**Aim 1: Development of InferNET: Inferring networks in data-poor species. *Rationale***. In this Aim, we propose to develop a network inference approach, ***InferNET*** that will ***infer*** regulatory networks in a data-poor target species, based on ***gene networks*** from several data-rich species. The InferNET approach is inspired by a variant of the *Robin Hood philosophy* -- "learning from the rich and giving to the poor". Inferred networks in the data-poor target (e.g. crop) species may be used to derive hypotheses and identify potentially important genes for validation testing. Most of the discussion below concerns co-expression networks, because, with the advent of Next-Gen sequencing, such data is now readily available and/or obtainable for many species. However, the methods we develop apply to other types of network edges (e.g. protein interaction [REF or NSF Rice protein interaction project web site (Vidal and Ecker)].), and we will apply them as data becomes available.

***Novelty***: ***InferNET*** differs from existing comparative network tools in plants, because InferNet uses data-rich species to ***infer*** regulatory networks in data-poor species. By contrast, the existing tools for comparing plant gene networks creates networks only based on experimental data, and then compares them post-hoc (e.g. CoP [Ogata 2010], Starnet [Jupiter 2009], ATTED-II [Obayashi 2011, and PlaNet [Mutwil 2011]). Also, with the exception of PlaNET, tools compare only two species at a time. InferNET, like PlaNet, uses the networks from multiple species simultaneously.

***Community Need:*** As the number of available genome sequences increases, it will be common to find a newly sequenced or poorly studied target species “*t*” that is phylogenomically similar (see phylogenetic tree in Fig. X) to those few “data-rich” species having a substantial body of experiments. For targets such as emerging crops and new “boutique” crop species, inferring networks will be particularly valuable.

**IMPLEMENTATION AND TESTING OF InferNET:**

**Species:** InferNet will mine the current 21 fully sequenced species (Fig. X) and we will add additional species to the InferNet analysis pipeline as their sequences become available. This would include species with fully sequenced and annotated genomes, and can also include species with fragmented “gene space” assemblies as are likely to be produced by next-generation sequencing technologies.

**Defining data-rich species**: To determine whether a species is indeed “data-rich”, we will use a technique analogous to Statistical Power Analysis [Hill, T. & Lewicki, P. (2007). STATISTICS: Methods and Applications. StatSoft, Tulsa, OK]. Mechanically, this consists of computing the p-values of large positive (r value >= 0.5) and large negative (r value <= -0.5) correlations within some species, for the experiments already done on that species. Next, to determine whether there is enough data for that species, we recompute the p-values of those same edges, assuming the same set of experiments had been repeated with the same results (of course the correlations won’t change, but the p-values will get smaller). If the number of p-values below a threshold of 0.05 increases by more than say 50% under this assumption, then we deem that the species is currently data-poor. Otherwise, it is data-rich. Admittedly, these thresholds are somewhat arbitrary, but they divide the 21 species reasonably.

For example, according to these metrics of the current fully sequence 21 species, species x,y,z are measured as data-rich and c,d,e are data-poor ) [NOTE: THIS AnalysisNeeds to be done]. It’s being done by Roberto. Should be done this week.

**Correlation networks**: The basic co-expression metric we use to identify correlation networks will be Pearson correlation, because it has been shown to be particularly useful in inferring functionality in current cross-species network studies [Mutwil 2011] [Usadel 2009], [[Klie et al., 2010](http://www.plantcell.org/content/23/3/895.full#ref-28)]. However, in the course of this study, we will also test our methods of analysis using other metrics including mutual information [Margolin 2006], Mutual Rank [Obayashi et al], and Spearman correlation [Hill 2007 reference below]. Our approach will be to train the InferNET algorithm using two or more data-rich source species (s1, s2, …), and then to apply the trained model to data-poor target species *t*. We validate the success of each in making valid predictions, using experimental data from data-rich species.

**Learning the Rules**: The InferNET training itself will be done as follows: Take several data-rich source species *s1, s2, …, sk,* and temporarily ignore the expression data from one of them, call it *v*. Next, using one of several machine-learning algorithms to be discussed below, we *learn* the parameters of a regression model that predicts co-expression edges in *v*. We then use that model learned in the data-rich species, to predict edges in the data-poor target species *t*.

***The input for the InferNET algorithm will be in the three formats described below.***

**orthotab: target species| target gene | source species | source gene | orthology val1 | orthology val2 …**: gives the gene-to-gene orthology value, according to several different orthology measures for example: Reciprocal best BLAST hits [Altschul 1997 Nuc Acid Resh], BLAST hits above a threshold, OrthologID [Chiu 2006 Bioinformatics], OrthoMCL [Li 2003 Genome Research], and Inparanoid [O’Brien 2005 Nuc. Acid Resh]. Our preliminary work used reciprocal best BLAST hits. Part of the machine-learning research will be used to determine which orthology method or methods will work best.

**edgetab: species | gene1 | gene2 | edgetype | strength | p-value**: gives the strength and the p-value of a given experimentally supported edge (e.g. by data including expression correlation). In our preliminary studies, we examine gene expression correlations that generally hold over all conditions. Edge relationships present only under certain conditions (e.g. drought conditions) or in certain tissues (“Gene Spaces”) can be retrieved using a focused set of experiments, as done in Aim 2. The machine learning stays the same, but the data can change. (We consider Microarray and RNA-seq data interchangeable under the proper normalization protocols[Bullard et al 2010].)

**speciestab (species1 | species2 | species similarity measure1 | species similarity measure2)**: measures sequence similarity of species according to several criteria (e.g. distance based, for example average percent identity of protein sequences, or through parsimony). Which similarity measure will work the best can be determined in the course of learning the coefficients of our Species Combining Rule.

**Machine Learning**: Now, to predict an edge between *g1* and *g2* in a data-poor target species *t*, we will combine evidence from edges in one or more data-rich source species s1, s2, …, as well as evidence from any experiments conducted in the data-poor target species *t* itself. The basic machine learning method will be Linear Regression and Regression Trees, with a penalty for complexity. For the sake of performance and robustness to noise, we will use one of the following three machine learning approaches:

1. **Random Forests [**Breiman 2001 Machine learning, Huynh-Thu 2010 PloS On**e]** Random forests are ensembles of decision trees which are constructed from random subsets of the data. They're fast to train, easy to parallelize, and perform extremely well.

2. **Large-Scale SVM Regression** [Bottou 2010] Bottou demonstrated that a stochastic gradient descent solver for a variety of learning problems (including support vector machine optimization) is able to scale to extremely large datasets, while converging to the predictive performance of traditional optimization algorithms.

3. **Large-Scale L-Regularized Learning [Shalev-Shwartz 2009]** Stochastic coordinate descent (a method related to stochastic gradient descent, but with a slightly different update rule), can be used to learn sparse regression models, with small training times, even for data sets where both the dimensionality and the number of training points is large.

**Validation testing of InferNet predictions**: The net effect of these machine-learning analyses will be to find the weighting of different factors (e.g. that correlation of source edges is more important than gene sequence orthology), that will lead us to estimate the correlation between two genes in some target species t. To determine which machine learning method is best, we will test them first on the data-rich species in “hide-the-answer” experiments. That is, we compare the predicted results (e.g. inferred edges in the target species t) that use no expression experiments (hide-the-answer) from the target species with the results from the experiments in the target species (see Table X). This analysis gives us precision and recall data to quantify the success of our prediction methods.

**Pilot study of InferNet**: In our pilot study, we tried to infer Pearson correlation edges in a “target” species Soy, knowing correlation edges in a “source” species Arabidopsis, trained using another data-rich species (Medicago), and the gene-by-gene orthology between genes in Arabidopsis and both Medicago and Soy (Fig. X). We selected these three species as an initial proof of concept because (i) there is ample and reliable Affymetrix data for each enabling us to validate our predictions, and (ii) Medicago and Soybean -- both legumes -- are quite closely related. The equation for network inference is trained using Arabidopsis and Medicago under an L-Regularized learning algorithm **[Shalev-Shwartz 2009].** Once we “learn” the rules for network structure using Arabidopsis and Medicago data, we applied this learned equation to infer edges in Soy. To test whether this approach worked, we evaluated the predictions using actual expression data from Soy.

This “InferNET” learning approach is superior to the Interolog approach which considers only BLAST scores to infer edges [Yu (2004) Genome Research,Annotation Transfer Between Genomes: Protein–Protein Interologs and Protein–DNA Regulogs ]. Under the Interolog approach, if (i) the co-expression edge between *g1* and *g2* in Arabidopsis has a certain correlation value *r,* (ii) *g1’* in the target (Soy) is the reciprocal top Blast hit for *g1*, and (iii) *g2’* is the reciprocal top Blast hit for *g2*, then the approach infers a correlation of *r* between *g1’* and *g2’*. As we show in the table below, InferNET has better recall (88% vs. 81%) and precision (77% vs. 69%) than the Interolog approach, even though Interolog by itself is quite informative.

***For our proof of concept study, the regression model had the following form:***

Estimated correlation in **target species *t*** (ECT) = a1\*Mean of Orthology values (MOv) + a2\*correlation of source pair (Cs) + a3\*p-value of correlation of source pair (Ps) (Fig. X).

This form of the regression model equation was chosen based on our expectation that the strength of correlation in the target species will depend on some statistic on the orthology assignments (a1\*MOv) and the strength and confidence in the correlation of expression in source species (a2\*Cs and a3\*Ps). For this study, mean of orthologous values is calculated as follows: if g1 and g2 are the source pair, and g1' and g2’ are the potential new target pair, and g1 and g1’ are reciprocally best blast hits (as are g2 and g2’), then we take the Mean of the Orthology values (MOv).. We chose the simple linear form of this equation for ease of understanding. The learned model also suggests why the Interolog approach [Yu 2004] works as well as it does. Whereas the MOv value and the correlation values both have absolute values between 0.5 and 1, the coefficient for correlation is 40 times greater than the correlation for orthology (1.2 vs. 0.03), implying that the specific value of orthology is unimportant for reciprocal Blast hits. Often, the correlation of the edge in the source species by itself predicts the correlation in the target.

Since there are a different number of experiments for each species and experiments from different sources, the distribution of correlation values can vary. So, we define two genes as “highly positively correlated”, if their correlation is in the top 5% of all measured correlations, and “highly negatively correlated”, if their correlation is in the bottom 5% (Table X). Thus, our machine-learning algorithm starts from the 5% most positively and negatively correlated pairs in Arabidopsis and infers positive or negative correlations about edges in the target (e.g. Soy) for reciprocal top blast hits of those elite pairs.

**We are still working on the exact numbers but the table should look something like this**:

**Predicting networks in Soy (Glycine Max)**

**Method | Positive Recall | Positive Precision | Negative Recall | Negative Precision**

InferNET | 88% (xxx/yyy) | 79% (xxx/yyy) | 73% (xxx/yyy) | 83% (xxx/yyy)

Interolog | 81% (xxx/yyy) | 69% (xxx/yyy) | 65% (xxx/yyy) | 78% (xxx/yyy)

**Predicting networks in Medicago**

**Method | Positive Recall | Positive Precision | Negative Recall | Negative Precision**

InferNET | XX% (xxx/yyy) | xx% (xxx/yyy) | XX% (xxx/yyy) | XX% (xxx/yyy)

Interolog | XX% (xxx/yyy) | XX% (xxx/yyy) | XX% (xxx/yyy) | XX% (xxx/yyy)

**Table X Caption**: Positive recall is the number of gene pairs in the target species correctly predicted to be positively correlated divided by the number of gene pairs that are positively correlated. Positive precision is the number of gene pairs correctly predicted to be positively correlated divided by the total number predicted to be positively correlated. Similar for negative correlation. The inferred correlation coefficients are XXXXXXXXX [TO BE FILLED IN].

**Limitations of the Proof-of-Concept Model and Planned Improvements of InferNet:**

**Orthology assignments**: In our future work, instead of using reciprocal top BLAST hits when inferring the correlation between some target pair g1’ and g2’, we will consider all gene pairs g11, g21; g12, g22 such that each g1i is above a similarity threshold GENESIM to g1’, and g2i is above the same similarity threshold GENESIM to g2’. This will imply that many gene pairs may be relevant to the prediction of a given target pair g1’ and g2’. This, in turn, implies the need for some form of aggregation over the potentially relevant gene pairs. We will include terms for mean, median, max, and min as the most representative aggregates. Each of the three machine learning mechanisms we will test will determine the weights for each term. We will also determine based on cross-validation the best gene orthology threshold, GENESIM.

**Incorporation of target species data**: In future development of InferNET, we will incorporate the limited expression data that is already available in the target species into the learning equation. The net result will be, for the edge g1’ between g2’, a term for an experimentally derived correlation and a term for the experimentally derived p-value.

**Use of additional species in training**: Further, in future development and testing of InferNET, we will be using more than two species for training. For example, based on available expression datasets we might train on Arabidopsis using data from two data-rich legume species (Soy and Medicago) and then apply the learned model on Cucumis (a data-poor species) or we would train on Rice using Maize and Sorghum as data-rich species and apply the model to Brachypodium, Setaria etc. In general, we might learn a model using *s1*, *s2*, *s3*, and *s4* and train on *s5*, then apply that model to a target species *t*. We will first create a model for each source-train species independently (e.g. from Arabidopsis to Glycine and then from Poplar to Medicago). Then we will form a “combining rule” consisting of a learned joint ranking of the several regression models weighted by phylogenomic similarity. The weights will be learned using one of the three machine learning methods above.

**Expected Outcomes of Aim 1 and future directions.**  Our goal in this Aim, is to construct a machine-learning model that can predict, with high recall and precision, the expression correlation of edges between genes in a little-studied “target” species, by inference from one or more data-rich “source” species. The success of the preliminary results suggest reason for optimism. We will apply the same techniques to other edge types (e.g. protein-protein interaction).