ABSTRACT

ANALYSIS OF GENE EXPRESSION DATA USING EXPRESSIONIST™ 3.1 AND GENESPRING™ 4.2

by

Indu Shrivastava

The purpose of this study was to determine the differences in the gene expression analysis methods of two data mining tools, Expressionist™ 3.1 and GeneSpring™ 4.2 with focus on basic statistical analysis and clustering algorithms. The data for this analysis was derived from the hybridization of *Rattus norvegicus* RNA to the Affymetrix RG34A GeneChip. This analysis was derived from experiments designed to identify changes in gene expression patterns that were induced in vivo by an experimental treatment.

The tools were found to be comparable with respect to the list of statistically significant genes that were up-regulated by more than two fold. Approximately 78% of this gene list was present in both tools. Expressionist™ 3.1 was capable of representing the different linkage methods of hierarchical clustering as average, complete and single, whereas in GeneSpring™ 4.2, the user could manipulate the separation ratio and minimum distance of the hierarchical tree.

ANALYSIS OF GENE EXPRESSION DATA USING EXPRESSIONIST™ 3.1 AND GENESPRING™ 4.2

by

Indu Shrivastava

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology And
Rutgers, The State University of New Jersey-Newark
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computational Biology

Federated Biological Sciences Department

January 2003

APPROVAL PAGE

ANALYSIS OF GENE EXPRESSION DATA USING EXPRESSIONIST™ 3.1 AND GENESPRING™ 4.2

Indu Shrivastava

Dr. Michael Recce, Thesis Advisor — Date
Associate Professor, Information Systems, New Jersey Institute of Technology

Dr. Jeffrey Liebman, Thesis Co-Advisor — Date
Computational Biologist, Novartis Pharmaceuticals

Dr. Peter Tolias, Committee Member — Date
Director, Center for Applied Genomics, Public Health Research Institute

BIOGRAPHICAL SKETCH

Author: Indu Shrivastava

Degree: Master of Science

Date: January 2003

Undergraduate and Graduate Education:

- Master of Science in Computational Biology,
  New Jersey Institute of Technology, Newark, NJ, 2003

- Bachelor of Arts in Biology,
  New Jersey Institute of Technology, Newark, NJ, 2002

- Bachelor of Science in Biology,
  Hawabagh Womens' College, Jabalpur, MP India, 1997

Major: Computational Biology

To My Mother, My Husband and the Rest of My Loving Family

ACKNOWLEDGMENT

I would like to express my heartiest gratitude to Dr. Michael Recce for being my advisor for this thesis and guiding me along every step of this endeavor. I would especially like to thank Dr. Jeffrey Liebman for granting me a Summer Internship at Novartis Pharmaceuticals and also serving as my Thesis Co-advisor, providing continuous support and valuable feedback. I would also like to thank Dr. Peter Tolias for participating as a committee member for this thesis.

My sincere appreciation goes out to Dr. Uwe Junker, Dr. Joseph Rahuel and Dr. John Rediske of Novartis for their special guidance, support and encouragement during my Internship. I would also like to thank the whole team at Novartis Pharmaceuticals, both in Basel, Switzerland and Summit, New Jersey who helped me in this endeavor. My appreciation also goes out to the technical support teams at GeneData and Silicon Genetics for helping me understand the functionality of the tools. A special note of thanks also goes to Dr. Marla Weetall for her support. In the end, I would like to thank Dr. Chak Tan and Dr. Michael Cody at the Center for Applied Genomics, and my friend, Purnima Suri, for their tireless efforts of proofreading my thesis.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF FIGURES

LIST OF FIGURES
(Continued)

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Microarrays and Gene Expression Analysis

Analyzing gene expression patterns to decipher information about biological processes leads to the discovery of innovative ideas regarding the mechanism of living beings. There are different methods to decipher patterns of gene behavior, some of which include application of statistical methods and algorithms to genomic data to discover genes that may be linked to specific diseases.

The gene code embodied in the DNA and RNA of an organism contains all the information required for protein synthesis. The study of messenger RNA (mRNA) expression may lead to the ultimate goal of understanding the expression of a gene. A new and powerful tool for analyzing gene expression, DNA microarray technology is being widely adopted at a rapid pace. The estimated increase in the entire DNA array market, including the actual arrays, as well as instruments and supplies is expected to grow from approximately $322 million in 2000 to about $1.2 billion in 2006, representing a compound annual growth rate of 24% [22].

DNA microarray technology allows analysis of thousands of genes simultaneously. This technology, along with others such as Southern and Northern Blotting is based on the process of hybridization. In Southern and Northern blotting, a small string of DNA, the oligonucleotide, is used to hybridize to complementary fragments of DNA by distributing the oligonucleotide probes over a gel containing samples of RNA or DNA. In microarrays, the oligonucleotides are immobilized on a

surface. This immobilization can be performed at micrometer distances and hence can be placed on a small single surface of one square centimeter. Microarray technology has transformed the concept of "one gene-one experiment". There are two major technologies available for gene expression analysis, namely, Affymetrix, Inc. and Spotted Arrays. Spotted arrays are custom made chips where a robot is used to spot cDNA or oligonucleotides on a glass slide [1]. The Affymetrix gene chip technology is being used for this DNA microarray study.

Microarray hybridization experiments begin with the extraction of mRNA and its conversion to complementary DNA by means of a reverse transcription reaction. The cDNA undergoes amplification and labeling, and then fragmentation and hybridization to 25-mer oligos (oligonucleotides) on the surface of the chip. After the unhybridized material is washed away, the hybridized strands are stained in a microfluidics station with biotin-labeled cRNA with Streptavidin–Phycoerythrin and then washed. The chip is then scanned in a confocal laser scanner; the signal is amplified with goat IgG and biotinylated antibody. The chip is then scanned again and the image analyzed by custom software [1]. The intensity of signal expression measured by laser scanners allows quantitative measurements of gene expression. It is these intensity values that are stored in the form of image files [2]. Expression arrays with different conditions as treatment, tissue and time can be analyzed simultaneously.

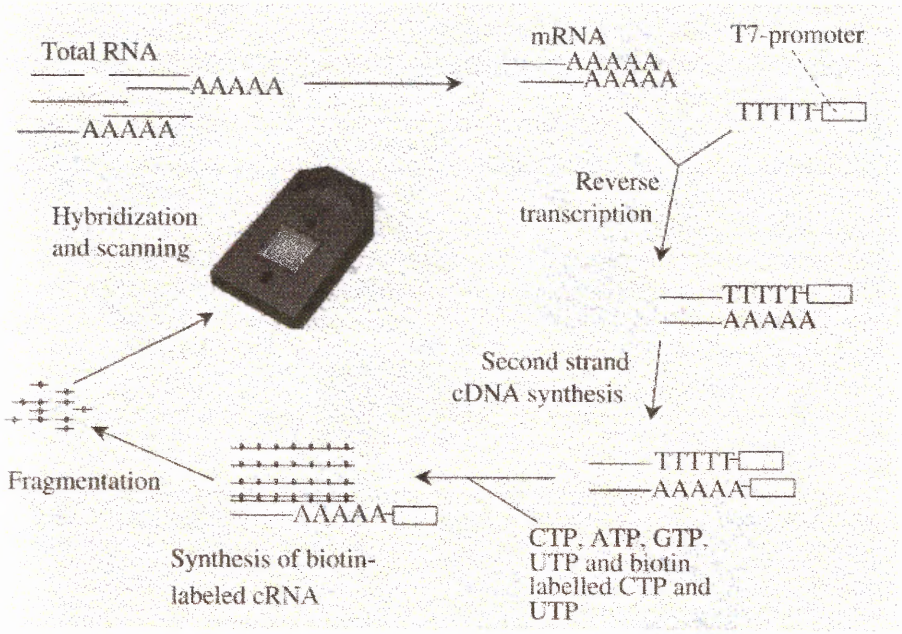

Figure 1.1 Preparation of sample for GeneChip Arrays. Courtesy: Christoffer Brothers [1].

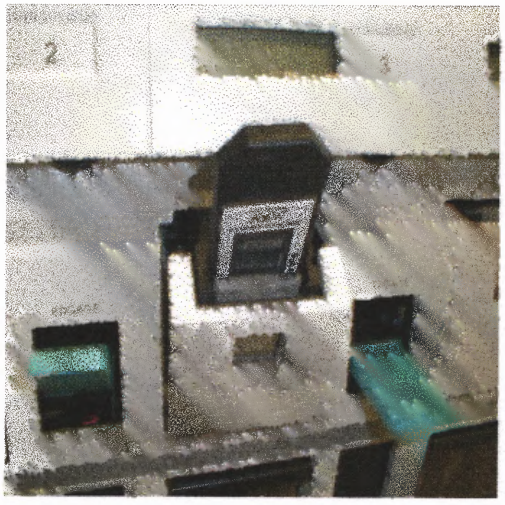

Figure 1.2 Fluidics Station for automation of staining and washing of the array (post hybridization). Courtesy: Center for Applied Genomics.

The basis for measuring changes in mRNA concentrations lies in the following concept that was developed by Affymetrix [6]:

A given region of gene DNA sequence is selected by Affymetrix, which consists of 11-20 oligos and are labeled as a perfect match (PM). As the name suggests, these are perfectly complementary to the mRNA of the sequence of interest. Another set of 11-20 oligos is taken, which are similar to the perfect match, except for the central (13^{th}) position, where one nucleotide has been changed to its complimentary nucleotide, i.e., a homomeric base change takes place. These are termed the mismatch oligos (MM). The concept of this lies in the understanding that MM oligos may be able to detect a non-specific or random cross hybridization to quantify weakly expressed mRNAs. The aim is to detect differences in mRNA concentrations, and not to quantitate the actual RNA concentrations [6].

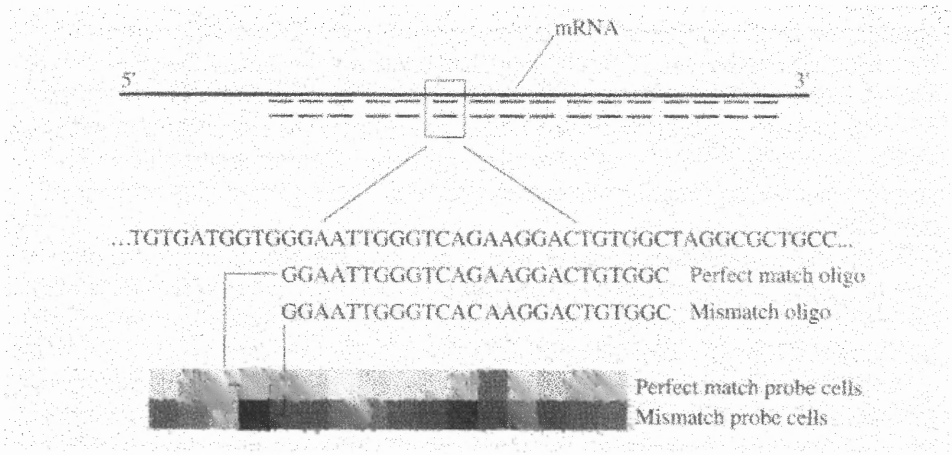

Figure 1.3 The Affymetrix GeneChip Technology.
Courtesy: Christoffer Brothers [1].

1.2 Quantitative Measurement of Microarrays

All of the data obtained from the above processes, both numeric and in raw image form, are stored in databases. The analysis of this data encompasses the whole field of RNA expression analysis. Study of the gene data involves expression profiling by performing expression analysis. Expression profiling implies the expression of every single measured gene over a number of conditions in order to predict its general behavior. This profiling takes place by the analysis of the expression of a gene by measuring the concentration of derived cRNA on the array.

The manipulation of data produced from these microarrays entails the implementation of different tests and analytical tools to determine the importance that a gene has for a particular disease. The integration of all the information retrieved from this data and the application of biological knowledge to decipher the gene's identity is termed as data mining. Thus, the process of data mining involves teasing the important information from large and "noisy" data sets. The systematic approach to understand the behavior of a gene is to study the change in gene expression from one condition to another. One of the methods to determine this change in expression is quantitative; hence the need for statistical analysis and algorithms emerged. These computational techniques help to predict the relation between the structure and function of a gene. The focus of this thesis involves the study of some of these computational techniques encapsulated in an application, and displayed in the form of a user-friendly interface.

The fastest way to view an initial global expression pattern for a given genome is to apply selected algorithms and statistical tests to the analysis of gene expression levels obtained from the signal intensities of the concentration of mRNA produced. The derived

patterns may depict a related biological function and hence aid in the identification of particular genes based on its behavior. This whole process brings together many practically complex components to reach a final conclusion about the gene information that is being sought by a biologist for a complete understanding of the genetic make up of an organism. Some of these include:

- Transfer of data sets from the scanned image file to files capable of being read by a quantitative analytical tool.
- Manipulation of the data sets obtained from the above converted files for an objective and uniform analysis.
- Implementation of many statistical tests to ensure the significance of the presence of certain experiments and genes.
- Implementation of different algorithms to group co-expressed genes together.
- Translation of this analysis into a visualizable, low-dimension format, capable of ready comprehension by the human mind.
- Representation of this visualization in a reproducible form capable of being understood by those not directly involved in the analysis procedure.
- Export of this information in another form for further subjective analysis if the need exists.

Genes obtained using microarray technologies are screened as an expression matrix where each row represents the behavior of a single gene over many experimental conditions and each column represents the attribute or experimental condition [9].

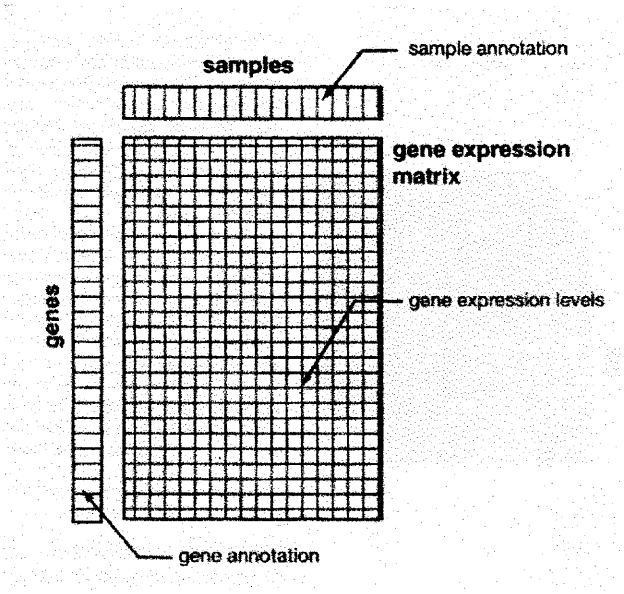

Figure 1.4 Conceptual view of gene expression data [9].

The urgent need for tools to rapid computations from these data has stimulated the production of software packages capable of performing concurrent calculations on gene expression data. Along with the emergence of basic tools, a plethora of competitive software tools has flooded the research market.

1.3 Available Resources for this Study (The Data Set)

The application of microarrays has become widespread at all levels of medical research pertaining to disease control and drug development. Some well-established areas of microarray research exist in oncology (the study of cancer), bone metabolism, cardiovascular diseases, respiratory diseases and immunology studies.

Arthritis is a class of disorders that affects joints and do not have one particular cause. This class of diseases represents perhaps the most common diseases of the modern age, affecting one person in seven in the U.S and Canada, or about 45 million people. The

most common arthritic disease is osteoarthritis, which affects 30 million people in North America. According to some estimates 80% of the people over the age of 70 suffer from osteoarthritis. About 3 million people in North America suffer from rheumatoid arthritis [24]. The cause of osteoarthritis is believed to be general wear and tear within the joint, and is associated with aging, while rheumatoid arthritis is an autoimmune disease that affects all ages. Because the various forms of arthritis have differing causes, different therapeutic approaches are required to determine the root cause of the disease. Blood tests for rheumatoid factor (RF), a marker for rheumatoid arthritis, are useful, but there may be no direct correlation between RF and the disease [24]. One approach to identify the possible biological cause is to analyze the change in gene expression caused by given experimental manipulations.

The data set used in this study was part of an experimental study during my Summer Internship at Novartis Pharmaceuticals and was performed in conformity with locally applicable animal welfare regulations. The aim was to identify genes that were up-regulated or down-regulated (in expression) in the development of musculo-skeletal effects resulting from a given experimental treatment. The experiments were performed in the general context of understanding the molecular and biochemical basis of arthritis.

Rats underwent an experimental treatment for seven and 21 days. Control animals were sacrificed at corresponding time points without having undergone this experimental treatment. At the end of the experiment, tissue samples were taken from paw connective tissue, liver, lung, skin and kidney, along with blood samples. After hybridization of *Rattus norvegicus* tissue RNA to the Affymetrix RG34A, the MAS^{TM} 4.0 (Affymetrix Microarray Suite) that contains empirical algorithms, was used to analyze the data.

The table depicting the number of replicates per control and treatment group is listed as follows:

Table 1.1 Number of Replicates (Chips) per Treatment Group

Tissue Type	Control 7-Day	Treatment 7-Day	Control 21-Day	Treatment 21-Day
Paw	5	3	4	5
Lung	5	5	4	5
Liver	5	5	4	5
Kidney	5	4	4	5
Blood	4	5	4	5
Skin	5	5	4	5

1.4 The Need to Compare Data Mining Tools

As cited by A. Brass, 2001 [11], "The bioinformatics tasks of microarray analysis can be divided into three linked activities -- data capture, data mining and visualization and interfaces". A lot of attention has been focused on data capture and new technologies to produce microarrays and techniques to increase the efficiency of capturing data from biological tissues. On the other hand, not much attention has been given to the data mining, visualization and interface areas of study. Even though many algorithms are being developed, the implementation of these algorithms in the form of computer programs is not progressing at the same pace. The availability of software packages for gene expression analysis has been a concern for researchers in recent years. The characteristics of each tool play a major role in the selection process. This is an important

issue with the limited amount of choices in the market. The selection should be based on the need of the company or individual and on the availability of servers and computers.

The convenience of a user-friendly statistical tool integrated with computer programs has eliminated the need for researchers to learn computer programming before pursuing gene expression analysis. However, a biologist with some statistical knowledge would be an ideal candidate to comprehend gene expression analysis.

The issue that lies in the selection of the tool is to obtain an accurate and precise understanding of the gene expression changes. The results of such analyses are critical since they may affect inferences regarding the expression of a given gene and thus the possible role of the gene in a particular disease.

Data mining tools should be selected on the basis of their merits and appropriateness for the given laboratory, independent of promotional information that may be disseminated by the vendor. It is thus beneficial to determine the efficiency of a particular software package before expending resources required for its implementation. The focus of this study is to analyze some of the features of the graphical user interface (GUI) of data mining software packages and the importance of this aspect in gene expression analysis studies. Since these packages aim to make results of analyses available to a wide range of users, thus, the implementation of the algorithms and statistical tests is favorable only if in a suitable, convenient and easy to use manner. This demonstration of comparison of features is explained by the implementation of some of the common statistical analyses and clustering techniques of two different tools by pursuing an analysis of mRNA expression data.

There have been a handful of researchers who have envisioned the need to compare different tools for the benefit of the users. Relatively few reports have described comparisons of microarray data analysis tools. A comparison of such tools may provide a prospective user with some initial guarantee of the enhanced quality of results, based on prior analysis of analysts, who have studied the details of the external and internal structure of the gene expression analytical tool. An ongoing study of microarray data analysis software by Y. F. Leung [23] is one of the more widely known, publicly available resources. This study compares the availability, and outlines the features of many packages, including information on vendor contacts for the tools and prices. Another large study of comparison of software packages including surveys from users was performed by the CSC, which is the Finnish IT Center for Science, owned by the Ministry of Education [19]. Since their study was based on some previous versions of the tools, there was a mixed response regarding the user-friendliness and ease of learning of GeneSpring, yet out of the 18 people surveyed, 17 suggested that they preferred continued use of the tool [20]. The Stanford Microarray Database group of Stanford University School of Medicine performed another study of the comparison of tools [21], which outlines some available tools with more expansion referred to the study by Y. F. Leung [23].

CHAPTER 2

COMPONENTS OF A GENE EXPRESSION ANALYSIS STUDY

2.1 Biological Aspect

Biologists often require information on particular genes that are induced or repressed in different treatment and disease conditions. This includes information, such as the function of the gene, similarity to other genes and how its expression correlates with the condition under study. A biologist takes this information about a gene and attempts to understand the function and effect of the gene on other biological mechanisms of the human body. DNA microarray technology allows biologists to detect the mRNA levels of thousands of genes in the cells at one time. Microarrays allow rapid gene expression monitoring and sequence analysis at the genomic level. The information obtained from this process is further used to assist the drug discovery process. Drug-human interactions can be explained after a comprehensive understanding of the change in gene expression by the effect of a certain drug treatment. Thus, the adverse as well as beneficial effects of a drug can be monitored by the differential expression in the corresponding genes. Since the information about these genes is computed via different algorithms, the computational and biological aspects of gene expression analysis studies are closely related.

2.2 Computational Aspect

Analysis of gene expression data can take place in several different ways, the most common of which are the statistical analysis and gene classification. The statistical analysis consists of implementation of various statistical tests such as the t-test, multiple testing correction, p-value, etc. Gene classification is basically of two main types, namely, supervised and unsupervised classification [17]. The unsupervised type of classification implies that the classification is not known a priori and hence needs to be discovered based on the pattern depicted by the data. Hence, new genes would be identified based on similar expression profiles and then a group of similar genes would be classified together. Examples of this type of classification are cluster analysis, class discovery and unsupervised pattern recognition. The supervised type of classification implies that the classes are predefined before data analysis and the aim is to determine the basis of classification from training or learning sets, which serve as a model to understand the behavior of similar genes. In this case, genes are grouped together according to some similar criteria, and the user has to understand the criteria of grouping, based on previous knowledge of the genes. Examples of this type of classification are k-means, discriminant analysis, class prediction and supervised pattern recognition [17].

The simple analysis involves the study of gene expression profiles after the data was filtered according to several concurrent criteria, so as to minimize the presence of spurious changes:

1. The fold changes (from the expression levels and normalized data) induced by the experimental treatment as compared with the corresponding control.
2. The mean expression levels for each group.

3. The statistical significance of any differences noted between the control and experimental groups.

Various gene lists were formed and compared (by MS Access) to identify genes that changed similarly under different conditions of time and tissue.

The basic analysis of gene expression serves the following two purposes:

1. Provides information about the general trend of the change in expression levels of different genes under different experimental conditions.
2. Acts as a pre-processing step for the data to be clustered for visualization of the relative expression of many genes at one time.

The basic analysis process includes some of the following steps, described as absolute call metrics, statistical significance, normalization methods and scaling.

2.2.1 Absolute Call Metrics

According to MAS^{TM} 4.0 (Affymetrix Microarray Suite), the criteria for determining the absolute call is based on the absolute difference measurements [6]. The absolute call, also known as the absolute measurement, is a criteria for consideration of a gene in some statistical and gene expression analyses. The three absolute calls are Present, Marginal and Absent and can be determined by the Average Difference calculation based on the perfect match and mismatch oligos [1]. These oligos are present on probe pairs and are used as a basis of comparison to calculate the non-specific hybridization of mRNA by calculating an Average Difference, which is calculated according to the formula:

$$Average\ Difference = \frac{\sum_N PM - MM}{N} \quad (2.1)$$

where N is the number of probe pairs, PM is number of perfect match signals and MM is the number of mismatch signal intensities [6].

As a way to reduce potential outliers in this calculation, only those probe pairs are used that deviate less than 3 standard deviations from the calculation [1]. Probe statistics are the basis for Affymetrix to decide the absolute call as Present, Absent and Marginal. The Average Difference is negative if the number of mismatch oligo probe pairs exceeds those of the present match oligos. This indicates that either the target is absent or the hybridization is non-specific [6].

Many studies have yielded different interpretations of these absolute calls and their exact influence on the decision to include the corresponding probe pairs as part of the calculation for final change in gene expression.

Both the data mining tools assessed in this study allow the use of any combination of absolute call measurements in the statistical analyses as desired by the user.

2.2.2 Statistical Significance

Statistical significance refers to the mathematical weight given to a particular gene in an analysis. It re-evaluates the importance of the presence of a gene and considers the probability that a gene would exist in a calculation by chance. Thus, it is a mechanism to increase our confidence that a change in expression is more than a simple variation corresponding to an array process. The reason for a given change in gene expression to occur by chance may involve experimental error, which could be caused by insufficient replicates of a particular sample. Statistical significance is calculated by the p-value

(probability value), which is directly proportional to the observed variation in genes by chance. The p-value is a transformation of the student's t-test that determines whether the mean intensity is statistically different from 1. The formula for this is:

$$t = \frac{\bar{X} - 1}{\frac{S_x}{\sqrt{n}}} \tag{2.2}$$

where '$\bar{X}$' is the Mean intensity, 'S_x' is the Standard deviation and 'n' is the number of replicates.

For example, while calculating the magnitude of gene expression changes, one may notice up-regulation of a gene by greater than three fold, i.e., the level of expression in one group may be three times the level of another group. This also indicates a fold change of 0.47 if the logarithmic forms of the initial levels are taken [1]. This gene may be an unstable gene and hence may tend to reveal widely differential values in separate conditions. Taking replicates of a sample and identifying consistently differentiated genes increases the probability that observed changes in gene expression are genuine, as opposed to arising from chance alone. This is where the test for significance comes into play. It helps to determine relative initial signals in all samples of the same condition and the corresponding change in signal in all samples of another condition. This can be calculated by determining the mean and variation of each gene across all samples. This can also be calculated by determining the standard deviation of each gene, then comparing the difference in expression levels between the two conditions with this standard deviation [6]. The more the change exceeds the standard deviation between replicates, the more significant it is [1].

2.2.3 Normalization Methods and Scaling

Normalization ensures that all the calculations and analysis from one gene chip to another are comparable to each other. This is a validation method to ensure that the level of mRNA measured on one chip is similar to that of on another replicate chip. Thus, normalization is a method to remove any variations or errors produced by the microarray technology process. One method to facilitate this comparison is to include "Housekeeping genes" on every chip and to use them as reference points for normalization. "Housekeeping genes" are genes whose expression levels are thought to depict consistent gene activity across any condition or treatment. A comparison of these genes could serve as a standard towards the initial calculation of bringing all the other genes on that chip to the same platform thus enabling a valid cross array comparison. This method is not commonly used any more. They have a limitation, since they may tend to be highly expressed and hence may not be representative of other genes of interest [8]. The more widely used method consists of normalization by the mean. Once the mean is calculated, and graphed as a normal distribution, a certain percentage (from the tails of the distribution) is clipped off, and the remaining is used as a scaling factor for the normalization.

The selection of an appropriate normalization method is based on the user's desired interpretation, and therefore the presence of a variety of normalization methods in comprehensive data mining tools is recommended.

2.3 Visualization of Data and Cluster Analysis

Easy visualization and interpretation of data is possible when the data is seen in collective groups of similar expression, function or general behavior. One of the most common methods to visualize data in this manner is by cluster analysis. The aim of clustering is to partition entities (genes) into groups based on given features of each entity to ensure that the groups are homogenous and well separated. Each group is called a cluster, and the partition is called clustering [29]. Clustering of data is also known to strengthen the signal when averages are taken within clusters of genes [17]. Clustering is a method of grouping genes that share similar expression patterns. This may also translate to similar biological function or structure, but basically depends on the interpretation by the biologists for such a conclusion. The results of clustering produce an aggregation of genes that portray the following properties [4]:

1. Reduced Intra-variability: This is also termed Homogeneity, as the elements in the same cluster are highly similar to each other.
2. Increased Inter-variability: This is also termed Separation, as the elements in different clusters show little similarity with each other.

Clustering algorithms are of two types, namely, Agglomerative and Divisive. The difference between the two can be illustrated in the following descriptions of both:

1. Agglomerative Clustering:

   This is also known as the 'bottom-up' approach, where:

   Input: Number of clusters = n.

   Output: Number of clusters = 1.

   Here 'n' is the number of single element sets containing each gene under study.

   Example: Hierarchical Clustering

2. Divisive Clustering:

   This is also known as the 'top-down' approach, where:

   Input: Number of clusters = 1.

   Output: Number of clusters = n

   Example: k-means Cluster, Self-Organizing Maps

   The type of clustering algorithm to be used depends on the biological problem [14]. Measures of similarity and linkage methods determine the mathematical criteria of clustering. Some of these are explained below:

2.3.1 Measures of Similarity

Measure of Similarity is a quantitative measure that determines the similarity between genes based on their expression profiles. These measures affect the clustering process, since some measure the similarity in expression levels, whereas others measure the similarity in expression patterns. This can be a critical step in the decision making process and depends solely on the requirements of the biologist. Also, some decisions are for the analyst to make whether the similarity measure should consider the effect of the outliers.

When each gene is plotted in a dimensional space using a standard x-y coordinate system, the distance between the genes can be used as a criterion to evaluate the measure of similarity of the genes.

As an attempt to reduce dimensionality for easy visualization of gene behavior, genes can be expressed in the form of vectors of a number of dimensions, where each dimension is represented by an experimental condition. The vector angle between two genes is also a commonly used distance for gene expression analysis as vectors demonstrate magnitude and direction similar to genes that demonstrate expression levels and pattern.

The Measures of Similarity can be divided into the following three types [4]:

1. Measures of Correlation
2. Measures of Distance
3. Measures of Confidence

Each of these is briefly defined below. In the figures corresponding to the following descriptions, the green line represents the gene from which a distance is calculated. The blue and black lines represent the genes closest to and furthest from the green gene respectively, according to the many different distance metrics described [3].

1. Measures of Correlation:

These are represented as correlation measures ranging from −1 to 1 and can be anywhere from exactly opposite to the exactly same respectively [4].

Correlation:

This method clusters genes that are closely related in terms of pattern. The expression level has no importance. Two genes are correlated when their expression values increase and decrease simultaneously (gene 1 and 4). They are anti-correlated when the converse is true (1 and 3). Genes must be correlated or anti-correlated to be clustered [3].

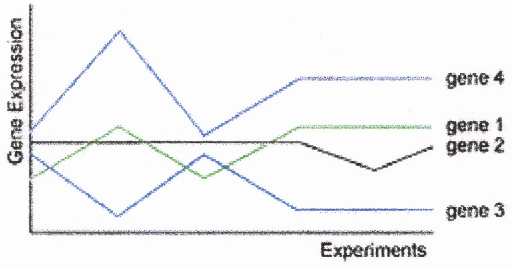

Figure 2.1 Correlation Metric.
Courtesy: Gene Data

Standard Correlation:

This measures the angular separation of expression vectors of two genes around zero. This method emphasizes the consistency of points where genes are over expressed [4].

The standard correlation coefficient (dot product of two normalized vectors) has been found to agree well to the concept of coexpressed genes. This could be due to the fact that the statistic considers similarity in shape (the gene pattern) as opposed to the magnitude of the two series of measurements (the expression levels) [17].

Positive Correlation:

This method clusters genes that are closely related in terms of pattern. The expression level has no importance. Two genes are positively correlated when their expression values increase and decrease simultaneously (gene 1 and 3).

This method is used mainly in gene expression analysis and is most useful for clustering [3].

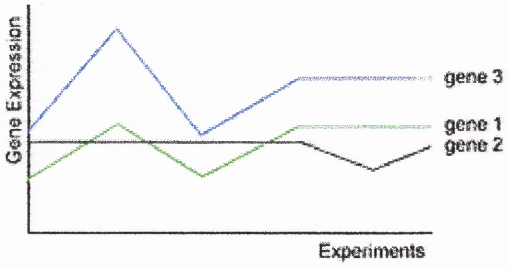

Figure 2.2 Positive Correlation.
Courtesy: Gene Data

Pearson correlation:

This method is similar to standard correlation, except that it measures the angle of separation of expression vectors of two genes around their mean expression levels instead of around zero. This method emphasizes consistency of both, over and under expression of genes. In log mode, Pearson and standard correlation are very similar [4].

Spearman Correlation:

This is similar to the Pearson correlation, except that the expression levels are replaced by their ranks. The method reduces the effect of large individual variations on the calculations (non parametric). The elements are ordered in the vector and then ranked, and new vectors are formed with ranked and ordered elements [4].

Change Correlation:

This method is applicable only for ordered condition points, and measures the change between each pair of elements (genes). The standard correlation is computed on these

change values. This method emphasizes consistency of change in gene expression levels, both upward and downward [4].

Up-regulated Correlation:

This method focuses on the upward change between each pair of conditions and then the standard correlation computed for the change values. This emphasizes periods when new RNA is being synthesized. This is applicable only for ordered condition points [4].

Smooth Correlation:

This method measures the agreement on smooth trends in data. This is measured by interpolating the average of each consecutive pair of elements (genes). A new value is inserted in each new value between the old values and then the standard correlation computed on the result [4].

2. Measures of Distance:

The distance measurement calculates the dissimilarity from 0 to infinity, by calculating the square root of the standard deviation. This is based on the measurement of the Euclidean distance between the expression profiles for gene A and gene B [4].

Euclidean (or L2):

This metric clusters genes that are closely related in terms of expression level. The method is sensitive to outliers. Thus, although gene 2 generally has expression values closer to gene 1, it is represented by a black line due to the effect of an outlier (circled

red) in one of the experiments. It is calculated as the sum of the squared distances of two vector values of two genes [3].

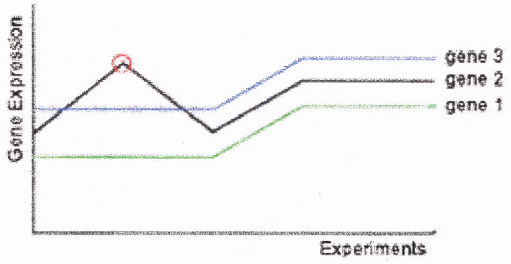

Figure 2.3 Euclidean Metric.
Courtesy: Gene Data

Normalized Euclidean:

This method of similarity clusters genes that are closely related in terms of expression pattern. The expression level has limited importance in this method. Two genes are similar when their expression values increase and decrease simultaneously by the same amount (gene 1 and 3). As a result, on a logarithmic scale, genes have exactly the same pattern [3].

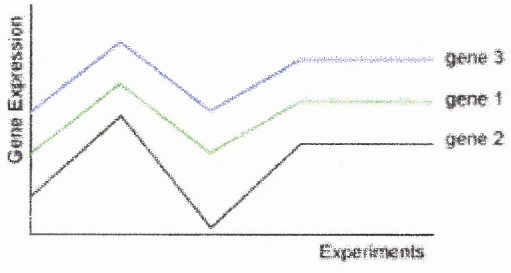

Figure 2.4 Normalized Euclidean.
Courtesy: Gene Data

L1 (or Manhattan):

This is a linear version of the Euclidean distance calculating the sum of the absolute distances of two vector values of two genes. The genes that are closely related in terms of

expression level are clustered together. This method is not sensitive to outliers. In this case the outlier (circled red) has less effect, and gene 2 is deemed closer to gene 1 and represented by a blue line [3].

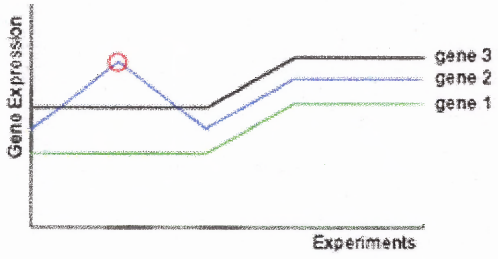

Figure 2.5 L1 Metric.
Courtesy: Gene Data

Maximum:

This metric clusters genes that are closely related in terms of expression level. This method is extremely sensitive to outliers. The calculated distance is the maximum distance between two genes. Therefore, only one outlier determines the calculation of the distance. In the figure, the blue line represents gene 3 since its maximum expression value (a) is lower than those of genes 2 and 4 (b and c) [3].

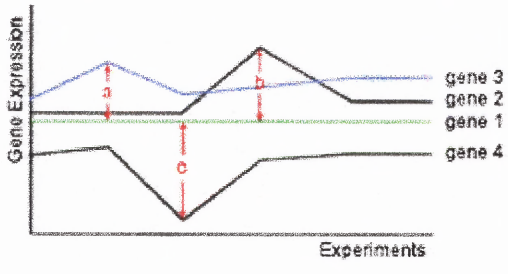

Figure 2.6 Maximum Metric.
Courtesy: Gene Data

3. Measures of Confidence:

The confidence results range from 0 to 1 representing a transition from no confidence to perfect confidence [4].

Spearman Confidence: If the Spearman correlation is represented as 'S' then the confidence is calculated as (1-P) where 'P' is the probability of getting an 'S' or higher sample correlation by chance alone, if the true correlation is zero [4].

Two-Sided Spearman Confidence: This is also calculated as (1-P), but 'P' here represents the probability of getting a sample correlation of |S| or higher, or –|S| or lower by chance alone if the true correlation is zero. This two sided test emphasizes genes that have either similar expression profiles or opposite profiles. Thus, genes with high two-sided confidence values contain similar as well as dissimilar genes [4].

2.3.2 Hierarchical Clustering

As described in Section 2.2, hierarchical clustering is a type of unsupervised type of classification, where the classes are defined by the data analysis and hence the clusters are not formed based on any previous knowledge of the behavior of genes. The main purpose of hierarchical clustering is to create a mock-phylogenetic tree also known as a dendrogram that places genes with similar expression patterns into nearby groups. The distance of one gene to the node connecting it to another gene indicates how closely the two genes are correlated. A shorter distance, which can be measured by the number of intervening nodes marks higher correlation between the genes. The distance travelled up a branch is directly proportional to how different the genes are.

Hierarchical clustering uses the agglomerative method of clustering algorithm as described in 2.3. The way the algorithm works is that each gene is considered as an individual set of elements of its own. The aim is to produce a dendrogram, or tree that

consists of all the genes under study to form a superset by the combination, or agglomeration of the entire single element sets.

2.3.2.1 The Algorithm. An outline of the algorithm of hierarchical clustering consists of the following steps [15]:

a) A pair wise distance matrix is constructed for all the genes in the study. This distance matrix is a table listing the genes along with the distance it bears with other genes. The distance between the genes is also determinative of the similarity in expression profiles of the genes and hence closer genes possess similar correlation coefficients.

b) The two closest genes are selected from this matrix that are also computationally, the least expensive to merge. These two single-gene clusters can be labeled C_1 and C_2 and clustered together to form a new cluster, which replaces the two single gene clusters. All the distances affected by the merge are recalculated.

c) This cluster is compared to other similar single element gene clusters and the distance between them is calculated.

This whole process is iterative from step (b) to (c) until the final result is a single cluster with smaller clusters analogous to a tree with many branches.

2.3.2.2 Linkage Methods. The Measures of Similarity described in Section 2.3.1 measure the many ways of calculating the similarity between genes in terms of distances between them. Once the genes merge to form clusters, the between-cluster dissimilarity measures are termed the linkage methods [15]. If two clusters are represented by x and y, then the number of gene sequences in each cluster is taken as S_x and S_y respectively. Based on the distance between these sequences of different clusters, there are basically three types of Linkage Methods [15,17]:

a) Single Linkage: This method uses the shortest distance between two sequences of different clusters as the total distance between two clusters. This is also known as the nearest neighbor or minimum method, since the sequence from one cluster seeks its closest neighbor (of minimum distance) from another cluster. The result of this linkage analysis is not very stringent, since the clusters are agglomerated on the basis of the least possible distance. The cost function of this method can be represented as:

$$\text{Cost function} = \min (\text{distance } (S_x , S_y)) \tag{2.3}$$

b) Average Linkage: This method uses the average distance between two sequences of different clusters as the total distance between two clusters. The cost function of this method can be represented as:

$$\text{Cost function} = \text{average (distance } (S_x , S_y)$$

$$= \frac{1}{| S_x | | S_y |} \sum_{S_x M x} \sum_{S_y M y} (\text{distance } (S_x , S_y) \tag{2.4}$$

The pair wise average linkage method is the most commonly used type of linkage method used for gene expression analysis studies, based on the results of Eisen et al. [17].

c) Complete Linkage: This method uses the greatest distance between two sequences of different clusters as the total distance between two clusters. This is also known as the furthest neighbor method since the sequence from one cluster seeks its farthest neighbor from another cluster. The clusters formed are close to each other due to the agglomeration of distant genes, thus reducing distances between different clusters. The cost function of this method can be represented as:

$$\text{Cost function} = \max (\text{distance } (S_x , S_y)) \tag{2.5}$$

Separation Ratio [4]:

This is how large the correlation difference between groups of clustered genes has to be for the groups to be considered discrete groups and not be joined together.

a) Increasing separation increases the "branchiness" of the tree.

b) This can range from 0 to 1.

c) At a separation ratio of 0, all gene expression profiles can be regarded as identical.

2.3.3 K Means Clustering

This is the divisive type of clustering technique, or the 'top-down' approach, where the user chooses the number of clusters desired. The process is based on the supervised type of classification, described in Section 2.2, since some prior knowledge regarding the gene expression pattern is known. The genes are randomly assigned into user-defined number of clusters (based on the different expression patterns displayed by the data set). K-means clusters are visualised in the form of graphs of expression profiles as opposed to dendrograms of hierarchical clustering. The steps of the algorithm include [15]:

a) Input of a user-defined number of clusters.

b) Random distribution of objects (genes or samples) into user-defined number of clusters.

c) The average expression vector is calculated for each cluster; this is also the centroid for each cluster.

d) This is used to determine the distance between clusters. The objects move according to the specification of being closer to a cluster's centroid.

e) The centroids are recalculated according to the new objects.

f) Thus, the final clusters contain objects that are closer to it, as compared to the previous cluster they were in.

g) The objects continue to move till the clusters formed are not stable enough to be altered.

This is an iterative process until (g) is no longer achieved. Until then, all steps from (d) to (g) are repeated. Thus during the run time of a k-means process, the number of genes per cluster continue to change based on the average radius (centroid) calculated for each iteration. The average radius is proportional to the tightness or stability of the gene to stay together, in a cluster and hence a greater correlation [4]. Random start implies potentially randomized clusters that need to be repeated to verify results. Improvements have been suggested which reduce the extent to which the search becomes trapped in local minima.

CHAPTER 3

THE DATA MINING TOOLS

3.1 Types of Data Mining Software

Microarray software can be of many types based on the area of analysis of the data. The different types include Image Analysis software, Data Mining software, SNP's Analysis software, Database/LIMS software, Public Expression Database, Primer Design and software for further data mining [23]. The scope of this thesis is limited to a study of data mining software. The purpose of a data mining software package for a research scientist is to be able to analyze the data based on a user-friendly interface of a robust computer program. This process of obtaining knowledge from information is the basis of existence of data mining software. This type of software aids in the transformation of a set of numbers into images to visualise changes in gene behaviour. It is dependent on the need of the user to focus in on a particular aspect of the image and analyse it to comprehend the aim of the particular study.

Data mining software can be divided into four basic types [23]:

1. Turnkey System: This is a computer system that is customized according to the need of the users. This includes all the requirements of an operating system, server software, database, client software, statistics software and even hardware. Examples: Rosetta resolver (Rosetta Biosoftware), Genetraffic (Iobion) and Expressionist (GeneData).

2. Comprehensive Software: This type of software can incorporate many different analyses for different stages in a single package. Examples: Cluster (Mike Eisen, LBNL), GeneMaths (Applied Maths) and GeneSpring (Silicon Genetics).
3. Specific Analysis Software: A software package of this type performs a couple of specific analyses. Examples: GeneCluster (Whitehead Institute Centre for Genome Research), SAM –Significance Analysis of Microarrays (Stanford University).
4. Extension/accessory of other software: This is an extension of another software's capability. Examples: Freeview, Arrayminer (Extension of GeneSpring).

3.2 Criteria for Selection of Data Mining Tools

The criteria for selection of data mining tools depend on the needs of the users, and not necessarily on those of the research scientists who design and carry out the experiments. The fact, that these users are not experienced computer programmers, statisticians and biologists has to be taken into consideration. The role of most users of the tools is to analyse the quantitative data, transform it into different forms of easy-to-understand visual representation and then hand it over to the biologists and scientists responsible for the detailed analysis of the results.

The requirements of the user include ease of use (in terms of user interfaces), management of the large amount of data, ability to access the data from different operating systems (and different workstations) and availability of the maximum number of options for different statistical analyses tests.

The majority of users may be familiar with some basic computer applications. Among these, the most commonly used is Windows. Interfaces with the 'click and select'

or 'drag and drop' methods are the easiest to work with for users lacking advanced computer technology expertise.

The major aspects of data mining software for gene expression analysis usually consists of:

1. Time Complexity: Time complexity refers to the run time of the algorithms being used, expressed as a function of the problem size. The exact run time of the algorithm depends on complexity of the algorithm and this, in turn, depends on the skill of the programmers who are responsible for writing the program code of these algorithms.
2. Preciseness and Accuracy of the Result: The computer programs should be robust with respect to the method of calculations, as the accuracy of results is the most crucial step towards the study of gene expression. The computational analysis of gene expression data from microarray technology is the first step towards producing a global perspective of the whole model of study. Results from these preliminary analyses are the next step to further validation by other laboratory methods and research studies. If these results were erroneous, subsequent studies could prove futile and waste time, money and resources. The precision of the results depends on the written code of the algorithm, which is required to be flawless in terms of coding, implying the absence of any type of bugs. In other words, the code should be such that it would work in any given environment and would not have to be debugged in case of any discrepancy.
3. Reduction in Dimension of Data for Visualisation: Reduction of higher orders of mathematical dimension improves visualisation of data by lower-dimensional human

conceptual and perceptual abilities. The images and data have to be adapted to levels compatible with human cognitive capacities, for easy interactive interpretation.

4. Import of Data: Access of the data is the first step towards data analysis and hence ease of database connectivity is a major issue for any data mining software package. The data then has to be transformed to a standardised data model consisting of rows of genes and columns of experiment attributes. Open Database Connectivity (ODBC) is a common method of abstracting a program from a database. JDBC is a Java Application Programming Interface (API) for executing SQL statements. By using the JDBC API, one can access almost any data source, from relational databases to spreadsheets to flat files.
5. Software Architecture: The programs have to be written in a language (e.g. Java) that is platform independent to make them compatible across many commonly used platforms at the same time. The most common platforms are workstation-based applications (as Macintosh, Windows, UNIX and Linux operating systems).
6. Probing of Visualisation: The results of the overall visualisation should be worthy of being probed further in order to focus in on a particular segment to facilitate a complete detailed analysis.
7. Cross Validation of Gene Properties by Inter-networking Links: The annotation of a particular gene can be confirmed by accessing other related databases. This is feasible if the software tool is capable of linking to external databases and has web access.
8. Accessibility and Availability: The accessibility and availability of the software influences the ease of use.

There are numerous software packages available, both for commercial use as well as those that are free for academia. The choice of selection depends on the frequency of use and the resources available as, size of server and/or speed of computers. If all the criteria for selection are available, then the selection of the applicable tool has to be made according to some important and standard criteria which involves the analytical tools embedded in the software package.

3.3 Criteria for Selection of Expressionist™ 3.1 and GeneSpring™ 4.2

The criteria for selection for the comparison of the two tools used in this study are based on the similarity of the functioning of the two commercially available tools. Both tools belong to a different category of classification of data mining software, yet the final results produced by both tools are very similar. Expressionist™ 3.1 is a turnkey type of software whereas GeneSpring™ 4.2 is categorized as a comprehensive type of software package. Besides the similarity in functioning, these tools were conveniently available on site at Novartis Pharmaceuticals during my Internship. Also, the two tools are Java based applications thus making them platform-independent to run on the operating system environments of Windows, Macintosh and Linux. Both tools are primarily accessible to other databases that increase their flexibility to work with in-house proprietary databases. This is a major issue with pharmaceutical companies and other proprietary research study fields in order to maintain the integrity and confidentiality of research data. Both tools also possess a user-friendly Java Applet viewer that makes it easy for anyone without a statistical or computer programming background to be able to perform the different analyses.

There are basically three methods that can be applied to microarray data, Classification, Clustering and Projection. The classification method comprises of Support Vector Machines (SVM) and Classification And Regression Trees (CART). The clustering includes k means and hierarchical trees. The projection method involves Multi Dimensional Scaling and Principal Component Analysis. Both Expressionist™ 3.1 and GeneSpring™ 4.2 contain most of these features and hence can be considered comparable on the grounds of the commonly used methods for gene expression analysis. The functionality of the tools utilize the techniques common to Windows of dragging and dropping, item selection and scrolling. Thus, the user interfaces are friendly and easy to learn. The usage model of both tools include [3]:

1. Data Upload and Structuring
2. Data Evaluation
3. Data Cleansing
4. Data Extraction and Analysis

The image representations and data export were reproducible directly from the data mining Tools.

MS Excel was also used in this analysis, primarily for easy visualization of gene lists and for summarized results. Results of some analysis were transferred to MS Excel to obtain a quantitatively global representation of the analysis. Data manipulation was particularly useful for sorting of lists of data and implementing user-preferred formulas present in MS Excel.

Table 3.1 Comparison of the System Requirements for the Use of the Selected Tools

Attribute	Expressionist™ 3.1	GeneSpring™ 4.2	MS Excel
Type of Data Mining Software	Turnkey (hence customized for an application)	Comprehensive (hence incorporates different analysis in a single package, universal for all)	Not applicable
Server	UNIX based computer with 1 GB RAM: Java 1.3VM	Not applicable	Not applicable
Client	Windows PC with a minimum of 128 MB RAM; web browser supporting Java Web Start	1. Client: Windows: Windows 95/98/NT/2000, 256 MB RAM (512 recommended) 40 MB of free disk space 1024 x 768 display Pentium II or better 2. Client: Macintosh: Mac OS 8.1 or higher (OSX Classic mode) 256 MB Ram (512 recommended) 40 MB of free disk space 1024 x 768 display MRJ 2.2.5 3. Client: Unix: Most common Unix OS's (Linux or Solarix recommended) A JVM installed that supports JDK 1.1 or later 256 MB Ram (512 recommended) 40 MB of free disk space 1024 x 768 display	Windows 95/98/NT Version 4.0 or later, 256 MB RAM (512 recommended), 40 MB of free disk space, 1024 x 768 display, Pentium 75 MHz or higher processor, Memory For Windows 95 or Windows 98: 16 MB of RAM for the operating system For Windows NT Workstation: – 32 MB of RAM for the operating system. 146 MB of available hard-disk space Display: VGA or higher resolution monitor
Database	Oracle 8.1.6 or higher	Not applicable	Not applicable

Table 3.2 Comparison of the Features of Expressionist™ 3.1 and GeneSpring™ 4.2

Feature	Expressionist™ 3.1	GeneSpring™ 4.2
Data Quality	Affymetrix Feature Quality	Affymetrix Feature Quality, Clontech, Axon, Biodiscovery, Incyte, Packard Biochip, Generic one-color and two-color
Data Handling and Display	Profile Display, Log-log Plot	Profile Display (can change horizontal and vertical axes or specify fold change intervals, colors, plot symbols and grid lines by using the Display Option Windows)
Statistical Analysis Tools	Histogram, Box-Plot, Tile Plot, Parallel Coordinate Plot	Bar Graph view, Correlation
Data Filtering	Valid Value Proportion, Average Expression, Variance, Highest Ratio	Average Expression, Fold Change, Statistical Group Comparison
Sample Comparison	N fold Regulation (Scatterplots), 2 Groups (Parametric tests), 2 Groups (Rank Test), 2 Groups (Absent/Present Search), K groups, K Ordered Groups	Scatterplots, Array layout view, Pathway views, Ordered List view
Similarity Search	Distance, Profile Distance Search, Group Characteristic	Similar genes according to Correlation coefficient and p-values
Clustering	Hierarchical, 2 D Hierarchical, Partitioning (k means and Self-Organizing Maps)	2 D Hierarchical, Partitioning (k -means and Self Organizing Maps)

3.4 Method of Comparative Analysis

The method of comparing two software data mining tools is by examining all external and internal features available in the tool. The external features refer to the user interface of the tool. A user-friendly interface is one that is easy to learn for a novice and has ample resources of learning, available with the tool. These resources could be in the form of online manuals (either separately, or in the form of hyperlinks from the main interface), separate handbooks, vendor supported tutorials or telephonic technical support by the vendor. Comparison of interfaces between two tools can be made by general use of the tool and outwardly appearance of the results produced. This consists of the quality and flexibility to modify the graphs and all such pictorial representation of the data.

The other basis of comparison consists of the internal features of the tool, including the actual algorithms and statistical tests being used by the tool. This type of a comparison is possible only after a thorough analysis of a data set, followed by comparison of the results from the two tools. Also the thorough analysis by the gene expression analysis tool could be followed by the validation of other independent methods of gene expression analyses, not involving microarray data analysis.

In this analysis, the data set described in Section 1.3 was used as a sample data set to determine the reliability of results of the two tools in study. This study was organized to compare the characteristic features of two data mining tools for gene expression analysis. Since all tools vary in external and internal features available, the aim of the study was to lay out the results of some statistical analysis involved in microarray data analysis. The goal was to reveal the observation of the importance of a data analysis tool in the final results of a study. Many user interface features as well as results of different

statistical algorithms and tests were compared between the two tools. For a proper analysis of the reliability of a data mining software and determination of whether it may comprise of all the essential algorithms, it is essential to understand the components required for a gene expression analysis study.

CHAPTER 4

RESULTS OF COMPARATIVE ANALYSIS

4.1 Comparative Analysis of the User Interface

Even though Java is used as the application-programming interface for both the tools in this study, there are some differences in the basic User Interface as follows:

1. Image Quality: In Expressionist™ 3.1, the image is of a bitmap form, whereas in GeneSpring™ 4.2, the image is a vector-based graphic [4]. Bitmap graphics are made up of bits or pixel, but vector-based graphics are formed from vector objects. Since the relationship between the vector objects is fixed, the size of vector graphics can be changed, but still look exactly the same. Vector graphics are resolution independent, which implies consistency in the looks of the image irrespective of the dots per inch used. On the other hand, the quality of bitmaps depends on the resolution since they are made up of individual pixels. Due to this characteristic of bitmaps, the image quality can be lost due to the change in size. Bitmap images also have a rectangular shape, which means they generally have a background [4].
2. The Java Applet viewer of GeneSpring™ 4.2 allows the user to view the loaded experiments as well as the results of the analysis in the same window; in Expressionist™ 3.1, the user has to switch between the Data and Results tab while changing any of the needed selections.
3. The "2 Groups Absent/Present Search" is present only in Expressionist™ 3.1. This test is only useful when the user has allowed all calls of "Absolute Call Measurements", i.e., Present, Absent and Marginal during the pre-processing analysis

of data. If the user starts an analysis, taking into consideration only Present calls, then the purpose of this test is nullified.

4. In GeneSpring™ 4.2, the user can modify the color settings of all visual representations. This includes the actual colors as well as the parameters to be colored by, for example, by expression, significance, classification and parameters. All color settings are fixed in Expressionist™ 3.1 and cannot be modified.

5. In Expressionist™ 3.1, the user can view individual gene expression profiles during the process of statistical filters as "Filter by N-Fold Regulation", "Filter by Expression Levels", "Filter by Variance", "Parametric Test", etc. This is not possible in GeneSpring™ 4.2, where the filtration of genes is shown by the reduction in the number of genes in the list.

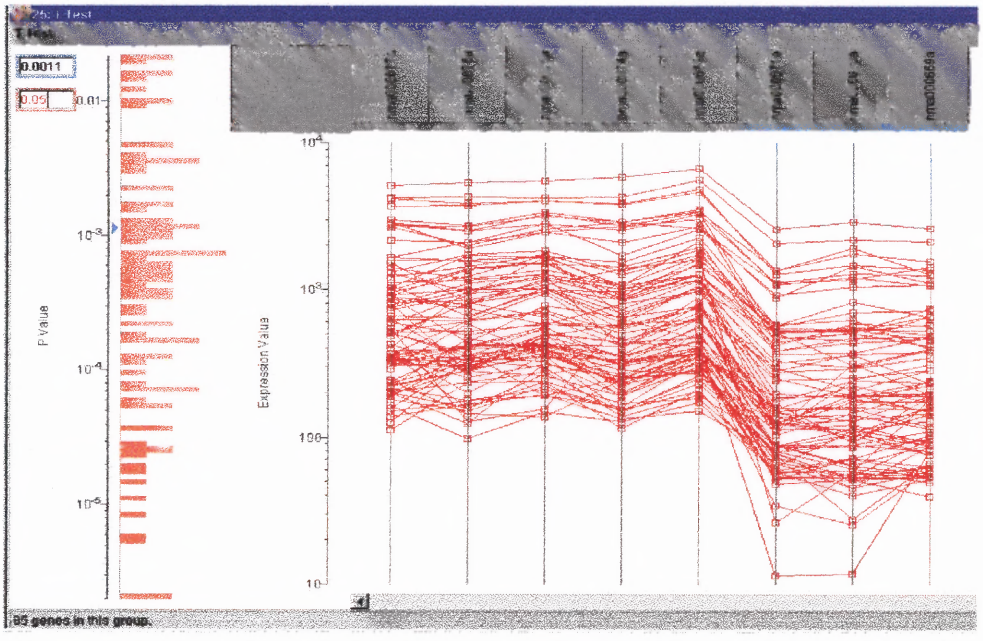

Figure 4.1 Screenshot of expression profiles of genes during a t-test in Expressionist™ 3.1.

6. The annotation of gene lists produced from GeneSpring™ 4.2 consists of different attributes as gene description, keywords, common name, different accession numbers and map positions on the chromosome. These can be changed in accordance to user preferences through the "Edit" drop-down menu. In Expressionist™ 3.1, the annotation consists of the accession number and gene description, based on the information from the local database. The format can only be changed in the Generic Data version after transfer of the file in a specified file type.
7. When details of a particular gene are required, the expression profile is selected to bring up a gene inspector window. In Expressionist™ 3.1, this window relates to information about the gene's expression levels in the different comparison groups and associated value of the analysis. In GeneSpring™ 4.2, besides the expression levels and associated value, the option of finding similar genes (similar in terms of correlation coefficients) is also present. Besides this, in both tools, links to other gene information is present. In the case of Expressionist™ 3.1, this consists of BioBench SRS, Entrez/NCBI, NetAffix, LocusLink and KEGG. In GeneSpring™ 4.2, this consists of Genbank, Gencards, Unigene, LocusLink, DDBJ, TIGR-TC and PubMed.
8. In GeneSpring™ 4.2, there is no method of determining a global mean or global median to be used as a reference value for normalization settings. This is not the case in Expressionist™3.1, where a mean or median value can be determined from a statistical representation of a histogram or boxplot. This value can then be used as a reference value for normalization.
9. The user interface to organize gene lists, experiment lists, gene trees, experiment trees, classifications, drawn genes and pathways into folders and sub-folders

according to personal requirements in GeneSpring™ 4.2 is based on a proper hierarchical tree structure. In Expressionist™3.1, this is just by a listing of the different gene lists or experimental groups.

10. The drop down menu for Experiments present in GeneSpring™ 4.2 allows the user to "Duplicate Experiments", "Merge/Split Experiments", "Change Experiment Parameters", "Change Experiment Interpretation" and Specifics of Global Error Model. In Expressionist™3.1, there is no option to "Merge Experiments"; the only method to perform this operation is to reload all samples. There is a way to split experiments in smaller experimental groups. There is no option for changing the interpretation of an experiment, i.e., to specify if groups of samples are replicates, continuous or non-continuous samples. The lists of experiments are linearly arranged as per order of creation or upload.
11. The physical position of a gene on the chromosome can be viewed in certain genomes in GeneSpring™ 4.2; there is no such visualization in Expressionist™3.1.

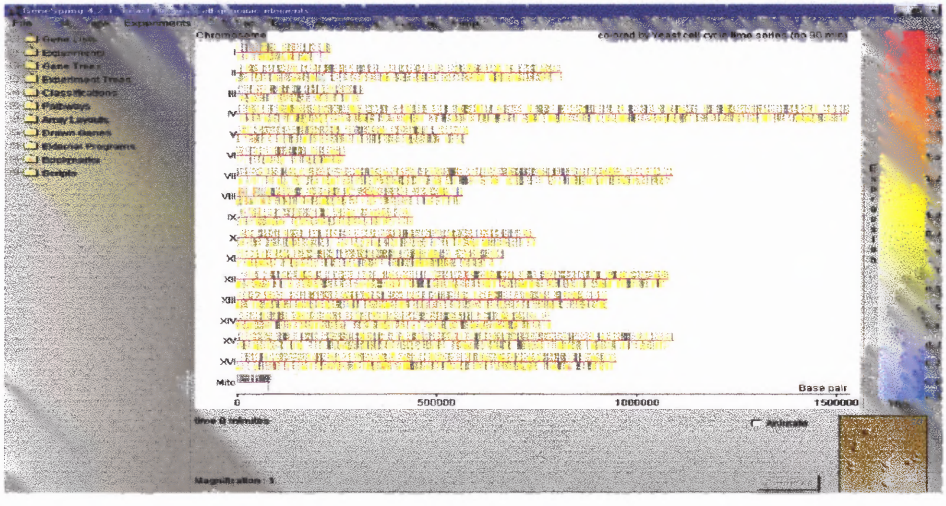

Figure 4.2 Screenshot of physical position of a gene on the chromosome in GeneSpring™ 4.2.

The choices of the different significance analysis in both tools are summarized below:

Table 4.1 Comparison of Statistical Tests of the Two Tools

Feature	Expressionist™ 3.1	GeneSpring™ 4.2
Parametric Test	1. Analysis of i) Genes or ii) Experiments 2.Number Of Best Scores (Default = 20) 3. Valid values per group (Default = 50% for genes and 20% for experiments)	1. Assuming equal variances (Student's T-test/ANOVA). 2.Not assuming equal variances (Welch's T test/ Welch ANOVA) 3. Global error model variances
Non-Parametric Test	Rank Test- 1. Wilcoxon test 2. Kruskal-Wallis test	Wilcoxon-Mann-Whitney test / Kruskal–Wallis test: application automatically selects the suitable of the two
Multiple Testing Correction	No Choice	1. Individual (genewise error rate) 2. Family-wise error rate 3. False Discovery Rate

Expressionist™ 3.1:

Parametric Test

This is also known as the 2 Groups (Parametric Test) [UM] – t-test, where the name also indicates the number of groups that are used for the analysis. This analysis uses the student's t-test assuming equal variances and also a normal distribution. It measures the extent to which the means of two groups are statistically different from each other. The analysis is appropriate to compare the means of two groups relative to the spread or variability of their values [3].

The results of this test allow the detection of genes that possess a maximum variation between the two groups, but a minimum variation between the genes within the same group. The increase in replicates per experiment group increases the efficiency of the result [3]. The different options within this analysis are:

1. Analysis of i) Genes ii) Experiments.

2. Number Of Best Scores (Default = 20): This specifies the number of items initially contained within the best scoring profile group.
3. Valid values per group (Default = 50% for genes and 20% for experiments): This specifies a threshold for a gene to be included in the analysis; it must have a valid value in at least the specified percentage of the genes (or experiments) that are being analyzed. A value is valid when it satisfies the quality settings (of Present, Absent or Marginal Absolute Calls).

Non-parametric test

Rank Test: This is also known as the 2 Groups (Rank Test) [UM] – Wilcoxon Test, an analysis that identifies genes with an expression level in one group which are statistically different from the other group, based on the ranks of the values and is not based on a normal distribution. Thus this test does not take into consideration the expression levels of the genes or experiment. Hence, it has the advantage of not being dependent on the normality of the data set distribution. This is efficient when used for very few numbers of replicates per group [3].

K Groups: As the name suggests, this analysis is used for more than two groups of experiments. The default test for this is the Kruskal-Wallis test. The k groups analysis identifies genes with an expression level that is statistically different in at least one of the selected groups, based on the ranks of the values. Hence, like the Wilcoxon test, this has the advantage of not being dependent on the normality of the data set distribution [3].

GeneSpring™ 4.2:

This tool has the options of different Multiple Testing Corrections within each group of significance analysis tests.

Parametric Test

The different options of the parametric test are:

1. Assuming equal variances (Student's t- test/ANOVA).
2. Not assuming equal variances (Welch's t-test/ Welch ANOVA).
3. Global error model variances: This is a two-component error model, which estimates measurement precision by combining variability of all genes. The global error model accounts for two types of error associated with microarrays, namely, measurement variation and between-samples variation [4]. The two variations to model are based on replicates and on deviation from 1 (in case of no replicates).

Non-parametric tests

Wilcoxon-Mann-Whitney test / Kruskal–Wallis test: There is no option to choose from the different methods, since the application automatically selects the most suitable method of statistical analysis based on the number of groups to be analyzed.

Multiple Testing Correction [4]:

This option is available only in GeneSpring™ 4.2 and is used is to adjust individual p-values to account for multiple comparisons and keep error rates less than or equal to user defined p cut-off values. There are different Multiple Testing Corrections available based

on the error rate, as individual error rate, family-wise error rate and false discovery rate [4]. These are classified along with the respective options to choose from as follows:

1. Individual (gene wise error rate):

   Selection: None

2. Family-wise error rate:

   Selection: Bonferroni / Bonferroni Step- down (Holm) / Westfall and Young Permutation

3. False Discovery Rate:

   Selection: Benjamini and Hochberg False Discovery Rate.

The Default is Benjamini and Hochberg False Discovery Rate for multiple testing correction with a p cut-off value of 0.05.

Listed below are some of the options in Expressionist™ 3.1 and GeneSpring™4.2:

Table 4.2 Normalization Methods in the Two Tools

	Expressionist™ 3.1	**GeneSpring™ 4.2**
1.	**Normalization** (i) None (ii) Logarithmic mean (iii) Arithmetic mean (iv) Median	**Per Spot** (i) Yes (ii) No
2.	**Reference Experiment:** To Be Specified	**Per Chip** (i) Positive control genes (ii) Median (iii) Constant values (iv) No Per Sample Normalization
3.	**Reference Value:** To Be Specified	**Per gene** (i) Median (ii) Particular Sample (iii) No Per Gene Normalization

Expressionist™ 3.1:

1. Normalization:
    a) None
    b) Logarithmic mean
    c) Arithmetic mean
    d) Median
2. Reference Experiment
3. Reference Value

The reference value takes precedence over other options of reference experiment and normalization method. This implies that if the user specifies a reference value and a reference experiment or normalization method is also specified, then the application automatically considers only the reference value and uses that as the normalization value. In order to determine a specific value, e.g. the global mean or median of all genes on all chips, the following method can be used to determine the reference value:

A histogram analysis of the experimental groups is performed with the "Include Summary" option selected. From the histogram, the "Info" option is selected and in red, appears the global mean and median of the genes and chips under study. This global mean or global median can be used as the reference value in the normalization settings.

GeneSpring™ 4.2:

1. Per Spot: This enables expression comparison on a relative scale.
2. Per Chip: This minimizes the differences from sample or array.
    i. Positive control genes: mostly "Housekeeping or spiked genes".
    ii. Median: by the 50th Percentile, assuming overall similar expression profile.
    iii. Constant values: for pre-normalized input data. This could be a custom defined value.
    iv. No per sample normalization.

Option: To use background correction or not, what values to use and which genes to use (absent, present, marginal).

3. Per gene: This enables expression comparison on a relative scale.
    i. Median: mostly used in the absence of a control sample.
    ii. Particular sample: When desired to use a designated sample.
    iii. No per gene normalization.

If experiments have been merged or split from a group of samples, then the option of starting with normalized values (from the original experiment) or of starting with the raw values can be made. In the latter case, the normalization can be made according to the above options.

4.1.1 The Learning Curve for a Biologist

The application in both tools is slightly different and hence the learning process also differs for both tools. This is a major factor for biologists who are entering the world of analyzing their own data, especially for the first time. Comprehension of various statistical terms and an understanding of the functioning of the algorithms that are responsible for the different analyses would take some time for a hard-core biologist. From this point of view, a user-friendly interface as well as accompaniment of sufficient manuals and guidance is necessary. Expressionist™ 3.1 has online documentation for registered users of the tool, whereas GeneSpring™ 4.2 has online documentation for anyone who requests permission from the vendor. Thus, prospective users can familiarize themselves with the tool before purchasing the license. GeneData, the vendor for Expressionist™ 3.1 provides technical support primarily through email correspondence, whereas SiliconGenetics, the vendor for GeneSpring™ 4.2 provides technical support primarily with a toll free technical support phone number as well as e-mail correspondence. The documentation and manuals for GeneSpring™ 4.2 provide detailed information and instructions on how to use the interface as well as the description of expression analysis studies and explanation of the relevant terms present as options in the tool. The user manual of GeneSpring™ 4.2 was more user-friendly and easier to learn. It was properly organized for an audience of first time users of microarray data analysis tools.

4.2 Implementation of Basic Analysis

The true comparative analysis of both the tools in this study can be brought about by implementation of the above statistical analysis on a set of data, details of which were specified in 1.3. The goal was to apply the same methods of statistical analysis in both tools and compare the final results. Ideally, the gene lists produced should be comparable with a minimum number of discrepancies, and those should be explainable.

The first step of the analysis involves the 'Import of Data' into the application packages. In this analysis, since both the tools being compared are commercial packages and hence are linked to the proprietary database of the affiliated company, the data is imported from the in-house databases. Expressionist™ 3.1 uploads the data from Affymetrix LIMS™ or MAS ™ and runs it through GeneData's Proprietary Data Quality Assurance Module [3]. In the case of GeneSpring™ 4.2, the data is uploaded from a file that is run through the proprietary database of the company that possesses the license for the tool. The data could also be imported in any other form; the autoloader feature recognizes many common formats of flat files originated directly from the gene chips [4]. In such cases, then formatting would be the next step to fulfill the requirements of the proper format for import of the data in accordance to the formatting guidelines.

The next step of the analysis involves 'Quality Control' of the chips as well as the genes. The purpose of the quality control step entails the removal of chips or genes that do not meet the threshold for a comparable analysis. In other words, it aids in the removal of poor quality samples. MAS™ 4.0 was the tool used for the quality control. The processes of 'Filter of genes' and 'Quality control' are supplemental to each other in this study.

This can be done in a variety of ways:

1. Direct implementation of basic statistical analysis and inference from the global results.
2. Global analysis by clustering, scatter plots, prediction of parameter values or condition inspector. Of these, the latter two features are available only in GeneSpring™ 4.2. Methods such as global representation in the form of box plots can be made, which is a feature present only in Expressionist™ 3.1.

The approach used in this study is the direct implementation of the basic statistical analysis and global analysis (by clustering and box plot representations) using both the tools. The gene expression analysis of the data set starts from a global representation in all six tissues of the paw connective tissue, liver, lung, skin, blood and kidney. The results of the quality control by direct implementation of basic statistical analysis is described only for Expressionist™ 3.1, whereas results of validation by global analysis by means of clustering and box plots is described for both tools.

4.2.1 Comparative Analysis

The first step of the quality control implementation involves normalization of the data. The normalization method used in this analysis is based on the global median of all the samples and genes under study. In Expressionist™ 3.1, this is performed as described in section 4.1 by determination of a specific reference value (in this case, the median), while in GeneSpring™ 4.2 it is performed by making the following selections:

1. Per Spot: No per spot normalization.

2. Per Chip: Median, by the 50th Percentile, assuming overall similar expression profile.

3. Per Gene: Median, mostly used in the absence of a control sample.

The second step of the quality control implementation involves filtering of the data based on absolute calls, statistical filters of t-test and fold changes. The criteria for choosing the appropriate measurement calls for all analyses is based on the fact that "Only Present calls" maybe more highly meaningful. Thus, in Expressionist™ 3.1, this is selected from the drop down menu of "Quality Settings", whereas in GeneSpring™ 4.2, it is selected by the drop down menu in "Change Experiment Interpretations" of the Experiment menu. The next step involves the implementation of different filters for fold changes, expression levels and statistical analysis. The fold changes for this analysis are taken as greater than or equal to two-fold change in expression from control to treatment; this includes induced as well as repressed genes. The criteria of greater than two fold is taken since there is a great variability in gene expressions, and fluctuations of small magnitude almost always suggest random changes as opposed to consistent change in gene expression. In Expressionist™ 3.1, this brought out by the "N-Fold Regulation" (in either the "Log-log plot" or "Profile representation") by selecting all genes that are differentially expressed by a factor greater than or equal to two. In GeneSpring™ 4.2 this is done by selecting "Add Filter on fold change" from the Tools menu and right clicking on the 'control' or 'treatment' interpretation of the experiment. For present purposes, expression levels were arbitrarily required as greater than or equal to 100 in the comparable group. This was done in Expressionist™ 3.1 by performing the "Filter by Average Expression" analysis, whereas in GeneSpring™ 4.2 it was performed by the "Expression Level Restriction" option from the Tools menu. The statistical group

comparisons used involve the parametric test employing the student's t-test that is the only option among parametric tests in Expressionist™ 3.1. The same test is implemented in GeneSpring™ 4.2 with the unequal variances option. The genes obtained after implementation of all the above analyses are said to be the genes that show significantly different expression between one condition and another, which in this case, is between the control and treatment groups.

The results of the above analysis are depicted in the following visual representations and tables:

Table 4.3 Number of Genes in All Tissues That Change in Expression Level according to Expressionist™ 3.1

Tissue	Number of Days	2-fold change, non-significant	2- fold change, significant
Paw	7 day	386	339
	21 day	42	11
Liver	7 day	23	7
	21 day	83	27
Lung	7 day	139	35
	21 day	76	32
Blood	7 day	29	7
	21 day	43	20
Skin	7 day	49	19
	21 day	142	66
Kidney	7 day	39	17
	21 day	61	21

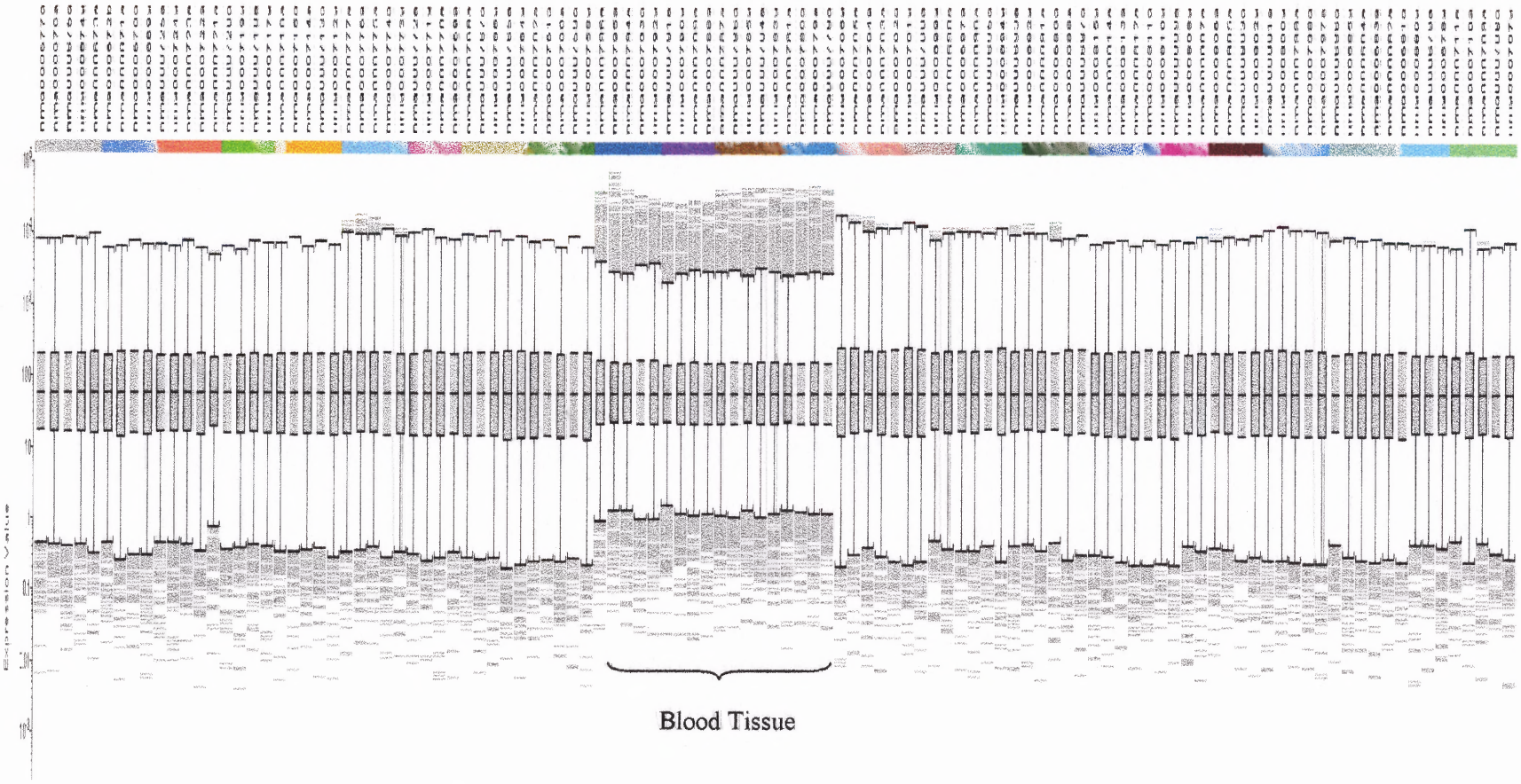

Figure 4.3 The Box plots of experiments to assure the medians are aligned together in Expressionist™ 3.1 (after normalization to 58.45).

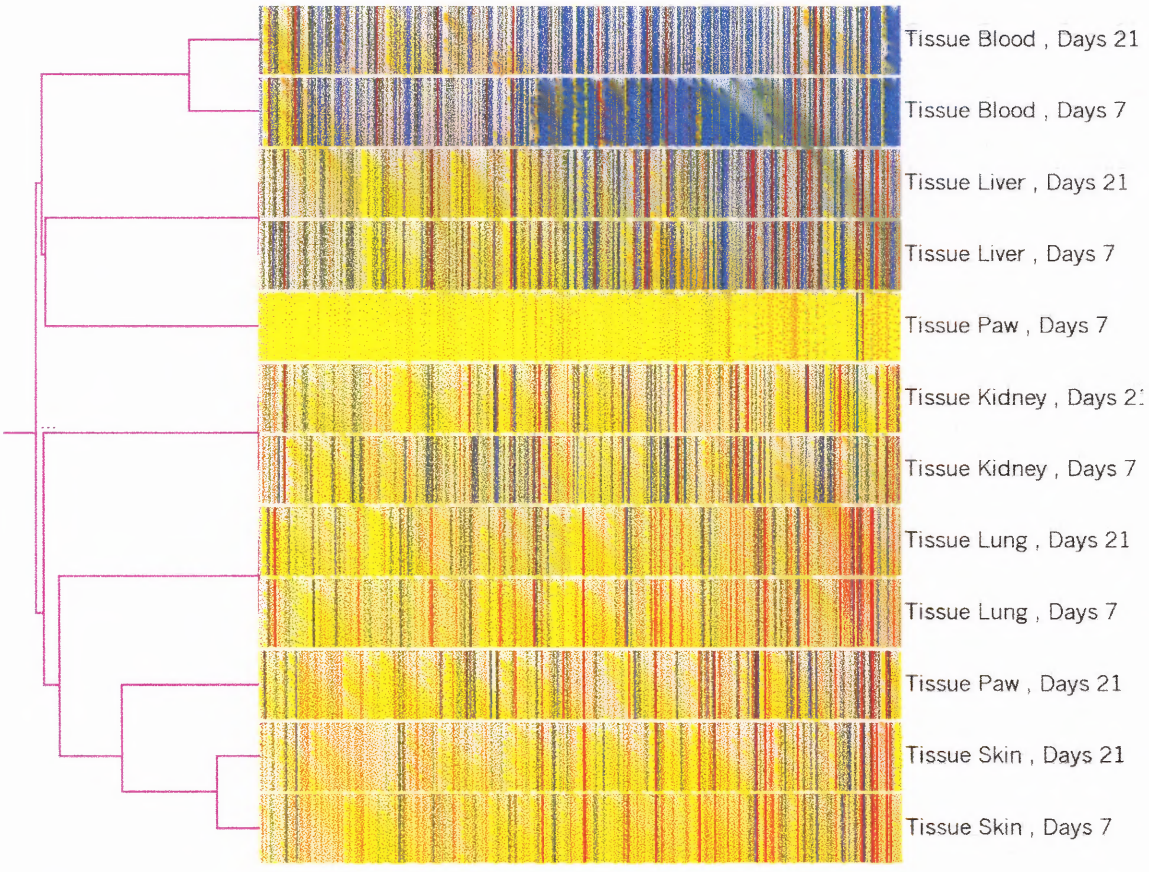

Figure 4.4 Experiment tree of all six tissues in GeneSpring™ 4.2.
(Expression level >100 in 6 out of 12 conditions)

This type of an analysis can also be performed in Expressionist™ 3.1 with the criteria of hierarchical clustering of experiments.

The above cluster representation suggests that the behaviour of genes in the blood tissue is very different from the rest of the tissues. The basic analysis of filter of genes of Expressionist™ 3.1 in Table 4.1 depicted a meager number of genes changing expression. Also, the box-plot representation in Expressionist™ 3.1 in Figure 4.3 shows how the blood tissue was not aligned with the remaining tissues. This is also visible in

hierarchical clustering of individual tissues as depicted in the Appendix by observing separate clusters for the control and treatment groups. Thus, different types of analyses in different tools suggested the same observation, hence validating the consistency in both tools to show similar results. These are global representations of the analysis. This does not, however, provide much information about individual genes.

Table 4.4 Number of Genes Differentially Expressed in Expressionist™ 3.1 (Expression levels of either control or treatment >100 and >2 fold change)

Time point	Direction of Regulation	Paw	Liver	Lung	Kidney	Skin
7	Up	109	2	24	3	1
	Down	135	0	13	1	1
21	Up	1	9	8	6	13
	Down	2	6	15	7	33

The above tables and figure demonstrate how the seven-day data of the paw is of more interest, based on the number of differentially expressed genes, both induced as well as repressed. Teasing apart the different clusters of each tissue and studying the relation between the control and treatment groups further validated this. The clusters of individual tissues in Appendix A show that the paw connective tissue is the only tissue seen to possess separate clusters for the control and treatment conditions. Thus, arbitrarily, the seven-day, paw connective tissue is used as a sample data set to compare the two tools.

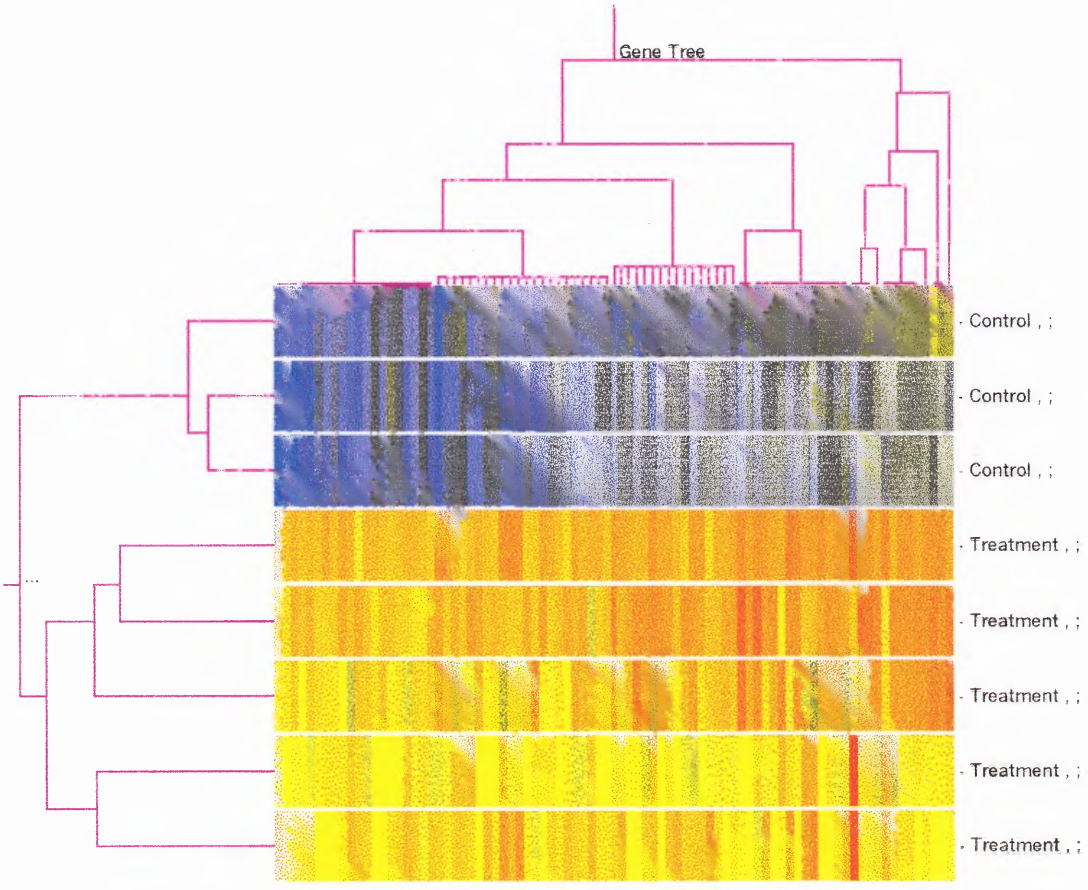

Figure 4.5 Experiment tree and gene tree of paw tissue at seven days in GeneSpring™ 4.2. (This shows genes up-regulated greater than or equal to two fold change in expression)

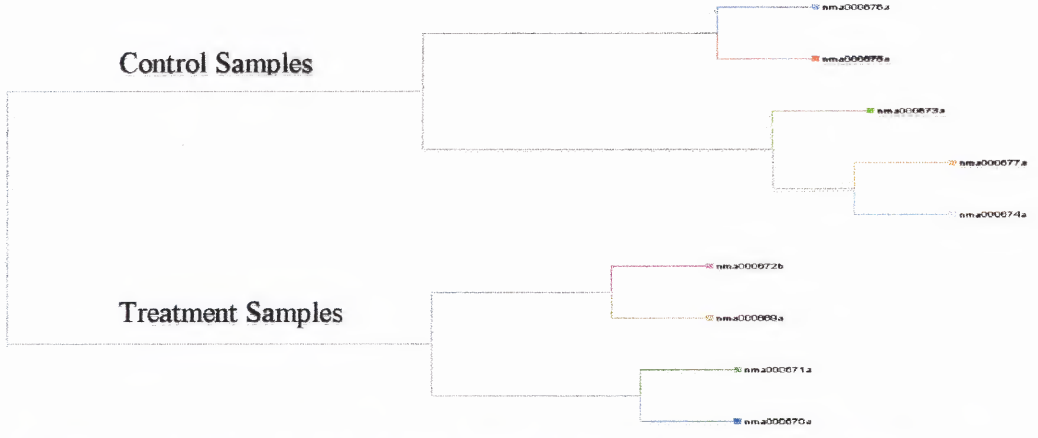

Figure 4.6 Experiment tree of paw tissue at seven days in Expressionist™ 3.1.

Among genes up-regulated in paw tissue, few or none were also up-regulated in other tissues at any time point. This was determined in Expressionist™ 3.1 by creating gene lists in all tissues, which consisted of genes up-regulated greater than two-fold. These gene lists of different tissues (but same time points) were compared and the following intersections were seen:

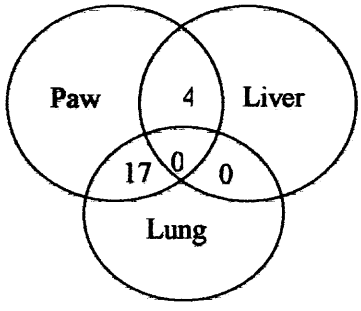

Figure 4.7 Intersections of gene lists from different tissues.

The remaining tissues were not represented in this Venn diagram since there were no other genes that were common with any other tissue. This type of Venn diagram analysis cannot be pictorially represented in Expressionist™ 3.1. GeneSpring™ 4.2 can generate perform Venn diagrams for up to three gene lists at a time. However, in Expressionist™ 3.1, the user can intersect an infinite number of gene lists at a time. After implementation of basic statistical analysis of genes up-regulated in paw connective tissue, a list of genes was prepared and sorted in descending order of fold changes and the following statistics regarding the two tools were gathered:

Table 4.5 Statistics of Comparison of the Two Tools After the Basic Statistical Analysis

Total Number in Expressionist™ 3.1 =111
Total Number in GeneSpring™ 4.2 =106

	Genes retaining their ranks in both tools	Number of genes present in Expressionist™ 3.1 but absent in GeneSpring™ 4.2	Number of genes present in GeneSpring™ 4.2 but absent in Expressionist™ 3.1	Number of genes present in both
Number of genes	2	26	21	85
As a % of total in GeneSpring™ 4.2	1.9%	24.5%	19.8%	80.2%
As a % of total in Expressionist™ 3.1	1.8%	23.4%	18.9%	76.6%

Table 4.6 The Two Genes that Retained Their Rank According to the Fold Changes in Each Tool

				GeneSpring™ 4.2		Expressionist™ 3.1	
Rank	Systematic Name	Common Name	Keywords	Fold Change	P-value	Fold Change	P Value
1	rc_AA893846_at		EST	20.4	0.00035	24.4	0.00037
18	L32132_at	Lbp	lipopolysacc-haride binding protein	4.8	0.00061	7.3	0.00068

4.2.2 A Desired Comparison

As could be seen from Table 4.5 above, approximately 85% of the genes satisfy the following selection and are present in both tools:

1. Up-regulation from control to treatment by greater than or equal to two fold.
2. Expression level of treatment group greater than or equal to 100.
3. p-value of compared groups less than or equal to 0.05.

This implies that the tools have different methods of determining some statistical tests. The following analysis is an attempt towards accounting for the discrepancy in the final list of genes that satisfy the above criteria. Thus, the two sets of genes are taken that are present in one tool, but not present in the other, and further analysed in both tools. The values in red, account for the difference. This may be because one of the above criterions may not be fulfilled by one tool. The results obtained are showed on the next page:

Table 4.7 The 21 Genes Present in GeneSpring™ 4.2, but Absent in Expressionist™3.1

			Raw Values				Fold Changes		Absolute Measurement Calls					
			Exp.	GS	Exp.	GS			Expressionist™3.1		GeneSpring™ 4.2		Exp.	GS
	Keywords	Systematic Name	Control	Control	Treatment	Treatment	Exp.	GS	Control	Treatment	Control	Treatment	P-value	P-value
1		D86041_at	37.65	37.67	106.0	106.0	2.8	2.8	A	P	P	P	-	0.00030
2	PS-PLA1	D88666_at	27.07	27.07	252.8	252.8	9.3	9.3	A	P	P	P	-	0.00174
3	PMP70	D90038_at	60.31	60.33	147.3	147.3	2.4	2.4	P	P	P	P	0.01355	0.00025
4	T-cell receptor	M18854_at	48.69	52.85	137.5	142.9	2.8	2.7	A	P	P	P	-	0.01333
5	EST	rc_AA851381_at	41.57	120.37	111.7	284.2	2.7	2.4	P	P	P	P	0.03075	0.00879
6	EST	rc_AA860057_at	18.58	14.97	104.1	125.4	5.6	8.4	A	P	P	P	-	0.00726
7	EST	rc_AA874990_at	38.37	45.07	176.5	108.5	4.6	2.4	A	P	P	P	-	0.03706
8	EST	rc_AA892849_at	43.27	54.87	111.6	118.7	2.6	2.2	A	P	P	P	-	0.03747
9	EST	rc_AI639338_at	16.45	42.63	140.6	103.6	8.5	2.4	A	P	P	P	-	0.00719
10	EST	rc_AI639365_at	20.71	45.40	106.9	105.8	5.2	2.3	A	P	P	P	-	0.00034
11	EST	rc_AI639401_at	29.75	14.50	199.4	209.3	6.7	14.4	A	P	P	P	-	0.02269
12	EST	rc_H31232_at	44.42	80.97	128.5	244.6	2.9	3.0	P	P	P	P	0.217	0.04812
13		S75435_i_at	15.56	41.57	112.5	111.7	7.2	2.7	A	P	P	P	-	0.00001
14		U09401_s_at	120.37	19.90	206.4	104.0	1.7	5.2	A	P	P	P	-	0.00513
15		U15550_at	14.97	38.37	125.4	176.5	8.4	4.6	A	P	P	P	-	0.00000
16		U31599_g_at	45.08	43.23	108.7	111.6	2.4	2.6	A	P	P	P	-	0.03992
17		U76714_at	54.89	21.25	118.7	140.5	2.2	6.6	A	P	P	P	-	0.00353
18		U76714_g_at	42.65	23.75	103.6	106.9	2.4	4.5	A	P	P	P	-	0.00113
19	insulin-like GF 1	X06107_i_at	45.38	32.60	110.6	186.9	2.4	5.7	A	P	P	P	-	0.03110
20	Fc-gamma receptor	X73371_at	11.02	44.43	209.3	128.5	19.0	2.9	A	P	P	P	-	0.01517
21	GH releasing hormone	Z34004exon_g_at	80.99	21.90	159.5	112.5	2.0	5.1	A	P	P	P	-	0.00174

Exp.= Expressionist™ 3.1
GS = GeneSpring™ 4.2

From Table 4.7 above, the p-values for the analysis in Expressionist™ 3.1 are missing for a majority of the genes. The reason for this absence can be attributed to the fact that once the absolute calls for the genes is determined to be 'Absent', they are automatically thrown out of the analysis, and hence the p-values for only 3 genes are shown in the table. These 3 genes have an absolute call as 'Present' and hence were a part of the analysis.

From a total of 21 genes, there were two genes that were not accounted for, i.e., they appeared to satisfy all the criteria for selection, yet were filtered from the analysis in Expressionist™ 3.1. These two genes were PMP70 (D90038_at) and an EST (rc_AA851381_at). The raw expression levels of the former gene are similar in both tools, but for the latter, even though the fold changes were comparable, yet there was a large discrepancy in the original raw expression values. This result cannot be accounted for at this point in the study.

The next table depicts the 26 genes that were present in Expressionist™ 3.1, but absent in GeneSpring™ 4.2.

Table 4.8 The 26 Genes Present in Expressionist™ 3.1, but Absent in GeneSpring™4.2

		Raw Values				Fold Changes		Absolute Measurement Calls					
		Exp.	GS	Exp.	GS			Expressionist™ 3.1		GeneSpring™ 4.2		Exp.	GS
	Systematic Name	Control	Control	Treatment	Treatment	Exp.	GS	Control	Treatment	Control	Treatment	P-value	P-value
1	AB012234_at	355.2	213.7	556.0	409.5	1.6	1.9	P	P	P	P	0.01095	0.00433
2	AF050214_at	238.5	234.2	443.7	448.3	1.9	1.9	P	P	P	P	0.00046	0.02086
3	AJ005394_at	449.4	439.5	877.2	886.0	2.0	2.0	P	P	P	P	0.00415	0.00651
4	AJ223355_g_at	83.9	82.3	154.3	156.4	1.8	1.9	P	P	P	P	0.00222	0.00068
5	D00753_at	76.9	75.1	140.1	141.4	1.8	1.9	P	P	P	P	0.00400	0.06429
6	L00191cds#1_s_at	1890.7	1849.5	3684.4	3717.3	1.9	2.0	P	P	P	P	0.00013	0.01196
7	L02529_at	191.9	188.4	372.4	376.9	1.9	2.0	P	P	P	P	0.00010	0.00090
8	M15562_g_at	932.0	914.1	1815.0	1834.3	1.9	2.0	P	P	P	P	0.01156	0.00522
9	M31837_at	116.8	115.0	221.6	224.0	1.9	1.9	P	P	P	P	0.00688	0.05770
10	M83678_at	223.9	219.4	435.6	438.6	1.9	2.0	P	P	P	P	0.00035	0.00021
11	rc_AA799340_at	1446.3	1418.6	2822.2	2843.6	2.0	2.0	P	P	P	P	0.00004	0.00001
12	rc_AA800844_s_at	789.0	773.3	1552.6	1564.6	2.0	2.0	P	P	P	P	0.00007	0.00011
13	rc_AA859757_at	302.5	296.1	593.0	600.3	2.0	2.0	P	P	P	P	0.00734	0.00337
14	rc_AA874848_s_at	513.3	502.6	969.0	980.1	1.9	2.0	P	P	P	P	0.01626	0.00879
15	rc_AA875023_at	132.2	129.6	249.0	251.1	1.9	1.9	P	P	P	P	0.00009	0.00004
16	rc_AA891204_s_at	1466.7	1440.4	2845.2	2871.0	1.9	2.0	P	P	P	P	0.00051	0.00950
17	rc_AA893702_s_at	228.4	223.9	409.6	441.5	1.8	2.0	P	P	P	P	0.01336	0.01676
18	rc_AA944422_at	475.0	465.5	898.9	903.7	1.9	1.9	P	P	P	P	0.00303	0.00214
19	rc_AI234060_s_at	281.2	275.7	527.8	533.1	1.9	1.9	P	P	P	P	0.00026	0.00006
20	S54008_i_at	299.7	293.9	557.4	561.8	1.9	1.9	P	P	P	P	0.00087	0.00029
21	U35776_at	142.9	140.2	269.4	272.9	1.9	1.9	P	P	P	P	0.00038	0.00755
22	X04979_at	4062.3	3978.3	7931.2	8008.0	2.0	2.0	P	P	P	P	0.00002	0.00009
23	X13044_g_at	1082.2	1060.7	2120.2	2143.9	2.0	2.0	P	P	P	P	0.00919	0.00408
24	X15512_at	92.1	90.3	179.5	182.2	1.9	2.0	P	P	P	P	0.00243	0.00093
25	X65454_g_at	107.8	105.8	217.0	218.9	2.0	2.1	P	P	P	P	0.00058	0.00032
26	X72914_at	831.7	815.4	1665.8	1680.6	2.0	2.1	P	P	P	P	0.00002	0.00001

Exp.= Expressionist™ 3.1
GS = GeneSpring™ 4.2

As can be seen from Table 4.8 above, there are a total of 16 genes, for which the discrepancies were not accounted for, i.e., they seem to satisfy all the criteria, yet were not present in the gene lists produced from both tools. This may imply that there was some other criterion by which the genes were filtered out of the analysis. Ideally, the fold changes are calculated from the normalized data. In this case, since the normalized values for each signal of gene expression were not accessible from Expressionist™ 3.1, a comparison of fold changes calculated from raw and normalized values were analyzed only in GeneSpring™ 4.2.

Table 4.9 Fold Changes in GeneSpring™ 4.2

	Systematic Name of Gene	Fold Change Calculated from Raw Values	Fold Change Calculated from Normalised values
1	AB012234_g_at	1.92	1.83
2	AF050214_at	1.91	1.86
3	AJ005394_at	2.02	1.96
4	AJ223355_g_at	1.90	1.82
5	D00753_at	1.88	1.89
6	L00191cds#1_s_at	2.01	1.97
7	L02529_at	2.00	1.94
8	M15562_g_at	2.01	1.92
9	M31837_at	1.95	1.93
10	M83678_at	2.00	1.93
11	Rc_AA799340_at	2.00	1.94
12	rc_AA800844_s_at	2.02	1.95
13	Rc_AA859757_at	2.03	1.94
14	rc_AA874848_s_at	1.95	1.86
15	Rc_AA875023_at	1.94	1.87
16	rc_AA891204_s_at	1.99	1.94
17	rc_AA893702_s_at	1.97	1.87
18	Rc_AA944422_at	1.94	1.85
19	rc_AI234060_s_at	1.93	1.87
20	S54008_i_at	1.91	1.84
21	U35776_at	1.95	1.89
22	X04979_at	2.01	1.95
23	X13044_g_at	2.02	1.91
24	X15512_at	2.02	1.92
25	X65454_g_at	2.07	1.99
26	X72914_at	2.06	1.99

This table accounts for filtration of all 26 genes during the analysis in GeneSpring™ 4.2, since the fold changes are less than two fold. This validates the fact that the normalized values are used to calculate fold changes in GeneSpring™ 4.2. The desired way to perform a comparison of this type would be to have some sort of comparison tool installed in these applications, where the raw data tables could be compared before any analysis is done. The embedded tool could be an analogy to MS Access that can compare two tables containing raw data for a full array of genes.

Once it is ensured that the same set of genes are being used for the analysis, then the same set of tests should be applied to the data set in both data mining tools. For example, GeneSpring™ 4.2 has the option to chose a Multiple Testing Correction, which is used to adjust individual p-values to account for multiple comparisons and keep error rates less than or equal to user defined p cut-off values. Since this option is not present in Expressionist™ 3.1, a comparative analysis of the precise nature and power of statistical significance tests of the two tools cannot be made.

Ideally, if the gene lists from one tool can be conveniently transported to the other, then the comparison of gene outliers and other variable conditions can be made. All such genes, which are commonly found to be outliers in both tools, can be set-aside during the preliminary analysis. Also, a comparison of genes after each step of analysis would be the ideal way to determine the significant genes that are obtained in all tools used for the analysis. A validation of the result by another tool at each step of the analysis ensures the significance and accuracy of the change in gene expression of a gene of interest.

A visualization tool such as viewing the physical position of a gene on the DNA of the organism is feasible only in GeneSpring™ 4.2, and hence the exact location of a gene cannot be compared in the two tools in this study.

4.3 Comparative Analysis of the Clustering Methods

In this section, the different clustering analyses are compared in Expressionist™ 3.1 and GeneSpring™ 4.2. This involves implementation of the different measures of similarity and clustering methods that are available in current versions of the two tools. It should be noted that comprehensive software tools are frequently updated with newer versions and the availability of features will differ with from version to version.

The gene list used for the comparative analysis consists of the 85 genes that were found to satisfy the criterion of being up-regulated after treatment as compared with control. These genes were statistically significantly different with respect to expression levels (p-value <0.05); the fold change was greater than or equal to two and raw expression levels were more than 100.

Some other features of the two tools that are not included in this thesis include SOM clustering, Principal Components Analysis and Regulatory Sequence Search.

4.3.1 Side-by-Side Comparison of Available Features

All the clustering options of the tools were compared and an attempt was made to use the corresponding features in both tools, so as to obtain results that would hold the same significance.

Some clustering differences explained between Expressionist™ 3.1 and GeneSpring™4.2:

i. The user can specify the Separation Ratio and Minimum Distance in GeneSpring™ 4.2. These measures control the difference in correlation between the objects being analysed for the dendrogram (in hierarchical clustering). This option is not present in Expressionist™ 3.1.

ii. The Valid Value Percentage of the number of objects (i.e., genes or experiments) to be included in an analysis can be specified by the user in Expressionist™ 3.1, whereas it is set to a default value of 50% in GeneSpring™ 4.2.

iii. In Expressionist™ 3.1, the entire dendrogram cannot be viewed if the length of the annotations of the genes exceeds a threshold number of permissible characters. This occurs primarily when the Experiment Name and Description are too long. The process to circumvent this problem is to export the data into a flat file from Expressionist™ 3.1, shorten the annotation in MS Excel, and re-import the file with an extension of '.abs' in the Expressionist™ 3.1 Generic Data server. This transport of data has to be done with great caution in order to prevent any discrepancies. This is because there are attributes associated with each gene that should be re-imported with the exact information as derived from the original Gene chip; this mainly includes the Absolute Call Measurement and the type of data, i.e., normalised or raw data.

iv. In GeneSpring™ 4.2 manipulation of labels for dendrograms and graphs can take place through the option of "Change Experiment Parameters" from the "Experiments"" menu, and hence the labelling of all the data is based on the user and not the structure of the database. Also, the type of label based on the annotation of the

genes can be specified from the "Gene Label" option under the "Preferences" drop down of "Edit". There is no such option in Expressionist™ 3.1.

Table 4.10 Hierarchical Clustering Options of the Two Tools

	Feature	Expressionist™ 3.1	GeneSpring™ 4.2
1	**Cluster by**	Experiments or Genes	Experiments or Genes
2	**Distance Metrics**	Positive (Default), Correlation, L1, Euclidean, Normalized Euclidean, Maximum	Standard Correlation (Default), Smooth, Change, Up-regulated, Pearson, Spearman Confidence, Two-sided Confidence, Distance
3	**Linkage Method**	Average, Single or Complete	Average
4	**Separation Ratio**	Cannot be defined	User Specified; default=0.5
5	**Minimum Distance**	Cannot be defined	User Specified; default=0.001
6	**Automatic Annotation**	No Choice	Choice of (i) yes or (ii) no
7	**Annotate with Standard Lists**	No Choice	Choice of (i) yes or (ii) no
8	**Valid values to be used**	User Specified	Default of either (i) Half (ii) All conditions
9	**Ontology Construction Tool**	No Choice	Can Construct an Ontological Classification

The following table shows some of the differences in the k-means clustering options of both tools:

Table 4.11 k-means Clustering Options of the Two Tools

	Features	Expressionist™ 3.1	GeneSpring™ 4.2
1	**Number of Clusters**	User Specified	User Specified
2	**Maximum Iterations**	Cannot be Specified	User Specified (Default=100)
3	**Distance Metric**	Positive (Default), Correlation, L1, Euclidean, Normalized Euclidean, Maximum	Standard Correlation (Default), Smooth, Change, Up-regulated, Pearson, Spearman Confidence, Two-sided Confidence, Distance
4	**Valid values to be Used**	User Specified (Default = 20%)	No Such Choice
5	**Start From Current Classification**	No Such Choice	Can be selected
6	**Animate Display While Clustering**	No Such Choice	Choice of (i) yes or (ii) no
7	**Test Additional Random Starting Clusters**	No Such Choice	User specified Number
8	**Discard genes with No Data for Half the Conditions**	No Such Choice	Choice of (i) Yes or (ii) No

4.3.2 Clustering Comparison

An ideal clustering comparison would consist of clustering of genes in different tools using the exact same tests. For example, the only linkage method available in GeneSpring™ 4.2 is the average linkage method, whereas Expressionist™ 3.1 also has options for single and complete linkages. Thus, the accuracy of results produced by a single or complete linkage clustering cannot be cross-validated between the two tools to determine which tool is more precise.

Also, the measures of similarity play a major role in the clustering process; hence only if the measures of similarity are the same in both tools, is it possible to compare

results and analyze the tool. In order to tease apart some computational information about the hierarchical clusters, the separation ratio and minimum distance are necessary in order to obtain gene-to-gene relationship. Since these features are not present in Expressionist™ 3.1, this analysis is not feasible and the efficient tool cannot be determined based on this test.

Another feature that is distinct from Expressionist™ 3.1 is the ease of making gene lists for any type of analysis or classification. In GeneSpring™ 4.2 there is a tool known as the Annotation and Ontology Construction Tool, which can automatically annotate any gene with public database information using GeneSpider. The user can also construct an ontological classification of the genome based on biological process, molecular function, cellular component, etc. This produces gene lists based on the above criteria within seconds. The same process in Expressionist™ 3.1 may become tedious due to constructing individual gene lists per category of the classification and then making a comparison or further analysis based on that classification. The Annotation and Ontology Construction Tool of GeneSpring™ 4.2 allows a simultaneous study of expression patterns in biological categories of genes by simply browsing through them and can be cross-referenced to new lists of genes.

In the case of k-means clustering, there is no valid comparison between the two tools, since the method used to display the number of initial clusters and number of iterations is very different. The number of clusters (unless user specified as in GeneSpring™ 4.2) and the number of genes per cluster changes with every iteration. In Expressionist™ 3.1, the user has to perform the k-means cluster analysis for an individual iteration and hence it is a very tedious process to observe the number of iterations that

takes place for convergence to occur. In the case of GeneSpring™ 4.2, the user can observe the total number of iterations that take place for convergence to occur in one step, but in order to observe the variation in number of clusters and number of genes per cluster, the same tedious process of performing a k-means cluster analysis for an individual iteration has to be performed. This can be proven by the results of the analysis:

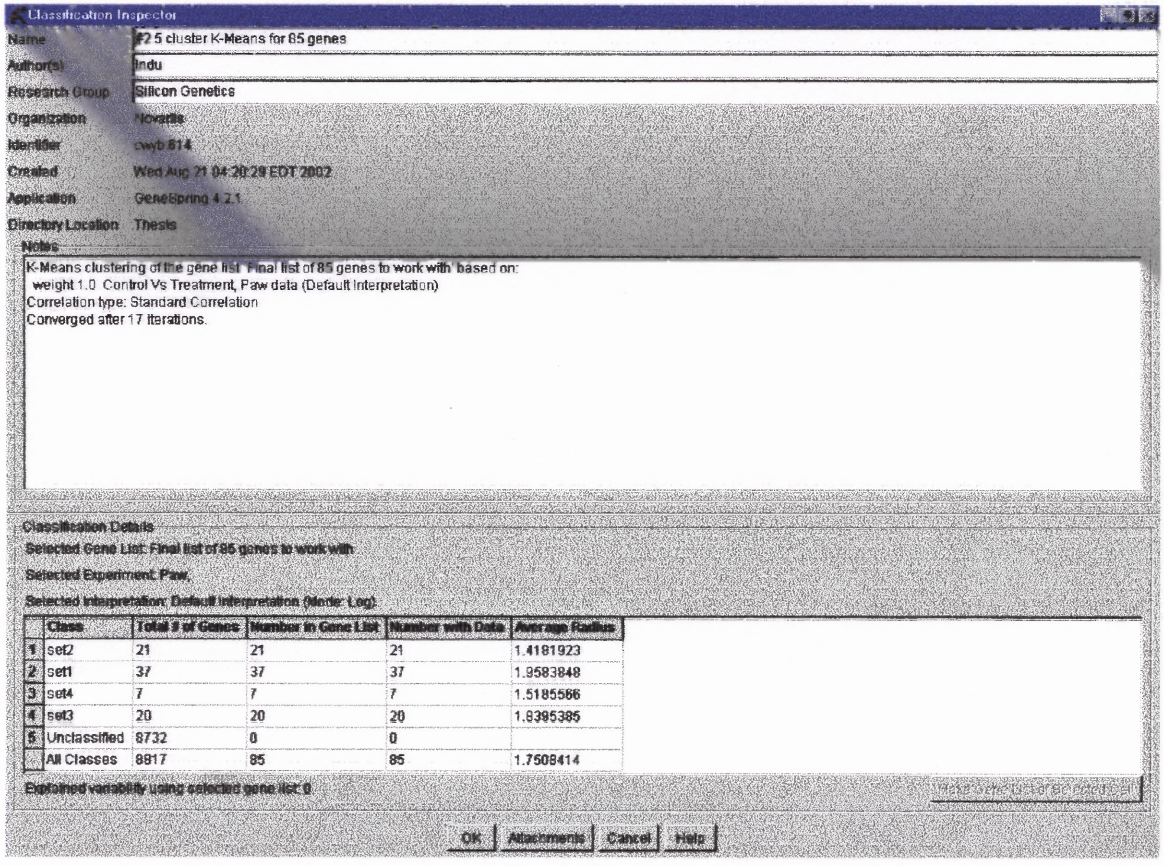

Figure 4.8 Screenshot of the Classification Inspector of GeneSpring™ 4.2.

Table 4.12 Variation in k-means Clusters in Expressionist™ 3.1

The Succession of Analysis	# of clusters	# in Cluster 1	Mean Distance	# in Cluster 2	Mean Distance	# in cluster 3	Mean Distance	# in Cluster 4	Mean Distance	# in Cluster 5	Mean Distance
1	4	4	0.03505	30	0.03363	29	0.023	22	0.02783	-	
2	5	16	0.0273	5	0.04123	37	0.02473	22	0.02657	5	0.01942
3	5	3	0.008616	5	0.03708	27	0.02369	18	0.02074	32	0.03421
4	3	22	0.0287	17	0.02361	46	0.04023	-		-	
5	5	36	0.02251	1	0	19	0.03339	25	0.02963	4	0.03505
6	5	4	0.03505	20	0.02808	29	0.023	31	0.0297	1	0
7	5	4	0.03505	23	0.02178	22	0.02988	23	0.02398	13	0.02347
8	4	20	0.02678	25	0.02963	6	0.04078	34	0.02468	-	
9	4	2	0.03339	4	0.03505	9	0.023	20	0.02808	-	
10	5	23	0.02972	26	0.02639	3	0.03792	31	0.02801	2	0.02051
11	3	20	0.02808	32	0.03337	33	0.03356	-		-	
12	4	14	0.02691	32	0.03337	27	0.02763	12	0.02555	-	
13	3	35	0.03393	5	0.04123	45	0.03136	-		-	
14	4	33	0.02467	24	0.02928	6	0.04078	22	0.02741	-	
15	5	31	0.0273	13	0.002578	22	0.02382	16	0.0273	3	0.03191
16	4	26	0.02755	17	0.04085	34	0.02606	8	0.02787	-	
17	5	4	0.03505	28	0.0292	32	0.02698	18	0.02218	3	0.008616
18	5	8	0.02305	17	0.02556	32	0.02336	24	0.02857	4	0.03505
19	5	37	0.02597	7	0.02523	18	0.02774	5	0.04123	18	0.02276
20	4	23	0.02717	5	0.04123	8	0.0256	49	0.0315	-	

CHAPTER 5

SUMMARY AND DISCUSSIONS

5.1 Summary of Comparative Analysis

As mentioned in Section 3.1, Expressionist™ 3.1 belongs to the turnkey type of software packages, whereas GeneSpring™ 4.2 is a type of comprehensive software [23]. This implies that because Expressionist™ 3.1 is customized for the application, the user would require a separate technical support team, who should be adept in the statistical as well as biological functioning and algorithmic translation of microarray experiments, so they can modify the analysis architecture by integrating their own analyses. Thus, the efficiency of the within-company technical support staff is critical for the quality of results produced. Genedata (vendor for Expressionist™ 3.1) can provide a number of value-adding services upon request. This includes features of integration with other software, either GeneData products or another existing software. GeneSpring™ 4.2 is a type of comprehensive software; hence it may not be able to accommodate new analyses developments, since the user would have to wait till the vendor releases a new version of the software with enhanced features.

From the biologist's perspective, the learning curve is shorter for GeneSpring™ 4.2 as explained in Section 4.11. The vendor also provides a free 30-day license for prospective users of the tool to test the efficiency and comfort level of the tool. GeneSpring™ 4.2 also generates various analyses easily [4]. The provision of free online telephone-web conference presentations of various topics also sets the foundation for new users of the tools. The tool can incorporate a scripting tool language for automated

process control; for example, a script may be written for calculating genes that are up-regulated by a factor of greater than two-fold and whose expression levels are above a threshold. Once a script has been written for such an operation, then the whole process is automated by selecting one test, which would perform the whole series of operations specified in the script.

Both the tools in this study are suitable primarily for commercial use and hence are not free. The data generated from these software packages may not be compatible with other contemporary software packages and hence the data need to be modified before being transported from one tool to another.

5.2 Development of Future Expression Analysis Tools

As John W. Tukey wrote, "Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone--as the first step." [23]. The simplification of microarray data analysis came from the emergence of data mining tools, which led to a new insight into the world of microarray expression technology. The emerging field of data analysis seems to produce fruitful results in the study of gene behaviour and drug discovery. The only way to improve the technology is to keep the old methods and tools as models, and produce modified versions, which will emerge as newer improved tools that are fit for the existing data.

Based on the background of microarray expression data and the detailed study of two software tools, future tools can be built according to the needs of the researchers and analysts. No matter whether the user is a biologist or statistician, the following provisions by a vendor facilitate the whole process of microarray data analysis:

1. Live web-based or telephonic adept technical support around the clock.
2. Easily accessible and easily available user manuals and documentation, either in hard copy-form separately, or in the form of hyperlinks in the application programming interface.
3. Regular on-site training sessions and online training sessions.

Besides this, the features, which when present in a data mining tool facilitates the whole learning process as well as the actual analysis of the gene data are:

1. User-friendly interface, such as drag and drop menus, gene lists and widgets, for users with no computer applications background.

2. Easy incorporation of data from other tools for cross-referencing and validation of results.
3. Easily manageable database for import of data into the tools, managed by local users of the tool.
4. Use of an application programming language, such as Java, which is platform independent.
5. Incorporation of as many statistical test options and clustering algorithms as has been discovered so far in data analysis studies. This provides flexibility to the user who may use any test that seems more appropriate for the particular analysis being performed. This will vary from user to user, as there is no real right or wrong when a biologist has to set certain criteria for his/her analysis.
6. Visualization tools for almost all analyses, so that the user can visualise the sequence of events and may be able to modify the process accordingly.
7. Presence of tools that can show intersection of many gene lists at a time.
8. Tools that may construct tables of genes showing gene-to-gene interaction in order to observe the change in expression pattern or level in the presence of another gene that is known (or unknown) to the user. This could supplement gene information for a better understanding of the gene behaviour in different conditions of presence of other genes in the system.
9. Ontology construction tools to classify the known genes in the study; further classification and cluster analysis of these ontology differentiated genes for a better interpretation of the genes.

10. Ability to build homology tables for an easy comparison of genes from one organism to another. This feature is available in GeneSpring™ 5.0.

The formation of small discussion groups and organizations responsible for standardization of gene expression tool requirements may prove very useful for the future development of such tools. This includes forums and conferences of computational biologists from around the world. Such discussions would bring about a more public and explanatory description of the requirements and specifications to be integrated in gene expression analysis tools.

APPENDIX

HIERARCHICAL CLUSTERING OF DIFFERENT TISSUES

This Appendix contains some images after hierarchical clustering in GeneSpring™ 4.2. The aim is to demonstrate the clustering pattern for the control and treatment groups in the different tissues of rat in this study.

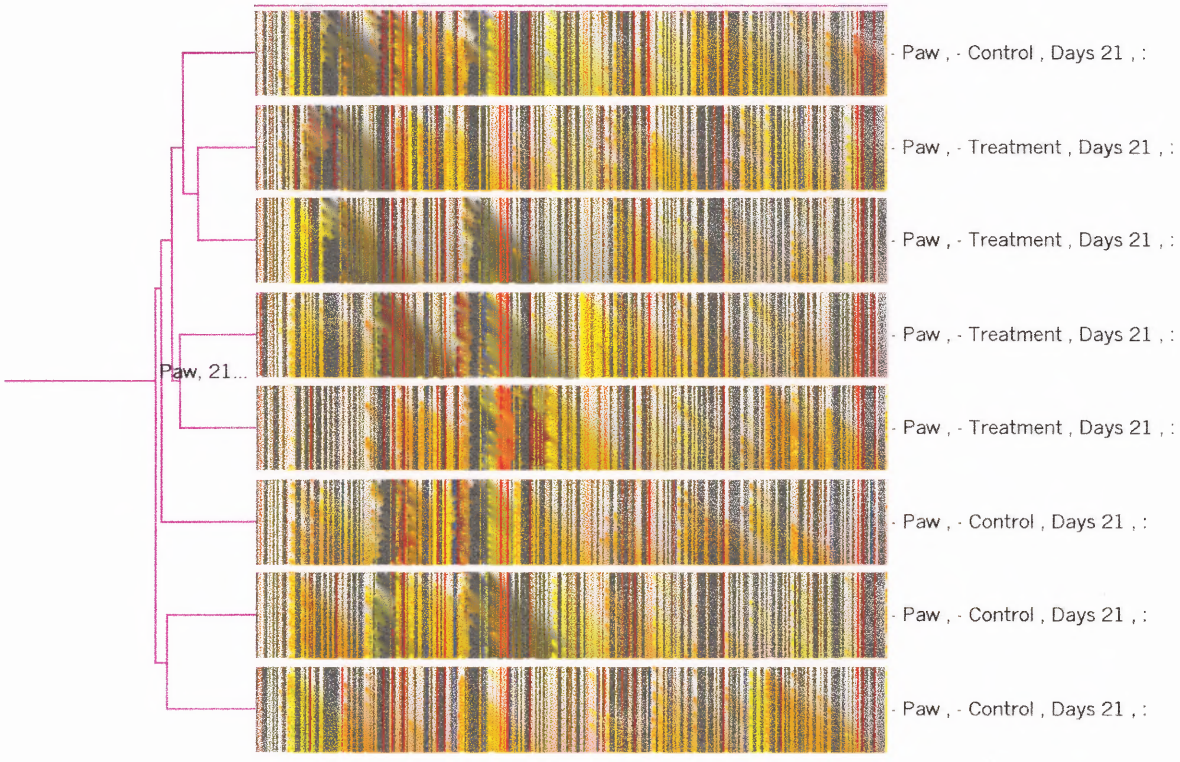

Figure A.1 Experiment tree of paw tissue at 21 days in GeneSpring™ 4.2.

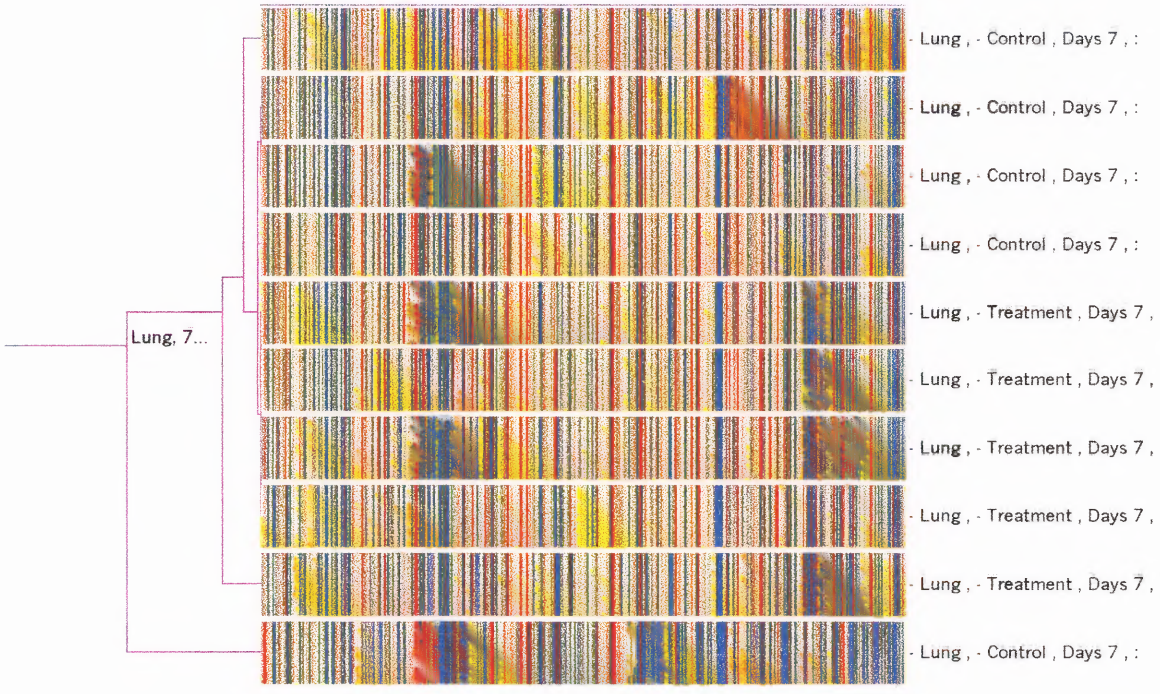

Figure A.2 Experiment tree of lung tissue at seven days in GeneSpring™ 4.2.

Figure A.3 Experiment tree of lung tissue at 21 days in GeneSpring™ 4.2.

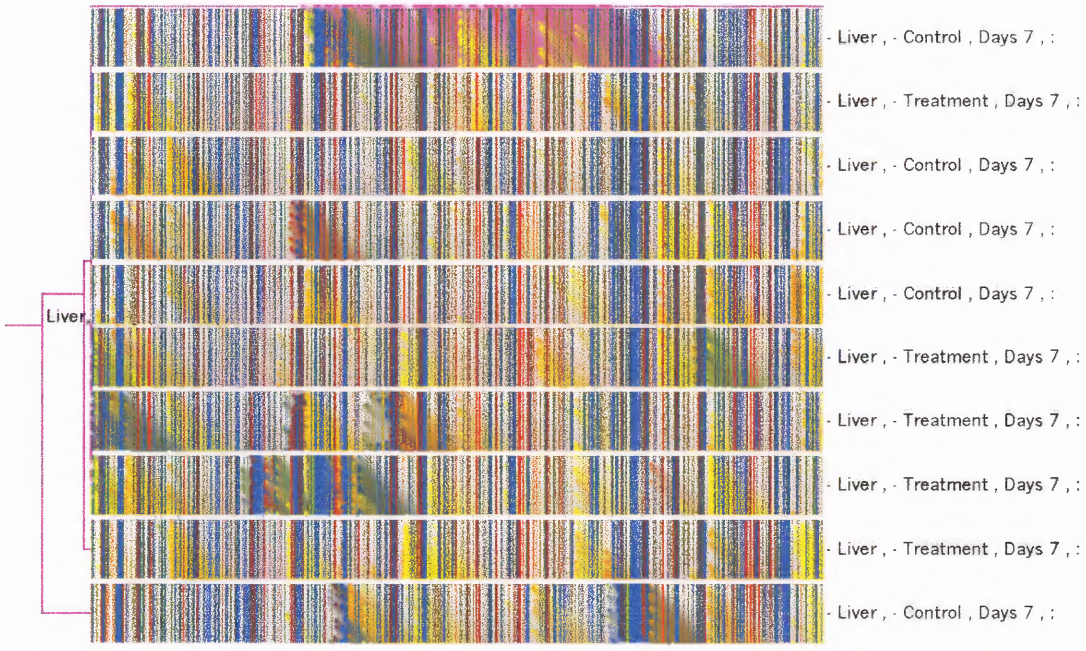

Figure A.4 Experiment tree of liver tissue at seven days in GeneSpring™ 4.2.

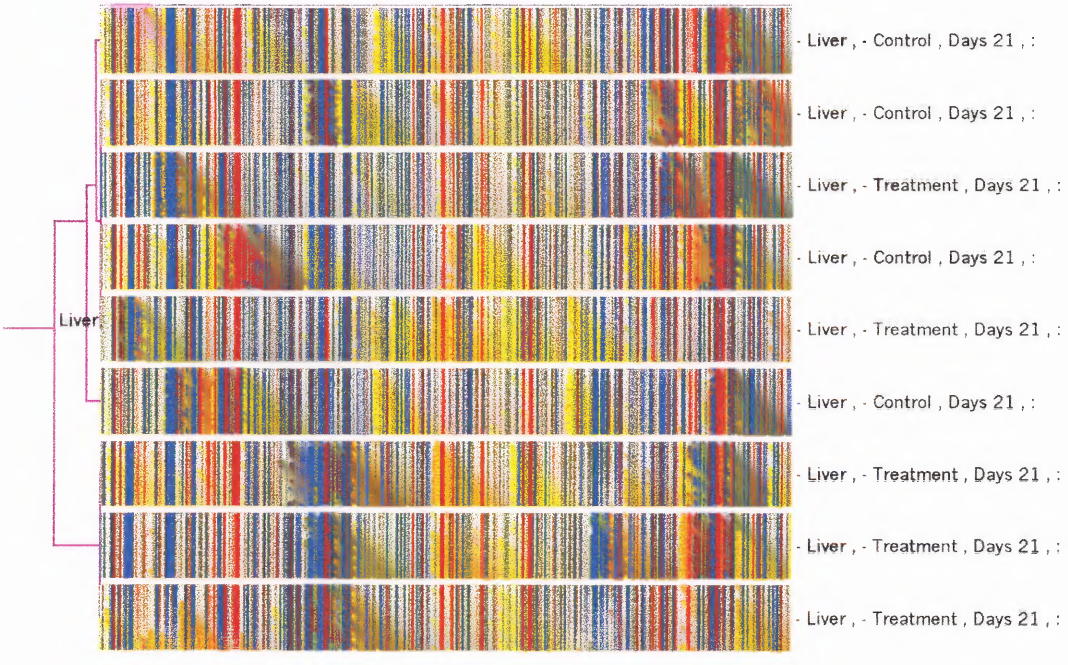

Figure A.5 Experiment tree of liver tissue at 21 days in GeneSpring™ 4.2.

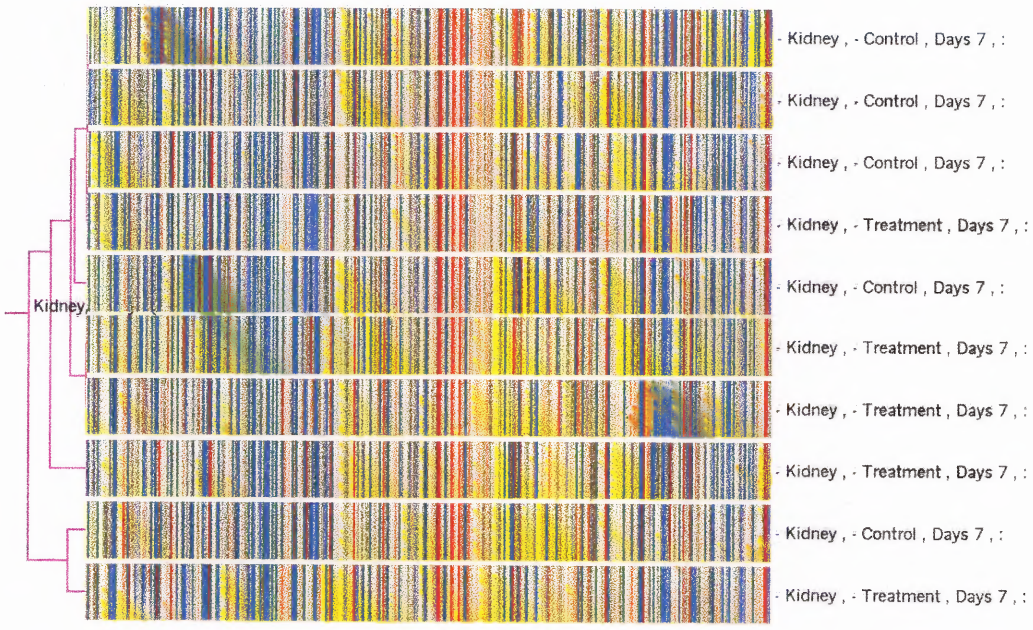

Figure A.6 Experiment tree of kidney tissue at seven days in GeneSpring™ 4.2.

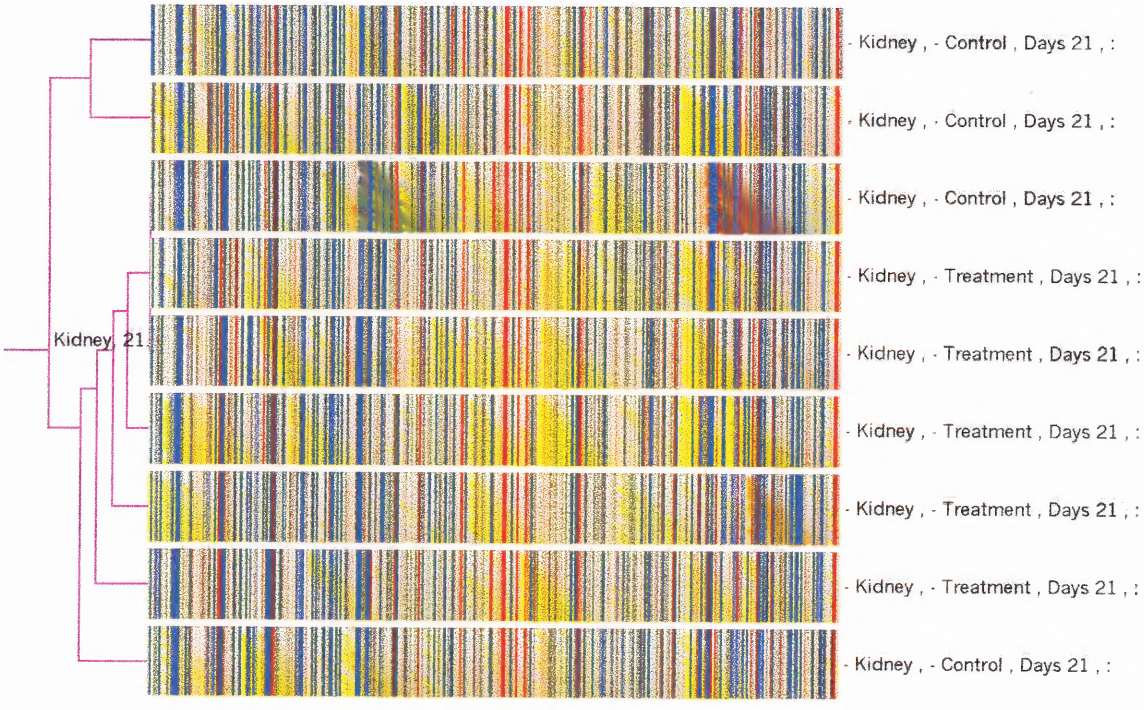

Figure A.7 Experiment tree of kidney tissue at 21 days in GeneSpring™ 4.2.

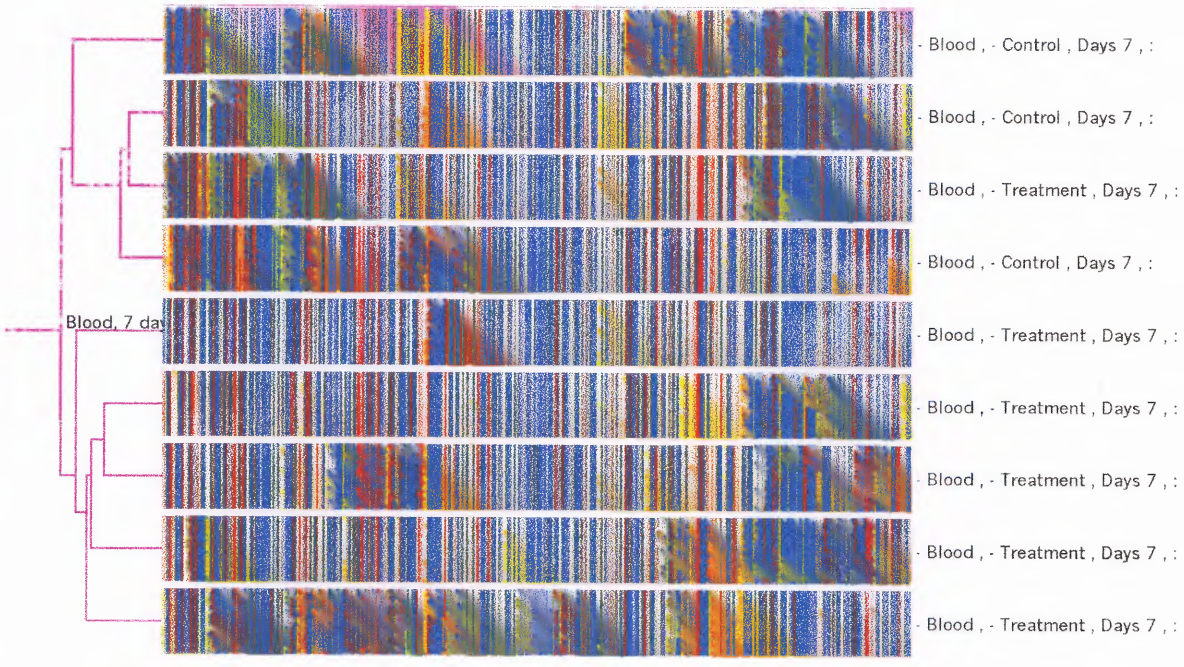

Figure A.8 Experiment tree of blood tissue at seven days in GeneSpring™ 4.2.

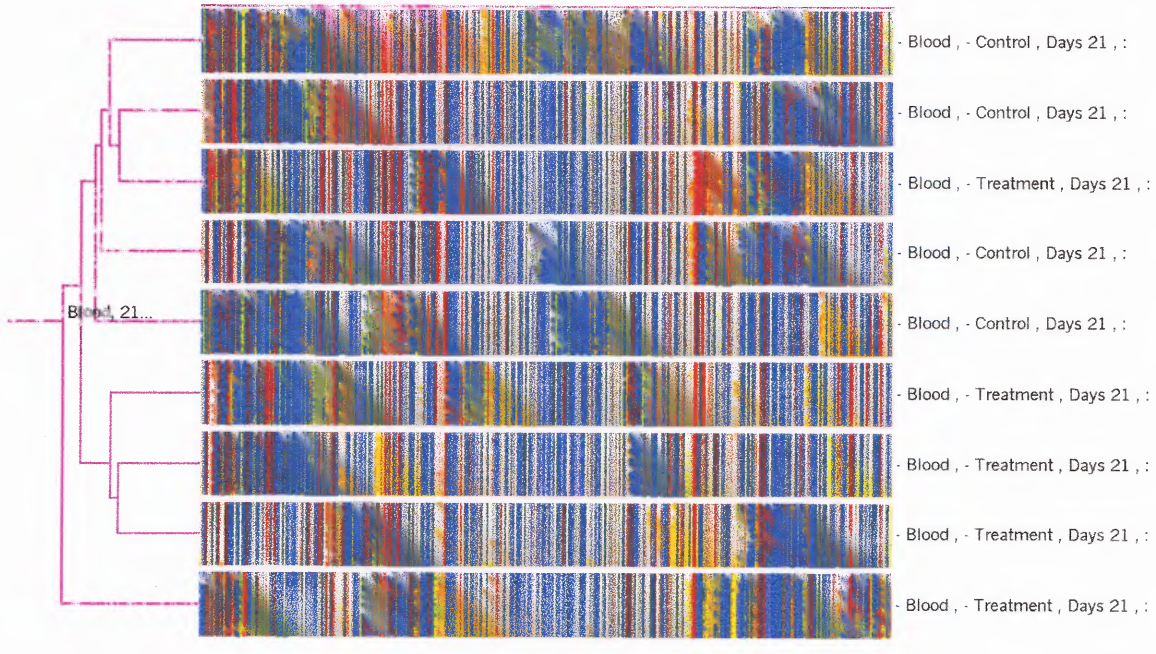

Figure A.9 Experiment tree of blood tissue at 21 days in GeneSpring™ 4.2.

REFERENCES

1. Knudsen, S. (2002). *A Biologist's Guide to Analysis of DNA Microarray Data*, John Wiley & Sons, Inc., New York.

2. Ramakrishnan, N. & Grama, A. (2001). "Mining Scientific Data", *Advances in Computers*, 55, 119-169.

3. Expressionist™ 3.1 User Manual (Furnished under License), GeneData, Basel, Switzerland.

4. GeneSpring™ 4.2 User Manual, Silicon Genetics, Redwood City, CA.

5. MS Excel (Version 1997) [Computer Software].

6. Affymetrix User Manual, MAS™ 4.0 (Microarray Analysis Suite).

7. Tusher, V.G., Tibshirani, R. & Chu, G. (2001), "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response", *Proceedings of the National Academy of Sciences, USA*, 98, 5119-2121.

8. Yang, Y.H, Dudoit, S., Luu, P. & Speed, T.P. (2001), "Normalization for cDNA Microarray Data", *SPIE Bios 2001, San Jose, California.*

9. Brazma, A., et.al (2001), "Minimum Information about a Microarray experiment (MIAME)-towards standards for microarray data", *Nature Genetics*, 29,365-371.

10. Aris, V., Tolias, P. & Recce, M. (2001), "Selective Expression algorithm for class separation for DNA Microarrays", *DIMACS Workshop on Analysis of Gene Expression Data, Rutgers University, NJ.*

11. Brass, A. (2002), "Microarray Analysis-A Bioinformatics Perspective", *Report, Genomics Microarrays*, 138-142.

12. Kohonen, T. (1997), *Self-Organizing Maps*, Springer, NY.

13. Smith, G.K., Yang, Y.H. and Speed, T. (2002), "Statistical Issues in cDNA Microarray Data Analysis", *Research Report, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia.*

14. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S.& Golub, T.R. (1999), "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation", *Proceedings of the National Academy of Sciences, USA*, 96, 2907-2912.

15. Quackenbush, J. (2001), "Computational Analysis of Microarray Data", *Review in Nature Genetics,* 2, 418-427.

16. Dudoit, S., Fridlyand, J. & Speed, T. (2000), "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data", *Journal of the American Statistical Association, Technical Report #576.*

17. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. (1998), "Cluster analysis and display of genome-wide expression patterns" *Proceedings of the National Science Academy of Science, USA,* 95, 14863-14868.

18. Leung, Y. F., Lam, D. S. C., Pang, C. P. (2001), "The miracle of microarray data analysis", *Genome Biology,* 2, 4021.1-4021.2.

19. CSC News, Finnish IT Center for Science. Microarray Data Handling. http://www.csc.fi/molbio/microarrays/CSCnews.pdf (25 October 2002).

20. CSC News, Finnish IT Center for Science. Summary of Survey Results of CSC's Microarray Services. http://www.csc.fi/molbio/microarrays/ (31 October 2002).

21. The Board of Trustees of Leland Stanford Junior University (2001), Stanford Microarray Database.
http://genome-www5.Stanford.edu/MicroArray/SMD/restec.html
(25 September 2002).

22. Cambridge Healthtech Institute Staff & Goodman, N. (2001), "New Tools and Approaches Revolutionizing Microarray Data Analysis", (Excerpt) *CHI report Bioinformatics: Getting Results in the Era of High-Throughput Genomics.*
http://www.chiresource.com/newsarticles/issue10_2.asp (25 September 2002).

23. Leung, Y. F. (2002), "My Microarray Software Comparison-Data Mining Software".
http://ihome.cihk.edu.hk/~b400559/arraysoft_mining.html (30 September 2002).

24. National Research Council (2001), Non-subjective Diagnosis of Arthritis.
http://www.ibd.nrc.ca/english/b_arthritis.htm (10 October 2002).

25. Yang, Y. H. , Speed, T. (2002), "Design issues for cDNA microarray experiments", *Review in Nature Genetics,* 3(8):579-588.

26. Nadon, R. & Shoemaker, J. (2002), "Statistical issues with microarrays: processing and analysis", *Trends Genetics,* 18(5), 265-71.

27. Wu, T. D. (2001), "Analyzing gene expression data from DNA microarrays to identify candidate genes", *Journal of Pathology,* 195(1), 53-65.

28. Brazma, A. & Vilo, J. (2000), "Gene expression data analysis", *FEBS Letters*, 25, 480(1), 17-24.

29. Ben-Dor, A., Shamir, R. & Yakhini, Z. (1999), "Clustering Gene Expression Patterns", *Journal Of Computational Biology*, 6 (3/4), 281-297.

30. Dougherty, E. R. et.al. (2002), "Inference from Clustering with Application to Gene-Expression Microarrays", *Journal Of Computational Biology*, 9,105-126.

31. Rocke, D. M. & Durbin, B. (2001), "A Model for Measurement Error for Gene Expression Arrays", *Journal Of Computational Biology*, 8, 557-569.