Colon Cancer Diagnosis Using NMR Spectra of Urine

by
Hannah T. Medford

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in the Department of Biomedical Engineering.

Chapel Hill
2006

Approved by:

Advisor : Dr. Jeffrey Macdonald
Reader : Dr. David Threadgill
Reader : Dr. Oleg Favorov

© 2006
Hannah T. Medford
ALL RIGHTS RESERVED

ABSTRACT

HANNAH T. MEDFORD: Colon Cancer Diagnosis Using NMR Spectra of Urine.

(Under the direction of Dr. Jeffrey Macdonald
.)

Colon cancer is the third most common cancer in people, and early diagnosis is critical to survival. This study investigates the efficacy of using metabolomic technology in diagnosing colon cancer in a mouse model. Urine from a genetically defined population of mice was analyzed by NMR spectrometry, after carcinogen exposure and categorization for tumor development based on histological examination. The NMR spectra were then analyzed by statistical methods of classification to determine if colon tumors result in changes to the metabolites secreted in urine that can be detected by NMR spectrometry. Different statistical analyses were also compared to determine which is most effective at retrieving information from the NMR data. Principal Component Analysis (PCA), a popular analysis for metabolomic data, is ineffective on this data set. Support Vector Machine (SVM) reveals six significant components, which when entered into PCA results in clear separation of normal and tumor-bearing mice.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

Cancer has become a serious public health issue. It's also one of the least understood disease processes. While a small handful of cancers have been determined to be familial, following a fairly obvious inheritance pattern, the vast majority appear sporadically in the general population. These cancers are suspected to be governed by an environmentally-induced genetic susceptibility.

Because of the uncontrollable variation in the environment, cancer susceptibility is very difficult to study in humans. Therefore, the need has arisen for strains of mice which show varying susceptibility to certain cancers, so that the development of these cancers can be studied in a laboratory setting. Such studies would provide insight into which genes play a part in cancer susceptibility and how they affect the resistance to carcinogenic effects in certain individuals.

Another barrier to this form of cancer studies is the difficulty in following the progression of most tumors over time. Most studies require that individuals be sacrificed to obtain information at different time-points. Obviously, once an individual has been sacrificed, no more information can be gathered.

Metabolomic techniques allow us to obtain data at many different time-points on a single individual non-invasively. Metabolomics will provide the scientific community with an enormous amount of invaluable data which must then be sorted and analyzed to extract the information. In fact, Metabolomic studies often yield so much data that they present a unique challenge. Often, the number of samples is quite small while the number of variables measured is enormous. Additionally, there can be unknown variables that lead to correlations between the observed variables adding to the complexity of the problem. Thus, it has become necessary to develop new and better statistical methods to extract usable information from the morass of data produced.

Chapter 2

Background

2.1 Cancer

Cancer is a general term used for many different diseases characterized by uncontrolled, abnormal growth of cells. A tumor may develop locally, spread into nearby tissues, or cancerous cells may spread through the blood stream or lymphatic system to other parts of the body(Tannock IF, Hill RP et al. 2005).

Normally, cells grow and divide only when the body needs new cells. Normal cells are programmed to grow old and die in an orderly process. When mutations occur within the genes that control the process of cell division and cell death, new cells form when the body does not need them, and old cells do not die when they should. The extra cells form masses called neoplasms or tumors. Tumors can be either benign or malignant. Malignant tumors can spread by invasion and a process called metastasis, while benign tumors tend to grow only locally.

Metastasis is when cells break off from the primary tumor, penetrate the lymphatic and blood vessels, and circulate to other parts of the body where they form tumors in normal tissue elsewhere in the body. When cancer cells spread to form a new tumor, it is called a secondary, or metastatic tumor, and its cells are like those in the original tumor. This means, for example, that if colon cancer metastasizes to the liver, the secondary tumor is made up of abnormal colon cells, not abnormal liver cells. The disease in the liver is metastatic colon cancer, not liver cancer.

A number of factors contribute to the process of metastasis. The abnormal cell must attach to and degrade the proteins of the extracellular matrix separating the tumor from surrounding tissues. Once the proteins of the matrix have been broken down, the cell can breach the extracellular matrix and is free to escape and migrate through the circulatory system. Another critical event required is the growth of a new network of blood vessels. This process of forming new blood vessels is called angiogenesis. Tumor angiogenesis actually starts with cancerous tumor cells releasing molecules that send signals to surrounding normal host tissue. This signaling activates certain genes in the host tissue that, in turn, make proteins to encourage growth of new blood vessels.

Cancers are classified based on the type of cell in which they originate. Adenomas originate from glandular tissue. Carcinomas originate in epithelial cells. Leukemia starts in the bone marrow stem cells. Lymphoma is a cancer originating in lymphatic tissue. Melanoma arises in melanocytes. Sarcoma begins in the connective tissue of bone or muscle. Teratoma begins within germ cells.

Early detection of cancer is critical to the clinical outcome in cancer treatment. For example, 15% of all cancer related deaths in the United States are do do colon cancer. Only 37% of colon cancer cases are detected early enough for standard treatments. Once colon cancer reaches the metastatic stage, only 7% of patients survive.

The current state-of-the-art standard for detecting colon cancer is white light endoscopy with gross visualization of the cancerous lesions. Unfortunately, the visual clues available to determine diseased states of lesions are very small. This is especially true in discriminating between benign, dysplastic growths, and malignant lesions.

2.2 Metabolomics

Metabolomics is the study of the entire complement of small molecules in an organism. It has become a rapidly expanding approach to biomedical and pharmaceutical research, and has a myriad of clinical applications.

Like the other "'omics," Genomics, Proteomics, etc, Metabolomics is a systems integration approach. Metabolomics describes the information gleaned from an endogenous survey of the metabolic profile or metabolome. Though some confusion exists over the creation of additional terms, such as "metabonomics" there can be no doubt of the value provided by the study of Metabolomics.

The small molecule complement in a cell provide an insight into that cell's status. It is the total result of all metabolic processes in the cell, both catabolic and anabolic. It also reflects absorption, distribution, and detoxification of materials, energy utilization, signal transduction, and regulation. It results from the expression of the genome and proteome in response to the cellular environment. While the Genome is representative of what might be, and the Proteome is what is expressed, it is the Metabolome that represents the current status of the cell or tissue.

A living cell responds to its environment quickly by utilizing and altering proteins whose effect is reflected in the metabolome. Thus, metabolomics also has a temporal component as some biomarkers reflect an immediate response while some reflect events that occured some time in the past.

When taken in its entirety it is not always necessary to know the identity of individual components in a metabolic profile. Systemic changes in pattern are indicative of specific states or of changes in status.

Just as it is possible to retrieve useful information from a comprehensive profile, it is also possible to focus on specific sub-systems (e.g. measurement of DNA adducts or protein oxidation products) to tease out information about specific states.

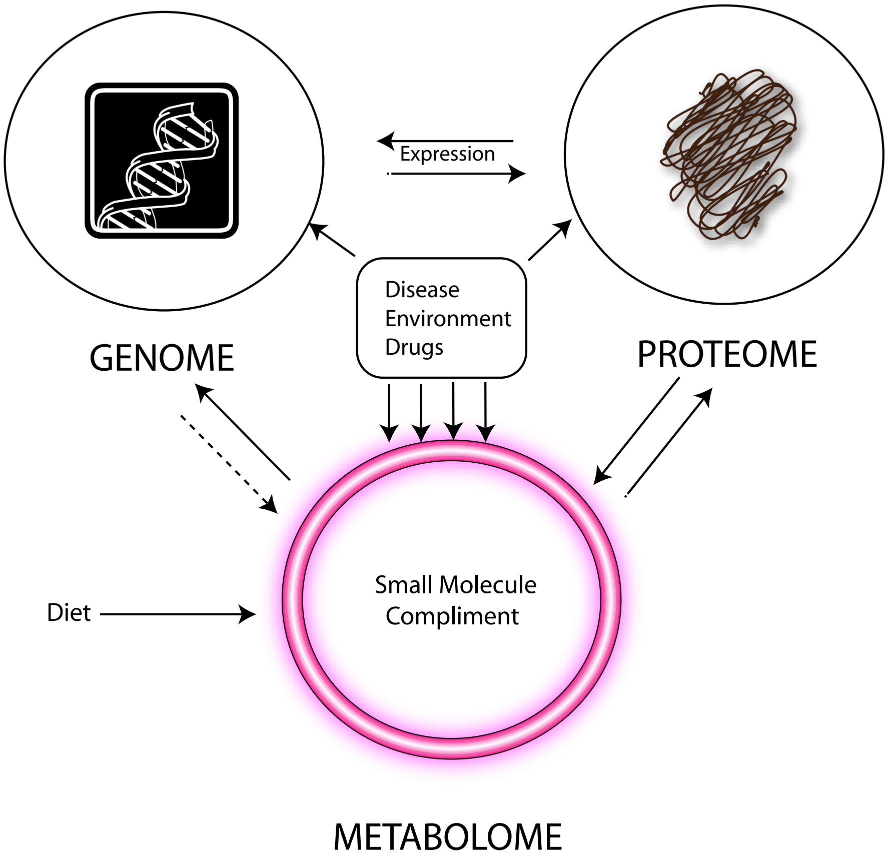

Figure 2.1: The Interactive System

The Metabolome is extremely sensitive to exogenous stimulation. For example, following the administration of a drug or toxin numerous pathways can be affected. Not only are changes seen in the system targeted but also in many other pathways.

High-throughput metabolomic techniques can be used for rapid profiling of large numbers of samples, while still being able to provide, to different extents, specific chemical information. Samples can be examined after solvent extraction (no derivatization required), or as intact tissues (magic angle spinning NMR), liquids or semi-solids (NMR)(Holmes, 2000).

Two major challenges have been recognized in the field of Metabolomics: Incredibly large data sets make it difficult to find relevant data on marker compounds. Detecting and identifying subtle changes can be problematic with the sensitivity required to identify such molecules.

2.3 Nuclear Magnetic Resonance

Nuclear magnetic resonance (NMR) is a physical phenomenon first described by Felix Bloch and Edward Mills Purcell in 1946. They shared the 1952 Nobel Prize for physics for their discovery. NMR is used as a spectroscopy technique to obtain physical, chemical, and electronic information about molecules. NMR is also the technology on which Magnetic Resonance Imaging (MRI) is based. Both NMR and MRI technologies have become invaluable tools in many aspects of science and medicine.

NMR involves the interaction of some atomic nuclei within an external magnetic field while being exposed to a second oscillating magnetic field. Not all nuclei experience this interaction, dependent on whether or not they possess a property called 'spin'. The property of spin can be thought of as a small magnetic field which causes the nuclei to produce an NMR signal. The magnetic conditions within the molecules are measured by monitoring the frequencies absorbed and emitted by the nuclei. NMR spectroscopy takes advantage of this phenomenon to obtain physical, chemical, and electronic properties of molecules and is the underlying technique of Magnetic Resonance Imaging (MRI).

In NMR, a sample of the material to be tested is placed inside a static external magnetic field formed by a strong electromagnet. An antenna is formed by a coil of wire around the sample, and is used to irradiate the sample with radio waves. At certain frequencies, atomic nuclei within the sample will absorb the radiation and enter an excited state. After a time, the nuclei will re-emit the radiation, which can be detected by the antenna. Finally, a measurement is taken of how much radiation is re-emitted, and when.

The Larmor equation can be used to determine the amount of energy needed for a given nucleus to resonate. The equation describes the relationship between the strength of the magnetic field, B0, and the precessional (Larmor) frequency, ω_0.

$\omega_0 = \gamma B_0$

The gyromagnetic ratio, γ, is the ratio of the magnetic moment to the angular momentum of a particle, and is constant for a given nuclei. For example, hydrogen (1H) has $\gamma = 4,258 Hz/G$.

In principle, proton (1H) NMR can detect any metabolites containing hydrogen. Signals can be assigned by comparison with libraries of reference compounds, or by two-dimensional NMR. 1H NMR spectra of urine are inevitably crowded not only because there is a large number of contributing compounds, but also because of the low overall chemical shift dispersion. 1H spectra are also complicated by spin-spin couplings which add to signal multiplicity, although they are an important source of structural information(Nelson, 2003).

2.4 Computational Methods of Classification

Generally, classification is the art of placing objects or concepts into groups based on a set of rules. In statistics, classification is a type of algorithm, which takes a feature representation of an object or concept and maps it to a label. Typically, a classification algorithm computes the probability of a class label based on the feature inputs that were observed.

There are many different approaches to solving classification problems. However every approach has the same goals, to imitate human decision-making behavior, but with greater consistency, to be able to handle a wide variety of problems and generalize when given enough data.

Multivariate data analysis is all about separating the information from the noise in a large dataset with a lot of variables. An effective analytical method should produce results that summarize the essential information in an easy to interpret format.

When the data set has many variables and the relationships between them are poorly understood, the problem becomes more and more difficult. This is especially true when the number of variables far exceeds the number of samples in a given data set.

2.5 Principal Component Analysis

Principal component analysis is designed to analyze the variance of a dataset in terms of the principal components. The principal components are defined as a set of variables that define a hyperplane that captures the maximum amount of variation in a dataset and is orthogonal to the previous principal component of the same dataset(Yeung, 2001). Essentially, PCA tries to find the parts of the dataset which are most important and defining, while simultaneously filtering out noise. This makes it easier to identify groupings and outliers and spot trends in the data.

PCA is a transform that chooses a new coordinate system for a data set such that the greatest variance by any projection of the data set comes to lie on the first axis, the first principal component, the second greatest variance on the second axis, and so on. PCA is used to reduce the dimensionality of a dataset and yet retain those characteristics of the dataset that contribute most to its variance by eliminating the lower principal components. These characteristics may be the 'most important', but this is not necessarily the case, depending on the application.

Unlike other transforms, PCA does not have fixed basis vectors, because in PCA the basis vectors are dependent on the data set.

If one assumes the empirical mean of the dataset is zero, the principal component (w1) of dataset (x) can be defined as:

$$w_1 = arg \max_{\|w\|=1} E\{(w^T x)^2\}$$

The k-th components can be found by subtracting the first k-1 components from x:

$$\hat{x}_{k-1} = x - \sum_{i=1}^{k-1} w_i w_i^T x$$

and by substituting this as the new dataset to to find a principal component in:

$$w_k = arg \max_{\|w\|=1} E\{(w^T \hat{x}_{k-1})^2\}$$

A simpler way to calculate the components wi uses the empirical covariance matrix of x, the measurement vector. By finding the eigenvalues and eigenvectors of the covariance matrix, we find that the eigenvectors with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the dataset. The original measurements are finally projected onto the reduced vector space.

2.6 Partial Least Squares

The method of Partial Least Squares (PLS) is known to be useful in some problems where the number of variables is equal to or less than the number of samples, and/or there are other variables that may lead to a correlation between variables(**?**). This method aims to identify the underlying factors, or linear combination of the independent variables, which best model the dependent variables. PLS has been applied in many different areas of research and technology, especially biotechnology and chemometrics.

PLS is similar to PCA, but instead of finding maximum variance hyperplanes, it is based on a linear regression. PLS forms a set of orthogonal components or factors from a large number of original variables. The main purpose of the method is to create a model between the factors rather the original data. So, the orthogonal factors are chosen so as to result in the greatest correlation. Sometimes the model is too strongly correlated such that it not only explains the data but also the noise! This is called "overfitting" and is one of the dangers of PLS because the model will appear to be very good, but is really useless in predicting samples that are not included in the training set(Denham, 1994).

The basic algorithm for PLS is described by the following equations:

$w = \frac{(X'y)}{\sqrt{norm(X'y)}}$

$t = Xw$

$p = \frac{(X't)}{norm(t)}$

$q = \frac{(y't)}{norm(t)}$

$X^{(k+1)} = X - tp'$

$y^{(k+1)} = y - qt'$

Here, N is the number of samples, M is number of variables, X[N,M] is the descriptor matrix, y[N] is the activity vector, w[M] is the auxiliary weight vector, t[N] is the factor coefficient vector, p[M] is the loading vector, and q is the scalar coefficient of relationship between factor and activity.

2.7 Support Vector Machine

A support vector machine (SVM) is a supervised learning technique first described by Vladimir Vapnik. An SVM is a maximum-margin hyperplane that lies in some space. Given training examples labeled either "yes" or "no", a maximum-margin hyperplane splits the "yes" and "no" examples, such that the distance from the closest examples to the hyperplane is maximized(Vapnik, V. 1995). So essentially, an SVM maximizes the distance between groups and minimizes the distance between points within a group.

The use of a maximum-margin hyperplanes is driven by the statistical learning theory. This provides a probabilistic test error bound which is minimized when the margin is maximized.

The original SVM was a linear classifier. However, Vapnik suggested using the kernel trick. In the kernel trick, every time a linear algorithm uses a dot product ,replace it with a non-linear kernel function. This causes the linear algorithm to operate in a different space. For SVMs, using the kernel trick makes the maximum margin hyperplane be fit in a feature space. The feature space is a non-linear map from the original input space, usually of much higher dimensionality than the original input space. In this way, non-linear SVMs can be created. If the kernel used is a radial basis function, the corresponding feature space is a Hilbert space of infinite dimension. SVMs are well regularized, so the infinite dimension does not alter the results.

The formulation of an SVM starts with a basic linear maximum-margin classifier. The performance of the classifier is measured in terms of classification error.

The decision making function of the classifier is $f(x, \lambda) = sgn(w\dot{x} + b)$

The kernel function of the SVM determines the margin, and the separability of the data. Different kernel functions may have differing levels of success in separating the data and maximizing the margin. The kernel function is:

$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$

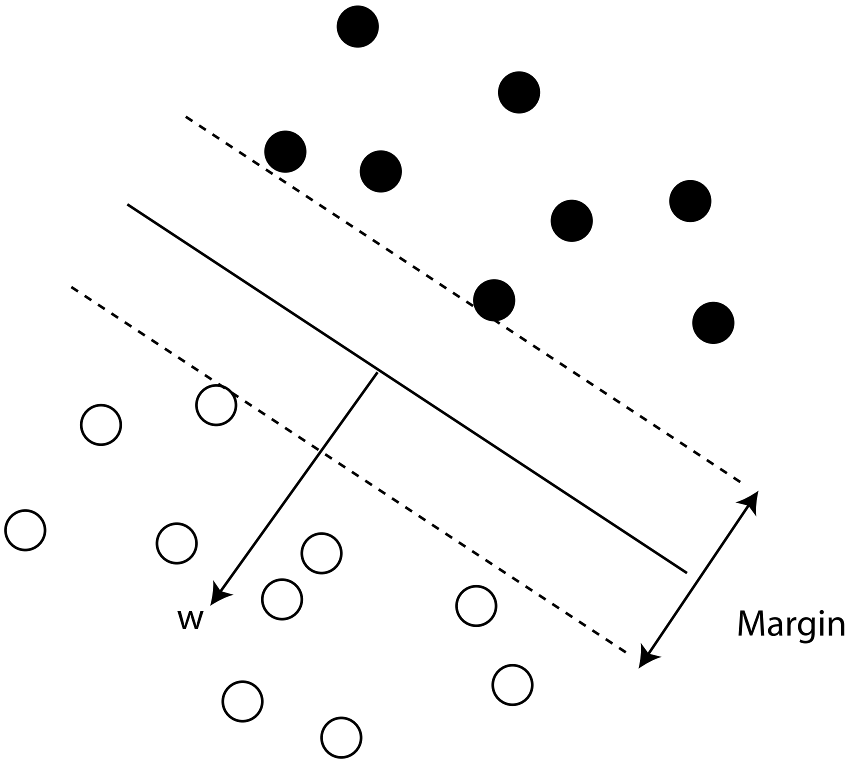

Figure 2.2: Decision Margin of Oriented Hyperplane

New kernels are being proposed by researchers all the time, but these are the four basic kernels used in the formulation of SVMs:

- linear: $K(x_i, x_j) = x_i^T x_j$
- polynomial: $K(x_i, x_j) = (\gamma * x_i^T x_j + r)^d, \gamma > 0$
- radial basis function (RBF): $K(x_i, x_j) = x_i^T x_j$
- sigmoid: $K(x_i, x_j) = x_i^T x_j$

SVM is a powerful method, but has not been utilized in the field of metabolomics.

2.8 Mutant Mouse Models

The development of transgenic and "knock out" mice has resulted in a revolution in our thinking. The mouse has emerged as the major research model for biology. The mouse has been referred to as the E. coli of modern biology and the surrogate for human biology. The mouse is inexpensive to maintain, easy to manipulate and its genome is syntetic with the human. The mouse has become the mammal of choice for the analysis of interaction of specific genes with the whole animal(Clarke, 2002).

In 1921, inbred strains of mice that were predisposed to tumor development were developed and disseminated among cancer researchers. In 1962, the discovery of a mutant mouse with low immunity led to human tumor transplantation. This development was valuable to cancer research. Then in the 1980's, transgenic mice were discovered. These mice have genes which have been altered to produce a desired characteristic.

The inbreeding of mice predisposed to developing cancer has led to a variety of specialized strains. In 1921, Leonell Strong established many inbred strains that frequently and spontaneously developed cancer. Serving as a virtually unlimited source of many types of tumors, these inbred mice have made it possible to study the growth and general characteristics of tumors.

In the late 1980s the methodology for engineering transgenic mice made it possible to create mice to address specific questions and problems. Transgenic mice result from genetically altered embryos: a gene or combination of genes is microinjected into developing oocytes. The genetic alteration affects the germ plasm, and subsequently can be transmitted to progeny. Through selective breeding, it then is possible to maintain a strain of mice consisting of individuals with particular traits of interest.

A specific trait, such as a predisposition to develop a particular type of tumor, can be introduced into a mouse strain by injecting into the embryo an oncogene, a gene that causes cancer. Transgenic mice permit the study of cancer in specific tissues, including initial tumor development.

Chapter 3

Methods

3.1 Animal Treatment

A population of recombinant inbred mice were treated with the carcinogen, azoxymethane (AOM). Each recombinant inbred strain has a different mix of the two parental strains. The A/J, which is highly susceptible to AOM induced tumors, and C57BL/6J, which is relatively resistant, inbred mouse strains were used to generate the AXB/BXA panel of recombinant inbred strains that were used in this experiment.

The mice were then sacrificed under anesthesia. Urine was collected from each individual, at the time of sacrifice, by inserting a needle into the bladder. The samples were immediately closed in tubes, frozen and stored at -80 degrees Celcius. The mice were then examined histologically for the presence of liver and colon tumors. The tumors were identified, scored, and recorded for each individual.

3.2 Sample Preparation

The urine samples were thawed to room temperature. The urine samples were prepared for NMR analysis by mixing 100 microliters of urine, 70 microliters of Phosphate Buffer (1mM TSP) and 530 microliters of D2O. Then the samples were centrifuged to separate any solid matter. The resulting supernatant was removed with a pipette and inserted into a clean NMR tube, which was capped and labeled.

3.3 NMR Analysis

Proton NMR spectra were obtained from each urine sample on a ANOVA 600 MHz NMR Spectrometer. A NOESY pulse sequence was used to collect the spectra, with a tau = 0, for greatest solvent suppression. Varian software was used to output relative peak intensities of the various resonances.

3.4 Data Processing

The freeware program Mestre-C was used to process the spectra and prepare them for statistical analysis. First, the spectra were imported from the Varian format. Then, a high-pass filter was applied to remove remaining solvent signal. The baseline of the spectra was then corrected using a point-by-point splines technique. Finally, the real parts of the processed spectra were outputted as a single column of intensities in ASCII format. Each spectra 240 bins, the total of all the bins was normalized to equal 1000. The magnitudes of the 240 bins created a histogram of numbers for each spectrum. The histograms for all spectra were combined to for a matrix which served as the input for the statistical analysis.

3.5 Statistical Analysis

The numerical text output from the Mestre-C program was loaded into Matlab and normalized so that mean = 0 and standard deviation = 1, and then used to form a matrix for input into two statistical analyses. The PLS toolbox plug-in for Matlab was used to run both sets of analyses.

First, PCA was used. The plot of Principal Components versus Eigenvalues (the Scree Plot) was used to determine the appropriate number of Principal Components for this analysis. These components were then mapped against one another to determine their usefulness in separating the data.

Chapter 4

Results

4.1 NMR

Each sample spectra was compared with histological data and labeled with either normal or sick, with degrees of metastasis. These are some representative sample spectra that resulted from our NMR spectrometry.

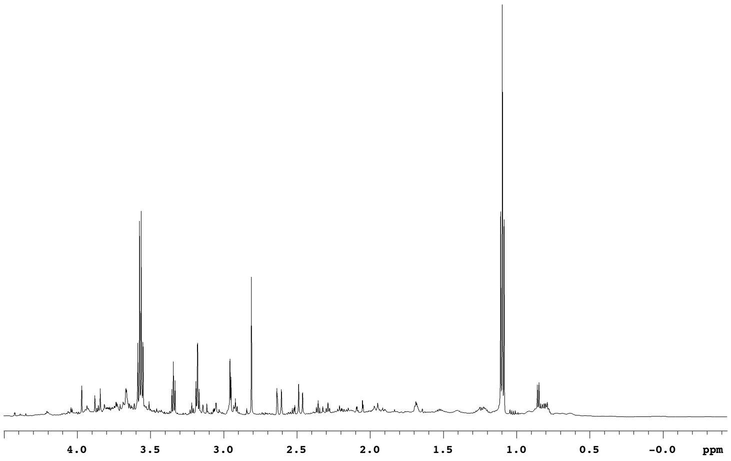

Figure 4.1: Normal

This individual, shown in figure 4.1, having no tumors, was determined to be of the normal phenotype. This spectra can be used for comparison when compared to the metabolomic spectra of animals in various stages of disease in this experiment.

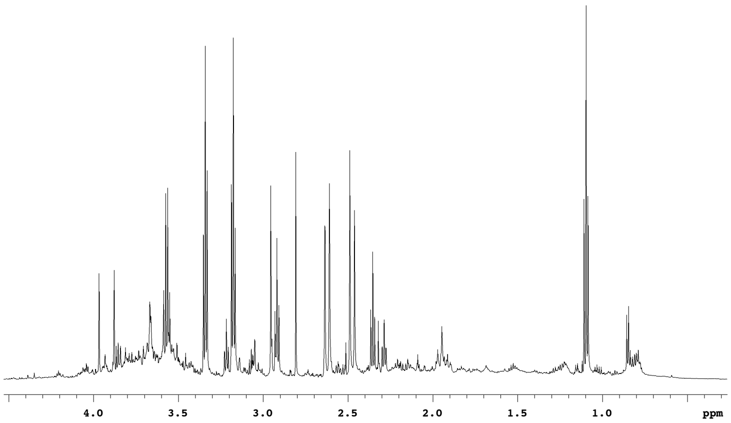

Figure 4.2: Colon Cancer

When histologically examined, this individual, shown in figure 4.2, was found to have significant tumors in the colon, but not elsewhere in the body.

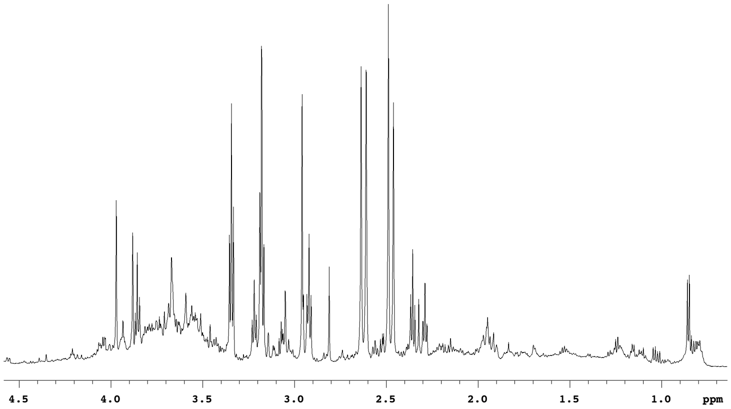

Figure 4.3: Liver and Colon Cancer

This unfortunate animal, shown in figure 4.3, was found to have tumors in both the colon and liver, and represents the most severe stage of the disease.

There are clearly differences in the spectra showing various stages of disease but these are too complex to be understood without the benefit of multivariate analysis.

4.2 Correlation Coefficient

Using the normalized matrix of data, the correlation coefficient was computed. The square of the correlation coefficient results in the coefficient of determination. This tells us how much of the variation in each bin is accounted for by the cancer score of the animal.

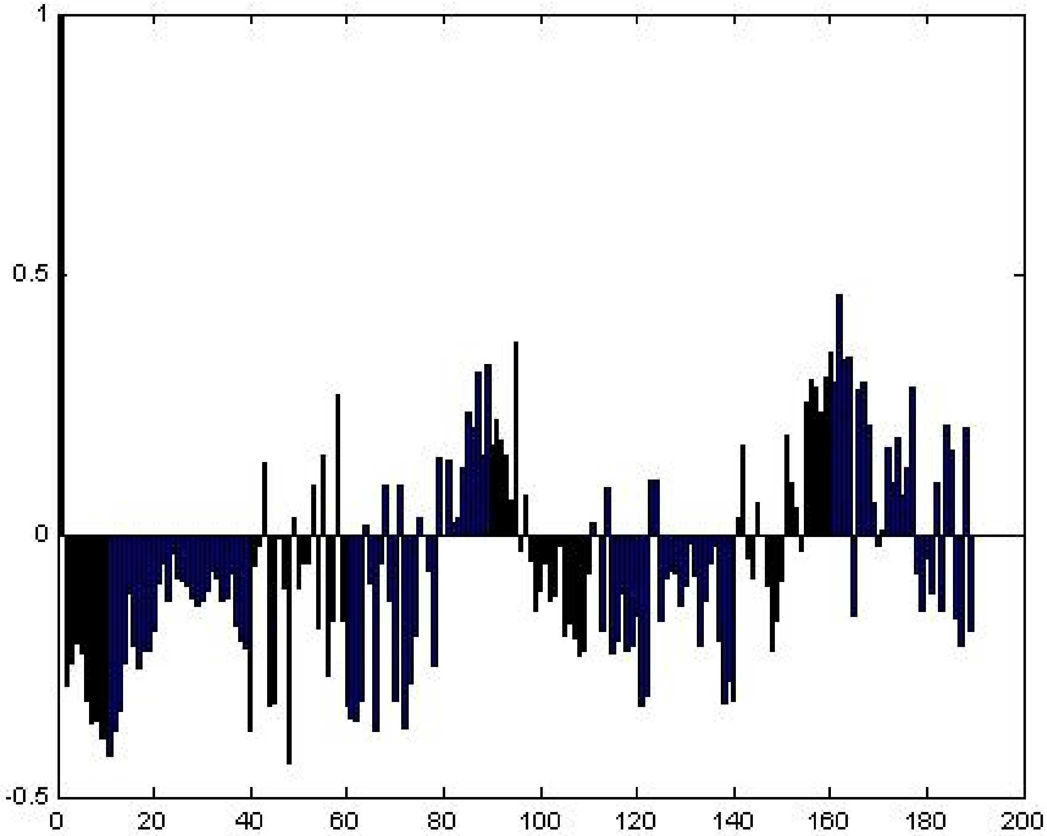

Figure 4.4: Correlation Coefficient

The results shown in 4.4 indicate that no significant correlation was found by this method.

4.3 PCA Results

The PCA resulted in a Scree Plot that indicated four significant principal Components.

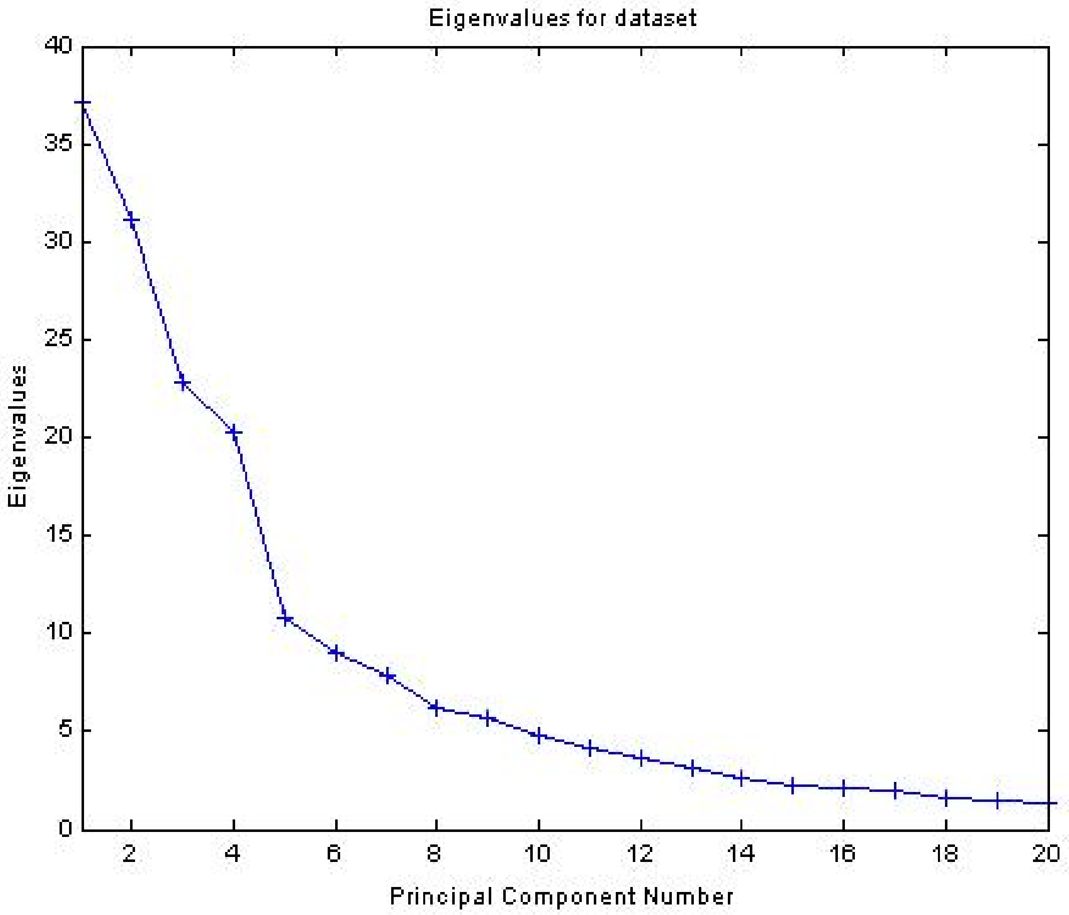

Figure 4.5: Scree Plot

When these components were plotted against one another, the analysis revealed that there was no tendency for clustering that would reflect the presence or the stage of cancer in the subject. The following figures show this data.

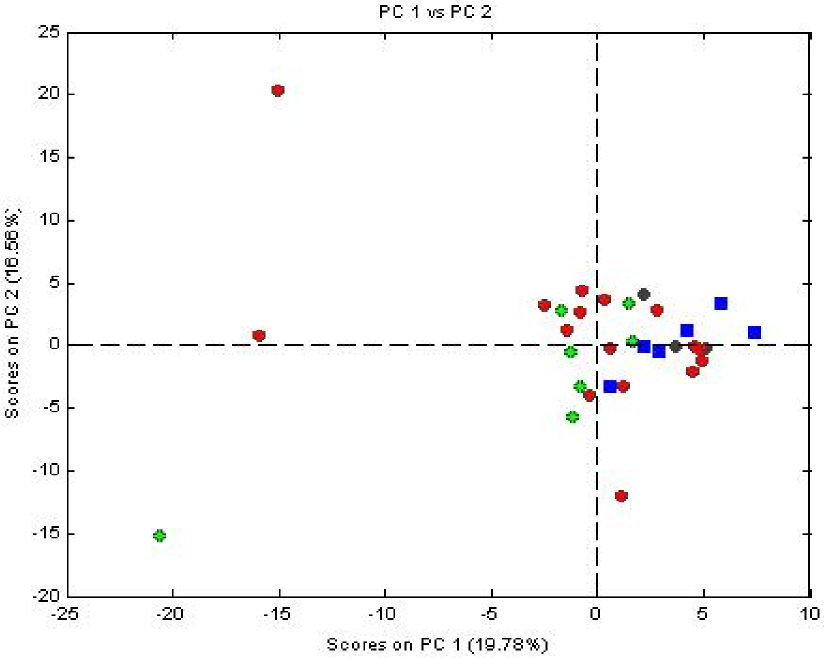

Figure 4.6: PC 1 versus PC 2

The plot of PC4 versus PC5, show in 4.18, shows slightly better separation of the classes than the remaining components. However, the separation is extremely weak and likely to be accidental.

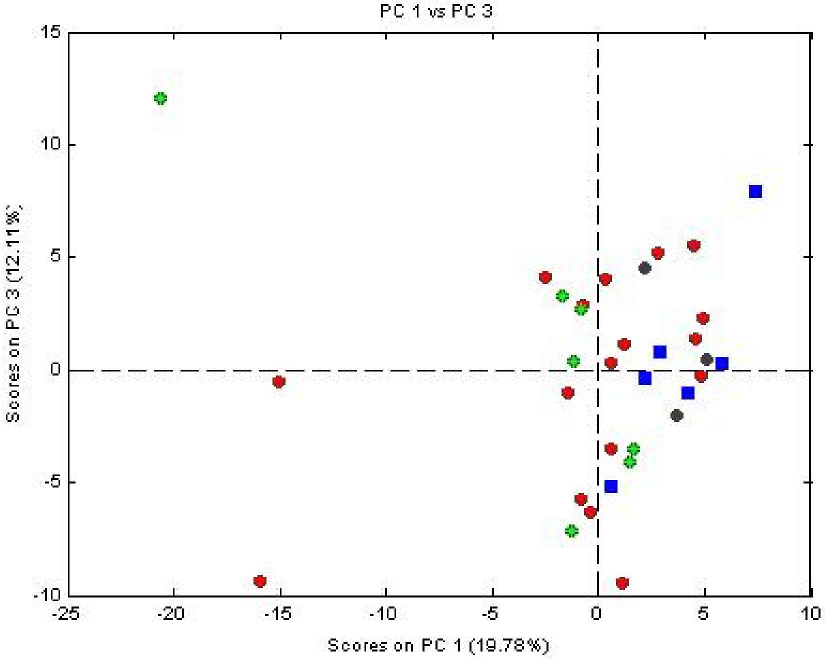

Figure 4.7: PC 1 versus PC 3

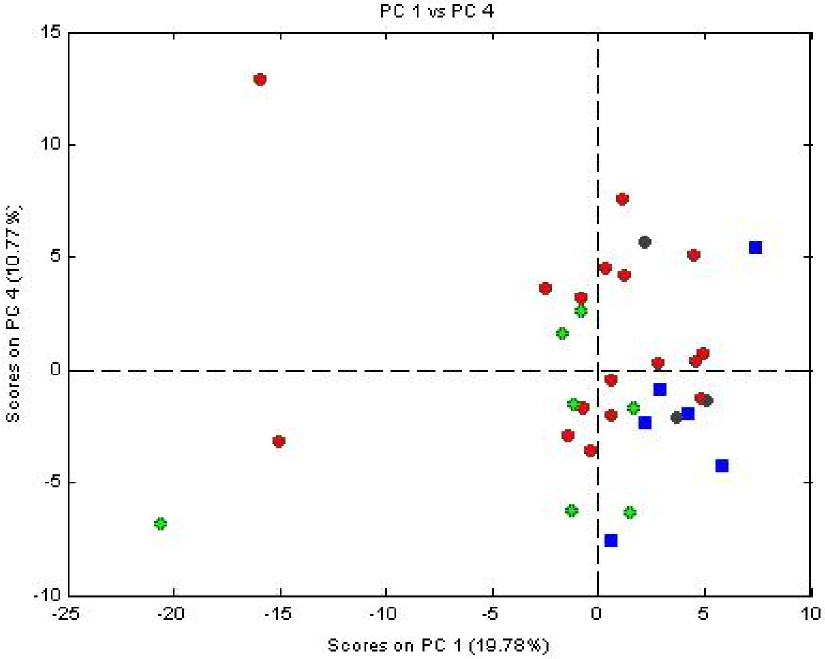

Figure 4.8: PC 1 versus PC 4

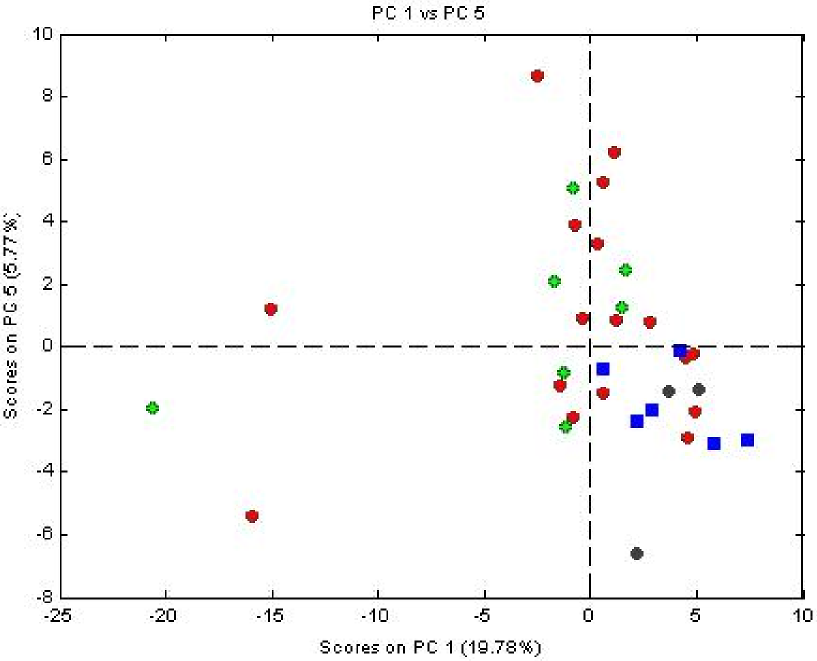

Figure 4.9: PC 1 versus PC 5

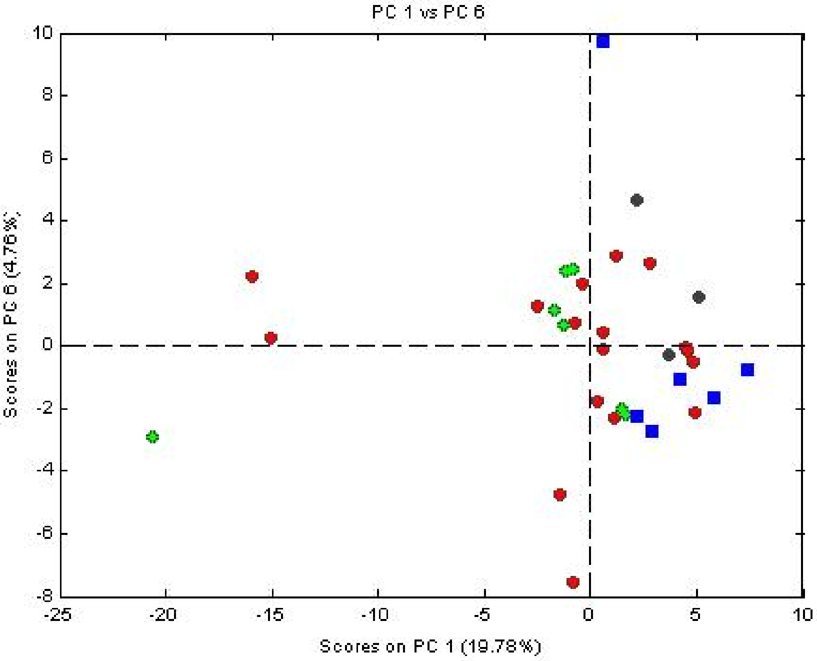

Figure 4.10: PC 1 versus PC 6

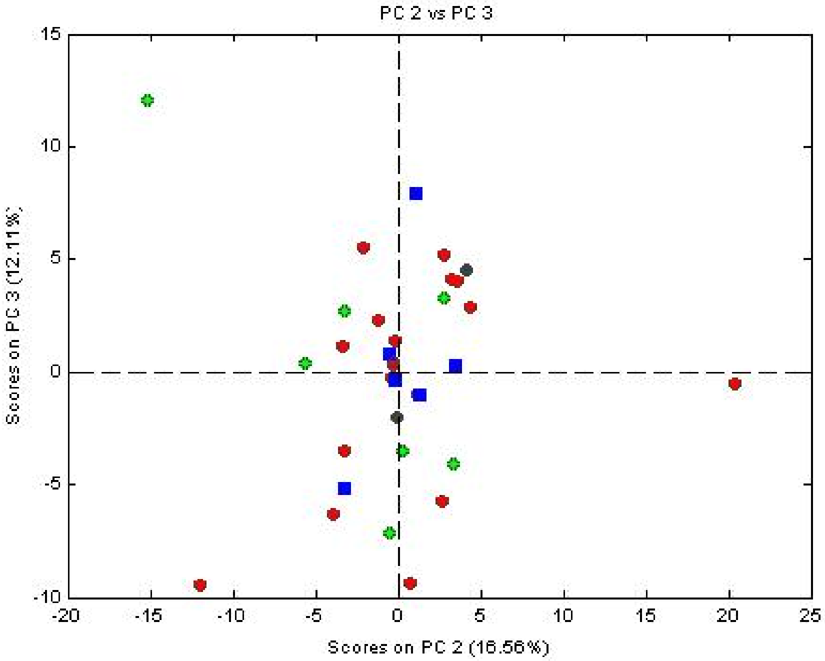

Figure 4.11: PC 2 versus PC 3

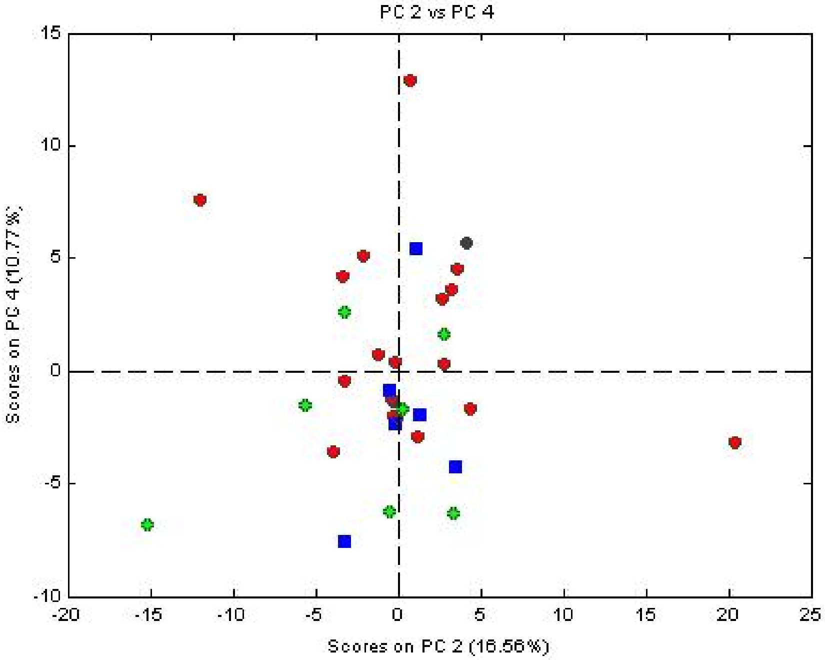

Figure 4.12: PC 2 versus PC 4

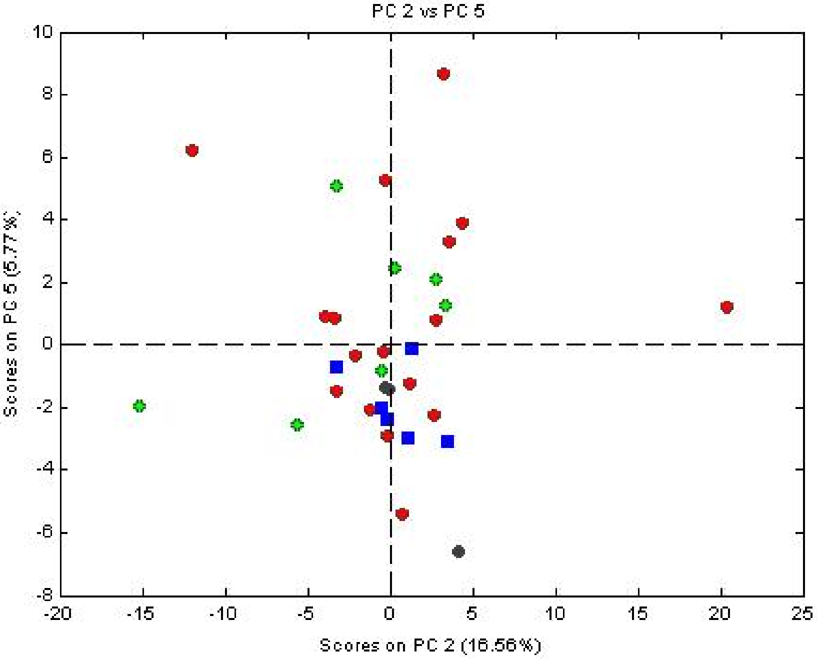

Figure 4.13: PC 2 versus PC 5

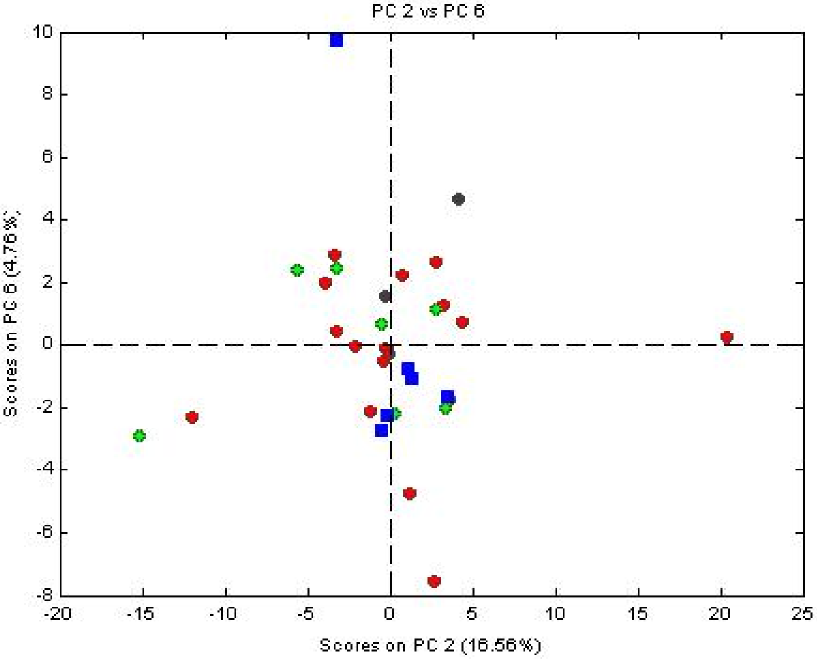

Figure 4.14: PC 2 versus PC 6

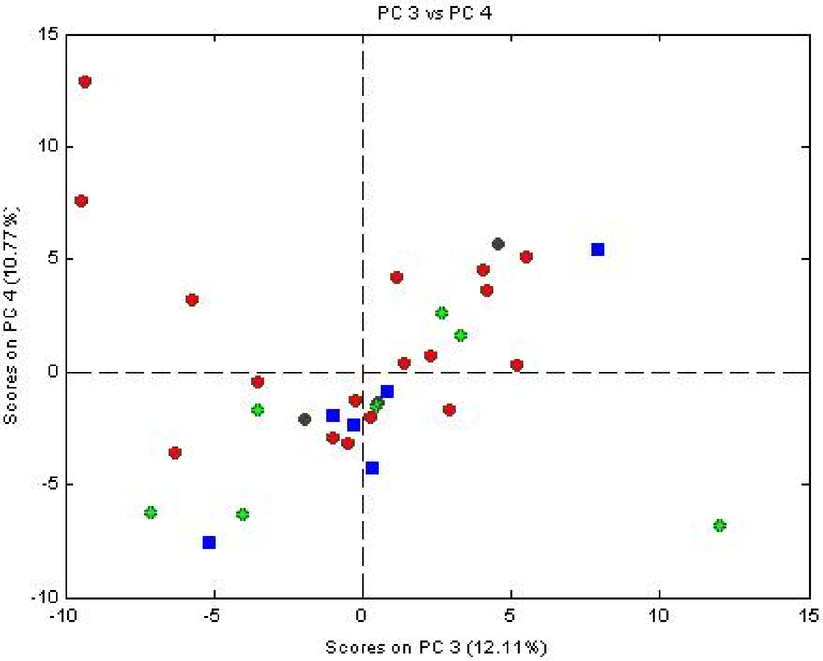

Figure 4.15: PC 3 versus PC 4

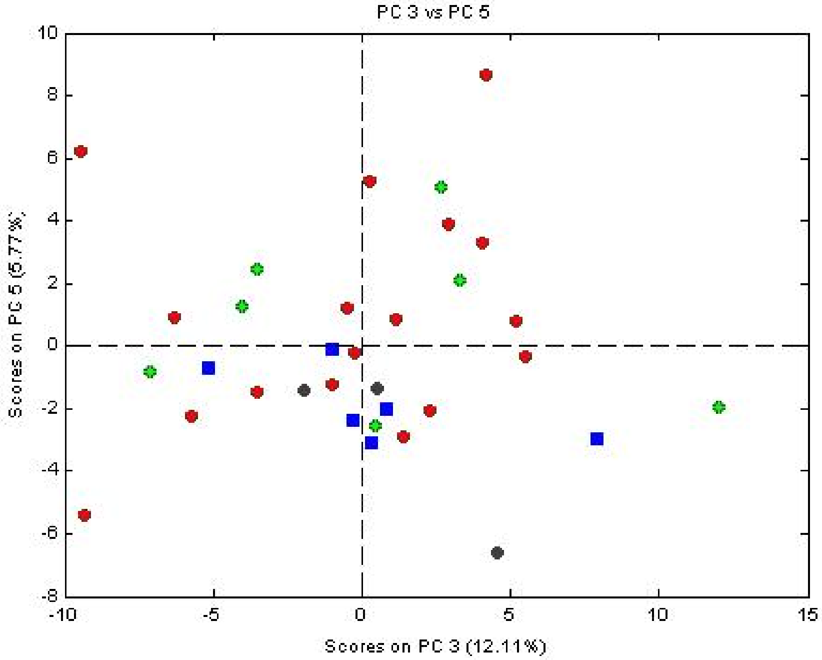

Figure 4.16: PC 3 versus PC 5

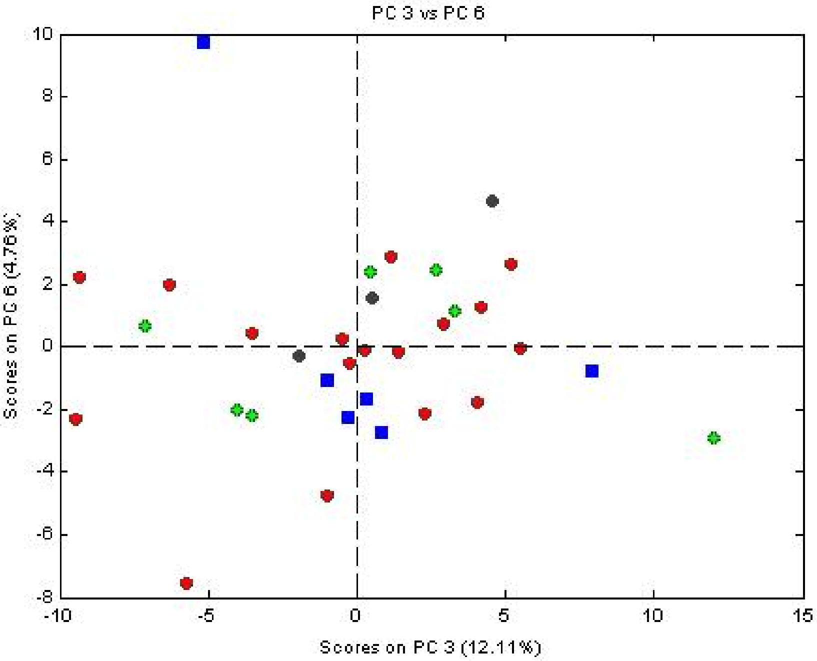

Figure 4.17: PC 3 versus PC 6

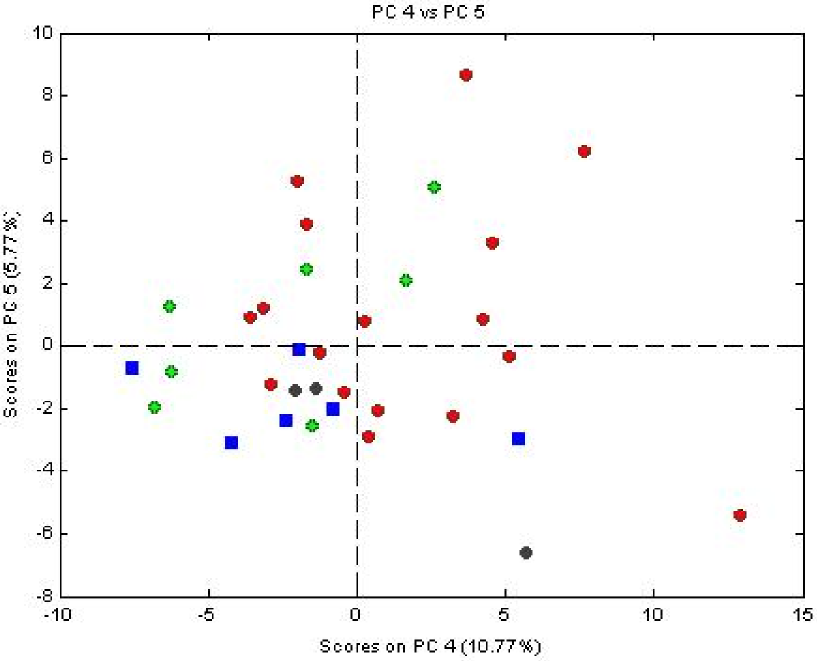

Figure 4.18: PC 4 versus PC 5

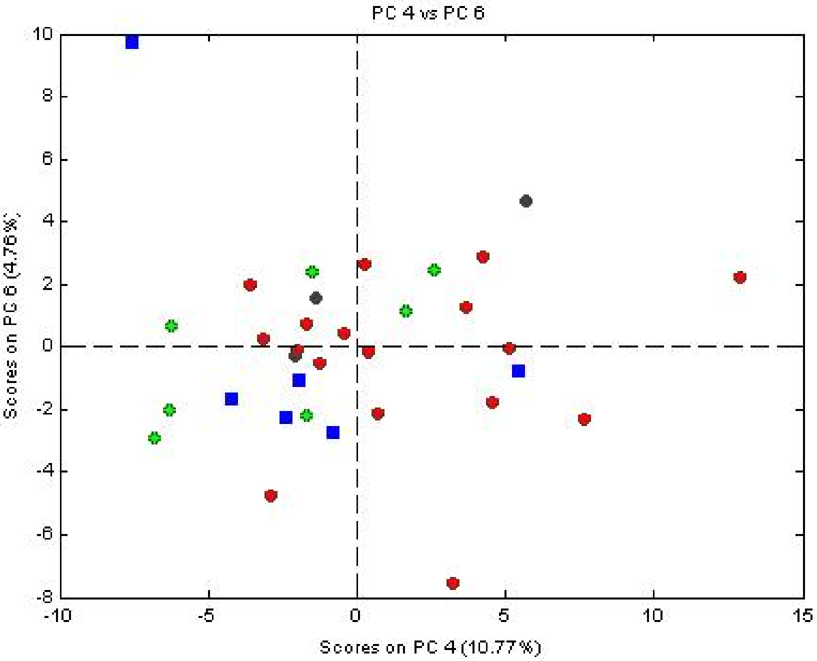

Figure 4.19: PC 4 versus PC 6

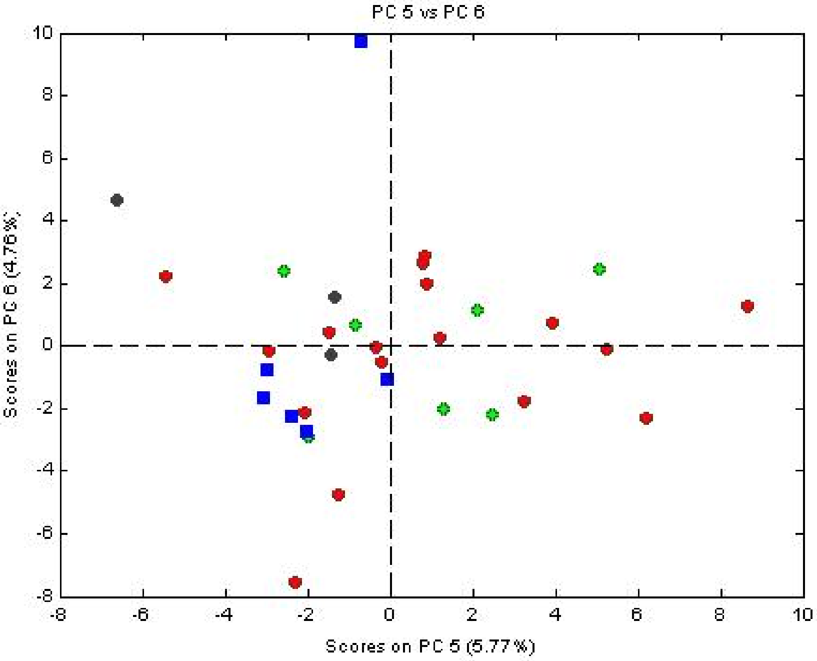

Figure 4.20: PC 5 versus PC 6

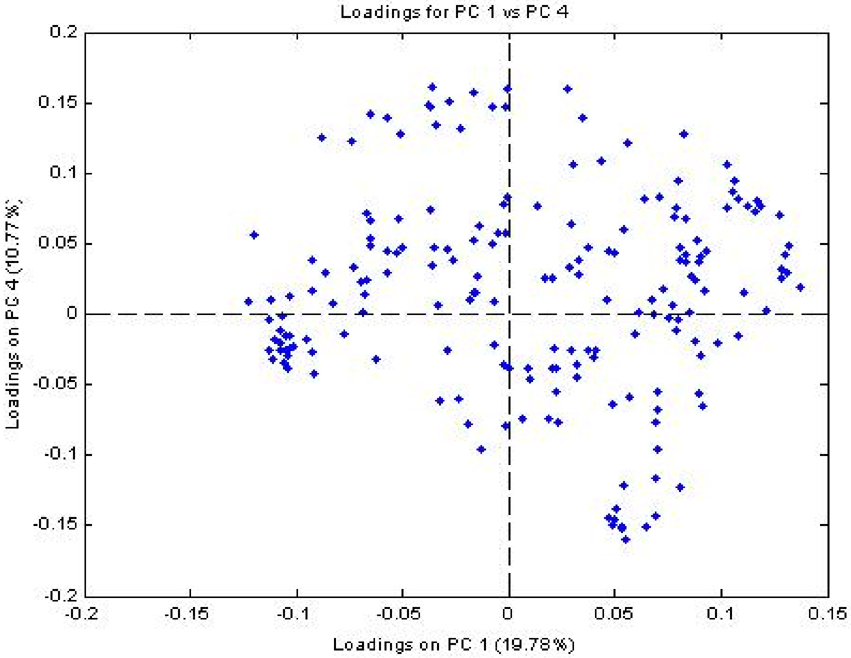

Figure 4.21: Loadings Plot for PC 1 versus PC 4

A Loadings Plot was created to show the loadings based on PC1 and PC4. This Loadings Plot, 4.21, also shows no tendency to significance of any of the variables.

4.4 Six Bin Analysis

PLS-DA is not appropriate in this case because the number of samples is so much smaller than the number of variables. A separate line of analysis using SVM has established that six of the bins are significant. PCA was run on the six bins and revealed clearer separation by class than the initial PCA.

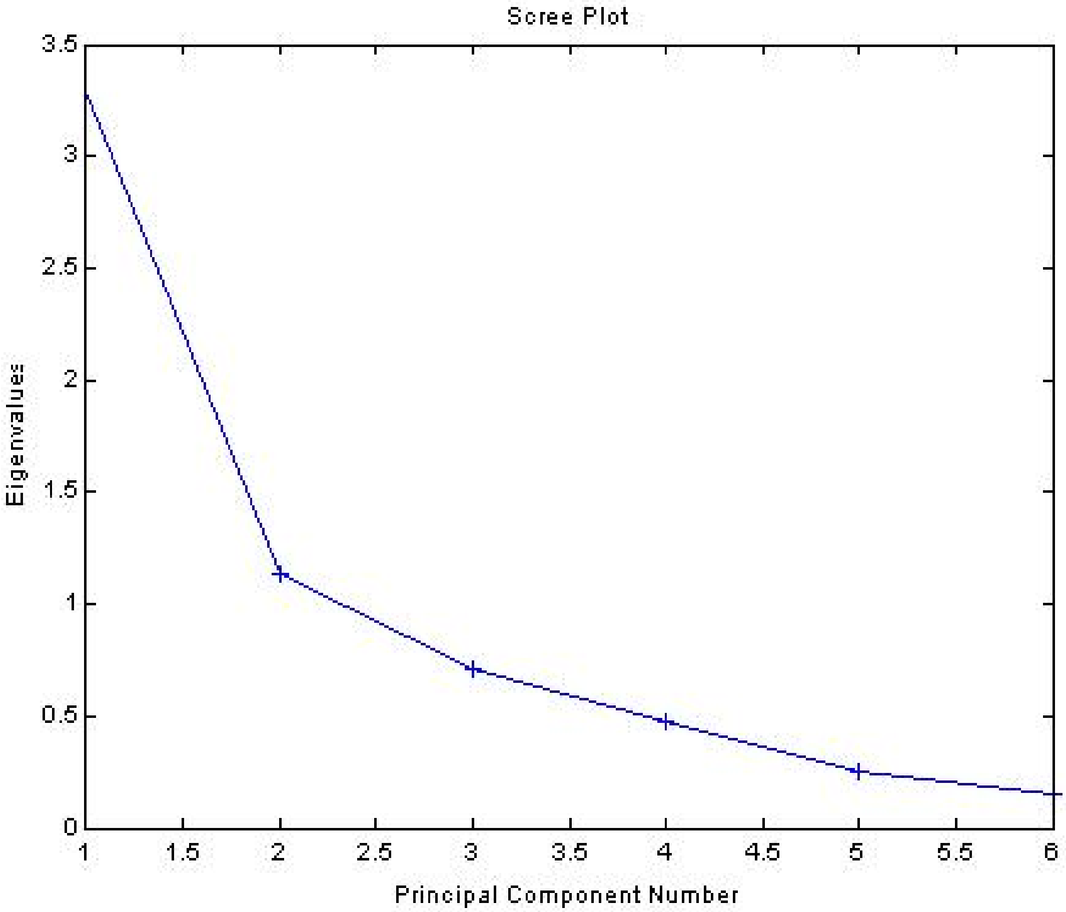

Figure 4.22: Scree Plot using only six variables

The Scree Plot for this analysis, shown in 4.22, reveals that all six principal components are significant. As was done before, these components were plotted against each other and evaluated for separation of data points.

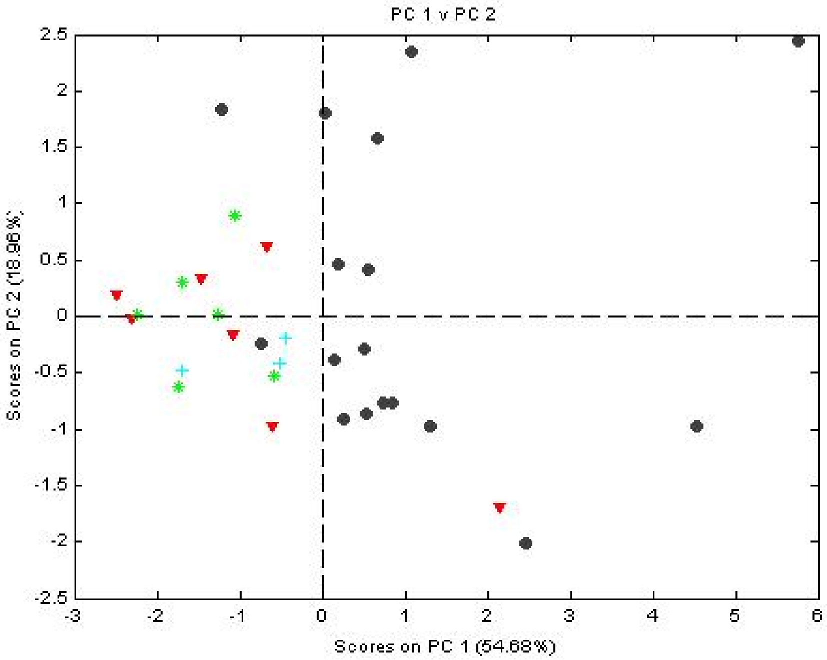

Figure 4.23: PC 1 versus PC 2

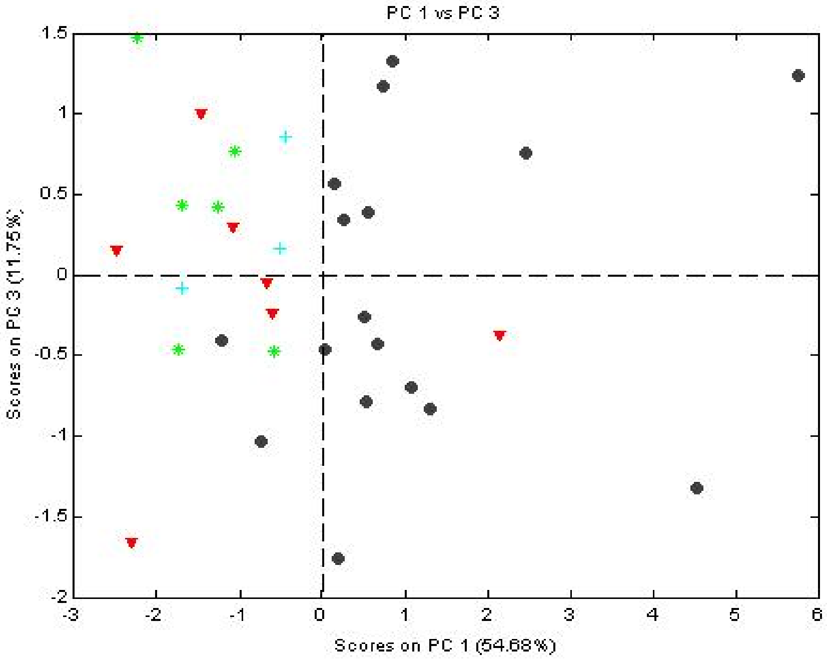

Figure 4.24: PC 1 versus PC 3

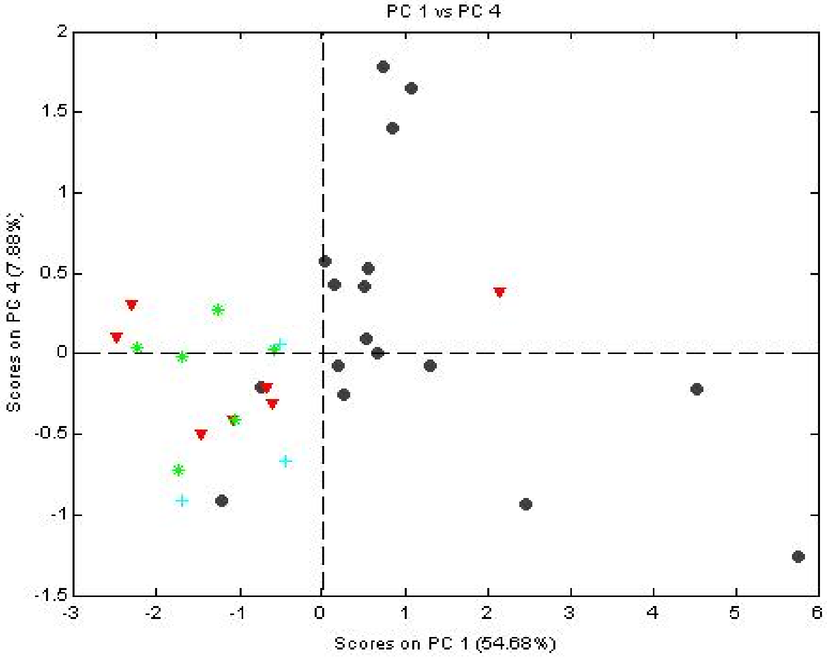

Figure 4.25: PC 1 versus PC 4

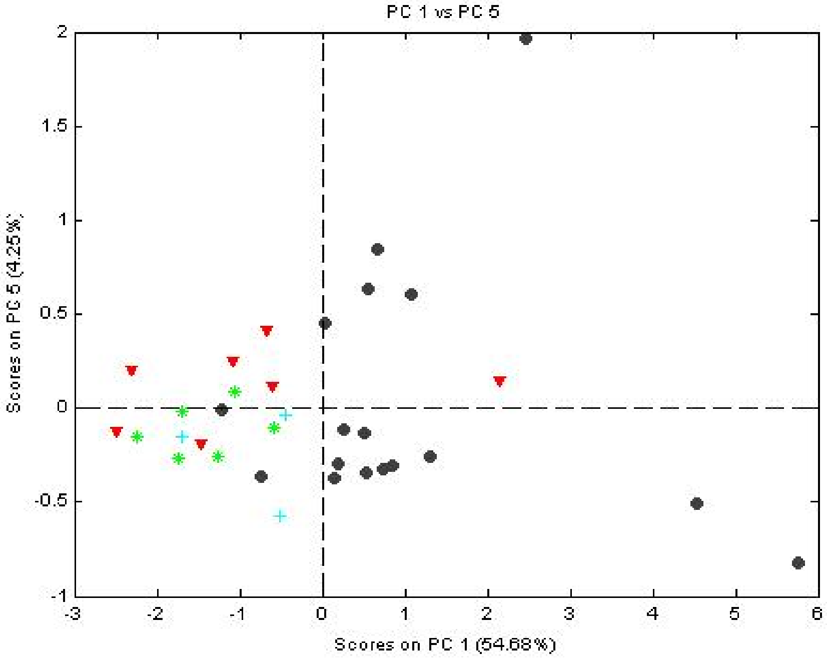

Figure 4.26: PC 1 versus PC 5

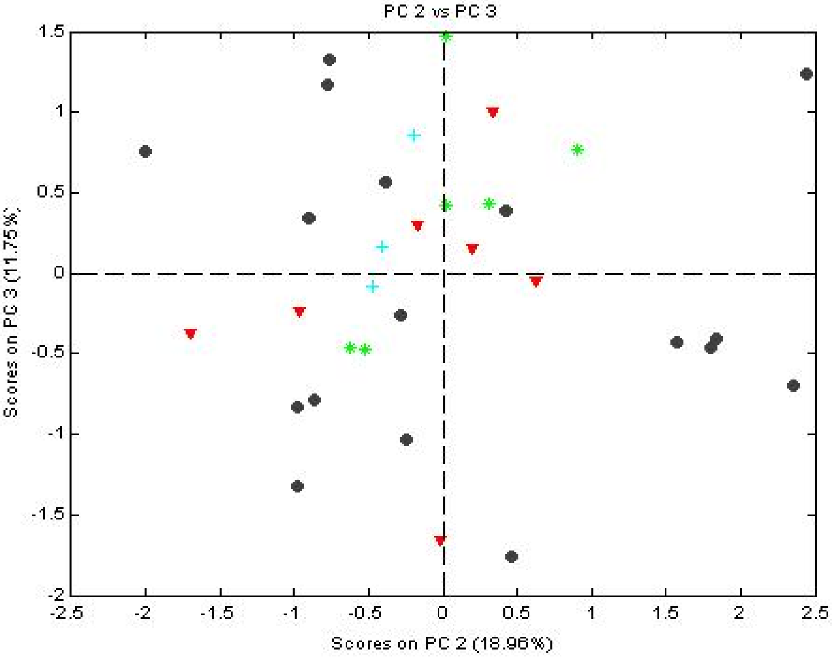

Figure 4.27: PC 2 versus PC 3

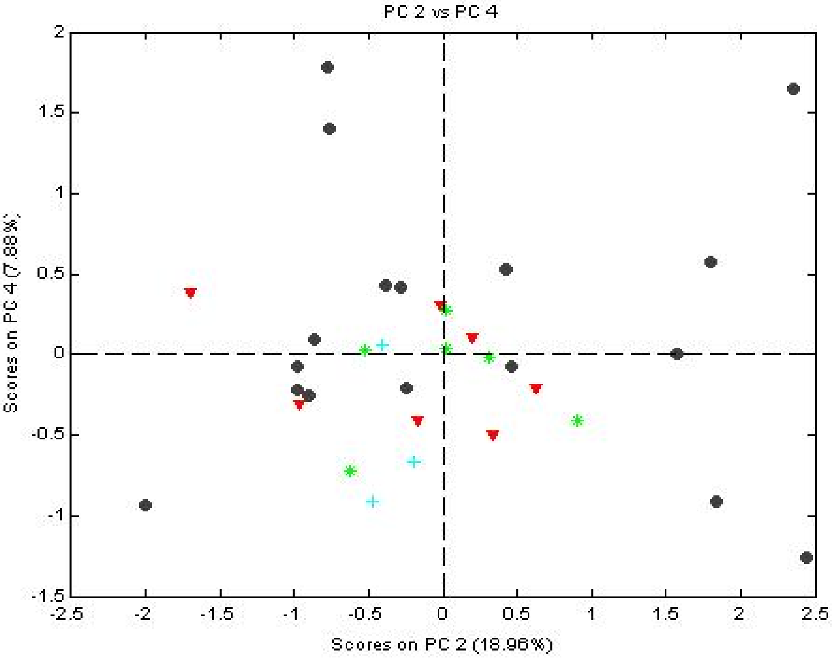

Figure 4.28: PC 2 versus PC 4

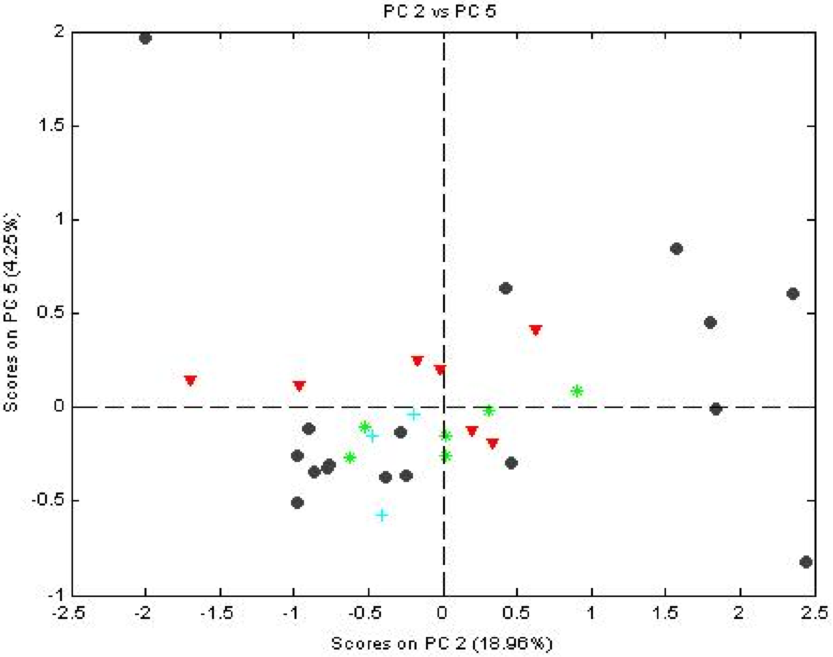

Figure 4.29: PC 2 versus PC 5

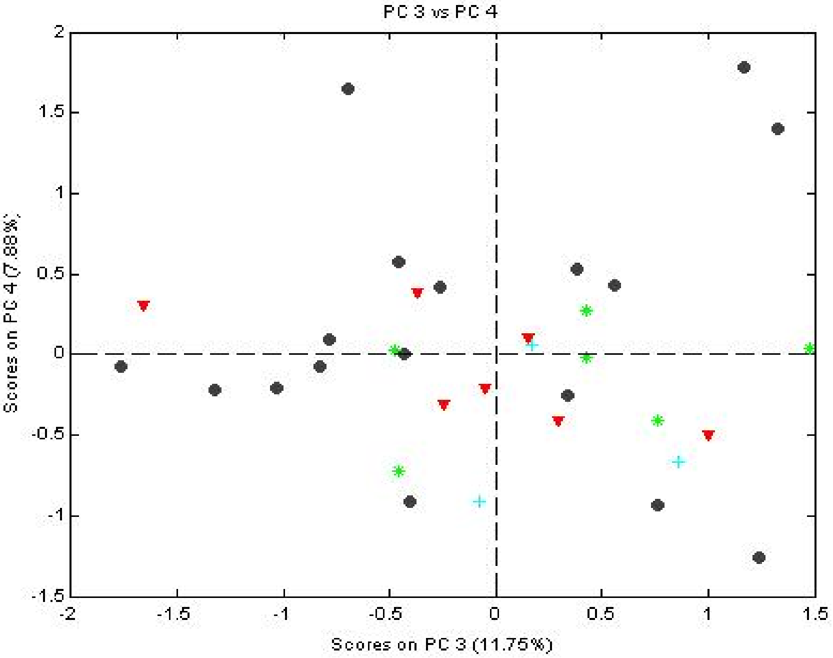

Figure 4.30: PC 3 versus PC 4

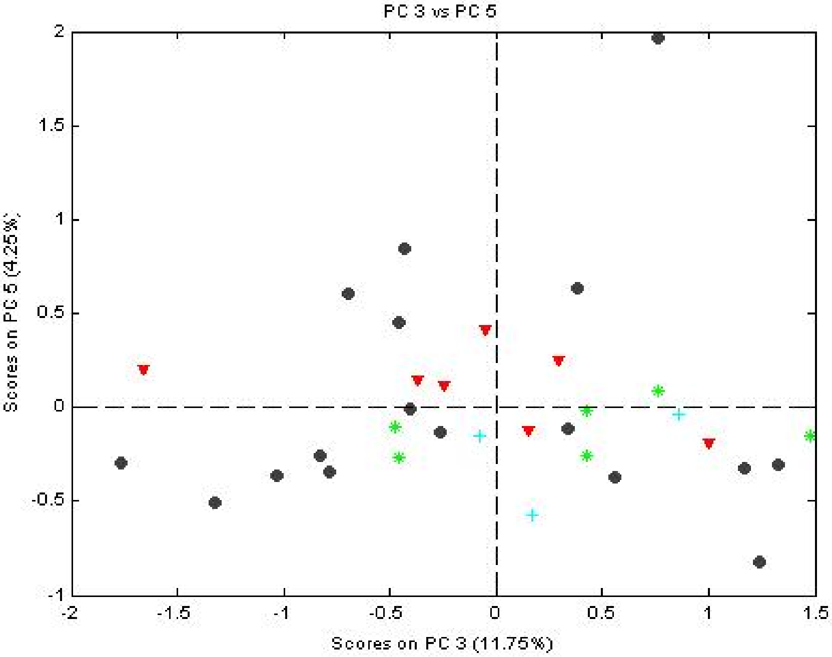

Figure 4.31: PC 3 versus PC 5

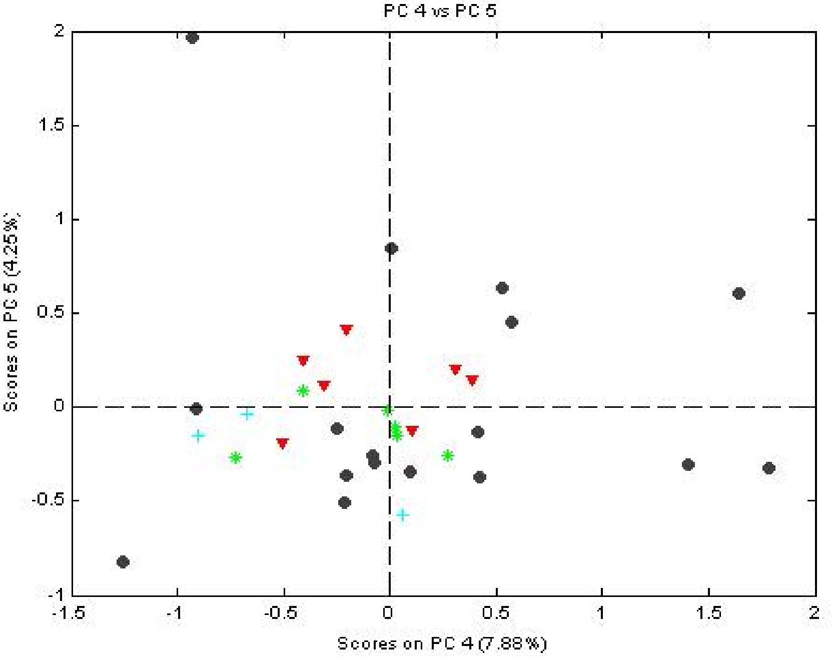

Figure 4.32: PC 4 versus PC 5

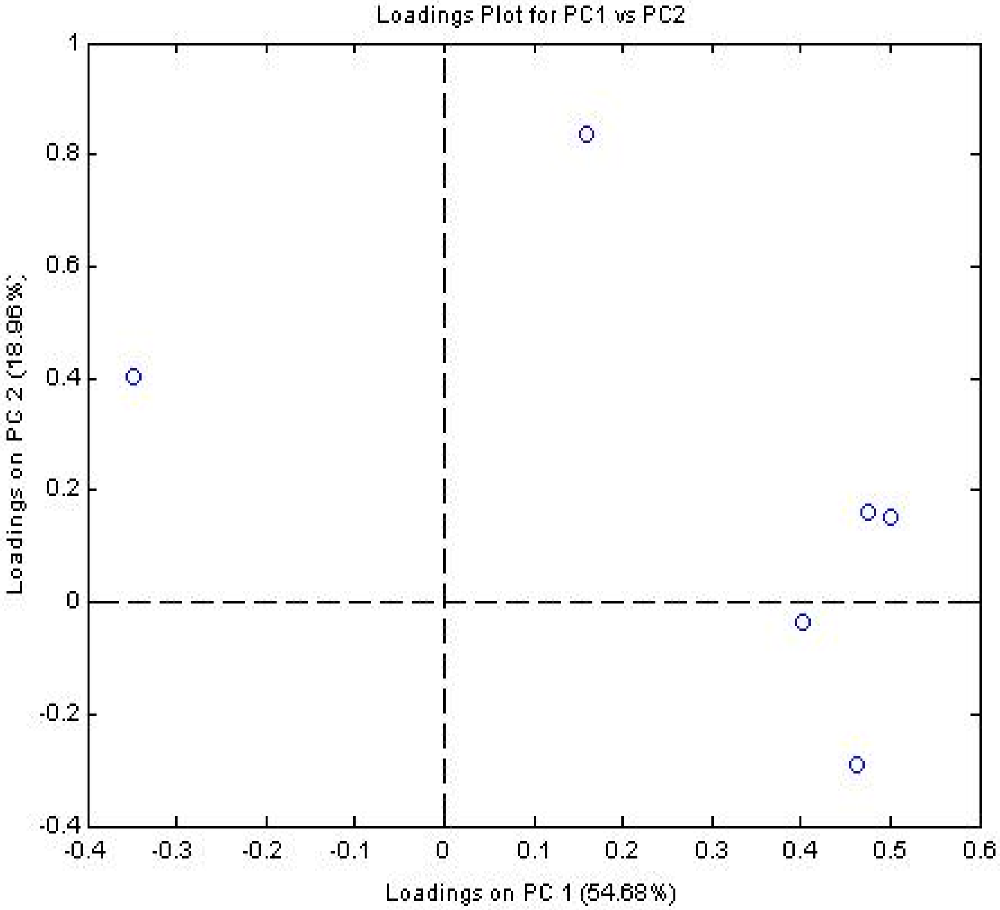

Figure 4.33: Loadings Plot for PC 1 and PC 2 using only six variables.

In this analysis, the best separation was shown by PC 1 and PC 2, 4.23. A Loadings Plot was created based on these components, shown in 4.33.

Chapter 5

Conclusions

Our results demonstrate that the one-pulse 1H NMR of mouse urine can be obtained within 30 minutes and then 240 peaks selected for input into the statistical algorithms. While the spectra themselves appeared subjectively to have differences according to the presence and stage of cancer in the subject, the correlations were too complex to be appreciated by the naked eye.

While PCA is currently the standard method of choice for metabolomic analyses, it proved to be wholly ineffective in this experiment. When the principal components were calculated and then plotted against one another, the data showed very little clustering at all. The method was too weak to sufficiently separate the data according to class, and this was probably due to the small number of samples relative to the number of variables. PLS-DA was not attempted because such an analysis would be even more unsuitable for such a small sample size. The question arises, is the lack of separation due to weakness in the method of statistical analysis or is NMR spectroscopy of urine not a sufficient vehicle for diagnosing colon cancer.

The SVM analysis revealed that six of the bins were more significant than the others. When PCA was run again using only those variables, eliminating the remaining bins, the data showed a much stronger tendency to cluster. The normal animals, with a cancer score of zero, were clearly separated from sick animals, with cancer scores of one or more. This tells us that the NMR spectra of urine does contain enough information to be useful in diagnosing cancer, but that PCA alone is not a sufficient analytical method for extracting the information.

This study is interesting because it does prove that the disease state of colon cancer results in characteristic changes in the small molecule metabolites secreted in urine. Further study is necessary to fully develop this technology and explore the use of new methods for multivariate analysis. In future experiments, it will be necessary to use a much larger data set, on the order of 500 samples, to yield better resolution. A larger sample group would result in a better predictive model, which could then be used to diagnose and potentially stage cancer in an individual not included in the training set. Such a model could be used in human clinical medicine as a non-invasive diagnostic tool, catching more cases of colon cancer at an early stage, increasing survival rates. Effective determination of the significance of each bin could also lead to potential biomarkers for colon cancer. Once these biomarkers have been identified, they may be explored for potential drug targets, which could lead to more effective, safer treatments for a serious human health threat.

Bibliography

Clarke, T (2002). Mice make medical history. *news@Nature(2002)*, December 2, 2002.

Denham, Michael C. (1994). Implementing Partial Least Squares *Statistics and Computing 1994*.

Gavaghan, CL. Holmes, E, Lenz, E Wilson, ID, Nicholson, JK (2000). An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences: application to the C57BL10J and Alpk:ApfCD mouse. *FEBS Letters 484(2000)*, 169-174.

Holmes, E., Nichohlls, A. Lindon, J.C. Connor, S.C. Connelly, J Haselden, J. Damment, S. J. P. Spraul, M. Neidig, P. and Nicholson, J K. (2000). Chemometric Models for Toxicity Classification Based on NMR Spectra of Biofluids. *Chem. Res. Toxicol.*, 13, 471-478.

Hoskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics 1988*, Volume 2, Number 3, pages 211–228.

Keun, HC, Ebbels, T, Antti, H, Bollard, M, Beckonert, O, Schlotterbeck, G, Senn, H, Niederhauser, U, Holmes, E, Lindon, J, and Nicholson, JK. (2002). Analytical Reproducibility in 1H NMR-Based Metabonomic Urinalysis. *Chem. Res. Toxicol.*, 15, 1380-1386.

Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, Feuer EJ, Thun MJ. (1995). Cancer statistics. *Cancer J Clin 2005*, pages 10–30.

Nelson, J. H. (2003). Nuclear Magnetic Resonance Spectroscopy. Upper Saddle River, NJ, Prentice Hall.

Nicholson JK, Connelly J, Lindon JC, Holmes E. (2002). Metabonomics: a platform for studying drug toxicity and gene function. *Nature Rev Drug Discov* 1: 153

Liu, M, Nicholson JK, Lindon JC. (1996). High-Resolution Diffusion and Relaxation Edited One- and Two-Dimensional 1H NMR Spectroscopy of Biological Fluids. *Anal. Chem.*, 68, 3370-3376.

Otto, Matthias. (1999). Chemometrics: statistics and computer application in analytical chemistry. Weinhim; New York.

Tannock IF, Hill RP et al (eds) (2005). The Basic Science of Oncology. McGraw-Hill.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, Chapter 5, New York.

Yeung and Ruzzo (2001). Principal component analysis for clustering gene expression data. *Bioinformatics* 17(9): 763-74.