University of Alberta

Spectral Processing Considerations for the Analysis of NMR Based Metabolomics Data

by

David Wai Ming Chang

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Process Control

Chemical and Materials Engineering

Fall, 2009

Edmonton, Alberta

Abstract

Employing a combination of biochemistry and chemometrics, the field of metabolomics has the potential to reveal some very significant insights into biological pathways related to drugs and diseases. This thesis explores this field in its depths; specifically focusing on Nuclear Magnetic Resonance (NMR) based methods. The thesis begins with an exploration of the quantum level relationships of molecules, and how these coupling patterns evolve into an NMR spectrum. The thesis will describe the development of a simplified spin simulation algorithm to predict NMR spin coupling patterns that are computed in fractions of a second and to build mathematically relevant basis functions. Later in the thesis, the issue of baseline distortions of real NMR experimental data is addressed by the development of an automated baseline correction algorithm. Data reduction techniques are further analyzed to understand the importance of the quality of the data used in advanced chemometric methods. For analysis of the data, the use of simple univariate techniques applied to NMR spectra of urine is explored to determine statistically significant biomarkers between disease states in asthma. More advanced statistics in the way of multivariate models, namely Partial Least Squares – Discriminant Analysis (PLS-DA), were used to build predictive models of *Streptococcus pneumoniae* pneumonia from NMR spectra of urine. Potential characteristics of the data that may invalidate assumptions required in our models were accounted for, such as ensuring the statistical normality of the *S. pneumoniae* pneumonia data by using log transformations. After the analysis, focus was given to the use of unique visualization techniques to further explore the complex relationships that exist between samples and variables, and relationships between variables. As will be made evident, this thesis deals with the basic physics of an NMR signal to building highly sophisticated models to help understand the NMR spectra from complex mixtures. All of these notions are important in the objective to garner the most information provided through an NMR experiment, as such to aid in the discovery of biochemical knowledge.

Acknowledgments

Many people need to be acknowledged for their insights and contributions to this work. First, the author would like to thank Ryan McKay and Brian Sykes for their discussions with on NMR theory. The author would like to acknowledge Kathryn Rankin for her discussions on NMR spin simulation. Aalim Weljie and Jack Newton provided many relevant discussions on Chenomx software and their contributions in chapter 4 on building statistical models were invaluable. Pascal Mercier and Cory Banack contributed significantly in their insights on NMR baseline correction. The author would like to thank Darryl Adamko and Eric Saude for sharing the asthma data for work discussed in chapter 5. Paige Lacy and Andriy Cheypesh performed all of the work on the mice model and provided all of the data in chapter 7. Also, Paige Lacy provided much needed cellular biology expertise in regards to the extracellular reaction for the asthma data in chapter 5. The author would like to thank Shana Regush and Bruce Lix for providing some of the NMR data on the normal and pneumonia patients in chapter 6. Carolyn Slupsky was a key part of this work in providing various discussions on the biochemical relevance of the metabolite data in chapters 5, 6, and 7. Carolyn also provided many details and wrote part of the discussion in chapter 6. Lastly the author would like to thank Sirish Shah for his guidance and patience as he supervised this new venture into metabolomics.

Table of Contents

List of Tables

List of Figures

1. Introduction

In recent years, with the advancement of computing technologies and biological sequencing technologies, there has been an ever growing interest in the area of bioinformatics. With the completion of the human genome project in 2003 (Human Genome Program, 2003), scientists have been trying to make sense of this enormous volume of genomic data. Mapping the human genome is but a small part of understanding human biology. A systems biology approach to this research is to look at the human genome, transcriptome, proteome, and metabolome in conjunction. The focus of this thesis is on the human metabolome, and the bioinformatics of the metabolome. The metabolome is described as the complete complement of all small molecules (<1500 Da) metabolites found in a specific cell, organ or organism (Wishart et al., 2007).

Metabolites are small molecules in the body, many of which are used in normal cellular functions. Detection and quantification of metabolites can give useful information as to the condition of the cells in the body. Metabolites can be measured from many biofluids (eg. cell cytoplasm, blood plasma, sputum, urine, etc.). The choice of sample type usually depends on factors such as specificity and availability. Measuring the metabolites within the cell cytoplasm should give results that are very specific to the condition of the particular cell samples. However, the extraction of cellular cytoplasm would normally require invasive methods and as such requires much effort, so availability of the samples is limited. Our research has mainly focused on metabolites found in urine.

Urine is relatively easy to obtain. The metabolite concentrations found in urine can be an indication of the health of a person. However, measuring metabolites from urine will offer challenges in specificity, as many changes in metabolites levels could be a systematic effect of the disease itself. It may be difficult find the root cause of the effects purely on metabolite levels.

After obtaining urine samples, there are many techniques that can be used for the detection and quantification of these metabolites. Along with various separation processes of the sample, many of these techniques fall into two broad categories: Nuclear Magnetic Resonance (NMR), and Mass Spectrometry (MS). Both techniques have their unique advantages and disadvantages. Our work is in collaboration with the Canadian National High Field Nuclear Magnetic Resonance Center (NANUC). Nuclear Magnetic Resonance is used as the primary detection technique in this work, however many of the modeling techniques can be extended to Mass Spectrometry. As will be discussed, the majority of the work involves getting the data to a point where meaningful statistical analysis can be done.

1.1. Understanding the 1D ^{1}H NMR Experiment

Nuclear Magnetic Resonance (NMR) spectroscopy is a highly evolved field of science. We first provide a brief description of NMR theory as it pertains to the generating the NMR spectrum. For a more detailed description of NMR theory, please refer to textbooks such as Evans (1995) and Goldman (1988). NMR as an experimental technique has evolved over the years to yield many powerful applications today. NMR experiments involve placing a sample inside a large uniform magnetic field. Among the various atomic nuclei in this sample, there are many isotopes that have an intrinsic angular momentum and a magnetic moment. The relationship between the two is proportional and called the magnetogyric ratio. When a sample is placed in this magnetic field, the Larmor theorem states that the motion of the magnetic moment will be as a precession around this magnetic field. This precession frequency is dependent on the field strength and nucleus of interest (Goldman, 1988). All of this assumes an isolated magnetic moment. There are many nuclei in biological samples that can give a NMR spectrum such as ^{1}H, ^{13}C, and some that cannot such as ^{12}C and ^{16}O. Since this research is focused on metabolites from biological systems, focus will be on the ^{1}H nucleus due to its abundance in organic compounds.

1.1.1. Free Induction Decay (FID)

The Free Induction Decay (FID) signal is a time series signal detected from the NMR spectrometer. Let us consider now a sample placed in a uniform magnetic field (B_o). (See Figure 1.1) Under thermal equilibrium the ^1H nuclei, which have a nuclear spin of ½, will either align itself with or against the applied magnetic field. The Schrödinger equation predicts that there will be a slightly higher population of spins aligned with the applied magnetic field than against it. Hence, an overall magnetic moment is in the z direction (See Figure 1.1). When a radio frequency (RF) pulse is applied to this system, energy is added to the system such that there is a slight shift in the population of magnetic moments towards the direction against the applied magnetic field (See Figure 1.1). The process by which this system returns back to equilibrium is described as an exponential decay and has a time constant known as T_1 (Evans, 1995).

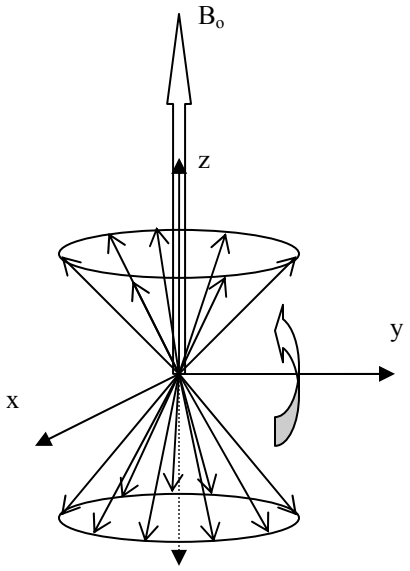

Figure 1.1 – Longitudinal (Spin-Lattice) relaxation process in rotating frame.

Another RF pulse can be designed such as to cause the magnetic moments to have phase coherence. To describe this process, imagine the frame of reference that is rotating with the overall precession frequency, to be called the rotating frame. In the rotating frame, under the assumption that there is a uniform magnetic field, all of the nuclei in the sample

will be precessing at the same frequency (ω_o). All of the nuclei however, will have their own phase, or a random starting point of rotation in space and time. The effect of this is a uniform or zero magnetic moment in the x-y plane (See Figure 1.2). When a RF pulse is applied to cause phase coherence, there is a shift in the magnetic moments such that they all have a similar phase. This shift, in the rotating frame, will have an effect that causes the magnetic moments to be on one side, say in the positive y-direction. The overall magnetic moment will then be non-zero in that direction (See Figure 1.2). Due to various relaxation pathways, that will not be discussed here, the magnetic moments will eventually return back to a uniform equilibrium. This relaxation is also described as a decaying exponential with a time constant known as T_2. Both T_1 and T_2, are relaxation mechanisms that are independent of each other (Evans, 1995).

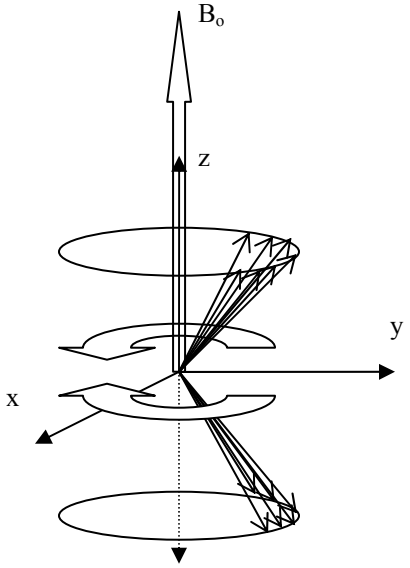

Figure 1.2 – Transverse (Spin-Spin) relaxation process in rotating frame.

In an NMR experiment, we are no longer in the rotating frame, but in a stationary frame called the laboratory frame. In the laboratory frame, the bulk magnetization moment in the z direction is unaffected by the two differing frame of references. However, in the laboratory frame, the magnetic moment that can be detected in the x-y plane will now be a sinusoid in the frequency of the precession frequency, with a decaying exponential

term. (See Equation 1.1) We finally come to the description of the Free Induction Decay (FID) signal (M) from a NMR experiment.

$$M = M_o \cdot \sin(\omega_o \cdot t) \cdot \exp(t / T_2) \quad \textbf{(1.1)}$$

Equation 1.1 gives a sinusoid with a decaying exponential term, observed in the laboratory frame. M_o is the amplitude of the signal directly proportional to the number of nuclei being observed by the probe.

1.1.2. Quadrature Detection

Quadrature Detection is a hardware and software technique used to measure FID signals. This technique offers some unique advantages over normally acquired (single phase) FIDs. One advantage is that this detector is phase sensitive. This means that not only can the frequency of spin be detected, but the direction of the spin can also be detected. From a hardware implementation, one can imagine having two detectors that are able to sense with a 90^o offset. The implication of working with quadrature detected signals is that a simple power spectrum is not appropriate for converting a FID signal into a spectrum in the frequency domain. A power spectrum is symmetric on both positive and negative axes. A quadrature detected signal will produce unique spectra on both sides of zero frequency. Thus, one could make use of this and store positive and negative relative frequencies from a central frequency, typically the frequency of protons on water molecules. This then allows the NMR equipment to store very high frequencies and work around the Nyquist sampling theorem. The Nyquist sampling theorem is discussed in detail in standard textbooks like Proakis and Manolakis (1996). Typical precession frequencies of protons placed in a high magnetic field are in the order of hundreds of megahertz, sampling at half that frequency is still a sampling rate in the order of nanoseconds. However, by using relative frequencies, one could use a smaller spectral bandwidth but shift the zero point to a central frequency. This will allow storage of high frequency data without a very fast sampling rate. Modern spectrometers use only one detector but split the signal into two channels, and perform a Hilbert transform on one of the channels. A detailed discussion on Fourier and Hilbert transforms is available in Proakis and Manolakis, 1996. The Hilbert transform in the frequency domain can be

simply described as follows: All negative frequencies of a signal get a +90^o phase shift and all positive frequencies get a -90^o phase shift (See Figure 1.3).

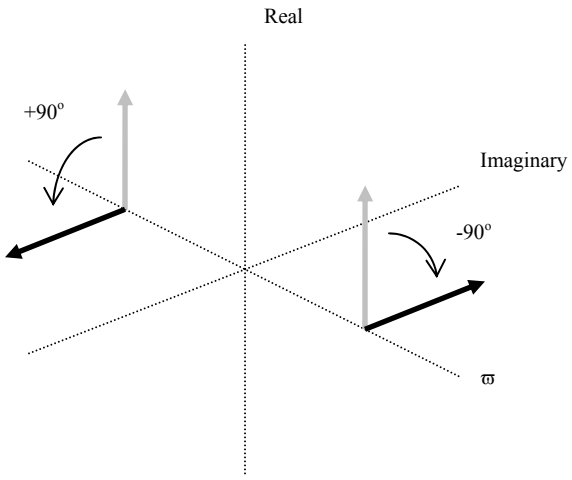

Figure 1.3 – Hilbert transform in the frequency domain.

In this implementation, the direction of spin cannot be known, since there is still only one detector. However, the use of relative frequencies still results in an efficient storage of data. The quadrature detected signal results in two time series signals, which can be combined into one complex FID. The central frequency is also chosen to be in the middle of frequencies of interest to maximize the observable frequencies. Note that there is still the chance for aliasing frequencies outside of the observable frequency bandwidth. Sampling frequency and acquisition time are still important parameters to consider, in order to avoid aliasing. The complex FID now needs to be converted to a useable spectrum in the frequency domain for spectroscopy. The algorithm for conversion of quadrature detected FID signals to a spectrum is described in detail in Figure 1.4.

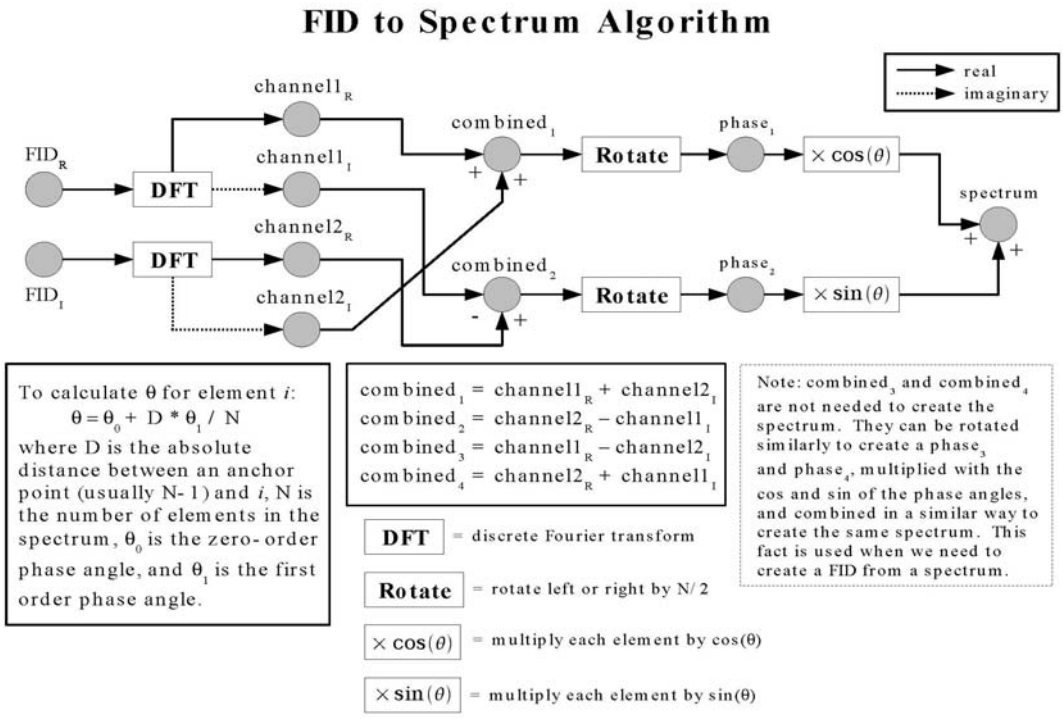

Figure 1.4 – Quadrature detected FID to spectrum conversion algorithm.

Note that the final projection of the spectrum onto the real axis requires a phase angle. This is a process known as phase correction or "phasing". Ideally the conversion of a quadrature detected signal should result in a spectrum with a pure absorption line shape in the real part, and a pure dispersion line shape in the imaginary part. As such the phase correction angle should be zero. However, practically this is never the case and a mixture of absorption and dispersion line shapes can occur. This phase angle (θ) is made up of a frequency independent (θ_0) and frequency dependent (θ_1) part (Hoch and Stern, 1996).

The second part of this process is to convert a spectrum back into a quadrature detected FID signal. Figure 1.5 describes this algorithm in detail. For this backward process, an unknown imaginary channel is estimated by a Hilbert transform. It is important to zero fill the time domain signal to twice the closest power of two in order to not incur any loss of information.

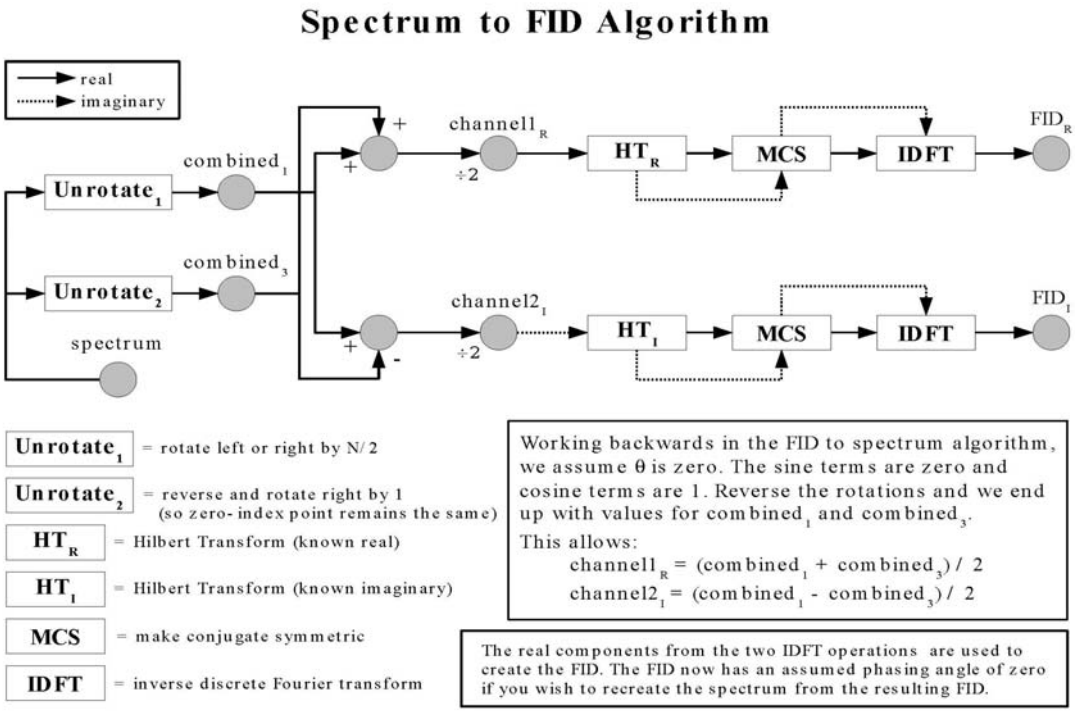

Figure 1.5 – Spectrum to quadrature detected FID conversion algorithm.

The above algorithms allow for the lossless conversion between time and frequency domain signals. Figure 1.6 shows the frequency domain FID signal of an NMR experiment. As can be seen in Figure 1.6, the FID consists of both a real (blue) and imaginary (green) channel. Figure 1.7 shows the frequency domain spectrum of the FID signal found in Figure 1.6. Note that neither the time domain FID nor the spectrum is referenced and therefore only show data points in the x-axis. Additional information collected during acquisition of the NMR data would be acquisition time. Acquisition time in the x-axis of the FID signal will allow us to calculate the sweep width in the frequency domain. Typically, NMR spectra are normalized in the frequency domain by the use of an internal standard. This standard is used as a zero reference, and all other signals are calculated to be parts per million (ppm) chemical shift from this reference.

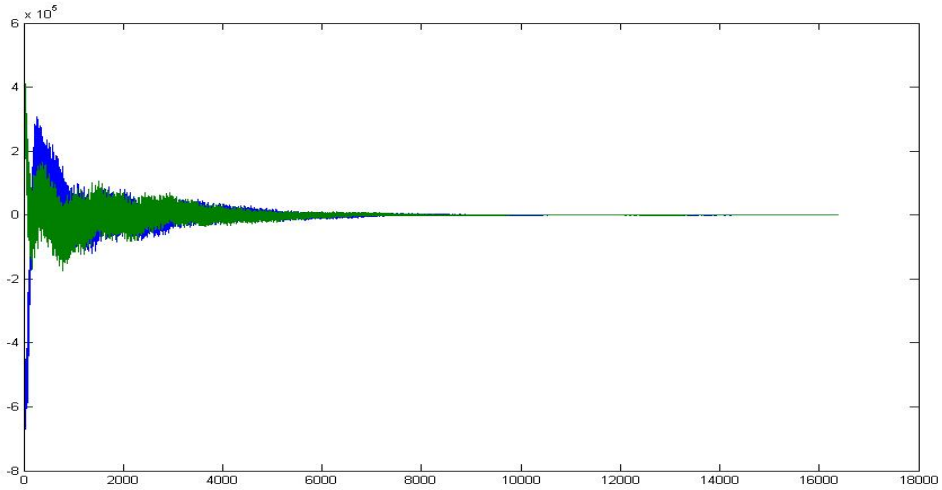

Figure 1.6 - FID of a typical NMR experiment showing real (blue) and imaginary (green) channels using quadrature detection.

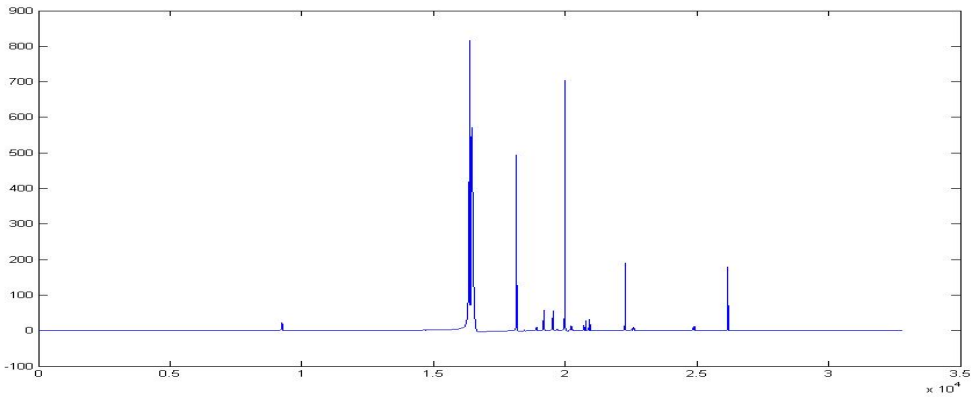

Figure 1.7 - Typical spectrum (frequency domain) of an NMR experiment.

We finally arrive at a complete description of an NMR experiment. This NMR experiment is the beginning of a workflow for NMR based metabolomics. This complex spectrum contains in it a wealth of information on the underlying chemical makeup of the mixture.

1.2. Objective of the Thesis

Understanding the NMR experiment, we now have the basis for exploring the information contained within this data. This thesis will journey into the various aspects of NMR based metabolomics, and the methods by which we can extract this useful information. Chapter 2 of the thesis will discuss the aspects of taking this complex spectrum from an NMR experiment to relevant biological models. Once mathematically modeled, the inherent data can be more easily understood and manipulated in a statistically relevant manner. Chapter 3 will address one of the major difficulties of working with the spectra of biological fluids, which is the presence of baseline distortions. Chapter 4 will explore the effects of data modeling discussed in Chapter 2 to building multivariate statistical models. These effects include over fitting and variable bias. Chapter 5 will discuss the different statistical models and techniques for analyzing metabolomics data, with examples in both asthma and S. *pneumoniae* infection. Chapter 6 will apply a unique workflow of combining the techniques discussed in previous chapters on the modeling and prediction of Streptococcus *pneumoniae* infection. Finally Chapter 7 will explore a novel technique of using normalized correlation difference maps.

2. NMR Data Preprocessing

In this chapter, we discuss the practical implications of using NMR data in a metabolomics workflow. Metabolomics is a combination of a powerful measurement technique such as NMR with chemometrics to extract biologically and statistically relevant results about metabolites from biological samples. To properly perform chemometrics on NMR data, one must properly preprocess the data. This chapter will discuss these methods of preprocessing NMR data, along with an in depth discussion on the targeted profiling technique, including the underlying spin simulated models used to identify and quantify the metabolites in the biofluid mixture.

NMR data can be viewed as a series of x-y data points. Along the x-axis lies the relative frequency at which the proton resonates, typically in the range of ~1000 Hz. A more standardized unit is "ppm", which is a magnet independent measurement of chemical shift, and is normalized to the magnetic field. Along the y-axis are the signal intensities. The intensities are dependent on the gain of the receiver and the area described by this data is directly proportional to the number of protons that are detected by the relaxation processes. At full resolution this data has a length that is a power of two; typical numbers are 32768 and 65536. Data under this representation, although highly detailed with a wealth of information, is not very practical in its use in the area of metabolomics. The study of metabolomics also combines NMR data from many different samples. Both the size of the arrays and the number of samples requires the consideration of data reduction. There is a practical limitation in working with large datasets, and the following sections will discuss the various data reduction techniques that can be used in metabolomics.

2.1. Binning

Binning, or sometimes referred to as bucketing, is a term coined by Nicholson et. al (1999), to describe the process of reducing the number of variables by grouping the intensities into evenly spaced "bins" or range of x variables (See Figure 2.1). Intensities are added together. This data is typically normalized with the total area of the spectrum. Aside from a reduction in the number of variables, another added benefit of binning a

spectrum is that small variations in chemical shifts of various peaks will be masked if a wide enough bin is chosen to cover the range of this variation. There are however, disadvantages to using this data reduction technique. As the width of the bins is increased, the less specificity there will be in determining what peaks are contained within each bin. Both the objectives of accounting for chemical shift variations and maintaining bin specificity are two opposing objectives. Determining an optimal bin width is not a trivial procedure. Another disadvantage in binning is in having evenly spaced bins. This often causes intensities from a single peak to be divided into two bins. Authors such as Lefebvre (2004) have proposed smart binning techniques which may help with the binning process.

2.2. Targeted Binning

One technique pioneered by Chenomx Inc. is termed targeted binning (www.chenomx.com). By making use of a database of resonance locations for a library of metabolites, one can create custom bins or integral areas (See Figure 2.1). These bins can be customized to account for variations in pH and ionic strength of the mixture. Specific proton clusters can be individually labeled bins. This technique of customized binning provides the flexibility to encompass all the resonances that belong to a specific metabolite into one variable. The issues of resonance overlaps are still a problem in this technique, which can lead to over estimates of integral areas.

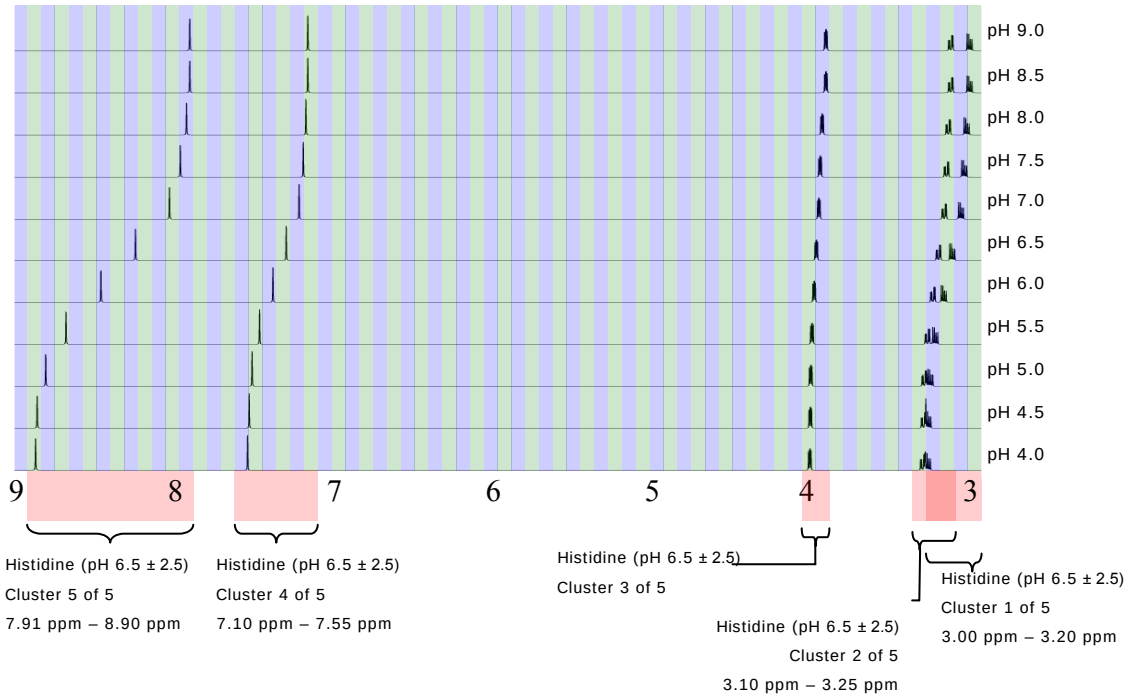

Figure 2.1 – Standard and targeted binning on histidine resonances.

2.3. Measuring Compound Concentrations

Another more elegant solution for data reduction of chemometric NMR data is to determine the compound concentrations *a priori*. By doing so, this ensures that the variables are unique and comparable across samples in a multi-sample experiment. One drawback of this methodology is in being able to both identify and quantify each compound in the spectrum accurately. This task is also non-trivial. Since any subsequent analyses of this data set is based on what compounds were actually identified, it is also possible to miss correlations that exists between compounds/variables that were not measured. If identification and quantification is done thoroughly however, this method is the better choice as there are no ambiguities in what the variables are measuring. It is in this author's opinion that this methodology is more powerful than a thorough analysis of ambiguous variables, as the end result will still be ambiguous.

Identification and quantification of compounds can be done in both the time and frequency domains. Review papers by Vanhamme et. al (2001) and Mierisová et. al (2001) give good overviews as to the current quantification techniques in both time and frequency domains. Quantification in the time domain can be summarized as fitting the parameters of the FID equation, much like equation 1.1. Equation 1.1 is for a single proton, a fit in the time domain will be a simultaneous fit of all protons known to be in the sample. In the frequency domain, there are two broad categories for quantification. One category is techniques that are based on the integration of peak areas. These techniques are very similar to binning described earlier, except that the range for integration can be chosen manually to avoid cutting a peak into two bins. With integration techniques there are still open issues with overlapping peaks. The other category of techniques is based on modeling the peak clusters in the frequency domain. Typically, these models are based on Lorentzian functions. Theoretically, modeling peak clusters as Lorentzians is equivalent to modeling equation 1.1 in the time domain. There are other functions that are used such as Gaussian or Voigt functions which are a mix between Lorentzian and Gaussian (Vanhamme et. al, 2001).

One of the prime methods used in this work for metabolite concentration measurements is termed targeted profiling. This is a technique again pioneered by Chenomx Inc. Chenomx provides software to identify and quantify metabolites (Weljie et. al, 2006). This software has the added benefit of a compound library database that relates clusters unique for each metabolite for a library of compounds. These built in constraints make identification and quantification of overlapping peaks possible.

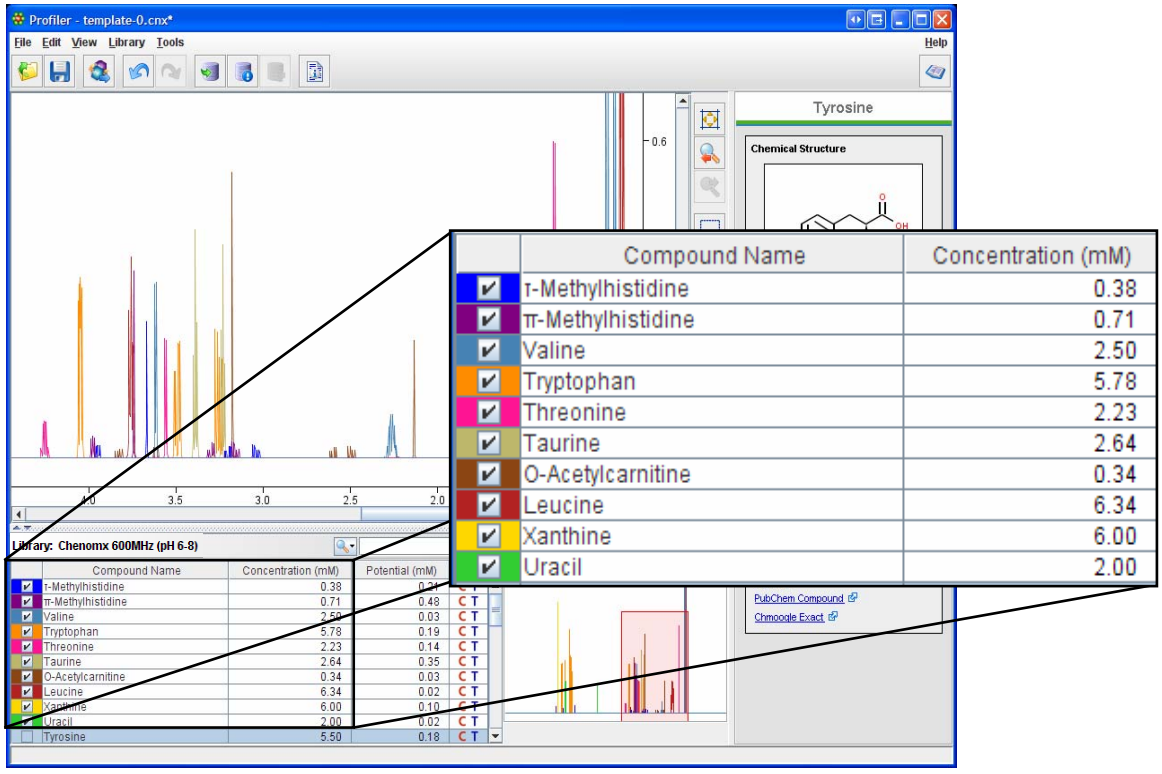

Figure 2.2 – Targeted profiling technique using Chenomx NMR Suite software.

Figure 2.2 shows a screenshot of Chenomx NMR Suite 4.6. Shown in this figure are the color coded resonances for a series of metabolites plus their associated concentrations determined using a reference peak.

2.4. Underlying Model to NMR for Quantification and Identification

While the preprocessing technique of targeted profiling is useful in a metabolomics workflow, the challenge to such a technique is the appropriate mathematical model used to represent the pure component spectrum in a mixture of metabolites. In this section we discuss the development of such a mathematical model for use in the targeted profiling technique.

Spin simulation of proton resonances have been developed from the early works of Roberts (1962). His work describes the spin-spin interactions between nuclei. From a quantum level description, Roberts describes a framework for predicting NMR transitions that form in the frequency domain. Each transition is a result of complex quantum energy interactions between different energy states of a proton. These interactions are complex, and every single nucleus in a system has a level of interaction with each other. The solution to large systems is very time consuming, even on the fastest computers of today.

A simpler set of equations was developed to estimate second order effects of only binary nuclei interactions. This simplified spin simulator was developed in collaboration with Chenomx Inc. The objectives of this work are to be able to simulate the FID signal of compounds and to be able to predict the spectral signature of compounds at varying field strengths. The basic principle behind the spin simulator is the FID equation. In order to simulate the FID signal of compounds, we have to simulate the FID signal of individual clusters of a compound. Since these signals are linearly additive, we can simply add the FIDs together to simulate the entire compound. Again, the FID equation is given as follows:

$$M = M_o \cdot \sin(\omega_o \cdot t) \cdot \exp\left(-t \cdot \left(\frac{1}{T_2} + R_x\right)\right) \quad \textbf{(2.1)}$$

Note that this equation is very similar to Equation 1.1, except an additional term R_x. This term was added to account for features in the spectra that was due to exchange of protons, not sufficiently explained by T_2 itself. The parameters T_2 and R_x are cluster specific and are measured either directly or indirectly from an acquired spectrum. The parameters M_o and ω_o are peak specific, and are calculated based on measurements of the cluster centers and j-coupling constants. In order to estimate the peak locations, the cluster center is sequentially split into each new peak location for every j-coupling constant. Therefore the total number of peaks for a cluster that has n j-coupling constants is 2^n. The amplitudes of each peak are not split evenly due to second order effects. Second order effects happen when two clusters are coupled to each other and have very close chemical shifts. For each split the amplitudes are modified based on an algorithm outlined in Table 2.1.

Table 2.1 – Amplitude modification for 2nd order effects.

$\delta\omega(i,k) <= 0$				$\delta\omega(i,k) > 0$			
$j(i,k)/abs(\delta\omega(i,k)) <= 1$		$j(i,k)/abs(\delta\omega(i,k)) > 1$		$j(i,k)/abs(\delta\omega(i,k)) <= 1$		$j(i,k)/abs(\delta\omega(i,k)) > 1$	
Lower Frequency Peak	Higher Frequency Peak	Lower Frequency Peak	Higher Frequency Peak	Lower Frequency Peak	Higher Frequency Peak	Lower Frequency Peak	Higher Frequency Peak
+ $(A_o/2)*(j(i,k)/abs(\delta\omega(i,k))$	- $(A_o/2)*(j(i,k)/abs(\delta\omega(i,k))$	+ $(A_o/2)*(abs(\delta\omega(i,k)/j(i,k))$	- $(A_o/2)*(abs(\delta\omega(i,k)/j(i,k))$	- $(A_o/2)*(j(i,k)/abs(\delta\omega(i,k))$	+ $(A_o/2)*(j(i,k)/abs(\delta\omega(i,k))$	- $(A_o/2)*(abs(\delta\omega(i,k)/j(i,k))$	+ $(A_o/2)*(abs(\delta\omega(i,k)/j(i,k))$

A_o in Table 2.1 are the amplitudes of the peak previous to the split. This modification is applied to cluster i, n number of times. k being the k^{th} split from 1 to n, since there are n J-coupling constants. These equations require the parameters j(i,k) and $\delta\omega(i,k)$. $\delta\omega(i,k)$ is calculated based on the knowledge of the absolute cluster centers. These equations were developed based on what was a good model to predict this affect. Figure 2.3 shows this effect on an AB spin system.

Other equations can be found in textbooks such as Harris (1983). These equations typically are based on knowledge of the weighted centers (by amplitude) of the clusters. Weighted centers require the knowledge of the amplitudes. Since the objective of the spin simulator is to predict the amplitudes, the amplitudes are not known *a priori*. An

iterative scheme is proposed in this thesis, where by the equations in Table 2.1 are used to first estimate the amplitudes in order calculate an estimated weighted center. Then equations based on weighted centers found in Harris (1983) were subsequently used iteratively to give a better estimate of the amplitudes and weighted centers. This algorithm was applied and only showed marginal improvement to the final amplitudes of the peaks. The final algorithm implemented in Chenomx NMR Suite do not use equations from Harris (1983).

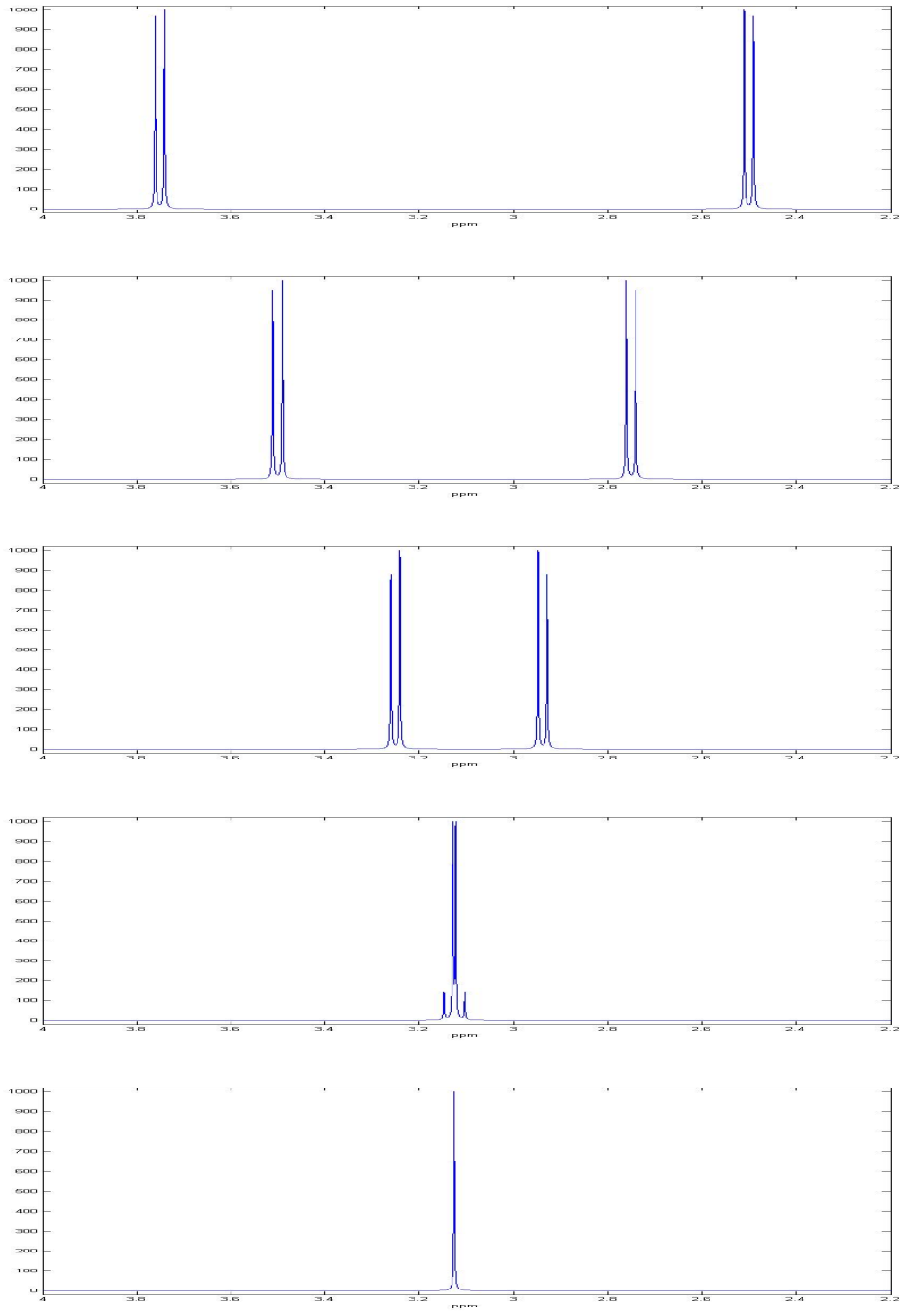

Figure 2.3 – 2nd order effects on an AB system, with decreasing distance between cluster centers.

Currently the cluster specific parameters of T_2 and R_x are not measured. Some reasonable values are used currently. Future developments will be to either measure the possible values for T_2 and R_x, or to estimate T_2 and R_x parameters by optimizing those parameters under the current framework to match a measured spectrum.

In order to test the spin simulation accuracy, a spectrum of tryptophan was used for the analysis. In Figure 2.4, we first obtain the model parameters using an 800 MHz acquired spectrum of tryptophan. In Figure 2.4, the parameters were adjusted such that the simulation matched the acquired spectrum perfectly. To check the robustness of the parameters that were estimated, a separate simulation of the same compound tryptophan was performed, but at the lower field strength. In Figure 2.5, the back calculated (predicted) spectrum at 500 MHz closely matches the acquired spectrum. In this case, the same parameters estimated previously were used. As can be seen in Figure 2.5, the predicted spectrum also matches the acquired spectrum quite well, including the second order effects.

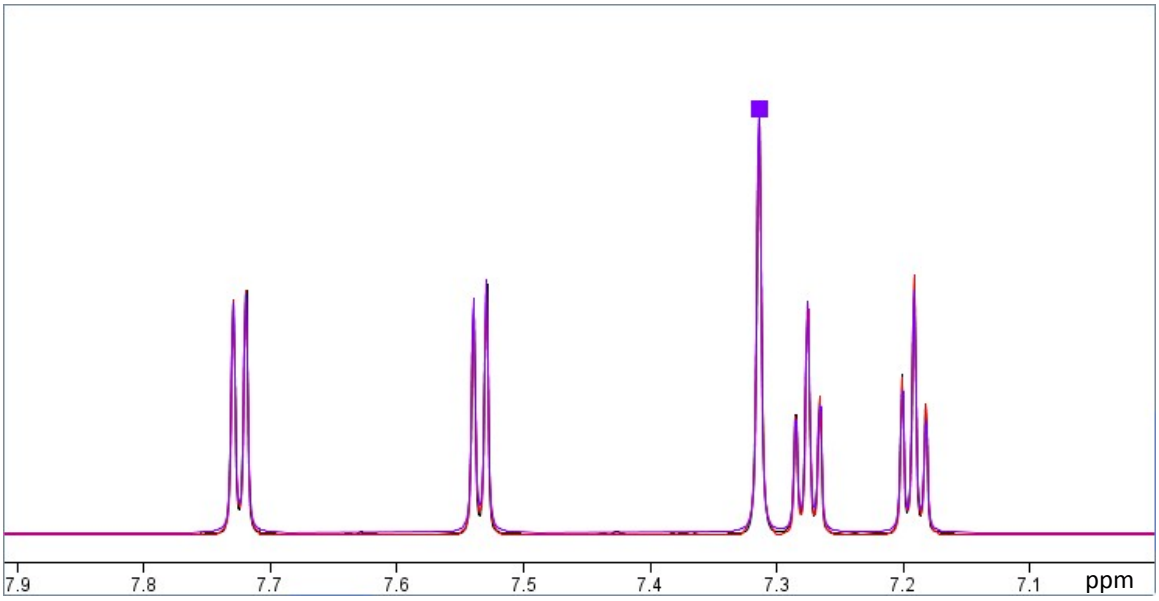

Figure 2.4 – Spin simulator results of aromatic Tryptophan clusters at 800 MHz. Purple is spin simulator results. Black is an acquired 800 MHz spectrum. Red is fitted spectra using Lorentzian curves to obtain parameters.

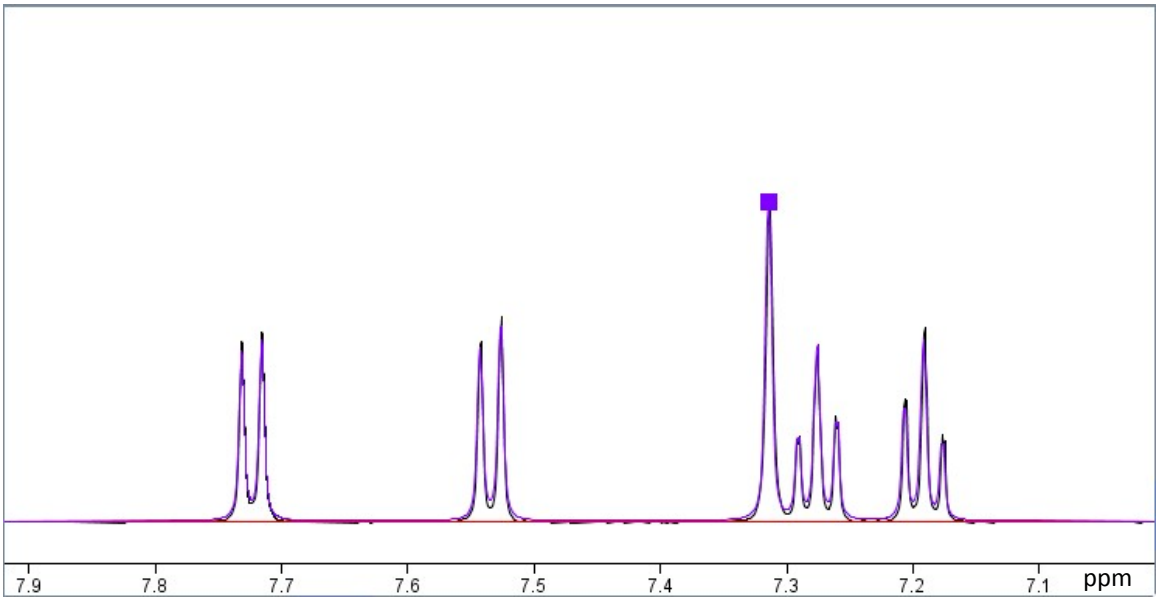

Figure 2.5 – Spin simulator results for aromatic Tryptophan clusters at 500 MHz. Purple is back calculated results, with parameters measured from 800 MHz. Black is an acquired 500 MHz spectrum.

The new modified spin simulation algorithm is shown to be able to reproduce the line shapes of real spectra. The innovation in the algorithm is in the time it takes to solve complex spin systems. In Table 2.2 we show the execution times of solving a standard spin simulation as described in Roberts (1962). We can compare these execution times with the modified algorithm, which has execution times <0.01 s for all the spin systems tested in Table 2.2. Given the major performance gains of this algorithm, the spin simulator is now implemented fully into Chenomx NMR Suite. The performance of this algorithm allows the simulation of spin-spin interactions in real time, as cluster centers are moving, or as J-coupling constants are changing.

Table 2.2 – Execution times for standard spin simulation algorithm.

# of Spins	Execution[1] Time	Largest Submatrix Size (n x n)	# of Transitions (computed)	# of Transitions[2] (visible)	RAM Usage[3]
2	< 0.01 s	4	4	4	128 B
3	< 0.01 s	9	15	9	432 B
4	< 0.01 s	36	56	42	1.59 KB
5	< 0.01 s	100	210	64	5.7 KB
6	< 0.01 s	400	792	286	21.7 KB
7	0.05 s	1,225	3,003	597	80 KB
8	0.3 s	4,900	11,440	1,124	306 KB
9	1.1 s	15,876	43,758	6,814	1.12 MB
10	7.6 s	63,504	167,960	22,449	4.33 MB
11	66 s	213,443	646,646	75,947	15.0 MB
12	*6.7 min*	853,776	2,496,144	*?*	63.7 MB
13	*39.6 min*	2,944,656	9,657,700	*?*	244 MB
14	*4 hours*	11,778,624	37,442,160	*?*	947 MB
15	*24 hours*	41,409,225	145,422,675	*?*	3644 MB

Notes:

1. Execution time does not include time needed to render transitions on-screen.
2. Many transitions that are computed have 0 height, and are thus invisible.
3. RAM usage assumes matrix operations are performed on Java ‘double-precision’ numbers (Size of double: 8 bytes, Size of a Transition: 24 bytes)

2.5. Conclusion

The results of the spin simulation show that the underlying model for generating NMR spectrum can be accurately modeled using estimated parameters from pure acquired spectra. Moreover, this modified spin simulation algorithm is significantly faster than a full spin system calculation. This algorithm is better suited for the simulation of spin systems in real time, and is an important step in the deconvolution of the spectrum into metabolite concentrations as mentioned earlier in this chapter. Also, we can conclude that there are several different methods of preprocessing an NMR spectrum such as spectral binning, Targeted Binning, and Targeted Profiling. The considerations for the method of preprocessing an NMR spectrum for statistical modeling will be explored in a later chapter.

3. Baseline Correction Problem and Some Automated Solutions

One of the reasons for doing one dimensional NMR experiments is to identify peaks that correspond to various protons and then to accurately estimate the concentration of the compound that these protons belong to. In order to do this properly, it is very important that the baseline of the spectrum is free from anomalies. Some of these anomalies can be caused by errors in the first few points in the FID. Another common cause of baseline problems is due to the signal from larger molecules. In urine, these are typically large lipids and proteins. In order to correct for these baseline signals, a possible solution is to model this baseline and subtract it from the original spectrum. A manual method for modeling a baseline is to manually select points along the spectra that are considered a baseline point. These points are then joined together by either straight lines or any order of a polynomial spline. A description of splines can be found in Ahlberg et. al. (1967). More advanced techniques use automated algorithms to find baseline points. What defines a baseline point is often a difficult task to determine. Several methods proposed are to model the baseline based on the mechanism that generated the baseline signal.

3.1. Automated Baseline Point Determination[1]

Experimental nuclear magnetic resonance (NMR) spectra tend to contain baseline distortions artifacts which can be caused by a variety of different sources, including instrument drifts and unwanted macro molecule signals. Metabolomics applications of NMR spectra often require the identification and quantitation of metabolites found in complex mixtures, since these mixtures can give a snapshot of the state of an organism. It is important to have a flat baseline in order to accurately quantify, hence the need for a good baseline correction algorithm. Systematic baseline distortions also add unwanted correlations in spectral binning data when building correlation models.

---

[1] Some of the material that appears in this section has been published in *Journal of Magnetic Resonance.* **2007**, 187(2), 288-292.

Early development on baseline correction algorithms includes work described by Pearson (1977), in which baseline correction can be broken into three steps. The first step is to determine the signal and baseline noise in the spectrum. The second is to use that information to build a model of the baseline. This model can be represented using interpolated line segments, or cubic splines if a smoother line is desired. Finally the third, and somewhat trivial step, is to "correct" the signal by subtracting the baseline model from the original signal. Further developments by Zolnai et. al (1988), Heuer and Haeberlen (1989), Gunter and Wuthrich (1991), and Bartels et. al (1995) all follow this standard pattern and have made significant contributions to each step.

While the problem of baseline correction in the realm of NMR signals is not new and there are some good solutions already available, in our experience the available methods work best on NMR spectra that do not have a very high signal density (Pearson, 1977, Gunter and Wuthrich, 1991, Bartels et. al, 1995). Many existing algorithms tend to be overly aggressive, often destroying the line shapes of prominent peaks in spectra with a wide dynamic range of peak shapes and sizes. The application of many of these methods to metabolomics data is therefore problematic, since NMR spectra of complex biofluids often result in very signal-dense spectra. In this chapter we propose a new algorithm for baseline correction which addresses this problem in a way that does not destroy the line shapes of prominent peaks. Our algorithm is designed for more densely populated spectra, but retains good performance in sparsely populated spectra as well. The algorithm was developed based on our combined knowledge of both NMR signals and of the baseline distortions that are common in the realm of complex mixtures. This chapter will focus on step one of the general three-step process: A systematic application of heuristic rules which can accurately determine the baseline points in a 1D NMR spectrum.

3.1.1. Approach

This section contains a detailed outline of our baseline correction algorithm, as implemented in Chenomx NMR Suite 4.6. The goal of the algorithm is to differentiate the regions of the spectrum (*S*) that are considered to be baseline noise from those that are considered to be signal. This determination is then stored in a Boolean vector known as a signal map (*SM*). *SM* has the same dimensions as *S*; each element contains information about whether the point corresponding in *S* is signal (true) or baseline noise (false).

The first and most important step in the algorithm is the high pass signal identification step. The objective here is to conservatively identify regions of the spectrum that are signal by looking at a modified version of *S* wherein all low frequency curves and rolls have been removed. Once the signal regions are identified, everything else can be considered baseline points. In order to accurately determine what is signal, the algorithm first attempts to calculate the standard deviation of the noise in *S*. This is a common step in other baseline correction algorithms (Pearson 1977, Gunter and Wuthrich 1991, and Bartels et. al 1995). However, the typical method for determining the standard deviation of noise by dividing the original spectrum (*S*) into multiple regions is insufficient. Rolling baselines and areas of high signal make it difficult to estimate the noise in a spectrum.

To overcome this problem, the algorithm proposed here chose to first use a high pass filter on the spectrum. Specifically, a moving average filter was used. This filter is designed to pass 0.5% of the highest frequency through the Nyquist frequency. The resulting signal is known as the high pass filtered spectrum (*HPFS*) and contains only the high frequency noise and signal. Figure 3.1 shows a spectrum before and after the high pass filter has been applied. Note this algorithm can be generalized to different nuclei spectra and therefore the author has removed the ppm x-axis to show the applicability to generalized sweep widths. From Figure 3.1, we can also see the resultant spectrum is highly distorted and not very useful in itself. However, the *HPFS* is still useful in obtaining a good estimate of the high frequency baseline noise, because rolls in the baseline have been removed, and signal dense areas have been narrowed.

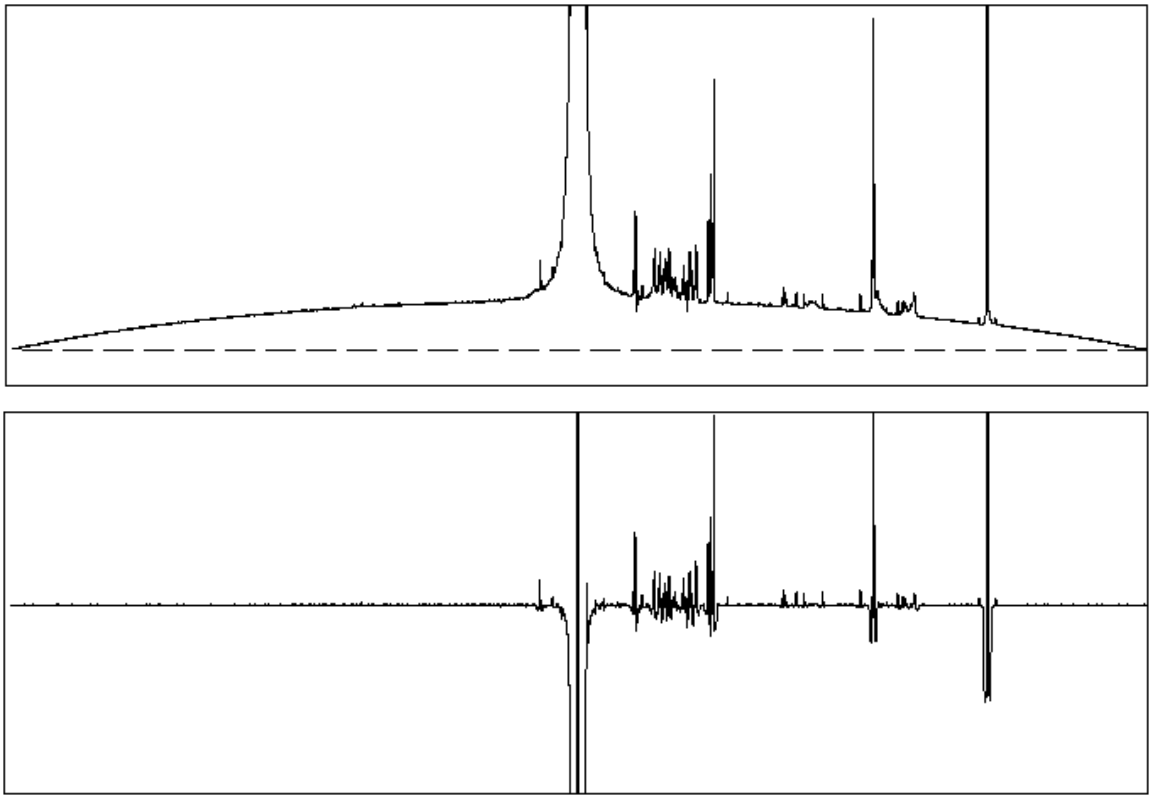

Figure 3.1 – Top: Original spectrum (*S*) with a noticeable baseline distortion. Bottom: High pass filtered spectrum (*HPFS*) showing the removal of the low frequency distortions. Full sweep width is shown.

At this point the *HPFS* is divided into evenly spaced segments, and the standard deviation of each segment is calculated. A percentage (*bfraction*) of the segments with the lowest intensities are assumed to be baseline signal, and the standard deviation of only the points contained within these segments is recalculated (*stdn*). *bfraction* can be adjusted based on the spectrum signal density. A *bfraction* value of between 0.2 and 0.5 was found to work well for complex mixtures.

Once *stdn* has been determined, the next step is to determine what percentage of the entire spectrum is signal. We continue to use the *HPFS* and consider all points with absolute intensities greater than two times the standard deviation of the noise to be signal. The indices of these absolute intensities are sorted based on the intensities themselves and then used in the signal windowing step.

The signal windowing step returns back to the original spectrum (*S*). Each signal point found in the previous steps is now used as the center of a signal window. The signal window width used is 0.2% of the total sweep width of the spectrum. Each point inside of the signal window is now also marked as signal in *SM*. Figure 3.2 shows a spectrum overlaid with the baseline points that were found after the signal windowing step.

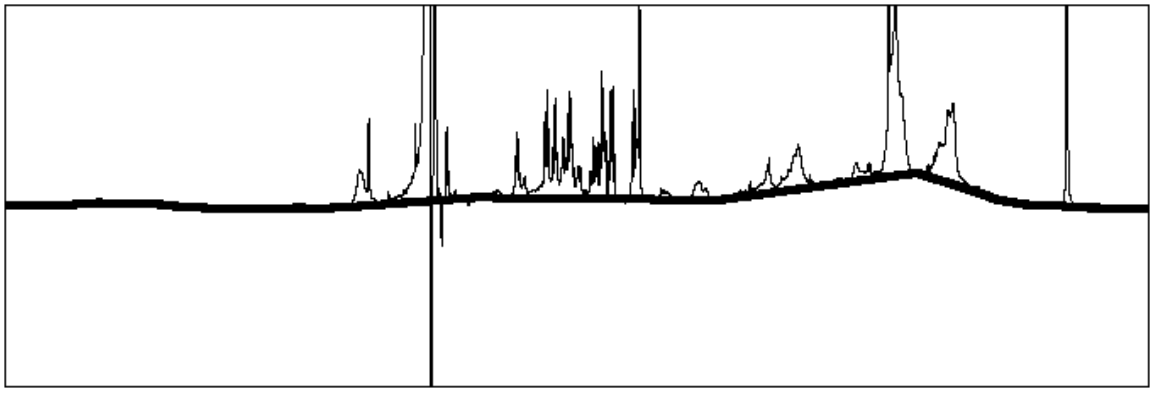

Figure 3.2 – Baseline points defined after signal windowing step. Full sweep width is shown.

The novelty of this algorithm is the use of a high pass filter. However, it is also a weakness: the high pass filter applied to very tall or large peaks in a spectrum will often misidentify the tails of these peaks as baseline in the signal map. In order to correct for this, a second step is applied. The objective of this step is to determine the most prominent Lorentzian peaks in the spectrum and guarantee that their tails are marked as signals in *SM*. This is because the tails of the most prominent peaks are often filtered out in the high pass filter, and misidentified as baseline signal due to their size relative to the signal window.

The first part of this step is to calculate the average or mean of the entire spectrum (S) in the frequency domain using only the positive values. Then, using an automatic peak picking algorithm, peaks that are twice the mean of the spectrum are located. The widths of the peaks are determined by walking halfway down both sides to find the half width of each peak. The peaks are then mathematically modeled as pure Lorentzian lineshapes and the central portion of *S* that contains 95% of their area is marked as signal in *SM*. Note that this often fixes regions that were erroneously marked as baseline in previous steps.

A 95% cutoff was needed because Lorentzian peaks have infinite tails. The algebraic model for a Lorentzian is:

$$L(x) = \frac{A \cdot w^2}{w^2 + 4 \cdot (x - c)^2} \tag{3.1}$$

Where, for any given position x {Hz}, width w {Hz}, center c {Hz}, and amplitude A, the function L is the intensity of the Lorentzian at x. Once these additional "signal points" are marked in *SM*, the determination of signal and baseline points is complete. Figure 3.3 shows the same spectrum as in Figure 3.2 with baseline points overlaid (in thick black) after correcting for prominent Lorentzian peaks. You will notice the correction of the misinterpreted baseline points (in thick black). The 95% regions from the picked Lorentzian peaks (A and B) and the region added to the SM (C) are also shown in Figure 3.3.

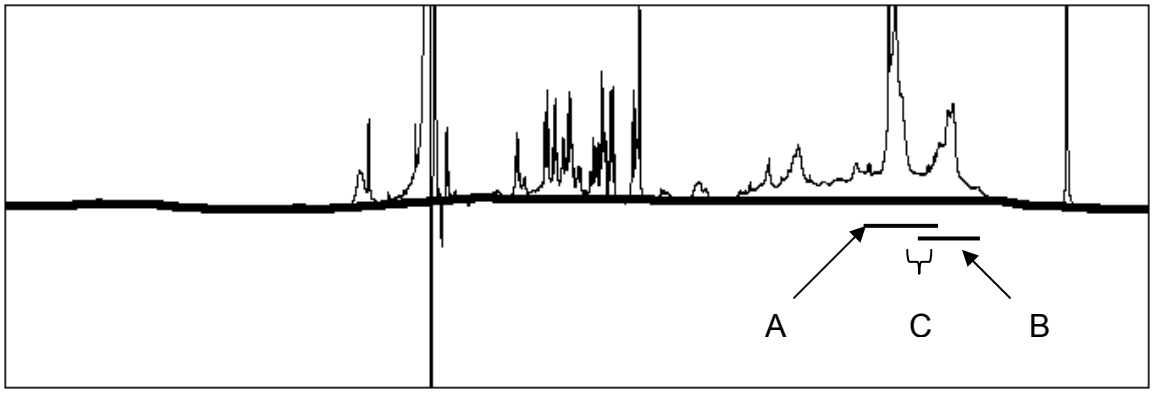

Figure 3.3 – Baseline points defined after correction for prominent Lorentzian peaks. Full sweep width is shown.

We showcase the algorithm's ability to accurately determine the baseline points. To model these points in a spectrum, the original baseline points were used and a simple linearly interpolated line was used to fill in the gaps between the baseline points. A more sophisticated natural cubic spline model is used in Chenomx NMR Suite 4.6.

3.1.2. Application

The performance of the algorithm is demonstrated in the following two examples, which were acquired from different NMR spectrometers and have different baseline distortion problems. As well, the first of these examples has a high signal density while the second example is sparse in signal density. These spectra were also chosen to clearly show the algorithm's ability to handle gross distortions in the baseline, while at the same time showing that it is able to non-destructively handle the more subtle baselines generated from the most advanced spectrometers today.

For our first example, we applied the algorithm to an NMR spectrum of an acidic plant extract. The exact details on the sample are unknown, however this spectrum was chosen for its obvious baseline distortion and signal density. This sample was run through an NMR flow system on a 400 MHz Varian spectrometer using a vast1d pulse sequence. Some of the older flow systems, which make use of the ssfilter VNMR command, do not always create straight baselines. As can be seen from the black line in Figure 3.4a, the

original spectrum had a fairly high signal density, as well as an obvious baseline distortion. The baseline points identified by the algorithm are shown, along with linearly interpolated points in between the gaps (in thick black) in Figure 3.4b. Finally, the baseline corrected spectrum (i.e. after subtraction) is displayed in Figure 3.4c.

Our second example uses another acid extract sample acquired on an older JEOL Spectrometer, which did not have digital filtering. The lack of digital filtering is probably the cause of this spectrum's pronounced baseline roll. This spectrum was acquired on a 500 MHz magnet using a single-pulse sequence. Figure 3.5a shows the original spectrum. Figure 3.5b shows the baseline points identified by the algorithm with linearly interpolated points in between the gaps (in thick black). Figure 3.5c shows again the high quality spectrum after the baseline distortion has been removed.

The baseline correction algorithm outlined was designed using characteristic distortions found commonly in spectra from complex mixtures. It follows the established three-stage template and aims at ensuring the accurate determination of baseline points without indentifying too many false positives. The result is a high quality baseline correction algorithm that can be used in a variety of metabolomics applications.

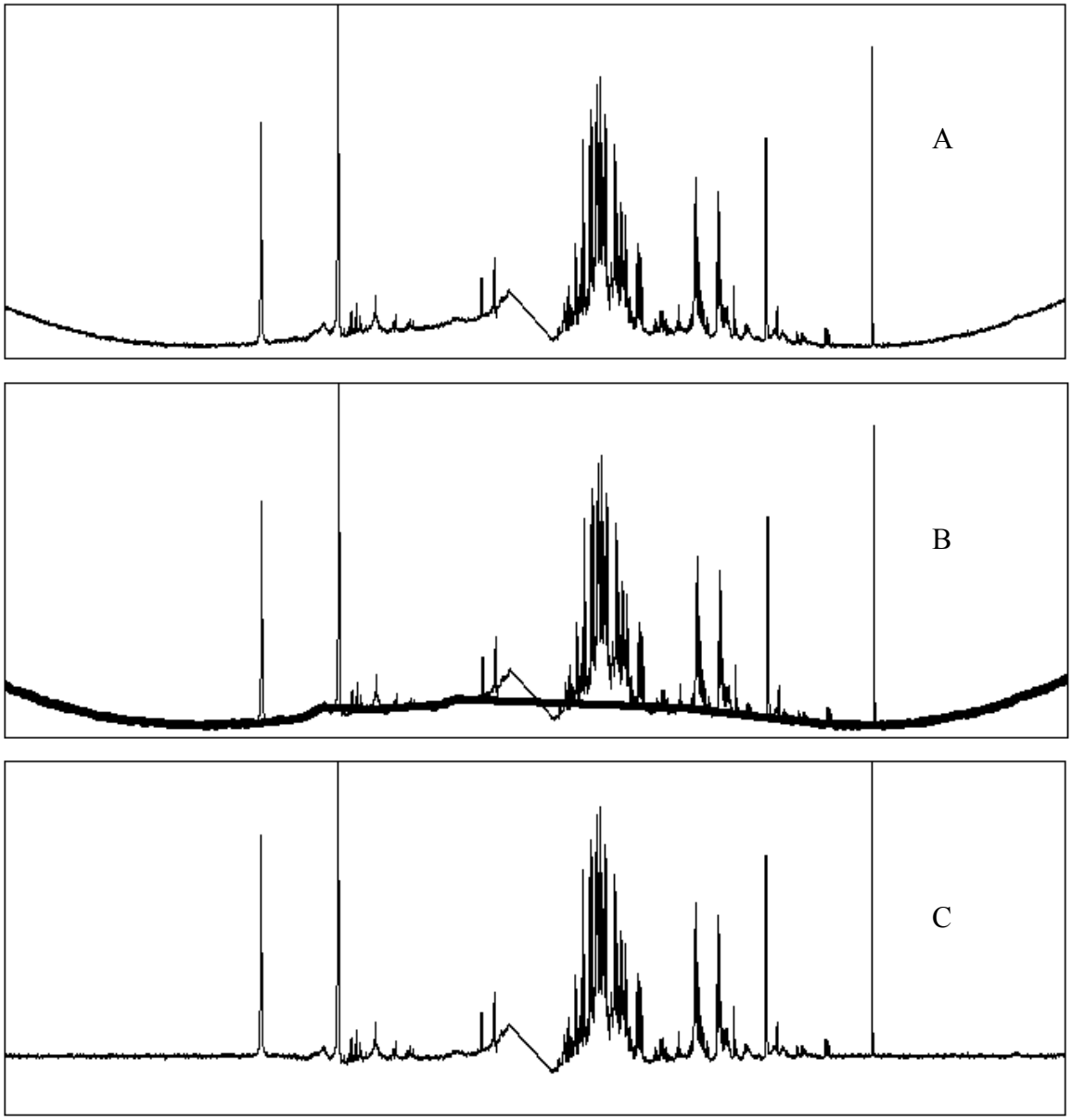

Figure 3.4 – A) Original spectrum of acidic plant extract. B) Baseline distortion model. C) Spectrum after baseline correction. Full sweep width is shown.

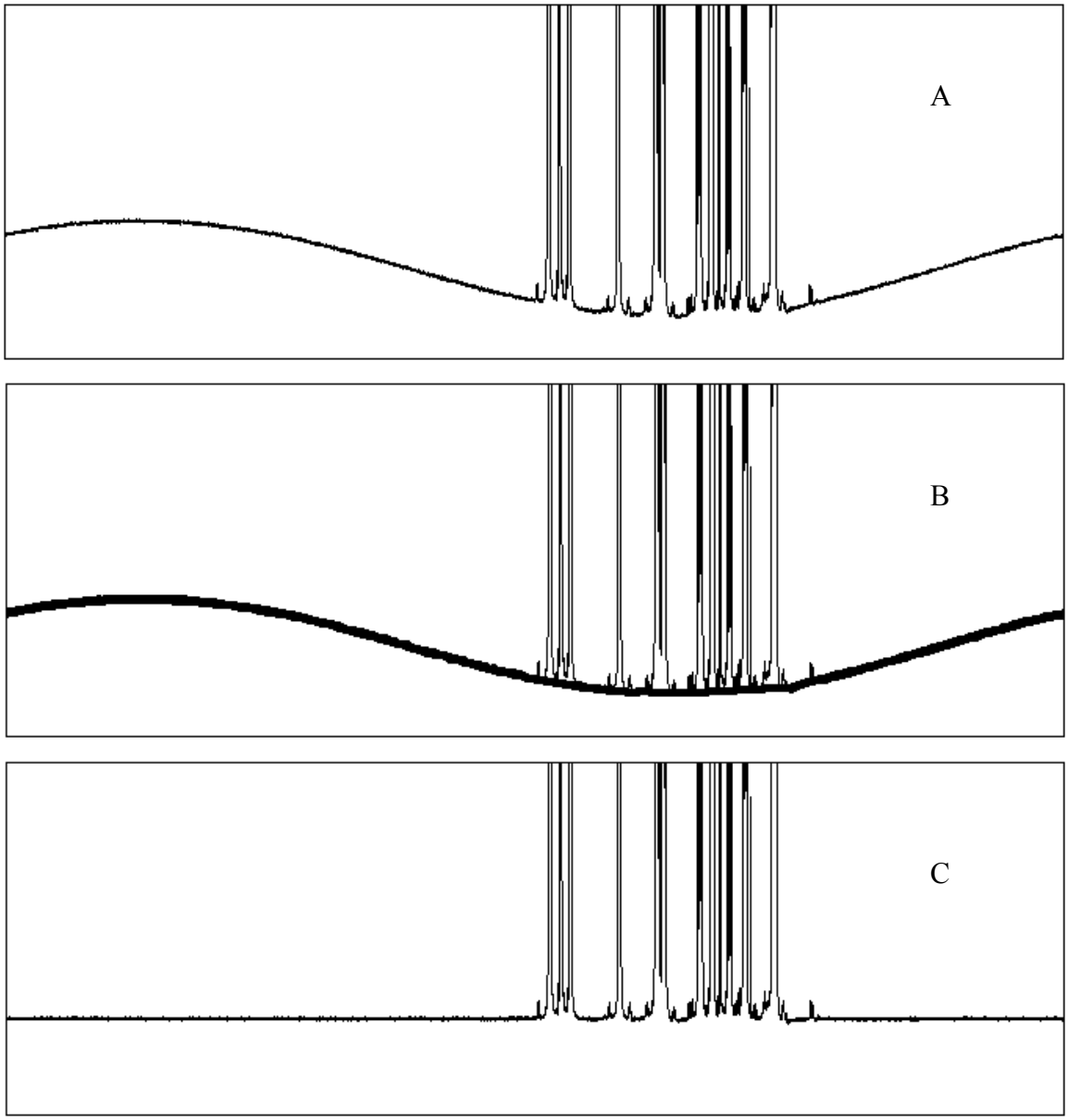

Figure 3.5 – A) Original acid extract spectrum. B) Baseline distortion model. C) Spectrum after baseline correction. Full sweep width is shown.

3.2. Modeling Lipid Distortions

In the second method, we propose to model the baseline in the frequency domain. In order to model the baseline of a signal caused by large macromolecules, we first look at the signal of the macromolecules themselves. Typically signal from these large macromolecules are very broad signals, due to the relatively fast decay rate (T_2). Figure 3.6 shows a ^{1}H NMR signal of a lipid extract from urine. Lipid extraction was done using a chloroform/methanol liquid/liquid separation similar to the one described in Khan et. al (2002).

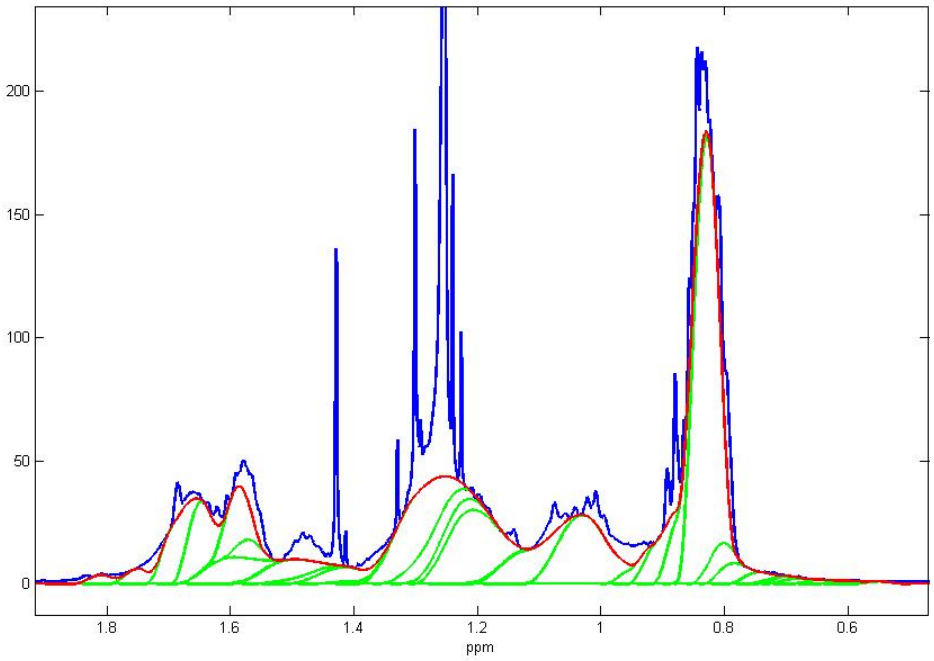

Figure 3.6 – Modeling of ^{1}H NMR signature of lipid extract in urine

Using this signature of lipid extract a series of broad Beta curves, which mimic the broad signals of a large macromolecule, were fitted to nearly match the shape of the baseline as per Figure 3.6. Beta curves were chosen to model these broad baseline signals. Beta curves are given by the following equation:

$$y = \frac{1}{B(a,b)} \cdot x^{a-1} \cdot (1-x)^{b-1} \cdot I_{(0,1)}(x) \qquad \textbf{(3.2)}$$

where a and b are parameters, and $B(a,b)$ is the Beta function. $I_{(0,1)}(x)$ makes this curve truncated between x values of 0 and 1. This property of truncated tails offers some numerical advantages. The Beta function is given by the following equation:

$$B(a,b) = \int_0^1 t^{a-1} \cdot (1-t)^{b-1} dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad \textbf{(3.3)}$$

where $\Gamma(x)$ is the gamma function.

The parameters *a* and *b* were chosen to be the same value to produce a symmetric function. The parameters used to fit the baseline were the area, span, and center for each Beta curve. Of these parameters, only the areas were automatically optimized using a non-linear optimizer in Matlab. After the lipid extract was fit, the same parameters were used to fit an unmodified urine spectra. Again in this case, the area was automatically optimized for each curve. Figure 3.7 shows an example of a baseline fit, using this algorithm. Figure 3.8 shows the resulting spectrum after the baseline curve was subtracted from the original spectra.

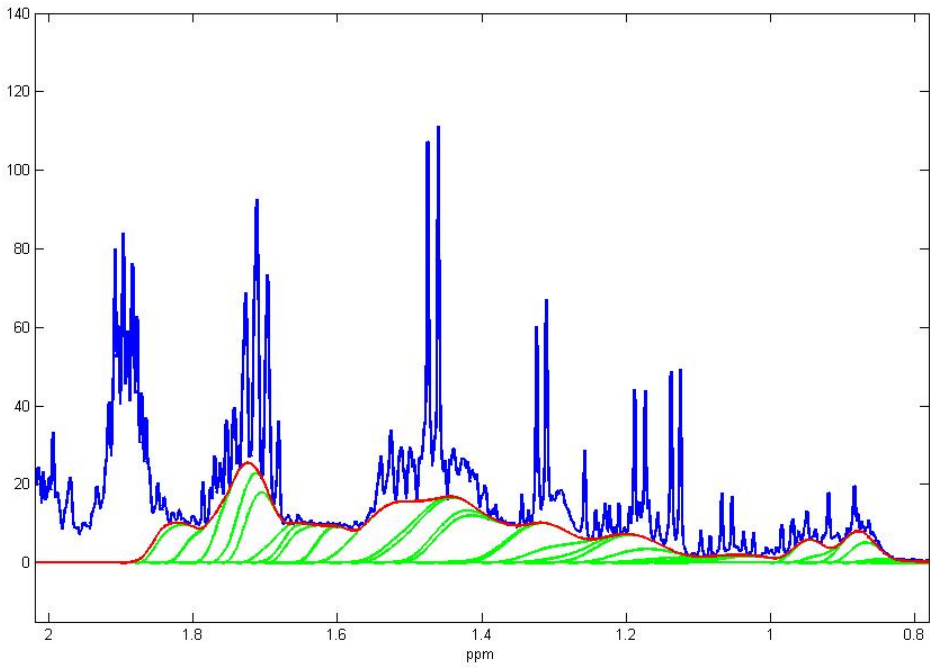

Figure 3.7 – Baseline model fit using a series of Beta curves in urine sample.

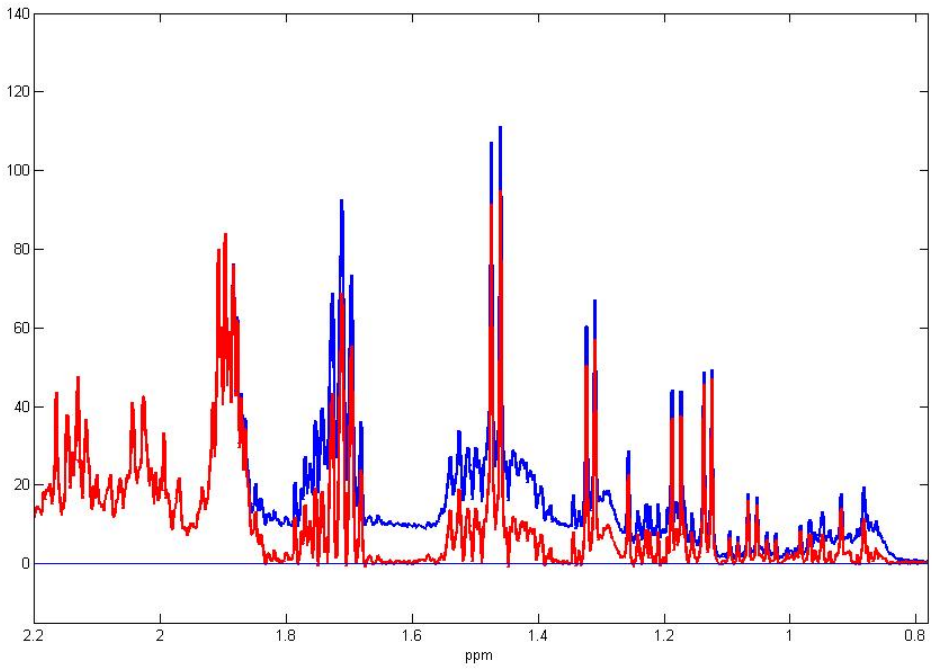

Figure 3.8 – Removal of baseline from original spectra. (Blue: original spectra, Red: spectra with baseline removed.)

3.3. Time Domain Baseline Correction

A third proposed method of performing baseline correction involves again the idea that these baseline signals are a result of macro molecules with small T_2 values. We make use of this information in the time domain. The idea is to remove the signal from the beginning of the time domain signal, since most of the signal from these large macro molecules will have decayed in the beginning of the time series. The time domain signal is truncated, and in order to retain the phase information, a linear backwards predictor is used to estimate the missing data without the macromolecule signal. First attempts at this idea involve a simple autoregressive (AR) model to perform this backwards predictor. Figure 3.9 shows some preliminary results of this procedure.

Here a serum sample was run twice. The first run was with a water suppression presaturation (presat) sequence (in blue), which involves a long low power radio frequency (RF) pulse used to saturate a specific frequency followed by a simple 90^{o}_{x} pulse. The second run was with a Carr-Purcell-Meiboom-Gill (CPMG) sequence (in green), which involves a 90^{o}_{x} pulse followed by successive 180^{o}_{y} pulses (Harris, 1983). As can be seen, the presat spectrum contains a very large baseline problem as compared to the CPMG spectrum. In a CPMG spectrum, there is typically intensity loss due to relaxation in the longitudinal (z) direction as the sequence of pulses is applied. This is naturally helpful in removing baseline, due to the fact that larger molecules also have very short T_1 times. So the intensities will first drop from these large molecules. One of the measures of the effectiveness of the baseline correction algorithm will be to compare it to a CPMG spectrum, since CPMG pulse sequences have been used by others (Van et. al, 2003) to remove lipid signals in metabolomics data in the past.

In this trial, the first 10% of the spectrum was removed from a 65535 data point FID. Again the backwards predictor was a simple AR model. An AR model can be described by the following equation:

$$x(n) = \sum_{k=1}^{p} a_k \cdot x(n-k) + w(n) \qquad \textbf{(3.4)}$$

This describes a process that can be modeled as a linear combination of past measurements of itself with the addition of white noise (Proakis and Manolakis, 1996). Additional complexities can be added such as a time delay term. In our trial, we used a second order AR model with a delay of one time step. The model parameters were estimated using a least squares algorithm. From Figure 3.9, we can see that this methodology does help with the baseline problem. However, there appears to be some frequency dependent phase distortions that are present, most probably caused by the backwards predictor. Further research needs to be done to look into the viability of this process and the removal of such distortions.

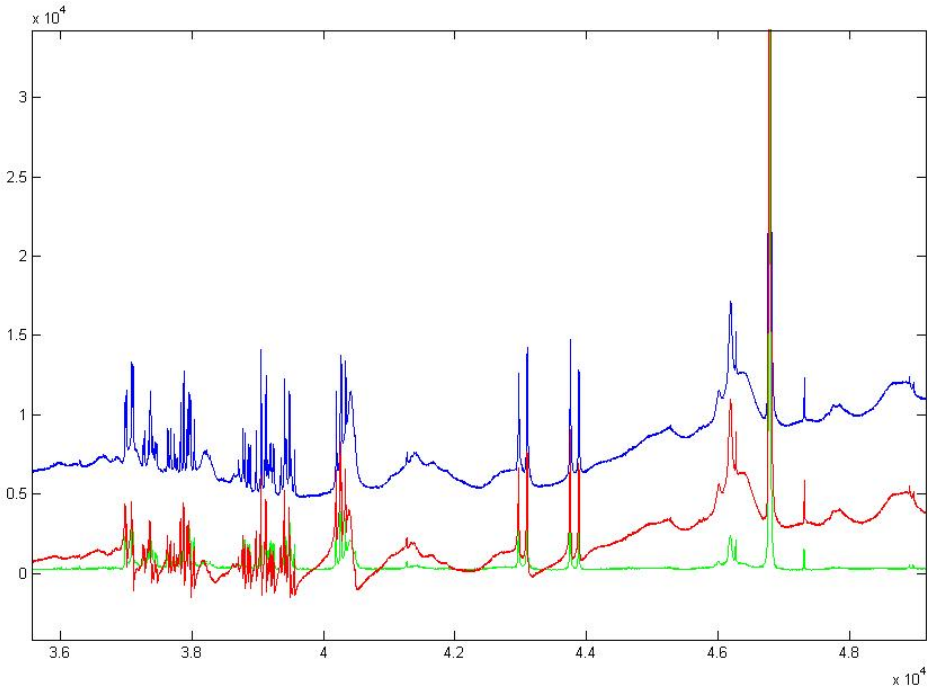

Figure 3.9 – Baseline correction scheme performed in the time domain. Blue is the preset experiment spectrum, red is the baseline corrected spectrum, and green is the same sample done under a CPMG experiment.

3.4. Conclusion

The three methods described in this chapter are distinct methodologies that are used to solve the issue of baseline distortions. The first of the three has the distinct advantage of being a general method of picking baseline points. The last two methods are more specific and used known mechanisms of baseline distortion to reproduce the baseline model. Both the automated baseline point algorithm, and lipid distortion beta curve methods work quite well, while more work needs to be done on the time domain method. Applying good baseline correction will clearly give better spectra for further use in statistical models and quantitative analysis.

4. Impact of Variable Reduction and Spectroscopic Distortions on Multivariate Statistical Models[2]

Chapter 2 describes in detail various types of preprocessing that can be done to NMR spectra for the purposes of variable reduction and Chapter 3 outlines several techniques that can be used to resolve baseline distortions. While these techniques help the overall quality of NMR spectra, the appropriateness and impact of such preprocessing techniques need to be explored in the context of metabolomics. The implications on building multivariate statistical models for the purposes pattern recognition and prediction will be addressed in this chapter.

Nuclear Magnetic Resonance (NMR) spectroscopy is a widely-used tool in the rapidly growing field of metabolomics, where the measurement of small molecule metabolites provides a chemical "snapshot" of an organism's metabolic state (Lindon et. al, 2003). NMR is inherently quantitative and non-selective and therefore a wealth of chemical information can be extracted from single NMR spectrum. Metabolomics studies often couple NMR spectral data with principal component analysis (PCA) and other pattern recognition techniques to uncover meaningful patterns in data sets (Holmes and Antti, 2002). Long-term goals of such computational model building include automation of data analysis as part of an integrated diagnostics platform (Wishart et. al, 2001) and personalized therapies (Clayton et. al, 2006). Building statistical models from NMR spectra can be problematic however, as spectral distortions present potentially confounding artifacts to techniques such as PCA (Defernez and Colquhoun, 2003, Halouska and Powers, 2006).

[2] Some of the material that appears in this chapter has been previously presented and published for the *Pacific Symposia on Biocomputing*. **2007**, 12, 115-126.

These distortions have an origin in the hardware (Siuda et. al, 1998), the type and nature of the sample, and choice of acquisition and processing parameters (Weljie et. al, 2006). For example, pre- and post-processing algorithms and the signal-to-noise (S/N) in the time domain will impact data quality. Metabolite signals in complex mixtures often span several orders of magnitude, thus requiring a significant dynamic range in the receiver. Furthermore, aqueous samples such as urine or plasma require suppression of the water solvent peak which is a 1000 times more concentrated than the metabolites of interest, resulting in distortions of the baseline and intensity of metabolite signals. Metabolites' resonance frequencies, lineshapes, and linewidths will vary between samples within an NMR metabolomics dataset irrespective of hardware considerations. Factors influencing these chemical modulations include sample pH, ionic composition, and inter-metabolite interactions (Lindon et. al, 2000). As a result, statistical analyses require some form of pre-processing or data reduction to ensure that the variables of interest are representative of the underlying chemical data (Webb-Robinson et. al, 2005).

In this chapter, the impact of spectral distortion on the quality of predictive statistical models built upon two alternative representations of preprocessed NMR data is assessed. A simulated dataset is used to model various types of spectral distortion in a systematic manner, and two techniques for dimensionality reduction, spectral binning and targeted profiling, are used to represent these simulated spectra. The results are assessed using the regression/classification extension of PCA, partial least squares for discriminant analysis (PLS-DA) (Umetrics, 2001). We validate our findings using a real-world data set of rat-brain extracts.

4.1. NMR Data Representations

In Chapter 1, a description of an NMR experiment was given. From the viewpoint of a single proton, the equation for an NMR experiment was also outlined. (See Equation 1.1). In Chapter 2 we describe a more complex situation of multiple protons that interact on a given molecule, which is modeled by a spin simulation. Finally in this chapter we add another layer of complexity where by different molecules, which many contain multiple

protons, are all present in a complex mixture. Couple that level of complexity with the fact that in metabolomics studies, one has many NMR experiments of different mixtures; clearly therefore the data that is required for metabolomics is quite complex and must be processed properly.

For a mixture, an NMR spectrum can be viewed as a linear combination of characteristic signals for each compound that is present in a given sample. As the concentration of a particular compound changes, the characteristic signal for that compound responds in a linear fashion. Thus, an NMR spectrum can be viewed from a theoretical perspective as follows:

$$\underset{[1\times n]}{d_{obs}} = \underset{[1\times k]}{c} \cdot \underset{[k\times n]}{\left[a \otimes s\right]} + \underset{[1\times n]}{u} + \underset{[1\times n]}{\varepsilon} \tag{4.1}$$

where d_{obs} is a [1xn] vector of the observed NMR data, c is a [1xk] vector representing the concentrations of k known compounds in the mixture, and s represents a matrix of the spectral signatures present in the solution. The variable a is a *spectrum calibration function* that is applied to each row of s to account for changes in the sample's pH, ionic strength, etc. While the variable u represents unknown contributions to the signal from unknown metabolites, lipoproteins, or any other contributions to the signal that are not explicitly modeled using s. Finally, the observed spectrum contains noise that is introduced by the NMR hardware and processing algorithms, ε.

4.1.1. Spectral Binning

Spectral binning (Holmes and Antti, 2002) is a widely-used technique where the spectrum is subdivided into a number of regions, and the total area within each bin is used as an abstracted representation of the original spectrum. The area encapsulated by a bin would ideally capture all of the area associated with a given resonance across all spectra in the dataset, thereby mitigating the effect of minor peak shift and line width variations for a compound across samples. A typical 64k NMR spectrum would be reduced using bin widths of 0.04 ppm, resulting in ~250 bin integral values. Spectral binning is agnostic of the underlying generative model described in Equation 4.1,

however it is commonly used due to the ease of implementation and complete spectral coverage.

4.1.2. Targeted Profiling

Targeted profiling (Weljie et. al, 2006) is a technique that leverages a reference spectral database to directly recover the concentration matrix *c* from Equation 1, which is then used as the input to pattern recognition techniques such as PCA or PLS-DA. Targeted profiling can be viewed as a method of recovering the latent variables in the form of underlying metabolite concentrations that generated the observed spectral data. Because of its reliance on a spectral database *s,* targeted profiling does not directly model or deal with the unknown term *u* in Equation 4.1. Since *u* may contain potentially important latent chemical information, it can be calculated directly as the residual from Equation 4.1, and spectral database-agnostic techniques such as spectral binning can be applied to *u* for subsequent analysis.

4.2. Methods

4.2.1. Synthetic Study

Several synthetic data sets were generated with specific characteristics to simulate, in a systematically controlled manner, some of the key challenges inherent in working with NMR data. The data for the synthetic study was generated using Chenomx NMR Suite 4.5 (Chenomx Inc., Edmonton, Alberta, Canada) compound database entries. Varying mixtures of twenty compounds, with the addition of DSS at 0.5 mM, were simulated. Compound concentrations for the following compounds were sampled randomly from a normal distribution: 2-oxoglutarate, acetate, acetone, alanine, betaine, carnitine, citrate, creatine, dimethylamine, fumarate, glucose, lactate, maleate, myo-inositol, taurine, tryptophan, tyrosine, urea, π-methylhistidine, τ-methylhistidine. Biologically viable population statistics of mean and standard deviation were used for each compound (Slupsky et. al, 2007) and these concentrations remained fixed from simulation to simulation.

Random uncorrelated noise was added to each spectrum in the frequency domain. Each spectrum was generated to have an equivalent amount of noise by an approximate signal to noise ratio (SNR) of 100:1.

The effect of pH variability was simulated by randomly varying compound resonance frequencies within an empirically validated range. This range reflects the compound's NMR frequency response to pH levels ranging from pH 4 to 9 as determined from pH curves of pure reference spectra. The magnitude of this range was controlled to test the effects of pH variation via a transform fraction parameter. A fraction of 1.0 allowed clusters to be transformed over the entire pH 4 to 9 range, while a fraction of 0.1 would allow for clusters to be transformed over 10% of the range, centered at pH 7.0. The actual pH range that this represents will be different for each compound depending on the relative pH sensitivity of the compound near pH 7.0.

In order to generate two classes of spectra, the population statistics of one or more metabolites were changed for each simulation. The parameters used in each simulation are outlined in Table 4.1.

Table 4.1 – Simulation Parameters for Synthetic Study.

Simulation #	Parameters	Value
1	Number of Files	200 (100 of each class)
	SNR	100
	Transform Fraction	0.1
	Group 1 Citrate/Tryptophan Mean ± Stdev (μmol)	2318 ± 1496 / 5 ± 2
	Group 2 Citrate/Tryptophan Mean ± Stdev (μmol)	1031 ± 945 / 10 ± 2
2	Number of Files	200 (100 of each class)
	SNR	100
	Transform Fraction	0.1
	Group 1 Maleate Mean ± Stdev (μmol)	30 ± 15
	Group 2 Maleate Mean ± Stdev (μmol)	60 ± 20
3	Number of Files	200 (100 of each class)
	SNR	100
	Transform Fraction	1
	Group 1 Citrate/Tryptophan Mean ± Stdev (μmol)	2318 ± 1496 / 5 ± 2
	Group 2 Citrate/Tryptophan Mean ± Stdev (μmol)	1031 ± 945 / 10 ± 2

4.2.2. Rat Brain Extracts

This real-world dataset is based on a previously published (McGrath et. al, 2006) dataset and was kindly provided by Dr. Brent McGrath and Dr. Peter Silverstone (Department of Psychiatry, University of Alberta). Twelve adult male Sprague-Dawley rats brains were dissected into frontal (fcx) cortex, temporal cortex (tcx), occipital cortex (ocx) and hippocampus (hipp) regions according to stereotaxic demarcation (McGrath et. al, 2006). For spectral binning, bins widths of 0.04 ppm were used, with the following dark regions defined: DSS (the internal standard): -0.1-0.1ppm, 0.6-0.7 ppm; methanol (a byproduct of the extraction process): 3.33-3.37 ppm; water: 4.5-5.5ppm; imidazole (the pH indicator): 7.13-7.5, 7.82-8.68 ppm.

The following compounds were identified and quantified using the targeted profiling technique (Weljie et. al, 2006) as implemented in Chenomx NMR Suite 4.5:

4-Aminobutyrate	Formate
Acetate	Hypoxanthine
Adenosine	Isoleucine
Alanine	Lactate
Aspartate	Leucine
Betaine	Lysine
Choline	Methanol
Citrate	N-Acetylaspartate
Creatine	Serine
Creatinine	Succinate
Formate	Taurine
Fumarate	Threonine
Glutamate	Tyrosine
Glutamine	Valine
Glycerol	Xanthine
Glycine	

4.3. Multivariate Statistical Modeling

All multivariate modeling was performed using SIMPCA-P+ 11.0 from Umetrics Inc. The following equations describe the statistics used to measure the quality of the models generated:

$$R_x^2 = 1 - \frac{\sum_{i=1}^{m}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{m}(x_i - \bar{x})^2} \quad \textbf{(4.2)}$$

$$R_y^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \quad \textbf{(4.3)}$$

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \textbf{(4.4)}$$

R^2_X and R^2_Y are calculated as the fraction of the sum of squares of all X and Y that the model can explain using the latent variables. Q^2 is the fraction of the total variation in Y that can be predicted ($\hat{Y}$) using the model via seven-fold cross-validation. Here m is the total number of observations in the model building set, n is the total number of observations in the hold out set, and $\bar{X}$ and $\bar{Y}$ represent the average over all observations in that set. Note that results shown in this chapter are also cumulative R^2_X, R^2_Y, and Q^2 values, which is the average of all 7 cross-fold validation sets.

Validation of the models was done using Permutations Tests. This test randomly permutes the class labels and builds new models based on these permuted labels using the same number of components. Permutations tests were performed using 100 permutations. Insignificant differences in the model quality parameters R^2_Y and Q^2 in the true model to the permuted models indicates that the true model was overfit with the currently available data and the number of components chosen.

4.4. Results

4.4.1. Synthetic Data

By systematically varying key properties of the synthetic data sets, several aspects of building statistical models on NMR data representations were assessed. The first issue assessed was the effect of noise on the spectra. Specifically, noise was added to the spectrum to see how robust both spectral binning and targeted profiling methods were at being able to recover the latent information in the data in the presence of noise. What was observed was that if the noise was completely uncorrelated, then both methods are very robust to varying noise levels.

The next issue we examined was the choice of variable scaling and normalization methods, since this can have a large impact on the quality of results obtained from multivariate statistical methods such as PLS-DA. Normalization for all spectral binning data was to the total area of the NMR spectrum. No normalization was necessary for the targeted profiling results, since direct quantification can be obtained with the addition of an internal standard. Both the spectral binning data and targeted profiling data were mean centered and were scaled using unit variance (UV) or Pareto scaling. UV scaling involves weighting each of the variables by the variables' group standard deviation, and has the advantage of not biasing statistical models towards large concentration compounds or high area bins. Pareto scaling involves weighting each of the variables by the variables' group variance, which minimizes the impact of noise. Data from simulation #1 was used to evaluate the effects of these two scaling procedures. This simulation encoded class differentiation through citrate, present at relatively high concentrations, and tryptophan, present at relatively low concentrations. Figure 4.1a demonstrates that PLS-DA on UV scaled data can recover differences in both tryptophan and citrate, while the loadings plot of Pareto-scaled data (Figure 4.1b) is only able to distinguish the intense citrate signal. UV scaling was superior to Pareto scaling in recovering a model that accurately reflected the variables of interest (both low- and high-concentration metabolites) for targeted profiling data.

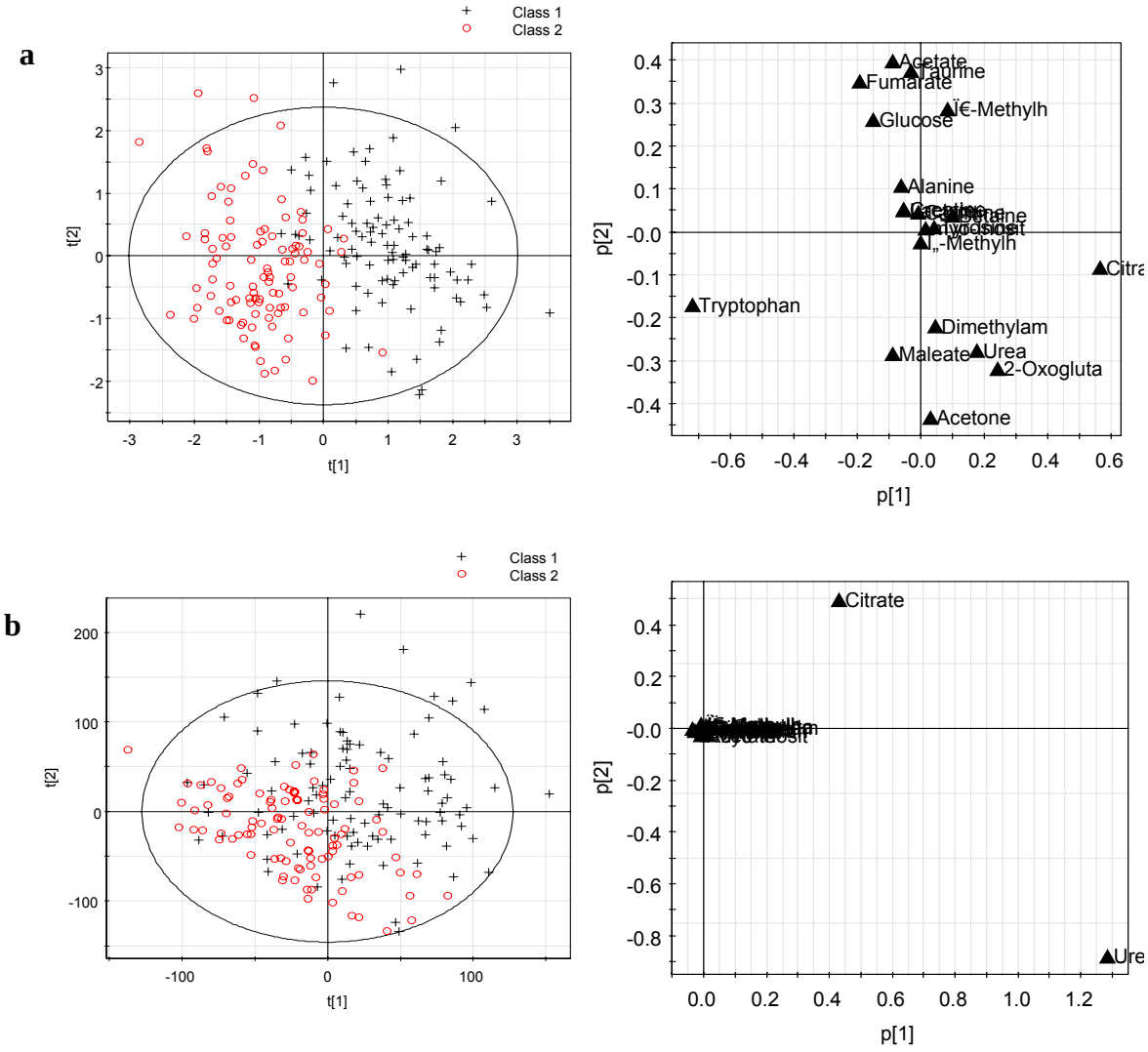

Figure 4.1 – PLS-DA models of simulation #1 (scores plot left, loadings plot right), showing targeted profiling data using a) unit variance scaling b) Pareto scaling.

Overlap of NMR resonances from different metabolites is another issue hampering the analysis of complex biofluid spectra. Further complications arise from compound overlap with dominant peaks such as urea, where low intensity peaks are often lost in traditional analyses due to the overwhelming magnitude of the urea signal. Simulation #2 generated a dataset in which a single metabolite, maleate, differentiates the two classes and overlaps with the high concentration urea signal, which varies randomly (i.e. urea does not encode class discrimination). Figure 4.2 shows the scores, loadings, and permutations tests for

spectral binning and targeted profiling methods. One can see from the loadings plot in Figure 4.2b, that targeted profiling methods identify maleate as a significant metabolite even under severe overlap conditions, while spectral binning shown in Figure 4.2a fails to distinguish the maleate bin as being significant. Spectral binning is also prone to generating highly overfit models as shown by the permutation test in Figure 4.2, whereas targeted profiling models show no signs of overfitting. Permutation tests help assess overfitting by randomly permuting class labels and refitting a new model with the same number of components as the original model. An overfit model will have similar R^2 and Q^2 to that of the randomly permuted data. Well fit models will have R^2 and Q^2 values that are always higher than that of the permuted data.

Sample matrix conditions such as pH and ionic strength can have profound effects on metabolites' NMR resonance frequencies. These shifts can directly influence the quality of the models that are generated using NMR data, and were modeled with simulation #3. Both spectral binning and targeted profiling gave rise to models that were able to separate the data in the latent variable space. However, the quality of the model generated with the spectral binning data was low and resulted in overfitting as shown in permutation plots (Figure 4.3). This is due to the large number of variable weights used in the loadings (Figure 4.3a). A large number of variables share similar weights because the same significant resonances are now migrating over adjacent bins due to pH/ionic strength variation. Models built on targeted profiling data, which accounts for the shifts in resonance locations directly in the modeling process, are able to separate the two groups and do not overfit the data (Figure 4.3b).

The final effect studied is the impact of limited sample sizes on predictive capacity, a typical problem in metabolomics studies. The effect of sample size was shown using a subset from Simulation #3. The size of the dataset was reduced from 100 to 20 samples in each class. Even with a limited sample size, the targeted profiling approach resulted in well fit PLS-DA models, as assessed by the permutations tests (Figure 4.4b). While the descriptive features of tryptophan and citrate are not as clearly distinguished in the loadings plot, the permutation plot indicates that even with a small number of samples the

data is not overfit. The results for spectral binning, however, are quite deceptive, as the PLS-DA model shows very good separation of classes in the scores plot (Figure 4.4a). However, the model generated has an extremely high degree of overfitting – the majority of the randomly permuted models generate Q^2 values higher than that of the non-permuted model (Figure 4.4a).

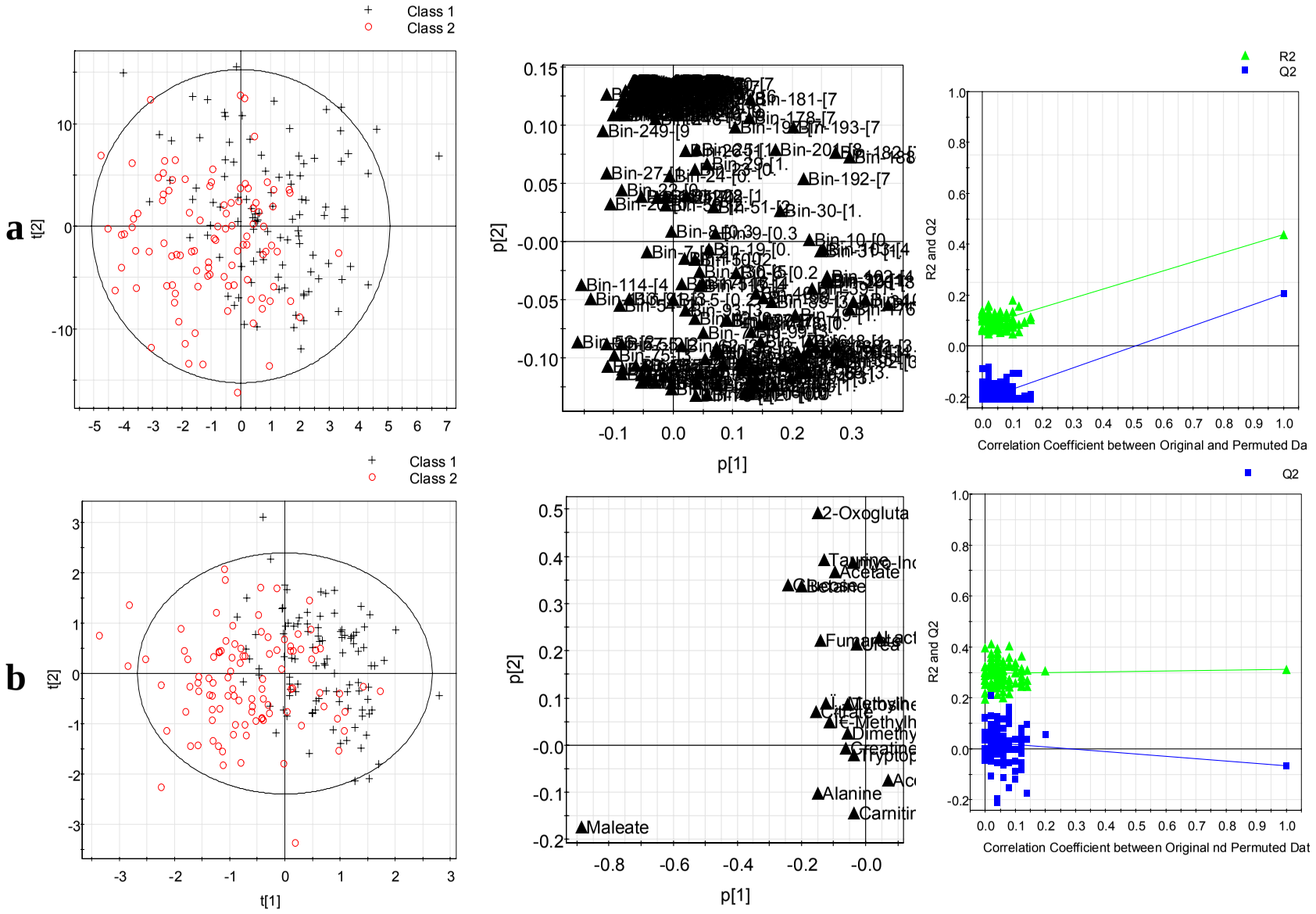

Figure 4.2 – PLS-DA models (scores plot left, loadings plot center, permutations plot right) for a) spectral binning and b) Targeted Profiling methods under conditions of highly overlapping clusters (simulation #2).

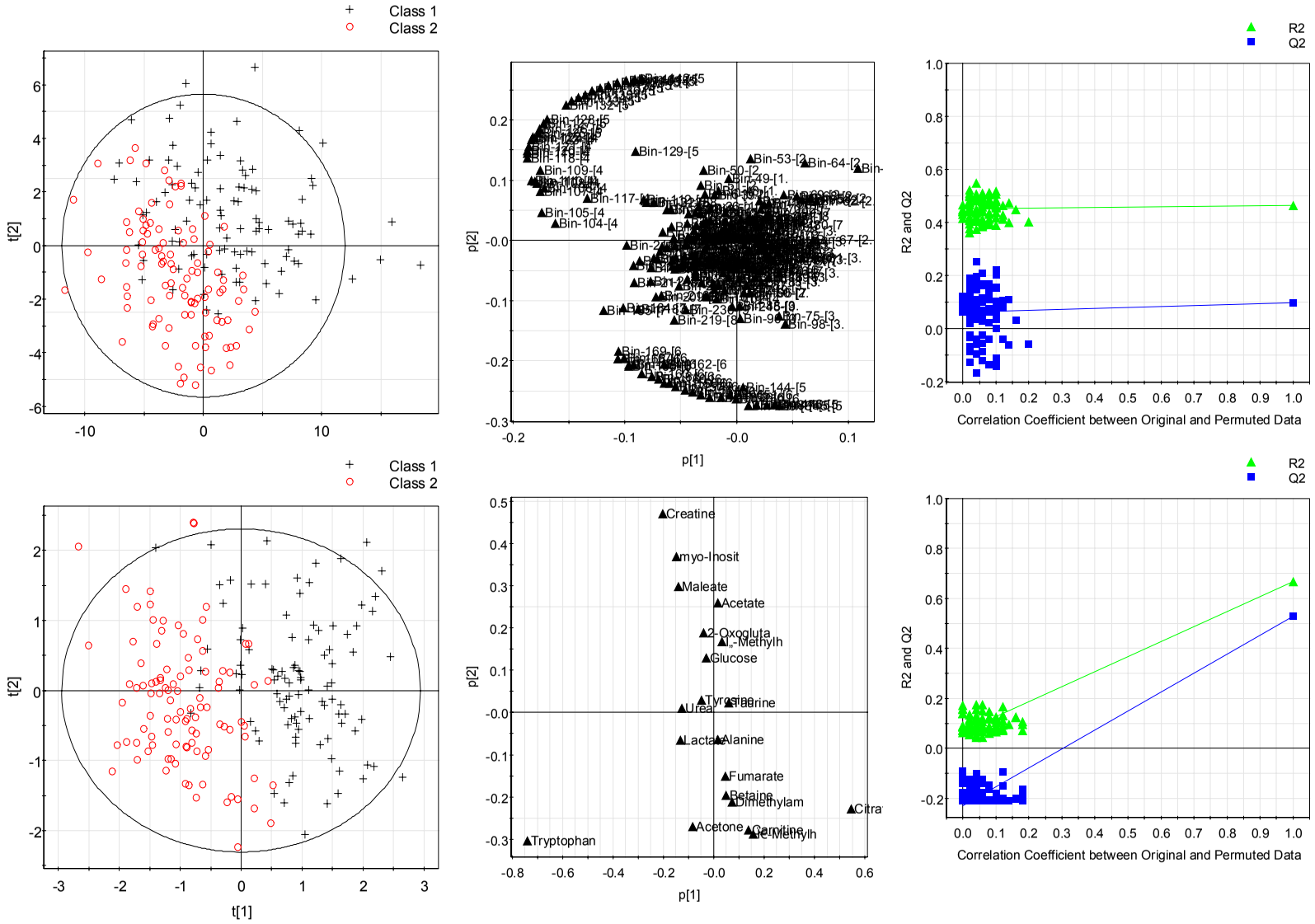

Figure 4.3 – PLS-DA models (scores plot left, loadings plot center, permutations plot right) for a) spectral binning and b) Targeted Profiling methods under conditions of varying pH (simulation #3).

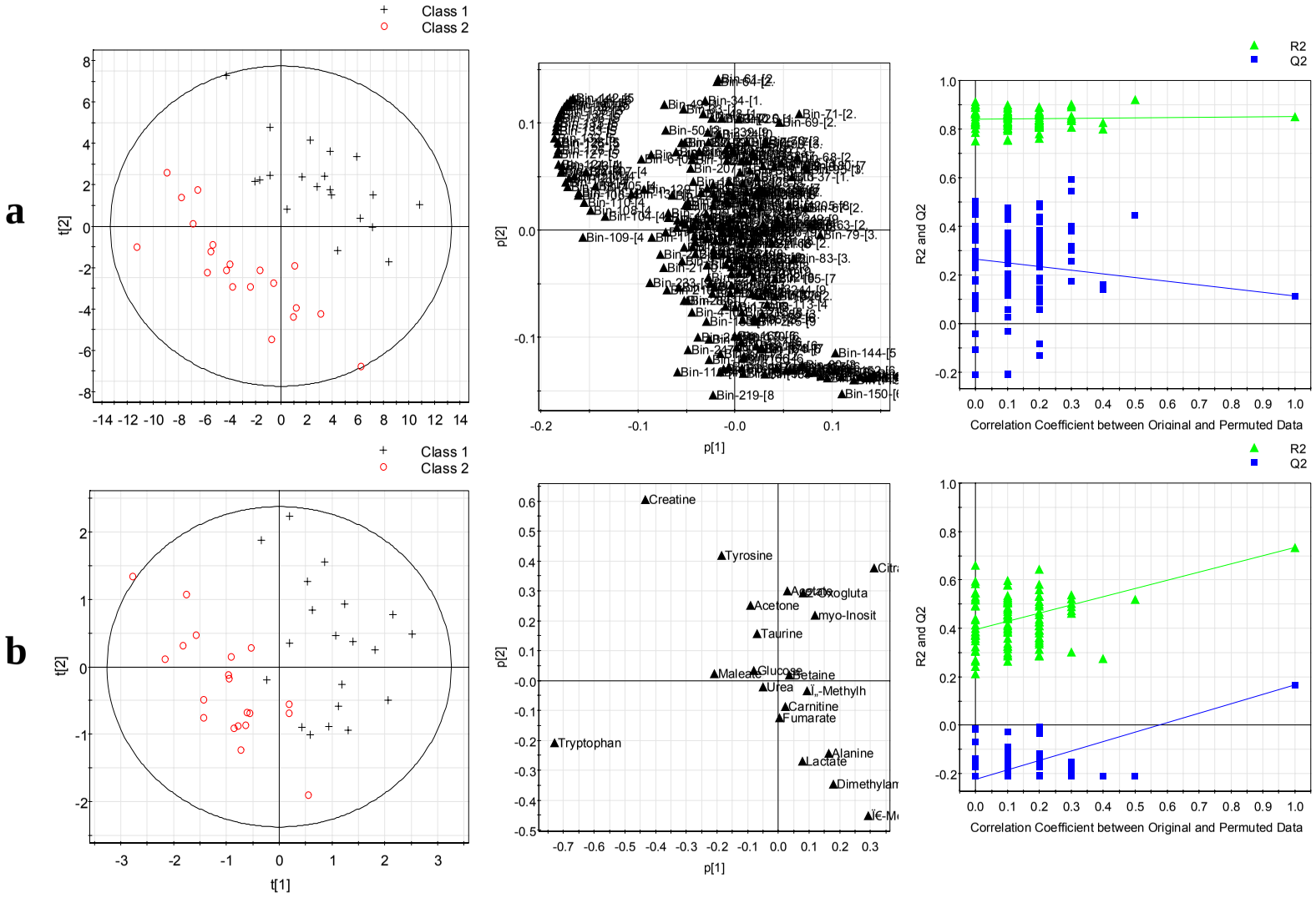

Figure 4.4 – PLS-DA models (scores plot left, loadings plot center, permutation plot right) for a) spectral binning and b) targeted profiling methods under conditions of varying pH and low sample size (simulation #3).

4.4.2. Rat Brain Extract

The rat brain extract dataset is a real-world dataset that exhibits many of the phenomena we have seen in the synthetic dataset. The spectra contain noise, have metabolite resonances that shift due to pH, and have low-concentration metabolites that are important in differentiating the different brain regions, thus making it a suitable model dataset to validate our findings from the synthetic dataset. This dataset was acquired at high resolution (800MHz) and contains ~30 NMR-visible compounds. We did not find that the choice of variable scaling affected the quality of the generated models for this dataset. We therefore used unit variance scaling for the results shown below.

We found that using spectral binning generated a model with slightly lower predictive accuracy than targeted profiling data: Q^2 for spectral binning was 0.468, whereas Q^2 for targeted profiling was 0.522.

As in our synthetic dataset, we found that spectral binning-based results were prone to overfitting. To test for overfitting, we randomly permuted the class labels for the PLS-DA analysis 100 times. With the spectral binning dataset, we found that some of the models generated with random permutations of the data had higher Q^2 and R^2 values than the non-permuted data. This is illustrated in Figure 4.5a. Internal validation of the model based on the targeted profiling representation of the NMR data do not exhibit the characteristics of an overfit model like that found in the spectral binning model, as shown in Figure 4.5b. The targeted profiling representation uses only 27 variables to represent the latent information in the dataset, thereby restricting the degrees of freedom available in the construction of a model, and reducing the capacity of the model to overfit the data.

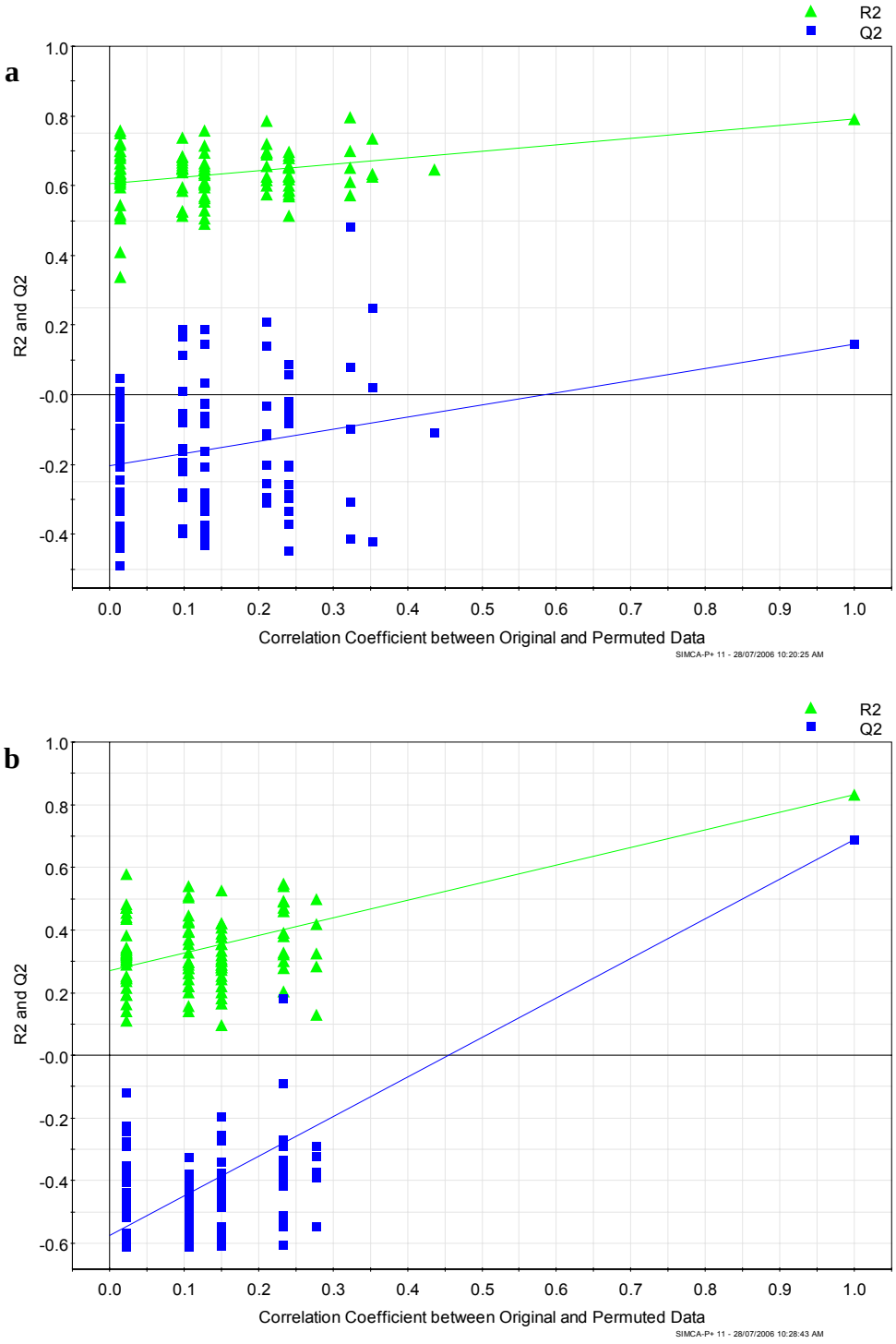

Figure 4.5 – a) Internal validation of spectral binning, showing clear evidence of overfitting with random permutations of the data generating better R2 and Q2 values than the non-permuted data. b) Internal validation of targeted profiling, showing clear decrease in performance on permuted data.

4.5. Conclusion

This chapter describes how the inherent properties of NMR spectroscopy can impact the predictive ability of models built upon spectral binning and targeted profiling representations of NMR data by using a novel method for synthetically generating NMR spectra. The quality of predictive models built was quantitatively assessed, as was the relative robustness of these two methods. Under the experimental design chosen, both methods are very robust with respect to noise. In contrast, variable scaling methods can affect both the quality and interpretability of the models generated. For targeted profiling data, unit variance scaling generates a more robust data representation. Targeted profiling was also found to be an effective dimensionality reduction technique that, overall, is more robust with respect to spectral distortions and high dynamic range metabolites than spectral binning, and is less prone to overfitting than spectral binning models. These findings were validated on a real-world dataset of rat-brain extracts consisting of ~30 NMR detectable metabolites, in which statistical models were less prone to overfitting based on a spectral profiling representation of the data. Spectral binning is a common method for data reduction due to the speed of analysis, while current targeted profiling implementations require interactive input and are relatively time-intensive. While the rat-brain extract study represents a relatively simple dataset, targeted profiling has successfully been applied to extensive studies of serum (Weljie et. al, 2007) and urine (Slupsky et. al, 2006). As increasingly automated methods for quantitative profiling of NMR data become available, we expect database-driven targeted profiling to become the data-reduction method of choice.

5. Characteristics of Targeted Profiling Data and the Implications for Extraction of Useful Biological Information

The topics discussed in previous chapters have been in the preprocessing and assessment of appropriate data representations of NMR metabolomics data. Given a biological fluid such as urine or serum, one can use NMR to arrive at a simple data representation that contains information about the metabolic content of the biological fluid. Using an appropriate experimental design, one can obtain a set of NMR data from a set of biological fluids. The data for a set of experiments is what is used as a basis for multivariate statistics and chemometrics. With Targeted Profiling data, the data representation is simply the concentrations of metabolites as extracted from the NMR spectrum. Having concentrations allows for more creative uses of this data, and allows for better extraction of information from a dataset. This chapter will outline the characteristics of Targeted Profiling data and how this data can be used to recover relevant biological information of the system that is being studied.

5.1. Metabolite Concentrations

One of the most simple and obvious uses of Targeted Profiling data is to look at metabolite concentrations directly. With a certified internal standard, Targeted Profiling will give accurate concentrations of free metabolites within a biological fluid. Using these concentration values, one can assess the metabolic state of a patient. Normal ranges of certain metabolites are well known to physicians. These ranges are often only accurate on specific analytical platforms and are not a measure of the absolute concentration of a particular metabolite within a biological fluid. However, using NMR and Targeted Profiling results, new ranges can be easily formed as a basis for normal ranges for a variety of metabolites. Figure 5.1 shows the metabolite levels in urine of three individuals. The five metabolites measured are tracked over a 30 day period. It can be seen in Figure 5.1 that metabolite levels are quite similar between different ages.

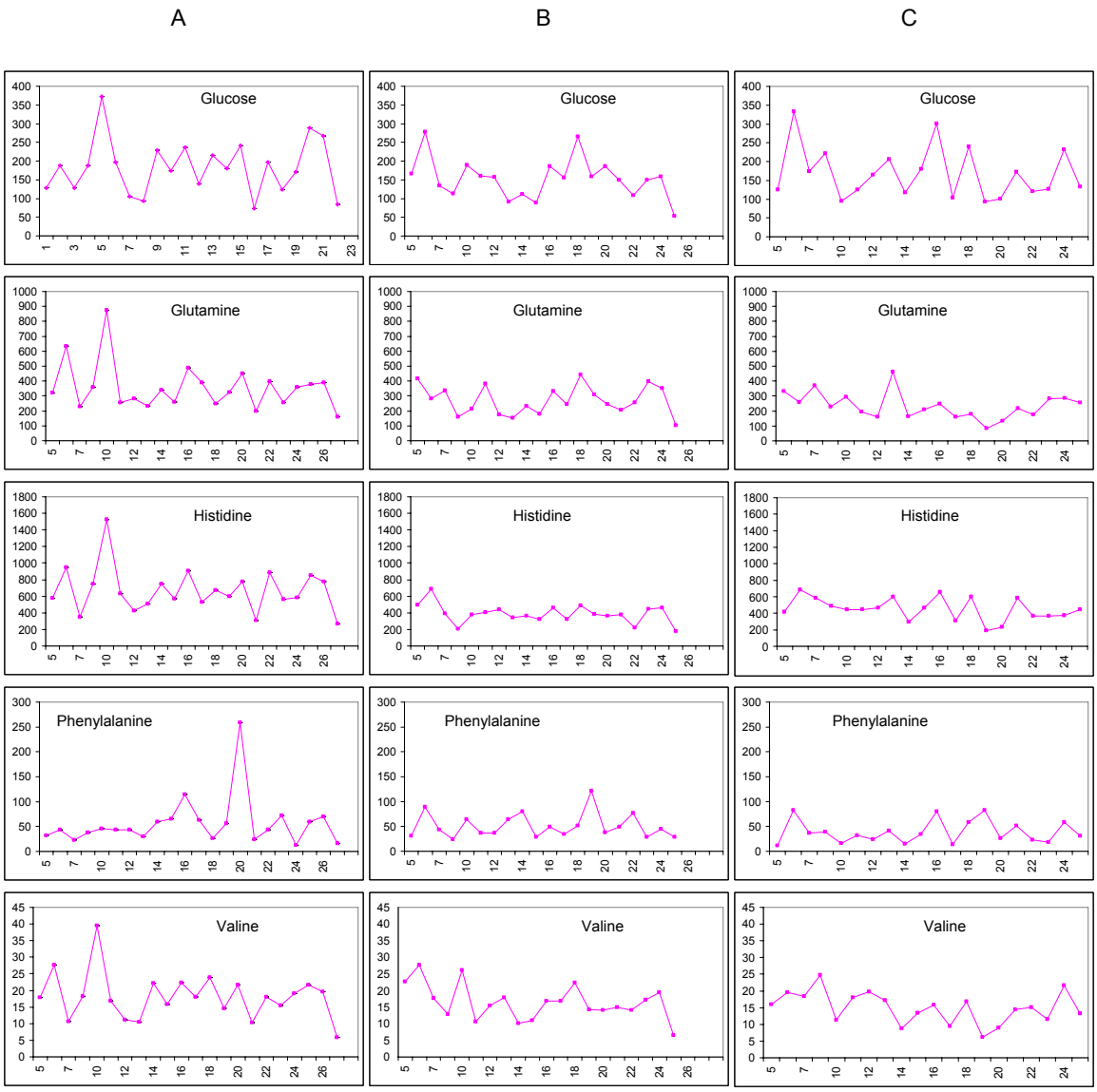

Figure 5.1 – Daily metabolite concentrations for 3 representative healthy men aged (A) 35 years, (B) 34 years, and (C) 20 years from morning urines. While daily fluctuation can be observed, it is clear that metabolite homeostasis is well regulated within specific ranges for these compounds

Using metabolite concentrations, one can compare differences between groups. This next example is a comparison between 21 patients that were admitted with acute asthma and 5

normal patients. The results of the NMR based metabolite measurements of their urine can be found in Figures 5.2 and 5.3.

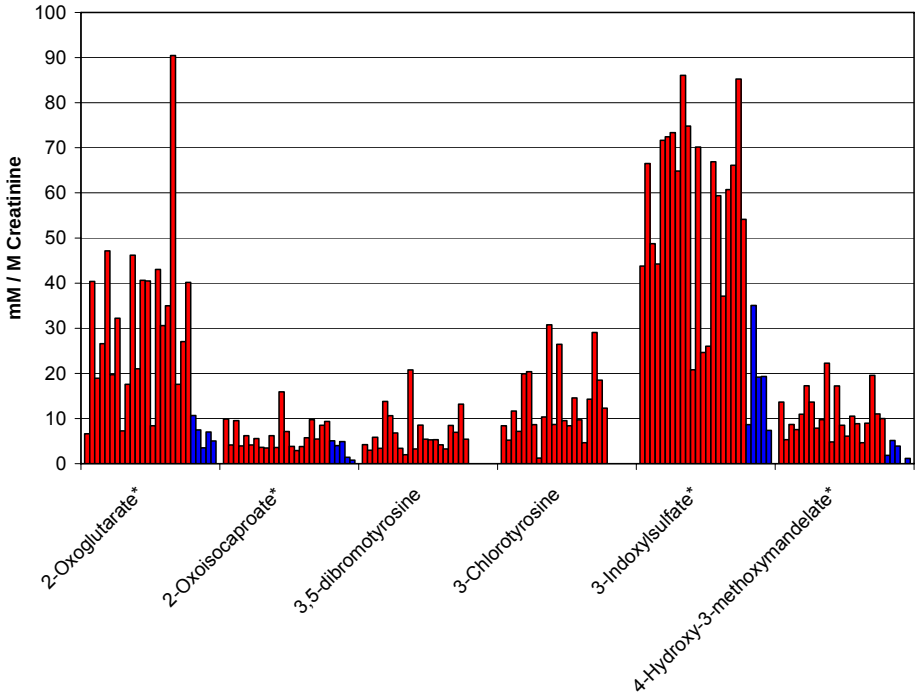

Figure 5.2 – Metabolite measurements (Set 1) of Asthma (red) and normal (blue) patients

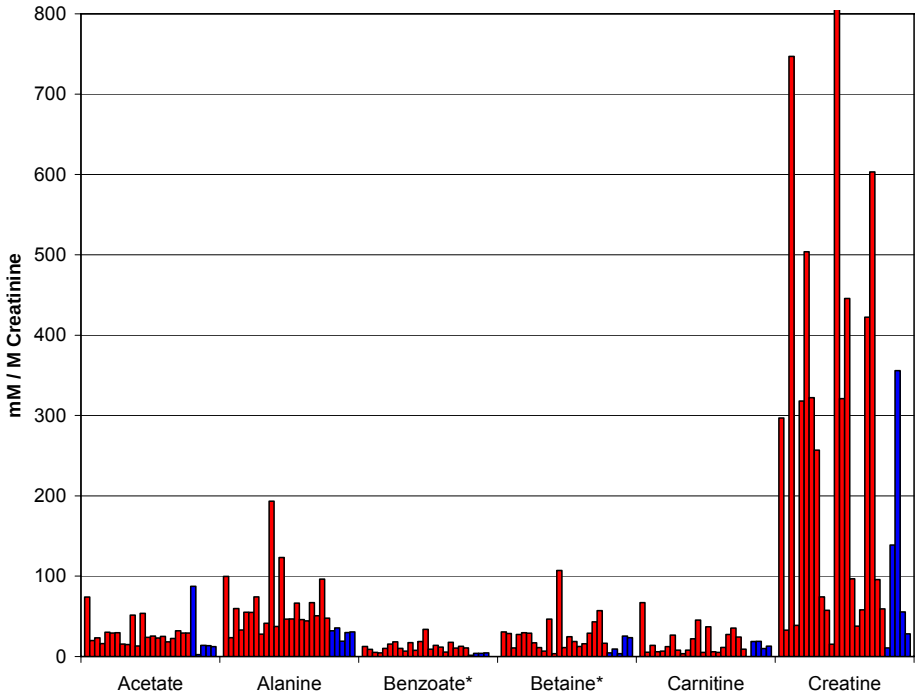

Figure 5.3 – Metabolite measurements (Set 2) of Asthma (red) and normal (blue) patients

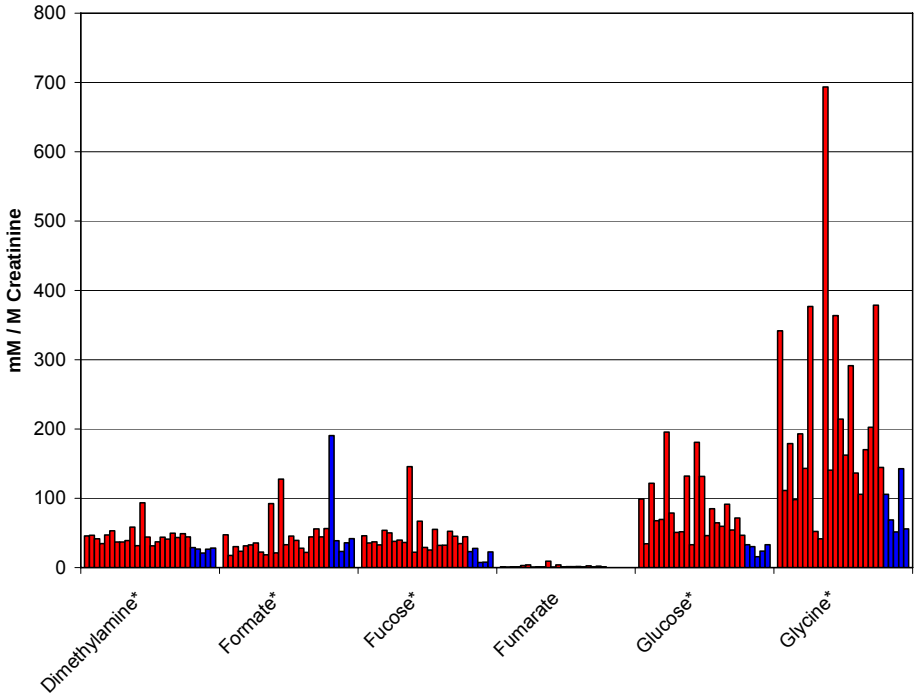

Figure 5.4 – Metabolite measurements (Set 3) of Asthma (red) and normal (blue) patients

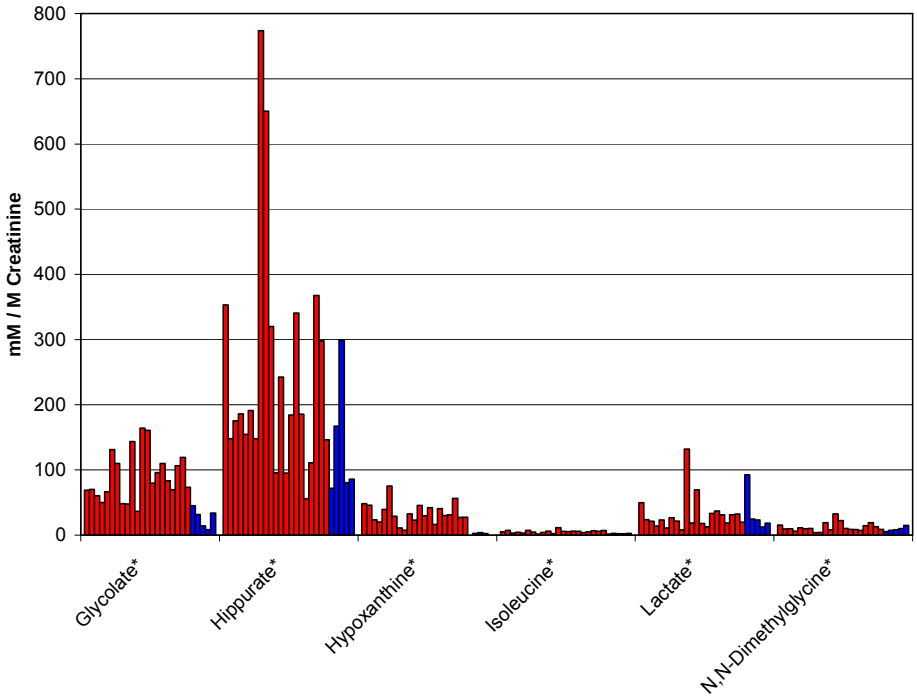

Figure 5.5 – Metabolite measurements (Set 4) of Asthma (red) and normal (blue) patients

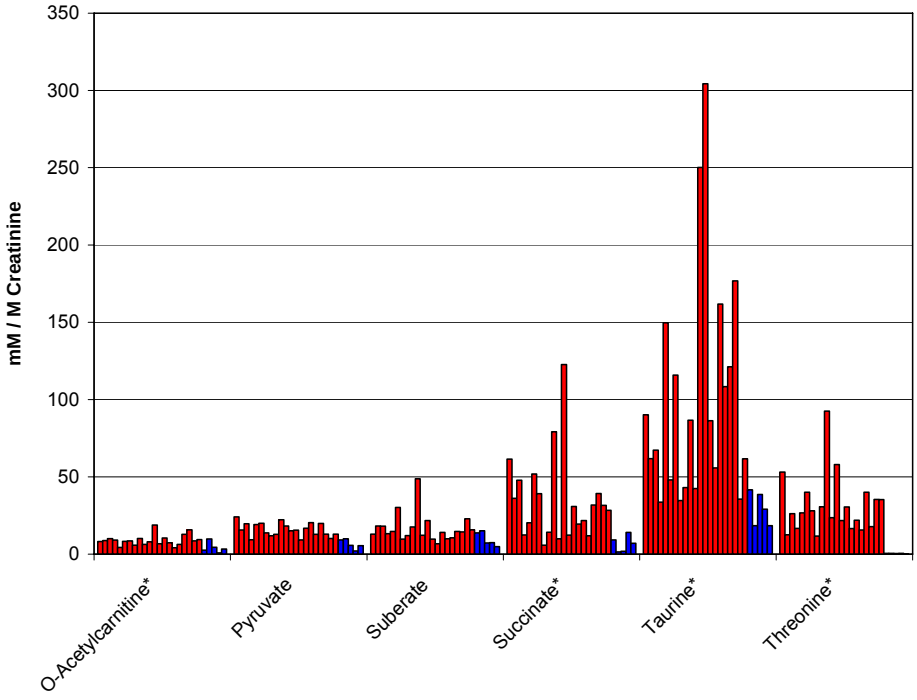

Figure 5.6 – Metabolite measurements (Set 5) of Asthma (red) and normal (blue) patients

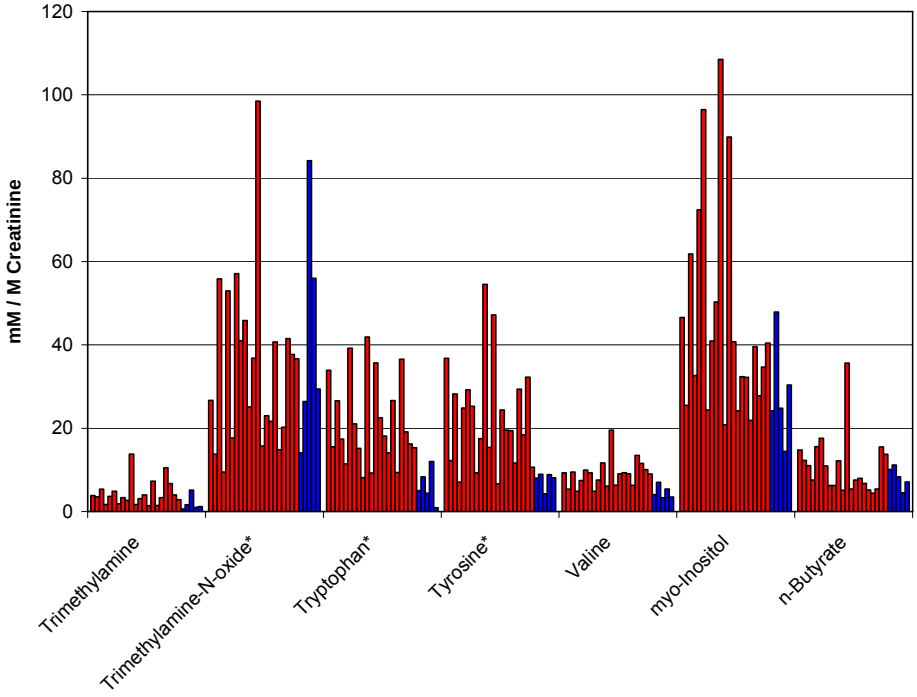

Figure 5.7 – Metabolite measurements (Set 6) of Asthma (red) and normal (blue) patients

As can be seen in Figures 5.2 to 5.7, there does appear to be group differences in metabolite concentrations for patients with acute asthma and those that are normal. Metabolites such as 2-oxoglutarate, succinate, 3-indoxylsulfate, 4-hydroxy-3-methoxymandelate all appear to increase in patients with acute asthma. While metabolites such as 3,5-dibromotyrosine and 3-chlorotyrosine appear in less than detectable levels in normal patients and are highly elevated in acute asthma patients.

5.2. Analysis of Variance

In the previous section, the example given by the asthma and normal patients show some differences between the two groups metabolite levels. This difference can be seen graphically. In order to test this hypothesis and test the statistical significance of differences between groups, one can perform an Analysis of Variance (ANOVA) test. An ANOVA test is simply a t-test when comparing two groups. To illustrate this, the next example is again of patients admitted with an acute asthma exacerbation. Urine samples were collected at the time of hospital admittance and again at a three-week follow-up visit (during which the corticosteroid Prednisone was prescribed). These two datasets were also compared with normal patients without asthma. We must also note that when testing multiple hypothesis of significance on the same dataset as there are multiple measurements of metabolites, it is likely to obtain false positive results of significance. A simple yet strict correction would be to apply a Bonferroni correction. Other methods and techniques can be used to help reduce the chances of false positives. In Figure 5.8 we show only four metabolites that were tested for significant differences between normal (blue), asthma patients before treatment (red), and asthma patients after treatment (green).

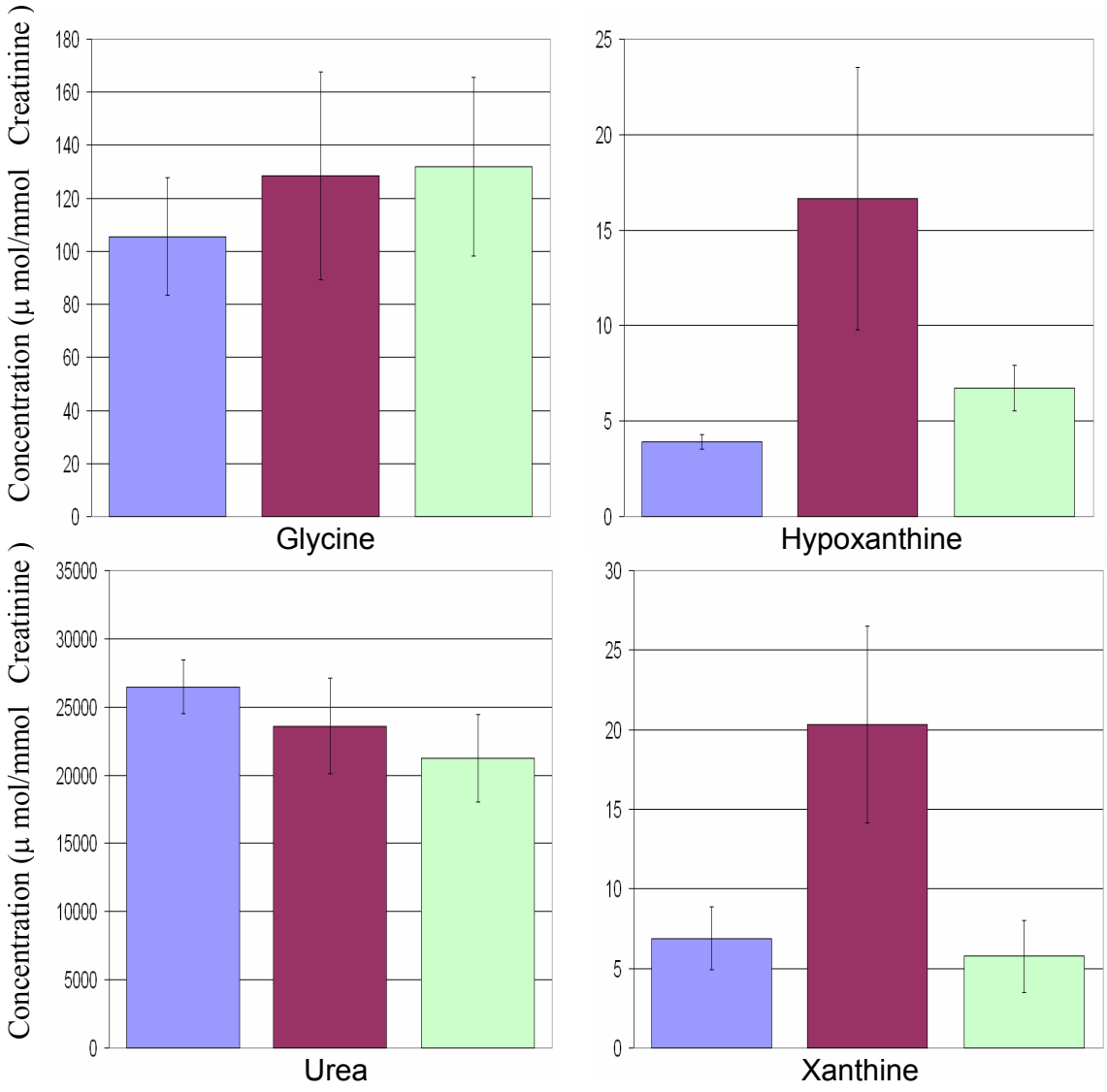

Figure 5.8 – Metabolite measurements of normal (Blue), asthma patients before treatment (purple), and asthma patients after treatment (green)

Shown in Figure 5.8 are two compounds Glycine and Urea, that showed no statistical difference between pre-treatment, treated, and normal patients. Two compounds that did show statistical significance of $\alpha<0.001$, were hypoxanthine and xanthine. From Figure 5.8 we can see that both hypoxanthine and xanthine are elevated in asthma patients compared to normal patients. As well, both levels of hypoxanthine and xanthine decrease to normal values after treatment. Looking deeper at the metabolic pathway of

hypoxanthine and xanthine could lead to explanation of the drug's action. In Figure 5.9, we show two possible pathways for the decomposition of hypoxanthine to xanthine. The first pathway is found in purine metabolism pathway. The second is an extracellular reaction caused by reactive oxidants. (Marnett et. al, 2003) Reactive oxygen species produced by activated immunological cells (NADPH oxidase) may increase production of hypoxanthine and xanthine. $O_2^{\cdot}$ produced by xanthine oxidase is not lipid soluble and cannot diffuse far from point of synthesis, but may be used by EPO to continue oxidative damage during pulmonary reperfusion.

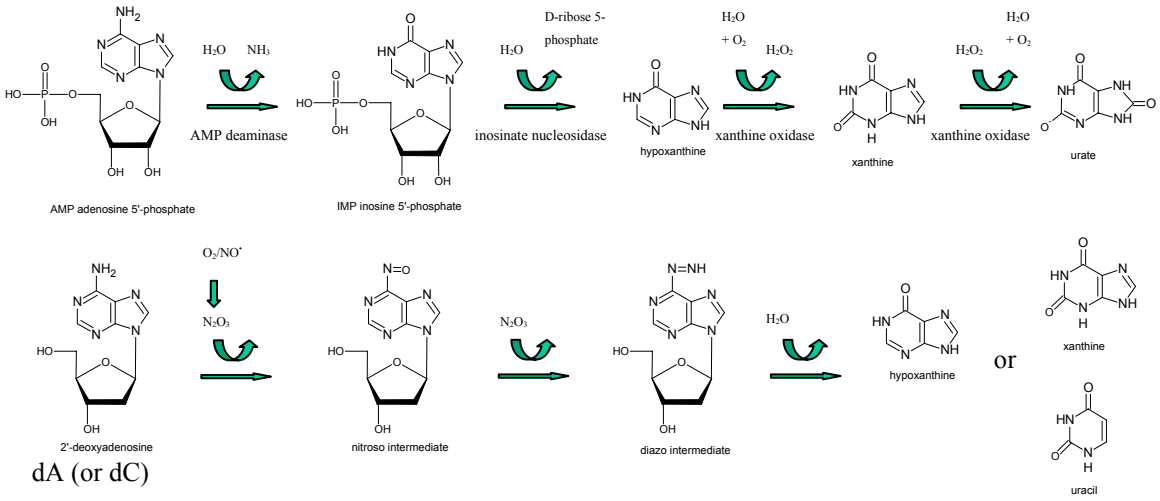

Figure 5.9 – Possible pathways for hypoxanthine and xanthine relationship

5.3. Metabolite Concentration Distributions

One consideration for performing ANOVA and other statistical tests is that the variables must follow a normal distribution. To show the distribution of metabolite concentrations, Targeted Profiling was performed on NMR spectra of urine samples from 59 normal patients. A total of 81 metabolites were identified and quantified. Histograms for each metabolite were plotted. It was clear that the concentrations did not follow a normal distribution. A log transformation was performed to data so the data fit a normal distribution more closely. Figure 5.10 shows the histograms of a few compounds both before and after log normalization.

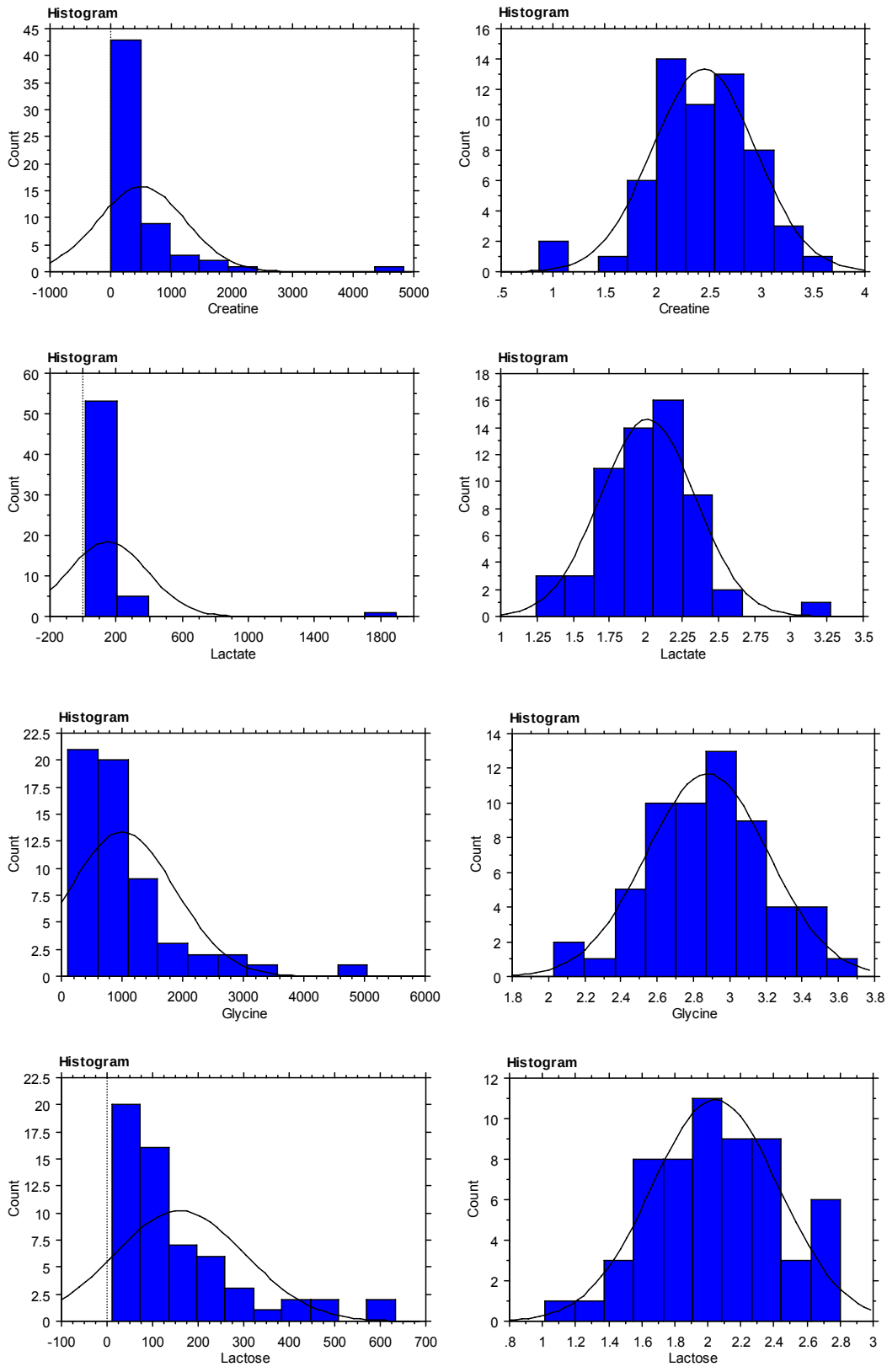

Figure 5.10 – Histograms of select metabolites before and after log transformation.

From Figure 5.10 we can see that in the population distribution of Creatine, Lactate, Glycine, and Lactose concentrations before log transformation is skewed. This is not surprising as there is a hard constraint to one side of the population (concentrations cannot be negative). The population distribution after performing a log transformation is much more normally distributed. (Figure 5.10) Using the log transformed data in subsequent statistical tests that are based on normally distributed assumptions is highly recommended. Although not all 81 of the metabolite distributions are shown in Figure 5.10, the majority of the population distributions are skewed and can take advantage of the log transformation. Only urea showed a normal distribution before transformation. This is probably due to large concentrations of urea found in urine, and the mean of the population being very far from the constraint at zero.

5.4. Dilution Normalization of Urine Metabolites

Another issue we wish to address is the question of normalization. Urine samples taken at various times of the day, often end up having differing degrees of dilution. Normalization without any prior knowledge typically involves mean centering and unit variance assumptions. However, since the objective is to remove the variance due to dilution of the urine, one way of normalization would be to normalize the concentration of the other metabolites by dividing the concentration of creatinine measured in the sample. This method is very common for urine measurements due to the fact that creatinine is very consistent with dilution levels.

Table 5.1 is a list of 15 compounds measured from urine samples obtained from a male population between the ages of 25-60. Table 5.1 shows the mean concentrations and the standard deviations of each of the compounds listed. Table 5.1 also shows the mean concentrations and the standard deviations of each of the compounds after it has been normalized to creatinine. As can be seen there is a marked improvement in the standard deviations of the measured compounds. Effectively, creatinine normalization removes the variance from creatinine or the variance due to the dilution of the sample.

The variance removed can help in classification when using unsupervised methods. However, when using supervised methods such as Partial Least Squares – Discriminant Analysis (PLS-DA), creatinine normalization is not necessary as the variance due to dilution will be in an orthogonal direction to that of the group classifier. For unsupervised methods however, the variance due to dilution effects can be the dominant factor and must be taken into consideration.

Table 5.1 – Metabolite concentration means and standard deviations before and after creatinine normalization.

Metabolite	Raw Data		Creatinine Normalized		
	Mean {μM}	Std. Dev. {μM}	Mean {μM/mM(Cr)}	Std. Dev. {μM/mM(Cr)}	% Improvement in Std. Dev.
Alanine	171.79	124.51	22.99	10.77	25.62%
Citrate	1339.70	1112.32	180.82	99.83	27.82%
Formate	140.24	103.74	20.22	11.60	16.59%
Glucose	164.48	107.04	21.95	9.40	22.28%
Glutamate	46.08	29.74	6.66	3.89	6.07%
Glutamine	286.29	188.71	38.49	16.20	23.83%
Glycine	738.84	633.73	101.78	69.66	17.33%
Hippurate	1336.35	1386.98	193.62	173.48	14.19%
Histidine	501.64	335.67	66.83	28.95	23.59%
Isoleucine	11.06	6.43	1.53	0.57	20.88%
Leucine	20.13	12.51	2.68	0.81	31.88%
Taurine	227.80	242.54	30.53	23.71	28.82%
Tiglylglycine	13.07	9.20	1.86	1.10	11.18%
TMAO	329.61	536.71	46.18	64.52	23.13%
Valine	19.13	12.12	2.54	0.72	34.95%

5.5. Conclusion

This chapter outlined some of opportunities for data analysis that is offered by working with Targeted Profiling data. Since the identification of the metabolites is done *a priori*, the interpretation of the results is much more meaningful and direct mapping onto metabolic networks is possible. It was also shown that there are some considerations that must be given when working with Targeted Profiling data, especially in the distribution of the data. Normality assumptions do not always hold when working with Targeted Profiling data. The log transformation of data is not typically done with Targeted Profiling data, and should be considered in all cases. Creatinine normalization for urine is well established, and when using unsupervised methods that make use of all the variance in the data, it was shown that the % variance attributed to creatinine variability or dilution can be anywhere between 6 to 35% for each metabolite. Examples of patients with Diabetes and Asthma showed that with careful management of data characteristics such as proper normalization and transformation, the statistical results can be very useful in metabolomics.

6. [3]Multivariate Models and Visualizations as Applied in a *Streptococcus pneumoniae* Study

This chapter applies multivariate statistical techniques on NMR derived metabolomics data to detect *Streptococcus pneumoniae* infections. New advanced multivariate visualization techniques are used, highlighting a number of cross variable interactions. *Streptococcus pneumoniae* is a major human pathogen causing life-threatening invasive diseases that can affect various parts of the body such as the lungs (pneumonia), blood (bacteremia) and meninges (meningitis), with high morbidity and mortality worldwide (Ridgway et. al, 1995). Furthermore, *S. pneumoniae* is one of the major causes of otitis media (middle ear infection), acute sinusitis (sinus infection), septic arthritis (joint infection), cellulitis (skin infection), upper respiratory tract infections, and causes other diseases including osteomyelitis (bone infection), peritonitis, endocarditis, and pericarditis (Dubost et. al, 2004, Kan et. al, 2006, Lopez et. al, 1999, Parada and Maslow, 2000). People at increased risk for infection include children under 2 years of age, those 65 years of age and older, people with compromised immune systems, people with chronic diseases such as diabetes, lung disease, cancer, kidney disease, people with impaired spleen function and smokers (Brandenburg et. al, 2000, Marrie, 2004, Mitchell, 2000, Redelings et. al, 2005). In addition, there are a large number of nasopharyngeal asymptomatic carriers of *S. pneumoniae* (Lopez et. al, 1999, Faden et. al, 2002, Gillespie and Balakrishnan, 2000, Peterson, 2006). It is also not unusual that pneumococci carried in the nasopharynx are resistant to one or more antibiotics (Peterson, 2006). Each year, *S. pneumoniae* infects millions of people in the United States alone resulting in more than 600,000 hospitalizations, with a mortality rate ranging from 5 – 22% (Mitchell, 2000, Bartlett et. al, 2000).

[3] Some of the material that appears in this chapter has been previously presented by Chang et. al at Metabolomics Society 2nd Annual Conference, June 25-29, 2006, Boston, MA, U.S.A..

Current diagnostics for *S. pneumoniae* infections require culture of a variety of specimens including sputum, bronchoalveolar lavage, cerebrospinal fluid, or blood, which can take several days for a positive result. To combat the lengthy diagnostic time, other tests have been developed such as the NOW test (Binax Inc.) which detects the cell wall polysaccharide of *S. pneumoniae* in the urine. However, the rate of detection for this test is only between 80 – 90% in bacteremic patients (Faden et. al, 2002), and since we don't have a gold standard for those who have non-bacteremic pneumococcal pneumonia, we don't really know if this test is of much value. Furthermore there is a 65% false positive rate in children who carry this microorganism in the nasopharynx (Faden et. al, 2002). An ideal diagnostic tool for *S. pneumoniae* infection would be something that is non-invasive, requires a minimal amount of a readily available sample that is not contaminated by carriage of the organism at the site from which it is obtained, can be done reasonably quickly, has a high specificity, and is technically simple to implement.

Recent advancements in the fields of genomics, transcriptomics, proteomics and metabolomics have led to proposals that a combinatorial or systems biology approach will lead to advanced diagnostics and therapeutics. Most "omics" studies rely on easily obtainable samples, such as urine or serum. However, the human biological system is complex, and human physiology is affected by many environmental factors such as diet, drugs, and symbiotic organisms (Nicholson et. al, 2005, Nicholson and Wilson, 2003). It is these factors that tend to complicate analyses and make interpretation difficult. Even so, much effort has been applied to find a few key differences between affected and unaffected individuals which may pave the way for earlier diagnostics and prognostics.

There have been a number of successful studies using ^{1}H NMR spectroscopy of urine to observe intestinal and urinary tract infections by various microbes (Gupta et. al, 2005, Van et. al, 2004, Wang et. al, 2004). Most NMR type analyses use raw NMR spectral data that provide no *a priori* information on the metabolites of interest to differentiate disease states. These types of analyses are difficult at best as ^{1}H NMR is very sensitive to sample conditions such as pH and ionic strength. In this paper, we have assigned and followed more than 80 metabolites in 59 patients testing positive for *S. pneumoniae* in

one or more of blood, sputum, bronchoalveolar lavage or endotracheal tube secretions, and 59 healthy controls to determine whether individuals with a pneumococcal infection may be differentiated from a healthy population based strictly on urinary metabolites as a first step toward creating a more robust diagnostic specific for this disease.

6.1. Experimental

Sample Collection and Preparation (note: sample collection was done through Dr. Marrie and Dr. Erik Saude):

Normal subjects: A total of 59 volunteer subjects, self identified as normal, constituted our control group. Urine samples were collected twice daily – once as the first void sample in the morning and the second around 1700 h.

Patients with *S. pneumoniae* infections: A total of 59 patients infected with *S. pneumoniae*, as determined through cultures of blood, sputum, cerebrospinal fluid, bronchoalveolar lavage samples, endotracheal tube secretions, ascites or a combination of any of these, constituted our pneumococcal infection group.

Written informed consent was obtained from each patient and normal subject before entering this study, and the study protocol was approved by the institutional ethics committee.

Sample processing: Upon acquisition of urine samples, sodium azide was added to a final concentration of approximately 0.02% to prevent bacterial growth. Urine was placed in the freezer and stored at -80 ˚C until ready for preparation and data acquisition.

Sample preparation: Urine samples from healthy individuals were prepared by adding 70 μL of internal standard (Chenomx Inc.) (consisting of ~5 mM DSS, 100 mM Imidazole, 0.2% sodium azide in 100% D_2O) to 630 μL of urine. Urine samples from the pneumococcal patients were prepared by adding 80 μL of the Chenomx internal standard to 820 μL of urine. Sample pH was adjusted to approximately 6.8 by the addition of small amounts of NaOH or HCl. 600 μL of sample was placed in a 5 mm NMR tube and stored at 4 ˚C until ready for data acquisition.

NMR data acquisition and processing: NMR spectra were acquired using the first increment of the standard NOESY pulse sequence on a 4-channel Varian INOVA 600 NMR spectrometer with triaxial-gradient 5 mm HCN probe. All spectra were recorded at 25 °C with a 12 ppm sweep width, 1 s recycle delay, 100 ms τ_{mix}, an acquisition time of 4 s, 4 dummy scans and 32 transients. ^{1}H decoupling of the water resonance was applied for 0.9 s of the recycle delay and during the 100 ms τ_{mix}. All spectra were zero-filled to 128k data points and multiplied by an exponential weighting function corresponding to a line-broadening of 0.5 Hz.

Concentration determination: Quantification of urinary components was achieved using the 600 MHz library from Chenomx NMR Suite 4.0 (Chenomx Inc., Edmonton, Canada), which uses the concentration of the added DSS to determine the concentration of metabolites. The Chenomx 600 database was validated against a set of known compound concentrations using the same NMR data collection parameters as used in this study and deemed accurate to better than 15% for all compounds reported.

6.2. Statistical Analysis

Partial Least Squares – Discriminant Analysis (PLS-DA) was performed using standard procedures as implemented in Simca P 11.0 (Umetrics, Umeå, Sweden). Input variables consisted of raw compound concentrations. Data were pre-processed by mean-centering and unit variance scaling prior to analysis. After building the PLS-DA models, a validation of these models was done by a permutation test. One hundred random permutations of class labels were performed, and the R^2 and Q^2 of these new models were compared with the original model before permutations. This test is a good indication of how well the original model was fit as compared to randomness.

ANOVA was done using the program StatView 5.0.1 (SAS Institute Inc., Cary, NC, USA). Each metabolite was subjected to a log transformation prior to analysis.

Correlation maps were created by calculating correlation matrices between each of the log-transformed metabolite concentrations. The correlation of each element was calculated using (Johnson and Wichern, 1998)**:**

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}} \cdot \sqrt{s_{kk}}} \quad \textbf{(6.1)}$$

where s_{ik} is the sample covariance between i and k, and was calculated as follows:

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right) \quad \textbf{(6.2)}$$

An appropriate color gradient was mapped onto the correlation values. A threshold was chosen to highlight important positive and negative correlations.

Heat maps were calculated from log-transformed metabolite concentration data as the deviation of each metabolite of the pneumococcal dataset from the mean of that metabolite of the control. The data were calculated as follows:

$$\frac{X_{PP} - \overline{X}_{Control}}{\sigma_{Control}} \quad (6.3)$$

6.3. Results

A total of 59 patients with pneumococcal disease, hereafter referred to as the pneumococcal group, ranged in age from 6 days to 92 years (Table 6.1). In addition to the positive cultures indicated in Table 6.1, *S. pneumoniae* was also isolated from the cerebrospinal fluid of 3 patients, and from ascites in 2 patients. The control group consisted of 29 males and 30 females ranging in age from 21 to 75 with a mean age of 43 ± 14 years.

Table 6.1 – Selected features of the 59 patients with *S. pneumoniae* infection

	Total Number (% of Total)	Survivors (%)	Non-Survivors (%)
Number of Patients	59 (100%)	48 (81%)	11 (19%)
Age (mean years ± SD)	56 ± 22	55 ± 22	59 ± 25
Male Gender (no, %)	35 (59%)	25 (52%)	10 (91%)
Diabetes as underlying chronic illness	9 (15%)	5 (10%)	4 (36%)
Bacteremia	38 (64%)	29 (60%)	9 (82%)
Pneumonia	36 (61%)	28 (58%)	8 (73%)
Bacteremic pneumococcal pneumonia	23 (39%)	17 (35%)	6 (55%)
Probable pneumococcal pneumonia*	5 (8%)	4 (8%)	1 (9%)
Possible pneumococcal pneumonia**	16 (27%)	15 (31%)	1 (9%)

*Probable pneumococcal pneumonia is defined as a positive culture in endotracheal tube secretions or bronchoalveolar lavage samples but not blood samples.

**Possible pneumococcal pneumonia is defined as *S. Pneumoniae* isolated from respiratory culture, but does not meet criteria for pneumococcal pneumonia.

Figure 6.1 depicts a comparison of typical ^{1}H NMR spectra obtained from sample urines of the control group (C,D) with approximate age and gender matched individuals from the pneumococcal group (A,B). Spectra were scaled according to the intensity of the creatinine resonances at approximately 3 and 4 ppm. While there are some differences between the spectra of the controls, there are major differences between the spectra of the controls versus the pneumococcal group. Citrate, which is a strong signal in the spectra from the control group, is very low in the spectra from the pneumococcal group. Carnitine and acetylcarnitine have much stronger signals in the spectra from the pneumococcal group versus the spectra from the control group. Of interest, the 58 year-old pneumococcal female patient had very high levels of creatine. In many patients, but not all, creatine was elevated significantly. Other major differences between the control and pneumococcal spectra may be observed in the 6 – 8 ppm range.

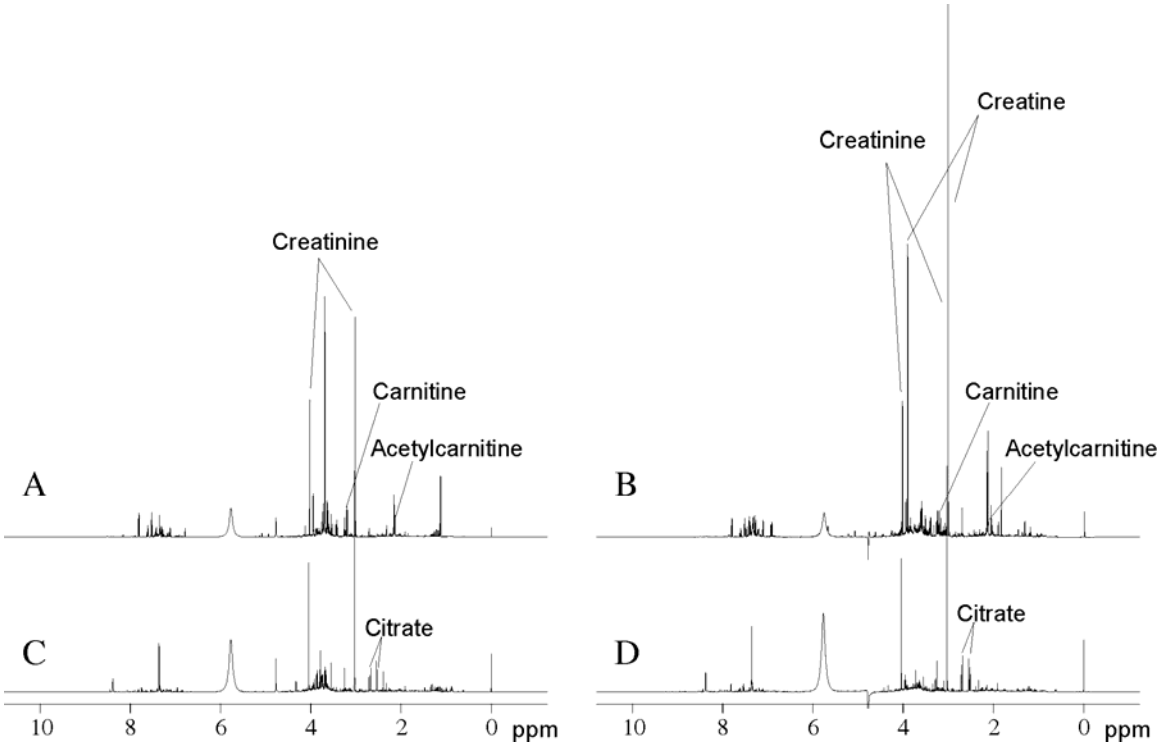

Figure 6.1 – 600 MHz ^{1}H NMR spectra obtained from (A) 26 year-old male with a possible case of pneumococcal pneumonia, (B) 58 year-old female with bacteremic pneumococcal pneumonia, (C) Healthy 26 year-old male, (D) Healthy 57 year-old female. None of these patients had diabetes.

Examination of the distribution of metabolite concentrations, measured from NMR spectra relative to the concentration of the added shift standard, revealed that very few metabolites exhibited a normal distribution (not shown). For example, several of the pneumococcal patients had diabetes, and thus had extremely high levels of glucose in their urine. This type of behavior is not unexpected and is generally inherent in these types of data, especially where a hard constraint of zero concentration is found at one end of the distribution. Upon log-transformation, histograms revealed a normal distribution for all metabolites. Thus for all data analysis presented herein using metabolite concentration data, the concentrations were subjected to log-transformation.

Univariate ANOVA was performed to test for significant differences in the means of each metabolite concentration between the control and pneumococcal group as well as the control and a subset of the pneumococcal group containing no diabetic patients. ANOVA revealed 37 out of the 82 measured metabolites had significantly different means, and removal of the diabetic patients had little or no effect on the results of the ANOVA. Some of the metabolites exhibiting major differences between the control and pneumococcal groups are summarized in Table 6.2 with their corresponding p-values.

Table 6.2 – Relative Metabolite Concentrations for patients with pneumococcal pneumonia.

	Pneumococcal patients - No Diabetics		All Pneumococcal patients	
Metabolite	Relative Concentration	p-value	Relative Concentration	p-value
Acetylcarnitine	↑↑↑	<0.0001	↑↑↑	<0.0001
Acetoacetate	↑↑	<0.0001	↑↑	<0.0001
Carnitine	↑↑	<0.0001	↑↑	<0.0001
Acetone	↑↑↑↑	<0.0001	↑↑↑↑	<0.0001
Fumarate	↑	<0.0001	↑	<0.0001
Valine	↑	<0.0001	↑	<0.0001
Trigonelline*	↓↓	<0.0001	↓↓	<0.0001
Tyrosine	↑	<0.0001	↑	<0.0001
Isoleucine	↑	<0.0001	↑	<0.0001
Acetate*	↑↑	<0.0001	↑↑	<0.0001
Fucose	↑	<0.0001	↑	<0.0001
Taurine*	↑↑	<0.0001	↑↑	<0.0001
Citrate	↓	<0.0001	↓	<0.0001
myo-inositol	↑↑	0.0001	↑↑	<0.0001
1-Methylnicotinamide*	↓	0.0001	↓	0.0001
Lactate*	↑↑	0.0003	↑↑	0.0001
Leucine	↑	0.0005	↑	0.0001
Dimethylamine	↑	0.0005	↑	0.0002
Threonine	↑	0.0006	↑	0.0005
Alanine	↑	0.0010	↑	0.0003
Hypoxanthine	↑↑	0.0014	↑↑	0.0015
Tryptophan	↑	0.0026	↑	0.0012
π-Methylhistidine*	↓	0.0078	↓	0.0064
Sucrose*	↑	0.0130	↑	0.0313
Creatine*	↑	0.0151	↑	0.0126

These results may also be presented in heat-map format. Figure 6.2 shows a heat map showing relative metabolite levels of the log-transformed metabolite concentrations for each pneumococcal patient to the average of the control group. Metabolites are colored according to the degree of difference between the average control concentration and each patient's individual concentration. The red color represents metabolites that increase whereas the green color represents metabolites that decrease. This figure is akin to a heat map used for gene microarrays as it illustrates the relationships between metabolite variables across the patient population in a similar manner as gene microarrays show the expression levels of genes. Figure 6.2 shows that acetylcarnitine, carnitine, acetone, acetoacetate and 3-hydroxybutyrate concentrations are elevated across virtually the entire pneumococcal patient population by at least 2 standard deviations. In contrast, citrate, 1-methylnicotinamide and trigonelline were found to decrease across the pneumococcal patient population. Of interest, other compounds such as lactate, acetate and glucose appeared to be somewhat elevated in the pneumococcal group as compared to the control group, but not to the extent of the ketones or carnitines. Interestingly, the patients with diabetes and pneumococcal disease did not appear much different from the patients with only pneumococcal disease. The x-axis in Figure 6.2 was sorted based on age and only for pneumococcal patients. However with such a heat map, one could compare between groups such as pneumococcal and tuberculosis by plotting additional data beside each other. The utility of this visualization technique is to quickly see metabolite patterns on an individual basis for a large number of samples.

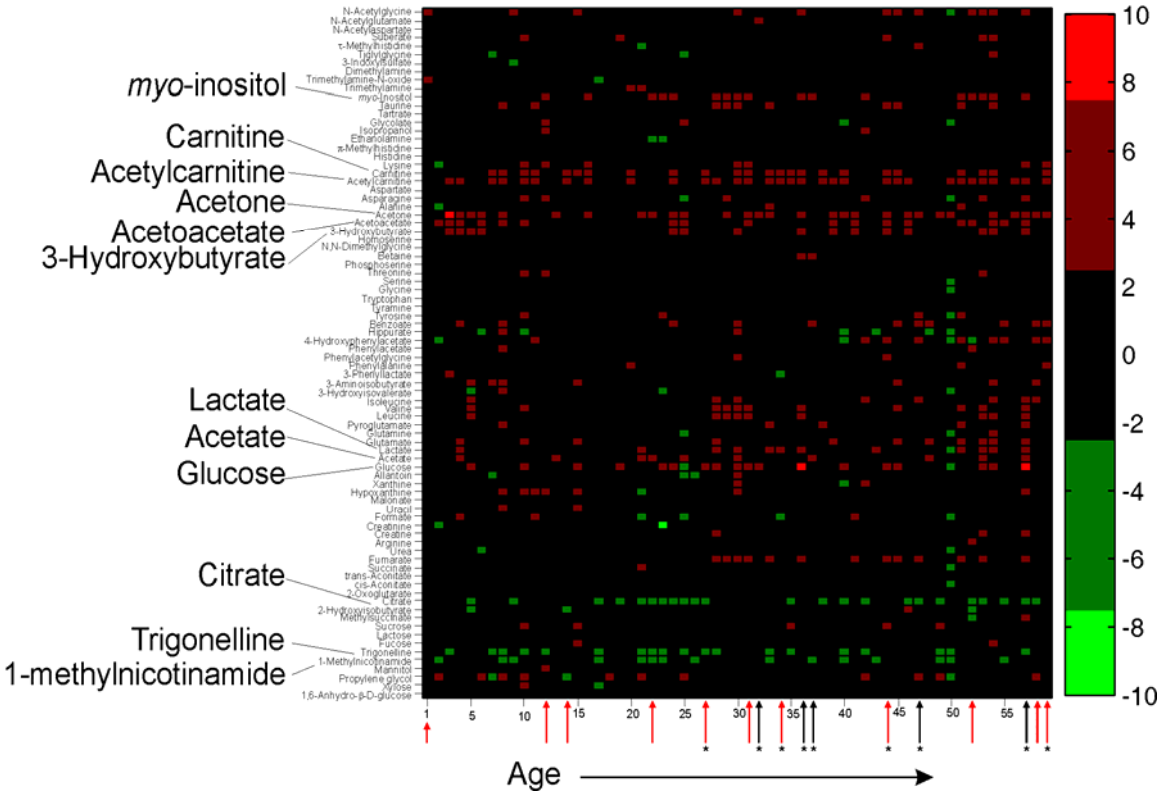

Figure 6.2 – Heat Map Representation of metabolite concentrations for pneumococcal patients. Each value was obtained after log-transformation by subtracting the average metabolite concentration determined from the control population from the pneumococcal patient metabolite concentration and dividing by the standard deviation of the control population. The coloring, representing the magnitude of the deviation, is shown as a side-bar. Those patients who died as a result of complications due to pneumococcal disease are indicated by the red arrows. Those patients who had diabetes are indicated by the asterisk. The patients are ordered from youngest (patient 1, 6 days old) to oldest (patient 59, 92 years old).

Multivariate PLS-DA was performed on the log-transformed pneumococcal and control datasets. A scores plot is shown in Figure 6.3A. A clear separation between the groups is observed. To test for significance of the model, a permutation test was done (Figure 6.3B) where each metabolite dataset was randomly assigned to either the control or pneumococcal groups. 200 permutations were performed, and none of the random assignments approached the R^2 or Q^2 determined for the model.

To be sure that the separation was not affected by potential co-morbidities, such as diabetes, we removed the diabetic patients from the PLS analysis (Figure 6.3D). As was observed with the univariate analysis, the multivariate analysis was affected very little by excluding patients with diabetes, and permutation testing indicated a valid model (Figure 6.3E). Inspection of the individual data points within the PLS-DA plots illustrates no further discrimination of the patient population based on age, gender, or death (data not shown).

Figures 6.3C and 6.3F are loadings plots for the PLS-DA shown in Figure 6.3A and 6.3D respectively. The key metabolites separating the pneumococcal group from the control group are: citrate, trigonelline, 1-methylnicotinamide, acetoacetate, acetylcarnitine, carnitine, acetate, and acetone. Interestingly, these were shown to be significant in the univariate ANOVA as well as heat map data.

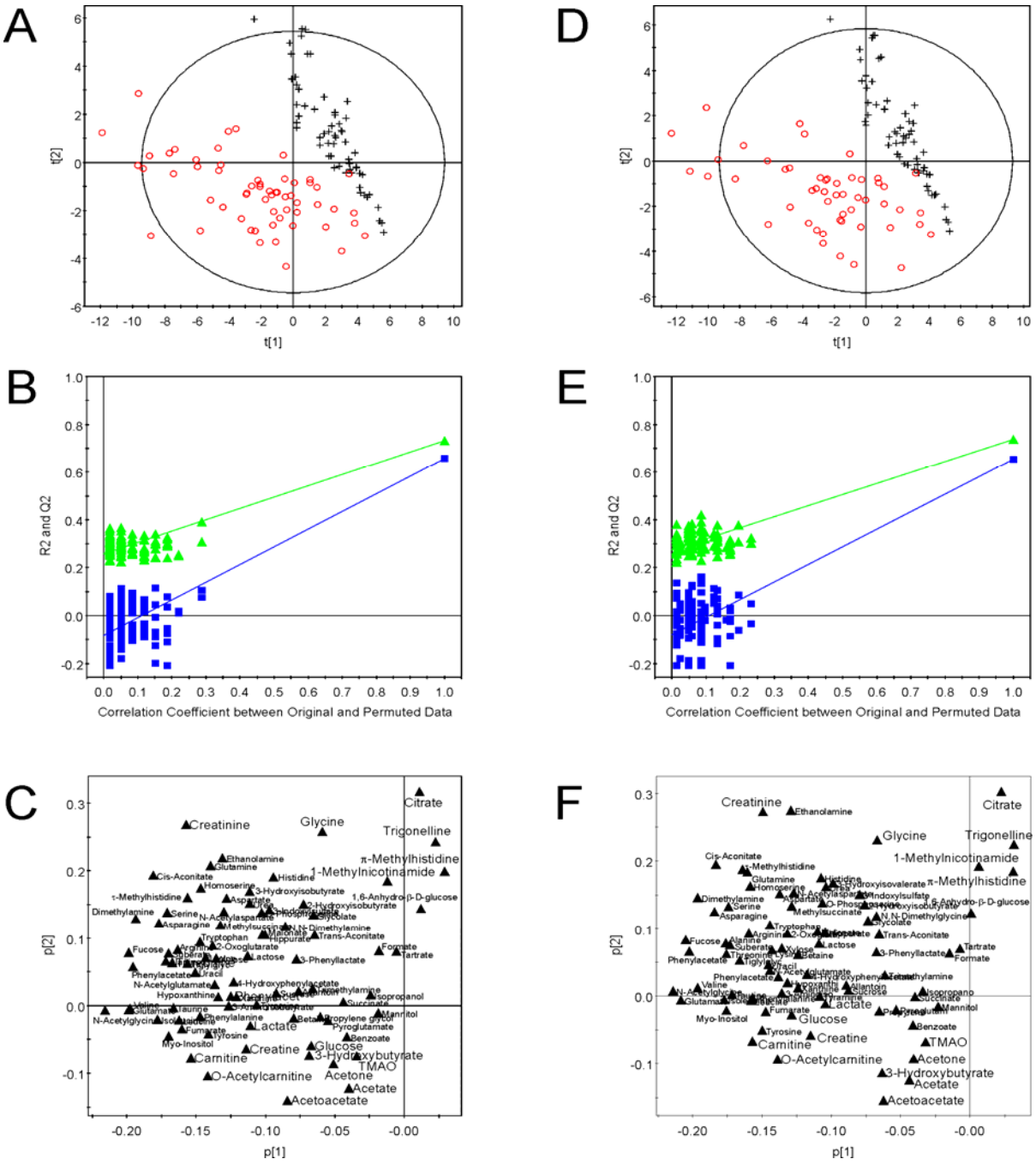

Figure 6.3 – (A) PLS-DA of the metabolite concentrations from all 59 pneumococcal patients and 59 healthy controls, (B) Permutations tests to validate model found in (A). (C) Loadings plot corresponding to (A). (D) PLS-DA of the metabolite concentrations from 50 non-diabetic pneumococcal patients (removal of the 9 diabetic patient data) and 59 healthy controls. (E) Permuation test to validate model found in (D). (F) Loadings plot corresponding to (B). Controls, black crosses; pneumococcal patients, red circles.

While ANOVA and heat maps provide us lists of metabolites that differ, they give no information on the (in)dependence of metabolites to one another. Correlation maps can provide a wealth of information about the independence between and/or co-dependence of variables (Tangirala et. al, 2005). The use of correlation maps described in this chapter is very similar to Statistical Total Correlation Spectroscopy (STOCSY) developed by Cloarec et. al (2005). The main difference between the correlation maps described here and STOCSY is that this method uses metabolite concentrations as the input data to the maps, while STOCSY uses binned areas as input. Figure 6.4 shows a comparison of the metabolite correlations between the control group and the pneumococcal group. In general, correlation maps are symmetric with one side of the diagonal mirroring the other side (the diagonal represents a 100% correlation of each metabolite to itself). Each cross-correlation yields interesting information about the metabolites. A positive correlation between two metabolites (red) indicates that as one metabolite increases in concentration, the correlated metabolite increases as well. A negative cross-correlation (blue) indicates that as one metabolite increases, the correlated metabolite decreases. Figure 6.4a and Figure 6.4b demonstrates striking differences between the control and pneumococcal correlation maps. In the control group, there are many metabolites that are positively correlated, whereas the correlation becomes less clear in the pneumococcal map. It can be observed that there are more significant negative correlations in the pneumococcal map as highlighted in a blue color gradient.

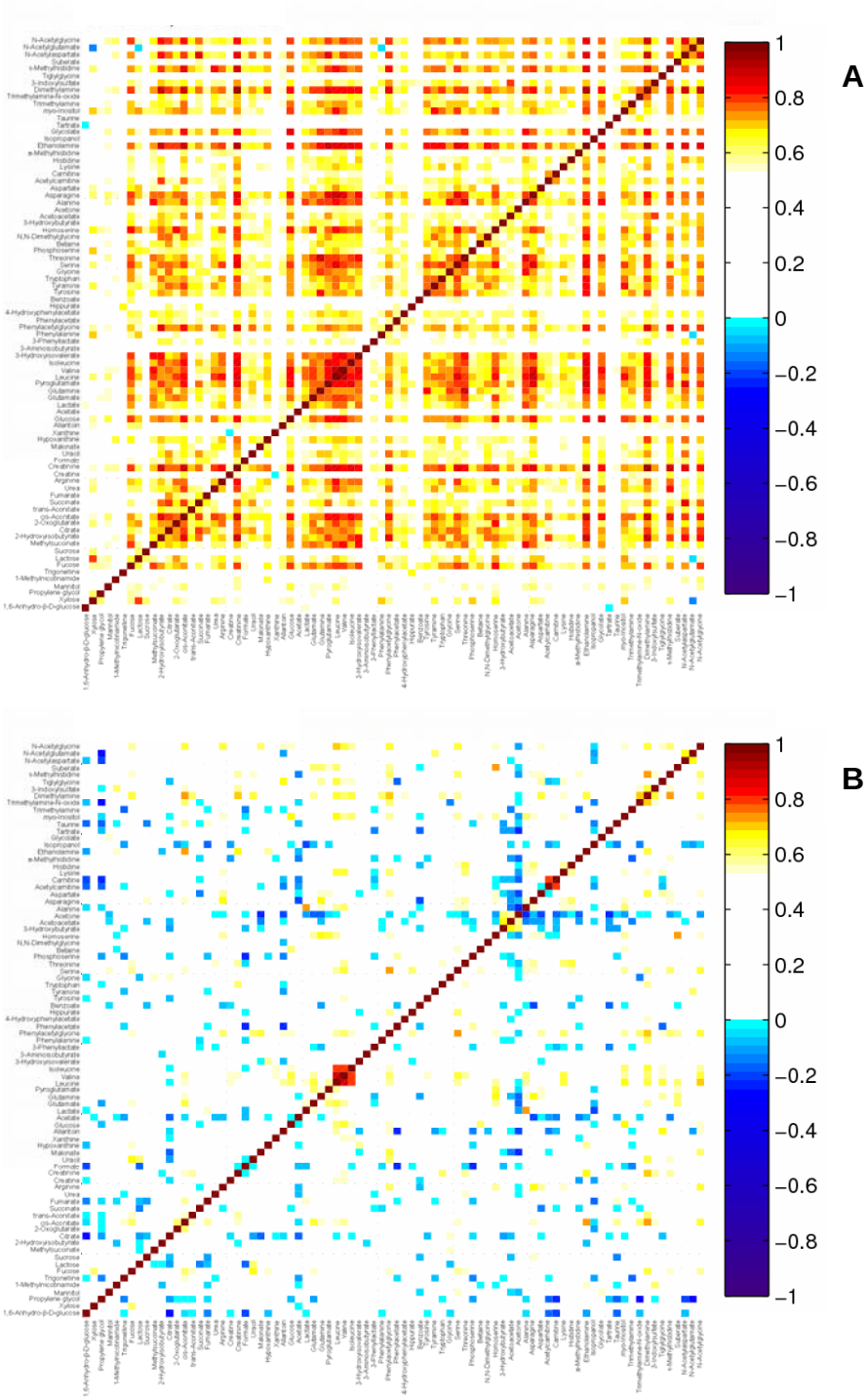

Figure 6.4 – (A) Compound correlation map of health controls. (B) Compound correlation map of pneumococcal patients. Red squares correspond to positive correlations and blue squares correspond to negative correlations. The diagonal represents 100% correlation of each metabolite to itself. Metabolites are indicated on each axis.

6.4. Discussion

We expect that our analysis of the urinary metabolome of pneumococcal patients will be a combination of the host metabolome, the bacterial metabolome, and the host response to the bacterial infection which might include organ injury, muscle damage, cytokine response or neutrophil activation (Gillespie and Balakrishnan, 2000). Furthermore, we expect that some of the differences we observe between healthy controls and pneumococcal patients may be due to other diseases such as diabetes or kidney disease. However, we expect that since only a subset of the pneumococcal patients will have these diseases, they should not be a discriminating factor.

Using both univariate and multivariate analysis techniques on metabolite concentrations determined for 82 compounds using ^{1}H NMR spectroscopy, it was determined that several compounds were responsible for separating the pneumococcal group from the healthy controls (Table 6.2). The elevated levels of glucose and ketone bodies (acetone, acetoacetate, and 3-hydroxybutyrate) in some patients may due to the fact that some are diabetic (9 patients in total), but may also be indicative of alterations in energy substrate-endocrine relationships. Indeed it has been found that the NMR spectra of urine from intra-operative patients showed increases in urinary excretion of alanine, ketone bodies, lactate and glucose over time that correlated specifically to the degree of surgical stress (Tatara et. al, 1999). Presumably, the ketone bodies indicate a shift to the use of fatty acids for energy production. As well, a rise in ketone body concentration coupled with a rise in acetylcarnitine concentration has also been shown in fasting healthy subjects (Bales et. al, 1986). Interestingly, we found increased levels of taurine which has been previously shown to be associated with liver damage (Kerai et. al, 1999). In a separate study, bacteremic pneumococcal patients were found to have high bililrubin concentrations (Shariatzadeh et. al, 2005). Some patients had high levels of trimethylamine-N-oxide (TMAO), dimethylamine, acetate, and lactate and low levels of citrate which has previously been shown to be related to kidney dysfunction (Bell et. al, 1991, Foxall et. al, 1993, Wishart, 2005).

In the majority of patients with pneumococcal disease, acetylcarnitine and carnitine were substantially elevated (Figure 6.2). Synthesis of acetylcholine has been shown to be stimulated by glucose, carnitine or acetylcarnitine (Nalecz et. al, 2004). Carnitine has also been shown to be an essential cofactor for the transport of fatty acyl groups into the mitochondrial matrix (Calvani et. al, 2000). In addition, carnitine has been found to be metabolized into trimethylamine which is absorbed, converted to trimethylamine-N-oxide in the liver and excreted in the urine (Rebouche, 2004). This may explain higher levels of trimethylamine and TMAO in some patients. Of interest, acetylcarnitine has also been shown to be used in the brain for the production of releasable glutamate rather than as an energy source (Nalecz et. al, 2004).

Univariate ANOVA suggests that any one of the 37 significant compounds should be a useful biomarker for pneumococcal disease. However, the best biomarker would only predict pneumococcal disease 80% of the time. This performance matches that of current testing technologies and it is not known at this time whether this biomarker would be specific for pneumococcal infection or for a number of other bacterial infections. Thus, multivariate analysis techniques, and in particular PLS-DA, was used for analysis. Out of 59 pneumococcal patients, only one overlapped with the control set (Figure 6.3A). Upon examination of the medical records, it was determined that this patient was admitted to the ER because of a drug overdose who tested positive for *S. pneumoniae* in a sputum sample, but did not test positive in a blood culture. Since there was no evidence of pneumonia on chest x-ray, this patient was likely colonized with *S. pneumoniae,* suggesting that carriers and infected individuals may be differentiated. However, many more carriers need to be studied before we can make this conclusion.

One question might be whether metabolites differentiating the control and pneumococcal groups are somehow interrelated, and if so, what the strength of the relationship between the variables is. Correlation maps (Figure 6.4) may provide clues as to metabolite relationships, affected metabolic pathways, and the source of the metabolite, whether from the human or bacterial metabolome. Furthermore, these relationships may well be

specific for different diseases due to differential immune responses, pathogen metabolites and stress metabolites, for example. Our results indicate that the difference between the pneumococcal group and the control group lies directly with the disease and may potentially be related to the nature of the infection and the resulting immunological host response. Interestingly, it has been shown that increased levels of creatine and π-methylhistidine are related to muscle injury (Hickson and Hinkelmann, 1985, Threlfall et. al, 1981, Threlfall et. all, 1984). However we see a negatively coupled relationship between creatine and π-methylhistidine; as creatine concentration increases, π-methylhistidine decreases. Clearly, more work needs to be done to define why these metabolites behave in this manner in pneumococcal patients and which metabolites appearing in urine may be specific to the bacterial infection, acute lung injury and/or neutrophil activation.

A survey of the pneumococcal patients who died (7 non-diabetics) revealed higher levels of lactate, leucine, and *myo*-inositol, and lower levels of 1-methylnicotinamide, citrate, acetylcarnitine, carnitine and taurine when compared to the survivors. One question that arises from this might be if the host response is somehow different in the patients who died, and whether we may be able to predict negative outcomes. Clearly, more data needs to be acquired to fully answer this question.

6.5. Conclusion

While the study of pneumococcal patients reveal distinguishing biomarkers from both univariate tests and multivariate statistical analyses, the validation of these biomarkers remains difficult. The patterns of endogenous metabolites show a superficial, yet important effect of metabolism. These effects are a good description of the phenotypic differences between diseased and non-diseased patients. A full systematic view of these effects still needs to be explored. Specificity of these biomarkers also depend very much on the control data used to building these models. Further work needs to be done to validate this data for specificity.

7. Normalized Correlation Difference Maps and the Study of Relationship between Variables

In previous chapters, we discussed the many tools needed for discovery of biologically relevant metabolite biomarkers using NMR. One of the more successful tools relied on multivariate models. PLS-DA models make use of the correlation structure between variables, both to reduce the model dimensions due to co-linearity and to use the correlation structure between classifying groups. Such models are powerful as predictors of class separation, but are often very difficult to interpret and visualize. Also seen in previous chapters, metabolites are ultimately what are generating the NMR data. These metabolites form a complex network of interactions that PLS-DA models can only infer, but are difficult to interpret. In this chapter, the use of normalized correlation difference maps will be shown as a novel method to indicate these co-relationships between metabolites. The goal here is not to facilitate prediction, as with building PLS-DA models, but to assist with the interpretation of meaningful metabolite relationships.

7.1. Mouse Model

In this chapter, we will be using data obtained from pneumococcal infected mice. The metabolites measured from mice urine corresponded well with the human study shown in previous chapters. The mouse model work was done by Paige Lacy and Andriy Cheypesh at the University of Alberta. The dataset included intra-tracheal injection of C57B16 mice with either bacterial growth media (sham) or growth media containing 107 cfu of S. *pneumoniae*. There was a total of 12 sham and 14 S. *pneumoniae* infected mice. 47 metabolites were then profiled using Chenomx NMR Suite 4.62.

7.2. Results and Discussion

Correlation maps were used to initiate this analysis, and are described in detail in Chapter 6. Using the 47 profiled metabolites as variables measured, correlation maps were generated showing the relationships between those variables for a specific group. In this case there were two groups, SHAM and infected mice. Figures 7.1 and 7.2 show the correlation maps for this data.

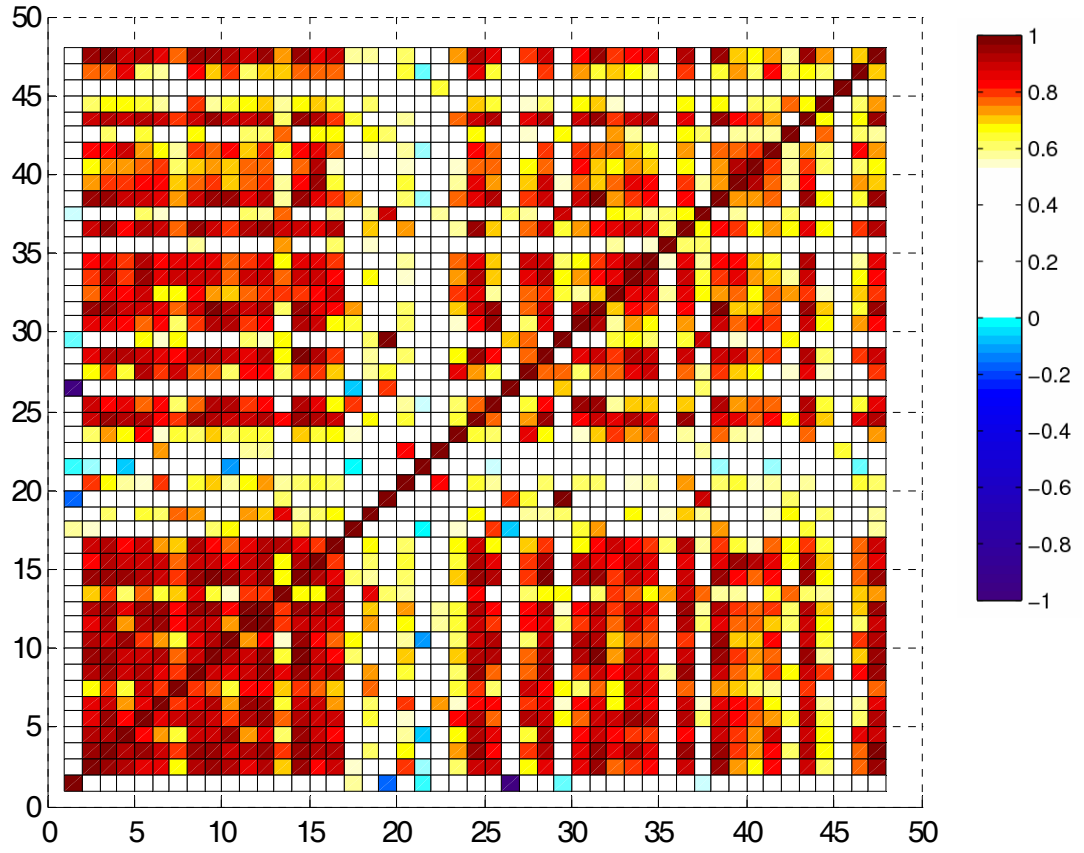

Figure 7.1 – Correlation map of SHAM mice showing relationships between measured concentration variables. (Red = Positive correlations, Blue = Negative correlations)

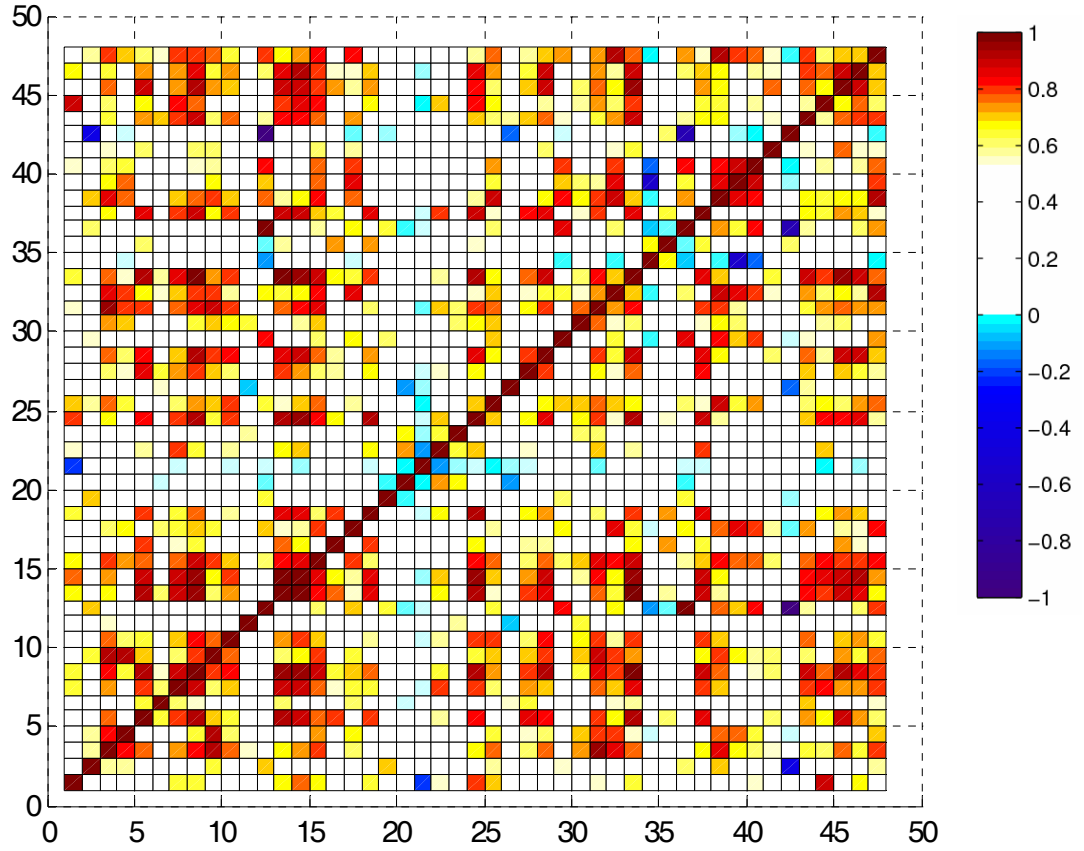

Figure 7.2 – Correlation map of infected mice showing relationships between measured concentration variables. (Red = Positive correlations, Blue = Negative correlations)

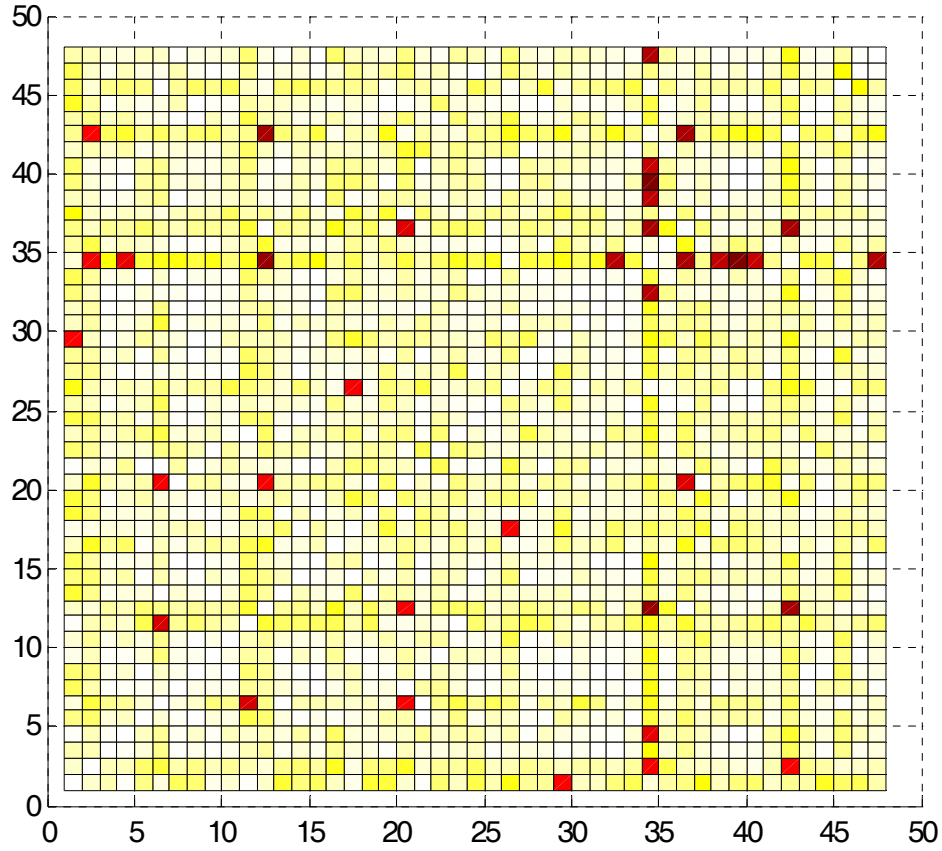

Figure 7.3 – Normalized correlation difference map, SHAM – infected. (Red: Absolute differences >1)

The normalized correlation difference (NCD) map (Figure 7.3) shows the difference (SHAM – infected) in correlation values between the two groups. Visually, this is an easy way to see which variable relationships (correlations) have changed after infection of S. *pneumonia*. Specifically, only the most significant differences between metabolites in SHAM and infected are highlighted in Figure 7.3.

Another way to understand the data is to sort through each of the data points on the correlation map. In Table 7.1, the correlation differences were sorted from highest to lowest, and correlation differences greater than one are shown. Also shown in Table 7.1 is the correlation value in each group of SHAM and infected, as well as the metabolites

that are associated to this correlation value. Only the difference values greater than one are shown on this list, as these correlations are guaranteed a sign change from SHAM to infected. That is not to say that there are not other correlations that have changed signs, however, these are the most dramatic changes after infection.

Table 7.1 – Sorted list of correlation differences

No.	xindex	yindex	Corr_sham	Corr_spneum	Diff value	Metabolite X	Metabolite Y
1	34	39	0.8621	-0.4038	1.2659	O-Acetylcarnitine	Trimethylamine
2	34	12	0.8874	-0.3009	1.1882	O-Acetylcarnitine	Citrate
3	42	12	0.6710	-0.4907	1.1617	Uracil	Citrate
4	42	36	0.7219	-0.4320	1.1539	Uracil	Succinate
5	34	36	0.8831	-0.2657	1.1489	O-Acetylcarnitine	Succinate
6	34	47	0.8879	-0.2421	1.1300	O-Acetylcarnitine	trans-Aconitate
7	34	32	0.8806	-0.2291	1.1096	O-Acetylcarnitine	N-Carbamoyl-β-alanine
8	34	38	0.8702	-0.2335	1.1038	O-Acetylcarnitine	Trigonelline
9	34	40	0.7566	-0.3125	1.0691	O-Acetylcarnitine	Trimethylamine N-oxide
10	11	6	0.9578	-0.0809	1.0387	Choline	Acetate
11	36	20	0.7909	-0.2210	1.0119	Succinate	Glucose

Using the results from Table 7.1, Box and Whiskers plots were generated for each of the metabolites in the list. These plots show the general distribution of the concentration values between the two groups. These plots will hopefully highlight the specific correlations between the variables and indicate the differences.

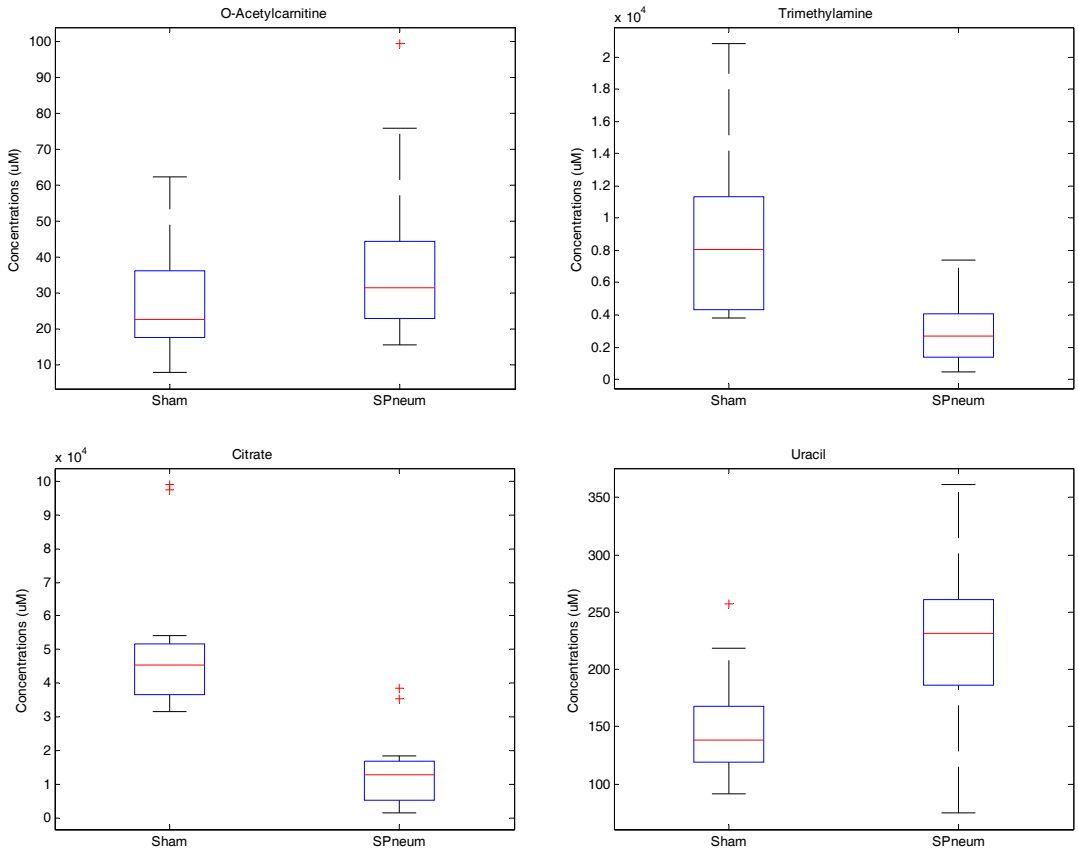

Figure 7.4 – Box and Whiskers plots of O-acetylcarnitine, trimethylamine, citrate, and uracil.

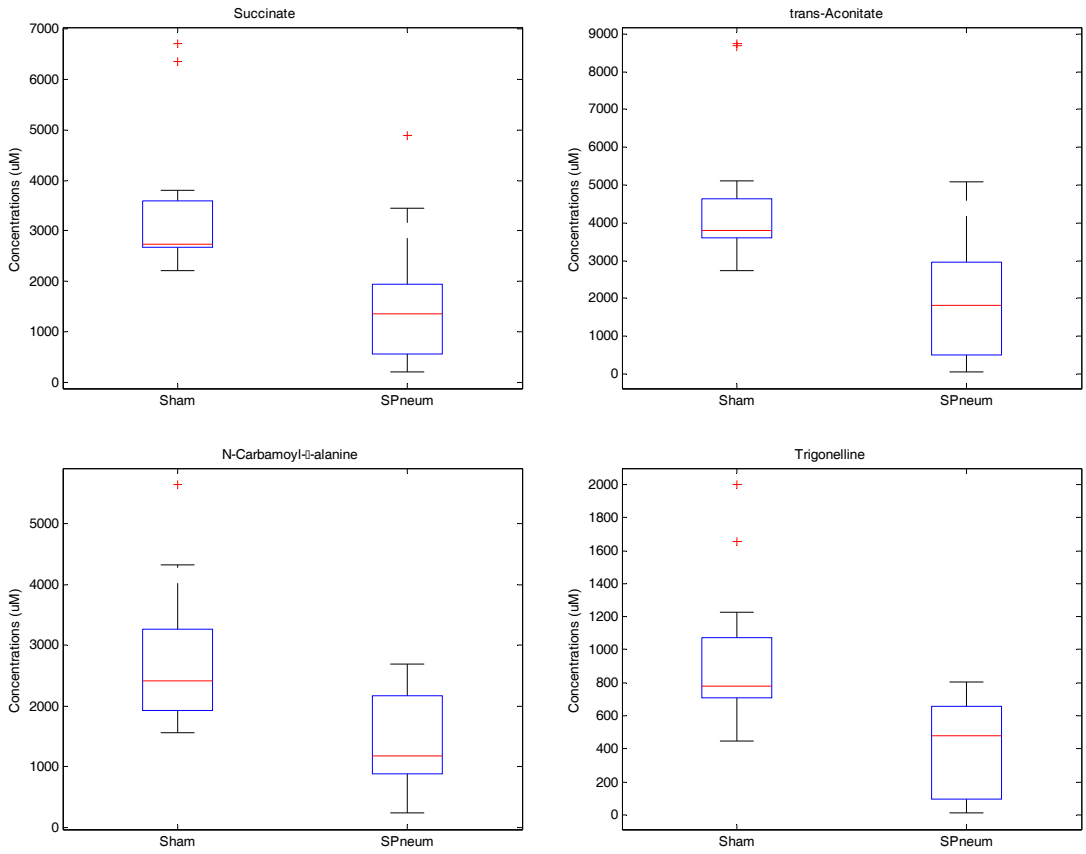

Figure 7.5 – Box and Whiskers plots of succinate, trans-aconitate, N-carbamoyl-β-alanine, and trigonelline.

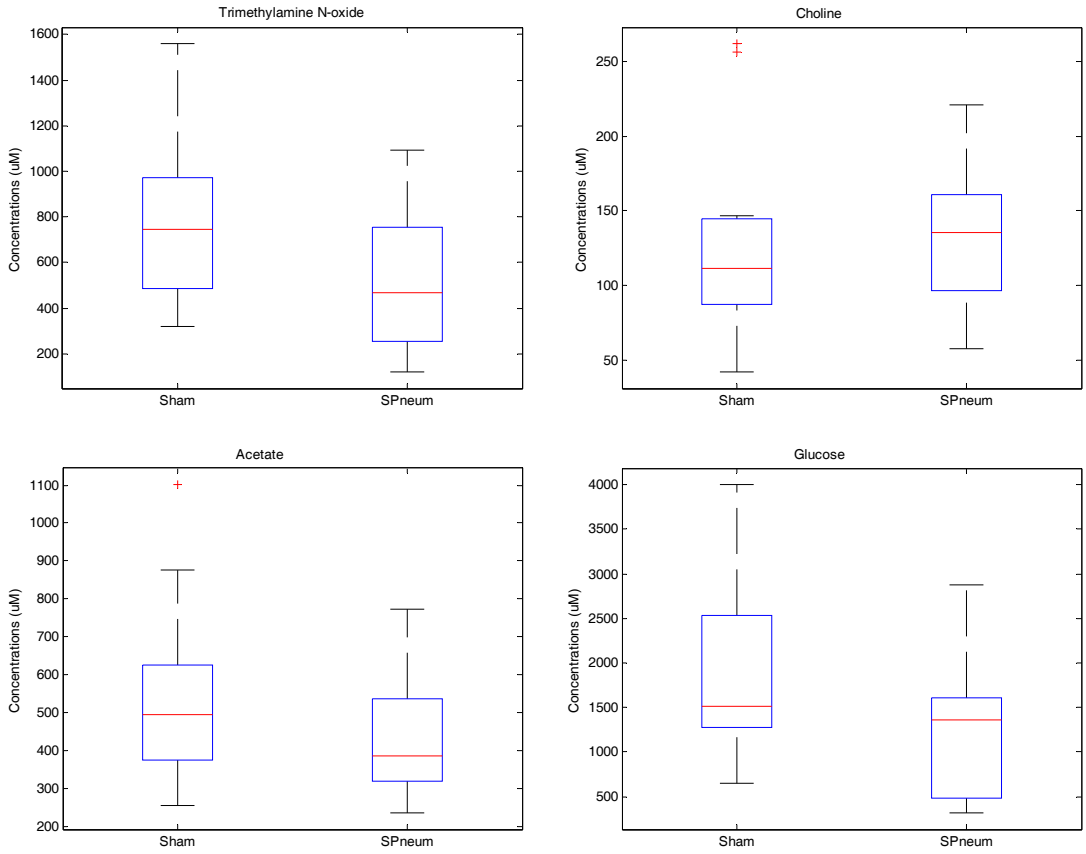

Figure 7.6 – Box and Whiskers plots of trimethylamine-N-oxide, choline, acetate, and glucose.

The utility of the NDC maps lie in the ability to visualize all the pair wise correlations between variables in a large data set. Using this visualization tool to reduce the variables of interest, we can easily find pairs of variables (metabolites) that are of importance when used to separate the two groups of SHAM and infected. The next plots in Figures 7.7 and 7.8 show the Box and Whiskers plots for ratios of two variables. These plots show a better distinction between the two groups than the individual metabolites themselves. The reason for this is because the pair wise correlations highlight the unique relationship between pairs of variables ignoring the variability within the variable due to different samples itself. A ratio of these variables will naturally highlight this relationship for all samples in the experiment.

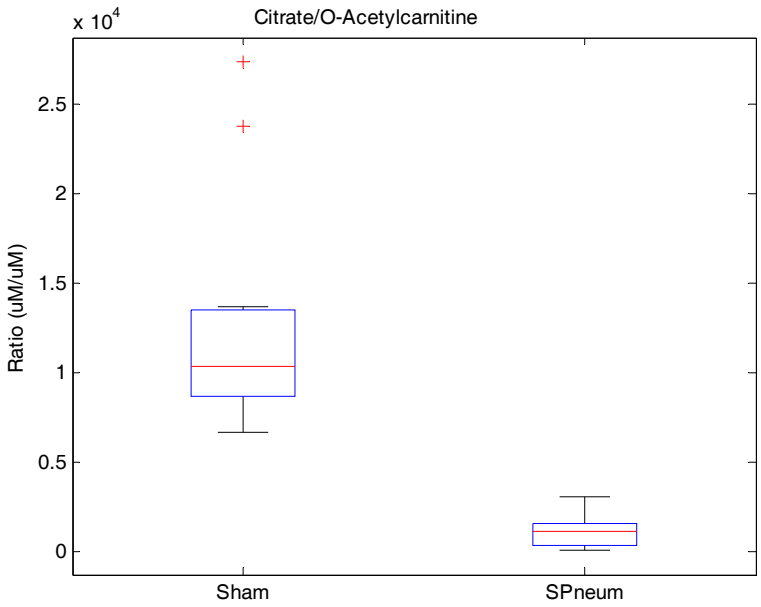

Figure 7.7 – Box and Whiskers plot for ratio of O-acetylcarnitine and citrate

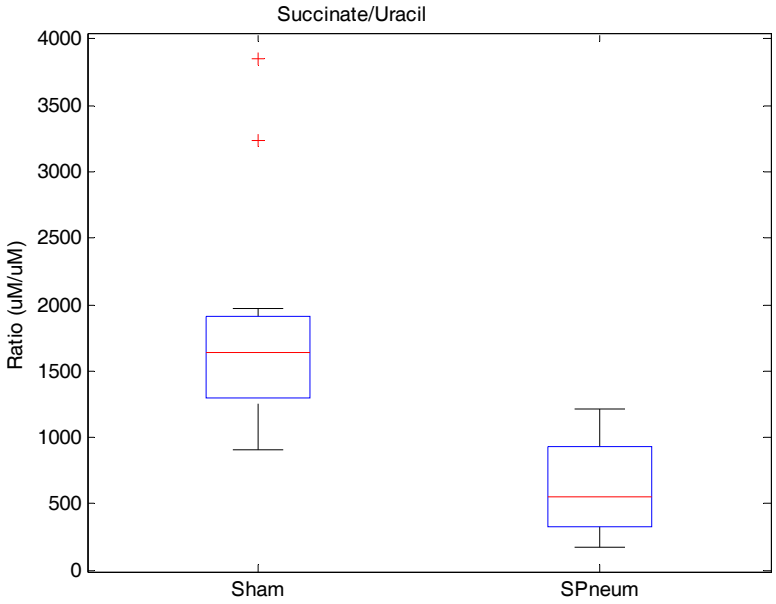

Figure 7.8 – Box and Whiskers plot for ratio of succinate and uracil

7.3. Conclusion

The mouse model data was used to highlight the technique of normalized correlation difference maps. Comparing these results with those of the human data in chapter 6, we see similar metabolites are highlighted due to S. *pneumonia* infection. Similar metabolites found to be markers in both human and mice were trimethylamine-N-oxide (TMAO), dimethylamine, acetate, and citrate. One metabolite that showed real significance in the mouse data and not in the human data is O-acetylcarnitine. Using normalized correlation difference maps to highlight relationship between variables also proved useful in identifying ratios of metabolites that better distinguished the disease state than single metabolites themselves.

8. Conclusions

NMR-based metabolomics starts with an NMR spectrum. This spectrum represents various levels of complexity. The most basic level is the quantum interactions between protons on a single molecule. One of the major contributions of this thesis included the development of a modified spin simulation algorithm. This was demonstrated to quickly simulate second order effects of proton coupling without the need for a computationally intensive full matrix diagonalization technique. This is important because a quantum level understanding of NMR spin-spin interactions was useful in simplifying the NMR spectrum to mathematical representations that are easy to interpret and reproduce. This development proved to be a major advancement for the Chenomx NMR Suite software. With this contribution, the metabolite database evolved from being a collection of peak (Lorentzian) based models to a collection of fully realized quantum models of metabolites.

We further addressed another common issue with NMR spectra, namely that of baseline distortions. An automated algorithm was developed to correct for such distortions. The application of this algorithm was illustrated on two real datasets and has been shown to be effective at removing these distortions. This is an important contribution to the field of NMR spectroscopy in general, and more specifically to metabolomics. This is due to the fact that the algorithm was specifically designed to handle signal-dense spectra such as those found in metabolomics. This algorithm provides a consistent baseline correction method for quantitative metabolite analysis. This algorithm is now integrated into Chenomx's NMR Suite software for metabolomics.

We also demonstrated how the inherent properties of an NMR spectrum can impact the predictive ability of models built upon spectral binning and targeted profiling representations of NMR data by using a novel method for synthetically generating NMR spectra. The quality of predictive models built was quantitatively assessed, as was the relative robustness of these two methods. Under the experimental design chosen, both methods are very robust with respect to noise. In contrast, variable scaling methods can

affect both the quality and interpretability of the models generated. We found for targeted profiling data, unit variance scaling generates a more robust data representation. Targeted profiling was also found to be an effective dimensionality reduction technique that, overall, is more robust with respect to spectral distortions and high dynamic range metabolites than spectral binning, and is less prone to overfitting than spectral binning models. These findings were validated on a real-world dataset of rat-brain extracts consisting of ~30 NMR detectable metabolites, in which statistical models were less prone to overfitting based on a spectral profiling representation of the data. Spectral binning is a common method for data reduction due to the speed of analysis, while current targeted profiling implementations require interactive input and are relatively time-intensive. As increasingly automated methods for quantitative profiling of NMR data become available, we expect database-driven targeted profiling to become the data-reduction method of choice.

Finally, after getting the full potential out of the data itself, we move on to appropriate multivariate techniques to best model and visualize this data. We explored multivariate approaches to modeling and visualizing this data through an application to analyzing spectra from patients infected with Streptococcus *pneumoniae*. We found solid discriminatory results from our PLS-DA models. We also were able to visualize the wealth of variables and samples through a simple heat map visualization approach. As well the interdependence of correlated variables was visualized through the use of compound correlation maps. All of this helped uncover some clues as to the defining metabolic changes from healthy individuals to the pneumococcal group.

Further, the use of correlation maps was explored to develop a new visualization technique known as normalized correlation difference (NCD) maps. The NCD maps allowed us to highlight the most significant metabolite relationships shared between two distinct populations. Using this new technique, we discovered that mice infected with S. *pneumoniae* also showed similar metabolic changes as humans after infection. As well NCD maps proved to be a useful tool to determine pair-wise ratios between metabolites.

These ratios transformed into new variables that have better classifying power than individual metabolites alone.

9. References

J.H. Ahlberg, E.N. Nilson, and J.L. Walsh. *The theory of splines and their applications.* Academic Press: New York, 1967.

J.R. Bales, J.D. Bell, J.K. Nicholson, and P.J. Sadler. *Magn Reson Med* **1986**, *3*, 849 - 856.

C. Bartels, P. Guntert, and K. Wuthrich. *Journal of Magnetic Resonance* **1995**, 117, 330-333.

J.G. Bartlett, S.F. Dowell, L.A. Mandell, T.M. File, D.M. Musher, and M.J. Fine. *Clin Infect Dis* **2000**, *31*, 347 - 382.

J.D. Bell, J.A. Lee, H.A. Lee, P.J. Sadler, D.R. Wilkie, and R.H. Woodham. *Biochim Biophys Acta* **1991**, *1096*, 101 - 107.

J.A. Brandenburg, J.A., T.J. Marrie, C.M. Coley, D.E. Singer, D.S. Obrosky, W.N. Kapoor, and M.J. Fine. *J Gen Intern Med* **2000**, *15*, 638 - 646.

M. Calvani, E. Reda, and E. Arrigoni-Martelli. *Basic Res Cardiol* **2000**, *95*, 75 - 83.

D. Chang, C.D. Banack, and S.L. Shah. *Journal of Magnetic Resonance.* **2007**, 187(2), 288-292.

D. Chang, A.M. Weljie, and J. Newton. *Pacific Symposia on Biocomputing*. **2007**, 12, 115-126.

D. Chang, K.N. Rankin, A. McGeer, S.L. Shah, T.J. Marrie, and C.M. Slupsky. **2006** Urinary Metabolite Profiles of *Streptococcus pneumoniae* Infection using ^{1}H NMR Spectroscopy. June 25-29, 2006, Boston, MA, U.S.A. Metabolomics Society 2nd Annual Conference.

T.A. Clayton, J.C. Lindon, O. Cloarec, H. Antti, C. Charuel, G. Hanton, J. P. Provost, J. L. Le Net, D. Baker, R. J. Walley, J. R. Everett, and J. K. Nicholson. *Nature* **2006**, 440, 1073.

O. Cloarec, M. Dumas, A. Craig, R. H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J.C. Lindon, E. Holmes, and J. Nicholson. *Analytical Chemistry* **2005**, 77(5), 1282 - 1289.

P. Comon. *Signal Processing* **1994**, 36, 287-314.

M. Defernez, and I. J. Colquhoun. *Phytochemistry* **2003**, 62, 1009.

J.J. Dubost, M. Soubrier, C.D. Champs, M.M. Ristori, and B. Sauvezie. *Joint Bone Spine* **2004**, *71*, 303 - 311.

J.N.S. Evans. *Biomolecular NMR Spectroscopy.* Oxford University Press: New York, 1995.

H. Faden, M. Heimerl, C. Varma, G. Goodman, and P. Winkelstein. *Pediatr Infect Dis J* **2002**, *21*, 791 - 793.

P.J. Foxall, G.J. Mellotte, M.R. Bending, J.C. Lindon, and J.K. Nicholson. *Kidney Int* **1993**, *43*, 234 - 245.

S.H. Gillespie, and I. Balakrishnan. *J Med Microbiol* **2000**, *49*, 1057 - 1067.

M. Goldman. *Quantum Description of High-Resolution NMR in Liquids.* Oxford University Press, Oxford, 1988.

P. Guntert, and K. Wuthrich. *Journal of Magnetic Resonance* **1992**, 96, 403-407.

A. Gupta, M. Dwivedi, G.A.N. Gowda, A. Ayyagari, A.A. Mahdi, M. Bhandari, and C.L. Khetrapal. *NMR Biomed* **2005**, *18*, 293 - 299.

S. Halouska, and R. Powers. *J. Magn Reson.* **2006**, 178, 88.

R.K. Harris. *Nuclear Magnetic Resonance Spectroscopy.* Longman Scientific and Technical, Longman Group: Essex, England, 1983.

A. Heuer, and U. Haeberlen. *Journal of Magnetic Resonance* **1989**, 85, 79-94.

J.F. Hickson, and K. Hinkelmann. *Am J Clin Nutr* **1985**, *41*, 246 - 253.

J.C. Hoch, and A.S. Stern. *NMR Data Processing.* Wiley-Liss Inc.: New York, 1996.

E. Holmes, and H. Antti. *Analyst* **2002**, 127, 1549.

Human Genome Program, U.S. Department of Energy, *Genomics and Its Impact on Science and Society: A 2003 Primer*, 2003.

R.A. Johnson, and D.W. Wichern. *Applied Multivariate Statistical Analysis*, Forth Edition. Prentice Hall Inc.: NJ, 1998.

B. Kan, J. Ries, B.H. Normark, F.Y. Chang, C. Feldman, W.C. Ko, J. Rello, D.R. Snydman, V.L. Lu, and A. Ortqvist. *Clin Microbiol Infect* **2006**, *12*, 338 - 344.

M.D.J. Kerai, C.J. Waterfield, S.H. Kenyon, D.S. Asker, and J.A. Timbrell. *Alcohol and Alcohol* **1999**, *34*, 529 - 541.

S.R. Khan, P.A. Glenton, R. Backov, and D.R. Talham. *Kidney International* **2002**, 62, 2002, 2062-2072.

D.D. Lee, and H.S. Seung. *Nature* **1999**, 401, No. 21, 788-791.

B. Lefebvre, R. Sasaki, G. Golotvin, and A. Nicholls. *Metabolic Profiling: Pathways in Discovery.* Lake Buena Vista, FL, Dec. 13-14, **2004**.

J.C. Lindon, E. Holmes, and J. K. Nicholson, *Anal. Chem.* **2003**, 75, 384A.

J.C. Lindon, J.K. Nicholson, E. Holmes, and J.R. Everett, *Concepts in Magnetic Resonance* **2000**, 12, 289.

B. Lopez, M.D. Cima, F. Vazquez, A. Fenoll, J. Gutierrez, C. Fidalgo, M. Caicoya, and F.J. Mendez. *Eur J Clin Microbiol Infect Dis* **1999**, *18*, 771 - 776.

L.J. Marnett, J.N. Riggins, and J.D. West. *J. Clin. Invest.* **2003**, 111, 583-593.

T.J. Marrie. *Chemotherapy* **2004**, *50*, 11 - 15.

B.M. McGrath, R. McKay, S. Dave, A.M. Weljie, C.M. Slupsky, C.C. Hanstock, A.J. Greenshaw, and P.H. Silverstone, *Neuroscience Research.* **2008**, 61:4, 351-359.

S. Mierisová, and M. Ala-Korpela. *NMR in Biomedicine* **2001**, 14, 247-259.

T.J. Mitchell. *Res Microbiol* **2000**, *151*, 413 - 419.

K.A. Nalecz, D. Miecz, V. Berezowski, and R. Cecchelli. *Mol Aspects Med* **2004**, *25*, 551 - 567.

J.K. Nicholson, E. Holmes, and I.D. Wilson. *Nat Rev Microbiol* **2005**, *3*, 431 - 438.

J.K. Nicolson, J.C. Lindon, and E. Holmes. *Xenobiotica* **1999**, 29, No. 11, 1181-1189.

J.K. Nicholson, and I.D. Wilson. *Nat Rev Drug Disc* **2003**, *2*, 668 - 676.

J.P. Parada, and J.N. Maslow. *Scand J Infect Dis* **2000**, *32*, 133 - 136.

G.A. Pearson. *Journal of Magnetic Resonance* **1977**, 27, 265-272.

L.R. Peterson. *Clin Infect Dis* **2006**, *42*, 224 - 233.

J.G. Proakis, and D.G. Manolakis. *Digital Signal Processing: Principles, Algorithms, and Applications,* Third Edition. Prentice-Hall Inc.: New Jersey, 1996.

C.J. Rebouche. *Ann N Y Acad Sci* **2004**, *1033*, 30 - 41.

M.D. Redelings, F. Sorvillo, and P. Simon. *Public Health Report* **2005**, *120*, 157 - 164.

E.J. Ridgway, C.H. Tremlett, and K.D. Allen. *J Infect* **1995**, *30*, 245 - 251.

J.D. Roberts. *An Introduction to the Analysis of Spin-Spin Splitting in High-Resolution Nuclear Magnetic Resonance Spectra.* W.A. Benjamin, Inc.: New York, 1962.

M.R. Shariatzadeh, J.Q. Huang, G.J. Tyrrell, M.M. Johnson, and T.J. Marrie. *Medicine* **2005**, *84*, 147 - 161.

R.H. Shumway, and D.S. Stoffer. *Time Series Analysis and Its Applications.* Springer-Verlag New York Inc.: New York, 2000.

R. Siuda, G. Balcerowska, and D. Aberdam. *Chemometrics Intell. Lab. Systems* **1998**, 40, 193.

C.M. Slupsky, K.N. Rankin, J. Wagner, H. Fu, D. Chang, A.M. Weljie, E.J. Saude, B. Lix, D.J. Adamko, S.L. Shah, R. Greiner, B.D. Sykes, and T.J. Marrie. *Analytical Chemistry* 2007, 79(18), 6995-7004.

A.K. Tangirala, S.L. Shah, and N.F. Thornhill. *J Process Control* **2005**, *15*, 931 - 941.

T. Tatara, Y. Iwao, J. Takeda, Y. Ishihara, S. Ohkochi, and H. Uedaira. *Clin Chim Acta* **1999**, *279*, 117 - 124.

C.J. Threlfall, A.R. Maxwell, and H.B. Stoner. *J Trauma* **1984**, *24*, 516 - 523.

C.J. Threlfall, H.B. Stoner, and C.S. Galasko. *J Trauma* **1981**, *21*, 140 - 147.

Umetrics AB. *Multi- and Megavariate Data Analysis: Principles and Applications*, Umeå, 2001.

Q.N. Van, J.R. Klose, D.A. Lucas, D.A. Prieto, B. Luke, J. Collins, S.K. Burt, G.N. Chmurny, H.J. Issaq, T.P. Conrads, T.D. Veenstra, and S.K. Keay. *Dis Markers* **2004**, *19*, 169 - 183.

L. Vanhamme, T. Sundin, P. Van Hecke, and S. Van Huffel. *NMR in Biomedicine* **2001**, 14, 233-246.

Q.N. Van, G.N. Chmurny, and T.D. Veenstra. *Biochemical and Biophysical Research Communications* **2003**, 301(4), 952-959.

Y. Wang, E. Holmes, J.K. Nicholson, O. Cloarec, J. Chollet, M. Tanner, B.H. Singer, and J. Utzinger. *Proc Natl Acad Sci USA* **2004**, *101*, 12676 - 12681.

B.J. Webb-Robertson, D.F. Lowry, K.H. Jarman, S.J. Harbo, Q.R. Meng, A.F. Fuciarelli, J.G. Pounds, and K.M. Lee. *J Pharm. Biomed. Anal.* **2005**, 39, 830.

A.M. Weljie, J. Newton, P. Mercier, E. Carlson, and C.M. Slupsky, *Anal. Chem.* **2006**, 78, 4430.

A.M. Weljie, R. Dowlatabadi, R.J. Miller, H.J. Vogel, F.R. Jirik. *Journal of Proteome Research* **2007**, 6(9),3456-3464.

D.S. Wishart. *Am J Transplant* **2005**, *5*, 2814 - 2820.

D.S. Wishart, L.M.M. Querengesser, B.A. Lefebvre, N.A. Epstein, R. Greiner, and J.B. Newton, *Clinical Chemistry* **2001**, 47, 1918.

D.S. Wishart, D. Tzur, C. Knox, R. Eisner, A.C. Guo, N. Young, D. Chen, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D.D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zang, G.E. Duggan, G.D. MacInnis, A.M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B.D. Sykes, H.J. Vogel, and L. Querengesser. *Nucleic Acid Research* **2007**, 35 (Database issue), D521-D526.

Z. Zolnai, S. Macura, and J.L. Markley. *Journal of Magnetic Resonance* **1989**, 82, 496-504.