**EXPERIMENTAL PLAN**

**Motivation & Significance**: Our ultimate goal is find a causal genetic network, effectively the circuit diagram underlying the regulation of genes in the N-assimilatory pathway. By analogy to an electrical network, such a gene regulatory network would enable us to infer the causal relationships between genes, thus predicting targeted interventions to reduce fertilizer usage in crops with implications for human health, energy and environment. Importantly, this work will illustrate an experimental/informatics approach to the discovery of the causal network for any gene (or potentially any trait) of interest a wide range problems in biology and medicine. In the true spirit of the Systems Biology cycle [Gutierrez 2005], the high through put data generated in Aim 1, will drive network modeling in Aim 2, and the resulting network models will in turn predict new experiments to perform in Aim 3. These predictions include the identification of new TF hubs for experimentation to refuel in the Systems Biology Cycle:

Aim 1. Experimental Innovation: Generation of high through put, high confidence TF🡪target datasets.

Aim 2. Computational Innovation: A Pipeline Machine Learning Approach to Finding Causal Links in regulatory networks using many kinds of genome-scale data.

Aim 3. The Systems Biology Cycle: The feedback from analysis to subsequent experiment.

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

**Aim 1. Generation of high throughput, high-confidence TF🡪target data for hubs in N-assimilatory networks.**

***Rationale*:**  In this Aim, we use a high throughput approach to perturb the expression of TF hubs associated with N-assimilatory networks and examine alterations in genome-wide target responses. This data will enable us to (i) identify TFs that control N-assimilatory genes in response to nitrogen sensing, (ii) identify biological processes and pathways that are coordinated with N-assimilation via these TFs, and (iii) produce high-confidence TF🡪target data to refine and train the machine learning approach to predict network states under untested conditions, the ultimate goal of Systems Biology.

**Approach**: Perturbations of TFs will be in a transient expression system in protoplasts using dexamethasone (DEX) to induce nuclear localization of the TF. DEX-induced nuclear localization of the TF will be performed in the presence or absence of cycloheximide (CHX), to distinguish activation of direct & indirect TF targets. We will analyze transcriptomic data as well as protein🡪DNA interaction data (ChIP-Seq) to enable us to identify direct and indirect targets of the TF. We call this approach “Network Walking.” The target gene data will be analyzed by clustering and GO term analysis tools we developed [Katari 2010] to identify which biological processes (e.g. bimodules) are controlled by each TF. This will enable us to identify how specific TFs coordinate N-assimilation with other biological functions in a plant, as an integrated system. For selected TFs, we will also test effects of TF perturbation in planta, using T-DNA mutants and/or overexpressors. The data from all TF perturbation studies will fuel the computational pipeline developed in Aim 2, to learn causal links in regulatory networks, and to generate predictive networks useful for targeted pathway intervention.

**Identification and prioritization of TFs mediating inorganic and organic-N signaling.** To date, our regulatory network models have identified TF hubs involved in regulating the N-assimilatory pathway genes in response to distinct N-signals [Krouk et al 2010][Gutierrez et al 2008]. Specific conditions for nitrogen pre-treatments and treatments have uncovered nitrate regulatory networks (growth in ammonium succinate, treatment with nitrate) or organic-N regulatory networks (growth in 1mM nitrate, and treatment with 40mM ammonium/nitrate) controlling genes in the N-assimilatory pathway [Wang 2004] [Gutierrez 2008]. The derived regulatory networks support a model in which inorganic-N (nitrate) signals induce TFs that activate genes involved in nitrate uptake, reduction, and assimilation into organic-N (Glu/Gln). These genes are in turn feedback-repressed by the organic-N products of N-assimilation (Glu/Gln). (This may represent an “energy conservation” mechanism to conserve ATP & reducing equivalents, and carbon skeletons required for nitrate reduction/assimilation into organic-N, when levels of organic-N are abundant.) These organic-N signals (Glu/Gln), in turn activate TFs and target genes involved in converting “reactive” Gln into “inert” Asn, a more carbon-efficient N-transport amino acid used for N-transport/storage (e.g. to seeds) (see Fig 1).

To identify the TFs associated with this N-regulatory network controlling N-assimilation, we used two types of network modeling approaches. TF hubs associated with nitrate-induction of N-assimilation genes were identified using time-series transcriptome data to drive a machine learning approach called State-space modeling [Krouk 2010]. The derived dynamic network models – which were validated using out of sample data - were used to prioritize TFs induced “early” and at the top in the hierarchy for experimental testing (e.g. HRS1, HHO1, HHO2, HHO3) (Table 1). TF hubs associated with organic-N regulation of N-assimilation genes (e.g. CCA1, GLK1/2, WRKY1, bZip1) were identified using the Arabidopsis multinetwork [Gutierrez 2007, 2008]. In this case, steady-state RNA data was used to predict TF🡪target edges based on correlation (>0.8) and representation of cis-elements [Gutierrez 2008]. Remarkably, the top-ranking TF hubs associated with nitrate-regulation (e.g. HRS1, HHO1, HHO2, HHO3) or organic-N regulation (e.g. GLK1 & GLK2) of the N-assimilatory pathway, comprise two sub-clades of a single myb gene family, for which one member (HRS1) was identified to be involved in phosphate signaling [Liu 2009]. This finding underscores the biological relevance of our network predictions, and suggests that members of this myb gene family may be involved in coordinating responses to nitrogen and phosphate nutrients in plants. This hypothesis is supported by our preliminary results of HRS1 expression in transient assays, where TF induction by DEX + N treatments, significantly induces genes involved in phosphate signaling (Aim 1B, Fig X). The finding that our DEX-inducible transient assay system in protoplasts recapitulates phenotypes observed in whole plants, attests to the biological relevance of the our system. We will test the TFs listed in Table X in this transient protoplast system to rapidly identify their genome-wide targets. Analysis of the TF targets data by transcriptome and ChIP assays, will enable us to (i) validate network predictions, (ii) identify biological processes co-regulated with N-assimilation, and (iii) provide TF🡪target data to feed our machine learning approach in Aim 2. In each of the subaims, we provide preliminary data using one of these TFs as a proof-of-principle example for the approach.

**Aim 1A. Generation of time-series transcriptome datasets for organic-N signaling.** Because causality moves forward in time, time-series experiments are a particularly promising source of structure for predictive network modeling. Our state-space modeling of time-series data from nitrate-treated plants, enabled us to generate a dynamic network model that was able to successfully predict the direction of change in expresion levels under untested conditions with 76% accuracy (e.g. whether a gene in the network would go up or down) [Krouk 2010]. By contrast, our network models for organic-N signaling were built using steady-state transcriptome data [Gutierrez 2008]. While those network models enabled us to identify and validated master regulatory genes (e.g. CCA1) involved in organic-N regulation of N-assimilation [Gutierrez 2008], we cannot use them to predict network dynamics or responses under untested conditions. To enable the generation of dynamic and predictive networks for organic-N networks, we will generate fine-scale, kinetic transcriptome data (0,3,6,9,12,15,20,25,30 min) using conditions shown to elicit an organic-N response (growth on 1mM nitrate, treatment with 40mM ammonium/nitrate vs. control KCl). This fine-scale time-series transcriptome data will be used to drive state-space modeling of regulatory networks in Aim 2, as described in [Krouk 2010]. This time-series data will also be combined with analysis of steady state and mutant data, to impart dynamics and predictive power to our organic-N regulatory network models, as described in Aim 2.

**Aim 1B. Network Walking: Generation of high through put, high confidence TF🡪target data for TF hubs.** As a functional test of the gene regulatory network predictions, our ultimate goal is to validate predicted direct and indirect targets of a TF genome-wide. We favor the DEX-inducible approach we have used to identify direct and functional targets genome-wide over Chromatin-IP (ChIP) for a number of reasons. While a Chromatin-IP (Chip-Seq) approach can confirm protein:DNA binding, it does not guarantee functional regulation [Eilers and Eisenman Genes & Dev. 2008. 22: 2755-2766]. An added advantage of the DEX-inducible TF system, is that one can identify the effect of regulatory components that may not bind directly to DNA [ Lee, J, He, K, Stolc, V, Lee, H, Figueroa, P, Gao, Y, Tongprasit, W, Zhao, H, Lee, I, and Deng, XW, *Analysis of transcription factor HY5 genomic binding sites revealed its hierarchical role in light regulation of development.* Plant Cell, 2007. **19**(3): p. 731-749.] Technically, our approach is an adaptation of the original technique [ Sablowski, RW and Meyerowitz, EM, *A homolog of NO APICAL MERISTEM is an immediate target of the floral homeotic genes APETALA3/PISTILLATA.* Cell, 1998. **92**(1): p. 93-103.] combined with transcriptome analysis described by [ Hanson, J, Hanssen, M, Wiese, A, Hendriks, MM, and Smeekens, S, *The sucrose regulated transcription factor bZIP11 affects amino acid metabolism by regulating the expression of ASPARAGINE SYNTHETASE1 and PROLINE DEHYDROGENASE2.* Plant J, 2008. **53**(6): p. 935-949.] [Wang, Y, Joshi, T, Zhang, XS, Xu, D, and Chen, L, *Inferring gene regulatory networks from multiple microarray datasets.* Bioinformatics, 2006. **22**(19): p. 2413-2420]. Our approach pushes the experimental concept further by studying; i) The effect of TF induction (by DEX) in the context of N-pre-programming and, ii) +/- CHX to compare both primary and secondary target relationships, to expand our regulatory network models starting from the beginning of the network, hence the name “Network Walking”.

***Approach*:** Conceptually our DEX transient assay approach to rapidly validate TF🡪target interactions genome-wide consists of the following; i) transient expression of a GR-TF fusion in protoplasts, ii) use of an inducible dexamethasone (DEX) system to artificially control the entry of the TF into the nucleus, iii) DEX induction of TF action +/- cycloheximide (CHX) to distinguish primary vs. secondary targets and iv) the use of nitrogen pre-treatments (e.g. prior to DEX induction of the TF) (see Fig. X). This last step allows us to identify the primary targets of the TF when acting alone or in combination with other TFs that are N-regulated transcriptionally or post-translationally. A global view of the DEX approach is summarized in Fig. X, and described more fully in Progress and in Bargmann et al 2012. As proof-of-principle, to apply this system to our nitrogen-regulatory networks, we employed this approach for the myb TF HRS1 (At1g1330), a top scoring TF hub of our “learned” nitrate-regulatory network, which was previously shown to be involved in phosphate signaling [Liu 2009, Journal of Integrative Plant Biology]. We hypothesize that HRS1 may also be involved in nitrogen sensing, as HRS1 is induced within 9 min of NO3- treatment, and in the inferred network, it is predicted to be the top most controlling and top most controlled gene [Krouk et al 2010]. Our initial analysis revealed four distinct clusters of genes that are influenced by a combination of HRS1 induction (by DEX treatment) and nitrate-status (Fig. X). Importantly, Cluster 4 genes reveal are direct targets of HRS1 (+DEX+CHX) that are only activated in the *presence* of NO3-. This set of genes are significantly overrepresented in the categories of phosphate transport (p-val 8.14 E-6) and cellular transport, which recapitulates the role of HRS1 in phosphate signaling observed in 35S::HRS1 overexpressing plants [Liu 2009]. This result suggests the rapid transient assay in protoplasts is relevant to whole plants. [Kranthi-This paragraph needs a description of what happened when we did the preliminary results on HRS1 gene +DEX and +/-CHX to show that direct targets are a subset of indirect] This transient approach for TF perturbation (+/- DEX, +/-CHX, +/-N) will be used for all the prioritized TF hubs listed in Table 1, which also lists our progress for each. Conditions for +/-N will be tailored to either nitrate responses (growth in ammonium succinate, treatment with nitrate) or organic-N responses (growth in 1mM nitrate, and treatment with 40mM ammonium/nitrate). Cluster analysis of transcriptomic data, followed by GO-term analysis using the Biomaps function in VirtualPlant [Katari 2010], will identify biological processes controlled by specific TFs. The transcriptome data will also be fed into the pipeline of machine learning algorithms developed in Aim 2, to fuel the generation and refinement of a causal predictive network.

**Aim 1C. Genome-wide validation of TF (protein)🡪targets (DNA) using Chip-Seq.** In Aim 1B above, we will identify genes whose transcription is activated in response to DEX-induced localization of a TF. In order to identify which of these regulated genes directly binds to the TF, we will perform ChIP-seq experiments, using anti-GR antibodies that will recognize the TF-GR fusion protein in our transient expression system. In order to conduct transcriptome studies as well as ChIP from the same sample, we performed a proof-of-principle study using a 35S::GR-BZip1 construct (shown in Table 1). We adapted the microChIP protocol from [Dhal and Collas, 2008], which requires a relatively small number of cells. After cell sorting, protoplasts were treated with 1% formaldehyde for 10 min and then quenched with 100mM glycine for 5 min. W5 buffer was then added to wash the cells before pelletting them by centrifugation at 2500 rpm for 2min. Cross linked cells were snap-frozen in liquid N and stored at -80C. The protocol for processing 1000 cells was then followed. Anti GR (sc-1002) from Santa Cruz Biotechnology Inc. was used to capture the GR::bZip1-DNA complexes and ChIP DNA was purified after reverse cross-linking using the MiniElute Reaction cleanup kit from QIAGEN.

NEED INPUT FROM YING AND KRANTHI HERE

The purified DNA was processed for Illumina sequencing…..and analyzed to identify peaks as follows……Preliminary analysis of the bZip1 data shows an enrichment of genes predicted to be bZIp1 targets in our organic-N regulatory networks including ……

**Aim 1D. Perturbations of TFs in whole plants: T-DNA mutants and 35S::Over-expressors.**

The transient expression assays in Aims 1B & C, are a rapid way to test how perturbations of TFs affect network targets, and will also enable us to overcome problems associated with functional redundancy in TFs. [Chen HW, Bandyopadhyay S, Shasha DE, Birnbaum KD, “[Predicting genome-wide redundancy using machine learning”.](http://www.ncbi.nlm.nih.gov/pubmed/21087504). BMC Evol Biol. 2010 Nov 18;10:357] [Cutler S and McCourt P (2005) Dude, Where’s my phenotype? Dealing with redundancy in Signaling Networks. Plant Physiology (2005) v. 138, pp558-9]. Owing to functional redundancy of TFs in signaling pathways, a negative result with a T-DNA mutant in the TFs listed in Table I have been inconclusive. We failed to see a molecular phenotype for predicted network targets in the cca1 or glk1 single mutants, but observed molecular phenotypes in 35S::CCA1 [Gutierrez et al 2008] and have also observed molecular phenotypes in double mutants for glk1/glk2 and cca1/lhy1, but not in the single mutants (Fig. X). However, in rare cases, we do see a molecular phenotype (misregulation of a target genes) in a TF T-DNA mutant. For example, TF WRKY1 (At2g04880), which is predicted to be a major hub in the N assimilatory network, is predicted to act as a “toggle switch” to induce expression of genes involved in nitrate reduction and assimilation into Gln, while simultaneously repressing expression of genes involved in converting Gln to Asn (Fig. X). Indeed WRKY1 T-DNA insertion mutants (SALK\_016954; SALK\_136009; SALK\_070989) revealed decreased expression of NIA2 and NRT2.1 (targets of WRKY1 activation), and increased expression bZIP1🡪ASN1 (both targets of WRKY1 repression), as predicted by the network model (Fig. X), while the opposite expression patterns are found in the DEX system where WRKY1 is over-expressed using methods described in Aim 1C. We will perform transcriptomic analysis on T-DNAs only if Q-PCR shows mis-regulation of a target gene in the N-assimilation network, for the TFs listed in Table I. Such T-DNA mutants in TFs will also be valuable for studying TF interactions (see Aim 3). Finally, based on results in the DEX-transient system, we will create transgenic plants with the Dex-inducible promoter for selected TFs to validate network targets in planta, and in distinct organs, or across distinct developmental time-points.

**Aim 2. Computational Innovation: A Pipeline Machine Learning Approach to Finding Causal Links in regulatory networks using many kinds of genome-scale data**

***Rationale***: Our ultimate goal is to employ machine-learning approaches 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) 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 transient expression experiments will support either primary or secondary TF🡪target relationships, some of which will be validated by ChIP-seq.

**Approach**: To improve the predictive power of our networks, we will create a pipeline to generate a predictive network model that makes use of four types of genomic data that are available to us, as generated in Aim 1: (i) steady state transcriptome data (e.g. N-treatments); (ii) TF perturbation transcriptome data (e.g. DEX data and T-DNA), (iii) time-series data (e.g. expression over time), 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 pipeline described below.

**The Network**: 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 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 tend to rise (increase in expression, because Z would be positive) as gene A rises (i.e., A to Z would be an inductive edge). 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].

**Pipeline/Consensus of Machine Learning Algorithms to integrate learning from distinct data-types.**

One reasonable approach is to establish a pipeline (Fig. X, figure dfg.eps) that analyzes different kinds of data using different algorithms. In one such pipeline, the MCZ algorithm (short for the Median Corrected Z-score method [Greenfield, A., Madar, A., Ostrer, H., Bonneau, R.: DREAM4: Combining Genetic and Dynamic Information to Identify Biological Networks and Dynamical Models. PloS one (2010]) is used for both the steady state and mutation data, followed by the DFG algorithm (short for Dynamic Factor Graph used for time-series data [Krouk,G.,Mirowski,P.,LeCun,Y.,Shasha,D.E.,Coruzzi,G.M.:Predictivenetworkmodeling of the high-resolution dynamic plant transcriptome in response to nitrate. Genome Biology 11(12), R123 (2010]). In this pipeline, the MCZ algorithm assigns weights to certain edges that are then processed DFG algorithm.

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

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

**Preliminary Results:** To test the value of pipelines vs. weighted consensus views for predicting edges in a network, we compare the results of our previous Dynamic Factor Graph (DFG) Approach [Krouk,G.,Mirowski,P.,LeCun,Y.,Shasha,D.E.,Coruzzi,G.M.:Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate. Genome Biology 11(12), R123 (2010] built on time-series data, with a pipeline or consensus “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) 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 consensus approach, we obtain z’’% of the gene expression directions correctly.

**Intended Approach**: Based on our preliminary results, we will iterate the following steps: use a consensus approach on nitrogen treatment steady state data to identify the genes relevant to nitrogen treatments (on the order of a few hundred), time-series data, and TF perturbation experiments (e.g. DEX experiments or T-DNA mutants). This will result in a regulatory network of the genes involved in N-assimilation.

(Dennis- We need to explain how we will define the subnetwork of genes involved in N-assimilation and the TFs that regulate them….OR are you intending to make a regulatory network using ALL TFs and ALL genes in the genome. We need to be clearer here. We do have a list of 57 target genes in the N-assimilatory pathway, including nitrate transporters…if that helps to anchor the network building.

DENNIS ANSWER: It will be a subnetwork consisting of the genes that could be relevant to nitrogen. See above)

RESPONSE TO DENNIS ANSWER:

Dennis- Last time, we identified N-regulated genes (all N-regulated TFs ~80TF) and genes in the N\_assimilation pathway that were regulated by nitrogen (9 genes). DFG then identified a subnetwork of 39 TFs that target the 8 genes in the N-assimilation pathway…..Is this the approach you are planning???? OR will we model regulation of ALL N-responsive genes which could be in the hundreds?\_

That in turn will suggest new genes on which to try DEX experiments (e.g. putative hubs). That will suggest a new network and so on.

As usual, our criterion of goodness will be the ability to predict well on out-of-sample data, both missing time points and missing mutations. Ultimately, we would like to learn the 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.

**DENNIS- PLEASE READ/EDIT AIM3**

**AIM 3.** **The feedback from analysis to subsequent experimentation: Testing new candidate TFs and TF interactions.**

**Rationale**: In the spirit of the Systems Biology cycle [Gutierrez 2005], the network models from Aim 2 will in turn predict new network perturbation experiments to perform in Aim 3. These predictions include identification of new TF hubs to test using the high throughput approaches described in Aim 1, including ones for which TF🡪target relationships are the hardest for our learned network to predict. In addition, our network models in Aim 2 will give suggestions for TF pairs that may act in concert (or antagonistically) on genes in the N-assimilatory network. Thus, we will also develop high throughput experimental approaches to test combinatorial functions of TFs in the learned network and these will enhance the network models leading to new predictions [Gloria we need a figure showing this cycle of experiment to modeling to more experiments to more modeling].

**Approach**: We will test TF🡪target relationships for new TFs identified in Aim 2 using the transient system described in Aim 1. We will also use two complementary approaches to identify network targets of TF-TF pairs: (i) Co-expression of TF1 and TF2 in the transient protoplast DEX-system, (ii) Testing TF pairs using a high throughput genetic approach. The high throughput genetic approach will use protoplasts from T-DNA deletion mutants for TF1 as a host for transient overexpression of TF2. [Gloria: the following sentence is inconsistent with the previous one. Also this one is VERY unclear. I think it could be dropped as the idea is explained below.] Differences in target gene expression between the transient expression of TF1 in WT compared with expression in protoplasts made from T-DNA mutants in TF2, will identify TF dependencies- including additive or synergistic effects on target genes in the networks.

**Aim 3A. Perturbation studies for new candidate TFs**: In this subaim, TFs will be targeted for perturbation studies described in Aim 1 based on network models derived in Aim 2. For example, new time-series data for organic-N signaling is likely to refine our current network models and identify new TFs which are activated early in the cascade (e.g. within 3-9 min), as we saw for the nitrate-responsive networks [Krouk et al 2010]. Also, the network models will identify which TFs may have the most influence based on their outdegree (“hubbiness”) in the network or based on the phenotypic importance of their targets. These will be the candidates to target for perturbation studies. [DENNIS\_ CAN YOU GIVE OTHER EXAMPLES HERE ABOUT HOW WE MIGHT IDENTIFY NEW TFS TO STUDY FROM THE NETWORK PREDICTIONS IN AIM 2? Please see above]

**Aim 3B. Experimental Testing of TF Pairs in the regulation of N-assimilation network.** Our new and refined network models generated in Aim 2 will suggest TF pairs that may function to regulate important target genes in the network. We will use two complementary approaches to test how these TF pairs may function in the regulation of target genes. In one approach, we will co-express two TFs using a single vector in the transient protoplast system. In a complementary approach, we will test the function of TF1 in a protoplast made from a TF2 T-DNA mutant. In both cases, we will compare activation of target genes to each single TF expressed in protoplasts from wild-type. Transcriptomic analysis will identify genes for which TF1/TF2 cooperation can explain additive or synergistic expression. Synergistic regulation of target gene expression would be functional evidence for TF cooperation, which could involve direct or indirect TF interactions.

**Co-expression of TF1 and TF2 and genetic perturbation:** In order to co-express two TFs (TF1 and TF2) in the same cell, we have used the MultiSite Gateway**®** Technology that allows one to simultaneously clone multiple DNA fragments in a defined order and orientation. To generate the expression construct, TF1 cDNA, TF2 cDNA and the “GR cassette”(that includes a terminator, 35S promoter and the GR sequence) were cloned in appropriate donor vectors and allowed to recombine with each other and with the destination vector pBob11. This results in the assembly of 2 consecutive expression cassettes in which each TF is fused to the GR sequence and the expression of each fusion is driven by a constitutive 35S promoter (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 size i.e. vectors containing small genes will be present in higher copy number than vectors containing long genes hence small genes might expressed at higher levels; (ii) when multiple vectors are transfected, each cell is likely to contain a random number of copies of each vector while the population of transfected cells will be homogeneous in its transgenic content; (iii) the expression vector harbors monomeric red fluorescent protein (mRFP) for positive fluorescence selection, therefore enabling the isolation of transfected protoplasts and avoiding confounding of the results by non-transformed cells. In addition, this co-expression vector will bypass the sorting of false positives due to fluorescence bleaching when multiple vectors with different fluorescence selections are used. We will compare activation of target genes to each single TF expressed in protoplasts from wild-type. As a complement to the multisite vector approach, we will also explore TF cooperation by expressing the single TF1 in a T-DNA mutant background for TF2. In both cases (double expression or expression of TF1 in tf2 mutant background), transcriptomic analysis will identify genes for which TF1/TF2 cooperation can explain additive or synergistic expression of target genes. Synergistic regulation of target gene expression would be functional evidence for TF cooperation- which could involve direct or indirect TF interactions. We will use this genetic approach for TFs that show a molecular phenotype in the single T-DNA mutant (e.g. WRKY1 T-DNA). For TFs that are functionally redundant, we will use double mutants (e.g. glk1/2, cca1/2) to explore transcriptional networks. For example, as glk1/2 are in the same myb family as HRS1, we will express HRS1 in a glk1/2 double mutant background, and determine the effects on target gene activation or repression. These perturbation results will be fed into the machine learning pipeline (Aim 2), to refine our network predictions.