**RESEARCH DESIGN**

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

***Rationale***. With the increasing number of genome sequences becoming available, it will be increasingly common to find a newly sequenced species *s,* that is phylogenomically similar to other species on which there are already available experiments. Many of those experiments will be genome-wide transcriptome expression measurements, which can used to infer a network of positive and negative expression correlation for the newly sequenced species *s*. We postulate that a network of expression correlation can be inferred in a species (target) having relatively little expression information by using experiments from one or more significantly better studied species (sources). Our approach will be to train the algorithm using the same several source species on a third species (trainer) where the trainer will be at a similar at a phylogenomic distance from the source species as the target and where the trainer has a “gold standard” in the form of a has moderate to high number of expression studies. Our method here focuses on expression data, but could be used for other kinds of network relationships.

Our neighboring species strategy for inferring edges between genes in species *s,* starts with pair-wise gene expression correlation data on other species. From that data, we will train a machine-learning algorithm to determine whether there will be correlation between two genes in species *s*. In addition to the simple Pearson correlation we use in this preliminary work, we will use related techniques such as mutual information [25], and Spearman correlation. (Note: It is a separate question to determine whether correlation signifies causality. If genes g1 and g2 correlate, g1 is a transcription factor, g2 is not, and g2 has a transcription factor-binding site that the protein associated with g1 can bind to, then this is some evidence for causality. The best test is time-series experiment and analysis [22,26-29], followed by a knock-out or over-expression experiment. As that data becomes available, we will use it as part of our network inference project.).

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

**orthotab: target species| target gene | other species | other gene | orthology val1 | orthology val2 …**: gives the gene-to-gene orthology value, according to several different orthology measures for example: reciprocal best blast [30] hits, OrthologID [4], OrthoMCL [31], and Inparanoid [32].

**edgetab: species | gene1 | gene2 | edgetype | strength | p-value | number of different experimental conditions**: gives the strength and the p-value (the probability it could arise by chance – we evaluate this using a non-parametric re-sampling approach) of a given experimentally determined edge. We consider only experimentally determined edges as an input to this inference algorithm to avoid circular inferences. 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 our preliminary work, we find correlations that generally hold over all conditions.

**species1 | species2 | species similarity measure1 | species similarity measure2**: measures sequence similarity according to one of a number of criteria (e.g. distance based, for example average percent identity of protein sequences, or through parsimony).

Now, to predict an edge between *g1* and *g2* in target species *s*, we will combine evidence from edges in one or more source species, as well as evidence from experiments in species *s* itself. The basic method will be regression and regression trees, with a penalty for complexity.

For the sake of performance and robustness to noise, ***we will use some mixture of the following three approaches*:**

1. **Random Forests** [33,34] 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** [35] 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** [36] 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 this analysis will be to find the weighting of different factors that will lead us to conclude that two genes in some species are correlated. Then, using available Arabidopsis time-series data [22], and other datasets that are currently being generated in our lab and others, we will combine correlation with time-series [22,26-29] and perturbation approaches using Graphical Lasso [37] to form causal networks.

**Preliminary Results.** In our initial case study, we consider steady-state data on three species Arabidopsis (A), Medicago (M), and Soy (G) (*Glycine max*) Fig. 4 & Table 2. We selected these three species as an initial test case because (i) there is ample and reliable Affymetrix data for each, and (ii) Medicago and Soybean -- both legumes -- are quite closely related (more so than Arabidopsis and Rice, as we discuss in the preliminary work for Aim 2). We tested the ability to infer Pearson correlation edges in a “target” species, knowing only correlation edges in a “source” species, and the gene-by-gene orthology between genes in the source species, and genes in the target species (Fig. 4). For this study, we analyze only those genes that are conserved across all three species - Arabidopsis, Medicago and Soybean.

**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 used stochastic gradient descent as the machine learning technique, by training a linear equation of the form:

Estimated correlation in target = a1\*mean of orthologous 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 confidence of correlation in the target species depends on the confidence in the orthology assignments(a1\*MOv), strength and confidence in the correlation of expression in source species(a2\*Cs and a3\*Ps) and a measure of the conservation of this correlation across various phylogenomic distances(a4\*Sv) Here, mean of orthologous values is calculated as follows: if g1 and g2 are the source pair, and g1' and g2’ are the potential 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.

However, this equation ignores relevant information because (i) it ignores experiments already done in the target species, and (ii) many gene pairs (besides reciprocal best blast hits) in the source species, may be relevant to the target pair g1 and g2, for example paralogs. For point (i), when some expression data about the target species is available, we will add a term of the form c\*prelimcorrelation(g1,g2), that takes into account the correlation between g1 and g2, based on the experiments performed so far, though we did not do that here. For point (ii), we may require some form of aggregation over the gene pairs of the source species that are orthologous above a threshold to g1 and g2. (Note: That is unnecessary in this preliminary study, where we focus on reciprocal best blast hits.) When using a threshold, cross-validation on a training set, would set the level of the threshold. Finally, once we have data on many pairs of species, we will include a4, that measures the similarity of species.

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%, 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, between, or negative) an edge in the target species is in. To assess the quality of the 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.

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.

**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.

When we train using Arabidopsis (A) and Medicago (M) data, we get values a1 = 0.0276, a2 = 1.2619, a3 = -0.8109. We then test this using Arabidopsis and Soy (G), to get 18,292 predicted highly positive correlations, 3,684 predicted highly negative correlations. This gives us a recall of 0.91, for highly positive correlations, with a precision of 0.96, and for highly negative correlations, we get a recall of 0.62, and precision of 0.89 (Table 2).

When we train using Arabidopsis (A) and Soy (G) data, we get values a1 = 0.0894, a2 =1.0571, a3 =-0.0063. We then test this using Arabidopsis (A) and Medicago (M), to get 21,384 predicted highly positive correlations, and 228 predicted highly negative correlations. This gives us a recall of 0.99 for highly positive correlations, with a precision of 0.98, and recall of 0.01 and precision of 0.8, for highly negative correlations. Recall is less for negative correlation values because the training set is smaller (Table 2).

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 2 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. To address this issue, we propose to assign confidence to the ortholog assignments based on expression data (NEW Aim 3). Additionally, using multiple species, with varying ploidy levels, at the training stage is expected to alleviate this apparent distortion in orthology assignments.

In this preliminary test, we only used one pair of species to train. As we develop this aim, we will train on several pairs of species, in which case coefficient a4\*species distance measure will be used in both training and predictions. Note also that this preliminary experiment makes predictions only about pairs in the target species whose members are highly orthologous to some pair in the source species. Our recall numbers would be much lower if we were measuring our success against identifying ALL correlation edges in the target species. Orthology helps and may identify some of the most important edges, but this technique complements rather than replaces in-species experimentation.

**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 species, by inference from a well-studied species. As more data about the species becomes available, we then apply the rest of our workflow to find a refined causal network.