Faculty of Science

CONNECTING THE DOTS: WORKING TOWARDS THE METABOLOMIC PROFILE OF *GRINDELIA SQUARROSA*

2017 | JASON MATHIUS MCFARLANE

B.Sc. Honours thesis – Chemical Biology

THOMPSON RIVERS UNIVERSITY

CONNECTING THE DOTS:
WORKING TOWARDS THE METABOLOMIC PROFILE OF *GRINDELIA SQUARROSA*

by

JASON MATHIUS MCFARLANE

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF SCIENCE (HONS.)
in the
DEPARTMENTS OF BIOLOGICAL AND
PHYSICAL SCIENCES
(Chemical Biology)

THOMPSON RIVERS UNIVERSITY

This thesis has been accepted as conforming to the required standards by:
Bruno Cinel (Ph.D.), Thesis Supervisor, Dept. Physical Sciences
Donald Nelson (Ph.D.), Co-supervisor, Dept. Biological Sciences
Jonathan Van Hamme (Ph.D.), Examining Committee member, Dept. Biological Sciences

Dated this 28th day of April, 2016, in Kamloops, British Columbia, Canada

ABSTRACT

Novel metabolomics methods using the NMR (nuclear magnetic resonance) spectrometer at Thompson Rivers University were developed. This will lead to better utilization of the NMR by opening up new applications of this powerful instrument. The method was applied to three different samples: a simple four compound mixture, a previously analyzed *Escherichia coli* lysate, and an extract of *Grindelia squarrosa*, which has not heretofore been analyzed using a metabolomics approach. The fractionated crudes extracts returned 597 compounds from the Biological Magnetic Databank, using HSQC peak chemical shifts. Results for the four-compound mixture were visualized as raw two-dimensional spectra, showing resolution of the peaks. The *E. coli* lysate provided invaluable insight into the necessity of high sensitivity as well as resolution in metabolomics. The future directions of this project are discussed, outlining the power of this technique for different mixtures and refining the experimental setup to reduce the necessity of relying on the currently available databases.

Thesis supervisors: Dr. Bruno Cinel and Dr. Donald Nelson

ACKNOWLEDGEMENTS

Thank you to my Honours supervisors, Dr. Bruno Cinel and Dr. Don Nelson for their guidance and support. I would also like to acknowledge Dr. Jon Van Hamme for taking time out of his busy schedule to evaluate my thesis. Funding provided by the Undergraduate Research Enhancement Award Program, UREAP.

TABLE OF CONTENTS

Abstract ii
Acknowledgements iii
Table of Contents iv
List of Figures vi
List of Tables vii
List of Acronyms viii
1 Introduction 1
1.1 Metabolomics 1
1.1.1 NMR-based metabolomics 1
1.2 Drug discovery 3
1.3 *Grindelia squarrosa* as a source of natural products 4
2 Materials and Methods 5
2.1 Composite mixture 5
2.1.1 Sample preparation 5
2.1.2 NMR acquisition parameters 5
2.2 *E. coli* lysate preparation 6
2.2.1 Growth curve 6
2.2.2 Sample Preparation 6
2.2.3 NMR acquisition parameters 6
2.3 Grindelia squarrosa 7
2.3.1 Sample collection 7
2.3.2 Sample preparation 7
2.3.3 Sample fractionation 8
2.3.4 NMR acquisition parameters 9
2.4 Computational processing 10
2.4.1 Peak deconvolution 10
2.4.2 Database querying 10
2.4.3 Constructing compound maps 10
3 Results and Discussion 11
3.1 Pulse sequences 11

3.2 Composite mixture 12
3.3 *E. coli* Lysate 15
3.3.1 Growth curve 15
3.3.2 NMR spectra 16
3.4 Grindelia squarrosa 18
3.4.1 Crude samples 18
3.4.2 Fractionated samples 21
4 Conclusions and Future Work 24
4.1 Complex mixture 24
4.2 *E. coli* lysate 25
4.3 *G. squarrosa* extracts 25
4.4 The direction of bioinformatics 27
4.5 Concluding remarks 27
5 Literature Cited 29
Appendix A 31
Appendix B 33

LIST OF FIGURES

Figure 1. The structures of the four compounds used in the composite mixture. The compounds are (A) coumarin, (B) limonene, (C) vanillin, and (D) menthol. The colours of the structure corresponds to the colour of the spectra in Figure 2, shown in the discussion. 5

Figure 2. (A), (B), (C), and (D) are the individual COSY spectra of 0.334 M coumarin, 0.289 M limonene, 0.285 M vanillin, and 0.289 M menthol in chloroform-d, respectively. The individual spectra were overlaid to give (E), which can be compared with (F), the actual COSY spectra produced by a mixture of all four compounds. 14

Figure 3. The growth curve of DH5-α bacteria grown in M9 minimal media at 37°C and 240 rpm. The data appears to level off at optical density of 0.2. 15

Figure 4. The spectra used to analyze the *E. coli* lysate. (A) shows a TOCSY spectrum, (B) a COSY spectrum, (C) an HSQC-TOCSY spectrum, and (D) an HSQC spectrum. 17

Figure 5. The connectivity network produced from six of the fractions showing the compounds present in each fraction. The large blue nodes represent the fractions, while the 594 unique compounds returned by the BMRB database are the small white nodes. Blue edges represent connections that result from a single fraction/compound pair, showing compounds that only appear in one fraction. The size of the nodes is proportional to the number of compounds returned by the database for each fraction, while the length of the edges is proportional to the inverse of the scoring value returned by the BMRB database referred to as 'Peak Match'. (A) shows the complete dataset from the BMRB database with no editing. (B) shows a simplified graph with only fractions 1, 2, and 3 shown. This allows for the easy visualization of compounds belonging only to 1 fraction (blue edges and on the peripherals), 2 fractions (grey edges and on the peripherals) and all three compounds (center of the graph). 22

Figure 6. This shows the same six fractions shown in Figure 5 with all compounds that had a Peak Match in the BMRB database equal or lower to 0.10. This filter removed 60 compounds from the dataset leaving 534 compounds. 36

LIST OF TABLES

Table 1. The eluent composition and fractions collected from each polarity solvent gradient. The eluent ranges from most polar (35% methanol:65% water) to least polar (100% methanol). The subfractions are volumes collected directly from the column. 8

Table 2. The range of subfractions pooled to give each fraction, which were analyzed by NMR spectroscopy. 9

Table 3. A selection of results returned from sample 15 from the BMRB database. The displayed compounds were chosen based on structural complexity and the presence of functional groups. 19

Table 4. A sample of the first 20 peaks with the highest intensity from the combination of the four trial samples (limonene, vanillin, coumarin, menthol). 31

Table 5. The first 28 rows of the output of NMR_reader.py using the four compound mixture as an input. 31

LIST OF ACRONYMS

BMRB	Biological Magnetic Resonance databank
COSY	Correlation Spectroscopy
csv	comma separated values
DMSO	Dimethyl sulfoxide
GC-MS	Gas Chromatography-Mass Spectrometry
HMBC	Heteronuclear Multiple Bond Correlation
HMDB	Human Metabolome Database
HSQC	Heteronuclear Single Quantum Correlation
LC-MS	Liquid Chromatography-Mass Spectrometry
MMCD	Madison Metabolomics Consortium Database
MS	Mass Spectrometry
NMR	Nuclear Magnetic Resonance
NS	Number of scans
OD	Optical Density
siRNA	small interfering Ribonucleic Acid
SW	Spectral width

TD	Time domain
TOCCATA	TOCSY Customized Carbon Trace Archive
TOCSY	Total Correlation Spectroscopy
TOCSY-HSQC	Total Correlation Spectroscopy-Heteronuclear Single Quantum Correlation
tsv	tab separated values

1 INTRODUCTION

1.1 METABOLOMICS

Metabolomics is the evaluation of all metabolites present in an organism. It is sometimes used interchangeably with metabonomics, which is the evaluation of metabolites as they change in concentration and composition, due to a change in a variable affecting an organism.[1] Due to the low concentrations of metabolites present in organisms, the full elucidation of the metabolites present is unfeasible with current instrumentation. Even so, a metabolomics approach allows for the identification of many compounds at once, and has the potential to reduce the number of rediscovered compounds, an issue in natural product chemistry. Additionally, it supports an untargeted approach, wherein the compounds present can be identified independently of their biological activity; as opposed to traditional isolation, where the active fraction must be constantly identified after each fractionation step in order to isolate a single active component. Two main metabolomics methods are used to identify the metabolites present. Gas or Liquid Chromatography-Mass Spectrometry (GC/LC-MS) is the most common, due to high sensitivity and ease of automation. However, Nuclear Magnetic Resonance (NMR) spectroscopy is also used; the advantages of NMR are the wealth of structural information gleaned as well as the ability to detect any organic compound.[2] Massive databases must be compiled of retention times and masses for GC/LC-MS and chemical shifts for NMR spectroscopy in order to get accurate results. Metabolomics also allows a library of compounds to be compiled for testing against multiple targets.

1.1.1 NMR-BASED METABOLOMICS

As NMR spectroscopy provides a relatively unique chemical shift value for each hydrogen, these values can be queried against an appropriate database of known compounds. However, due to the complexity of most natural product libraries, a 1-dimensional spectrum quickly becomes too crowded to parse out individual chemical shift values. Some of the techniques that researchers have developed to get around this problem are bucketing,[3,4] multivariate analyses,[5,6] and the use of 2-dimensional NMR spectroscopy.[7]

One of the main advantages of NMR-based metabolomics is the inherent ability to glean structural information of the metabolites present. NMR-based metabolomics has not gained the same following that chromatographic methods have, especially as there is no industry standard

for NMR spectroscopy methods. This is part due to the many approaches that can be used. Beyond the different types of pulse programs, which probe different interactions between nuclides, there are many different methods of sample preparations and data processing.

Some of the different pulse programs commonly used for metabolomics are: 1D proton, J-resolved proton, 1D $\{^{1}H\}^{13}C$, 2D ^{1}H-^{1}H COSY, 2D ^{1}H-^{13}C HSQC, 2D ^{1}H-^{13}C HMBC, and 2D ^{1}H-^{13}C TOCSY-HSQC. 1D proton views all of the individual chemical and magnetic environments of a proton (^{1}H) in a molecule. Its downfall is the limited range over which the protons are spread out from (typically 0 to 12 ppm). This leads to many overlapping peaks when used for analyzing complex mixtures. One way of dealing with this, without adding too much complexity, is the use a J-resolved proton pulse sequence. This sequence separates a 1D proton spectra based on the strength of the coupling constant, J, which causes the peaks produced by nuclides to split into higher and lower energy states based on the spin state of neighboring hydrogens. This allows protons that appear at identical chemical shifts, but have different coupling constants (which are typically on the range of 1-6 Hz), to be differentiated from each other. 1D $\{^{1}H\}^{13}C$ is a carbon-13 spectra that is decoupled from proton couplings. This decoupling simplifies the carbon-13 spectra, giving a single peak for each environment. This fact, along with the greater range that a carbon-13 spectrum is spread over (typically 0 to 200 ppm) increases the likelihood that each peak is at a unique chemical shift.

While one-dimensional spectra have decent sensitivity within a 20-minute time period, 2-dimensional pulse sequences trade some of that sensitivity for high resolution of peaks. In essence, each peak that appears in the 2-dimensional plot represents the interaction of nuclides that neighbor each other on a molecule. As a consequence, it is possible to determine two chemical shifts that are theoretically on the same molecule. This increases the certainty of chemical identification. Two-dimensional NMR spectroscopy can be divided into two categories: homonuclear and heteronuclear. Homonuclear experiments look at coupling between the same types of nuclides. For example, ^{1}H-^{1}H COrrelated SpectroscopY (COSY) looks at hydrogens that are within three bonds of each other on a molecule. This gives symmetrical spectra with peaks along the diagonal (the same x and y coordinates) representing protons coupling with themselves and off-diagonal peaks representing the interaction of two different protons. This type of spectra can be “walked” along, going from off-diagonal peak to off diagonal peak along the same x or y

coordinate, to determine substructures of entire spin systems. Unfortunately, wherever a tertiary carbon appears, there is a break in the coupling hydrogens, making it hard to determine the entire structure of a molecule. At the tradeoff of sensitivity, the distance of coupling detected can be increased with a TOtal Correlation SpectroscopY (TOCSY) experiment. This pulse sequence increases the time that the protons can influence each other through coupling, as well as represses the signal of 3-bond coupling. An additional way to get around tertiary carbons is to use a ^{13}C-^{13}C COSY experiment. This detects 1 bond ^{13}C-^{13}C couplings, similar to how the proton variation detects three bond correlations. Unfortunately, carbon-13 has such a low natural abundance that the chance of two neighboring carbon atoms being carbon-13 is 1: 10,000. To be functional, the metabolites must be isotopically-labeled with carbon-13, by introducing labeled feedstock to the growing organism. This is particularly useful for determining metabolic pathways that metabolites were produced from, as there is high sensitivity for molecules that incorporate the feedstock while other molecules are fainter.[8,9]

The other type of 2-dimensional NMR spectroscopy experiment is heteronuclear. This is the detection of coupling between two different nuclides, usually ^{1}H and ^{13}C, though ^{15}N can also be used instead of carbon-13 if the sample has been isotopically-labeled. Heteronuclear single quantum coherence spectroscopy (HSQC) probes one bond ^{1}H - ^{13}C interactions. One useful feature of heteronuclear experiments is that, because the coupling is between different nuclides, there is no peak from self-coupling. That is to say, there is no large diagonal mass of peaks, which obscure the peaks produced by the coupling of protons of similar chemical shift. As well, HSQC is an asymmetrical spectra, removing the redundancy present in COSY type spectra. HSQC has additional simplicity, because each proton and carbon theoretically produces only a single peak in the spectra. This is complicated somewhat from noise due to very large peaks, such as solvent peaks. Heteronuclear Multiple Bond Correlation spectroscopy (HMBC) looks at the coupling between carbon-13 and protons within three bonds by suppressing the signal from single bond coupling. A similar pulse program, TOCSY-HSQC, also includes a TOCSY mixing time to allow spin coupling to spread further throughout the molecule.

1.2 DRUG DISCOVERY

As the field of medicine is developing and more of a focus is being placed on personalized medicine, there is a large push in the field of chemical biology to develop new therapeutic agents

to treat a variety of ailments. In order to do this, large collections of chemical libraries must be screened against biological targets. These chemicals can then be synthesized in lab in various diversity- and target-oriented synthesis approaches. These approaches have the potential to produce wide varieties of chemical scaffolds; however, they are limited by the synthetic methodologies available, are time consuming to synthesize, and (particularly in diversity-oriented synthesis) produce many compounds without biological activity. In contrast, organisms synthesize natural product libraries, so many complicated organic reactions are catalyzed by enzymes in high yield, reducing time and money.[10] Additionally, biological molecules are considered optimized through evolution to act within biological systems and on biological targets. However, these advantages are offset by the need to identify and isolate the natural products from a complex mixture. Unfortunately, the traditional method of activity directed isolation, wherein the extract is screened against a biological target and any active components isolated, suffers from a high incidence of rediscovery. This is due to the compounds only being characterized after isolation. As mentioned in section 1.1, one method that has the potential to deal with this issue is untargeted metabolomics.

If constituents in an extract can be identified in a metabolomics experiment before or concurrent with assaying for biological activity, then previously discovered active compounds can be identified and priority placed on novel compounds. This opens up new possibilities from sources that have previously been investigated using traditional methods. Using metabolomics allows for new compounds to be detected, and the activity of compounds that are in low concentration to be elucidated.

1.3 *GRINDELIA SQUARROSA* AS A SOURCE OF NATURAL PRODUCTS

Exploring plants used in traditional medicine as a source of natural products could lead to the development of new therapeutics. However, many plants have already been thoroughly characterized. Either new plants must be investigated or new approaches, such as metabolomics, used to characterize more compounds present in the natural product "library". *G. squarrosa* is a common plant in the southwest interior of British Columbia. It has a yellow, bulbous flower and is covered in a resinous exudate. *G. squarrosa* is known in indigenous medicine as an expectorant; that is, it induces mucous to be expelled from the lungs. Recent research reports the crude extract to have modest antibiotic activity. This makes the plant a promising source of

bioactive molecules. Previous research has been done to identify the components of the essential oil of *G. squarrosa* by GC-MS,[11] though little work has been done on the more polar constituents, or using a metabolomics approach.

2 MATERIALS AND METHODS

2.1 COMPOSITE MIXTURE

2.1.1 SAMPLE PREPARATION

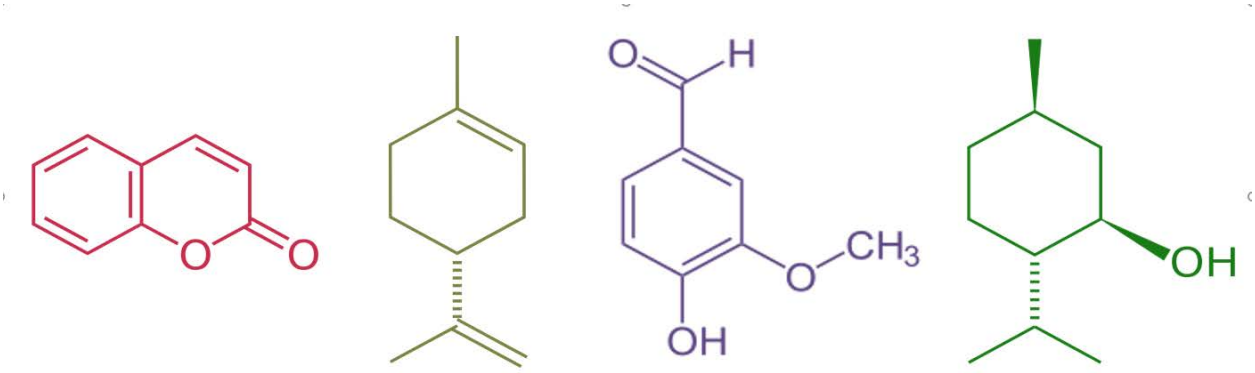


Figure 1. The structures of the four compounds used in the composite mixture. The compounds are (A) coumarin, (B) limonene, (C) vanillin, and (D) menthol. The colours of the structure corresponds to the colour of the spectra in Figure 2, shown in the discussion.

Solutions of 0.0197 g limonene, 0.0226 g menthol, 0.0217 g vanillin, and 0.0244 g coumarin in 0.5 mL chloroform-d were prepared to give solutions of 0.289 M, 0.289 M, 0.285 M, and 0.334 M, respectively. NMR analyses were done with the parameters described in section 2.1.2, then the solutions were mixed pair-wise (limonene and menthol, vanillin and coumarin) in a 1:1 ratio. These samples were analyzed before being mixed in equal volumes to give a final concentration of 0.0723 M limonene, 0.0723 M menthol, 0.0713 M vanillin, and 0.0835 M coumarin. This final solution was also analyzed by NMR spectroscopy.

2.1.2 NMR ACQUISITION PARAMETERS

1D proton (number of scans (NS)=16), and 2D COSY spectra (NS=8) for each sample in 0.5 mL chloroform-d were obtained on a Bruker Ultrashield™ Plus Avance III 500 MHz NMR spectrometer with a 5 mm PATXI $^{1}H/D-^{13}C/^{15}N$ Z xgradient probe. Each spectrum was coloured and overlaid to show the unique cross peaks between samples.

2.2 *E. COLI* LYSATE PREPARATION

2.2.1 GROWTH CURVE

DH5-α *E. coli* were grown in 20 mL M9 minimal media in a 100-mL side armed flask at 37°C and 240 rpm by inoculating with 2 mL of an LB overnight broth. Optical density at 600 nm (OD) measurements were taken every 30 minutes to determine the time required for culture saturation. As the OD failed to reach 3.0, as previously reported in the literature,[12] the growth curve of *E. coli* grown in LB broth was determined to establish a theoretical maximum.

2.2.2 SAMPLE PREPARATION

DH5-α *E. coli* were grown in 1 L of M9 minimal media in four 250 mL portions each in a 1-L Erlenmeyer at 37°C and 240 rpm. The culture was split into four 250 mL portions, which were each inoculated with 10 mL LB broth overnights. The cultures were grown for 24 hours, then the cells were collected by centrifugation at 8000 xg and 4°C. The pellets were combined and washed with three 15 mL portions of 50 mM phosphate buffer (pH 7.0). The cells were pelleted and resuspended in 10 mL deionized water. This suspension was frozen for 2 hours at -20°C, then thawed. This was repeated two more times, before the cell fragments were collected by centrifuging at 16,000 xg at 4°C. The supernatant was retained and 10 mL of methanol was added, then 10 mL of chloroform. The container was agitated, then left for 12 hours for the organic and aqueous layers to separate. The chloroform was removed by transfer pipette and the methanol removed by 1 hour of nitrogen blowdown. The water was removed under vacuum centrifugation for 3 hours at 30°C. The residue was resuspended in 1 mL of D_2O and centrifuged for 5 minutes at 14,000 xg to pellet undissolved compounds.

2.2.3 NMR ACQUISITION PARAMETERS

1D proton (NS=16), 2D COSY spectra (NS=16, time domain (TD)=4096x256, offset 4.7 ppm, spectral width (SW)=10.9920 ppm*10.9920 ppm), 2D TOCSY(NS=8, TD=4096x2048, offset 4.7 ppm, SW=10.9920*10.9920, mixing time=0.090 s), 2D HSQC-TOCSY (NS=128, TD=2048x512, proton offset 4.7 ppm, carbon-13 offset=85.0 ppm, SW=10.9920 ppm*170.9149 ppm, mixing time=0.080 s), and 2D HSQC (NS=64, TD=2048x512, proton offset 4.7 ppm, carbon-13 offset=85.0 ppm, SW=10.9920 ppm*170.9149 ppm) were used to analyze the aqueous portion of the extract.

2.3 GRINDELIA SQUARROSA

2.3.1 SAMPLE COLLECTION

Samples were collected from the side of the Red-Tailed Hawk trail in Kenna Cartwright Park in Kamloops, BC. Initial samples (numbered 1-18) were collected by cutting the stem three inches below the sepals, labeling with masking tape and placing in a brown paper bag to dry for three weeks. For subsequent samples, the entire plant, including roots, was removed from the soil and placed in a dark box to dry, also for three weeks.

2.3.2 SAMPLE PREPARATION

The dried *G. squarrosa* flowers were weighed and ground with a mortar and pestle. 2 mL of hexanes (ACS grade, BDH) were added, and the grinding continued until the plant matter appeared homogenous. The sample was transferred to a 50-mL Erlenmeyer flask, 8 mL more hexanes added for a total volume of 10 mL, sealed with parafilm and extracted for 24 hrs. The hexanes fraction was filtered through a coarse filter paper, and the solid returned to the Erlenmeyer flask. 10 mL of acetone (ACS grade, BDH) was added, the flask was sealed with parafilm, and the flask was left for 24 hrs to extract. The acetone fraction was filtered through a coarse filter paper, and the solid returned to the Erlenmeyer flask. 10 mL of methanol (HPLC grade, BDH) was added, the flask was sealed with parafilm, and the flask was left for 24 hrs to extract. The methanol fraction was filtered through a coarse filter paper, and the solid returned to the Erlenmeyer flask. The acetone and methanol fractions were evaporated to dryness by nitrogen blow down and combined.

For the sample fractionation, 3.72 grams of plant material (2.97 g flower heads and 0.75 g leaves) were prepared as described above except that the sample was dried by rotary evaporation instead of nitrogen blow down. This gave a total fractionated mass of 0.4465 g, or 12% yield. The acetone and methanol fractions were resuspended in 20 mL of methanol and stored at 4°C for 1 week. The methanol was dried by nitrogen blow down to produce a dark green oil. The oil was suspended in 50:50 methanol to water by ultra-sonication bath for one minute and partially dried by nitrogen blow down. 10 mL methanol was added and the oil was resuspended and transferred to a 50-mL round bottomed flask.

2.3.3 SAMPLE FRACTIONATION

Preparative C-18 125 Å 55-105 µm Waters silica solid phase was added to the resuspended crude extract form a loose slurry. The mixture was dried to a thick paste, and then transferred to the top of a 3x10 cm C-18 column. As shown in Table 1, the sample was eluted with gradients from 35:65 methanol: water to 100% methanol and subfractions collected in approximately 12 mL portions. Methanol was applied until the visible dark band was eluted, with these later fractions collected in 100 mL portions.

Table 1. The eluent composition and fractions collected from each polarity solvent gradient. The eluent ranges from most polar (35% methanol:65% water) to least polar (100% methanol). The subfractions are volumes collected directly from the column.

% Methanol	% Water	Volume (mL)	Subfraction range
35	65	100	1-7
45	55	100	8-13
55	45	100	14-19
65	35	100	20-26
70	30	100	27-32
75	25	100	33-38
80	20	100	39-44
90	10	100	45-49
100	0	800	50-71

Fractions were combined based on a visual comparison of colour and fluorescence based loosely around the polarity of the eluents. Details as to which fractions were combined are shown in Table 2. Each combined fraction was evaporated to dryness by rotary evaporation. The fractions were transferred to test tubes with five 2 mL portions of methanol, apart from fraction 1, which was relatively insoluble in methanol. As such, fraction 1 was transferred with three 2 mL portions of methanol and three 2 mL portions of deionized water. The combined fractions were then evaporated to dryness by nitrogen blow down, weighed, and stored at 4°C until analysis.

Table 2. The range of subfractions pooled to give each fraction, which were analyzed by NMR spectroscopy.

Fraction number	Subfraction range	Mass (g)
1	1-8	0.2227
2	9-14	0.0138
3	15-20	0.0144
4	21-28	0.0552
5	29-34	0.0233
6	35-45	0.0414
7	46-50	0.0281
8	51-55	0.0222
9	56-60	0.0106
10	61-65	0.0049
11	66-71	0.0099

With the exception of fraction 1, each fraction was dissolved in 1 mL of methanol-d_4 and a drop of tetramethylsilane (TMS) was added as a standard. The test tube was ultra-sonicated to suspend metabolites adhered to the glass, and then centrifuged for five minutes to remove undissolved particles. The supernatant was then transferred to an NMR tube and analyzed by NMR spectroscopy. Fraction 1 was treated like the rest of the fractions, except that it was dissolved in 1 mL of DMSO-d_6 instead of methanol-d_4.

2.3.4 NMR ACQUISITION PARAMETERS

1D proton (NS=16), 2D COSY spectra (NS=128, TD=8192x256, offset 4.7 ppm, SW=10.9920 ppm*10.9920 ppm), 2D HMBC (NS=128, TD=4096x256, proton offset 4.7 ppm, carbon-13 offset=85.0 ppm, SW=10.9920 ppm*172.9149 ppm), and 2D HSQC (NS=64, TD=2048x512, proton offset 4.7 ppm, carbon-13 offset=85.0 ppm, SW=10.9920 ppm*172.9149 ppm) were used to analyze all fractions and crude extracts.

2.4 COMPUTATIONAL PROCESSING

2.4.1 PEAK DECONVOLUTION

A list of peaks was compiled by the peak picking command in Bruker Topspin v2.1, with the lowest contour level visually set to just above the baseline. These peaks were exported to a comma separated values (csv) file with four columns: Peak, containing an arbitrary unique identifier; ?(F2) [ppm], containing a list of the F2 chemical shifts; ?(F1) [ppm], containing a list of the F1 chemical shifts; and Intensity [abs], containing the peak integration values. A sample of the output is shown in Appendix A, Table 4.

In order to compile sets of multiple peaks belonging to the same substructure, a Python 3.7 program was developed in Spyder v3.13 using the Pandas library. The script is included in Appendix B as NMR_reader.py. In essence, the NMR_reader.py program took a table of two-dimensional chemical shifts and condensed the redundant chemical shifts in one column by appending the corresponding chemical shifts in the other column to one, unique value. This simplified the dataset by taking a list of approximately 2000 chemical shifts pairs and reducing it to a list of approximately 500 unique entries without loss of information.

2.4.2 DATABASE QUERYING

The NMR_reader.py program was executed on the sets of COSY data from the composite mixtures. The sets of substructures from the NMR_reader.py program were converted to a csv format, and then the individual substructures were inputted into the TOCCATA database.[12]

The csv peak list of the HSQC experiments was manually converted to a tab separated value (tsv) file with two columns (?(F2) [ppm] and ?(F1) [ppm]) by uploading the csv files of the HSQC experiments to Google sheets and exporting the two desired columns as a tsv file. This tsv file was uploaded to the BMRB database website[13] in the “Search 2D HSQC lists” submenu. A list of potential compounds was returned for each peak list uploaded, which was then downloaded as a tsv file, and converted to a csv file.

2.4.3 CONSTRUCTING COMPOUND MAPS

The csv files from the BMRB database of fractions 1-4, 6, and 7 were concatenated. The entries in the first column were changed to the fraction name that the compound came from without spaces (e.g. “Fraction1”). The name of the first column was changed to “Fractions”. The column that contained the compound names was renamed to “compounds”. The scoring value supplied

by the BMRB database, “Peak Match”, was renamed to “weight” in order to align better with the graphing program. A new spreadsheet was made that contained one column, “Nodes”, that contained all entries from the compound column and a single entry for each fraction.

These csv files were used as inputs for a network parsing Python program, NMR_network.py. Refer to Appendix B for details; but in brief, this program dealt with all duplicate values and parsed the data into a format that could be read by Cytoscape, a network visualization platform.[14] NMR_network.py wrote an xml file, which was inputted into Cytoscape as a network file. In Cytoscape, the layout was set to Edge-weighted Spring Embedded Layout, based on the weight parameter. Then from the Tools drop down menu, Workflow was opened and the “Analyze selected networks and create custom styles” was selected. This option returned metadata about the networks as well as adapted the network to highlight key features. In order to clean up the number of edges displayed, Layout | Bundle Edge | All Nodes and Edges was selected.

An additional map for clearer understanding was constructed from fractions 1, 2, and 3. As well, a simplified complete map was made by deleting all edges that had a weight value equal or less than 0.10.

3 RESULTS AND DISCUSSION

3.1 PULSE SEQUENCES

In order to optimize peak resolution and sensitivity many pulse sequences were investigated over the course of this project. The TOCSY and TOCSY-HSQC were initially chosen due to their integration with the TOCCATA database.[12] These pulse sequences are advantageous because they have very narrow peaks in the indirect dimension on the spectra, meaning that horizontal one dimensional “slices” of the spectra should only contain peaks from one spin system. The result is a one-dimensional spectrum that contains only the peaks from one spin system. Unfortunately, the spectra produced from the *E. coli* lysate did not appear to be well resolved or sufficiently sensitive to isolate individual spin systems from the raw spectra, as shown in Figure 4. The COSY pulse program also provides connectivity information, and is more sensitive than TOCSY. Unfortunately, COSY shows less information than TOCSY, so additional computational steps must be done to isolate complete spin systems. COSY spectra are more powerful than 1D proton NMR spectra, clustering more consistently in statistical treatments[15].

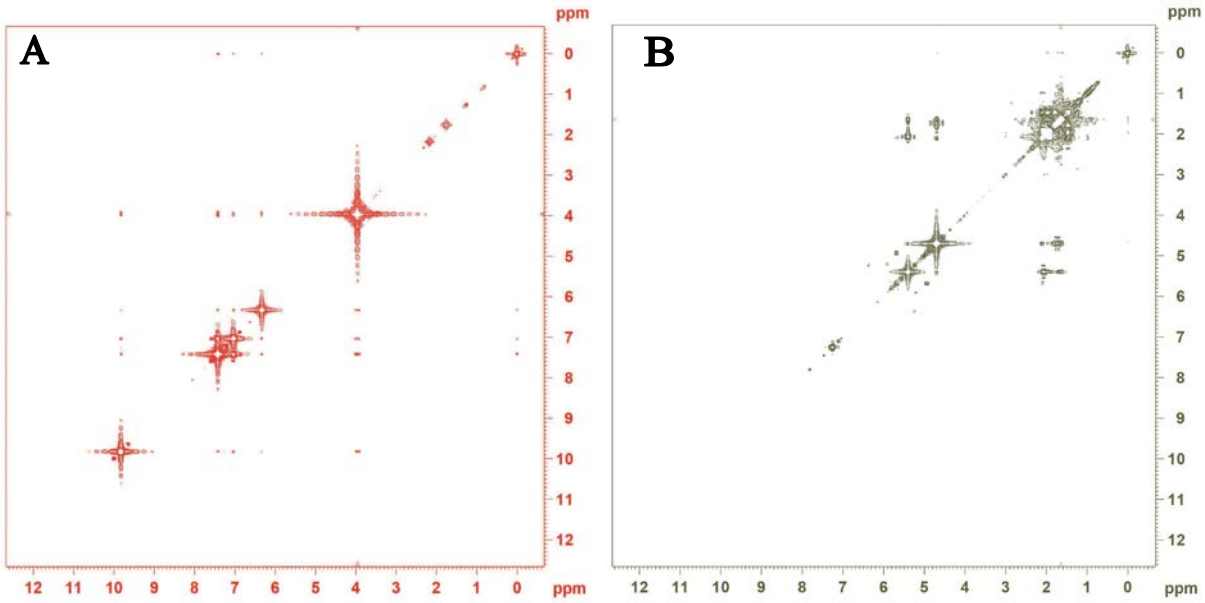
The computer program, NMR_reader.py attempted to take the high sensitivity of COSY spectra and convert it into the structural information provided by TOCSY slices.

HSQC has lower sensitivity than heteronuclear experiments by virtue of the low sensitivity of the carbon-13 nuclide. This is offset by the high resolution afforded by the expanded carbon-13 dimension compared to the proton dimension.[16] HSQC is also advantageous as every nuclide appears only once in the two-dimensional spectrum. One interesting advantage of HSQC is it is possible to determine qualitative amounts of metabolites, by varying delay times and extrapolating backwards to determine peak intensity at t = 0.[17]

3.2 COMPOSITE MIXTURE

Compounds were selected because recent research identified some of them in the essential oil of *G. squarrosa* (limonene, menthol)[11] and others were similar in structure to acetylsalicylic acid (vanillin). The analysis of the composite mixture allowed any peaks that may arise from the pulse sequence to be identified and be considered in more complex mixtures. In Figure 2, the off diagonal peaks are clearly resolved between compounds, while the diagonal "self-coupling" peaks show a large degree of overlap. Querying the COSY peak list of the 4 compounds against the BMRB database returned a list of 815 compounds. Unfortunately, the BMRB database was not setup to allow the querying of multiple related COSY peaks, so each compound was based on a single peak assignment. This probably contributed to the extremely high incidence of false positives. The expected compounds were present in the data output; they appeared in the first 100 compounds with high peak match score with the exception of limonene, which only appeared as limonene oxide. This is probably due to a limitation of the database, because limonene is not present in the dataset. However, limonene is also prone to autooxidation in the presence of oxygen, so this could be another contributing factor. Further work need to be done in order to increase the fidelity of the results. Querying the output of NMR_reader.py against the TOCCATA database in order to reference more peaks was unsuccessful due to the limitations of the TOCCATA database. Though the database allows for substructures gleaned from multiple bond coupling relationships in TOCSY-type experiments, the compounds contained in it are derived from the human metabolome database (HMDB). As such, exogenous compounds are not present in the reference dataset. This limits the applicability of the TOCCATA database to natural product extracts.

Other databases were considered. NAPROC-13[18] has an excellent selection of natural products.[19] Unfortunately, the database relies exclusively on carbon-13 assignments, and only a single substructure at a time can be queried. This makes it ideal for two-dimensional ^{13}C-^{13}C COSY experiments with carbon-13 labelled feedstock, but requires elaborate processing to make it useable with carbon-13 in natural abundance. The Madison Metabolomics Consortium Database allows many parameters to be inputted at once, including two-dimensional COSY.[20] When the composite mixture was queried, only six compounds were returned, a large improvement over the BMRB database; however, only one compound (coumarin) was returned that was in the composite mixture.



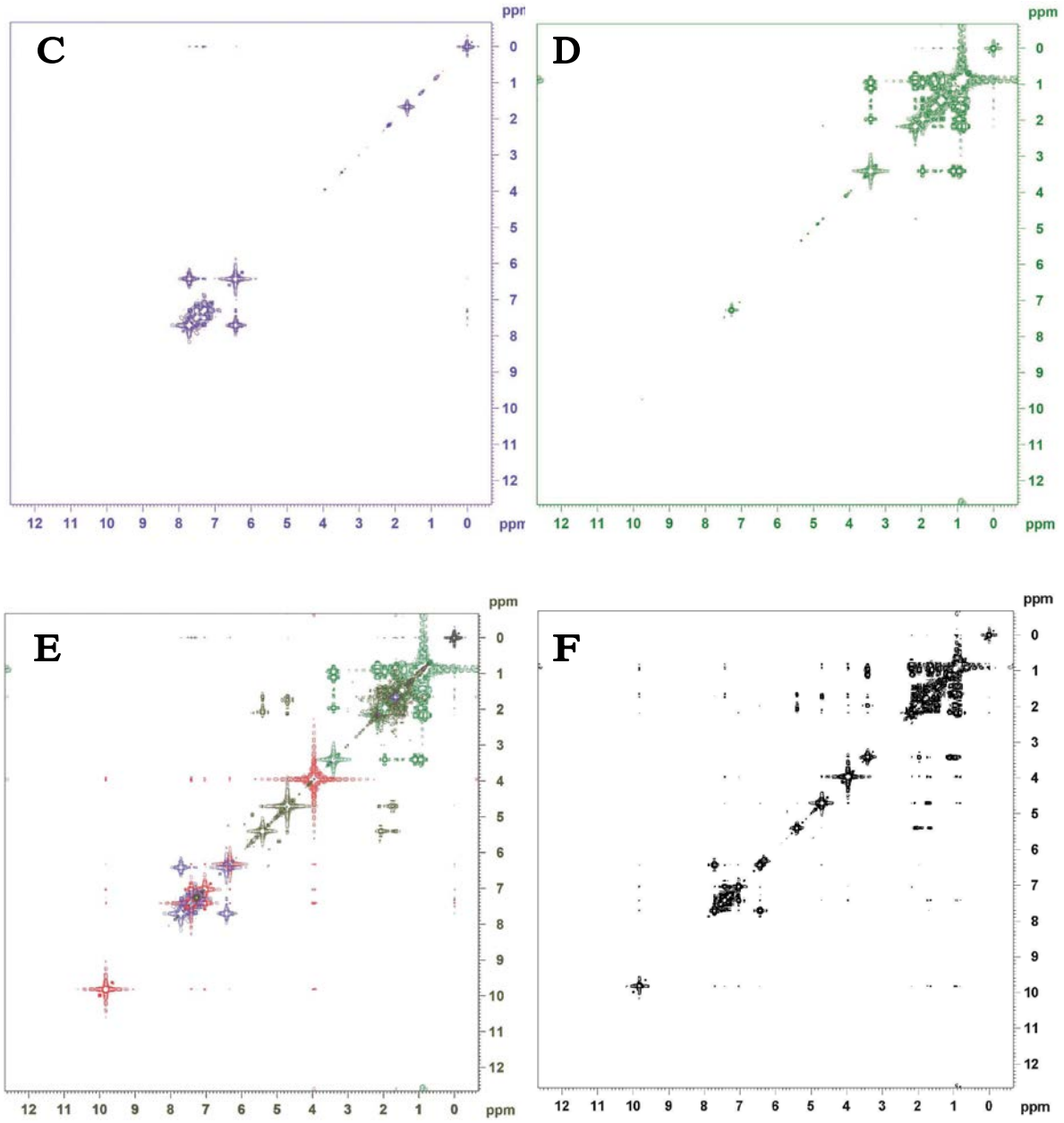


Figure 2. (A), (B), (C), and (D) are the individual COSY spectra of 0.334 M coumarin, 0.289 M limonene, 0.285 M vanillin, and 0.289 M menthol in chloroform-d, respectively. The individual spectra were overlaid to give (E), which can be compared with (F), the actual COSY spectra produced by a mixture of all four compounds.

The spectra produced from the four compounds are shown in Figure 2. Figure 2 A-D shows the individual spectra of each compound. In retrospect, it may have been advantageous to select some compounds with more complex chemical scaffolds, as these compounds would produce more off diagonal peaks in the COSY spectra. The compounds that were chosen had few

neighboring hydrogens, which are what a COSY pulse sequence detects. This is particularly noticeable in Figure 2 C, vanillin, which only shows one off-diagonal peak produced by neighboring hydrogens.

3.3 *E. COLI* LYSATE

3.3.1 GROWTH CURVE

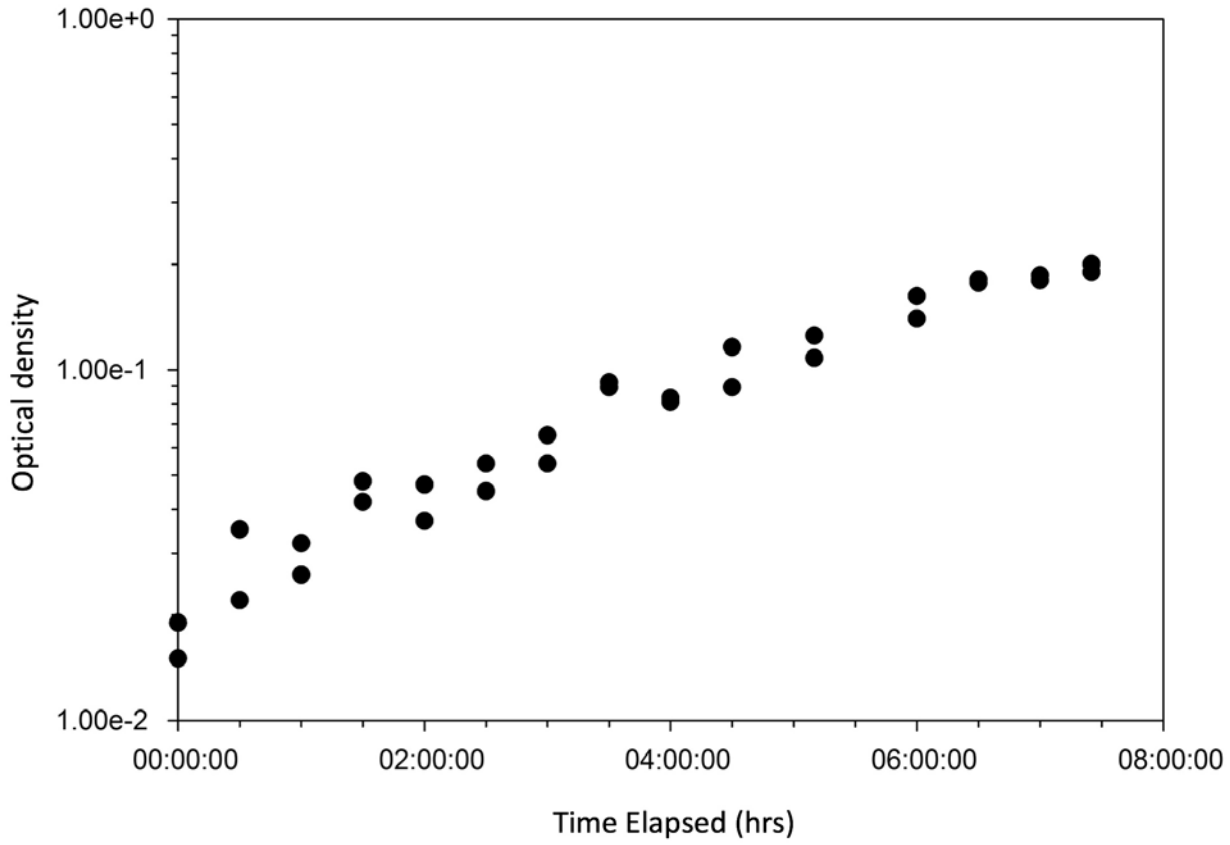


Figure 3. The growth curve of DH5-α bacteria grown in M9 minimal media at 37°C and 240 rpm. The data appears to level off at optical density of 0.2.

As shown in Figure 3, the DH5-α *E. coli* grew to ~0.2 optical density (OD) after 7 hours of growth. This culture saturation was considerably lower than reported in the method followed, with an absorbance of 3.0. This contributed to low sensitivity in the NMR spectra. It is possible that the cultures needed to be incubated for longer, in order to maximize cellular concentrations and, more importantly, maximize the production of secondary metabolites. The low OD may also be due to low oxygen levels, which could be alleviated by increasing the shaker speed and further splitting the culture into more Erlenmeyer flasks. Changing the concentration of glucose

did not seem to affect the final OD, meaning that glucose was not the limiting component of the media. Changing the media to LB broth doubled the final OD, but it did not approach the order of magnitude increase that was reported in the literature, indicating potential discrepancies in the methods of Bingol et al.[12] The two main possibilities are that the culture was measured with a path length of 10 cm instead of 1 cm, or culture saturation was simply assumed to have an OD of 3 without any sort of spectrophotometric measurements to confirm. As contacting the researcher received no response, the analysis of the 0.3 OD 1-L culture was done.

3.3.2 NMR SPECTRA

Figure 4 shows the four different spectra produced by NMR analysis of the *E. coli* lysate. Note the poor sensitivity displayed in Figure 4 C, which causes it to have the same or fewer peaks than the corresponding near coupling spectra, while the opposite should be the case. The TOCSY spectrum (Figure 4 A) has the opposite issue, where there is so much signal, that almost no resolution can be observed. Sensitivity was low, in part due to the low cell concentrations already mentioned, but also because the parameters were poorly optimized for the *E. coli* lysate. The number of scans should be increased, and the resolution in the time domain decreased. There were also issues with the HSQC-TOCSY pulse sequence, and to a lesser extent the TOCSY pulse sequence, where the sample became hot over the course of the experiment. Part of this can be attributed to the exothermic mixing time inherent in TOCSY pulse programs; however, the HSQC-TOCSY was an adiabatic variant, due to some corruption in the parameter files of the basic pulse program. What effect, if any, this had on the increase in temperature is unclear. Heat is undesirable due to averaging of signals that are separate at room temperature (which most of the database spectra have been collected). Additionally, the change in temperature can decompose analytes and cause solvent to rapidly evaporate. A CryoProbe can be used to keep the temperature of an NMR experiment constant, but our instrument used is not equipped with that probe. This lead to a preference for lower energy pulse sequences–such as HSQC, HMBC, and COSY–in the rest of the analyses done by NMR spectroscopy.

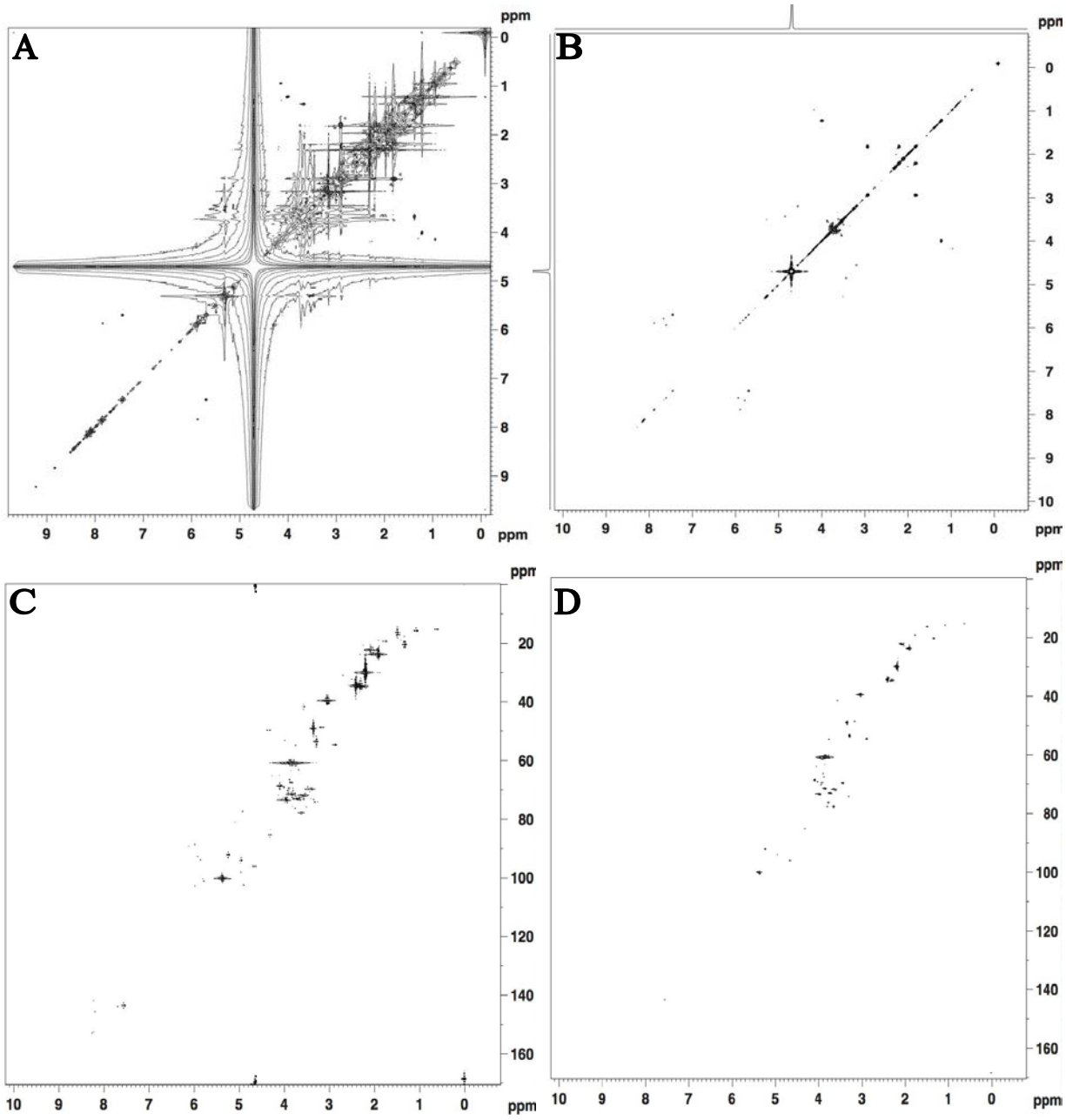


Figure 4. The spectra used to analyze the E. coli *lysate. (A) shows a TOCSY spectrum, (B) a COSY spectrum, (C) an HSQC-TOCSY spectrum, and (D) an HSQC spectrum.*

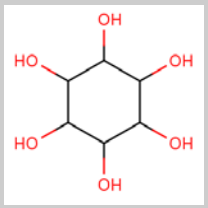
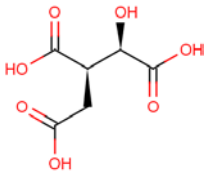
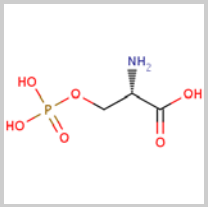
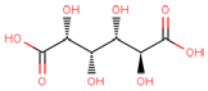
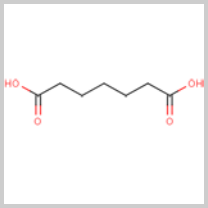
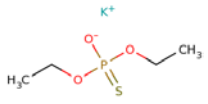
3.4 GRINDELIA SQUARROSA

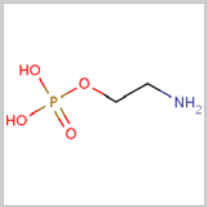
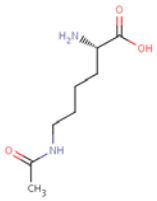
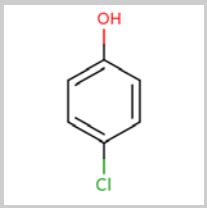
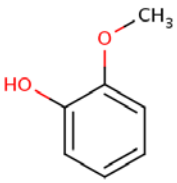
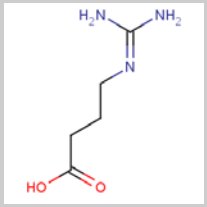
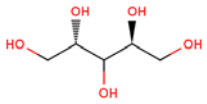
3.4.1 CRUDE SAMPLES

The extraction with hexanes gave a viscous, yellow oil that had a strong, resinous scent that was completely miscible in hexanes. The acetone extraction was a vivid yellow, with a faint green hue that varied in intensity based on the amount of photosynthetic material included in the plant sample. The methanol fractions were an off-yellow colour that was quite weak in intensity.

Solvent-wise, the DMSO-d_6 dissolved the most extract. However, this was mitigated somewhat by the large, broad solvent peak produced by DMSO that overwhelmed the weaker peaks next to them. The 1:1 D_2O: acetone-d_6 dissolved the least amount of the extract and had the added issue of two solvent peaks, one broad peak from HOD, and one intense peak from the six equivalent protons in acetone. These peaks produced unwanted noise and overlapped with low intensity analyte peaks. Methanol-d_4 showed a compromise between these two extremes. Most compounds were soluble in the methanol, as that was one of the two solvents that comprised. The methanol has two solvent peaks; one at 3.31 ppm and the OH proton at 4.78 ppm. The hydrocarbon portion allows methanol to dissolve moderately nonpolar compounds. The OH group allows for hydrogen bonding with compounds with hydrogen bond donors and acceptors. The OH group also allows the solvent to be buffered to increase the reproducibility of the chemical shifts. Unfortunately, buffering was neglected in this experiment, which may have led to inconsistencies in the chemical shifts. Querying the BMRB database returned between 217 and 477 compounds. These compounds were compared against the results of the networks produced from the fractionated compounds (Figure 5). The sample dissolved in methanol displayed 54 new compounds that were not present in any of the fractions analyzed; this is probably due to the incomplete analysis of the fractions, but may indicate that the BMRB database requires very few compounds in order to reduce the number of false positives.

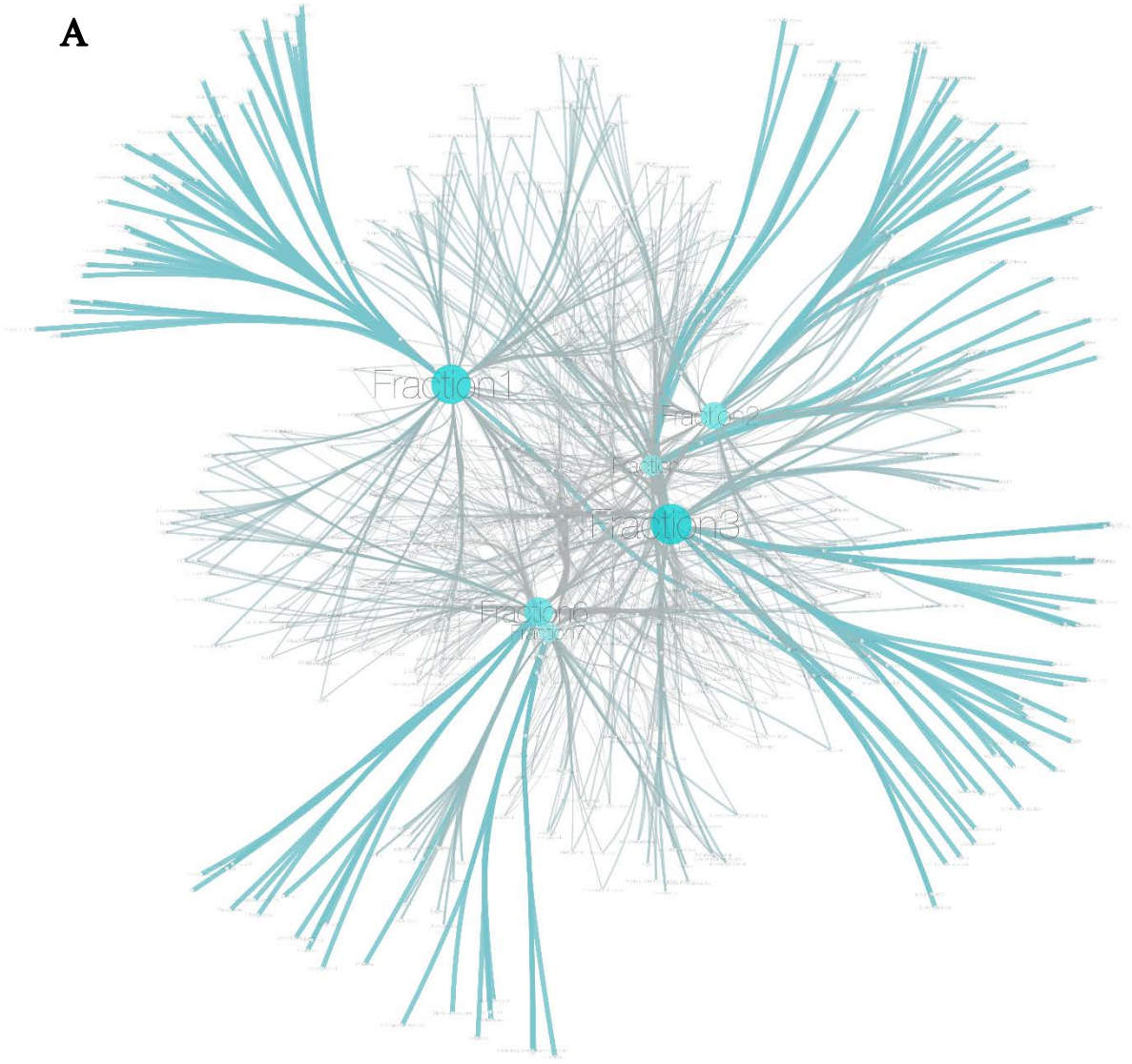
Table 3. A selection of results returned from sample 15 from the BMRB database. The displayed compounds were chosen based on structural complexity and the presence of functional groups.

Peak match	Structure	Name
1		scyllo-Inositol
0.8		D-(+)-Threo-isocitric acid
0.78		O-Phospho-L-serine
0.75		D-Saccharate
0.73		Pimelic acid
0.67		O,O-diethyl thiophosphate

0.57		2-Aminoethyl dihydrogen phosphate
0.53		Nepsilon-Acetyl-L-lysine
0.5		4-Chlorophenol
0.5		Guaiacol
0.5		4-Guanidinobutyric acid
0.5		L-(-) Arabitol

3.4.2 FRACTIONATED SAMPLES

The fractions were unevenly split in the mass of the isolated compounds, varying from 0.0049 g to 0.2227 g, as shown in Table 2. This indicates, at a very superficial level, that the way that the subfractions were pooled needs to be reevaluated. Ideally each fraction would contain an equal number of compounds, and those compounds would be in relatively equal concentrations.



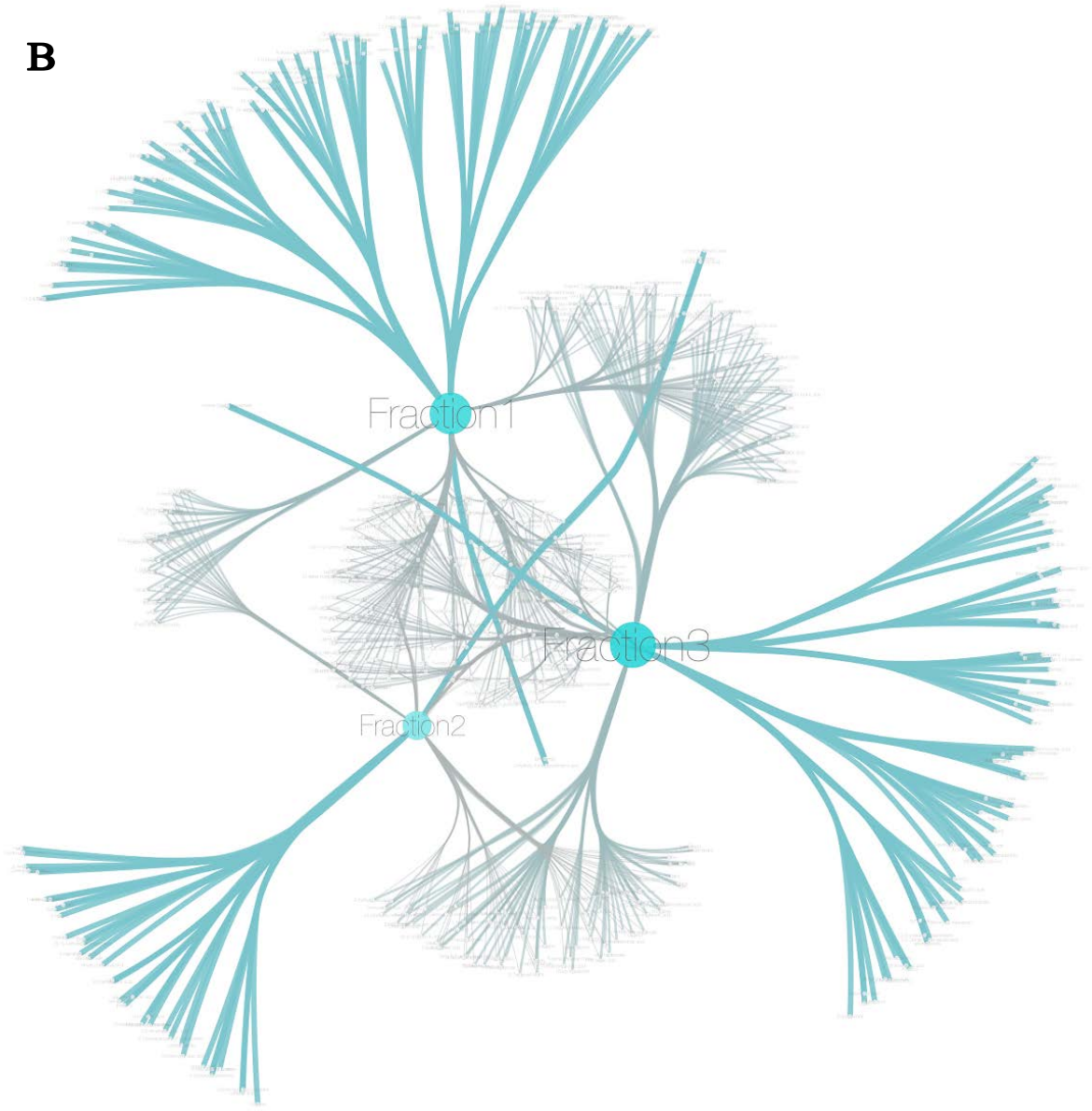


Figure 5. The connectivity network produced from six of the fractions showing the compounds present in each fraction. The large blue nodes represent the fractions, while the 594 unique compounds returned by the BMRB database are the small white nodes. Blue edges represent connections that result from a single fraction/compound pair, showing compounds that only appear in one fraction. The size of the nodes is proportional to the number of compounds returned by the database for each fraction, while the length of the edges is proportional to the inverse of the scoring value returned by the BMRB database referred to as 'Peak Match'. (A) shows the complete dataset from the BMRB database with no editing. (B) shows a simplified graph with only fractions 1, 2, and 3 shown. This allows for the easy visualization of compounds belonging only to 1 fraction (blue edges and on the peripherals), 2 fractions (grey edges and on the peripherals) and all three compounds (center of the graph).

Ideally, each fraction should contain at most 20 compounds. One way to do this would be to

keep the 71 subfractions separate and characterize each one individually. However, analyzing 71 subfractions at 20 hours a fraction would take over two months. One way to minimize the analysis time would be to reduce the acquisition time to only 20 minutes, which would require a more manageable 30-hour timeframe. This may be feasible for the smaller number of compounds present in each fraction, as they may require less sensitivity to detect each compound.

The large number of common compounds clearly visualized in Figure 5 B, and suggested in Figure 5 A, indicates poor separation of the crude extract. This could be due to overloading the C-18 chromatography column. Either decreasing the amount of extract that is separated or increasing the amount of column packing material to increase resolution of compounds should be done. Unfortunately, decreasing the amount of extract loaded on the column would greatly impact sensitivity. Therefore, the more prudent choice would be to increase the amount of column material, even though there will be more solvent to remove before analysis.

As shown in Figure 5, the networks allow one to decide upon a different order to pool fractions. Fraction 1 and 3 are quite large nodes, indicating a large number of compounds being returned from the database. If these fractions were separated in more, smaller fractions, the compounds would be easier to identify. Additionally, fractions 6 and 7 are quite small and close together on the network. This seems to indicate that they contain similar compounds because similar forces from the compounds affect them. They should be pooled together to increase sensitivity and reduce network complexity.

Ideally, the network would be constructed from the peaks, rather than compound assignments from the database. This would remove some of the uncertainties and limitations of the databases, such as very limited amounts of compounds in the databases and the uncertain assignment of compounds that are in the database. Additionally, incorporating only the peaks reduces analyst bias, where compounds that are not expected to be bioactive are dismissed out of hand. Unfortunately, due to slight changes in temperature and solvent composition, there is minor variation in the chemical shift of the same peak between runs. This means that a single peak coordinate cannot be used as a common compound identifier between fractions. Instead, a bucketing approach[3,4] must be used, dividing the two dimensional spectra into a grid. The increments in NMR bucketing are typically 0.04 ppm for a one-dimensional spectrum, though due to the approximately 10-fold increase in ppm range for ^{13}C, it seems reasonable to believe

that increments of 0.4 would provide sufficient resolution for a ^{13}C spectra. Thus, a 2D HSQC spectra, such as the one used for the compilation of the networks, consisting of 10.992 ppm in the ^{1}H dimension and 173 ppm in the ^{13}C dimension would consist of a 275 ^{1}H buckets X 433 ^{13}C buckets for a total of 119075 data points. If the apex of a peak falls in a bucket, the corresponding intensity is considered the value for the entire bucket, and the peak can be related to peaks in other fractions that fall in the same range. One of the downsides of this method is when multiple peaks appear in the same bucket. In this case, some sensitivity is lost as the peak intensity are added together, but as such deviation could not be detected by a database due to the database treating each peak as a range, such points are academic at best. The bioinformatics programs to do this sort of bucketing already exist for LC-MS,[21] where the two dimensions are the retention times and the masses of the compounds being detected instead of ^{1}H and ^{13}C chemical shifts. Therefore, it would be relatively simple to adapt these programs for two-dimensional NMR spectra.

4 Conclusions and Future Work

4.1 Complex mixture

There is still more interpretively useful data to be gleaned from simple, few compound mixtures, both of the compounds discussed here and of other mixtures. A dilution series of the mixture should be done to determine the limits of detection and optimize the instrumental parameters to give a balance between sensitivity and experiment time. An experiment that was not done was an HSQC analysis of the composite mixture. This would allow the effects of querying multiple peaks on the certainty of assignment to be empirically determined. Additionally, compounds that vary only slightly (alcohol group versus methoxide group) should be examined to explore whether this technique can differentiate between analogues of the same compounds, or if the technique is limited to only determining classes of compounds. However, even if NMR metabolomics can only identify classes of compounds, the technique is still a powerful tool for characterizing natural product extracts.

The complex mixture also provides a useful platform for examining the effects of buffering the NMR solvent on chemical shifts. This is important because chemical shifts of nuclides can vary up to 1 ppm for ^{1}H and up to 10 ppm for ^{13}C depending on whether the molecule is in the protonated, acidic form, or the deprotonated, basic form.[22] Different solutions of methanol-d_4 can

be made with varying concentrations and pH values. Chemical shifts of known compounds could be compared at these different conditions. Ideally, the buffering of methanol would prove unnecessary, as buffering either requires expensive deuterated buffers, or introduces a large water peak into the NMR spectra, which obscures many finer peaks. Typically, the suggested pH is around 7.3, probably due to that being close to physiological pH.

4.2 *E. COLI* LYSATE

The *E. coli* lysate provided valuable insight into the importance of pulse program selection, as well as clarifying the effects of the TD and NS parameters on spectra sensitivity and resolution. For the purpose of this experiment, it was only included in order to produce a defined, complex mixture to confirm the database results from. Due to the poor sensitivity displayed, the spectra gathered did not display enough information to query the peaks correctly. However, beyond this experiment, examining the metabolic profile of bacteria allows perturbations, be it genetic based or chemical, to be examined on a metabolomics level. This could be applied to examining the biosynthesis of natural products from bacterial transformants in order to optimize feedstock and minimize side reactions. Introducing carbon-13 labeled feedstock could increase sensitivity. This would increase sensitivity 100-fold in experiments involving carbon, as well as distinguishing the metabolites of the feedstock from the rest of the metabolome. Changing the focus of the lysate to examine more nonpolar constituents is also a possible direction. To do so, size exclusion chromatography packing is introduced to the bacterial culture. Exuded small molecules are adsorbed by the packing material, and after the bacterial culture is grown to saturation, the small organic molecules are extracted by methanol.[23] The gathered compounds are more nonpolar than those gathered through methanol/chloroform extraction. These compounds can then be analyzed through metabolomics experiments to compile a natural product library.

4.3 *G. SQUARROSA* EXTRACTS

One of the first things that still needs to be done with the *G. squarrosa* extract is to analyze the remaining fractions by NMR spectroscopy, as only fractions 1-7 and fraction 11 were analyzed. The database results of all the fractions should be visualized on a network. This network can be used to determine which fractions have more compounds–and thus should be further split into more fractions–and which fractions display very similar compounds–which should be pooled. This information should guide the subfraction pooling of a second reverse phase column

chromatography with identical parameters to the first, described in section 2.3.3.

The new fractions should be analyzed by NMR spectroscopy and queried against the BMRB database. The results would then be visualized in a network; this network would be used to determine the uniqueness and number of constituents of each fraction. The list of unique compounds from each column chromatographic separation should be compared to indicate the reproducibility of both the database results and the plant composition.

Once that is done, the fractions should be tested for activity. The purpose of this step is to gain biological profile, rather than looking for specific bioactive constituents. Several different approaches exist for biological profiling. One such method is to compare the expression levels of several reported genes upon treatment with the fractions.[24] The expression levels can be compared with the expression levels upon treatment with compounds of known modes of action, or upon treatment with siRNAs (genetic knockdown). In either case, similar expression levels between fraction and known perturbagen is interpreted as a compound that acts on the same pathway as the known perturbagen.

In a similar way, cell morphology can be visualized using fluorescent dyes.[21,23,25] Changes in morphology can be statistically correlated between fractions, as well as with known perturbagens. One large advantage of this platform is that it already designed to integrate with a network compiled from MS-based metabolomics. The network changes from visualizing the relationship between compounds and fraction to visualizing the relationship between compounds and fraction activity. This process is improved through increasing the number of fractions analyzed, as the fewer compounds present in a fraction to be tested for activity the easier it is to identify the single compound causing that activity.

Even without added biological activity, the fraction networks can be further analyzed. If the weight of the lines were changed to peak intensity, normalized to the TMS added in a constant concentration, the edge lengths would be related to concentration, rather than the peak match value returned by the database. Unfortunately, the peaks in two-dimensional spectra are not directly related to concentration, as in one-dimensional proton NMR. Other factors, such as magnitude of the coupling constant, J, and NMR parameters such as using a decoupling irradiation pulse, and changing mixing time in TOCSY type experiments also influence the intensity of the peaks. However, as long as the instrument parameters are constant, the influences

other than compound concentration should also remain constant. When linking metabolomic profile with biological profile, having some indication of molecular concentration allows a dose response relationship to be observed, helping to narrow in on the biologically active compound.

4.4 THE DIRECTION OF BIOINFORMATICS

The programs developed over the course of this project were limited in function. Ideally, there is more functionality that should be included in each program, particularly NMR_reader.py. Noise can be filtered from HSQC type spectra by incorporating the intensity of each peak into the output of NMR_reader.py. Then the highest intensity peak would be retained, discarding the smaller, noise peaks. Noise can also be visualized in COSY spectra. If the spectra are not symmetrized, noisy peaks show up as intense vertical lines. Any peaks in the output of NMR reader that have more than four peaks in it should be discarded, as they are most likely the result of noise. Another feature that would be useful is to add a way to "walk" along related peaks by referencing entries in the list of F2 chemical shifts to the key of the reversed library, eventually compiling entire substructures in one table entry.

Ideally, the NMR data should be manipulated so as to return excellent results in the MMCD.[20] This database contains the largest number of compounds and allow multiple inputs to be queried at once, increasing the validity of the results.

4.5 CONCLUDING REMARKS

Metabolomics methods using the NMR spectrometer were developed. After considering several pulse programs, HSQC showed the greatest promise for producing well resolved and simple spectra. Querying the list of peaks produced from the HSQC NMR spectra of the fractionated *G. squarrosa* extract against the HMBC database returned 597 unique compounds. Querying the crude extract in the same manner returned between 217 and 477 compounds. The four-compound mixture returned 815 compounds from the querying of one-dimension of the COSY spectrum. This large number of compounds returned from a relatively simple mixture raised questions about the quality of the database assignments. To improve the assignment, more peaks known to be on the same molecule need to be queried against the database. As the databases are not configured to accept multiple related sets of peaks at a time, relating peaks with biological activity, and only analyzing those subsystems that correlate with biological activity is proposed as a potential work around. This reduces the number of peaks that need to be analyzed. This

integration would result in a powerful technique that is complementary to the currently used MS-based metabolomics, further exploring chemical space.

5 LITERATURE CITED

(1) Dona, A. C.; Kyriakides, M.; Scott, F.; Shephard, E. A.; Varshavi, D.; Veselkov, K.; Everett, J. R. *Comput. Struct. Biotechnol. J.* **2016**, *14*, 135–153.

(2) Kurita, K. L.; Linington, R. G. *J. Nat. Prod.* **2015**, *78* (3), 587–596.

(3) Smolinska, A.; Blanchet, L.; Buydens, L. M. C.; Wijmenga, S. S. *Anal. Chim. Acta* **2012**, *750*, 82–97.

(4) Leenders, J.; Frédérich, M.; de Tullio, P. *Drug Discov. Today Technol.* **2015**, *13*, 39–46.

(5) Larive, C. K.; Barding, G. A.; Dinges, M. M. *Anal. Chem.* **2015**, *87* (1), 133–146.

(6) Öman, T.; Tessem, M.-B.; Bathen, T. F.; Bertilsson, H.; Angelsen, A.; Hedenström, M.; Andreassen, T. *BMC Bioinformatics* **2014**, *15* (1), 413.

(7) Mahrous, E. A.; Farag, M. A. *Drug Discov.* **2015**, *6* (1), 3–15.

(8) Kruger, N. J.; Ratcliffe, R. G.; Roscher, A. *Phytochem. Rev.* **2003**, *2* (1–2), 17–30.

(9) Krishnan, P.; Kruger, N. J.; Ratcliffe, R. G. *J. Exp. Bot.* **2005**, *56* (410), 255–265.

(10) Hu, Y.; Potts, M. B.; Colosimo, D.; Herrera-Herrera, M. L.; Legako, A. G.; Yousufuddin, M.; White, M. A.; MacMillan, J. B. *J. Am. Chem. Soc.* **2013**, *135* (36), 13387–13392.

(11) Veres, K.; Roza, O.; Laczkó-Zöld, E.; Hohmann, J. *Nat. Prod. Commun.* **2014**, *9* (4), 573–574.

(12) Bingol, K.; Bruschweiler-Li, L.; Li, D.-W.; Brüschweiler, R. *Anal. Chem.* **2014**, *86* (11), 5494–5501.

(13) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Kent Wenger, R.; Yao, H.; Markley, J. L. *Nucleic Acids Res.* **2008**, *36* (Database issue), D402-408.

(14) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. *Genome Res.* **2003**, *13* (11), 2498–2504.

(15) Féraud, B.; Govaerts, B.; Verleysen, M.; de Tullio, P. *Metabolomics* **2015**, *11* (6), 1756–1768.

(16) Xi, Y.; de Ropp, J. S.; Viant, M. R.; Woodruff, D. L.; Yu, P. *Anal. Chim. Acta* **2008**, *614* (2), 127–133.

(17) Rai, R. K.; Sinha, N. *Anal Chem* **2012**, *84*, 10005–10011.

(18) López-Pérez, J. L.; Therón, R.; del Olmo, E.; Díaz, D. *Bioinformatics* **2007**, *23* (23), 3256–3257.

(19) Johnson, S. R.; Lange, B. M. *Front. Bioeng. Biotechnol.* **2015**, *3*, 22.

(20) Cui, Q.; Lewis, I. A.; Hegeman, A. D.; Anderson, M. E.; Li, J.; Schulte, C. F.; Westler, W. M.; Eghbalnia, H. R.; Sussman, M. R.; Markley, J. L. *Nat. Biotechnol.* **2008**, *26* (2), 162–164.

(21) Kurita, K. L.; Glassey, E.; Linington, R. G. *Proc. Natl. Acad. Sci.* **2015**, *112* (39), 11999–12004.

(22) Platzer, G.; Okon, M.; McIntosh, L. P. *J. Biomol. NMR* **2014**, *60* (2), 109–129.

(23) Schulze, C. J.; Bray, W. M.; Woerhmann, M. H.; Stuart, J.; Lokey, R. S.; Linington, R. G. *Chem. Biol.* **2013**, *20* (2), 285–295.

(24) Potts, M. B.; Kim, H. S.; Fisher, K. W.; Hu, Y.; Carrasco, Y. P.; Bulut, G. B.; Ou, Y.-H.; Herrera-Herrera, M. L.; Cubillos, F.; Mendiratta, S.; Xiao, G.; Hofree, M.; Ideker, T.; Xie, Y.; Huang, L. J.; Lewis, R. E.; MacMillan, J. B.; White, M. A. *Sci. Signal.* **2013**, *6* (297), ra90.

(25) Woehrmann, M. H.; Bray, W. M.; Durbin, J. K.; Nisam, S. C.; Michael, A. K.; Glassey, E.; Stuart, J. M.; Lokey, R. S. *Mol. Biosyst.* **2013**, *9* (11), 2604–2617.

APPENDIX A

Table 4. A sample of the first 20 peaks with the highest intensity from the combination of the four trial samples (limonene, vanillin, coumarin, menthol).

Peak	(F2) [ppm]	(F1) [ppm]	Intensity [abs]
1	1.6561	7.0413	31298
2	7.0456	1.6568	31298
3	5.7015	5.7114	31338
4	5.7146	5.6984	31338
5	1.2646	1.5134	31366
6	1.5126	1.2657	31366
7	1.1211	1.396	31396
8	1.3951	1.1223	31396
9	1.6953	7.4063	31554
10	7.411	1.6959	31554
11	0.7557	1.5264	31558
12	1.5256	0.7572	31558
13	2.0085	2.2956	31632
14	2.2955	2.0088	31632
15	7.4241	7.8235	31756
16	7.8286	7.4194	31756
17	0.8731	3.9514	31970
18	3.9529	0.8745	31970
19	1.7475	7.6931	32008
20	7.6981	1.7481	32008

Table 5. The first 28 rows of the output of NMR_reader.py using the four compound mixture as an input.

F1 (ppm)	F2 (ppm)					
0.9006	0.9006					
0.9137	0.9137					
-0.0118	-0.0118	7.7062	7.1065	7.4194	7.5237	7.4976
7.2629	7.2629	7.3411	6.4285			
-0.1422	-0.1422					
1.6438	1.6438					
0.8224	0.8224	7.2108	7.2629	7.289		

3.9775	3.9775					
1.6698	1.6698					
4.6945	4.6945	6.598				
4.0035	4.0035	2.1783				
7.4194	7.4194	9.6227				
0.7963	0.7963	7.6149	7.4324			
9.8182	9.8182					
1.722	1.722					
0.9267	0.9267	7.0934	7.4976			
0.1444	0.1444					
0.1053	0.1053					
0.0793	0.0793					
0.001	0.001	1.722				
0.8745	0.8745	3.9775				
0.8094	0.8094	4.7076	9.6227			
0.8615	0.8615	7.5497	7.4715			
2.1392	2.1392					
2.1913	2.1913					
2.1653	2.1653	1.6307	0.8094	7.4194	1.722	9.6227
1.5264	1.5264					
1.5916	1.5916	4.4207	4.4729	1.6568		

APPENDIX B

```
#NMR_reader.py

'''To compile peaks with the same chemical shift as one entry in a dictionary'''
import pandas as pd
import numpy as np


def F1(table, number):
  '''defines the F1 direction in an easy to reference way'''
  return table.loc[number, '?(F1) [ppm]']

def F2(table, number):
  '''defines the F2 direction in an easy to reference way'''
  return table.loc[number, '?(F2) [ppm]']


def dict_maker(file):
  '''converts the table of peaks into a dictionary, where the F2 peak is the
  key and all F1 chemical shifts attached to the key are compiled into a list.
  This removes redundant F2 peaks'''
  spectra = pd.read_csv(file)

  peaks = {}
  for number in range(spectra.shape[0]):
    peak = [F2(spectra, number)]

    connected = []

    true_spectra = spectra.isin(peak)

    if F2(spectra, number) not in peaks:
      for x in range(number, spectra.shape[0]):
        if F2(true_spectra, x) == True:
          connected.append(F1(spectra, x))

        peaks[F1(spectra, number)] = connected

  return peaks

def dict_maker_rev(file):
  '''converts the table of peaks into a dictionary, where the F1 peak is the
  key and all F2 chemical shifts attached to the key are compiled into a list.
```

```
    This removes redundant F1 peaks"'
    spectra_rev = pd.read_csv(file)
    peaks_rev = {}
    for n in range(spectra_rev.shape[0]):
        #returns the reverse of the dict_maker loop
        peak_rev = [F1(spectra_rev, n)]
        connected_rev = []
        true_spectra = spectra_rev.isin(peak_rev)
        if F1(spectra_rev, n) not in peaks_rev:
            for x in range(n, spectra_rev.shape[0]):
                if F1(true_spectra, x) == True:
                    connected_rev.append(F2(spectra_rev, x))
                peaks_rev[F1(spectra_rev, n)] = connected_rev

    return peaks_rev


"'runs both the dict_maker and dict_maker_rev functions on one data set"'
file = input("What is your filepath? ")
product = input("Where should this be written? ")
product_rev = input("Where should the reverse be written? ")

lysate = dict_maker(file)

lysate_rev = dict_maker_rev(file)

output = open(product, "w")

for key in lysate:
    print (key, lysate[key], file = output)
output.close()

output_rev = open(product_rev, "w")

for key in lysate_rev:
    print (key, lysate_rev[key], file = output_rev)
output_rev.close()
```

```
# NMR_network.py

import networkx as nx
import pandas as pd

# makes empty graph and empty nodes list
G = nx.Graph()
listything = []

# open csv file that contains the list of nodes as a dataframe
# make sure to label the column containing the nodes as "Nodes"
fraction_1 = pd.read_csv("/Users/jasonmcfarlane/Downloads/nodes.csv")

'''iterates through the "Nodes" column and inputs each entry into the empty
nodes list. G.add_nodes_from()creates a nodes table from this list, ignoring
any redundant entries'''
for n in range(len(fraction_1.index)):
  s = fraction_1["Nodes"]
  listything.append(s[n])
G.add_nodes_from(listything)

'''Makes a dataframe from a table of edges. The first column contains a list
of fractions and is labelled "Fractions". The second column contains the
compounds returned in each fraction. The weight column is a value returned from
the database, and represents the certainty of the compound assignment. A "1" is
a high likelihood, while "0" is low likelihood.'''
edges_list = pd.read_csv("/Users/jasonmcfarlane/Downloads/yuple_edit10%.csv")

for n in range(len(edges_list.index)):
  first = edges_list["Fractions"]
  second = edges_list["compound"]
  weight = edges_list["weight"]
  G.add_edge(first[n], second[n], weight=float(weight[n]))

'''Writes the graph to an xml file for easy visualization using Cytoscape'''
nx.write_graphml(G, "/Users/jasonmcfarlane/Desktop/weightedgraph_edit10%.xml")
```

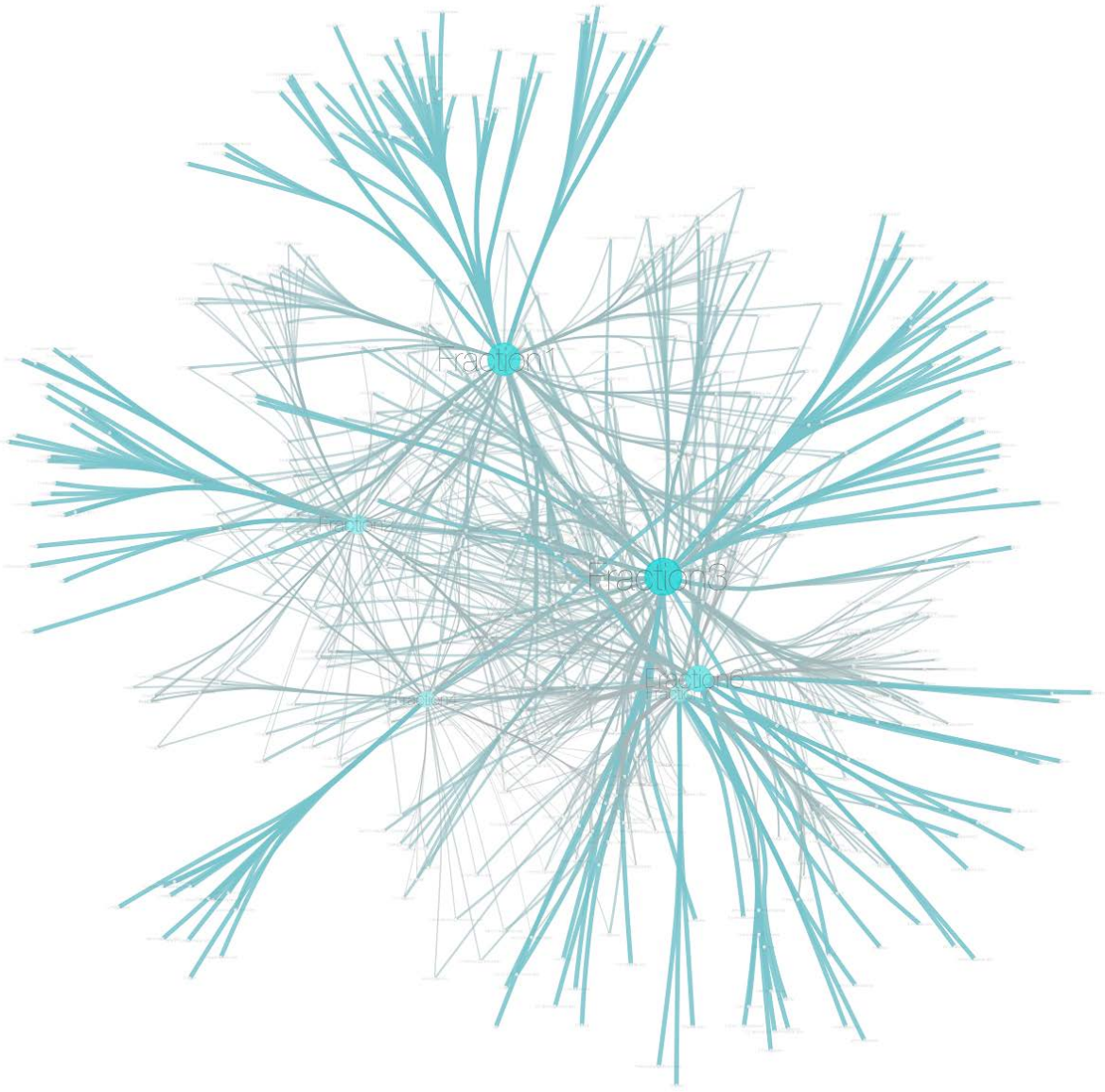


Figure 6. This shows the same six fractions shown in Figure 5 with all compounds that had a Peak Match in the BMRB database equal or lower to 0.10. This filter removed 60 compounds from the dataset leaving 534 compounds.