**Aim 1: Development of the Phylogenomic Network Inference (PNI) model on Expression data**

***Rationale***. In this Aim, we propose to develop phylogenomically-informed network inference approaches to *learn* regulatory networks in a data-poor target species (whether crop or non-crop), based on information from several data-rich species. This Phylogenomic Network Inference (PNI) approach is inspired from the *Robin Hood philosophy* -- "learning from the rich and giving to the poor". Such inferred networks in the target species may then be used to identify potentially important genes in that species. Most of the discussion below concerns co-expression networks because such data is available for several species, but the methods apply to other network types and we will apply them as data becomes available.

***Novelty***: Our Phylogenomic Network Inference (PNI) approach differs from existing comparative network approaches in plants, because PNI uses data-rich species to ***learn*** regulatory networks, and ***infer*** them in data-poor species. By contrast, the existing tools create networks from data-rich species and compare them post-hoc (e.g. CoP [Ogata 2010], Starnet [Jupiter 2009], ATTED-II [Obayashi 2011, and PlaNet [Mutwil 2011]. Most existing network tools compare two species at a time. PNI, like PlaNet, uses the networks from multiple species simultaneously though, again, PNI infers networks whereas PlaNet assumes the pre-existence of networks.

***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 to those few “data-rich” of the 21 (and growing) fully sequenced species for which there is already a substantial body of experiments (see phylogenetic tree in Fig. X).

**Data-types**: Due to the advent of Next-Gen sequencing technologies, much of the experimental data for expression atlases (and new species) will come in the form of genome-wide transcriptome expression measurements. In our approach, this data can used to train (using the data-rich species) a set of rules and infer, using those rules, a network of positive and negative expression correlation for the target species *t*. While we will initially develop the methods using RNA expression datasets, our methods will also be used for data supporting other kinds of network relationships such as protein-protein relationships, as they become available for multiple species. The next likely candidate for protein interaction data is rice thanks to the NSF Rice protein interaction project [REF].

The basic co-expression metric we use initially 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 [REF]. Our approach will be to train the PNI 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 approach using left out data as described in detail below.

**Defining data-rich species**: To determine whether a species is “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, 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 the species is currently data-poor. Otherwise, it is data-rich. Admittedly, these thresholds are somewhat arbitrary, but they divide the 21 species reasonably (i.e. 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 ) THIS ANALYSIS NEEDS TO BE DONE….

Learning the Rules: The PNI training itself will be done as follows. Take the data-rich source species *s1, s2, …, sk,* and temporarily ignore the expression data from one of them, call it *v*. Choose *v* so that its phylogenomic distance (measured from the phylogenetic tree shown in Fig. X) from the other source species is approximately the same as the distance between *t* (the target) and the other source species. 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 PNI 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], OrthologID [Chiu 2006 Bioinformatics], OrthoMCL [Li 2003 Genome Research], and Inparanoid [O’Brien 2005 Nuc. Acid Resh]. Our preliminary work used BLAST. At this point, we don’t know which orthology method or methods will work best. Part of the machine learning research will be used to determine this.

**edgetab: species | gene1 | gene2 | edgetype | strength | p-value | number of different experimental conditions**: gives the strength and the p-value of a given experimentally supported edge (e.g. by data including expression correlation). The p-value will be evaluated using a non-parametric re-sampling approach called “shuffling” as described, for example, in [Statistics is Easy! Dennis Shasha and Manda Wilson Synthesis Lectures on Mathematics and Statistics 2008 (doi:10.2200/S00142ED1V01Y200807MAS001) Morgan&Claypool Publishers)

In our preliminary studies, we examine correlations that generally hold over all conditions. However, certain edge relationships may be present only under certain conditions (e.g. drought conditions for plants). The tools we propose to build could be used just for the conditions of interest, in which case, we would choose the subset of edgetab corresponding to those conditions. That is, focusing on one or more conditions changes the data and possibly the results (e.g. we may find edges that apply only in certain conditions), but not the method.

**speciestab (species1 | species2 | species similarity measure1 | species similarity measure2)**: measures sequence similarity according to several criteria (e.g. distance based, for example average percent identity of protein sequences, or through parsimony). Again, we don’t know *a priori* which similarity measure or measures will work the best until we do the research, but we can determine the measures that work best in the course of machine learning. That is, we will include all measures and then will learn the weights of each, where a weight that is high in absolute value constitutes importance.

**Technology**: In our work to date, we have not distinguished between NextGen and microarray data. Results have shown that the two measurements are consistent under the correct normalization protocol [Bullard et al 2010]. When sufficient data of both kinds is available, we will also try to treat the two kinds of data separately.

Dennis- Are we really going to mix the 2 data types? Is this very dangerous as it may be susceptible to critique. Can we show that once you convert next gen and microarray to correlation you get the same results? I believe Manny has microarray and Next gen both from the same RNA samples in Arabidopsis from a N-treatment experiment for us to test this. Ok we should discuss this but Kranthi can show.

**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 any evidence from the small (if any) experiments conducted in the target species *t* itself. The basic machine learning method will be 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 with 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**: 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 infer that two genes in some target species t are co-expressed. To determine which machine learning method is best, we will test them first on the data-rich species in “leave-out” experiments. As in the preliminary work discussed below, we will attempt to predict the co-expression edges of a data-rich species such as Medicago, using two other data-rich species (e.g. Arabidopsis and Soy). We then test the accuracy of our predictions using left out data from Medicago.

In leave out data validation, to assess the quality of our predictions, we compare the predicted results (e.g. inferred edges in the target species t) that use no expression experiments from the target species, with the results from the experiments in the target species (see Table 2). This gives us precision and recall data to quantify the success of our prediction methods.

**Preliminary Results.** In our initial case study to begin to test our methods, we consider expression data of three “data-rich” species Arabidopsis (A), Soy (G) (*Glycine max*) and Medicago (M) Fig. 4 & Table 2. We selected these three species as an initial proof of concept because (i) there is ample and reliable Affymetrix data for each, and (ii) Medicago and Soybean -- both legumes -- are quite closely related (see phylogenetic tree Fig. X). (NEED KRANTHI HERE- I THINK HE SAID THESE 3 SPECIES ARE EQUIDISTANT. NEED TO INTRODUCE HOW WE WILL CALCULATE PHYLOGENETIC DISTANCE). That is above

**Fig. 4:** Fig 4. **Fig 4.** **Phylogenomic Network Inference Model.**  **Panel A**, describes the equation used on the training data to determine the coefficients (a1, a2, a3..), which are then used for predicting the correlation edges in **Panel B**. Panel B shows an example where the model is trained (e.g. coefficients are determined) using correlation data in Arabidopsis (A) and Soy (G, Glycine max) as well as orthology data between A and G. Then, the model is used to predict correlated edges in M (Medicago) (a neighbor species of G), given the coefficients determined in training, and orthology between genes in A and M and correlations in A. When training on several pairs of species, coefficient a4 (species distance measure) will be used in training and predictions.

We first tried to infer Pearson correlation edges in a “target” species Medicago, knowing correlation edges in a “source” species Arabidopsis trained using another data-rich species (Soy), and the gene-by-gene orthology between genes in Arabidopsis and Medicago (Fig. 4). The equation for network inference is trained using Arabidopsis and Soy under an L-Regularized learning algorithm **[Shalev-Shwartz 2009].** Once we “learn” the rules for network structure using Arabidopsis and Soy daya, we applied this learned equation to infer edges in Medicago, and then test the quality of our predictions using “left out” expression data from Medicago.

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

Estimated correlation in **target species *t*** = a1\*mean of orthology values + a2\*correlation of source pair + a3\*p-value of correlation of source pair (Fig. 4A). This form of the 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 the proof of concept 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, in this case percent identity (BLAST? Gloria: I think so. Let’s check with Manny), between g1 and g1', and between g2 and g2'. We chose the linear form of this equation because such equations are easy to understand and entail discovering just a small handful of coefficients. Surprisingly the results are quite good (see Table 2).

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%, and “in between” otherwise (Table 2). Thus, our machine-learning algorithm predicts which of these three categories (positive, in between, or negative) an edge in the target species is in.

The NEW table 2 should be: (REPLACE TABLE 2 with the following values)

**Table 2: Phylogenomic Network Inference between Arabidopsis (A), Medicago (M), and Soy (G, *Glycine max*).** The table is separated into two parts – (Left) Coefficients obtained from training and (RIGHT) The precision and recall of the correlation predictions. The analysis was performed reciprocally, using A🡪 M for training, and then predicting G, or using A🡪 G as training, and M for test. Recall is less for negative correlation values because the training set is smaller.

We first learned coefficients to the PNI linear equation using Arabidopsis (A) and Soy (G, *Glycine max*). Then, we use those “learned” coefficients to infer edges in Medicago (M), based on edges in Arabidopsis (Figure 4B) Then, we will do another test in which Soy and Medicago reverse roles. Results from these tests are summarized in Table 2.

DENNIS IS THIS A->G train, A->M prediction/validation?????? Yes

Positive Precision: 43971/47572 (%92.43)

Positive Recall: 43971/61247 (%71.79)

Harmonic mean (2\* precision \* recall/(precision + recall)) 81%

Negative Precision: 24628/41904 (%58.77)

Negative Recall: 24628/28229 (%87.24)

Calculated weights: 0.1382 0.6705 0.7203

When we train using Arabidopsis (A) and Soy (G) data, we get values a1 = 0.1382, a2 =0.6705, a3 =-0.7203. When we train using Arabidopsis (A) and Medicago (M) data, we get values a1 = 0.0776, a2 = 0.6303, a3 = -0.6407.

**A->M train, A->G prediction/validation**

Positive Precision: 39494/43435 (%90.93)

Positive Recall: 39494/47634 (%82.91)

Harmonic mean: 87%

Negative Precision: 24808/32948 (%75.29)

Negative Recall: 24808/28749 (%86.29)

Calculated weights: 0.0776 0.6303 -0.6407

The two training sets provide different weights for the coefficients, which can be summarized as a shift in reliance on the orthology value (a1) to the confidence in correlation in source (a2+a3) when we replace Soy(G) with Medicago(M) as the trainer. This shift in reliance may be explained by the fact that Soy has gone through a recent whole genome duplication, and hence often has two paralogs for each Arabidopsis gene, of which only one might still maintain the correlation. Hence, the estimation for correlation between these two species might be more sensitive to the orthology assignment being correct.

**Limitations of the Proof-of-Concept Model and Improvements over the course of this Research**

**Orthology assignments**: In our future work, instead of using reciprocal top BLAST hits, 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’. Unlike in the proof-of- concept study, 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 and median, as the two 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, 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 the machine learning approach, we will be using more than two species for training. For example, based on currently available expression datasets in the dicots, we might train Arabidopsis using data from two data-rich legume species (Soy and Medicago) and then apply the learned model on Cucumis (a data-poor species). In the Monocots- we would train on rice and ?????………. (KRANTHI NEEDS TO REVIEW AND FILL IN THESE SPECIES BASED ON AVAILABLE DATA- WE ALSO NEED TO DO THIS BASED ON SOME PHYLOGENETIC METRIC- NOT JUST BIOLGIST INTUITION LOOKING AT THE TREE……)

Sometimes we will learn from multiple source species. For example, 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 xx and then from Poplar to xx) (KRANTHI FILL IN). Then we will form a joint ranking of the several regression models weighted by genome orthology. The weights going into the joint ranking will also have to be learned.

**(Gloria does not understand this part and where it fits in. Did I give it the correct title?)**

**Parameters to Learn**: Mathematically, for each potential target edge between *g1’* and *g2’*, let *G1\_s* be genes from a source species *s* such that each gene in *G1\_s* has an orthology similarity value to *g1’* at least as large as some threshold value GENESIM. (If there are several orthology methods, then there will be a different similarity threshold for each.) Define *G2\_s* analogously with respect to *g2’*. Now find edges *E12* in *s* between genes from *G1\_s* and *G2\_s* that are in the top k% of all correlation values (we used 5% in our proof of concept, but this will be a discoverable parameter). Then, we will infer an expression for each target edge that is a linear expression in the mean of orthology values for the genes in *E12*, the median orthology value of those genes, the mean and median correlation of the edges in E12, and the mean and median p-value of the edges in *E12*. Other coefficients will be evaluated having to do with other edge types. For example, protein-protein edges in Arabidopsis may help predict co-expression edges in other species. Thus, there will be at least six coefficients to learn. It will also be necessary to learn the best value of k by using cross-validation on the training species. Finally, it will be necessary to assign weights to the predictions from each source species. This will again be achieved by learning.

**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 well-studied “source” species. To summarize the challenge, each regression model will have to fit six coefficients (coefficients on different orthology metrics and strength of correlation the parameter k, the weights of different species based on their phylogenetic distance, and finally the inference algorithm to use. We are optimistic that we will succeed, because the preliminary results have worked out surprisingly well. So far, we have mainly discussed inferring co-expression networks, but we will also infer protein-protein interaction networks as training sets from other data-rich species including rice become available. The techniques are similar and we anticipate that the quality of the results is as high based on the results of [Debodt et al].