A Bayesian Network Approach To Building Gene Regulatory Networks

Jeremy Burrell, Chao-Jen Wong, Joan Chen

San Diego State University

Abstract

Continuous abiotic stresses (such as extreme temperatures, high winds and edaphic conditions) can have adverse effects on plant life. This can become a major constraint in crop production. In order to alleviate these problems, it is important to understand the cold stress mechanisms at the molecular and cellular levels, regarding the signaling pathways from cold perception to activation of gene expression. We took a system biology approach, aiming to build the gene regulatory network using the cold stress microarray data that have been obtained, to gain a better understanding of plant responses to cold stress at the molecular level. The results generated from this approach will be utilized to generate hypothesis which can be tested by other experimental approaches.

To discover the gene regulatory network in response to cold stress on plants, we have employed a new proposed reverse engineering method implemented in the "GeneNet" R package (Opgen-Rhein and Strimmer 2007). This method is a heuristic algorithm applying an extended graphical Gaussian model and dynamical correlation shrinkage estimators to the inference of partial directed causal networks from high-dimensional time series expression data. Here, we analyze the expression data of 3, 376 interesting genes collected at 7 different time points (0, 0.5, 1, 3, 6, 12, 24 hours) from a microarry experiment. The resulting network shows 150 significant edges connecting 45 nodes.

Background

Plants respond to various environmental stresses such as cold stress by activating the expression of a large number of genes through a series of signal transduction cascades. These signal transduction cascades, often times interact with each other, compose the signaling networks, which eventually govern the transcriptional regulation of expression of genes which are involved in protecting plants from cold damage. Thus understanding the molecular mechanisms of the signal transduction network will greatly help in plant's stress tolerance, which in turn, will have a big economic influence on agricultural business.

Methods and Materials

The GeneChip® microarray data was downloaded from publicly available database (http://www.arabidopsis.org/info/expression/ATGenExpress.jsp). Plant total RNA were extracted from leaves of 18-day old plants treated with cold (40 C) for the indicated time periods, according to the standard protocol. cRNA labeling, hybridization and scanning was described in Hannah et al (2006). The Arabidopsis ATH1 whole genome microarray from Affymetrix was used for this set of cold stress experiments. All the data analysis was performed in R. The raw data file (.cel) files from chip hybridization were read directly into R. Robust Multi-Chip Analysis (RMA) algorithm was used for chip background correction, normalization and to obtain gene expression estimates.

Local-Pooled Error (LPE) Test

Using the Local-Pooled Error (LPE) test we estimate the significance of each genes differential expression. LPE test is based on pooling errors within genes and between replicate arrays for genes in which expression values are similar (Jain et. al. 2003).
LPE is most useful when dealing with a low number of replicates (ie 2-3). This is one reason for using the LPE test rather than the the traditional 2-sample t-statistics. Also, within-gene estimates of variance doe not provide a reliable hypothesis testing framework.

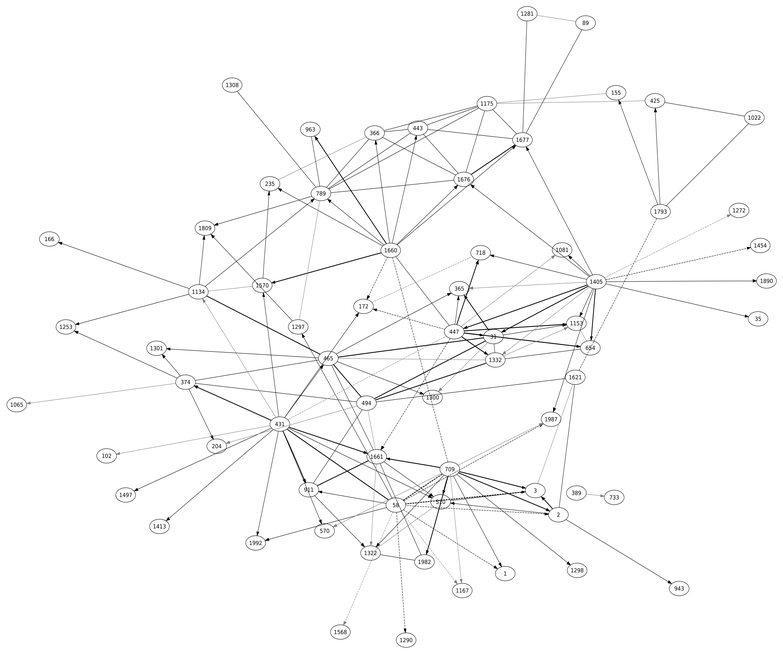
Procedures for LPE and how genes were selected to be used in GeneNet:

1. Evaluate baseline error distribution for each gene between replicates at each time point.

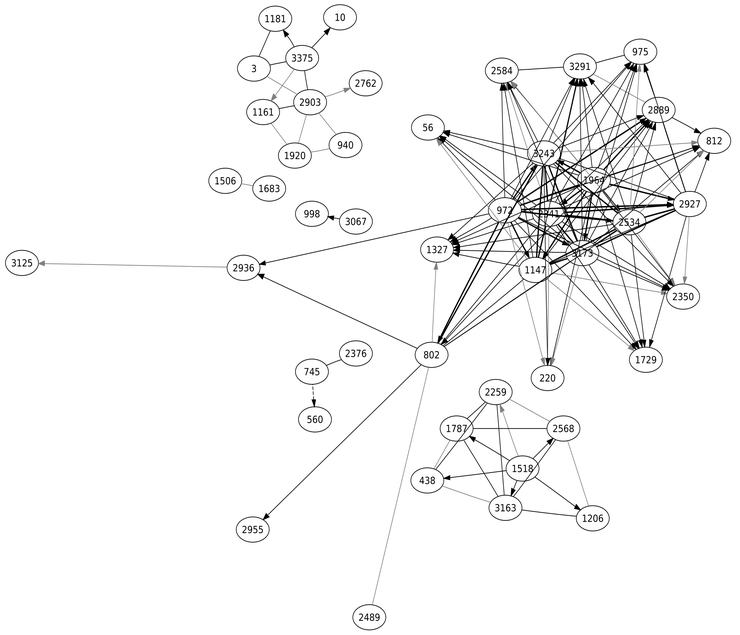
2. Calculate z-statistics for each gene. Here we allow positive z-scores denote up-regulation of genes and negative z-scores denote down-regulation.

3. Select genes with p-value less than or equal to 0.01

4. Once we have our up and down regulated genes with p-value less than or equal to 0.01, we further select genes with log2-fold change greater than or equal to 1.



Up Regulated Network



Up and Down Regulated Network

Node Number	Probeset	Target Description
1	244904_at	hypothetical protein
2	244977_at	cytochrome b/f
3	245015_at	large subunit of riblose-1,5-bisphosphate carboxylase/oxygenase
31	245397_at	auxin-responsive protein IAA1
35	245422_at	putative protein
58	245638_s_at	F5A9.10 unknown protein
89	245928_s_at	vegetative storage protein Vsp1
102	246103_at	putative protein
155	246687_at	putative protein proline-rich protein APG
166	246860_at	putative protein various predicted proteins
172	246888_at	putative protein
204	247252_at	unknown protein
235	247463_at	embryo-specific protein - like embryo-specific protein 3
365	248801_at	homeobox-leucine zipper protein-like
366	248807_at	pectin methylesterase-like
374	248888_at	potassium channel protein KAT1
389	249065_at	putative protein
425	249614_at	putative protein predicted proteins
431	249645_at	thionin Thi2.2
443	249813_at	acyltransferase
447	249895_at	male sterility 2-like protein
465	250012_x_at	auxin-induced protein-like
494	250327_at	putative serine rich protein
510	250446_at	nucleoid DNA-binding protein cnd41
570	251221_at	putative protein ER6 protein
654	252204_at	putative protein
709	252957_at	glycine-rich protein 2 (GRP2)
718	253066_at	1-aminocyclopropane-1-carboxylate synthase - like protein
733	253203_at	arginine decarboxylase SPE2
789	253753_at	glycine-rich protein like glycine-rich protein
911	254882_s_at	putative protein various predicted reverse transcriptases/transposons
943	255403_at	putative GH3-like protein
963	255782_at	transcription factor
1022	256597_at	acidic ribosomal protein P2b (rpp2b)
1065	257254_at	salicylic acid carboxyl methyltransferase
1081	257642_at	putative HLH DNA-binding protein
1134	258196_at	hypothetical protein
1153	258399_at	early auxin-induced protein, IAA19
1167	258545_at	putative GTPase
1175	258675_at	putative nonspecific lipid-transfer protein
1253	259391_s_at	delta 9 desaturase
1272	259583_at	hypothetical protein
1281	259640_at	beta-glucosidase
1290	259713_at	unknown protein similar to phosphate translocators
1297	259766_at	unknown protein
1298	259767_s_at	unknown protein
1300	259784_at	auxin-induced protein
1301	259787_at	auxin-induced protein
1308	259871_at	nodulin-like protein
1322	260007_at	unknown protein
1332	260152_at	putative IAA6 protein
1405	260957_at	delta 9 desaturase
1413	261046_at	flavonol 3-o-glucosyltransferase
1454	261450_s_at	O-methyltransferase
1497	262003_at	unknown protein
1568	262797_at	putative sugar transporter protein
1570	262819_at	putative cytochrome P450
1621	263509_s_at	hypothetical protein
1660	263979_at	En/Spm-like transposon protein
1661	263981_at	unknown protein
1676	264146_at	hypothetical protein
1677	264147_at	receptor-like protein glossy1 (gl1)
1793	265441_at	unknown protein
1809	265698_at	hypothetical protein predicted by genscan
1890	266606_at	putative AP2 domain transcription factor
1982	267545_at	unknown protein
1987	267595_at	putative glucanse
1992	267644_s_at	unknown protein

Node Number	Probeset	Target Description
3	244932_at	PSI 9KDa protein
10	245002_at	PSII D2 protein
56	245317_at	hypothetical protein
220	246302_at	Ca2+/H+-exchanging protein-like
438	247717_at	nonspecific lipid-transfer protein precursor
560	248528_at	putative protein
745	249895_at	male sterility 2-like protein
802	250199_at	putative protein
812	250287_at	AP2 domain containing protein RAP2.6
940	251379_at	putative protein ABP130 protein
972	251428_at	beta-glucosidase-like protein
975	251438_s_at	putative protein
998	251625_at	beta-1,3-glucanase 2 (BG2) (PR-2)
1147	252570_at	isovaleryl-CoA-dehydrogenase precursor (IVD)
1161	252743_at	guanine nucleotide-exchange - like protein
1181	252957_at	glycine-rich protein 2 (GRP2)
1206	253073_at	cytochrome P450 monooxygenase - like protein
1327	253829_at	Medicago nodulin N21-like protein
1506	254819_at	pEARLI 1-like protein
1518	254915_s_at	cysteine proteinase contains similarity to cysteine proteinase RD21A
1683	256252_at	glucosyl transferase
1729	256589_at	cytochrome P450
1787	256994_s_at	limonene cyclase
1920	258097_at	transport protein
1964	258327_at	unknown protein
2259	260130_s_at	beta-glucosidase
2341	260668_at	unknown protein
2350	260741_at	hypothetical protein
2376	260957_at	delta 9 desaturase
2489	261684_at	hypothetical protein
2534	262047_at	hypothetical protein
2568	262260_at	major latex protein (MLP149)
2584	262373_at	hypothetical protein
2762	263654_at	unknown protein
2889	264524_at	tat-binding protein
2903	264625_at	hypothetical protein
2927	264777_at	unknown protein similar to L-allo-threonine aldolase
2936	264809_at	superoxidase dismutase
2955	264923_s_at	peroxiredoxin
3067	265837_at	unknown protein
3125	266165_at	putative copper/zinc superoxide dismutase
3163	266353_at	unknown protein
3173	266393_at	late embryogenesis abundant M17
3243	266901_at	hypothetical protein
3291	267168_at	putative alcohol dehydrogenase
3375	AFFX-Athal-25SrRNA_s_at	Arabidopsis thaliana

GeneNet

GeneNet implements a statistical learning algorithm proceeding in two steps: (i) transform correlation network into a partial correlation network, which is an undirected graph displaying the linear associations, and (ii) convert the undirected graph into a partially directed graph by estimating pairwise log-ratios of standardized partial variances.
Consider a linear regression with Y as response and the set of vectors $\{X_1, \ldots, X_k, \ldots, X_K\}$ as covariates. The regression coefficient estimator of Y is defined by

$$\beta_k^\gamma = \underbrace{\tilde{\rho}_{\gamma k}}_{\mathcal{A}} \underbrace{\sqrt{\frac{SPV_\gamma}{SPV_k}}}_{\mathcal{B}} \underbrace{\sqrt{\frac{\sigma_\gamma^2}{\sigma_k^2}}}_{\mathcal{C}},$$

where $\tilde{\rho}_{\gamma k}$ is the partial correlation between Y and X_k, and $SPV_\gamma = \tilde{\sigma}_\gamma^2/\sigma_\gamma^2$ called standardized partial variance.

$\mathcal{A}$: Establish edges between nodes. An edge is drawn if $\mathcal{A} \neq 0$ determined by a multiple testing method (Efron 2004).
$\mathcal{B}$: Determine the directionality of edges of a causal network. If $\log \mathcal{B} \neq 0$ evaluated by the multiple testing, edges are directed in a fashion that direction of arrow points from the node with larger standardized partial variance to the node with smaller standardized partial variance.

Summary and Future Research

We have built a gene regulatory network in plant cold stress response, based on the time point microarray data. We have identified a number of master regulatory genes which might play important roles in transducing cold stress signals to other genes. The network built in this work sets up a framework for bench scientists to design experiments to further test the importance of those genes, which might be important in cold stress response.

References

[1] Efron, B. (2004), Large-scale simultaneous hypothesis testing: the choice of a null hypothesis , ***J Amer Statist Assoc,*** 99:96-104.
[2] Hannah MA, Wiese D, Freund S, Fiehn O, Heyer AG, Hincha DK. Natural genetic variation of freezing tolerance in Arabidopsis. ***Plant Physiol***. 2006 Sep;142(1):98-112. Epub 2006 Jul 14.
[3] Jain, N. et. al. (2003), Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrys, ***Bioinformatics***, 19, 15:1945-1951.
[4] Opgen-Rhein, R. and Strimmer, K. (2007), From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data, ***BMC Systems Biology,*** 1:37.

Acknowledgements

Supported as a trainee by NIH RoadMap Initiative award T90 DK07015