**(a) Significance:** Modeling predictive Gene Regulatory Networks (GRNs) from comprehensive gene expression sets is both a major goal *and* challenge in fundamental and applied research in Systems Biology. Our overall goal in this grant, is to infer a causal genetic network, effectively the circuit diagram underlying the regulation of nitrogen-use efficiency in response to nitrogen sensing in plants, through the iteration (loop) between high throughput experimentation and machine-learning approaches to network inference. These “learned” network models are then tested for their ability to infer 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 (e.g. presence of too few time-point measurements, many genes, measurement errors and noise) [Jaeger and Monk 2010][ De Smet and Marchal (2010) Nature Reviews: Microbiology, vol 8, p 717-729] Despite these caveats, machine learning approaches have successfully been used to predict network states under untested conditions in microbes [Bonneau Cell 2007] and, more recently in plants [Krouk 2010]. In the current NIH cycle, we applied a machine learning approach to time-series transcriptome data, and the derived networks accurately predict gene expression states on left-out experimental data [Krouk 2010]. We now propose to build and improve on the predictive power of our inferred networks, using an innovative and iterative combination of high throughput experimentation and computation. On the experimental front, we have developed a high-throughput approach to TF perturbation that enables us to rapidly validate TF🡪target interactions genome-wide (within weeks) and to overcome problems of TF redundancy. In this system, TF🡪target edges are validated with transcriptomics and ChIP-seq data. These TF🡪target datasets will in turn train and validate a new network inference pipeline. The novelty of this pipeline, compared to current network inference methods, is that it is capable of analyzing distinct genomic data-types (e.g. time series, steady state, and TF perturbation data) using a combination of algorithms. Evidence that this pipelined approach will improve the predictive power of our inferred networks is suggested by preliminary studies. In our application, the inferred N-assimilatory networks can suggest targeted interventions to reduce nitrogen fertilizer use, with implications for human health, energy and the environment. More broadly, our work will illustrate a combined experimental/informatics approach to the discovery of causal networks for any gene, metabolic pathway, process, or trait of interest, with applications across a wide range of problems in biology and medicine.

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

**Experimental: “NetWalk”: A rapid and reliable method to validate TF🡪targets genome-wide (Aim 1).** TF perturbation studies are essential to generate TF🡪target data needed to train and validate network inference approaches. The “NetWalk” transient expression system enables us to perturb the function of a TF and rapidly discover its effect (induction or repression) on its immediate targets. This system that has the following advantages compared to traditional reverse genetic studies: (i) *High throughput*: it accelerates TF perturbation studies (to weeks) compared to classical genetic approaches (4-6 months), (ii) *Circumvents TF Redundancy*: it avoids problems with TF redundancy encountered using mutants [Chen BMC Evol Biol. 2010] [Cutler & McCourt P (2005)], because it gives the direct effects of a TF on its children. (iii) TF🡪target *Genome-wide data*: it experimentally validates TF🡪 target data for any gene. [Gloria: I don’t like this genome-wide adjective because each TF🡪target data has to do with immediate descendants]. This TF🡪target validation data will be used both as validation for the networks inferred in Aim 2 and the networks will suggest which TFs to test.

**Computational: “Pipelined Network Inference”: An integrated network inference approach (Aim 2).** Current network inference methods are designed to accept specific genomic data-types - *either* steady state *or* time-series experiments. To improve the predictive power of our inferred networks, we will develop a computational pipeline that learns from multiple genomic data types (e.g. time-series data, steady state data, and TF perturbation data). [Bonneau (2007) Cell] [ Bonneau Genome Biol 2006] [Wang Y,. Bioinformatics 2006] [Shimamura. BMC Syst Biol 2009]. This pipeline will combine the DFG algorithm (Dynamic Factor Graph), a form of “state space” analysis we used previously for analysis of 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 used for 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. Importantly, as this approach is iterative, the networks inferred from this new computational pipeline will drive a new round of experimentation e.g. in **Aim 3 (e.g.** the testing of new TFs and TF interactions) which will refuel our network learning, in a true Systems Biology cycle [Gutierrez 2005].

**The N-assimilatory Network: A paradigm for metabolic regulatory network inference.** The N-assimilatory network is a valuable system to develop methods and models for metabolic regulatory networks, owing to its exquisite transcriptional regulation by nutrient cues, and its importance to plant growth and development [Ruffel S, 2010]. Inorganic-N taken up from the soil as nitrate, NO3, is assimilated into organic-N as Glu/Gln which also act as signals of N-status. Nitrate signaling 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 Glu/Gln signaling networks [Gutierrez 2008] [Rawat 1999], but the receptor(s) remain elusive [Hsieh (1998)][ Lam (1998)]. To identify the downstream components of inorganic (nitrate) or organic-N (Glu/Gln) sensing, we derived networks for organic-N regulated networks from steady state transcriptome data [Gutierrez 2007,2008; Nero 2009a,b], or dynamic network models based on machine-learning analysis of time-series transcriptome data for nitrate-regulated networks [Krouk 2010]. These networks support a regulatory model in which inorganic-N (nitrate), induces expression of TFs (e.g. HRS1, HHO1,2,3) that activate genes involved in nitrate uptake, reduction, and assimilation into organic-N (Glu/Gln) (Fig. XA). In turn, the N-assimilation products, Glu/Gln, act as signals to repress expression of TFs (e.g. CCA1 GLK1,WRKY1) that activate genes involved in Gln synthesis, thus constituting a negative feedback loop (Fig. X). Further, Glu/Gln induces expression of TFs (e.g. bZip1) partially through derepression, which in turn activates genes involved in the conversion of Gln🡪Asn, a more carbon-efficient amino acid used for N-transport/storage (Fig. XB). This regulation in response to N-metabolite signals may represent an “energy conservation” mechanism to save ATP, reducing equivalents, and carbon skeletons required for nitrate reduction and assimilation into Glu/Gln, when levels of organic-N are abundant. Indeed, altering transcriptional regulation of genes in the N-assimilatory pathway can be used to effect changes in N-use efficiency [Oliveira (2002)][ Lam, (2003)]. Thus, our inferred N-assimilatory networks may enable us to predict interventions to optimize N-use efficiency in plants.

**PROGRESS REPORT**: This report highlights progress most relevant to our Research Strategy. Our successes in Aim 4, “Creating a dynamic and predictive network model for the control of N-assimilation”, and in Aim 2 “Developing a high throughput method to validate TF🡪target interactions genome-wide”, have influenced approaches used in the other aims and therefore lead the report. ***Summary*:**  Our networks predicted based on gene expression data, aim to model the circuit by which the N-assimilatory network is regulated in response to N-metabolic cues. Our first network models were based on steady-state transcriptome data analyzed in the context of an Arabidopsis multinetwork [Gutierrez 2007], where TF->target “edges” were predicted based on correlation and *cis*-element representation [Gutierrez 2008][Nero 2009b]. That approach enabled us to successfully model a network controlling the N-assimilatory pathway in response to organic-N sensing, and to test TF🡪 target predictions using transgenics [Gutierrez 2008] or mutants [Obertello 2010] (Aim 1). A new approach developed during the current NIH cycle, was the generation of time-series data (Aim 2B), which enabled us to use machine learning approaches to model dynamic and predictive network models (Aim 4) for the control of N-assimilation in response to nitrate sensing [Krouk 2010]. Predictions from this “learned” network, were validated by predicting the expression of genes in an experiment that wasn’t used in the inference of the network and then comparing the predictions with that left-out experiment [Krouk 2010]. To accelerate our validation of TF🡪target predictions genome-wide, we developed a rapid high through-put system to transiently perturb TF expression, and to monitor changes in target gene expression within 2 weeks time (Aim 2A) [Bargmann et al, 2012, submitted]. This TF perturbation system involves transient TF overexpression (as a 35S::GR::TF fusion) in protoplasts, selection of successful transformants by FACS [Bargmann 2009], selective triggering of TF nuclear localization with Dexamethasone (DEX), followed by assays of target gene activation (by Q-PCR or transcriptome). This system successfully uncovers changes in expression of direct TF targets, as validated using transcriptome analysis of a TF (ABI3) for which Gene Regulatory Networks are well known [Vernoux 2011]. We recently applied this system to validate TF🡪target predictions for TF hubs involved in nitrate [Krouk et al 2010] or organic-N [Gutierrez 2008] regulation of the N-assimilation pathway (Aim 1). The DEX system underpins our “Network Walk” method in this grant. Lastly, our studies of post-translational responses to nitrate and Glu/Gln sensing were conducted collaboratively as part of a Fogarty Award t0 R. Gutierrez [Vidal 2010]. These studies are highlighted below.

**[Gloria: you can leave this out. Already discussed] Prior Aims 4 & 2B. Create a time-dependent dynamic network model for the control of N-uptake/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:***  Completion of Aims 2B and 4, enabled us to achieve one of the main goals of Systems Biology – predicting network states under untested conditions. Fine-scale time-series transcriptome data generated in (Aim 2B), analyzed using a machine learning approach called “State Space modeling” (Aim 4), generated a predictive N-assimilatory network model validated (using left-out data) experimentally [Krouk et al 2010].

**[Gloria: again, you can leave this out. Nobody cares about these details] Prior Aim 2B. Time-course analysis for validating causal TF-target relationships and beyond.** In this aim, we generated fine-scale time-series transcriptome data to drive machine learning of regulatory networks. Time-points assayed correspond to first time of sentinel gene induction (15- 20 min), and the preceding time points (0, 3, 6, 9, 12, min). Linear modeling identified 550 genes whose expression was regulated by nitrate, as a function of time [Krouk 2010]. Importantly, this list included >200 nitrate-regulated genes (some induced as early as 3 min) not previously identified in studies of steady state mRNA (20 min) [Wang 2004]. This time-series transcriptome data was used for network inference in Aim 4.

**Prior Aim 4. Create causal regulatory network based on time-course microarray data**. Because causality moves forward in time, time-series experiments are a particularly promising source of structure to derive predictive networks. Thus, time-series data, was analyzed using a machine-learning, network inference approach (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 genes at the previous time point) [Mirowski, P., et al. Clin Neurophysiol, 2009]. In a departure from previous 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). 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 up to three or four) (Fig. XA). 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 the20 min transcriptome. Surprisingly, based on these relatively few time points, the coherence of the regulatory network model generated is good enough that it is able to predict the direction of gene change (up or down regulation) 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.

Next, we 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) (Fig. XB) for validation testing. As our studies of T-DNA mutants were hampered by issues of functional redundancy, we 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 role of SPL9 over-expression evaluated over time, validated the vast majority of the genes in the regulatory network that we tested [Krouk 2010]. A set of TF hubs (e.g. HRS1, HHO1, 2 & 3) induced early in the cascade, and at the top of the hierarchy, were prioritized for TF perturbation in transient assays described below, and preliminary evidence supports their role in nitrate signaling (see Aim 1B, Research Strategy).

**Prior Aim 2A. Use genome-wide approaches to validate and discover causal TF->target relationships mediating N-regulation of N-uptake/assimilation pathway genes. *Relevant 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. This system was developed and validated using a well-studied TF (ABI3), for which targets identified in our system could be validated by external data.

**Background**: A critical component of Gene Regulatory Network (GRN) inference is experimental validation of TF🡪target predictions. Chromatin immunoprecipitation followed by sequencing analysis (ChIPseq) 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)]. Therefore, ChIP analyses are often combined with genome-wide transcriptional analysis of samples in which TF function is perturbed [Zhu 2012]. Perturbing TF function using reverse genetics is problematic, as functional redundancy of TFs [Cutler and McCourt 2005] appears to be a built-in feature of GRNs that makes them robust. As an alternative approach, 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 or regulated nuclear entry of the TF [*e.g*. Hachez 2011, Bustos 2010]. Indeed, regulated nuclear translocation of a TF, combined with the use of protein synthesis inhibitors, has previously been used to eliminate the effects of secondary transcriptional regulators, and to filter direct from indirect TF targets [Bustos]. Genes whose expression is affected by TF perturbation can then be validated as direct targets using *cis*-element analysis and ChIPseq to build network views of TFs and their targets [Zhu 2012].

**The “NetWalk” System**: To conditionally “activate” a TF using a high-throughput system, we adapted our transient expression system [Bargmann et al 2009] which uses a positive fluorescent selection marker (RFP) to drive Fluorescence Activated Cell Sorting (FACS) of successfully transformed protoplasts. The use of a GatewayTM expression vector make it suited for high-throughput analyses. This system was adapted to overexpress any TF, selectively induce its translocation into the nucleus, and identify changes in network target genes expression based on RNA analysis (Q-PCR or transcriptome). Technically, protoplasts are transfected with the expression vector a TF fused to the rat glucocorticoid receptor (GR), which harbors the TF-GR protein in the cytoplasm, and treatment with the GR-ligand dexamethasone (DEX) induces nuclear localization [REF]. RNA analysis of +DEX/-DEX performed on using 35S::TF-GR constructs vs. empty vector controls, identifies genes whose expression is affected (induced or repressed) following TF nuclear localization with DEX. The effects of the translocation of the TF into the nucleus by DEX can be measured in the presence of the translational inhibitor cycloheximide (CHX), to identify 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 “NetWalk” transient TF perturbation system was used to investigate the genome-wide targets of the well-studied TF, ABI3 (ABSCISIC ACID INSENSITIVE3). ABI3 is a master regulator in the ABA signaling pathway, and can validiate the TF🡪target data generated using NetWalk. Following the transformation of protoplasts with a 35S::ABI3-GR expression vector, and FACS selection of positive transformants [Bargmann 2009], sequential treatments (+/-CHX, to block translation) followed by (+/-DEX, to effect nuclear localization) were performed. RNA analysis identified genes whose expression was altered by DEX treatment in the ABI3::GR constructs, compared to empty vector controls. This study showed that known ABI3 target genes (e.g. CRU1, PER1) are direct targets (e.g. up-regulation of RNA levels in +DEX/+CHX), where CHX treatment potentiates the induction of these direct targets, potentially by eliminating repressors (Fig. X). Further, genome-wide transcriptome analysis confirmed ABI3 direct targets to be induced by +DEX/+CHX compared to empty vector based on: (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 the identification of direct ABI3 TF targets using this transient expression “NetWalk” system, suitable for systematic GRN analyses of any TF in *Arabidopsis* (or other species).

**Prior Aim 1. Test hypotheses for TF-Target interactions controlling N-uptake and assimilation. *Revelant 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. ***Summary*:**  In the previous cycle, we identified regulatory networks and TFs associated with N-assimilation data using a multinetwork analysis [Gutierrez 2007], where TF🡪target edges were predicted based on correlation of expression (>0.8) and over-representation of *cis*-elements [Gutierrez 2008][Nero 2009a; Nero 2009b]. The regulatory network model, shown in (Fig. X) indicates a TF regulatory mechanism by which organic-N can repress further synthesis of Gln, and induce synthesis of Asn, used for N-storage. While some components were validated using overexpresors (e.g. CCA1), by and large, our reverse genetic studies of predicted TFs in this N-assimilatory network have been hindered by issues of TF functional redundancy [Obertello 2010] [Krouk, unpublished]. Experimental confirmation for the TF🡪target relationships has required transgenic overexpressors and double mutants as follows: CCA1 (35S::CCA1) [Gutierrez 2008], GLK1 (glk1/2), WRKY1 (wrky1, and 35S::WRKY1), and bZip1 (transient expression) [BaenaGonzales XXX]. Because constitutive overexpression in transgenic plants cannot be used to identify direct TF targets (see Prior Aim2A above), we prioritized these TFs for perturbation studies in the transient “NetWalk” system, and preliminary results are in Aim 1B of the 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 also employs 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 an miRNA-TF module was regulated in response to sensing N-metabolites in the environment (e.g. inorganic nitrate) or internally (e.g. assimilated N, Glu/Gln) [see Vidal 2010]. The N15 tracer experiments - are currently being conducted in collaboration with Drs. Ruffel, G. Krouk and Pascal Tillard at INRA (Montpellier, FR) using a XXXXXX analyzer to determine if perturbations in TFs associated with the N-assimilation pathway alters N-use efficiency.

**Research Strategy: Overview**: This plan builds on our prior success in inferring regulatory networks able to predict gene responses under untested conditions [Krouk 2010]. We will expand and improve the predictive power of our regulatory networks by generating new high throughput data, followed by new computational modeling, in an iterative cycle. In Aim 1, we will perturb expression of TFs using the high throughput transient expression system, and validated TF🡪target datasets will drive a new machine learning pipeline in Aim 2. This pipeline is novel, in that it combine analysis of distinct genomic datasets (time series, steady state and TF perturbation data) into a single combined network inference analysis, to improve the predictive power of our networks, and to identify new targets and TF interactions for perturbation studies in Aim 3. Those data will turn drive a new cycle of analysis, model refinement, in the true systems Biology cycle (Fig X). Our Aims are:

**Aim 1. High throughput experimentation**: “NetWalk”: Generation of high throughput, high confidence TF🡪target datasets for TF hubs in the N-assimilatory network.

**Aim 2. Computational Modeling**: “Pipelined Network Inference”: A novel 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.** **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: “NetWalk”: Generate high throughput, high confidence TF🡪network target datasets to drive network inference**. ***Rationale***: In this aim, we will generate new types of data to fuel and validate our new network inference pipeline in Aim 2. The novelty of this approach is that it will combine three distinct data-types into one learning pipeline; (i) time-series data, (ii) steady state transcriptome data, and (iii) TF->target data from perturbation. For the TF perturbation studies, we will use “NetWalk”, a transient, high throughput system to transiently perturb TF activity and monitor targets genome-wide using transcriptomics and ChIP-seq. In each subaim, we provide preliminary data for one TF as a proof-of-principle example for the feasibility of the approach. For selected TFs, we will also examine perturbations *in planta* using mutants and/or transgenics. The diverse datasets supporting TF🡪target relationships in Aim 1 will be combined to fuel a single network inference pipeline in Aim 2.

**Aim 1A. Generate new time-series transcriptome datasets for organic-N signaling.** Our time-series transcriptome data from nitrate-treated plants (i) identified >200 new nitrate regulated genes (compared to steady state data) [Wang 2004], and (ii) enabled us 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]. This new time-series data for organic-N signaling, should enable us to uncover new genes including TFs that are induced early (e.g. within minutes), and to generate dynamic network models [Krouk 2010]. This data will be used in the machine learning pipeline of Aim 2 to build a network that can identify TF hubs among the first TFs to be induced.

**Aim 1B. “NetWalk”: Generation of high throughput, high confidence TF🡪network target data. Prioritization of TF hubs for transient perturbation studies.**  We previously identified TF hubs associated the N-assimilatory network response to inorganic (nitrate) [Krouk 2010] or organic-N [Gutierrez 2008] signals. Identification of TF hubs (CCA1, GLK1/2, WRKY1, bZip1) controlling the organic-N network, were based on TF🡪target correlations (>0.8) using steady-state transcriptome data, and representation of *cis*-elements enabled us to identify and validate some of our predictions [Gutierrez 2008] [Nero 2009a]. By contrast, time-series data analyzed using machine learning methods enabled us to generate predictive networks, using left-out data [Krouk 2010]. To increase our throughput validation testing of the network, we selected TF hubs (e.g. HRS1, HHO1,HHO2, HHO3) induced “early” and at the top in the hierarchy [Krouk 2010]. Surprisingly, the TF hubs associated with nitrate vs. organic-N signaling, correspond to two distinct subclades of a single myb gene family, a member of which was previously associated with phosphate signaling [Liu 2009] (Fig. XA). This unexpected finding suggests biological relevance in nitrogen signaling, supported by preliminary perturbation studies detailed below.

**“NetWalk”: Generation of TF🡪network target data using transient TF perturbation.** To identify network targets of a TF genome-wide, we will use the transient [Gloria: we want NetWalk to be new otherwise reviewers will say there isn’t anything new] system described more fully in Progress (Prior Aim2A) and in [Bargmann et al 2012.] Briefly, “NetWalk” consists of the following components; i) transient expression of a 35S::GR-TF fusion in protoplasts (the vector expresses RFP as transfection control allowing FACS sorting of the transfected cells), ii) use of an inducible dexamethasone (DEX) system to artificially control the timing of the translocation of the TF into the nucleus, iii) use of the translation inhibitor cycloheximide (CHX), to identify primary targets of the TF. This approach is an adaptation of the original DEX-induction technique [ Sablowski, 1998.] combined with transcriptome analysis as described in [ Hanson, 2008] [Wang, 2006.]. In addition, we push the DEX-concept further, using N-treatments to precondition cells to allow N-regulated post-translational modification, as well as N-regulated transcriptional activation of cofactors. This system overcomes the issue of TF redundancy encountered in reverse genetic approaches (see Aim 1D). In addition, the “inducible” nature of the system enables identification of direct TF targets, not possible in 35S::TF transgenic plants (see discussion in Prior Aim2A). The TF🡪target data from transcriptome analysis will be complementary to Chromatin-IP (ChIP) data we collect (in Aim 1C) and has several advantages: (i) ChIP (coupled with Chip-Seq) can confirm protein:DNA binding, but it does not guarantee 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].

We previously describe validation of the “NetWalk” system for a well-studied TF ABI3 (Prior Aim2A) [Bargmann 2012]. Here, we describe a ***proof-of-principle*** example of a TF in the nitrate-regulatory network, which shows that (i) nitrate-preconditioning affects target gene activation and (ii) targets uncovered in this transient assay have relevance to studies in whole plants.We initially hypothesized a role for HRS1 in nitrogen signaling, as its expression 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 in the N-assimilation network [Krouk et al 2010]. Interestingly, HRS1 was previously associated with phosphate signaling, based on phenotypes of transgenic 35S::HRS1 plants [Liu 2009]. Our initial analysis of transcriptome data from 35S::GR-HRS1 activation 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). [The following is not a sentence:] Cluster 4 genes, defined as primary targets of HRS1 (e.g. activated by +DEX+CHX) as specifically induced only in the *presence* of NO3-. Interestingly, GO term analysis shows significant overrepresentation of genes involved in phosphate transport (p-val 8.14 E-6) in the HRS1 network targets, recapitulating the phosphate signaling phenotype observed in 35S::HRS1 overexpressing plants [Liu 2009]. This result is strong evidence that our results in the transient DEX protoplast system have relevance to whole plants.

The ***prioritized TFs*** will be subjected to transient perturbation studies using the NetWalk system, and Fig. X shows progress in Gateway cloning and DEX-induction. The conditions for +/-N pre-treatments (e.g. prior to DEX-induction of nuclear translocation) will address how nitrate signaling (growth in ammonium succinate, treatment with nitrate) or organic-N signaling (growth in 1mM nitrate, and treatment with 40mM ammonium/nitrate) affects TF activation of target genes. Transcriptome analysis will identify specific genes, gene clusters and biological processes controlled by each TF (see Fig X). Direct targets (+CHX +DEX) will also be confirmed by ChIP studies (in Aim 1C).

**Aim 1C. Genome-wide validation of TF🡪network targets using Chip-Seq.** To test the feasibility of performing ChIP-Seq and transcriptome analysis on the same samples, we performed a ***proof-of-principle*** study. Analysis of this data enabled us to identify a map of TF binding sites for bZip1, using chromatin-immunoprecipitation in the transient expression system. bZip1 was selected for these studies, as it was validated to directly activate predicted target genes in our network (e.g. ASN1) (Fig. X), as well as in independent studies [Baena-Gonzales]. We adapted the micro-ChIP protocol from [Dhal and Collas, 2008], which requires a relatively small number of cells (1,000 cells) to conduct both transcriptome and ChIP analysis from the same samples. After transformation of protoplasts with the 35S::GR::bZip1 construct, transformed protoplasts selected by FACS (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 from 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 to facilitate library preparation from 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 (the DNA without immunoprecipitation) as control. The libraries were sequenced on the Illumina GAIIx platform.

***Bioinformatic analysis*** of deep-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 (Fold change >2) compared to the background (input DNA), indicating significant binding (p<1E-8) of bZIP1. To validate that the ChIP samples were *direct* bZip1 targets, we performed an analysis of cis-regulatory motifs. 500bp immediately upstream of the 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 of bZip1🡪target data from ChIP studies, we intersected this list of genes with the bZip1 targets predicted from our multinetwork analysis (e.g. based on correlation > o.8 and over-representation of 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: CCA1, GLK1, WRKY1 whose regulation by bZip1 is depicted in Fig. X. [Gloria: it is really not clear what CHIP brings to the party. Why spend resources on this when we already get the direct targets]

**Aim 1D. TF perturbations in whole plants.** For selected TFs, we will validate TF🡪targets using TF perturbations *in planta* using mutants or transgenics, as follows. **T-DNA mutants**: As mentioned above, TF redundancies underlie the robustness of the N-assimilation network, soT-DNA mutants in TFs could not be used to validate TF🡪target relationships for bzip1 [Obertello 2010], glk1 [Gutierrez 2008], hrs1, hho1, 2 and 3 [Medici and Krouk, unpublished] (Table X). In the case of WRKY1, a TF predicted to act as a toggle switch between activating genes involved in nitrate assimilation vs. its conversion into Asn (see Fig. X), three independent T-DNA wrky1 mutant lines (SALK\_016954; SALK\_136009; SALK\_070989) showed decreased expression 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). Interestingly, the opposite expression patterns are observed when WRKY1 is transiently over-expressed in protoplasts, see Aim 1B. Single and double mutants (e.g. glk1/2, hho1/2, cca1/lhy1) that show alteration regulation of targets in the N-assimilation pathway, will be used in TF interaction studies described in Aim 3. **Transgenic lines**: 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], 35::GLK1 [XXXX], 35S::HRS1 [Liu 2009]. For selected TFs, 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)].

**Data integration and analysis**: The distinct sets of data generated in Aim 1 (i) time-series transcriptome, (2) steady state transcriptome, (iii) data from TF perturbations, and (iv) 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” 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). For each subaim (Aim 1B, C, and D) we have working examples to demonstrate the feasibility of the approach, so we do not anticipate technical problems in data generation or interpretation. In addition, the coordinated analysis of distinct data sets (e.g. TF🡪target identification based on transcriptome (Aim 1B) vs ChIP seq (Aim 1C) will cross-validate our TF🡪targets identified.

**Aim 2. Computational Innovation: “Pipelined 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. 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 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 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 Pipelined Network Inference approach 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 Pipelined 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, (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 (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 2012], we show using data from the DREAM benchmark (Dialogue for Reverse Engineering Assessments and Methods) [Schaffter, (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 Pipelined.eps, 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 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 of the Pipelined Network Inference Approach:** 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 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 regulated members of the 66 genes in the N-assimilatory pathway, and 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 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). 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 outcome Aim 2 will a “learned regulatory network” for TFs controlling the N-assimilatory pathway, built/trained on a pipeline analysis of data sets (from Aim 1) that will improve 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. 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 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.

**Lessons from Machine Learning: (Gloria likes manny’s version better [Ok, we will decide later).**

**Manny version**: From machine learning, we will discover which data type (time-series, steady state, TF perturbation) is most useful for network inference, and also which is not necessary. 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.

**DENNIS version: For example, our preliminary** results show that closely spaced time series experiments can lead (through appropriate machine learning algorithms) to a causal network model having quite accurate predictions about untested conditions. Adding in steady state data helps a bit, but not that much, suggesting that causal edges are difficult to tease out of pure steady state experiments. [we should probably try this] We believe that Dex data will help us more, because .... [fill in after we try this]. Steady state data still plays a role as we have found in previous work in our lab.It can lead to inferences about the importance of particular genes that can then be validated. So the usefulness is to give

qualitative hints of the form: try gene x.

**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 spirit of the iterative Systems Biology approach [Gutierrez 2005], the network models from Aim 2, will in turn predict new/refined network models, and new TF perturbation experiments to perform in Aim 3. Our network models in Aim 2 will also indicate whether TFs may act in concert (or antagonistically) in controlling the expression of the genes in the N-assimilatory network. In the iterative cycle, the results from Aim 3 will in turn be fed back to refine the network models in Aim 2.

**Approach**: We will test TF🡪target relationships for new candidate TF hubs identified in Aim 2 using the transient “NetWalk” DEX-inducible system described in Aim 1B. To test for TF1-TF2 cooperation, we will use two complementary approaches: (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, as described below.

**Aim 3A. Perturbation studies for new candidate TFs**: In this subaim, we will functionally validate TF🡪targets (using approaches described in Aim 1) for new TFs identified in Aim 2. These may include TFs that are hubs in steady state or time series experiments or TFs that are believed to affect critical genes. As we observed in our nitrate studies, [Krouk et al 2010] time series experiments for organic-N networks may uncover TFs which were not found to be regulated in steady state studies.

**Aim 3B. Testing TF cooperation in the N-assimilatory network.** Our refined network models generated in Aim 2 will suggest TF pairs that may cooperate to regulate target genes in the network. We will use two complementary approaches to test TF cooperation: (i) co-expression of TF1/TF2 using the DEX-protoplast system (see Aim 1B) & (ii) using DEX-TF1 in combination with genetic perturbation of tf2, as described below.

 **Co-expression of TF1 & TF2:** We will select examples where two TFs cooperate (additively or synergistically) to induce each other or control a target gene (e.g. GLK1🡨🡪CCA1,) or antagonistically (e.g. WRKY1--|bZip1) (see Fig. X), by identifying target genes whose expression is regulated significantly differently in response to transient activation by DEX using the TF1/Tf2 double expression vector, compared to either TF expressed alone in a single vector. In order to co-express two TFs (TF1 and TF2) in the same cell, we are using the MultiSite Gateway**®** Technology that allows for simultaneously 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) were cloned in appropriate Gateway donor vectors, and allowed to recombine with each other and with the destination vector pBob11 [Bargmann 2009]. 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.

**Perturbation of TF1 activity in a tf2 mutant background**: As a complement to the multisite vector approach, we will explore TF1/TF2 interactions by transiently expressing the single 35S::GR:TF1 in protoplasts made from a mutant in tf2, and will compare target gene activation to TF1 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) will be interpreted as evidence for TF1 and TF2 cooperation and/or interaction. This functional evidence based on target gene regulation could involve direct or indirect TF interactions including (e.g. protein-DNA interaction, e.g. one TF is the target of the other). We will use this genetic approach for TFs that show a molecular phenotype (e.g. loss of target gene expression) in the single T-DNA mutant (e.g. wrky1) to explore possible TF cooperation (e.g. wrky1/CCA1) or antagonism (e.g. wrky1/bZip1). We may also use protoplasts made from double mutants (e.g. glk1/2) to explore TF cooperation in transcriptional networks (e.g. glk1/2/CCA1).

**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 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 (e.g. to quadratic equations) in cases where the TF1/TF2 interaction data shows synergistic effects for a target gene, 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 an additive or synergistic fashion.

**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.

**TIMELINE**

**Year 1-2**

**Year 3-4**