Paper DP04

Statistical Analysis of Gene Expression Micro Arrays

John Schwarz, University of Louisville, Louisville, KY

ABSTRACT

Advancements in genetic research have led to increased amounts of data often without efficient analysis techniques. One such area of genetics that has developed a great deal in the past several years has been micro arrays. One area of micro array experimentation is that of gene expression. Gene expression uses arrays of thousands of genes with one to two targeted strands of DNA that are fluorescently tagged and used to identify which genes are expressed. This experiment can identify which conditions cause certain genes to be activated in different cells. It can also be used to track certain cellular changes through the expression of genes.

The data gathered from the micro array experimentation is contained in a large matrix containing the data results from gene expression analysis on a micro array plate. Some of the standard analysis techniques and programs can be inefficient and inconclusive. The purpose of this analysis is to arrive at conclusive results with efficient methods. The primary statistical software used in the analysis was SAS/Stat system version 8.2. A series of statistical tests was applied to the data set to determine the meaningfulness of the results and the efficiency of the tests.

INTRODUCTION

Every living organism is made up of a single cell or many groups of cells. The identity of these different cells and the function of these cells are determined by genes. Genes are segments of DNA providing the code for producing proteins. Different organisms contain different numbers of genes. For example, a human being contains 30,000 genes as estimated and the fruit fly contains only about 13,000 genes. The identification of genes depends upon the knowledge of the DNA sequence made up of different alleles (different forms of a gene). The human genome has recently been fully mapped. However, not every gene has been fully expressed in the DNA alleles. Certain organisms have been fully mapped; this refers mainly to smaller organisms with a fewer number of different genes. The advantages of having the knowledge of the full sequenced genes will be discussed later with micro arrays. Understanding genes is important; unfortunately, genes can often be hidden within compacted DNA strands and are not easily identifiable.

Genes also interact in different ways; some have similar properties and responsibilities while other genes may be totally different. Genes that are different, however, may be involved in the same reactions inside a cell and, equivalently, genes that have similar properties may not be involved in the same reactions. Micro arrays offer an efficient method of comparing multiple genes quickly and easily. Micro arrays, however, require several analyses upon completion of an experiment. The analysis of micro arrays offers substantial evidence of genes that may or may not be related in a cell.

BACKGROUND

Gene expression is important in cellular identification and gene function. With new technologies and research, gene expression and identification have become an ever growing area in biotechnologies with the opportunity for new, more efficient analyses available. The field of cellular genetics has shown that changing pH and temperature causes certain genes to be expressed and not expressed. It is possible to alter these settings in a lab and the expressed genes can be identified. Most genes are known by the proteins they produce and the function of these proteins. It is possible to analyze large groups of proteins as well as genes. This process will be discussed later. Analyzing different genes expressed can determine when certain reactions take place in the body, or determine what processes some of the genes are responsible for. Understanding which genes are expressed is important to different gene interactions and cellular identities. According to Campbell,

> ìIn all organisms, the expression of specific genesis are most commonly regulated at the level of transcription by DNA-binding proteins that also interact with other proteins and often with external signals. For that reason, the term *gene expression* is often equated with gene activity- that is, transcription- for both prokaryotes (cells lacking membrane enclosed nucleus and membrane enclosed organelles) and eukaryotes (cell with membrane enclosed nucleus and membrane enclosed organelles). However, the greater complexity of eukaryotic cell structure and function provides opportunities for controlling gene expression at additional stagesî (Campbell 2002, p. 362).

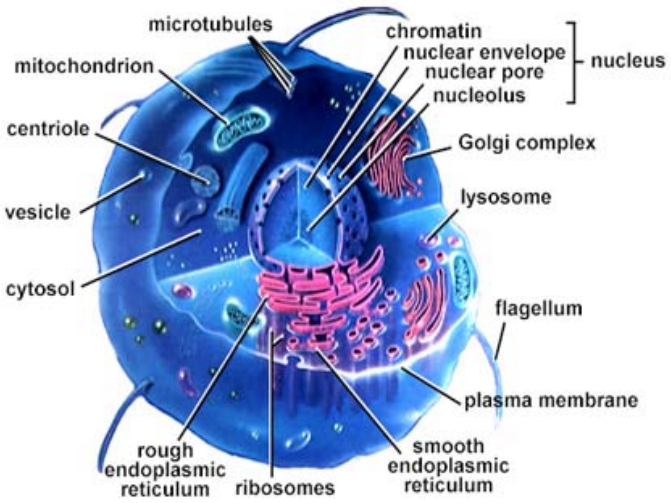


Figure 1- Cellular Structure (Brazma 2003, p. 1)

Gene expression can be broken down into several different stages and identified at any time during one of the specific stages. These range from gene to functional protein activity. The first stage in expression is the unpacking of chromatin (complex of DNA and proteins that makes up a chromosome) DNA (Figure 1). In chromatin form, DNA is closely packed and regulatory proteins used in the transcription process often are unable to access certain portions of the DNA, and consequently certain genes are not expressed. The differences in chromatin packing vary in each different cell type and thus, some cells inhibit or allow the expression of certain genes. The chromatin packing is known as regulatory function for gene expression.

Even though entire DNA strands exist in chromatin form at times, they can be unpacked to allow access to certain strands for protein production and replication.

Methylation occurs, meaning the process when the DNA is unpacked and methyl groups (CH_3) groups are placed on the ends of the DNA strand for the gene being used, and identifies the start and finish of the gene to replication compounds used for protein construction. Transcription occurs and the DNA is copied into RNA to change certain alleles and is then copied to mRNA for protein production outside the nucleolus. The mRNA is moved to the cytoplasm where protein translation is accomplished. Polypeptide (polymer chain in which amino acids are linked together with the peptide bonds) groups are created from the mRNA code and, after cleaving and modification, are known as proteins. From this stage the proteins are sent to certain areas of the cell for purposes identified by the type of protein. The mRNA strand breaks down after replicating. Usually several proteins in the cytoplasm and the allele groups are absorbed and recycled for later protein production. This whole process is known as the gene expression, and again any stage of this process can be used to identify the gene expression (Campbell 2002, p. 364).

Cellular identification is determined by the genes expressed and the proteins produced. Different cells have different roles in the body and in turn produce a number of different proteins. By learning which conditions activate certain genes and deactivate other genes, more can be understood about certain cellular identities. Furthermore, much can be understood about cellular life and death and whether death occurs as a consequence of time or as a malfunction (Campbell 2002). Cells generally have the ability to regenerate and reproduce themselves. These processes are marked with the expression of certain genes. The understanding of when and how these processes take place, and more importantly, why some cells do not go through this process can be better understood through gene expression analysis.

The main objective in micro array analysis is to identify gene expression and the comparison of several genes at once. The process of identifying and comparing the genes expressed in a cell or culture is a complex process resulting in large amounts of data. The process can be simplified into six main steps (See Figure 2):

- Selecting the cell culture for analysis
- Identifying the specific DNA gene sequence
- Radioactively tagging the DNA sequences
- Hybridization of the array (Hybridization is the process in which the fluorescently tagged cDNA is applied to the array)
- Laser intensity readings from the plate
- Interpreting the results of the hybridized array

Of course, only fully sequenced genes can be used in this experimentation since known DNA sequences are hybridized to an array of thousands of different genes at one time. The idea behind creating this array is to identify genes at certain points in an expression and isolate conditions for the certain genes to be expressed.

There are several different types of micro array experimentations. The first type is gene expressions using DNA. This method uses the DNA sequences of specific genes that are applied to an arrayed, cultured plate with the goal of identifying genes in the cells and comparing their interactions. Initially, the mRNA is used in this experiment. The

mRNA, messenger RNA (synthesized DNA), is used in the production of proteins. The mRNA is copied directly from DNA in the cell. When the mRNA is copied, different gene alleles are used that differ from normal DNA. The mRNA sequences are known to be unstable and to deteriorate after a short amount of time, making them useless for hybridization reactions. For this reason cDNA (DNA copied from mRNA with enzyme reverse transcriptase) is copied from the mRNA for its ease of use and compatibility (Fortina 2000).

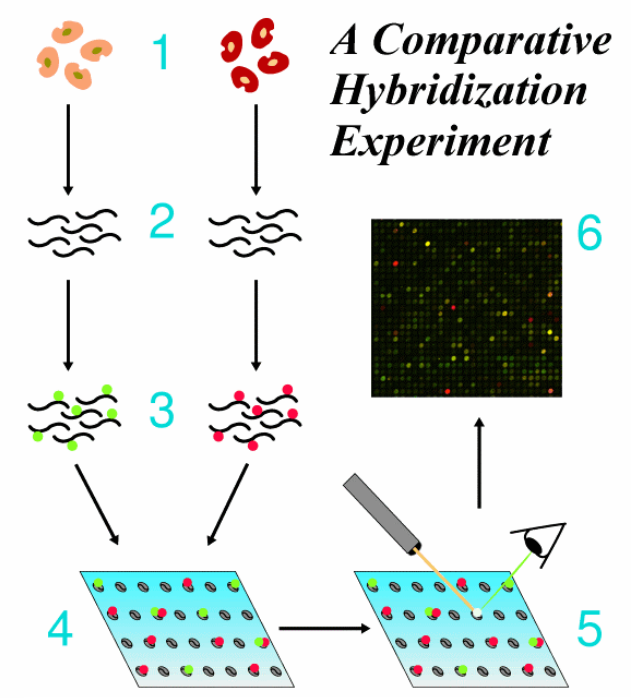


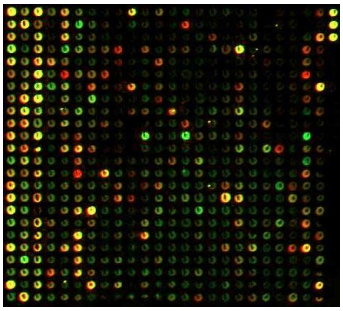
Figure 2- Microarray process (Buhler 2003, p. 1)

The cDNA, complementary DNA, is copied from the mRNA and uses the original alleles as DNA. The cDNA is then tagged with fluorescent markers and applied to an array of cellular cultures. Another type of micro array experimentation uses DNA sequences. A DNA sequence, again fluorescently tagged, can be used so that the same sequence can be identified in different cells across an array. This identification can be used to understand evolutionary traits across species. The third process uses proteins. Proteins are tagged fluorescently and applied to a slide consisting of an environment with protein-carrying structures. This method can be used to determine the relationships of many proteins at once and to show how they interact with one another in specific areas of the body. Protein analysis can help to gain an understanding for which proteins have certain relationships or even tracing it back to certain genes that produce the proteins.

The process of the creation of an array plate generates thousands of closely contained DNA segments. The DNA segments are placed within a small region on a glass slide and glued to prevent washing away during the gene reaction process (Campbell 2002). The DNA segments are placed in a grid pattern for identification once the hybridization reaction has taken place. The use of certain genes depends upon what is known about the organism. Some organisms have fully sequenced genes so that all the genes can be used in the expression. Organisms that have a great number of genes may require the use of several array plates since not all the genes can fit on a single array.

Continuing with the other steps involved in micro array, the analysis will be described using the techniques of gene expression micro arrays. Hybridization is the next step to micro array.

> ìIn addition, realization of such developments will facilitate early diagnosis and evaluation of treatment strategies. The assay is template-dependent and involves primer extension by a radioactive- or dye-labeled dideoxynucleotide terminator (ddNTP) with the tag of the incorporated base revealing the identity of the template complementary nucleotide imme-diately 3' to the primer. Solution-phase extensions followed by electrophoresis on gels as well as microtiter plate-based assays have been described (Fortina 2000, p. 884).î



Each spot on the array contains more than one sample strand of DNA. The ability for more than one interaction at each spot is possible. Thus the intensity of the fluorescent marking can differ from spot to spot and cannot be used to determine how accurately the strength of the binding is to the sequence. The different sequences are marked with different colored fluorescent tags. Using different colors allows for multiple detection of gene expression (Figure 3). The different colors can also determine similar genes and gene interactions.

Figure 3- Microarray plate(Leming 2003, p. 1)

Following the hybridization process, a washing stage is carried out. The purpose of the washing is to eliminate any fluorescently labeled sequences that did not hybridize to gene spots. That is why it is important to secure the spots before the hybridization process. This security leads to a clearer interpretation of which gene spots hybridize with the applied marked gene sequences.

The fluorescent tags can be identified by computer analysis. Since hybridization can occur on only a single sequence of DNA in each spot, visual detection cannot be accurate. Furthermore, the fluorescent tags used in the hybridization process cannot be visually identified unaided. The tags used require certain frequencies provided by laser treatment after hybridization and washing has occurred. Consequently, there are thousands and thousands of data results from each micro array hybridization experiment. Simple visual analysis in most cases is time consuming and inefficient. Analysis by computer is the most efficient and accurate option. For this reason, different statistical methods must be employed to understand the relation between different genes. Several different programs exist currently designed to analyze the large sets of data produced by the micro array process. Some of these include the simple frequency counts from each of the different experiments and genes. Other methods include simple statistical tests to understand small portions of the large array.

METHODS

The first step in the analysis process was to examine the properties of each gene that reacted with the micro array plate. The method used for modeling the data was kernel density estimation. Two-dimensional graphical representations of kernel density estimators resemble a smoothed histogram. The kernel density estimator approximates a probability density function allowing for specific values to be accentuated. The kernel density estimator uses all the values in the data. This differs from the histogram in that data are separated into certain sections and allow for the maximum value to be emphasized, not just the section in which this maximum occurs. When dealing with such large data sets for visual analysis, certain portions of the data are shortened by bound adjustments. This allows for the most prominent portions of the data to be observed clearly. The equation used for kernel density estimation is as follows:

$$\hat{f}(x) = \frac{1}{n a_n} \sum_{j=1}^{n} k\left(\frac{x - X_j}{a_n}\right)$$

The value n is the sample size for the dataset, a_n is a constant based on the sample size. Inside the summation, k is a known density function, and x is any value within the domain of the density function f.

The kernel density can be analyzed visually to allow for preliminary relationships to be determined. Since this process is visual, optimal graphical representations must be achieved. Optimal graphical representation is done through bandwidth adjustment. Several different methods for bandwidth adjustment are available through SAS, and each method produces different results. For determining the best bandwidth adjustment, the different methods are tested individually at different percentage levels.

Using SAS and kernel density estimation, the KDE procedure returns output in table form. The table lists 401 rows of the density function, data position, and count. From the table form, the kernel density model can be plotted and then visually examined. The specified SAS code is as follows:

```
Proc kde data=_proj_.John gridl=0 gridu=954 method=SROT
bwm=1.00 out=outkde1;
Var CJ_BaP1;
Run;
```

The first section of the code designates procedure. Next, the lower and upper bounds are declared for the data. Then the method of bandwidth adjustments is specified. The second line begins with the percentage for the bandwidth adjustment and finishes with the location of the output file where the analysis will be stored. The third line declares the variable that is analyzed.

For this data set, the method of bandwidth was Silvermanís Rule of Thumb Method. This method is modeled by the following equation:

$$h = 0.9 \min[\hat{\sigma}, (Q_1 - Q_3)/1.34] n^{-1/5}$$

(SAS Institute 2003, p. 1)

In this equation, n is the sample size and $\hat{\sigma}$ is the sample standard deviation. The Q values represent the first and third quartiles. This method proved visually optimal for the data set at 100%. The 100% bandwidth gives a smooth representation of the data which with multiple graphs allows for a simple distinct comparison. The lower bound was set at zero and the upper bound was determined based on the analysis of each individual variable. Each upper bound had the same density endpoint value.

RESULTS

The resulting kernel density analysis produced a series of tables with three columns and four hundred rows plus one. For graphical analysis, the first two columns, value and density, were used to plot solutions. The graphs from each different variable could then be combined and simple visual interpretations could be made. Each single variable plot was similar in pattern but maximum regions often differed from variable to variable. Figure 4 is an example of a plotted graph using the variable of CJ_BaP1.

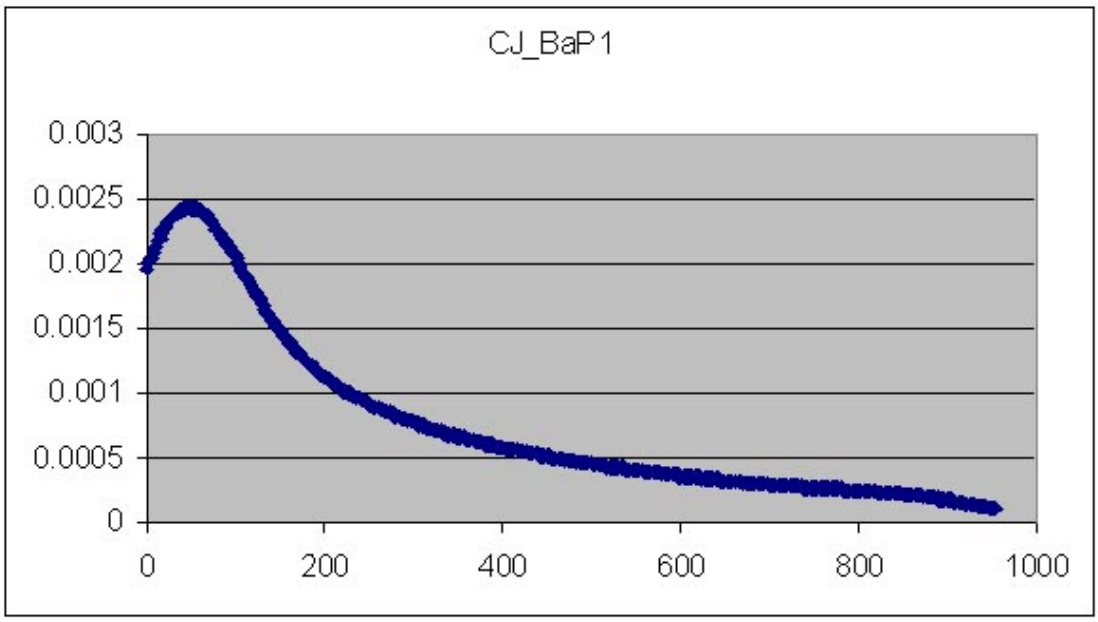


Figure 4-Gene CJ_BaP1 Kernel Density Plot

There were several different methods involving the comparison of the separate graphs. The first was to plot all of the variables on one graph. The resulting plot is illustrated in figure 5. The y axis represents the probability density and the x axis defines the frequency of the light recorded. The boxes indicate the different concentrations of genes.

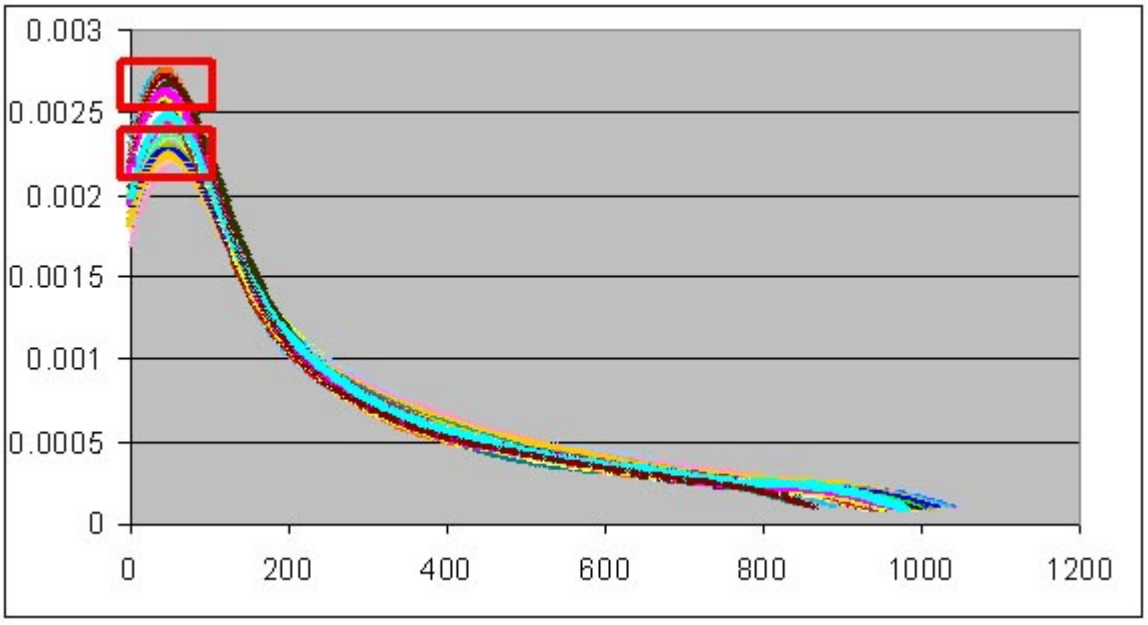


Figure 5-Comparison of all genes using kernel density.

Upon visual analysis, it can be determined that there are two major concentrations of variables that are the different genes run through the experiment, and several outliers. The upper concentration is made up of 26 gene variables. The lower concentration is made up of 14 variables and the remaining 2 variables look to be outliers. From this point, several different comparisons were made. The concentrations represent possible groups of related genes. First each concentration was examined more thoroughly by removing the other variables from the plot. This was repeated several times, and smaller concentrations throughout the group could be examined individually. An example of a smaller concentration is illustrated in figure 6 and the clusters within are identified by the boxes on the

plot.

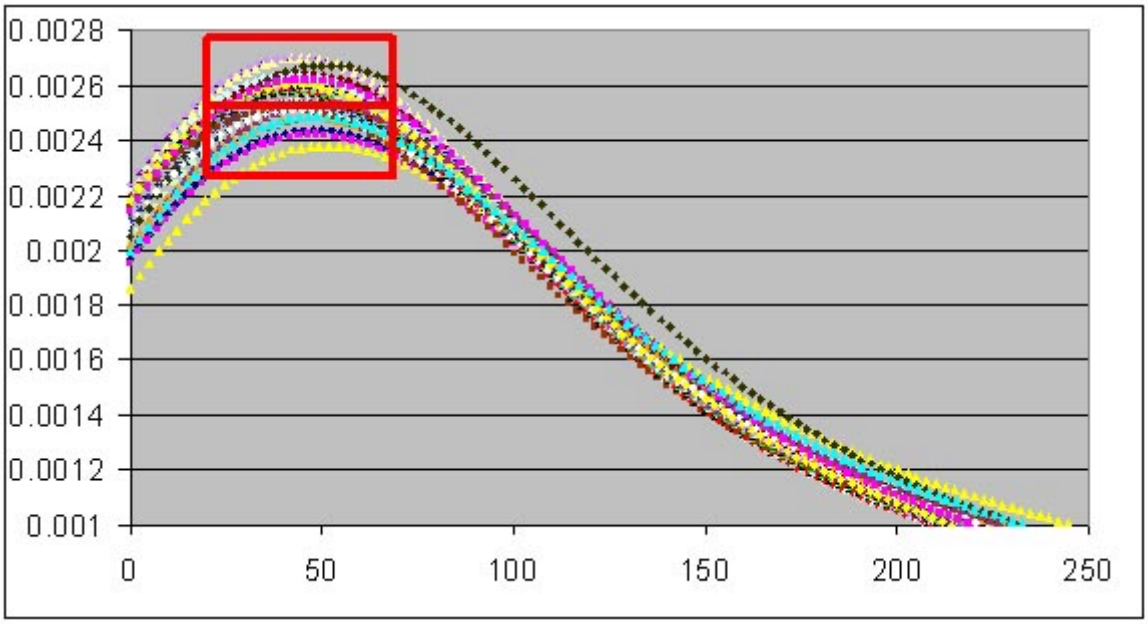


Figure 6-Upper concentration of genes analyzed.

Similarly the analysis of the concentration was repeated for the lower concentration and again smaller clusters could be identified and interpreted. An example from the lower concentration is illustrated here figure 7 and the clusters identified by the boxes..

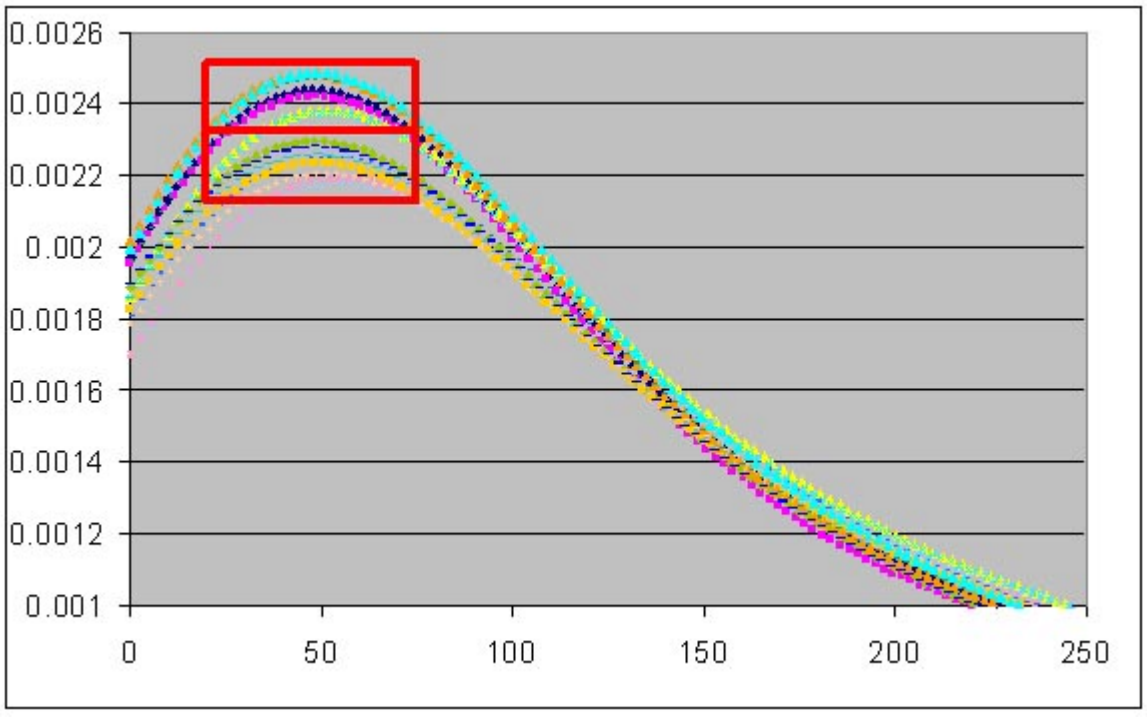


Figure 7-Lower concentration of genes analyzed.

Next, an examination of the difference in the concentrations and outliers was undertaken. Several variables were selected from each cluster nearly at random, and plotted to observed differences in position and behavior. Plotting variables against one another helped to distinguish outliers and cluster groups. Figure 8 is an example of this analysis.

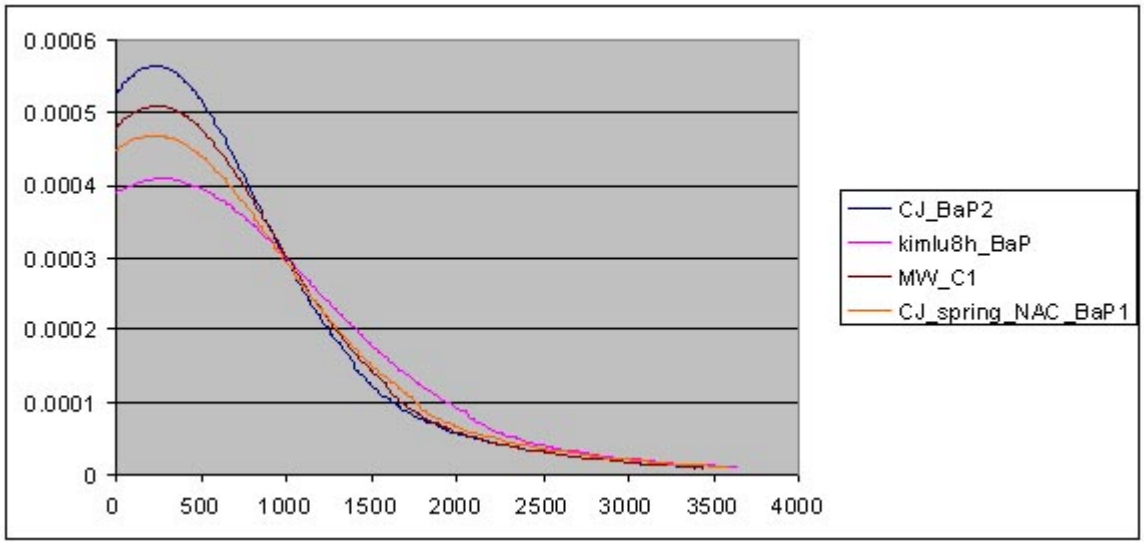


Figure 8- Example of genes that differ in expression.

The difference in concentrations was the initial step to understand gene relationships and which genes could possibly be associated with one another. Using this information, the individual clusters could help identify more exactly which genes had a higher probability of being related to one another. This is demonstrated in figure 9.

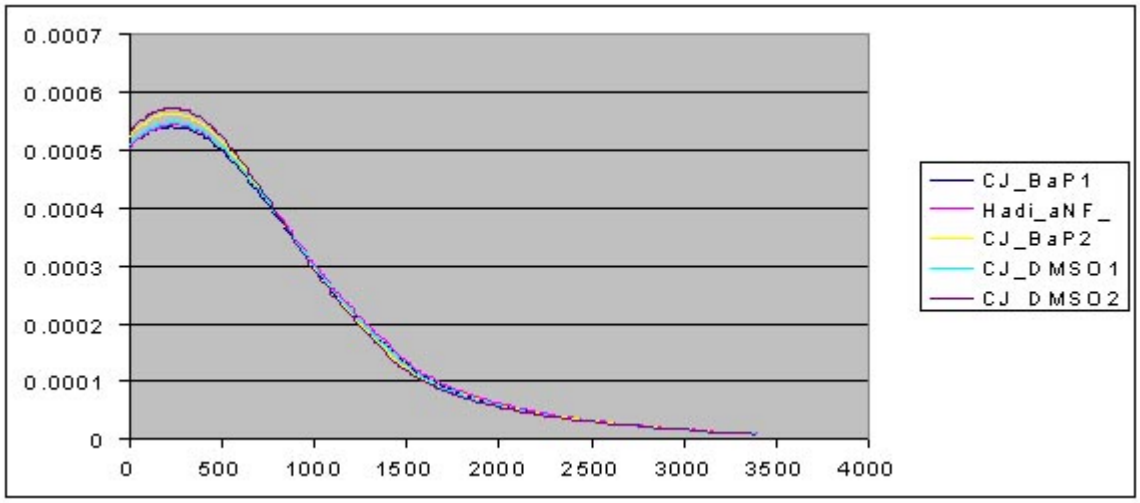


Figure 9-Example of genes that are similar in expression.

CONCLUSION

The visual comparison has allowed for a greater understanding of the relations between the genes. As discussed before in the results section, there were two major concentrations of the gene variables when analyzed by SAS kernel density estimator (Fig 2). The larger concentration contained 26 different genes and when examined again was shown to contain 2 smaller clusters within the concentration (Fig. 3). The larger of the clusters contained 14 separate gene variables and the smaller cluster contained 12 different gene variables. These results suggest that the groups of genes contained in each cluster have relations within the cell and may have similar expression properties. The smaller concentration that contained 14 from the original gene variables also contained two smaller clusters (Fig 4). Each of these smaller clusters contained 7 gene variables each. Again, these results suggest that the groups of genes contained in each cluster have a chance of being related and share similar expressions. The two gene variables that appear to be outliers were not clustered together; thus they appear to have no relation with any other gene run in the experiment.

The kernel density estimation allowed for easier visual interpretation of the larger data set. By using this method, comparing multiple variables at once becomes much easier and possible. This also has allowed for easier presentation of the results. The bandwidth gave the ability to have a clear presentation and easier interpretation.

Further analysis of the data using SAS could determine exact relationships and even confidence intervals within the different clusters and groups of the data. More advanced techniques such as mixed models could also be explored

as options to analyze the data set and identify relations between different genes and their expression within the cell.

REFERENCES

Brazma, Alvis; Parkinson, Helen; Schlitt, Thomas; Shojatalab, Mohammadreza. ìQuick Introduction to Elements of Biologyî http://www.ebi.ac.uk/microarray/biology_intro.html (December 2003).

Buhler, Jeremy. ìAnatomy of a Comparative Gene Expression Studyî http://www.cs.wustl.edu/~jbuhler/research/array/ (December 2003).

Campbell, Neil A.; Reece, Jane B. ìBiology.î Pearson Education, San Francisco, Ca. 2002.

Fortina, Paolo; Delgrosso, Kathleen; Sakazume, Taku; Santacroce, Rosa; Moutereau, Stephane; Su, Hung-Ju; Graves, David; McKenzie, Steven; Surrey, Saul. ìSimple two-color array-based approach for mutation detectionî European Journal of Human Genetics; Nov2000, Vol. 8 Issue 11, p884, 11p.

Leming Shi, PHD. ìDNA Microarray.î http://www.gene-chips.com/ (March 2004).

SAS Institute Inc. 1999. SAS Online Doc , Version 8.2, SAS Institute Inc., Cary, North Carolina, USA 2003.

ACKNOWLEDGMENTS

I acknowledge the help from my advisor Dr. Patricia Cerrito.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

John Schwarz
University of Louisville
1416 Sylvan Way
Louisville, KY 40205
Work Phone: 502-235-5230
Email: jcschw02@louisville.edu