**(a) Significance:**  Our goal is to model a causal genetic network, effectively the circuit diagram underlying the pathway controlling nitrogen-use efficiency in plants. Our approach is to iterate (loop) between experimental work and machine-learning approaches to network inference. These “learned” network models will be tested for their ability to infer network states under untested conditions including perturbations. A practical goal is to guide targeted interventions in important pathways in biology, agriculture, or medicine. Despite its promise, determining predictive Gene Regulatory Networks (GRNs) from comprehensive gene expression sets, remains both a major goal and a major challenge in fundamental and applied research. 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, Advantages and limitations of current network inference methods

Riet De Smet and Kathleen Marchal in Nature Review]. Despite these caveats, machine learning approaches have successfully been used to predict network states in microbes [Bonneau Cell 2007], and more recently in higher eukaryotes including plants [REFS we should use our own paper]. During the past NIH funding cycle, we applied a machine learning approach applied to time-series transcriptome data, and to infer a network that succesfully predicted gene expression states under untested conditions [Krouk 2010]. In this renewal, we propose to build on and improve the predictive power of our inferred networks, using a combination of innovations in 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 and that overcomes problems with TF redundancy. This data is crucial to the training and validation of our new network inference pipeline. Individual network inference methods are designed to accept specific genomic data-types - either steady state or time-series experiments. To address this, we will create a pipeline that can analyze multiple types of genomic data (e.g. time series, steady state, and mutant) using a combination of algorithms, which should improve predictive power of our inferred networks, as suggested by our preliminary studies. Importantly, our combined innovations in experimental and computational are inherently iterative: high throughput experimentation feeds computational learning, which feeds experimentation, and so on.In our study, the resulting N-assimilatory network can suggest targeted interventions to reduce nitrogen fertilizer usage, with implications for human health, energy and the environment. More broadly, this 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’s Courant.

**Experimental: A rapid and reliable method to validate TF🡪targets genome-wide (Aim 1).** TF perturbation studies are essential to generate TF🡪target data to train and validate network inference approaches. Our transient system described in Progress and Aim 1 of the plan, will enable us to rapidly validate TF🡪targets geneome-wide with the following advantages: (i) *Genome-wide data*: generation of genome-wide TF🡪target data to “train” and validate machine learning of regulatory networks, (ii) *High throug put*: it accelerates TF perturbation studies (to weeks) compared to (4-6 mos) for transgenics, and (iii) *Circumvents Redundancy*: it surmounts problems with TF redundancy [Chen BMC Evol Biol. 2010] [Cutler & McCourt P (2005)]. This genome-wide TF🡪target validation data will be used to drive and validate the networks inferred in Aim 2.

**Computational: “Pipelined Network Inference” an integrated network inference approach (Aim 2).** In this renewal, 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 MCZ algorithm (short for the Median Corrected Z-score method) a simple but successful approach to infer the effects of TF perturbations [Greenfield & BOnneau ??], with the DFG algorithm (short for Dynamic Factor Graph), a form of “state space” analysis we used previously [Krouk et al 2010] for analysis of time-series transcriptome data. Preliminary results suggest this new pipelined approach will increase the quality of our predictions, as judged using out of sample data. Importantly, our approach is iterative cycle of computation and experimentation, as the new networks inferred from this computational pipeline, will drive a new round of experimentation in **Aim 3**, that will refuel our network learning, in a true Systems Biology cycle [Gutierrez 2005]. [Aim 3 is not described here]

**PROGRESS REPORT**: This report highlights progress that is most relevant to our Experimental Plan. Our successes in “Creating a dynamic and predictive network model for the control of N-assimilation” (Aim 4), and in developing a high through put method to validate TF🡪target interactions genome-wide (Aim 2), have influenced approaches used in Aims 1 and 3, and the flow of the report.

**Summary:**  During this past cycle, our combined advances in experimentation and computation have enabled us to accomplish a major goal in Systems Biology – predicting network states under untested conditions. Our studies pertain to modeling how nitrogen metabolites act as signals of inorganic-N (nitrate) or organic-N (Glu/Gln) to regulate genes in the N-assimilatory network in response to N-availability. As reference, our first network models, based on steady-state transcriptome data, were 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]. That approach enabled us to derive and test hypotheses for TFs involved in regulation of N-assimilation pathway in response to nitrate [Nero 2009b] or organic-N sensing [Gutierrez 2008], several of which were validated. In the current cycle, we used time-series transcriptome data (Aim 2B), and a machine-learning approach (State-Space Modeling) (Aim 4) to create a dynamic, time-dependent model for transcriptional control of the N-assimilatory pathway in response to nitrate sensing [Krouk 2010]. Predictions from this “learned” network were validated both *in silico* (using out-of-sample data) and experimentally (for selected TFs) [Krouk 2010]. To accelerate our validation of TF🡪target predictions genome-wide, and to overcome problems with TF redundancy, we developed a rapid high through-put system to perturb TF expression and monitor changes in target gene expression within weeks (Aim 2A) [Bargmann et al, 2012, submitted]. This system enables one to transiently overexpress any TF of interest (as a 35S::TF::GR fusion) in protoplasts, to select successful transformants by FACS [Bargmann 2009], and to selectively induce TF nuclear localization upon treatment with Dexamethasone (DEX). Transcriptome analysis identifies changes in expression of TF targets, and a TF (ABI3) for which Gene Regulatory Networks are well known [Vernoux 2011], was used to validate this system [Bargmann et al 2012]. We have recently applied this system to validate TF🡪target predictions for hubs involved in nitrate [Krouk et al 2010] or organic-N [Gutierrez 2008] regulation of the N-assimilation pathway (Aims 1A/B), enabling us to overcome problems with TF functional redundancy encountered in our previous studies of T-DNA mutants [Obertello 2010]. Finally, our studies of miRNA-TFs (Aim 3) also uncovered evidence for post-translational control by nitrate or organic-N signals [Vidal 2010].

**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:***  Our 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), was analyzed using a machine learning approach called “State Space modeling” (Aim 4), to generate a model for the control of N-assimilatory network able to predict gene responses under untested conditions. These predictions were validated (using left-out data) and experimentally [Krouk et al 2010].

**Aim 2B. Time-course analysis for validating causal TF-target relationships and beyond.** In this subaim, we produced time-series transcriptome data to drive machine learning. Time points assayed, corresponded to the first time point at which sentinel target genes in the N-assimilation pathway were induced (15- 20 min), and the preceding time points (0, 3, 6, 9, 12, min). Linear modeling identified 550 genes whose expression was regulated in response to nitrate, as a function of time, as detailed in [Krouk 2010]. Importantly, this list included >200 nitrate-regulated genes (induced as early as 3 min), not previously identified in steady state studies (20 min) [Wang 2004]. This data was used for network inference in Aim 4.

**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 generated in Aim2B, 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-reduction 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. X). 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 from Aim 2B (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 “left-out” data (e.g. transcriptome data from 20 min). Surprisingly, the coherence of the regulatory 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 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 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, the “most influential TFs” in the network (e.g. the one that is predicted to influence the most genes in the network) were selected for validation testing. Owing to problems we encountered with functional redundancy using T-DNA mutants, we turned to TF overexpression constructs. For the master TF gene, SPL6, we examined overexpression of mRNA in 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 (see Additional file 5) [Krouk 2010]. As transgenics for other TFs in the network were not widely available, we developed a high throughput method for TF perturbation in Aim 2A.

**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 paper describes a rapid transient expression system that enables identification of direct TF targets in genome-wide within 2 weeks. This system was tested using a well-studied TF (ABI3), for which detected targets 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]. As redundancy of TFs appears to be a built in feature of GRNs (that makes them robust), reverse genetic studies of the TF effectors are notoriously plagued by functional redundancy issues [Cutler and McCourt 2005]. Alternately, transcriptomic analysis of transgenic TF over-expressors (OE) can give an indication of the targets [Suzuki et al, Reeves] but it is not clear whether changes in transcript levels are a direct consequence of TF manipulation, or caused by indirect or possibly pleiotropic effects. A better approach, is to use an inducible form of the TF, either through conditional expression or regulated nuclear entry of the TF [*e.g*. Hachez 2011, Bustos 2010], which can give a more reliable indication of the direct targets. Use of protein synthesis inhibitors, together with regulated nuclear translocation of a TF, has also been used to eliminate the effects of secondary transcriptional regulators and to filter direct from indirect TF targets [Bustos]. Analysis of direct targets by Gene Ontology (GO) studies, cis-element analysis, as well as ChIPseq can be used to build network views of TFs and their targets and learn about their function. We have developed a high through put version of this system.

**The System**: We adapted our transient expression system [Bargmann et al 2009] to overexpress a TF, selectively induce its entrance into the nucleus, and identify its targets based on transcriptome analysis. When performed in the presence of a protein translation inhibitor CHX, only direct TF targets are activated. This molecular assessment of these direct TF targets is then complemented with computational analysis with respect to the putative regulation of target genes by shared *cis*-regulatory elements (CREs). As a test case, we investigate the GRN of the canonical abscisic acid (ABA) signalling pathway under the influence of the transcription factor ABSCISIC ACID INSENSITIVE 3 (ABI3). Technically, plant protoplasts are transfected with a vector that expresses the TF-of-interest fused to the rat glucocorticoid receptor (GR), which allows the controlled entry of the chimeric TF into the nucleus by addition of the GR-ligand dexamethasone (DEX) [REF]. Furthermore, the plasmid contains a separate expression cassette with a positive fluorescent selection marker (red fluorescent protein [RFP]) used to drive Fluorescence Activated Cell Sorting (FACS)- of successfully transformed protoplasts (Fig X) [Bargmann, 2009]. This cell-sorting step, selects successfully transfected cells allowing reliable qPCR or transcriptomic analysis of multiple independent transfections. Cloning of a TF-of-interest into the plasmid is achieved through GatewayTM recombination, suited for high-throughput analyses. Lastly, the effect of induction by DEX is measured in the presence or absence of the translational inhibitor cycloheximide (CHX), allows for the distinction of direct target genes of the TF-of-interest (+DEX/+CHX). (for details of this concept see [ Lloyd et al (1994) *Science*.] [Sablowski RW, Meyerowitz EM (1998). *Cell*]. Controls are empty vector

**The Results**: As a proof-of-principle, this transient system was used to investigate the genome-wide targets of the well-studied network targets of the TF, ABSCISIC ACID INSENSITIVE (ABI3) a master gene in the ABA signaling network [Vernoux *et al*, 2011 Mol Syst Biol. 2011 Jul 5;7:5089). Following transient expression of a 35S::AB13-GR construct and FACS selection of positive transformants [Bargmann 2009], treatments (+/-DEX= to induce nuclear localization) and, (+/-CHX) to block translation. Transcriptome analysis was performed to identify genes whose expression was altered by DEX treatment. Controls are empty vector. This analysis showed that known ABI3 target genes (e.g. CRU1) are direct targets (e.g. up-regulated in +DEX/+CHX) (Fig. X). An analysis of all direct target genes (e.g. induced in +DEX/+CHX) compared to empty vector show: (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). Additionally, promoter analysis using MIME [Bailey and Elkan, 1994) ], yields *de novo* recovery of putative ABI3 CREs, including the ABRE as a top candidate. These findings demonstrate and validate the application of this novel technique for an investigation of the ABI3 network at a genomic level. Moreover, the experimental method and analysis employed should prove a useful tool for further systematic GRN analyses in *Arabidopsis* as well as other organisms.

**Aim 1. Test hypotheses for TF-Target interactions controlling N-uptake and assimilation. Revelant Publications: Nero et al 2009a, 2009b, Obertello et al 2010. Summary:**  In the previous cycle, we identified networks involved in organic-N regulation of N-assimilation pathway genes, using steady state transcriptome data in a multinetwork analysis where TF🡪target edges were predicted based on correlation of expression (.0.8) and representation of cis-elements [Gutierrez 2008]. In this organic-N regulatory network, the TF hubs (CCA1, WRKY1 and GLK1) are postulated to activate genes involved in assimilating nitrate into Gln (e.g. NiR, Gln1.3), and to repress TFs (e.g. bZip1) predicted to activate genes involved in Gln🡪Asn conversion (e.g. ASP3, ASN1) (Fig. X). Our analysis of T-DNA mutant uncovered functional redundancies [Obertello 2010], and we have validated the TF🡪target predictions as follows: 35S::CCA1 [Gutierrez 2008], glk1/2 [Para-Gaillo, unpublished], transient expression (bZip1) [Baena-Gonzalez 2009], and in rare cases using single T-DNA mutants (for WRKY1) [Colon, unpublished]. In each case, Q-PCR of a predicted target gene was used for validation. To advance our studies of the TF🡪targets to the genomic level, we have implemented the DEX-transient system described in Aim 2B (above), see (Table X) for progress.

**Aim 3. Determine the role(s) of post-transcriptional mechanisms in mediating N-regulatory network. Relevant publications: Vidal et al 2011. Summary**: This aim exploited the ability of the multinetwork database to integrate post-transcriptional datasets (miRNA) and proteomic data into our analysis of transcriptional regulatory networks. We employ 15N labeling to monitor N-assimilation products over time in wild-type and in transgenic lines altered in regulatory components. We have conducted miRNA experiments in collaboration (R. Gutierrez, Chile), a Fogarty Award winner connected to this grant. The resulting networks identified miRNAs involved in regulating changes in root architecture in response to nitrate vs Glu signaling (see Vidal 2011). The N15 labelling experiments are currently being conducted in collaboration with G. Krouk (Montpellier, FR) and are in progress.