Effects of Drought Stress on Gene Expression Patterns in the Needles of Loblolly Pine Trees: Towards a Problem-Solving Environment for the Analysis of Microarray Data.

Lenwood S. Heath, Naren Ramakrishnan, Ronald R. Sederoff, Leonel van der Zyl, Dawei Chen, Ying-Hsuan Sun, Boris I. Chevone, Shun-Hwa Li, Keying Ye, Ross Whetten, and Ruth Grene Alscher , Departments of Computer Science (LSH, NR, DC), Plant Pathology, Physiology, and Weed Science (RGA, BIC), Dairy Science (S-H Li), Statistics (KY),Virginia Tech, Blacksburg , VA 24061 and Forest Biotechnology Group (RRS,Y-H Sun, RW), Department of Forestry, North Carolina State University, Raleigh, NC 27695

Introduction

Our long-term biological goals are to study how plants adjust their metabolism, through global changes in gene expression, to changes that induce oxidative stress, and to determine how exposure to oxidative stress affects wood production in particular, and development in stems and needles in general. Several stress-responsive genes have been identified that are activated during the juvenile to mature transition in differentiating loblolly xylem tissue (Sun et al, in preparation).

Our bioinformatics goal is to build a problem-solving environment (PSE) for analyzing microarray data, facilitating gene discovery through clustering algorithms, inductive logic programming, and data mining using *a priori* knowledge. A long-term bioinformatics goal is to model cellular regulatory networks. This PSE will be flexible enough to allow the biology researcher to configure the analysis process with a selection of tools at every step of the analysis; to visualize the results of one or many experiments along a variety of dimensions; and to assist in the exploration of hypotheses that might be supported by data already available. While PSEs are becoming increasingly accepted in domains such as cell biology and genome information management, there does not exist a comparable comprehensive facility for microarray data analysis, despite the intense attention currently paid to microarray technology (http://www.bsi.vt.edu/ralscher/gridit).

Some preliminary data, documenting drought-mediated changes in gene expression in the needles of loblolly pine seedlings, are shown here.

Materials and Methods

Plant material and physiological treatments. 2 year-old seedlings of a loblolly pine clone "D" were obtained from North Carolina State University in mid June, 1999, and allowed to adapt to greenhouse conditions at Blacksburg, VA, for 6 weeks. On July 29, 1999, trees

were divided into three groups, of four plants each, for moisture stress treatments. The three treatment groups were 1) moist (watered every 2-3 days), 2) intermediate (watered when needle water potential reached -10 to -12 bars) and 3) dry (watered when needle water potential reached -16 to -18 bars). Water potential was measured periodically on a subset of needle fascicles from each treatment using the pressure bomb technique (Soilmoisture Equipment Corp., Santa Barbara, CA). Seedlings were watered when the designated water potential value was attained. In the intermediate water stress treatment, trees were subjected to four drought cycles lasting 7 to 8 days each and in the severe treatment, trees were subjected to three cycles lasting l1 to 12 days each. Seedlings were harvested September 2, 1999, one day after the final watering (hydrated state).

*Array*s. A printing robot was built at NCSU, from the instructions given at (http://cmgm.stanford.edu/pbrown/mguide/index.html). Four pins in the Stanford configuration were used to spot 384 amplified cDNAs from differentiating xylem, shoot tips or pollen cones (Allona et al, 1998) chosen on the basis of function and/or previously documented response to drought stress (Table 1). Each cDNA was spotted a total of 16 times in eight replicated sets of randomized designs in two separate arrays (4 x 24 x 16 per array) on glass slides (ArrayIt, Telechem, Sunnyvale, CA, Figure 1). After printing, slides were post-processed according to the manufacturers' instructions, and dried. Slides were stored desiccated at room temperature.

Table 1. Responses to mild drought stress of clones represented on loblolly pine est microarray

Category	*Description*	*Total in funcrtional category*	*Increase*	*Un affected*	*Decrease*
1	Phenylpropanoid pathway	33	12	20	1
1.1	Lignin biosynthesis	17	5	12	1
1.2	Oxidases	7	2	5	0
1.3	Non-lignin	2	2	0	0
2	Cell Wall related	40	11	27	2
2.1	Cellulose	3	1	2	0
2.2	Hemicellulose	0	0		0
2.3	Pectin	4	1	3	0
2.4	Arabinose, xylose	12	3	9	0
2.5	Sucrose metabolism	0	0		0
2.6	Extensins	11	2	9	1
2.7	Other cell wall proteins	8	3	5	1
2.8	Oxidases	0	0		0
3	Signal Transduction	19	4	15	1
3.1	Calcium-associated	5	2	3	0
3.2	Kinases	3	1	2	0
3.3	Transcription factors	2	1	1	1
4	Plant Growth Regulation	10	4	6	0

4.1	ABA	1	0	0	0
4.2	Auxin	3	2	1	0
4.3	Ethlyene	4	1	3	0
4.4	Development	1	1	0	0
4.4.1	Seed	1	1	0	0
4.4.2	Senescence	0	0		0
5	Environment	40	18	19	3
5.1	Abiotic	33	15	15	3
5.1.1	Heat	11	5	6	1
5.1.2	Drought	5	2	3	0
5.1.3	NADPH/Ascorbate/Glutathione scavenging pathway	11	5	6	2
5.1.4	Xenobiotics	3	2	1	0
5.1.5	Cold	0	0		0
5.1.6	Non-plant	2	1	1	0
5.2	Other antioxidant processes	4	2	2	0
5.3	Biotic	2	1	1	0
6	Respiration, N metabolism, nucleic acids	5	1	4	0
7	Protease-associated	21	9	12	0
7.1	Ubiquitin-associated	17	8	9	0
7.2	Other proteases	1	0	0	0
8	Chloroplast-associated	20	5	15	1
8.1	Thylakoid-associated proteins	6	2	4	1
8.2	Photosynthetic carbon metabolism	12	3	9	0
8.3	Pigments	1	0	0	0
8.4	ATP synthesis	1	0	0	0
9	Nucleus	5	1	4	0
9.1	Chromosome-associated	5	1	4	0
10	Gene expression	26	15	9	1
10.1	Translation	4	2	2	0
10.1.1	Regulation of Translation	1	0	0	0
10.2	Transcription	7	4	3	0
10.2.1	Regulation of Transcription	6	3	3	0
10.3	Post-transcriptional processing	0	0		0
10.4	Post-translational processing	15	9	6	1
10.4.1	Oxidoreductases	0	0		0
10.4.2	Chaperonins	4	1	3	1
10.4.3	Thiols	6	5	1	0
11	Carbon metabolism	19	6	13	1
11.1	Glycolysis	5	0	5	1
11.2	Oxidaitve pentose phosphate pathway	0	0		0
11.3	RPPP	1	0	0	0
11.4	Sucrose metabolism	7	4	3	0
11.5	Starch metabolism	0	0		0
11.6	Glyoxylate cycle	1	0	0	0
12	Cell Structure	24	9	15	0
12.1	Transport protein, including aqauporin	24	9	15	0
12.2	Other membrane proteins	0	0		0
13	Mitochondrion	2	2	0	0
14	Peroxisome and Glyoxysome	0	0		0
15	Cytoskeleton	6	1	5	0
15.1	Tubulin	3	0	3	0

15.2	Actin	3	1	2	0

Figure 1. Illustration of randomized microarray design with replications

Put Randomized Design Drawing here

Probe Synthesis. Total RNA was isolated from stems and needles by the method of Chang et al (1993), with modifications, and stored at -70°. cDNAs were synthesized using SuperScript II. RNA isolated from the needles of intermediate stress and control trees were the source of the probe pair used to produce the preliminary data shown here. Duplicate hybridizations were carried out. Hybridization and labeling with Cy3 or Cy5 was carried out using the Genisphere 3DNA Expression Detection Kit, according to manufacturers' instructions, using 20ug RNA as starting material. Hybridized slides were scanned in ScanArray 4000 (GSL Luminomics).

Computational tools Microarray Suite (Scanalytics, Fairfax, VA) running on a Macintosh G3 was used to grid the overlaid Cy3 and Cy5 image files for the four microarrays and to extract a calibrated ratio for each spot. Microarray Suite employs techniques described in Chen et al. (1993) to differentiate spot pixels from background pixels and to calculate a calibrated ratio. The remaining analysis was done with tools that we are developing under Unix and Linux operating systems for our problem solving environment. Perl scripts were employed throughout: to generate the randomized design; to convert that design to instructions to drive the TECAN robot; to extract clone annotation from a variety of sources including GenBank, the Computational Biology Center at the University of Minnesota (http//:www.cbc.umn.edu), and our local resources; to represent and manipulate the complicated associations between each clone and its spots on the microarray; to extract the spot statistics from the Microarray Suite output while organizing them by clone; and to analyze the expression levels evident for each clone. We employed Matlab and Excel for data visualization, and Postgres for the inductive logic programming (ILP) that searched for associations between a category and the expression levels of clones in that category.

Results

Data Analysis. Statistical analysis of the relative fluorescence intensities for each of the 384 clones (Cy3/Cy5, stress/moist), each represented by 16 log calibrated ratio values, revealed that, with only 3 exceptions, the standard deviations of the relative intensities are clustered around 0.3 and there is no correlation between mean log ratio and

standard deviation. It is also clear that the 16 log ratios for a typical clone does not follow a normal distribution, but rather show a large, relatively even spread. This suggests that traditional statistical analysis that employs the mean and standard deviation will be overly pessimistic in identifying clones that are up- or down-expressed. Indeed, there are only 3 clones for which the absolute value of the mean exceeds the standard deviation. Using the even spread of the log ratios, we assume that a clone whose expression is not affected by intermediate drought stress will show a uniform distribution with mean log ratio of 0.

With a uniform distribution, the probability that any particular log ratio is positive (or negative) is 0.5. The number of positive (or negative) log ratios follows a binomial distribution with parameters 16 and 0.5. Under our assumption, the probability that the number of positive log ratios (or negative log ratios) of a clone whose expression was unaffected by drought stress is 12 or more is only 0.0384064. The corresponding probability for 11 or more is 0.105057. Hence we can identify a clone with 12 or more positive log ratios as being up-expressed with a probability of about .96 and a clone with 11 as being up-expressed with a probability of at least .89. This analytical approach illustrates the power of the 16 replicates and suggests the use of more replicates to establish higher levels of confidence. By these criteria, 74 clones either up-expressed or down-expressed at the .96 level, and 153 clones at the .89 level. In the future, we intend to use more refined assumptions about the distribution to better estimate confidence levels.

Effects of Drought Stress on individual genes. The stressed trees had been subjected to four drought cycles over a period of 6 weeks, effectively functioning in the presence of elevated reactive oxygen species throughout that time. The clones whose expression appears to have been altered in needles of the stressed trees primarily fall into three groups, each of which is associated with antioxidant or defense functions. The three major groups consist of phenylpropanoid metabolism, and other cell wall proteins), antioxidant defense genes (Alscher et al. 1997), and genes associated with thiol-containing molecules and metabolic regulation (glutaredoxin, periredoxin) and/or post-translational processing (protein disulfide isomerase). The latter class is, in effect, another group of defense molecules, protecting cellular constituents that have been oxidized as a consequence of stress imposition. The expression of genes in these classes appears to have increased in the drought-stressed samples.

Effects of drought stress on genes grouped according to functional categories as revealed by ILP. Once we determined for each of the 384 clones whether it was up-expressed, down-expressed, or unchanged, we represented the expression of each clone as facts in a logic programming language. These facts were combined with our functional classification of the clones to establish a basis for inductive logic programming. Using

inductive logic programming, we automatically derived 14 rules that held with a high level of confidence for that basis (Table 2).

Table 2. Rules derived by Inductive Logic Programming (ILP)

Category	*Implication*	*Positive evidence*	*Negative evidence*
1	Unchanged	20	13
1.3	High	2	0
2	Unchanged	28	13
3	Unchanged	17	5
4	Unchanged	6	4
6	Unchanged	4	1
7	Unchanged	12	9
8	Unchanged	14	6
9	Unchanged	4	1
10	High	15	11
11	Unchanged	12	7
12	Unchanged	12	9
13	High	2	0
15	Unchanged	5	1

In continuing with this approach, we will identify more refined rules using additional data to be obtained in the next round of microarray comparisons. With a semi-structured Postgres database, we will be able to combine expression data from numerous experiments using ILP, visualization, and data mining techniques.

References

I. Allona ,M. Quinn,E. Shoop, K. Swope, S. St Cyr, J. Carlis, J. Riedl, E. Retzel, M. M. Campbell, R. Sederoff, and R. Whetten (1998) Analysis of xylem formation in pine by cDNA sequencing, Proc Natl Acad Sci USA 95(16), 9693--8.

S. Chang, J. Puryear, and J. Cairney (1993) A simple and efficient method for isolating RNA from pine trees, Plant Mol. Biol. Rep. 11, 113-116.

Y. Chen, E.R. Dougherty, and M.L. Bittner (1997) Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images, J. Biomed Optics 2, 364--374.

R. G. Alscher , J. L. Donahue, and C. L. Cramer (1997) Reactive oxygen species and antioxidants: Relationships in green cells, Physiol. Plantarum 100, 224--233.

Acknowledgments

We express our gratitude to Ina Hoeschele for her help with the overall statistical design and to David Clapham for an initial version of the functional categories.