**(a) SIGNIFICANCE.** Modeling predictive Gene Regulatory Networks (GRNs) from comprehensive genomic data is both a major goal *and* challenge in fundamental and applied research in [Systems Biology Albert, R., and Barabasi, A.L. (2002). Statistical mechanics of complex networks. Rev. Mod. Phys. 74: 47-97.]. In this grant, our overall goal is to infer a causal genetic network, effectively the circuit diagram underlying the metabolic regulation of the N-assimilation pathway in plants, to predict network states under untested conditions and in response to perturbation, a holy grail of Systems Biology. Computational inference of biological networks is an inherently difficult problem, mainly due to limitations of genomic data-types (e.g. presence of too few time-point measurements, many genes, measurement errors and noise) [Jaeger and Monk 2010]. However, progress has been made using machine learning approaches [ De Smet and Marchal (2010) Nature Reviews: Microbiology, vol 8, p 717-729] to predict network states under untested conditions in microbes [Bonneau Cell 2007] and yeast [REFS], as well as in higher eukaryotes including animals [REF???] and plants [Reka Albert (2007), The Plant Cell November 2007 vol. 19 no. 11 3327-3338 Network Inference, Analysis, and Modeling in Systems Biology] [Krouk 2010]. In the current cycle of this NIH grant, we applied a machine learning approach to time-series transcriptome data to model how the N-assimilatory network is controlled in response to nitrate-signaling, and these learned networks have enabled us to accurately predict gene expression states on “left-out” experimental data [Krouk 2010]. In this renewal, we propose to build on and improve the predictive power of our inferred networks, by developing *NetWalk: An iterative high throughput experimental/machine learning approach to infer causal networks.* The experimental power of this approach relies on a high-throughput TF perturbation system to rapidly validate TF🡪 targets genome-wide. This data will in turn support and validate a new network inference approach, whose novelty resides in integrating diverse genomic datasets from time-series, steady state, and TF-perturbation experiments into a single pipeline. A limitation of current machine learning frameworks, is that they can only analyze each data type separately. We show in preliminary studies, that integrating analysis of diverse genomic data in a single NetWalk pipeline improves prediction power, compared to separate analyses. Broadly applied, the NetWalk pipeline will illustrate a combined experimental/informatics approach to discover causal networks for any gene, metabolic pathway, process, or trait, with applications across a wide range of problems in biology, medicine and agriculture. In our example, the inference of regulatory networks controlling the N-assimilation pathway in plants can suggest targeted interventions to reduce nitrogen fertilizer use, with implications for health, energy and the environment.

**(b) INNOVATIONS** from a close collaboration between Biologists and Computer Scientists at NYU Courant.

**Experimental: NetWalk Targets. A high-throughput method to validate TF🡪network targets.**  This approach provides a systems view of a TF function *in vivo*, based on transient TF perturbation*.* The system involves TF overexpression, coupled with inducible TF nuclear import, to identify effects on TF targets genome-wide. Its advantages are: i) *High throughput*: acceleration of TF perturbation studies to weeks, compared to stable transgenics (~4-6 mos), ii) *TF Redundancy*: transient overexpression and TF activation overcomes problems of TF redundancy encountered in T-DNA mutants, iii) *Validation*: TF🡪 targets can be validated genome-wide using transcriptome *and* ChIP-seq approaches. The TF🡪target data generated in Aim 1, will train and validate our the NetWalk pipeline inference approach in Aim 2, as below.

**Computational:** **NetWalk Inference. A computational pipeline for network inference combining multiple data-types.** Current network inference methods are limited in that they can each only accept specific transcriptome data-types (e.g. steady state, time-series, or TF mutant) Bonneau (2007) Cell] [ Bonneau Genome Biol 2006] [Wang Y,. Bioinformatics 2006] [Shimamura. BMC Syst Biol 2009]. This road-block to data integration prevents the identification of “high confidence” TF🡪target edges required to train and validate machine learning approaches. To overcome this problem, we will develop a computational network inference pipeline with the capacity to “learn” from multiple genomic data types (e.g. time-series data, steady state data, and TF perturbation data), to validate and identify “high confidence” edges in our networks. This pipeline will combine the DFG algorithm (Dynamic Factor Graph), a form of “state space” analysis we used previously on time-series transcriptome data [Krouk et al 2010], with the MCZ algorithm (Median Corrected Z-score method), a simple but successful approach to infer the effects of TF perturbations on targets using steady state mutant data [Greenfield & Bonneau]. Preliminary results suggest that this new pipelined approach to network inference will increase the quality of our predictions, as judged using left-out data. As the NetWalk approach is iterative, the networks inferred from this new computational pipeline in Aim 2, will drive a further round of experimentation, on new TFs and TF interactions in Aim 3, in a true Systems Biology cycle [Gutierrez 2005].

**Application: N-assimilatory networks.** The N-assimilatory pathway is well-suited for modeling regulatory networks, owing to its exquisite transcriptional regulation by nutrient cues, and its importance to plant growth and development [Ruffel S, 2010]. Our studies suggest that the N-metabolite inputs and outputs of the pathway (nitrate vs Glu/Gln) also serve as N-signals, to regulate pathway genes according to N-supply and demand [Gutierrez 2008] [Krouk 2010] [Ruffel 2011]. Using transcriptome to model regulatory networks [Krouk 2010 [Gutierrez 2008] enabled us to identify TFs that mediate this N-signaling (Fig. X). In this model inorganic-N (nitrate) induces genes involved in nitrate uptake, reduction and assimilation into Glu/Gln (Fig. X). The organic-N products, in turn feed back repress genes involved in Gln synthesis, and activate genes involved in synthesizing Asn, an inert carbon-efficient amino acid used to transport/store N (e.g. in seeds). This may represent an “energy sensing” mechanism to save ATP, reducing equivalents, and C-skeletons when levels of organic-N are abundant, and there is support for this hypothesis [Baena-Gonzales]. We have shown that transcription of key genes in the N-assimilation pathway (GLN1 or ASN1) affects N-use in transgenic plants [Oliveira (2002)][Lam, (2003)], thus, our inferred network models may predict targeted interventions to alter N-use efficiency, a goal of significant import to agriculture, energy, and the environment.

**PROGRESS REPORT**: ***Overview:*** During the current cycle, we have made significant advances in building predictive network models of N-assimilatory networks. In the previous round, network models built using steady-state transcriptome data analyzed using an Arabidopsis multinetwork [Gutierrez 2007, 2008] uncovered TFs associated with organic-N regulation of the N-assimilatory network [Gutierrez 2008]. During the current funding cycle, we used a machine learning approach to advance our network models to become dynamic and to predict network states under untested conditions – the ultimate goal of systems biology. To enable this, we analyzed time-series data (Aim 2B) using state-space modeling, to infer networks that were validated *in silico* (e.g by predicting gene expression states in an experiment that wasn’t used in the inference of the network, and then comparing our predictions with the data from that “left-out” experiment), and experimentally by mutation of a TF hub [Krouk 2010]. To accelerate the experimental validation of our TF🡪 target predictions (Aim 2A), we developed a high through-put system to perturb (e.g. induce) TF function and monitor changes in the expression of its targets [Bargmann et al, 2012, submitted]. This transient TF perturbation system successfully uncovers changes in expression of direct TF targets, as validated using ABI3, a well-studied TF for which network targets are known [Vernoux 2011]. The transient TF perturbation system can significantly accelerate our studies, and overcome problems we have encountered with TF redundancies, and we are currently using it to validate other TF hubs in the N-assimilatory network that respond to nitrate [Krouk et al 2010] or organic-N [Gutierrez 2008] signals (Aim 1). Finally, our studies of post-translational responses to N-signals uncovered miRNA:TF pairs regulated in response to nitrate and Glu/Gln sensing, and were conducted collaboratively as part of an NIH Fogarty Award t0 R. Gutierrez [Vidal 2010].

***Note****: This Progress Report specifically highlights progress most relevant to our Research Strategy.*

**Prior Aim 1. Test hypotheses for TFs controlling targets in N-assimilation. *Publications*:** Nero (2009a) [*In silico* Evaluation of Predicted Regulatory Interactions in Arabidopsis thaliana.](http://www.ncbi.nlm.nih.gov/pubmed/20025756?itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum&ordinalpos=1) ***BMC Bioinformatics***, 10(1): 435; Nero (2009b) “[A system biology approach highlights a hormonal enhancer effect on regulation of genes in a nitrate responsive "biomodule".](http://www.ncbi.nlm.nih.gov/pubmed/19500399?itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum&ordinalpos=4) ***BMC Syst Biol***., 3:59; Obertello M (2010) “Modeling the global effect of the basic-leucine zipper transcription factor 1 (bZIP1) on nitrogen and light regulation in Arabidopsis.” ***BMC Syst Biol***., 4:111; Ruffel et al (2011) *Nitrogen economics: Distinct systemic signaling for nitrogen supply vs. demand. Ruffel et al. (2011)* ***PNAS*** *108; 18524-18529*.

***A regulatory network controlling N-assimilation in response to nitrogen sensing*.** Our regulatory networks model the mechanisms by which the N-assimilation pathway, is regulated in response to inorganic and organic-N signals. Interestingly, a split-root system designed to measure systemic N signaling, revealed that genes involved in nitrate uptake and reduction are also regulated in response to systemic signals of N-supply and demand [Ruffel 2011]. Signaling by nitrate is well documented [Wang et al 2004] [Krouk, (2010) Current Opin Plant Biology] and recent studies have identified a nitrate “transceptor” [Ho 2009,] **[Gojon** 2011]. There is also ample, though less direct, evidence to support signaling by the N-assimilation products Glu/Gln [Gutierrez 2008] [Rawat 1999], but the receptor(s) remain elusive [Hsieh (1998)][ Lam (1998)] [Forde & Lea 2007, J Exp Bot. 58 pp2339]. To identify the downstream components of such inorganic and organic-N signaling mechanisms, we modeled regulatory networks controlling N-assimilation genes. Our first network models used an Arabidopsis multinetwork approach [Gutierrez 2007], to generate regulatory networks from steady state transcriptome data from plants exposed to nitrate [Nero et al 2009a,b] or organic-N [Gutierrez 2008] treatments. To advance our networks models from static to dynamic, we generated time-series transcriptome data (Aim 2B) and machine learning to infer how nitrate-signaling controls the N-assimilatory networks in roots [Krouk 2010] (See Aim4, below). These network studies support a model in which nitrate induces expression of TF hubs (e.g. HRS1, HHO1,2,3) that activate genes involved in nitrate uptake, reduction, and assimilation into organic-N (Glu/Gln) (Fig. XA) [Krouk et al 2010]. In turn, the organic-N products (Glu/Gln) of N-assimilation repress expression of TF activators (CCA1, GLK1, WRKY1) of Gln synthesis, in a negative feed back loop. This results in derepression of bZip1 an activator of the Gln-dependent synthesis of Asn, an inert, C-efficient amino acid used for N-transport/storage (Fig. XB). This regulation may represent an “energy conservation” mechanism to save ATP, reducing equivalents, and carbon skeletons, when levels of organic-N are abundant. Indeed, there is experimental support for bZip1🡪ASN1 in an “energy sensing” mechanism [Baena Gonzales]. ***Validation of TF🡪target predictions***: Our validation of this regulatory model (Fig. X) has involved TF perturbations in T-DNA mutants and transgenic plants. **For organic-N signaling**, the TF🡪target relationships shown in Fig. X, have been confirmed using the following lines: CCA1 (35S::CCA1) [Gutierrez 2008], GLK1 (glk1/2) [Para, unpublished]. Single locus WRKY1, T-DNA mutants confirm its role as “toggle” regulating the switch between Gln and Asn synthesis. Three WRKY1 T-DNA lines (SALK016954; SALK136009; SALK070989) each showed decreased expression of genes predicted to be targets of WRKY1 activation (NIA2 & NRT2.1), and increased expression of genes predicted to be targets of WRKY1 repression (bZIP1 & ASN1) (Fig. X). A bZIP1 T-DNA mutant showed altered regulation of genes controlled by nitrogen and light [Obertello 2010], and the bZIP1🡪ASN1 regulatory edge was confirmed independently by transient expression studies [Baena-Gonzales], and by preliminary bZip1 ChIP-seq studies (see Aim 1C of Plan). **For nitrate signaling**: Single and double mutants in the myb family (HRS1, HHO1, HHO2, etc) are being conducted by our collaborator, Dr. Krouk (INRA, France). These T-DNA mutants will be useful in our studies of TF interactions described in Aim 3 of the Plan.

**Prior Aims 4 & 2B. Creation of a time-dependent dynamic network model for the control of N- assimilation. *Relevant Publication***: Krouk *et al.* (2010) “Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate.” ***Genome Biology,*** 11 (12), R123.

***Summary:*** The analysis of fine-scale time-series transcriptome data generated in (Aim 2B), using a machine learning approach called “State Space modeling” (Aim 4), enabled us to predict network responses under untested conditions- the ultimate goal of Systems Biology. Network predictions were validated using left-out data and experimentally for select TFs, as described in [Krouk et al 2010] and below.

***A predictive network model for nitrate-control of the N-assimilation network.*** Because causality moves forward in time, time-series experiments are a valuable source of structure to derive predictive networks. We thus generated time-series transcriptome data from nitrate-treated plants (0,3,6,9,12,15,20 min) (Aim 2B) and identified 550 genes whose expression was regulated by nitrate, as a function of time. This analysis identified >200 genes (some induced as early as 3 min) not previously identified in studies of steady state transcriptome data [Wang 2004]. This time-series data of nitrate-regulated genes was then used to drive a machine-learning, network inference approach called “State-space analysis”, with several adaptations. 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) [Mirowski, P., et al. Clin Neurophysiol, 2009]. In a departure from previous state-space frameworks, we implemented a noise-mitigation approach that uses hidden variables to represent an idealized, “true” sequence of gene expressions **z**(*t*) that would be measured if there were no noise (e.g. in transcriptome data) (Fig. XA). The goal is to *learn* the function ***f***, that determines the change in expression of a target gene as a linear (or if needed non-linear, to account for TF interactions) combination of the expression of a relatively small number of TFs (typically under 10) (Fig. XB). To test the ability of “State Space” to generate accurate *predictive* regulatory networks, we used the 0, 3, 6, 9, 12, 15 min transcriptome data (as a training set), and then used the “learned” network to *predict* the direction of gene expression change from 15🡪20 min, and validated these predictions using the “left-out” 20 min transcriptome data. Surprisingly, based only on these relatively few time points, the coherence of the regulatory network model generated is good enough that it is able to predict gene expression states (e.g. induced or repressed) on future data points. State Space predictions of gene regulation at the 20 min time-point, were correct for 74% of the genes in a sub-network of 76 genes associated with the N-assimilation pathway. As comparison, the *"naive trend forecast"* test was correct for only 52% of the genes, just slightly better than random, p-val<0.006. When compared with other network inference approaches [Bonneau, Genome Biol, 2006. **7**(5): p. R36] [Bonneau, Cell, 2007. **131**(7): p. 1354-65.] [Wang, Bioinformatics, 2006.] [Shimamura, BMC Syst Biol.], our State-Space method showed a slight improvement in accuracy, and had a better signal-to-noise ratio using the same data. Further, our adapted 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. We next used the model to predict the “most influential TFs” in the network (e.g. ones that are predicted to influence the most genes in the network), and prioritized them for validation testing (Fig. XC). As studies of T-DNA mutants were hampered by issues of functional redundancy, we thus turned to TF overexpression constructs. For the master TF gene, SPL6, we examined overexpression of mRNA using an spl9 mutant impaired in miRNA binding [Wang Cell 2009, 138:738-749]. The effect of SPL9 over-expression evaluated over time, validated the vast majority of the genes in the regulatory network that we tested [Krouk 2010]. To accelerate testing all TFs hubs, we developed a transient expression approach described below.

**Prior Aim2A. Use of genome-wide approaches to validate causal TF->target relationships mediating N-regulation of N-assimilation networks. *Publication*: “***A transient transformation system for genome-wide transcription factor target assessment”.* Bargmann et al., 2012. (Submitted) **Summary**: This rapid transient expression system enables identification of direct TF targets in genome-wide within 2 weeks. We developed this system using a well-studied TF (ABI3) as a proof-of-principle study, so that targets uncovered in our transient expression system could be validated using known ABI3 network targets.

**Background**: A critical component of Gene Regulatory Network (GRN) inference is experimental validation of TF🡪target predictions. While TF🡪targets can be identified by Chromatin immunoprecipitation, TF perturbation studies have advantages because: (i) Chromatin-IP can confirm protein:DNA binding but it does not guarantee functional activation/repression [(Zheng *et al*, 2009)], and (ii) TF perturbation studies can identify the effect of TFs that may not bind directly to DNA. TF perturbation can be accomplished using T-DNA mutants or overexpressors, but both approaches are problematic. Functional redundancy of TFs is a built-in feature of GRNs that makes them robust, and as such molecular phenotypes in T-DNA mutants of TFs are rare [Cutler and McCourt 2005]. As alternative, transcriptomic analysis of transgenic TF over-expressors (35S::TF) can suggest the TF🡪 targets [Suzuki et al, Reeves], but it cannot reveal whether changes in transcript levels are a *direct* consequence of TF manipulation, or caused by indirect or possibly ectopic effects. A better approach is to use an inducible system that relies either on conditional expression of a TF, or regulated entry of the TF into the nucleus to activate its function [*e.g*. Hachez 2011, Bustos 2010]. Indeed, regulated nuclear translocation of a TF (using DEX), combined with the use of protein synthesis inhibitors (cycloheximide, CHX), has previously been used to eliminate the effects of secondary transcriptional regulators, and to identify direct from indirect TF targets, based on changes in target gene expression [Bustos]. We thus adapted this transient DEX-induction system to induce TF action in plant protoplasts, as described in [Bargmann 2012] and below.

**The System: A DEX-inducible transient TF perturbation system**: To accelerate validation of TF🡪target relationships using TF perturbation, we adapted our transient expression system [Bargmann 2009], to overexpress any TF, selectively induce its translocation to the nucleus, and identify changes in target gene expression (Q-PCR or transcriptome analysis). Technically, protoplasts are transfected with a 35S-driven Gateway expression vector harboring a TF fused to the rat glucocorticoid receptor (GR). In this system, HSP90 binding to GR, restrains TF-GR location to the cytoplasm, until the addition of the GR ligand dexamethasone (GR), which displaces HSP90 and allows nuclear import of the TF (Fig. X). Following transfection and DEX treatment, a positive fluorescent selection marker (RFP) on the vector allows isolation of successfully transformed protoplasts using Fluorescence Activated Cell Sorting (FACS) [Bargmann et al 2009]. Subsequent RNA analysis (by Q-PCR or transcriptome) can identify genes whose expression is affected (induced or repressed) upon TF nuclear localization induced by DEX. If the effects of DEX are measured in the presence of the translational inhibitor cycloheximide (CHX), only direct targets of the TF of interest (for details of this concept see) [ Lloyd et al (1994) *Science*.] [Sablowski RW, Meyerowitz EM (1998). *Cell*].

**The Results:** As ***a proof-of-principle***, this transient TF perturbation system was used to investigate the targets of the well-studied TF, ABI3 (ABSCISIC ACID INSENSITIVE3), a master regulator of ABA signaling networks [Vernoux *et al*, 2011 Mol Syst Biol. 2011 Jul 5;7:5089) where known ABI3 direct targets can be used to validate our system. Protoplasts were transformed with a 35S::GR::ABI3 expression vector, and sequential treatments of +/-CHX, and +/- DEX (to induce nuclear localization of TF) were performed. FACS selection of positive transformants [Bargmann 2009] and subsequent RNA analysis enabled us to identified genes whose expression was altered by DEX treatment in the protoplasts expressing the GR::ABI3 constructs, compared to empty vector controls. This analysis showed that known direct ABI3 target genes (e.g. CRU1, PER1) are specifically induced by ABI3 nuclear localization (e.g. +DEX), and this DEX-induction was enhanced under +CHX conditions, possibly by eliminating repressors (Fig. X). Further confirmation that other genes induced under +DEX +CHX are direct ABI3 targets includes: (i) significant over-representation of GO-categories known to be controlled by ABI3 (e.g. “response to abscisic acid stimulus”, “seed development”, “dormancy process”) and (ii) significant overrepresentation of *cis*-elements known to directly bind to ABI3 (e.g. the RY-repeat motif (CATGCA)) identified using (CISPROM) or de novo using MIME [Bailey and Elkan, 1994)]. These findings validate that this transient expression system can be used to identify direct TF targets. This transient assay system is adapted to high through put testing of TFs in the N-assimilatory network (see Plan). **Prior Aim 3. Determine the role(s) of post-transcriptional mechanisms in mediating N-regulatory network. *Relevant publications*:** Vidal EA, Araus V, Lu C, Parry G, Green PJ, Coruzzi GM, Gutiérrez RA (2010). “[Nitrate-responsive miR393/AFB3 regulatory module controls root system architecture in Arabidopsis thaliana.](http://www.ncbi.nlm.nih.gov/pubmed/20142497?itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum&ordinalpos=1)” ***Proc. Natl. Acad. Sci.*** ***USA***, 107(9): 4477-82. ***Summary***: This aim explored the role of miRNAs in mediating nitrogen-metabolite signaling, and the use of 15N-tracers to monitor N-use efficiency. The miRNA studies were conducted in collaboration (R. Gutierrez, Chile), a Fogarty Award recipient connected to this NIH grant. That study identified how a miRNA-TF module was regulated in response to sensing inorganic nitrate (externally) or N-assimilated internally (Glu/Gln) [see Vidal 2010].

**RESEARCH STRATEGY:**  Our strategy builds on our prior success in inferring N-regulatory networks able to predict gene responses under untested conditions – the ultimate goal of Systems Biology [Krouk 2010]. We now aim to expand on and improve the predictive power of our learned networks using an approach called “*NetWalk”: an iterative high throughput experimental/machine learning approach to infer causal networks.* This approach is novel, in that it combines diverse types of genomic datasets (time series, steady state and TF perturbation data) generated in Aim1, into a pipelined network inference analysis in Aim 2, to improve the predictive power of the networks, and to infer the next TFs and TF pairs for perturbation in Aim 3, in an iterative Systems Biology cycle of high through put experimentation, analysis/modeling, and validation.

**Aim 1. NetWalk Targets: High-throughput validation of TF🡪network targets**. ***Rationale***: We will generate diverse genomic datasets in Aim1, that will drive a the new NetWalk network inference pipeline in Aim 2. Combining diverse datasets into one analysis pipeline, will enable us to identify “high confidence” edges to validate and refine our network inference pipeline and to increase its predictive power. The data we will generate consists of: i) new time-series transcriptome data (Aim 1A), ii) steady state transcriptome data, and iii) TF perturbation data in which TF->targets will be validated by transcriptomic (Aim 1B) and ChIP-seq (Aim 1C) analysis. In each subaim, we provide preliminary data for one TF, as proof-of-principle for the feasibility of the approach.

**Aim 1A. Generation of time-series transcriptome datasets for organic-N signaling.**  Time-series transcriptome data from nitrate-treated plants, was very valuable as it enabled us to (i) identify >200 new nitrate regulated genes (including TFs induced as early as 3 min) compared to steady state data [Wang 2004], and (ii) to derive dynamic, predictive network models based on State-space machine learning [Krouk 2010]. We will now use the same approach to generate time-series transcriptome data (0, 3, 6, 9, 12, 15, 20, 25, 30 min) under conditions shown to elicit an organic-N response (growth on 1mM nitrate, transient treatment with 40mM ammonium/nitrate vs. control KCl) [Gutierrez 2008]. The analysis of this new time-series data, for organic-N signaling, should enable us to uncover new genes including TF and generate network models (using machine learning approaches [Krouk 2010] and using the new pipeline in Aim 2, to prioritize TF hubs among the first TFs to be induced, for validation studies described below.

**Aim 1B. “NetWalk”: Generation of high throughput, high confidence TF🡪network target data.** *Prioritization of TF hubs for testing.* Our current network models have identified TF hubs predicted to regulate genes in the N-assimilatory pathway by nitrate [Krouk 2010] or organic-N [Gutierrez 2008] signals (Fig. X). TF hubs in the organic-N networks (CCA1, GLK1, WRKY1, bZip1), were identified from steady state data, and TF🡪targets were based on correlation (>0.8) and representation of *cis*-elements (Fig. X) [Gutierrez 2008]. In a machine learning approach, time-series data and state-space modeling, enabled us to generate dynamic, predictive networks for nitrate regulation of the N-assimilation pathway, and to identify TF hubs induced early (e.g. HRS1, HHO1, HHO2, HHO3) and at the top of the cascade (Fig. X) [Krouk 2010]. Unexpectedly, the TF hubs predicted to mediate either nitrate or organic-N regulation of N-assimilation - and identified in separate studies - comprise two subclades of a single myb gene family (see Fig. X), a member of which was previously associated with phosphate signaling [Liu 2009]. This finding underscores the biological relevance of these prioritized TFs in nutrient sensing, which is supported by preliminary studies below.

*“NetWalk”: Generation of TF🡪network target data using transient TF perturbation.*To identify network targets of the prioritized TFs genome-wide, we will adapt the transient system described in Progress (Prior Aim2A) [Bargmann 2012]. Briefly, this TF perturbation system currently comprises the following components; i) transient expression (using 35S CaMV promoter) of a GR-TF fusion in protoplasts (the vector expresses RFP as transfection control allowing FACS sorting of successfully transfected cells), ii) use of an inducible Dexamethasone (DEX) system to inducibly control translocation of the TF into the nucleus, iii) use of the translation inhibitor Cycloheximide (CHX), to identify primary targets of the TF. We will now add an additional layer to the DEX-concept, by iv) including N-treatments to precondition cells (prior to TF nuclear localization by DEX) to allow N-regulated post-translational modifications of TFs, and/or N-regulated transcriptional activation of potential TF cooperators. The ***prioritized TFs*** will be subjected to transient perturbation studies and Fig. X shows progress in Gateway cloning and transient expression. For each TF, the N-pretreatments will be tailored for either nitrate signaling (growth in ammonium succinate, treatment with nitrate) or organic-N signaling (growth in 1mM nitrate, and treatment with 40mM ammonium/nitrate). Indeed, we show using a ***proof-of-principle*** example TF (HRS1) that (i) nitrate-preconditioning affects DEX-induced target gene activation and (ii) that the TF🡪targets uncovered in this transient assay have relevance to whole plants.We initially hypothesized a role for the myb TF HRS1 in nitrate signaling, as its expression is induced within 9 min of NO3- treatment, it is predicted to be the top most controlling and top most controlled TF hub in the N-assimilation network [Krouk et al 2010]. Intriguingly, HRS1 was previously associated with phosphate signaling, based on phenotypes of transgenic 35S::HRS1 plants [Liu 2009]. Our preliminary analysis of transcriptome data from 35S::GR-HRS1 activation in the transient DEX system, revealed four distinct gene clusters whose expression is influenced by a combination of HRS1 nuclear import (by DEX) and nitrate-treatment (Fig. X). Genes in Cluster 4, defined as primary targets of HRS1 (e.g. activated by +DEX+CHX), are specifically induced in +DEX/+NO3 conditions. Further, GO-term analysis of these HRS1 nitrate-dependent HRS1 direct targets, reveals significant overrepresentation of genes involved in phosphate transport (p-val 8.14 E-6), reminiscent of the phosphate signaling phenotype observed in 35S::HRS1 plants [Liu 2009]. This result provides evidence that TF🡪targets identified in the transient DEX protoplast system have relevance to whole plants, and further suggest that some of the TFs in our nitrate-regulated networks may coordinate responses to diverse nutrients. Following this example, transcriptome analysis will identify specific genes, clusters and biological processes controlled by each prioritized TF. The TF🡪target data from transcriptome analysis will be complementary to the Chromatin-IP (ChIP) data (in Aim 1C) and has several advantages: (i) ChIP can confirm protein:DNA binding, but it does not indicate functional regulation [Eilers 2008], and (ii) transcriptome analysis from the DEX-inducible TF system allows one to identify the effect of regulatory components that may not bind directly to DNA [ Lee, J, 2007].

**Aim 1C. Genome-wide validation of TF🡪network targets using Chip-Seq.** Chromatin immunoprecipitation followed by sequencing analysis (ChIP-seq) can reveal the binding of a TF to the promoter of a target gene, but does not indicate if this results in actual gene activation/repression [(Zheng *et al*, 2009)]. Thus, we will perform ChIP analyses on protoplasts from TF-perturbation studies, and compare TF🡪targets identified by ChIP, with target genes identified by transcriptional analysis (Aim 1B) [Zhu 2012]. To test the feasibility of performing ChIP-Seq and transcriptome analysis on the same samples, we performed a ***proof-of-principle*** study that identified a genome-wide map of TF binding sites for the TF bZip1, using ChIP seq on samples from the transient protoplast expression system. We adapted the micro-ChIP protocol from [Dhal and Collas, 2008], which requires a relatively small number of cells (1,000 cells). After transformation of protoplasts with a 35S::GR::bZip1 construct, transformed protoplasts (~5-8,000 cells), were treated with 1% formaldehyde for 10 min, quenched with 100mM glycine for 5 min. Cells were washed in W5 buffer, pelleting by centrifugation (2,500 rpm for 2 min) and snap-frozen at -80C. Anti GR (sc-1002) antibodies (Santa Cruz Biotechnology Inc.) were used to capture the GR::bZip1 protein:DNA complexes, and ChIP DNA was purified after reverse cross-linking using the MiniElute kit (QIAGEN). For ***Illumina sequencing*** **of the ChIP-DNA**, paired-end libraries were constructed using the sample prep guide (Illumina, San Diego, CA), with the following modifications for low amounts of starting DNA (~1ng): (i) adaptor oligo mix was further diluted 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 from the immunoprecipitated DNA and the input DNA as control. The libraries were sequenced on the Illumina GAIIx platform. ***Bioinformatic analysis* of ChIP Seq reads** was used to identify network targets of bZIP1 binding genome-wide. The sequence reads were filtered for quality, trimmed to remove adapter sequences, and aligned to the TAIR10 assembly of the *A. thaliana* genome. Genomic regions significantly enriched in the immunoprecipitated sample relative to the input DNA, were identified using the QuEST peak-calling algorithm [Valouev et al., Nature methods 5, 829-834 (2008)]. This analysis identified approximately 300 genic regions that show increased peak heights (individual peak heights **>4** fold, overall genic region **>2** fold) compared to the background (input DNA), indicating significant binding (p<1E-8) of bZIP1 binding sites in these 300 genes. **To validate that the ChIP samples were *direct* bZip1 targets**, we performed an analysis of cis-regulatory motifs. 500bp immediately upstream of the transcription start site (TSS) of genes with the highest normalized read count for bZIP1 binding, were analyzed with MEME [Bailey and Elkan, Proc. of ISMB,1994 pp28-36]. The most significant cis-motif found (p<1.9e-6) in this set was G[C/A]CACGT[G/C] which includes the G-box motif (CACGTG), a known bZIP1 binding site [Kang et.al. *Mol. Plant (2010) 3 (2): 361-373*]. As further validation, we intersected of bZip1🡪target data from ChIP studies with the bZip1 targets predicted from our multinetwork analysis (e.g. based on TF🡪target correlation > o.8 and over-representation of bZIP1 cis-binding elements) [Gutierrez 2008]. Of the 30 predicted bZip1 targets in this organic N-regulatory network, 16 show significant enrichment in the bZip1-ChIP studies, and the significance of this overlap is p<0.001. These bZIP1 targets include the prioritized TFs in our networks: CCA1, GLK1, WRKY1 whose regulation by bZip1 is depicted in Fig. X., as well as the N-assimilation target gene ASN1.

**Outcome of Chip-Seq**: **Identifying gold-standard TF🡪target paire and predicting TF interactions:** The TF🡪targets identified by ChIP seq data will be intersected with targets identified by transcriptome analysis (Aim 1B). This will identify “gold standard” TF🡪target examples where direct TF binding results in gene activation (or repression). This high-quality information will be used to train and validate our inference pipeline in Aim 2. In addition, as we perform ChIP studies across the TFs in our network, the collective ChIP studies will provide information about TF1 and TF2 binding (e.g. promoter location, target genes etc), that will help to direct our studies on TF interactions (in Aim 3B). Once we have Chip-Seq data from several TFs in the network, we can start looking at combinatorial regulation. This approach is fully described in Aim 3.

**Data integration and analysis**: The distinct sets of data generated in Aim 1 (i) time-series transcriptome, (i) steady state transcriptome, (iii) data from TF perturbations, (a) transcriptome and (b) ChIP-Seq, will be combined into a single pipeline for network inference in Aim 2.

**Expected results, limitations and alternate approaches (Aim 1)**: Aim 1 will produce distinct datasets to “feed” and validate the Pipelined Network Inference approach in Aim 2. Some new outcomes will be, for example, the identification of new TFs invoked to operate early in the network based on new time-series data for organic-N signaling (Aim 1A), and validations of direct TF1🡪TF2 relationships the effects of which can be further explored in Aim 3. TF interactions can also be predicted from ChIP seq experiments performed across several TFs. For each subaim (Aim 1B, C, and D), we have proof-of-principle examples to demonstrate the feasibility of the approach, so we do not anticipate technical problems in data generation or computational interpretation. In addition, the coordinated analysis of the distinct data sets (e.g. based on transcriptome (Aim 1B) vs ChIP seq (Aim 1C) will cross-validate our TF🡪targets identified.

**Aim 2: NetWalk Inference: A computational pipeline for network inference combining multiple data-types. *Rationale***: Our ultimate goal is to build 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 a small (76 gene) 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 quality and predictive power of such networks, by feeding a network learning pipeline both new transcriptome data from new time-series experiments (for organic-N signaling, Aim 1A), as well as new TF perturbation data. The new transcriptome data generated from these TF perturbation experiments will uncover direct TF targets (Aim 1B), which will also be validated by ChIP-seq (Aim 1C). The refined network models derived from this pipeline in Aim 2, will in turn suggest new TFs and TF pairs on which to perform perturbation experiments (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 Pipelined Network Inference approach to generate a predictive network model that makes use of four types of genomic data which we have or will generate in Aim 1: (i) steady state transcriptome data (e.g. N-treatments); (ii) time-series data (e.g. expression over time), (iii) TF perturbation data in which TF🡪target binding is validated by transcriptome analysis and (iv) ChIP seq. No single algorithm for machine learning can be used on all four kinds of data, so we will use several algorithms together in a Pipelined Network Inference approach to improve learning by integrating data analysis, as described below.

**The Network Model**: The nodes in the causal network will be genes and the edges between genes will represent positive and negative influences. A positive coefficient associated with an edge implies that the edge is inductive. 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 roughly ten, but this still means that even for 100 genes, our algorithms would have to find the proper 1,000 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, our initial model will be “linear” as opposed to quadratic. (A quadratic model would include terms of the form d\*B\*C, where B and C would be gene expressions of TF pairs that interact in a multiplicative fashion. We will refine our initial models with a quadratic model for TF interactions in Aim 3, but only in special cases. Surprisingly, linear models can explain behavior well in network predictions. A model explains behavior if it can predict the state of one gene given the state of other genes at previous time points [Krouk et al 2010].

**Building a 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 we subjected to preliminary validation testing, the GENIE3 algorithm is used on steady state data (resulting from N-treatments), the MCZ algorithm (short for the Median Corrected Z-score method [Greenfield, (2010]) is used for TF perturbation data, followed by the DFG algorithm (short for Dynamic Factor Graph) for time-series data [Krouk (2010]. In this pipeline, the GENIE3 and MCZ algorithms assign initial weights to certain edges that are then refined by the DFG algorithm. In preliminary work, done for a book we have written called *Network Inference in Molecular Biology: a hands-on approach*, [Lingeman/Shasha 2012], we show using data from the DREAM benchmark (Dialogue for Reverse Engineering Assessments and Methods) [Schaffter, (2011)] that the GENIE3🡪MCZ🡪DFG pipeline is better than using any single machine learning algorithm by itself.

**Step 1: GENIE: Steady state data.** The GENIE3 (**GE**ne **N**etwork **I**nference with **E**nsemble of Trees) [Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE 5(9):e12776. doi:10.1371/journal.pone.0012776] algorithm uses a collection of regression trees (which are generalized decision trees) to perform network inference. The algorithm creates a collection of regression trees, ranks the edges from each ones, and then creates a global ranking.

**Step 2: MCZ: TF perturbation data**. In the MCZ algorithm, if gene xi influences gene xj, then perturbing gene xi should change the value of gene xj in a significant way. We will measure the significance based on the number of standard deviations from the median value of gene xj over all non-perturbation experiments. This approach has successfully identified genes whose expression is significantly altered when analyzing steady state transcriptome data mutants [Greenfield 2010].

**Step 3: DFG: Time series data**. The DFG (Directed Factor Graph Approach) is a version of state space modeling that we have successfully used to perform network inference time-series data [Krouk 2010]. (see Previous Aim 4, for an explanation of this network inference method).

**Preliminary Results of the Pipelined Network Inference Approach:** To test the value of pipelines for predicting edges in our N-assimilatory networks, we compare the results of our previous Dynamic Factor Graph (DFG)/State Space Modeling Approach [Krouk,G 2010] built on time-series data with the GENIE3🡪MCZ🡪DFG pipeline. 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. On the other hand, using just the data up to 12 minutes, we were able to predict the values at 15 minutes correctly only 57% of the time. As the following table shows, new data and the computational pipeline improves these results at least for this small sample of genes and using new perturbation data for only one TF (HRS1).

|  |
| --- |
| **TABLE X: NETWALK MACHINE LEARNING PIPELINE (Preliminary Results)** |
| **STEPS** | **1. GENIE** | **2. MCZ** |  **3. DFG** | Prediction Accuracy | Prediction Accuracy |
| **DATA** | **Steady State** | **TF perturbation** (35S::HRS1 +DEX) | **Time Series** | **12-15 min** | **15-20 min** |
| **Test 1** | No | No | **Yes** | 57% | 74% |
| **Test 2** | **Yes** | No | **Yes** | 66% | 74% |
| **Test 3** | **Yes** | **Yes** | **Yes** | 68% | 76% |

**Caption**: The 12-15 min column shows the correctness of predicting whether expression will increase or decrease between 12 and 15 minutes depending on training the data up to 12 minutes. Using time series data alone gives just 57% accuracy, this increases substantially with steady state data to 66%, and still further to 68% given just one TF perturbation dataset (for 35S::HRS1 in the transient DEX system). The improvements from 15 to 20 minutes are less impressive, partly because the model is already so good (for these 76 genes in the network) with the time series data alone. The single TF perturbation gene (HRS1) helped, even though that experiment gave the direct targets of that gene alone.

**Intended Approach**: Based on our preliminary results, we will iterate the following steps: (i) identify the genes relevant to nitrogen treatments based on steady state and time-series transcriptome data resulting in just over 200 genes (including regulated members of the 66 genes in the N-assimilatory pathway as well as TFs), (ii) develop a Pipelined 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 predicted 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 pertain to the most important genes or which will clarify inaccuracies in our model.

**In addition to the pipeline approach**, we will also test a crowd-sourcing approach to apply several algorithms that each predicts edges and then use a “voting” approach to determine the best one. In that scheme, illustrated in the figure Pipelined.eps [Gloria: do we want figures for both or just the pipeline?], in addition to the Median-Corrected Z-score and Dynamic Factor Graph algorithms, we use the NIR (Network Identification by Multiple Regression [Gardner, (2003)]) and CLR (Context Likelihood of Relatedness [Faith, (2007)]) on steady state data. We also use BANJO (Bayesian Inference with Java Objects [Yu,J (2004)]), Time-Delay ARACNE [Zoppoli, (2010)], and Inferelator [Greenfield, (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) 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).

**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). 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)**: Aim 2 will yield a “learned regulatory network” for TFs controlling the N-assimilatory pathway, built/trained on a pipeline analysis of data sets (from Aims 1 and 3). Predictions will be performed using out-of-sample data, and we will test a number of machine learning methods together to identify which combination is best. The main possible problems we may encounter, have to do with the fact that the number of possible edges is much larger than the number of experimental data points we will have. We nevertheless believe that the approach will work, because we have already seen successful results in predictive modeling using much less data, and a single machine algorithm [Krouk et al 2010]. Moreover, our preliminary results for combining machine learning approaches in a pipeline shown in Aim 2 suggest we can improve on this already successful approach to network inference.

**An additional benefit of our approach**, is that we will discover which data type (time-series, steady state, TF perturbation) is most useful for network inference. For example, we may learn that the TF perturbation data is more informative than the steady state data and Chip-seq data combined, or that time-series is just as informative (and less expensive) compared to Chip-seq for regulatory network modeling. These findings will allow our lab and possibly others to save money and time, by focusing on generating the most informative data type(s) for network inference.

**Aim 3: NetWalk Coordination: Experimental testing of transcription factor interactions.**

***Rationale***: The refined network models in Aim 2 will suggest additional TF perturbation experiments to perform. This includes: i) new TF hubs and ii) TF interactions. These studies have the potential to identify TFs and network motifs (e.g. TF pairs) that globally regulate the N-assimilatory pathway in response to inorganic- (nitrate) and organic-N (Glu/Gln) signals. These findings will be tested further in transgenic plants.

**Aim 3A. Perturbation studies for new candidate TFs**: We will functionally validate TF🡪targets for new TFs identified in Aim 2 using the transient DEX-system, as described in Aim 1B-C. With this approach, we will test new TFs hubs acting early and at the top of the cascade as newly discovered in new time series experiments (Aim1A) and/or TFs that affect key metabolic genes in the N-assimilatory pathway (e.g. GLN1 and ASN1) whose expression is known to affect N-use efficiency [Lam 2003][Oliveira 2002].

**Aim 3B. Testing TF interactions in the control of the N-assimilatory network**

**Identification and Prioritization of TF pairs for testing**: The linear models used in the NetWalk pipeline to “learn” and infer regulatory networks in Aim 2, will identify all TFs predicted to individually and additively control expression of a target gene. However, to avoid overfitting on the relatively small amount of data, our algorithm does not explicitly model TF interactions (see explanation Aim 2). However, in a second phase of analysis (e.g. after the initial modeling identifies a set of potential TF regulators), we can adapt our linear network model to predict interactions among these TFs by creating a reasonably parsimonious quadratic model as follows: for each target gene X in the linear model, we will try to fit a model for X testing pairwise combinations of X’s TF inputs. Thus, the linear model (in Aim 2) can suggest which TFs are relevant to X. Now, in Phase II, the linear model is then refined into a quadratic model on those relevant TF genes and incorporated into our machine learning model (in Aim 2). In fact, we already have evidence for TF interactions from our current network models. For example, interaction of BZip1 and WRKY1 in the regulation of ASN1 (WRKY1|--|bZIP1🡪ASN1) has experimental support including regression modeling and from bZIP1 binding to WRKY1 and ASN1 promoters (see Aim 1C). As we generate additional ChIP-Seq data for the prioritized TF list (Fig. X), we will use algorithms, such as SpaMo [Whitington et. al., NAR 2011], designed to identify tightly linked cis-motifs, whose spatial distribution could suggest a functional and/or mechanistic basis for TF1:TF2 protein interactions. Based on the networks generated in Aim 2 and our added quadratic models, we will prioritize testing of TF pairs that: i) are predicted by our networks to have regulatory interactions (e.g. TF1🡨🡪TF2), ii) whose binding sites are found to co-occur in Chip-Seq studies, and/or iii) that share the most targets in the network and/or target key N-assimilatory genes shown to control N-use (e.g. GLN1 and ASN1) [Lam 2004][Oliviera].

**Approach**: We will test potential TF interactions using the transient DEX-inducible system described in Aim 1B using two complementary approaches: (i) Co-expression of TF1 and TF2 in a single vector, (ii) expression of TF1 in protoplasts from tf2 mutant background. In each case, synergistic activation (or repression) of target genes when TF1 and TF2 are both present, compared to either one alone, will be evidence for interaction.

 ***Co-expression of TF1 & TF2*:** To test how TF interactions affect target gene expression *in vivo*, we will co-activate the TFs in the DEX transient assay system using a TF1/TF2 double expression vector, and identify genes whose expression differs synergistically compared to either TF expressed alone in a single vector. The control is the empty vector. TF target genes whose expression in the TF1/TF2 double vector experiments will be compared to each single TF. If target gene expression is synergistically activated (e.g. more than the additive effect of its expression levels in response to TF1 alone + TF2 alone), this will considered evidence for TF interaction. In order to co-express two TFs (TF1 and TF2) in the same cell, we will use the MultiSite Gateway**®** Technology that allows for simultaneous cloning of 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) will be cloned in appropriate Gateway donor vectors, and allowed to recombine with each other and with the destination vector pBob11 [Bargmann 2009]. This results 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.

***Perturbation of TF1 activity in a tf2 mutant background***: As a complement to the multisite vector approach, we will use sensitized genetic backgrounds to examine effects of TF1/TF2 interactions on target gene expression *in vivo*. Here we will transiently express the single 35S::GR:TF1 in protoplasts made from T-DNA mutant in TF2. This will enable us to assess TF1 function in the absence of TF2. We will compare the effect on target gene expression to that observed when TF1 is expressed in protoplasts from wild-type plants (where native TF2 is present). Loss or change in target gene regulation by TF1 in a tf2 mutant protoplast (compared to wild type), could indicate TF1 and TF2 cooperation or interaction. We will use this genetic approach for TFs whose mutants show a molecular phenotype (e.g. alteration in target gene expression). But we will also explore the possibility that a T-DNA mutant that does not exhibit alteration in expression of a target gene (e.g. due to TF redundancy), could provide a sensitized background to uncover evidence (in the DEX transient system) for their interaction.

**Aim 3C. Testing effects on N-assimilation of TF perturbations in whole plants.** For selected TFs, we will study changes in genome-wide responses to nitrogen treatments (e.g. nitrate or Glu/Gln) in plants perturbed in TF expression (T-DNA) or transgenic overexpressors. A list of current T-DNA and transgenic lines is in Fig. X. We will perform transient treatments with conditions shown to elicit a nitrate response (growth on ammonium succinate/treatment with nitrate) [Krouk 2010] or conditions shown to elicit an organic-N response (growth on nitrate/treatment with high ammonium nitrate or Glu) [Gutierrez 2008], and monitor changes in transcriptome responses. Transcriptome data from the TF perturbations in planta, will also be used to refine our network models (Aim 2).

**Data integration and analysis**: The transcriptome 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 analyze/integrate the transcriptome data supporting TF1/TF2 interactions (Aim 3B) into our machine learning pipeline in Aim 2, we will modify the model for TF regulation of target gene expression to include TF interactions as quadratic terms in cases where the TF1/TF2 interaction data shows synergistic effects for a target gene expression, compared to the single TF data.

**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 improve the quality of edges in the inferred network of Aim 2. It will also enable us to test how TF interaction affects target genes in the network, by studying whether two TFs affect target gene expression in a synergistic fashion *in vivo*, using the transient assay system.

**Aim 1 Postscript**: **Effects of TF perturbations on N15-assimilation in whole plants**: For TFs validated to affect the N-assimilatory pathway based on our transient assays, we will collaborate with Dr. Gabriel Krouk & Sandrine Ruffel at the INRA Biology & Biochemistry of Plants on studies related to optimizing 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 plants for several TFs for which changes in N-assimilation target genes are observed in the corresponding T-DNA mutants or transgenic overexpressors (e.g. WRKY1 T-DNAs, 35S::CCA1). This collaboration between our genomics lab and a world-renowned plant physiology lab, will be the ultimate test of our ability to use system biology approaches to predict targeted interventions in the N-assimilatory pathway, with significance to plant biomass and N-use efficiency.

**TIMELINE**

**Year 1-2**

**Year 3-4**

**CUT SECTION:**

Our aims are:

Aim 1: NetWalk Targets: High throughput experimental validation of TF🡪target genome-wide.

Aim 2: NetWalk Inference: A computational pipeline for network inference combining multiple data-types.

Aim 3: NetWalk Coordination: Experimental testing of transcription factor coordination.

].

**Aim 1D. Effects of TF perturbations in whole plants.**  ***Transgenic Studies***: Based on results in our transient assays (Aims 1B and C), we will select TFs for perturbations *in planta*. For some TFs, we already have data from stable 35S::TF transgenic lines which support TF🡪target relationships predicted in our network models, e.g. 35S::CCA1 [Gutierrez et al 2008], 35S::GLK1 [XXXX], 35S::HRS1 [Liu 2009], so transcriptomic analysis to identify genes whose expression is activated (or repressed) compared to wild-type controls, will be straight forward. However, such lines will not allow us to identify direct targets. Therefore, in selected examples prioritized by our transient studies, we will produce transgenic Arabidopsis plants that express a TF::GR fusion under the control of the TF native promoter?, to enable inducible expression and identification of target genes. To do this, the genomic region encompassing the coding sequence and the promoter (1Kb upstream the ATG or up to the upstream gene) of the TF will be cloned into a pENTR vector and then transferred into a plasmid called “pDEX”, a Gateway destination vector we created (to enable rapid TF-GR cloning) by inserting the GR sequence downstream the Gatewaycassette of pMDC 99 [Brand et al.,2006. [Plant Physiol. 141:1194-1204](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=16896232)].

**BACTERIA**

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)

**YEAST**

Cantone, I. et al. A **yeast** synthetic network for in vivo assessment of reverse-engineering and modeling approaches. Cell 137, 172–181 (2009)

Ye, C., Galbraith, S. J., Liao, J. C. & Eskin, E. Using network component analysis to dissect regulatory networks mediated by transcription factors in **yeast**. PLoS Comput. Biol. 5, e1000311 (2009).

Zhu, J. et al. Integrating large-scale functional genomic data to dissect the complexity of **yeast** regulatory networks. Nature Genet. 40, 854–861 (2008).

Lee, I., Date, S. V., Adai, A. T. & Marcotte, E. M. A probabilistic functional network of **yeast** genes. Science 306, 1555–1558 (2004).Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S.,

**ANIMALS**

Margolin, A. A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a **mammalian** cellular context. BMC Bioinformatics 7, S7 (2006).

Basso, K. et al. Reverse engineering of regulatory networks in **human B cells**. Nature Genet. 37, 382–390 (2005).

**CANT DETERMINE ORGANISM**

Gregoretti, F., von Belcastro, Di Bernardo, D., Oliva, G.: PLoS ONE: A Parallel Implemen- tation of the Network Identification by Multiple Regression (NIR) Algorithm to Reverse-Engineer Regulatory Gene Networks. PloS one (2010)

Heckerman,D.,Geiger,D.:LearningBayesianNetworksTheCombinationofKnowledgeand Statistical Data. Machine Learning (1995)

Huynh-Thu,V.A.,Irrthum,A.,Wehenkel,L.,Geurts,P.:InferringRegulatoryNetworksfrom Expression Data Using Tree-Based Methods. PloS one 5(9), e12,776 (2010)

Marbach,D.,Schaffter,T.,Mattiussi,C.:Generatingrealisticinsilicogenenetworksforper- formance assessment of reverse engineering methods. Journal of Comp. Biology 16(2), 229– 239 (2009)

Yu,J.:AdvancestoBayesiannetworkinferenceforgeneratingcausalnetworksfromobserva-tional biological data. Bioinformatics 20(18), 3594–3603 (2004)

Zavlanos, M.M., Julius, A.A., Boyd, S.P., Pappas, G.J.: Inferring Stable Genetic Networksfrom Steady-State Data. Automatica 47, 1113–1122 (2011)

Zoppoli, P., Morganella, S., Ceccarelli, M.: TimeDelay-ARACNE: Reverse engineering ofgene networks from time-course data by an information theoretic approach. BMC Bioinfor- matics (2010)

**PLANT NETWORKS:** Mostly these use steady state data or boolean models,

Network Inference, Analysis, and Modeling in Systems Biology Reka Albert

Published online before print November 2007, The Plant Cell November 2007 vol. 19 no. 11 3327-3338

Lan, H., Carson, R., Provart, N.J., and Bonner, A.J. (2007). Combining classifiers to predict gene function in Arabidopsis thaliana using large-scale gene expression measurements. BMC Bioinformatics 8: 358

Locke, J.C., Millar, A.J., and Turner, M.S. (2005). Modelling genetic networks with noisy and varied experimental data: The circadian clock in Arabidopsis thaliana. J. Theor. Biol. 234: 383-393

Mendoza, L., and Alvarez-Buylla, E.R. (1998). Dynamics of the genetic regulatory network for Arabidopsis thaliana flower morphogenesis. J. Theor. Biol. 193: 307-319

Zeilinger, M.N., Farre, E.M., Taylor, S.R., Kay, S.A., and Doyle III, F.J. (2006). A novel computational model of the circadian clock in Arabidopsis that incorporates prr7 and prr9. Mol. Syst. Biol. 2: 58

Chang, W.C., Li, C.W., and Chen, B.S. (2005). Quantitative inference of dynamic regulatory pathways via microarray data. BMC Bioinformatics 6: 44.

Li et al., 2006 Li, S., Assmann, S.M., and Albert, R. (2006). Predicting essential components of signal transduction networks: A dynamic model of guard cell abscisic acid signaling. PLoS Biol. 4: e312

Albert et al., 2007a Albert, R., DasGupta, B., Dondi, R., Kachalo, S., Sontag, E., Zelikovsky, A., and Westbrooks, K. (2007a). A novel method for signal transduction network inference from indirect experimental evidence. J. Comput. Biol. 14: 927-949 (discusses net-synthesis)

Ma and Zeng, 2003; Ma, H.W., and Zeng, A.P. (2003). The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics 19: 1423–1430

Ma'ayan et al., 2005

Diaz and Alvarez-Buylla, 2006 A model of the ethylene signaling pathway and its gene response in Arabidopsis thaliana: Pathway cross-talk and noise-filtering properties. Chaos 16: 023112

Espinosa-Soto, C., Padilla-Longoria, P., and Alvarez-Buylla, E.R. (2004). A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. Plant Cell 16: 2923-2939

Geisler-Lee, J., O'Toole, N., Ammar, R., Provart, N.J., Millar, A.H., and Geisler, M. (2007). A predicted interactome for Arabidopsis. Plant Physiol. 145: 317-329

Jen, C.H., Manfield, I.W., Michalopoulos, I., Pinney, J.W., Willats, W.G., Gilmartin, P.M., and Westhead, D.R. (2006). The Arabidopsis co-expression tool (act): A www-based tool and database for microarray-based gene expression analysis. Plant J. 46: 336-348

Li et al. (2006) Systems Biology Update: Cell Type-Specific Transcriptional Regulatory Networks

Plant Physiol. 2010 152: 411-419

Systems Biology Update: Cell Type-Speciﬁc Transcriptional Regulatory Networks

Li Pu and Siobhan Brady\* Department of Plant Biology and Genome Center, University of California, Davis, California 95616

**General reference to systems science:**

Albert, R., and Barabasi, A.L. (2002). Statistical mechanics of complex networks. Rev. Mod. Phys. 74: 47-97