Expresso - A Problem Solving Environment for Bioinformatics: Finding Answers With Microarray Technology

Ruth G. Alscher*, Boris I. Chevone*,
Lenwood S. Heath†, and Naren Ramakrishnan†
Department of Plant Pathology, Physiology, and Weed Science*
Department of Computer Science†
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061
Contact email: heath@cs.vt.edu

Keywords: Bioinformatics, problem solving environments (PSEs), microarray technology, gene expression, data mining.

Abstract

This research addresses the need for high performance computational tools for capturing, storing, analyzing, and visualizing large quantities of gene expression data, as produced by microarray technology. We describe the design of the Expresso PSE, which aids in visualizing analytical results along multiple dimensions, searching for patterns in the data, and modeling functional relationships among genes. The usefulness of the PSE is demonstrated by its development and use in parallel with a set of microarray experiments addressing hypotheses about stress response and wood formation in loblolly pine, with results of potential economic benefit to forestry.

1. INTRODUCTION

Microarrays (sometimes referred to as *gene chips* or *DNA chips*) are a relatively new tool in biotechnology, inspired by miniaturization trends in micro-electronics [12]. Microarray technology is an experimental approach to study all the genes in a given organism simultaneously; it has rapidly emerged as a major tool of investigation in experimental biology. The basic idea is to 'print' DNA templates (*targets*) for all available genes that can be expressed in a given organism, as spots on a high-density 2D array in a very small area on a solid surface[1] (typically a microscope slide or a membrane). The goal is then to determine the genes that are expressed when cells are exposed to experimental conditions, such as drought stress, ozone pollution, extreme heat, or extreme cold. To accomplish this, RNA molecules (*probes*) are extracted from the exposed cells and 'transcribed' to form complementary DNA (cDNA) molecules. These molecules are then allowed to bind (*hybridize*) with the targets on the microarray; they adhere only with the locations on the array corresponding to their DNA templates. Typically such cDNA molecules are tagged with fluorescent dyes, so the expression pattern can be extracted as an image by a laser scanner tuned to the frequency of fluorescence. Intensity differences in spots will then correspond to differences in expression levels for particular genes, between a *control* and a *treated* sample. Using this approach, one can 'measure transcripts from thousands of genes in a single afternoon' [12]. Microarrays thus constitute an approach of great economic and scientific importance, one whose methodologies are continually evolving to achieve higher value and to fit new uses.

The successful application of microarray technology requires many steps: (1) extracting genetic material to use as probes from different tissues subjected to different experimental conditions; (2) labeling probes from different experimental conditions with different fluorescent dyes (typically two) that fluoresce at different frequencies; (3) selection of hundreds or thousands of DNA sequences as sources of targets; (4) biochemical amplification or synthesis of these targets to obtain sufficient material for many experiments; (5) printing thousands of these targets on glass slides via a specialized robot (the *arrayer*) to obtain a tiny grid of targets (the *microarray*); (6) hybridizing the probes to the targets on the microarray; (7) washing away any probes that do not hybridize to a target on the microarray; (8) scanning the microarray with a specialized scanner that can detect the different fluorescent frequencies of the two dyes; and (9) analyzing the resulting scanned images for differences in gene expression levels. Applications utilizing microarrays are characterized by the need to couple experimental design with data-driven analyses, to integrate wet-lab pipetting, robotic manipulation, and image readers at the software level, and to track and identify data and physical entities as they proceed through the laboratory and computational pipeline.

[1] 'Array' here refers merely to a rectangular arrangement; the 'axes' do not have any biological interpretation.

2. APPLICATIONS IN BIOLOGY

Accomplishments in biology and medicine brought about through the use of microarrays are already considerable [2, 9, 16]. Microarrays have been used to identify transcript profiles characteristic of particular tumor types [8, 10, 11, 17] cell cycle regulatory mechanisms [6, 7]; and oxidative stress responses [13, 15]). Life scientists are not surprised by the complexity of gene expression patterns that are being revealed with this new technology, the number of genes involved, or the involvement of expected and unexpected aspects of cellular function in regulation and metabolism. As a discipline, however, there does not exist established methodologies to process this volume of data. The amount of data generated is such that a major change in the treatment and analysis of data is clearly in order. The development of new computational tools for data analysis is thus a major focus area of research.

3. CHALLENGES AND OPPORTUNITIES INVOLVING COMPUTATION

Successful application of microarray technology requires the experimenter to address challenges that affect the quality and repeatability of microarray experiments. One class of challenges stems from the need to identify and control sources of experimental variability. Microarray slides are subject to various anomalies that must be addressed. A second class of challenges relates to successfully analyzing the data — extracting the maximum information from the data while avoiding unsupportable conclusions. Deciding on whether differential expression has occurred is not a binary decision but might result in outcomes such as "meaningless" or "increased expression with probability $> 57\%$". There is also the challenge of designing an experiment, or series of experiments, to obtain the maximum amount of information toward the end of the computational pipeline. Additional replication of data contributes to greater statistical confidence but limits the number of distinct cDNAs that can appear on a given slide. A final frontier is being able to integrate experiment layout with sophisticated forms of data mining, to intelligently 'close-the-loop' in microarray design and analysis. To achieve this goal, high fidelity models from robotics, physical chemistry, and molecular biology have to be 'pushed' deeper into the computational pipeline.

Our research group has recognized the need to address the microarray experiment life cycle as a whole and to fashion a computational system that integrates the needed design, analysis, and data management tasks as well as the laboratory and computational components. The Expresso PSE is designed to support all microarray activities including experiment generation, data acquisition, image processing, statistical analysis and data mining. Currently, the latter three stages are completely automated and integrated within our implementation.

Expresso will be able to take high-level specifications from the biologist (e.g., 'Design a new experiment using the layouts from 1999, the dye concentrations used by Mark, and the conditions of mild drought stress,' or 'Use the same experimental setup as EXPT-99-Pine-Drought, but with a signal threshold of 0.6') to compose programmatic descriptions of new experiment sequences. After the experiment completes, results from image and data analysis will be used to model interactions among genes, refine hypothesized gene regulatory networks as well as suggest candidates for further experimentation and/or update earlier decisions, thus building a complex analysis and design support system.

4. DESIGN OF EXPRESSO

Our starting point for microarray design and analysis is provided by a microarray experiment language (MEL), a semi-structured data format that is used to provide specifications of experiment sequences. For example, Fig. 1 provides a definition of a basic microarray experiment in MEL. Notice that these descriptions only serve to model the experimental process; data and images generated at every stages are committed to traditional database storage.

A particular advantage of this approach is that experiment and scenario evaluation can be viewed as a sophisticated form of database querying. For example, the SQL query

```
SELECT Expression(*)
FROM Array A, Stress Level B
WHERE Hybridization(A,B) >= Acceptable;
```

evaluates the expression patterns for experiments where hybridization levels were acceptable. The key technical issue here is to use query optimization algorithms to selectively 're-order' costly selection operations and/or detect when experiments have to be conducted. For example, in the above query, if entries in Array A have not been subjected to stress levels represented in B, then the query could act as an interface to Expresso, invoking the pipeline on-the-fly, to gather the requested data. This feature also supports the provision of iteration, aggregation, and composition operators by introducing minimal overhead in implementation. For example, compositions can be achieved by relational joins, aggregation by user-defined VIEWs, and iteration by 'index striding' on domain-specific records (such as stress levels).

Change detection and processing in semi-structured data is crucial in Expresso, where objects of interest change formats, stations, and schema rapidly. A variety of approaches are available for detecting meaningful changes [4]. Many of them not only introduce operators (such as subtree copy and subtree modification) to describe the changes but also automatically compute a 'minimum-cost' edit script that can be used to wrap around the formats and maintain consistency in the database. For example, if a given clone's classification is changed from the 'thiol' category to the 'cell wall' category, the edit script

```
EXPERIMENT PINE_DROUGHT_GROWTH May-August,2000 "384 clones"
...
DYE CY5 "Genisphere Kit"
...
PRINTING_ROBOT NCSU_FBC "Brown-type robot at NCSU"
...
PRINTING_CONFIGURATION Stanford4x16x24  4 16 24 QUADRANTS
...
TISSUE D4M      D4      Needles Unstressed (Control)
```

Figure 1: Example description of a microarray experiment in MEL which includes the name of the experiment, the dyes used, description of the printing robot, printing formats, and the tissue configurations. The lines are in a self-describing format where the start field identifies the nature of information modeled.

can be used to update descriptions of all experiments that will be affected by this change. Such edit scripts, in turn, can be visualized using browsing tools such as eXeMeL and programmatically invoked in the pipeline.

One of the important analysis steps supported by Expresso involves the modeling of clusters of genes in the language of first-order predicate logic. Statements in this language refer to encodings of experimental conditions, gene clusters, observations and other properties useful for forming high level representations. For example, '*activation(expt1,gene-cluster1,0.5)*' asserts that the genes in *gene-cluster1* are moderately activated under the conditions of experiment *expt1*. In turn, *gene-cluster1* will have definitions obtained from the clustering process. Inductive logic programming (ILP) [14] techniques are then used to learn rules that specify interactions between the various predicates and their parts. This also enables problem-specific background knowledge (such as the fact that the activation of a certain group of genes are known to be inversely correlated with expression data of a different group) to be incorporated into the mining process. In addition, domain specialists can also understand the induced concept descriptions, as opposed to generalized function approximators such as neural networks. The example rule: *activation(E1,G,-1) :- activation(E2,G,-0.5), stresslevel(E1,S1), stresslevel(E2,S2),* $S2 > S1+4$. models the implication that genes in a cluster (*G*) go from 'moderately repressed' (*E2*) to 'heavily repressed' (*E1*) by increasing stress levels by more than four steps of magnitude. Since the rules induced are horn clauses (that contain at most one predicate in their consequent side), they allow us to perform 'What if' analyses of various design scenarios, such as correlating effect of varying levels of stress conditions.

ILP is often prohibitively expensive; the standard practice is to restrict the hypothesis space to a proper subset of first-order predicate logic. We make use of several techniques to control the complexity. First, we incorporate domain specific restrictions (syntactic as well as semantic) for the induction of rules. An example of a syntactic constraint would be that rules include levels of drought stress in the antecedents and that they do not generalize across the space of stresses. An example of a semantic constraint would be consistency checks between groups of genes and known functions. Second, we incorporate a generality ordering to guide the induction of rules, to prune the search space for generating plausible hypotheses, and to aid in abduction (the process of constructing a rule that needs to be justified further). Finally, the software architecture of our PSE is augmented with a natural database query interface, so that we can utilize this aspect to provide meta-patterns for rule generation. A schematic overview of the functioning of the PSE is provided in Fig. 2.

5. INITIAL RESULTS

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

Current Work

Our work on effects of drought stress on loblolly pine is ongoing (Figs. 3 and 4). We conducted preliminary experiments to determine the extent of intraclonal variation in pine and have used those data to estimate the net effect of mild drought stress on gene expression patterns (Fig. 4). We are developing new analytical algorithms, incorporated in flexible software tools, to uncover the number, strength, role, and mechanisms of molecular interactions manifested in DNA microarray data. The application of these new algorithms will serve to initially identify groups of coordinately regulated genes as well as overlapping networks of coordinately regulated genes. As successive iterative strategies for the discovery of genetic interactions are employed, more complex tools will capture the nature of these interactions at our current level of understanding and allow us to alter and refine that understanding further. At this time, we are building some of the data filters and data analysis modules that will go into the pipeline. These are the mathematical and statistical tools that we feel will give us the greatest immediate benefit.

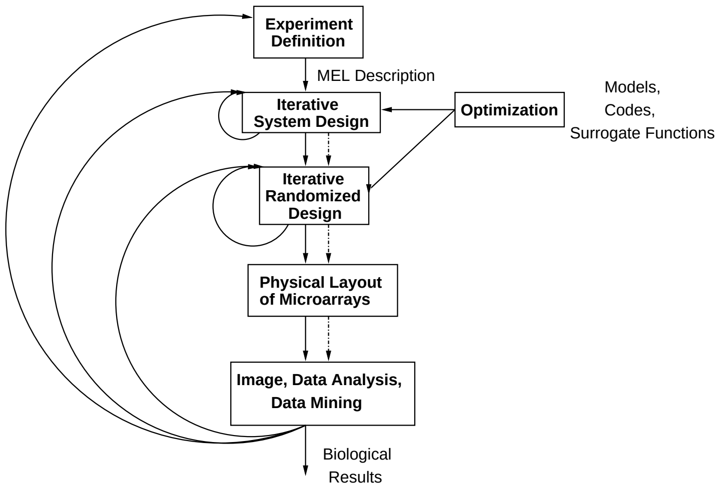

Figure 2: Execution and flows in Expresso. The solid lines indicate computational flows; the dashed lines indicate realizations of the pipeline in devices such as robots and image readers.

Materials and Methods

Plant material and physiological treatments: Two-year old seedlings of a loblolly pine clone 'D' were obtained from North Carolina State University (NCSU) 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).

Arrays: A printing robot was built at NCSU, from the instructions given in [1]. Four pins in the Stanford configuration were used to spot 384 amplified cDNAs from differentiating xylem, shoot tips or pollen cones chosen on the basis of function and/or previously documented response to drought stress. 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 3). After printing, slides were post-processed according to the manufacturers' instructions, and dried. Slides were stored desiccated at room temperature.

Probe Synthesis: Total RNA was isolated from stems and needles by the method of Chang et al. [3] 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. It employs techniques described in Chen et al. [5] 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 Expresso. Perl scripts were employed extensively in Expresso: 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 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. Expresso employed Matlab and Excel for data visualization,

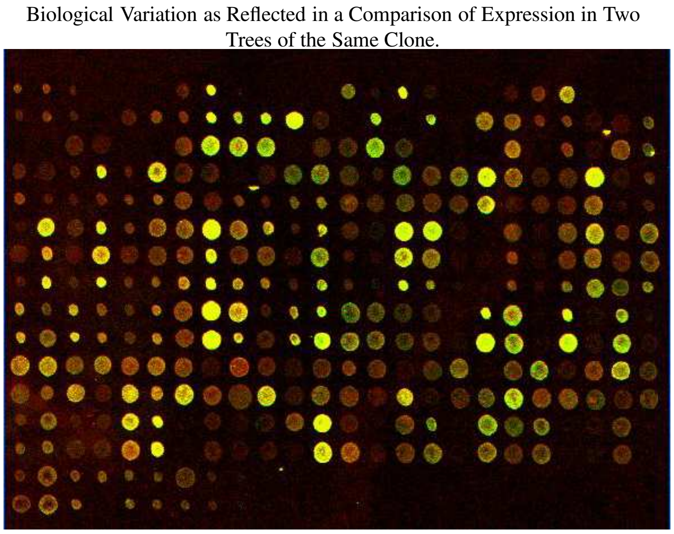

Figure 3: A test of variability in gene expression with a loblolly pine clone. 1,408 cDNAs (ESTs) from differentiating xylem of loblolly pine were printed by the gridding robot and hybridized to Cy3 and Cy5 labeled cDNAs pooled from RNA from actively growing stem tissue of 2 trees from the same clone. Signals were imaged in the GSI laser scanner. 128 GFP cDNAs were included in the printing. Labeled GFP was also included in the hybridization mixture of each labeled sample pair. The array was printed in four 24×16 sub-quadrants, of which one is shown here (courtesy Y.-H. Sun).

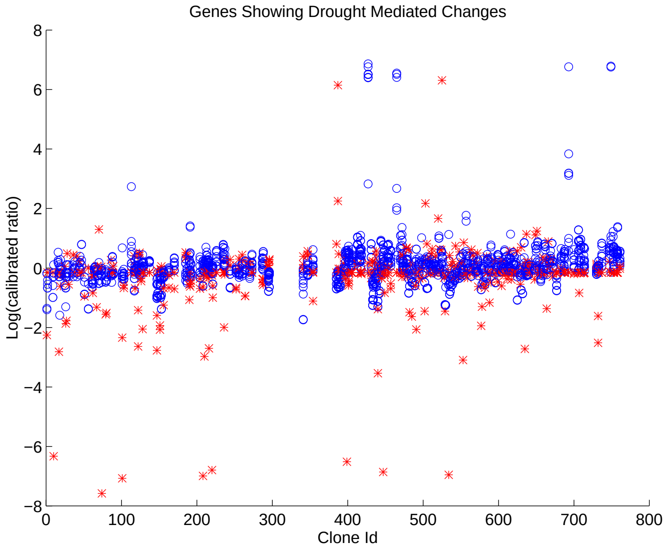

Figure 4: 265 genes out of 768 responded to three cycles of mild drought stress. Data presented are from four individual tests of intraclonal variation (blue spots) and one drought stress experiment (red spots).

and Aleph for the ILP algorithm that searched for associations between a category and the expression levels of clones in that category.

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 zero-centered 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, 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, Expresso represents the expression levels as facts in logic programming notation suitable as input to Aleph. language. These facts were combined with our functional classification of the clones to establish a basis for inductive logic programming. An example rule mined by ILP is the inter-comparison pattern: *level(A,MvsD, negative) :- level(A,MvsI,positive).* This rule states: "If the moist versus intermediate comparison was positive for a clone (represented as *A*), the moist versus drought comparison was negative." The confidence of the rule was 95.8%.

6. FUTURE WORK

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 intend to combine expression data from numerous experiments using ILP, visualization, and other data mining techniques. Our proposed activities focuses on automating many aspects of the tools considered here to build a high performance experiment management system for microarray bioinformatics. The contributions from our research will not only support microarray bioinformatics but also serve as a unifying framework to study data-driven, exploration-based, and experiment-centric environments in other application domains.

Acknowledgements: The authors acknowledge helpful discussions with D. Chen, R.R. Sederoff, R. Whetton, Y.-H. Sun, I. Hoeschele, S.-H. Li, L. van Zyl, K. Ye, C. Struble, and V. Jouenne.

References

[1] P. Brown. A Complete Guide to Microarraying. URL: *http://cmgm.stanford.edu/pbrown/mguide/index.html*, 2000.

[2] C.B. Epstein et al. Microarray Technology - Enhanced Versatility, Persistent challenge. *Curr Opin Biotechnol*, Vol. 11(1):pp. 36–41, Feb 2000.

[3] S. Chang, J. Puryear, and J. Cairney. A Simple and Efficient Method for Isolating RNA from Pine Trees. *Plant Molec. Biol. Reporter*, Vol. 11:pp. 113–116, 1993.

[4] S. Chawathe and H. Garcia-Molina. Meaningful Change Detection in Structured Data. In *Proceedings of the ACM-SIGMOD Conference on Management of Data, Tucson, Arizona, USA*, pages 26–37, 1997.

[5] Y. Chen, E.R. Dougherty, and M.L. Bittner. Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images. *J. Biomed Optics*, Vol. 2:pp. 364–374, 1997.

[6] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, and R.W. Davis. A Genome-Wide Transcriptional Analysis of the Mitotic Cycle. *Mol Cell*, Vol. 2:pp. 65–73, 1998.

[7] S Chu, J DeRisi, M Eisen, J Mulholland, D Botstein, PO Brown, and I Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282:699–705, Oct 1998.

[8] C.M. Perou et al. Distinctive Gene Expression Patterns in Human Mammary Epithelial Cells and Breast Cancers. *Proc Natl Acad Sci USA*, Vol. 96:pp. 9212–9217, Aug 1999.

[9] J.C. Cushman and H.J. Bohnert. Genomic Approaches to Plant Stress Tolerance. *Curr Opin Plant Biol*, Vol. 3:pp. 117–124, Apr 2000.

[10] J. Khan et al. cDNA Microarrays detect Activation of a Myogenic Transcription Program by the PAX3-FKHR Fusion Oncogene. *Proc Natl Acad Sci U S A*, Vol. 96(23):pp. 13264–13269, Nov 9 1999.

[11] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, D.C. Bloomfield, and E.S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, Vol. 286:pp. 531–537, 1999.

[12] H. Hamadeh and C.A. Afshari. Gene Chips and Functional Genomics. *American Scientist*, Vol. 88:pp. 508–515, November-December 2000.

[13] S. A. Jelinsky and L. D. Samson. Global Response of Saccharomyces Cerevisiae to an Alkylating Agent. *Proc Natl Acad Sci USA*, Vol. 96(4):pp. 1486–1491, Feb 16 1999.

[14] S. Muggleton. Scientific Knowledge Discovery Using Inductive Logic Programming. *Communications of the ACM*, Vol. 42(11):pp. 42–46, 1999.

[15] S.A. Amundson et al. Fluorescent cDNA Microarray Hybridization reveals Complexity and Heterogeneity of Cellular Genotoxic Stress Responses. *Oncogene*, Vol. 18(24):pp. 3666–72, Jun 17 1999.

[16] C. Somerville and S. Somerville. Plant Functional Genomics. *Science*, Vol. 285:pp. 380–383, Jul 1999.

[17] G.P. Yang et al. Combining SSH and cDNA Microarrays for Rapid Identification of Differentially Expressed Genes. *Nucleic Acids Res*, Vol. 27(6):pp. 1517–1523, Mar 15 1999.