**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.

With the increasing number of genome sequences becoming available, it will be common to find a newly sequenced or poorly studied target species “t” *[Gloria: I’ve gone back to t, because our method does not depend on being newly sequenced, just being data poor. I was going to say p, but poor is too negative.]* that is phylogenomically similar to those few of the 21 (and growing) sequenced species on which there is already a substantial body of experiments (see phylogenetic tree Fig. X).

Much of the experimental data will come in the form of genome-wide transcriptome expression measurements, which can used to infer a network of positive and negative expression correlation for the target species *t*. (Our methods will also be used for data supporting other kinds of network relationships such as protein-protein relationships, as they become available.) Our basic co-expression metric will be Pearson correlation because it has been shown to be particularly useful in inferring functionality [PlaNet paper and these references from that paper Usadel 2009, [Klie et al., 2010](http://www.plantcell.org/content/23/3/895.full#ref-28)], though we will also support mutual information [Margolin 2006], and Spearman correlation. Our approach will be to train the algorithm using two or more data-rich source species (s1, s2, …) and then to apply the trained model to target species t.

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, recompute the p-values of those same edges assuming the same set of experiments had been repeated with the same results. 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. x,y,z are measured as data-rich and c,d,e are data-poor [*needs to be done*]).

The 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 *si*. Choose *si* so that its phylogenomic distance from the other source species is approximately the same as the distance between *t* and the other source species. 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 *si*. Use that model to predict edges in *t*.

***The input for our 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 [Altschul 1997 Nuc Acid Resh] hits, OrthologID [Chiu 2006 Bioinformatics], OrthoMCL [Li 2003 Genome Research] , and Inparanoid [O’Brien 2005 Nuc. Acid Resh]. At this point we don’t know which orthology method or methods will work best. Part of the machine learning results will be to determine this. Our preliminary work used BLAST.

 (DENNIS- AGAIN- HOW WILL YOU DECIDE WHICH WORKS BEST? *Gloria: that will be part of the research. Please see the sentences I added.*)

**edgetab: species | gene1 | gene2 | edgetype | strength | p-value | number of different experimental conditions**: gives the strength and the p-value (evaluated using a non-parametric re-sampling approach [Statistics is Easy! Dennis Shasha and Manda Wilson Synthesis Lectures on Mathematics and Statistics 2008 (doi:10.2200/S00142ED1V01Y200807MAS001) Morgan&Claypool Publishers) of a given experimentally determined edge. We consider only experimentally determined edges as an input to the inference algorithm to avoid circular inferences. In our preliminary work, we find correlations that generally hold over all conditions. Note that certain edge relationships may be present only in certain conditions (e.g. drought conditions for plants). In that case, the tools we propose 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, but not the method. DO YOU WANT TO SAY HOW YOU WILL EXPAND TO CONDITION SPECIFIC EXPRESSION? *Done*

**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 whichever measure receives the most weight is the most useful, though several may in fact be useful. [NEED MORE DETAIL HERE ON HOW YOU WILL ASSIGN THIS VALUE AND HOW YOU WILL DETERMINE WHICH IS THE BEST METRIC. *Gloria: we need to do the research. See above sentences.]*

In our work to date, we have not distinguished between NextGen and microarray data. Results [Kranthi has reference?] have shown that the two measurements are consistent except at high intensities. When sufficient data of both kinds is available, we will also try to treat the two kinds of data separately.

Now, to predict an edge between *g1* and *g2* in a target species *t*, we will combine evidence from edges in one or more 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.

The net effect of these analyses will be to find the weighting of different factors that will lead us to conclude that two genes in some species are co-expressed. To determine which method is best, we will test them 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 other data-rich species.

To assess the quality of our predictions, we compare the predicted results (that use no expression experiments in the target species), with the results from the experiments in the target species.

**Preliminary Results.** In our initial case study, we consider steady-state expression data of three species Arabidopsis (A), Medicago (M), and Soy (G) (*Glycine max*) 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). We tested the ability to infer Pearson correlation edges in a “target” species Medicago, knowing only correlation edges in a “source” species Arabidopsis, and the gene-by-gene orthology between genes in Arabidopsis and Medicago (Fig. 4). The equation for inference is built using Arabidopsis and Soy under an L-Regularized learning algorithm **[Shalev-Shwartz 2009]** and tested using Medicago. *Note that the figure must change to eliminate a4 and Sv*

**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.

For our proof of concept study, the regression model had the following form: Estimated correlation in target species = a1\*mean of orthology values + a2\*correlation of source pair + a3\*p-value of correlation of source pair, and + a4\*species distance measure (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). [*We have to take this part away: and a measure of the conservation of this correlation across various phylogenomic distances (a4\*Sv).*] 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, 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.

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% (KRANTHI COMMENTS THAT YOU NEED TO FILL IN ABSOLUTE TERMS *Gloria: yes, but I don’t think it’s necessary as the proof is in the pudding*) of all measured correlations, and “highly negatively correlated”, if their correlation is in the bottom 5%, and “in between” otherwise. Thus, our machine-learning algorithm predicts which of these three categories (positive, inbetween, or negative) an edge in the target species is in.

The table should be:

**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.

Positive Precision: 43971/47572 (%92.43)

Positive Recall: 43971/61247 (%71.79)

Negative Precision: 24628/41904 (%58.77)

Negative Recall: 24628/28229 (%87.24)

Calculated weights: 0.1382 0.6705 0.7203

A->M train, A->G test

Positive Precision: 39494/43435 (%90.93)

Positive Recall: 39494/47634 (%82.91)

Negative Precision: 24808/32948 (%75.29)

Negative Recall: 24808/28749 (%86.29)

Calculated weights: 0.0776 0.6303 -0.6407

We have assigned coefficients to the linear equation using Arabidopsis (A) as source species, and Soy (G, *Glycine max*) as the target. Then, we use those 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.

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

 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). 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 Preliminary Model and Future Work

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. The machine learning mechanism will determine the weights for each term. We will also determine based on cross-validation the best gene orthology threshold GENESIM.

Further, we will incorporate whatever limited expression data is already available in the target species. 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.

Further, we will be using more than two species for training. For example, we might use Arabidopsis, Poplar, and xx as source species, learn a model using Arabidopsis and Poplar these on xx, and then apply the learned model on yy [*Gloria: Please fill in reasonable species*]. We will create a model for each source species independently (e.g. from Arabidopsis to xx and then from Poplar to xx). Then we will form a joint ranking of the several regression models weighted by genome orthology. That weight will also have to be learned.

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 GENESIM. (If there are several orthology methods, then there will be a different similarity metric 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). The 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. 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.**  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, the parameter k, the weights of different species, and finally the inference algorithm to use. We are optimistic that we will succeed, because the preliminary results have worked out so well.

WE NEED SOME DISCUSSION ABOUT POSSIBLE PROBLEMS, ALTERNATE APPROACHES AND SOLUTIONS. *None of the references you have sent proposed alternative approaches.*