Computational Analysis of Metabolomic Toxicological Data Derived from NMR Spectroscopy

Benjamin J. Kelly
Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all
Part of the Computer Sciences Commons

Repository Citation
Kelly, Benjamin J., "Computational Analysis of Metabolomic Toxicological Data Derived from NMR Spectroscopy" (2009). Browse all Theses and Dissertations. 275.
https://corescholar.libraries.wright.edu/etd_all/275

This Thesis is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu, library-corescholar@wright.edu.
Computational Analysis of Metabolomic Toxicological Data Derived from NMR Spectroscopy

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

By

BENJAMIN J. KELLY
B.S. Biological Sciences and Chemistry, Wright State University, 2005

2009
Wright State University
I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY BENJAMIN J. KELLY ENTITLED COMPUTATIONAL ANALYSIS OF METABOLOMICS TOXICOLOGICAL DATA DERIVED FROM NMR SPECTROSCOPY BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE.

_____________________________________
Michael L. Raymer, Ph.D.
Thesis Director

_____________________________________
Thomas A. Sudkamp, Ph.D.
Department Chair

Committee on
Final Examination

_____________________________________
Michael L. Raymer, Ph.D.

_____________________________________
Travis E. Doom, Ph.D.

_____________________________________
Nicholas V. Reo, Ph.D.

_____________________________________
Joseph F. Thomas, Jr., Ph.D.
Dean, School of Graduate Studies
Computational Analysis of Metabolomic Toxicological Data Derived from NMR Spectroscopy.

Nuclear magnetic resonance (NMR) spectroscopy is a non-invasive method of acquiring metabolic profiles from biofluids. The most informative metabolomic features, or biomarkers, may provide keys to the early detection of changes within an organism such as those that result from exposure to a toxin. One major difficulty with typical NMR data, whether it come from a toxicological, medical or other source, is that it features a low sample size relative to the number of variables measured. Thus, traditional pattern recognition techniques are not always feasible. The "curse of dimensionality" is an important consideration in selecting appropriate statistical and pattern recognition methods for the identification of potential biomarkers.

In this thesis, several alternatives for isolating biomarkers are evaluated on NMR-derived toxicological data set and results are compared: the fold test, univariate ranking, the unpaired t-test, and the paired t-test are examined. Potential biomarkers were inspected for differences based on several subjective criteria including ability to identify consistent differences between treatment and control samples and distinguish potential vehicle effects, those effects caused by the method of delivery performed on both treated and control animals.

Based on these results, the paired t-test method is preferred, due to its ability to attribute statistical significance, to take into consideration consistency of a single subject over a time course, and to mitigate the low sample, high dimensionality problem. A protocol for the paired t-test is also proposed to remove potential vehicle effects and identify toxic responses above the vehicle effects.

Due to the large number of variables to be considered, a correction for multiple testing must be employed. In this thesis, several methods of correction for multiple test is evaluated. An acceptable p-value cutoff for each correction is proposed so that the most appropriate correction can be applied based on the purpose of the metabolomic toxicology experiment.

Also in this thesis, a more complex method for identifying biomarkers, Orthogonal Projection to Latent Structures Discriminator Analysis (O-PLS-DA), is compared to the t-test using synthetic data sets based on the characterization of experimental NMR spectra. The ranking of potential biomarkers produced by both methods is compared to the ranking of features used to create the synthetic data. In addition, an O-PLS-DA
A permutation test method of determining an important feature cutoff is evaluated using the synthetic data. The variable-at-a-time t-test method using a p-value threshold is also evaluated for comparison. Based on these results the O-PLS-DA permutation test was not consistent or stable enough to distinguish truly responding biomarkers.

The benefits of O-PLS-DA, including its ability to deal with correlated variables, removal of unwanted systematic variation, and the ability to deal with some amount of missing data, make it sufficient for identifying potential biomarkers. It is determined that O-PLS-DA does not rank potential biomarkers differently than the t-test nor does it classify new samples significantly better or worse than a majority-vote based t-test classifier.
# Contents

1 Introduction ................................................. 1
   1.1 Overview .................................................. 1
   1.2 Contributions ........................................... 2

2 Background and Literature Survey ......................... 4
   2.1 Background .................................................. 4
      2.1.1 Genes, Proteins, and Metabolites ...................... 4
      2.1.2 Toxicology ............................................ 7
      2.1.3 Genomics and Toxicology .............................. 9
      2.1.4 Proteomics and Toxicology ............................ 10
      2.1.5 Metabolomics and Toxicology ......................... 11
      2.1.6 Nuclear Magnetic Resonance (NMR) Spectroscopy ...... 12
         2.1.6.1 NMR-based Metabolomics .......................... 15
   2.2 Literature Survey ......................................... 16
      2.2.1 NMR Analysis .......................................... 16
         2.2.1.1 Partial Least Squares Regression .................. 19
         2.2.1.2 Orthogonal Projection to Latent Structures ....... 26
      2.2.2 Microarray Analysis .................................... 32
      2.2.3 Multiple Test Corrections .............................. 35
      2.2.4 Validation ............................................. 38
         2.2.4.1 Cross-validation .................................... 38
         2.2.4.2 Permutation Test ................................... 41
         2.2.4.3 Randomization t-test ............................... 42

3 Preliminary Biomarker Feature Identification Analyses .. 44
   3.1 Experimental Data Set .................................... 44
      3.1.1 Animals ............................................... 44
      3.1.2 Urine Sample Preparation ............................. 45
      3.1.3 1H NMR Spectroscopy .................................. 46
      3.1.4 Data Preprocessing .................................... 46
   3.2 Methods ................................................... 47
      3.2.1 Fold Test .............................................. 47
      3.2.2 The t-test ............................................. 48
         3.2.2.1 Removal of Vehicle Effect ......................... 49
3.2.3 Multiple Test Correction Comparison .......................... 50
  3.2.3.1 Bonferroni Correction ......................................... 50
  3.2.3.2 Holm’s Step-down Correction ................................. 52
  3.2.3.3 Benjamini-Hochberg Correction .............................. 53
  3.2.4 Visualizing Bins .................................................. 54

3.3 Results ................................................................. 55
  3.3.1 Comparison of Methods .......................................... 55
    3.3.1.1 Removal of Vehicle Effects ................................. 57
  3.3.2 Comparison of Multiple Test Corrections ....................... 59

3.4 Discussion ............................................................ 63
  3.4.1 Comparison of Methods .......................................... 63
  3.4.2 Comparison of Multiple Test Corrections ....................... 68

3.5 Conclusion ............................................................ 75

3.6 Computational Hardware and Software ............................ 76

3.7 Data Acquisition Acknowledgement .................................. 77

4 Biomarker Feature Selection Method Analyses ....................... 78
  4.1 Synthetic Data Sets .................................................. 78
    4.1.1 Characterization of Real-World Spectra ....................... 79
    4.1.2 Creation of Synthetic Control and Treatment Data Sets ....... 79
  4.2 Methods ............................................................... 81
    4.2.1 Data Preprocessing .............................................. 81
    4.2.2 Comparison as a Ranking Method ............................... 82
    4.2.3 Comparison as a Method of Determining Important Variables .. 84
    4.2.4 Comparison as a Classifier .................................... 89
  4.3 Results ................................................................. 91
    4.3.1 Comparison as a Ranking Method ................................ 91
    4.3.2 Comparison as a Method of Determining Important Variables .. 93
    4.3.3 Comparison as a Classifier .................................... 95
  4.4 Discussion ............................................................ 96
    4.4.1 Comparison as a Ranking Method ................................ 96
    4.4.2 Comparison as a Method of Determining Important Variables .. 99
    4.4.3 Comparison as a Classifier .................................... 100
  4.5 Conclusion ............................................................ 101
  4.6 Computational Hardware and Software ............................ 104

5 Conclusion ............................................................... 105
  5.1 Summary ............................................................... 105
  5.2 Contributions ........................................................ 107
  5.3 Future Work .......................................................... 107
    5.3.1 Further Exploration of O-PLS-DA-based Permutation Test ....... 108
    5.3.2 Expansion of O-PLS-DA ........................................... 108
    5.3.3 Expanded Real-life Data Sets ................................... 108
    5.3.4 Additional Biomarker Identification Methods .................. 109
    5.3.5 Additional Preprocessing Methods .............................. 109

Bibliography ............................................................. 111
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>DNA base pairs and the double helix.</td>
<td>5</td>
</tr>
<tr>
<td>2.2</td>
<td>Single stranded RNA and double stranded DNA.</td>
<td>6</td>
</tr>
<tr>
<td>2.3</td>
<td>The four levels of protein structure.</td>
<td>8</td>
</tr>
<tr>
<td>2.4</td>
<td>The informational flow of the three in biological processes and their respective areas of study.</td>
<td>9</td>
</tr>
<tr>
<td>2.5</td>
<td>A NMR spectrum.</td>
<td>13</td>
</tr>
<tr>
<td>2.6</td>
<td>A Lorentzian peak.</td>
<td>14</td>
</tr>
<tr>
<td>2.7</td>
<td>A graphical representation of the PCA loadings (A) and scores (B).</td>
<td>17</td>
</tr>
<tr>
<td>2.8</td>
<td>Permutation plot of $R^2$ or $Q^2$ for random class assignments.</td>
<td>41</td>
</tr>
<tr>
<td>3.1</td>
<td>Example of time course plot</td>
<td>55</td>
</tr>
<tr>
<td>3.2</td>
<td>Example of mean with error bar plot</td>
<td>56</td>
</tr>
<tr>
<td>3.3</td>
<td>Example of a time point removed due to vehicle effects</td>
<td>58</td>
</tr>
<tr>
<td>3.4</td>
<td>Example of a time point not removed due to vehicle effects because of an response beyond the apparent vehicle effect</td>
<td>60</td>
</tr>
<tr>
<td>3.5</td>
<td>PCA goodness plot of p-values with no multiple test correction</td>
<td>61</td>
</tr>
<tr>
<td>3.6</td>
<td>PCA goodness plot of p-values with Bonferroni correction</td>
<td>62</td>
</tr>
<tr>
<td>3.7</td>
<td>PCA goodness plot of p-values with Holm’s correction</td>
<td>62</td>
</tr>
<tr>
<td>3.8</td>
<td>PCA goodness plot of p-values with Benjamini-Hochberg correction</td>
<td>63</td>
</tr>
<tr>
<td>3.9</td>
<td>Example of bin/time point significant by two-fold test</td>
<td>65</td>
</tr>
<tr>
<td>3.10</td>
<td>Example of a response beyond a possible vehicle effect</td>
<td>66</td>
</tr>
<tr>
<td>3.11</td>
<td>Example of missed vehicle effect by the unpaired t-test</td>
<td>67</td>
</tr>
<tr>
<td>3.12</td>
<td>Example of paired t-test identifying consistency among subjects</td>
<td>69</td>
</tr>
<tr>
<td>3.13</td>
<td>Example 1 of a bin/time point considered significant by all three methods</td>
<td>70</td>
</tr>
<tr>
<td>3.14</td>
<td>Example 2 of a bin/time point considered significant by all three methods</td>
<td>71</td>
</tr>
<tr>
<td>3.15</td>
<td>Example 3 of a bin/time point considered significant by all three methods</td>
<td>72</td>
</tr>
<tr>
<td>3.16</td>
<td>Example 4 of a bin/time point considered significant by all three methods</td>
<td>73</td>
</tr>
<tr>
<td>4.1</td>
<td>Permutation test plot of data with no responding variables</td>
<td>87</td>
</tr>
<tr>
<td>4.2</td>
<td>Permutation test plot of data with 10 responding variables</td>
<td>88</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Comparing the differentially responsive bin results from the paired t-test (PTT), unpaired t-test (UTT), and two-fold test (2F) ................................................................. 57
3.2 Number of bin/time points significant at different p-value thresholds .......................... 59
4.1 Results from ranking comparison of O-PLS-DA loadings and t-test p-values on binned synthetic data .......................................................... 92
4.2 Results from ranking comparison of O-PLS-DA loadings and t-test p-values on peak synthetic data .......................................................... 93
4.3 Results from O-PLS-DA permutation test and t-test p-value cutoff test on binned synthetic data .......................................................... 94
4.4 Results from O-PLS-DA permutation test and t-test p-value cutoff test on peak synthetic data .......................................................... 95
4.5 Results from classification comparison of the O-PLS-DA and majority vote-based classifiers on binned synthetic data .......................................................... 96
4.6 Results from classification comparison of the O-PLS-DA and majority vote-based classifiers on peak synthetic data .......................................................... 97
I would like to thank my advisors, Dr. Michael Raymer and Dr. Travis Doom, and my remaining committee member, Dr. Nicholas Reo, for their assistance, guidance, and support throughout this entire process. A special thanks also goes out to Dr. Nicholas DelRaso from the Biosciences and Protection Division of the Air Force Research Laboratory at Wright-Patterson Air Force Base.

I would also like to extend my gratitude to the Dayton Area Graduate Studies Institute (DAGSI) for their financial assistance.

Additionally, I would like to thank the members of Dr. Reo’s lab at the Department of Biochemistry and Molecular Biology and the Boonshoft School of Medicine including Mike Kent, Deirdre Mahle, Dan Homer, Andy Neuforth, and Bill Couch.

Finally my current and former associates in the Bioinformatics Research Group (BiRG) deserve a great deal of credit for their assistance and encouragement. Those associates include Amanda Hanes, Doug Raiford, Al Herner, C.J. Maynard, Michael Peterson, and Gina Cooper. A special thanks to lab associate Paul Anderson for his work on the synthetic data and his ideas on the permutation test used in this thesis.
Dedicated to
my wife and family
for their love and encouragement.
Introduction

1.1 Overview

Early detection of systemic changes within an organism, whether due to disease, treatment, or something as simple as a change in diet, is an important component in monitoring and protecting the well-being of the organism. Non- or low-invasive options to identify variations in systemic tissue or organ function can be valuable tools for a variety of medical and life science applications.

Biomarkers, or measurable items that indicate changes within an organism, offer a chance at early detection of disease or treatment. Biomarkers can take the form of genes, proteins, metabolic products or intermediates, or any other measurable biomolecule. One increasingly studied group of biomarkers is metabolites. Metabolites, products of basic life processes within cells, that behave differently due to stresses or treatments can be measured from biofluids using nuclear magnetic resonance (NMR). Toxicology is one field in which metabolic biomarkers are becoming more and more relevant.

The problem of dimensionality has been identified as a complication when attempting isolate important variables, such as biomarkers (Raudys and Jain, 1991). For NMR analyses, the spectrum is often decomposed into a series of variables, each associated with the concentration of one or
1.2. CONTRIBUTIONS

more metabolites. As a result, traditional pattern recognition techniques such as linear discriminant analysis may be misleading due to noise and collinearity among the variables.

Toxicology experiments commonly have the benefit of controls, internal, external, or both. Separate organisms are almost always kept with every parameter except the parameter of interest remaining constant as an external control. In addition, the treated organisms are commonly measured before a treatment is administered as an internal or self control. These controls have the potential to help identify responses attributable to the vehicle in which the toxin or treatment was delivered into the body or environmental factors such as temperature, as opposed to a response due to the toxin or treatment itself.

1.2 Contributions

In this thesis, several statistical and pattern recognition alternatives will be evaluated on an NMR-derived toxicological data set. The fold test, utilized in microarray data analysis, along with the Student’s t-test, used in both NMR and microarray data analysis, will be performed on the same data set and results will be compared based on several subjective criteria. These tests are performed on each variable individually, thus mitigating the problems of high dimensionality and low sample size. The performance of the methods based on these criteria will be described and characteristics preferred by one method over another will be noted. Several options for correcting for multiple tests will also be investigated.

These variable-at-a-time methods are contrasted with tests like linear discriminant analysis (LDA), principal component analysis (PCA), and both projection to latent structures discriminant...
1.2. CONTRIBUTIONS

analysis (PLS-DA) and orthogonal projection to latent structures discriminant analysis (O-PLS-DA), two of the most common methods of NMR analysis used today. These methods look at the entire set of variables at once. Synthetic data sets created using the characterization of real-life NMR spectra (Anderson et al., in preparation to be published), will be utilized to evaluate O-PLS-DA and the t-test. Feature rankings created by O-PLS-DA and the t-test will be compared to the true feature ranking used to create the synthetic spectra. The classification accuracy of the synthetic spectra will also be evaluated for O-PLS-DA and a majority-vote based classifier based on the t-test.

In addition, an O-PLS-DA permutation test for determining an acceptable cutoff for responding features will be performed. This cutoff will be compared to the true number of responding features used to create synthetic data sets. An alternate method involving a t-test’s p-value as a cutoff will also be included for comparison.

This thesis assumes a knowledge of commonly employed techniques such as the Student’s t-test, PCA, and LDA. Formulas for these techniques were omitted from this thesis, but are included in the associated citations. Assumptions and parameters will be discussed, as the focus of this thesis is to bring as much information to someone attempting to analyze similar data sets so that these techniques may be reproduced.
Background and Literature Survey

2.1 Background

2.1.1 Genes, Proteins, and Metabolites

Genes are the elements of heredity, determining what inherited traits a parent will transmit to the offspring during reproduction. They are encoded within double stranded, double-helix deoxyribonucleic acid (DNA) molecules, the building blocks of life (Watson and Crick, 1953). Each strand is made up of a lengthy sequence, or polymer, of four possible DNA nucleotides: guanine (G), adenine (A), thymine (T), and cytosine (C). The opposite strand’s sequence is fixed and complementary, based on the fact that guanine binds only with cytosine and adenine binds only with thymine. These bindings, formed by relatively weak hydrogen bonds, form what are referred to as base pairs (Figure 2.1) (Hartl and Jones, 2005, pg. 9).

The DNA sequence that makes up a gene is transcribed into single stranded ribonucleic acid (RNA). RNA is likened to single stranded DNA, except that the nucleotide thymine is replaced with uracil (U) (Figure 2.2). Triplicates of RNA nucleotides, called codons, are translated into amino acids, which are pieced together and folded around each other to form proteins (Figure
2.1. BACKGROUND

Figure 2.1: DNA base pairs and the double helix.

Adapted from (AccessExcellence, 2007).
URL: http://www.accessexcellence.org/RC/VL/GG/dna2.html
2.1. BACKGROUND

Figure 2.2: Single stranded RNA and double stranded DNA.

Adapted from (AccessExcellence, 2007).
URL: http://www.accessexcellence.org/RC/VL/GG/rna.html
2.1. **BACKGROUND**

2.3). Proteins perform the functions of cells. The specific conformations of the proteins heavily influence function and these conformations can be altered by things such as pH, temperature, salt concentration, or other environmental stresses (Campbell et al., 1999, pp. 70-76).

An organism’s metabolism is the totality of all its chemical processes, which are performed by proteins. Any small molecule that is an intermediate or product of metabolism is referred to as a metabolite. Metabolites are constantly moving throughout the fluids within an organism and are often found excreted as waste. Changes in cellular function will invariably change the metabolites within the cell and within the organism.

### 2.1.2 Toxicology

Toxicology is the study of the effects of toxins, or poisoning chemicals, upon an organism. An important aspect of toxicology is the study of the molecular interactions between the organism and the toxin including the biological mechanisms affected. Toxicology also includes the study of the symptoms expressed by and treatment of an organism being poisoned. One of the more interesting questions explored within toxicology is that of detection. How can scientists detect that an organism is affected by a toxin? Which toxin is affecting the organism? How early the toxin can be identified so that treatment can begin as soon as possible? All of these questions are pertinent to the study of organisms from bacteria to rats to humans.

Toxicological studies, like many biological, biochemical, and biomedical disciplines, can be grouped into three ‘omics,’ as the flow chart in Figure 2.4 shows.
2.1. BACKGROUND

Figure 2.3: The four levels of protein structure.

Adapted from (AccessExcellence, 2007).
URL: http://www.accessexcellence.org/RC/VL/GG/protein.html
2.1. BACKGROUND

Figure 2.4: The informational flow of the three in biological processes and their respective areas of study.

2.1.3 Genomics and Toxicology

Genomics is the study of an organism’s genome, or its entire set of genes including the interactions of these genes with one another and with environmental factors. It more specifically deals with the DNA sequence that makes up the genes within the genome, the organization of these genes, the function of these genes, and the evolution of these genes.

The main intersection of genomics and biological areas such as toxicology is referred to as functional genomics. Functional genomics is the study of how a stimulus, such as a toxin, effects gene function. This can be done through both forward and reverse genomics. Forward genomics refers to ”guessing” what gene or genes is affected in the process of interest and testing for changes in that gene. It requires some knowledge of the pathway disrupted so as to not taking shots in the dark. Reverse genomics requires no knowledge. Instead, markers are placed throughout regions of the genome. A marker may be a site-directed mutation, or change, in a gene. It may also be a deletion of a gene or of a whole section of the genome. If one of these markered regions is
2.1. BACKGROUND

statistically linked to the effects a toxin has on the organism, that region is explored more carefully as the effected gene or genes is likely to be found in that region (Hartl and Jones, 2005, pg. 524).

Expression of genes is often tested through the use of DNA microarrays. Microarrays are a common method of detecting active genes in a biological sample. Microarrays are small plates of glass to which the strands of complimentary DNA (cDNA) corresponding to a large number of known genes from the organism of interest are attached. The strands are complementary, meaning the entire strand is composed of the corresponding base pairs necessary for the single stranded RNA from the sample to bind. This creates a stable, double-stranded molecule. Typically two RNA samples are labeled with different fluorescent dyes, mixed in equal amounts, and washed onto the microarray. The amount of binding between the cDNA and the samples thus show the genes currently being expressed and their relative abundance in comparison to each other. These relative abundances are quantitatively measured by a scanner. Analyses of microarray data will be discussed using a literature survey in Section 2.2.2 (Hartl and Jones, 2005, pg. 524).

2.1.4 Proteomics and Toxicology

The second of the three areas, proteomics, is the study of an organism’s proteome, or the entire set of proteins produced and utilized by the organism. One advantage and difficulty of proteomics is that each organism only has one genome (each cell has the same DNA copy within it), but because each organism has organism-specific proteins and each cell has cell-specific proteins, an organism has many proteomes (Liebler, 2002).

Proteomics and toxicology intersect, much like genomics, on how the toxin effects protein function and levels. Proteomic analysis also often involves protein digestion, protein microarrays,
2.1. BACKGROUND

protein separations, gel electrophoresis, mass spectrometry, and other techniques beyond the scope of this thesis (Liebler, 2002).

2.1.5 Metabolomics and Toxicology

The third area, metabolomics, is the study of the metabolites present in and produced by an organism. The main advantage of metabolic study is how comprehensive it can be. In the cases where a toxin directly affects gene regulation or gene expression levels, either genomics or proteomics may be appropriate. However, a toxin of interest may interact at a different level of an organism, thus rendering a genomic or proteomic analysis less informative. A toxin affecting any level of an organism will alter the concentrations of biochemicals in the organism’s biofluids, such as blood, sweat, urine, and breast milk. It often does not matter where within the organism a toxin is introduced for biofluid compositions throughout to be altered. The toxin may interact specifically with the biochemical whose concentration has changed, or may interact higher in the metabolic pathway. Biochemicals within tissue or biofluid are part of a dynamic equilibrium and disturbances, such as those caused by toxins, will manifest in affected biochemical concentrations. Thus, unlike proteomics and genomics that assess snapshots of intermediate products from specific parts of the organism that may or may not contribute to variation in cellular function, metabolomics dynamically assesses the end products of cellular function, and in turn gene expression, throughout the whole organism (Nicholson et al., 1999).


2.1. BACKGROUND

2.1.6 Nuclear Magnetic Resonance (NMR) Spectroscopy

Nuclear Magnetic Resonance (NMR) was discovered simultaneously by Bloch et al. (1946) and Purcell et al. (1946). The phenomenon that is studied by NMR occurs when a static magnetic field is applied to the nuclei of certain atoms, followed by a second oscillating field. Any nuclei with an odd number of protons or neutrons can be measured using NMR, though $^1$H and $^{13}$C are the most common. These nuclei exhibit specific rotating electrical charges, or spins properties.

There are three steps to a NMR spectroscopy experiment. The first, polarization, is the interaction between the set of spin properties and a static magnetic field. The result is a balanced state. The second step is the resonance step. An exciting radio frequency (RF) field supplies energy to the spin system, perturbing the previously mentioned balanced state. The RF field is applied in short pulses, simultaneously exciting the spin system into transitions between energy states. Finally, relaxation occurs immediately after the RF pulse when the spin system returns to the original balanced state. As these transitions relax, a current is induced in an RF coil, which picks up all of the transitions from all of the nuclei simultaneously. This current or signal decays exponentially, because of this relaxation. This free induction decay (FID) signal is then Fourier transformed, resulting in a spectrum as shown in Figure 2.5 (Franconi et al., 2000, pp. 1–26).

The shape of the peaks in a Fourier-transformed NMR spectrum are Lorentzian in shape, as shown in Figure 2.6. This shape can be described by the Lorentzian function (Weisstein, 2004):

\[
L(x) = \frac{1}{\frac{1}{\pi} \left( \frac{1}{2} \Gamma \right)^2 + \left( x - x_0 \right)^2}
\]

where $\Gamma$ is the peak width at half height and $x_0$ is the center of the peak.
2.1. BACKGROUND

Figure 2.5: A NMR spectrum.
2.1. BACKGROUND

While two peaks in two different spectra may correspond to the same molecule, their placement in their respective spectrum may be shifted and their shape may be slightly varied due to differences in the experimental conditions (such as temperature, pH, instrumental inconsistencies, or other factors). These differences have been shown to have an affect on pattern recognition techniques, and therefore must be mitigated (Siuda et al., 1998). There are two commonly utilized pre-analyzation techniques to deal with these differences. The first, peak alignment, is where the same peaks from different spectra are moved, or aligned, so that the highest point of the peaks are at the same position (Forshed et al., 2005). Several different methods of peak alignment have been proposed including the use of partial linear fit (Vogels et al., 1996), a genetic algorithm (Forshed et al., 2003), a beam search genetic algorithm (Lee and Woodruff, 2004), and a PCA-based method (Stoyanova et al., 2004). The second, binning or bucketing, is where the spectrum is divided into
2.1. BACKGROUND

evenly sized bins, often 0.04-0.07 ppm wide. The information in each bin is then combined in some fashion, often by integration of the area underneath the curve or by summing the values contained in the bin. Binning is common because of its simplicity and its effectiveness in mitigating peak fluctuations in shape and position. This simplicity comes at the cost of sensitivity, as smaller, subtle changes may be masked by larger changes in nearby peaks. However, this loss of sensitivity did not prevent successful results for Nicholls et al. (2001) or Holmes et al. (1998).

A common artifact of NMR spectra is a distorted baseline, including the raised baseline that can be found in Figure 2.5 between 2000 and 2500 Hz. The peaks in this range overlap too much to allow a return to the normal baseline near zero intensity. Baseline distortions also may arise from hardware and software sources, including too large of a gain, or signal amplification given to the FID. This is commonly corrected by software provided by the NMR manufacturer, though other baseline corrections have been proposed such as the Continuous Wavelet Derivative transform in combination with penalized least squares method (Cobas et al., 2006).

2.1.6.1 NMR-based Metabolomics

Unlike various other proteomic, genomic, and metabolomic analyses, NMR spectroscopy is non-invasive, non-destructive, and requires little sample preparation. Biofluids can often be extracted with no surgical extraction. The samples are not altered or destroyed in any way and may be saved for later use. It is typical for only a solvent to be added before a sample is ready for testing. _In vivo_ testing, testing of samples from within an organism as opposed to from a test tube, is often slow and expensive at the genomic and proteomic levels. NMR spectroscopy usually requires minimal run time as 200-300 samples per day is possible with modern NMR spectrometers (Reo, 2002). There
2.2. LITERATURE SURVEY

are also no expensive, disposable components necessary as there are in microarray experiments. Due to these reasons, exploratory analysis for biomarker discovery is where NMR spectroscopy excels in the field of biochemistry and toxocology (Nicholson et al., 1999).

2.2 Literature Survey

The desire to isolate potential biomarkers while handling the high dimensionality/low sample size problem has been discussed across metabolomics and genomics. The following subsections survey the use of several techniques in determining metabolites or genes indicating a biological response. Also surveyed are statistical considerations involved in such a study.

2.2.1 NMR Analysis

Initial studies in the field of NMR-based metabolic biomarkers began being published in the 1980's. Gartland et al. identified metabolites found in rat urine, noting their time-course over 48 hours (Gartland et al., 1989). Changes from potential biomarker metabolites were confirmed through conventional biochemistry methods with confidence measures provided by t-tests.

This non-computational method for classification and identifying markers from NMR data was soon followed by the use of statistical methods directly on the data derived from the NMR data. Gartland et al. performed an analysis using non-linear mapping (NLM), principal component analysis (PCA), and hierarchical cluster analysis (Gartland et al., 1991).

PCA is a method of data reduction in which a data matrix $X$ is essentially decomposed into a sum of the products of two vectors, a score $t$ and a loading $p$ (Hotelling, 1933; Jolliffe, 1986).
2.2. LITERATURE SURVEY

The loadings ($p_1, p_2$) place the PC vector in the multidimensional (in this case 2-d) space (A). The scores ($t_1, t_2$) represent the position of each sample on the PC vector (B). Reproduced from Geladi and Kowalski (1986).

Each product of a scores vector and a loadings vector make up a principal component (PC). A graphical representation of the loadings and scores from a two variable PCA model is shown in Figure 2.7. Each score represents a sample’s placement on the PC. Each loading represents a variables contribution towards the PC’s direction in space. PC’s can be computed iteratively as in the NIPALS method (discussed further in Section 2.2.1.1) or all at once using eigenvectors. PCA performed feature selection and has no knowledge of any classification.

Gartland et al. (1991) concluded that NLM, PCA, and hierarchical cluster analysis adequately reduced and classified the data for the several toxins administered. Nicholson and colleagues identified discriminatory metabolites in two publications via a combination of NLM, PCA, Fisher weights, and paired t-tests (Holmes et al., 1992; Anthony et al., 1994). In these cases, NLM and PCA were used as exploratory data analysis methods to visually inspect clusters arising between a set of control animals and animals exposed to two different toxins over a time period in order
2.2. LITERATURE SURVEY

to observe the time course of a toxic episode. Fisher weights were then calculated between time intervals for both treatments and controls and large weights were supported by significance values from paired t-tests. This series of pattern recognition techniques provided some of the first metabolic and toxicological biomarkers derived from NMR spectroscopy.

Another method that has been used in NMR-based metabolomics is linear discriminant analysis (LDA). Lisboa et al. compare the results of LDA versus a multi-layer perceptron (MLP) in determining responding metabolites in Howells’ et al.’s (1992) rat tumor data (Lisboa et al., 1998). LDA, unlike PCA, is a data classifier. It attempts to create a decision boundary which separates the classes as much as possible. The knowledge of class assignments greatly increases the power of LDA. However, this power comes at a price. Under circumstances of low sample size relative to the dimensionality of the data, LDA may be too aggressive in finding a decision boundary based solely on noise when no real differences between the classes are present (?).

The MLP is a non-linear neural network, described in more detail by Bishop (1995), that is trained by minimizing an error function on training data. This can lead to overfitting if too many weights are included in the network. Consideration was taken to avoid this overfitting problem, which can falsely lead to results which may be due to artifacts in the data as opposed to an biological response. The LDA method of feature selection was able to identify more responding metabolites than the MLP method, including two responses that could not be attributed to a known metabolite. It is possible that these responses were a result of LDA having too much power.
2.2. LITERATURE SURVEY

2.2.1.1 Partial Least Squares Regression

More recently, the practice of using Partial Least Squares (PLS) regression has become popular for the analysis of NMR spectra. Originally proposed by Wold (1966) for the field of economics, PLS was adopted for chemometrics, the use of mathematical or statistical applications to chemical data such as NMR spectroscopic data, by Kowalski et al. (1982). NMR analysis is an example of a chemometric application. Sometimes called ”Projection to Latent Structures”, PLS employs features of PCA and Multiple Linear Regression (MLR) and is useful to predict a set of dependant variables from a large set of independent variables. This is the situation we must deal with for NMR analysis. We have a large set of dependant variables (spectral data) and an independent class variable, which we may wish to predict.

PLS has the goal of predicting a dependent variable(s) \( Y \) from independent variables \( X \) while at the same time produce a meaningful relationship between \( X \) and \( Y \). A PLS tutorial, from which the following description is based upon, is presented by Geladi and Kowalski (1986).

MLR, described in more detail by Mardia (Mardia et al., 1980), attempts to find sensitivities \( b \) which minimize residuals \( e \) in the following equation:

\[
Y = XB + E
\] (2.2)

where \( Y \) is a matrix of dependent variables \( y \) for each sample, \( X \) is a matrix with the number of columns equal to the number of independent variables and the number of rows equal to the number of samples, \( B \) is a matrix of sensitivities \( b \) (or regression coefficients), and \( E \) is a matrix of residuals \( e \). MLR has two major problems associated with it. The first is that if there are more
2.2. LITERATURE SURVEY

independent variables than samples, there are an infinite number of possible solutions for \( b \). The second lies in a popular choice for minimizing the residuals, termed the "least-squares method". The least-squares method is defined as:

\[
b = (X'X)^{-1}X'y
\]  

(2.3)

The problem is that \((X'X)^{-1}\) may not exist. There are two reasons for this. One is that \( X \) may not have full rank. The other occurs when there are linear or near-linear relationships among \( X \) variables and is termed collinearity, multicollinearity, or singularity. If a least-squares method is necessary, removing collinear variables from the model is one possible remedy (Weisberg, 1985, pg. 197). A pseudo-inverse of \((X'X)\) is also sometimes used. This is simply an approximation of the inverse, and may not be appropriate in all situations. In cases of severe rank deficiencies pseudo-inverse solutions may include imaginary numbers. LDA suffers from these same downfalls Eriksson et al. (2006). Alternatively, different methods of data reduction like PCA Lilien et al. (2003) and PLS Nguyen and Rocke (2002) have been used prior to LDA in an attempt to circumvent these issues.

As mentioned in Section 2.2.1, the Principal Components (PC) from Principal Components Analysis (PCA) can be calculated iteratively or all at once. A common iterative procedure, likely because of its simplicity, is the Nonlinear Iterative PArtial Least Squares (NIPALS) method, proposed by Wold (1966). NIPALS calculates the score \( t_1 \) and the loading \( p_1 \) from the matrix \( X \). The product \( t_1p_1' \) is subtracted from \( X \). The residual \( E_1 \) is then used to calculate \( t_2 \) and \( p_2 \):

\[
E_1 = X - t_1p_1', E_2 = E_1 - t_2p_2', ..., E_h = E_{h-1} - t_hp_h'
\]  

(2.4)
2.2. LITERATURE SURVEY

until $E_h = 0$ or some criterion has been met. The NIPALS algorithm for calculating $t_2$ and $p_2$ is as follows:

1. take a vector $x_j$ from $X$ and call it $t_h : t_h = x_j$

2. calculate $p'_h : p_h = \frac{t'_h X}{t'_h t_h}$

3. normalize $p'_h : p'_{h-new} = \frac{p'_{h-old}}{||p'_{h-old}||}$

4. calculate $t'_h : t_h = \frac{X p_h}{p'_h p_h}$

5. compare the $t_h$ from step 2 to that of step 4, if equal stop, if different go to step 2

Principal Component Regression (PCR) uses the PCA scores matrix $T$, described above as essentially a representation of $X$, to repropose the MLR equation with these representations of $X$:

$$Y = TB + E$$ \hspace{1cm} (2.5)

where

$$T = XP$$ \hspace{1cm} (2.6)

and

$$B = (T'T)^{-1}T'Y$$ \hspace{1cm} (2.7)

The column vectors $t$ of $T$ have a property that the column vectors $x$ (variables) of $X$ do not. They are orthogonal, thus ensuring $(T'T)^{-1}$ exists. Orthogonal variables will not change
the regression estimates or standard errors. Thus, adding collinear variables orthogonal to all other variables cannot negatively affect the regression estimates (Belsley et al., 1980, pp. 107-108). PCR also allows for some noise (random error) reduction as much of the inherent noise will be found in the lesser column vectors of T, which are normally removed from a PCR model.

PLS combines elements of PCA and MLR by extending PCR for both X and Y. PLS essentially consists of a regression between the scores of X, T, and the scores of Y, U. During the regression, information is exchanged between the parts of the algorithm that create and manipulate these scores. This exchange of information allows for a stronger relationship between X and Y within the PLS model.

There are several algorithms which are used to create a PLS model. NIPALS, again one of the most common again because of its simplicity, involves calculating each latent vector (LV) (analogous to a principal component in PCA) one at a time. After each LV is calculated, it is subtracted from both X and Y and the procedure is repeated until X is a null matrix or some criterion has been met. The NIPALS algorithm requires an extra matrix W be involved in the computation. The vector w replaces all of the instances of the the vector p, including:

\[ t = Xp \] (2.8)

which becomes:

\[ t = Xw \] (2.9)

Each vector w is referred to as an X-weight and the substitution is necessary because the PLS
2.2. LITERATURE SURVEY

NIPALS algorithm changes the order of the steps from the PCA NIPALS algorithm. Thus the X-scores (t) are not guaranteed to be orthogonal. This is a necessary requirement to ensure collinearity is not a problem. For a pseudocode version of the PLS algorithm, see Geladi’s PLS tutorial (Geladi and Kowalski, 1986).

It is often not advantageous to iterate the algorithm until X is null, as the lesser LV’s often include noise and collinearity. There are several methods for determining the number of LV’s to include in a PLS model. One method is to calculate the residual between the actual Y and the predicted Y. Once this residual falls below some small, previously established error, stop iterating the algorithm.

Another method is to calculate a Prediction REsidual Sum of Squares (PRESS) statistic. PRESS is defined as follows:

\[
PRESS = \sum_{i=1}^{n} (e_i^2) \tag{2.10}
\]

where \(e_i\) is the residual between the predicted and actual Y during a leave-one-out cross validation and \(i\) is the sample left out of the model building for validation. A few large residuals could dominate a PRESS statistic (Walpole et al., 1998, pp. 445-447). The model with the minimum PRESS determines the number of LV’s that should be included in the model. The randomization t-test, proposed by (van der Voet, 1994) in another method of determining the number of LV’s to include. More on this method can be found in Section 2.2.4.3.

PLS is also useful is determining which of the independent variables are most important in predicting dependant variables. One method of interpretation is to analyze the X-weights, which
2.2. LITERATURE SURVEY

represent the correlation between the independent variables and the Y-scores. Numerically large weights are important independent variables and similar weights show variables which are performing similarly. The X-loadings typically give similar information.

Variable Influence on Projection (VIP) is another method of identifying important independent variables. The PLS weights \( w_{ak} \), where \( k \) and \( a \) represent the X-variable and LV respectively). VIP measures X-variables for their relative importance for all Y-variables and all latent vectors. It is calculated as follows:

\[
VIP_k = \left[ \sum_{a=1}^{A} (w_{ak})^2 \cdot SSY_a \cdot \frac{K}{SSY_{tot,expl} \cdot A} \right]^{1/2} \tag{2.11}
\]

where \( SSY_a \) is the sum of squares of Y explained in the \( a^{th} \) latent vector, \( SSY_{tot,expl} \) is the total sum of squares of Y explained in all A latent vectors, and \( K \) is the total number of variables in X.

A general rule of thumb for VIP values is that X-variables with VIP’s less than 0.8 are unimportant and X-variables with VIP’s greater than 1 are important (Wold et al., 1993). The VIP method will not converge on a set of important variables, however. It is not possible to cull all variables with large VIPs and rerun the VIP method on the smaller set of variables, as the number of important variables will continue to get smaller, eventually converging to 1.

The X-scores and Y-scores produced by PLS contain information about the samples and their similarities and dissimilarities with respect to the given model. A plot of the X-scores for the first two LV’s shows how samples from different classes cluster together in a particular PLS model.
2.2. LITERATURE SURVEY

The regression coefficients, $b_{\text{pls}}$, are calculated as follows:

$$b_{\text{pls}} = w \ast \text{inv}(p' \ast w) \ast \text{diag}(b) \ast c'$$  (2.12)

Regression coefficients are used to predict $Y$ from $X$ and can show positive and negative association between independent variables and dependent variables.

PLS Discriminant Analysis (PLS-DA) is a version of PLS where the dependent variable is a categorical class assignment for the sample, rather than some continuous variable(s). Otherwise, the PLS implementation and interpretation remain unchanged. This PLS-DA method is commonly used in NMR analysis where samples belong to some number of classes (some number of treatments and controls).

Bertram et al. performed a NMR spectroscopic analysis of steer blood plasma obtained from a vein, a portal, and an artery both before and after infusion of short-chain fatty acids in a vehicle (treatment) and only the vehicle (control) (Bertram et al., 2005). They first performed PCA in order to explore any clustering behavior between the groups. Clustering was observed in a PCA scores plot of PC1 versus PC2. PLS-DA was then performed in order to investigate the metabolic differences between the three sources of blood plasma (two PLS-DA models were created, one between the artery and the portal and another between the portal and the vein). Metabolic differences were identified in both PLS-DA models via loading plots and specific metabolites were identified based on their position in the spectra. Additionally, Bertram et al. created PLS regression models in order to predict metabolite concentrations based on independent variables from the NMR spectra. Relative error rates were reported as ranging from 3% to 39%, which they determined to be promising.
2.2. LITERATURE SURVEY

This use of PCA to examine clustering behavior followed by PLS-DA to select metabolites contributing to the difference between groups is a common sequence of methods for NMR analysis. Griffin et al. used PCA and PLS-DA to examine differences between rats with central nervous system lesions caused by two different adenoviruses and a set of control rats (Griffin et al., 2004). Several metabolite differences between the 3 groups were identified and a leave-one-out cross-validation of the predictive ability of their PLS-DA models resulted in a predictive power of 79%. Samuelsson et al. explored the metabolic differences between rainbow trout exposed to synthetic estrogen and those which were not (Samuelsson et al., 2006). Important metabolites in the separation between treatment and control were identified using PLS-DA regression coefficients. Bertram et al. use PCA and PLS-DA to reveal changes in blood plasma and urine between pigs on a whole grain rye and those on a non-whole grain wheat diet (Bertram et al., 2006). Both urine and plasma samples separated via a PLS-DA scores plot and differences in metabolite concentrations between the two diets using the regression coefficients.

2.2.1.2 Orthogonal Projection to Latent Structures

Orthogonal Projection to Latent Structures (O-PLS), or Orthogonal Partial Least Squares, is an extension of PLS that has recently become popular among scientists performing NMR analyses. Initially proposed by Trygg and Wold in 2002, it combines a technique called Orthogonal Signal Correction (OSC) (Wold et al., 1998) with PLS with the goal of removing systematic variation within X not correlated with Y (or orthogonal to Y) from the model in order to make the model more interpretable (Trygg and Wold, 2002). A bonus of O-PLS is that removed Y-orthogonal components of X can be separately analyzed and may reveal some insight into the sources of the
X-only variation. The following description of O-PLS will primarily follow the original O-PLS article by Trygg and Wold (2002) and the original OSC article by Wold et al. (1998).

In order to perform any orthogonal data-filtering, there are 3 criteria which the Y-orthogonal component of X needs to meet. The orthogonal component should involve large systematic variation in X, it must be predictive by X, and it must be orthogonal to Y. OSC meets these three requirements in a PCA-related solution. The PCA part of the solution meets the first two criteria inherently, as the PC’s produced describe as much of X as possible and each PC is predictive by X. The third criterion is met using the NIPALS algorithm, which allows for the vector w to be modified in any way such that the steps of NIPALS remain in order. Thus, w can be modified such that in the equation:

\[ t = Xw \]  \hspace{1cm} (2.13)

\( t \) contains only X variation orthogonal to Y. There is an internal iterative procedure where w is modified in order to make \( t \) as close to \( t^* \) as possible where \( t^* \) is orthogonal to Y and defined as follows:

\[ t^* = (1 - Y(Y'^{-1}Y')^{-1})t \]  \hspace{1cm} (2.14)

OSC has the problem that it tends to overfit the data and cross-validation is difficult due to implementation and time issues.

Gavaghan et al. explored the use of OSC to remove intrinsic physiological variation between subjects, specifically rat strain and diurnal variations in NMR metabolomic data (Gavaghan et al., 2018).
2.2. LITERATURE SURVEY

PCA and PLS-DA were performed both before and after OSC-filtering the data. Before data-filtering results showed differences between urine samples between two strains of rats and showed differences between morning and evening urine samples within the same rat strain. The predictive ability of the PLS-DA model was improved by filtering. First, diurnal cycle variation was removed by OSC and PLS-DA predictive accuracy was raised from 90% to 95%. Alternatively, strain variation was removed via OSC and PLS-DA accuracy was raised from 90% to 100%. Predictive accuracy was calculated by a leave-10%-out cross validation. The authors conclude that OSC improves the effectiveness of PCA and PLS-DA by removing confounding factors (in this case strain or diurnal cycles) from the data.

Brindle et al. came to a similar conclusion when studying the metabolic differences between severe, moderate, and mile coronary heart disease (CHD) in NMR-derived human serum samples (Brindle et al., 2002). The authors claim OSC-filtering improved both the PCA and PLS-DA analyses, though no results were shown. Differences between CHD severities were clear in both PLS-DA scores plots and regression coefficients indicated the most differentiating metabolite was LDL-cholesterol.

O-PLS borrows from OSC and PLS, creating an orthogonal correction algorithm which only requires a simple modification to the NIPALS algorithm. Like OSC, the goal is to filter out the Y-orthogonal X-variation. PLS is then performed on the filtered data. The O-PLS algorithm starts by creating a PLS component (consisting of vectors: \( p, w, t, u, c \)). \( w_{osc} \), the \( w \) vector from OSC described above, can be computed as the difference between \( p \) and \( w \). The OSC vectors \( t_{osc} \) and
2.2. LITERATURE SURVEY

$p_{\text{osc}}$ can then be calculated using $w_{\text{osc}}$, and the filtered data $X_{\text{opls}}$ matrix is calculated as follows:

$$X_{\text{opls}} = X - t_{\text{osc}}p_{\text{osc}}'$$

(2.15)

where $X$ is the original data matrix.

At this point, $X_{\text{opls}}$ can either be used as the final, filtered data or the algorithm can be reiterated with $X_{\text{opls}}$ replacing $X$. If at any time the difference between $p$ and $w$ is close to zero, all of the $Y$-orthogonal $X$-variation has been removed from the data and the PLS component (after performing PLS on the filtered data) $t$ will have maximal correlation to $y$.

The major benefit of O-PLS is that resulting PLS models are more parsimonious and are easier to interpret because the variation not correlated to $Y$ is removed. The result is often that the $Y$-correlated data can be explained in one PLS latent vector (LV). Additionally, because the NIPALS PLS algorithm is followed so closely, cross-validation is implementable, thus reducing the chance of overfitting. O-PLS is robust enough to deal with data sets that are comprised totally of $Y$-orthogonal $X$-variation (PLS will return no LV’s) and data sets with no $Y$-orthogonal $X$-variation. O-PLS can also deal with an amount of missing data because it follows the NIPALS method. For a step-by-step outline of the O-PLS algorithm, see Trygg and Wold’s original O-PLS article (Trygg and Wold, 2002).

Cloarec et al. evaluate the limitations of O-PLS-DA with respect to chemical shift variation (Cloarec et al., 2005). Both full and binned NMR spectra were evaluated with O-PLS-DA using simulated data so that shifts and the source of discrimination could be varied. The authors determined that peak position variation negatively affects predictive ability in both full and binned
simulated spectra, but in most situations only slightly. The information from this positional variation can be modeled by the Y-orthogonal components, allowing for additional insight about the cause of the variations. In addition to their conclusions on peak position variation, the authors offer a method for visualizing the O-PLS-DA loadings plot such that they model the shape of the original spectrum. The loadings are back-scaled by multiplying them by the variance of that particular variable. The important variables are then color-coded according to their weights in the O-PLS-DA model. This means that larger weights, corresponding to important variables for discrimination, are easily identifiable.

Like PLS-DA, Orthogonal Projection to Latent Structures Discriminant Analysis (O-PLS-DA) extends O-PLS to a situation where the one dependent variable is a class assignment value for the sample, rather than some continuous variable(s). The steps of the O-PLS algorithm otherwise remain unchanged and PLS-DA is performed on the filtered data.

Stella et al. use O-PLS-DA to study the effects of dietary modulation (Stella et al., 2006). Specifically subjects were put on three diets (vegetarian, low meat, and high meat) randomly for 15 days each diet. Urine samples were collected and NMR spectroscopy was performed. PCA results were initially inconclusive as individual sample variation dominated the analysis. O-PLS-DA was applied to remove some X variation, and a scores plot showed a separation between the three groups that PCA (before filtering) could not attain. The authors used the Cloarec et al. (2005) method of visual interpretation described above, which clearly illustrated the parts of the spectra different between the high meat group and the vegetarian group, corresponding to important metabolic differences. Some of the metabolites responsible for these differences were not products of human metabolism, but of the gut microflora, indicating the bacterial composition within the
2.2. LITERATURE SURVEY

digestive system changes based on diet.

This series of PCA and O-PLS-DA was also performed by Coen et al. on metabolic changes in rats due to galactosamine toxicity (Coen et al., 2007) and Rezzi et al. on healthy humans’ dietary preferences and how they effect that basal metabolic state and gut microbiome (Rezzi et al., 2007). Both articles begin by performing PCA in order to determine whether samples cluster in an unsupervised environment, followed by O-PLS-DA to identify the metabolites strongly correlated with class. Back-scaled loading plots depicted spectral differences in both articles and 7-fold cross-validations were performed to establish the robustness of the model.

Weljie et al. were able to create a metabolic biomarker pattern from NMR spectroscopic data identifying rheumatoid arthritis in mice (Weljie et al., 2007). The authors first pare down the NMR spectra with the method referred as ‘targeted profiling’ first proposed by Weljie et al. (2006). Targeted profiling involves comparing NMR spectra of a set of pure, known metabolites and the sample NMR spectra in order to identify peaks of interest. These peaks are then used for the remainder of the analysis. O-PLS-DA was performed on the 88 remaining metabolites and to prevent overfitting, the authors performed a permutation test with test statistics $R^2$ and $Q^2$. More on these test statistics and permutation plots can be found in Section 2.2.4.1. Those metabolites with O-PLS-DA model VIP values greater than one were chosen for a Multivariate ANalysis Of Variance (MANOVA) which selected 18 significant metabolites for the metabolic profile. In addition, the authors performed a pathways analysis and results suggested biochemical pathway differences due to arthritis.
2.2. LITERATURE SURVEY

2.2.2 Microarray Analysis

Many of the issues found in NMR data analysis are also found in microarray data analysis, most notably the curse of dimensionality. When discussing features, we now are referring to genes instead of bins or peaks associated with metabolites. One feature-at-a-time method that has been employed in the microarray community is the fold test. The simple nature of the test is one of the reasons for its use. The fold ratio is calculated as follows:

$$\text{fold ratio} = \log(\bar{y}_{11}) - \log(\bar{y}_{12})$$

(2.16)

where $\bar{y}_{11}$ and $\bar{y}_{12}$ are the means of the two groups to be tested. A two-fold difference is a common cutoff to determine an interesting difference; however, no statistical significance can be attributed to this cutoff.

Schena et al. monitored expression patterns in *Arabidopsis thaliana* (Schena et al., 1995). Several situations were introduced, including differential gene expression and single gene overexpression, and the authors determined that the fold test on microarray data could identify all of the gene expression situations tested. Schena et al. also tested the ability of microarrays in performing large-scale human gene discovery and used the fold test to analyze human cell response to heat shock (Schena et al., 1996). DeRisi et al. investigated genome wide expression associated with the metabolic shift from fermentation to respiration *Saccharomyces cerevisiae* using the fold test (DeRisi et al., 1997). The authors claim that their results demonstrate both feasibility and utility of DNA microarrays for comprehensive exploration of expression trends.

Several more recent microarray comparison/survey articles have compared the fold test to
more complicated tests, addressing both the benefits and drawbacks. In one such survey, Draghici compares the unusual ratio test, univariate statistical tests (specifically the t-test), a microarray specific ANOVA model, and a modified ANOVA called the noise sampling model, along with the fold test (Draghici, 2002). The unusual ratio test is similar in simplicity to the fold test. The test looks for treatment:control ratios a set number of standard deviations, typically 2, away from the mean treatment:control ratio. The advantage over the fold test is that the unusual ratio test will automatically adjust its cutoff threshold, whereas the fold test has a user-defined, arbitrary cutoff. The major drawback is that it will invariably select 5% of the variables as differentially responding, whether many more or many less than 5% are actually responding. The ANOVA, specifically using the Kerr and Churchill model, is intriguing in that it can account for any variability in the measurements, so long as each source of variance has been built into the model. Unfortunately it also requires that the experiment is carried out in a rather specific manner including having a sufficient number of degrees of freedom available. The noise sampling model is interestingly different from a more standard ANOVA in that it does not require such a specific experimental design. Instead, an estimation of the noise is calculated along with confidence levels for variable response. The noise sampling model can only be used to provide estimates of the log ratios of the variables, but is more sensitive than the fold or unusual ratio tests.

Zhu and Hastie compare univariate ranking and recursive feature elimination (RFE) as methods to select variables responsible for differences between controls and treatments (Zhu and Hastie, 2004). Penalized logistic regression (PLR) and support vector machines (SVM) are described and used as classifiers. Zhu and Hastie acknowledge that many times the variables of interest may be more important than classifying the samples. Univariate ranking and recursive feature elimination
are evaluated using three different real world data sets. One univariate ranking method, proposed by Golub et al. (1999), ranks genes in a two-class problem and is defined as follows:

\[
\rho_j = \left| \frac{\bar{x}_j^{(1)} - \bar{x}_j^{(2)}}{\sigma_j^{(1)} - \sigma_j^{(2)}} \right|
\]  

(2.17)

where \( j \) is a particular gene, \( \bar{x}_j^{(1)} \) is the mean expression level for gene \( j \) across all samples in classification group 1, \( \bar{x}_j^{(2)} \) is the mean for gene \( j \) across all samples in classification group 2, \( \sigma_j^{(1)} \) is the standard deviation for gene \( j \) across all samples in classification group 1, and \( \sigma_j^{(2)} \) is the standard deviation for gene \( j \) across all samples in classification group 2. Univariate rankings, in general, assume orthogonality across genes (Zhu and Hastie, 2004).

RFE, proposed by Guyon et al. (2002), calculates a ranking criterion for each variable approximating the sensitivity of a loss function to the variable based on the coefficient corresponding to that variable. A number of variables are then removed from consideration after every iteration of the RFE (in Zhu’s case the lowest 10%). The univariate ranking method of variable reduction added to classification errors for both the PLR and SVM classifiers when compared to RFE. Between the three data sets, the univariate ranking method selected fewer features than the RFE method for one set, while selecting more features than the RFE method for the other two sets. Zhu and Hastie did not elaborate as to why this happened, though it appears to be related to the data sets themselves. Zhu did conclude that the PLR classifier combined with the RFE was the method of choice. PLR, while performing comparably to the SVM, adds estimates of the underlying probability. PLR with RFE tends to reduce more feature than the SVM with RFE.
2.2. LITERATURE SURVEY

2.2.3 Multiple Test Corrections

Many methods in both NMR and microarray analysis lead to repeating many statistical tests over the entire set of features. This repeating of tests must be accompanied by some type of multiple test correction. We would expect by random chance that a small percentage, 5% when using a standard 0.05 threshold, will reject the null hypothesis even if no significant differences are present. The more tests performed, the more likely this Type I error (false positive error) will become an issue (Shaffer, 1995).

A common and easily implemented correction was proposed by Bonferroni (1936). Its formula is simple and defined as follows:

\[ p_c = k p_o \]  \hspace{1cm} (2.18)

where \( p_o \) is the original p-value of the test, \( k \) is the number of hypotheses one wishes to test, and \( p_c \) is the corrected p-value of the test. The correction changes all p-values by the same amount. Rejecting a test’s null hypothesis becomes difficult as the number of tests becomes large, thus creating large corrected p-values. Few (if any) p-values, whether they are actually significant or significant by chance, will fall below the desired significance level. This is the reason the Bonferroni correction is considered conservative. The Bonferroni correction makes no assumptions about model or distribution beyond independence of the tests. It has, even with its conservative nature, been used in microarray analyses (Wayne and McIntyre, 2002).

A similarly simple multiple test correction was proposed by Holm (1979) and is referred to as a Holm’s step-down correction. This means that the p-values are adjusted from the smallest
2.2. LITERATURE SURVEY

p-value to the largest. Each p-value adjustment is incrementally smaller than the previous. The Holm’s method is defined as follows:

\[
P_{c1} = kp_o_1\quad (2.19)
\]
\[
P_{cj} = \max(p_{c(j-1)}, (k - j + 1)p_o_j) \quad \text{for} \quad 2 \leq j \leq k \quad (2.20)
\]

where \(p_o\) is the original p-value of the test, \(k\) is the number of hypotheses one wishes to test, and \(p_c\) is the corrected p-value of the test. Sometimes referred to as the Bonferroni step-down method, it is like Bonferroni in that it assumes independence of the test statistics but makes no assumptions about the model or distribution. While less conservative than the Bonferroni method due to the decremented adjustments, it is still quite conservative and similar issues of large p-values and difficulty rejecting the null hypothesis arise.

Westfall and Young (1993) proposed another step-down method defined as follows: (1) Calculate original p-values for each variable and rank them. (2) Permute which samples are treatment and which are control. (3) Calculate new p-values for each variable in the permuted data. (4) The new p-values should be adjusted such that each new p-value is less than or equal to the next in order. This is accomplished by starting at the last p-value, comparing it to the next to last, and changing the next to last to the minimum of the two p-values compared. These new p-values are then compared to the original p-values. (5) Steps 1-4 are repeated a large number of times, ideally equal to the number of possible permutations of sample labels. The proportion of permuted sets where the new p-value was less than that of the original p-value becomes the adjusted p-value.

The major advantage to the Westfall-Young method of multiple test correction is that it takes
2.2. LITERATURE SURVEY

into account some dependance between the tests, giving more power and creating a less conservative correction than Bonferroni or Holm’s methods (Dudoit et al., 2002).

Sankoh et al. (1997) discuss the statistical concerns associated with common corrections, including the Bonferroni, Holm’s step-down, Hommel, and Tukey et al. (TCH) methods. Hommel’s method, proposed by Hommel (Hommel, 1989), is a step-down method slightly less conservative than Holm’s step-down method and is defined as follows: Order the p-values such that:

\[ p_1 >= p_2 >= ... >= p_K \]

with corresponding null hypotheses \( H_{01}, H_{02}, H_{0K} \). Find the largest \( m \) for which

\[ p_1 > \alpha; p_1 > \alpha, p_2 > \alpha/2; p_1 > \alpha, p_2 > 2\alpha/3, p_3 > \alpha/3; ...; p_1 > \alpha, p_2 > \alpha(m-1)/m, p_3 > \alpha(m-2)/m, ..., p_m > \alpha/m. \]

Then reject \( H_{0K} \) for which \( p_k < \alpha/m \). Sankoh et al. continue by describing the \textit{ad hoc} procedure proposed by Tukey et al. (Tukey et al., 1985). Little theoretical performance analysis has been done on \textit{ad hoc} procedures, but simulation studies have suggested them for strongly correlated data. The TCH method is defined as follows:

\[ \alpha_k = 1 - (1 - \alpha)^{1/K} \]  \hspace{1cm} (2.21)

After a simulation study based on \( K=7 \) tests and a desired \( \alpha \) of 0.05, the Bonferroni correction was determined to be the most conservative, with the lowest number of rejected null hypotheses. The Hommel and Holm’s step-down corrections were slightly less conservative. The TCH correction was the most liberal, rejecting the same number of null hypotheses as the unadjusted \( \alpha \) value. In conclusion, they determine that the step-down methods perform well for weakly correlated data and a small number of tests, but become more conservative as the number of tests increase or the correlation of the data increases. The \textit{ad hoc} TCH procedure perform well for strongly correlated data, but not otherwise.
Any reported results and findings must be validated in either their predictive accuracy or their ability to be reproduced. Several methods have been proposed in association with the tests discussed earlier in this section. A subset of those have been selected as potential methods for validation of the results of this thesis.

2.2.4.1 Cross-validation

Svante Wold proposed a cross-validation method of selecting the optimal number of principal components for PCA such that the predictive properties are maximized (Wold, 1978). The validation centers around creating a linear predictive model using the PCA components. The model’s formula is as follows:

$$y_{ik} = \alpha_i + \sum_{a=1}^{A} B_{ia} T_{ak} + \epsilon_{ik}$$  \hspace{1cm} (2.22)

where $i$ represents a $Y$-variable from 1 to the total number of variables, $k$ represents a sample from 1 to the total number of samples, $a$ represents a PC from 1 to $A$ (the total number of PC’s), $y_{ik}$ is the $y$ data value for a given $Y$-variable and sample, $B_{ia}$ is a regression coefficient for a given $Y$-variable and PC, $T_{ak}$ is a $X$-score for a given PC and sample, and $\epsilon_{ik}$ is the residual for a given $Y$-variable and sample.

The first part of the validation requires a cross-validation of the $\alpha_i$ term. In some cases $\alpha_i$ will be zero, such as the validation of PLS or O-PLS where the model being tested is already in a linear, predictive form. In these cases, $\alpha_i$ does not need to be validated. The samples are divided
2.2. LITERATURE SURVEY

into a number of groups (between 3 and 5, Wold suggests). The first group is left out and the difference of the $y_{ik}$ average between the remaining groups and the left out groups is calculated. This difference is then used as the partial PRESS, and the process is iterated for each group not yet left out. The partial PRESS values are added to form an overall PRESS and a ratio $R$ is calculated as follows:

$$ R = \frac{PRESS}{S_\alpha} \quad (2.23) $$

where $S_\alpha$ is the residual sum of squares for the model with no samples left out calculated as the sum of the squares of the differences between each $y_{ik}$ and the mean of all $y_{ik}$. If $R < 1$, then the predictions are improved by including an $\alpha$ term. Otherwise no such term is added.

The samples are then redivided into some number of groups (between 4 and 7) and a leave-one-group-out cross validation is performed. For each group left out, the first PC is computed using the non-left-out data. The $Y$-variables are then predicted for the left-out samples using the above described model. The validation repeats until each group has been left out once. The differences between the predicted and actual classes for each group left out are used to compute the ratio:

$$ R = \frac{PRESS}{S_\epsilon} \quad (2.24) $$

where $S_\epsilon$ is the residual sum of squares for the current model. If $R < 1$, then the predictions are improved by including this first PC. Otherwise, the PC did not add to the predictive ability and the best number of PC’s is zero. This process is repeated including an additional PC each iteration. The validation stops once an iteration’s ratio $R$ is less than 1 and the optimal number of PC’s is
one less than the current iteration.

This method of validating the number of components to be included in the model is used by Trygg and Wold in their validation of O-PLS with experimental data (Trygg and Wold, 2002).

$R^2$, the coefficient of multiple determination, is a measure of the adequacy of a fitted model. It explains what percentage of the total variation in $Y$ is explained by the model. It is calculated as follows:

$$R^2 = \frac{SSR}{SSY} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

(2.25)

where $\bar{y}$ is the $y$ mean across all samples, $\hat{y}_i$ is the predicted value for a sample $i$, and $y_i$ is the $y$ value for sample $i$ (Walpole et al., 1998, pp. 427-428). $R^2$ is often easily calculated from an ANOVA table, a common output of many popular statistical software packages. It is commonly utilized because it is easy to interpret and only requires the sum of squares to calculate (Weisberg, 1985, pg. 19).

$Q^2$, often referred to as the $R^2$ of prediction, reflects the prediction performance. It is calculated as follows:

$$Q^2 = 1 - \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^{n} (e_i^2)}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

(2.26)

where PRESS is the Predicted REsidual Sum of Squares (see Section 2.2.1.1), $\bar{y}$ is the $y$ mean across all samples, and $y_i$ is the $y$ value for sample $i$ (Walpole et al., 1998, pg. 447). The closer to 1 the $Q^2$ value is, the more predictive capability the model exhibits. Any $Q^2$ value less than 1 shows the model has no predictive power Weljie et al. (2007).
2.2. LITERATURE SURVEY

Figure 2.8: Permutation plot of $R^2$ or $Q^2$ for random class assignments

The $X$-axis in a permutation plot is the correlation coefficient between the original data and the permuted data, the $Y$-axis is the $R^2$ or $Q^2$ value, and the data points are the $R^2$ and $Q^2$ values for a given permutation of the $Y$ matrix (class assignments). Reproduced from Chang et al. (2007).

2.2.4.2 Permutation Test

Chang et al. employed a permutation test to test their O-PLS-DA models for overfitting (Chang et al., 2007). The authors randomly permuted the class labels for their testing samples, then calculated both an $R^2$ and $Q^2$ (discussed in Section 2.2.4.1). To permute the class labels means to randomly shuffle the class labels while maintaining the number of samples in each class found in the original data. This permutation is performed several times (perhaps up to 100 times) in order to create a plot of $R^2$ and $Q^2$ values for the various permutations (see Figure 2.8).

The $y$-axis represents the $R^2$ and $Q^2$ values and each point represents a permutation. The $x$-axis represents a correlation coefficient between the permuted class label vector and the original
class label vector. The larger the correlation coefficient, the closer to the original class assignments this particular permutation is. The point with a correlation coefficient of 1.0 represents the original class labels remaining unaltered.

A trend with a positive slope indicates that the original class labeled data performs better in the model than the permuted class-label data. A negative or zero slope indicates that the permuted data performs at least as well as the original data and that the model is likely to be overfit. Weljie et al. employed this method of testing for an overfit model (Weljie et al., 2007).

### 2.2.4.3 Randomization t-test

Hilko van der Voet proposed a randomized t-test for comparing the predictive accuracy of two methods (van der Voet, 1994). This particular validation method is applicable both for comparing two completely different methods and for comparing the same method with different numbers of components. For each comparison, square differences are computed as follows:

\[ d_i = e_{Ai}^2 - e_{Bi}^2 \]  

(2.27)

where \( i \) is an observation between 1 and the total number of observations, and \( e_{Ai} \) and \( e_{Bi} \) are the errors between the actual \( y \) values and the predicted \( y \) values for observation \( i \) from model A and model B respectively. For some number of iterations (199 or more, suggests van der Voet), random signs (+ or -) are assigned to each of the \( d_i \) values. The absolute value of the mean of these random signed differences are compared to the absolute value of the mean of the original differences. A p-value is computed by adding the number of iterations in which the randomly-signed mean was greater than or equal to the original mean and dividing by the number of iterations. A p-value
2.2. LITERATURE SURVEY

greater than some significance threshold indicates that the two models do not have a significantly different predictive power.

The author uses this randomized t-test first to compare two different Multiple Linear Regression (MLR) models and a PLS model. The validation model was able to show that the two different MLR models were not significantly different, even though the Mean Squared Error of Prediction (MSEP) showed a difference that may be construed as meaningful. The PLS model was significantly different ($p \leq 0.005$) from the first MLR model even though its MSEP was closer the first MLR model than the second MLR model was. van der Voet suggests that a simple inspection of the MSEP values may mislead comparisons between methods.

Secondly, van der Voet used this randomized t-test to determine the number of components, or latent vectors, to use in a PLS model. Models with components from 0 to 10 were constructed. The model with the minimum MSEP, in this case the 9 component model, was compared to the other 10. The PLS models with 6, 7, 8, and 10 components were all considered not to be significantly different than the 9 component model. Thus, the model with 6 components would be chosen because it is less complex and does not have significantly less predictive ability than the model with the optimal MSEP.
Preliminary Biomarker Feature

Identification Analyses

3.1 Experimental Data Set

All live animal data collection was performed by the lab of Dr. Nicholas DelRaso at the Air Force Research Laboratory’s Human Effectiveness Directorate. All preprocessing was performed by the lab of Dr. Nicholas Reo at the Wright State University’s Department of Biochemistry and Molecular Biology and Boonschoft School of Medicine.

3.1.1 Animals

All protocols for handling laboratory animals were approved by the Wright-Patterson Air Force Base Institutional Animal Care and Use Committee (IACUC) and meet appropriate Federal guidelines. Fisher 344 rats (approx. 250 g) were chosen for this study because of their inbred characteristics. Animals were obtained from Charles Rivers Laboratory equipped with jugular vein catheters and were allowed to acclimate for seven days prior to the start of the study. Animals were then
3.1. EXPERIMENTAL DATA SET

placed into either a control or chemical exposed group and housed individually in metabolism
cages for the duration of the experiment. All rats were given ad libitum access to food (Purina Cer-
tified Rat Chow # 5002) and fresh conditioned reverse osmosis water. The housing environment
was maintained on a 12 hour light-dark cycle at 25 °C, and all animals were examined by Vivarium
personnel twice daily to ensure their health and well being. Treated animals were administered a
single, sub-lethal, 50 mg/kg dose of α-naphthylisothiocyanate (ANIT) in corn oil vehicle via gav-
age, while control animals were given an identical volume of corn oil vehicle only. Urine samples
were collected pre-dose and at 24, 48, 72, and 96 hours post-dose for the duration of the experi-
ment using metabolism cages with collection containers chilled on dry ice. All urine samples were
stored at -40 °C prior to analysis by NMR spectroscopy.

3.1.2 Urine Sample Preparation

Urine samples for NMR analysis were prepared as described by Robertson et al. and modified
as follows (Robertson et al., 2000). Urine samples were thawed at 4 °C overnight then allowed
to come to room temperature just prior to NMR sample preparation. A 600 µL aliquot of urine
was transferred to a 1.5 mL Eppendorf tube, mixed with 300 µL of a phosphate buffer (0.2 M
monosodium phosphate and 0.2 M disodium phosphate, pH 7.4), and allowed to equilibrate for ten
minutes. Samples were then centrifuged at 5000 rpm for ten minutes to remove any particulate
matter and 550 µL of supernatant was transferred to a 5mm NMR tube. An internal standard
consisting of 150 µL of trimethylsilylpropionic (2, 2, 3, 3 d₄) acid (TSP) dissolved in deuterium
oxide was added at a final concentration of 2 mM.
3.1. EXPERIMENTAL DATA SET

3.1.3 $^1$H NMR Spectroscopy

$^1$H NMR spectra were acquired on a Varian INOVA operating at 600 MHz at a probe temperature of 25 °C. Water suppression was achieved using the first increment of a NOESY pulse sequence, which incorporated saturating irradiation (on-resonance for water) during the relaxation delay (2 s presaturation) and the mixing time. Samples acquisition parameters included a mixing time of 38 ms, a 4.0 s acquisition time, and data were signal averaged over 64 transients using an interpulse delay of 9.05 sec.

3.1.4 Data Preprocessing

NMR spectral data were processed using Varian software and employing exponential multiplication (0.3 Hz line-broadening), Fourier transformation, and baseline flattening (fifth-order polynomial and spline fitting routines). Spectra were subdivided into 280 regions (bins) of 0.04 ppm width for integration using Varian Binning software. The residual water signal (∼4.8 ppm) and the urea signal in urine spectra (∼5.8 ppm) were excluded from the analyses. The integrated bin areas were then normalized to the TSP signal intensity. A second normalization was performed by summing 256 bins over the entire metabolite range, which excludes the region of TSP signal (0.53 to -0.28 ppm).

The resulting data were comprised of a set of 256 features (bins) for each of 5 time points (predose, 24 hr post-dose, 48 hr post-dose, 72 hr post-dose, and 96 hr post-dose) for 5 toxin-injected treatment rats and 13 vehicle-only-injected control rats.
3.2 METHODS

3.2 Methods

As a method of determining which variable-at-a-time tests isolate desirable bin/time point combinations, bin/time point combinations selected by the tests are analyzed. The three variable-at-a-time methods compared are the fold test, the unpaired t-test of differences, and the paired t-test with vehicle removal. The performance of the three methods are compared based on the following subjective criteria: ability to identify consistent differences between treatment and control samples, ability to identify consistency across multiple time points, ability to identify vehicle effects found in both treatment and control animals, and ability to identify responses found in treatment samples above those attributed to vehicle effects.

3.2.1 Fold Test

The fold test is often used for identifying differentially expressed genes in microarray analysis and is applicable for NMR data (Cui and Churchill, 2003; Marchal et al., 2002). It is a direct comparative approach in which the log ratio of the two conditions are evaluated and the resulting ratio is compared to an arbitrary cut-off. The equation for the fold test is defined as follows:

\[
fold\ ratio = \log(\bar{y}_{i1}) - \log(\bar{y}_{i2})
\]

(3.1)

where \(\bar{y}_{i1}\) and \(\bar{y}_{i2}\) are the means of the two groups to be tested. A two-fold difference is a common cutoff to determine an interesting difference; however, no statistical significance can be attributed to this cutoff.

For this data set, the log ratios of pre-dose measurements to each of the 24, 48, 72, and 96
3.2. METHODS

hour measurements of the same bin were calculated. The means of these ratios were compared to a cut-off ratio of 2. All bin/time point combinations that exhibited a two-fold ratio are retained for visual comparison between methods. These significant bins will be referred to as part of the 2F (two-fold) group.

3.2.2 The t-test

The simplest statistical method of identifying responsive bins is the t-test. A two-sample t-test between the pre-dose measurements and each subsequent time point is not applicable for this data set. The t-test’s assumption of independent sets of data cannot be met. The fact that the pre-dose measurements and the 24, 48, 72, and 96 hour time points come from the same animal creates a situation where the samples are not independent.

Instead, a two-sample t-test versus a set of controls using the differences between the pre-dose measurements and each subsequent time point is performed. These differences are assumed to be nearly normally distributed (Clarke, 1994). With such a small sample size ($n < 15$), it may not be safe to assume that the treatment measurement and the control measurements have equal variances. Thus, a two-sample t-test assuming unequal variances is used.

The differences between pre-dose measurements and the 24 hour measurements for each bin are calculated for both the treatments and the controls. This process is repeated for the 48, 72, and 96 hour measurements. These differences are then compared using a two-sample t-test assuming unequal variances. The bin/time point combinations are then ranked by p-value for comparison. Any bin/time point combination p-value that fell outside a p-value threshold value of 0.01 is used for the comparison between methods. These significant bins will be referred to as part of the UTT.
3.2. METHODS

An alternative to this unpaired t-test of differences is the paired t-test. A paired t-test between the pre-dose measurements and each following time point takes advantage of the fact that the pre-dose measurements and the subsequent time point measurements are taken from the same subject. Any inherent variability between subjects is presumably mitigated by each subject serving as its own control (Marchal et al., 2002).

3.2.2.1 Removal of Vehicle Effect

In a previously presented method (Kelly et al., 2007), a change to the standard paired t-test was proposed to account for the effect of the corn oil vehicle within which the toxin is delivered into the animal. These vehicle effects have the ability to confound results, causing a bin/time point to be considered significantly different from the pre-dose measurement when that significant difference has nothing to do with the injected toxin. To account for this effect, the vehicle control subjects are administered the same vehicle injections without any toxin.

For this data set the pre-dose and the 24 hour measurements from treated rats are compared using a paired t-test and the p-values are calculated for each bin. This process is repeated for the 48, 72, and 96 hour measurements from treated rats, along with the 24, 47, 72, and 96 hour measurements from control rats. Any bin/time point that is significantly different from the pre-dose measurement in either group is considered to be exhibiting a possible vehicle effect. In the event that a bin/time point exhibits both a toxin response and a vehicle effect simultaneously, an unpaired t-test between the treatment and control groups at that bin/time point is performed. If there is a significant difference (p-value threshold of 0.01) between the treatment and control groups at that
bin/time point, then the bin/time point is not removed because there appears to be a toxin response beyond the vehicle effect. If there is not a significant response, the bin/time point is removed from further analysis. These remaining bin/time point combinations are then ranked by p-value for comparison. In addition, a threshold value of 0.01 is used for visual comparisons between methods. These significant bins will be referred to as part of the PTT (paired t-test) group.

3.2.3 Multiple Test Correction Comparison

The nature of measuring the same sample at several time intervals and testing hundreds of hypotheses on these measurements requires an adjustment to the p-value. Without an adjustment, the p-values do not necessarily carry any statistical significance and can only be used as a method of ranking. Several options for multiple test correction have already been discussed in 2.2.3.

In order to assess relative conservativeness among methods, representing groups which control for different error rates, paired t-test p-values from the data described in Section 3.1 are corrected by the Bonferroni correction, Holm’s step-down method, and Benjamini-Hochberg correction. For the sake of comparison non-corrected p-values are also included in the comparison.

3.2.3.1 Bonferroni Correction

One simple and common approach is to perform a Bonferroni correction. The Bonferroni correction is considered conservative and assumes independence of the test statistics (Dudoit et al., 2002). If independence cannot be assumed the test correction becomes extremely conservative. In this particular data set, along with similar data sets, independence cannot be assumed. Many metabolites will react similarly to biological changes, as they may come from the same or related
3.2. METHODS

pathways. Moreover, many metabolites are represented by multiple bins. Citrate for example is found in this particular NMR spectrum in bins 194 and 198, due to multiple types of hydrogen atoms in the molecule. Therefore in this data set the assumption of independence cannot be met. The Bonferroni correction may still be used in situations of dependence, but the correction becomes highly conservative and type-II (false negative) error is likely to increase (Sankoh et al., 1997).

The Bonferroni correction, because of its conservative nature, is difficult to use when a large number of tests are being performed. The adjusted p-value becomes exceedingly large, making the null hypothesis nearly impossible to reject. It was chosen to be part of this comparison because it is a standard among multiple test corrections. When multiple test corrections are compared, the Bonferroni correction is almost invariably included (Draghici, 2002; Dudoit et al., 2002).

The Bonferroni correction is applied to the paired t-test p-values. The data set includes 4 time points (24, 48, 72, 96 hrs) and 240 bins (excluding any bins that where each sample’s value is 0) for a total of 960 paired t-tests. Therefore, for the Bonferroni correction formula:

\[ p_c = kp_o \]  

(3.2)

where \( p_o \) is the original p-value of the test, \( k \) is the number of hypotheses one wishes to test, and \( p_c \) is the corrected p-value of the test becomes:

\[ p_c = (960)p_o \]  

(3.3)

for this data set.
3.2. METHODS

3.2.3.2 Holm’s Step-down Correction

Another simple approach to correcting for multiple tests is the Holm’s step-down method. Holm’s method has the same assumption of independence of the test statistic as the Bonferroni correction. The difference between the two corrections is that the Holm’s correction decreases the amount of correction as the p-value gets larger. The Holm’s correction, therefore, is slightly less conservative. This may only be observed in a situation where a large number of the tests performed have small, statistically significant p-values. Thus the Holm’s correction is still difficult to use in most situations.

The Holm’s correction is applied to the paired t-test (without vehicle removal) p-values. The data set includes 4 time points (24, 48, 72, 96 hrs) and 240 bins for a total of 960 paired t-tests. Therefore, for the Holm’s step-down correction formula:

\[ p_{c1} = kp_{o1} \]  \hspace{1cm} (3.4)

\[ p_{cj} = \max(p_{cj-1}, (k - j + 1)p_{oj}) \quad \text{for} \quad 2 \leq j \leq k \]  \hspace{1cm} (3.5)

for this data set where the p-values are sorted from smallest to largest, \(p_o\) is the original p-value of the test, \(k\) is the number of hypotheses one wishes to test, and \(p_c\) is the corrected p-value of the test becomes:

\[ p_{c1} = (960)p_{o1} \]  \hspace{1cm} (3.6)

\[ p_{c2} = (959)p_{o2} \]  \hspace{1cm} (3.7)

\[ p_{c3} = (958)p_{o3} \]  \hspace{1cm} (3.8)
and so on.

### 3.2.3.3 Benjamini-Hochberg Correction

The last method to be discussed is the Benjamini-Hochberg multiple test correction, defined as follows:

\[
p_{c_1} = p_{o_1} \left( \frac{1}{k} \right) \quad (3.9)
\]
\[
p_{c_j} = p_{o_j} \left( \frac{j}{k} \right) \quad \text{for} \quad 2 \leq j \leq k \quad (3.10)
\]

where the p-values are sorted from smallest to largest, \( p_o \) is the original p-value of the test, \( k \) is the number of hypotheses one wishes to test, and \( p_c \) is the corrected p-value of the test. The correction then states to start from the \( k \)th p-value and check if the corrected p-value is less than the desired \( \alpha \). If it is not, move on to the \( k-1 \) th p-value and check if it is less than the desired \( \alpha \) and so on. If it is, then all null hypothesis with that corrected p-value or less are rejected (Benjamini and Hochberg, 1995). The correction assumes independent or positively dependant tests (i.e. no negative correlation).

The motivation behind the Benjamini-Hochberg correction is to control False Discovery Rate (FDR), as opposed to Familywise Error Rate (FWER). FWER is the probability of making at least one false positives among all tests being performed. FDR is the expected proportion of false positives over all positives results. Controlling for FDR makes a multiple test correction less conservative and more powerful at the expense of allowing more Type I error (false positives) (Shaffer, 1995). More classical methods like the Bonferroni correction and the Holm’s correction
control for FWER, making them inherently more conservative and less powerful, especially for a large number of tests.

The Benjamini-Hochberg correction is applied to the paired t-test p-values. The data set includes 4 time points (24, 48, 72, 96 hrs) and 240 bins for a total of 960 paired t-tests. Therefore, for the Benjamini-Hochberg correction formula becomes:

\[
p_{c1} = p_{o1} \left( \frac{960}{1} \right)
\]

\[
p_{c2} = p_{o2} \left( \frac{960}{2} \right)
\]

\[
p_{c3} = p_{o3} \left( \frac{960}{3} \right)
\]

and so on for this data set.

### 3.2.4 Visualizing Bins

Two methods for visualizing bins across all time points are used in this analysis. The first, an example can be seen in Figure 3.1, is a simple plot of the treatment animals, each represented by a different marker, where the time course of each of these animals can be followed across all 5 time points (pre-dose and after 24, 48, 72, and 96 hours). The y-axis represents the bin-intensity values of each animal at each time point and the x-axis represents the time at which the bin-intensity was measured.

The second, an example can be seen in Figure 3.2, is a plot of the mean intensity of the treatment (or control) animals and includes error bars representing two standard errors away from the mean. The plot covers the time course of the mean and error bars across 5 time points (pre-dose...
3.3. Results

3.3.1 Comparison of Methods

The differences between the fold test, unpaired t-test of differences, and the paired t-test with vehicle removal are analyzed using subjective criteria by comparing the differentially responsive bin/time points produced by each method. The subjective criteria included the ability to identify differences between treated and pre-dose (or control) samples, the ability to identify consistency and after 24, 48, 72, and 96 hours). The y-axis represents the bin-intensity values of each animal at each time point and the x-axis represents the time at which the bin-intensity was measured.
Figure 3.2: Example of mean with error bar plot

3.3. RESULTS

across time points and the ability to identify both vehicle effects and responses above and beyond the vehicle effects. Table 3.0(a) shows that the paired and unpaired methods result in approximately the same number of responsive bin/time points; however, the two-fold method found fewer responsive bin/time points. Table 3.0(b) shows that there is a considerable overlap in the specific bins selected by the paired and unpaired methods, and that only 6 bin/time points selected by all three methods. Table 3.0(c) shows that both the unpaired and paired methods select bin/time points that the fold test method does not consider significant. These differences are examined in Section 3.4.1.
3.3. RESULTS

(a) Cardinality

<table>
<thead>
<tr>
<th>PTT</th>
<th>UTT</th>
<th>2F</th>
</tr>
</thead>
<tbody>
<tr>
<td>86</td>
<td>74</td>
<td>10</td>
</tr>
</tbody>
</table>

(b) Intersection

<table>
<thead>
<tr>
<th>PTT ∩ UTT</th>
<th>PTT ∩ 2F</th>
<th>UTT ∩ 2F</th>
<th>PTT ∩ UTT ∩ 2F</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>8</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

(c) Set difference (row - column)

<table>
<thead>
<tr>
<th></th>
<th>PTT</th>
<th>UTT</th>
<th>2F</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTT</td>
<td>0</td>
<td>26</td>
<td>78</td>
</tr>
<tr>
<td>UTT</td>
<td>14</td>
<td>0</td>
<td>68</td>
</tr>
<tr>
<td>2F</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.1: Comparing the differentially responsive bin results from the paired t-test (PTT), unpaired t-test (UTT), and two-fold test (2F).

PTT, UTT, and 2F are the sets of differentially responsive bin/time points using the paired t-test, unpaired t-test, and two-fold test methods, respectively. Table 3.0(a) shows the number of bin/time points found with each method. Table 3.0(b) shows the overlap between methods in the number of bin/time points found. Table 3.0(c) shows the difference between methods in the number of bin/time points found.

3.3.1.1 Removal of Vehicle Effects

Of the entire set of 960 bin/time points, 7.7% (74) are identified as vehicle effects because both the treatment group and control group’s time point measurement are significantly different from the pre-dose measurement. Of these 74 bin/time points, 64.9% (48) are removed from analysis due to vehicle effect. An example is shown in Figure 3.3. The 24 hour time point is removed from analysis because both the treatment’s paired t-test (p = 0.0030) and control’s paired t-test (p = 1.18e-05) are significant and the difference between the two groups is not significant (p = 0.669).

The remaining 35.1% (26) are not removed because the treatment group is significantly different from the control group, suggesting a response above and beyond a vehicle effect. An example is shown in Figure 3.4, where the 24 hour time point is not removed from analysis despite an ap-
3.3. RESULTS

Figure 3.3: Example of a time point removed due to vehicle effects

Bin 201, 24 hr is likely exhibiting a vehicle effect. The bin/time point would be removed by paired t-test (treatment p-value = 0.00302, control p-value = 1.182e-05, difference between treatment and control p-value = 0.6686).
3.3. RESULTS

<table>
<thead>
<tr>
<th></th>
<th>0.1</th>
<th>0.05</th>
<th>0.01</th>
<th>0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>No correction</td>
<td>430</td>
<td>296</td>
<td>134</td>
<td>84</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Holm’s</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Benjamini-Hochberg</td>
<td>187</td>
<td>73</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3.2: Number of bin/time points significant at different p-value thresholds

parent vehicle effect. Both the treatment’s paired t-test (p = 0.00062) and control’s paired t-test (p = 0.00021) at the 24 hour time point are significant, however the difference between the two groups is also significant (p = 0.0015). Thus the response cannot be attributed solely to the vehicle and the bin is be removed.

3.3.2 Comparison of Multiple Test Corrections

The differences between the the Bonferroni, Holm’s step-down, and Benjamini-Hochberg multiple test corrections are analyzed using the paired t-test. Table 3.2 shows the number of bin/time points considered significant at several significance levels (p = 0.1, 0.05, 0.01, 0.005) by each correction, along with the number considered significant without a correction. From these results, it can be inferred that the Benjamini-Hochberg correction is rather less conservative than the Bonferroni and Holm’s corrections for this data set.

In order to determine a suggested p-value threshold for each multiple test correction, a formula for determining how differentiated the treatment and control groups are when only considering the bin/time points that meet the threshold should be used. For this, PCA was used to determine the separation of the groups by measuring the within-class scatter ($S_w$) and between-class scatter ($S_b$).
3.3. RESULTS

Figure 3.4: Example of a time point not removed due to vehicle effects because of an response beyond the apparent vehicle effect

Bin 174, 24 hr is likely exhibiting a response above and beyond a vehicle effect. The bin/time point would be not removed by paired t-test (treatment p-value = 0.00062, control p-value = 0.00021, difference between treatment and control p-value = 0.0015).
3.3. RESULTS

Figure 3.5: PCA goodness plot of p-values with no multiple test correction

\[ S_w \text{ and } S_b \text{ is then used to calculate a PCA Goodness of separation:} \]

\[
PCA \text{ Goodness} = \frac{(S_b)}{N} \tag{3.14}
\]

where \( N \) is the number of bin/time points considered significant by a multiple test correction at a significance level. Figure 3.5 shows a plot of the PCA goodness where no correction has been implemented. The plot is reproduced in Figure 3.6, Figure 3.7, and Figure 3.8 for the data corrected by the Bonferroni, Holm’s, and Benjamini-Hochberg methods respectively. The x-axis for the PCA goodness plots is the desired p-value significance level and the y-axis is the PCA goodness. Further discussion of these results can be found in Section 3.4.2.
3.3. RESULTS

Figure 3.6: PCA goodness plot of p-values with Bonferroni correction

Figure 3.7: PCA goodness plot of p-values with Holm’s correction
3.4 Discussion

3.4.1 Comparison of Methods

The fold test performed as expected. Only the most extreme differences between the pre-dose and later time point measurements are selected. The fold test includes no sensitivity for consistency and has no ability to attribute statistical significance to any selected bin/time points. The 72 hour time point for bin 82, shown in Figure 3.9, is an example of a bin/time point considered significant by the two-fold test, but not by either other method. It is apparent by this bin/time point, along with several others, that the fold test is not sensitive to the amount of variability between animals. This is not a good result for identifying biomarkers. Large amounts of variability between samples is likely to be nothing more than just that, variability. With few samples, any substantial variability
3.4. DISCUSSION

diminishes confidence in that bin.

The unpaired t-test considers consistency of the group (treatment or control) and attributes statistical significance to the results. This is important for isolating potential biomarkers as a clear difference between treatment and control samples would indicate that a change has occurred in one group of animals and not the others. Given that the conditions these two groups endure are identical, the only difference between them is the injected toxin. For the data explored here, this not only dramatically increases the number of selected bin/time points as compared to the fold test but also increases the quality bin/time points as measured by statistical significance.

The paired t-test is able to provide the same improvements over the fold test as the unpaired t-test while also providing a more effective method of vehicle removal and considering consistency among individual samples (when all samples move in the same direction by a similar amount), not only groups. Toxin responses above and beyond a vehicle effect (the situation described in Section 3.3.1.1), are identified by the paired t-test more often than the unpaired t-test. A potential biomarker could be affected by both the vehicle and by the toxin. In this situation, it is possible that the treatment response, (through the sum of both the vehicle and toxin effects) is significantly different than the control response caused only by the vehicle. The response is different between the two groups and therefore should not be removed from consideration. Figure 3.10 demonstrates this situation at the 48 hour time point. Several instances of the unpaired t-test failing to remove a suspected vehicle effect also occur, as in Figure 3.11 at the 24 hour time point.

Another situation where the paired t-test appears to have an advantage over the unpaired t-test of differences is shown in Figure 3.12. The average difference between the 24 hr time point and the pre-dose measurement are small in the treatment and control for different reasons. The
Bin 82, 72 hr is considered significant by the two-fold test (fold ratio = \(-3.84\)), but not by the unpaired t-test \((p = 0.0920)\) or paired t-test \((p = 0.0318)\) at a 0.01 p-value threshold.
3.4. DISCUSSION

Figure 3.10: Example of a response beyond a possible vehicle effect

Bin 171, 48 hr is not considered significant by the unpaired t-test ($p = 0.0129$) but is by paired t-test ($p = 0.00317$) at a 0.01 $p$-value threshold because of a response beyond the vehicle effect (control $p$-value = 7.475e-05, difference between treatment and control $p$-value = 0.000350).
3.4. DISCUSSION

Figure 3.11: Example of missed vehicle effect by the unpaired t-test

Bin 166, 24 hr is likely exhibiting a vehicle effect, but is still considered significant by the unpaired t-test (p = 0.00727). The bin/time point has been removed by paired t-test (p = 0.000468) because of vehicle effect (control p-value = 6.311e-06, difference between treatment and control p-value = 0.0141).
3.4. DISCUSSION

control groups measurements are highly varied, resulting in a average near 0 while the treatment
groups measurements are small, but upward and uniform. The unpaired t-test does not recognize
this consistency across the time points where the paired t-test does. This consistency consideration
is also important for potential biomarker identification. A true biomarker of toxicity will affect
all samples similarly, producing a consistent response from pre- to post-dose measurements. Even
small changes must be considered as any metabolite could be the best for determining toxicity of
an subject.

Examples of bin/time points that were identified by the two-fold test, unpaired t-test, and
paired t-test can be seen in Figures 3.13 - 3.16.

3.4.2 Comparison of Multiple Test Corrections

Without a multiple test correction, a large number of bins are considered statistically significant
(134 at a 0.01 p-value threshold). In Figure 3.5, the largest jump in PCA goodness occurs at a
desired p-value threshold of 0.0001. This signifies a point at which the PCA cluster of treatment
samples and the PCA cluster of control samples are tightly compacted, but maintain separation
from each other.

The Bonferroni correction and Holm’s step-down method performed similarly throughout the
comparison as the number of statistically significant bin/time points remained the same in Table 3.2
and the PCA goodness plots in Figures 3.6 and 3.7. The conservative nature of the two corrections
is evident by the fact that only 2 bin/time points being statistically significant at a 0.01 p-value
threshold. Also, the p-value threshold in which the PCA goodness first indicates good clustering
is 0.125, a high p-value.
Figure 3.12: Example of paired t-test identifying consistency among subjects

Bin 32, 24 hr is not considered significant by the unpaired t-test (p = 0.665) but is by the paired t-test (p = 0.00184) at a 0.01 p-value threshold.
Figure 3.13: Example 1 of a bin/time point considered significant by all three methods

Bin 194, 48 hr is considered significant by the two-fold test (fold ratio = -2.314), unpaired t-test (p = 3.738e-06) and paired t-test (p = 0.000129) at a 0.01 p-value threshold.
3.4. DISCUSSION

Figure 3.14: Example 2 of a bin/time point considered significant by all three methods

Bin 201, 48 hr is considered significant by the two-fold test (fold ratio = -3.369), unpaired t-test (p = 1.291e-05) and paired t-test (p = 0.000133) at a 0.01 p-value threshold.
Figure 3.15: Example 3 of a bin/time point considered significant by all three methods

Bin 194, 48 hr is considered significant by the two-fold test (fold ratio = -2.286), unpaired t-test (p = 4.991e-08) and paired t-test (p = 3.159e-05) at a 0.01 p-value threshold.
Figure 3.16: Example 4 of a bin/time point considered significant by all three methods

Bin 194, 48 hr is considered significant by the two-fold test (fold ratio = -2.070), unpaired t-test \( (p = 1.074\times10^{-9}) \) and paired t-test \( (p = 9.829\times10^{-6}) \) at a 0.01 p-value threshold.
3.4. DISCUSSION

The Benjamini-Hochberg correction is clearly different from the previous two corrections. While only 8 bin/time points are statistically significant at a 0.01 threshold in Table 3.2, 73 are significant at a threshold of 0.05. This is in clear contrast to the conservative 6 bin/time points the Bonferroni and Holm’s corrections produce. In addition, the p-value threshold of a first good PCA goodness value in Figure 3.8 is 0.01, less than the 0.125 threshold of the previous two corrections.

Considered jointly, the evidence in Table 3.2 and Figures 3.5 - 3.8 provide insight to the relative conservativeness of the corrections compared. In order to achieve the same PCA goodness, the p-value threshold for the Bonferroni or Holm’s corrected paired t-test must be over 10 times larger. These differences are an indication that a researcher must decide whether to control for familywise error rate or for false discovery rate before performing a multiple test correction. If the nature of the experiment is such that a single false positive is to be avoided at all costs, such as a drug that will help a patient with a certain metabolic makeup but will harm a patient with any other metabolic makeup, the Bonferroni or Holm’s correction may be most appropriate. However, in situations where a researcher would like to minimize the false positives to an acceptable number, a less conservative correction such as Benjamini-Hochberg may be sufficient and perhaps more informative.

For this metabolomic toxicology data set, two clear alternatives for which correction to perform are present. The intent of the experiment is likely either to sample a larger number of potential biomarkers for further analysis or to select a small number of biomarkers that are assuredly altered by the toxicity of the samples. A larger number of potential biomarkers will be considered significantly differentiated from pre-dose to post-dose (or control to treatment) if the Benjamini-Hochberg correction is applied to the p-values. This comes at the price of allowing some amount
3.5. **CONCLUSION**

of false positives to likely be included. For the intent of capturing all the potential biomarkers in an experiment, this is likely an acceptable consequence. It is more important to include as many potential biomarkers as possible in order to capture as many of the true, yet unknown, biomarkers of toxicity. A smaller set of potential biomarkers will be statistically significant if the Bonferroni or Holm’s corrections is used. However, a higher confidence that these significant potential biomarkers are true positive results make it the probable better choice for the intent of isolating a potential biomarker with high certainty.

### 3.5 Conclusion

The criteria for what makes a biomarker identification method in metabolomic toxicology more closely matches the criteria in which the paired t-test with vehicle effect removal was preferred. A biomarker should respond consistently across the samples from the pre-dose measurement to a post-dose measurement. A biomarker should respond differently in the treatment samples when compared to the control samples. Finally, a vehicle effect should be identified and removed from consideration, unless there is a response which exceeds the vehicle effect. The fold test fails to reliably meet any of the criteria. The unpaired t-test was effective at meeting all three criteria. The paired t-test is more able to identify responses above the vehicle effect and can more effectively identify consistency among individual samples. Thus, the paired t-test with vehicle effect removal appears the more complete of the three tests, having distinct advantages over the other two methods, for which examples have been shown. This conclusion helps answer the question of how the rat is being effected by the toxin. The paired t-test with vehicle effect removal gives insight into which metabolites may be altered by exposure.
3.6. COMPUTATIONAL HARDWARE AND SOFTWARE

Multiple test correction is necessary when performing the number of tests needed to completely analyze the large number of bins contained in this data set. This is so that results caused by random chance are not incorrectly rejected or accepted. It is often the case that the ultimate goal of a metabolomic toxicology experiment is to isolate all potential biomarkers of toxicity. In this situation a single bin labeled a potential biomarker incorrectly is not likely to be a problem. Thus it is not necessary to control for the familywise error rate. A wider number of statistically significant bins will be identified by the t-test with a Benjamini-Hochberg correction and a suggested p-value threshold of 0.01, thus leaving a wider set of potential biomarkers to investigate. This will also give a broader picture of the metabolic pathways affected by the toxin. It may also be the case that a higher-certainty indicator of toxicity is desired, where a single error could prove detrimental to an experiment. In this situation, a more conservative Bonferroni or Holm’s correction should be chosen, with a suggested p-value threshold of 0.125. A smaller range of bins will be identified as statistically significant, but the chance of a false positive is also reduced significantly.

3.6 Computational Hardware and Software

All analyses in this chapter were performed using the following computers:

1. Laptop running Microsoft(R) Windows Vista(R) Home Premium 64-bit Service Pack 1, Intel(R) Core2(R) Duo P8400 (2.26 GHz), 4 GB RAM, 320 GB HDD

2. Desktop running Microsoft(R) Windows XP(R) Professional 32-bit Service Pack 3, Intel(R) Pentium(R) 4 (3.73 GHz), 2 GB RAM, 60 GB HDD

3. Desktop running Ubuntu 7.04, Intel(R) Pentium(R)4 (3.73 GHz), 2 GB RAM, 60 GB HDD
3.7 DATA ACQUISITION ACKNOWLEDGEMENT

4. Desktop running Microsoft(R) Windows Vista(R) Ultimate 64-bit Service Pack 1, Intel(R) Core2(R) Duo E6400 (2.13 GHz), 4 GB RAM, 250 GB HDD

All analyses were performed using the following programs:

1. Matlab(R) R2007b for Windows
2. Microsoft(R) Excel(R) 2007
3. Ruby 1.8.6-26

3.7 Data Acquisition Acknowledgement

The live animal data acquisition described in this thesis was supported in part by grants from the Air Force Research Laboratory (AFRL), Mantech Environmental Technology, Inc. (Contract No. ManTech/WBI-002), Alion Science and Technology (Contract No. SUB1174146RB), and The Henry M Jackson Foundation for Advancement of Military Medicine, Inc. (Contract No.132633).
Biomarker Feature Selection Method

Analyses

4.1 Synthetic Data Sets

The correct answer to what information a data set contains must be known for proper, statistical comparison of methods. This is not always possible in a real data set such as the one described in Section 3.1. It is often more beneficial to use a synthetic data set in which the design of the data set includes a “known” answer. A synthetic data set should emulate a real life data set as closely as possible. It is often difficult to create data that encompasses all of the unique characteristics of real-world data. Thus simple data sets are often used to compare methods. Results using simple data sets must be presented with the caveat that the methods were not tested on the type of complex data normally found in the real world. This is the motivation behind the yet unpublished manuscript by Anderson et al.: Characterization of 1H NMR spectroscopic data and the generation of synthetic validation sets. The manuscript describes a method of deconstructing real-world NMR spectra into their component parts (peaks, baseline, and noise), then using the information about these parts to create new synthetic spectra with desired, real-world characteristics.
4.1. SYNTHETIC DATA SETS

4.1.1 Characterization of Real-World Spectra

The data set described in Section 3.1 yielded 22 pre-dose spectra. These are the real-world, experimental spectra used by Anderson et al. to construct the synthetic data sets used for the analyses described and performed in this chapter. All peaks from each spectra are modeled using Gaussian-Lorentzian functions. Peak centers, Gaussian standard deviations/Lorentzian widths at half height, proportions of peak Lorentzian, and peak magnitudes are characterized algorithmically by solving a nonlinear curve-fitting problem. A baseline was likewise fit using a nonlinear curve-fitting algorithm. The peaks and baseline are then combined into a global model, which is modified until a normally distributed residual with a mean of zero is achieved.

Once fit, the peaks are classified into 3 groups: baseline, background, and foreground. Baseline peaks are broad, and generally only distort the baseline. They are considered to vary from spectrum to spectrum. Foreground peaks are those which are clearly observable and are assumed to be observable throughout all the spectra in a data set. Distributions for the peak parameters described above, distances between peaks, baseline intensities, and relationships between peaks and the baseline are calculated, while distributions for the variability of peaks between spectra are estimated by first matching peaks using a simple algorithm, then characterizing the differences between matched peaks.

4.1.2 Creation of Synthetic Control and Treatment Data Sets

Each synthetic data set is based upon a single base spectrum. A base spectrum is constructed by generating a number of signal peaks (foreground and background peaks) and a baseline, both
4.1. SYNTHETIC DATA SETS

of which are based on the distributions calculated above. Finally, baseline peaks and noise are added to the spectra. Variability of peak parameters based on the distributions described above are introduced to all synthetic spectra based on one base spectrum. Likewise, variability in the baseline is added to each synthetic spectra.

In order to evaluate methods for identifying differences between treatment and control samples, synthetic treatment data sets will also need to be created. Each base spectrum is used as the model for a synthetic treatment base spectrum. Each peak height within the base spectrum is changed by a percentage pulled from a distribution of percent differences. This distribution has a mean of 0% difference and a standard deviation of 80%. The percent differences are bound by a minimum of -100% and a maximum of 200%. From this point, several different treatment base spectra can be created by adjusting the percentage of peaks not responding. Consider a treatment base spectrum with 50% of peaks not responding. The heights of 50% of the peaks in this base spectrum are maintained at the same magnitude as the ”control” base spectrum (the base spectrum the treatment base spectrum was based upon). At this point, the treatment spectra are created using the treatment base spectrum in the same way that the control spectra are created using the control base spectrum. The treatment and control spectra are similar in the number of peaks, but differ in peak heights and contain the variation from spectrum to spectrum and noise.

A total of 30 different synthetic control data sets were created, each containing 20 ”control” samples, using preliminary characterization distributions. Four treatment data sets based on each synthetic data set were created with 0, 50, 75, and 90 percent of the signal peaks not changed from the corresponding control data set. Each of the 120 treatment data sets contained 20 ”treatment” samples.
O-PLS-DA is a more complex, all-variables-at-a-time method for predicting a dependant variable (or variables) from a set of independent variables and for determining the relative importance of these independent variables. A simple, bin-at-a-time method like the Student’s t-test can perform the same tasks, but how does the performance of O-PLS-DA and the t-test compare? A comparison of the two methods’ ability to rank the importance of variables and classify new samples are performed on the synthetic NMR data sets described in Section 4.1. In addition a permutation method for determining which variables are important and which are not using O-PLS-DA are compared to a t-test p-value cutoff using synthetic NMR data.

### 4.2.1 Data Preprocessing

The comparisons described above are performed on the synthetic data preprocessed in two distinct ways. The first method is uniform binning, as described in Section 3.1.4. The second is what can best be described as the ideal case. It assumes that any spectra can be deconstructed into its component signal peaks. Then a perfect peak-matching algorithm is applied which knows what peaks match across all of the spectra. A large number of variables left, but there is no loss of sensitivity that occurs when combining peaks into bins. It is an unrealistic expectation that any method of preprocessing could produce the type of pre-processed data.

To calculate the value of a bin in a spectrum, the intensity at every data point (spaced 0.00041684 ppm apart) within the 0.04 ppm boundary is summed. To maintain the scale of the data across spectra, the mean and standard deviation of the control bins are calculated and all bins (treatment or
4.2. METHODS

current) are scaled by subtracting the mean and dividing by the standard deviation. The result becomes the value for that bin. A bin is considered to be showing a response if at least one responding peak falls within its boundaries. The criteria for a peak falling within a bin was that a point 1.5 widths at half height from the peak’s center must be within that bin’s boundary. This takes into consideration the situations where a peak falls close enough to a boundary to effect the bins on both sides. This determination of responding bins was performed on the control and treatment base spectra, such that variation and noise did not negatively influence the true answer for which bins are showing a response. Data sets preprocessed in this manner will from now on be referred to as binned data sets.

To calculate the value of a peak in a spectrum, the area underneath each peak is calculated using the magnitude, Gaussian standard deviation/Lorentzian width at half height, and proportion Lorentzian. The 0.00041684 ppm spacing between data points for the peak is maintained and the data points for the peak span 3 Gaussian standard deviations/Lorentzian widths at half height to the left and right of the peak center. The intensities at these data points are then summed and the result becomes the value for that peak. If a control base spectrum magnitude was different from the treatment base spectrum magnitude for a given peak, that peak was considered responding. Data sets preprocessed in this manner will be from now on be referred to as peak data sets.

4.2.2 Comparison as a Ranking Method

O-PLS-DA provides information about the relative importance of variables through the loadings (p). Loadings with high absolute values represent the variables most influential to the O-PLS-DA model. Loadings with similar absolute values influence the model similarly. Therefore, a simple
4.2. METHODS

ranking of variables by the absolute value of their loadings will be sufficient to evaluate O-PLS-DA as a method of ranking variables by their importance.

O-PLS-DA is performed as described by Trygg and Wold (2002). More specifically, steps 1 through 12 of the one dependent variable (in this case, the class label) outline of the O-PLS algorithm by Trygg and Wold found on page 121 are performed. At least one O-PLS component is filtered from the original data. To determine if any additional O-PLS components should be removed, the method of selecting the optimal number of components proposed by Wold (1978) and described in Section 2.2.4.1 is performed. A leave-one-out cross validation is performed after each O-PLS component is removed to determine if removing that component improved the predictive ability of the model. If not, the previous model without that component removed is restored and no further components are removed.

The Student’s t-test provides a simple method for comparing the importance of variables, though in more of an indirect way. A t-test assesses the difference between two groups and calculates a p-value that represents the chance of a result as extreme as the result found assuming that the null hypothesis is true. In the case of two groups of samples, each with a large number of variables, a p-value will represent the chance of finding a difference as large as the tested difference in a true null hypothesis situation. Therefore, the smaller the p-value, the larger the difference between the two groups of samples for that single variable, and the more important the variable in assessing the differences between the two groups. Thus, a ranking of the p-values for each variable in increasing order will be sufficient to evaluate the t-test as a method of ranking variables by their importance.

In order to determine a correct ranking for the binned data sets, the absolute differences between the binned synthetic control base spectrum and the binned treatment base spectrum are
4.2. METHODS

calculated and sorted in descending order. As described in Section 4.1, the control base spectrum
does not contain noise, peak variability, or baseline variability and the treatment base spectrum
only contains peak height differences from the control base spectrum and no noise or other vari-
ability. The correct ranking for the peak data sets are determined in the same fashion.

For this comparison the O-PLS-DA and t-test methods described above are performed on
the synthetic data sets described in Section 4.2.1. More specifically, for each of 30 data sets, 20
synthetic control samples are tested against 20 synthetic treatment samples with 0 percent of peaks
not responding. This is repeated for each data set with treatment samples with 50, 75, and 90
percent of peaks not responding. Each combination of control samples and treatment samples is
also repeated with 10 samples each and 5 samples each. A total of 360 different O-PLS-DA and
t-test tests are performed on the binned data. All 360 tests are also performed on the peak data sets
described in Section 4.2.1.

4.2.3 Comparison as a Method of Determining Important Variables

Another important piece of information that would be helpful for a researcher studying NMR
spectra is which variables (bins or peaks) are truly the important ones. Inherently, O-PLS-DA will
not distinguish the important variables from the non-important variables. The loadings results will
simply show their relative importance to each other. However, when no variables show discernable
differences between the classes, the loadings do not differ from each other as significantly. A
permutation test may be able to exploit this fact. A permutation test permutes the labels of the
samples, then compares an O-PLS-DA loadings results for these permuted-label samples to the
O-PLS-DA loadings results for the correctly label samples. If there are no responding variables,
4.2. METHODS

then an O-PLS-DA test using permuted labels should not be significantly different than an O-PLS-
DA test using correct labels. If there are responding variables present, the results of the correctly
labeled test should reflect this, and the responding variables should have larger loadings than those
not responding. The more responding the variables, the larger the difference in loadings will be.
For those tests performed on permuted-label samples, no significant differences in the loadings
should be expected, as the differences between samples no longer coincide with the classes of the
samples.

On this basis, a permutation test for O-PLS-DA which is meant to determine which variables
are important and which are not is proposed. The steps to the test are as follows:

1. Perform O-PLS-DA on treatment and control samples with correct labels. Get loadings
   rankings.

2. Start with the $n = n_{\text{start}}$ least important variables as determined by the correctly-labeled
   loadings rankings.

3. Run O-PLS-DA with correct labels using only these $n$ worst variables.

4. Permute the labels such that the same number of control and treatment samples are the same
   as the original numbers. Run O-PLS-DA with these permuted labels using only the $n_{\text{start}}$
   worst variables.

5. Repeat Step 4 $N$ times, creating a different permutation each time.

6. Calculate $R^2$ for each permutation. If $b$ permutations have an $R^2$ greater than $r$, add $n_{\text{step}}$ to
   $n$ and go to Step 3.
4.2. METHODS

7. If not, stop here. The worst \( n - n_{\text{step}} \) variables are considered non-important variables. The remaining variables are important.

The parameters \( n_{\text{step}}, b, \) and \( r \) will all be determined by the best performance of the permutation test on a set of test data sets. \( n_{\text{start}} \) and \( N \) will be set based upon the characteristics of the data set being tested. The \( R^2 \) calculation above is a goodness of fit measure. It is discussed further in Section 2.2.4.1 for a true fitting situation. In this case, it is used to determine whether a permutation line comprised of \( n \) points (1 point per variable included) closely fits the correctly labeled line comprised of the same \( n \) points. It is performed as follows:

\[
R^2 = \frac{SSR}{SSY} = \frac{\sum_{i=1}^{n}(\hat{p}_i - \bar{p})^2}{\sum_{i=1}^{n}(p_i - \bar{p})^2} \quad (4.1)
\]

where \( \bar{p} \) is the mean of loadings across all variables, \( \hat{p}_i \) is the permuted line’s loading for a variable \( i \), and \( p_i \) is the loading value for correctly labeled line’s variable \( i \).

The overall goal of the permutation is to determine the point at which the shape of the correctly labeled loadings line no longer is similar to the permuted label loadings line, and to consider all of the variables, with lesser loadings, tested before that point not important. An example of a permutation test on a set of variables where no variables are responding can be found in Figure 4.1. This figure is a plot of the permuted label lines (in blue) and the correctly labeled line (in red). The Y-axis is the absolute value of the loading of that variable. The X-axis is the ranking of that variable’s loading when compared to the other variables used in that test. The larger, more important variables will thus have a low X value. Notice that the correctly labeled line falls within the cluster of permuted label lines. An example of a permutation test on a set where several
responding variables are present can be found in Figure 4.2. Notice how the shape of the correctly labeled line is different than the permuted label lines. The most important difference between the shapes of the lines is that the larger loadings in the first few ranked variables are significantly larger than the remaining variables. This is found in the correctly labeled line, but not in the permuted label lines.

The t-test, because it is a bin-at-a-time test, does not lend itself to a permutation in the same manner. It does not matter how many variables are included in a test, because each variable will be tested separately. Therefore, a simple p-value threshold will be used for comparison. Variables with p-values greater than the threshold are not considered important.
4.2. METHODS

Figure 4.2: Permutation test plot of data with 10 responding variables
4.2. METHODS

For this comparison the O-PLS-DA permutation and t-test methods described above are trained on the binned synthetic data sets described in Section 4.2.1. More specifically, for each of 3 data sets, 20 synthetic control samples are tested against 20 synthetic treatment samples with 50 percent of peaks not responding. This is repeated for each data set with treatment samples with 75 and 90 percent of peaks not responding. The methods are trained to determine which parameter combinations determined the number of important variables most closely to the actual number of important variables. The O-PLS-DA permutation test parameters that are trained are $n_{\text{step}}$, $b$, and $r$. The t-test parameter trained is the p-value threshold. Using these parameters are then tested on 15 binned data sets, where 20 synthetic control samples are tested against 20 synthetic treatment samples with 50 percent of peaks not responding. This is repeated for each data set with treatment samples with 75 and 90 percent of peaks not responding. The training and testing of the two methods is also performed on the peak data sets described in Section 4.2.1.

4.2.4 Comparison as a Classifier

One of the intended uses of an O-PLS-DA model is to classify new, yet-unseen data. An O-PLS-DA model produces regression coefficients, one for each variable. To predict a class from these coefficients, a simple multiplication is performed:

$$\hat{y} = x \ast b_{\text{pls}}$$  \hspace{1cm} (4.2)

where $\hat{y}$ is the predicted class of the $x$ data, a new sample. One stipulation is required, however. Any new $x$ data must be corrected (scaled, orthogonal components removed, and/or other tech-
4.2. METHODS

Techniques) before prediction. The resulting $\hat{y}$ will also be corrected, thus a decision boundary of 0 is used for a two-class problem. A result closer to -1 is predicted to be in class 0 and a result closer to 1 is predicted to be in class 1.

O-PLS-DA is performed as described in 4.2.2. After an O-PLS-DA model is established, any unseen $x$ data for prediction is required to go through the same corrections as the original $X$ data from the samples that the model is created with. The $x$ data is first scaled by subtracting the original $X$ mean and dividing by the original $X$ standard deviation. As detailed by Trygg and Wold (2002), the original $X$ data has some amount of estimated orthogonal variation removed. Any new $x$ data must also have this variation removed. Therefore, the new $x$ data is next corrected by the following step:

$$x_c = x_o - \left( \frac{x_o \ast w_{osc}}{w_{osc}' \ast w_{osc}} \right) \ast p_{osc}$$ (4.3)

where $x_c$ is the corrected $x$ data, $x_o$ is the original new sample’s $x$ data before correction, and $w_{osc}$ and $p_{osc}$ are the weights and loadings respectively from the orthogonal signal correction step of O-PLS. Should more than one orthogonal component be removed, this correction would be applied using each component’s $w_{osc}$ and $p_{osc}$. The first component’s $x_c$ would then be used as the second component’s $x_o$.

Performing a t-test on a series of variables does not inherently provide a method for classifying new samples. However, a common technique is to create a majority vote-based classifier based on the t-test’s p-value results (Qiu et al., 2008). The first step of a vote-based t-test classifier is to perform a t-test between the two classes for every variable. The mean of each class for each variable is then computed and the midpoint between the two classes means becomes a decision
4.3. RESULTS

boundary for future samples. Rather than using every variable, a subset of the variables is given voting privileges. For each variable voting, the value of that variable for the new $X$ is compared to the decision boundary. Which side of the boundary determines which class the vote is cast for. Whichever class has the majority after all voting variables have done so is the predicted class for the new $X$. As a note, this majority voting classifier will only work when there are two classes.

For this comparison the O-PLS-DA and t-test majority voting classifying methods described above are performed on the synthetic data sets described in Section 4.2.1. More specifically, for each of 15 data sets, 20 synthetic control samples (split into 10 training samples and 10 testing samples) are tested against 20 synthetic treatment samples (split into 10 training samples and 10 testing samples) with 50 percent of peaks not responding. This is repeated for each data set with treatment samples with 75 and 90 percent of peaks not responding. Each combination of control samples and treatment samples was also repeated with 10 samples each (split into 5 training samples and 5 testing samples). The O-PLS-DA and t-test majority voting classifying methods are performed on both binned and peak data sets described in Section 4.2.1.

4.3 Results

4.3.1 Comparison as a Ranking Method

Ranking results from O-PLS-DA loadings, t-test p-values, and the absolute difference between treatment base and control base spectra are analyzed via correlation coefficients. A common method of assessing correlation, the Pearson product-moment correlation coefficient, may not applicable here as it requires the assumption that the variables (loadings, p-values, and differences)
4.3. RESULTS

<table>
<thead>
<tr>
<th>Percent of peaks</th>
<th>Samples</th>
<th>OPLSDA and true ranking correlation</th>
<th>t-test and true ranking correlation</th>
<th>OPLSDA and t-test correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>not changing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>0.440</td>
<td>0.441</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.571</td>
<td>0.572</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.668</td>
<td>0.669</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>avg</td>
<td>0.560</td>
<td>0.561</td>
<td>0.996</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0.278</td>
<td>0.279</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.399</td>
<td>0.400</td>
<td>0.991</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.513</td>
<td>0.519</td>
<td>0.985</td>
</tr>
<tr>
<td></td>
<td>avg</td>
<td>0.397</td>
<td>0.399</td>
<td>0.991</td>
</tr>
<tr>
<td>75</td>
<td>5</td>
<td>0.181</td>
<td>0.180</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.247</td>
<td>0.251</td>
<td>0.981</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.358</td>
<td>0.367</td>
<td>0.963</td>
</tr>
<tr>
<td></td>
<td>avg</td>
<td>0.262</td>
<td>0.266</td>
<td>0.979</td>
</tr>
<tr>
<td>90</td>
<td>5</td>
<td>0.071</td>
<td>0.070</td>
<td>0.991</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.126</td>
<td>0.124</td>
<td>0.963</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.173</td>
<td>0.178</td>
<td>0.925</td>
</tr>
<tr>
<td></td>
<td>avg</td>
<td>0.123</td>
<td>0.124</td>
<td>0.960</td>
</tr>
</tbody>
</table>

Table 4.1: Results from ranking comparison of O-PLS-DA loadings and t-test p-values on binned synthetic data.

being compared are jointly normally distributed. A non-parametric method of assessing correlation, the Spearman rank correlation coefficient makes no such assumptions (Spearman, 1904). The correlation is measured on the rankings of the values, instead of on the values themselves. The O-PLS-DA loadings are ranked by importance, from largest loading to smallest. The t-test p-values are ranked by importance as well, from smallest p-value to largest. The base spectra differences are ranked from largest difference to smallest. These rankings are performed on both binned synthetic data and peak data. Average correlation coefficients across the 30 binned data sets can be found in Table 4.1. Average correlation coefficients across the 30 peak data sets can be found in Table 4.2. The results of these tables are examined in Section 4.4.1.
4.3. RESULTS

<table>
<thead>
<tr>
<th>Percent of peaks not changing</th>
<th>Samples per group</th>
<th>OPLSDA and true ranking correlation</th>
<th>t-test and true ranking correlation</th>
<th>OPLSDA and t-test correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>0.363</td>
<td>0.363</td>
<td>0.999998</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.422</td>
<td>0.422</td>
<td>0.999997</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.463</td>
<td>0.463</td>
<td>0.999997</td>
</tr>
<tr>
<td>avg</td>
<td></td>
<td>0.417</td>
<td>0.417</td>
<td>0.999997</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.517</td>
<td>0.517</td>
<td>0.999973</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.619</td>
<td>0.619</td>
<td>0.999950</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.695</td>
<td>0.695</td>
<td>0.999951</td>
</tr>
<tr>
<td>avg</td>
<td></td>
<td>0.613</td>
<td>0.613</td>
<td>0.999958</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.431</td>
<td>0.431</td>
<td>0.999958</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.517</td>
<td>0.517</td>
<td>0.999906</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.581</td>
<td>0.581</td>
<td>0.999778</td>
</tr>
<tr>
<td>avg</td>
<td></td>
<td>0.512</td>
<td>0.512</td>
<td>0.999879</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.286</td>
<td>0.286</td>
<td>0.999909</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.352</td>
<td>0.352</td>
<td>0.999857</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.401</td>
<td>0.401</td>
<td>0.999640</td>
</tr>
<tr>
<td>avg</td>
<td></td>
<td>0.348</td>
<td>0.348</td>
<td>0.999799</td>
</tr>
</tbody>
</table>

Table 4.2: Results from ranking comparison of O-PLS-DA loadings and t-test p-values on peak synthetic data.

4.3.2 Comparison as a Method of Determining Important Variables

The O-PLS-DA permutation test is performed as described in Section 4.2.3 and the parameters \( n_{step} \), \( b \), and \( r \) are selected based on the training data sets. The parameter \( n_{step} \) is trained from 2 different step sizes, 10 and 25. The parameter \( b \) is trained from 3 different numbers of close permutations, 3, 5, and 7. The parameter \( r \) is trained from 4 different \( R^2 \) values, 0.90, 0.92, 0.94, and 0.96. The combination of these parameters that minimized the number of errors is chosen to be used on the testing data sets. An error occurred when a bin or peak is labeled responding or non-responding differently by the permutation test and the true results. Any ties in the number of errors are resolved by which combination minimized the number of extreme results, such as where all of the bin/peaks or none of the bin/peaks are determined to be responding. The best performing
4.3. RESULTS

<table>
<thead>
<tr>
<th>percent of peaks not changing</th>
<th>avg errors</th>
<th>t-test standard deviation</th>
<th>percent of total bins</th>
<th>avg errors</th>
<th>OPLSDA standard deviation</th>
<th>percent of total bins</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>117.4</td>
<td>15.1</td>
<td>34.5%</td>
<td>188.5</td>
<td>110.9</td>
<td>55.4%</td>
</tr>
<tr>
<td>75</td>
<td>131.7</td>
<td>10.4</td>
<td>38.7%</td>
<td>180.4</td>
<td>131.9</td>
<td>53.1%</td>
</tr>
<tr>
<td>90</td>
<td>134.1</td>
<td>10.8</td>
<td>39.4%</td>
<td>141.0</td>
<td>81.3</td>
<td>41.5%</td>
</tr>
</tbody>
</table>

Table 4.3: Results from O-PLS-DA permutation test and t-test p-value cutoff test on binned synthetic data.

combination on the peak training data sets is determined to be a $n_{step}$ of 10, a $b$ of 7, and an $R^2$ of 0.94. The parameters $n_{start}$ and $N$ are chosen to be 25 bins and 100 permutations respectively. The average number of errors from the permutation test using these parameters on the 15 binned test data sets can be found in Table 4.3. The best performing combination on the peak training data sets is determined to be a $n_{step}$ of 10, a $b$ of 5, and a $R^2$ of 0.94. The parameters $n_{start}$ and $N$ are chosen to be 100 peaks and 100 permutations respectively. The average number of errors from the permutation test using these parameters on the 45 peak test data sets can be found in Table 4.4. The results of these tables are examined in Section 4.4.2.

The t-test p-value threshold method is performed as described in Section 4.2.3. A range of potential p-value thresholds (between 0.0025 and 0.25) are used on the training data sets and the threshold which minimized the number of errors is chosen to be used on the testing data sets. Any ties in the number of errors are resolved by averaging the tied thresholds. The best performing p-value thresholds are 0.23625 and .02 for the binned and peak training data sets respectively. The average number of errors from the t-test method using these p-value thresholds on the 45 binned test data sets can be found in Table 4.3. The average number of errors from the t-test method using these p-value thresholds on the 45 peak test data sets can be found in Table 4.4. The results of these tables are examined in Section 4.4.2.
4.3. RESULTS

| percent of | avg | t-test | percent of | avg | OPLSDA |
| changing | errors | standard | total peaks | errors | standard | total peaks |
| 50 | 932.2 | 9.9 | 45.1% | 878.0 | 752.3 | 42.5% |
| 75 | 477.2 | 7.5 | 23.1% | 691.5 | 555.0 | 33.4% |
| 90 | 205.2 | 5.8 | 9.9% | 778.9 | 528.0 | 37.7% |

Table 4.4: Results from O-PLS-DA permutation test and t-test p-value cutoff test on peak synthetic data.

4.3.3 Comparison as a Classifier

O-PLS-DA was performed on the training samples (5 treatment and 5 control samples, then 10 treatment and 10 control samples). The resulting models produced a $b_{pls}$ which was used to classify the testing samples (5 treatment and 5 control, then 10 treatment and 10 control) as described in Section 4.2.4. Testing samples are corrected first by scaling, then by removing orthogonal variation via $w_{osc}$ and $p_{osc}$. The decision boundary for the models is 0. A sample is predicted to be a control sample if the resulting $\hat{y}$ is less than 0. Conversely a $\hat{y}$ greater than 0 is predicted to be a treatment sample.

The majority vote-based classifier based on the t-test is performed on the same training samples as O-PLS-DA (both 5 and 10 samples each treatment and control). Means of each class for each variable are calculated and depiction boundaries are established. For a threshold of which variables receive a vote and which do not, the p-value threshold found to perform best at identifying the most important variables in Section 4.3.2 is used. For binned synthetic data, that threshold is 0.23625. For peak synthetic data, that threshold is 0.020. Any variable with a p-value less than or equal to that threshold is given a vote.

The number of control and treatment binned data samples misclassified with percentages for
### 4.4 Discussion

#### 4.4.1 Comparison as a Ranking Method

Table 4.1 illustrates that the coefficients measuring the correlation between the rankings of the O-PLS-DA loadings and t-test p-values on the binned synthetic data are high and consistent across different numbers of samples and different numbers of responding bins. This suggests that the loadings and p-values produce a similar ranking of the potential biomarkers. There is a slight
### 4.4. DISCUSSION

<table>
<thead>
<tr>
<th></th>
<th>5 train/test samples</th>
<th>10 train/test samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>number</td>
<td>percent of total</td>
</tr>
<tr>
<td>OPLSDA control</td>
<td>1</td>
<td>0.33%</td>
</tr>
<tr>
<td>samples missed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPLSDA treatment</td>
<td>1</td>
<td>0.33%</td>
</tr>
<tr>
<td>samples missed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPLSDA total</td>
<td>2</td>
<td>0.33%</td>
</tr>
<tr>
<td>samples missed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-test control</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>samples missed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-test treatment</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>samples missed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-test total</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>samples missed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.6: Results from classification comparison of the O-PLS-DA and majority vote-based classifiers on peak synthetic data.

A decrease in the correlation coefficients as the number of samples increases.

The coefficients measuring the correlation between the rankings of the O-PLS-DA loadings and the true rankings on the binned synthetic data have an inverse relationship with the percentage of peaks not responding. Thus, the more bins not containing differences, the less the correlation there is between the loadings and the true rankings. Two potential explanations for this exist. The random variation could cause O-PLS-DA to notice the separation between classes in one non-responding bin as more significant than the separation in another non-responding bin. Another explanation is that for the bins where differences between the treatment and control base spectra bins are 0 (non-responding bins), the Spearman’s rank correlation coefficient method states that an average of the rankings be used as the rank for each bin. If bins 1, 2, 3, and 4 all have the same values and are the next to be ranked, then they will each be given a ranking equal to the average of the next 4 rankings. If the next 4 rankings are 24, 25, 26, and 27, then bins 1 receives the ranking 25.5 or the average of those 4 next rankings. This will cause a large number of bins to
4.4. DISCUSSION

be ranked the same in the true rankings. It is unlikely for O-PLS-DA to assign precisely the same loading to several bins given the variation and noise introduced to the sample spectra. This would contribute to differences between the loadings rankings and the true rankings and an increase in the number of non-responding bins would lead to an increase in the number of differences in the rankings from this cause. The coefficients measuring the correlation between the rankings of the t-test’s p-values and the true rankings on the binned synthetic data also have an inverse relationship with the percentage of peaks not responding.

The correlation between the rankings of the loadings and the true rankings on the binned synthetic data increase as the number of samples increases. This is also true of the correlation between the p-values and the true rankings. Thus, the more samples included, the less influence each sample has on the model and the better the model represents the true population that these samples are drawn from.

The results from the peak data ranking comparison, found in Table 4.2, show many of the same trends as the binned data. An increase in the number of samples increased the correlation between the O-PLS-DA loadings and the true ranking. This was also true for the correlation between the t-test p-values and the true ranking. A slight decrease in the correlation between the loadings and the p-values occurs as the number of samples increases. As the percent of peaks not responding increases, the correlations decrease. The only exception between the trends in the binned and peak data sets was that the correlations between the loadings (and the p-values) and true ranking for the 0 percent of peaks not changing data set are lower than those of the other data sets. It is possible that the 0 percent peaks not responding data, where a difference exists between treatment and control samples at every peak, has a substantial number of peaks that fall below the limit of
4.4. DISCUSSION

detection. At some point the differences between the classes will be below the average variation between spectra. Neither the t-test nor O-PLS-DA will be able to detect any differences past this limit.

The loss of sensitivity when reducing the number of variables from approximately 2000 peaks to 340 bins is evident in the differences between bin and peak correlations. The binned data correlations for the between the loadings and true rankings, p-values and true rankings, and loadings and p-values are less in nearly all data sets. The peak data correlations are lower for only the data sets where all of the peaks are responding.

4.4.2 Comparison as a Method of Determining Important Variables

Table 4.3 reveals differences between the t-tests and permutation tests’ ability to identify the bins containing differences between treatment and control samples. While the difference between the number of mismarked bins is not great, it is significant for the 50 and 75 percent of peaks not responding data sets (p < 0.001 for each). It is not significant for the 90 percent of peaks not responding data sets (p = 0.131). One of the major reasons for the significance of the differences is due to the vastly different standard deviations of the errors. The standard deviations of the t-test errors are roughly 10 times smaller than that of the permutation test errors. This likely signifies an instability of the permutation test. While the average error seems reasonable, this average is the product of a large number of results where the permutation test determined that nearly all of the bins are significant or that nearly all of the bins are not significant.

The peak results found in Table 4.4 showed similar results. The number of differences are again not large, but are significant for the 50, 75, and 90 percent of peaks not changing data sets (p
4.4. DISCUSSION

< 0.001 for each). Also, the standard deviations of the errors remained higher for the permutation test. The number of errors for the t-test is lowered with fewer responding peaks, as did the standard deviations of the error. Unlike the binned data, the number of responding variables diminishes at a larger rate from 50 percent of peaks not responding to 90. This was not observed in the permutation test and is again likely due to the instability of the permutation test.

There are a large number of errors for both tests on the binned data sets. Binned data has an obvious loss of sensitivity, but additionally many small differences will be masked both by spectra to spectra variability and in the combining of smaller and larger peaks within a bin. A portion of the bins considered responding due to responding peaks lying within the bin are not going to be visible to O-PLS-DA, the t-test, or any other method. The goal is not to achieve 100% accuracy, rather, it is for a comparison of these two methods. The accuracy was improved in the peak test data sets for the t-test method, but the permutation test results are unable to provide any further information for the comparison between the binned and peak data sets.

4.4.3 Comparison as a Classifier

Table 4.5 reveals differences in classification results with respect to classification method and training and testing sample size. O-PLS-DA classifies new samples with lower error percentage when trained with more samples on the binned synthetic data sets. As the number of samples used to create the model increases, the important variables that best distinguish between the classes become more clear. This is expected to be the case for nearly any classifier. The t-test’s majority vote-based classifier’s accuracy, however, does not increase with sample size. It remains at a consistent level. There is no intuitive reasoning for this result and thus it would not be expected to be the case for all
4.5. CONCLUSION

majority vote-based classifier based upon t-tests. The O-PLS-DA classifier outperforms the t-test classifier on the 10 training, 10 testing sample data set \( (p < 0.0001) \). There is not a significant difference between the classifiers in the 5 training, 5 testing sample data set \( (p = 0.2199) \).

The peak synthetic data results in better classification percentages, shown in Table 4.6. There are no errors for the t-test classifier, and a minimal number for the O-PLS-DA classifier \( (< 1\%) \). The differences between the classifiers is not significant for either the 5 training, 5 testing sample data set \( (p = 0.1590) \) or 10 training, 10 testing sample data set \( (p = 0.1238) \).

The differences in performance on binned data and peak data are again evident. The O-PLS-DA and t-test error in classification rates increases from a 0-1% range in the peak data to a 9-19% range in the binning data. Important information necessary to accurately classify new samples is lost with the data reduction of binning.

4.5 Conclusion

Synthetic spectra created through the characterization of real-world NMR spectra proved to be an effective means of comparing O-PLS-DA and the t-test in the areas of ranking variable importance, distinguishing the important variables from non-important variables, and classification of unseen samples. Knowledge of what bins/peaks are different and the magnitude of these differences is the only proper way to determine whether a particular method is performing adequately or to compare the relative performance of methods.

In both the ranking and classifier comparison results, there were no distinct advantages or disadvantages to using either the t-test or O-PLS-DA. The ranking comparison showed a correlation
4.5. **CONCLUSION**

between the loadings and p-values of nearly 1. A t-test comparing classes one variable at a time is as good of a result as possible when variation and noise are present. O-PLS-DA achieved nearly an identical ranking. This suggests that if the ultimate goal is to rank the bins/peaks by importance, O-PLS-DA will achieve results as good as the t-test. Classification results were also similar between the two methods. O-PLS-DA classifier error rates were slightly better than the majority vote-based classifier based on the t-test results, though not significantly. For classifying new NMR spectral samples or ranking potential biomarkers either method proved effective. The only major difference between the methods is that the t-test is a variable-at-a-time method, while O-PLS-DA takes all of the variables into consideration at the same time. Thus, the t-test may be the better choice for ranking and classification, as it avoids the "curse of dimensionality." Either method, O-PLS-DA or the t-test helps to determine both whether or not the rat has been exposed to the toxin (classification comparison) and how the rat was affected by exposure (ranking comparison).

Should the permutation test have been able to approach the results of the t-test, perhaps even exceed the results, the permutation could have been considered a viable method for determining the number of responding variables in a spectral data set. In its current incarnation, however, the permutation test is inconsistent and unstable. While good results on several individual data sets were observed, the permutation test too often determined that either no responding variables were present or nearly all variables were responding. This is evident in the standard deviation results for both binned and peak test data sets. A p-value threshold produced statistically better results. This is still a less than ideal solution, however, as training data sets were necessary to determine the optimum threshold. In real-world NMR data sets it is unlikely that the necessary knowledge, namely the correct number of responding bins/peaks, will be known. Therefore neither the t-test
4.5. CONCLUSION

nor the permutation test can currently be recommended as a method for determining the correct number of potential biomarkers without this prior knowledge. The ability to determine the number of responding variables, or metabolites, would have helped to determine exactly how the rat was being affected by exposure to the toxin. With this information we may have been able to determine exactly which metabolites are changing by measurable amounts, what metabolic pathways those changes are attributable to, and therefore which bodily functions are altered by exposure.

A common theme throughout these three comparisons is that the often-used preprocessing step of binning data reduces the information contained in the spectral data. Both ranking and classification results were improved significantly in the peak data as compared to the binned data, save for the correlation between the rankings produced by the two methods. The classification results showed a marked difference across the various data sets. A 0% error rate on peak data rose to nearly 20% by reducing the number of variables through uniform binning. A smarter method of binning, such as variable width binning, or some method of peak alignment or matching could lessen the amount of reduction in sensitivity and thus would be a better alternative to uniform binning. The peak data, which was meant as a perfect-world situation where every peak can be completely distinguished from others and every peak can be matched across both control and treatment spectra is an unrealistic model. However, the closer a data set’s preprocessing step is to this model, the more likely it is to maintain as many true differences between samples as possible. This would allow for the most complete potential biomarker identification possible.
4.6. COMPUTATIONAL HARDWARE AND SOFTWARE

4.6 Computational Hardware and Software

All analyses in this chapter performed using the following computers:

1. Laptop running Microsoft(R) Windows Vista(R) Home Premium 64-bit Service Pack 1, Intel(R) Core2(R) Duo P8400 (2.26 GHz), 4 GB RAM, 320 GB HDD

2. Desktop running Microsoft(R) Windows XP(R) Professional 32-bit Service Pack 3, Intel(R) Pentium(R) 4 (3.73 GHz), 2 GB RAM, 60 GB HDD

3. Desktop running Ubuntu 7.04, Intel(R) Pentium(R)4 (3.73 GHz), 2 GB RAM, 60 GB HDD

4. Desktop running Microsoft(R) Windows Vista(R) Ultimate 64-bit Service Pack 1, Intel(R) Core2(R) Duo E6400 (2.13 GHz), 4 GB RAM, 250 GB HDD

All analyses were performed using the following programs:

1. Matlab(R) R2007b for Windows

2. Microsoft(R) Excel(R) 2007

3. Ruby 1.8.6-26
Conclusion

5.1 Summary

With the ultimate goal of identifying potential biomarkers, both simple and more complex methods for analyzing spectral data from NMR-derived toxicological experiments were compared. Three simple methods, the two-fold test, the unpaired t-test, and the paired t-test with vehicle effects removal, were inspected using pre-determined subjective criteria in order to characterize their similarities and differences. The most comprehensive of the three, the paired t-test, displayed an ability to reward an consistency across time points and to identify both responses due to vehicle effect and those responses above and beyond vehicle effects. These criteria closely match the criteria desired for the isolation of potential biomarkers in a metabolomic toxicology data set.

Multiple test corrections are necessary to compensate for the multitude of tests performed when NMR spectra are analyzed one variable at a time. Three different multiple test corrections were performed, representing methods which control for different error rates and have different assumptions. The Bonferroni and the Holm’s corrections are conservative, and control for the familywise error rate. The Benjamini-Hochberg correction, on the other hand, controls for the false discovery rate and is less conservative. Therefore, if a single false positive is to be avoided, as in
5.1. SUMMARY

the case when a fool-proof potential biomarker is sought, then the more conservative corrections would be more appropriate. If capturing a wide number of potential biomarkers is the focus, than the less conservative Benjamini-Hochberg correction would be the proper choice. A PCA goodness measure was used to determine a suggested p-value threshold for each multiple test correction.

The more complex O-PLS-DA was compared to the variable-at-a-time t-test in 3 categories using synthetically created data: the ability to rank variables by importance, the ability to determine which of the variables are important and which are not, and the ability to correctly classify new data as coming from treatment or control samples. O-PLS-DA did not differ significantly from the t-test in the ranking and classifying tests. An O-PLS-DA permutation test was not stable enough to consistently distinguish between the correct number of responding and non-responding variables. Additionally, how the loss of sensitivity incurred by binning the synthetic spectra affects the performance of the t-test and O-PLS-DA was illustrated.

O-PLS-DA is a sufficient method for determining the relative importance of potential biomarkers for distinguishing between treatment and control samples, performing as well as the t-test. The O-PLS-DA and t-test are also sufficient classifiers of new spectral samples given a group of known samples. The t-test, which is not susceptible to the ”curse of dimensionality,” may be the safer choice for ranking and classification. Neither method could be recommended for determining the correct number of responding variables.
5.2 Contributions

Several contributions are submitted as part of this thesis. Three variable-at-a-time methods are compared using experimental toxicological data. These methods identify changes from treatment animals to control animals one variable, or metabolite, at a time, attempting to answer the question of what metabolites are being affected by exposure to a toxin. In addition, three multiple test corrections are compared for their relative conservativeness and p-value thresholds are suggested for this particular experimental data.

One variable-at-a-time method, the t-test, is compared to O-PLS-DA, which considers all the variables at once in attempting to identify differences between treatment and control animals. These methods are compared using synthetic data in which the answers are known, in three capacities: the ability to rank important variables, the ability to classify new samples, and the ability to determine the number of responding variables. A permutation test based on O-PLS-DA is proposed for determining the number of responding variables, or metabolites. The ranking comparison and the comparison of the ability to determine the number of responding variables attempt to identify which method is more effective at determining how the animal is being affected by the toxin. The classification comparison attempts to identify which method is more effective at determining whether or not an animal has been exposed.

5.3 Future Work

While this work discusses and compares several methods for identifying biomarkers from NMR-derived toxicological data, there are several areas in which further investigation beyond this thesis
5.3. FUTURE WORK

could be continued.

5.3.1 Further Exploration of O-PLS-DA-based Permutation Test

As stated in Section 4.5, the current incarnation of the permutation test was not consistent or stable enough to determine which bins or peaks were responding and which were not. The permutation test itself may not be the problem, rather just a part of it. It may be that with a different combination of parameters, chosen from a wider set of training parameters, or with a method different from the $R^2$ for determining a difference in shape between the permuted and non-permuted lines, that the permutation test could become a viable method determining responding variables.

5.3.2 Expansion of O-PLS-DA

Given the time-course and multiple dose NMR data described in Section 3.1, O-PLS-DA may be able to be modified so that instead of the $Y$ being simply 1 for treatment and 0 for control, $Y$ is either the dose or time-point. Thus it may be possible to predict a sample’s time, dose, or perhaps both if a particular toxin has been completely characterized for multiple doses and multiple time points.

5.3.3 Expanded Real-life Data Sets

The types of evaluation performed in this thesis could also be expanded to data sets featuring different toxins and/or different animals, so long as the types of controls (internal and/or external) are consistent with those found in this data set.
5.3. FUTURE WORK

5.3.4 Additional Biomarker Identification Methods

Given the ability to create synthetic data sets, a series of comprehensive comparisons of potential biomarker identification methods would be informative. The strengths and weaknesses of additional methods’ performance in ranking of potential biomarkers, testing cutoffs, and classification for various numbers of

There are several other methods used in the microarray community that may have different strengths when compared to the methods addressed in this thesis. Baldi and Long have implemented a Bayesian probabilistic framework for microarray analysis (Baldi and Long, 2001) that may also be appropriate for NMR analysis. Several other methods were mentioned in the literature survey including penalized logistic regression (Zhu and Hastie, 2004), SAM (Marchal et al., 2002), ANOVA (Cui and Churchill, 2003), noise sampling model (Draghici, 2002), and an ANOVA-based bootstrap method (Marchal et al., 2002).

5.3.5 Additional Preprocessing Methods

This thesis illustrated some of the problems associated with the uniform binning of spectra, most notably how the reduction in sensitivity affects the performance of ranking, classifying, and biomarker determination methods. There are several more flexible methods of data reduction. They include the gaussian (Anderson et al., 2008) and adaptive (Davis et al., 2007) binning methods and genetic algorithm (Forshed et al., 2003), PCA (Stoyanova et al., 2004), and fuzzy warping (Wu et al., 2006) based methods for peak alignment, among others. A comparison between these methods and the perfect spectra deconstruction and peak matching method using synthetic data (described in Sec-
5.3. FUTURE WORK

Evaluation 4.2.1) would give a good indication of their relative strengths and weaknesses. The method which most closely replicates the results of the ideal situation would be the suggested method of preprocessing for similar, real-world NMR data.
Bibliography


BIBLIOGRAPHY


112
BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


