**MOTIVATION AND NOVELTY**

This NSF Plant Genome application aims to enable plant biologists to infer and exploits gene networks across a wide variety of plant genomes. The overall approach to enabling this research is to develop machine learning methods that enable us to (i) exploit Next-Gen sequence data across a large number of plant species to construct the gene networks in data-rich species and infer them in data poor species, to (ii) exploit data associated with traits in crop species to predict networks and testable hypotheses in model species, and (iii) to develop a pipeline X-Net to enable network construction and the identification of BioModules in any species of interest, as data becomes available. These cross-network approaches will promote hypothesis derivation in crops and will enhance the ability to predict the “translatability” of gene functionality from models to crops, and between crops. By making the X-Net tool “user friendly”, we will also empower the plant biologists to generate and evaluate the significance of modules conserved across plant species, a task beyond the expertise of computer scientists.

**BACKGROUND:**

**Gene interaction networks and Arabidopsis**. Analyzing gene expression data in the context of gene networks has enabled hypothesis-driven gene discovery, for example, to reveal transcriptional responses to external stimuli and to uncover coordinate expression of different genes [Usadel 2009 Plant Cell Environ] [Lee 2010, Nature Biotech] [Gutierrez 2007 Genome Biology]. Such studies have shown that functionally related genes tend to be transcriptionally coordinated (i.e., co-expressed) [[Stuart et al., 2003](http://www.plantcell.org/content/23/3/895.full#ref-56) Science; [Persson et al., 2005](http://www.plantcell.org/content/23/3/895.full#ref-49) PNAS]. Using these “guilt-by-association” approaches, coexpression network analyses have proved valuable for rapid inference of gene function, subcellular localization of gene products, and biological pathway discovery [[Wei et al., 2006](http://www.plantcell.org/content/23/3/895.full#ref-64) Plant Physiol; [Yonekura-Sakakibara et al., 2008](http://www.plantcell.org/content/23/3/895.full#ref-65) Plant Cell; San Clemente et al., 2009 ; [Usadel et al., 2009](http://www.plantcell.org/content/23/3/895.full#ref-63) Plant Cell Environment; [Klie et al., 2010](http://www.plantcell.org/content/23/3/895.full#ref-28) J. Computational Biology]. In Arabidopsis, network tools developed to integrate co-expression data with other sources of gene interaction data (e.g. protein-DNA, protein-protein and other interactions) have also been deployed to study gene function. For example, “AraNet” – a probabilistic network tool [Lee et al 2010-Nature Biotech]- has been used to identify genes associated with traits using early seedling development as a test case. Likewise, the Arabidopsis Multinetwork [Gutierrez et al 2007 Genome Biology, Gutierrez et al 2008 PNAS, and Katari et al 2010] - developed in under previous NSF funding of this grant - has been used to derive and validate biological hypotheses for gene regulatory networks involved in carbon and nitrogen sensing [Gutierrez et al Genome Biol 2007, Gutierrez PNAS 2008] and in the nitrogen regulation of lateral root development [Gifford et al 2008], as examples. With the advent of Next-Gen sequencing, it is now possible to generate transcriptome data for a large number of species. The challenge now is to build on the successes in the data rich models to chart a path to generating gene networks and candidate genes across plant species, especially crops.

**Enabling gene interaction networks across plant species**: In an attempt to transfer knowledge derived from Arabidopsis gene regulatory networks to crops, a number of studies have compared species-specific expression networks in Arabidopsis to a crop species, to identify the “best” ortholog for a given Arabidopsis gene. The idea that a number of studies have explored is that the best ortholog for a gene must have a similar “network neighborhood” as well as a similar sequence. Such analysis platforms include PlaNet [Mutwil 2011 Plant Cell], Starnet [Jupiter 2009 BMC Bioinformatics], CoP [Ogata 2010 Bioinformatics], and ATTED-II [Obayashi 2011 Plant and Cell Physiology]. All of these platforms share the commonality of building co-expression networks within each species, and comparing networks across species post-hoc. One of the limitations of all of these approaches is that they *assume* the existence of enough data in the target (e.g. crop) species to construct reliable co-expression (and potentially other) networks under conditions of interest. This assumption can lead to two problems: (i) there may be insufficient data in the target species to support the conclusions, and (ii) the discovery of network modules is limited in scope. Because most of the newly sequenced species will be “data poor” (compared to the models), we propose a novel approach that takes advantage of data-rich species to learn and infer networks in data-poor species. [Dennis doesn’t love these last few sentences. It might just get us enemies and may not be true:] Additionally, the current approaches to comparative networks are mostly limited to pair-wise comparisons of species; Starnet [Jupiter 2009 BMC Bioinformatics], CoP [Ogata 2010 Bioinformatics], and ATTED-II [Obayashi 2011 Plant and Cell Physiology], with some exceptions such as PlaNet [Mutwil et al 2011]. By contrast, our cross-network method aims to mine correlation data from multiple species simultaneously to learn and infer networks across any species of interest.

**Overview of Aims**: In this application, we propose to develop phylogenomically-informed network inference approaches to LEARN regulatory networks in a data-poor target species (e.g. crop), based on information from several data-rich species. This Phylogenomic Network Inference (PNI) approach is based on the *Robin Hood philosophy* - "robbing/learning from the rich and giving to the poor" (Aim 1). Such inferred networks in the data-poor species may then be used to find “best” orthologs and “ortholog neighborhoods” in the networks, as in the approches described above [Mutwil 2011, Plant Cell]. A second application of network learning in our proposal, uses data from the target species (e.g. crop) to inform network discovery and testing in an associated model species. Here, the idea is to use expression data from a target species (e.g. crop) to identify ALL genes that may be associated with a trait, use networks in Arabidopsis to identify the best ortholog neighborhoods, and then to test their functionality using mutagenesis or over-expression in Arabidopsis, where such experiments are relatively easy (Aim 2). This approach should enhance the translatability of acquired knowledge from Arabidopsis to crop species. Finally, we will develop a user friendly pipeline called “X-Net”, that will enable plant biologists to perform multi-network analysis for any species of interest, keeping the networks up-to-date with respect to their data sources. The X-Net pipeline will enable researchers to combine correlation network data –either actual or inferred -- from a target (e.g. crop) species of interest with validated gene interaction data from data-rich species such as Arabidopsis and Rice (e.g. protein:protein, DNA:protein; miRNA-target data) to derive testable hypotheses. Interpreting the target crop species correlation network data in the context of validated interaction data in the data-rich models will also enhance the ability to transfer acquired knowledge from Arabidopsis to crops. (Aim 3).

**The novel features of our network phylogenomic approach to gene regulatory networks include an the use of data rich model species to infer networks in data poor crop species, the use of trait studies in crop species to identify good target genes in model species, and an ongoing data collection pipeline to construct multinetworks (i.e. networks having multiple edge types) as data sources change. This interplay among species will enhance transfer of knowledge between crops and models.**

**[Dennis finds the rest here too redundant to the discussion of aims, so suggests deleting from here down]**

**The Robin Hood approach to inferring networks in data-poor target (e.g. crop) species**: We use a machine learning approach in which data-rich model species are used to train a regression model which then leads to inferred networks in data-poor crop species (Aim 1). We call this the *Robin Hood* *Approach* to network inference. This inference approach learns network rules by analyzing multiple species simultaneously, not by comparing them post hoc as current programs do either pair-wise (CoP) [Ogata et al 2010], Starnet [Jupiter et al 2009, BMC Bioinformatics] or multispecies (PlanNet) [Mutwil et al 2011]. The existing tools can also only create networks from data-rich species. In our approach, we LEARN from data rich species and infer on data poor species. Thus, this Robin Hood approach to network inference should greatly expand the network approach to hypothesis generation across plant biodiversity.

**Using crops to improve transfer of knowledge and testing in models**. We next exploit gene-to-trait associations obtained from crop species, sequence orthology to Arabidopsis, and network analysis in Arabidopsis to identify targets for mutagenesis. Here, expression data from the crop species is driving discovery and experimentation in Arabidopsis, where it is easier to do experiments and then the results will be applied to the crop species (**Aim 2**).

**X-Net:** The novelty of the X-Net approach is to enable multinetwork analysis in any species of interest.