METHODS FOR DIFFERENTIATION AND PHYTOCHEMICAL INVESTIGATION OF CRATAEGUS BY NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

by

Jensen Allan Lund

B.Sc. (Hons.), The University of British Columbia, 2013

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

MASTER OF SCIENCE

in

THE COLLEGE OF GRADUATE STUDIES

(Biochemistry and Molecular Biology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

April 2016

© Jensen Allan Lund, 2016
The undersigned certify that they have read, and recommend to the College of Graduate Studies for acceptance, a thesis entitled:

Methods for Differentiation and Phytochemical Investigation of Crataegus by Nuclear Magnetic Spectroscopy

Submitted by Jensen Allan Lund in partial fulfillment of the requirements of

The degree of Master of Science

Dr. Paul Shipley, Chemistry, UBCO
Supervisor, Professor

Dr. Susan Murch, Chemistry, UBCO
Supervisory Committee Member, Professor

Dr. Soheil Mahmoud, Biology, UBCO
Supervisory Committee Member, Professor

Dr. Daniel Durall, Biology, UBCO
University Examiner, Professor

External Examiner, Professor

April 25, 2016
(Date submitted to Grad Studies)

Additional Committee Members include:

Dr. Paula Brown, NRG, BCIT
(Please print name and faculty/school in the line above)
Abstract

*Crataegus* (hawthorn) is a genus of flowering, fruit-bearing, small-to-medium-sized trees native to northern temperate zones. Hawthorn has an extensive ethnobotany with numerous examples of use in the food and medicine of the Chinese, North American Aboriginals, and Europeans. Modern hawthorn natural health products (NHPs) are used for cardiovascular ailments and have shown promise for adjunctive therapy in treatment of chronic heart failure. Although many NHPs containing *Crataegus* are licensed for sale in Canada, and there is extensive ethnobotanical use, little is known about the potentially therapeutic phytochemistry of North American hawthorns. NHP monographs allow three *Crataegus* spp. for use in NHPs. Regulatory requirements demand that specifications for NHPs include identity of the source material. Hence, it is important that there are methods available to differentiate species and this is the rationale for the work in this thesis. The first objective of this thesis was to explore differentiation of *Crataegus* spp. by nuclear magnetic resonance (NMR) spectroscopy and multivariate data analysis (MVDA) of leaf extracts. Two European and two North American species were harvested in late summer 2008 from the same test farm and used in this study. The results suggest that we may best differentiate *Crataegus* spp. by 1D $^1$H PRESAT NMR-acquired data that correlate to phenolic compounds. Multiple significance analysis methods were used to provide lists of potential biomarkers in the phenolics spectral region that may give insight to differential therapeutic effects of hawthorn NHPs. Machine learning was applied to chemometrics data and results correlated with and outperformed model accuracy in traditional methods of MVDA. The second objective of this thesis was to quantify flavonoids in the same samples. I found that vitexin and its derivatives were significantly more concentrated in the European leaves and
rutin significantly more concentrated in the North American leaves. The concentrations of rutin and naringenin reported in this study are the highest reported for *Crataegus*. The results were validated by comparison to high performance liquid chromatography, which produced generally consistent results of mean concentrations as measured in the leaf extracts.
Preface

The literature necessary to write the traditional Chinese medicine section in Chapter 1.5.1 was translated with the assistance of Ying Liu.


The HPLC analysis presented in Chapter 3 is based on work conducted at BCIT’s Natural Health and Food Products Research Group laboratory in Burnaby, BC with the assistance of Dr. Paula Brown and Elizabeth Mudge. The validated method for HPLC analysis of hawthorn leaf flavonoids was developed by this group (Mudge and Brown, 2016). I was responsible for participating in extraction optimization, preparation of hawthorn leaf extracts for analysis, analysis by HPLC, and quantification of compounds based upon chromatographic data.
# Table of Contents

Abstract........................................................................................................................................................................... iii

Preface................................................................................................................................................................................. v

Table of Contents ........................................................................................................................................................ vi

List of Tables ........................................................................................................................................................................ x

List of Figures ....................................................................................................................................................................... xiv

List of Abbreviations ......................................................................................................................................................... xix

Acknowledgements ............................................................................................................................................................ xxii

Dedication .......................................................................................................................................................................... xxiv

## Chapter 1: Introduction .................................................................................................................................................. 1

1.1 Botany and ecology of *Crataegus* .............................................................................................................................. 1

1.2 Classical taxonomy and identification of *Crataegus* spp. ........................................................................................ 3

1.3 Molecular and chemical techniques for taxonomy and species identification of *Crataegus* ........................................... 12

1.3.1 Chemotaxonomy ...................................................................................................................................................... 16

1.3.2 Metabolomics and its relationship to chemotaxonomy ........................................................................................ 16

1.3.3 Localization of specialized metabolites with respect to plant anatomy .............................................................. 18

1.4 Phytochemistry of *Crataegus* ................................................................................................................................ 20

1.5 Ethnobotany and modern natural health products (NHPs) of *Crataegus* ................................................................. 29

1.5.1 Ethnobotanical use of *Crataegus* ....................................................................................................................... 29

1.5.2 Modern use of *Crataegus* in natural health products (NHPs) and dietary supplements (DSs) ............................................ 31

1.6 Regulatory requirements of natural health products (NHPs) and dietary supplements (DSs) ........................................... 33

1.6.1 Identification of botanical materials for use in natural health products (NHPs) and dietary supplements (DSs) ................. 34

1.6.2 Evaluation of strength and potency of botanical materials for use in natural health products (NHPs) and dietary supplements (DSs) ................................................................................................. 35

1.6.3 Regulatory requirements for *Crataegus* natural health products (NHPs) and dietary supplements (DSs) ......................... 35
Chapter 2: An evaluation of nuclear magnetic resonance data acquisition and multivariate data modeling for differentiation of Crataegus spp. 53

2.1 Synopsis .............................................................................................................. 53
2.2 Research objectives ............................................................................................. 54
2.3 Experimental ......................................................................................................... 54
   2.3.1 Plant materials ............................................................................................... 54
   2.3.2 Experimental materials ................................................................................ 55
   2.3.3 Sample preparation ...................................................................................... 55
   2.3.4 Data acquisition ............................................................................................ 57
   2.3.5 Spectral data processing ............................................................................... 58
   2.3.6 Exploratory data analysis ............................................................................. 59
   2.3.7 Metabolite significance analysis .................................................................. 61
2.4 Results and discussion ........................................................................................ 63
   2.4.1 Exploratory data analysis ............................................................................. 63
   2.4.2 Supervised models ...................................................................................... 67
   2.4.3 Machine learning ......................................................................................... 73
   2.4.4 Metabolite significance analysis .................................................................. 75
   2.4.5 Summary and applications of findings ......................................................... 93
Chapter 3: Quantification of *Crataegus* flavonoids by nuclear magnetic resonance spectroscopy

3.1 Synopsis .............................................................................................................................. 96
3.2 Research objectives .......................................................................................................... 97
3.3 Experimental ..................................................................................................................... 97
  3.3.1 Plant materials ............................................................................................................. 97
  3.3.2 NMR experimental materials ...................................................................................... 98
  3.3.3 NMR sample preparation ............................................................................................ 98
  3.3.4 NMR data acquisition ................................................................................................. 99
  3.3.5 NMR spectral data processing .................................................................................. 99
  3.3.6 NMR instrumental detection limits .......................................................................... 100
  3.3.7 HPLC experimental materials .................................................................................. 101
  3.3.8 HPLC sample preparation ......................................................................................... 101
  3.3.9 Chromatographic instrumentation and conditions .................................................... 101
  3.3.10 Chromatographic data processing ......................................................................... 102
  3.3.11 Statistical data processing ....................................................................................... 103
3.4 Results and discussion .................................................................................................... 103
  3.4.1 Quantification of flavonoids in *Crataegus* leaf extracts ........................................... 103
  3.4.2 Implications of quantification results for *Crataegus* utility ..................................... 119
  3.4.3 Summary and applications of findings ......................................................................... 121

Chapter 4: Conclusion ............................................................................................................ 122

References ............................................................................................................................... 126

Appendices .............................................................................................................................. 142

Appendix A: Optimization of *Crataegus* leaf extraction for quantitative analysis by nuclear magnetic resonance spectroscopy ............................................................................................................. 142

Appendix B: Selection of T<sub>1</sub> relaxation time for quantitative analysis by nuclear magnetic resonance spectroscopy .................................................................................................................. 145

Appendix C: Signals used for quantitative analysis by nuclear magnetic resonance spectroscopy ........................................................................................................................................ 147

Appendix D: Standard curves used for quantitative analysis by high performance liquid chromatography – diode array detection platform .......................................................................................... 151
Appendix E: Summary of significance analysis of phytochemical differences among four *Crataegus* spp................................................................. 152
Appendix F: Standard curve for determination of NMR instrument detection limits and linearity ................................................................................................................................. 156
List of Tables

Table 2.1: *Crataegus* spp. and sample IDs used for generation of chemometric models in untargeted experiments throughout chapter 2 and quantification of flavonoids in chapter 3........................................................................................................................................55

Table 2.2: Summary of random forest (RF) classification performance for four *Crataegus* spp. Classification performance is compared for two NMR acquisition techniques and with four spectral regions. Classification performance is specified as out-of-bag (OOB) error. Species misclassified in confusion matrices are noted with individual within class error (WCE) probabilities.................................................................74

Table 2.3: Summary of significant signals for differentiation of four *Crataegus* spp. from entire spectrum for pJRES. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM d(i) values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed. Putative identifications are listed where possible. VT=vitexin.................................................................................................................................79

Table 2.4: Summary of significant signals for differentiation of four *Crataegus* spp. from entire spectrum for $^1$H 1D PRESAT. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM d(i) values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed. ..........................................................................................80

Table 2.5: Summary of significant signals for differentiation of four *Crataegus* spp. from organic acids region for pJRES. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM d(i) values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed.........................................................................................81

Table 2.6: Summary of significant signals for differentiation of four *Crataegus* spp. from organic acids region for $^1$H 1D PRESAT. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM d(i) values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed.................................................................82
Table 2.7: Summary of significant signals for differentiation of four *Crataegus* spp. from sugars region for pJRES. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM $d(i)$ values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed................................. 83

Table 2.8: Summary of significant signals for differentiation of four *Crataegus* spp. from sugars region for $^1$H 1D PRESAT. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM $d(i)$ values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed................................. 84

Table 2.9: Summary of significant signals for differentiation of four *Crataegus* spp. from phenolics region for pJRES. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM $d(i)$ values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed. Putative identifications are listed where possible. QD=quercetin; CA=chlorogenic acid; HP=hyperoside; IV=isovitexin; VT=vitexin; VR=vitexin-$2''$-O-rhamnoside; RT=rutin. .............................................................. 85

Table 2.10: Summary of significant signals for differentiation of four *Crataegus* spp. from phenolics region for $^1$H 1D PRESAT. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM $d(i)$ values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed. Putative identifications are listed where possible. QD=quercetin; CA=chlorogenic acid; HP=hyperoside; IV=isovitexin; VT=vitexin; VR=vitexin-$2''$-O-rhamnoside; RT=rutin. .............................................................. 87

Table 2.11: Summary of significant phenolic signals for differentiation of closely related *Crataegus* spp. by pJRES. Significant signals are listed for pJRES NMR acquisition technique by cross-validation (CV) performance in models generated by support vector machine (SVM) learning. Spectral bin significance is listed for differentiation of either *C. monogyna* from *C. laevigata* or *C. douglasii* from *C. okanaganensis*. Importance measures are reported as both frequency of selection (FoS) and mean importance measure (MIM). Putative identifications are listed....... 89
Table 2.12: Summary of significant phenolic signals for differentiation of closely related *Crataegus* spp. by 1D $^1$H PRESAT. Significant signals are listed for 1D $^1$H PRESAT NMR acquisition technique by cross-validation (CV) performance in models generated by support vector machine (SVM) learning. Spectral bin significance is listed for differentiation of either *C. monogyna* from *C. laevigata* or *C. douglasii* from *C. okanaganensis*. Importance measures are reported as both frequency of selection (FoS) and mean importance measure (MIM). Putative identifications are listed where possible.

Table 2.13: Summary of significant phenolic signals for differentiation of European and North American *Crataegus* spp. by 1D $^1$H PRESAT and pJRES. Significant signals are listed for 1D $^1$H PRESAT and pJRES NMR acquisition technique by cross-validation (CV) performance in models generated by support vector machine (SVM) learning. Spectral bin significance is listed for differentiation of *C. monogyna* and *C. laevigata* from *C. douglasii* and *C. okanaganensis*. Importance measures are reported as both frequency of selection (FoS) and mean importance measure (MIM). Putative identifications are listed where possible.

QD=quercetin; CA=chlorogenic acid; HP=hyperoside; IV=isovitexin; VT=vitexin; VR=vitexin-2″-O-rhamnoside; RT=rutin.

Table 3.1: Instrumental limits of detection for compounds measured by quantitative NMR in Chapter 3. Concentrations are expressed as mg/g dry leaf equivalents. LOD is instrument limit of detection and LOQ is instrument limit of quantification.

Table 3.2: Concentrations of flavonoids in leaf extracts of four *Crataegus* spp. as measured by NMR (eight compounds measured) and HPLC-DAD (five compounds measured) (n=9 per species, per method [except *C. laevigata* n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material and expressed as mg/g dry leaf. Error represented as standard error of the mean. ‘ND’ indicates compound not detected in sample. ‘<LOQ’ indicates compound was detected in sample, however below the instrumental limit of detection.
Table A-1: Experimental schematic for half-factorial analysis of four factors on extraction efficiency.

Table C-1: Compounds of interest in *Crataegus* and their associated signals. Compound and associated proton environments that correspond to signal for quantification shown in left column. Example spectra shown in right column with chemical standard (top) and hawthorn extract sample (bottom).
List of Figures

Figure 1.1: Distribution of *Crataegus monogyna* around the world. Biodiversity occurrence data according to the Global Biodiversity Information Facility (GBIF, 2015). ..........................7

Figure 1.2 Distribution of *Crataegus laevigata* around the world. Biodiversity occurrence data according to the Global Biodiversity Information Facility (GBIF, 2015). 8

Figure 1.3 Distribution of *Crataegus douglasii* around North America. Biodiversity occurrence data according to the Global Biodiversity Information Facility (GBIF, 2015) .......................................................... 9

Figure 1.4 Distribution of *Crataegus suksdorfii* around North America. Biodiversity occurrence data according to the Global Biodiversity Information Facility (GBIF, 2015) .......................................................... 10

Figure 1.5 Distribution of *Crataegus okanaganensis* around North America. Biodiversity occurrence data according to the Global Biodiversity Information Facility (GBIF, 2015) .......................................................... 11

Figure 1.6 Chemical structures of one hydroxycinnamic acid and eight flavonoids present in various *Crataegus* spp. .......................................................... 24

Figure 1.7: Reports of quantitative NMR spectroscopy in scientific literature as indexed by Web of Science on November 27, 2015. Search terms are as follows: (qNMR) OR ("quantitative NMR") OR ("quantification by NMR") OR ("quantitation by NMR"). .......................................................... 49

Figure 2.1: Site map depicting geographical location of *Crataegus* samples used for generation of chemometric models in untargeted experiments throughout Chapter 2 and quantification of flavonoids in Chapter 3............................... 56

Figure 2.2: PCA scores plots generated from NMR spectra of *Crataegus* extracts. Plots A, C, E, and G represent models generated from spectra acquired by pJRES. Plots B, D, F, H represent models generated from spectra acquired by 1H 1D PRESAT. Plots A, B were generated from the whole spectrum. Plots C, D were generated from the organic acid region of the spectrum. Plots E, F were generated from the sugars region of the spectrum. Plots G, H were generated from the phenolics
region of the spectrum. Ellipsoids around classes represent 95% confidence intervals within class…………………………………………………………………………………………………….. 65

Figure 3.1: Concentration of vitexin-2-O-rhamnoside in leaf extracts of four Crataegus spp. as measured by NMR and HPLC-DAD (n=9 per species, per method [except C. laevigata n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material. Upper whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol (⊕). Median concentration is represented by horizontal line. Boxplot not shown for C. douglasii and C. okanaganensis as measured by NMR since concentrations are below instrument LOQ. ………………………………………………………………………………………………………….. 107

Figure 3.2: Concentration of vitexin in leaf extracts of four Crataegus spp. as measured by NMR and HPLC-DAD (n=9 per species, per method [except C. laevigata n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material. Upper whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol (⊕). Median concentration is represented by horizontal line……………….. 109

Figure 3.3: Concentration of isovitexin in leaf extracts of four Crataegus spp. as measured by NMR and HPLC-DAD (n=9 per species, per method [except C. laevigata n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material. Upper whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol (⊕). Median concentration is represented by horizontal line………………………………………………………………………………………………………………………………………110

Figure 3.4: Concentration of rutin in leaf extracts of four Crataegus spp. as measured by NMR and HPLC-DAD (n=9 per species, per method [except C. laevigata n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material. Upper
whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol (⊕). Median concentration is represented by horizontal line.

Figure 3.5: Concentration of hyperoside in leaf extracts of four *Crataegus* spp. as measured by NMR and HPLC-DAD (n=9 per species, per method [except *C. laevigata* n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material. Upper whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol (⊕). Median concentration is represented by horizontal line.

Figure 3.6: Concentration of naringenin in leaf extracts of four *Crataegus* spp. as measured by NMR (n=9 per species, per method [except *C. laevigata* n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material. Upper whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol (⊕). Median concentration is represented by horizontal line.

Figure 3.7: Concentration of chlorogenic acid in leaf extracts of four *Crataegus* spp. as measured by NMR (n=9 per species, per method [except *C. laevigata* n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material. Upper whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol (⊕). Median concentration is represented by horizontal line.

Figure A-1: Main effects plot resulting from half-factorial experiment for extraction optimization. Results expressed as mean concentration of chlorogenic acid extracted from leaf samples. Highest points correspond to factor variable with greatest level of extraction. Largest influences correlate to steepest slopes observed.

Figure A-2: Results of time dependent extraction efficiency analysis. Results expressed as mean concentration of chlorogenic acid (A) or total vitexin (B) extracted from
leaf samples. Minimum extraction duration is 15 minutes and maximum duration is 8 hours. Test samples derived from bulk sample of *C. monogyna* leaf (542a)..... 143

Figure A-3: Results of sample mass dependent extraction efficiency analysis. Results expressed as mean concentration of total vitexin extracted from leaf samples. Minimum sample mass is 10 mg and maximum sample mass is 250 mg. Test samples derived from bulk sample of *C. wilsonii* leaf (698)...................... 144

Figure B-1: Results of inversion recovery experiment for $T_1$ relaxation delay determination. Region of spectrum shown for relevant signals to quantitative analysis in Chapter 3 ($\delta$ 6-8 ppm). Test sample derived from sample of *C. monogyna* leaf (491)................................. 145

Figure D-1: Standard curve of five flavonoids as detected by HPLC-DAD. Vitexin concentration range: 1.1-210 ppm. Vitexin-2-O-rhamnoside concentration range: 0.9-435 ppm. Isovitexin concentration range: 1.0-206 ppm. Hyperoside concentration range: 1.1-530 ppm. Rutin concentration range: 1.0-198 ppm. Data shown here used in HPLC-DAD quantification of flavonoids in *Crataegus* leaf extract samples throughout Chapter 3............................... 151

Figure E-1: Tukey simultaneous 95% confidence intervals for vitexin-2"-O-rhamnoside content in leaf extracts of four *Crataegus* spp. Differences along X-axis indicate differences of means for the compound evaluated here. If an interval between two species does not contain zero, the mean concentrations of the compound are significantly different................................................................. 152

Figure E-2: Tukey simultaneous 95% confidence intervals for vitexin content in leaf extracts of four *Crataegus* spp. Differences along X-axis indicate differences of means for the compound evaluated here. If an interval between two species does not contain zero, the mean concentrations of the compound are significantly different................................................................. 153

Figure E-3: Tukey simultaneous 95% confidence intervals for isovitexin content in leaf extracts of four *Crataegus* spp. Differences along X-axis indicate differences of means for the compound evaluated here. If an interval between two species does not contain zero, the mean concentrations of the compound are significantly
different. NOTE: concentrations used in this analysis are below instrument LOQ.

Figure E-4: Tukey simultaneous 95% confidence intervals for rutin content in leaf extract of four Crataegus spp. Differences along X-axis indicate differences of means for the compound evaluated here. If an interval between two species does not contain zero, the mean concentrations of the compound are significantly different.

Figure E-5: Tukey simultaneous 95% confidence intervals for hyperoside content in leaf extracts of four Crataegus spp. Differences along X-axis indicate differences of means for the compound evaluated here. If an interval between two species does not contain zero, the mean concentrations of the compound are significantly different.

Figure E-6: Tukey simultaneous 95% confidence intervals for chlorogenic acid content in leaf extracts of four Crataegus spp. Differences along X-axis indicate differences of means for the compound evaluated here. If an interval between two species does not contain zero, the mean concentrations of the compound are significantly different.

Figure F-1: Standard curve of chlorogenic acid standard as detected by NMR. Detector response of chlorogenic acid signal (7.02 ppm, d) normalized to DSS internal standard signal at -0.018 ppm. Linearity described by R2. Standard error of the estimate, Syx, is also noted.
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D</td>
<td>one-dimensional</td>
</tr>
<tr>
<td>2D</td>
<td>two-dimensional</td>
</tr>
<tr>
<td>AHP</td>
<td>American Herbal Pharmacopoeia</td>
</tr>
<tr>
<td>AHPA</td>
<td>American Herbal Products Association</td>
</tr>
<tr>
<td>ANOVA</td>
<td>analysis of variance</td>
</tr>
<tr>
<td>AUC</td>
<td>area under the curve</td>
</tr>
<tr>
<td>BOLD</td>
<td>Barcode of Life Database</td>
</tr>
<tr>
<td>BP</td>
<td>British Pharmacopoeia</td>
</tr>
<tr>
<td>CE</td>
<td>capillary electrophoresis</td>
</tr>
<tr>
<td>CBOL</td>
<td>Consortium for the Barcode of Life</td>
</tr>
<tr>
<td>COSY</td>
<td>homonuclear correlation spectroscopy</td>
</tr>
<tr>
<td>CSV</td>
<td>comma separated value</td>
</tr>
<tr>
<td>CV</td>
<td>cross validation</td>
</tr>
<tr>
<td>DAD</td>
<td>diode array detection</td>
</tr>
<tr>
<td>DMAPP</td>
<td>dimethylallyl pyrophosphate</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DS</td>
<td>dietary supplement</td>
</tr>
<tr>
<td>DSS</td>
<td>4,4-dimethyl-4-silapentane-1-sulfonic acid</td>
</tr>
<tr>
<td>ESI</td>
<td>electrospray ionization</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>GACP</td>
<td>good agricultural and collection practices</td>
</tr>
<tr>
<td>HCA</td>
<td>hierarchical cluster analysis</td>
</tr>
<tr>
<td>HerbPro</td>
<td>Naturally Grown Herb and Spice Producers Cooperative</td>
</tr>
<tr>
<td>HMBC</td>
<td>heteronuclear multiple-bond correlation</td>
</tr>
<tr>
<td>HPLC</td>
<td>high performance liquid chromatography</td>
</tr>
<tr>
<td>HSD</td>
<td>honest significant difference</td>
</tr>
<tr>
<td>HSQC</td>
<td>heteronuclear single quantum coherence</td>
</tr>
<tr>
<td>IPP</td>
<td>isopentenyl pyrophosphate</td>
</tr>
<tr>
<td>IR</td>
<td>infrared</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>ITS</td>
<td>internal transcribed spacer</td>
</tr>
<tr>
<td>ITS2</td>
<td>nuclear ribosomal internal transcribed spacer 2</td>
</tr>
<tr>
<td>JRES</td>
<td>J-resolved NMR experiment</td>
</tr>
<tr>
<td>LOD</td>
<td>limit of detection</td>
</tr>
<tr>
<td>LOQ</td>
<td>limit of quantification</td>
</tr>
<tr>
<td>LV</td>
<td>latent variable</td>
</tr>
<tr>
<td>MCCV</td>
<td>Monte-Carlo cross validation</td>
</tr>
<tr>
<td>MEP/DOXP</td>
<td>2-C-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate</td>
</tr>
<tr>
<td>MS</td>
<td>mass spectrometry</td>
</tr>
<tr>
<td>MVDA</td>
<td>multivariate data analysis</td>
</tr>
<tr>
<td>NHP</td>
<td>natural health product</td>
</tr>
<tr>
<td>NHPID</td>
<td>Natural Health Products Ingredients Database</td>
</tr>
<tr>
<td>NMR</td>
<td>nuclear magnetic resonance</td>
</tr>
<tr>
<td>NNHPD</td>
<td>Natural and Non-prescription Health Products Directorate</td>
</tr>
<tr>
<td>nrDNA</td>
<td>nuclear ribosomal deoxyribonucleic acid</td>
</tr>
<tr>
<td>OOB</td>
<td>out-of-bag</td>
</tr>
<tr>
<td>OPC</td>
<td>oligomeric proanthocyanidin</td>
</tr>
<tr>
<td>PC</td>
<td>principal component</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
</tr>
<tr>
<td>PFTE</td>
<td>polyfluorotetraethylene</td>
</tr>
<tr>
<td>Ph. Eur.</td>
<td>European Pharmacopoeia</td>
</tr>
<tr>
<td>pJRES</td>
<td>skyline proton projection apparent decoupled J-resolved spectrum</td>
</tr>
<tr>
<td>PLS</td>
<td>partial least squares</td>
</tr>
<tr>
<td>PLS-DA</td>
<td>partial least squares-discriminant analysis</td>
</tr>
<tr>
<td>PPRC</td>
<td>Pharmacopoeia of the People's Republic of China</td>
</tr>
<tr>
<td>PRESS</td>
<td>predicted residual sum of squares</td>
</tr>
<tr>
<td>PURGE</td>
<td>presaturation utilizing relaxation gradients and echoes</td>
</tr>
<tr>
<td>RF</td>
<td>random forest</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>ROC</td>
<td>receiver operator characteristic</td>
</tr>
<tr>
<td>ROS</td>
<td>reactive oxygen species</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>RPC</td>
<td>reverse phase chromatography</td>
</tr>
<tr>
<td>SAM</td>
<td>significance of microarrays</td>
</tr>
<tr>
<td>SLV</td>
<td>single laboratory validation</td>
</tr>
<tr>
<td>SVM</td>
<td>support vector machine</td>
</tr>
<tr>
<td>TCM</td>
<td>traditional Chinese medicine</td>
</tr>
<tr>
<td>TLC</td>
<td>thin layer chromatography</td>
</tr>
<tr>
<td>TSS</td>
<td>total sum of squares</td>
</tr>
<tr>
<td>UBC</td>
<td>University of British Columbia</td>
</tr>
<tr>
<td>USP</td>
<td>United States Pharmacopoeia</td>
</tr>
<tr>
<td>UV</td>
<td>ultraviolet</td>
</tr>
<tr>
<td>VIP</td>
<td>variable importance in projection</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
</tbody>
</table>
Acknowledgements

I would like to first thank my supervisor, Dr. Paul Shipley for giving me the opportunity to work on this project, explore the field of NMR spectroscopy and natural health products, and for his enduring support and mentorship over the past two-and-a-half years. I would also like to extend my deep gratitude to my supervisory committee members, Dr. Paula Brown, Dr. Susan Murch, and Dr. Soheil Mahmoud for their outstanding direction, feedback, and advice throughout the duration of my project.

A huge thanks is owed to Dr. Paula Brown’s Natural Health and Food Products Research Group at BCIT—especially Elizabeth Mudge and Hazrah Moothoo—for continual support throughout this project and for helping me to develop skills necessary to optimize plant extractions and analyze my samples by HPLC. I’d also like to thank the entire Murch lab at UBC Okanagan for providing me with a functional workspace and for being a genuinely fun group to spend time with.

I’d like to thank those who’ve supported the hawthorn project, particularly Dr. Tim A. Dickinson of the Royal Ontario Museum Green Plant Herbarium and his research group for specimen identification, and Jeanette Lee of HerbPro for maintaining a beautiful hawthorn trial farm.

My wallet and I owe a enormous thank you to the agencies that have helped fund my education throughout my project: UBC Okanagan, BC Proteomics Network, the Natural
Health Products Research Society of Canada, University of Ottawa, the Canadian Institutes of Health Research, and the Waters Corporation.

I’d also like to thank UBC Okanagan for providing me with an excellent undergraduate and graduate education.

I’d like to extend a thank you to those fellow grad students and office mates who’ve tolerated my sense of humor over the past years and provided much insight: Dr. Eric Dennis, Doug Whitelaw, Dr. Carla Meints, Dr. Caitlyn Makins, Ryan Moss, Dr. Andrew Veale, Osei Fordwour, Matthew Glover, Yann André, and Dr. Adéline Delcambre.

I thank my family and friends for moral support throughout my education.

A special thanks is owed to my wife and lab mate, Dr. Ying Liu, who not only provided me with continual emotional support, but also assisted me with numerous aspects of my project, and endured many, many disgusting jokes.
In dedication to the memory of my late grandfather, Garry Lund.
Chapter 1: Introduction

1.1 Botany and ecology of *Crataegus*

*Crataegus*, more commonly known as hawthorn, is a diverse genus of flowering, fruit-bearing shrubs or small trees that are native to northern temperate zones, including areas of North America, Europe, and Asia (Edwards et al., 2012). The genus *Crataegus* is a member of the family Rosaceae, which includes 95 genera and many economically important crops, such as apples, cherries, raspberries, and roses (Watson and Dallwitz, 1992). Another genus belonging to Rosaceae, *Mespilus*, also known as medlars, is a small genus of only two known species that are very closely related to *Crataegus* — the two genera so closely resemble each other that they were often misidentified with each other into the 19th century (Phipps et al., 2003). Recent nuclear and chloroplast phylogeny data suggest that the critically endangered North American medlar, *Mespilus canescens*, may actually be a hybrid originating from another North American species of *Crataegus* (Lo et al., 2007).

Species of *Crataegus* are characterized as shrubs or small trees that typically grow to a height of 2-10 meters (Ozcan et al., 2005) but in extreme cases, may be as small as 1 meter or as tall as 15 meters (Phipps et al., 2003). Hawthorn trunks and major branches are often thorned, which has led to their utility as hedges to demarcate property lines, or prevent crossing of cattle and wildlife. Leaves of *Crataegus* are typically elliptic to ovate, but may manifest into triangular shapes. The leaf margins are always toothed, and the teeth may be either long and sharp or smaller and blunted. Some species of *Crataegus* possess deeply lobed leaves. The leaf lobes are indented into multiple pinnate (central vein with secondary veins stemming from the mid-vein) lobes. Even though the venation pattern of *Crataegus* leaves are for the most part pinnate, they are useful in identification, since some species
possess craspedodromous venation (secondary veins directly reach leaf margin), and others possess camptodromous venation (secondary veins do not directly reach leaf margin). One complication of leaf-based identification of *Crataegus* spp. is a result of variation in leaf shapes in individual trees, however this variation is much more pronounced in extension-shoot leaves, therefore one should use short-shoot leaves for identification (Phipps et al., 2003).

Flowers of *Crataegus* are arranged into inflorescences (clusters of flowers) typically comprised of 5-30 flowers. The flowers are approximately 10-25 mm in diameter and petals may be coloured white, pink, or red. Individual hawthorn flowers have five rounded petals, five sepals, stamens in multiples of five, and 1-5 styles (Phipps et al., 2003). Following pollination, hawthorns begin to develop fruit that, when fully developed, are a variety of colours depending upon the species. Fruits of hawthorn, sometimes termed ‘haw’, may be yellow, orange, pink, red, purple, or black. They are typically 6-16 mm in diameter, spherical to ellipsoid, and may be smooth or hairy. The apex of the fruit has evident remains of flower components including the calyx, sepals, stamens, and styles. Although fruits of *Crataegus* are small and berry-like, they are technically a pome and contain between 1 to 5 very hard, seeded pyrenes (Phipps et al., 2003; Spellenberg et al., 2014).

The reproductive biology of *Crataegus* is quite complex and has, in part, led to a great deal of diversity in the genus. Diploid hawthorns may partake in sexual breeding—either from self or other individuals. Many other hawthorns such as some individuals of *C. suksdorfii* and *C. douglasii* are polyploids and are able to reproduce both sexually and via apomixis (asexual formation of seeds). Apomixis in hawthorns is often observed as gametophytic apomixis in which the egg parthenogenetically develops into an embryo.
Interestingly, some hawthorns may need pollination to trigger apomictic reproduction even though the pollen doesn’t fertilize the egg—a phenomenon termed ‘pseudogamy’ (Lo et al., 2009a; Phipps et al., 2003). Numerous genera, including *Crataegus*, that are able to undergo gametophytic apomixis are also almost always associated with polyploidy and hybridization (Lo et al., 2009a) and so, hybridization and presence of agamic complexes in *Crataegus* leads to a large amount of genetic and morphological variation even within a single species, and consequently, the genus is quite confusing for taxonomists (Camp, 1942).

Hawthorns may be found growing naturally as individual trees, groups of trees, or thickets across much of the Northern Hemisphere, except for temperate deserts or high montane or boreal forests. Because hawthorns are often used as ornamentals, there are some species of *Crataegus*, such as *C. monogyna*, that have been naturalized into regions of the Southern Hemisphere (Phipps et al., 2003). Hawthorns exert a very large impact on the ecology of areas in which they are endemic. These trees are home to many different species of small mammals, birds, reptiles, and insects. Hawthorn’s thorny foliage offers protection and excellent nesting sites to its inhabitants, the fruits provide food for birds and small mammals, and the flowers are important to nectar-feeding insects (Phipps et al., 2003).

Despite the positive ecological impacts of hawthorn in many areas, in others, they may be considered a noxious invasive species, such is the case of *C. monogyna*’s invasion of some ecosystems in Oregon and British Columbia (CABI, 2015).

### 1.2 Classical taxonomy and identification of *Crataegus* spp.

Hawthorn has a long history in several cultures worldwide, both in terms of its use and its integration in culture. Hawthorn has been the subject of many legends in Celtic, Gaelic, and English tradition (Culpeper, 1850; Phipps et al., 2003). The plant also has some
religious significance and is viewed as sacred by traditionally Catholic cultures (Phipps et al., 2003). Hawthorns were first described in written botanical literature approximately 2300 years ago as three species in Theophrastus's Historia Plantarum (Thanos, 2011). With a very rich history, it is not surprising that the genus *Crataegus* was rapidly described in the context of classical Linnaean taxonomy. French botanist Joseph Pitton de Tournefort was the pioneer of the modern concept of a genus (Stuessy, 2009) and his description of the genus *Crataegus* was revised and added to Linnaeus’ Genera Plantarum (1754), which defined the genus *Crataegus* comprising nine species at the time (Dönmez, 2004).

Further botanical exploration of North American and Asian areas led to the taxonomic classifications of many new hawthorns in the late 18th and the 19th centuries. Notable botanists involved in *Crataegus* research at the time include Meriwether Lewis and William Clark, David Douglas, and Thomas Nuttall (Dickinson, 2000). Toward the late 19th and the early 20th century, there was a very large increase in descriptions of new North American *Crataegus* spp., which had eventually led to some question of how systematic botanists had missed hundreds of *Crataegus* spp. for so many years (Standish, 1916). Standish (1916) proposed that several pieces of evidence suggest that the species-multiplication in the genus *Crataegus* are not due to discovery of novel species, but in fact represent polyploid hybrids of already-classified *Crataegus* spp. Despite this evidence and that of others (Longley, 1924), taxonomic confusion persisted for decades and led to Camp (1942) to write about the troubled state of *Crataegus* taxonomy in his paper titled “The *Crataegus* problem.”

Because of the presence of agamic complexes and hybridization in *Crataegus*, the concept of speciation in the genus has been difficult and controversial. Although dated
estimates suggest that there may be over a thousand *Crataegus* spp. (Palmer, 1925), current estimates place the number of hawthorn species at approximately 200 members (Edwards et al., 2012). To complicate matters further, traditional methods of species identification applied to *Crataegus* spp. is notoriously difficult and herbarium and arboretum specimens are often misidentified (Phipps et al., 2003).

Traditional methods for identification and discovery of new plant species typically involve careful examination of the specimen’s morphology. With many years of training and use of extensive reference collections of *Crataegus*, a taxonomist will use morphological features of the flowers, fruit, and foliage to identify a hawthorn specimen. Most species of *Crataegus* are divided into three ‘sections’ that have some distinct morphological features. Section *Crataegus* is a group of approximately 30 species including the notable species *C. monogyna*. Members of this section typically have leaves that are deeply dissected (cut deeply into relatively narrow lobes), leaf venation that reaches the sinuses, flowers with 20 stamens, and thorns of variable length. Section *Sanguineae* includes 15 species native to Asia and central Europe. Members of this section have larger leaves that have relatively shallowly lobed leaves, leaf venation that only reaches deep sinuses near the leaf base, flowers with 20 stamens, fruit containing laterally rough nutlets, and short thorns. Section *Coccineae* includes most North American species. Members of this section have relatively shallowly lobed leaves, do not have leaf venation that reaches the sinuses on short-shoot leaves, flowers with 10 or 20 stamens, fruit containing laterally smooth nutlets, and long thorns. In addition to these three major sections, there are numerous other smaller sections. Because the evolutionary relationships between the three large sections and numerous smaller sections are somewhat ambiguous, the genus *Crataegus* is alternatively split into approximately forty
small and relatively homogenous groups termed ‘series’. If in taxonomic classification both series and section are used, then section is considered to be the more inclusive rank (Phipps et al., 2003).

Notable *Crataegus* spp. discussed throughout the remainder of this thesis include *C. monogyna*, *C. laevigata*, *C. douglasii*, *C. suksdorfii*, and *C. okanaganensis*. In Linnaeus’ *Species Plantarum* (1753), *Crataegus oxyacantha* is described for a species of hawthorn growing in Northern Europe and the name was used to describe several similar species including *C. monogyna* and *C. laevigata* until 1974 when Byatt published a paper suggesting that *C. oxyacantha* is a source of confusion and that it should be rejected in a movement to correctly differentiate some of the species grouped under this name. Twelve years later, *C. oxyacantha* was formally rejected by the International Botanical Congress (Brummitt, 1986). Despite this decision, the name *C. oxyacantha* is still in use in some natural health product (NHP) monographs and NHPs (AMR, 2010).

*Crataegus monogyna* Jacq., also known as the one-seeded hawthorn, is native to Eurasian regions and some of Africa, however may be found in other areas of the world, such as North America as an invasive species or ornamental (Figure 1.1) (CABI, 2015; Phipps et al., 2003). Members of *C. monogyna* are characterized by alternate, simple, ovate leaves that are deeply dissected and glabrous (smooth, glossy leaves), with veins reaching the sinuses. The flowers may be white or red depending on maturity, with 15 or 20 stamens, and pinkish-purple anthers. As the common name suggests, *C. monogyna* may be characterized by a small, reddish, ellipsoid to globose pome containing a single nutlets (Phipps et al., 2003; Spellenberg et al., 2014). As a result of its value as an ornamental crop for landscaping
purposes *C. monogyna* has also been used in selection of numerous cultivars (Phipps et al., 2003; Spellenberg et al., 2014).

![Distribution of Crataegus monogyna around the world.](image)

**Figure 1.1: Distribution of *Crataegus monogyna* around the world.** Biodiversity occurrence data according to the Global Biodiversity Information Facility (GBIF, 2015).

*Crataegus laevigata* (Poir.) DC., also known as the midland or woodland hawthorn, is native to European regions and some of North Africa, however unlike *C. monogyna*, *C. laevigata* has not spread extensively as an invasive species (Figure 1.2) (Phipps et al., 2003; USDA and ARS, 2015). Members of *C. laevigata* are characterized by alternate, ovate leaves that are very shallowly dissected into blunt lobes, and glabrous except on veins that reach sinuses. The flowers may be white or pale pink, with 20 stamens, and purple anthers. The fruit is a small, reddish, ellipsoid to globose pome containing two nutlets, which along with leaf morphology differences, may be used in differentiating *C. laevigata* from *C. monogyna* (Phipps et al., 2003; Spellenberg et al., 2014).
Figure 1.2 Distribution of *Crataegus laevigata* around the world. Biodiversity occurrence data according to the Global Biodiversity Information Facility (GBIF, 2015).

*Crataegus douglasii* Lindl., also known by the common name black hawthorn, is native to a relatively large area of the Pacific Northwest of the United States and southern British Columbia, however may be found sporadically in the Northeastern Great Lakes area (Figure 1.3) (Phipps et al., 2003; Spellenberg et al., 2014). Members of *C. douglasii* are short bushes (in comparison to *C. monogyna* or *C. laevigata*) characterized by alternate, simple, elliptic to obovate leaves that are very shallowly lobed with prominent sharply-toothed leaf margins. The flowers are white with 10 stamens and pink anthers. As *C. douglasii*’s common name suggests, the ellipsoid fruit are dark purple to black at maturity and typically contain 3-4 nutlets (Phipps et al., 2003; Spellenberg et al., 2014).
Figure 1.3 Distribution of *Crataegus douglasii* around North America. Biodiversity occurrence data according to the Global Biodiversity Information Facility (GBIF, 2015).

*Crataegus suksdorfii* (Sarg.) Kruschke, also known as Suksdorf’s hawthorn, is native much of the Pacific Northwest of the United States (Figure 1.4) and is highly related to *C. douglasii* and at one point was known as *Crataegus douglasii* Lindl. var. *suksdorfii* Sarg. (Giblin, 2015). Although the overall morphology of *C. suksdorfii* is similar to *C. douglasii*, there are some notable differences that led to its classification as a unique species. *C. suksdorfii* has flowers with 20 stamens, rather than the 10 observed in specimens of *C. douglasii* (Brunsfeld and Johnson, 1990). *C. suksdorfii* is typically diploid and self-sterile and *C. douglasii* is tetraploid self-compatible. However, examples of polyploidy in *C. suksdorfii* do exist in specimens with triploid or tetraploid cytotypes (Lo et al., 2009a). The flowering times and fruit development timelines also differ between the two species (Brunsfeld and Johnson, 1990).
Crataegus okanaganensis (J.B. Phipps & O’Kennon), also known as the Okanagan hawthorn, is a recently discovered species with the first specimen discovered in 1994 near the UBC Okanagan campus along Highway 97 (Phipps and O’Kennon, 1998). The species has a narrow distribution ranging from British Columbia to Washington (Figure 1.5), and also sporadic specimens present in northwest Montana (Phipps et al., 2003). Members of C. okanaganensis are relatively large shrubs that grow to approximately 8 meters in height. The leaves are medium size, ovate to rhombic, and may either be broad with sharp lobes or narrow with very shallow lobes. The flowers are white or pale pink with 10 stamens. In August, the fruit is red with 2-3 nutlets, however the fruit continues to deepen in colour to a deep purple when fully mature (Phipps et al., 2003). This species remained undescribed into the late 20th century due to its high morphological similarity to other specimens of the region, particularly Crataegus williamsii (Phipps and O’Kennon, 1998). C. okanaganensis has glossier foliage with a more leathery texture and broader lobes than that of C. williamsii. The
pattern of seasonal fruit colour change also varies slightly between the two species (Phipps and O'Kennon, 1998).

Figure 1.5 Distribution of *Crataegus okanaganensis* around North America. Biodiversity occurrence data according to the Global Biodiversity Information Facility (GBIF, 2015).

Throughout the history of taxonomy and identification of hawthorns, there has been a considerable amount of ambiguity in determining relationships between various *Crataegus* spp. and determination of species identification by simple morphological and cytological studies. Recent developments in techniques such as molecular phylogenetics, DNA barcoding, and chemotaxonomy have been applied in an attempt to solve some of these issues.
1.3 Molecular and chemical techniques for taxonomy and species identification of *Crataegus*

Molecular phylogenetics was born from theoretical frameworks developed in the 1960s (Suarez-Diaz and Anaya-Munoz, 2008) and evolved from early applications of protein electrophoresis to determination of evolutionary relationships (Ahlquist, 1999). The development of DNA–DNA hybridization by Charles Sibley and Jon Ahlquist which allowed for comparisons of genotypes of two organisms was a substantial development in furthering the field of molecular phylogenetics (Ahlquist, 1999). With the advent of Sanger sequencing, molecular phylogenetics has replaced DNA-DNA hybridization with DNA sequencing which allows an analyst to identify similarities or differences between organisms based upon the degree of genetic similarity between their DNA sequences (Sanger and Coulson, 1975; Sanger et al., 1977; Voet and Voet, 2010).

Entire genome sequencing is time-consuming and expensive, so sequencing of a specific areas, or loci, of a chromosome of a given organism is often performed. This requires much fewer numbers of base pairs to be sequenced and the characteristic pattern in a given loci corresponds to an organism’s haplotype. Variations in the base pair sequence give rise to new haplotypes. These haplotypes are compared and contrasted in molecular systematic analyses to determine evolutionary relationships (Buntjer et al., 2005).

Molecular phylogenetics has been applied to numerous studies of *Crataegus* taxonomy. The earliest studied nuclear ribosomal DNA (nrDNA) internal transcribed spacer (ITS) sequences to attempt to elucidate relationships between members of *Crataegus* (Dickinson et al., 2000). One study using four chloroplast and five nuclear regions demonstrated some evidence that there is close genetic associations between section
Sanguineae, of East Asian origin, and section Douglasianae, of Western North American origin, leading to suggested ancient trans-Beringian migrations of Crataegus spp. (Lo et al., 2009b). There is also molecular evidence that European hawthorns migrated into Asia, evidenced by phylogenetic association between section Crataegus of Europe and the Asian species C. hupehensis, C. songarica, and C. pinnatifida (Lo et al., 2009b). Another molecular study examined pollen flow within populations of two similar hawthorns that differ in ploidy level and breeding system, Crataegus crusgalli (tetraploid) and Crataegus punctata (sexual diploid). Through microsatellite analysis of five loci, the authors determined that pollen flow (the extent of geographical distance of which an individual disperses pollen) in C. crusgalli is relatively limited and that the species has lower genotypic diversity than C. punctata (Lo et al., 2010b).

Another study used microsatellite and chloroplast sequence markers to determine population structure and genetic variability in the two closely related North American black-fruited hawthorns, C. douglasii and C. suksdorfii (Lo et al., 2009a). Within-population genetic diversity was found to be greatest in the diploid sexual C. suksdorfii individuals but lowest in triploid apomicts of the same species. Genetic evidence in C. douglasii suggests that there is frequent gene flow between populations of C. douglasii despite large geographical distances between populations. In contrast, C. suksdorfii gene flow appears to be limited by distance between populations and particularly by apomixis in polyploid individuals of C. suksdorfii. This poses a large barrier for different populations of C. suksdorfii to transfer genetic information between each other. For these reasons, the authors speculate that future allopatric speciation may divide C. suksdorfii into separate taxa (Lo et al., 2009a).
A subsequent investigation of *C. suksdorfii* and *C. douglasii* employed multiple approaches in elucidation of reticulate evolution in these species and to examine how reticulate evolution and allopatric speciation contribute to speciation of some polyploid North American hawthorns. Markers from nuclear and chloroplast sequences were used in generation of phylogenetic trees and networks compared in the study and demonstrated that both approaches are complementary in inferring evolutionary history of the *Crataegus* polyploids investigated (Lo et al., 2010a). The results demonstrated that there are four possible routes of polyploid formation and reticulate evolution in *C. suksdorfii* and *C. douglasii*: first, autotriploidy may give rise to 3x individuals of *C. suksdorfii*; second, allotriploidy, resulting from fusion between reduced gametes of 2x *C. suksdorfii* and 4x *C. douglasii*, may result in 3x individuals of *C. suksdorfii*; third, backcrossing between unreduced gametes from 3x *C. suksdorfii* and reduced gametes from 2x *C. suksdorfii* fuse to produce stable tetraploid members of *C. suksdorfii* (also known as a “triploid bridge”); and fourth, in a site with an overlapping niche between 4x *C. douglasii* and 4x *C. suksdorfii*, interbreeding between the tetraploids of the two species may occur (Lo et al., 2010a).

Another investigation of the North American black-fruited hawthorns, *C. douglasii* and *C. suksdorfii* demonstrated geographical parthenogenesis in the two species. Flow cytometry and nuclear DNA was used in the study to determine the ploidy level and breeding system of 282 black-fruited hawthorns distributed across sites in the Pacific Northwest, Cypress Hills, and upper Great Lakes basin. The genetic characterization was combined with environmental data from the regions in which the trees were grown, and the authors observed that the tetraploid *C. douglasii* inhabits a wider range of geographical areas and tolerates
more extreme temperatures and soil moisture regimes than the diploid *C. suksdorfii* (Lo et al., 2012).

Prompted by failure to authenticate hawthorns by DNA barcoding using the nuclear ribosomal internal transcribed spacer 2 (ITS2) (explored in detail in Chapter 1.7.2), the most recent study of evolution in North American black-fruited hawthorns established further indication of reticulate evolution by using evidence from ITS2 and plastid sequences (Zarrei et al., 2014). Furthermore, the authors explored reasons why ITS2 was of limited use in hawthorn authentication. The authors determined that in the phylogenetic tree generated from the data matrix of ITS2 sequences, ribotypes of the western North American hawthorns studied were divided into two clades—one comprised of diploid members of section *Douglasia* (with partial contribution from section *Sanguineae*) and the other comprised of diploids and polyploids belonging to section *Coccineae*. Both clades were determined to contribute to the ribotypes observed in polyploid members of section *Douglasia* and therefore provide evidence for repeated hybridization between the two clades and explained previous accounts of phenetic intermediacy in some members of section *Douglasia* (Zarrei et al., 2014).

Using similar foundational techniques to molecular phylogenetics, the aim of DNA barcoding is to identify an unknown organism using a specific DNA locus to match it to a database of identified organism’s sequences of the same locus by measuring genetic similarity between the compared loci. This differs from molecular phylogenetics in that the goal is not to determine evolutionary relationships, but rather to identify an unknown organism as a member of a known species (Kress et al., 2005). DNA barcoding is further discussed in the context of *Crataegus* in Chapter 1.7.2. Since both morphological and
molecular studies in _Crataegus_ often do not supply information necessary to unambiguously
make inferences about phylogenetic relationships and species identity, chemotaxonomy may
be employed as a tool to aid these activities.

**1.3.1 Chemotaxonomy**

Chemotaxonomy, from the terms ‘chemical’ and ‘taxonomy,’ is the study of the
classification of organisms, typically plants, by utilizing chemical components to observe and
make inferences about evolutionary relationships. Based on a broad definition of
chemotaxonomy, this method or field of study has ancient origins, although very crude by
modern standards (Stace, 1992). Herbalists and early pharmacists have been documenting
information associating that morphologically similar plants sometimes have similar effects
when administered as a therapeutic agent (indirectly associating chemical content of the
plant) for hundreds of years (Stace, 1992). The advent of technologies such as
electrophoresis and chromatography have allowed for rapid growth of chemotaxonomy in
recent decades. An early example would be the use of electrophoresis in the 1960s by John
Vaughan and his research team to clarify taxonomic relationships in _Brassica_ (Vaughan and
Denford, 1968; Vaughan and Waite, 1967a, b). Many other early studies of chemotaxonomy
employed various other chromatographic methods (Cooperdriver and Swain, 1977; Floyd,
1979).

**1.3.2 Metabolomics and its relationship to chemotaxonomy**

Systems biology, alternatively ‘-omics’ technologies, are at the forefront of chemical
and biological sciences—each targeted towards study of a specific field in a biological
system, such as proteomics or genomics. An organism will typically have many thousands of chemical compounds within its metabolome—the entirety of an organism’s metabolites aside from biomacromolecules like DNA, RNA, and proteins. Metabolomics is the study of biochemical processes within an organism concerning its metabolites (Fiehn, 2002). This can be targeted in some way to observe one or several compounds of interest or untargeted to observe the entire metabolome (Wang and Chen, 2013). Metabolomics has been revolutionary to the expansion of chemotaxonomic research. Through applications of advanced analysis techniques such as mass spectrometry and nuclear magnetic resonance spectroscopy, large amounts of detailed chemical information about a sample can be acquired, and this data can be used with appropriate statistical analysis in chemotaxonomic studies to determine relationships among species (Reynolds, 1996).

Metabolites included within the metabolome involve primary and specialized metabolites, hormones, intermediate chemical compounds, and many other small-molecule chemical compounds. The use of metabolites as markers of taxonomic relationships is quite intuitive. The genome of an organism provides a reference template for the transcriptome, which is expressed as the proteome. The proteins that make up the proteome lead to production of the metabolome (Oliver et al., 1998). Because the genome gives rise to metabolites within an organism, it makes sense that the metabolome may able to provide information about the evolutionary relationships between species. In chemotaxonomic studies, it is important to consider which chemical constituents of the plant are relevant to elucidating evolutionary relationships between species (Stace, 1992). Primary metabolites are those vital to growth and development, metabolism, and reproduction of an organism. Primary metabolites are quite ubiquitous—this makes them somewhat non-valuable in
chemotaxonomic studies (Stace, 1992). Specialized metabolites are generally regarded as non-vital to life and include a much more diverse spread of chemical compounds, especially in plants. Some of these metabolites are phenolics, terpenoids, alkaloids, and amines (Stace, 1992). Because specialized metabolites are more diverse and unique within a single species than primary metabolites, they serve as much better chemotaxonomic markers (Stace, 1992). Leaf phenolics have been used in chromatographic studies of hybridity of some *Crataegus* spp. and hybrids have demonstrated chemical intermediacy with respect to their putative parents (Dickinson, 2003; Wells and Phipps, 1989). A more recent study used a combination of 1D- and 2D- NMR, UV, and HPLC- electrospray ionization (ESI)-MS data to investigate *Crataegus* taxonomy. The authors identified and quantified numerous flavonoids in three European hawthorns and determined that 4″-acetylvitexin-2″-O-rhamnoside, isoorientin, orientin, and 8-methoxykaempferol-3-O-glucoside may serve as potential chemotaxonomic markers for differentiation of *C. monogyna* and *C. pentagyna* from *C. laevigata* (Prinz et al., 2007). *Crataegus* has a wide variety of phytochemicals that have been extensively studied and reviewed (Edwards et al., 2012), which may lead to many potential avenues for species differentiation and identification by its metabolome.

### 1.3.3 Localization of specialized metabolites with respect to plant anatomy

To ensure survival, plants must be able to interact with its environment. This includes protective roles such as responding to environmental stressors, plant defense against herbivory, and protection from pathogens. Environmental interactions may also include those that have reproductive roles such as plant-animal mutualism.
Specialized metabolites have a wide range of roles and must be localized to different areas of the plant where the intended function may take place. Flavonoids have a wide range of function and in general, serve to protect plant tissues from ultraviolet radiation and damage from reactive oxygen species, and so are found in most cell types (Agati et al., 2012). These compounds are produced on the cytoplasmic surface of the endoplasmic reticulum and are subsequently translocated to various cell compartments such as the cell wall or vacuoles (Petrussa et al., 2013). High accumulation of flavonoids is most apparent in vacuoles of both leaf surface epidermal cells and glandular trichomes where they are thought to protect against UV-B radiation (Agati et al., 2012). Flavonoids are also among the most important floral pigments, and so, accumulate in significant concentrations in both the vacuoles and cytoplasm of epidermal cells on the surface of flower petals (Markham et al., 2001). Essential oil terpenoids are another very important group of specialized metabolites with important functions in pollinator attraction and repelling herbivores. These compounds are produced in many plant parts but are especially concentrated in epidermal cells, glandular trichomes, and secretory cells (Bakkali et al., 2008). In *C. monogyna* and *C. laevigata* glandular trichomes may or may not be present on the leaf surface, however if they are, they tend to form around the leaf margin and along veins (AHP, 1999b). In flowers of the same species, the sepals and hypanthium internal surface are densely covered in trichomes (AHP, 1999a). One notable specialized metabolite of *Crataegus* is trimethylamine, which is responsible for the characteristic fishy-like smell of its flowers (Robertson et al., 1993). The compound’s presence in the flowers functions to attract pollinators that are attracted to carrion-like odours, such as flies (Pammel, 1892).
1.4 Phytochemistry of *Crataegus*

The leaves, fruits, and flowers of *Crataegus* contain a diverse profile of chemical classes including sugars and sugar alcohols, organic and phenolic acids, terpenes, essential oils (including mixtures of terpenoids and phenylpropanoids), and phenylpropanoids (including hydroxycinnamic acids, lignans, and flavonoids). Numerous compounds within these chemical classes are implicated in prevention and treatment of ailments in pharmacological studies (Edwards et al., 2012).

Leaves of *Crataegus* produce sugars via photosynthetic reactions and these sugars are transported to the fruit and accumulate throughout the growing season. Sorbitol is an important phloem-translocated sugar alcohol in produced in many members of Rosaceae, including *Crataegus* (Edwards et al., 2012; Suzuki and Dandekar, 2014) and accumulates in significant concentrations in the fruits where a large portion is converted to fructose by sorbitol dehydrogenase (Aguayo et al., 2013). Consequently, fructose is the most abundant sugar (on average) present in fruits of *Crataegus*, however other sugars and sugar alcohols are present in minor concentrations (Edwards et al., 2012).

Plant materials of *Crataegus* have a variety of organic and phenolic acids that have functions in primary metabolism, cellular protection, and metabolic regulation. The most abundant of these in fruits include citric acid, malic acid, and quinic acid. The fruits also contain ascorbic acid, tartaric acid, and protocatechuic acid, which all have antioxidant activity and are important to cellular protection against reactive oxygen species within the plant (Edwards et al., 2012).

Terpenes are a diverse class of compounds biosynthesized from units of isoprene in the activated forms isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate
(DMAPP). These compounds are derived from either the classical mevalonate pathway or the mevalonate-independent 2-C-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate pathway (MEP/DOXP pathway) (Eisenreich et al., 2004). Numerous terpenes have been identified in *Crataegus*. Two triterpenes, oleanolic acid and ursolic acid have been described in *C. pinnatifida* (Cui et al., 2006; Edwards et al., 2012). In addition to molecular structural functions in the plant, some terpenes are also implicated in humans to have therapeutic potential (Edwards et al., 2012). Another study of terpenes in the leaves of *C. pinnatifida* described eight known monoterpen glycosides and a novel monoterpen glycoside, (3S,5R,6R,7E,9R)-3,6-epoxy-7-megastigmen-5,9-diol-9-O-β-D-glucopyranoside (Edwards et al., 2012; Gao et al., 2010). Recently, one new sesquiterpene, (1α,4αβ,8αα)-1-isopropanol-4a-methyl-8-methylenedecahydronaphthalene, was discovered in seed extracts of *C. pinnatifida* that demonstrated antithrombotic activity *in vitro* comparable to acetylsalicylic acid (Zhou et al., 2014). In another recent report, (6S,7Z,9R)-roseoside, a terpene isolated from *C. pinnatifida*, demonstrated significant antithrombotic activity in a dose dependent manner in an *in vitro* platelet count model and was determined to inhibit thrombus formation in an *in vivo* animal model (Li et al., 2015).

Hydroxycinammic acids are a class of phenylpropanoid compounds derived from cinnamic acid. The class is comprised of many compounds that are quite prevalent throughout the plant kingdom (Edwards et al., 2012; Teixeira et al., 2013). Chlorogenic acid, sinapic acid, coumaric acid, and ferulic acid are all phenylpropanoids that have been identified in *Crataegus* and all have antioxidant activity and are important to cellular protection against reactive oxygen species (ROS) within the plant (Edwards et al., 2012). A recent study reported the discovery a novel hydroxycinnamic acid, threo-2-(4-hydroxy-3,5-
dimethoxyphenyl)-3-(4-hydroxy-3-methoxyphenyl)-3-ethoxypropan-1-ol, in *C. pinnatifida* and tested it for antithrombotic activity *in vitro*, however the authors concluded it had no significant effect (Zhou et al., 2014). Hydroxycinnamic acids are implicated in contributing to therapeutic effects of some NHPs. They are also so abundant in a wide range of plants that humans consume, hydroxycinnamic acids may account for approximately one-third of the phenolic compounds consumed in the human diet (Teixeira et al., 2013). Chlorogenic acid is an ester of caffeic acid and quinic acid, and has garnered interest as due to its inhibition of reactive oxygen species formation, antihypertensive effects resulting from improvements of endothelial function and nitric oxide bioavailability in the arterial vasculature, and inhibition of factors causing DNA damage (Edwards et al., 2012; Zhao et al., 2012).

Lignans, formed by dimerization of cinnamic alcohols, are quite diverse in chemical structure and play a role in plant defense (typically functioning as an antioxidant) and may also regulate plant growth (Peterson et al., 2010). In humans, these compounds, when ingested from plant material, are converted to phytoestrogens and other lignan compounds compounds by intestinal flora (Axelson et al., 1982; Borriello et al., 1985; Heinonen et al., 2001). Lignan and its derived phytoestrogens are thought to have some positive impact on human health in prevention of cardiovascular disorders, osteoporosis, and hormone-related cancers (Adlercreutz, 2007; Korkina, 2007; Peterson et al., 2010). Seven lignan glycosides have been identified in *C. pinnatifida* (Edwards et al., 2012; Gao et al., 2010). Subsequent to that study, eight new dihydrobenzofuran neolignans were recently discovered in *C. pinnatifida*, and were named for the plant as pinnatifidanin C I-VIII. The authors also performed *in vitro* studies on the isolated compounds and concluded that they have potent antioxidant activity (Huang et al., 2013).
Like other phenylpropanoids in plants, flavonoids are formed by the shikimate pathway in which bioconversions of phenylalanine and tyrosine ultimately result in a large number of unique phytochemicals. Flavonoids have a C₆-C₃-C₆ structure with two phenyl rings (A, B rings) flanking a central heterocyclic ring (C ring). The profile of flavonoids produced by plants depends on several factors including genetics, climate, soil type, and geography (Edwards et al., 2012). The class includes over 6500 identified compounds, although thousands more are likely still to be discovered (Ververidis et al., 2007). Although flavonoids are specialized metabolites, they are involved in a number of biochemical roles in a plant including pigmentation, defense from pathogens, insects, and herbivores, response to environmental stresses such as ultraviolet radiation or wounding, antioxidants to counter reactive oxygen species, photoreceptors, and developmental signaling and regulation (Ververidis et al., 2007). They are also responsible for many of the positive impacts on human health from consumption of plants, plant extracts, and plant-based NHPs and have been shown to have numerous pharmacological activities including antioxidant activity, antioncogenic action, cardioprotective effects, antimicrobial effects, and anti-inflammatory action (Ververidis et al., 2007).

Notable flavonoids present in *Crataegus* discussed throughout the remainder of this thesis include naringenin, quercetin, hyperoside, rutin, vitexin, isovitexin, and vitexin-2″-O-rhamnoside (Figure 1.6).

Naringenin is a flavonone that is the precursor of many other flavonoids including quercetin, hyperoside, rutin, vitexin, isovitexin, and vitexin-2″-O-rhamnoside (Edwards et al., 2012). Naringenin has been previously detected in very low concentrations in leaves and flowers of various *Crataegus* spp. (Edwards et al., 2012; Keser et al., 2014). This compound
Figure 1.6 Chemical structures of one hydroxycinnamic acid and eight flavonoids present in various *Crataegus* spp.
has been shown to reduce cholesterol by inhibition of hepatic acyl coenzyme A:cholesterol O-acyltransferase and 3-hydroxy-3-methylglutaryl-coenzyme A reductase (Lee et al., 1999), prevents dyslipidemia, apolipoprotein B overproduction, and hyperinsulinemia in insulin-resistant mice (Mulvihill et al., 2009), and inhibit apolipoprotein B–dependent hepatitis C virus secretion both in vivo and in vitro (Nahmias et al., 2008).

Quercetin and its glycosylated derivatives have been identified in the fruit, flowers, and leaves of a variety of Crataegus spp. (Edwards et al., 2012) and has been more recently quantified in leaves (300.8 µg/g) of C. gracilior (Hernández-Pérez et al., 2014), detected in Crataegus x mordenensis (Aladedunye et al., 2014), and detected in an unidentified Serbian hawthorn (Ţugić et al., 2014). Formed by hydroxylations of naringenin at the 3’ and 3 positions and a dehydrogenation, quercetin may be glycosylated to produce numerous other compounds that have demonstrated therapeutic efficacy (Edwards et al., 2012). However, quercetin dihydrate (aglycone form; precursor to quercetin) alone has little evidence to support in vivo therapeutic effects as it undergoes numerous biotransformations when consumed and therefore is quite difficult to argue therapeutic uses from in vitro data (Barnes et al., 2011). There are, however, several studies that suggest efficacy of quercetin in vivo as an anti-diabetic compound and a recent study elucidates the mechanism of action as induction of insulin secretion by activation of L-type calcium channels of pancreatic β-cells (Bardy et al., 2013). Although the European Food Safety Authority does not recognize quercetin as effective in treatment of prevention of any specific disease, Health Canada does allow indications of the compound as “an antioxidant” or “a capillary/blood vessel protectant” (EFSA, 2011; NNHPD, 2012).
Hyperoside, the 3-O-galactoside of quercetin, is found in plants belonging to the genera *Crataegus* and *Hypericum* (Zou et al., 2004), as well as many other genera (Calzada et al., 1999; Li et al., 2005; van der Watt and Pretorius, 2001). Hyperoside has been quantified in the fruit, flowers, and leaves of many different *Crataegus* spp. with concentrations ranging from less than 0.1 mg/g to approximately 40 mg/g depending primarily on the plant material—leaves tend to have higher amounts, particularly in *C. monogyna* and *C. pinnatifida* (Edwards et al., 2012). Hyperoside is a widely studied phytochemical and numerous therapeutic effects have been attributed to it. Hyperoside demonstrates antioxidant activity (Wen et al., 2015; Zou et al., 2004), which has been credited as one of the mechanisms by which *Crataegus* extracts may protect cells against DNA damage by ROS (Mustapha et al., 2014). It has been demonstrated that hyperoside inhibits vascular inflammation caused by high blood glucose both *in vivo* and *in vitro* (Ku et al., 2014a) as well as systemic inflammation via inhibition of the high-mobility group box 1 signaling pathway (Ku et al., 2015). The compound is also partially responsible for antiviral and antimicrobial activities in *Crataegus* extracts (Orhan et al., 2007), antiplatelet and antithrombotic activity (Ku et al., 2014b), and may suppress inflammatory reactions via suppression of nuclear factor-κB activation and production of tumor necrosis factor, interleukin-6, and nitric oxide (Kim et al., 2011b).

Rutin, the 3-O-rutinoside of quercetin, has been identified in several plants (Lucci and Mazzafera, 2009) and has been quantified in numerous *Crataegus* spp. in trace amounts up to approximately 4 mg/g in fruit, flowers, and leaves (Edwards et al., 2012; Hernández-Pérez et al., 2014; Keser et al., 2014). Overall, the compound tends to accumulate in greater concentrations in flower and fruit, however this varies dramatically among species (Edwards
et al., 2012). Rutin has been shown to strongly inhibit formation of blood clots in animal models of thrombosis by inhibition of a protein disulfide isomerase and the authors of the study indicate the compound’s potential use for prevention and treatment of strokes, cardiomyopathies, deep venous thrombosis, and pulmonary embolism (Jasuja et al., 2012). The compound is also an antioxidant (Metodiewa et al., 1997) and demonstrates anti-inflammatory activity both in vivo and in vitro (Guardia et al., 2001; Jung et al., 2007).

Vitexin, the 8-C-glucoside of apigenin, is fairly ubiquitous in hawthorn and may be found in significant concentrations in fruit, flowers, and leaves (Edwards et al., 2012). The plant material with greatest accumulation of vitexin is highly variable among different Crataegus spp. (Edwards et al., 2012). If the glycosylation occurs at the 6-C position, isovitexin (apigenin-6-C-glucoside) is produced. Vitexin is widely studied for potential health effects, however a recent study of hepatic, gastric, and intestinal first-pass effects of vitexin in a rat model shows that vitexin has very low oral bioavailability (Xue et al., 2014). Results of a recent study suggest that vitexin protects brain against cerebral ischemia/reperfusion injury by regulation by mitogen-activated protein kinase and apoptosis signalling pathways, however vitexin was administered intravenously in the mouse model used (Wang et al., 2015) and therefore is unlikely to reflect effects from human consumption of hawthorn. Another study also displays that vitexin may protect against cardiac hypertrophy via inhibition of calcineurin and calmodulin kinase II signalling pathways (Lu et al., 2013), but is also unlikely to explain any health benefits of hawthorn consumption as vitexin is administered intraperitoneally in the study. Like other flavonoids of Crataegus, vitexin also demonstrates antioxidant activity (Bahorun et al., 1994).
Vitexin-2″-O-rhamnoside is to be one of the main flavonoid glycosides of some *Crataegus* spp. (Wei et al., 2014), and has been quantified in the fruit, flowers, and leaves of many different members of *Crataegus*. Leaves tend to accumulate higher amounts, particularly in *C. monogyna* and *C. pinnatifida* (Edwards et al., 2012). In addition to flower and leaf material of *C. monogyna*, Health Canada also recognizes *C. laevigata* as a source material of vitexin-2″-O-rhamnoside in NHPs (NNHPD, 2015c). In one recent study, vitexin-2″-O-rhamnoside demonstrated significant antithrombotic activity in a dose dependent manner in an *in vitro* platelet count model and was determined to inhibit thrombus formation in an *in vivo* animal model (Li et al., 2015). The compound has also exhibited protection against ROS leading to anti-oxidative and anti-apoptotic activity *in vitro* (Wei et al., 2014), and also contributes to antimicrobial and antiviral activities of *Crataegus* extracts (Orhan et al., 2007).

The late Dr. Rudolf Fritz Weiss suggested that the combination of active constituents in hawthorn plant materials is not simply additive or synergic, but rather a beneficial potentiative interaction—that is, the health benefits of hawthorn consumption are most positive when administered as a whole herbal product, rather than as individual chemical compounds (Weiss, 2001). It is evident that many flavonoids found in *Crataegus* spp. alone are associated with numerous health benefits, however the complete effect of hawthorn consumption is a result of the entire phytochemical profile of hawthorn as a whole botanical preparation.
1.5 Ethnobotany and modern natural health products (NHPs) of *Crataegus*

Hawthorn has been used in the traditional medicine of numerous cultures for many centuries with uses ranging from digestive aids to foods and tools. Not surprisingly, the plant has transitioned into the modern era as a prominent NHP and dietary supplement (DS), most prominently attributed to its benefits to cardiovascular health (Edwards et al., 2012).

1.5.1 Ethnobotanical use of *Crataegus*

*Crataegus* is widely distributed across the planet and hawthorn is also reported in the ethnobotany of many cultures. There are numerous reports of use in traditional European societies, North American First Nations, and in traditional Chinese medicine (TCM).

*C. monogyna*, *C. laevigata*, and other European hawthorns have been widely incorporated in European folklore and religion and have been used for medicine and food for centuries. Cardiovascular benefits of hawthorn supplementation was first reported by Greek herbalist Pedanius Dioscorides (c. 40-90 AD) in his extensive pharmacopeia *De Materia Medica* and later by Swiss physician Paracelsus (1493–1541) (Weihmayr and Ernst, 1996). Other literature reports the use of hawthorn tinctures for cardiovascular ailments in nineteenth century Western medicine (Anschutz, 1900; Phipps et al., 2003). Although the fruit of *C. monogyna* and *C. laevigata* are particularly unpalatable, they have been used as a food source in times of starvation and may be prepared into jellies and jams (Morgenstern, 2015).

Many different First Nations groups of North America have used hawthorn for a variety of uses including medicine, food, and tools. Groups reported in North American ethnobotanical literature to have used hawthorn include the Thompson First Nation,
Cherokee, Meskwaki, Blackfoot, Ojibwa, Potawatomi, Okanagon, Okanagan-Colville, Iroquois, Algonquin, Abenaki, Micmac, and Malecite (Edwards et al., 2012; Moerman, 1998, 2009). Fruit of *C. douglasii* has been used for food and beverages by Thompson First Nations, Okanagon, Okanagan-Colville, and Salish First Nations and the bark, shoots, leaves, and fruit have also been used in medicinal preparations for treatments of gastrointestinal ailments (Moerman, 1998, 2009; Steedman, 1929; Turner and Bell, 1971; Turner et al., 1980). The fruit of *C. chrysocarpa* has been used for food by Blackfoot, Lakota, Ojibwa, and Omaha First Nations (Gilmore, 1913; Hellson, 1974; Reagan, 1928; Rogers and Buechel, 1980) and the fruit and roots have also been used in medicinal preparations for treatments of gastrointestinal ailments by the Blackfoot, Ojibwa, and Potawatomi (Hellsen, 1974; Hoffman, 1891; Smith, 1933). Other North American *Crataegus* spp. have been used for bark infusions to treat heart ailments (Hamel and Chiltoskey, 1975) and the sharp spines of various North American hawthorns have been used by some First Nations groups to fashion dermatological probes, awls, and fish hooks (Edwards et al., 2012; Royer and Dickinson, 2007). Known in traditional Mexican medicine as Tejocote, hawthorn has also been reported to have been used for treatment of respiratory ailments (Arrieta et al., 2010; Edwards et al., 2012).

Hawthorn, in particular *C. pinnatifida*, has a very extensive history of traditional use in China. The plant is first described as an herbal medicine in text by Chinese herbalist Su Jing in his work *Tang BenCao* (659 AD), translated roughly as “*Tang Materia Medica*” named for the dynasty (Tang) in which it was written. Other ancient Chinese texts also describe hawthorn for use in treating gastrointestinal ailments (Lan, 1450; Wu, 1332). The *Compendium of Materia Medica* (*BenCao GangMu*) written by Li Shizhen (1593) is
considered to be the most complete and detailed TCM reference ever written. This work outlines the preparation of general hawthorn extracts by cooking fruit until juice is excreted. The compendium indicates hawthorn for a very wide variety of applications including clearance of bodily toxins, treatment of dysentery, shampoo for healthy hair, treatment of skin ulcers and sores, back pain relief, digestive aid, assistance of pancreatic function, and treatment for symptoms of hernias of the small intestine. The text also describes a cooked preparation of hawthorn fruit juice with sugar to be administered orally for treatment of abnormal lochia (postpartum hemorrhage). Another preparation of nutlets of hawthorn fruit combined with alcohol and other herbs is indicated for childbirth complications (Li, 1593). Despite hawthorn’s history of use in Western herbalism for treatment of cardiac disorders, most ancient Chinese TCM literature on hawthorn do not refer to cardiovascular treatments, however, the compendium specifies that hawthorn reduces blood viscosity and increases blood flow (Li, 1593). Tanghulu, roughly translated as “candied bottle gourd,” is a candied hawthorn food item that one of the most common and recognizable hawthorn products in China and was first introduced during the Song dynasty (960-1279 AD) (Fu-Cha, 1961). To prepare tanghulu, hawthorn fruits are skewered onto bamboo skewers and dipped in molten sugar syrup that is then left to harden—creating a street food that may be eaten simply as a dessert, but also is used as a digestive aid after meals (Fu-Cha, 1961).

1.5.2 Modern use of *Crataegus* in natural health products (NHPs) and dietary supplements (DSs)

Hawthorn is a common botanical source used in production of many modern NHPs and DSs and is widely studied in evidence-based medicine for potential as an adjunctive
treatment for various cardiovascular ailments including congestive heart failure, angina, arrhythmia, and hypertension (Edwards et al., 2012; Fong and Bauman, 2002). The primary species used for cardiovascular health NHPs and DSs are *C. laevigata* and *C. monogyna* and there are a wide variety of commercial hawthorn preparations under many brands, such as Cratamed®, Cratagutt®, Esbericard®, Eurython®, and Oxicant® (Phipps et al., 2003). In addition to these commercial preparations, there are also hawthorn teas, creams, and other supplements (Phipps et al., 2003).

Consumption of hawthorn has shown promising benefits for adjunctive therapy in treatment of chronic heart failure in a recent Cochrane Review (Guo et al., 2008). The systematic review of fourteen clinical trials provided evidence that treatment of hawthorn improves maximal cardiac workload, exercise tolerance, decreases cardiac oxygen consumption, and reduces shortness of breath and fatigue when compared to placebo (Guo et al., 2008). The cardioprotective effects of hawthorn have often been attributed to the phenolic compounds such as flavonoids and oligomeric proanthocyanidins present in leaf and berry material (Chang et al., 2002; Schussler et al., 1995), however many other phytochemicals are likely to contribute to the total effect of hawthorn supplementation (Weiss, 2001). Hawthorn supplementation is generally recognized as safe and adverse effects of hawthorn consumption are considered to be mild and uncommon, including nausea, dizziness, headaches, gastrointestinal discomfort, and cardiac palpitations (Daniele et al., 2006; Guo et al., 2008).

Several recent studies of *Crataegus* extracts have demonstrated antithrombotic activity when administered *in vivo* (Arslan et al., 2015; Li et al., 2015; Song et al., 2012) and when applied to *in vitro* models (Li et al., 2015; Zhou et al., 2014). All of these studies use *C.*
pinnatifida except one in which extracts from *C. monogyna* and *C. davisii* are evaluated (Arslan et al., 2015). Another recent *in vivo* study of healthy volunteers to investigate antithrombotic activity as a result of *C. laevigata* leaf and flower supplementation concluded that there was no significant effect when compared with placebo (Dalli et al., 2011), which points to the significance of choosing an appropriate botanical material to reach a desired effect. Optimization of species is also an important consideration in the goal of reaching a therapeutic effect. A potential approach is to use chemical data from plant extracts to generate reasonable hypotheses about possible therapeutic effects (based on chemical compounds with known activity), then test these hypotheses *in vitro* and *in vivo*.

Because the phytochemical profile of a botanical product varies from species to species and also varies with different locations due to environmental factors, regulation of NHPs and DSs are required to ensure that a desired medicinal effect is achieved.

### 1.6 Regulatory requirements of natural health products (NHPs) and dietary supplements (DSs)

Manufacturers of NHPs and DSs are required by governmental agencies (respective to the market in which they intend to sell) to provide evidence for safety, efficacy, and quality of botanical products used as active constituents of their products. The strength and potency of the NHP or DS is often evaluated and identification of botanical materials is also of concern. The sources of information for minimum specifications of these products are monographs, compiled in various pharmacopoeias that are published by governmental agencies or medical/pharmaceutical organizations. Monographs are specifications for production of an NHP or DS and may describe preparation of a single ingredient or describe
a formulation of multiple medicinal ingredients for a given therapeutic indication. For example, Health Canada’s Natural Health Products Ingredients Database (NHPID) contains a “single ingredient monograph” for *Crataegus monogyna* (NNHPD, 2013b), but also contains another monograph for TCM preparations which may include select members of *Crataegus* when indicated for an ailment in the paradigm of TCM (NNHPD, 2015b).

Health Canada’s Natural and Non-prescription Health Products Directorate (NNHPD) oversees regulation of NHPs and DSs sold in Canadian markets and provide monographs for numerous natural products on the NHPID, however the NNHPD also recognizes other major pharmacopoeias and international standards such as the United States Pharmacopoeia (USP), the British Pharmacopoeia, and the European Pharmacopoeia (NNHPD, 2015d).

1.6.1 Identification of botanical materials for use in natural health products (NHPs) and dietary supplements (DSs)

The NNHPD recommends that manufacturers follow Good Agricultural and Collection Practices (GACP) such as those outlined by the World Health Organization (WHO, 2003) and the American Herbal Products Association-American Herbal Pharmacopoeia (AHPA-AHP). The GACPs provide guidelines for proper cultivation, identification, collection, and handling of medicinal plants.

Botanical identity may be established through a variety of means. The AHPA-AHP GACP guide suggests use of reference guides, comparison to voucher specimens, and use of trained botanists or taxonomists as potential methods of identification of plants (AHPA-AHP, 2006). In addition to these techniques, identity may be established through: microscopic
examination of botanical materials; molecular techniques, such as DNA fingerprinting; or chemical techniques, such as chromatography (NNHPD, 2015d).

1.6.2 Evaluation of strength and potency of botanical materials for use in natural health products (NHPs) and dietary supplements (DSs)

Monographs for botanical materials often include specifications with respect to phytochemical content in the finished product (Edwards et al., 2012). Quantitative tests for particular phytochemicals in a botanical extract allow for the potency of the product to be reported (NNHPD, 2015d). Examples of quantification methods for evaluation of potency in NHPs and DSs include high performance liquid chromatography (HPLC), thin layer chromatography (TLC), and more commonly, spectroscopic methods (Edwards et al., 2012; NNHPD, 2015d). The NNHPD states that the tolerance limits for the quantity of medicinal phytochemical in a NHP/DS must conform to the relevant pharmacopoeial standard referenced or must meet 80% to 120% of the amount stated on the product label (NNHPD, 2015d).

1.6.3 Regulatory requirements for Crataegus natural health products (NHPs) and dietary supplements (DSs)

As of August 2015, there are approximately 650 licensed hawthorn NHPs approved for sale in Canadian markets (NNHPD, 2015a). The NHPI contains three monographs that pertain to use of Crataegus in NHPs covering four species. Two single ingredient monographs are in the database for C. laevigata (Poir.) DC. and C. monogyna Jacq., however they are nearly identical. The monographs specify that these species may be used for
maintenance of cardiovascular health and may be prepared as an NHP from either berry or leaf and flower. Hydroalcoholic berry extracts must be standardized to a minimum of 0.93% oligomeric proanthocyanidins (OPCs) and hydroalcoholic leaf and flower extracts must be standardized to 18.75% OPCs (as epicatechin) and/or 2.2% flavonoids (as hyperoside). Tablets prepared from leaf and flower material must be standardized to 15-20 mg OPCs and 6-7 mg flavonoids (NNHPD, 2013a, b). The Canadian monographs do not specify an analytical platform or procedure to be used for these analyses. The “Traditional Chinese Medicine Ingredients” monograph specifies that *Crataegus cuneata* and *C. pinnatifida* may be used for NHPs based on the paradigm of TCM (NNHPD, 2015b).

The United States Pharmacopeia and The National Formulary (USP-NF) guide lists two monographs for either whole hawthorn leaf and flower (dried tips of flower-bearing branches) or powdered hawthorn leaf and flower. The USP monographs state that *C. monogyna* Jacq. (Lindm.) or *C. laevigata* (Poir.) DC., also known as *C. oxycantha* Linné (Fam. Rosaceae) may be used in production of NHPs (USP, 2012). The USP monographs specify that the NHP is to contain no less than 0.6% C-glycosylated flavones (as vitexin) and no less than 0.45% O-glycosylated flavones (as hyperoside), as calculated from dry botanical materials and determined by a TLC identification test with UV detection or a liquid chromatography platform coupled to a UV detector (USP, 2012).

The AHP also publishes two monographs for either *Crataegus* spp. berry or leaf with flower (AHP, 1999a, b). The AHP monographs allow for use of any species or hybrids of *Crataegus* so long as they meet the chemical requirements set out in the monograph. Leaf and flower material of *Crataegus* must contain no less than 1.5% flavonoids (as hyperoside)
and berries must contain no less than 1.0% procyanidins (as cyanidin chloride), as calculated from dry botanical materials (AHP, 1999a, b).

The British Pharmacopoeia (BP) contains three *Crataegus* monographs reproduced from the European Pharmacopoeia (Ph. Eur.) for hawthorn berries (P. Eur. Monograph 1220), hawthorn leaf and flower (P. Eur. Monograph 1432), and hawthorn leaf and flower dry extract (P. Eur. Monograph 1865) (BP, 2009). The BP/P. Eur. monograph for berries allows for fruits of *C. monogyna* Jacq. (Lindm.), or *C. laevigata* (Poir.) D.C. (*C. oxyacantha* L.) or their hybrids, identified by one or more of the following: macroscopic features; microscopic features; or identification by TLC. The fruit material contains no less than 1.0% procyanidins (as cyanidin chloride), as determined by a non-chromatographic spectrophotometric test. The BP/P. Eur. monograph for leaf and flower allows for hawthorn leaf and flower (dried tips of flower-bearing branches) of *C. monogyna* Jacq. (Lindm.), or *C. laevigata* (Poir.) D.C. (*C. oxyacantha* L.) or their hybrids, or other European *Crataegus* species including *C. pentagyna* Waldst. et Kit. ex Willd., *C. nigra* Waldst. et Kit., *C. azarolus* L., with identification of leaf material by TLC. The leaf material contains no less than 1.5% flavonoids (as hyperoside), determined by a non-chromatographic spectrophotometric test. The hawthorn leaf and flower dry extract monograph allows for preparations of the species outlined in the leaf and flower monograph, however dried aqueous extracts must contain no less than 2.5% flavonoids (as hyperoside) and dried hydroalcoholic extracts must contain no less than 6.0% flavonoids (as hyperoside), determined by a non-chromatographic spectrophotometric test (BP, 2009).

The Pharmacopoeia of the People's Republic of China (PPRC), compiled by the Pharmacopoeia Commission of the Ministry of Health of the People's Republic of China
deals exclusively with the Chinese hawthorns, *C. pinnatifida* Bge. and *C. pinnatifida* var. major N.E. Br. The compendium contains two monographs for hawthorn including one for leaf and flower (Crataegi folium; dried flower-bearing branches) and one for fruit (Crataegi fructus) for use in TCM. The PPRC monographs specify that dry leaf and flower must contain no less than 7.0% flavonoids (as rutin) and no less than 0.050% hyperoside, determined by TLC with UV detection. The monograph also specifies that dry fruit must contain no less than 5.0% organic acids (as citric acid), determined by titration (PCPRC, 2010).

### 1.7 Methods for evaluation of potency and authentication of *Crataegus* natural health products (NHPs) and dietary supplements (DSs)

The pharmacopoeias outlined throughout Chapter 1.6 specify requirements of botanical identities and evaluation of potency for *Crataegus* NHPs. As discussed in Chapter 1.5.2, the botanical identity can have an effect on therapeutic outcomes, since phytochemistry is likely to differ between species. Because the chemical content of botanical materials may be inherently inconsistent due to factors outside of speciation (e.g. climate and geography), the potency is often also evaluated for phytochemicals thought to have a therapeutic impact.

The NHP industry has come under a large amount of scrutiny in recent months over an investigation of authenticity of numerous NHPs sold in US markets. Spearheaded by New York Attorney General (NY AG) Eric T. Schneiderman, a commission was launched ordering DNA barcoding testing for botanical DSs sold in major US retailers including GNC, Walmart, Walgreens and Target. The results of this commission found that there were unlisted ingredients or no herbal DNA present in 79% of products tested. This resulted in
extensive negative media coverage towards the NHP industry and ordering these retailers to halt sales of “affected” supplements (Smith, 2015).

Aside from numerous problems with the authentication method used by the NY AG’s commission (to be discussed in more detail in Chapter 1.7.2), the fiasco started by the NY AG brought concerns of NHP authentication and quality control to the public eye. There is now imminent concern for development of reliable methods for evaluation of NHP potency and authenticity.

1.7.1 Evaluation of potency of *Crataegus* botanical materials

Evaluation of NHP/DS potency is often specified in monographs for botanical materials. The USP monograph, for example, requires that hyperoside and vitexin are detected by either TLC or HPLC (USP, 2012). TLC analyses use sizes, fluorescence behaviour, and relative position of spots for identification and quantification of compounds. The method of analysis has seen a large amount of use in routine purity testing due to the low cost of analysis and ease of use (Bart, 2005; Sherma and Fried, 2004). However, TLC is plagued by several limitations, including overlapping spots (co-elution) due to restricted separation efficiency and low number of theoretical plates, inability to distinguish between some isomers, and poor detection limits. Overall, TLC is technologically inferior to HPLC in virtually all possible applications (Bart, 2005; Sherma and Fried, 2004). For these reasons, HPLC analyses are considered to be the gold standard, and so the HPLC method has been increasing in use (Edwards et al., 2012; Sherma and Fried, 2004).

The USP monographs for hawthorn outline an HPLC method that separates hyperoside and vitexin in a hawthorn sample matrix via reverse phase chromatography
(RPC) on a C\textsubscript{18} column with UV detection, and subsequent quantification by a standard curve from an external reference standard (Edwards et al., 2012; USP, 2012). These methods do not, however, provide reliable information for authentication of Crataegus botanical materials (Edwards et al., 2012).

### 1.7.2 Authentication of Crataegus botanical materials by morphology and molecular methods

As reviewed in Chapter 1.6.1, examination of morphological characteristics are acceptable from a regulatory standpoint for authentication of botanical materials, however as discussed in Chapter 1.1 and 1.2, many *Crataegus* spp. are notoriously difficult to identify by morphological characteristics. AHP publishes a reference for microscopic characterization of botanical materials including *C. monogyna* and *C. laevigata* (Upton et al., 2011).

DNA barcoding allows for identification of plants by comparison of DNA sequence similarity to that of identified organisms in a database (Zarrei et al., 2015). The Consortium for the Barcode of Life (CBOL) oversees administration of the Barcode of Life Database (BOLD) which is a public reference library of DNA barcode data that allows researchers to identify unknown specimens as belonging to a known species (CBOL, 2015). CBOL specifies two plastid loci, matK and rbcL, as the barcode regions to be used in generation of a database of identified DNA sequences for vascular plants (CBOL et al., 2009). Additional loci may be used to improve resolution at the species level (Zarrei et al., 2015).

Improper use and limitations of DNA barcoding techniques often calls into question results of these studies. Numerous issues with the use of DNA barcoding in the NY AG’s investigation of NHP manufacturers led to the results obtained, and due to negative press and
product recalls, affected retailers and manufacturers have paid the price. A major issue with the use of DNA barcoding for authentication of NHPs is that the method is unsuitable for testing of many finished products—particularly botanical extracts, since the source plant DNA is likely no longer present or present in only minute quantities. DNA barcoding cannot distinguish different parts of a particular plant since the DNA barcode does not differ in different parts. Additional problems with the study used were use of only one method, use of an inexperienced and unqualified laboratory, no method validation, no release of experimental protocols used, and lack of reproduction of results in other laboratories (Smith, 2015).

DNA barcoding is also often criticized for poor resolution at the species level, and *Crataegus* is no exception here. A recent publication used three plastid regions, matK, rbcLa, and psbA-trnH, and the internal transcribed spacer 2 (ITS2) of nrDNA to discriminate *Crataegus* spp. by DNA barcoding, however the technique only differentiated a handful of *Crataegus* spp. that could already be recognized by morphological studies (Zarrei et al., 2015). The authors indicate that hybridization and polyploidy in *Crataegus* make DNA barcoding of limited use for NHP authentication for hawthorns and that the poor utility of DNA barcoding is also mirrored in other plants that have substantial reproductive complexities (Hollingsworth et al., 2011).

1.7.3 Potential alternative techniques for authentication of *Crataegus* botanical materials

Chemical fingerprinting techniques are currently considered to be the most reliable methods for authentication of botanical materials using techniques such as HPLC, TLC, and
capillary electrophoresis (CE) (Smillie and Khan, 2010). An assortment of relevant marker compounds are used in this technique to authenticate a botanical material as belonging to a particular species and simultaneously provides an indication of potency of the material (Smillie and Khan, 2010), however existing quantitative analytical methods as applied to *Crataegus* are insufficient (Edwards et al., 2012).

Unlike with DNA barcoding, chemical studies are reliable for authentication of NHPs, since much of the chemistry is conserved during processing of plant material into a finished product. Additionally, in contrast to DNA barcoding, since chemical composition is variable among different plant parts, we may use chemical-based authentication to distinguish different parts from the same plant. With chemical identification techniques, one measures the very thing that elicits a biological response. This is in contrast to genomics, where if there is no gene expression, then a biologically active compound will not result. Similarly, proteomics does not necessarily guarantee that a biologically active compound will result if a protein is not part of a functional pathway.

Analytical methods for botanical materials must be validated to ensure that the method used is fit to measure the analytes it is intended to measure and/or observe. There are three levels of validation: single laboratory validation (SLV) in which a method is validated for a single laboratory, technician, and equipment; peer-verified studies which validate an analytical method to 2-7 laboratories; and full collaborative studies which validate an analytical method to 8 or more laboratories (Brown, 2015; Brown and Lister, 2014). For an analytical method to be validated it must meet specific criteria regarding precision, accuracy, specificity, robustness, ruggedness, and reproducibility, and must be selective and provide reasonable limits of quantification (LOQ) and limits of detection (LOD) (Smillie and Khan,
Another concern of these studies include availability of reference standard materials in the form of either chemical standards or botanical materials (Brown, 2015; Smillie and Khan, 2010).

Chemometrics is the measurement and use of chemical data to provide information about a sample and how we present and make inferences about such data (Wold, 1995). Chemometric studies are often complex and may tie into several fields of study (e.g. chemistry, statistics, mathematics, computer science, biology) to answer a particular question. These studies frequently examine complicated systems, which are inherently multivariate, and so, multivariate statistical data analysis (MVDA) techniques are often characteristic of chemometric studies (Wold, 1995).

Chemometrics is often applied to solving predictive problems like that of plant material authentication, providing a theoretical basis for chemical fingerprinting. These fingerprinting techniques incorporate metabolomic data sourced from platforms such as mass spectrometry (MS), nuclear magnetic resonance (NMR) spectroscopy, and infrared (IR) spectroscopy (Smillie and Khan, 2010; Turi et al., 2015).

1.7.4 Nuclear magnetic resonance (NMR) spectroscopy

NMR spectroscopy is a powerful analytical technique that takes advantage of the difference in energy of spin-active nuclei (e.g. $^1\text{H}$, $^{13}\text{C}$, $^{31}\text{P}$) placed in a strong magnetic field between aligning with and opposed to the field (Wider, 2000). Advantages of this technique are that it doesn’t require specialized chromatographic techniques (although there are applications of LC-based methods to NMR), plant extracts can be directly analyzed without extensive preparative steps, short run-times (for one-dimensional (1D) $^1\text{H}$ NMR), and it does
not rely upon external chemical calibration standards (Wang and Chen, 2013). A
disadvantage of this method is that it is not as sensitive as MS instruments, and therefore is
appropriate for analysis of compounds in relatively high concentrations (Wang and Chen,
2013). 1D NMR is a rapid technique, however two-dimensional (2D) NMR techniques may
be required to solve complex structures or reduce signal congestion. Some examples of 2D
NMR techniques include \(J\)-resolved spectroscopy (JRES), heteronuclear single quantum
coherence (HSQC), correlational spectroscopy (COSY), and heteronuclear multiple-bond
correlation spectroscopy (HMBC) (Wider, 2000).

1.7.5 Nuclear magnetic resonance spectroscopy-based metabolomics

NMR metabolomics has numerous applications including monitoring treatment and
disease states in biological fluids in human and animal studies. The technique has also seen a
lot of recent use in studies of plant chemistry. The plant metabolome is dependent upon its
genotype, developmental stage, and environment and external factors. This allows an analyst
to identify metabolic patterns or markers that correspond to plant identity and anatomy, stage
of development, and disease or stress states (Kim et al., 2011a). As discussed in Chapter
1.7.4, NMR spectroscopy has numerous advantages as an analytical platform which
translates to many advantages for its utility as a metabolomics tool. Ideally, a metabolomics
data source would provide as much information as possible about a wide range of
metabolites—NMR is an ideal match here since it allows for a comprehensive view of
metabolites in a sample under given conditions (Kim et al., 2011a).

Chemical data may be coupled with use of MVDA techniques such as hierarchical
clustering analysis (HCA), principal component analysis (PCA), or partial least squares-
discriminant analysis (PLS-DA), which allows one to evaluate a full chromatogram (or other metabolomic dataset) or targeted segment of a chromatogram and compare it to models generated from authenticated botanical materials, allowing for authentication without the necessity of identifying all signals in the chemical dataset (Smillie and Khan, 2010; Ward et al., 2007).

1.7.6 Special considerations in nuclear magnetic resonance spectroscopy and its application to multivariate statistical modeling

Since NMR spectroscopy provides a large amount of chemical information with minimal sample preparation (Kim et al., 2011a), it is not surprising that it has become a very important tool in the field of metabolomics. Using NMR spectroscopy, one can take a rapid snapshot of the entire metabolome in a plant sample extract and use that data for both identification and comparison of phytochemicals in a sample. A $^1$H NMR spectrum provides a large amount of detail about a range of phytochemicals in a plant extract sample. By delimiting ranges of the spectrum, an analyst is able to target specific classes or compounds, such as organic acids, sugars, and phenolics. An analyst can then compare signals from spectra of standards of known compounds and use these to identify and quantify signals of interest in the plant sample spectrum. One can also take the whole spectrum or delimited section of the spectrum of a plant sample and apply the dataset to MVDA techniques to provide untargeted or semi-targeted metabolomics data.

A particular concern in NMR studies of biological samples is of exchangeable protons in the sample. In high concentrations, these protons can result in a very intense signal that attenuates desired data from the NMR spectrum and may produce several minor, broad
peaks from exchangeable protons on groups such as amines, amides, and alcohols, leading to baseline distortions. To resolve this issue, there are a variety of solvent suppression techniques that may be applied in the NMR experiment. The most appropriate and effective solvent suppression technique for NMR metabolomics is presaturation utilizing relaxation gradients and echoes (PURGE) because it is highly selective and has a superior ability to suppress signals without causing baseline distortions which would have a confounding effect on a dataset derived from the NMR spectrum (Simpson and Brown, 2005).

1D $^1$H NMR experiments are the simplest NMR acquisition techniques for metabolomics studies, however there is sometimes an issue of signal congestion in spectra due to scalar coupling. The $J$-resolved (JRES) experiment separates chemical shifts and scalar coupling into two dimensions. The result is a spectrum with reduced spectral congestion, which simplifies the spectrum and allows for easier identification and quantification of signals in crowded areas of the NMR spectrum (Viant, 2003). Overlap of signals may result in overrepresentation of some signals and loss of others, which when exported as a dataset for MVDA, may skew results. A one-dimensional spectrum produced from in silico projection of the two-dimensional JRES spectrum is referred to as a skyline proton projection apparent decoupled 1D (pJRES) spectrum. The pJRES spectrum is an alternative manner for preparation of NMR data for use in metabolomics studies which has the advantage of producing a spectrum that resembles a standard 1D $^1$H NMR spectrum. However, the pJRES spectrum contains less signal congestion than a standard $^1$H NMR experiment because only chemical shift is represented, yielding a dramatically simplified spectrum. pJRES may lead to more representative results in some metabolomics studies with better separation in multivariate models (Kim et al., 2011a; Ludwig and Viant, 2010).
technique is not without disadvantages, though—the total experiment time is often much longer than a standard $^1$H NMR experiment; line widths are broadened; phase-twisted lineshapes may be produced; spectra may not be directly used in absolute quantification studies; increased technical variability; and strongly-coupled spin systems, such as those of some sugars, may produce spectral artifacts (Foxall et al., 1993; Ludwig and Viant, 2010; Nicholson et al., 1995; Viant, 2003).

Data scaling techniques allow variables to be more comparable in generation of multivariate models and allow an analyst to modify the variance structure in an experimental data set with the goal of emphasizing biologically-relevant variance in subsequent MVDA applications (van den Berg et al., 2006). The largest spectral features are not necessarily the most informative ones, and so scaling methods are often applied to allow for an analyst to examine as much relevant variation as possible (Keun et al., 2003). Pareto scaling, in which the dataset is mean-centered and each variable is divided by the square root of its standard deviation, is one scaling technique that is appropriate for datasets where there is large signal intensity variance, such as in the case of NMR spectroscopy (van den Berg et al., 2006). The result of this scaling technique is that larger-fold changes become reduced in intensity to a greater degree than small-fold changes, and therefore the large-fold changes are not as dominant as they are in the original spectrum. This reduces the relative importance of larger values in multivariate models, allows small-fold changes to be better represented, and does a good job of preserving the original data structure (Teng, 2013; van den Berg et al., 2006).

1.7.7 Quantification by nuclear magnetic resonance spectroscopy

Although its use as a tool for quantification is limited in comparison with traditional analytical platforms like HPLC, NMR spectroscopy has been gaining popularity as tool for
chemical quantification as evidenced by a roughly exponential increase of associated terms in reported literature (Figure 1.7). The platform has numerous advantages including non-destructive analysis, minimal solvent requirements, high reproducibility and precision (chronologically and across different instruments) and an ability to see a “global snapshot” of all detectable metabolites present in a sample. However, there are numerous disadvantages that have limited its use including lack of sensitivity, potential accuracy issues, and signal congestion (particularly in biological samples; note, this may also be an issue in chromatographic separation) (Bharti and Roy, 2012; Holzgrabe, 2008; Malz, 2008; Malz and Jancke, 2005).

Plant flavonoids and other phenolic compounds have been associated numerous times with beneficial health effects. These compounds are in relatively high concentrations in plant material and so may be detected by NMR spectroscopy. Several considerations must be made to ensure accurate quantification by NMR spectroscopy: a stable internal standard of accurately known concentration should be used; the internal standard must have a well-isolated signal; the signal of interest must also be isolated from other resonances; all proton environments of interest must be fully relaxed prior to subsequent pulses (i.e. pulse delay should exceed 5xT₁ relaxation time for slowest-relaxing resonance); baseline should be free of distortions; and samples should be well-shimmed to ensure that the external magnetic field is homogenous (Bharti and Roy, 2012; Holzgrabe, 2008; Malz, 2008; Malz and Jancke, 2005).
Figure 1.7: Reports of quantitative NMR spectroscopy in scientific literature as indexed by Web of Science on November 27, 2015. Search terms are as follows: (qNMR) OR ("quantitative NMR") OR ("quantification by NMR") OR ("quantitation by NMR").
1.8 Deficiencies in medicinal and phytochemical studies of *Crataegus*

A bibliographic search of various *Crataegus* spp. outlines the arguable underuse of North American *Crataegus* in modern NHPs. To determine if literature on medicinal properties of three common North American hawthorns is available (*C. douglasii*, *C. suksdorfii*, *C. chrysocarpa*), a search of Web of Science was performed in late November 2015. A search of the terms “(crataegus OR hawthorn) AND (douglasii OR suksdorfii OR chrysocarpa) AND (antioxidant OR health OR medicine OR cardio*)” produce no records of academic literature. In contrast, the same search replacing “(douglasii OR suksdorfii OR chrysocarpa)” with a search term “(pinnifitada)” referring to the well-researched Asian hawthorn, *C. pinnifitada*, produces 123 records. Alternatively, the same search replacing “(douglasii OR suksdorfii OR chrysocarpa)” with a search term “(monogyna OR laevigata)” referring to the two oft-used European species, *C. monogyna* and *C. laevigata*, produces 116 records.

Despite an extensive history of ethnobotanical use in First Nations traditional medicine, North American *Crataegus* spp. has not been often referenced to in herbal medicine literature. In fact, one authoritative reference of herbal medicine states that North American hawthorns have no role in medicine and are sometimes used to adulterate other hawthorn NHPs (Weiss, 2001).

Existing NHP/DS monographs also do not provide any information or specifications on the use of North American *Crataegus* spp. Furthermore, these monographs do not differentiate *C. monogyna* and *C. laevigata* in the context of medicinal use (Edwards et al., 2012). The AHP *Botanical Pharmacognosy - Microscopic Characterization of Botanical Medicines* reference guide states that the differentiation of *C. monogyna* and *C. laevigata* is
of little practical importance (Upton et al., 2011), however results from research discussed throughout this thesis suggest that there are chemical differences between the species, and so, it is possible that there may be some differences (although likely minor) in therapeutic effect. This research cannot, however, make inferences about the effects in vivo or in vitro, rather may only be used to generate reasonable hypotheses that may direct future experiments.

Quantitative analytical methods for evaluation of potency and species authentication with application to Crataegus NHPs are currently lacking and are in need of revision. There are many potential therapeutic phytochemicals that are not covered by the existing USP assay and methods published in other monographs (Edwards et al., 2012). It is also evident that further exploration into phytochemical composition and medicinal potential of North American Crataegus is warranted.

1.9 Research hypotheses and objectives

Adulteration in Crataegus is very rare (Wichtl, 2004) with no known specific adulterants (Bone and Mills, 2013). Despite this fact, regulatory agencies demand that NHP specifications establish botanical identity (AHP-AHP, 2006). It is therefore important to have methods to discriminate one species from another. As discussed in Chapter 1.7.5, NMR spectroscopy is a very useful tool in the study of metabolomics in plant materials since it can be used to observe the presence of a very wide range of phytochemicals (Kim et al., 2011a), and so, may be useful in exploratory chemical studies. The hypotheses of the work guiding this thesis are as follows: NMR spectroscopy is a useful tool in building models to discriminate species of Crataegus (Chapter 2); and NMR spectroscopy is a quantitative tool that allows one to examine a wider range of compounds in Crataegus leaf extracts without
external chemical calibration (Chapter 3). These hypotheses were explored through the following specific objectives:

**Chapter 2**

a) To explore differentiation of *Crataegus* spp. by two different NMR acquisition techniques.

b) To perform untargeted exploration of the data and to also explore the influence of chemical classes on model performance in a semi-targeted approach.

c) To investigate machine learning algorithms as a modern technique for chemometrics data analysis.

d) To determine spectral bins that are significant in species differentiation in the exploratory models.

**Chapter 3**

a) To optimize a method of extraction of hawthorn leaf, suitable for application to quantification by NMR.

b) To quantify flavonoids in North American and European *Crataegus* leaf by NMR.

c) To compare quantitative data from NMR to a validated orthogonal method of analysis using HPLC instrumentation.
Chapter 2: An evaluation of nuclear magnetic resonance data acquisition and multivariate data modeling for differentiation of *Crataegus* spp.

2.1 Synopsis

Regulatory requirements demand that specifications for natural health products (NHPs) include identity (based on genus and species level) of the source material used in NHP production (AHPA-AHP, 2006). Hence, it is important that there are methods available to differentiate species. Since very little is known regarding the phytochemistry of North American *Crataegus*, it is also desirable to determine how these hawthorns differ from their European counterparts in terms of potentially bioactive compounds. This would be useful for hypothesis generation in future *in vitro* and *in vivo* evaluation of therapeutic potential. Four specific objectives were examined: (1) to investigate differentiation of *Crataegus* spp. by two different NMR acquisition techniques; (2) to perform untargeted exploration of the data and to also explore the influence of chemical classes on model performance in a semi-targeted approach; (3) to investigate machine learning algorithms as a modern technique for chemometrics data analysis; and (4) to determine spectral bins that are significant in species differentiation in the exploratory models. The results presented here show that the differences between the two acquisition methods are small, however a standard 1D $^1$H experiment has a small overall advantage in ability to discriminate species over spectra acquired from the *J*-resolved (JRES) experiment and also has the advantage of a shorter experimental duration. The results also indicated that we may best differentiate *Crataegus* spp. by NMR-acquired data that correlates to phenolic compounds. Random forest, a machine learning algorithm, was applied to chemometrics data and the results correlated with and outperformed
traditional data analysis methods, such as partial least squares discriminant analysis (PLS-DA). Lastly, lists of important chemical markers for differentiation of species as determined by a combination of significance analysis methods was also generated.

### 2.2 Research objectives

I hypothesize that NMR spectroscopy is a useful tool in building models to discriminate species of *Crataegus*. The current study was designed to address the following objectives in order to investigate the hypothesis: (1) to determine the influence of two different NMR acquisition techniques on discrimination of *Crataegus* spp.; (2) to explore the data as a whole or in a semi-targeted fashion to determine how each chemical class differs among species; (3) to investigate how discriminatory and classification models may be built from the data; and (4) to determine those signals that are most influential in the differentiation of species.

### 2.3 Experimental

#### 2.3.1 Plant materials

Leaf material from twelve individual hawthorn trees—three plants for each of four species were collected by Jeanette Lee on August 22, 2008 from the Naturally Grown Herb and Spice Producers Cooperative (HerbPro) agroforestry demonstration farm in Edgewood, BC, Canada (Table 2.1; Figure 2.1) and species identified by Dr. Timothy A. Dickinson of the Royal Ontario Museum with vouchers stored at Herbarium of the Royal Ontario Museum. Three technical replicates were derived from each biological replicate.
Table 2.1: *Crataegus* spp. and sample IDs used for generation of chemometric models in untargeted experiments throughout chapter 2 and quantification of flavonoids in chapter 3.

<table>
<thead>
<tr>
<th></th>
<th><em>C. monogyna</em></th>
<th></th>
<th><em>C. douglasii</em></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample IDs:</td>
<td>Wkhgc491</td>
<td>Sample IDs:</td>
<td>Wkhgc492</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wkhgc493</td>
<td></td>
<td>Wkhgc488</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wkhgc489</td>
<td></td>
<td>Wkhgc494</td>
<td></td>
</tr>
<tr>
<td><em>C. okanaganensis</em></td>
<td></td>
<td><em>C. laevigata</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample IDs:</td>
<td>Wkhgc490</td>
<td>Sample IDs:</td>
<td>Wkhgc496</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wkhgc495</td>
<td></td>
<td>Wkhgc498</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wkhgc497</td>
<td></td>
<td>Wkhgc499</td>
<td></td>
</tr>
</tbody>
</table>

### 2.3.2 Experimental materials

4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS; 97%) NMR reference standard and 99% deuterated methanol (methanol-d<sub>4</sub>) were purchased from Sigma-Aldrich (St. Louis, MO, USA). A Bellco Cellector™ Tissue Sieve (30 mesh, 520 µm screen) was purchased from PhytoTechnology Laboratories (Shawnee Mission, KS, USA). 3 mL Luer-Lok™ tip syringes, 1.7 mL microtubes, 2.0 mL microtubes, and 0.45 µm polyfluorotetraethylene (PFTE) syringe filters were purchased from Fisher Scientific Company (Ottawa, ON, Canada). Aldrich ColorSpec NMR tubes (7 in. L x 5 mm diam., 0.38 mm wall, 400 MHz) were purchased from Sigma-Aldrich.

### 2.3.3 Sample preparation

250 mg of ground lyophilized hawthorn leaf was added to a 1.7 mL microcentrifuge tube and extracted sequentially with three aliquots (1.0 mL, 0.5 mL, 0.5 mL) of deuterated
Figure 2.1: Site map depicting geographical location of *Crataegus* samples used for generation of chemometric models in untargeted experiments throughout Chapter 2 and quantification of flavonoids in Chapter 3.
methanol (methanol-$d_4$) containing 5 mM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) as an internal standard by 15 seconds of vortex mixing and 5 minutes of sonication (Fisher Scientific FS20 sonicator). After sonication, the mixtures were centrifuged for 5 min at 16,000 x g, and the supernatants collected and combined in a 2.0 mL microcentrifuge tube. This extract was filtered through a 0.45 µm PFTE syringe filter prior to transfer into an NMR tube for analysis.

2.3.4 Data acquisition

A Varian MercuryPlus 400 MHz NMR spectrometer (Varian, Inc., Palo Alto, CA, USA) equipped with an automated triple broadband probe with a pulse field gradient generator ($^1$H at 400.14 MHz) was used to acquire spectral data. Software control of the instrument was performed with Varian VnmrJ version 2.2 for Linux.

1D $^1$H NMR experiments were acquired at 25°C over a spectral width of 4201.7 Hz with an observe pulse of 11.80 µs (90°) using a PURGE water suppression sequence. A relaxation delay of 2.000 s was elapsed prior to subsequent pulse sequence execution. Total acquisition time was 3.899 s yielding 16,000 points. 128 transients were collected from each spectrum.

2D $J$-resolved (JRES) experiments were acquired at 25°C using a two-pulse sequence incorporating a gradient spin echo for removal of strong coupling artifacts and the PURGE pulse sequence as in the $^1$H spectral acquisitions. Data in the directly-detected F2 dimension were obtained with a spectral width of 4201.7 Hz, eight transients, and a relaxation decay of 1.375 s, with an acquisition time of 0.625 s for a dataset size of 2000 points. Data in the
indirectly-detected F1 dimension were obtained with a spectral width of 64 Hz and 128
increments in t1.

2.3.5 Spectral data processing

Spectra were processed using MestReNova version 10.0.1 (Mestrelab Research S.L.,
Santiago de Compostela, Spain) and were manually phased, baseline corrected using a
Whittaker Smoother base point detection and spline fitting algorithm with the filter set to 16
and the smooth factor set to 20,000. The spectra were referenced to the DSS chemical shift at
-0.018 ppm, binned to chemical shift widths of 0.005 ppm, and the DSS reference peak
normalized to 100. A transposed data matrix of chemical shifts and the intensity values of the
superimposed spectra was exported as a comma separated value file and imported to Excel
(Microsoft, Inc., Redmond, WA, USA) for formatting of the exported data matrix prior to
import into multivariate data analysis (MVDA) software. Bins that correlate to the methyl
peak of the internal standard DSS were excluded prior to multivariate analyses (δ 0.15 -
0.18 ppm). Bins that correlated to the presaturated area of spectrum were also removed prior
to multivariate analyses (δ 4.80 – 4.64 ppm).

2D spectra were also processed using MestReNova. The spectra were tilted 45° and
symmetrized around the F1 axis. The F2 dimension was extracted to produce a skyline
proton projection apparent decoupled 1D spectrum (pJRES). The spectra were referenced to
the DSS chemical shift at -0.018 ppm, baseline corrected using a Whittaker Smoother base
point detection and spline fitting algorithm with the filter set to 16 and the smooth factor set
to 20,000, binned to chemical shift widths of 0.005 ppm, and the DSS reference peak
normalized to 100. The pJRES skyline spectra were superimposed upon each other prior to
.data export. pJRES chemical shifts and intensity values of the superimposed spectra were exported as a transposed data matrix comma separated value file and imported to Microsoft Excel for formatting of the exported data matrix prior to import into MVDA software.

2.3.6 Exploratory data analysis

The spectral data were imported into one of two multivariate statistical analysis software suites: Solo version 7.3.1 (Eigenvector Research, Inc., Wenatchee, WA, USA) and MetaboAnalyst 3.0 (Xia et al., 2012; Xia et al., 2009; Xia et al., 2015), developed from the R programming language (R Foundation for Statistical Computing, Vienna, AT) and an assortment of R packages (R Core Team, 2014).

Data preprocessing

The exported data matrix was formatted and imported into MetaboAnalyst or Solo as a comma separated value (CSV) file. Classes were defined for each sample according to its species identification. Data was normalized by Pareto scaling prior to analysis.

Principal component analysis

Principal component analysis (PCA) models were generated using MetaboAnalyst with the prcomp R package and the Rscript chemometrics.R. Pareto scaling was applied prior to analysis. Scores plots and loadings plots were generated for JRES spectral data and 1D 1H PRESAT spectral data and analyzed as whole spectrum (-0.5–10 ppm; except excluded regions as per Chapter 2.2.5), organic acid region (0.8–2.99 ppm), sugar region (3–5.5 ppm; except excluded regions as per Chapter 2.2.5), and phenolics region (6–8 ppm) to observe
separation among *Crataegus* spp. by phytochemical profiles. Ellipsoids surrounding classes in scores plots are calculated from 95% confidence intervals.

*Partial least squares discriminant analysis*

Partial least squares discriminant analysis (PLS-DA) models were generated using MetaboAnalyst with the pls R package and the further classification and cross-validation by the caret R package. Pareto scaling was applied prior to analysis. Scores plots and loadings plots were generated for pJRES spectral data and 1D $^1$H PRESAT spectral data and analyzed as whole spectrum (-0.5–10 ppm), organic acid region (0.8–2.99 ppm), sugar region (3–5.5 ppm), and phenolics region (6–8 ppm) to provide visualization of separation among *Crataegus* spp. by phytochemical profiles. Ellipsoids surrounding classes in scores plots are calculated from 95% confidence intervals. The models were also used in determination of significant metabolites.

*Cross-validation and evaluation of PLS-DA models*

The PLS-DA models generated from the pJRES spectral data and 1D $^1$H PRESAT spectral data were cross-validated by 10-fold cross validation (CV) and PLS-DA classification performance using the top 4 latent variables was evaluated by prediction accuracy of classification during CV, $R^2$ values, and $Q^2$ values.

For additional rigour of model evaluation, a permutation test was performed on all PLS-DA models following CV to ensure that separation of classes as shown in the PLS-DA model is a statistically significant result. The ratio of the between sum of squares and within sum of squares (B/W ratio) was calculated for class assignment predictions in the model and
was plotted for 1000 permutations. p-Values were calculated from the results of the permutation tests.

**Random forest models**

Random forest (RF) models were generated using MetaboAnalyst with the randomForest R package (Liaw and Wiener, 2002). Confusion matrices and out-of-bag (OOB) error values were generated for JRES spectral data and 1D $^1$H PRESAT spectral data and analyzed as whole spectrum (-0.5–10 ppm), organic acid region (0.8–2.99 ppm), sugar region (3–5.5 ppm), and phenolics region (6–8 ppm) to provide model performance measures of classification among *Crataegus* spp. by phytochemical profiles.

### 2.3.7 Metabolite significance analysis

A total of four methods were employed in determination of significant metabolites within total datasets and subsets of the dataset as grouped by organic acid, sugar, or phenolics region. Important bins are determined by variable importance in projection (VIP) of PLS-DA models, significance of microarrays (SAM) analysis, RF algorithms, and receiver operator characteristic (ROC) curves generated from support vector machine (SVM) algorithms. p-Values and false discovery rates (FDR) for the significant bins were calculated by using a Kruskal-Wallis test, as a non-parametric analysis of variance analysis. Where bins can be identified as corresponding to known standards, putative chemical identity is given.
**Variable importance in projection of PLS-DA models**

PLS-DA models were produced as described in Section 3.2.6. A list of VIP scores were generated from the first and second latent variables of PLS-DA models using MetaboAnalyst and VIP values exceeding 1 were considered significant and compared with other metabolite significance methods.

**Significance of microarrays analysis**

The significance threshold value, $\Delta$, was automatically selected to minimize FDR (below 0.01) by the analysis software. A list of the important features as defined by highest observed $d(i)$ values (exceeding $\Delta$ threshold) were generated by multi-class SAM using MetaboAnalyst with the siggenes R package (Schwender, 2012) and was later compared with other metabolite significance methods.

**Random forest analysis**

RF models were produced as described in Section 3.2.6. A list of the top 50 variables of importance were generated from the RF algorithm by measuring the OOB error increase associated with variables during permutations (mean decrease accuracy; MDA) using MetaboAnalyst with the randomForest R package (Liaw and Wiener, 2002). Results from this analysis were subsequently compared with other metabolite significance methods.

**Cross-validation performance in support vector machine (SVM) algorithms**

Monte-Carlo cross validation (MCCV) with multiple iterations are used in CV of models generated from SVM algorithms. MCCV randomly selects a fraction of the data to
form a training set and assigns the remaining points as a test set. This process is repeated multiple times (Dubitzky et al., 2007). MCCV is used to compute ROC curves and to calculate confidence intervals of the computed area under the curve (AUC) values. The ROC curves generated from SVM algorithms are not suitable for comparison of all four classes simultaneously, so the CV performance in SVM learning models were used for identification of biomarkers which separate the most closely related species from each other. This technique is suitable for application to comparison of two classes, therefore important spectral bins in the phenolics region (6—8 ppm) are identified for separation of *C. monogyna* from *C. laevigata* and *C. douglasii* from *C. okanaganensis* and also for binary comparison of European *Crataegus* and North American *Crataegus*.

### 2.4 Results and discussion

#### 2.4.1 Exploratory data analysis

Principal component analysis (PCA) is an unsupervised MVDA technique that allows an analyst to visualize large datasets and explore patterns and variation between classes. By this technique, the dimensionality of large datasets can be dramatically reduced into principal components that represent a certain proportion of variance represented in the original dataset. Models are generated by orthogonal transformation of data matrices derived from binned spectral data into these principal components, which represent eigenvectors and the corresponding eigenvalues of the covariance matrix. The principal components are used to build dimensions of a scores plot, in which the variance between groups is best described by the first principal component, and decreasing amounts of variance are described by higher
principal components. A loadings plot allows for visualization of influence of spectral bins in generation of the principal components.

The PCA scores plots for the first two principal components (PCs) derived from the whole spectrum acquired as either pJRES or 1D $^1$H PRESAT data were compared. While *C. laevigata* is separated in the first two PCs of the PCA model generated using pJRES, the other three species are not. The PCA model generated using 1D $^1$H PRESAT shows greater separation of *C. laevigata* and although it does not entirely separate the other three species, separation of these classes in the PCA model generated using 1D $^1$H PRESAT is more pronounced (Figure 2.2 A, B). For the organic acids portion of the spectrum, the first two PCs of the PCA model generated using pJRES separates *C. laevigata* a relatively large extent from the other groups, however the other three species are not separated. This pattern is very similar for the PCA model generated using 1D $^1$H PRESAT (Figure 2.2 C, D). In examination of the sugars portion of the spectrum, the PCA model generated using pJRES does not differentiate species and the difference within groups is roughly equal to the difference between groups. The PCA model generated from the first two PCs of the using 1D $^1$H PRESAT separates both *C. laevigata* and *C. monogyna* from *C. douglasii*, however these classes are overlapped by the confidence interval of *C. okanaganensis* (Figure 2.2 E, F). The phenolics region of the spectrum was delimited for a PCA model generated from the first two PCs using pJRES, which separated *C. douglasii* from the other three species, however, there is partial overlap in the confidence intervals of the other three species and they are not completely separated in the model.
Figure 2.2: PCA scores plots generated from NMR spectra of *Crataegus* extracts. Plots A, C, E, and G represent models generated from spectra acquired by pJRES. Plots B, D, F, H represent models generated from spectra acquired by $^1$H 1D PRESAT. Plots A, B were generated from the whole spectrum. Plots C, D were generated from the organic acid region of the spectrum. Plots E, F were generated from the sugars region of the spectrum. Plots G, H were generated from the phenolics region of the spectrum. Ellipsoids around classes represent 95% confidence intervals within class.
The first two PCs of the PCA model generated using 1D \(^{1}H\) PRESAT is quite similar to the PCA model generated using pJRES, however \(C.\) \(douglasii\) and \(C.\) \(okanaganensis\) are separated and the confidence intervals of the two European species overlap (Figure 2.2 G, H).

Upon visual inspection of the scores plots, the phenolics region of the spectra provides the largest extent of class separation in both acquisition techniques. Overall separation in PCA models appears to be superior in \(Crataegus\) extracts acquired by 1D \(^{1}H\) PRESAT in comparison with pJRES. Interestingly, \(C.\) \(laevigata\) was relatively well separated from other species in models generated from other regions of the spectrum. A possible reason for this is that the \(C.\) \(laevigata\) trees on the farm trial plot did not fruit easily. Consequently, sugars and organic acids in the leaves may not have been translocated away to the fruit to the same extent as trees from the other three species. Phenolic compounds are specialized metabolites, which unlike sugars and organic acids, can vary quite significantly between different species or even individuals of the same species grown under different environmental conditions (Stace, 1992; Ververidis et al., 2007). In the case of the samples used in this dataset, environmental conditions such as geographical location, watering frequency, and sunlight are held somewhat constant, so it is likely that variation in phenolic profiles as a result of speciation is being observed here. Although pJRES is often considered to be superior to 1D \(^{1}H\) methods for metabolomics data, the acquisition technique does not appear to improve unsupervised class separation in this \(Crataegus\) dataset.
2.4.2 Supervised models

Partial least squares discriminant analysis (PLS-DA) is a supervised MVDA technique that allows for creation of models of classification and also allows for selection of important features. This technique produces visual models analogous to PCA, however in PLS-DA, the principal components are transformed by partial least squares regression in a way that maximizes separation between classes and allows an analyst to better understand which variables are most responsible for separation of classes. Separation in the visual representation of the PLS model does not necessarily indicate model performance, which is later discussed in the context of cross-validation. Supervised PLS-DA models were constructed to observe how metabolic profiles in the extracts drive variance between the species.

The PLS-DA scores plots for the first two latent variables (LVs) derived from the whole spectrum or sections of the spectrum acquired as either pJRES or 1D $^1$H PRESAT data were compared. Further LVs (up to LV4) were also explored in pairwise scores plots to determine if change in separation patterns were visually notable by combinations of other LVs (data not shown).

For the entire spectrum, clustering of both European species and both North American species are observed in models generated from both pJRES and 1D $^1$H PRESAT spectra (Figure 2.3 A, B). Separation is not any larger in two-component comparison of further LVs. For PLS-DA models generated from organic acid regions of the NMR spectrum, C. laevigata is well separated in pJRES acquired data with partial separation of C. monogyna from the two North American species’ confidence intervals. Models from the same region generated from 1D $^1$H PRESAT spectra widen separation of the North American hawthorns,
Figure 2.3: PLS-DA scores plots generated from NMR spectra of *Crataegus* extracts. Plots A, C, E, and G represent models generated from spectra acquired by pJRES. Plots B, D, F, H represent models generated from spectra acquired by $^1$H 1D PRESAT. Plots A, B were generated from the whole spectrum. Plots C, D were generated from the organic acid region of the spectrum. Plots E, F were generated from the sugars region of the spectrum. Plots G, H were generated from the phenolics region of the spectrum. Ellipsoids around classes represent 95% confidence intervals within class.
but clustering of the European hawthorns is observed in the model (Figure 2.3 C, D). The majority of variance is represented in the first two LVs of the organic acid PLS-DA models and separation is no more pronounced in two-component comparison of further LVs.

For the sugars portion of the spectrum, there is overlap of the confidence intervals for all species, however *C. laevigata* is significantly different than the North American species in the model generated from pJRES spectra. Clustering of both European species and both North American species are observed in models generated from 1D ¹H PRESAT spectra (Figure 2.3 E, F). A majority of variance is represented in the first two LVs of the sugars PLS-DA models and separation is no more pronounced in two-component comparison of further LVs. For PLS-DA models generated by the phenolic region of pJRES spectra, *C. okanaganensis* is completely separated from other groups, however the other groups are mostly separated with small confidence interval overlap. *C. douglasii* is significantly different from *C. douglasii* in the model. For models generated by the phenolic region of 1D ¹H PRESAT spectra, *C. okanaganensis* and *C. douglasii* are completely separated from other groups and the two European species cluster with small intersection of their confidence intervals (Figure 3.2 G, H). Although a majority of variance is not represented in the first two LVs of the phenolics PLS-DA models, separation is no more pronounced in two-component comparison of further LVs.

As PLS-DA is a supervised method, it is very eager to please and tends to provide over-optimistic data models, so it is important that models are well-validated. Cross-validation is often used as an estimate of model validity in evaluation of PLS-DA models when there is a lack of available samples for external validation. The purpose of validity assessments is to determine how well the data will generalize to a new and independent set of
data. k-Fold cross validation of PLS-DA models is performed by randomly dividing the original dataset into a number of subsamples, k, one of which is retained as a test set to be applied to testing the model and the k-1 remaining subsamples, are used in training the model. This CV process is repeated excluding each subsample individually to form parallel models and the average difference between actual and predicted responses are calculated for the models to form the predicted residual sum of squares (PRESS). Q^2 is an expression of the PRESS ratio to the total sum of squares (TSS; total variance that a regression model is able to explain) of the response, as calculated by the equations below:

\[ PRESS = \sum (\text{experimental response} - \text{predicted response})^2 \]

\[ Q^2 = 1 - \frac{PRESS}{TSS} \]

Whereas R^2 is a goodness of fit, Q^2 is a goodness of prediction, and provides a measure of model predictability during CV. Q^2 is the de facto standard in evaluation of PLS-DA models during CV (Eriksson et al., 2006).

PLS-DA model performance is calculated for three criteria for all models generated (Figure 2.4)—Q^2, R^2, and accuracy. Here, Q^2 values are used to evaluate model performance—an ideal model will have a Q^2 value approaching 1. Generally, Q^2 values above 0.5 are acceptable however higher values are desirable (Triba et al., 2015). The pJRES-acquired sugars region models performed the worst in CV (Q^2 = 0.04 for 2 LVs; 0.46 for 4 LVs). PLS-DA model performance with phenolics data was excellent for both pJRES-acquired data (Q^2 = 0.82 for 2 LVs; 0.88 for 4 LVs) and 1D \textsuperscript{1}H PRESAT-acquired data (Q^2 =
0.87 for 2 LVs; 0.94 for 4 LVs). For all other models, performance in CV was poor for 2 LVs, except in the model generated from the whole 1D $^1$H PRESAT-acquired spectra ($Q^2 = 0.78$). Overall, phenolics models from 1D $^1$H PRESAT-acquired data are the best-performing in CV.

Permutation tests were performed on the models to ensure that original classification of samples in the original PLS-DA model is significantly better than random classification of samples to arbitrary groups. The class labels for the samples in the original predictive model are randomly permuted. A new model is then calculated with randomized, incorrect class labels. The assumption is that the new model will not have acceptable class prediction as measured by $Q^2$. By repeating the permutation test many times, a $H_0$ distribution of assumedly poor-performing models are created. The original non-permuted model should have a result that is significantly separated from the $H_0$ distribution created by the random permuted models (Westerhuis et al., 2008). This significance is expressed as a p-value and the value is related to model performance in 1000 permutations.

In permutation testing of the classification models, prediction error of the original models were found to be zero and significantly separated in comparison with the permutation test under a $H_0$ distribution of no effect. Therefore, the observed p-value < 0.001 for all PLS-DA models with 1000 permutations each, as measured by predictive ability, $Q^2$. 
Figure 2.4: PLS-DA cross-validation (CV) model performance generated from NMR spectra of *Crataegus* extracts. Plots A, C, E, and G represent models generated from spectra acquired by pJRES. Plots B, D, F, H represent models generated from spectra acquired by $^1$H 1D PRESAT. Plots A, B were generated from the whole spectrum. Plots C, D were generated from the organic acid region of the spectrum. Plots E, F were generated from the sugars region of the spectrum. Plots G, H were generated from the phenolics region of the spectrum. Dark green bars represent model accuracy during CV. Peach bars represent $R^2$ calculated during CV. Light green bars represent $Q^2$ calculated during CV. Performance measures are given for up to four latent variables. Red star denotes optimal $Q^2$ value.
2.4.3 Machine learning

Random forest (RF) algorithms were explored as a relatively new method for supervised learning of models in metabolomics datasets. RF is a subset of decision tree learning, which belongs to an emerging field of computer science known as machine learning. Developed from studies of computational learning theory in artificial intelligence models, machine learning is responsible for algorithms that are used in a wide variety of applications, such as internet search engines, self-driving cars, and speech recognition.

Decision tree learning can be used to generate predictive models for a wide variety of data types, however like PLS-DA, it is prone to overfitting. RF algorithms correct for the issue of overfitting. RF is a supervised machine learning algorithm that is suited to datasets with numerous dimensions, such as with chemometrics data. This algorithm grows a large number of classification trees, each of which is constructed from a random feature selected from a bootstrap sample from the original dataset. In order for the algorithm to classify a given object from an input vector, the vector is applied to each tree in the ensemble and then each tree votes for a class. The majority vote of the ensemble determines object classification.

Unlike PLS-DA or decision tree learning, the RF algorithm does not overfit data, and there is no need for CV, and instead an unbiased estimate of model error is given as out-of-bag (OOB) error. As previously noted, the trees are grown from bootstrap samples, however approximately one-third of samples are left out of the dataset used in generation of classification trees—these samples comprise the OOB data. The OOB data is used as a test set to be applied to the model and the results are summarized in a confusion matrix with an
associated OOB error—a measure of RF classification performance. This process is analogous to cross-validation in PLS models.

1000 trees were grown and cumulative error rates were plotted for RF classification and used in creation of confusion matrices for the RF models (Appendix 3.x). RF models were generated for pJRES and 1D $^1$H PRESAT data from whole spectrum, organic acids region, sugars region, or phenolics region of the spectra. The total error associated with misclassification in the RF models is expressed as a single overall out-of-bag (OOB) error for the model (Table 3.2).

Table 2.2: Summary of random forest (RF) classification performance for four *Crataegus* spp. Classification performance is compared for two NMR acquisition techniques and with four spectral regions. Classification performance is specified as out-of-bag (OOB) error. Species misclassified in confusion matrices are noted with individual within class error (WCE) probabilities.

<table>
<thead>
<tr>
<th></th>
<th>OOB error</th>
<th>Misclassified species, WCE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pJRES</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole spectrum</td>
<td>0.0278</td>
<td><em>C. okanaganensis</em> (0.11)</td>
</tr>
<tr>
<td>Organic acids</td>
<td>0.0556</td>
<td><em>C. okanaganensis</em> (0.11); <em>C. douglasii</em> (0.11)</td>
</tr>
<tr>
<td>Sugars</td>
<td>0.0833</td>
<td><em>C. okanaganensis</em> (0.22); <em>C. monogyna</em> (0.11)</td>
</tr>
<tr>
<td>Phenolics</td>
<td>0</td>
<td>None misclassified</td>
</tr>
<tr>
<td><strong>1D $^1$H PRESAT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole spectrum</td>
<td>0</td>
<td>None misclassified</td>
</tr>
<tr>
<td>Organic acids</td>
<td>0</td>
<td>None misclassified</td>
</tr>
<tr>
<td>Sugars</td>
<td>0</td>
<td>None misclassified</td>
</tr>
<tr>
<td>Phenolics</td>
<td>0</td>
<td>None misclassified</td>
</tr>
</tbody>
</table>

OOB error is zero for all RF models created from 1D $^1$H PRESAT data. For RF models generated from pJRES data, only the phenolics region produced a model with an OOB error of zero. Of the three other spectral regions examined for pJRES, the sugars region has the worst classification accuracy. These results mirrored supervised learning performance.
in PLS-DA models—the phenolics region in both acquisition methods result in good prediction accuracy; 1D $^1$H PRESAT outperformed pJRES; and the sugars region of pJRES data provided the most inferior classification performance.

Overall results of the exploratory data analysis suggest that in both unsupervised learning and supervised learning, 1D $^1$H PRESAT outperformed pJRES NMR acquisition for this *Crataegus* dataset. In particular, the phenolics region allowed for optimal class separation in unsupervised learning and was used to provide optimal models for supervised learning algorithms. It should be noted that the whole spectrum provided acceptable models, however when delimited to sugars or organic acids, significant performance losses were noted in predictive capability. Phenolic compounds are specialized metabolites, and as such, should provide better differentiation between species than more conserved chemical classes, like sugars and organic acids. The overarching question of this thesis is about how one can use chemical data from hawthorn leaf extracts to differentiate at a species resolution. It is then pertinent to explore whether we may use these models to discover new biomarkers. Some of these biomarkers may also provide clues for subsequent generation of hypotheses about differential therapeutic effects among hawthorn species—leading into studies that allow for well-informed choices in selection of hawthorn species for use in NHPs. Following the analysis of the supervised and unsupervised models, significant spectral bins in classification accuracy were identified and summarized.

### 2.4.4 Metabolite significance analysis

As noted by Broadhurst and Kell (2006), it is a large risk in metabolomics studies that persuasive, but spurious metabolic markers may be found, and so measures to avoid making
false discoveries must be made. For this reason, SAM and a univariate significance test (Kruskall-Wallis) were used in parallel with the multivariate (PLS-DA) and machine learning (RF) tests discussed earlier. In order to determine those bins that are significant, PLS-DA VIPs, SAM significant features, and RF significant features will be compared. VIP analysis is one method in which important variables may be identified. VIP scores are generated from a weighted sum of squares of the PLS loadings and correspond to the amount of variation that the bin represents.

SAM analysis is a statistical technique originally developed for identification of differentially expressed genes in large datasets of DNA microarray data (Tusher et al., 2001). Like DNA microarray data, metabolomics datasets are very large and are high-dimensional in nature, so SAM statistical methods have been applied to discovery of significant metabolites in plant extracts (Broadhurst and Kell, 2006; Brown et al., 2012a; Brown et al., 2012b). SAM calculates a test statistic, $d(i)$, expressing the relative difference in metabolite concentration based upon permutation analysis of the binned spectral data, relative to each bin’s standard deviation in repeated measurements. In the permutation analysis, the expected $d(i)$ is compared to the observed $d(i)$ and the significance of difference is determined by an adjustable threshold, $\Delta$. Those bins with a $d(i)$ exceeding $\Delta$ are deemed significant by SAM and they are compared to the distribution estimated by random permutations of class labels. For each threshold $\Delta$, some proportion of the bins in the permutation set will be found to be significant by chance—this is the foundation upon which FDR is calculated for SAM. The associated false discovery rate (FDR) is calculated by in SAM by taking the average of the number of falsely called bins (false-positive calls) and dividing it by the number of bins called significant (Schwender, 2012). This FDR represents the proportion of metabolites
likely to have been identified by chance as significant. Here, FDR was calculated for each significance threshold, \( \Delta \), selected and \( d(i) \) values exceeding the threshold for a FDR \( \leq 0.01 \) were deemed significant and reported. Note that FDR values reported in tables in this chapter are not calculated by SAM, rather the magnitude of bin significance in SAM here is expressed by \( d(i) \) values that have exceeded the selected significance threshold, \( \Delta \), where FDR \( \leq 0.01 \).

A list of variables of importance is also generated using the RF algorithm by measuring the OOB error increase associated with variables during permutations (mean decrease accuracy; MDA). Fuzzy matching algorithms (automated approximate string matching of text or numbers) applied to these lists will determine spectral bins which are deemed important by at least two of the aforementioned models in order to ensure each important bin has had its importance validated by an orthogonal significance analysis technique.

Because of the multiple comparisons problem, a Bonferroni-corrected p-value and FDR is given for each bin, as determined by a Kruskal-Wallis test (non-parametric equivalent to one-way analysis of variance) to further ensure that only statistically significant bins are deemed important. The Bonferroni-corrected p-value is reported along with the FDR in order to have a numerical concept of both significance and type I error rate, respectively, for the data, as analyzed by the Kruskal-Wallis test. These results are tabulated for the entire spectrum (Table 2.3; Table 2.4), organic acids region (Table 2.5; Table 2.6), sugars region (Table 2.7; Table 2.8), and the phenolics region (Table 2.9; Table 2.10). It should be noted that a significant limitation in data analysis of NMR spectra for use in identification of biomarkers is that multiple signals in the spectrum will represent a single compound,
generating a dataset with large covariance. As mentioned in Chapter 2.2.7, it is another limitation that multiple spectral bins will give rise to a single signal in the original spectrum. A third consideration is that protons may have equivalent chemical shifts and therefore may overlap, where a single bin may represent more than one metabolite. From observations of the hawthorn extract raw spectral data, it is quite difficult to parse out identities of sugars and organic acids due to these overlap issues, however the signals that arise from the phenolics region (6-8 ppm) are well-separated and quite interpretable. Putative chemical identification is given for bins correlating to known flavonoids where possible. Absolute concentrations of many compounds corresponding to significant bins discussed here are provided in Chapter 3.

For the entire spectrum, 18 bins (2 identified) from the pJRES-acquired dataset and 4 (none identified) bins from the 1D ¹H PRESAT-acquired dataset were deemed significant. For the organic acids region, 19 bins (none identified) from the pJRES-acquired dataset and 15 (none identified) bins from the 1D ¹H PRESAT-acquired dataset were deemed significant. For the sugars region, 17 bins (none identified) from the pJRES-acquired dataset and 6 (none identified) bins from the 1D ¹H PRESAT-acquired dataset were deemed significant. For the phenolics region, 35 bins (7 identified) from the pJRES-acquired dataset and 24 (5 identified) bins from the 1D ¹H PRESAT-acquired dataset were deemed significant. Identified significant bins in the phenolics region correlated to quercetin, chlorogenic acid, hyperoside, isovitexin, vitexin, vitexin-2″-O-rhamnoside, and rutin.

As evidenced by the multivariate models (Figure 2.2; Figure 2.3), it is more simple in general to distinguish the two European species, *C. monogyna* and *C. laevigata*, from the two North American species, *C. okanaganensis* and *C. douglasii*, than it is to separate the two European species from each other and the two North American species from each other based
Table 2.3: Summary of significant signals for differentiation of four *Crataegus* spp. from entire spectrum for pJRES. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM $d(i)$ values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed. Putative identifications are listed where possible. VT=vitexin.

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>Putative compound ID</th>
<th>VIP score</th>
<th>$d(i)$</th>
<th>MDA</th>
<th>p-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.613999</td>
<td></td>
<td>3.17</td>
<td>44.43</td>
<td>2.11E-03</td>
<td>4.58E-06</td>
<td>2.79E-04</td>
</tr>
<tr>
<td>0.619128</td>
<td></td>
<td>2.77</td>
<td>33.78</td>
<td>2.32E-03</td>
<td>4.12E-06</td>
<td>2.79E-04</td>
</tr>
<tr>
<td>0.885811</td>
<td></td>
<td>1.07</td>
<td>41.12</td>
<td>3.15E-03</td>
<td>1.18E-05</td>
<td>3.21E-04</td>
</tr>
<tr>
<td>1.07044</td>
<td></td>
<td>1.50</td>
<td>34.64</td>
<td>1.69E-03</td>
<td>6.29E-06</td>
<td>2.79E-04</td>
</tr>
<tr>
<td>1.14736</td>
<td></td>
<td>1.09</td>
<td>62.41</td>
<td>1.65E-03</td>
<td>2.02E-06</td>
<td>2.79E-04</td>
</tr>
<tr>
<td>1.15249</td>
<td></td>
<td>1.78</td>
<td>76.52</td>
<td>2.07E-03</td>
<td>1.66E-06</td>
<td>2.79E-04</td>
</tr>
<tr>
<td>1.15762</td>
<td></td>
<td>1.91</td>
<td>282.13</td>
<td>3.21E-03</td>
<td>2.27E-06</td>
<td>2.79E-04</td>
</tr>
<tr>
<td>1.13148</td>
<td></td>
<td>1.08</td>
<td>163.28</td>
<td>2.50E-03</td>
<td>1.42E-05</td>
<td>3.53E-04</td>
</tr>
<tr>
<td>1.33199</td>
<td></td>
<td>2.35</td>
<td>101.65</td>
<td>1.93E-03</td>
<td>1.82E-04</td>
<td>1.48E-03</td>
</tr>
<tr>
<td>1.69099</td>
<td></td>
<td>1.27</td>
<td>35.51</td>
<td>1.91E-03</td>
<td>8.73E-05</td>
<td>9.87E-04</td>
</tr>
<tr>
<td>2.71156</td>
<td></td>
<td>1.43</td>
<td>53.99</td>
<td>2.59E-03</td>
<td>1.54E-06</td>
<td>2.79E-04</td>
</tr>
<tr>
<td>3.24493</td>
<td></td>
<td>1.56</td>
<td>82.87</td>
<td>1.67E-03</td>
<td>1.90E-04</td>
<td>1.53E-03</td>
</tr>
<tr>
<td>4.34243</td>
<td></td>
<td>1.85</td>
<td>138.16</td>
<td>1.76E-03</td>
<td>3.05E-05</td>
<td>5.29E-04</td>
</tr>
<tr>
<td>4.48603</td>
<td></td>
<td>1.80</td>
<td>36.35</td>
<td>1.87E-03</td>
<td>4.68E-06</td>
<td>2.79E-04</td>
</tr>
<tr>
<td>6.59385</td>
<td></td>
<td>2.47</td>
<td>55.75</td>
<td>2.47E-03</td>
<td>3.87E-06</td>
<td>2.79E-04</td>
</tr>
<tr>
<td>7.46569</td>
<td></td>
<td>2.37</td>
<td>38.77</td>
<td>2.01E-03</td>
<td>8.04E-06</td>
<td>2.79E-04</td>
</tr>
<tr>
<td>7.96829</td>
<td>VT</td>
<td>2.89</td>
<td>39.31</td>
<td>1.59E-03</td>
<td>6.23E-06</td>
<td>2.79E-04</td>
</tr>
<tr>
<td>7.97342</td>
<td>VT</td>
<td>3.35</td>
<td>42.60</td>
<td>2.78E-03</td>
<td>4.50E-06</td>
<td>2.79E-04</td>
</tr>
</tbody>
</table>
Table 2.4: Summary of significant signals for differentiation of four *Crataegus* spp. from entire spectrum for $^1$H 1D PRESAT. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM $d(i)$ values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed.

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>VIP score</th>
<th>$d(i)$</th>
<th>MDA</th>
<th>p-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$^1$H 1D PRESAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.94202</td>
<td>2.62</td>
<td>148.25</td>
<td>2.76E-03</td>
<td>1.39E-05</td>
<td>1.13E-04</td>
</tr>
<tr>
<td>3.25212</td>
<td>2.31</td>
<td>136.70</td>
<td>2.35E-03</td>
<td>8.21E-05</td>
<td>3.74E-04</td>
</tr>
<tr>
<td>4.21717</td>
<td>1.26</td>
<td>222.19</td>
<td>2.01E-03</td>
<td>6.81E-07</td>
<td>6.45E-05</td>
</tr>
<tr>
<td>7.64232</td>
<td>2.41</td>
<td>198.21</td>
<td>2.03E-03</td>
<td>1.54E-06</td>
<td>6.45E-05</td>
</tr>
</tbody>
</table>
Table 2.5: Summary of significant signals for differentiation of four *Crataegus* spp. from organic acids region for pJRES. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM $d(i)$ values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed.

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>VIP score</th>
<th>$d(i)$</th>
<th>MDA</th>
<th>p-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>pJRES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.890939</td>
<td>1.28</td>
<td>60.48</td>
<td>9.61E-03</td>
<td>8.77E-07</td>
<td>7.86E-05</td>
</tr>
<tr>
<td>0.896068</td>
<td>1.62</td>
<td>53.08</td>
<td>9.29E-03</td>
<td>4.74E-07</td>
<td>7.86E-05</td>
</tr>
<tr>
<td>0.906325</td>
<td>1.54</td>
<td>42.44</td>
<td>5.01E-03</td>
<td>3.35E-06</td>
<td>1.03E-04</td>
</tr>
<tr>
<td>0.911453</td>
<td>1.13</td>
<td>48.96</td>
<td>4.57E-03</td>
<td>6.82E-06</td>
<td>1.39E-04</td>
</tr>
<tr>
<td>1.09095</td>
<td>3.86</td>
<td>35.84</td>
<td>1.06E-02</td>
<td>1.05E-06</td>
<td>7.86E-05</td>
</tr>
<tr>
<td>1.14736</td>
<td>1.21</td>
<td>62.41</td>
<td>5.74E-03</td>
<td>2.02E-06</td>
<td>7.86E-05</td>
</tr>
<tr>
<td>1.15249</td>
<td>2.29</td>
<td>76.52</td>
<td>7.11E-03</td>
<td>1.66E-06</td>
<td>7.86E-05</td>
</tr>
<tr>
<td>1.15762</td>
<td>2.99</td>
<td>282.13</td>
<td>4.82E-03</td>
<td>2.27E-06</td>
<td>8.12E-05</td>
</tr>
<tr>
<td>1.16275</td>
<td>1.67</td>
<td>68.53</td>
<td>5.16E-03</td>
<td>1.52E-06</td>
<td>7.86E-05</td>
</tr>
<tr>
<td>1.21404</td>
<td>2.31</td>
<td>45.14</td>
<td>5.34E-03</td>
<td>2.38E-05</td>
<td>2.22E-04</td>
</tr>
<tr>
<td>1.31148</td>
<td>1.56</td>
<td>163.28</td>
<td>3.49E-03</td>
<td>1.42E-05</td>
<td>1.79E-04</td>
</tr>
<tr>
<td>1.31661</td>
<td>2.43</td>
<td>107.46</td>
<td>5.51E-03</td>
<td>7.30E-06</td>
<td>1.41E-04</td>
</tr>
<tr>
<td>1.32173</td>
<td>4.38</td>
<td>81.15</td>
<td>3.77E-03</td>
<td>1.22E-05</td>
<td>1.66E-04</td>
</tr>
<tr>
<td>1.70124</td>
<td>1.23</td>
<td>204.94</td>
<td>4.52E-03</td>
<td>1.78E-04</td>
<td>7.53E-04</td>
</tr>
<tr>
<td>1.83971</td>
<td>1.19</td>
<td>134.92</td>
<td>3.96E-03</td>
<td>4.03E-05</td>
<td>3.09E-04</td>
</tr>
<tr>
<td>1.88587</td>
<td>2.29</td>
<td>34.88</td>
<td>5.70E-03</td>
<td>4.86E-06</td>
<td>1.30E-04</td>
</tr>
<tr>
<td>2.7013</td>
<td>1.07</td>
<td>46.53</td>
<td>3.90E-03</td>
<td>2.71E-06</td>
<td>8.94E-05</td>
</tr>
<tr>
<td>2.71156</td>
<td>2.06</td>
<td>53.99</td>
<td>5.15E-03</td>
<td>1.54E-06</td>
<td>7.86E-05</td>
</tr>
</tbody>
</table>
Table 2.6: Summary of significant signals for differentiation of four *Crataegus* spp. from organic acids region for $^1$H 1D PRESAT. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM $d(i)$ values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed.

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>VIP score</th>
<th>$d(i)$</th>
<th>MDA</th>
<th>p-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.857016</td>
<td>1.09</td>
<td>8.61</td>
<td>4.13E-03</td>
<td>1.17E-04</td>
<td>4.43E-04</td>
</tr>
<tr>
<td>0.887017</td>
<td>1.47</td>
<td>13.80</td>
<td>5.72E-03</td>
<td>8.11E-06</td>
<td>9.53E-05</td>
</tr>
<tr>
<td>0.902018</td>
<td>1.77</td>
<td>16.69</td>
<td>5.69E-03</td>
<td>2.70E-06</td>
<td>7.46E-05</td>
</tr>
<tr>
<td>0.907018</td>
<td>2.21</td>
<td>22.21</td>
<td>4.89E-03</td>
<td>5.94E-07</td>
<td>7.46E-05</td>
</tr>
<tr>
<td>0.912018</td>
<td>1.50</td>
<td>10.25</td>
<td>4.14E-03</td>
<td>2.18E-06</td>
<td>7.46E-05</td>
</tr>
<tr>
<td>0.95202</td>
<td>1.22</td>
<td>9.94</td>
<td>5.08E-03</td>
<td>3.22E-05</td>
<td>2.25E-04</td>
</tr>
<tr>
<td>0.962021</td>
<td>1.20</td>
<td>9.93</td>
<td>6.37E-03</td>
<td>3.68E-06</td>
<td>8.71E-05</td>
</tr>
<tr>
<td>1.15703</td>
<td>1.74</td>
<td>9.55</td>
<td>4.93E-03</td>
<td>5.01E-06</td>
<td>9.53E-05</td>
</tr>
<tr>
<td>1.16203</td>
<td>1.28</td>
<td>17.04</td>
<td>6.56E-03</td>
<td>1.86E-06</td>
<td>7.46E-05</td>
</tr>
<tr>
<td>1.16703</td>
<td>1.17</td>
<td>12.58</td>
<td>3.69E-03</td>
<td>1.66E-04</td>
<td>5.55E-04</td>
</tr>
<tr>
<td>1.69705</td>
<td>1.01</td>
<td>11.61</td>
<td>6.60E-03</td>
<td>3.73E-05</td>
<td>2.39E-04</td>
</tr>
<tr>
<td>1.88706</td>
<td>1.61</td>
<td>8.48</td>
<td>5.69E-03</td>
<td>1.45E-05</td>
<td>1.31E-04</td>
</tr>
<tr>
<td>2.7021</td>
<td>1.31</td>
<td>27.84</td>
<td>6.73E-03</td>
<td>6.54E-07</td>
<td>7.46E-05</td>
</tr>
<tr>
<td>2.7121</td>
<td>1.82</td>
<td>36.69</td>
<td>4.68E-03</td>
<td>2.04E-05</td>
<td>1.55E-04</td>
</tr>
<tr>
<td>2.7221</td>
<td>1.41</td>
<td>26.16</td>
<td>5.49E-03</td>
<td>9.19E-06</td>
<td>9.88E-05</td>
</tr>
</tbody>
</table>
Table 2.7: Summary of significant signals for differentiation of four *Crataegus* spp. from sugars region for pJRES. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM $d(i)$ values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed.

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>VIP score</th>
<th>$d(i)$</th>
<th>MDA</th>
<th>p-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>pJRES</td>
<td>3.13723</td>
<td>2.00</td>
<td>50.06</td>
<td>3.72E-03</td>
<td>9.06E-06</td>
</tr>
<tr>
<td></td>
<td>3.2398</td>
<td>1.08</td>
<td>70.49</td>
<td>5.83E-03</td>
<td>1.23E-04</td>
</tr>
<tr>
<td></td>
<td>3.24493</td>
<td>1.80</td>
<td>82.87</td>
<td>4.73E-03</td>
<td>1.90E-04</td>
</tr>
<tr>
<td></td>
<td>3.32185</td>
<td>1.63</td>
<td>24.10</td>
<td>3.43E-03</td>
<td>4.48E-05</td>
</tr>
<tr>
<td></td>
<td>3.47571</td>
<td>1.58</td>
<td>24.87</td>
<td>6.04E-03</td>
<td>3.66E-05</td>
</tr>
<tr>
<td></td>
<td>3.48084</td>
<td>1.63</td>
<td>21.31</td>
<td>5.88E-03</td>
<td>7.49E-05</td>
</tr>
<tr>
<td></td>
<td>3.98856</td>
<td>1.37</td>
<td>28.35</td>
<td>4.05E-03</td>
<td>2.04E-05</td>
</tr>
<tr>
<td></td>
<td>4.33217</td>
<td>1.45</td>
<td>51.05</td>
<td>3.36E-03</td>
<td>1.46E-04</td>
</tr>
<tr>
<td></td>
<td>4.3373</td>
<td>2.46</td>
<td>62.40</td>
<td>6.25E-03</td>
<td>7.07E-05</td>
</tr>
<tr>
<td></td>
<td>4.34243</td>
<td>2.41</td>
<td>138.16</td>
<td>4.43E-03</td>
<td>3.05E-05</td>
</tr>
<tr>
<td></td>
<td>4.34756</td>
<td>1.64</td>
<td>26.69</td>
<td>5.74E-03</td>
<td>1.26E-04</td>
</tr>
<tr>
<td></td>
<td>4.40397</td>
<td>1.18</td>
<td>61.62</td>
<td>2.96E-03</td>
<td>1.85E-05</td>
</tr>
<tr>
<td></td>
<td>4.4091</td>
<td>1.06</td>
<td>66.28</td>
<td>5.24E-03</td>
<td>1.51E-04</td>
</tr>
<tr>
<td></td>
<td>4.48603</td>
<td>1.69</td>
<td>36.35</td>
<td>1.13E-02</td>
<td>4.68E-06</td>
</tr>
<tr>
<td></td>
<td>4.99375</td>
<td>1.73</td>
<td>30.09</td>
<td>8.65E-03</td>
<td>6.74E-06</td>
</tr>
<tr>
<td></td>
<td>4.99888</td>
<td>1.36</td>
<td>21.10</td>
<td>7.03E-03</td>
<td>1.05E-05</td>
</tr>
<tr>
<td></td>
<td>5.07068</td>
<td>1.62</td>
<td>25.79</td>
<td>4.57E-03</td>
<td>6.56E-05</td>
</tr>
</tbody>
</table>
Table 2.8: Summary of significant signals for differentiation of four *Crataegus* spp. from sugars region for $^1$H 1D PRESAT. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM $d(i)$ values, and RF mean decrease accuracy (MDA). Bonferroni-corrected $p$-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed.

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>VIP score</th>
<th>$d(i)$</th>
<th>MDA</th>
<th>$p$-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$^1$H 1D PRESAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.14212</td>
<td>1.98</td>
<td>37.15</td>
<td>4.76E-03</td>
<td>2.76E-06</td>
<td>4.84E-05</td>
</tr>
<tr>
<td>3.81215</td>
<td>1.25</td>
<td>90.98</td>
<td>4.89E-03</td>
<td>4.33E-06</td>
<td>5.29E-05</td>
</tr>
<tr>
<td>4.21717</td>
<td>1.11</td>
<td>110.56</td>
<td>6.52E-03</td>
<td>6.81E-07</td>
<td>4.66E-05</td>
</tr>
<tr>
<td>4.48718</td>
<td>2.11</td>
<td>102.46</td>
<td>6.84E-03</td>
<td>9.12E-07</td>
<td>4.66E-05</td>
</tr>
<tr>
<td>4.49218</td>
<td>2.04</td>
<td>76.89</td>
<td>4.09E-03</td>
<td>1.73E-06</td>
<td>4.66E-05</td>
</tr>
<tr>
<td>5.43722</td>
<td>1.84</td>
<td>67.97</td>
<td>4.85E-03</td>
<td>5.40E-06</td>
<td>5.79E-05</td>
</tr>
</tbody>
</table>
Table 2.9: Summary of significant signals for differentiation of four *Crataegus* spp. from phenolics region for pJRES. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM $d(i)$ values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed. Putative identifications are listed where possible. QD=quercetin; CA=chlorogenic acid; HP=hyperoside; IV=isovitexin; VT=vitexin; VR=vitexin-2″-O-rhamnoside; RT=rutin.

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>Putative compound ID</th>
<th>VIP score</th>
<th>$d(i)$</th>
<th>MDA</th>
<th>p-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.14766</td>
<td></td>
<td>1.05</td>
<td>10.84</td>
<td>6.22E-03</td>
<td>8.78E-05</td>
<td>6.73E-04</td>
</tr>
<tr>
<td>6.15279</td>
<td></td>
<td>1.33</td>
<td>75.37</td>
<td>8.40E-03</td>
<td>9.43E-05</td>
<td>6.83E-04</td>
</tr>
<tr>
<td>6.15792</td>
<td></td>
<td>1.21</td>
<td>52.01</td>
<td>7.08E-03</td>
<td>1.51E-04</td>
<td>9.35E-04</td>
</tr>
<tr>
<td>6.16305</td>
<td></td>
<td>1.18</td>
<td>34.47</td>
<td>5.69E-03</td>
<td>6.18E-05</td>
<td>5.36E-04</td>
</tr>
<tr>
<td>6.16818</td>
<td></td>
<td>1.35</td>
<td>45.43</td>
<td>9.40E-03</td>
<td>7.28E-05</td>
<td>5.77E-04</td>
</tr>
<tr>
<td>6.17331</td>
<td>QD</td>
<td>1.30</td>
<td>30.68</td>
<td>4.72E-03</td>
<td>2.72E-06</td>
<td>1.37E-04</td>
</tr>
<tr>
<td>6.18356</td>
<td>QD</td>
<td>1.34</td>
<td>20.33</td>
<td>4.33E-03</td>
<td>1.45E-05</td>
<td>1.93E-04</td>
</tr>
<tr>
<td>6.37845</td>
<td>QD</td>
<td>1.31</td>
<td>20.34</td>
<td>4.32E-03</td>
<td>1.48E-05</td>
<td>1.93E-04</td>
</tr>
<tr>
<td>6.58359</td>
<td></td>
<td>1.77</td>
<td>25.42</td>
<td>2.95E-03</td>
<td>2.26E-05</td>
<td>2.59E-04</td>
</tr>
<tr>
<td>6.59385</td>
<td></td>
<td>2.77</td>
<td>55.75</td>
<td>8.01E-03</td>
<td>3.87E-06</td>
<td>1.37E-04</td>
</tr>
<tr>
<td>6.59897</td>
<td></td>
<td>2.67</td>
<td>12.16</td>
<td>9.80E-03</td>
<td>5.98E-06</td>
<td>1.37E-04</td>
</tr>
<tr>
<td>6.72719</td>
<td></td>
<td>1.15</td>
<td>19.01</td>
<td>2.77E-03</td>
<td>4.69E-05</td>
<td>4.40E-04</td>
</tr>
<tr>
<td>6.79386</td>
<td>CA</td>
<td>1.05</td>
<td>30.87</td>
<td>5.21E-03</td>
<td>1.55E-04</td>
<td>9.35E-04</td>
</tr>
<tr>
<td>6.8554</td>
<td>HP</td>
<td>2.44</td>
<td>12.12</td>
<td>9.86E-04</td>
<td>6.30E-05</td>
<td>5.36E-04</td>
</tr>
<tr>
<td>6.87591</td>
<td></td>
<td>2.22</td>
<td>41.26</td>
<td>4.19E-03</td>
<td>5.48E-06</td>
<td>1.37E-04</td>
</tr>
<tr>
<td>6.92207</td>
<td></td>
<td>4.86</td>
<td>31.34</td>
<td>5.94E-03</td>
<td>9.73E-06</td>
<td>1.47E-04</td>
</tr>
<tr>
<td>6.9272</td>
<td></td>
<td>4.43</td>
<td>37.20</td>
<td>6.01E-03</td>
<td>7.90E-06</td>
<td>1.37E-04</td>
</tr>
<tr>
<td>6.93746</td>
<td>IV or VT or VR</td>
<td>1.89</td>
<td>16.04</td>
<td>4.51E-03</td>
<td>5.10E-06</td>
<td>1.37E-04</td>
</tr>
<tr>
<td>7.28619</td>
<td>Naringenin</td>
<td>1.06</td>
<td>30.25</td>
<td>7.19E-03</td>
<td>5.37E-05</td>
<td>4.77E-04</td>
</tr>
<tr>
<td>7.30158</td>
<td></td>
<td>1.42</td>
<td>12.54</td>
<td>3.02E-03</td>
<td>1.55E-04</td>
<td>9.35E-04</td>
</tr>
<tr>
<td>7.46056</td>
<td></td>
<td>1.29</td>
<td>32.89</td>
<td>5.12E-03</td>
<td>7.30E-06</td>
<td>1.37E-04</td>
</tr>
<tr>
<td>7.46569</td>
<td></td>
<td>2.40</td>
<td>38.77</td>
<td>5.14E-03</td>
<td>8.04E-06</td>
<td>1.37E-04</td>
</tr>
</tbody>
</table>
Table 2.9: Summary of significant signals for differentiation of four *Crataegus* spp. from phenolics region for pJRES. Significant signals are listed for two NMR acquisition techniques. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM $d(i)$ values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed. Putative identifications are listed where possible. QD=quercetin; CA=chlorogenic acid; HP=hyperoside; IV=isovitexin; VT=vitexin; VR=vitexin-2″-O-rhamnoside; RT=rutin.

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>Putative compound ID</th>
<th>VIP score</th>
<th>$d(i)$</th>
<th>MDA</th>
<th>p-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>pJRES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.50159</td>
<td></td>
<td>1.82</td>
<td>23.88</td>
<td>5.27E-03</td>
<td>1.16E-05</td>
<td>1.62E-04</td>
</tr>
<tr>
<td>7.59903</td>
<td>HP</td>
<td>3.10</td>
<td>59.76</td>
<td>8.18E-03</td>
<td>3.85E-06</td>
<td>1.37E-04</td>
</tr>
<tr>
<td>7.60416</td>
<td>HP</td>
<td>3.43</td>
<td>46.00</td>
<td>6.86E-03</td>
<td>1.96E-06</td>
<td>1.37E-04</td>
</tr>
<tr>
<td>7.64519</td>
<td>RT or QD</td>
<td>1.76</td>
<td>9.56</td>
<td>9.69E-04</td>
<td>1.81E-04</td>
<td>9.99E-04</td>
</tr>
<tr>
<td>7.67596</td>
<td>RT</td>
<td>1.64</td>
<td>12.03</td>
<td>2.50E-03</td>
<td>6.22E-04</td>
<td>2.97E-03</td>
</tr>
<tr>
<td>7.79905</td>
<td></td>
<td>1.29</td>
<td>14.49</td>
<td>3.75E-03</td>
<td>1.01E-04</td>
<td>7.20E-04</td>
</tr>
<tr>
<td>7.80417</td>
<td></td>
<td>1.48</td>
<td>37.07</td>
<td>6.48E-03</td>
<td>5.78E-06</td>
<td>1.37E-04</td>
</tr>
<tr>
<td>7.81956</td>
<td></td>
<td>1.56</td>
<td>14.30</td>
<td>3.25E-03</td>
<td>9.09E-05</td>
<td>6.79E-04</td>
</tr>
<tr>
<td>7.82469</td>
<td>IV</td>
<td>1.68</td>
<td>23.83</td>
<td>3.53E-03</td>
<td>4.58E-05</td>
<td>4.40E-04</td>
</tr>
</tbody>
</table>
Table 2.10: Summary of significant signals for differentiation of four *Crataegus* spp. from phenolics region for $^1$H 1D PRESAT. Spectral bin significance is listed for PLS-DA variable importance in projection (VIP), SAM $d(i)$ values, and RF mean decrease accuracy (MDA). Bonferroni-corrected p-values and false discovery rates (FDR) by Kruskal-Wallis testing are listed. Putative identifications are listed where possible. QD=quercetin; CA=chlorogenic acid; HP=hyperoside; IV=isorvitexin; VT=vitexin; VR=vitexin-2″-O-rhamnoside; RT=rutin.

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>Putative compound ID</th>
<th>VIP score</th>
<th>$d(i)$</th>
<th>MDA</th>
<th>p-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.15225</td>
<td></td>
<td>1.09</td>
<td>86.93</td>
<td>3.27E-03</td>
<td>1.66E-04</td>
<td>3.93E-04</td>
</tr>
<tr>
<td>6.15725</td>
<td></td>
<td>1.24</td>
<td>146.82</td>
<td>8.18E-03</td>
<td>1.71E-06</td>
<td>3.03E-05</td>
</tr>
<tr>
<td>6.19225</td>
<td></td>
<td>2.79</td>
<td>62.71</td>
<td>3.46E-03</td>
<td>1.35E-06</td>
<td>3.03E-05</td>
</tr>
<tr>
<td>6.58727</td>
<td></td>
<td>2.27</td>
<td>54.82</td>
<td>5.75E-03</td>
<td>2.22E-06</td>
<td>3.03E-05</td>
</tr>
<tr>
<td>6.59227</td>
<td></td>
<td>2.67</td>
<td>73.03</td>
<td>3.92E-03</td>
<td>4.74E-07</td>
<td>3.03E-05</td>
</tr>
<tr>
<td>6.86228</td>
<td>HP</td>
<td>2.37</td>
<td>72.87</td>
<td>4.77E-03</td>
<td>3.08E-06</td>
<td>3.03E-05</td>
</tr>
<tr>
<td>6.93229</td>
<td>IV or VT or VR</td>
<td>3.69</td>
<td>54.51</td>
<td>2.41E-03</td>
<td>5.49E-06</td>
<td>3.52E-05</td>
</tr>
<tr>
<td>7.2773</td>
<td>Naringenin</td>
<td>1.03</td>
<td>72.46</td>
<td>4.03E-03</td>
<td>3.19E-05</td>
<td>1.10E-04</td>
</tr>
<tr>
<td>7.45231</td>
<td></td>
<td>1.13</td>
<td>74.34</td>
<td>4.81E-03</td>
<td>2.53E-06</td>
<td>3.03E-05</td>
</tr>
<tr>
<td>7.45731</td>
<td></td>
<td>1.41</td>
<td>96.37</td>
<td>2.34E-03</td>
<td>4.19E-06</td>
<td>3.11E-05</td>
</tr>
<tr>
<td>7.49731</td>
<td></td>
<td>1.73</td>
<td>119.20</td>
<td>3.58E-03</td>
<td>3.40E-06</td>
<td>3.03E-05</td>
</tr>
<tr>
<td>7.50231</td>
<td></td>
<td>1.65</td>
<td>93.60</td>
<td>4.33E-03</td>
<td>2.29E-06</td>
<td>3.03E-05</td>
</tr>
<tr>
<td>7.57731</td>
<td></td>
<td>1.91</td>
<td>55.00</td>
<td>2.86E-03</td>
<td>6.98E-06</td>
<td>3.59E-05</td>
</tr>
<tr>
<td>7.58731</td>
<td>HP or CA</td>
<td>2.40</td>
<td>61.22</td>
<td>3.04E-03</td>
<td>6.10E-06</td>
<td>3.59E-05</td>
</tr>
<tr>
<td>7.60732</td>
<td>HP</td>
<td>2.17</td>
<td>105.76</td>
<td>5.32E-03</td>
<td>2.59E-06</td>
<td>3.03E-05</td>
</tr>
<tr>
<td>7.61232</td>
<td>QD</td>
<td>1.60</td>
<td>87.88</td>
<td>3.60E-03</td>
<td>3.12E-06</td>
<td>3.03E-05</td>
</tr>
<tr>
<td>7.61732</td>
<td>QD</td>
<td>1.85</td>
<td>245.10</td>
<td>8.80E-03</td>
<td>1.01E-06</td>
<td>3.03E-05</td>
</tr>
<tr>
<td>7.63732</td>
<td>QD</td>
<td>1.55</td>
<td>66.26</td>
<td>4.06E-03</td>
<td>4.10E-06</td>
<td>3.10E-05</td>
</tr>
<tr>
<td>7.83233</td>
<td>IV</td>
<td>1.74</td>
<td>64.22</td>
<td>2.36E-03</td>
<td>3.71E-06</td>
<td>3.03E-05</td>
</tr>
<tr>
<td>7.95233</td>
<td></td>
<td>1.06</td>
<td>55.86</td>
<td>3.13E-03</td>
<td>6.74E-06</td>
<td>3.59E-05</td>
</tr>
<tr>
<td>7.96233</td>
<td>VT</td>
<td>2.68</td>
<td>58.92</td>
<td>3.59E-03</td>
<td>1.74E-06</td>
<td>3.03E-05</td>
</tr>
<tr>
<td>7.98233</td>
<td>VT</td>
<td>3.10</td>
<td>71.59</td>
<td>3.59E-03</td>
<td>4.99E-06</td>
<td>3.45E-05</td>
</tr>
<tr>
<td>7.98733</td>
<td>VT or VR</td>
<td>1.48</td>
<td>77.33</td>
<td>4.12E-03</td>
<td>2.62E-06</td>
<td>3.03E-05</td>
</tr>
</tbody>
</table>
upon the phenolics chemical data. The biomarkers discussed earlier contributed to separation of all four species in an overall approach.

CV performance in support vector machine (SVM) learning models (suitable for comparison of two groups; MCCV with multiple iterations—as discussed in Chapter 2.2.7) was used to identify phenolic biomarkers that separate the most closely related species (Table 2.11; Table 2.12). Additionally, the markers that differentiate European *Crataegus* from North American *Crataegus* are examined by this method (Table 2.13). Phenolic biomarkers were chosen in this approach because models from this chemical class provided the highest resolution of species differentiation. Important features are selected and numerically expressed as frequency of selection (FoS) and mean importance measure (MIM). FoS is an expression of how often a feature is selected as being important in the training set during CV of the SVM model and as the value approaches 1, the bin is selected more iterations. FoS shows the stability of the rank of a given bin’s importance—naturally, this measure is quite robust and insensitive to outliers. The same cannot be said for MIM, however MIM is able to provide a quantitative measure of a bin’s importance as selected from training data during CV of the SVM model. The value MIM is the proportion that a given feature contributes to overall model accuracy during CV (Xia et al., 2013). Since this is a binary comparison, relative intensities (correlated here to concentration) of bins are listed for each group compared. Absolute concentrations of many compounds corresponding to significant bins discussed here are provided in Chapter 3.

17 bins (2 identified) for pJRES and 6 bins (none identified) for 1D ¹H PRESAT were found significant in separating *C. monogyna* from *C. laevigata*. Identified significant bins correlated to quercetin. The results presented here suggest that the concentration of
Table 2.11: Summary of significant phenolic signals for differentiation of closely related *Crataegus* spp. by pJRES. Significant signals are listed for pJRES NMR acquisition technique by cross-validation (CV) performance in models generated by support vector machine (SVM) learning. Spectral bin significance is listed for differentiation of either *C. monogyna* from *C. laevigata* or *C. douglasii* from *C. okanaganensis*. Importance measures are reported as both frequency of selection (FoS) and mean importance measure (MIM). Putative identifications are listed where possible. QD=quercetin; HP=hyperoside; RT=rutin.

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>Putative compound ID</th>
<th>FoS</th>
<th>MIM</th>
<th>Concentration in <em>C. laevigata</em></th>
<th>Concentration in <em>C. monogyna</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>6.8195</td>
<td>1</td>
<td>7.49E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>7.80418</td>
<td>1</td>
<td>3.56E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>6.82463</td>
<td>1</td>
<td>3.51E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>6.17844</td>
<td>QD</td>
<td>2.79E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>6.69642</td>
<td>0.98</td>
<td>4.22E-04</td>
<td>High</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>6.1169</td>
<td>0.98</td>
<td>3.97E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>6.83489</td>
<td>0.98</td>
<td>2.73E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>7.55801</td>
<td>0.98</td>
<td>2.66E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>6.17331</td>
<td>QD</td>
<td>3.00E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>7.79905</td>
<td>0.94</td>
<td>2.86E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>6.70155</td>
<td>0.88</td>
<td>2.96E-04</td>
<td>High</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>7.55288</td>
<td>0.84</td>
<td>2.23E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>6.37332</td>
<td>0.82</td>
<td>2.49E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>6.84002</td>
<td>0.82</td>
<td>2.09E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>6.90156</td>
<td>0.78</td>
<td>2.72E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>7.11183</td>
<td>0.76</td>
<td>2.73E-04</td>
<td>Low</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>6.75796</td>
<td>0.76</td>
<td>2.38E-04</td>
<td>High</td>
<td>Low</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.11: Summary of significant phenolic signals for differentiation of closely related *Crataegus* spp. by pJRES. Significant signals are listed for pJRES NMR acquisition technique by cross-validation (CV) performance in models generated by support vector machine (SVM) learning. Spectral bin significance is listed for differentiation of either *C. monogyna* from *C. laevigata* or *C. douglasii* from *C. okanaganensis*. Importance measures are reported as both frequency of selection (FoS) and mean importance measure (MIM). Putative identifications are listed where possible. QD=quercetin; HP=hyperoside; RT=rutin.

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>Putative compound ID</th>
<th>FoS</th>
<th>MIM</th>
<th>Concentration in <em>C. douglasii</em></th>
<th>Concentration in <em>C. okanaganensis</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>7.77341</td>
<td></td>
<td>0.98</td>
<td>5.10E-04</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>7.80418</td>
<td></td>
<td>0.98</td>
<td>4.67E-04</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>6.72719</td>
<td></td>
<td>0.9</td>
<td>4.57E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>6.17844</td>
<td>QD</td>
<td>0.86</td>
<td>3.80E-04</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>7.77853</td>
<td></td>
<td>0.82</td>
<td>4.31E-04</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>6.73232</td>
<td></td>
<td>0.82</td>
<td>3.86E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>6.41435</td>
<td>RT or HP</td>
<td>0.78</td>
<td>4.11E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>6.18357</td>
<td>QD</td>
<td>0.76</td>
<td>3.61E-04</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>
Table 2.12: Summary of significant phenolic signals for differentiation of closely related *Crataegus* spp. by 1D $^1$H PRESAT. Significant signals are listed for 1D $^1$H PRESAT NMR acquisition technique by cross-validation (CV) performance in models generated by support vector machine (SVM) learning. Spectral bin significance is listed for differentiation of either *C. monogyna* from *C. laevigata* or *C. douglasii* from *C. okanaganensis*. Importance measures are reported as both frequency of selection (FoS) and mean importance measure (MIM). Putative identifications are listed where possible.

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>Putative compound ID</th>
<th>FoS</th>
<th>MIM</th>
<th>Concentration in <em>C. laevigata</em></th>
<th>Concentration in <em>C. monogyna</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>6.11225</td>
<td></td>
<td>1</td>
<td>4.48E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>7.40731</td>
<td></td>
<td>0.98</td>
<td>4.65E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>7.42731</td>
<td></td>
<td>0.96</td>
<td>3.29E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>7.37731</td>
<td></td>
<td>0.76</td>
<td>3.28E-04</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>7.00729</td>
<td></td>
<td>0.76</td>
<td>3.23E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>6.00724</td>
<td></td>
<td>0.76</td>
<td>3.17E-04</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>

1D $^1$H PRESAT: *C. douglasii* vs. *C. okanaganensis*

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>Putative compound ID</th>
<th>FoS</th>
<th>MIM</th>
<th>Concentration in <em>C. douglasii</em></th>
<th>Concentration in <em>C. okanaganensis</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>7.52231</td>
<td></td>
<td>1</td>
<td>1.89E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>7.2523</td>
<td></td>
<td>0.98</td>
<td>1.81E-04</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>7.38731</td>
<td></td>
<td>0.94</td>
<td>1.80E-04</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>7.75232</td>
<td></td>
<td>0.94</td>
<td>1.66E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>7.38231</td>
<td></td>
<td>0.9</td>
<td>1.47E-04</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>
Table 2.13: Summary of significant phenolic signals for differentiation of European and North American *Crataegus* spp. by 1D 'H PRESAT and pJRES. Significant signals are listed for 1D 'H PRESAT and pJRES NMR acquisition technique by cross-validation (CV) performance in models generated by support vector machine (SVM) learning. Spectral bin significance is listed for differentiation of *C. monogyna* and *C. laevigata* from *C. douglasii* and *C. okanaganensis*. Importance measures are reported as both frequency of selection (FoS) and mean importance measure (MIM). Putative identifications are listed where possible. QD=quercetin; CA=chlorogenic acid; HP=hyperoside; IV=isorvitexin; VT=vitexin; VR=vitexin-2″-O-rhamnoside; RT=rutin.

### PJRES: European *Crataegus* vs. North American *Crataegus*

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>Putative compound ID</th>
<th>FoS</th>
<th>MIM</th>
<th>Concentration in European <em>Crataegus</em></th>
<th>Concentration in North American <em>Crataegus</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>7.64006</td>
<td>RT</td>
<td>0.78</td>
<td>4.59E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>7.97855</td>
<td>VT or VR</td>
<td>0.64</td>
<td>4.39E-04</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>7.38877</td>
<td>QD</td>
<td>0.38</td>
<td>4.13E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>7.59904</td>
<td>QD</td>
<td>0.32</td>
<td>3.99E-04</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>6.93746</td>
<td>IV or VT or VR</td>
<td>0.26</td>
<td>3.92E-04</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

### 1D 'H PRESAT: European *Crataegus* vs. North American *Crataegus*

<table>
<thead>
<tr>
<th>Spectral bin (ppm)</th>
<th>Putative compound ID</th>
<th>FoS</th>
<th>MIM</th>
<th>Concentration in European <em>Crataegus</em></th>
<th>Concentration in North American <em>Crataegus</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>7.5303</td>
<td>QD</td>
<td>0.92</td>
<td>2.26E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>7.91733</td>
<td>QD</td>
<td>0.92</td>
<td>2.26E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>7.63232</td>
<td>RT</td>
<td>0.92</td>
<td>2.13E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>7.49731</td>
<td>CA</td>
<td>0.92</td>
<td>1.63E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>6.10225</td>
<td>QD</td>
<td>0.54</td>
<td>1.60E-04</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>7.76232</td>
<td>QD</td>
<td>0.54</td>
<td>1.55E-04</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>6.36226</td>
<td>IV or VT or VR</td>
<td>0.4</td>
<td>1.48E-04</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>6.94229</td>
<td>IV or VT or VR</td>
<td>0.36</td>
<td>1.44E-04</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>7.05229</td>
<td>QD</td>
<td>0.32</td>
<td>1.46E-04</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>
quercetin is significantly higher in samples of *C. monogyna* in comparison to *C. laevigata*, however the method used here is not an absolute quantification. Previous studies have also detected significant concentrations of quercetin in *C. monogyna* (Nikolov et al., 1982).

8 bins (3 identified) for pJRES and 5 bins (none identified) for 1D $^1$H PRESAT were found significant in separating *C. douglasii* from *C. okanaganensis*. Identified significant bins correlated to quercetin and rutin or hyperoside.

7 bins (5 identified) for pJRES and 10 bins (5 identified) for 1D $^1$H PRESAT were found significant in separating European *Crataegus* from North American *Crataegus*. In general, vitexin and its derivatives were in relatively higher concentrations in European hawthorns and were significant in separating them from North American hawthorns. Conversely, rutin was in relatively higher concentrations in North American hawthorns and were significant in separating them from European hawthorns. Other identified significant bins correlated to quercetin and chlorogenic acid.

### 2.4.5 Summary and applications of findings

The current study demonstrated the utility of NMR spectroscopy for generation of exploratory models that discriminate four species of hawthorn. Overall, 1D $^1$H PRESAT NMR acquisition is superior for construction of both supervised and unsupervised multivariate models in comparison with pJRES-acquired spectra for examination of sugars and organic acids regions. Both 1D $^1$H NMR or pJRES acquisition are appropriate for models generated from phenolics regions of spectra, and based on the data, phenolics compounds do significantly vary among *Crataegus* spp.. Combinations of significance analysis methods were used to provide lists of potential biomarkers that may give insight to how we may best
authenticate hawthorn NHPs or make hypotheses about differential therapeutic effects of hawthorn NHPs. The results of the experiments suggest that we may best differentiate *Crataegus* spp. by NMR-acquired data that correlate to phenolic compounds. It should also be noted that a majority of known therapeutic compounds in hawthorn are phenolics. With this in mind, we also applied SVM in an attempt to provide lists of potential biomarkers in the phenolics spectral region that separate the most closely related species. For this dataset, machine learning algorithms were applied to the chemometrics data and results from RF correlated with and outperformed PLS-DA, an aging and more traditional method.

It is worth noting a similar work to this chapter by Prinz et al. (2007) also uses chemical data for identification/differentiation of various *Crataegus* spp., however the authors were unable to objectively and comprehensively distinguish species by the targeted flavonoids since no univariate or multivariate statistical tools were applied for this purpose. In the current study, exploratory multivariate analyses allowed for a much more comprehensive overview of the metabolites present in the leaf extracts. With the combination of supervised models, machine learning algorithms, and univariate statistics, significant metabolites for species differentiation were discovered in a powerful and objective manner.

As discussed in more detail in Chapters 1.6 and 1.7, regulatory requirements for NHPs/DSs demand that specifications for NHPs include identity (based on genus and species level) of the source material used in NHP production (AHPA-AHP, 2006). Hence, it is important that there are methods available to differentiate species. Due to the aftermath of NHP authentication-related scandals reported widely in the media, most recently the DNA-barcoding scandal of 2015, spearheaded by the New York Attorney General, there has been a push for chemical authentication technologies. As discussed in Chapter 1.7, DNA barcoding
is simply unsuitable alone for authentication of finished NHPs due to several factors including extensive DNA cross-contamination and lack of DNA in many extracts and tinctures. Even with pure hawthorn samples, DNA barcoding provides poor resolution at the species level in the genus *Crataegus* (Zarrei et al., 2015). Chemical techniques allow for reliable species authentication and provide more information than DNA barcoding alone, such as the quality of bioactive compounds in a sample.

Despite the optimistic results of the current study, a larger dataset and external validation would be required for application to authentication problems and samples originating from different ecotypes. Also, multivariate data models of fruit material or leaf and flower combinations would be necessary for NHP authentication, as these represent a majority of *Crataegus* NHP formulations.

Future hypotheses may also be guided by the results of this exploratory study. The results demonstrate that discrimination in the exploratory models is driven by differences in signals that correspond to phenolic compounds that are associated with health benefits (as discussed in Chapter 1.4). Consequently, it is possible that there would be differential therapeutic effects attributed to two species that differ in the content of these phytochemicals. Hypotheses such as these could guide future *in vitro* investigations that, if promising, could be used to guide hypotheses for *in vivo* studies. Additionally, based upon the data here, it is reasonable to hypothesize that those chemical compounds found to drive variance in discriminatory models will have significantly different chemical concentrations when compared as absolute concentrations among samples from the same biological replicates studied here.
Chapter 3: Quantification of *Crataegus* flavonoids by nuclear magnetic resonance spectroscopy

3.1 Synopsis

Flavonoids are the primary bioactive compounds in *Crataegus* and so they hold great potential for use as an authentication marker or quality marker. Typically two of these compounds are specified for quantification in hawthorn monographs (BP, 2009; NNHPD, 2013a, b; PCPRC, 2010; USP, 2012). Quantification of the flavonoids in North American *Crataegus* (and comparison to the commonly-used European species) is a necessary first step toward future study of the therapeutic potential of these hawthorns and future development of authentication models. There are many published methods for the determination of flavonoids in *Crataegus* (Cui et al., 2006; Melikoglu et al., 2004; Prinz et al., 2007; Ringl et al., 2007; USP, 2012), however few are validated and most rely on use of either surrogate markers or specific chemical reference materials. There is a need for and a lack of appropriate reference standards and proper validation of chemical analysis methods for application to botanical materials (Betz et al., 2011; Brown et al., 2014; Brown and Lister, 2014). Application of nuclear magnetic resonance spectroscopy (NMR) is a possible solution to these problems. NMR enables multi-component analysis without the requirement for expensive, unavailable, or unstable chemical calibration standards. NMR is used in order to elucidate the content of seven flavonoids (naringenin, quercetin, hyperoside, rutin, vitexin, isovitexin, and vitexin-2"-O-rhamnosome) and one hydroxycinnamic acid (chlorogenic acid) in two European hawthorns, *C. monogyna* and *C. laevigata*, and two North American hawthorns, *C. douglasii* and *C. okanaganensis*. Quantification results presented here are validated by an orthogonal analytical platform, high performance liquid chromatography –
diode array detection. NMR was not useful for detection and quantification of flavonoids in relatively low concentrations due to relatively poor detection limits, however a wider range of compounds could be examined without chemical standards with reasonable accuracy (as evaluated by comparison to HPLC-DAD results). The results presented here demonstrate the potential of NMR for quantification of a wider range of compounds with a relatively high concentration that are difficult to find reference materials for or are costly to purchase.

3.2 Research objectives

I hypothesize that NMR spectroscopy is a quantitative tool that allows one to quantify a wider range of compounds in *Crataegus* leaf extracts without external chemical calibration. The current study was designed to address the following objectives in order to investigate the hypothesis: (1) to fine-tune extraction and instrument parameters to get reliable quantification data; (2) to measure the quantity of the significant compounds in the hawthorn samples and determine how they differ among four *Crataegus* spp.; and (3) to compare the quantitative NMR results to a validated orthogonal technique performed on the same biological samples.

3.3 Experimental

3.3.1 Plant materials

Leaf material from twelve individual hawthorn trees representing four species were collected by Jeanette Lee on August 22, 2008 from the Naturally Grown Herb and Spice Producers Cooperative (HerbPro) agroforestry demonstration farm in Edgewood, BC, Canada and species identified by Dr. Timothy A. Dickinson of the Royal Ontario Museum
with vouchers stored at Herbarium of the Royal Ontario Museum (as described in Chapter 2.2.1; Table 2.1; Figure 2.1). Three technical replicates were derived from each biological replicate.

3.3.2 NMR experimental materials

4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) (97 %) NMR reference standard and 99% deuterated methanol (methanol-$d_4$) were purchased from Sigma-Aldrich (St. Louis, MO, USA). A Bellco Collector™ Tissue Sieve (30 mesh, 520 µm screen) was purchased from PhytoTechnology Laboratories (Shawnee Mission, KS, USA). 3 mL Luer-Lok™ tip syringes, 1.7 mL microtubes, 2.0 mL microtubes, and 0.45 µm polyfluorotetraethylene (PFTE) syringe filters were purchased from Fisher Scientific Company (Ottawa, ON, Canada). Aldrich ColorSpec NMR tubes (7 in. L x 5 mm diam., 0.38 mm wall, 400 MHz) were purchased from Sigma-Aldrich.

3.3.3 NMR sample preparation

The method for NMR sample preparation is optimized as described in Appendix A. 100 mg of ground lyophilized hawthorn leaf was added to a 1.7 mL microcentrifuge tube and extracted sequentially with three aliquots (1.0 mL, 0.5 mL, 0.5 mL) of deuterated methanol (methanol-$d_4$) containing 5 mM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) as an internal standard by 30 seconds of vortex mixing and 8 total hours of extraction (2h40m per aliquot) by water bath at 60°C (Boekel Grant PB-600; Feasterville, PA, USA). After water bath extractions, the mixtures were centrifuged for 5 min at 16,000 x g, and the supernatants
collected and combined in a 2.0 mL microcentrifuge tube. This extract was filtered through a 0.45 μm PFTE syringe filter prior to transfer into an NMR tube for analysis.

3.3.4 NMR data acquisition

A Varian MercuryPlus 400 MHz NMR instrument (Varian, Inc., Palo Alto, CA, USA) equipped with a 400 MHz automated triple broadband probe equipped with a pulse field gradient generator (1H at 400.14 MHz) was used to acquire spectral data. Software control of the instrument was performed with Varian VnmrJ version 2.2 for Linux.

1D 1H NMR experiments were acquired at 25°C over a spectral width of 4201.7 Hz with an observe pulse of 11.80 μs (90°) using a PURGE solvent suppression sequence. A relaxation delay of 40.000 s was used between acquisitions. Total acquisition time was 3.899 s yielding 16,000 points. 64 transients were collected from each spectrum.

3.3.5 NMR spectral data processing

Spectra were processed using MestReNova version 10.0.1 (Mestrelab Research S.L., Santiago de Compostela, Spain) and were manually phased, baseline corrected using a Whittaker Smoother base point detection and spline fitting algorithm with the filter set to 16 and the smooth factor set to 20,000. The spectra were referenced to the DSS chemical shift at -0.018 ppm.

Signals of interest to be integrated for quantification of correlated compounds were determined by pure chemical standard spectra where the signals are relatively well-isolated and possible to integrate accurately and precisely on hawthorn extract spectra. The compounds quantified and associated signals for quantification are outlined in Appendix C.
For a peak identified as correlating to a compound of interest, the peak was normalized to its proton count and the normalized value of DSS was recorded. This normalized DSS value was then used to calculate concentration (molarity) of the compound of interest in the extract by the following equation:

\[
[\text{compound of interest}] = \frac{9 \times [\text{DSS}]}{\text{normalized DSS integration}}
\]

The concentration was then used to calculate an approximation of milligrams compound of interest per gram of leaf material.

3.3.6 NMR instrumental detection limits

Since NMR signals are Lorenzian lines, S/N ratios alone cannot be used to calculate limits of detection (LOD) and limits of quantification (LOQ) for a NMR instrument (Holzgrabe, 2008). To determine LOD and LOQ for the NMR instrument used in this study, a 10-point (equidistant) standard curve of was generated from ≥95% chlorogenic acid (Sigma-Aldrich; St. Louis, MO, USA) in methanol-\textit{d}\textsubscript{4} containing DSS internal standard, ranging from 5.64 mM to 0.56 mM chlorogenic acid. Signal used for quantification is described in Appendix C. The normalized detector response was used in generation of the curve for calculation of instrument LOD and LOQ for any NMR signal. Instrument LOD was calculated based on slope and standard error of the regression line where \( \text{LOQ}=3.3\sigma/s \); likewise, instrument LOQ was calculated by \( \text{LOD}=10\sigma/s \) (\( \sigma \) is standard deviation of the response and \textit{s} is the slope of the standard curve).


3.3.7 **HPLC experimental materials**

Five flavonoid chemical reference standards: vitexin-2″-O-rhamnoside (purity 94%), isovitexin (purity 96.6%), vitexin (purity 95.9%), rutin (purity 89.3%), and hyperoside (purity 99.1%), were purchased from ChromaDex (Irvine, CA, USA). Reference standards were stored at room temperature and desiccated. Water used in experiments is nanopure grade and HPLC-grade methanol was purchased from VWR International (Mississauga, ON, Canada). Reference standard stock solutions (1000 µg/mL) were prepared by weighing 10 mg of standard material into individual 10 mL volumetric flasks and made up to volume with 70:30 methanol:water. Standards were stored at 4ºC until use.

3.3.8 **HPLC sample preparation**

Samples were prepared according to validated method (Mudge and Brown, 2016). 200.0 ± 5.0 mg of ground lyophilized hawthorn leaf was added to a 50 mL centrifuge tube and extracted with 15 mL methanol/water/acetic acid (50:48:2, % v/v/v) by 1 minute of vortex mixing followed by 60 minutes on a Burrell Scientific Model 75 wrist-action shaker (Pittsburgh, PA, USA). After shaking, the mixture was centrifuged for 12 minutes at 1467 x g, the supernatant collected and filtered through a 0.45 µm PFTE syringe filter prior to transfer to a HPLC vial for analysis.

3.3.9 **Chromatographic instrumentation and conditions**

Chromatographic parameters are described as optimized by the Natural Health & Food Products Research Group (NRG) at British Columbia Institute of Technology (BCIT) (Mudge and Brown, 2016). Samples were analyzed on an Agilent 1200 series liquid
chromatograph equipped with a binary pump, in-line degasser, temperature-controlled column compartment, and diode array detector (DAD; Mississauga, ON, Canada). Software data acquisition and workup was performed with Agilent Chemstation (Mississauga, ON, Canada). Chromatographic separation was achieved by a Phenomenex Kinetex C\textsubscript{18} analytical column, 2.6 \(\mu\)m, 150 x 4.6 mm id. (Torrance, CA, USA) with column temperature controlled to 25ºC. Sample injection volume was 3 \(\mu\)L and flow rate was set to 0.4 mL/min. Mobile phase conditions are as follows: 0.01% phosphoric acid in water:tetrahydrofuran/acetonitrile/2-propanol (8:4:1, v/v/v); 1.0-12.0 minutes, 85:15-82:18 v/v; 12.0-22.0 minutes, 82:18-80:20 v/v; 22.0-23.0 minutes, 80-20-25:75 v/v; 23.0-25.0 minutes, 25:75 v/v; 25.0-25.5 minutes, 25:75-85:15 v/v; 25.5-30.5 minutes, 85:15 v/v. Experiment run time is 30.5 minutes with 5.0 minute post-run. DAD data was acquired at 270, 340, and 360 nm. Quantification was performed on data acquired at 340 nm.

### 3.3.10 Chromatographic data processing

Peaks of interest were integrated in Agilent Chemstation and these values recorded. Standard curves of the external chemical standards vitexin-2\(^{-}\)O-rhamnoside, isovitexin, vitexin, rutin, and hyperoside were created in Microsoft Excel and used in determining concentration of these compounds from peak integrals (Appendix D). The concentration was then used to calculate an approximation of milligrams compound of interest per gram of leaf material.
3.3.11 Statistical data processing

Quantitative data from both HPLC and NMR were imported into Microsoft Excel (Redmond, WA, USA) for conversions from integral data to concentrations of compounds of interest. These data were then imported into Minitab 17 (State College, PA, USA) for statistical comparisons between platforms and among species examined.

Minitab 17 was also used to generate boxplots which represent the distribution of concentrations of compounds measured in *Crataegus* samples. Here, the box represents the interquartile range of the data (middle-50% of data distribution) and the upper and lower whiskers represent the upper and lower 25% of the distribution, respectively. Mean concentrations are indicated by sun cross symbol (⊕). Median concentrations are represented by horizontal line. An asterisk (*) indicates an outlier sample.

Comparisons between analytical platforms were performed by a two-sample t test at α=0.05. Comparisons among four *Crataegus* species were performed by analysis of variance (ANOVA) with Tukey’s honest significant difference post-hoc test at α=0.05 (Appendix E).

3.4 Results and discussion

3.4.1 Quantification of flavonoids in *Crataegus* leaf extracts

For the current study, extractions were optimized to ensure complete extraction of flavonoid compounds; it was ensured that all solvents used contained an accurately known concentration of internal standard; signals corresponding to compounds of interest were carefully chosen for minimal signal overlap; PURGE suppression was used to suppress unwanted resonances from exchangeable protons (as discussed in Chapter 2.3); an
appropriate pulse delay was chosen (Appendix B); the baseline was filtered; and all samples were shimmed manually. The relaxation delay of 40 seconds was selected for a sufficient duration that ensured full relaxation for peaks of interest as determined by an inversion recovery experiment (at approximately 8 seconds, all resonances of interest have exceeded $T_1$; Appendix B).

Instrument LOD and LOQ for the NMR was determined from chlorogenic acid standard curve data (Appendix F) using the formulae described in Chapter 3.2.6. The instrument LOQ for signals of interest is 0.58 mM and the instrument LOD is 0.19 mM. These limits are converted for mg/g leaf equivalents for each compound and summarized in Table 3.1. Linearity of the chlorogenic acid standard curve is $R^2=0.9988$ (Appendix F).

**Table 3.1: Instrumental limits of detection for compounds measured by quantitative NMR in Chapter 3.** Concentrations are expressed as mg/g dry leaf equivalents. LOD is instrument limit of detection and LOQ is instrument limit of quantification.

<table>
<thead>
<tr>
<th></th>
<th>LOD (mg/g leaf equivalents)</th>
<th>LOQ (mg/g leaf equivalents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vitexin-2″-O-rhamnoside</td>
<td>0.88</td>
<td>2.68</td>
</tr>
<tr>
<td>Vitexin</td>
<td>0.66</td>
<td>2.01</td>
</tr>
<tr>
<td>Isovitexin</td>
<td>0.66</td>
<td>2.01</td>
</tr>
<tr>
<td>Rutin</td>
<td>0.93</td>
<td>2.83</td>
</tr>
<tr>
<td>Hyperoside</td>
<td>0.71</td>
<td>2.15</td>
</tr>
<tr>
<td>Naringenin</td>
<td>0.41</td>
<td>1.26</td>
</tr>
<tr>
<td>Quercetin</td>
<td>0.51</td>
<td>1.57</td>
</tr>
<tr>
<td>Chlorogenic acid</td>
<td>0.54</td>
<td>1.64</td>
</tr>
</tbody>
</table>
A limitation of the experiment should be noted—the HPLC analysis includes the full set of samples as discussed in Chapter 3.2.1, however sample 498 is absent in the NMR analysis due to lack of sample quantity. This sample represents one of *C. laevigata*, and so the quantitative results by NMR for this species may not be as representative as the other three species that have a full complement of samples.

Quantification of six flavonoids (vitexin, vitexin-2″-O-rhamnoside, isovitexin, rutin, hyperoside, quercetin, and naringenin) and one hydroxycinnamic acid (chlorogenic acid) was attempted from the NMR spectra of hawthorn leaf extracts. HPLC data was also acquired from samples of the same trees to provide an orthogonal platform for comparison to the NMR data. The HPLC method was used to separate and quantify vitexin, vitexin-2″-O-rhamnoside, isovitexin, rutin, and hyperoside in the hawthorn leaf extracts. Results of quantitative data from both platforms are compared where applicable—the HPLC method does not separate quercetin, naringenin, or chlorogenic acid, and so they are only quantified by NMR here.

Vitexin-2″-O-rhamnoside (Figure 1.6; Chapter 1.4) is found in numerous species of *Crataegus* and is believed to have numerous therapeutic effects (Li et al., 2015; Orhan et al., 2007; Wei et al., 2014). Concentrations of vitexin-2″-O-rhamnoside were compared among four *Crataegus* spp., as quantified from NMR and HPLC-DAD data (Table 3.2; Figure 3.1). There were no significant differences found between mean concentrations of the compound in comparing measurements from NMR and HPLC-DAD data. The mean concentrations of vitexin-2″-O-rhamnoside (as determined by NMR) in the leaf extracts of two European species, *C. monogyna* (11.6 mg/g leaf) and *C. laevigata* (14.4 mg/g leaf) are not significantly different from each other. The compound was not detected in *C. okanaganensis* or
Table 3.2: Concentrations of flavonoids in leaf extracts of four *Crataegus* spp. as measured by NMR (eight compounds measured) and HPLC-DAD (five compounds measured) (n=9 per species, per method [except *C. laevigata* n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material and expressed as mg/g dry leaf. Error represented as standard error of the mean. ‘ND’ indicates compound not detected in sample. ‘<LOQ’ indicates compound was detected in sample, however below the instrumental limit of detection.

<table>
<thead>
<tr>
<th></th>
<th>Vitexin-2&quot;-O -rhamnoside</th>
<th>Vitexin</th>
<th>Isovitexin</th>
<th>Rutin</th>
<th>Hyperoside</th>
<th>Quercetin</th>
<th>Naringenin</th>
<th>Chlorogenic acid</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. monogyna</strong></td>
<td>11.6 ± 1.7</td>
<td>&lt;LOQ</td>
<td>ND</td>
<td>&lt;LOQ</td>
<td>14.5 ± 0.6</td>
<td>ND</td>
<td>ND</td>
<td>17.8 ± 1.2</td>
</tr>
<tr>
<td><strong>C. laevigata</strong></td>
<td>14.4 ± 1.5</td>
<td>&lt;LOQ</td>
<td>ND</td>
<td>ND</td>
<td>9.4 ± 0.4</td>
<td>ND</td>
<td>ND</td>
<td>10.1 ± 0.4</td>
</tr>
<tr>
<td><strong>C. douglasii</strong></td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>4.8 ± 0.5</td>
<td>14.2 ± 0.9</td>
<td>ND</td>
<td>2.9 ± 0.3</td>
<td>16.6 ± 1.3</td>
</tr>
<tr>
<td><strong>C. okanaganensis</strong></td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>10.1 ± 0.8</td>
<td>6.4 ± 0.6</td>
<td>&lt;LOQ</td>
<td>ND</td>
<td>24.3 ± 1.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Vitexin-2&quot;-O -rhamnoside</th>
<th>Vitexin</th>
<th>Isovitexin</th>
<th>Rutin</th>
<th>Hyperoside</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. monogyna</strong></td>
<td>9.03 ± 2.08</td>
<td>0.16 ± 0.01</td>
<td>0.43 ± 0.10</td>
<td>0.36 ± 0.12</td>
<td>17.15 ± 1.16</td>
</tr>
<tr>
<td><strong>C. laevigata</strong></td>
<td>14.74 ± 0.42</td>
<td>0.95 ± 0.13</td>
<td>0.37 ± 0.02</td>
<td>0.05 ± 0.03</td>
<td>7.24 ± 0.42</td>
</tr>
<tr>
<td><strong>C. douglasii</strong></td>
<td>0.12 ± 0.01</td>
<td>ND</td>
<td>0.08 ± 0.02</td>
<td>4.94 ± 0.25</td>
<td>17.76 ± 1.00</td>
</tr>
<tr>
<td><strong>C. okanaganensis</strong></td>
<td>0.24 ± 0.03</td>
<td>ND</td>
<td>ND</td>
<td>9.55 ± 0.39</td>
<td>10.24 ± 0.61</td>
</tr>
</tbody>
</table>
Figure 3.1: Concentration of vitexin-2-O-rhamnoside in leaf extracts of four *Crataegus* spp. as measured by NMR and HPLC-DAD (n=9 per species, per method [except *C. laevigata* n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material. Upper whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol (⊕). Median concentration is represented by horizontal line. Boxplot not shown for *C. douglasii* and *C. okanaganensis* as measured by NMR since concentrations are below instrument LOQ.
*C. douglasii* by NMR, however was detected in trace amounts by HPLC-DAD (0.12 mg/g leaf in *C. douglasii* and 0.24 mg/g leaf in *C. okanaganensis*).

Vitexin (Figure 1.6; Chapter 1.4) is also found in numerous species of *Crataegus* and like its rhamnoside derivative, is associated with various cardioprotective effects and antioxidant activity (Bahorun et al., 1994; Edwards et al., 2012; Lu et al., 2013). Concentrations of vitexin were compared among four *Crataegus* spp., as quantified from NMR and HPLC-DAD data (Table 3.2; Figure 3.2). Vitexin was detected by NMR in leaf of *C. monogyna* and *C. laevigata*, however the compound was below instrument LOQ. The compound was not detected in *C. okanaganensis* or *C. douglasii* by either analytical platform. An analysis of the same biological samples by HPLC-DAD quantified the mean concentrations of vitexin in the leaf extracts of two European species, *C. monogyna* (0.16 mg/g leaf) and *C. laevigata* (0.95 mg/g leaf).

Isovitexin (Figure 1.6; Chapter 1.4) has been previously described in hawthorn (Prinz et al., 2007; Ringl et al., 2007) and other plant genera (Fu et al., 2007; Peng et al., 2008). Some therapeutic effects have also been attributed to its presence (Lin et al., 2002; Peng et al., 2008). Concentrations of isovitexin were compared among four *Crataegus* spp., as quantified from HPLC-DAD data (Table 3.2; Figure 3.3). Isovitexin was not detected by NMR, however the compound was detected in trace amounts by HPLC-DAD in *C. monogyna* (0.43 mg/g leaf), *C. laevigata* (0.37 mg/g leaf), and *C. douglasii* (0.08 mg/g leaf).

The quantification results of the vitexins presented here suggest that NMR is a poor choice for authentication or quality assessment purposes of *C. monogyna* and *C. laevigata* since this class of compounds is often used for this purpose (PCPRC, 2010; USP, 2012) and the detection and quantification of vitexin and isovitexin would be severely limited by the
Figure 3.2: Concentration of vitexin in leaf extracts of four *Crataegus* spp. as measured by NMR and HPLC-DAD (n=9 per species, per method [except *C. laevigata* n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material. Upper whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol (\(\oplus\)). Median concentration is represented by horizontal line. Boxplot not shown for species as measured by NMR since concentrations are below instrument LOQ.
Figure 3.3: Concentration of isovitexin in leaf extracts of four *Crataegus* spp., as measured by NMR and HPLC-DAD (n=9 per species, per method [except *C. laevigata* n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material. Upper whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol ( ). Median concentration is represented by horizontal line. Asterisk (*) indicates outlier sample. Boxplot not shown for species as measured by NMR since concentrations are below instrument LOQ.
relatively poor detection limits of NMR instruments. Consequently, HPLC-based analyses may be more appropriate for this purpose.

Rutin (Figure 1.6; Chapter 1.4) is common to several genera of plants and has been detected in trace amounts in *Crataegus*. The compound has been noted for its antithrombotic activity (Jasuja et al., 2012), antioxidant potential (Metodiewa et al., 1997), and anti-inflammatory activity (Jung et al., 2007). Concentrations of rutin were compared among four *Crataegus* spp., as quantified from NMR and HPLC-DAD data (Table 3.2; Figure 3.4). There were no significant differences found between mean concentrations of the compound in comparing measurements above LOQ (*C. douglasii* and *C. okanaganensis*) from NMR and HPLC-DAD data. The mean concentrations of rutin (as determined by NMR) in the leaf extracts of two North American species, *C. douglasii* (4.8 mg/g leaf) and *C. okanaganensis* (10.1 mg/g leaf), are significantly different from each other. The compound was below instrument LOQ in *C. monogyna* and not detected in *C. laevigata*, however could be quantified by HPLC-DAD (0.36 mg/g leaf in *C. monogyna* and 0.05 mg/g leaf in *C. laevigata*).

Hyperoside (Figure 1.6; Chapter 1.4) is a flavonoid found in several genera of plants including *Crataegus* (Calzada et al., 1999; Edwards et al., 2012; Zou et al., 2004), and like other flavonoids in hawthorn, have numerous beneficial therapeutic effects (Ku et al., 2015; Mustapha et al., 2014; Wen et al., 2015). Concentrations of hyperoside were compared among four *Crataegus* spp., as quantified from NMR and HPLC-DAD data (Table 3.2; Figure 3.5). There were no significant differences detected between mean concentrations of the compound in comparing measurements from NMR and HPLC-DAD data for *C. monogyna* and *C. douglasii*. Mean concentration of hyperoside in *C. laevigata* and *C. okanaganensis*
Figure 3.4: Concentration of rutin in leaf extracts of four *Crataegus* spp. as measured by NMR and HPLC-DAD (n=9 per species, per method [except *C. laevigata* n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material. Upper whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol (⁺). Median concentration is represented by horizontal line. Boxplot not shown for *C. monogyna* and *C. laevigata* as measured by NMR since concentrations are below instrument LOQ.
Figure 3.5: Concentration of hyperoside in leaf extracts of four *Crataegus* spp. as measured by NMR and HPLC-DAD (n=9 per species, per method [except *C. laevigata* n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) or methanol/water/acid (HPLC) dry leaf extracts of ground leaf material. Upper whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol (+). Median concentration is represented by horizontal line.
determined by NMR) in the leaf extracts of *C. monogyna* (14.5 mg/g leaf) and *C. douglasii* (14.2 mg/g leaf) are not significantly different from each other, but are significantly different from both *C. laevigata* and *C. okanaganensis*. *C. laevigata* (9.4 mg/g leaf) and *C. okanaganensis* (6.4 mg/g leaf) are also significantly different from each other.

Naringenin (Figure 1.6; Chapter 1.4) is a flavonoid ubiquitous to the plant kingdom (Kennedy, 2014), however has only been found in trace concentrations in select samples of *Crataegus*—0.006 mg/g leaf of *C. microphylla* (Melikoglu et al., 2004) and 0.004 mg/g leaf of *C. monogyna* (Keser et al., 2014). Concentrations of naringenin were compared among four *Crataegus* spp., as quantified from NMR data (Table 3.2; Figure 3.6). The mean concentration of naringenin (as determined by NMR) in the leaf extracts of *C. douglasii* (2.9 mg/g leaf) is significantly higher than all other species compared, in which the compound was not detected.

Quercetin (Figure 1.6; Chapter 1.4) has been previously described in some species of *Crataegus*, however is typically found in relatively low concentrations (up to approximately 0.3 mg/g leaf) (Aladedunye et al., 2014; Hernández-Pérez et al., 2014; Žugić et al., 2014). Concentrations of quercetin were evaluated for four *Crataegus* spp., as quantified from NMR data. Signals corresponding to quercetin were only detected in the leaf extracts of *C. okanaganensis*, however the corresponding signal is far below the instrument’s LOQ, and cannot be reliably quantified.
Figure 3.6: Concentration of naringenin in leaf extracts of four *Crataegus* spp. as measured by NMR (n=9 per species [except *C. laevigata* n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) dry leaf extracts of ground leaf material. Upper whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol (+). Median concentration is represented by horizontal line. Boxplot not shown for *C. monogyna*, *C. laevigata*, and *C. okanaganensis* as measured by NMR since concentrations are below instrument LOQ.
Chlorogenic acid (Figure 1.6; Chapter 1.4) is a hydroxycinnamic acid found in a great number of plants, including *Crataegus*, and has been widely studied for its purported therapeutic benefits (Edwards et al., 2012; Zhao et al., 2012). Concentrations of chlorogenic acid were compared among four *Crataegus* spp., as quantified from NMR data (Table 3.2; Figure 3.7). The mean concentrations of chlorogenic acid (as determined by NMR) in the leaf extracts of *C. monogyna* (17.8 mg/g leaf) and *C. douglasii* (16.6 mg/g leaf) are not significantly different from each other, but are significantly different from both *C. laevigata* and *C. okanaganensis*. *C. laevigata* (10.1 mg/g leaf) and *C. okanaganensis* (24.3 mg/g leaf) are also significantly different from each other in this respect.

NMR spectroscopy and HPLC-DAD exhibit a great deal of consistency in measurements of flavonoid compounds of the *Crataegus* spp. discussed here, however some significant differences in measurement do exist. In the case of hyperoside, there were significant differences between measurements of the compound by NMR and HPLC-DAD. In *C. monogyna*, *C. douglasii*, and *C. okanaganensis* the compound was significantly less concentrated as measured by NMR, however, paradoxically the compound is significantly more concentrated as measured by NMR in *C. laevigata*. This pattern is not consistent with typical explanations related to extraction differences, such as solvent polarities (which do indeed differ here between the NMR and HPLC-DAD methods). It is possible that since the compliment of NMR samples was lower for *C. laevigata* (n=6) than all other species (n=9) examined, we do not observe a difference in hyperoside concentration between NMR and HPLC-DAD that is consistent with the other species. If all sample sizes were equal, it may be more possible to make a conclusive statement regarding this difference.
Figure 3.7: Concentration of chlorogenic acid in leaf extracts of four *Crataegus* spp. as measured by NMR (n=9 per species [except *C. laevigata* n=6 for NMR]). Concentrations are shown for deuteromethanolic (NMR) dry leaf extracts of ground leaf material. Upper whiskers represent the upper 25% of the distribution and lower whiskers represent the lower 25% of the distribution. Mean concentrations are indicated by sun cross symbol (⊕). Median concentration is represented by horizontal line.
Since conventional NMR spectroscopy does not involve any chromatographic steps, it is more likely that signal congestion will lead to issues in quantification of compounds in biological samples than with chromatographic platforms. Despite careful selection of signals that do not overlap others, it is a possibility that hidden overlapping signals were present, and if true, would lead to quantification errors in NMR data. Despite these differences, the biological differences among species exceed that of instrumental differences.

It is very apparent from the data that the relatively poor detection limits of NMR would pose a challenge for analysis/authentication of certain *Crataegus* species, such as *C. monogyna* where isovitexin is often used as a marker compound. However, since the method could be used to detect and quantify a wider range of compounds, other interesting findings could be made for compounds in relatively high concentrations, such as with the case of naringenin in *C. douglasii*—the results here suggest that this compound may serve as a good marker compound for authentication of the species.

From these data, it is not universally true that concentrations measured by NMR are more precise than that of HPLC-based analytical platforms. In some instances, there was greater variability within NMR-measured concentrations than HPLC-measured concentrations, and in other instances, the converse was true. One compound (isovitexin) was not detectable by NMR, however was quantifiable by HPLC-DAD. In other cases (vitexin, quercetin), reliable quantification of signals by NMR could not be made due to limitations of instrument detection.

An assessment of quantification results acquired by both NMR and HPLC-DAD reveals consistency in measurements of mean concentrations of vitexin-2″-O-rhamnoside and rutin (in cases where measurements were above LOQ for both instruments). As noted above,
hyperoside was an exception. Since NMR has poor detection limits in comparison to HPLC, evaluations could not be made for those compounds (vitexin, isovitexin) that could not be detected or reliably quantified by NMR.

Overall, where comparison was possible, the quantitative data acquired by NMR is in agreement with HPLC-DAD data, and although is far less sensitive, may be used to analyze a wider range of compounds above its detection limits without a need for expensive or difficult to find chemical calibration standards.

3.4.2 Implications of quantification results for Crataegus utility

To my knowledge, the current published body of scientific literature does not contain any quantitative analyses of the phytochemistry of North American Crataegus, and so the results presented here show some chemical differences between the species discussed and provide direction for future investigations of North American Crataegus.

Vitexin derivatives (vitexin, vitexin-2″-O-rhamnoside, and isovitexin) are significantly more concentrated in the leaves of European hawthorns, C. monogyna and C. laevigata, than in the North American hawthorns, C. douglasii and C. okanaganensis. Conversely, the leaves of North American hawthorns contain significantly higher concentration of rutin than those of the European hawthorns. These results suggest that the vitexin derivatives and rutin may be useful in differentiation of European hawthorns from North American hawthorns. Moreover, the results presented in the significance analysis presented in Chapter 2 reflect these data.

Interestingly, the rutin concentrations measured here are the highest concentrations ever quantified in plant materials of Crataegus. The differences between North American
and European *Crataegus* in concentrations of vitexin derivatives and rutin could lead one to hypothesize that there may be different therapeutic effects between the pairs of species since both vitexin derivatives and rutin carry a myriad of potential therapeutic effects (Jasuja et al., 2012; Jung et al., 2007; Lu et al., 2013; Metodiewa et al., 1997; Wang et al., 2015; Wei et al., 2014). This result underlines the need for further investigation of the phytochemistry and therapeutic effects of North American *Crataegus*.

All species examined here have considerable quantities of hyperoside and chlorogenic acid, however the mean concentrations of both compounds are significantly different within the European hawthorns and within the North American hawthorns. Consequently, there is potential for utility of these compounds in differentiation of these two groups of closely related species. The results of the significance analysis presented in Chapter 2 are consistent with that hypothesis.

Similar to rutin, the naringenin concentrations measured here are the highest concentrations ever quantified in plant materials of *Crataegus*. *In vivo* studies of naringenin supplementation in rats suggest beneficial health effects, but the supplementation is multiple magnitudes higher than that of concentrations previously described in *Crataegus*—approximately 1 mg/g equivalent for cholesterol reduction (Lee et al., 1999) and 10-30 mg/g equivalent for prevention of dyslipidemia, apolipoprotein B overproduction, and hyperinsulinemia (Mulvihill et al., 2009). The concentrations of naringenin in the leaf extracts of *C. douglasii* (2.9 mg/g leaf), as measured here, approach potentially therapeutic concentrations, and so future studies of therapeutic effects of *C. douglasii* may lead to promising results—however future studies would be required to draw any conclusions.
3.4.3 Summary and applications of findings

The results presented here suggest that quantitative analyses of hawthorn leaf extracts are reasonably accurate as measured by NMR spectroscopy, however with relatively poor detection limits. HPLC-DAD analysis of the same plant samples was performed as an orthogonal platform for validation of results, and overall, the measurements are consistent between the two analytical platforms where compounds were within detection limits of the NMR instrument. The European hawthorns studied accumulate higher concentrations of vitexin derivatives in leaf material than the North American hawthorns. Conversely, the North American hawthorns accumulated significantly higher concentrations of rutin in their leaves, as compared to European hawthorns. Hyperoside and chlorogenic acid are present in relatively high concentrations in plant materials of all species but do differ among them. Quercetin could not be quantified in any samples. The North American hawthorn leaves analyzed here contain the highest concentrations of rutin and naringenin ever reported for \textit{Crataegus} in the literature, to my knowledge. As emphasized throughout Chapter 3.3.1, there are instances where NMR would be an unsuitable analytical platform for authentication/quality assessment of \textit{Crataegus} due to poor sensitivity, and so HPLC is more suitable where trace/low concentration compounds are used as markers. In spite of this, we can detect and quantify a much wider range of compounds with NMR without the need for expensive or difficult-to-obtain chemical calibration standards, which in many applications, may outweigh the disadvantages. Consequently, rational thought should be put into the decision-making process of choosing an analytical platform with the detection limits and target compounds in mind.
Chapter 4: Conclusion

Little is currently known regarding the phytochemical content of North American *Crataegus* species. The phytochemical profile is highly related to potential therapeutic effects upon ingestion of a given plant material. Because of this relationship, it is important that we understand what is contained within the plant material in order to allow one to make educated decisions in evaluating potential of a given plant for future *in vitro* and *in vivo* studies on therapeutic effects. Additionally, authentication of NHPs has become current concern for the general public and so warrants further development of methods for establishing identity of botanical materials. The phytochemical profile often differs among species, so we may use these data along with appropriate statistical techniques to make predictions about sample identity—forming a basis for authentication of botanical materials.

I describe two bodies of work in this thesis with the goal of exploring methods to differentiate *Crataegus* spp. by chemical content and quantifying important flavonoids of North American *Crataegus*.

The first objective of this thesis was to explore chemometric methods for differentiation of *Crataegus* spp. and determine the compounds that were significant in differentiating species. NMR data acquisition techniques were evaluated for utility in achieving this objective. The results demonstrated that the differences between acquisition methods are small, however standard 1D $^1$H experiments has a slight overall advantage in ability to discriminate species and also has the advantage of a shorter experimental duration. The results also indicated that we may best differentiate *Crataegus* spp. by NMR-acquired data that correlates to phenolic compounds. Lastly, machine learning algorithms may be applied to chemometrics data and the results correlate with and outperform traditional data.
analysis methods, such as PLS-DA. The data presented in Chapter 2 provide framework for future multivariate analyses of Crataegus extracts by NMR spectroscopy including those that may lead to robust authentication models. Additionally, a majority of known therapeutic compounds in hawthorn are phenolics. With this in mind, combinations of significance analysis methods were used to provide lists of potential biomarkers in the phenolics spectral region that may give insight to differential therapeutic effects of hawthorn NHPs.

The second objective of this thesis was to quantify important phytochemicals (relevant to differentiation and therapeutic potential) in North American and European Crataegus spp. This objective was carried out in order to quantify flavonoids of select North American Crataegus and to evaluate the utility of NMR spectroscopy for this purpose. An orthogonal analytical platform, HPLC-DAD, is used to validate the results. The results presented in this chapter demonstrate the similarities and differences of flavonoid content among North American and European hawthorns. Similar to the biomarkers discussed in Chapter 2, the vitexin derivatives are particularly concentrated in European Crataegus and conversely, rutin is significantly more concentrated in North American Crataegus.

Comparison of quantification results by NMR to HPLC-DAD shows consistency in measurement of mean concentrations of vitexin-2″-O-rhamnoside and rutin (when comparing measurements above LOQ for both instruments). In spite of this result, some differences exist—such is the case in measurements of hyperoside, where the compound was significantly less concentrated as measured by NMR (except for one species with imbalanced replicates). Since NMR has an instrumental LOD/Q multiple magnitudes higher than HPLC, comparisons could not be made for those compounds (vitexin, isovitexin) that could not be detected or reliably quantified by NMR. Although poor detection limits proved to hinder the
application of NMR in some instances, the ability to detect a wider range of compounds without the need for expensive or difficult-to-find chemical calibration standards was a distinct advantage of the platform. This investigation underlines that future quantification studies of flavonoids using NMR must consider both its advantages and limitations—the platform allows one to learn more about the unknown chemistry without reference materials, however, at the cost of sensitivity.

Taken together, the results presented in this thesis may be used to generate reasonable hypotheses regarding the therapeutic potential of North American *Crataegus*. Since North American *Crataegus* leaf extracts contain high concentrations of rutin and low concentrations of vitexin derivatives, as compared to their European counterparts, the therapeutic potential of each plant may differ. As discussed in detail in Chapter 1.4, vitexin derivatives and rutin are both associated with anti-thrombotic function, however rutin is thought to act as a protein disulfide isomerase inhibitor which disrupts platelet aggregation and fibril formation (Jasuja et al., 2012) and vitexin derivatives are thought to act via an adenine diphosphate pathway where they disrupt a G-protein-mediated signaling pathway that leads to thrombus formation (Li et al., 2015). It is therefore possible that administration of North American *Crataegus* leaf may have different clinical implications than European *Crataegus* leaf since the pathways that the characteristic compounds act upon to achieve anti-thrombotic activity differs. Additionally, clinical implications may differ here due to other reported activities, such as anti-inflammatory properties of rutin (Guardia et al., 2001; Jung et al., 2007) and anti-apoptotic activity of vitexin derivatives (Wei et al., 2014). These hypotheses would need to be tested with *in vitro* and *in vivo* models to make reasonable statements about any therapeutic effect of North American *Crataegus*. Additionally, the
methods for data acquisition and exploration evaluated in this thesis may be used to guide the creation of future models for authentication of hawthorn plant materials and NHPs.
References


Bakkali, F., Averbeck, S., Averbeck, D., Idaomar, M., 2008. Biological effects of essential oils – A review. *Food and Chemical Toxicology* 46, 446-475.


CBOL, 2015. What Is DNA Barcoding?


Fu-Cha, D., 1961. YanJing SuiShi Ji. Beijing Publishing Group, Beijing, CN.


Linnaeus, C., 1753. Species Plantarum. Laurentius Salvius, Stockholm, SE.

Linnaeus, C., 1754. Genera Plantarum. Laurentius Salvius, Stockholm, SE.


NNHPD, 2015a. Health Canada Licensed Natural Health Products Database.


NNHPD, 2015d. Quality of Natural Health Products Guide.


Rogers, D. J., Buechel, E., 1980. Lakota Names and Traditional Uses of Native Plants by Sicangu (Brule) People in the Rosebud Area. Rosebud Educational Society, St. Francis, SD.


139


Appendices

Appendix A: Optimization of Crataegus leaf extraction for quantitative analysis by nuclear magnetic resonance spectroscopy

A deuterated solvent suitable for extraction of flavonoids in plant material was selected. Research from previous students and the scientific literature suggest that methanol (in this case, the deuterated equivalent—methanol-d4) is suitable for this purpose (Anokwuru et al., 2011; De Sotillo et al., 1994; Falleh et al., 2012; Goli et al., 2005; Jakopi et al., 2009; Jaroszynska, 2003). Extractions were performed sequentially (in three aliquots: 0.5 mL, 0.5 mL, 1.0 mL) with a constant total solvent volume of 2.0 mL. Several factors were identified as being most likely to affect extraction efficiency of flavonoids in Crataegus leaf including: test portion weight, extraction device, extraction time, and radiation effects. A half-factorial was designed (Table A-1) and results analyzed (Figure A-1) in Minitab 17 (State College, PA, USA) for initial optimization.

Table A-1: Experimental schematic for half-factorial analysis of four factors on extraction efficiency

<table>
<thead>
<tr>
<th>Run</th>
<th>Weight of test portion (mg)</th>
<th>Water bath at 60°C or sonication</th>
<th>Extraction Time (min)</th>
<th>Radiation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>water bath</td>
<td>15</td>
<td>dark</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>water bath</td>
<td>60</td>
<td>dark</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>sonicate</td>
<td>15</td>
<td>light</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>sonicate</td>
<td>60</td>
<td>light</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>water bath</td>
<td>15</td>
<td>light</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>water bath</td>
<td>60</td>
<td>light</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>sonicate</td>
<td>15</td>
<td>dark</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>sonicate</td>
<td>60</td>
<td>dark</td>
</tr>
</tbody>
</table>
From the results of the half-factorial experiment, samples were kept dark (although this factor was vary minor and likely not significant) and extracted in a 60°C water bath in subsequent experiments. Sample mass and extraction time were further optimized and the results were plotted (Figure A-2; A-3).

**Figure A-1: Main effects plot resulting from half-factorial experiment for extraction optimization.** Results expressed as mean concentration of chlorogenic acid extracted from leaf samples. Highest points correspond to factor variable with greatest level of extraction. Largest influences correlate to steepest slopes observed.

From the results of the half-factorial experiment, samples were kept dark (although this factor was vary minor and likely not significant) and extracted in a 60°C water bath in subsequent experiments. Sample mass and extraction time were further optimized and the results were plotted (Figure A-2; A-3).

**Figure A-2: Results of time dependent extraction efficiency analysis.** Results expressed as mean concentration of chlorogenic acid (A) or total vitexin (B) extracted from leaf samples. Minimum extraction duration is 15 minutes and maximum duration is 8 hours. Test samples derived from bulk sample of *C. monogyna* leaf (542a).
An extraction duration of 2 hours was determined sufficient for complete extraction of chlorogenic acid, however vitexin derivative extraction efficiency plateaued after 6 hours. An extraction duration of 8 hours was selected from these results (Figure A-2) for the quantitative analyses in Chapter 3. Sample quantity was limited and solvent saturation was a concern so the minimum sample quantity to achieve maximum extraction efficiency was determined. A sample mass of 100 mg was selected from these results (Figure A-3) for the quantitative analyses in Chapter 3. Using the overall results described here, the method as described in Chapter 3.2.3 was developed. To ensure complete extraction of leaf materials, three post-extraction samples were vortex mixed for 30 seconds and re-extracted overnight in a fresh 1 mL aliquot of deuterated methanol. No detectable amount of flavonoid compounds of interest was found in these samples (data not shown).
Appendix B: Selection of $T_1$ relaxation time for quantitative analysis by nuclear magnetic resonance spectroscopy

An inversion recovery experiment, also known as a $T_1$ relaxation experiment, was performed to determine the delay time sufficient for accurate integration of signals that correspond to compounds of interest in the hawthorn leaf extracts. The results of the experiment in the relevant region of the spectra (approximately $\delta$ 6-8 ppm) is shown in Figure B-1.

![Figure B-1: Results of inversion recovery experiment for $T_1$ relaxation delay determination. Region of spectrum shown for relevant signals to quantitative analysis in Chapter 3 ($\delta$ 6-8 ppm). Test sample derived from sample of $C$. monogyna leaf (491).]
After a delay time of 8 seconds had elapsed, all signals have reached or exceeded $T_1$ relaxation (as determined by recovery of signal into positive Y space). As described in Chapter 3.2.4, a relaxation delay of 5 times the $T_1$ relaxation time for the slowest-relaxing signal should be selected to ensure all proton environments of interest are fully relaxed prior to acquisition of subsequent transients (Bharti and Roy, 2012; Malz and Jancke, 2005)—therefore, the relaxation delay for quantification experiments described in Chapter 3 is 40 seconds.
Appendix C: Signals used for quantitative analysis by nuclear magnetic resonance spectroscopy

Table C-1 outlines signals used for quantification of compounds of interest, chemical compounds and protons associated with these signals, and example spectra depicting the signals of interest.

**Table C-1: Compounds of interest in *Crataegus* and their associated signals.** Compound and associated proton environments that correspond to signal for quantification shown in left column. Example spectra shown in right column with chemical standard (top) and hawthorn extract sample (bottom).

<table>
<thead>
<tr>
<th>Compounds of Interest</th>
<th>Chemical Environments</th>
<th>Example Spectra</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vitexin-2″-O-rhamnoside</strong></td>
<td></td>
<td><img src="image" alt="Vitexin-2″-O-rhamnoside" /></td>
</tr>
<tr>
<td>Vitexin</td>
<td>δ 7.97 ppm</td>
<td><img src="image" alt="Vitexin" /></td>
</tr>
</tbody>
</table>

Note: Left side of peak integrated only to prevent overlap (representing 1 eq. H)

Note: Right-most side of peak integrated only to prevent overlap (representing 1 eq. H)
Table C-1: Compounds of interest in *Crataegus* and their associated signals. Compound and associated proton environments that correspond to signal for quantification shown in left column. Example spectra shown in right column with chemical standard (top) and hawthorn extract sample (bottom).

<table>
<thead>
<tr>
<th>Compound</th>
<th>Signal</th>
<th>Spectrum 1</th>
<th>Spectrum 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Isovitexin</strong></td>
<td>$\delta 6.50$ ppm</td>
<td><img src="image1" alt="Isovitexin structure" /></td>
<td><img src="image2" alt="Isovitexin spectrum" /></td>
</tr>
<tr>
<td><strong>Rutin</strong></td>
<td>$\delta 7.63$ ppm</td>
<td><img src="image3" alt="Rutin structure" /></td>
<td><img src="image4" alt="Rutin spectrum" /></td>
</tr>
</tbody>
</table>
Table C-1: Compounds of interest in *Crataegus* and their associated signals. Compound and associated proton environments that correspond to signal for quantification shown in left column. Example spectra shown in right column with chemical standard (top) and hawthorn extract sample (bottom).

<table>
<thead>
<tr>
<th>Compound</th>
<th>Associated Proton Environments</th>
<th>Example Spectra</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hyperoside</strong></td>
<td>δ 7.81 ppm</td>
<td><img src="image" alt="Spectra" /></td>
</tr>
<tr>
<td></td>
<td>Galactose δ 7.27 ppm</td>
<td></td>
</tr>
</tbody>
</table>

**Naringenin**

<table>
<thead>
<tr>
<th>Associated Proton Environments</th>
<th>Example Spectra</th>
</tr>
</thead>
<tbody>
<tr>
<td>δ 7.27 ppm</td>
<td><img src="image" alt="Spectra" /></td>
</tr>
<tr>
<td>δ 7.27 ppm</td>
<td></td>
</tr>
</tbody>
</table>

Note: Peak shown is for hawthorn extract sample only. Right-most side of peak integrated only to prevent overlap (representing 1 eq. H). No reference spectra available. Peak determined from literature spectra (d, J=8.5 Hz) (Azimova and Vinogradova, 2013; Estork et al., 2014; Maltese et al., 2009)
Table C-1: Compounds of interest in *Crataegus* and their associated signals. Compound and associated proton environments that correspond to signal for quantification shown in left column. Example spectra shown in right column with chemical standard (top) and hawthorn extract sample (bottom).

<table>
<thead>
<tr>
<th>Compound</th>
<th>δ ppm</th>
<th>Example Spectra</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quercetin</td>
<td>7.71</td>
<td><img src="image" alt="Quercetin Peak" /></td>
</tr>
<tr>
<td>Chlorogenic acid</td>
<td>7.02</td>
<td><img src="image" alt="Chlorogenic Acid Peak" /></td>
</tr>
</tbody>
</table>

Note: Peak shown is for compound reference standard only. Compound not detected in hawthorn extract samples.
Appendix D: Standard curves used for quantitative analysis by high performance liquid chromatography – diode array detection platform

Figure D-1 displays standard curves of hyperoside, vitexin-2″-O-rhamnoside, vitexin, isovitexin, and rutin, as measured by HPLC-DAD. Linearity was also evaluated for vitexin-2″-O-rhamnoside and hyperoside up to 435 ppm and 530 ppm, respectively, however data is not shown in Figure D-1 (y = 4.1314x + 9.6032 and R² = 0.99983 for vitexin-2″-O-rhamnoside 0.9-435 ppm; y = 3.6135x + 7.1368 and R² = 0.99983 for hyperoside 1.1-530 ppm).

![Figure D-1: Standard curve of five flavonoids as detected by HPLC-DAD. Vitexin concentration range: 1.1-210 ppm. Vitexin-2″-O-rhamnoside concentration range: 0.9-174 ppm. Isovitexin concentration range: 1.0-206 ppm. Hyperoside concentration range: 1.1-212 ppm. Rutin concentration range: 1.0-198 ppm. Data shown here used in HPLC-DAD quantification of flavonoids in Crataegus leaf extract samples throughout Chapter 3.](image-url)
Appendix E: Summary of significance analysis of phytochemical differences among four *Crataegus* spp.

ANOVA with a Tukey’s honest significant difference post-hoc test was used to detect differences in compound concentrations (as measured by $^1$H NMR) among four *Crataegus* spp. at a 95% confidence level. The differences of means from the Tukey post-hoc results were plotted for each compound in order to visualize these data (Figure E-1-E-7).

**Figure E-1:** Tukey simultaneous 95% confidence intervals for vitexin-2″-O-rhamnoside content in leaf extracts of four *Crataegus* spp. Differences along X-axis indicate differences of means for the compound evaluated here. If an interval between two species does not contain zero, the mean concentrations of the compound are significantly different. Concentrations measured by $^1$H NMR.
Figure E-2: Tukey simultaneous 95% confidence intervals for vitexin content in leaf extracts of four *Crataegus* spp. Differences along X-axis indicate differences of means for the compound evaluated here. If an interval between two species does not contain zero, the mean concentrations of the compound are significantly different. Concentrations measured by $^1$H NMR.

Figure E-3: Tukey simultaneous 95% confidence intervals for isovitexin content in leaf extracts of four *Crataegus* spp. Differences along X-axis indicate differences of means for the compound evaluated here. If an interval between two species does not contain zero, the mean concentrations of the compound are significantly different. NOTE: concentrations used in this analysis are below instrument LOQ. Concentrations measured by $^1$H NMR.
Figure E-4: Tukey simultaneous 95% confidence intervals for rutin content in leaf extracts of four \textit{Crataegus} spp. Differences along X-axis indicate differences of means for the compound evaluated here. If an interval between two species does not contain zero, the mean concentrations of the compound are significantly different. Concentrations measured by $^1$H NMR.

Figure E-5: Tukey simultaneous 95% confidence intervals for hyperoside content in leaf extracts of four \textit{Crataegus} spp. Differences along X-axis indicate differences of means for the compound evaluated here. If an interval between two species does not contain zero, the mean concentrations of the compound are significantly different. Concentrations measured by $^1$H NMR.
Figure E-6: Tukey simultaneous 95% confidence intervals for chlorogenic acid content in leaf extracts of four Crataegus spp. Differences along X-axis indicate differences of means for the compound evaluated here. If an interval between two species does not contain zero, the mean concentrations of the compound are significantly different. Concentrations measured by $^1$H NMR.
Appendix F: Standard curve for determination of NMR instrument detection limits and linearity

Using the method described in Chapter 3.2.6, a standard curve was created from a 10-point standard dilution of chlorogenic acid in methanol-d4 (Figure F-1).

**Figure F-1:** Standard curve of chlorogenic acid standard as detected by NMR. Detector response of chlorogenic acid signal (7.02 ppm, d) normalized to DSS internal standard signal at -0.018 ppm. Linearity described by $R^2$. Standard error of the estimate, $S_{yx}$, is also noted.