**EXPERIMENTAL PLAN**

**Motivation & Significance**: Our ultimate goal is find a causal genetic network, effectively the circuit diagram regulating the N-assimilatory pathway in plants. By analogy to an electrical network, such a gene regulatory network would enable us to infer the causal relationships between genes, the ultimate goal of Systems Biology. The nitrate assimilation pathway in Arabidopsis is an ideal system to develop methods to model metabolic regulatory networks in higher eukaryotes, owing to its exquisite transcriptional regulation in response to nutrient and environmental factors. The resulting inferred networks can in turn suggest targeted interventions to reduce fertilizer usage with implications for human health, energy and the environment. This work will also illustrate an experimental/informatics approach to the discovery of the causal network for any gene (or potentially any trait) of interest across a wide range of pathways in biology and medicine. In the true spirit of Systems Biology [Gutierrez 2005], our three Aims describe is an iterative cycle of high-through put experimentation (Aim 1), analysis & modeling (Aim 2), which drives new experimentation (Aim 3), as follows:

**Aim 1. Experimental Innovation**: “Network Walking”: Generation of high through-put, high confidence TF🡪target datasets for TF hubs in the N-assimilatory network.

**Aim 2. Computational Innovation**: “Consensus Network Inference”: A pipelined machine learning approach to find causal links in regulatory networks using many kinds of genome-scale data (steady state, time series and TF perturbation.

**Aim 3. Re-fuel the Systems Biology Cycle**: The feedback from analysis to subsequent experiment.Testing new candidate TF hubs and TF cooperation suggested by new/refined network models.

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

**Aim 1. Experimental Innovation: “Network-Walking”: Generation of high through put, high confidence TF🡪target datasets for TF hubs in the N-assimilatory network.**

***Rationale*:** Our network studies have identified TF hubs associated with regulating the N-assimilatory pathway genes in response to inorganic (nitrate) vs. organic-N signals (Glu/Gln).In this aim, we will generate four types of data which will be used to test and refine our network models which will be “learned” in Aim 2 from a consensus analysis of four types of data: (i) time-series transcriptome data, (ii) steady state transcriptome data, TF perturbation studies including (iii) transcriptome and (iv) ChIP seq. For the TF perturbation studies, we will use a “Network Walking” approach, a transient, high throughput approach to perturb the expression of TF hubs associated with N-assimilatory networks, and examine alterations in target responses genome-wide. For selected TFs we will also examine effects of perturbations in planta.

**Nitrogen nutrient sensing and the regulation of the N-assimilation pathway**. To date, our regulatory network models for the N-assimilation pathway model in which inorganic-N (nitrate) signals activate TFs and target genes involved in nitrate uptake, reduction, and assimilation into organic-N (Fig. X). These genes are feedback-repressed by the organic-N products of N-assimilation (Glu/Gln). (This may represent an “energy conservation” mechanism to conserve ATP & reducing equivalents, and carbon skeletons required for N-assimilation, when levels of organic-N are abundant.) The organic-N signals (Glu/Gln) in turn activate target genes involved in converting “reactive” Gln into “inert” Asn, a more carbon-efficient N-transport amino acid used for N-transport/storage (e.g. to seeds) (see Fig. X). Our network models [Krouk 2010, Nero 2009b, Gutierrez 2008] have enabled us to identify TFs involved in mediating the responses to nitrate and organic-N signals (see Fig. 1, Table X), which are the focus of our initial perturbation studies.

**Prioritization of TF hubs for perturbation studies.** We used two different types of network modeling approaches to identify the TFs associated with N-regulation of this N-assimilation network. Time-series transcriptome data and state-space modeling (e.g. Dynamic Factor Graphs- DFG) was used to derive predictions network models for nitrate control of N-assimilation genes [Krouk 2010]. Network predictions – which were validated using out of sample data –was used to prioritize TFs induced “early” and at the top in the hierarchy for experimental testing (e.g. HRS1, HHO1, HHO2, HHO3) (Table 1). By contrast, TF hubs associated with organic-N regulation of N-assimilation pathway (e.g. CCA1, GLK1/2, WRKY1, bZip1) were identified based on TF🡪target correlation (>0.8) and representation of cis-elements [Gutierrez 2008] [Nero 2009a]. Remarkably, the TFs identified to be in nitrate vs. organic-N networks comprise two sub-clades of a single myb gene family, for which one member (HRS1) was also identified to be involved in phosphate signaling [Liu 2009] (Table X). This finding underscores the biological relevance of our approach, and suggests a role for these TF hubs in coordinating responses to nutrient signaling.. This hypothesis is supported by our preliminary results of HRS1 expression in transient protoplast assays (see Aim 1B), and attests to the biological relevance of the transient system. The TFs as prioritized in Table X will be subjected to perturbation studies below. In each subaim, we provide preliminary data using one of these prioritized TFs as a proof-of-principle example for the approach. The data generated in Aim 1 will fuel an iterative cycle of model refinement (in Aim 2), that will identify new TFs for experimentation, leading to model refinement, as part of the iterative Systems Biology cycle.

**Aim 1A. Generation of time-series transcriptome datasets for organic-N signaling networks.** In this subaim, we will generate time-series transcriptome data (0, 3, 6, 9, 12, 15, 20, 25, 30 min) for plants treated with conditions shown to elicit an organic-N response (growth on 1mM nitrate, treatment with 40mM ammonium/nitrate vs. control KCl) [Gutierrez 2008]. This type of fine-scale time-series data enabled us to identify >200 new nitrate-regulated genes compared to previous steady state studies (e.g. at 20 min) [Wang 2004], and was used to generate dynamic, predictive networks [Krouk 2010]. By contrast, our current organic-N network models were built using steady-state transcriptome data analyzed in the context of the Arabidopsis multinetwork [Gutierrez 2007, 2008]. While those network models enabled us to identify and validate master TF genes involved in organic-N regulation of the N-assimilation pathway (e.g. CCA1) [Gutierrez 2008], we cannot use these static network models to predict network dynamics or responses under untested conditions, an important goal of Systems Biology. The time-series data which will fuel a network inference pipeline in Aim 2, and will enable us to identify organic-N regulated TF hubs induced early in the cascade for TF perturbation studies below.

**Aim 1B. “Network Walking”: Generation of high through put, high confidence TF🡪target data.** In this subaim, we will use TF perturbation to validate predicted TF targets genome-wide. This data will be used to functionally test and refine our gene regulatory network inference pipeline described in Aim 2. Our TF-expression system (reviewed in Progress) employs transient transformation of Arabidopsis protoplasts with TF-GR (glucocorticoid receptor) fusions, and inducible nuclear re-localization of TFs by dexamethasone (DEX) treatment. Transcriptomic analysis then enables us to investigate the network targets of a given TF genome wide. This system is rapid and this allows the discovery of network targets any given TF in less than 2 weeks. Importantly, it also overcomes problems encountered with TF redundancy encountered in reverse genetic approaches (see Aim 1D). This DEX-inducible approach also has several advantages over Chromatin-IP (ChIP) as follows: (i) Chromatin-IP (Chip-Seq) can confirm protein:DNA binding, but it does not guarantee functional regulation [Eilers and Eisenman Genes & Dev. 2008. 22: 2755-2766], and (ii) the DEX-inducible TF system allows one to identify the effect of regulatory components that may not bind directly to DNA [ Lee, J, He, K, Stolc, V, Lee, H, Figueroa, P, Gao, Y, Tongprasit, W, Zhao, H, Lee, I, and Deng, XW, *Analysis of transcription factor HY5 genomic binding sites revealed its hierarchical role in light regulation of development.* Plant Cell, 2007. **19**(3): p. 731-749.] Technically, our approach is an adaptation of the original technique [ Sablowski, RW and Meyerowitz, EM, *A homolog of NO APICAL MERISTEM is an immediate target of the floral homeotic genes APETALA3/PISTILLATA.* Cell, 1998. **92**(1): p. 93-103.] combined with transcriptome analysis described in [ Hanson, J, Hanssen, M, Wiese, A, Hendriks, MM, and Smeekens, S, *The sucrose regulated transcription factor bZIP11 affects amino acid metabolism by regulating the expression of ASPARAGINE SYNTHETASE1 and PROLINE DEHYDROGENASE2.* Plant J, 2008. **53**(6): p. 935-949.] [Wang, Y, Joshi, T, Zhang, XS, Xu, D, and Chen, L, *Inferring gene regulatory networks from multiple microarray datasets.* Bioinformatics, 2006. **22**(19): p. 2413-2420]. Our approach pushes the DEX-concept further, using N-treatments to precondition cells as described below.

***The “Network Walking*** ***Approach”*:** Conceptually, our DEX transient assay approach to rapidly validate TF🡪target interactions genome-wide consists of the following components; i) transient expression of a 35S::GR-TF fusion in protoplasts using FACS selection of successfully transfected cells (based on RFP selection on vector), ii) use of an inducible dexamethasone (DEX) system to artificially control the entry of the TF into the nucleus, iii) DEX induction of TF nuclear import is performed +/- cycloheximide (CHX), to identify primary and secondary targets of the TF (hence the name “Network Walking”, and iv) the use of nitrogen pre-treatments (e.g. prior to DEX induction of the TF) (see Fig. X). This last step allows us to identify the primary targets of the TF when acting alone or in combination with other TFs that are N-regulated transcriptionally or post-translationally. A global view of the DEX approach is summarized in Fig. X, and described more fully in Progress and in Bargmann et al 2012.

***As proof-of-principle***, we demonstrate that the DEX-inducible activation of a TF (HRS1) enabled us to uncover network targets with relevance to whole plants. HRS1 (At1g1330), a top scoring TF hub of our “learned” nitrate-regulatory network [Krouk 2010], was previously shown to be involved in phosphate signaling [Liu 2009, Journal of Integrative Plant Biology]. We hypothesize that HRS1 may also be involved in nitrogen signaling, as HRS1 is induced within 9 min of NO3- treatment in the dynamic inferred network, and it is predicted to be the top most controlling and top most controlled TF hub controlling the N-assimilation pathway [Krouk et al 2010]. Our initial analysis of transcriptome data from 35S::GR-HRS1 expressed in the transient DEX-protoplast system, revealed four distinct gene clusters influenced by a combination of HRS1 nuclear import (by DEX) and nitrate-treatment (Fig. X). Cluster 4 genes, which are primary targets of HRS1 (+DEX+CHX) and activated by DEX only in the *presence* of NO3- , show significant overrepresentation of phosphate transport genes (p-val 8.14 E-6). Thus, the targets of HRS1 observed in the transient protoplast system, recapitulates the phosphate signaling phenotype observed in 35S::HRS1 overexpressing plants [Liu 2009]. These results support the notion that TF->target relationships validated in the rapid & transient protoplast system, are relevant to whole plants. [Kranthi-This paragraph needs a description of what happened when we did the preliminary results on HRS1 gene +DEX and +/-CHX to show that direct targets are a subset of indirect] We will subject the prioritized TFs to the perturbations in this protoplast system (+/- DEX, +/-CHX, +/-N), and Table X lists our progress. Conditions for +/-N treatments will be tailored to either nitrate signaling (growth in ammonium succinate, treatment with nitrate) or organic-N signaling (growth in 1mM nitrate, and treatment with 40mM ammonium/nitrate). Transcriptome analysis will identify gene targets, clusters and biological processes controlled by specific TFs. This TF perturbation transcriptome data will also fuel the Consensus Network Inference machine learning pipeline in Aim 2 ..

**Aim 1C. Genome-wide validation of TF (protein)🡪targets (DNA) using Chip-Seq.** In this subaim, we will determine use Chip-Seq to identify TF->target interactions in the transient DEX-TF expression system, and compare these results to gene activation (transcriptome). This will enable us to determine and compare (i) network targets to which the TF is bound, and (ii) network targets that are transcriptionally activated in response to TF activation (by DEX). Because 35S:: GR-TF (glucorticoid receptor) constructs are expressed in this transient assay system, we can use anti-GR antibodies to bind to the TF-GR fusion protein. To demonstrate that we can perform ChIP and transcriptome studies and from the same sample, we performed a ***proof-of-principle*** study using a 35S::GR-BZip1 construct (Table X). As our transient system for TF expression typically results in 8-12,000 successfully transfected protoplasts (sorted by FACS), we adapted the micro-ChIP protocol from [Dhal and Collas, 2008], which requires a relatively small number of cells. After transformation with the Gateway-TF construct and FACS cell sorting (to select successfully transformed cells), protoplasts were treated with 1% formaldehyde for 10 min and then quenched with 100mM glycine for 5 min. W5 buffer was then added to wash the cells before pelletting by centrifugation at 2,500 rpm for 2 min. Cross-linked cells were snap-frozen in liquid N and stored at -80C. The protocol for processing 1,000 cells was then followed. Anti GR (sc-1002) antibodies from Santa Cruz Biotechnology Inc. were used to capture the GR::bZip1:DNA complexes, and ChIP DNA was purified after reverse cross-linking using the MiniElute Reaction cleanup kit from QIAGEN. ChIP DNA was prepared for illumine sequencing as follows. Pair-end ChIP-seq libraries were constructed in accordance with the Illumina ChIP-seq sample prep guide (Illumina, San Diego, CA), with the following modifications to facilitate library preparation from low amounts of starting DNA (~1ng): (i) adaptor oligo mix was further diluted by 3-fold to maintain a proper adaptor to DNA insert ratio; (ii) Solid Phase Reversible Immobilization (SPRI) magnetic bead-based technology was used to size-select the library after adaptor ligation, instead of the common agarose gel size selection, to minimize DNA loss. (iii) SYBR gold agarose gel size selection was performed after PCR enrichment to remove adaptor self-ligation product from the library. Libraries were separately constructed for the immunoprecipitated DNA and the input DNA (the DNA without immunoprecipitation) to serve as control. The libraries were sequenced on the Illumina GAIIx platform. The obtained sequence reads were filtered for quality and trimmed to remove adapter sequences. These filtered reads were then aligned to the TAIR10 assembly of the Arabidopsis thaliana genome. Genomic regions that are significantly enriched in the immunoprecipitated sample relative to the input DNA were identified using a peak-calling algorithm called QuEST [Valouev et al., Nature methods 5, 829-834 (2008)]. The genes immediately adjacent to the identified peak positions are considered putative targets of bZIP1. KRANTHI- RESULTS OF ANALYSIS bZIP1???????

**Aim 1D. TF perturbations in whole plants: T-DNA mutants and 35S::Over-expressors.**  In this aim, we will identify TF🡪targets using TF perturbations in whole plants. Owing to functional redundancy of TFs [Chen HW, Bandyopadhyay S, Shasha DE, Birnbaum KD, “[Predicting genome-wide redundancy using machine learning”.](http://www.ncbi.nlm.nih.gov/pubmed/21087504). BMC Evol Biol. 2010 Nov 18;10:357] [Cutler S and McCourt P (2005) Dude, Where’s my phenotype? Dealing with redundancy in Signaling Networks. Plant Physiology (2005) v. 138, pp558-9]., we have failed to see a molecular phenotype (e.g. altered expression of a predicted target gene) in T-DNA mutants for most of the prioritized TFs listed in Table X. For example, while 35S::CCA1 plants show alterations in predicted target genes [Gutierrez et al 2008], we fail to see molecular phenotypes in single mutants, but have seen alterations in regulation of predicted targets in double mutants (e.g cca1/lhy1 and glk1/glk2). In rare cases, misregulation of predicted TF🡪target genes were observed in a single TF T-DNA mutant. For example, our network models [Gutierrez 2008] predict that the TF WRKY1 (At2g04880), acts like a “toggle switch”, to induce expression of genes involved in nitrate reduction and assimilation into Gln, while simultaneously repressing expression of genes involved in converting Gln to Asn (Fig. X). Indeed all three WRKY1 T-DNA insertion mutants (SALK\_016954; SALK\_136009; SALK\_070989) revealed decreased expression of NIA2 and NRT2.1 (targets of WRKY1 activation), and increased expression bZIP1🡪ASN1 (both targets of WRKY1 repression), as predicted by the network model (Fig. X), while the opposite expression patterns are found in the DEX transient-assay system where WRKY1 is over-expressed using the approach in Aim B). We will perform transcriptomic analysis on T-DNA mutants for TFs where Q-PCR analysis shows mis-regulation of a predicted target gene. Such T-DNA mutants in TFs will be useful to test models for TF-TF cooperation (see Aim 3B). We will also create stably transformed 35S::GR-TF plants, for TFs prioritized based on transient protoplast assays. These stably transformed plants will allow us to explore how TF->network targets vary across organs, or developmental time-points – where the context of TF partners may affect network targets.

**Data integration and analysis**: The four sets of data generated in Aim 1- (i) time-series transcriptome, (2) steady state transcriptome, and data from TF perturbations (iii) transcriptome, and (iv) ChIP-Seq will be combined into a single pipeline for machine learning (Aim 2), so that predicted regulatory network connections are derived based on a consensus amongst the datasets.

**Expected results, limitations and alternate approaches (Aim 1)**: The overall goal of Aim 1 will be to produce distinct datasets to validate TF🡪target relationships predicted in the regulatory network models in Aim 2. For each subaim, we provide a working example to demonstrate the feasibility of the approach, so we do not anticipate technical problems in data generation.

**Aim 2. Computational Innovation: Consensus Network Inference: A pipelined machine learning approach to find causal links in regulatory networks using many kinds of genome-scale data**

***Rationale***: Our ultimate goal is to model a causal genetic network, effectively the circuit diagram underlying the regulation of genes in the N-assimilatory pathway. To date, we have used a machine learning approach (Dynamic Factor Graph State Space Modeling, DFG for short) to generate a predictive network model for nitrate control of N-assimilatory pathway based on time-series data from wild-type plants [Krouk 2010]. In this aim, we develop methods and approaches to refine and improve the predictive power of such networks, by feeding our network learning algorithms experimental data derived from cells (or plants), in which we have perturbed expression of the TF and monitored effects on target genes. The transcriptome data generated from these TF perturbation experiments will support either primary or secondary TF🡪target relationships, some of which will also be validated by ChIP-seq. In an iterative cycle of experimentation – computation- experimentation, the refined network models derived from this pipeline in Aim 2 will in turn suggest new TFs for experimentation (in Aim 3), refueling the iterative systems biology cycle of model building, experimentation and model refinement.

**Approach**: To improve the predictive power of our networks, we will create a Consensus Network Inference pipeline to generate a predictive network model that makes use of four types of genomic data that are available to us, as generated in Aim 1: (i) steady state transcriptome data (e.g. N-treatments); (ii) time-series data (e.g. expression over time), (iii) TF perturbation transcriptome data (e.g. DEX data and T-DNA), , and (iv) TF🡪target binding validated by ChIP seq. No single algorithm is best for machine learning using all four kinds of data, so we will use several algorithms together in a Consensus Network Inference pipeline described below.

**The Network Model**: The causal network will consist of nodes that are genes and edges between genes that are labeled with coefficients. A positive coefficient corresponds to an inductive edge. A negative coefficient corresponds to a repressive edge. For a given target gene Z, these coefficients will be reflected in the form of an equation gene Z = c1\*A + c2\*B + c3\*C …. , where A, B, and C are expression levels of transcription factor genes. So, if c1 is positive, then gene Z will rise (increase in expression, because Z would be positive) as gene A rises (i.e., A to Z would be an inductive edge) assuming the other genes are held constant. If c2 were negative, then an increase in expression of B would cause a decrease in Z, all else being equal. Thus, B to Z would be a repressive edge. Typically, machine learning algorithms to infer such equations will include a regularization factor that will limit the number of additive products on the right hand side to three or four, but this still means that even for 100 genes, our algorithms would have to find the proper 400 edges out of a possible 5,000 and assign coefficients to those edges. Regularization is a form of parsimony: we want to find the simplest model that explains the behavior. Simpler models tend to be more robust to noise because they avoid overfitting. In fact, as part of our quest for simplicity, we prefer “linear” models to quadratic ones (in which there would be terms of the form d\*B\*C, where B and C would be gene expressions). Surprisingly, the resulting models explain behavior well. A model explains behavior if it can predict the state of one gene given the state of other genes at the same or previous time points. For example, we used this approach to successfully predict regulatory edges in a network generated from time-series data in nitrate-treated roots [Krouk et al 2010].

**Building a Consensus Network Inference Pipeline of Machine Learning Algorithms to integrate learning from distinct data-types.**

One reasonable approach to combining multiple forms of evidence to derive network edges is to establish a machine learning pipeline (Fig. X, figure dfg.eps) that analyzes different kinds of data using different algorithms. In one such pipeline, the MCZ algorithm (short for the Median Corrected Z-score method [Greenfield, A., Madar, A., Ostrer, H., Bonneau, R.: DREAM4: Combining Genetic and Dynamic Information to Identify Biological Networks and Dynamical Models. PloS one (2010]) is used for both the steady state and mutation data, followed by the DFG algorithm (short for Dynamic Factor Graph used for time-series data [Krouk,G.,Mirowski,P.,LeCun,Y.,Shasha,D.E.,Coruzzi,G.M.:Predictivenetworkmodeling of the high-resolution dynamic plant transcriptome in response to nitrate. Genome Biology 11(12), R123 (2010]). In this pipeline, the MCZ algorithm assigns initial weights to certain edges that are then refined by the DFG algorithm. The basic idea of the MCZ algorithm is that if gene \[x\_j] influences \[x\_i], then perturbing \[x\_j] should change the value of \[x\_i] in a significant way. We will measure the significance based on the number of standard deviations from the median value of \[x\_i] over all non-perturbation experiments.

In preliminary work, done for a book we have written called “Network Inference in Molecular Biology”, [Jesse Lingeman and Dennis Shasha 2012 *Network Inference in Molecular Biology -- a hands-on framework*, Springer 2012], we show using data from the DREAM benchmark (Dialogue for Reverse Engineering Assessments and Methods) [Schaffter, T., Marbach, D., Floreano, D.: GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics 27(16), 2263–2270 (2011)] that the MCZ🡪DFG pipeline is better than using any single machine learning algorithm by itself.

Alternatively, we can posit a collection of algorithms that each predicts edges and then use a “voting” approach to determine the best one. In that scheme, illustrated in the figure consensus.eps, in addition to the Median-Corrected Z-score and Dynamic Factor Graph algorithms, we use the NIR (Network Identification by Multiple Regression [Gardner,T.S.,DiBernardo,D.,Lorenz,D.,Collins,J.J.:Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. Science 301(5629), 102–105 (2003)]) and CLR (Context Likelihood of Relatedness [Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., Gardner, T.S.: Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biology 5(1), e8 (2007)]) on steady state data. We also use BANJO (Bayesian Inference with Java Objects [Yu,J.:AdvancestoBayesiannetworkinferenceforgeneratingcausalnetworksfromobservational biological data. Bioinformatics 20(18), 3594–3603 (2004)]), Time-Delay ARACNE [Zoppoli, P., Morganella, S., Ceccarelli, M.: TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. BMC Bioinfor- matics (2010)], and Inferelator [Greenfield, A., Madar, A., Ostrer, H., Bonneau, R.: DREAM4: Combining Genetic and Dynamic Information to Identify Biological Networks and Dynamical Models. PloS one (2010)]. Because each machine-learning algorithm ranks the TF🡪Target edges, we “weight” each vote for a TF🡪target edge depending on its rank in each of these programs. Highly ranked edges acquire a weight close to 1, and lowly-ranked edges acquire weights near 0, where the weight drops off exponentially. Algorithms on steady state data from wild-type and pertubation data (e.g. DEX-TF and T-DNA) assign greater or lesser weights to edges. Those weights are refined by one of the time-series based algorithms (e.g. Inferelator, Dynamic Factor Graphs, or Time Delay ARACNE).

**Preliminary Results:** To test the value of pipelines for predicting edges in a network, we compare the results of our previous Dynamic Factor Graph (DFG)/State Space Modeling Approach [Krouk,G.,Mirowski,P.,LeCun,Y.,Shasha,D.E.,Coruzzi,G.M.:Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate. Genome Biology 11(12), R123 (2010] built on time-series data, with a pipeline “weighted” approach to network inference. Our criterion for quality is how well the resulting network predicts out-of-sample data. In our previous paper [Krouk 2010], we showed that using the training data consisting of time-series data from the time of a perturbation, 3 minutes later, 6 minutes later, 9, 12, and 15, we were able to use a Dynamic Factor Graph Approach to predict the direction of expression change of 76 genes associated with the N-assimilation network between 15 and 20 minutes 74% of the time correctly. By contrast, a naïve trend forecasting method, which predicted the direction of expression change of genes between 15 and 20 minutes as being the same as between 12 and 15 minutes, was correct only 52% of the time, having a prediction accuracy marginally better than chance.

Using the new data from perturbation of TF function (using Dex, in Aims 1B and 1C) and the steady state data along with the time series data from wild-type plants, we can predict z% of the gene directions correctly. Using only the steady state data along with time series data, we can predict z’% of the gene directions correctly. Using the pipeline approach, we obtain z’’% of the gene expression directions correctly. [this paragraph should be rendered as a table. Gloria could you set one up that has five columns with headers: Steady State, Genetic Perturbation, Time Series, Prediction accuracy, p-value]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Steady State | TF perturbation | Time Series | Prediction Accuracy | P-Value |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**Intended Approach**: Based on our preliminary results, we will iterate the following steps: (i) identify the genes relevant to nitrogen treatments (on the order of a few hundred) based on steady state and time-series transcriptome data resulting in just over 200 genes (including nitrogen-regulated members of the 66 genes in the N-assimilatory pathway, and TFs), (ii) develop a pipeline Consensus Network Inference approach based on steady state, TF perturbation experiments (e.g. DEX experiments or T-DNA mutants), and time-series data.. This will result in a regulatory network of the genes controlling the N-assimilation pathway. That regulatory network will suggest new genes on which to try DEX experiments (e.g. putative TF hubs) (in Aim 3). Those validations will in turn be used to refine a new network in an iterative approach. As usual, our criterion of goodness (for network predictions) will be the ability to predict well on out-of-sample data, both missing time points and missing TF perturbations. Ultimately, we would like to learn the network model well enough that we choose the next DEX experiment to try based on which TF🡪target relationships will be the hardest for our learned network to predict.

**Data integration and analysis**: Aim 2 integrates the four sets of data generated in Aim 1- (i) time-series transcriptome, (2) steady state transcriptome, and data from TF perturbations (iii) transcriptome, and (iv) ChIP-Seq into a single pipeline for machine learning (Aim 2). The “learned” network connections are based on a consensus amongst the datasets. We will test the various approaches for their ability to correctly predict network states using out of sample data, as exemplified for the preliminary study shown in Table X.

**Expected results, limitations and alternate approaches (Aim 2)**: The overall goal of Aim 2 will be to develop a machine learning pipeline to integrate all genomic data (from wt and TF perturbation studies) to generate a predictive network model. We provide a working example which demonstrates that combining these data sets improves the accuracy of our predictions. Our accuracy of predictions is performed using out-of-sample data, and we will test a number of machine learning methods to identify which combination is best. Dennis- can you anticipate any problem issues or alternative approaches? The main possible problems have to do with the fact that the number of possible edges is much larger than the number of experimental data points we have. We nevertheless believe that the approach will work because we have already seen promising results using less data and a single algorithm.

**Aim 3. The Systems Biology Cycle: The feedback from analysis to subsequent experiment. Testing new candidate TF hubs and TF cooperation predicted from refined network models.**

**Rationale**: In the iterative nature of the Systems Biology cycle [Gutierrez 2005], the network models from Aim 2 will in turn predict new TF perturbation experiments to perform in Aim 3. These include TF hubs to test using the high throughput approaches described in Aim 1B and 1C. These may include, for example, new TFs for which TF🡪target relationships are the hardest for our learned networks (in Aim 2) to predict. Our network models in Aim 2, will also give suggestions for TF partners that may act in concert (or antagonistically) on genes in the N-assimilatory network. Thus, in Aim 3, we will use high throughput approaches to test the cooperation of TFs, and these results will in turn feed back to refine the network models. [Gloria we need a figure showing this cycle of experiment to modeling to more experiments to more modeling].

**Approach**: We will test TF🡪target relationships for new candidate TF hubs identified in Aim 2, using the transient DEX-system for TF activation described in Aim 1B. We will also use two complementary approaches to test for TF1-TF2 cooperation: (i) Co-expression of TF1 and TF2 in the transient protoplast DEX-system, (ii) Testing of TF cooperation using a genetic approach in the transient expression system. For the latter approach, transcriptomic data from TF1 subjected to DEX-activation in wild-type protoplasts, will be compared to activation of TF1 in protoplasts made from tf2 T-DNA mutants. Changes in target gene activation between these transcriptomes, will suggest TF1-TF2 dependencies in wild-type (e.g. including additive or synergistic effects on target genes in the networks), that are lost in the tf2 mutant background will suggest TF1/2 cooperation (additive) or interaction (synergistic) effects on targets in the network.

**Aim 3A. Perturbation studies for new candidate TFs**: In this subaim, refined network models derived in Aim 2, will be used to identify new TFs for targeted for perturbation studies as described in Aim 1B and C. For example, new time-series data for organic-N signaling generated in Aim 1A, should identify new TFs that are activated early in the cascade (e.g. within 3-9 min). The prediction that we will find new TFs involved in organic-N signaling is based on our time-series analysis of nitrate-responsive networks where we identified >200 new nitrate-responsive genes in a 3-9 min window [Krouk et al 2010], not previously observed in steady state studies (at 20 min) [Wang 2004]. The kinetic network models will also identify which TFs may have the most influence based on their out-degree (“hubbiness”) in the network, or based on the phenotypic importance of their targets. These will be among the new candidate TFs to target for perturbation studies.

**Aim 3B. Testing TF cooperation in the N-assimilatory network: Co-expression of TF1 and TF2 and genetic perturbations.** Our refined network models generated using the pipeline in Aim 2, will suggest TF pairs that may cooperate to regulate key target genes in the network. We will use two complementary approaches to test how selected TF pairs may cooperate in the regulation of target genes in the network. In one approach, we will co-express the two TFs in the transient DEX-protoplast system using a single vector, and compare regulation of target genes to those observed for expression of each single TF. Target genes regulated differently in the TF1/TF2 double expression vector (e.g. additive or synergistic effects on target gene expression), when compared to expression of each TF alone, will be evidence for cooperation. In a complementary high through-put genetic approach, we will express TF1 in protoplasts made from a tf2 T-DNA mutant using the DEX-transient activation system described in Aim 1B. Evidence for TF1/2 cooperation, will be alterations in target regulation by TF1 in a tf2 mutant background, compared to wild-type.

**Co-expression of TF1 & TF2:** In order to co-express two TFs (TF1 and TF2) in the same cell, we are using the MultiSite Gateway**®** Technology that allows one to simultaneously clone multiple DNA fragments in a defined order and orientation. To generate the co-expression construct, TF1 cDNA, TF2 cDNA and the “GR cassette” (35S promoter - GR sequence- 3’ terminator) were cloned in appropriate Gateway donor vectors, and allowed to recombine with each other and with the destination vector pBob11 [Bargmann BO, Birnbaum KD.

Plant Physiol. 2009 Mar;149(3):1231-9]. This resulted in the assembly of two consecutive expression cassettes 35S::GR:TF1 and 35S::GR:TF2 (Fig. X). Expressing both TFs from the same vector has a few advantages: (i) it eliminates the problem of dealing with multiple vectors that could have different transfection efficiencies due to their inert TF size; (ii) when multiple vectors are transfected, each cell is likely to contain a random number of copies, while a single co-expression vector will result in homogeneous transgenic content of protoplasts. Where target gene expression is synergistically regulated in the TF1/2 vector, compared to either single vector will identify targets of TF cooperation.

**Perturbation of TF1 activity in a tf2 mutant background**: As a complement to the multisite vector approach, we will explore TF1/2 cooperation by transiently expressing the single 35S::GR:TF1 in protoplasts made from a T-DNA mutant in tf2, compared to expression in wild-type protoplasts (where TF2 is present). Loss, or change in target gene regulation by TF1 in a tf2 mutant protoplast (compared to wild type), would indicate cooperation of TF1 and TF2 (in wild-type). This would provide functional evidence for targets of TF cooperation- that could involve direct or indirect TF interactions. We will use this genetic approach for TFs that show a molecular phenotype in the single T-DNA mutant (e.g. WRKY1 T-DNA). For TFs that are functionally redundant, we will use double mutants (e.g. glk1/2, cca1/2) to explore TF cooperation in transcriptional networks. For example, as glk1/2 are in the same myb family as HRS1, we will express HRS1 in a glk1/2 double mutant background, and determine the effects on target gene activation or repression, compared to activation of HRS1 in wild-type protoplasts. These perturbation results will be fed into the machine learning pipeline (Aim 2), to refine our network predictions.

**Data integration and analysis**: The transcriptome and ChIP-seq data from TF perturbation experiments in Aim 3A will feed back into the machine learning pipeline in Aim 2, to refine network predictions. In order to analyse/integrate the transcriptome data supporting TF1/2 interactions (Aim 3B) into our machine learning pipeline in Aim 2, we will have to modify the equation for TF regulation of target gene expression to include TF interactions (e.g. to quadratic equation) whenever the interaction data shows synergistic effects compared to the single TF data. (DENNIS- IS THIS TRUE? DO YOU WANT TO REWRITE THIS?)

**Expected results, limitations and alternate approaches (Aim 3)**: The results of Aim 3 will enable us to test predictions for new TFs identified in Aim 2, and to validate target edges for TFs that the machine learning program is not very good at predicting. It will also enable us to test how TF interaction or cooperation affects target genes in the network, by studying whether two TFs affect target gene expression in an additive or synergistic fashion. The results from Aim 3, will feed back to Aim 2, to further refine our network models.

**TIMELINE**

**Postscript**: For TFs to validated to affect the N-assimilatory pathway, we will collaborate with Dr. Gabriel Krouk & Sandrine Ruffel at the INRA Biology & Biochemistry of Plants on studies related to N-use efficiency. Using N15 as a tracer, the rate and amount of NO3 incorporation into total N can be measured/dry weight. We are currently testing several TFs for which changes in N-assimilation target genes are observed in the mutants or transgenics (e.g. WRKY1 T-DNA, 35S::CCA1). These collaborative studies between our genomics lab and a world-renouned plant physiology lab, will be the ultimate test of our ability to use system biology approaches to predict targeted interventions in this pathway, with significance to plant biomass and N-use efficiency.