Using Interaction Networks for Biomarker Discovery (Dennis version)

Note that in this version, I’m using Pearson and Spearman correlation (Ulises suggested this) to establish the p-value of the relevance of a gene to Nitrogen efficiency. This requires less data than logistic regression approaches and let’s us use new ecotype data easily. Otherwise most of the preceding framework stays the same.

I address several important questions at the end.

Goals:

1. Determine a network(s) of genes that can be used as markers for Nitrogen use efficiency.
2. As per NSF director, it would be great to show that the Arabidopsis interaction network is useful.

**Data:**

1) Nitrogen expression data from different ecotypes. Data is available from Miriam’s 7 ecotypes study

2) A measure of the NUE for the different ecotypes. The NUE information will be provided by Ying.

3) Potentially split root data that measures features like the transport of hormones or other chemical entities.

**General Strategy:**

Use a subset of ecotypes to identify networks of genes that can be used to predict NUE. Then use these networks as input to classifiers to predict NUE of the remaining ectoypes.

**The Algorithm:**

1) Start by determining p-value for each gene where the null hypothesis is that the expression of the gene does not change when looking at the nitrogen effect. Do this by computing the correlation (both Pearson and Spearman) of gene expression per ecotype in low nitrogen settings and in high nitrogen settings with Nitrogen efficiency as measured by N15 uptake or biomass. Then use a permutation test to determine p-values.

2) Calculate Paired-score for **ALL** genes (not just the ones with an edge) in the Arabidopsis interaction network. ***Paired-score*** will be calculated by taking the sum of the –log(pvalue) for each gene in pair (this can be for triplets as well). This will result in a null hypothesis distribution of “Paired-scores” all possible paired

3) Identify significant protein interactions. Paired-score of each pair in the interaction network will be compared to the Paired-score null distribution to determine its significance. The Paired-score should also be adjusted for multiple hypothesis testing based on "Statistical significance for genomewide studies" by John Storey and Robert Tibshirani PNAS August 5, 2003 9440-9445

4) Create networks by joining pairs that share a common gene.

5) For each network, use the expression values from the training dataset to build a regression model. Test the model by trying to predict the NUE using expression data from test dataset.

6) Repeat 1-5 leaving out one ecotype at a type and keep track of the pairs and their performance for each run. The performance of the pair is reflected by the performance of the network it is in.

**Which Other Ecotypes Do We Need**

I think we can determine how many more ecotypes we need by seeing how much better our model is as we add more of the ecotypes we have. Thus, if we have 7 ecotypes, we imagine training with 3 and see how good our predictions are on each of the remaining 4 (in a leave out manner). Then we train with 4 and see how we do with the remaining ones and so on. As long as we’re improving substantially each time, we need to continue. This does not tell us how many more we need (unless our accuracy plateaus before we reach 7), but it may tell us which are the most useful (whether extremal ones in terms of nitrogen efficiency or less extreme ones).

**Which Genes to Target for Treatments**

Highest priority should be the ones that the regression model gives the highest positive or negative coefficients to. A gene gets extra weight if it’s conserved across species. Ditto for gene pairs.

**How to Tell Whether Protein-Protein Interaction Networks Helped**

Look at the accuracy of the regression model both with and without.