21.

5. Gifford, M.L., et al., *Cell-specific nitrogen responses mediate developmental plasticity.* Proc Natl Acad Sci U S A, 2008. **105**(2): p. 803-8.

6. Thum, K.E., et al., *An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in Arabidopsis.* BMC Syst Biol, 2008. **2**: p. 31.

7. Nero, D., et al., *In silico evaluation of predicted regulatory interactions in Arabidopsis thaliana.* BMC Bioinformatics, 2009. **10**: p. 435.

8. Wang, R., et al., *Genomic analysis of the nitrate response using a nitrate reductase-null mutant of Arabidopsis.* Plant Physiol, 2004. **136**(1): p. 2512-22.

9. Gutierrez, R.A., et al., *Insights into the genomic nitrate response using genetics and the Sungear Software System.* J Exp Bot, 2007. **58**(9): p. 2359-67.

10. Gutierrez, R.A., et al., *Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis.* Genome Biol, 2007. **8**(1): p. R7.

11. Gutierrez, R.A., et al., *Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1.* Proc Natl Acad Sci U S A, 2008. **105**(12): p. 4939-44.

12. Warde-Farley, D., et al., *The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.* Nucleic Acids Res, 2010. **38 Suppl**: p. W214-20.

13. Moreno-Risueno, M.A., W. Busch, and P.N. Benfey, *Omics meet networks-using systems approaches to infer regulatory networks in plants.* Curr Opin Plant Biol, 2009.

14. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nature Genetics., 2000. **25**: p. 25-29.

15. Mewes, H.W., et al., *MIPS: analysis and annotation of proteins from whole genomes.* Nucleic Acids Res, 2004. **32**(Database issue): p. D41-4.

16. Poultney, C.S., et al., *Sungear: interactive visualization and functional analysis of genomic datasets.* Bioinformatics, 2007. **23**(2): p. 259-261.

17. Ferro, A., et al., *NetMatch: a Cytoscape plugin for searching biological networks.* Bioinformatics, 2007. **23**(7): p. 910-2.

18. Krouk, G., et al., *High resolution dynamic transcriptome of Arabidopsis roots in response to nitrate: Molecular physiology and predictive modeling.* submitted, 2010.

19. Mirowski, P. and Y. LeCun, *Dynamical Factor Graphs for Time Series Modeling.* Lecture Notes in Artificial Intelligence, 2009. **5782**: p. 128-143.

20. Craigon, D.J., et al., *NASCArrays: a repository for microarray data generated by NASC's transcriptomics service.* Nucleic Acids Res, 2004. **32**(Database issue): p. D575-7.

21. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic Acids Res, 2004. **32 Database issue**: p. D277-80.

22. Ding, X., et al., *A rice kinase-protein interaction map.* Plant Physiol, 2009. **149**(3): p. 1478-92.

23. Popescu, S.C., et al., *MAPK target networks in Arabidopsis thaliana revealed using functional protein microarrays.* Genes Dev, 2009. **23**(1): p. 80-92.

24. Rohila, J.S., et al., *Protein-protein interactions of tandem affinity purified protein kinases from rice.* PLoS One, 2009. **4**(8): p. e6685.

25. de Folter, S., et al., *Comprehensive interaction map of the Arabidopsis MADS Box transcription factors.* Plant Cell, 2005. **17**(5): p. 1424-33.

26. Popescu, S.C., et al., *Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays.* Proc Natl Acad Sci U S A, 2007. **104**(11): p. 4730-5.

27. Gholami, A.M. and K. Fellenberg, *Cross-species common regulatory network inference without requirement for prior gene affiliation.* Bioinformatics, 2010. **26**(8): p. 1082-90.

28. Mutwil, M., et al., *GeneCAT--novel webtools that combine BLAST and co-expression analyses.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W320-6.

29. Yu, H., et al., *Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.* Genome Res, 2004. **14**(6): p. 1107-18.

30. Geisler-Lee, J., et al., *A predicted interactome for Arabidopsis.* Plant Physiol, 2007. **145**(2): p. 317-29.

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

32. Masoudi-Nejad, A., et al., *KEGG bioinformatics resource for plant genomics research.* Methods Mol Biol, 2007. **406**: p. 437-58.

33. Zhang, P., et al., *MetaCyc and AraCyc. Metabolic pathway databases for plant research.* Plant Physiol, 2005. **138**(1): p. 27-37.

34. O'Brien, K.P., M. Remm, and E.L. Sonnhammer, *Inparanoid: a comprehensive database of eukaryotic orthologs.* Nucleic Acids Res, 2005. **33**(Database issue): p. D476-80.

35. Li, L., C.J. Stoeckert, Jr., and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* Genome Res, 2003. **13**(9): p. 2178-89.

36. Tatusov, R.L., E.V. Koonin, and D.J. Lipman, *A genomic perspective on protein families.* Science, 1997. **278**(5338): p. 631-7.

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

38. Barrett, T., et al., *NCBI GEO: mining tens of millions of expression profiles--database and tools update.* Nucleic Acids Res, 2007. **35**(Database issue): p. D760-5.

39. Davuluri, R., et al., *AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors.* BMC Bioinformatics, 2003. **4**(1): p. 25.

40. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Res, 2006. **34**(Database issue): p. D140-4.

41. Gustafson, A.M., et al., *ASRP: the Arabidopsis Small RNA Project Database.* Nucleic Acids Res, 2005. **33**(Database issue): p. D637-40.

42. Lu, C., et al., *Elucidation of the small RNA component of the transcriptome.* Science, 2005. **309**: p. 1525.

43. Michaelewicz, Z. and D. Fogel, *How to Solve It: Modern Heuristics*. 2004: Springer Verlag.

44. Wang, L., et al., *A dynamic gene expression atlas covering the entire life cycle of rice.* Plant J, 2009.

45. Murphy, K. and S. Mian, *Modelling Gene Expression Data using Dynamic Bayesian Networks*, in *Technical report, Computer Science Division*. 1999, University of California and Life Sciences Division, Lawrence Berkeley National Laboratory.

46. Efron, B., et al., *Least angle regression.* Annals of Statistics, 2004. **32**(2): p. 407-451.

47. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 2005. **67**: p. 301-320.

48. Tibshirani, R., *Regression shrinkage and selection via the Lasso.* Journal of the Royal Statistical Society Series B-Methodological, 1996. **58**(1): p. 267-288.

49. Bonneau, R., et al., *A predictive model for transcriptional control of physiology in a free living cell.* Cell, 2007. **131**(7): p. 1354-65.

50. Bonneau, R., et al., *The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo.* Genome Biol, 2006. **7**(5): p. R36.

51. Shimamura, T., et al., *Recursive regularization for inferring gene networks from time-course gene expression profiles.* BMC Syst Biol, 2009. **3**: p. -.

52. Wang, Y., et al., *Inferring gene regulatory networks from multiple microarray datasets.* Bioinformatics, 2006. **22**(19): p. 2413-2420.

53. Reiss, D.J., N.S. Baliga, and R. Bonneau, *Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks.* BMC Bioinformatics, 2006. **7**: p. 280.

54. Chiu, J.C., et al., *OrthologID: automation of genome-scale ortholog identification within a parsimony framework.* Bioinformatics, 2006: p. btk040.