University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

12-2006

Analysis of NMR spectra using digital signal processing techniques.

Jason Aaron Gearheart
University of Louisville

Recommended Citation

Gearheart, Jason Aaron, "Analysis of NMR spectra using digital signal processing techniques." (2006). *Electronic Theses and Dissertations.* Paper 483.
https://doi.org/10.18297/etd/483

ANALYSIS OF NMR SPECTRA
USING DIGITAL SIGNAL PROCESSING TECHNIQUES

By

Jason Aaron Gearheart
B.S., University of Louisville, 2005

A Thesis
Submitted to the Faculty of the
University of Louisville
J. B. Speed School of Engineering
in Partial Fulfillment of the Requirements
for the Professional Degree of

MASTER OF ENGINEERING

Department of Computer Engineering and Computer Science

December 2006

ANALYSIS OF NMR SPECTRA
USING DIGITAL SIGNAL PROCESSING TECHNIQUES

Submitted by: ________________________________
Jason Aaron Gearheart

A Thesis Approved on

____
(Date)

by the Following Reading and Examination Committee:

____
Dr. Ahmed Desoky, Thesis Advisor

____
Dr. Memhed Kantardzic

____
Dr. Andrew Lane

DEDICATION

To Christy:

I love you.

To Mom and Dad:

Thanks for all of your encouragement and
support through my many years away from home.

To Ted:

This project would not have been possible
without your guidance and understanding of the science.

To Dr. Desoky:

Your knowledge, persistence, and patience
made this thesis a reality.

ABSTRACT

Since its development, nuclear magnetic resonance (NMR) has become one of the primary methods of chemists for structure elucidation, which is the determination of a compound's molecular structure. Current software packages enable scientists to visualize the raw data produced by the spectrometer so that they can manually determine a compound's component parts. This is accomplished by manually comparing the spectrum of the mixture with various reference materials believed to be present.

But these software packages can be expanded to do even more. It is the purpose of this thesis to provide an automated analysis package capable of analyzing a mixture spectrum for the components contained within it and at what concentration. Future enhancements include the development of a centralized database can can archive spectral information for known reference materials and expansion of the proposed method from one dimension to multiple dimensions. The reference materials used tend to be pure compounds.

NOMENCLATURE

^{1}H	=	Protium (the most common isotope of Hydrogen)
^{2}H	=	Deuterium
^{13}C	=	Carbon-13
ASCII	=	American Standard Code for International Interchange
DSP	=	Digital Signal Processing
DSS	=	2,2-Dimethyl-2-silapentane- 5-sulfonate sodium salt
FID	=	Free Induction Decay
NMR	=	Nuclear Magnetic Resonance
PPM	=	Parts Per Million
TMS	=	Tetramethylsilane ($(CH_3)_4Si$)
1D NMR	=	One-dimensional Nuclear Magnetic Resonance
2D NMR	=	Two-dimensional Nuclear Magnetic Resonance

TABLE OF CONTENTS

APPENDIX

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Throughout the years, NMR has become the primary form of analysis for chemists. The methods used today for performing NMR experiments have improved greatly since NMR was discovered. However, is it required that trained spectroscopists still do much of the work. Even though several software packages are available for visualizing NMR spectra, trained spectroscopists must still perform visual analysis of the spectra. In other words, the vast majority of software packages focus on manipulating the data for visualization, not analysis.

However, there are a few software packages capable of determining the molecular structure of a compound, known as structure elucidation. This thesis establishes a new method that is capable of both determining the compounds that comprise a mixture and calculating the concentration of those compounds. To accomplish this, several digital signal processing (DSP) techniques are used to analyze data produced by the NMR spectrometer for its component parts.

Three different methods of component analysis are considered, one from the time domain and two from the frequency domain. In the time domain, convolution is believed to be the best method for analysis. In the frequency domain, the first method considered compares the energy in a reference spectrum against the energy in a mixture spectrum; the second method uses square-waves combined with comparison of the energy in each spectra. Analysis shows that the second frequency domain method using square-waves proved to produce the best results.

This thesis also provides several future enhancements to be researched, such as a storage method for reference compounds that will make analysis easier and expansion of the developed method into multiple dimensions. The Biological Magnetic Resonance Data Bank (BRMB) is a library of raw NMR data.

II. LITERATURE REVIEW

Nuclear magnetic resonance (NMR) spectroscopy was first discovered by two physicist research groups; (1) F. Bloch, W.W. Hansen, and M. E. Packard and (2) E. M. Purcell, H. C. Torrey, and R. V. Pound. (Friebolin, 1998, pg. 1) Four years later, it was discovered that the transitions between the energy levels of a compound were dependent upon the environment. This led to the discovery of the chemical shift, as well as the use of NMR for molecular structure elucidation. (Nelson, 2003, pg. 1) NMR spectroscopy has since morphed itself into "the most direct and general tool for identifying the structure of both pure compounds and mixtures as either solids or liquids." (Lambert and Mazzola, 2004, pg. 1)

2.1. Current Analysis Methods

Before a sample may be processed by a NMR spectrometer, it must first be dissolved in a solvent chosen by the spectroscopist. The selected solvent must have no resonance in the expected frequency region of the sample. This frequency range is highly dependent upon the type of spectrum being obtained, either hydrogen (^{1}H) or carbon (^{13}C). Since NMR experiments can only check for one type of nucleus at a time, a deuterated solvent produces the best results when obtaining a ^{1}H or ^{13}C spectrum because it provides a deuterium (^{2}H) lock signal. (Lambert and Mazzola, 2004, pg. 33) Once the sample has been dissolved, the solution is placed in the spectrometer for analysis.

2.1.1. FID to Spectrum

Free induction decays (FID), or time-domain signals, represent the raw data produced by a NMR spectrometer. The intensity of the signal decreases with time due to spin-spin relaxation in the sample. (Macomber, 1988, pg. 26) An example of an FID is shown in Figure 2.1.

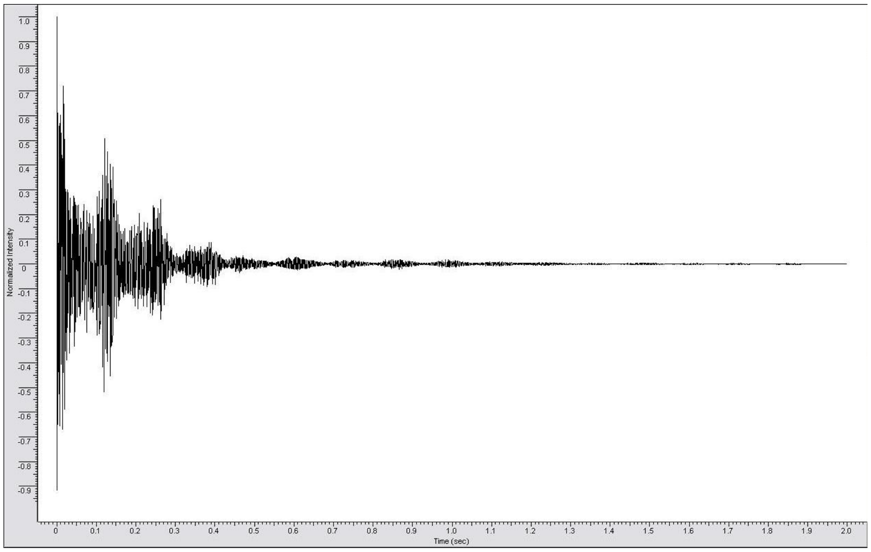

FIGURE 2.1 – An example FID of tryptophan.

Once a FID has been acquired, it may undergo preprocessing before further analysis can be performed. There are two main parts to the FID; the front part, on the left, contains the majority of the intensities to be observed, while the tail, on the right, contains the majority of the resolution. The first step in preprocessing is to apply a weighting function based on the primary concern of the experiment: sensitivity or resolution. (Lambert and Mazzola, 2004, pg. 48)

In ^{13}C NMR, sensitivity is typically the primary concern and involves an exponential weighting function (Figure 2.2). An exponential function suppresses the tail of the FID while emphasizing the front. This results in the loss of resolution, but improved signal to noise ratio. As a consequence of the loss of resolution, this type of weighting results in some line-broadening. As such, these functions are commonly referred to as line-broadening functions. (Lambert and Mazzola, 2004, pg. 48) An exponential function is the most common form of a sensitivity enhancement weighting function, but any function that increases with time could be used.

Alternatively, in ^{1}H NMR, resolution is the primary concern. Because of the increased complexity, the weighting function (Figure 2.3iii) is a combination of

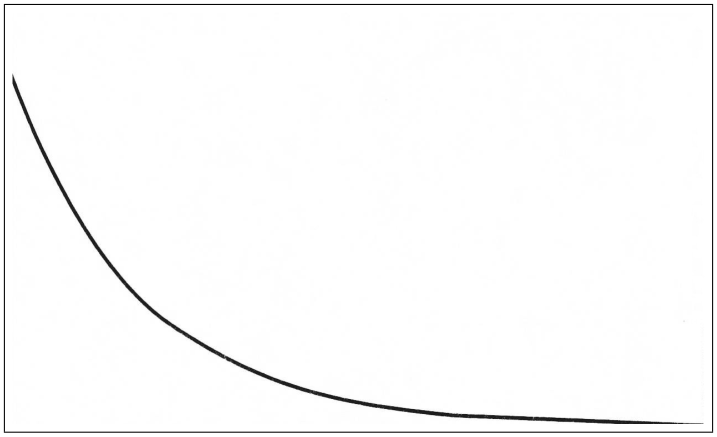

FIGURE 2.2 – Sensitivity enhancement function. (Lambert and Mazzola, 2004, pg. 49)

two key functions: typically a negative line-broadening function (Figure 2.3i) and a Gaussian function (Figure 2.3ii) are used to create this complex function. Contrary to the sensitivity weighting function, resolution is improved at the expense of signal to noise ratio. This type of signal is also known as a double exponential, and can be calculated by $e^{t/k} \times e^{-t^2/l}$ where k and l are found empirically for all spectra. (Nelson, 2003, pg. 36) The weighting function in 2.3iii is only an example of one of the many weighting functions that can be used to enhance the resolution. Many, however, do employ a shifted Gaussian function in them. Also, all resolution enhancement weighting functions must have an amplitude of one at $t = 0$.

Because of the difficulty involved with interpreting a FID, scientists would prefer to examine the frequency components involved in the FID rather than the raw time-domain signal. To accomplish this, the FID must be transformed to a frequency domain spectrum using a Discrete Fourier transformation (see Figure 2.4). (Lambert and Mazzola, 2004, pg. 13) Zero filling is used to double the length of the FID, by adding a zero signal to the tail of the FID. This is required to ensure that no experimental data is lost in the Fourier transform, as Fast Fourier transforms, which are most commonly used, require the number of data points to be a power of two. Also, if the FID has been truncated, doubling the length using zero filling ensure the

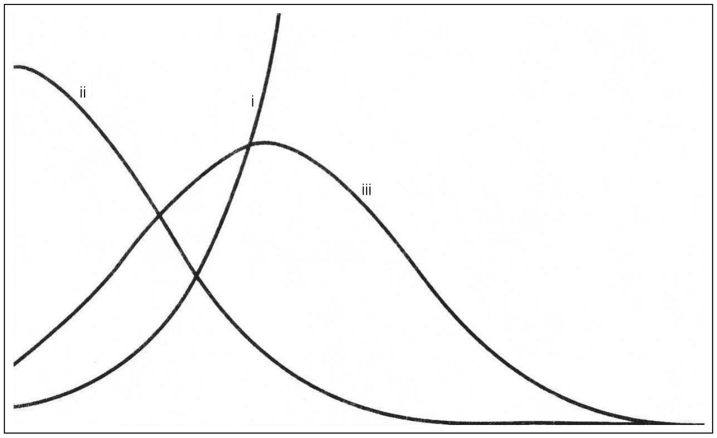

FIGURE 2.3 – Resolution enhancement function containing (i) a negative line-broadening function, (ii) a Gaussian function, and (iii) the resultant function by combining (i) and (ii). (Lambert and Mazzola, 2004, pg. 49)

imaginary part has not been lost.(Lambert and Mazzola, 2004, pg. 49)

Once the transformation is complete, the spectrum usually must be phase corrected. There are two types of phase correction. If the phasing error is almost constant throughout the spectrum, a zero-order phase correction is applied. If the phasing error is dependent upon the frequency of the spectral line (error increases as frequency increases), then a first-order phase correction is applied. In most cases, both types of phase correction will are required. (Lambert and Mazzola, 2004, pg. 53)

Once phase correction has been completed, the spectrum is ready for analysis. In order to properly analyze the spectrum, a common unit of measurement must be used. Leaving the spectrum in *hertz* is a poor choice because the applied frequency must be given for the graph to have any meaning. Therefore, the δ, or part per million (ppm), scale is used. The δ-scale is a dimensionless scale that provides the ability to normalize all spectra to a common unit of measurement. (Silverstein et al., 1991, pg. 173) Conversion of a spectrum from hertz to ppm is accomplished through Equation 2.1.

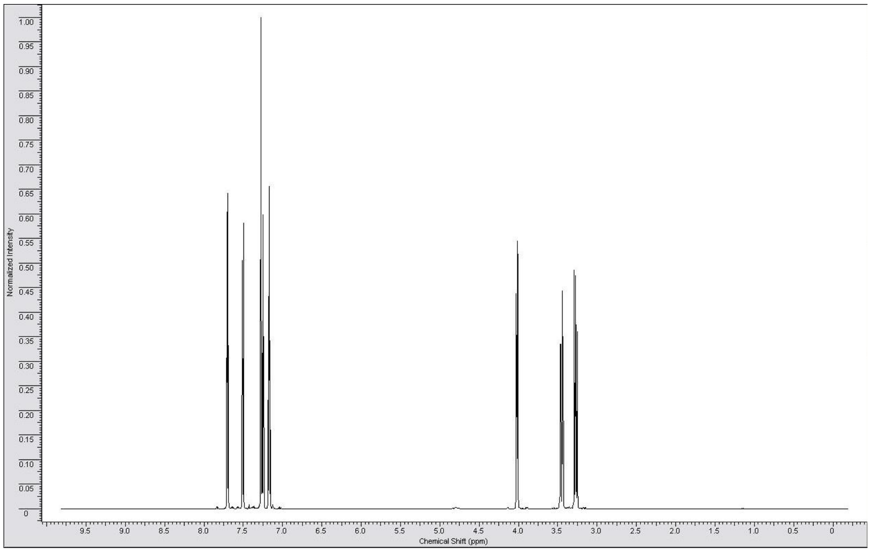

FIGURE 2.4 – An example spectrum of tryptophan.

$$\delta = \frac{\nu}{\text{frequency applied to obtain FID}} \times 10^6 \tag{2.1}$$

where:

$$\nu = \text{Frequency to convert}$$

2.1.2. Analysis Technique Considerations

Several considerations need to be made when performing analysis. First, a zero point of reference for the δ-scale must be established. While the zero point may drift with respect to absolute frequency, it will always be in relation to the remainder of the spectrum. (Macomber, 1988, pg. 42) The most commonly used reference for the zero point is tetramethylsilane (TMS) in organic compounds. In most cases, TMS is ideal because it has such a low frequency that few materials exist to its right in a spectrum. (Lambert and Mazzola, 2004, pg. 55) TMS is also suitable because it is “soluble in most organic solvents, is unreactive, and is volatile.” (Lambert and Mazzola, 2004, pg. 6) However, with the introduction of more sensitive spectrometers, TMS is now being used as a secondary reference. (Lambert and Mazzola, 2004, pg. 55) However, another compound, 2,2-Dimethyl-2-silapentane- 5-sulfonate sodium salt (DSS), is better to use for aqueous solutions since TMS is not soluble in water. (Markley et al., 1998, pg. 130)

The appearance of multiplets - a signal with two or more peaks - is caused by a phenomenon known as the spin-spin coupling. The number of peaks in a given multiplet is determined by adding one to the number of equivalent nuclei coupled to the current nucleus. The intensity of the peaks in the multiplet follows the coefficients of the binomial distribution, also known as Pascal's Triangle. When a nucleus, n, is coupled with multiple groups of other nuclei, it will be split multiple times based on the aforementioned rule. The order in which n is split is based on the coupling constants. The largest coupling constant splits first, followed by the remaining coupling constants in descending order. (Macomber, 1988, pg. 98) The number of peaks resulting from the splitting of n follows the formula $2s + 1$, where s is the spin state of the coupling atom. At lower resolutions, accidental equivalence can cause several different atoms to appear equivalent, thereby reducing the multiplets to a single peak. This can be reconciled by increasing the field strength and operating frequency of the spectrometer. (Macomber, 1988, pg. 47)

Second-order effects can also affect the interpretation of data. These effects cause asymmetry in multiplets in which inner lines, those closest to the other coupled multiplet, have a higher intensity than outer lines. In order to reduce the second-order effects, a higher field strength can be used when acquiring the original FID. If using a higher field strength is not possible, then it is possible to classify spin systems as either weakly coupled or strongly coupled by comparing the difference in the chemical shift between the two multiplets, $\Delta\nu$ (in Hz), and the absolute value of the coupling constant, J (in Hz). Weakly coupled systems, defined by $\Delta\nu/J > 10$, exhibit less second-order effects than strongly coupled systems, defined by $\Delta\nu/J < 10$. Strongly coupled systems require additional attention to account for the increased probability of second-order effects. (Macomber, 1988, pg. 122)

The intensity of the peak is not equal to its amplitude. Instead, the intensity of a peak is equivalent to the integrated area under the curve. This can be approximated by multiplying the amplitude of the peak by the full-width at half-height of the peak. Alternatively, to obtain a more precise measurement, most modern spectrometers have the ability to integrate any peak within the acquired spectrum. (Macomber, 1988, pg. 43)

By visually comparing the spectrum of a reference material to the spectrum of a mixture, spectroscopists can estimate what compounds compose the mixture. Alternatively, spectroscopists can identify a compound by comparing the shielding parameters of groups of peaks with a database of known shielding parameters. The parameters can give the expected chemical shift of a hydrogen within ±0.2 ppm. The spectroscopists can then compare the findings with the actual spectra of these structures to confirm the identification. (Macomber, 1988, pg. 56)

2.2. Current Software Packages

There are several software packages currently used to analyze NMR spectra. All are capable of performing basic operations, such as the application of a weighting function, zero-filling, Fourier transformation, and phase correction on both 1D and 2D NMR. However, some programs offer additional functionality. The proceeding sections evaluate three of these applications:

- NMRPipe (`http://spin.niddk.nih.gov/NMRPipe/`)
- MestRe-C (`http://www.mestrec.com/`)
- ACD/Labs (`http://www.acdlabs.com/`)

2.2.1. NMRPipe

NMRPipe is a free, Linux-only application. It does not perform many advanced functions, but does has the ability for batch processing through command line execution. It can also convert between several common formats, including ASCII. Additionally, NMRPipe can produce a peak-picked list of the spectrum, which is useful for isolating noise from actual data. This peak list can be exported as an ASCII file, so that it can be read by virtually any program. NMRPipe is also able to create batch processing files that make it easier to repeat the processing if required.

However, the documentation for NMRPipe provides little insight into output parameter meanings. NMRPipe can also be complicated to both install and use. For example, the menu system is counter intuitive by requiring the right mouse button be used to access the menu as opposed to the left mouse button being typically used by the operating system settings. Finally, it is important to note that some application components are not installed by default.

2.2.2. MestRe-C

MestRe-C is a Windows®-only application that provides much of the same functionality as NMRPipe. Similar to NMRPipe, it is capable of reading several different data formats and of producing a peak-picked list of the spectrum.

However, MestRe-C is more subjective in the selection of parameters for analysis. With NMRPipe, one can specify definitive thresholds which to test against. Conversely, MestRe-C requires the user to specify thresholds by using the mouse to set the limits. This introduces a variability between materials that can potentially distort the analysis. Since MestRe-C provides all of the functionality of NMRPipe but is more complex to use, NMRPipe is recommended over this product, assuming there is no bias to a given operating system.

2.2.3. ACD/Labs

ACD/Labs is the most powerful tool tested. It is a Windows®-only application capable of completing all NMRPipe and MestRe-C functionality, as well as several additional functions. For example, ACD/Labs is capable of producing a peak-picked list of the spectrum, compute the shifts and coupling constants, calculate the integrated area under the curve, and compare the clusters in the spectrum to a database of known clusters in order to predict the molecular structure of the compound. Additionally, it has the ability to predict a spectrum, its shifts, and its coupling constants from a molfile. A molfile is a special file format containing information about the compound, such as atoms included and their bonds. (MDL, 2006) Finally, ACD/Labs enables the user to create a unique compound and predict its spectrum.

With all of its capabilities, ACD/Labs is complex and confusing to use. For example, there are several options a user can select in computing the Fourier transform of an FID. ACD/Labs will also automatically normalize the data so that the maximum peak in either the FID or spectrum has an amplitude of one. Finally, as to be expected, ACD/Labs has a significantly greater cost for its additional functionality over the previous two packages.

2.2.4. Other Software Packages

There are several other applications available for NMR processing, each offering a variety of features. These packages were not evaluated, and should be considered in determining the appropriate application in ones particular needs. Some of these applications include AcornNMR, XWIN-NMR, matNMR, Azara, and Chenomx.

III. ANALYSIS PROCEDURES

Before analysis can be performed, it must be determined if it will be completed in the time or frequency domain. Both domains have strengths and weaknesses. Within the time domain, analysis could theoretically be performed faster. Additionally the raw FID produced by the spectrometer is already in that domain. However, it is unknown how to adjust for physical attributes, such as pH and temperature, within the time domain. Conversely, while the frequency domain is slower and requires that each sample be transformed before analysis, it is known how to adjust for physical attributes. Furthermore, spectroscopists are accustomed to working within the frequency domain. Regardless of which domain is selected, both require the use of signal processing techniques to aid in analysis.

Because analysis in the time domain is faster and requires less computational effort, it is considered the ideal analysis domain. As such, it was initially selected and tested with the convolution method. However, since analysis proved to be too difficult, the frequency domain was considered the only viable alternative. The energy analysis method and the square-wave analysis method were both tested. To test each of the three methods, a known mixture of amino acids, with concentrations given in Table 3.1, is used. If the method can predict within 10 percentage points each of the references involved, then the method is considered plausible for analysis. All three of these methods are discussed in the proceeding sections.

3.1. Time Domain Analysis

In the time domain, several different methods seemed probable. Correlation was not used in the time domain because it compares two signals to determine if they are equal to or related to one another. Pearson's correlation coefficient (Equation 3.1 (Woolson, 1987, pg. 270)) is a value ranging between negative one and one that

TABLE 3.1

The mixture provided is known to be made up of glutamine (gln), lysine (lys), and tryptophan (trp) in the concentrations given.

Reference	**Concentration**
Gln	48.0%
Lys	27.0%
Trp	6.0%

shows the extent to which two signals are linearly related. A zero value means that the signals are completely unrelated, one means the signals are positively related, and negative one means the signals are negatively related.

$$r = \frac{\sum_{i=1}^{N} \left(X[i] - \bar{X}\right)\left(Y[i] - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{N} \left(X[i] - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{N} \left(Y[i] - \bar{Y}\right)^2}} \tag{3.1}$$

where:

X and Y represent the signals being correlated.

When comparing raw FIDs, the correlation coefficient is useful in determining if a reference material exists in a mixture material. However, it is unable to give the concentration. This led to the use of convolution.

3.1.1. Method 1: Time domain Convolution

The standard deviation of the convolution of a reference spectrum and a mixture spectrum divided by the standard deviation of the mixture spectrum will produce a close approximation of the percentage of the reference material in the mixture. Since the frequency of the reference is constant throughout, it is not necessary to use the entire FID for analysis, therefore making it possible to use a smaller number of data *points when computing the convolution. Normalization of the two spectra may be required to obtain reliable results.*

Equation 3.2 (Hsu, 1984, pg. 115) calculates the convolution of the mixture (X_1) and reference (X_2) spectra. The standard deviations (Equation 3.3 (Woolson, 1987, pg. 20)) of both the resulting convoluted signal and the mixture spectrum are then computed. The concentration of the reference material in the mixture can then be computed by Equation 3.4.

$$F[i] = \sum_{k=-\infty}^{\infty} X_1[k] \times X_2[i-k] \tag{3.2}$$

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{N} \left(X[i] - \bar{X}\right)^2}{N-1}} \tag{3.3}$$

where:

X is a signal in either the time domain or the frequency domain.

$$Concentration = \frac{\sigma_{convolution}}{\sigma_{mixture}} \tag{3.4}$$

Testing this method using the known amino acid mixture does not produce accurate results. This is because the peak amplitudes of the mixture spectrum are considerably larger than those in the reference spectrum. This holds true even when only one reference contributes to the peak. Therefore, the amplitudes of the spectra must be normalized so the variances of the spectra are equal to one. Normalization is completed by Equation 3.5. However, even after the spectra are normalized, convolution still does not provide an adequate means of approximating the concentration of a reference material in a mixture.

$$X'[i] = (X[i] - \bar{X})/\sigma_X \tag{3.5}$$

where:

X is the signal being normalized.

TABLE 3.2

Calculated concentrations of the three known references in the mixture using the convolution method in the time domain. These concentrations were calculated using three different subsets of the reference spectra.

Reference	**Calculated Concentration**		
	500 data points	5000 data points	all data points
Gln	0.13%	0.31%	0.34%
Lys	0.07%	0.23%	0.24%
Trp	0.07%	0.23%	0.26%

3.2. Frequency Domain Analysis

After convolution failed, it was decided to test methods in the frequency domain. The frequency domain provides more opportunities for analysis. Calculating the correlation coefficient (Equation 3.1) was thought to be able to produce an estimate of the reference concentration in a mixture. However, testing shows that while the correlation coefficient indicates whether the reference material is in the mixture, it can not indicate at what concentration.

3.2.1. Method 2: Frequency domain Energy Analysis

By removing the noise in both the mixture and reference spectra, a more accurate calculation of the energy in each spectrum can be computed. Then, by comparing the energy in the reference to the energy in the mixture, it is believed to be possible to acquire a close approximation of the reference material concentration in the mixture.

The resolution of the spectra causes noise from surrounding peaks to produce skewed results. To adjust for this, a simple high-pass filter was applied to the spectrum to remove the noise. This is done by incrementing through each data point in the reference spectrum, and outputting the corresponding data point in the mixture spectrum into a new reference spectrum if it is above a certain threshold. This

TABLE 3.3

Calculated concentration results for the three known references in the mixture using the energy analysis method in the frequency domain.

Reference	**Calculated Concentration**
Gln	42.97%
Lys	9.77%
Trp	12.57%

threshold is set at approximately 17.5% of the maximum amplitude of the spectrum. This threshold value was found through trial and error and can be adjusted if needed. This threshold must also be applied to the mixture to remove noise. Once the new reference and mixture spectra are computed, the energy of both can be calculated and compared using Equation 3.6. (Hsu, 1984, pg. 45) Ideally, the percent composition can then be determined by Equation 3.7.

$$Energy = \sum_{i=-\infty}^{\infty} X[i]^2 \tag{3.6}$$

where:

X is a spectrum.

$$Concentration = \frac{Energy_{reference}}{Energy_{mixture}} \tag{3.7}$$

As can be seen in Table 3.3, this method produces a closer, but still incorrect, estimate of the concentration ratios. While still incorrect, the threshold analysis and masking the mixture using the reference, produces more refined results and thus could be a key component in final analysis method.

3.2.2. Method 3: Frequency Domain Square-Wave Analysis

Generate a square-wave, with maximum amplitude of one, based on the positions of the peaks in the reference spectrum. Multiply the square-wave by the mixture *spectrum to isolate the reference material in the mixture. Using this new reference spectrum, compare the energy in the reference and mixture spectra using Equation 3.7. This produces a close approximation of the reference material concentration in the mixture.*

Instead of using a high-pass filter, a square-wave is generated such that for each peak in the reference spectrum, a square encompassing the entire peak with an amplitude of one is generated (see Figure 3.1). This square-wave is then multiplied by the mixture spectrum to produce the contribution of the reference spectrum toward the mixture spectrum. Comparing the energy in the resultant and original mixture spectra using Equation 3.7 produces the concentration within ±6% of the actual concentration. This is considered to be close enough for the spectroscopists, who can then refine the results into more accurate information.

In order to remove the effects of noise on the testing and ease computation, a third-party program (ACD/Labs, see Section 2.2.3) is used to produce a peak list for both the mixture and reference spectra. These peak lists are then used to recreate the spectra assuming a Lorentzian line shape (see Equation 3.8) for the peaks. An example of a Lorentzian line shape is shown in Figure 3.2 with $\alpha_L = 0.5$, $\nu_0 = 6$, and $V = 2$.

$$\phi(\nu) = \frac{1}{\pi} \frac{\alpha_L * V}{(\nu - \nu_0)^2 + \alpha_L^2} \tag{3.8}$$

where:

$$\begin{aligned} \alpha_L &= \text{Half-width at half-height} \\ \nu_0 &= \text{Center point of the peak} \\ V &= \text{Area under the peak} \end{aligned}$$

Since a peak list is provided, peaks from the reference spectrum are matched with a one-to-one relation to peaks in the mixture spectrum using the nearest neighbor technique within a certain threshold. A new spectrum following the Lorentzian line

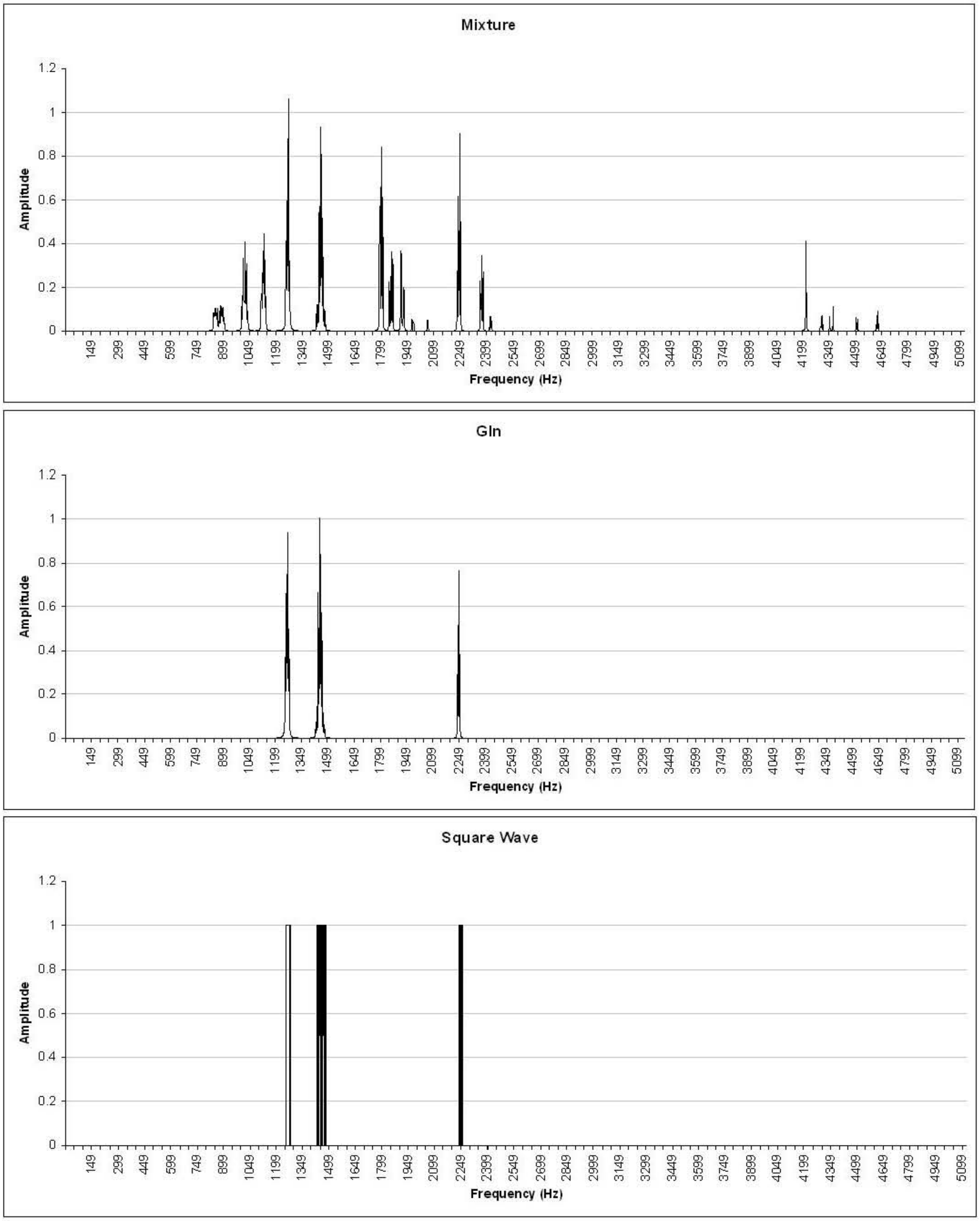

FIGURE 3.1 – Example of how a square-wave would look if generated by glutamine. By comparing the spectrum of gln to the mixture spectrum, it is apparent that gln does exist in the mixture.

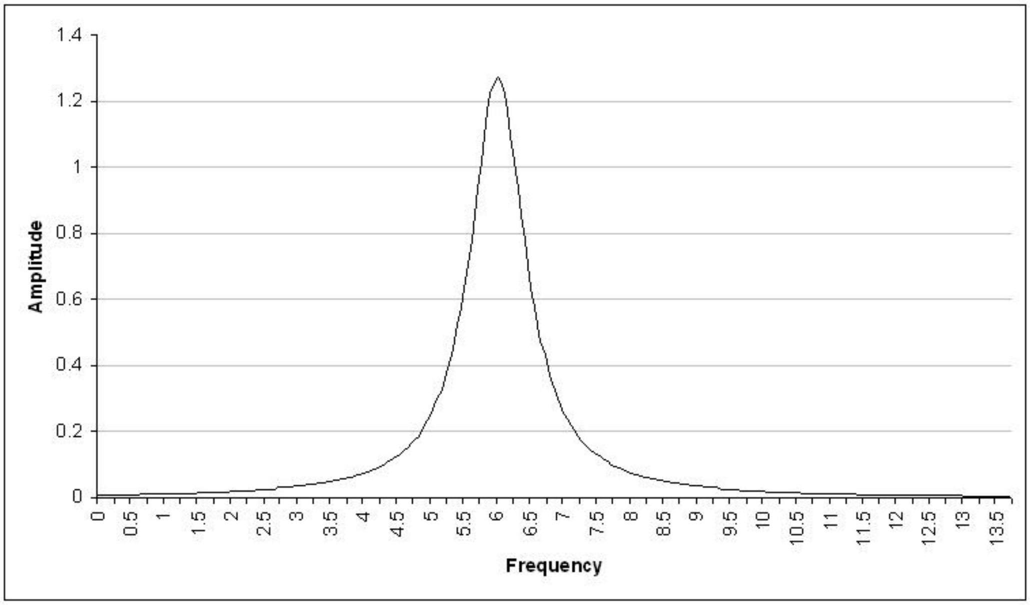

FIGURE 3.2−Lorentzian line shape with $\alpha_L = 0.5$, $\nu_0 = 6$, and $V = 2$.

shape is generated using the matched peaks. Increased accuracy is achieved because peaks in the mixture that are not in the reference are removed, and therefore have no influence on the new spectrum.

The aforementioned method produces accurate results for spectra that are aligned, but generates incorrect results for spectra that are not aligned before analysis. In order to properly align the spectra, the mixture spectrum is shifted and then compared to the reference spectrum. Since it is unnecessary for the mixture spectrum to be shifted throughout the entirety of the reference spectrum, a window of size 400 Hz is used. This increases computational speed as well as lowers the chance of a false positive.

For each comparison, the mixture spectrum is shifted by 1Hz through the entire window range, ±200 Hz of the original position. A high-pass filter is then applied to both the mixture and the reference spectra. After the filter has been applied, the correlation coefficient (Equation 3.1) is calculated and stored along with the shift value from the original point that produces it. Once the mixture spectrum

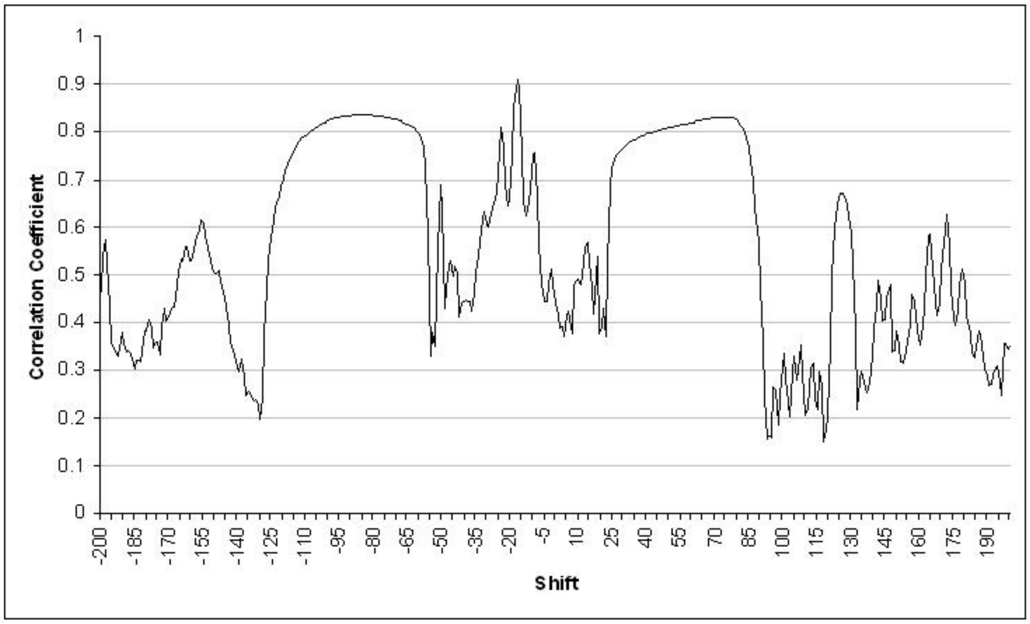

FIGURE 3.3 – Correlation coefficients related to shift.

has completed all comparisons, the mixture is shifted by the value that corresponds with the maximum correlation coefficient. This ensures that the mixture and reference spectra are properly aligned when the square-wave is generated and multiplied by the mixture spectrum. Figure 3.3 shows a plot of the correlation coefficients for each shift when comparing glutamine (Gln) to the known amino acid mixture. The maximum correlation coefficient of Gln with the mixture is 0.9084 and corresponds with the shift value of -17.

As shown in Table 3.4, this method produces results much closer to the actual concentrations, with the exception of tryptophan (Trp). This is due to the fact that the physical attributes, such as temperature and pH balance, at which the FID of Trp was collected are significantly different than those at which the mixture was collected.

However, this method can produce false positives. As in the case of ornithine (Orn) and phosphoarginine (Parg) shown in Table 3.4, results show their presence in the mixture, when in actuality they do not exist. Therefore, it is important to consider the percentage of peaks in the reference that match peaks in the mixture.

TABLE 3.4

Preliminary results of the square-wave analysis method. Calculating the concentration alone does not provide enough information to determine the composition of a mixture.

	Reference	**Calculated Concentration**
Actual Components	Gln	54.47%
	Trp	0.54%
	Lys	30.42%
Other Components	Orn	16.03%
	Pyr	0.10%
	Nad	0.00%
	Gsh	0.46%
	Parg	11.49%

Only reference materials with a high percentage of matched peaks, with a threshold of 60% or greater, should be considered. The new results produced by using this metric are shown in Table 3.5. A false positive is still possible if the number of peaks in the reference is low.

In conclusion, this method consists of three steps: preprocessing, direct analysis, and additional considerations. In the preprocessing step, the spectra are generated from the peak list using a Lorentzian line shape and aligned to the point of maximum correlation. In the direct analysis step, the square-wave is generated and applied to the mixture spectrum to remove all peaks that do not exist in the reference spectrum. Finally, in the additional considerations step, the percentage of matched peaks is used to reduce the number of false positives produced by this method.

3.3. Analysis Refinements

Even though the frequency domain square-wave analysis method has the po-

TABLE 3.5

Results for Frequency domain Square-wave Analysis with Matched Peaks

	Reference	**Calculated Concentration**	**Matched Peaks**
Actual Components	Gln	54.47%	100.0%
	Trp	0.54%	56.5%
	Lys	30.42%	84.6%
Other Components	Orn	16.03%	22.2%
	Pyr	0.10%	20.9%
	Nad	0.00%	0.0%
	Gsh	0.46%	7.7%
	Parg	11.49%	43.9%

tential to accurately analyze mixtures, it needs some additional refinement. These adjustments include modifying the spectra to account for physical attributes such as pH balance and temperature, removing noise, and adjusting the frequencies of the spectra.

3.3.1. Physical Attributes

When the physical attributes of a mixture spectrum differ from those of a reference spectrum, all of the peaks will not align correctly. A variation in the physical attributes causes the spectrum to be more spread out. The spreading is a result of an increase in the coupling constants with respect to frequency as opposed to an increase in the center point of the peak with respect to frequency.

This preceding method looks only at the spectrum in its entirety. It is unaware of how the peaks are produced, and therefore unaware of the clusters that are created. As such, it is unable adjust for the coupling constant. For more information on the coupling constant, see Section 2.1.2.

3.3.2. FID Processing

Currently, a third-party software package is required to remove the noise in the spectrum. Eventually, an integrated software tool could be able to take a raw FID and perform all of the required preprocessing to produce a usable spectrum. This preprocessing includes line smoothing, noise reduction and transforming the FID into a spectrum in the frequency domain.

3.3.3. Adjusting Frequencies

If a reference material and a mixture are sampled at different frequencies, the reference spectrum will not align with the mixture spectrum. This can result in either a false positive or a false negative. To correct this, NMR spectroscopists use the δ scale as opposed to hertz (see Section 2.1.1).

When the FID is produced by the spectrometer, it will also produce a file listing the attributes of the experiment, such as temperature, applied frequency, and pH balance. Unfortunately, there is no readily available documentation as to how to properly interpret this file. A spectroscopist must adjust the spectra prior to using any analysis method.

3.3.4. Peak matching

During the peak matching process, peaks are assigned on a first-come, first-served basis; when a peak in the mixture spectrum is assigned to a peak in the reference spectrum, it is no longer considered for any other peak. This can lead to inaccurate analysis by eliminating peaks too early. By rechecking each peak in the mixture to ensure it is matched with the closest peak in the reference, a more accurate analysis can be obtained.

By aligning the reference and mixture spectra at the point of maximum correlation, this problem occurrence is reduced. However, because there is no adjustment for physical attributes, this can cause erroneous analysis calculations.

3.3.5. Overlapping Peaks

It is possible for a mixture peak to exist in multiple references. It is left to the spectroscopist to find and correct this error.

IV. FUTURE ENHANCEMENTS/RESEARCH

This thesis developed an analysis method for determining the concentration of a reference in a mixture in the frequency domain. Two key enhancements are the creation of a central repository of references and the modification of the analysis method to process multidimensional spectra.

4.1. Repository of References

The creation of a central reference database in a standardized format could improve research. In order to effictively create this database, reference spectra must be subdivided into functional groups. A functional group is the multiplet created from an atom or group of chemically equivalent atoms (see Section 2.1.2). This eases the manipulation of the spectrum to account for variations in physical attributes such as pH and temperature, as well as assists in isotopomer analysis. Using a molfile and the third-party software package ACD/Labs (Section 2.2.3), these functional groups can be generated relatively easily.

A molfile is a standardized file format that holds all information related to the chemical structure of a compound. This format was introduced and standardized by Elsevier MDL, and consists of three major sections: the Counts Line, the Atom Block, and the Bonds Block. The Counts Line contains basic information about the compound as well as the version number of the molfile. The Atom Block contains the information about the individual atoms, such as symbol, charge, and hydrogen count. Finally, the Bonds Block contains information about the bonds between atoms. For more information, the format description is available from Elsevier MDL's website. (MDL, 2006)

Using the molfile, ACD/Labs is capable of predicting the spectrum for the compound, as well as a list of coupling constants and chemical shifts. However, this

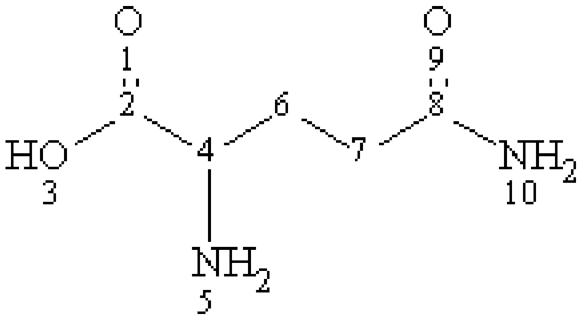

FIGURE 4.1 – Molecular structure of glutamine

prediction does not always mimic actual compounds. Therefore, the predicted coupling constants need to be compared to an actual compound spectrum and adjusted accordingly. Once the adjustment is complete, the chemical shift and the coupling constants associated with each atom in the Atom Block becomes a functional group. Figure 4.1 shows the structure of glutamine: Tables 4.1 & 4.2 show the coupling constants and the chemical shifts corresponding with the molfile in Figure 4.2, as predicted by ACD/Labs.

Once all functional groups have been created, it is possible to store each reference as a combination of the molfile representing that compound and a list of corresponding functional groups. While these functional groups are created under ideal circumstances, they are easily manipulated to correct for varying physical attributes such as pH and temperature. These corrections could then be completed by modifying the chemical shifts and coupling constants of the functional groups, and letting the software package recreate the spectrum, as opposed to modifying each data point.

These functional groups can also be modified to perform isotopomer analysis. In isotopomer analysis, one hydrogen atom is replaced with deuterium, a hydrogen isotope with a spin of 1 instead of $\frac{1}{2}$. When the spectrum is produced, the multiplet of the deuterium will be a triplet instead of a doublet. This enables the scientist follow the specific atom through other experiments.

TABLE 4.1

Coupling constants for glutamine

Jn	Nucleus 1	Nucleus 2	Value (Hz)	Error
J2	3,5	3,5	2.88	1.45
J3	3,5	4	8.10	0.00
J4	3,5	4	0.00	0.00
J4	3,5	6<’>	0.50	0.00
J4	3,5	6<’’>	0.50	0.00
J3	4	6<’>	7.60	0.00
J3	4	6<’’>	2.60	0.00
J4	4	7	-1.00	1.80
J2	6<’>	6<’’>	14.10	0.00
J3	6<’>	7	7.30	0.00
J3	6<’’>	7	7.30	0.00
J2	7	7	-16.55	2.36
J4	7	10	0.50	0.00
J2	10	10	2.67	0.13

```
 10  9  0  0  0  0  0  0  0  0  1 V2000
    5.0038   -3.9826    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    5.0038   -5.8311    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.4162   -6.7662    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    6.6130   -6.7444    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    6.6348   -9.0279    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    8.2223   -5.8311    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    9.8316   -6.7444    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   11.4191   -5.8311    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   11.4191   -3.9826    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   13.0284   -6.7662    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
  6  4  1  0  0  0  0
  2  4  1  0  0  0  0
  5  4  1  0  0  0  0
  7  6  1  0  0  0  0
  1  2  2  0  0  0  0
  3  2  1  0  0  0  0
  8  7  1  0  0  0  0
 10  8  1  0  0  0  0
  9  8  2  0  0  0  0
M  END
```

FIGURE 4.2 – Molfile for glutamine.

TABLE 4.2

Shift values for glutamine

Nucleus	nH	Shift Value(ppm)	Error
3,5	3	6.78	2.65
4	1	3.83	0.33
6<’>	1	1.86	0.14
6<’’>	1	1.94	0.11
7	2	2.17	0.02
10	2	6.93	0.25

4.2. Multidimensional Spectral Processing

A second enhancement to the frequency domain square-wave analysis method would be the expansion to multiple dimensions. Spectroscopists prefer to use 2D NMR because of the higher resolution and greater amount of information provided over 1D NMR. In order to expand this method into multiple dimensions, there are some additional considerations that need to be made.

Most programs that are capable of generating a peak list, such as those listed in Section 2.2, are also capable of reading and analyzing 2D NMR data. The methods for importing data need to be altered to allow for the additional axis. Also, since there are two center points and two full-width at half-height measures, the peak variables need to be expanded to hold the additional data.

Generating the spectrum, applying the threshold, and shifting the mixture, matching the peaks need to be completed in both dimensions. Correlation can still be used, but with slight modification; the data needs to be in either row major order or column major order, and the two spectra would need to be the same size in both dimensions. If the spectra are different sizes, zero-padding can be used to eliminate this difference.

Finally, the energy comparison needs to change. Instead of calculating the area under the curve, the energy comparison needs to calculate the volume under the surface. This is a relatively small change, as most software packages already calculate both the area and volume of the peak.

4.3. Convolution

It is believed to be possible to perform analysis using convolution within the time domain. Convolution is capable of calculating the amount of overlap between two signals. Because the mixture FID must contain all of the reference FIDs that compose it, it should theoretically be possible to use convolution to calculate the overlap between a reference FID and the mixture FID and to deduce the concentration of the reference contained within.

V. CONCLUSION

Of the three methodologies considered, only one method could accurately produce close estimations of the actual concentrations. This method, frequency domain square-wave analysis, consists of four primary steps:

1. Produce the peak-picked lists of the spectra.
2. Align the mixture and reference spectra.
   (a) Generate the spectrum using the peak-picked list.
   (b) Apply a threshold to the spectrum, with all values below the threshold set to zero.
   (c) Determine the maximum shift window.
   (d) Compute the correlations throughout the range of the shift.
   (e) Shift the mixture to the point of maximum correlation.
3. Match the peaks in the reference to the peaks in the mixture.
4. Compare the energy of the matched peaks to the energy of the mixture to obtain the concentration of the reference material in the mixture.

With this method, it is possible to produce results within ±6% of the actual concentrations. Unfortunately, the method also produces several false positives. To reduce the number of false positives, an additional metric needs to be applied. Using a threshold of the percent of peaks matched from the reference, the number of false positives s reduced dramatically. Table 3.5 shows the results of three compounds known to be in the mixture and five compounds that were not in the mixture.

Even with the additional metric, this method still lacks a few key attributes. For example, the method does not neutralize the effects of differing physical attributes between the reference spectrum and mixture spectrum. There is also no allowance included to adjust for a difference in the measured frequency of the reference material and mixture. Finally, the method is currently applicable to only 1D NMR. It is speculated that adjustments for these factors will increase the accuracy of this method.

Additionally, there are several enhancements that would allow the method to be used as a stand-alone program. Currently, a third-party program is required to process the FID, including zero-filling, Fourier transformation, phase correction, and baseline correction, and to generate a peak list. Combining this additional functionality with the current method into a program would enable all analysis processing to be completed within a single application.

REFERENCES

Friebolin, Horst. 1998. *Basic one- and two-dimensional NMR spectroscopy.* Weinheim ; New York: Wiley-Vch, 3rd edition.

Hsu, Hwei P. 1984. *Applied Fourier Analysis.* Harcourt Brace Jovanovich college outline series. San Diego: Harcourt Brace Jovanovich, 1st edition.

Lambert, Joseph B. and Mazzola, Eugene P. 2004. *Nuclear magnetic resonance spectroscopy : an introduction to principles, applications, and experimental methods.* Upper Saddle River, N.J.: Pearson Education.

Macomber, Roger S. 1988. *NMR spectroscopy : basic principles and applications.* Harcourt Brace Jovanovich college outline series. San Diego: Harcourt Brace Jovanovich, 1st edition.

Markley, John L., Bax, Ad, Arata, Yoji, Hilbers, C. W., Kaptein, Robert, Sykes, Brian D., Wright, Peter E., and Wuthrick, Kurt. 1998. Recommendations for the Presentation of NMR Structures of Proteins and Nucleic Acids. *Pure & Applied Chemistry* 70:117–142.

MDL. 2006. Elsevier MDL :: Download a File. Available from: `http://www.mdli.com/downloads/public/ctfile/ctfile.jsp` [cited Oct. 31, 2006].

Nelson, John H. 2003. *Nuclear magnetic resonance spectroscopy.* Upper Saddle River, NJ: Prentice Hall.

Silverstein, Robert Milton, Bassler, G. Clayton, and Morrill, Terence C. 1991. *Spectrometric identification of organic compounds.* New York: Wiley, 5th edition.

Woolson, Robert F. 1987. *Statistical methods for the analysis of biomedical data.*

Wiley series in probability and mathematical statistics. Applied probability and statistics,. New York: Wiley.

APPENDIX I. AMINO ACIDS

The 20 amino acids responsible for protein synthesis.

Abbreviation		Name
Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic Acid
Cys	C	Cysteine
Gln	Q	Glutamine
Glu	E	Glutamic Acid
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Ser	S	Serine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine

APPENDIX II. REFERENCE DATA

Reference	**FID**			**Spectrum**
	No. of Data Points	Frequency (Hz)	Temperature (°C)	No. of Peaks
mixture	12001	599.76	20.00	81
Gln	12001	599.76	20.00	15
Trp	12001	599.76	20.00	23
Lys	12001	599.76	20.00	39
Orn	13999	599.76	20.00	45
Pyr	13999	599.76	20.00	2
Nad	12000	599.76	20.00	50
Gsh	12000	599.76	20.00	39
Parg	12000	599.76	20.00	41