**(a) Significance.**

* Importance of Problem or critical barrier to progress.
* How will project improve scientific knowledge, technical capability in one or more broad fields.
* How concepts, methods, technologies, that drive this field will be changed if proposed aims are achieved.

**(a) Significance:** Our goal is to model a causal genetic network, effectively the circuit diagram underlying the regulation of genes in the N-assimilatory pathway by applying machine-learning approaches to genomic data. By analogy to an electrical network, such a gene regulatory network would enable us to infer network states under untested conditions and mutations. Such a capability is the holy grail of Systems Biology, as such predictions can be used to suggest targeted interventions in pathways and processes in medicine and agriculture.

Currently, inference of biological networks is an inherently difficult problem due to the presence of few time-point measurements, many genes, measurement errors, and random fluctuations in the environment [ Jaeger and Monk 2010 Jaeger J, Monk N: Reverse engineering of gene regulatory networks. In Learning and Inference in Computational Systems Biology. Edited by: Lawrence N, Girolami M, Rattray M, Sanguinetti G. Cambridge MA: MIT Press; 2010.] Our proposal seeks to improve success in network inference, through a combination of innovations in both experimental and computational approaches. Our experimental innovation “Network Walking”, a rapid method to generate reliable high through put data for TF🡪target interactions, is used to validate and refine our machine learning algorithms for network inference. The novelty of our machine learning pipeline, is that it integrates analysis of multiple types of genomic data (steady state, time series, and perturbation) into a combined analysis. A further novelty is that our approach is inherently iterative: experiment feeds learning which feeds experiment and so on.

Our advances in predictive modeling of regulatory networks controlling N-assimilation, will suggest candidates genes for 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, biological process (or potentially any trait) of interest with applications across a wide range of problems in biology and medicine.

**(b) Innovation**

* + How does the application challenge and seek to shift current research paradigms
  + Describe novel theoretical concepts, approaches or methodologies being used and any advantage over existing
  + Explain any refinements, improvements or new applications of theoretical concepts, approaches or methodologies.

**(b) Innovation.** Our work integrates innovations in a combined experimental & computational approach have been developed as a result of a close interaction between biologists and computer scientists at NYU’s Courant Institute of Mathematical Science.

**1. Experimental Innovation: “Network Walking: A rapid and reliable high through-put method for detecting TF🡪target interactions genome-wide (Aim 1).**  To accelerate experimental validation of TF🡪target predictions genome-wide, and to overcome problems of TF functional redundancy in reverse genetic approaches in plants, [Chen HW, Bandyopadhyay S, Shasha DE, Birnbaum KD, “Predicting genome-wide redundancy using machine learning”.. BMC Evol Biol. 2010 Nov 18;10:357] [Cutler S and McCourt P (2005) Dude, Where’s my phenotype? Dealing with redundancy in Signaling Networks. Plant Physiology (2005) v. 138, pp558-9.], we have developed a rapid and high through-put protoplast system to identify network targets of a TF genome wide within 2 weeks [Bargmann et al, 2012, submitted]. This approach, called “Network Walking”, employs a transient protoplast system to overexpress any TF of interest (as a TF-GR glucocorticoid receptor fusion) and to selectively induce TF nuclear localization upon treatment with dexamethasone (+DEX). When DEX treatment is performed in the presence of cycloheximide (+ CHX) only primary targets of the TF are transcriptionally activated. Direct TF🡪target binding is confirmed using anti-GR antibodies in ChIP-Seq experiments. The genome-wide data from this system is used to refine and train a pipeline of machine learning methods for predictive network modeling, as discussed below and in the Research Plan.

**2. Computational Innovation: Using a combination of different machine learning algorithms for analysis of steady state data, mutation data, and time-series data to generate predictive networks (Aim 2).**  As reviewed by Jaeger, inference of biological networks is an inherently difficult problem due to the presence of [ Jaeger and Monk 2010 Jaeger J, Monk N: Reverse engineering of gene regulatory networks. In Learning and Inference in Computational Systems Biology. Edited by: Lawrence N, Girolami M, Rattray M, Sanguinetti G. Cambridge MA: MIT Press; 2010.] because of the experimental limitations (e.g. few time-point measurements, many genes, measurement errors, and random fluctuations in the environment). Because of these limitations, [Gloria: I suggest we start here as the above sentences have already been said:] methods for computational inference of gene regulation networks can be crudely divided into two approaches: (i) using non-linear or state-space based modeling of the complex interactions between a restricted number of genes (typically ten) with hidden states for unknown values for protein abundance of transcription factors (using a relatively small set of time-series data); or (ii) simpler, but linear, models of transcription factor-gene interactions, relying on larger (hundreds to thousands) numbers of microarray measurements from steady state mRNA data [Bonneau et al (2007) Cell vol. 131;1354] [ Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V: The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biol 2006, 7:R36.] [Wang Y, Joshi T, Zhang XS, Xu D, Chen L: Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics 2006, 22:2413-2420] [Shimamura T, Imoto S, Yamaguchi R, Fujita A, Nagasaki M, Miyano S:Recursive regularization for inferring gene networks from time-course gene expression profiles. BMC Syst Biol 2009, 3:41.]. In our previous network models, we successfully employed the state space machine-learning approach to model time-series transcriptome data for N-regulatory networks affecting N-assimilatory pathway genes [Krouk et al 2010]. Separately, we have generated N-regulatory networks using steady state data [Gutierrez et al 2008]. In this proposal, we will develop a computational pipeline that enables us to combine time-series data, with steady state data, and perturbation data into a single pipeline for machine learning. This pipeline uses the MCZ algorithm (short for the Median Corrected Z-score method) for the analysis of steady state data (from wild-type vs. mutant data), a simple but successful approach, with the DFG algorithm (short for Dynamic Factor Graph, a form of state space analysis) which is driven by time-series data.

The regulatory networks derived from this pipeline will drive a new round of experimentation, an inherent trait of the Systems Biology cycle.