Dennis- THESE POINTS STILL NEED TO BE ADDED to AIM 1

1. Comment on ability to use “Gene space” assemblies that might be available in future, in lieu of whole genomes. Discuss measures of completeness of such assemblies. (Kranthi?) [Dennis has done this in the context of subconditions. See if you like this]

------------------------------------------------------------------------------------------------------------

**Aim 1: Development of InferNET: A machine learning approach to “learn” networks in data-rich species and infer 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 (whether crop or non-crop), based on ***gene networks*** from several data-rich species. The InferNET approach is inspired by the *Robin Hood philosophy* -- "learning from the rich and giving to the poor". Such inferred networks in the data-poor target (e.g. crop) species may then be used to derive hypotheses and identify potentially important genes in that species 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), and we will apply them as data becomes available for at least a few species.

***Novelty***: ***InferNET*** differs from existing comparative network tools in plants, because InferNet uses data-rich species to ***learn*** regulatory networks in data-poor species. By contrast, the existing tools for comparing plant gene networks creates networks only for data-rich species and then compares them post-hoc (e.g. CoP [Ogata 2010], Starnet [Jupiter 2009], ATTED-II [Obayashi 2011, and PlaNet [Mutwil 2011]). Additionally, most existing network tools – with the exception of PlaNET- 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 thanks to Next-Gen sequencing, 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). InferNET capitalizes on the vast wealth of accumulated knowledge about gene interactions across species to learn the rules for network inference in poorly studied target species. This will be particularly valuable for new and emerging crops and new “boutique” crop species.

**Data-types**: Much of the experimental data for expression atlases (and new species) will come in the form of genome-wide transcriptome expression measurements. In the InferNET 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*. 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 beyond Arabidopsis is rice thanks to the NSF Rice protein interaction project [REF].

**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 reference above]. 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 left-out data from data-rich species, as described in detail below.

**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 (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 ) [Needs to be done]……………. It is conceivable that certain species might have many experiments that explore very few conditions. In that case, we could use the diversity-finding technique advocated by PlaNet [Mutwil 2011]. In practice however, this is not an issue, because different researchers tend to have different interests which will drive the selection of experiments to analyze.

**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*. Choose species *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 species) 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 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. At this point of our preliminary analysis, 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**: 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. However, certain edge relationships may be present only under certain conditions (e.g. drought conditions) or in certain tissues (Gene Spaces). The proposed tools could be used for all experiments or just for the conditions of interest, in which case, we would choose the subset of “edgetab” corresponding only to those specific conditions, as well as a set of “control” experiments under standard growth conditions to filter out genes whose expression does not change (e.g. housekeeping genes). That is, focusing on one or more conditions or tissues 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 of species according to several criteria (e.g. distance based, for example average percent identity of protein sequences, or through parsimony). We don’t know *a priori* which similarity 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 in our “combining rule”, where a weight that is high in absolute value suggests importance.

**RNA Expression Technology**: In our work to date, we have considered NextGen and microarray data from Affymetrix as separate datasets. However, 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 and together to compare the results.

**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 data-poor 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 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 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. 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 analysis gives us precision and recall data to quantify the success of our prediction methods.



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 Soy (Fig. 4). 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. 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 the expression data from Soy.

This “InferNET” learning approach is superior to the Interolog approach which considers only BLAST scores to infer edges [Insert Interolog reference]. 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*** = 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?), 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. As mentioned above, the results are quite good (see Table 2 for details). The learned model also reveals why interolog 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. Instead, the correlation of the edge in the source species by itself predicts the correlation of the edge 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%, and “in between” otherwise (Table 2). 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) from pairs of genes that are 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 network for Medicago**

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

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

Interolog | 81% (xxx/yyy) | 69% (xxx/yyy) | 65% (xxx/yyy) | 78% (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 [TO BE FILLED IN].

**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 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’. 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 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**: (Kranthi needs to work on this section).

Further, in future development and testing of the machine learning approach we will use in InferNET, 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 THIS SECTION ABOVE 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 “combining rule” consisting of a learned joint ranking of the several regression models weighted by genome orthology.

**Technical Discussion of the Learning Problem**: 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 learn weights to the predictions from each source species to form a combining rule.

**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].