Determination of a robust metabolic barcoding model for chemotaxonomy in Aizoaceae species: Expanding morphological and genetic understanding.

Amelia Hilgart

A thesis presented for the degree of
Doctor of Philosophy

Department of Chemistry
Science Faculty
University of Cape Town
South Africa
January, 2016

Abstract

The use of metabolic fingerprints as taxonomic markers is becoming more common. Many studies have found that by comparing the vast metabolic fingerprints of closely related species to each other, secondary metabolites tend to be unique to the samples of individual species and are identified in clustering algorithms as the variables responsible for species-specific clustering. A holistic approach to metabolic fingerprinting was thus employed to assess the stability of various metabolomic markers and finally to distinguish taxonomically difficult Aizoaceae species.

Many secondary metabolites are not constitutively produced. Because at least some Aizoaceae species facultatively use crassulacean acid metabolism (CAM), there was a potentially interesting molecular switch that could be monitored for transitions in metabolic fingerprints. In order to contextualise the changes in carbon uptake, 20 different climate, nutrient, physiological, and other variables were monitored over the course of 12 months to build up a store of species-specific information to use in model optimisation across 5 Aizoaceae species (*Galenia africana, Aridaria noctiflora, Carpobrotus edulis*, *Ruschia robusta*, and *Tetragonia fruticosa*) using two Crassulaceae species as CAM controls (*Cotyledon orbiculata* and *Tylecodon wallichii*).

Metabolic fingerprints of the leaves of various Aizoaceae species were generated using LC/TOFMS, following which Principal Components Analysis (PCA) was used to identify the LC-MS ions which distinguished the species from each other, or in statistical terms, were informative. Once isolated, this subset of informative data was established as metabolic barcodes for the identification of the study species. A machine learning algorithm, Random Forest, was used to build a classification model based on the metabolic barcodes which was then trained on various trends from the factors monitored over the year. The use of these trends in the development of a classification model based on metabolic barcodes resulted in a highly robust classification model for species identification. Clustering analysis of a subset of ions which corresponded to compounds previously isolated from Aizoaceae species did not show species-specific clustering and was inevitably biased by compounds from species with a greater number of studies focusing on compound isolation.

Ideally, this model should be expanded to include other species from the Aizoaceae family to further check robustness of the model. Application of this model to these and other species could facilitate not only species identification and distribution, but also the identification of novel chemical constructs associated with particular species.

Declaration

I, Amelia Annie Hilgart, hereby declare that the work on which this thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I authorise the University to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature:

Date: October 8, 2015

Acknowledgements

I would like to thank the University of Cape Town's Stable Light Isotope Laboratory for their technical support, Dr. Kurt van der Heyden (University of Cape Town) for time on his computer cluster, and Dr. Sam Feagley (Texas Agriculture and Mining University) for his assistance in the interpretation of the soil data. I would also like to thank Associate Professor David Gammon and Professor Jill Farrant for teaching me the true meaning of good supervision. I would also like to thank Keren Cooper for her unerring support and honesty when I needed it the most as well as her extraordinary assistance in generating some of my more complex figures.

More personally, I would like to thank my mother who taught me how to love science, my father who taught me attention to detail and persistence, and my husband who convinced me to learn computer programming and fed me while I did it.

Contents

List of Figures

List of Tables

Table of Abbreviations

Abbreviation	Definition
2D	Two dimensional
AWC	Average water content
BH	Bolus Herbarium
C3	Metabolic pathway for carbon fixation which converts carbon dioxide and ribulose bisphosphate (RuBP, a 5-carbon sugar) into 3-phosphoglycerate
CAM	Crassulacean acid metabolism
DNA	Deoxyribonucleic acid
EI	Electron ionisation
EIC	Extracted ion chromatogram
ESI	Electrospray ioniser
GC	Gas chromatography
GC-MS	Gas chromatography coupled to mass spectrometry
GMD	Golm Metabolome Database
HCA	Hierarchical clustering
HPLC	High performance liquid chromatography
ICP-MS	Inductively coupled plasma-mass spectrometry
KMG	Kimberley McGregor Museum Herbarium
LC	Liquid chromatography
LC-TOF-MS	Liquid chromatography coupled to a time-of-flight mass spectrometry analyser
LC-MS	Liquid chromatography coupled to a mass spectrometry analyser
LC-MS/MS	Liquid chromatography coupled to a tandem mass spectrometer
MS	Mass Spectrometry
MSTFA	N-Methyl-N-(trimethylsilyl) triuoroacetamide
NIST	National Institute of Standards and Technology
NMR	Nuclear magnetic resonance spectroscopy

NPC	Natural products chemistry
OOB	Out-of-bag error rate
PAR	Photosynthetically-active solar radiation
PC	Principal component
PCA	Principal components analysis
PEPC	Phosphoenolpyruvate carboxylase
PLS-DA	Partial least squares discriminant analysis
RAM	Random-access memory
RubisCO	Ribulose-1,5-bisphosphate carboxylase/oxygenase
SANBI	South African National Biodiversity Institute
SVD	Singular value decomposition
TIC	Total ion chromatogram
UCT	University of Cape Town
UV	Ultraviolet radiation
UV (scaling)	Unit variance scaling
UV-Vis	Ultraviolet to visible wavelengths

Chapter 1

Introduction

1.1 Using chemical profiling to better understand plant genetics

South Africa has an extraordinary floral diversity and extensive modern ethnobotanical practices, and with these a paramount need for botanical conservation and taxonomic assessment of its plant species (for a comprehensive review of these issues, see Nortje (2011); Dyubeni and Buwa (2012); Semenya et al. (2012); Wheat (2014)). While modern conservation and taxonomic assessments are based on morphological and DNA sequence data, the presence of unique metabolites is an accepted method of taxonomic identification, and is known as chemotaxonomy. Further, the use of metabolic fingerprints as chemotaxonomic markers is becoming more prevalent with fairly recent improvements in high-throughput compound separation and identification techniques.

Most studies employing a chemotaxonomic approach consider only a small subset of compounds such as those found in an in-house analytical standards library or those which have shown specific distributions between species in the past when generating metabolic fingerprints (a review of these follows). While levels of these compounds measured in various ways often show species specificity, very few authors consider the stability of those concentrations over time when generating metabolic fingerprints. Concentrations of specific compounds have been used for quality control in the industrial preparation of herbal medicine (Govindaraghavan et al., 2012), but species are still mostly distinguished using DNA markers.

In the following thesis, the use of LC-MS metabolic fingerprint data is explored as a way of creating an inexpensive, unique and highly specific method for the identification of closely related species. The intensity of the analytical signatures which were most important in species distinction were used to assess the stability of species identifications across time as compared to changes in various climate, nutrient, and biological factors. Once it was established that the subset of important analytical signatures were consistent over time, a classification model was generated to classify samples based on reduced metabolic fingerprints or "metabolic barcodes".

Considering these results, it appears reasonable that a taxonomic assessment of species based on metabolic barcodes could be used to assess a species' conservation status and distribution using the described method. Further, compounds identified by clustering algorithms of metabolic fingerprints tend to be secondary metabolites. Thus employing the method developed in this thesis could ultimately allow us a better understanding of secondary metabolite diversity and distribution across evolutionary lineages if employed on a greater scale.

This study focuses on the Aizoaceae family which is particularly species rich in South Africa, but is also a relatively young taxonomic family (Klak et al., 2003). Its recent divergence makes typical DNA sequencing techniques less viable as there has not been sufficient time for genetic divergence at the DNA sites typically tested (Klak et al., 2007).

1.2 Plant taxonomy

The central unit of taxonomy is the species and the pinnacle of taxonomy is associating a specific unequivocal name with a given species by which it can be classed (Padial et al., 2010). Traditional plant taxonomy bases the identity of a species on a variety of features, firstly from the morphology of its flowers and fruits and then from the greater plant. While the methods of this analysis have changed slightly, this has been the basis of taxonomy for over 250 years since Linnaeus introduced the binomial nomenclature system in 1753 (Padial et al., 2010).

Current taxonomy generally combines morphological characterisations of the past with phylogenetic markers in order to differentiate the relatedness of species. While total genome analysis is becoming more and more common for plants (Soltis et al., 2013), it is much more common to look at specific regions of DNA which are easier to isolate and significantly cheaper to work on. This practice has been formalized in the characterization of DNA barcodes where consistent pieces of DNA are considered over many species which allows for rapid species assessments and identification (Mankga et al., 2013) as well as greater evolutionary understanding (Goldstein and Desalle, 2010).

However, in cases where the variation between species is less than the variation within a species for these specific markers, this is no longer a viable option. As mentioned above, the Aizoaceae family represents such a case. In this family there has been rapid speciation over a short and recent evolutionary time frame, thus, to distinguish species which are morphologically quite different from each other many DNA regions have to be examined (Klak et al., 2003, 2007, 2013).

Phylogenetics and morphology are also used in systematics to determine evolutionary relationships between species, based on the postulate that the more closely related two species are, the more similar they will be at morphological and genetic levels. The use of unique chemical features aids this process as it adds additional phenotypic information for a species to generated models of speciation as was done in Incerti et al. (2013).

1.3 Secondary metabolism in plants

Plants are sessile and their defensive mechanisms are dependent upon an innate response both to biotic and abiotic factors (Jones and Dangl, 2006). These defence mechanisms come in many forms and may include physical barriers, such as spines or thorns, structural adaptations such as elevated silica incorporation, or chemical defences such as anti-palatability complexes and toxins in the form of secondary metabolites (Rasmann and Agrawal, 2009).

In plants, secondary metabolites are defined as "compounds produced by plants that are not directly essential for basic photosynthetic or respiratory metabolism..." (Theis and Lerdau, 2003). While the inability to produce secondary metabolites may not result in the death of a plant, it may impede overall fitness in the form of reduced fecundity and/or defensive capability (D'Auria and Gershenzon, 2005; Field et al., 2006).

As is shown in the Figure 1.1, secondary metabolite classes are synthesised via specific pathways in primary metabolism. It is important to note that because secondary metabolites are derived from the products of primary metabolism, it is often energetically quite expensive to produce them (Theis and Lerdau, 2003). Thus, secondary metabolites are often not constitutively produced over time, which means that not only are their concentrations not constant, but that their presence is not necessarily constant (Bourgaud et al., 2001).

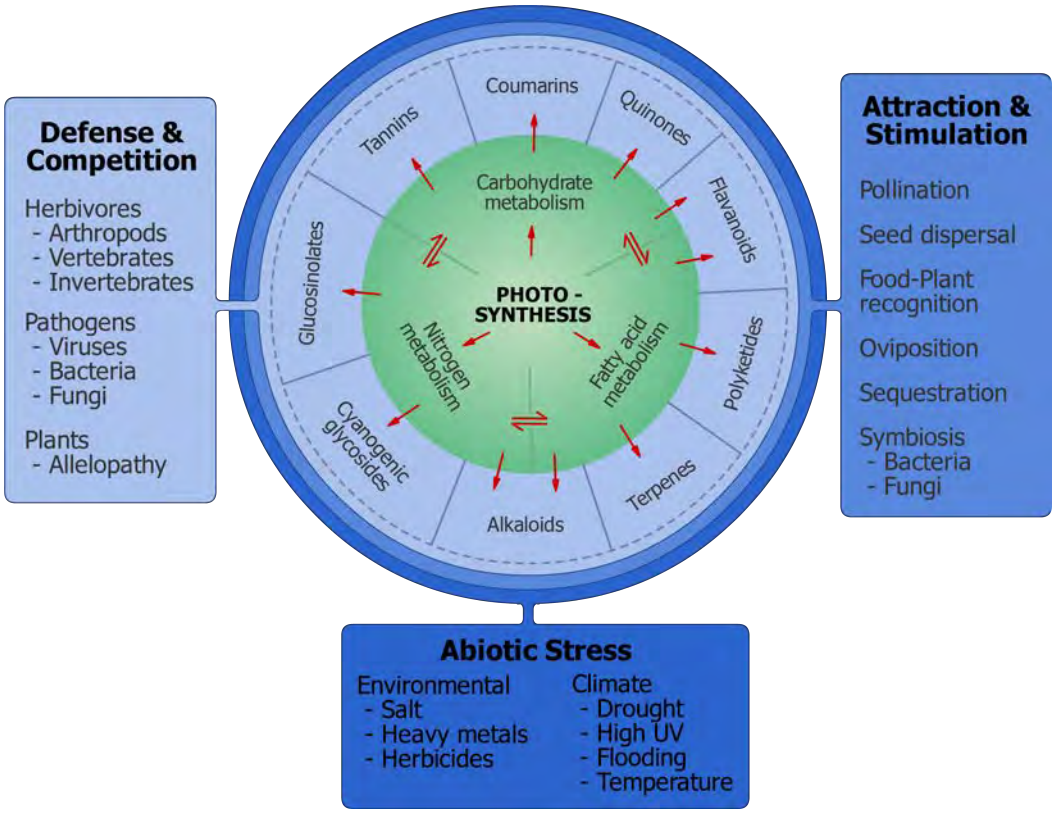


Figure 1.1: **Secondary metabolite production.** Figure adapted from Hartmann (2007).

To date, researchers have discovered over 200,000 plant secondary metabolites, although predictions suggest that this is only a fraction of what exists globally (Santos Pimenta et al., 2013). A large part of this diversity is attributed to the fact that very small changes in the genetic code for enzymes has led to large variations in secondary metabolite chemistry over time (Lewinsohn and Gijzen, 2009). Plants also show a high degree of enzyme promiscuity which allows for exceptional metabolic flexibility (Moore et al., 2013) which again leads to a high degree of diversity within secondary metabolite classes. As an example, analysis of the

Arabidopsis thaliana genome revealed that 25% of its genes are utilised in secondary metabolite production, indicating a significant evolutionary imperative in this species to manufacture secondary metabolites (Field et al., 2006).

The more closely related two species are, the greater proportion of total genetic material they share. While there are many instances of convergent evolution leading to the production of the same metabolite in completely unrelated species, it is far more common to see secondary metabolic pathways arise along directed evolutionary lineages (Pichersky and Gang, 2000).

1.4 Natural products chemistry and chemotaxonomy

Natural products are loosely defined as chemical products from organisms. This may include an entire organism, an extract of an organism, a single compound from an organism, or anything in between although it generally refers to a single biosynthesized compound with a molecular weight of less than 2,000 amu (as defined inSarker and Nahar (2012)).

Most reviews of NPC state that it began early in human history. This is evidenced by the appearance of the first known written record on a clay tablet from Mesopotamia dating from BC 2600 which describes, among other plant-derived medicines, the use of licorice and poppy capsule latex (Borchardt, 2002). The first extraction of a phytochemical as a single entity is credited to a German pharmacist, Friedrich Wilhelm Sertürner (1805), who isolated morphine from poppy capsule latex (as reviewed in Yun et al. (2012); Ji et al. (2009); Li and Vederas (2009)).

With the ability to isolate and characterise single compounds, pharmaceutical companies then came to prefer the purified compounds for the use as ingredients in medicines rather than crude extracts, as this allowed for more rigorous control of the formulations and monitoring of structure-activity relationships (Ji et al., 2009; Newman et al., 2000). The elucidation of the structures of these compounds, facilitated by significant advances in spectroscopy, allowed scientists to prepared synthetic analogues in order to improve upon the medical efficacy of these naturally occurring scaffolds (Ji et al., 2009; Newman et al., 2000). This has strongly influenced the modern NPC paradigm which is based around the isolation of single or a few biologically active small molecule secondary metabolites from an organism.

There are many advanced chromatographic and other separation techniques that are commonly utilised to expedite the separation process. These include various kinds of advanced liquid chromatography systems that can work on milligram (semi-preparatory HPLC) to gram scales (advanced solvent extraction systems) across ranges of polarity volatility, and solubility. For an extensive review, see Sarker and Nahar (2012).

Once it has been established that a plant extract contains a chemical entity of interest, the starting point in understanding its production is its isolation. When the chemistry of the molecule is unique and there is little information on the plant species or its close relatives as to what types of secondary metabolites are common along its genetic lineage, untargeted NPC techniques are typically employed to elucidate its chemical identity.

The use of NPC in Chemotaxonomy is not a novel concept. It was promoted strongly by Greshoff (1909), "Since plants are no longer classified according to a single character... It appears clear that chemistry and botany should co-operate in the study of the plant world..., so that one might demand that every accurate description of a new genus or of a new species should be accompanied by a short chemical description of the plant." In the text, Greshoff describes the many different species he worked on while visiting collaborators at the Royal Botanical Gardens in Kew and the struggle involved in chemical identification. Before the invention of DNA sequencing technology, determining the presence of chemical classes such as tannins, alkaloids, hydrocyanic acid, and saponins was a more viable option. He admits, however, that the isolation and identification of specific molecules was the crux of what needed to be done, but with the limitations of the technology of the time even he admitted that efforts would have to be targeted in order to be effective.

There are indeed many examples in plant taxonomy of plants being named after their chemical constituents. The name of the genus *Oxalis* from the family Oxalidaceae comes from the elevated oxalic acid content of its species. The Solanaceae family has a genus *Atropa* known for the high concentration of the tropane alkaloid atropine amongst all of its species.

Today, more traditional chemotaxonomic approaches concentrate on one or a few known secondary metabolites that have been isolated and identified within genetic lineages and use them more to identify species than to define new or resolve taxa. Recent examples of this approach include that of Liu et al. (2013) where various phenolics were used to identify plant species and subspecies in traditional Chinese medicine and those of Zafar et al. (2010) and Ahmad et al. (2010) who have used various flavonoids to determine misidentified plant species in markets dealing in traditional remedies in the Middle East.

1.5 Metabolomics and natural products chemistry

A bridge has formed between the chemistry sub-discipline, NPC, and the molecular biology sub-discipline, metabolomics, which offers new approaches to address rapid compound identification of molecules from biological organisms. Generally, they are distinguished by the extent to which unknown compounds are identified, where metabolomics seeks to determine unique chemical signals towards a specific biological function and NPC seeks to isolate novel compounds.

1.5.1 The kinds of questions metabolomics answers about plants

The field of metabolomics explores the extractable small molecule metabolites of an organism (Sarker and Nahar, 2012; Dieterle et al., 2011). Metabolomics analyses are generally carried out by one of two strategies, targeted or untargeted. In Table 1.1 these strategies are further detailed as described by Goodacre et al. (2004) and reviewed in Ernst et al. (2014) and Monteiro et al. (2013).

Table 1.1: **Strategies used in plant metabolomics experiments.** The classification scheme is taken from Goodacre et al. (2004) and describes the experiments that most commonly are applied in plant research.

Metabolomics strategy	Analysis	Definition	Example
Targeted metabolomics approaches	Metabolite target analysis	Targets specific metabolites for identification and quantification	Determination of the concentration of specific primary metabolites which would be affected by the introduction of a toxin
	Metabolite profiling	Targets and quantifies metabolites of specific chemical class or the metabolites of a specific metabolic pathway	Flavanoid analysis, amino acid analysis, the metabolites of the tricarboxylic acid cycle
Untargeted metabolomics approaches	Metabolic fingerprinting	Classification of samples based specific chemical signals which distinguish sample classes from each other	Defining specific chemical signatures that are unique between different plant species
	Metabonomics	Measures the general metabolic responses of an organism to specific perturbations such as disease, a toxin, or genetic modification	The general response of a plant species to a test pesticide
	Metabolomics	A comprehensive analysis of as many metabolites as possible under a given set of conditions	Profiling all of the metabolites of a plant species that is desiccation tolerant to determine its metabolic strategies for tolerance

Each of these regimes serves to answer different questions about the state of the biochemistry of an organism. As such, each requires specific analytical approaches and statistical methods.

1.5.1.1 Targeted Approaches

Generally, targeted metabolomics analyses are used to investigate the concentration and range of specific metabolites within and between samples. Targeted analyses are considered older approaches as they were commonly used before "metabolomics" became a discipline (Monteiro et al., 2013). They are also at least partially dependent on the use of analytical reference standards for the identification of specific metabolites. Targeted studies are approached in two ways, the first focuses on specific metabolites for which reference standards are available such as in a profiling analysis of common secondary metabolites across plant samples (*metabolite target analysis*). The second approach is to focus around classes of metabolites which have similar extraction methods such as flavanoids or were derived from the same metabolic pathways (*metabolite profiling*).

Before recent advances in analytical techniques, the identification of unknown compounds in a high-throughput manner was quite challenging. Because targeted studies follow known chemistry, various properties of additional metabolites of interest or of unknown metabolites can be extrapolated. This process is common in chemotaxonomic studies, as discussed above, where specific metabolites have been identified in specific genetic lineages.

1.5.1.2 Untargeted Approaches

Untargeted analyses determine the overall chemical state of an organism across many classes of chemicals. Due to the types of questions being asked and the broad scope, these analyses inevitably deal with a significant proportion of unknown metabolites although they are typically guided by analytical standards to a greater or lesser extent. *Metabolomics* as an analysis serves to identify as many compounds as possible within a sample, such as analysing the chemical composition of a biofilm. *Metabonomics* compares metabolites between an experimental class and a control class to try to determine unique chemical features within the experimental class. This type of analysis is common in medical studies where researches are trying to identify metabolites as disease biomarkers. Finally, there is *metabolic fingerprinting* which is used to identify metabolites which distinguish sample classes from each other. Metabolic fingerprinting is the untargeted metabolomics strategy most commonly employed in chemotaxonomy to distinguish species from each other.

While targeted metabolomics has well defined methods in terms of compound identification, in many cases untargeted metabolomics is still in its infancy as a strategy and faces many challenges in this regard. The problems stem from difficulties in identifying unknown compounds; firstly, because metabolomic databases are inadequately annotated resulting in low numbers of compounds being easily identifiable without the use of analytical standards (Matsuda et al., 2009), and secondly, the statistical tools necessary to comprehend data on such a large scale are still being refined (Ernst et al., 2014; Khatri et al., 2012; Karsai and Kampis, 2010; Prill et al., 2010).

Untargeted metabolomics pushes at the boundaries of known biochemical space to understand global metabolic change. In the context of disease or stress in an organism, or in chemotaxonomic studies such as this one, multivariate statistics help to reduce the pool of metabolites from thousands to a smaller and highly specific subset of metabolites which are significant to one samples' class or another, without the preconception of what one would normally expect to see. Untargeted metabolomics approaches will always be limited by the fact that there is no one way to extract the global metabolome as metabolites show large variations in polarity and solubility (van den Ouweland and Kema, 2012; Dieterle et al., 2011; Allwood and Goodacre, 2010).

The lack of database annotations presents further difficulties, although many metabolites are primary metabolites that have been well characterised chemically and can be identified. In cases where there are relatively few metabolites that are unidentified, these can be isolated and characterised using NPC techniques. This is where NPC techniques become invaluable.

1.6 High throughput analytical techniques and their application in chemotaxonomy

Just as no single solvent system is currently suited to extract every small molecule from an organism, it is also true that no single analytical system or collection of systems is capable of identifying all metabolites (Allwood and Goodacre, 2010). Thus the selection of an appropriate analytical system is always a compromise between speed, chemical selectivity and instrument sensitivity (Sumner, 2006).

A combination of chromatographic instruments coupled to spectroscopic instruments allows for the separation and identification of metabolites from extracted metabolite pools. The most common systems for this process are gas chromatography (GC) or liquid chromatography (LC) coupled with mass spectrometry (MS). The GC and LC separation systems allow for a single or relatively few compounds to be passed to the mass analyser and the high resolution of modern MS analysers allow for correlation of a mass with a particular combination of elements, or a molecular formula. This is thus a highly valuable tool which contributes to the identification and quantification of metabolites.

Nuclear magnetic resonance spectroscopy (NMR) can also be coupled with LC, which gives a different method for compound identification. An ideal systems combines a chromatographic system with a variety of detection systems, such as LC-UV-MS-MS or LC-UV-NMR-MS systems, which give a maximum amount of chemical information per run and thus more information for unknown compound identification. Other spectroscopic methods, such as infra-red spectroscopy, have also been used (Sandasi et al., 2013, 2012).

1.6.1 GC-MS analysis

GC-MS has more consistently annotated databases than LC-MS due in part to its being an older technique, but also because it uses hard ionisation prior to mass detection. Hard ionisation (most commonly electron ionization (EI) at an electron energy of -70 eV) results in the fragmentation patterns which give structural information about a compound and further allows for accurate identification when combined with chromatographic retention data (Dunn et al., 2013). Because a standard ionisation technique is consistently applied, fragment ions are generated at specific retention times, making annotations for GC-MS databases are fairly complete for known compounds (Monteiro et al., 2013).

However, this technique has a few limitations. Because compounds must be volatile in order to be carried in the gas phase, there are significant limits to which compounds can be analysed using this technique. This is partly solved by derivatisation, where polar functional groups, such as hydroxyl, amine, and carboxylic acid groups, are derivatised to form nonpolar functional groups (Yi et al., 2014). However, even with derivatisation, the number of compounds that can be studied at one time using this technique is limited to approximately 1,000 "components", where at least some of the chemical signatures are differentially derivatised compounds (Sumner, 2006).

Metabolites reported from these experiments typically include those that have been identified using various online databases (NIST, GMD, etc.) or in-house databases which give a basis for qualitative comparisons of concentration. The use of analytical standards is still the most commonly accepted way to reliably determine the identity and concentration of a compound. An advantage of GC-MS is that it is relatively cheap as it uses an inert carrier gas such as He rather than the expensive analytical solvents needed in HPLC applications. GC also tends to have less chromatogram drift than LC and thus is more reproducible across experiments (see Sumner (2006) for review).

1.6.1.1 GC-MS and chemotaxonomy

Because GC-MS produces reliable, reproducible ion fragmentation patterns and because it is an older technology, it has a longer history of use as a chemotaxonomic method. GC-MS chemotaxonomy studies tend to be approached in two ways, the first is through untargeted analysis focusing on plant volatiles such as in Radulović and Dekić (2013); Sandasi et al. (2013); Lorenz et al. (2012); Xue et al. (2012). Alternatively, a more classic targeted analysis is also common, looking at specific compounds or classes of compounds that were previously isolated from various species in a family. This is done by using molecular ions and fragmentation patterns to identify new molecules of the same class (Calderón et al., 2013; Berkov et al., 2012; El Bazaoui et al., 2011).

1.6.2 LC-MS analysis

For various reasons great strides have recently been made in the use of LC-MS in the identification of small molecules. The chemical information coming from LC-MS has significantly increased because of advances in both LC and MS. These include advances in mass spectrometry which allow for more precise accurate masses to be attained, but also advances in column structure and pump pressures which allow for more efficient separation of individual metabolites in LC.

Because separation of compounds via HPLC depends on solubility and polarity rather than volatility, a greater range of compounds can be processed through such a system rather than through GC-MS. Additionally, separation based on polarity also provides information about the nature and class of the compound being analysed. The increase in separatory power also means that more useful information comes from the UV-Vis detectors normally attached to the LC system.

From the MS side, mass detectors have increased in sensitivity and machines are now capable of obtaining accurate masses beyond 5 ppm, the standard mass detection for publication (Allwood and Goodacre, 2010), into the pmol or fmol range (Sumner, 2006). The mass sensitivity of this technique makes it invaluable in cases where plant material is scarce as optimally it only requires 100-200 mg of fresh plant material but in more extreme cases 30 mg can be used (Tolstikov et al., 2007).

In LC-MS, soft ionisation techniques are used, usually electrospray ionisation (ESI), so that the parent ion correlates directly to the mass and molecular formula of the compound. By comparing the accurate mass recorded to masses in databases, possible identities are established (see below). These experiments can be done in negative and positive ionisation modes which result in $M-1$ and $M+1$. This is further complicated by the formation of adducts, such as $[M+H]^+$ and $[M+Na]^+$ in positive mode and $[M-H]^-$ and $[M+Cl]^-$ in negative mode (Dunn et al., 2013).

Even though accurate masses of parent ions are achieved, the LC-MS databases are relatively new and under-annotated which makes identification of high dimensional data quite difficult (Monteiro et al., 2013; Lange et al., 2008). This is compounded by adduct formation and the poor reproducibility of fragmentation patterns and ionisation (Ernst et al., 2014). Even when a mass is associated with a specific compound, there are many combinations of the same atoms which result in the same masses but with different structures (isomers, enantiomers, etc.). The major limitation of these systems thus is the correct identification of metabolites within an extract which tends to be only around 30% (Zhou et al., 2012).

This limitation is addressed to some extent by tandem MS experiments where a second MS analyser is linked to the first. For example, in experiments with quadrupole-time of flight instruments (Q-TOF/MS), a quadrupole is used to filter an ion of a particular mass to a collision cell, the precursor ion of interest is then fragmented before being passed to a TOF detector for accurate mass determination of the fragments. In this way, tandem MS instruments have two functions, the first is to scan intact parent ions and the second is to analyse the resulting fragmentation ions (Allwood and Goodacre, 2010). Depending on the detection system, MS^n can be achieved, such as in orbitrap MS. Even so, at least two independent and orthogonal data types such as retention time and mass spectrum, accurate mass and tandem mass spectrum, etc., are needed relative to an authentic compound analysed under identical experimental conditions to verify a putative metabolite identification (Dunn et al., 2013; Zhou et al., 2012; Sumner et al., 2007). Further, because there are so many different mass detectors used in both simple MS experiments as well as tandem MS experiments, the resulting databases are difficult to annotate as molecules will have different ion patterns across all of the different systems.

There are also many analytical limitations associated with LC-MS/MS especially with respect to the ionisation of target compounds, ion source transformation of target compounds, and the selection of target compound ions which to some degree must be dealt with on a compound by compound basis(Vogeser and Seger, 2010). To address these issues, many groups are starting to establish and employ custom reference libraries of chemical standards which allow true identification and quantification of specific metabolites (Kingston, 2011; Lange et al., 2008; Monteiro et al., 2013). This reduces identification ambiguity considerably as it results in specific retention times, molecular weights, preferred adducts, and in-source fragments of compounds, as well as their associated MS/MS spectra for the specific systems utilised. Because error can be propagated at many steps in the LC-MS experiment and in the preprocessing of the LC-MS data (for a comprehensive review, see Vogeser and Seger (2010)), internal standards are applied at various steps to ensure accuracy of the molecular weights of known and unknown compounds.

While LC-MS is useful for producing metabolic fingerprints and for determination of relative size, solubility, and polarity of molecules, it must be combined with natural products isolation methodologies or the use of analytical standards to get true identification of novel or unusual compounds.

1.6.2.1 LC-MS and chemotaxonomy

As with GC-MS, LC-MS has recently been employed in targeted metabolomics analyses to identify specific molecules from plants which have been previously recorded in a particular evolutionary lineage such as in Kanfer and Patnala (2013); Liu et al. (2013). LC-MS has also become an important metabolic fingerprinting technique where fingerprints can more generally be used as molecular markers (Elvira et al., 2014; Messina et al., 2014; Safer et al., 2011).

1.6.3 NMR analysis and chemotaxonomy

NMR is arguably the single most important and versatile technique for structure determination, with the advantage over X-ray diffraction of not requiring crystalline samples. While it provides rich structural information, NMR on its own tends to be used less frequently as a metabolomics technique as it is has much lower resolution than detection techniques such as MS and signal overlap in spectra of complex mixtures presents a significant challenge. These challenges are being overcome, to some extent, by the development and use of high field instruments (600MHz is increasingly common) and an array of 2D NMR techniques which contribute to the resolution of the array of signals.

As a metabolic fingerprinting technique, ^{1}H NMR is the most commonly used due to the inherent sensitivity of the ^{1}H nucleus, the relatively short times for spectrum acquisition, and the reasonable resolutions achieved (Kim et al., 2011). For these reasons ^{1}H NMR makes for an excellent comparative analysis framework for metabolic fingerprinting as it allows for consistent and rapid expansion of databases for metabolite identification. These factors make comparative analyses of ^{1}H metabolic profiles ideal for identification of unknown metabolites and once a novel set of signals are identified, further structural elucidation using ^{13}C NMR, and 2D NMR can then be achieved on the same sample (Halabalaki et al., 2014). Paring NMR with HPLC greatly aids in the structural elucidation as fewer signals can be compared at a time(Kim et al., 2011).

Current work involving NMR in chemotaxonomy for metabolic fingerprinting generally uses NMR on its own (Kim et al., 2012; Safer et al., 2011; Kim and Choi, 2010), although LC-NMR is becoming more common (Halabalaki et al., 2014).

1.6.4 Summary

Each of the approaches above has its benefits and weaknesses, with the ultimate goal of metabolic fingerprinting remaining to analyse as many compounds as possible in a single run. GC-MS is limited in the kinds of compounds that it can be used to analyse, but is highly reproducible. LC-MS can be used to analyse an extraordinary range of compounds, but is not as reproducible and the results are challenging to interpret. NMR gives beautiful structural information but has a generally low detection threshold.

1.7 The approach in the following thesis

The intention of this study was to undertake a metabolic fingerprinting analysis on standardised plant extracts for chemotaxonomic purposes. In selecting appropriate tools for this analysis, and based on the forgoing summary, it was clear that GC-MS was too limited in scope, and in our context routine access to NMR was limited to a 400 MHz machine with its inherently limited sensitivity and resolution. LC-MS therefore became the method of choice for generating metabolic fingerprints for its accessibility as well as its high sensitivity. It was then clear that the most significant challenge lay in how to handle the very large data sets generated by LC-MS, and for this reason, multivariate methods were required to achieve holistic understanding of the data.

1.7.1 Clustering for dimension reduction

In many cases hierarchical clustering is used to determine if the relevant analytical signals adequately identify plant sample classes. This is an agglomerative undirected technique and if the plant samples do cluster as expected, then multivariate statistical approaches are employed to determine which analytical signals are responsible for sample clustering.

Principal components analysis (PCA) and partial-least squares discriminant analysis (PLS-DA) are, by far, the most common multivariate techniques employed in metabolomic studies for dimension reduction across data from all of the different spectroscopic analyses commonly employed in metabolomics experiments (Elvira et al., 2014; Messina et al., 2014; Aliferis and Cubeta, 2013; Liu et al., 2013; Kim et al., 2012). They differ in that PCA is undirected, where no information about the sample classes is given prior to the analysis being run. The “discriminant” aspect of PLS-DA arises as sample classes are established ahead of classification, hence “supervised”. The difference between these techniques is thus that PCA establishes clusters of samples around the variability of the variables whereas PLS-DA clusters samples around the variability of the samples.

In either case, variance is described in two matrices, the score and the loading matrices, where the variance of the samples is described in the score matrix and the variance of the variables is described in the loading matrix. These can be superimposed to form a biplot so that the clustering of the samples and variables can be compared.

Because the purpose of PLS-DA is to identify the variables which distinguish sample classes, this and other forms of discriminant analysis are inappropriate for establishing if analytical signals can be used to determine sample classes. Moreover, in cases where the number of variables greatly outnumber the number of samples, discriminant analyses are prone to over fitting; this is especially

the case with megavariate data which Rubingh et al. (2006) define as cases where there are more than ten times the number of variables as there are observations. For this reason, only unsupervised techniques were employed in dimension reduction.

In PCA, singular value decomposition (SVD) is applied to a normalised matrix to generate two covariance matrices for samples and variables respectively. In a taxonomical fingerprinting application, analytical signals are the variables. The analytical signals which show change in intensity between plant samples are highly informative and distinguish species samples from each other. If that variation is less in samples of the same species than it is between species, then the samples representing each species will cluster together.

In principle, once the scores and loadings are overlaid the analytical signals responsible for sample clustering can be visualised. In cases where clustering is applied only to a subset of data from analytical signals that have been identified as specific compounds using analytical standards, this is a fairly straightforward process where the analytical signals which represent the standards analysed are visually assessed or a threshold, such as a Hotelling T-test with a 95% confidence oval, is determined for importance. For identification purposes, where analytical signals from known compounds are not used exclusively, the same data reduction process is used to identify interesting analytical signals accompanied by attempts to identify the relevant compounds using online databases or by further characterisation from the spectroscopic data. More commonly a combined approach is utilised where there are specific compounds of interest tracked using analytical standards along with some tentatively identified compounds. Most of these studies revolve around previously understood chemotaxonomy where there are specific compounds which are known to be different between species.

In completely undirected studies, where nothing is assumed about the chemotaxonomic separation of the species, because of the high-dimensionality of the data, a simple biplot is not particularly informative. Thus a subsequent step was employed to identify more precisely which chemical signals are responsible for species specific clustering.

1.7.2 Information theory- turning ions into information

In information theory, the information given by each variable is quantified. As was shown in Yip et al. (2014), information theory can be applied directly to a PCA loading matrix to determine which parts of a data set are specifically informative by ranking the loading values of the individual peaks. Once the amount of information for each ion is knows, an information threshold can be established and the ions responsible for explaining a threshold variance can be subset into a new data matrix. In Chapter 5 the details of this are given more specifically.

In essence, this work has established a method whereby metabolomic fingerprints can be generated by establishing a subset of highly informative analytical signals that in turn can be used as a metabolomic barcode for species identification. This concept mirrors the use of DNA fingerprinting in DNA barcoding and serves as an accompanying approach for taxonomic purposes.

Once this subset of analytical signals was established as an identification fingerprint for each species, it was necessary to assess the stability of those analytical signals in samples collected over time to ensure future reproducibility. It was also important to establish a model for future species identification.

1.7.3 Applications of machine learning

Machine learning is a statistical method used for model generation and has two functions, classification and prediction. Once a model is established it can be used to classify unknown samples, or to predict the outcome of a sequence of events. While classical modelling techniques make assumptions about data distribution and then predict or classify based on assumed patterns, machine learning builds models by sampling the data, in most cases over and over again, and combing those models into an average model based on average model features generated.

While it is easy to see if all of the values of one variable across many samples are distributed along a line, the assumption made in linear regression, it is difficult to see this across hundreds or thousands of variables. In the epoch of big data which is now beginning, machine learning is becoming a more common tool used to make sense of data which is not visually accessible and does not readily permit the formulation of generalized assumptions.

Machine learning is relatively new to metabolic fingerprinting, but has been used successfully by De Bruyne et al. (2011) who used random forest and support vector machines to classify bacterial species from LC-MS/MS fingerprints. Interestingly, their classification model was stable across 49 bacterial strains with a 94-98% accuracy even with different sample preparation methods. Many other authors are using machine learning for fingerprinting purposes (Boccard et al., 2010; Scott et al., 2010; Gao et al., 2012). More specifically Howley et al. (2006) describe the use of PCA to reduce the dimensionalitly of spectral data. As a classification system across many different types of machine learning, they found that by reducing the number of variables using PCA, they were able to increase species identification rates.

The methods utilized in this work included the information theory method described in Yip et al. (2014) to reduce the number of variables from the PCA loading matrix into a set of informative ions or metabolomic barcodes. Those barcodes were then applied to the machine learning classification modelling algorithm random forest as explored in De Bruyne et al. (2011) to assess their stability as chemotaxonomic markers for the identification of the various Aizoaceae species studied and to generate a model for species identification.

1.8 Expanding biological understanding with metabolic fingerprints

In the present study where secondary metabolites were analysed as a potential chemotaxonomic tool, it was deemed important to have an understanding of the physiological responses of the species under study to the environmental conditions in which they naturally occur. A few examples follow.

The importance of contextualizing metabolite profiles was demonstrated by Aliferis and Cubeta (2013) who explored the use of 189 metabolites in the chemotaxonomic classification of various fungal species which were previously identified using genetic and morphological markers. They were then able to further identify the same species grown on different substrates using a combination of targeted and untargeted metabolomics techniques.

While soil nutrients and environmental conditions obviously play a significant role in metabolite production, Messina et al. (2014) showed that for the *Olearia phlogopappa* (Asteraceae) species complex, there was not a significant difference between the metabolic profiles of leaf material grown in the greenhouse and leaf material collected from the field. Furthermore, populations that have proven to be particularly problematic to separate with genetic markers, but were morphologically different, were successfully distinguished using metabolic fingerprints. By using a more holistic approach, the nature of the metabolites produced by an organism, as well as what stimulates their production, can be identified.

While it is of importance to ultimately assess the robustness of a chemotaxonomic classification by monitoring changes in the intensity of analytical signals over time, few studies have as yet reported this in natural field conditions in wild populations. In the present study, distinct seasonal variation prevalent in the study area (reviewed below) was deemed to be likely to have the greatest impact on metabolic variation within any one species. Thus samples were collected once a month for 12 months (from April 2011 to March 2012) for fingerprinting analysis.

To our knowledge, there have been no studies to date on the variation over time across the entire metabolome of individual plant species, and assessment of the effect this has on metabolic fingerprint-based chemotaxonomy.

1.8.1 The approach utilised in the present study

The present study was performed on five Aizoaceae species (*Galenia africana, Aridaria noctiflora, Carpobrotus edulis*, *Ruschia robusta*, and *Tetragonia fruticosa*) using two Crassulaceae species as controls (*Cotyledon orbiculata* and *Tylecodon wallichii*). The study was guided by previous taxonomic work using phylogenetic markers by Klak et al. (2013, 2007, 2003), which established the difficulty in the separation of these species using the typical nuclear and plastid DNA regions. These difficulties suggested that the use of chemical markers might be particularly valuable at the very least to give additional phenological markers to consider when analysing these species.

An initial NMR study was attempted using a locally available 400 MHz machine, but the resolution was not high enough for this to be used as a metabolic fingerprinting technique (data not shown). Due to the inherent insensitivity of the technique, only the compounds that are most prevalent in the sample are detectable, and while this is in itself a potentially interesting pursuit, the most abundant compounds are not necessarily different between species. Thus these highly abundant compounds were insufficient as molecular markers. Ultimately, LC-MS was chosen as the only fingerprinting technique for its sensitivity as well as its coverage of chemical space in terms of molecular weight, polarity and solubility.

The goal of this study was to determine the stability of the metabolic fingerprints of five Aizoaceae species over 12 months and then to determine if those fingerprints could be further divided into a subset of metabolites which could accurately distinguish the species from each other. This study combines the ideas from NPC, metabolic fingerprinting, and chemotaxonomy by comparing species that have been shown to be related via phylogenetic and morphological studies with tests that indicate shifts in metabolite production. To this end, biological profiles were first created including isotopic, metobolic, nutrient, phenological, and climate (see Chapter 3) to provide a context to understand metabolite production across 5 Aizoaceae species as derived from LC-MS metabolic fingerprints.

Then, the undirected multivariate clustering method PCA was applied in order to group species and the ions from LC-MS and thus identify the analytical signatures responsible for species-specific clustering. Leverage scoring was applied to identify the ions responsible for the clustering and those ions were turned into metabolic barcodes. The barcodes were tested for stability by comparing their ion intensities over time. Finally, the machine learning technique, random forest, was employed to generate a classification model which was trained and tested on plant samples in various ways to insure that stability.

Chapter 2

Contextualising the ecology

Every study must be contextualised in the space in which it was conducted. This study utilised plants, soil, climate data, and knowledge from a community in Namaqualand, South Africa (see Figure 2.1). This study was also associated with and facilitated by a long-term ecological survey of the area by the director of the University of Cape Town's Plant Conservation Unit, Timm Hoffman. This meant that there was a long-term ecological record for the study area, a well trained para-ecologist on site, and an on site research station at which the field work could be conducted.

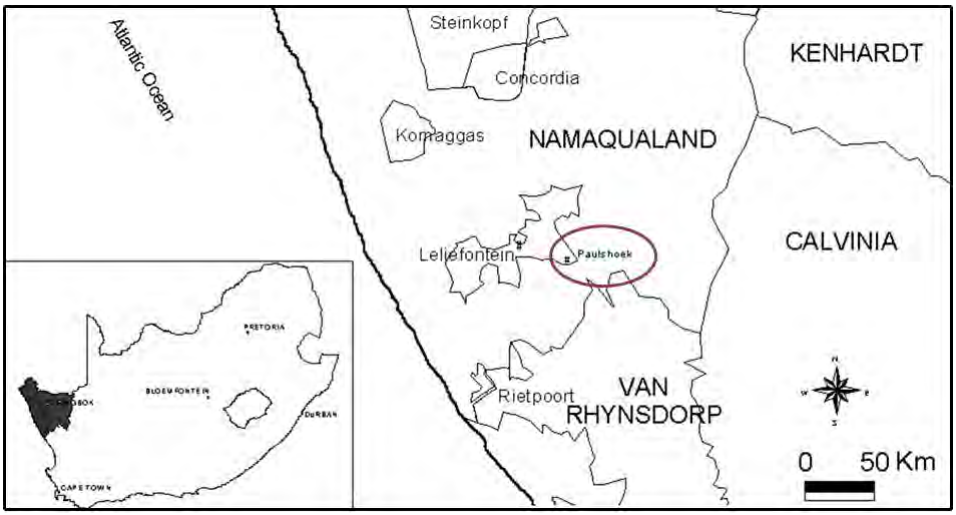


Figure 2.1: **Location of study site.** Figure adapted from Rohde and Hoffman (2008) indicating historical land division. Field work occurred in the village of Paulshoek, shown circled in red, which forms part of the old Leiliefontein communal reserve.

2.1 Namaqualand

Namaqualand is a desert ecosystem characterised by physical boundaries; to the east, the Kamiesburg mountain range, to the south, the Olifants River and to the north, the Orange River with the Atlantic Ocean to the west. In total, it covers about 50,000 km^2. The soil composition is mostly granite sands although mineral diversity as well as altitude changes contribute greatly to plant species diversity (Hoffman et al., 2007). It is considered to be the most botanically diverse desert in the world with an estimated 3,500 distinct plant species across 135 plant families. Incredibly, 25% of this diversity is endemic to Namaqualand. These factors together make it one of only two deserts in the world to be considered biodiversity hotspots, the other being the Horn of Africa (Desmet, 2007). A "hotspot" is defined by a combination of two characteristics the first is that there are high levels of floral endemism, which is defined as greater than 1,500 endemic plant species, and the second, that the ecosystem has to have lost at least 70 percent of its original habitat (Myers et al., 2000). The Succulent Karoo Biome is estimated to have once covered about 112,000 km^2, but is currently estimated to cover about 30,000 km^2 which is about 26% of its original area (Myers et al., 2000).

2.2 Climate

Located at S 30 21 58.0, E 18 15 14.5, Paulshoek falls ecologically within the arid and semi-arid winter rainfall region of Namaqualand, within the greater Succulent Karoo Biome (Figure 2.1). The region receives roughly 200 mm of rainfall per annum (Rohde

and Hoffman, 2008). Floristically, it is shrub-land, dominated by leaf succulents and deciduous-leafed woody shrubs. Succulents represent 35% of the floral diversity in the region with 500 recognised Aizoaceae species (Desmet, 2007). The soil is characterised by granite sands with 50% of the particles being greater than 0.3 mm diameter and overall nutrition being variable depending on the grazing habits of the particular area (Allsopp, 1999).

2.2.1 Climate data

Climate data for the study area from the periods 1998-2003 were supplied by Timm Hoffman in his report, Hoffman (2005). Methods of collection and analysis of the data are given below. Similar data was not collected during the present study due to storm damage to the monitoring equipment in 2010, and the difficulties associated with replacing and maintaining the equipment. Current data was thus interpreted in the light of trends from 1998-2003.

2.2.1.1 Temperature

The records available from 1998-2003 are displayed in Figure 2.2 and show that the temperature varies within a small range. The trends in high and low temperatures are displayed as monthly averages and appear to be consistent between the years. The highest standard deviation of the mean is 3.3°C and the range extended from 2.68°C to 31.15°C. Temperatures fall within the range previously reported by Desmet (2007).

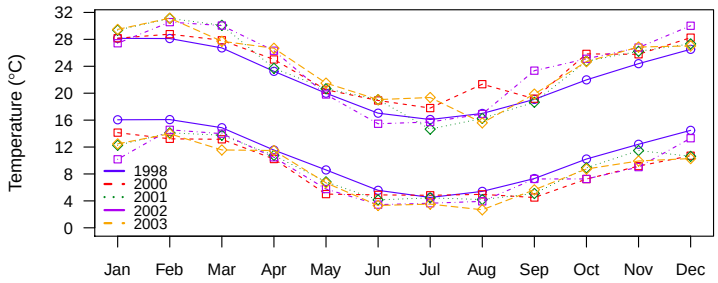


Figure 2.2: **Monthly temperature distribution in Paulshoek.** Daily temperature (min and max) were determined from data collected hourly from May of 1998 to December of 2003 and are presented as monthly averages.

2.2.1.2 Photosynthetically active solar radiation (PAR)

PAR measures wavelength intensities 400 to 700 nm, which is the photosynthetically usable wavelength range. Daily changes in PAR, or photoperiodism, are seasonal and play many important regulatory roles in plants including leaf abscission, reproduction, and metabolite response (Jones, 2013). Monthly averages of PAR from 1998-2003 are displayed in Figure 2.3.

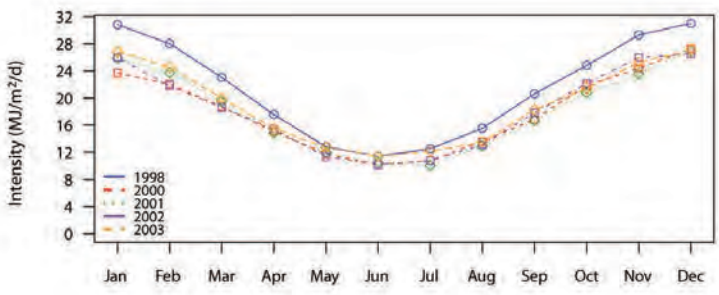


Figure 2.3: **Monthly photosynthetically active solar radiation distribution.** PAR was calculated in mvolts every hour and converted to mega joules per meter squared and summed for the day to give units of MJ/m^2/d. These values were then averaged over each month and displayed by year. Final analyses used an average of the 5 years (1998-2003) with a maximum standard deviation of 2.13 MJ/m^2/d and a range of 6.29-27.73 MJ/m^2/d.

2.2.1.3 Vapour pressure

Vapour pressure plays a role in stomatal opening which is critical for plants in arid environments (Agam and Berliner, 2006; Kanniah et al., 2012; Zhang et al., 2014). Thus plants utilising C3 metabolism will experience higher rates of transpiration on days with high vapour pressure deficit as compared to plants utilising CAM metabolism. Monthly vapour pressure averages are shown in Figure 2.4.

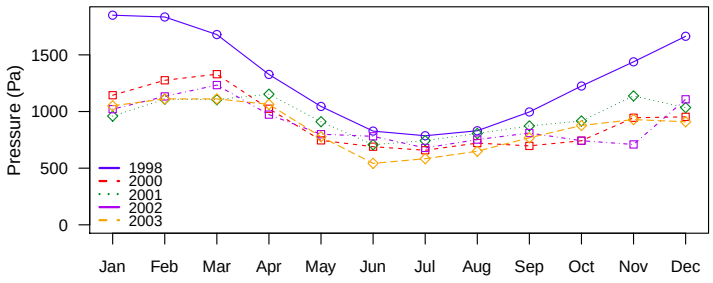


Figure 2.4: **Average monthly vapour pressure distribution.** Average monthly vapour pressure was calculated from the hourly temperature and relative humidity and averaged over the month. Final analyses used monthly average across the five years of data with a maximum standard deviation of 299.47 Pa and a pressure range between 236.81 and 1292.67 Pa.

2.2.1.4 Rainfall

Total annual rainfall was monitored every day by a Paulshoek local and summed over the month for monthly totals. This data was collected for the duration of the current study in Paulshoek and site 3 and is shown relative to the 15 year average for the area (see Figure 2.5).

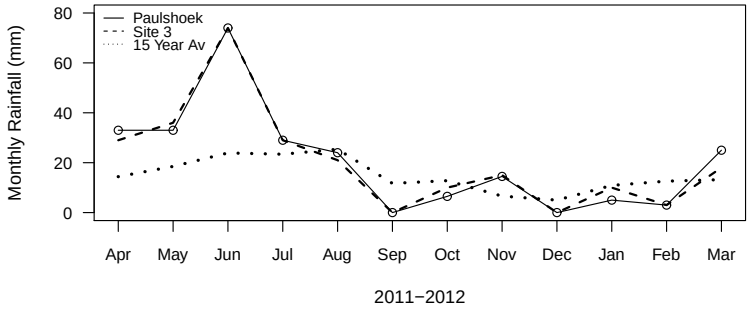


Figure 2.5: **Monthly rainfall totals from Paulshoek.** Rainfall totals were monitored over the study period from both field sites and are plotted with the 15 year average.

Rainfall was fairly consistent between Paulshoek (247 mm) and Field Site 3 (245 mm) although in years past this has not always been the case (Timm Hoffman, unpublished work). Both sites received more rain than across the 15 year average (178mm) especially in the early months of the study from April-June of 2011 (see Figure 2.6).

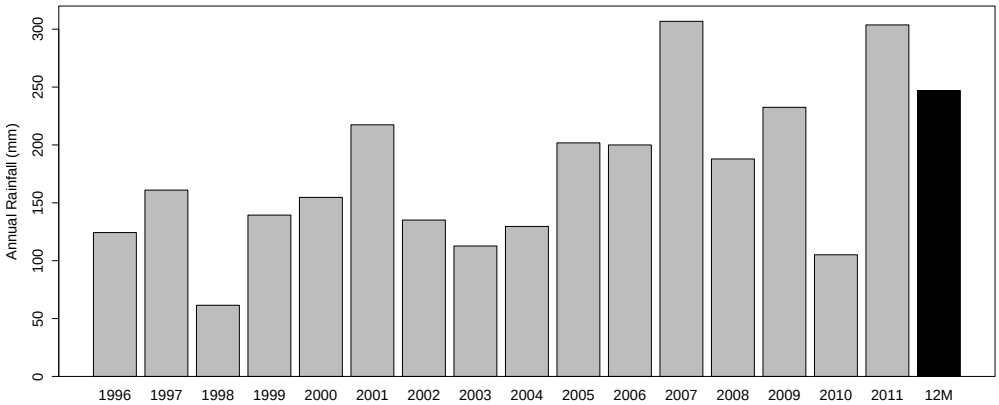


Figure 2.6: **Total annual rainfall over a period of 15 years.** Previous years (1996-2011) are indicated in grey and the combined 12 months of the study period is indicated in black.

Rainfall averages were relatively high over the study period (12M) especially in the late summer months at the beginning of the year, although late winter/early summer averages were on the low side. Over the 12 month time span this study was conducted, the total rainfall ranked as the third highest rainfall average recorded in the last 15 years.

2.2.2 Summary

Considering the above combination of climate variables, it was decided that the 12 month collection period could be separated into two seasons: winter, which lasted from April 2011 to August 2011 and summer, which lasted from October 2011 to March 2012. September was possibly a transition month between the two seasons and so was left out of the seasonal analysis conducted in Chapter 5. A comparison of these trends with plant-based analyses will be covered in Chapter 4.

2.3 Species selection

The main objective this study was to determine the stability of the metabolome of various plant species over a one-year period and to see how that variation related to previous taxonomic studies. The species selection, therefore, needed to be particularly rigorous in order to capture species with the greatest potential for variability. Outlined in this section are the various criteria used for this selection including high density distribution at the field site, a range of carbon uptake mechanisms, phylogenetic diversity, previously recorded biological activity associated with a plant or its extract, and impact on the local economy.

Ultimately five Aizoaceae species were selected to represent a family of more than 500 species in the family reported to grow in Namaqualand (Desmet, 2007). These include *Galenia africana*(L.), *Aridaria noctiflora* (L.) Schwantes var. noctiflora, *Carpobrotis edulis* (L.) Bolus, *Ruschia robusta* (L.) Bolus, and *Tetragonia fruticosa* (L.). As a contrast and a baseline for various experiments, two species from the Crassulaceae family were also selected. These include *Cotyledon orbiculata* (L.) var. orbiculata and *Tylecodon wallichii* (Harv.) Tolken ssp. wallichii.

Plant collection was authorised by the Northern Cape Province and the Kamiesberg Municipality under permit "FLORA 043/2011" in 2011 and was updated in 2012 for the duration of the study period. Voucher information can be found in Table 2.2.

2.3.1 Metabolic diversity

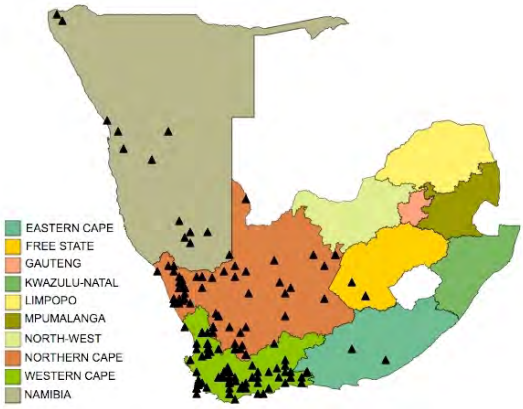
The primary selection criterion for the Aizoaceae family was its metabolic plasticity as species in this family exhibit facultative CAM photosynthetic metabolism (facultative CAM). This is the ability of a plant to switch between C3 and CAM carbon uptake (Niewiadomska et al., 2011; Silvera et al., 2010; Herrera, 2008; Libik et al., 2005). Presently, few studies have explored which of the Aizoaceae species utilise this mechanism.

Interestingly, a study by Vogt et al. (1999) showed that certain secondary metabolites from *Mesembryanthemum crystallinum* (L.) were only produced when the plant was utilising CAM metabolism. If this were the case for secondary metabolite production more generally across many Aizoaceae species, it might pose complications when attempting to account for stable chemical markers. Because of this, two Crassulaceae species, which are obligate CAM, were selected as markers for CAM metabolites and their potential changes over season.

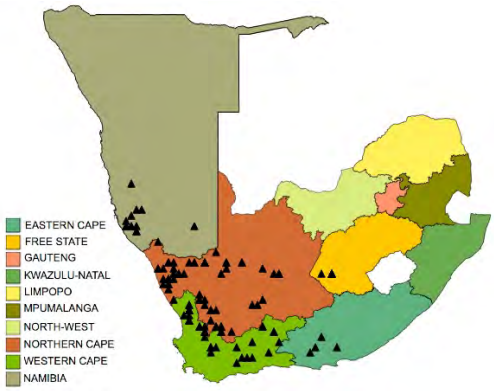
2.3.2 Distribution

All of the species utilized in the current study are rated with the status "least concerned" by the South African National Biodiversity Institutes (SANBI) 2011 Red Book List. As noted by Driver et al. (2009), "A taxon is considered as Least Concerned when it has been evaluated against the five IUCN criteria and does not qualify for the categories Critically Endangered, Endangered, Vulnerable and Near Threatened, or the South African categories Critically Rare, Rare or Declining. Widespread and abundant taxa are typically listed in this category". This factor was crucial as it allowed for repeat bulk collection (700 g) to take place over the course of 12 months.

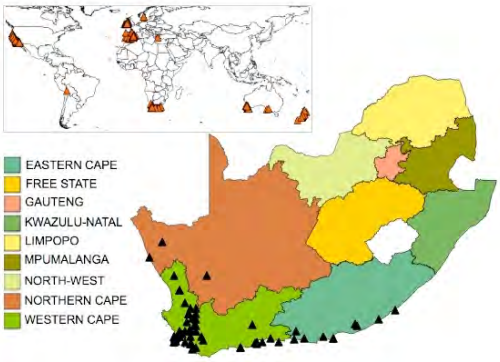
All of the study species have southern African distribution as is shown in Figures 2.7 and 2.8 except for *C. edulis*. *C. edulis* occurs in the coastal regions of every continent on the planet capable of hosting vegetative tissue. This is, in part, due to its historic use as a sand dune stabiliser. The effectiveness of *C. edulis* in coastal dune habitats makes it a highly invasive species that easily out-competes native dune plant populations (Novoa and González, 2014). There are also two South African endemics: *T. wallichii* is endemic to Northern Cape, Western Cape, and Eastern Cape provinces (see Figure 2.8b), while *R. robusta* is found only in the Succulent Karoo (see Figure 2.7d).



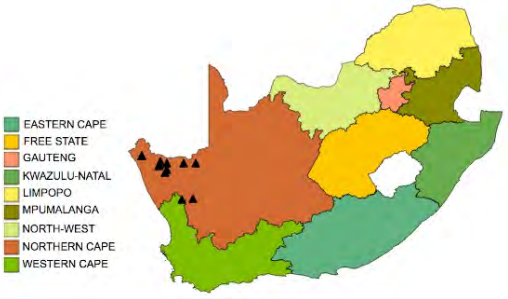
(a) *G. africana*



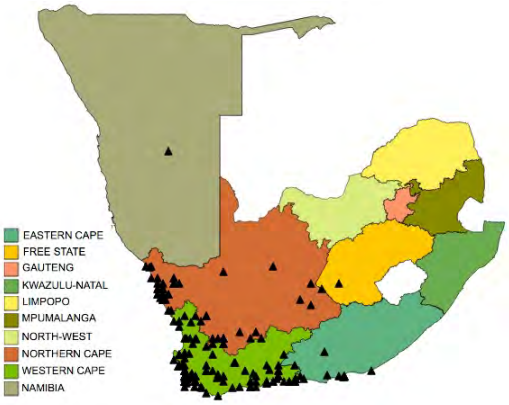
(b) *A. noctiflora*



(c) *C. edulis*

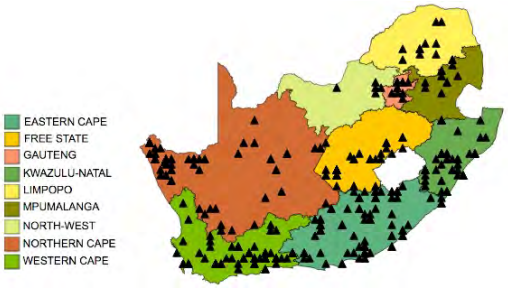


(d) *R. robusta*

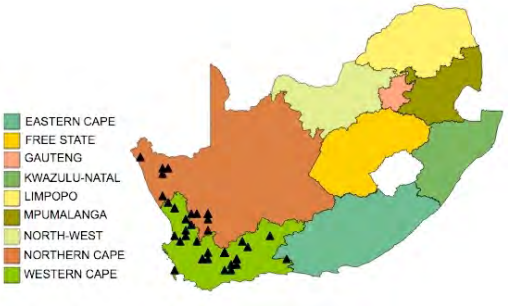


(e) *T. fruticosa*

Figure 2.7: **Distribution of Aizoaceae species.** South Africa's provinces are shown and the grey mass appearing at the top left in only some of the maps represents Namibia. *C. edulis* is the only species with a global distribution.



(a) *C. orbiculata*



(b) *T. wallichii*

Figure 2.8: **Distribution of control Crassulaceae species.** The colours delineate South Africa's provinces.

Table 2.1 indicates the distribution of species in Paulshoek which are impacted by the level of grazing in a particular area as estimated by the director of UCT's Plant Conservation Unit, Timm Hoffman in October of 2013 based on a visual assessment.

Table 2.1: **Estimated distribution of study species within Paulshoek.** "% Cover" represents the percentage of that species in the global species diversity of that area and "Density" is the number of plants per hectare. "Outside" refers to the area outside the fence-line of the "Campsite". *C. edulis* is not mentioned in this table as it grows in riparian environments only and does not appear at most stock posts or in or around the "Campsite".

	Campsite		Outside		Slooitjiesdam		Stockposts	
Species	% Cover	Density	% Cover	Density	% Cover	Density	% Cover	Density
G. africana	5	100	30	> 1000	7	300	40	> 5000
A. noctiflora	5	50	< 1	< 10	0	0	0	0
R. robusta	<1	5	0	0	50	>1000	1	< 100
T. fruticosa	5	50	1	10	1	<50	0	0
C. orbiculata	2	30	< 1	< 10	0	0	0	0
T. wallichii	3	100	5	200	0	0	0	0

As is seen in Table 2.1 the "Campsite" area, which is surrounded by a fence and protected from grazing, hosts *G. africana* at roughly 1/10th of the density it has outside the fence-line. This trend follows in areas which are consistently grazed which tend to have a much higher population density of *G. africana* and *R. robusta* than the other areas. The absence of *A. noctiflora* and *T. fruticosa* in these areas can be attributed to their high degree of palatability (see Figure 2.3). *C. edulis* was not included in this table as the ecosystem that it grows in is not normally associated with stock posts and is not located within the "Campsite" area, however, it does not seem to be affected by grazing pressure.

The lack of *C. orbiculata* and *T. wallichii* in the area around stock posts is thought to be due to active removal by farmers due to the common knowledge of their toxicity and their inability to repopulate after removal (Kellerman, 2009).

G. africana and *R. robusta* are pioneer species, which tend to be the first plants to reclaim highly disturbed areas such as after intensive grazing pressure. This ultimately leads to increases in their distribution and density and is especially the case with *G. africana* (Simons and Allsopp, 2007; Riginos and Hoffman, 2003; Carrick, 2003; Todd and Hoffman, 1999).

2.3.3 Taxonomic diversity

Species were selected from three of the four Aizoaceae sub-families (Klak et al., 2003) as is seen in Figure 2.9. No species were selected from the fourth Aizoaceae subfamily Sesuvioideae, due to their general lack of availability at the study site. Klak et al. (2003) regrouped Aizooideae and Tetragonioidea into the same subfamily, Aizooideae, so species were selected to represent this previous grouping for comparison. In Klak et al. (2003, 2007, 2013) the subfamily Ruschioidea were particularly difficult to separate using typical plastid and nuclear DNA markers and so two species from this subfamily were also selected.

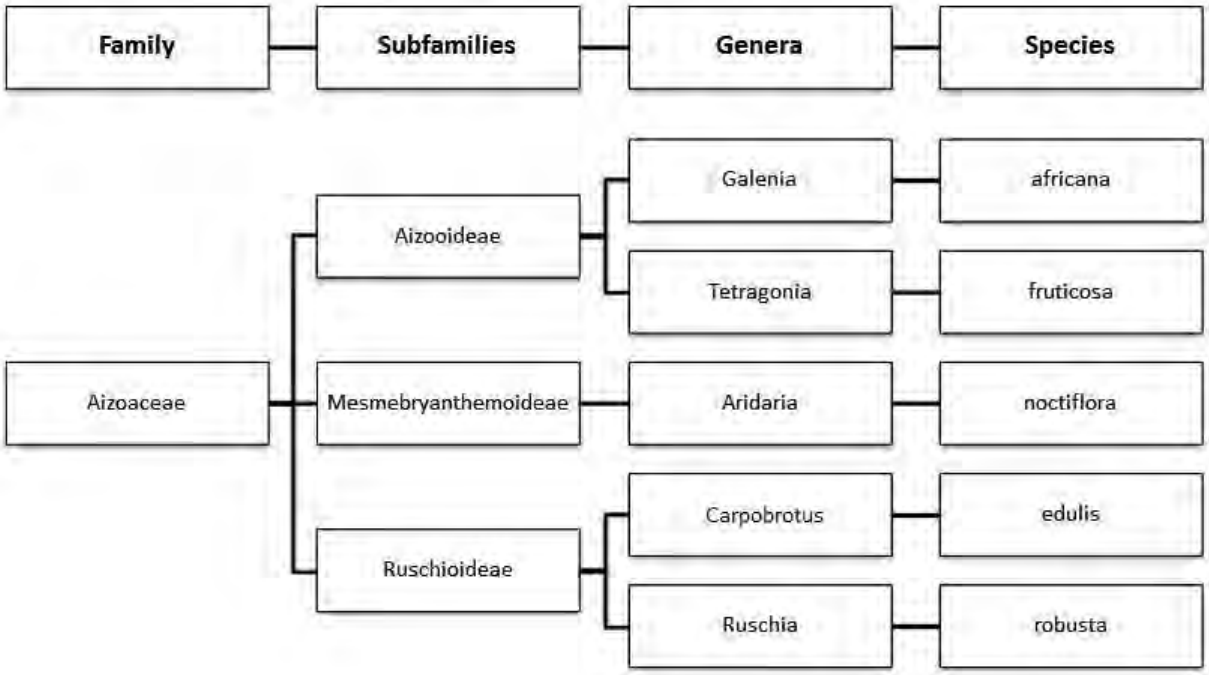


Figure 2.9: **Taxonomy of selected Aizoaceae species.** Figure adapted from Klak et al. (2003), and is not an accurate representation of phylogenetic distances.

C. edulis was considered particularly important as, due to its global distribution (see Figure 2.7c), much work has already been done on its metabolism. This was critical as it added a number of metabolites to the Aizoaceae secondary metabolite database as will be described in the following chapter.

2.3.4 Ethnobotanical background

South Africa has a rich ethobotanical heritage that plays a huge role in its modern society. As such, there are many well documented cases of plants used for medicines, and a summary of this information for the species under investigation in this study are given in Table 2.2.

Of the plants studied, four are used widely in traditional medicine (*G. africana*, *C. edulis*, *C. orbiculata*, and *T. wallichii*), one has some local use (*T. fruticosa*), and two are not known to have any medicinal use (*A. noctiflora*, *R. robusta*). The latter have been included in order, potentially, to highlight components of the metabolic profile which may be responsible for biological activity.

Of the seven species, three are also widely considered to be toxic (*G. africana*, *C. orbiculata*, and *T. wallichii*). Previous studies of *C. orbiculata* and *T. wallichii*, have shown a seasonal trend in the production of their toxic principal cotyledocide (Botha et al., 2001). Local accounts by herders as seen in Table 2.3, suggest an understanding of these seasonal fluctuations and that *G. africana* may also have seasonally dependent toxicity.

Of the five Aizoaceae species, only *G. africana* and *C. edulis* have been chemically profiled to any extent. A review of compounds that have previously been isolated from *G. africana* and *C. edulis* as well as any other Aizoaceae species can be found in Appendix A.

Table 2.2: **Plant species selected for this study, together with a summary of ethnobotanical information and herbarium voucher numbers.** "Voucher number" refers to the herbarium vouchers lodged in the Bolus Herbarium (BH) at the University of Cape Town, South Africa and the Kimberley McGregor Museum Herbarium (KMG) in Kimberley, South Africa. A review of metabolites previously isolated and characterised from Aizoaceae species can be found in Appendix A.

Scientific name	Common name(s)	Medicinal use	Toxic properties	Voucher Number
Galenia africana (L.)	Kraalbos, geelbos	External wounds/infections, eye infections, toothache, bladder infections, ringworm, burns, dandruff (Nortje, 2011; Knowles, 2005)	Waterpens: cirrhosis and severe ascites (Botha and Penrith, 2008; Bath et al., 2005; Kellerman et al., 1988)	BH58371 KMG35911
Aridaria noctiflora (L.) Schwantes var. noctiflora	Vyebos	No known uses	No known toxicity	BH58374 KMG35910
Carpobrotus edulis (L.) Bolus	Suurvy, perdevy, vyerank, ghaukum, sour fig	Infections of the mouth and throat, dysentery, digestion, tuberculosis, diuretic, external wounds/infections, earache, ringworm, diphtheria, treatment of stings and bites, oral thrush, ulcers, delayed labor, wart removal, teething problems, sunburn and diabetes (Nortje, 2011)	No known toxicity	BH58375 KMG35909
Ruschia robusta (L.) Bol.	Large Xhouroe, swart t'nouroebos, swartstamvyebos	No known uses	No known toxicity	BH58377 KMG35908
Tetragonia fruticosa (L.)	Persleinbos, slaaibos, waterslaaibos, kinkelbossie, kinkelklappers, roosmaryn, kleinsaadklappiesbrak	Additive in herbal remedies (Wheat, 2014)	No known toxicity	BH58376 KMG35907
Cotyledon orbiculata (L.) var. orbiculata	Plakkies, kouterie, varkoor, pig's ear, skapiesbos, ppbos	Epilepsy, toothaches, earaches, mouth ulcers, removal of corns, syphilis, intestinal parasites, burns, cracked lips, band aid(Nortje, 2011; Steyn et al., 1986; Dyubeni and Buwa, 2012)	Krimpsiekte: general paralysis and weakness (Kehoe, 1912; Bath et al., 2005; Kellerman, 2009; Botha and Penrith, 2008; Botha et al., 2003; Mabona and Van Vuuren, 2013)	BH58372 KMG35906
Tylecodon wallichii (Harv.) Tlken ssp. wallichii	Nenta, kandelaarsbos, kriempsietebos, kandelaarbos	Plantar warts and abscesses (Nortje, 2011)	Krimpsiekte: general paralysis and weakness (Bath et al., 2005; Kellerman, 2009; Botha and Penrith, 2008)	BH58373 KMG35905

2.3.5 Economic impacts

The economic impact of highly distributed species cannot be underestimated. For example, the toxic cardiac glycosides produced by *C. orbiculata* and *T. wallichii* in southern Africa are estimated to cost South African farmers about R 23.5 million per year (Botha and Penrith, 2008). *G. africana* has been shown to have lethal effects post-ingestion in livestock by many researchers (Pool et al., 2009; Botha and Penrith, 2008; Bath et al., 2005). Although the direct impact on economic aspects of domestic herds has not been assessed, this is becoming more problematic as *G. africana* is a pioneer plant in an increasingly disturbed area (Allsopp et al., 2007; Riginos and Hoffman, 2003; Allsopp, 1999).

The plants considered in this study range in palatability (Table 2.3) and have been described by herders in interviews as well as by various investigators (Hendricks et al., 2002) as follows; *R. robusta*, *A. noctiflora* and *T. fruticosa* are the most palatable with *C. edulis* and *G. africana*, being the least palatable. Surprisingly, *C. edulis* proved to be particularly unpalatable for a non-toxic species and falls in the same palatability range as the toxic species. This is made evident by large sprawling intact plants that while easily accessible to animals, are trampled, but not eaten.

Table 2.3: **Palatability of Species.** Palatability is based on a score from 0-5 where 0 is the least palatable and 5 is the most palatable (Timm Hoffman, unpublished work). Notes taken directly from interviews with local farmers.

Species	Score	Notes
G. africana	0.3	Animals will eat it when dry. Get "waterpens" (water-belly) when veld is "skars" (nothing to eat). The Bushmanland species considered it a good bush. "Maak vee vet" (Fattens livestock).
A. noctiflora	2.6	Donkeys will "kap" (trample or destroy) it out and eat stems when green.
C. edulis	0.0	Animals will not eat. Fruits eaten by people.
R. robusta	3.9	Green fruits are eaten when young. Donkeys pull this species out. Good for rams, especially when leaves are turgid.
T. fruticosa	3.7	
C. orbiculata	0.1	Also gives krimpsiekte (shrinking sickness), especially when "verlep" (wilted).
T. wallishii	0.0	Only dangerous at certain times of the day. Very dangerous if dew has fallen.

The palatable species are also critical in terms of dry-land herding practices as local species are adapted to grow in arid environments where as other typical feed crops are not. As overgrazing decimates populations of palatable non-toxic species, the food supply becomes limited for existing herds and the probability of livestock consuming toxic species increases.

Chapter 3

Methodology

This chapter outlines the rationale and methodologies utilised to catalogue the phenotypical changes of the five Aizoaceae species in response to changing environmental conditions from month to month together which metabolite production over the same period.

As outlined in Chapter 1, there are many factors which might contribute to the metabolism of secondary metabolites. In order to assess the impact of these factors in the present study, a set of factors which could indicate potential metabolic switches were selected for analysis. Changes in abiotic factors that would affect all of the plants in the field were monitored and a number of biotic as well as abiotic factors were considered to help determine when seasons occurred. These include a variety of climate factors, soil nutrient analyses, plant nutrient analyses, plant phenology, plant physiology markers, and markers of carbon uptake transition. The season determination for the purposes of model building will be further discussed in Chapter 5.

3.0.5.1 Key consideration in the selection of field sites and study materials

All of the field sites selected were those that were minimally impacted by grazing animals in order to discount potential effects of plant-vertebrate herbivore interactions. Invertebrate herbivore and pathogen interactions were much more difficult to control. To mitigate these effects, any material with mechanical damage or which showed signs of infection was specifically excluded from collection.

3.0.5.2 Phenology

To understanding species growth and development, phenology was monitored together with leaf water content, carbon uptake method, and macronutrient and micronutrient content.

3.0.5.3 Abiotic stress

The climate variables considered are described in Chapter 2, and soil analyses were performed to determine if there were environmental factors which were directly contributing to health and response of study species and thus potentially influencing secondary metabolite pathways.

3.1 Field sites

Due to the nature of bulk sampling, it was necessary to locate sites where sufficient plant material could be collected without destroying the local population. This led to the selection of the following three field sites:

3.1.1 Field Site 1

Field site one was located in a ravine within the Paulshoek village proper (Figure 3.1). This is the preferred habitat of *C. edulis*. This site is rocky with fine dark sand and is comparatively moist. Floristically, the landscape is dominated by unpalatable shrub species, and in particular, by *G. africana* as well as various Crassulaceae species. The area is openly exposed to grazing livestock although *C. edulis* appears not to be grazed.

Figure 3.1: **Field site 1.**

3.1.2 Field Site 2

Field site two was also located within the village of Paulshoek and was the major collection site, also known as the "camp site" (Figure 3.2). This site was chosen because four of the five species of interest grew there and were protected from grazing. The area is approximately 6 ha in size and has been fenced off since 1996. Topographically, it is a south facing rocky slope with soil consisting of large-grained granite sands. Floristically, the landscape is dominated by a variety of succulent shrubs and is the most botanically diverse of the three sites. Field sites 1 and 2 are less than one kilometre from each other. While two *R. robusta* shrubs grow in the "camp site", they were not large enough to continuously collect 700 g of material, so a third field site was ultimately selected.

Figure 3.2: **Field site 2.**

3.1.3 Field Site 3

Field site three is several kilometres away from the village and is often referred to as "Slooitjiesdam" (S30 23 01.8 E18 19 43.1, altitude 1048.3m). This site is in an open communal grazing area and is flat with large-grained granite sands. Floristically, the area is dominated by *R. robusta* and various species of succulent subshrubs. The collection area was relatively far from the nearest stock post so the plants were mostly undisturbed (see Figure 3.3).

Figure 3.3: **Field site 3.**

3.2 Processing and analysis of plant material

Because the goal was to assess the stability of metabolite production for its potential use as a chemotaxonomic marker, and because secondary metabolites may not be constitutively produced, it was deemed necessary for the plants to be collected from wild populations. The climate and environment of Namaqualand are quite extreme, with wide temperature changes, strong winds, high altitude, etc., and it was therefore not possible to mimic the ecosystem in the available growth chambers. As the hypothesis was that the seasonal transition would give us the most extreme changes in metabolite production, it was decided that sample collection would have to occur in the native environment.

3.2.1 Collection of plant material

In the event that a particularly interesting metabolite was identified, and for future additional analyses, about 700 g of leaves were collected at a time for potential compound isolation. Because several of the species, *A. noctiflora*, *T. fruticosa*, and *T. wallichii*, drop their leaves in the summer months, leaves were only collected when available (see Figure 3.4 for specifics).

Figure 3.4: **Illustration of seasonal variation of aerial organs of selected species from April 2011 to March 2012.** [A] Due to the very small size of *G. africana* flower buds, flowers and fruits the data presented are rough estimates. [B] As the fruits of both Aizoaceae species and Crassulaceae species are capsules, immature fruits are presented. [C] Fruits were collected and eaten by local children. [D] Flower buds indistinguishable from leaves until in the lab.

In total, leaf material was collected at sunrise in the first week of each month over the period of one year from April 2011 to March 2012.

Only plant materials that did not display signs of mechanical damage or site specific discolouration, as would be associated with insect damage or microbial infection, were sampled. In cases such as the plant on the right in Figure 3.5, where betalain accumulation as defence against UV exposure is more evident in one plant than the other (Vogt et al., 1999), leaves were collected at a rate proportional to their distribution in the field. The material was then placed in black plastic bags and kept at room temperature until arrival at the university where material was kept at 4°C until being further processed.

Figure 3.5: **Two *R. robusta* shrubs from field site 3.**

Aspects of plant micro-environment (e.g differences in slope/gradient, sunlight and nursery effects, *inter alia*) were highly variable. In an attempt to rectify this, and to attain 700 g of leaves, leaf material was collected from at least 10 individuals within the population. Leaf material was then pooled resulting in the average fingerprint for any particular species in the field at that time point.

3.2.2 Laboratory processing of plant material

Once all of the leaf material was removed from the collected branches, 60-70 g of leaf material were selected at random for drying for 48 hr at 70°C. Large succulent leaves were slashed laterally to promote drying. Absolute leaf water content was measured during this time from the difference in mass upon drying. Dried leaves were milled by hand in a mortar and pestle to a fine powder and then mixed again. Milled leaves were sent to two analytical labs for analysis. Elemental micronutrients and macronutrients were determined by Bemlab (16 Van der Berg Cresent, Grant's Centre, Strand, 7137, South Africa; www.bemlab.co.za) as is briefly described in the following section. Also described below is the process used in the Stable Light Isotope Laboratory (Dept. of Archaeology, UCT) where leaf and soil samples were analysed for carbon isotope ratios as well as carbon and nitrogen content.

Approximately 500 g of the remaining leaf material from a specific collection were then shredded in a blender in enough ethanol (99.9%, Chemix, South Africa) to cover the plant material by several centimetres. The resulting slurry was then rung through cheese cloth and filtered through filter paper. The remaining solids were then re-suspended in ethanol and filtered again via the aforementioned process until the ethanol ran clear. The eluent was collected in clean recycled 2.5 L brown glass solvent bottles and stored at 4°C until used.

When flowers, or fruits of each species were available along with leaves, specimens were also collected for herbarium submission. Upon return from the field, specimens were frozen for 24 hr at −20°C and pressed in a Bolus Herbarium issued plant press. Blotting paper was replaced every 4-8 hr for the first two weeks to prevent mould contamination. Voucher specimens were identified by and deposited at the Bolus Herbarium, University of Cape Town, South Africa as well as at the McGregor Museum Herbarium in Kimberley, South Africa. For voucher information, see Table 2.2.

While ethanol is not as commonly used as methanol and chloroform in metabolomics experiments, it is still relatively common (for review of extraction methods see (Mushtaq et al., 2014)). As the extraction process eventually used over 400 L of solvent, it was deemed necessary to use a cheap and relatively environmentally friendly solvent. In addition, the project took place in the wider context of a wider project in collaboration with the Global Institute for Bio-Exploration (GIBEX, `http://www.gibex.org/`) which advocated the use of ethanol as a solvent suitable for use in field-deployable kits for analysis of the bio-activity associated with plants. The possible formation of metabolite artefacts arising from the use of ethanol, such as ethyl esters or glycosides, is explored later in this chapter and the limitations of ethanol in solubilising various metabolites is explored in Chapter 4.

3.2.3 Absolute water content(AWC)

Leaf water content directly reflects the effect of water availability to a plant. Because Namaqualand is a desert ecosystem, this was considered one of the more important markers of plant stress in the summer months.

Leaf samples were weighed before and after oven drying at 70°C for 48 hr and absolute water content determined using the following equation:

$$AWC = [\frac{Mass_{Wet} - Mass_{Dry}}{Mass_{Wet}}] \times 100 \quad (3.1)$$

3.3 Soil and leaf analyses

Elemental nutrient analyses on leaf and soil samples were performed by Bemlab agricultural analytical laboratory in Strand, South Africa. The following methods are a brief description of their protocols.

3.3.1 Macro and micronutrient analyses of leaf material

Milled leaf material was ashed at 480°C, then shaken in a 32% HCl solution before passing through filter paper (Kalra, 1998; Miller, 1998). The cation (K, Ca, Mg and Na) and micronutrient (B, Fe, Zn, Cu, Mn) content of the extracts were measured using a Varian inductively coupled plasma optical emission spectrometer (ICP-OES). Total Nitrogen content of the ground leaves was determined through total combustion in a Leco N-analyser.

3.3.2 Soil analyses

Because three study sites were used, soil analyses were conducted to determine general soil composition and to assess whether soil factors contributed to plant stress during the study period. Because of the halophytic nature of at least some Aizoaceae species (Winter et al., 1976), there was particular attention paid to salt concentrations as these might contribute to CAM transition in summer months. The methods are thus described in Table 3.1 were performed by Bemlab analytical laboratory.

Table 3.1: **Soil analyses as described by Bemlab.** Soil analyses were run on four samples collected from a depth of approximately 10 cm. Aizoaceae species tend to root in the top 50 cm of soil (Carrick, 2003). *G. africana* has a deeper tap root system (Carrick, 2003). Four sample were taken at a 10 cm depth.

Analysis	Method
Total P in soil	Total P was extracted with a 1:1 mixture of 1N nitric acid and hydrochloric acid at 80°C for 30 minutes. P concentration in the extract was then determined with a Varian ICP-OES optical emission spectrometer.
pH, P Bray II, and organic C	Soil was air dried, sieved through a 2 mm sieve and analysed for pH (1.0 M KCl), P (Bray II) and organic matter by means of the Walkley-Black method (Non-Affiliated Soil Analysis Work Committee, 1990). The extracted solutions were analysed with a Varian ICP-OES optical emission spectrometer.
Total NH_4^+ and NO_3^- concentration in soil	Ammonia and nitrate were extracted from soil with 1 N KCl and their concentrations determined colorimetrically on a SEAL AutoAnalyzer 3.
Extractable cations	Soil was sieved through a 2 mm sieve. Total extractable cations, namely K, Ca, Mg and Na, were then extracted at pH 7 with 0.2 M ammonium acetate (Non-Affiliated Soil Analysis Work Committee, 1990). The extracted solutions were analysed with a Varian ICP-OES optical emission spectrometer.
Soil texture (% clay, silt and sand) and Water Holding Capacity	Chemical dispersion was analysed using sodium hexametaphosphate (Calgon, Pennsylvania, USA) and three sand fractions were determined through sieving (Non-Affiliated Soil Analysis Work Committee, 1990). Silt and clay were determined via sedimentation rates at 20°C, using an ASTM E100 (152H-TP) hydrometer (Seta, Surrey, UK). Soil water holding capacity was determined mathematically from the soil texture using a calculation model adapted from Saxton and Rawls (2006).

3.3.3 Stable isotope analysis

Carbon and nitrogen isotope analysis and C and N content were assessed by the University of Cape Town's Stable Light Isotope Facility in the Department of Archaeology. These analyses were done in order to assess various biological parameters associated with carbon and nitrogen content, and to assess isotope distribution for the analysis of accurate masses as determined by MS. The following is a brief description of the standard protocols used.

3.3.3.1 Leaf preparation

Finely milled dry leaves were analysed directly for carbon and nitrogen content and carbon and nitrogen isotope ratios.

3.3.3.2 Soil preparation

Soils were analysed for organic and inorganic carbon content as well as total carbon and nitrogen content and carbon and nitrogen isotope ratios. Four soil samples were collected at 10 cm depth from each site. Approximately 250 mL of soil from each soil sample was dried down at 70°C for 48 hr. Samples were then filtered through a 1 mm sieve and a portion set aside for total nitrogen analysis as well as isotope analysis. For measurement of organic carbon, approximately 100 g of dried and filtered sample were washed three times in 1 M HCl (150 mL) at 4°C. For measurement of inorganic carbon, approximately 100 g of dried and filtered sample were washed three times in 3.5% sodium hypochlorite solution at 4°C. The inorganic and organic carbon samples were then dried down again at 70°C for 48 hr. The nitrogen, inorganic carbon, and organic carbon samples were then measured directly and independently.

3.3.3.3 Analysis

Dried samples were weighed into tin cups to an accuracy of 1 microgram on a Sartorius micro balance. Sealed samples were combusted in a Flash EA 1112 series elemental analyzer (Thermo Finnigan, Milan, Italy). The gases were passed to a Delta Plus XP IRMS (isotope ratio mass spectrometer) (Thermo electron, Bremen, Germany), via a Conflo III gas control unit (Thermo Finnigan, Bremen, Germany).

The in-house standards consisted of the following:

1. Sucrose (Australian National University, Canberra, Australia)
2. Merck Gel (Merck)
3. Lentil (Dried lentils from local supermarket)

All of the in-house standards had been calibrated against International Atomic Energy Agency (IAEA) standards. Nitrogen content is expressed in terms of its value relative to atmospheric nitrogen, while carbon is expressed in terms of its value relative to Pee-Dee Belemnite. Carbon isotope ratios were determined using the following formula:

$$\delta^{13}C = [\frac{\frac{^{13}C}{^{12}C}Sample - \frac{^{13}C}{^{12}C}Std}{\frac{^{13}C}{^{12}C}Std}] \times 10^3 \tag{3.2}$$

where sample refers to the ratio of ^{13}C to ^{12}C of sample being analysed and "Std" refers to the ratio of ^{13}C to ^{12}C reference standard value.

3.4 Generating LC-MS fingerprints

High performance liquid chromatography (HPLC) coupled with mass spectrometry (MS) has become one of the most common techniques in analysis of metabolomics data (Zhou et al., 2012). The following section describes the method used for LC/TOFMS experiments on crude ethanolic extracts of plant material from all of the species. This section will also explain the data pretreatment work flow. Unfortunately, the computational power available was unable to process the Aizoaceae species and the Crassulaceae species at the same time, therefore only the LC-MS profiles of the Aizoaceae species will be discussed.

3.4.1 Sample preparation

15 mL of each extract was centrifuged for 15 min at 20,000 rpm on a Beckman Avanti J-E ultracentrifuge (Palo Alto, USA) at 4°C and the supernatant was removed to fresh tubes. 500 μL aliquots were then dried down over 10 hours in pre-weighed 2 mL eppendorf tubes and reweighed to determine the soluble solid concentrations of each sample. 500 μL aliquots were then adjusted to 20 mg/mL with additional ethanol based on their soluble solid concentrations.

3.4.2 HPLC

A liquid phase gradient was optimised for *G. africana* due to its unique compound distribution and was universally applied to all samples for comparison (see Figures 4.10). The methodology was based on previous experiments by Farag et al. (2012); Li et al. (2011); Falleh et al. (2011b) and was performed on an Agilent 1290 Infinity series high performance liquid chromatography system with the following settings.

Table 3.2: **HPLC method.**

Part	Setting
Injection volume	5 μL
Solvent A	Reverse osmosis water made to analytical grade using a Millipore Milli-Q water system (Bedford, MA) + 0.1% analytical grade formic acid (Fluka, Switzerland)
Solvent B	Analytical grade acetonitrile (Honeywell Burdick and Jackson, USA) + 0.1% analytical grade formic acid (Fluka, Switzerland)
Mobile phase	3% solvent "B" held for 2 min, gradient to 90% "B" over 30 min, held for 2 min, and equilibrated for the next run with a gradient to 3% "B" over 2 min.
Solid phase	Agilent Poroshell 120 EC-C18 threaded column - 4.6 mm by 150 mm with 2.7 μm C18 particles (with compatible C18 Agilent guard column)
Flow rate	0.3 mL/min
UV-Vis λ monitored	210, 230, 250, 270, 280, and 340 nm

3.4.3 Mass spectrometry

The eluent from HPLC was then passed to an Agilent JetStream electrospray ioniser (ESI) with the following settings.

Table 3.3: **Electrospray ioniser specifications.**

Part	Setting
Nozzle voltage	1kV
Desolvation gas flow	8 L/min (300°C)
Sheath gas flow	11 L/min (350°C)

Ionised compounds were then passed to a time of flight (TOF) mass analyser with the following settings.

Table 3.4: **Mass analyser specifications.**

Part	Setting
Detector type	TOF
Ion mode	(+)
Mass range for collection	200-1700 m/z
Reference masses	121.050873 m/z (Purine)
	922.009798 m/z (HP-0921)
Scan rate	1/sec

Compound m/z ratios were finally recorded in centroid mode rendering a 0.7 Gb file for each sample.

3.5 Pretreatment of LC/TOF-MS data

Pretreatment was carried out using the open source platform, MZmine. Full details of this pretreatment are included here because preprocessing and data handling is critical for retrieval of suitable data from MS traces and is a process which is poorly documented in the literature for untargeted metabolomics approaches. The protocol followed is outlined in Figure 3.6. Additionally the method here developed was used to process the LC-MS data in the theses of Wheat (2014) and Dace (2014).

In addition to MZmine, selected R packages were utilised for various MZmine aplications: for baseline correction, “rJava”, “ptw”, and “gplots” were used and for the generation of peak lists and peak annotations, “xcms”, and “CAMERA” were used.

Processes indicated in blue in Figure 3.6 will be discussed in this section, final processing and a review of the data generally will be examined in Chapter 4 and an analysis of results in light of the total study aims will be covered in Chapter 5.

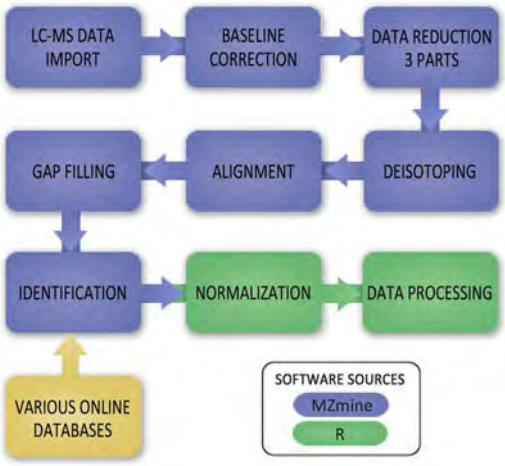


Figure 3.6: **LC-MS raw data preprocessing work flow.** The figure above displays the LC-MS data preprocessing work flow schematic as adapted from Want and Masson (2011). Method development was also guided by Katajamaa and Oresic (2005); Katajamaa et al. (2006); Pluskal et al. (2010, 2012).

3.5.1 LC-MS data import

In the first step, the LC-MS data files were converted from Agilent Technologies’ proprietary “.d” format to the open source “.mzdata” file format. This was done using Agilent Technologies’ MassHunter Workstation Qualitative Analysis software version B.05.00. Post-conversion, files were loaded into MZmine version 2.10. All operations were carried out on a 64-bit Windows 7 Enterprise operating system.

Due to the high RAM use of MZmine, a computer cluster was needed to analyse all of the Aizoaceae leaf samples. Samples were analysed on a 56 core, 128 GB RAM, 7 machine, rocks cluster. Even with the cluster, attempts to add the Crassulaceae samples proved impossible.

Total Ion Chromatograms (TICs) cover 34 min of run time and contain 2053 mass spectroscopy readings (see Figure 3.7). The TICs represent the sum of the ion intensities of each MS scan with a line drawn between each scan.

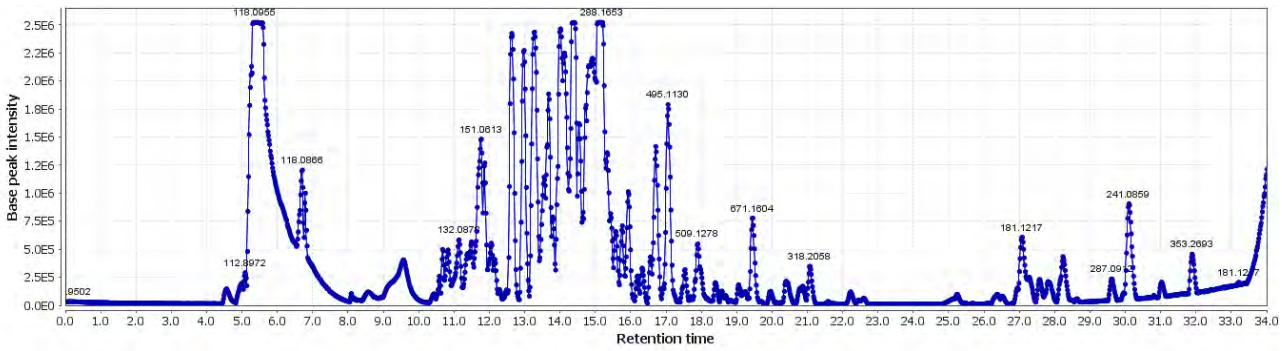


Figure 3.7: **Representative TIC.** Where each blue dot represents a scanning point across 34 min of run time.

3.5.2 Baseline correction

Once the files were loaded into MZmine, LC-MS data preprocessing was initiated with baseline correction. This was done in order to increase resolution especially in the later part of the runs where baseline tapering consistently occurs. Baseline correction was conducted using the following smoothing and binning parameters.

3.5.2.1 Smoothing

Smoothing was conducted to reduce noise from a measured spectrum. MZmine utilises the Savitzky-Golay filter for this purpose which reduces the total noise through the preservation of high-frequency components, this also helps to maintain peak shape (Zhou et al., 2012). The higher the smoothing value, the smoother the baseline. A baseline of 1,000,000 was selected.

3.5.2.2 Binning

Binning is used to increase the signal to noise ratio. To this end, each scan was divided into bins of 0.1000 Da as was reported by Tautenhahn et al. (2008). This was a particularly computationally heavy step, but when attempts were made to correct the baseline without binning, it retained much of the noise previously seen.

3.5.3 3 part data reduction

In order to use the mass spectral peaks they must be sorted from a TIC to extracted ion chromatograms (EIC) which breaks the full chromatogram into signals representing individual chemical components.

3.5.3.1 Mass detection

The mass detection algorithm separates each scan into an ion list. The centroid algorithm was used to process the data as ion detection was run in centroid mode. The noise level was selected at an intensity of 200 as visual assessment across samples from each species revealed that this removed the majority of noise from the data. Figure 3.8 serves as an example where the noise threshold is visualised as ions above (red) and below (blue) the acceptance threshold.

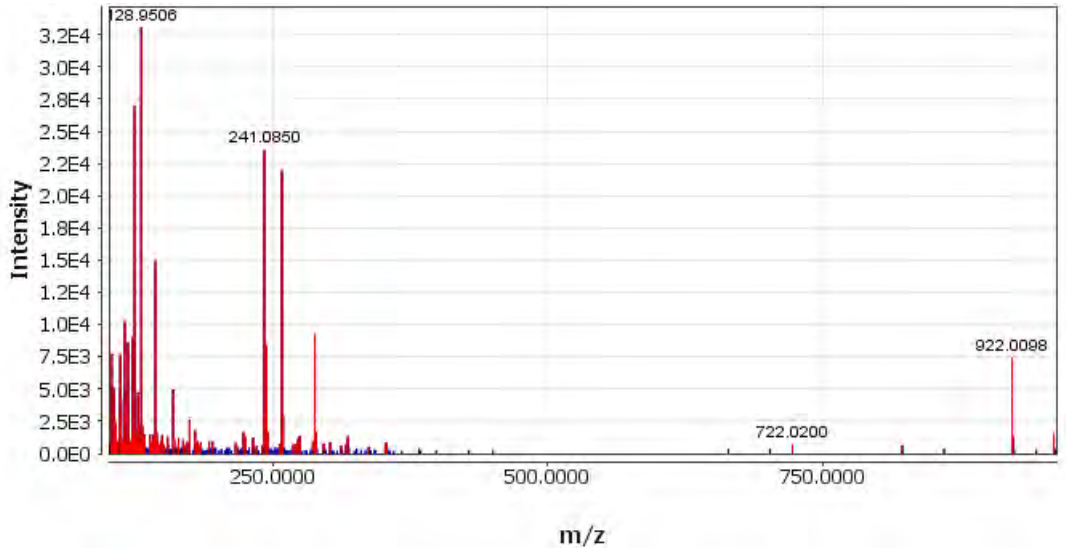


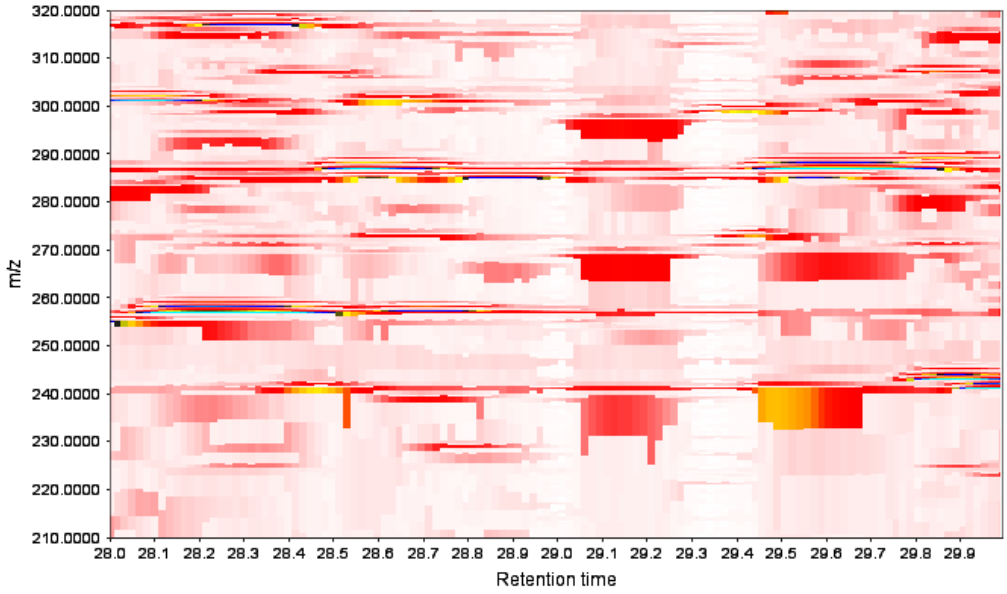
Figure 3.8: **Representative mass detection of a single scan of a leaf sample.** The red lines represent ion intensities above the threshold and the blue lines those signals below the threshold; the latter were removed in this step.

3.5.3.2 Chromatogram builder

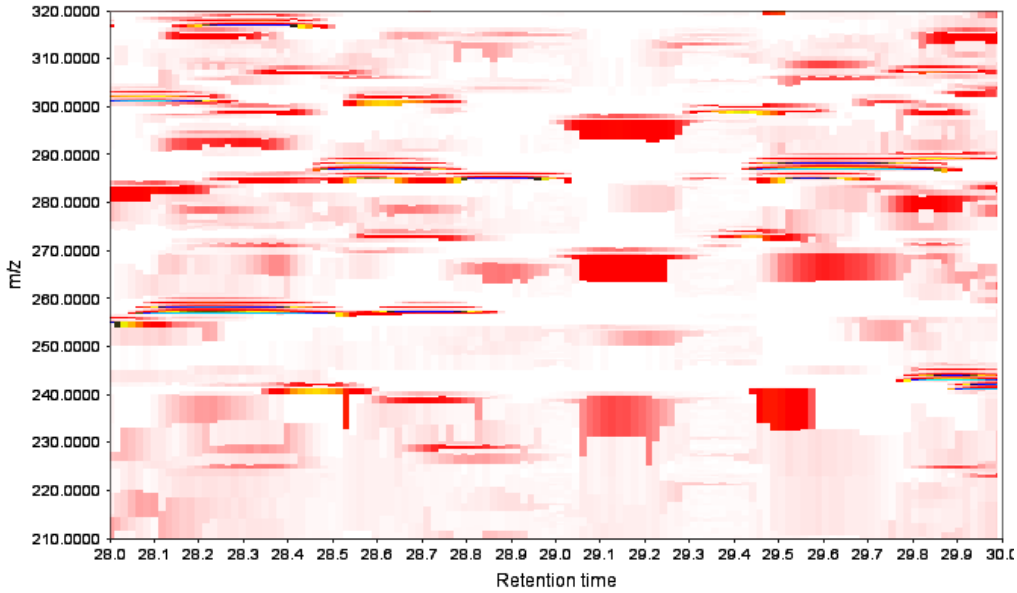
The chromatogram builder takes the lists generated in the mass detection step and creates chromatograms based on the continuous appearance of ion masses over consecutive scans.

The minimum time span was established at 0.1 min. While ideally this value would be slightly lower, anything below 0.1 min was too computationally expensive to run. This value was also established as sufficiently significant by comparing chromatographic peaks in a two dimensional comparison. The minimum peak height was selected at an intensity of 200, and any peak below this intensity was discarded. The maximum difference between chromatograms for them to be considered the same, was 5.0 ppm.

In Figure 3.9a, peaks are almost continuous across the represented retention time range with areas of greater and lesser intensity. This 2D TIC is also quite noisy as is represented by an almost continuous pink background. In Figure 3.9b, peaks have become isolated as single entities and the pink haze has mostly been removed from the background indicating that only continuous signals above the intensity threshold remain.



(a) *Before chromatograms were linked.*



(b) *After chromatograms were linked*

Figure 3.9: **Peak shape as determined by chromatogram building.** A section of the TIC is presented as an example of the chromatogram before (A) and after (B) chromatogram building. Blue regions represent the highest ion intensity and red the lowest ion intensity while white represents the absence of a signal.

3.5.3.3 Chromatogram deconvolution

Chromatogram deconvolution was then utilised to develop extracted ion chromatograms (EICs) by separating peaks from each other using a local minimum search. The threshold for removing noise was 65% of the chromatographic intensity which means that ions which were below 65% of the intensity of the highest intensity peak with the same mass were removed. A retention time range was limited to peak minimums of at least 0.1 min. Peak height was constrained as a minimum 5% of the total height with an absolute height threshold at an intensity of 50. To assist in cases where the chromatogram was not smooth, or to reduce noise, the minimum ratio between each peak's height and base width was set to 2.

In Figure 3.10, the highlighted peaks are the highest intensity ion peaks within the mass and retention time thresholds. The ions which are below 65% of peak intensity are not highlighted and were removed in this step.

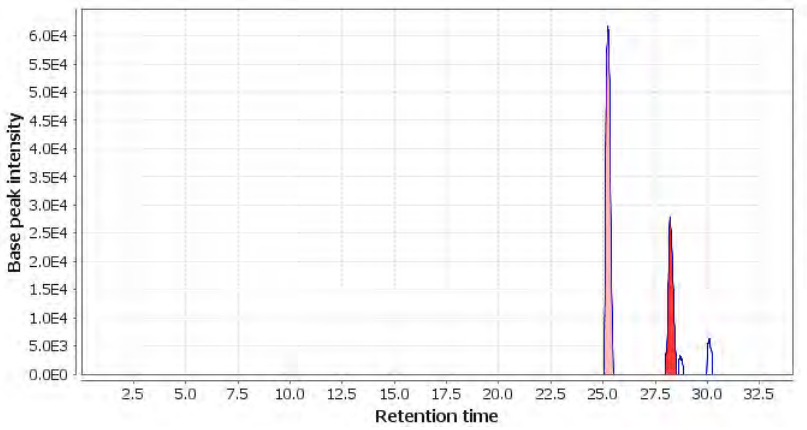


Figure 3.10: **Representative deconvolution of ion spectrum from leaf sample.** Shaded peaks were included in the resulting EIC list while non-shaded peaks were removed.

3.5.4 Deisotoping

This algorithm detects ion peaks with different masses which represent isotope variation and groups them together into single ion peaks. As seen in the isotope analysis in Figures 4.4a and 4.4b, ^{13}C and ^{15}N concentrations varied in abundance in the leaf tissues. In particular, ^{15}N in *C. edulis* was between 22-31%, which indicates that any N containing compounds in this species in particular would have substantial ^{15}N ion peaks. Compounds with identical molecular formula, and varying only in the isotopic composition of the constituent elements, should have identical retention times. Deisotoping parameters were set with a 0.5 min retention time range and a maximum Da change of 2 Da. The final output were monoisotopic masses for each ion.

3.5.5 Summary of ions prior to total alignment

Prior to aligning the profiles, the ion distribution of the extracts for each plant species was established as is shown in Figure 3.11. If sample preparation had been ineffective between collections, the ion number between samples would vary greatly. However, as can be seen from the low overall variability between samples of each species as determined by the standard error of the mean (see error bars in Figure 3.11), sample standardisation appears to have been effective.

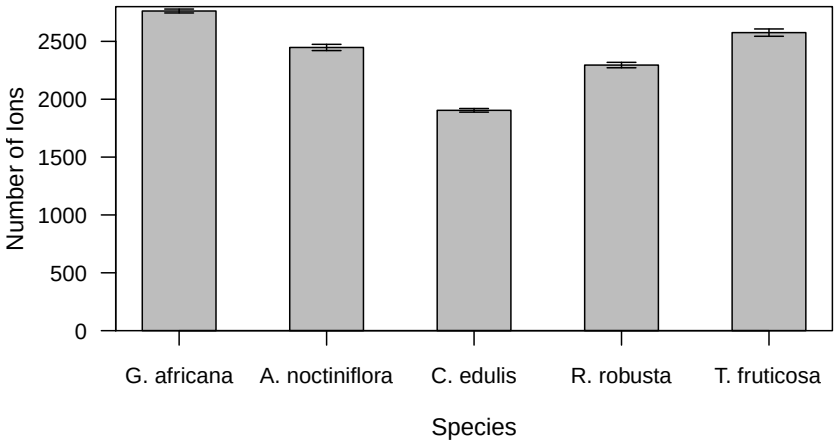


Figure 3.11: **Comparison of total number of ions in each species and across all Aizoaceae leaf samples.** The above error bars were established from the standard error of the mean (SE). The highest SE was found in *T.fruticosa* at 32 ions and the least from *C. edulis* at 16 ions.

The greatest number of ions was found in *G. africana* with an average of 2763 and the least from *C.edulis* with 1905. These numbers are in agreement with the range of values expected from previous reports in the literature for other species. Total variation over time was negligible which suggested that total ion comparison in the form of gap-filling was appropriate.

3.5.6 Alignment of ions across samples and species

In order to compare ion profiles of individual plant extracts to each other, the ion peaks of each had to be aligned. To determine if ion peaks across all of the samples were identical, each comparison was given a score determined by the mass and retention time

of each peak using pre-set tolerances (Katajamaa and Oresic, 2005). As was previously done, mass tolerance was selected at 5.0 ppm and weighted at 50%. The retention time tolerance was set rather loosely at 3.0 min to allow for chromatogram drift with the weighting set at 10%. As part of the consideration for alignment, the same charge state was also required.

The product of this step is somewhat disconcerting as the ion numbers change from at most about 3,000 ions in each species to 23,000 ions across all of the species. The alignment parameters were left at 1 min, which is fairly generous as far as retention time is concerned, and there were peaks in the list that apparently should have combined, but do not appear to have. Due to previously established methodologies, the mass tolerance was left at 5 ppm which is admittedly strict. As this is the standard by which compound IDs were judged, it was necessary to weight the mass heavily.

3.5.7 Filling in the gaps at or below threshold

The aligned ion lists indicate the presence or absence of each ion in each sample at the specified thresholds. Gap filling searches the aligned ion data in an attempt to locate peaks below the set threshold. This is critical in work looking for unique ion peaks as the difference between metabolites produced at a low concentration and no production of a metabolite is statistically significant and different. Mass tolerance was set to 5.0 ppm with a retention time tolerance of 3 min. The maximum allowed deviation from the peak shape, or the intensity tolerance, was set at 20%.

3.5.8 Analysis of internal reference standards

As part of the in-house referencing system built into the MS, internal reference standards were continuously injected into the mass detector along with the sample being analysed. The UCT internal reference standards have masses of 121.0509 and 922.0089 Da. From a general mass search of the TIC (representative Figure 3.12 from all G. africana samples), it is evident that the internal reference standard masses were present in the samples across the entire retention time range at relatively low intensities.

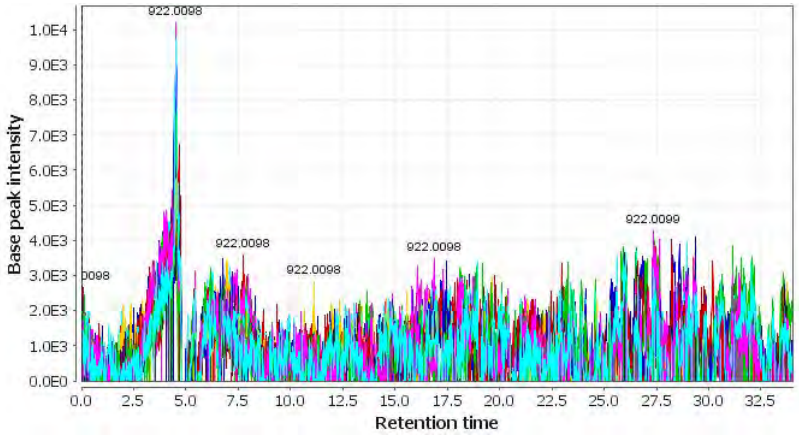


Figure 3.12: **TIC of the mass 922.0089 representing the HP-0921 standard.** To show that this is common to all samples, all *G. africana* leaf sample TICs at this mass are overlaid.

From the resulting ion list, there were 26 masses that appeared within 1 ppm of 121.0509 from a RT range of 4.0 min to 32.6 min. There were also 22 masses within 1 ppm of 922.0089 from a RT range of 4.1 to 30.5 min. The constant presence of these internal reference masses and their accurate mass recordings before and after the data was preprocessed confirmed that the mass analyser was working and that preprocessing was effective.

3.6 Database of compounds from Aizoaceae literature and plant primary metabolism

Not surprisingly, it was found that when the final ion list was inserted into online databases for compound identification, one mass would result in many hits. While some of these identifications were obviously not relevant, such as those associated with synthetic pharmaceuticals, there were still too many hits for the same mass to analyse. To increase the probability of achieving true compound hits, a database of compounds that have been previously identified in any Aizoaceae species was created. Over 30 papers focusing on metabolite composition of Aizoaceae species were considered. This compound list primarily consists of secondary metabolites isolated in NPC studies. A second list of common plant primary metabolites was also constructed consisting of amino acids, organic acids, carbohydrates, and sugar alcohols.

Because the plant metabolomes were extracted in ethanol, there is a possibility that esterification occurred in metabolites containing carboxylic acid groups. There is also the possibility of ethyl glycosides forming from reduced sugars.

Typically primary metabolites such as amino acids, organic acids and carbohydrates are so polar that they move with the solvent front through a reverse phase column prior to detection. If esterification did occur, the resulting ethyl esters would be significantly less polar, have a higher mass, and might be detectable in the LC-MS profile.

In order to determine if this had happened, appropriate primary and secondary metabolites were esterified *in silico* and their appropriate adduct masses were added to the compound list (see Figure 3.13). Tables representing the metabolites and their adducts data are presented in Appendix A.

Identification of the plant metabolites was carried out using the guidelines defined by the Metabolomics Standards Initiative (see "Proposed minimum metadata relative to metabolite identification- Nomenclature for non-novel metabolites" in Sumner et al. (2007)). The accurate masses of the "M+H" and the "M+EtOH" ions of the compound library (in Appendix A, Table A.1 and Table A.2) were compared to the masses of the of the ions in the final processed ion list generated from the ethanol extracts of the plants. In cases where extract ion masses fell within 5 ppm of an expected compound mass, and the retention time of the ion fell within an expected range for that compound, compound identifications were considered likely. In cases where extract ion masses fell within 10 ppm, identifications were considered possible.

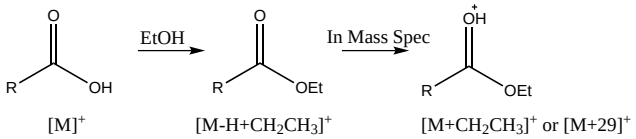


Figure 3.13: **Example of ions expected from the formation of ethyl esters.**

According to Sumner et al. (2007), compounds identified " without chemical reference standards, based upon physiochemical properties and/or spectral similarity with public/commercial spectral libraries" fall into the category of putatively annotated compounds. By this classification system (explored also in Dunn et al. (2013)) these would be type 2 identifications. A table of the compounds and the confidence of their identifications can be found in Appendix A.

Chapter 4

Results and discussion

The results of the analyses in Chapter 3 are described in the following chapter.

4.1 Leaf analyses

4.1.1 Phenology

Various features of plant growth and development were monitored over time including leaf and shoot growth, leaf yellowing and abscission, as well as sexual maturation, development, and seed dispersal (see Table 4.1 with accompanying key, Table 4.2). Due to the very small nature of *G. africana* and *R. robusta* flowerbuds, flowers, and fruits, it was difficult to distinguish the phenophases of the sexual organs, and the indicated phenophases are therefore estimates. While the phenological data was considered important for the contextualisation of various nutrient and physiological data, the imprecision of the measurement makes it impractical to use statistically. Binary designations were considered for statistical analysis (1 for present, and 0 for absent), but these were ultimately rejected because of the imprecision of the phenological measurements.

Generally, leaf and shoot growth are seen at the beginning of the rainy season (see Figure 2.5). *T. fruticosa* was the first species to form flower buds (June). *G. africana* had the longest period of sexual maturation (seven months, July 2011 - January 2012) as compared to the other study species. *R. robusta* had the shortest sexual maturation with flower buds forming and fruits maturing within two months. *A. noctiflora* began to develop sexual organs in September around the same time that it transitioned from C3 to CAM metabolism (see Figure 4.4a). Leaf yellowing also coincides with reductions in C and N content (see Figures 4.2e and 4.2f).

Table 4.1: **Phenology across the study period for all of the species.** Phenology was monitored in the field from April 2011 to March 2012.

Plant	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar
G. africana	LG, SG	LG, SG	LG, SG, LY	LG, SG, FB, LY	LG, SG, FB, LY	LG, SG, FB, LY	LG, SG, FB, LY	FB, LY, LA	FB, FL, LY, LA	FL, IF, LY, LA	MF, SD, LY, LA	LY, LA
A. noctiflora	LG, SG, MF, SD	LG, SG	LG, SG	LG, SG	LG, SG	LG, SG, FB	LG, SG, FB. FL, LY	FL,IF, LY, LA	IF, LY, LA	MF, SD	MF, SD	SD
C. edulis	LG, SG	LG, SG	LG, SG	LG, SG	LG, SG	LG, FB	LG, FB, FL, IF	LG, FB, FL, IF	LG, FB, FL, IF	LG, MF	LG,	LG
T. fruticosa	LG, SG	LG, SG	LG, SG, FB	LG, SG, FB, FL	LG, SG, FB, FL	LG, SG, FB, FL, IF	FB, FL, IF, LY	IF, MF, SD, LY	IF, MF, SD, LY, LA	MF, SD, LY, LA	MF, SD, LY, LA	MF, SD, LA
R. robusta	LG, SG, MF, SD	LG, SG, MF, SD	LG, SG, MF, SD	LG, SG, MF, SD	LG, SG, MF, SD	LG, SG, MF, SD	LG, MF, SD	LG, FB, FL, IF, MF, SD, LY	IF, MF, SD, LY	MF, SD, LY, LA	MF, SD, LY, LA	MF, SD, LY, LA
C. orbiculata	LG, FL, IF, MF, SD, LY	LG, FB, FL, IF, MF, SD, LY	LG, FB, FL, IF, MF, SD, LY	LG, SG, FB, FL, IF, MF, SD, LY	LG, SG, FB, FL, IF, MF, SD, LY	LG, SG, FB, FL, IF, MF, SD, LY	LG, IF, MF, SD, LY	SG, FB, MF, SD, LY	FB, FL	FB, FL, IF, LY, LA	FB, FL, IF, LY, LA	FB, FL, IF, MF, SD, LY, LA
T. wallichii	LG, MF, SD	LG, MF, SD	LG	LG	LG	LG	LY	SG, FB, LY	FB, LY	FB, FL, IF, LY, LA	FL, IF, LY, LA	IF, LA

Table 4.2: Phenology key.

Leaf growth	Shoot growth	Flower buds	Flowers	Immature fruits	Mature fruits	Seed Dispersal	Leaf yellowing	Leaf abscission
LG	SG	FB	FL	IF	MF	SD	LY	LA

4.1.2 Leaf water content

While leaf material was collected around the same time of day at each collection, the error bars shown in Figure 4.1 may be due normal water content fluctuations as were seen in *M. crystallinum* (Winter et al., 1978). It is common in deciduous species to see decreases in the overall leaf water content before the leaves abscise, which was the case in *T. fruticosa* which went through total leaf abcision, and in *R. robusta* and *G. africana* which went through partial leaf abcision. The most succulent species, *A. noctiflora*, *C. edulis*, *C. orbiculata*, and *T. wallichii* retained water more effectively, although *A. noctiflora* did experience complete complete leaf abcision.

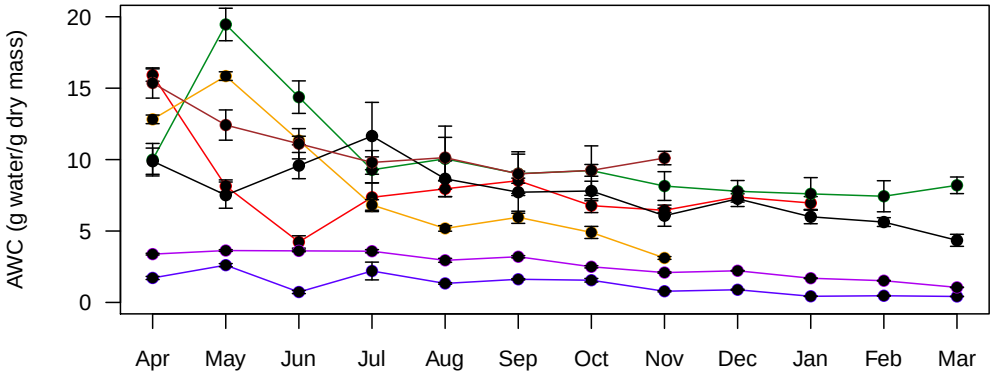


Figure 4.1: **Absolute water content of leaf material across all study species from 2011-2012.** Data sets for the Aizoaceae species are designated by differently coloured lines, blue for *G. africana*, red for *A. noctiflora*, green for *C. edulis*, purple for *R. robusta*, orange for *T. fruticosa*, and the Crassulaceae species are black for *C. orbiculata* and brown for *T. wallichii*. Error bars represent the standard deviation between replicates ($n \geq 7$).

4.1.3 Macronutrients and micronutrients from leaves

4.1.3.1 Elemental macronutrient analysis of leaves

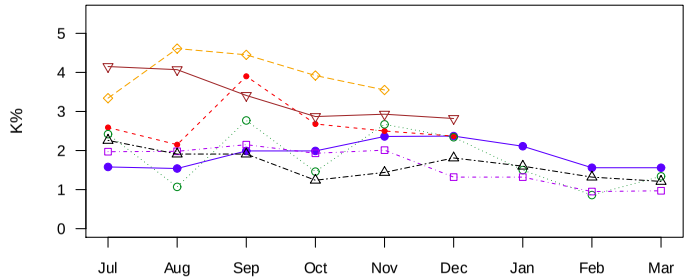
Due to the relatively high overall concentration of macronutrients in plants, these levels were determined on a % dry mass basis. Because micronurients tend to be present at trace levels only, they are expressed in parts per million (ppm). Of the data presented in Figures 4.2 and 4.3, all of the analyses were run by Bemlab starting in July 2011 with the exceptions of C and N which were determined by UCT's Stable Light Isotope Lab beginning in April 2011.

For the Aizoaceae species, all of the macronutrient and micronutrient values measured fell within generic plant values with the exception of Mn (Figure 4.3b) and Na (Figure 4.3e) where usual ranges are reported to be between 10-100 ppm for Mn (Hänsch and Mendel, 2009) and above 2500 ppm for halophytes (Pilon-Smits et al., 2009). Mn and Na uptake may be selective as they can be substituted for K (Pilon-Smits et al., 2009). Na can help to facilitate nitrate uptake by acting as a counterion and considering the phenological phase that the plants were at at the time of peak Na uptake, i.e. rapid leaf and shoot growth, this appears to be a possibility. Several Aizoaceae species are known to be halophytes, which may explain Na tolerance (Winter et al., 1976) allowing for $> 0.25\%$ of the mass of a plant to be Na (Pilon-Smits et al., 2009). This appears to be ubiquitous in the Aizoaceae species and within them across multiple months, but not the Crassulaceae species. These observations suggest that these concentrations are unique to the two months where it occurs rather than being an artefact of analysis at different times. Interestingly, *C. edulis* which grew in the most saline soil (see Figure 4.4), had the third highest concentration of Na further suggesting that Na is selectively taken up by the Aizoaceae species.

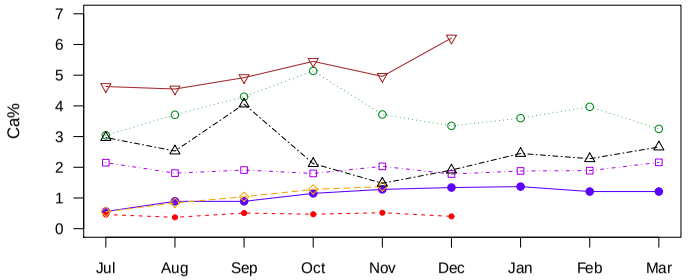
The least succulent species of Aizoaceae, *G. africana* and *R. robusta* appear to have the highest overall carbon content (see Figure 4.2e). Interestingly, the Crassulaceae species show higher overall carbon content than the Aizoaceae species with a similar absolute water content. This is probably due to the carbon distribution throughout the succulent matrix of the Crassulaceae leaves in contrast to the Aizoaceae species where the succulent matrix is mainly water.

The leaves of the species that abscise completely, *viz. A. noctiflora*, *T. fruitcosa*, and *T. wallichii* (see Table 4.1), do not completely dehydrate prior to abscission (see Figure 4.1), the reduction in carbon content immediately before onset of leaf abscission is probably an indication that the plants are reclaiming chlorophyll and other carbon structures in preparation for leaf abscission (see Figure 4.2e). This is consistent with the extent of leaf yellowing occurring at the same time (see Table 4.1).

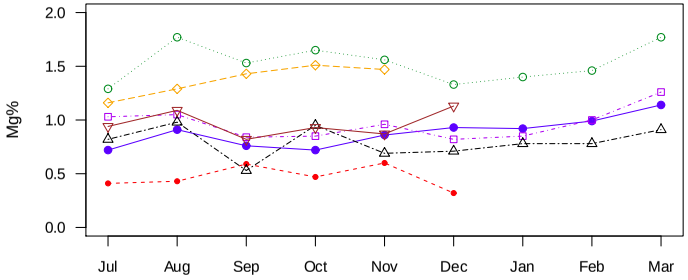
High nitrogen content is associated with high levels of nutrition for herbivores and therefore herbivore preference. *T. fruticosa*, which is considered the most palatable plant in the study (see Table 2.3), contains the most nitrogen (see Figure 4.2f) (Mattson, 1980).



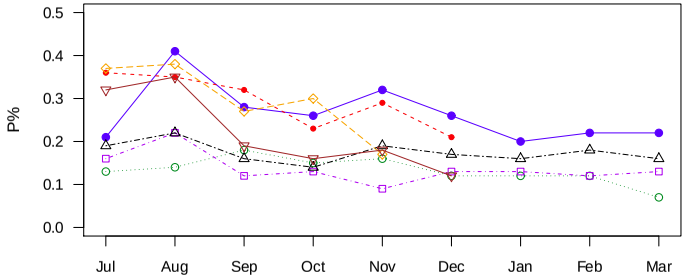
(a) Potassium content of leaves(% dry mass).



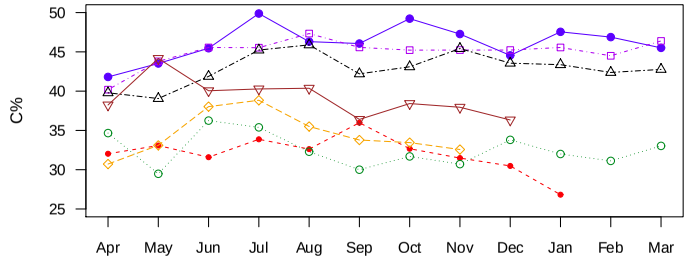
(b) Calcium content of leaves(% dry mass).



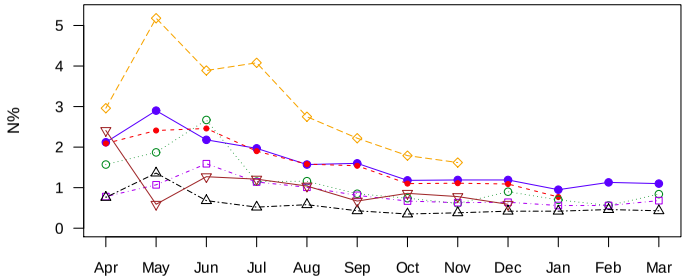
(c) Magnesium content of leaves(% dry mass).



(d) Phosphorous content of leaves(% dry mass).

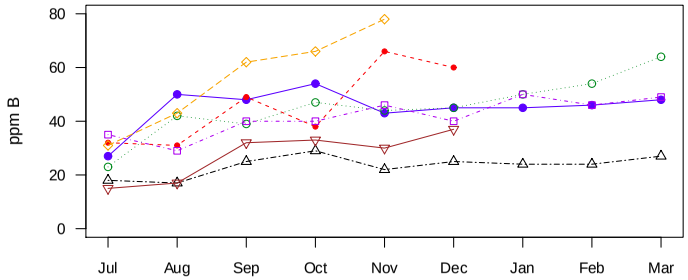


(e) Carbon content of leaves(% dry mass).

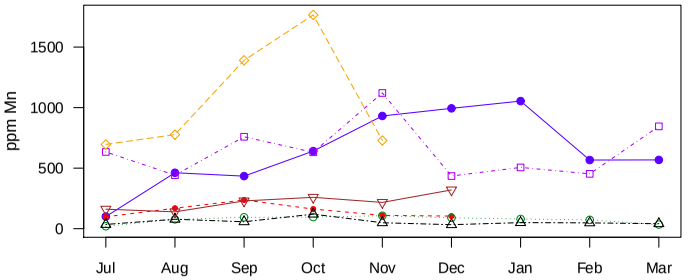


(f) Nitrogen content of leaves(% dry mass).

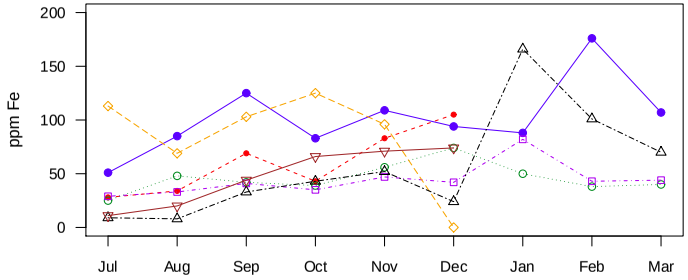
Figure 4.2: **Leaf macronutrients across all study species from 2011-2012.** Data sets for the different Aizoaceae species designated by differently coloured lines: are blue for *G. africana*, red for *A. noctiflora*, green for *C. edulis*, purple for *R. robusta*, orange for *T. fruticosa*, and the Crassulaceae species are black for *C. orbiculata* and brown for *T. wallichii*.



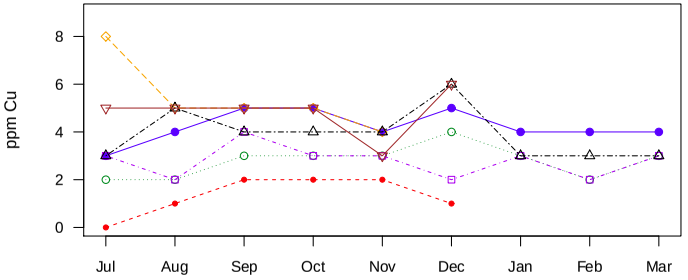
(a) Boron content of leaves(ppm dry mass).



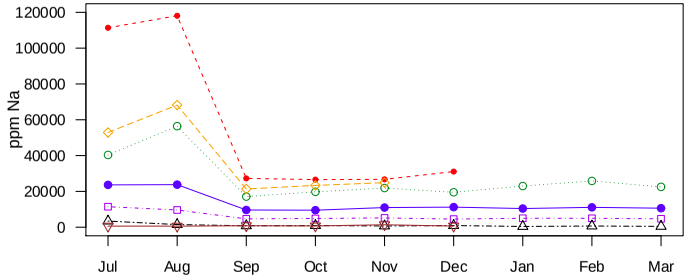
(b) Manganese content of leaves(ppm dry mass).



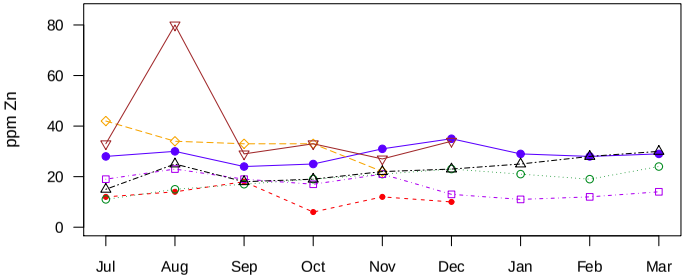
(c) Iron content of leaves(ppm dry mass).



(d) Copper content of leaves(ppm dry mass).



(e) Sodium content of leaves(ppm dry mass).



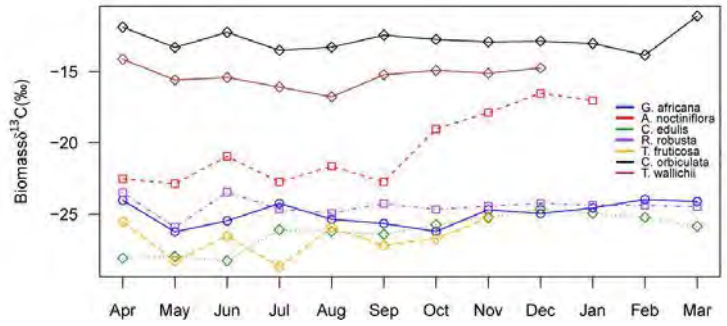
(f) Zinc content of leaves(ppm dry mass).

Figure 4.3: **Leaf micronutrients across all study species from 2011-2012.** Data sets for the different Aizoaceae species designated by differently coloured lines: blue for *G. africana*, red for *A. noctiflora*, green for *C. edulis*, purple for *R. robusta*, orange for *T. fruticosa*, and the Crassulaceae species are black for *C. orbiculata* and brown for *T. wallichii*.

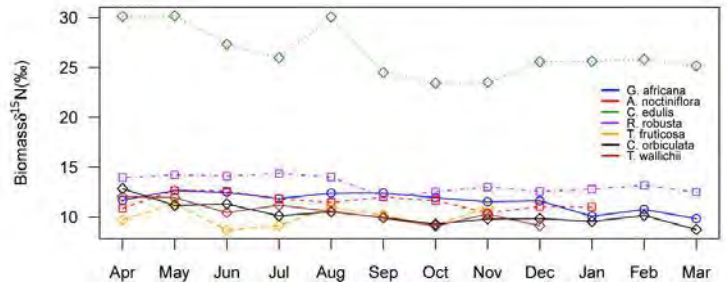
4.1.4 Stable light isotope analysis

Some members of the Aizoaceae family are known to utilise facultative CAM, which means that they can use two different carbon fixation methods. Phosphoenylpyruvate carboxylase (PEPC), the enzyme that fixes carbon in CAM uses ^{13}C isotopes preferentially in comparison to Ribulose bisphosphate carboxylase oxygenase (RubisCO), the enzyme that fixes carbon in C3, thus providing a mechanism for determining carbon fixation (Smith, 1972). *C. orbiculata* and *T. wallichii* are obligate CAM species and therefore consistently show a higher level of ^{13}C isotope fixation and in this instance were used as field controls for the other species to facilitate the determination of transition (see Figure 4.4a).

Of all the Aizoaceae species studied, *A. noctiflora* appears to be the only one which transitions towards CAM carbon fixation, suggesting that it is a facultative CAM species. This trend mirrors the findings of Winter et al. (1978) on wild *M. crystallanum* populations, in terms of when it transitions as well as the time taken to move from ^{13}C isotope ranges indicative of C3 carbon uptake to ^{13}C ranges indicative of CAM carbon uptake (Winter et al., 1976). On the other hand, previous studies have shown *C. edulis* to also be capable of facultative CAM (Winter et al., 1976; Herrera, 2008), so it was surprising that it did not appear to transition as suggested by data from the current study. This may be due to the riparian environment that it grows in where water levels are consistently higher than the rest of the field sites (data not shown), a particular factor over the study period due to the higher than average rainfall (see Figure 2.6).



(a) δ^{13}C Isotope discrimination across all species.



(b) δ^{15}N Isotope distribution across all species.

Figure 4.4: **Leaf isotope ratios across all study species.** Isotope ratios are expressed as δ^{13}C, or δ^{15}N respectively, in ppm with respect to aforementioned standards.

Because the vast majority of compounds in living organisms are carbon based, changes in the carbon isotope levels will ultimately change the accurate masses of many compounds, especially if they transition between different carbon uptake mechanisms. This may ultimately bias statistical analyses of carbon-based molecules which are based on high-resolution mass analysis, with some lower-intensity ions appearing to suggest lower concentrations of a component in a particular sample.

Unlike δ^{13}C, there are no reports of any explicit role for δ^{15}N in plant metabolic processes. However, because accurate mass is later used to identify compounds from metabolic fingerprints, it was thought useful to determine δ^{15}N content as could have a profound effect on the exact mass of of nitrogen containing compounds.

Typical δ^{15}N levels in plant material is 3-5%, so the δ^{15}N levels in the study species as indicated in Figure 4.4b, and especially in *C. edulis*, were extraordinarily high (Samson Chimphango, University of Cape Town, Biological Sciences Department, private communication, 2011). This is in line with the findings of Heaton (1987) who established the unusual N isotope ratios for plant material found along the western coast of South Africa.

4.2 Soil analyses

To ensure that there was no soil nutrient deficit stress or nutrient toxicity, various analyses were conducted. The results of the soil analyses were assessed with the assistance of Professor Sam Feagley of the Department of Soil and Crop Science at Texas

Agriculture and Mining University as well as assessment provided by Bemlab analytical laboratory.

4.2.1 Soil texture - % clay, silt, and sand

Mechanical analysis in Figure 4.3 revealed that the majority of soil particles across the three study sites were classified as sand with relatively large-grained particles. This is consistent with a previous analysis of the area by Allsopp (1999). Large grained sands are less likely to be able to hold onto water resulting in a more rapid onset of water deficit stress. From this, all nutrient value norm ranges are determined for sandy soil.

Table 4.3: **Mechanical analysis of soil samples.**

Site	Clay%	Silt%	Fine Sand%	Medium Sand%	Large Sand%	Class
1	12.67±0.94	4.00±2.82	31.70±1.87	23.93±4.33	28.10±0.86	Sand
2	12.67±0.94	6.67±0.94	31.07±1.36	19.53±1.31	30.47±1.76	Sand
3	10.00±0.00	2.00±0.00	30.83±4.50	23.03±0.74	34.53±3.77	Sand

4.2.2 Macronutrients, micronutrients, and total phosphorous

The generic plant norms represented in Table 4.4 were issued with the report from Bemlab and apply specifically to analysis of soils on wine farms. Because the plants used in this study are wild species, there are no generic norms for them, thus nutrient levels discussed are assessed first in light of the nutrient recommendations for wine farms and then by the nutrient concentrations in the leaves. As discussed in Hänsch and Mendel (2009), various nutrient concentrations are necessary for normal enzyme activity and general metabolism in plant leaves. The following discussion revolves around the specific nutrient levels that fell outside of wine farm soil norms in light of the nutrient concentration needed for metabolism. In the cases where values were found outside the recommendations of Bemlab results could not be examined in light of leaf levels, results were also reviewed externally by Professor Samuel Feagley.

Table 4.4: **Soil Data analysed from the three study sites.** Data shown are the average n=4±SD. Recommended values were described by a Bemlab analyst ("Normal Range") for soils encountered on wine farms. Values highlighted in red indicate concentrations above the recommended values for that measurement and those highlighted in blue represent concentrations below recommended values. Columns "1", "2", and "3" represent field sites 1, 2, and 3 respectively. Comments under "Diagnosis" suggest the potential implications of nutrient toxicity or deficit as indicated by red and blue highlights respectively.

Nutrient	Normal Range	1	2	3	Diagnosis
Cu	5-25 ppm	1.64±0.47	1.09±0.13	0.48±0.11	Cu deficiency
Zn	>1 ppm	5.73±2.75	2.63±0.95	1.17±0.26	Normal
Mn	5-60 ppm	71.567±34.02	114.07±56.35	72.60±16.14	Normal
B	1-3 ppm	2.33±1.36	0.47±0.03	0.34±0.06	B deficiency
Fe	50-150 ppm	81.94±48.46	127.47±34.66	33.15±5.93	Normal
Cl	<350	1661.093±70.90	20.33±7.08	13.24±2.41	Cl toxicity (salinity)
P	20-150 ppm	885.53±91.18	274.89±83.89	88.99±6.14	Decreased Zn uptake
K	70-120 ppm	276.00±120.41	141.67±37.28	105.00±19.44	Decreased Mg uptake
pH	5.5-6.5	7.500 ±0.37	5.17±0.97	5.30±0.46	Normal
Resist. (Ohm)	>400	83.33±4.71	2253.33±592.19	7350.00±1140.21	High salinity
C%	0.8-1.5%	0.52±0.32	0.41±0.04	0.20±0.09	Normal
$NO_3^- + NH_4^-$	6-15 ppm	3.97±1.40	3.10±0.58	2.58±0.13	Poor vegetative growth
Na^+	<1.2 cmol/kg	7.07±0.84	0.21±0.05	0.11±0.01	Soil compaction
K^+	0.1 cmol/kg	0.70±0.31	0.36±0.10	0.27±0.05	Normal
Ca^{2+}	2-6 cmol/kg	19.68±6.23	2.61±0.26	0.96±0.20	Potential salt stress
Mg^{2+}	0.5-2 cmol/kg	7.84±0.42	1.68±0.09	0.82±0.07	Salinity
Na%	<10%	20.61±3.90	3.96±0.74	4.68±0.10	Soil compaction
K%	4-6%	2.12±0.98	7.02±2.08	11.41±2.89	K requirement, Decreased Mg uptake
Ca%	65-85%	54.53±7.47	49.94±3.80	39.42±3.92	Ca deficiency
Mg%	15-20%	22.73±2.87	32.15±0.95	34.20±1.07	Soil compaction
T-value	<15%	35.29±6.33	5.22±0.32	2.42±0.28	Sodium brackish

Of the values tested, most fell within generic plant norms indicated by Bemlab. Cu concentrations were a little low (<2 ppm) which might suggest Cu deficit. Hänsch and Mendel (2009) indicate that the Cu range inside the plant tissues should be between 1-20 ppm for normal enzyme incorporation, which is true for all samples across all species suggesting that Cu concentration in the soil is sufficient for these species (see Figure 4.3d).

Boron levels should be between 1-3 ppm in the soil which is the case for field site 1 but <0.5 ppm for sites 2 and 3 making the soil in these areas slightly boron deficient. Hänsch and Mendel (2009) suggest that the B concentration for proper enzyme incorporation is 3-100 ppm in plant tissues. Figure 4.3a shows that all samples across all plant species have sufficient B uptake for enzyme use indicating that B concentration in the soil for these species is sufficient.

Cl concentrations at site 1 are about 5 times higher than they should be which is normally toxic to plants and is a marker of high soil salinity. P concentration $> 150 ppm$, as was also high at site 1 which can inhibit Zn uptake, but Zn concentrations in the leaves of *C. edulis* appear within the normal range (Hänsch and Mendel, 2009) suggesting that enzyme incorporation of Zn is sufficient at this soil P concentration (see Figure 4.3f).

4.2.3 Physical and chemical characteristics

Table 4.4 shows that pH at site 1 is higher than the normal value listed for wine farms, but is still within an acceptable range for plant growth generally (Samuel Feagley, personal communication, June 2013). The Resistance at site 1 is quite low, indicating high salt levels. %C, and NO_3^- levels were within normal ranges (Samuel Feagley, personal communication, June 2013).

4.2.4 Exchangeable cations (Ca^{2+}, K^{+}, Mg^{2+}, and Na^{+}) and base saturation (Ca%, K%, Mg%, and Na%)

Table 4.4 also indicates that the typical cation exchange capacity (CEC) of the four cations combined should be around 2-3 $cmol/kg$. Where study sites 2 and 3 are within the normal cation range, study site 1 has about 10 times higher levels than usual. This is due to elevated Ca^{2+}, Mg^{2+}, and Na^{+}, and clay content under 5% (see Table 4.3). The elevated exchangeable cation concentrations in soil 1 confirm the high salinity markers in the previous analyses.

Base saturation appears to be stable in sites 2 and 3, but is about 10 times higher than normal in site 1. As a percent of cation exchange, the T-value should be $< 15.00 cmol/kg$. This supports the previous diagnosis indicating that the soil at collection site 1 is highly saline while soil from sites 2 and 3 fall within normal ranges.

4.2.5 Stable light isotope analysis of soil

From Table 4.5, it is evident that organic carbon is below optimal at sites 1 and 3 (<0.5) which may result in increased soil compaction. The organic carbon content of field site 2 is within the ideal range (between 0.8 - 1.5%). Nitrogen content is slightly low (where normal values are between 0.06 - 0.15%) at all three field sites, and confirms the low $NO_3^- +$ and NH_4^- levels from the Bemlab analytical analysis seen in Table 4.4, however elemental N concentrations are not indicative of the bio-available N in soil. The ratio of inorganic to organic carbon at field sites 1 and 3 is almost 1 suggesting that the turnover of vegetative tissue is not particularly high in these areas. The ratio of organic to inorganic carbon at field site 2 is about 2 suggesting that field site 2 represents more stable vegetative turnover. This is further reflected in the C:N ratio where the nitrogen levels are slightly higher than ideal for effective microbial breakdown of the strata but are fairly typical of generic sandy soil (Saxton and Rawls, 2006).

Table 4.5: **Mass spectral analysis of soil C and N.** %C is the total C present in the sample and total organic C (TOC) and total inorganic C (TIC) are the % organic and inorganic C contents. Error measurements are based on the standard deviation of the samples, where $n = 3 \pm SD$ at each field site. Ideal soil values for total carbon, TOC, total N and C:N ratios were established be Bemlab analytical lab and confirmed by Samuel Feagley. The C and N isotope ratios were discussed in Trudell et al. (2004) and Peterson and Fry (1987).

Site	%N	$\delta^{15}N/^{14}N$	%C	% TOC	% TIC	C:N ratio
Ideal value	0.06-0.15%	−4-15	0.8-1.5%	0.5-1%	0.3-0.5%	10-15
1	0.035±0.019	19.833±3.092	0.518±0.236	0.332±0.180	0.374±0.093	15.144±1.635
2	0.050±0.006	13.406±0.620	0.666±0.051	0.993±0.665	0.425±0.083	13.558±1.670
3	0.021±0.005	15.893±0.225	0.220±0.095	0.149±0.039	0.161±0.106	10.136±1.844

Carbon isotope ratios fall within the range of C3 plants reflecting the carbon cycling of these species through the soil, as plants take in carbon from the surrounding atmosphere, we would not expect soil levels to reflect leave levels. Soil nitrogen isotope ratios roughly reflect those seen in the leaf tissues, indicating the source of the elevated ^{15}N in those tissues.

4.2.6 Discussion of soil analysis

Study sites 2 and 3 appear to have typical soil macronutrient and micronutrient profiles that fall into the ideal ranges for plant growth. While some of the values for sites 2 and 3 fall below or above the generic wine farm norms, none of them are so extreme as to prevent plant growth (Samuel Feagley, personal communication). This is further indicated by the values of these nutrients in leaf tissues of the plants at these sites (see Figure 4.2 and Figure 4.3).

Study site 1, however, has several values outside normal ranges, including resistance, base saturation, Cl concentration, exchangeable cation measurements, and base saturation. These values together confirm that the salinity of this site is far above generic norms. This could potentially reduce Mg uptake, which could potentially be detrimental as plants require 0.3-1.0% Mg for enzyme activity (Maathuis, 2009). However, all of the leaf samples of all species studied have Mg levels above this range as is shown in Figure 4.2c. Interestingly, *C. edulis*, the only species collected at field site 1, the only site effected by elevated salt concentrations, contains the highest concentration of Mg (Figure 4.2c). *C. edulis* is a known halophyte (Winter et al., 1976) and thus may have adapted alternate strategies of Mg uptake to compensate for the high salinity environments it grows in. Of the biomarkers of salinity stress tested none appear to have shown negative impacts on the growth of *C. edulis*.

While field sites 2 and 3 are 10 km apart, and field sites 1 and 2 are less than 1 km apart, field sites 2 and 3 are more similar in terms of the size and texture of the soil as well as their nutrient composition. Field site 1 is located within a ravine which gives it different water collection and distribution properties than sites 2 and 3. Field site 1 is saline and rates as brackish when all of the different measurements are considered which will place an increased burden on plants which are already experiencing water deficit stress.

4.3 Statistical analysis and discussion

To assess how the different factors or indices noted above relate to each other, correlation matrices were created for each of the Aizoaceae species. The analysis returned a matrix of Pearson's correlation values, from -1 to 1. Positive values indicate correlations and negative values indicate inverse correlations. Values greater than 0.6 or less than -0.6 are considered significant, and values greater than 0.9 or less than -0.9 are considered highly significant.

Measurement uncertainty was determined by the standard deviation of each measurement series. Variables where the standard deviation was greater than an order of magnitude of the mean were removed from consideration.

In the subsequent statistical analysis, the climate data was used as follows. For temperature, the monthly high temperature average (Hi) was used, and for (SR) and vapour pressure (VP), the values used are averages across the five years for each month.

Consistency checks were performed in relation to various variables. In the first instance, two independent measures of nitrogen concentration were made. With regard to the determination of nitrogen, "N" represents levels measured via elemental combustion analysis and "pN" levels measured via inductively coupled plasma-mass spectrometry (ICP-MS). The significant correlation between these analytical techniques over the course of the study for all of the species can be seen in each matrix, although the correlation is lower for *A. noctiflora* and *T. fruticosa*, because there is less data available for these species due to leaf abscission. Secondly, with regard to temperatures, it is noted that the highest average temperatures of the year are associated with highest levels of solar radiation. As a result, these should be highly correlated in all of the matrices and these indicators suggest that the analysis is robust and is making logical associations.

4.3.1 *G. africana*

Analyses of *G. africana* leaves across the study period are shown in Figure 4.5. The data for Mg, Mn, Fe, Cu, B, and "Rain" were removed from the analysis due to high measurement uncertainty. The independent measures of nitrogen, "N" and "pN", were highly correlated at 0.94. These N measures also show consistently high correlations with Na levels and leaf water content, and inverse correlations with Ca, temperature highs, and solar radiation, indicating that the analysis is reliable.

It is not surprising that climate factors associated with Namaqualand summer months (high average temperature and high solar radiation) play a role in trends associated with physiological indicators of water deficit stress, leaf water content and electrolyte content. For example, leaf water content is inversely correlated with high temperature (-0.93) and solar radiation (-0.89) confirming that these variables have a strong influence on leaf water loss.

Nitrogen content is highest during periods of high rain (0.88), low temperatures (-0.80), and most significantly, low levels of solar radiation (-0.91). The highly significant correlation between nitrogen content and Na content (0.91) suggests that Na may be involved in nitrogen uptake, particularly as concentrations of the cation calcium (-0.92) is significantly inversely correlated with nitrogen content. Unsurprisingly, the combination of these factors suggests that the Namaqualand climate plays a significant role in *G. africana's* life cycle.

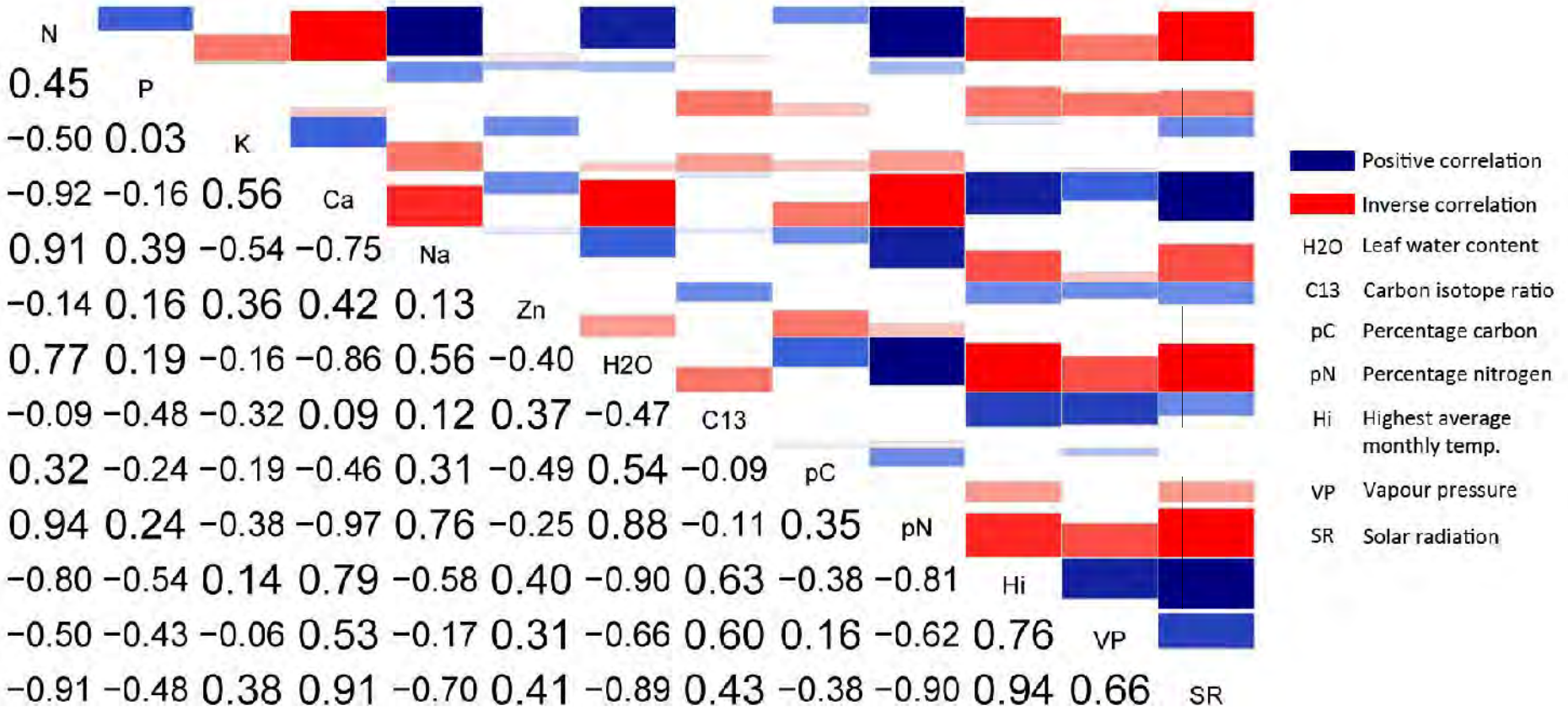


Figure 4.5: **Correlation matrix for *G. africana* using values from April 2011 to March 2012.** Analyses are labelled by a short code on the diagonal, with reference to the key at the right of the matrix. All labels are references to elemental nutrients with the following exceptions, "H20" is the leaf water content, "C13" - $\delta^{13}C/^{12}C$ ratio, "pC" and "pN", the carbon and nitrogen content respectively as measured using ICP-MS, "Rain", the total monthly rainfall, "Hi", the highest average monthly temperature and "SR", the average photosynthetically active solar radiation. The Pearson's correlations are indicated below the diagonal and the boxes above the diagonal are an alternate representation where blue boxes are positive correlations and red boxes are inverse correlations and the size of the box is proportional to the absolute value of the correlation.

4.3.2 *A. noctiflora*

In the analyses for *A. noctiflora* shown in Figure 4.6, the data for Fe, B, and Rain were removed from consideration due to high measurement uncertainty. The independent measurements of N concentration had a correlation of 0.82 and similar signification correlations to the other variables. The decreased significance between the N measurements (0.82 in *A. noctiflora* vs. 0.94 in *G. africana*) is a sign of the scarcity of data in this particular series. As *A. noctiflora* leaves abscised early in the dry season only leaf samples which were collected could be analysed. It was decided to continue with the analysis despite this due to the relatively small measurement uncertainty presented in the variables.

δ^{13}C levels in *A. noctiflora* are highly correlated with high temperatures (0.94) and solar radiation(0.96) and inversely correlated nitrogen (-0.98) and phosphorous (-0.86) levels. This corresponds with previous studies suggesting that that elevation in the δ^{13}C ratio is related to water deficit stress.

In contrast to *G. africana*, while there is still a positive correlation between N and Na levels (0.73), there is no statistically significant inverse correlation between N and any of the other cations. N is however correlated highly with leaf water content (0.96), P (0.88), and C (0.79) and inversely correlated with high temperature (-0.98) and solar radiation (-0.96).

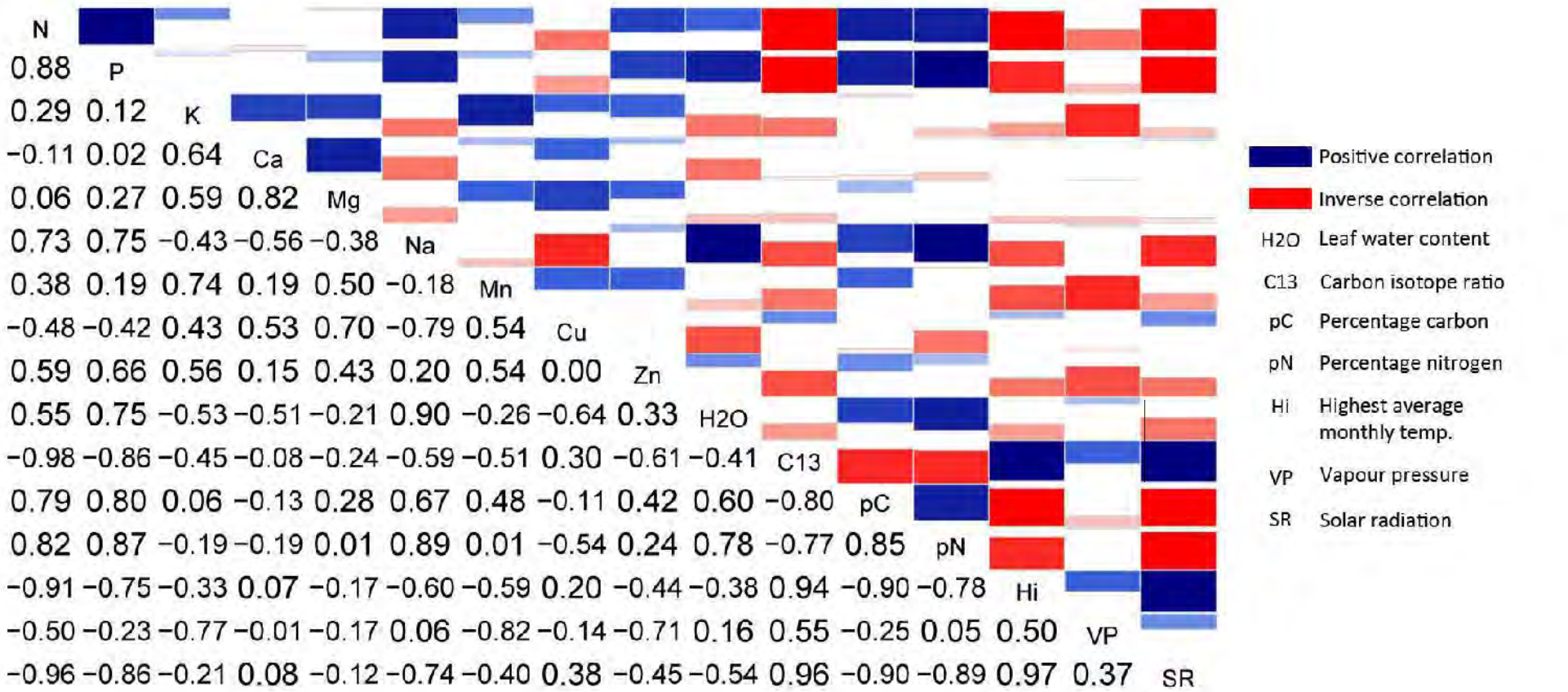


Figure 4.6: **Correlation matrix for *A. noctiflora* using values from April 2011 to March 2012.** The labels for each analysis are represented by the alphabetical code on the diagonal. Analyses are labelled by a short code on the diagonal, with reference to the key at the right of the matrix. All labels are references to elemental nutrients with the following exceptions, "H20" is the leaf water content, "C13" - $\delta^{13}C/^{12}C$ ratio, "pC" and "pN", the carbon and nitrogen content respectively as measured using ICP-MS, "Rain", the total monthly rainfall, "Hi", the highest average monthly temperature and "SR", the average photosynthetically active solar radiation. The Pearson's correlations are indicated below the diagonal and the boxes above the diagonal are an alternate representation where blue boxes are positive correlations and red boxes are inverse correlations and the size of the box is proportional to the absolute value of the correlation.

4.3.3 *C. edulis*

In the data for *C. edulis*, shown in Figure 4.7, the values for Mn, Fe, B, and Rain were removed from consideration due to high measurement uncertainty. N and pN are highly correlated (0.95), and each is consistently correlated with the other variables.

As was shown with *G. africana*, there is a similarly high correlation for *C. edulis* between N concentrations and selective Na uptake. Also, as was shown in *A. noctiflora*, carbon isotope ratios for *C. edulis* also appear to be related to water deficit stress.

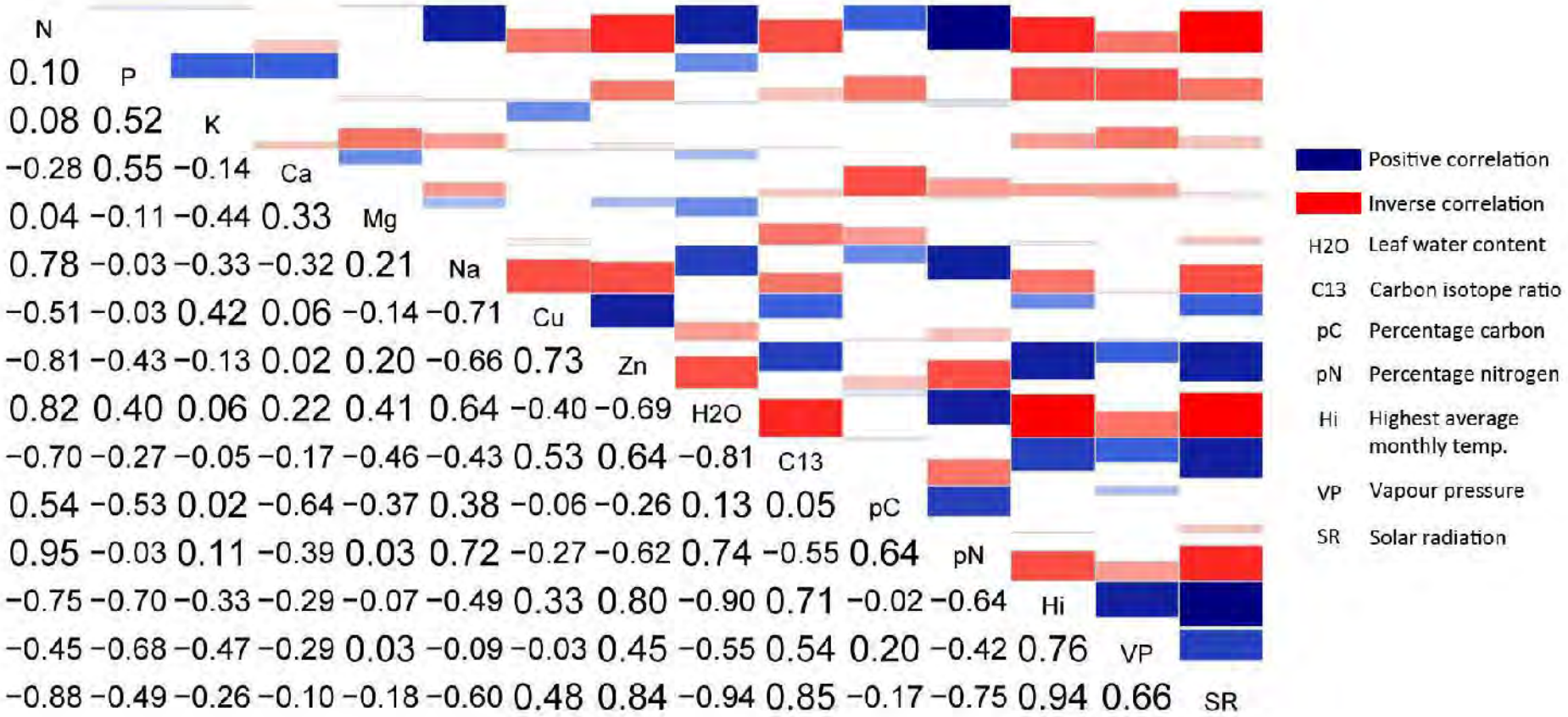


Figure 4.7: **Correlation matrix for *C. edulis* using values from April 2011 to March 2012.** Analyses are labelled by a short code on the diagonal, with reference to the key at the right of the matrix. All labels are references to elemental nutrients with the following exceptions, "H20" is the leaf water content, "C13" - $\delta^{13}C/^{12}C$ ratio, "pC" and "pN", the carbon and nitrogen content respectively as measured using ICP-MS, "Rain", the total monthly rainfall, "Hi", the highest average monthly temperature and "SR", the average photosynthetically active solar radiation. The Pearson's correlations are indicated below the diagonal and the boxes above the diagonal are an alternate representation where blue boxes are positive correlations and red boxes are inverse correlations and the size of the box is proportional to the absolute value of the correlation.

4.3.4 *R. robusta*

In the data for *R. robusta*, shown in Figure 4.8, the values for Mn and Rain were removed from consideration due to high measurement uncertainty. The independent measures of nitrogen content are highly correlated at 0.96. *R. robusta* appears to be similar to *C. edulis* in terms of N response to solar radiation(-0.83), high temperatures (-0.71), and $\delta^{13}C/^{12}C$ ratio (-0.77). As was seen in the other species, *R. robusta* also appears to lose water in its leaves at a rate inversely correlated to increases in temperature (-0.94) and solar radiation (-0.88).

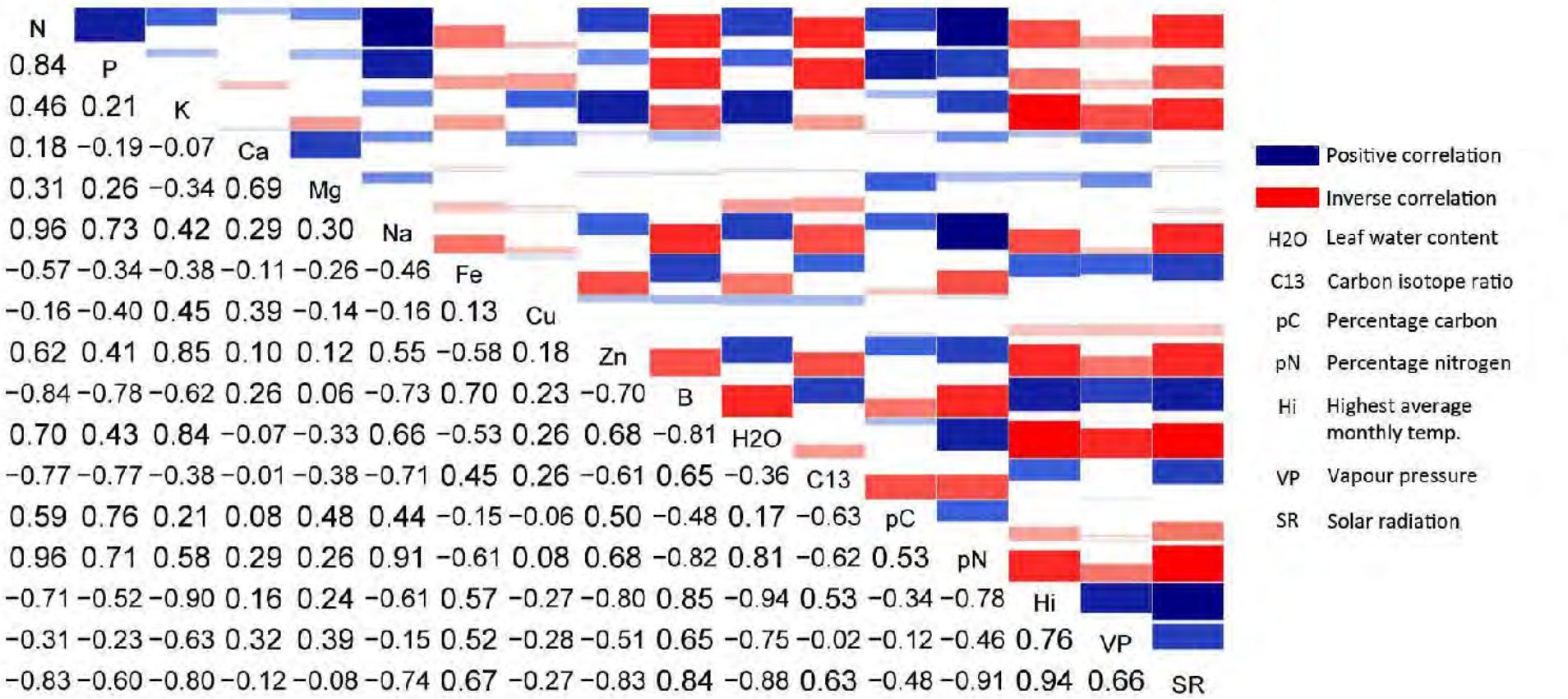


Figure 4.8: **Correlation matrix for *R. robusta* using values from April 2011 to March 2012.** Analyses are labelled by a short code on the diagonal, with reference to the key at the right of the matrix. All labels are references to elemental nutrients with the following exceptions, "H20" is the leaf water content, "C13" - $\delta^{13}C/^{12}C$ ratio, "pC" and "pN", the carbon and nitrogen content respectively as measured using ICP-MS, "Rain", the total monthly rainfall, "Hi", the highest average monthly temperature and "SR", the average photosynthetically active solar radiation. The Pearson's correlations are indicated below the diagonal and the boxes above the diagonal are an alternate representation where blue boxes are positive correlations and red boxes are inverse correlations and the size of the box is proportional to the absolute value of the correlation.

4.3.5 *T. fruticosa*

In the data for *T. fruticosa*, shown in Figure 4.9, the values for Ca, Mn, Fe, Cu, B, and Rain were removed from consideration due to high measurement uncertainty. As is seen in Figure 4.9, almost every variable is correlated to the others. Due to early leaf abscission (see Figure 4.1), there were not enough collections for meaningful analysis.

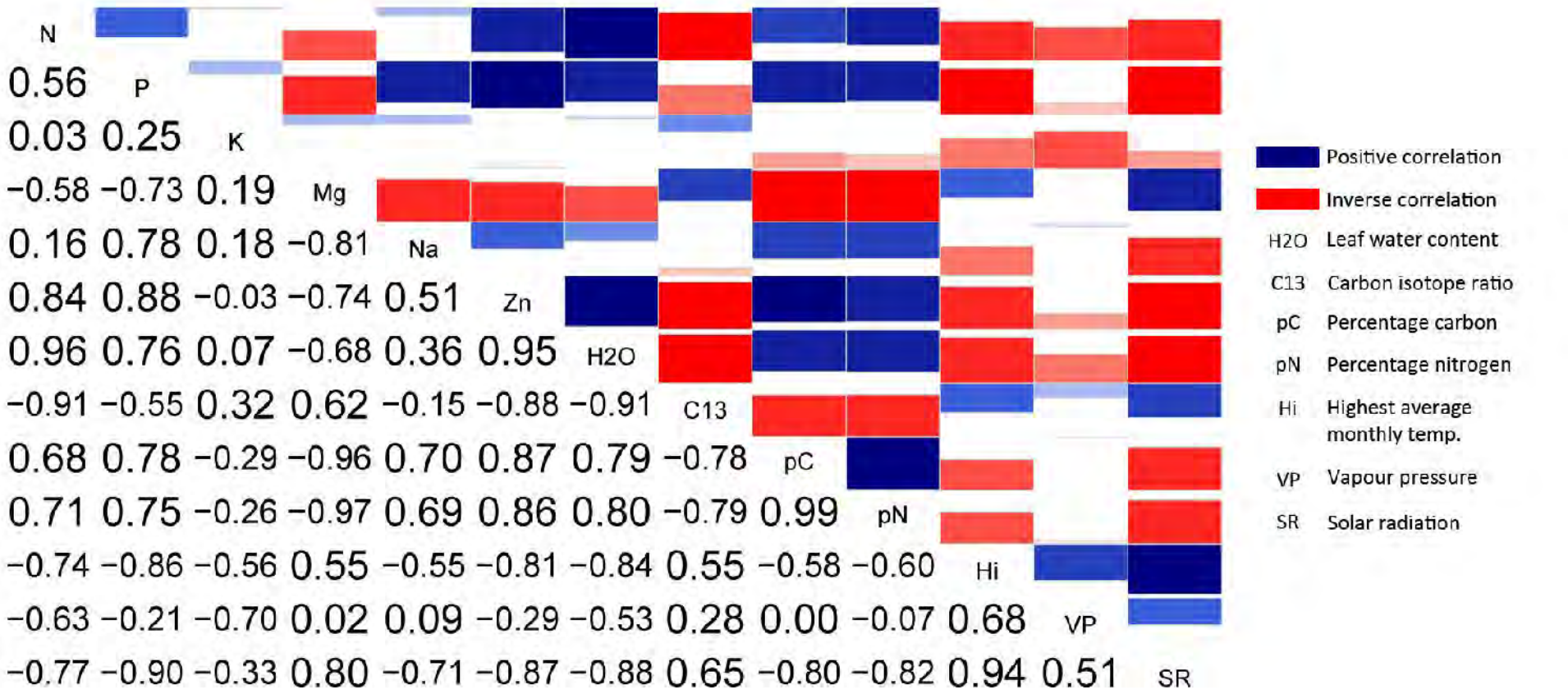


Figure 4.9: **Correlation matrix for *T. fruticosa* using values from April 2011 to March 2012.** Analyses are labelled by a short code on the diagonal, with reference to the key at the right of the matrix. All labels are references to elemental nutrients with the following exceptions, "H20" is the leaf water content, "C13" - $\delta^{13}C/^{12}C$ ratio, "pC" and "pN", the carbon and nitrogen content respectively as measured using ICP-MS, "Rain", the total monthly rainfall, "Hi", the highest average monthly temperature and "SR", the average photosynthetically active solar radiation. The Pearson's correlations are indicated below the diagonal and the boxes above the diagonal are an alternate representation where blue boxes are positive correlations and red boxes are inverse correlations and the size of the box is proportional to the absolute value of the correlation.

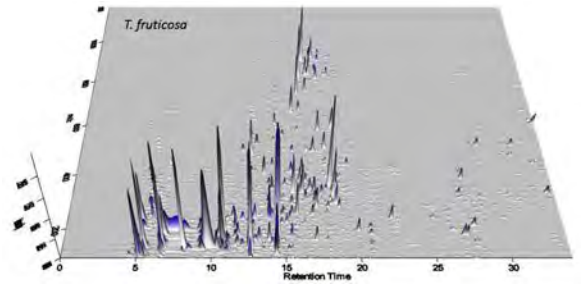
4.3.6 Summary of correlation analysis

In all species N and Na concentrations and leaf water content appear to be inversely correlated with $\delta^{13}C/^{12}C$ ratios, average high temperatures, and solar radiation. These factors also seem to be the most consistently significant variables studied in all of the plant species. Therefore, these factors confirm the season definitions described in Chapter 2 as winter, from April 2011 to August 2011 and summer, from October 2011 to March 2012 with September being excluded as a transition season. These definition were carried into the analyses as described in Chapter 5 for the analysis of metabolic barcode stability across seasons.

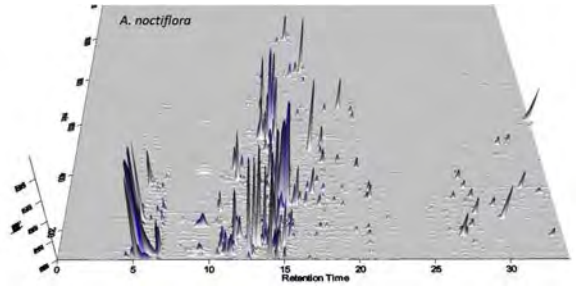
4.4 LC-MS

Representative, three dimensional TIC LC-MS profiles of all five Aizoaceae species are presented in Figure 4.10. The highest molecular weight compounds elute mid-run in all of the species and there are particularly high intensity peaks in all chromatograms at 5 min. The characteristic late eluting compounds which distinguish *G. africana* from the other species can be seen in Figure 4.10e. The sparse ion number described in Chapter 3 in Figure 3.11 is further indicated by Figures 4.10c and 4.10d, for *C. edulis* and *R. robusta* respectively.

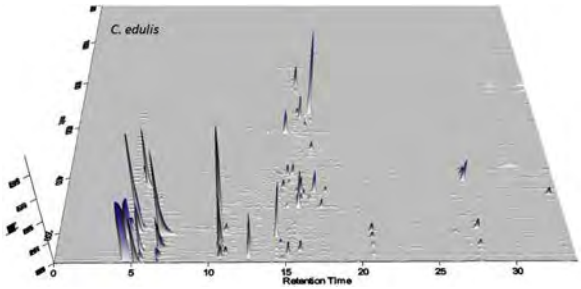
On the basis of phylogenetic analyses, *G. africana* (Figure 4.10e) and *T. fruticosa* (Figure 4.10a) have been placed in the same subfamily, where the distribution of ions in extracts from these two species show significant differnces whereas those of *A. noctiflora* (Figure 4.10b) and *T. fruticosa* actually appear more similar, at least in terms of the molecular weight and polarities of their constituents.



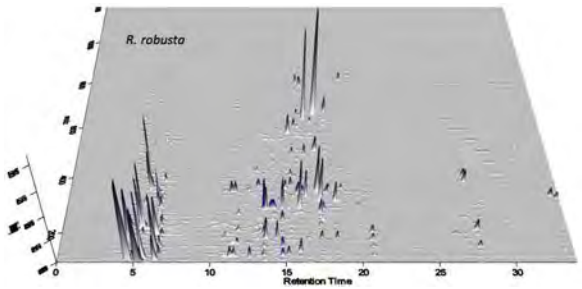
(a) *T. fruticosa* TIC



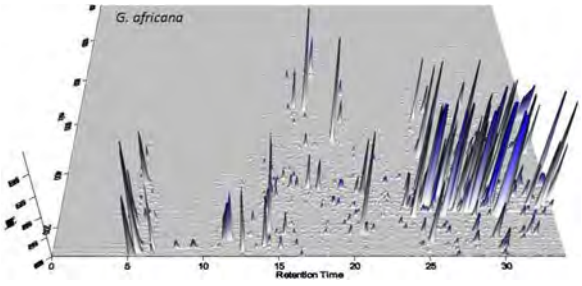
(b) *A. noctiflora* TIC



(c) *C. edulis* TIC



(d) *R. robusta* TIC



(e) *G. africana* TIC

Figure 4.10: **Representative TICs of all ethanol extracts of Aizoaceae species.** The x-axis is the retention time, the y-axis is the *m/z*, and the z-axis is the ion intensity.

4.4.1 Exploration of compounds previously found in Aizoaceae species

From a survey of literature which reports structures of secondary metabolites found in Aizoaceae species, a database of 108 compounds was compiled (Appendix A, Table A.1). A further 60 common plant primary metabolites were also included for a total of 168 metabolites (Appendix A, Table A.2). As shown in Table 4.6, 32 of the previously described secondary metabolites could be

identified on the basis of their relevant ion masses being within 5 ppm of the calculated value while another 15 had molecular ions within 10 ppm of caluculated values, thus allowing for putative identification of 47 compounds. In addition, 24 of the 60 primary metabolites were also identified in the ion data either in the protonated $[M+H]^+$ or esterified $[M+EtOH]^+$ form, 11 within 5 ppm and 13 within 10 ppm of calculated values.

Table 4.6: **Overview of identified metabolites.** Secondary metabolites were "Expected" or "Unexpected" to be found in the processed ion data based on the fact that they were extracted in ethanol and separated with the HPLC method described in Chapter 3. "Lit" are compounds found in the literature, "Sugar" are various carbohydrates, "Other" metabolites that did not quite fit the general category were Fructose-6-phosphate and 2-ketoglutarate. "Primary" is the sum of the values for all of the different classes of primary metabolites, and "Total" is a combination of all of the primary metabolites and expected secondary metabolites. For the row headers, "Compounds considered" are the total metabolites of each class, "Found" are the total found compounds of each class within 10 ppm of the exact adduct mass, % indicates the percentage of each class found. Adduct masses considered were [M+H] and [M+EtOH].

	Lit	Expected	Unexpected	Amino Acid	Organic Acid	Sugars	Other	Primary	Total
Compounds considered	108	71	37	21	9	28	2	60	168
5 ppm	32	32	0	8	5	1	1	15	47
10 ppm	15	15	3	7	0	1	1	9	27
Found	47	47	3	15	5	2	2	24	74
%	43.5	66.2	8.1	71.4	55.6	7.1	100.0	40.0	44.1

Consideration was also made for the fact that the compounds used to build the Aizoaceae compound library come from a wide variety of extraction techniques and detection methods and thus represent a wide range of polarities and solubilities. Because our extraction solvent was ethanol not all of these compounds should appear in the processed data, such as in the case of terpenoids or fatty acids. For this reason, these compounds were not expected to be seen in the processed ion data, whereas more polar metabolites were expected. Primary metabolites are somewhat challenging to study generally using HPLC because due to their size and polarity they typically elute from of the stationary phase with the solvent front. However, unlike their more nonpolar counter parts, they would be extracted in ethanol, thus they are more likely to be seen in the MS data.

43.5% of the all of the compounds considered were identified in the plant extracts on the basis of their molecular ions being within 10 ppm of the calculated values, with 40.0% of the primary metabolites and a 66.2% of the expected secondary metabolites being identified. If one removes the hydrophobic compounds from the secondary compound list, there are 65 compounds that remain. None of the hydrophobic compounds (terpenoids or fatty acids) were found in the ion data. This means that of the compounds reportedly identified in Aizoaceae species, which were most likely to be detected with the system utilised, 66.2% were actually detected and putatively identified.

4.4.2 Potential esterification of metabolites

As discussed in Chapter 3, because of the use of ethanol as the extraction solvent, there is a possibility that organic acids would be esterified to some extent, and potential ethyl glycosides formed from reducing sugars. Four ions corresponding to esterified secondary metabolite were detected, 1 within 5 ppm and 3 within 10 ppm of calculated values (in total 5.63% of identified compounds). In addition, 11 ethyl esters or ethyl glycosides of primary metabolites were detected, 5 within 5 ppm and 6 within 10 ppm of calculated values (in total 18.33% of compounds found). This is aligned with the fact that 83.33% of the primary metabolites contained functional groups prone to form ethyl esters or glycosides whereas only 34.75% of the secondary metabolites reported in the literature contained requisite functional groups, and the majority were too hydrophobic to have been extracted in ethanol.

Chapter 5

Generating models and metabolic barcodes

Metabolic barcodes were conceived of as a feature selection method to reduce the 23,000 ions from the LC-MS metabolic fingerprints into a subset of ions which could be used to classify the Aizoaceae species studied. This was important because less significant associations between the samples (the covariance of low intensity ions) forces over-fitting in clustering methods as it also would in a classification model. Additionally, the analysis of 23,000 ions is computationally intensive and if it is possible to reduce the number of features considered, the model would be more widely applicable.

5.1 Analysis of metabolic fingerprints

The processed LC-MS data (as described in Chapter 4) rendered a metabolic fingerprint for each sample of each species. Before metabolic barcodes could be established, various analyses needed to be conducted in order to understand the variation and spread of the ion intensities of the metabolic fingerprints. To begin, the LC-MS data was exported from MZmine as a ".csv" file and loaded into the statistical programming language R. Multivariate statistical analysis was applied in order to understand the global metabolite data (Liland, 2011). Commonly multivariate approaches are used to reveal how samples and metabolites are distributed and to ensure that the groupings are biologically meaningful thus indicating correct preprocessing of data (Nobeli et al., 2003; Xia et al., 2009; Sato et al., 2008). All scripts used can be found in Appendix B.

Prior to statistical analysis, the data was centred and scaled to prevent scale bias in downstream analyses. This is traditionally approached using unit variance scaling method which uses the standard deviation of a variable as the scaling factor. In the end, this ensures that all of the variables are considered with equal weight and that ions with high intensity are not considered more important statistically than ions with lower intensities. Importantly however, the m/z ratio threshold for ion detection were set quite low so that in the gap-filling step low intensity peaks would be detectable(see Chapter 3). Because of this, it is possible that some of the ions selected represent noise and that using this method would give these variables equal weight to ions with intensities above the noise threshold. Thus, the Pareto scaling method, which uses the square root of the standard deviation as the scaling factor, was compared to the more traditional UV scaling method to determine which would work best. Theoretically, the Pareto method results in a reduction of the relative importance of metabolites with high intensities by decreasing large fold changes more than small ones but does not make them equal as is the case in UV scaling(van den Berg et al., 2006) thus making Pareto scaling less sensitive to outliers.

5.2 Final analysis of chromatogram consistency

Data were centred and scaled using both the UV and Pareto scaling methods and compared using hierarchical clustering (HCA). As an indication of the accuracy of the technique, species-specific clustering with either method suggests that the data preprocessing, and spectrum alignment in particular, were relevant.

To generate the dendrograms, a distance matrix was first created using the standard pre-sets in R's "stats" package using the "dist" function. An agglomerative hierarchical model was then applied to the distance matrix using the Ward method. Agglomerative models are based on comparing values on an observation by observation basis, where the observations are placed at terminal nodes, and then branches are built backwards until all of the terminal nodes have been incorporated into the model. An analysis of variance approach is used to determine the distance between clusters and these are then grouped accordingly (Nugent and Meila, 2010; Hastie et al., 2008). Using this approach, the dendrograms shown in Figure 5.1 and Figure 5.2 were generated.

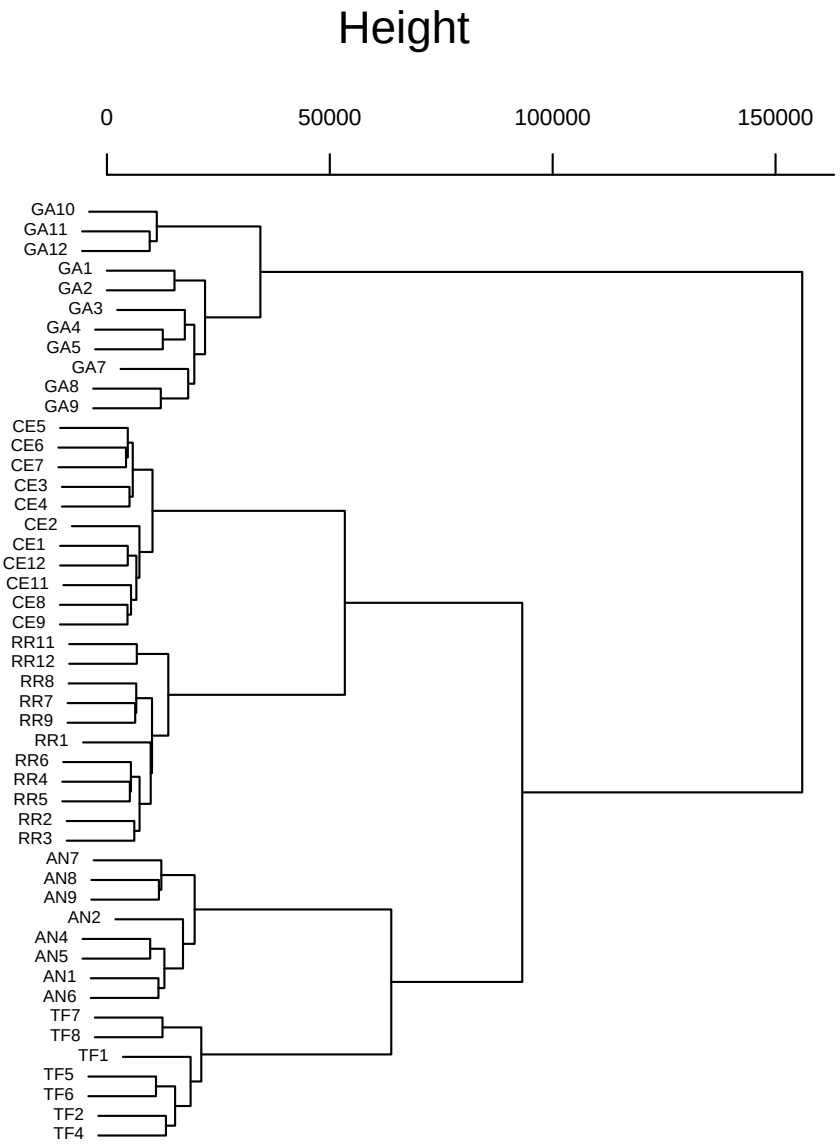


Figure 5.1: **Hierarchical clustering of Aizoaceae leaf sample ion data using UV scaling.** The species are represented here as "GA" for *G. africana*, "AN" for *A. noctiflora*, "CE" for *C. edulis*, "RR" for *R. robusta*, and "TF" for *T. fruticosa*, and the numbers represent the collection month with 1 in April of 2011 and 12 in March of 2012.

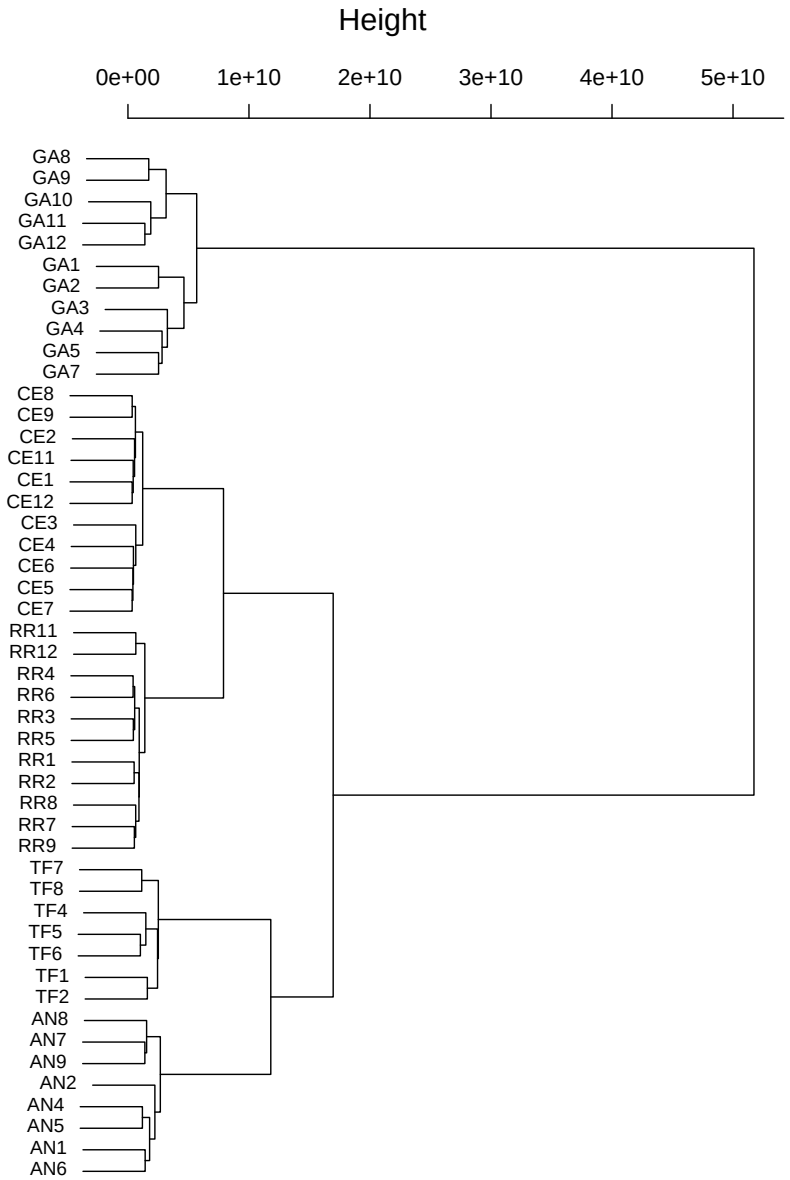


Figure 5.2: **Hierarchical clustering of Aizoaceae leaf sample ion data using Pareto scaling.** The species are represented here as "GA" for *G. africana*, "AN" for *A. noctiflora*, "CE" for *C. edulis*, "RR" for *R. robusta*, and "TF" for *T. fruticosa*, and the numbers represent the collection month with 1 in April of 2011 and 12 in March of 2012.

Both Figure 5.2 and Figure 5.1 show species-specific clustering. In addition, it is interesting to note that in both dendrograms the *A. noctiflora* ("AN") and the *T. fruticosa* ("TF") samples cluster more closely together than the *G. africana* ("GA") and *T. fruticosa* ("TF") samples which are phylogenetically classified in the same subfamily (see Figure 2.9, (Klak et al., 2007)). On the other hand, *C. edulis* ("CE") and *R. robusta* ("RR") cluster together indicating the closeness of their metabolic profiles but their significant difference from other species under analysis. In Klak et al. (2007) these species were indistinguishable using a few typical genetic markers, until in 2013 when a larger number of DNA sequence regions were employed (Klak et al., 2013). Even so, many of the species in this subfamily remain difficult to distinguish phylogenetically.

As can be seen when comparing the distances between the nodes in Figure 5.2 and Figure 5.1, the nodes of Figure 5.1 are about 5 orders of magnitude smaller than those in Figure 5.2. This indicates that less of the variance in the data is covered in this analysis, thus the Pareto scaling method was carried forward in the analysis.

5.3 Determination of metabolic barcodes

Once it was established through HCA that the preprocessing gave biologically logical results, focus could be placed on the analysis of the ions for the generation of barcodes.

5.3.1 Establishing principal components

To determine which ions distinguish species from each other, data reduction was initiated using PCA of ion data from all of the Aizoaceae samples. PCA is an undirected clustering algorithm commonly used in multivariate analysis of large data sets and is commonly employed in data reduction in metabolomics analysis (see Chapter 1). Principal components (PCs) were determined using R's singular value decomposition (SVD) algorithm. In SVD a normalised ion matrix X is decomposed into three parts:

$$X = UDV \tag{5.1}$$

where the product of U and D form the score matrix T, and V becomes the loading matrix P:

$$X = (UD)V \rightarrow X = TP \tag{5.2}$$

A biplot of the first two PCs was generated (Figure 5.3) to determine how the plant species clustered based on all of the ions from the monthly collections. All of the *G. africana* samples cluster to the top left quadrant (green), while the *A. noctiflora* and *T. fruticosa* samples cluster together in the bottom right quadrant, and the *C. edulis* and *R. robusta* samples cluster in the top right quadrant.

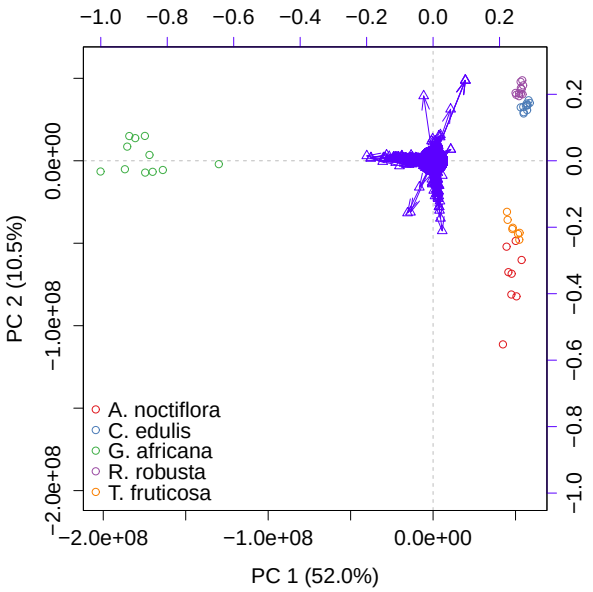


Figure 5.3: **PCA biplot of PCs 1 and 2 of ions from all Aizoaceae leaf samples.** The circles represent the monthly plant samples and the triangles represent the ions.

The level of the variance described by PC 1 between the groupings of *G. africana* and the other Aizoaceae species, as shown in Figure 5.3, is so extreme that the other plant samples are barely distinguishable from each other, but they do, however, separate quite well along PC2. This suggests that they are significantly more similar metabolically to each other than they are to *G. africana*. This includes *T. fruticosa* with which *G. africana* shares the most similarity in terms of phylogenetic markers as noted earlier. This is also consistent with the analysis of the UV-Vis and TICs of the extracts of the various species where those of *G. africana* had a significantly different profile than the other species.

For illustrative purposes, the SVD was also run on the data scaled using the UV scaling method and PC1 and PC2 were also plotted in Figure 5.4.

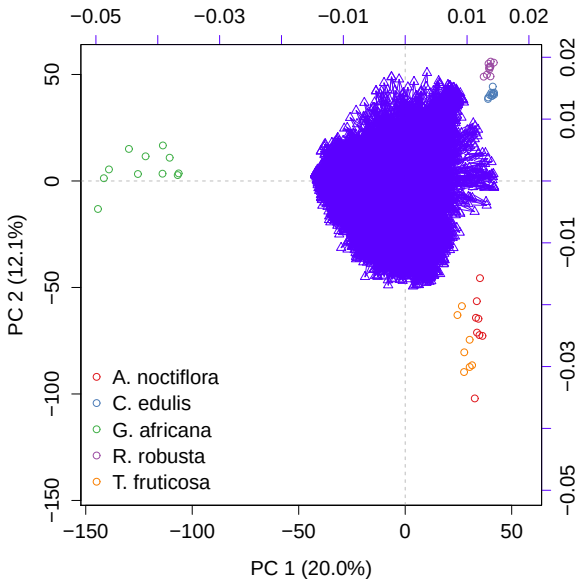


Figure 5.4: **PCA biplot of PCs 1 and 2 of ions from all Aizoaceae leaf samples.** The circles represent the monthly plant samples and the triangles represent the ions.

Because the species clustered effectively in both the hierarchical clustering and the PCA, it was clear that there were specific ions responsible for the clustering. To determine which these were, PCs that covered 90% of the variance in the data were selected. The first PC covers the most variance and every PC thereafter covers less and less of the total variance.

5.3.2 Variance covered by each PC

The purpose of PCA in this analysis was dimension reduction, and the next step was thus the identification of PCs covering the majority of the variance in the data set. In order to make this assessment the amount of variation explained by each PC (λi) was determined (Wehrens, 2011):

$$\lambda i = \frac{d_i^2}{n-1} \tag{5.3}$$

where d are the diagonal elements of matrix D (Equation 5.1) and n is the total number of observations. The variation of each of the PCs was then summed together and the percent of the variance covered by each PC was determined by dividing each PC by the total variance and multiplying that by 100. The fraction of the variance accounted for can then be expressed as a percentage:

$$V\% = \frac{\lambda_i}{\sum_{j=1}^{a}(\lambda^j)} * 100 \tag{5.4}$$

In Figure 5.5, the majority of the variance is seen in the first few PCs (see expanded plot "B") which is expected from PCA analysis (Wehrens, 2011).

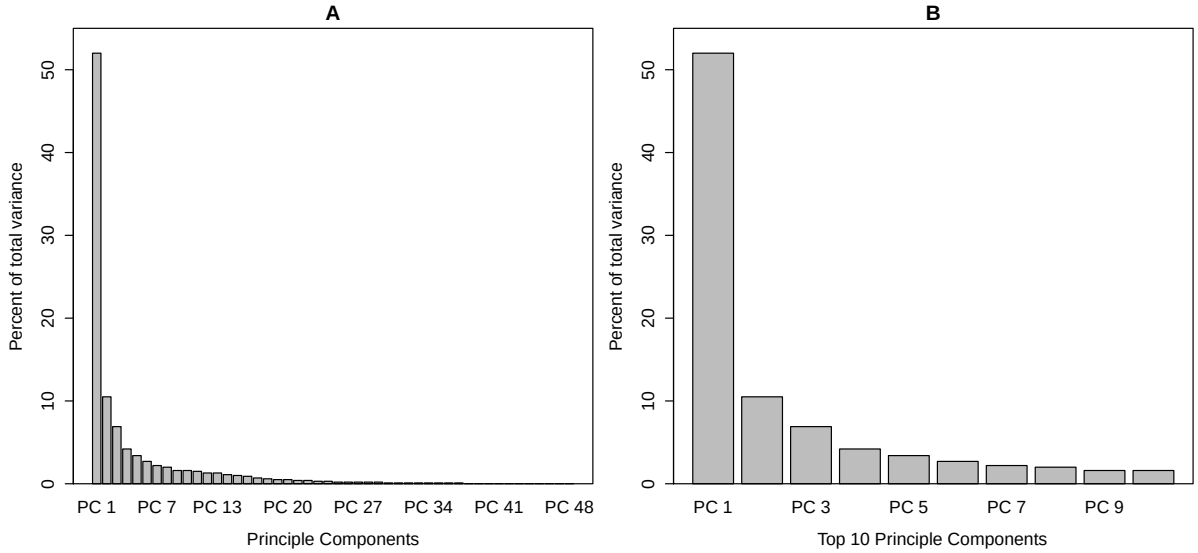


Figure 5.5: **Relative variance of the PCs generated from all Aizoaceae sample ions.** "A" shows the percent variance covered by all of the PCs and "B" shows the percent variance covered by the top 10 PCs.

In order to determine which PCs cumulatively covered 90% of the total variance, the PCs were ordered by the amount of variance each contained, starting with the most variance covered, and consecutively summed until the PCs covering at least 90% of the variance were reached. As is shown in Figure 5.6 ("B"), the first 12 PCs cover 90% of the total variance in the data. After this determination, matrix P (Equation 5.2) was reduced to 12 columns, thus forming matrix P' to reflect this finding.

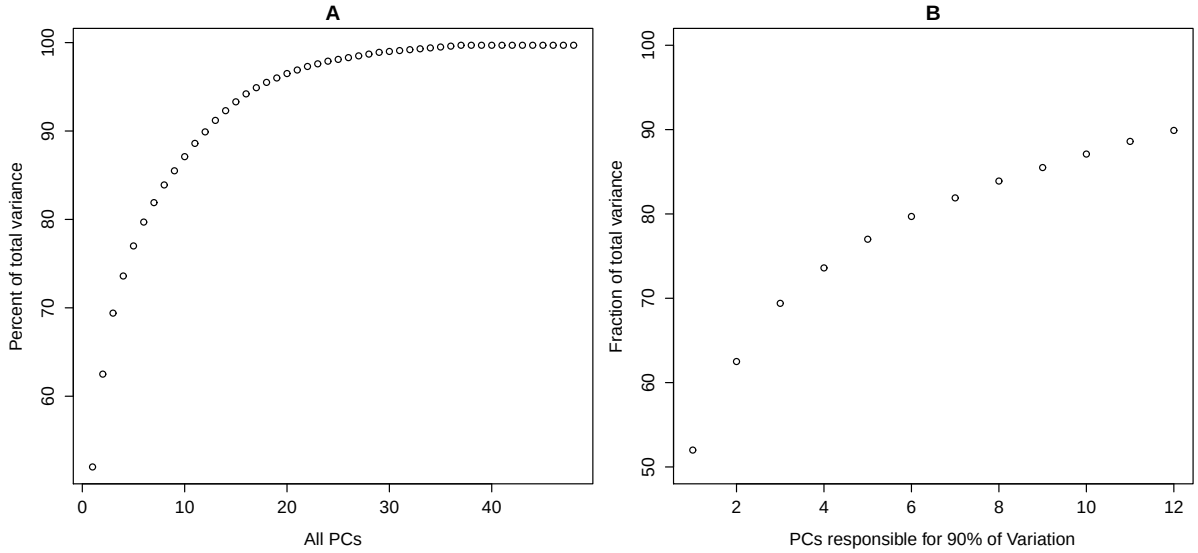


Figure 5.6: **PCs in order of greatest contribution to variance.** "A" shows all 48 PCs (or matrix P) and "B", the 12 PCs which cover 90% of the variance in the ion data (or matrix P').

5.3.3 Weighting the PCs

Each value in each column of P' represents the amount of variance described by each ion for that PC. To ensure that each ion contributed the percent variance indicated by the variance covered by each PC, each PC (column in matrix P') was then weighted to reflect the amount of variance that it described (see Figure 5.5):

$$PW = P' \times \lambda i \tag{5.5}$$

Where P' is the shortened Loading matrix and λi (see Equation 5.3) is a vector containing the variance described by each PC.

The correction can be seen in the increased linearity of the data. Starting with the unweighted PCs in Figure 5.7 "A", the influence of PC2 on the ion data is indicated by the distribution of the ions between both axes . This is corrected in Figure 5.7 "B" where the ions align more strongly along the x-axis.

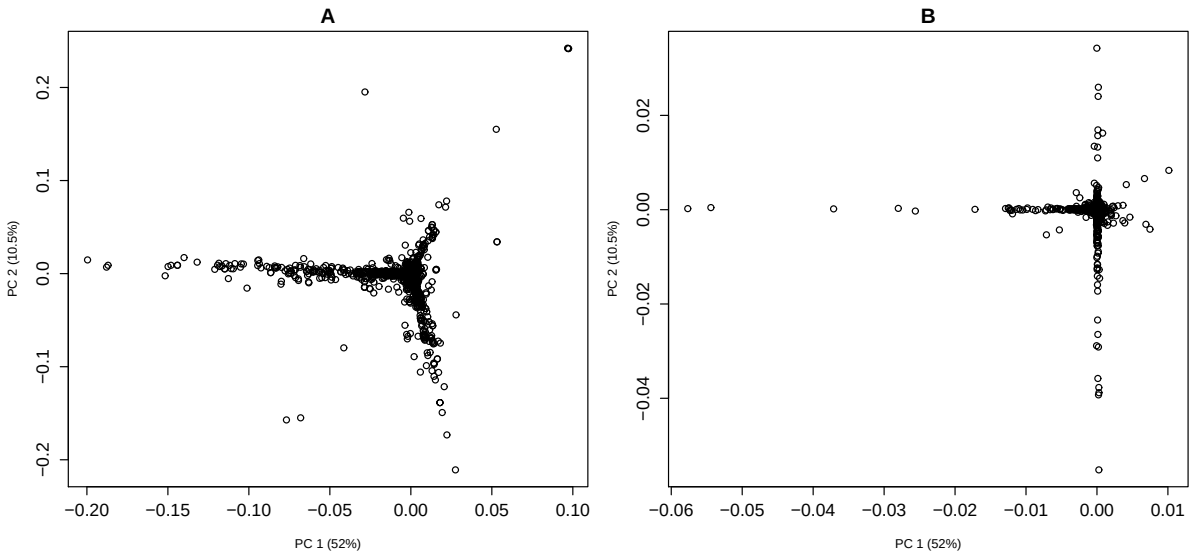


Figure 5.7: **The effects of weighting on PCs 1 and 2.** "A" is before and "B" is after weighting the PCs.

5.3.4 Determining leverage scores

The concept of leverage scores arises from the discipline of information theory. It is based on the idea that only a fraction of the variables in a data set are distinctly different between observation classes (MacKay, 2003). In this case, leverage scores were used in order to distinguish which ions separated the species from each other in PCA, and were thus informative ions. Leverage scores were determined using a slight modification of the method used in Yip et al. (2014). The weighted matrix PW was converted into leverage scores by summing the variance represented by each PC for each ion and dividing this by the total variance of the ions as represented by the following formula:

$$P_\lambda = \frac{\sum_{j=1}^{k}(PW_{\lambda,j})^2}{\sum_{j=1}^{k}\sum_{\lambda=1}^{m}(PW_{\lambda,j})^2} * 100 \tag{5.6}$$

where $PW_{\lambda,j}$ represents the values from each of the 12 PCs for each ion j for the total number of ions k across PC 1 (j) to PC 12 (m) (Yip et al., 2014). Resulting leverage scores (PW_λ) are represented as a percent of variance covered by each ion as a function of the total variance.

The leverage scores of the individual ions were arranged from greatest to least. In Figure 5.8, they were then summed in a cumulative manner such that the fewest number of ions needed to cover a respective percentage of variance is displayed. As is shown in Figure 5.8, coverage of 99% of the total variance in the data is achieved by about 700 ions which is significant in that it represents only about 3% of the approximately 23,000 ions in each metabolic fingerprint.

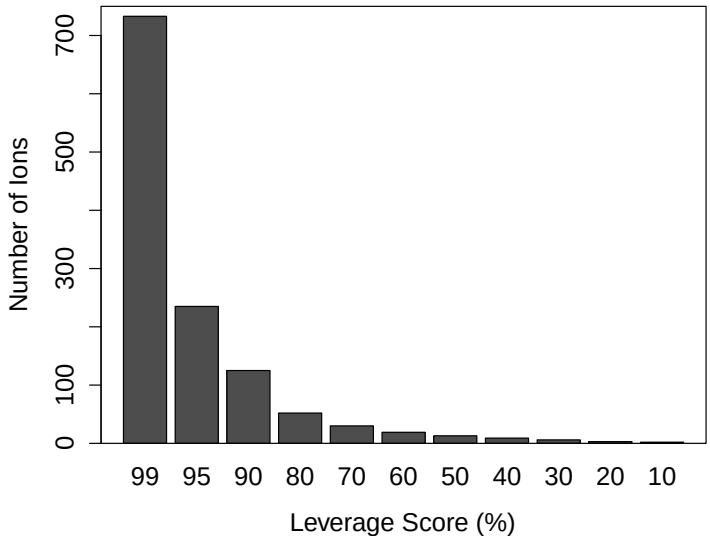


Figure 5.8: **The number of informative ions per leverage score across all Aizoaceae samples.** Leverage scores display the number of specific ions used to explain the total variance of the data set.

To see how this reduced ion pool might change the clustering of the individual samples, HCA was again performed as previously described. Every 10 percentage points of total variance were tested for species-specific clustering. This revealed, as shown in Figure5.9, that as few as the 19 ions that cover 60% of the variance could still achieve species-specific clustering. However, as the intention of this work was to compare species that were not closely related to each other, it was thought important to include a larger sample of ions in the fingerprint to compensate for instances of convergent evolution and to make the classification model more robust. For this reason, ions that represent 90% of the variance of the data were selected to represent metabolic barcodes.

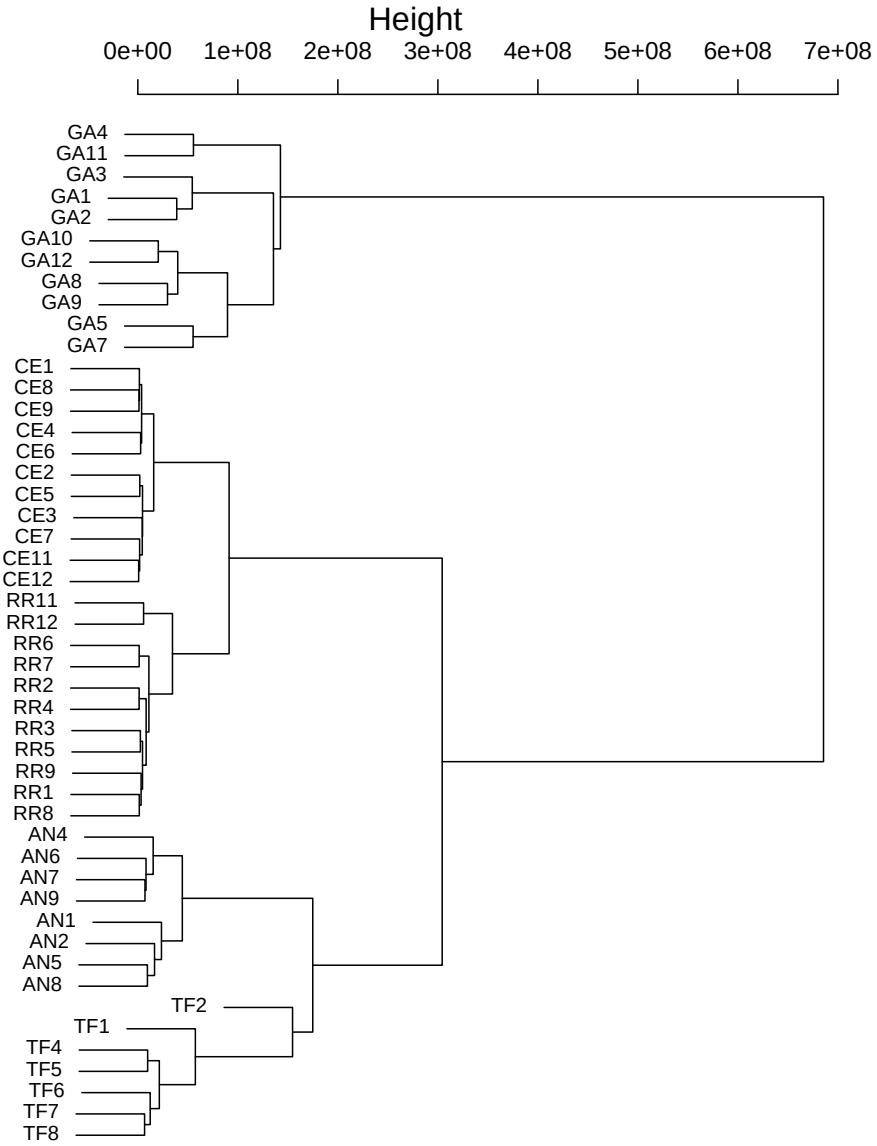


Figure 5.9: **Hierarchical clustering of Aizoaceae samples with variables covering 60% of the total variation.** The species are represented as "GA" for *G. africana*, "AN" for *A. noctiflora*, "CE" for *C. edulis*, "RR" for *R. robusta*, and "TF" for *T. fruticosa*, and the numbers represent the collection month with 1 in April of 2011 and 12 in March of 2012.

HCA with the ions covering 90% of the total variance is displayed in Figure 5.10 for comparison, specifically to show that ions which cover 90% of the total variance add an order of magnitude to the separation in the distance between sample clusters. This indicates that there is much greater separatory power for classification when additional informative ions are included rather than just more data. As compared to the total metabolic fingerprint in Figure 5.2, total cluster distance is reduced by a single order of magnitude.

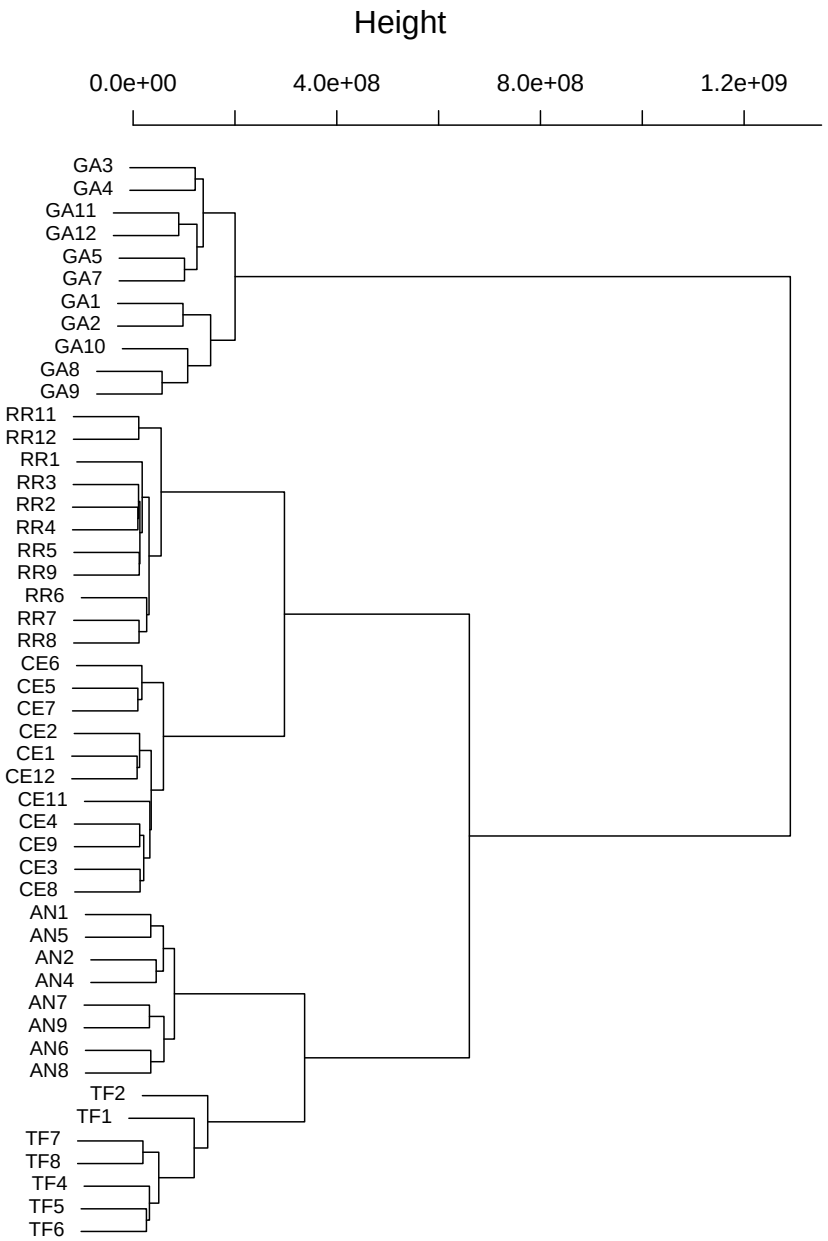


Figure 5.10: **Hierarchical clustering of informative ions across all Aizoaceae samples.** The species are represented as "GA" for *G. africana*, "AN" for *A. notiniflora*, "CE" for *C. edulis*, "RR" for *R. robusta*, and "TF" for *T. fruticosa*, and the numbers represent the collection month with 1 in April of 2011 and 12 in March of 2012.

PCA was performed as previously described on the ions covering 90% of the total variance as an independent method of non-specific species classification and to determine how species clustering of only the barcode ions compared to the species clustering of total dataset (see Figure 5.3). The PCA demonstrated in Figure 5.11 is still mostly able to separate species into individual clusters in a manner reflecting what was seen in the total dataset. PC1 and PC2 cover 59.8% variance between samples which is only 3% less than the amount of variance covered by the total dataset (see Figure 5.3). We would expect this to be the case as each additional variable added to the data set will ultimately contribute to the covariance in the PCs. By reducing the number of compounding variables we are reducing this affect. Because of the clear clustering in both the HCA and the PCA of the species when only using the ions covering 90% of the total variance, the significance of the barcodes appears well established.

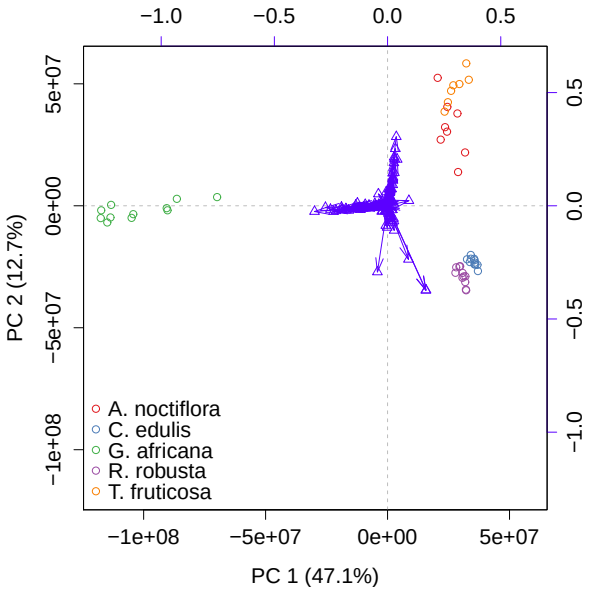


Figure 5.11: **PCA biplot of only the barcode ions from the Aizoaceae leaf samples.** Where circles represent the plant samples and the triangles represent the variables.

5.4 Stability of metabolic barcodes

It was then important to assess if the metabolic barcodes would represent stable markers for future study. In studies of metabolic fingerprints for chemotaxonomy by Incerti et al. (2013) and Farag et al. (2012), the ions which were important for species' clustering tended to be secondary metabolites. Because secondary metabolites may not be constitutively produced, their use as chemotaxonomic markers needed to be tested.

The use of the gap-filling step and the generally low thresholds set for ion intensity during the preprocessing steps of the LC-MS data (see Chapter 3) made it possible for ions representing noise to be significant in the model. To test if this was the case, the maximum ion intensities for the 125 barcode ions determined. The minimum value of the ion intensity maximums across the 125 barcode ions was then determined to be 2,818,060. This indicates that even in the sample with ions of the lowest intensity, all of the barcode ions had intensities well above the noise threshold.

In the HCA and PCA analyses described in Figure 5.10 and Figure 5.11, the species clustered together as expected, although both HCA and PCA are sensitive to outliers. To gain insight into how the ion intensities of the barcode ions were distributed across the metabolomic barcodes of the samples of the same species as well as to see how they were distributed between different species, a heatmap with accompanying dendrograms for the plant extract samples and the ions covering 90% of the variance was generated in Figure 5.12.

Generally, Figure 5.12 demonstrates the species' samples cluster (vertical dendrogram) shown in Figure 5.10 with specific ion clusters represented in the horizontal dendrogram. There are two ions in particular which stand out significantly as potential outliers for the model; the ion with m/z 137.0482 (dark purple) which has a high intensity in the *T. fruticosa* sample from collection 2 but a much lower intensity in the other samples from that species and the ion with m/z 593.2741 (mid purple, in the middle of the *G. africana* ion data on the left side) which has a high intensity in the *G. africana* sample from collection 2 but also a much lower intensity across the rest of the samples from all species.

It is also important to note that the vast majority of the variance is represented in the *G. africana* samples (clustered at the top of the heatmap) but that there are significant ion clusters for each species. It is also important to note that most of the ions change in intensity across all of the samples of a species, indicating change in total concentration between collections.

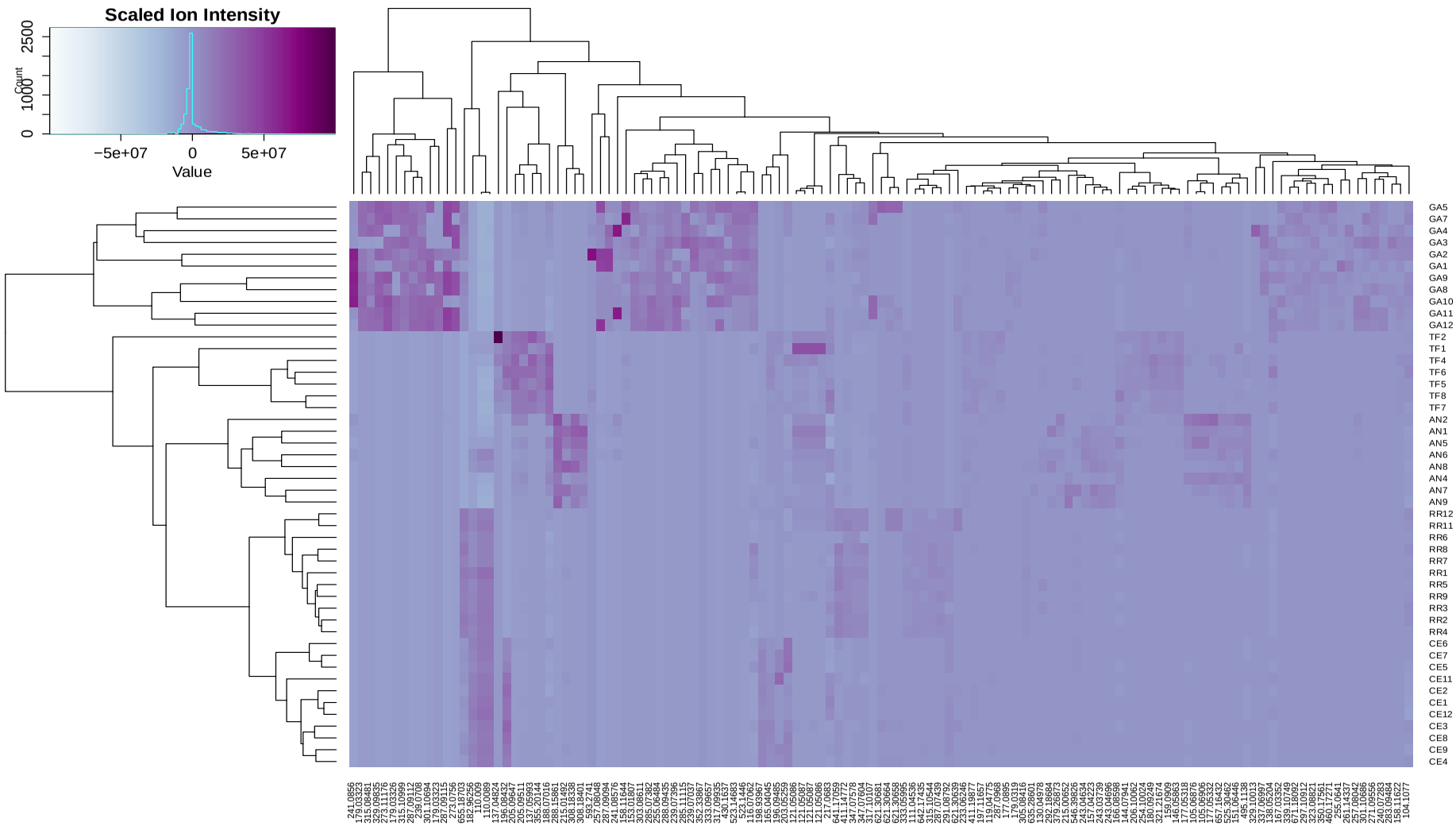


Figure 5.12: **Heat map of informative ions and their groupings with individual species' samples.** The species are represented here as "GA" for *G. africana*, "AN" for *A. noctiflora*, "CE" for *C. edulis*, "RR" for *R. robusta*, and "TF" for *T. fruticosa*, and the numbers represent the collection month with 1 in April of 2011 and 12 in March of 2012. The heat map indicates the ion intensity for a the barcode ions in a given sample. The dendrogram at the top represents the relationship of those ions to each other and the dendrogram to the left shows the relationship of the samples to each other.

Because it was obvious in Figure 5.12 that the intensity of the ions changes between plant samples, and in some cases, more drastically than in others, change of ion intensity compared to change in the various factors described in Chapter 4 was then considered. To understand how those variables might have influenced ion intensities, correlation matrices were then generated comparing the ion intensities of the ions covering 60% of the variance with the climate, nutrient, and physiological variables found to be most important in the previous chapter (N and Na content, leaf water content, carbon isotope ratios, and high temperatures and solar radiation). Henceforth, these variables will be referred to collectively as seasonal classifiers. Figure 5.13 is a representative matrix including at least one Pearson's correlation value > 0.7 between the ion intensities and the seasonal classifier values. Only *G. africana* was selected for this analysis as this species' samples have a much higher degree of variation than the other species. *G. africana* was also one of three species who's leaves did not abscise and thus had a full year's collection of data.

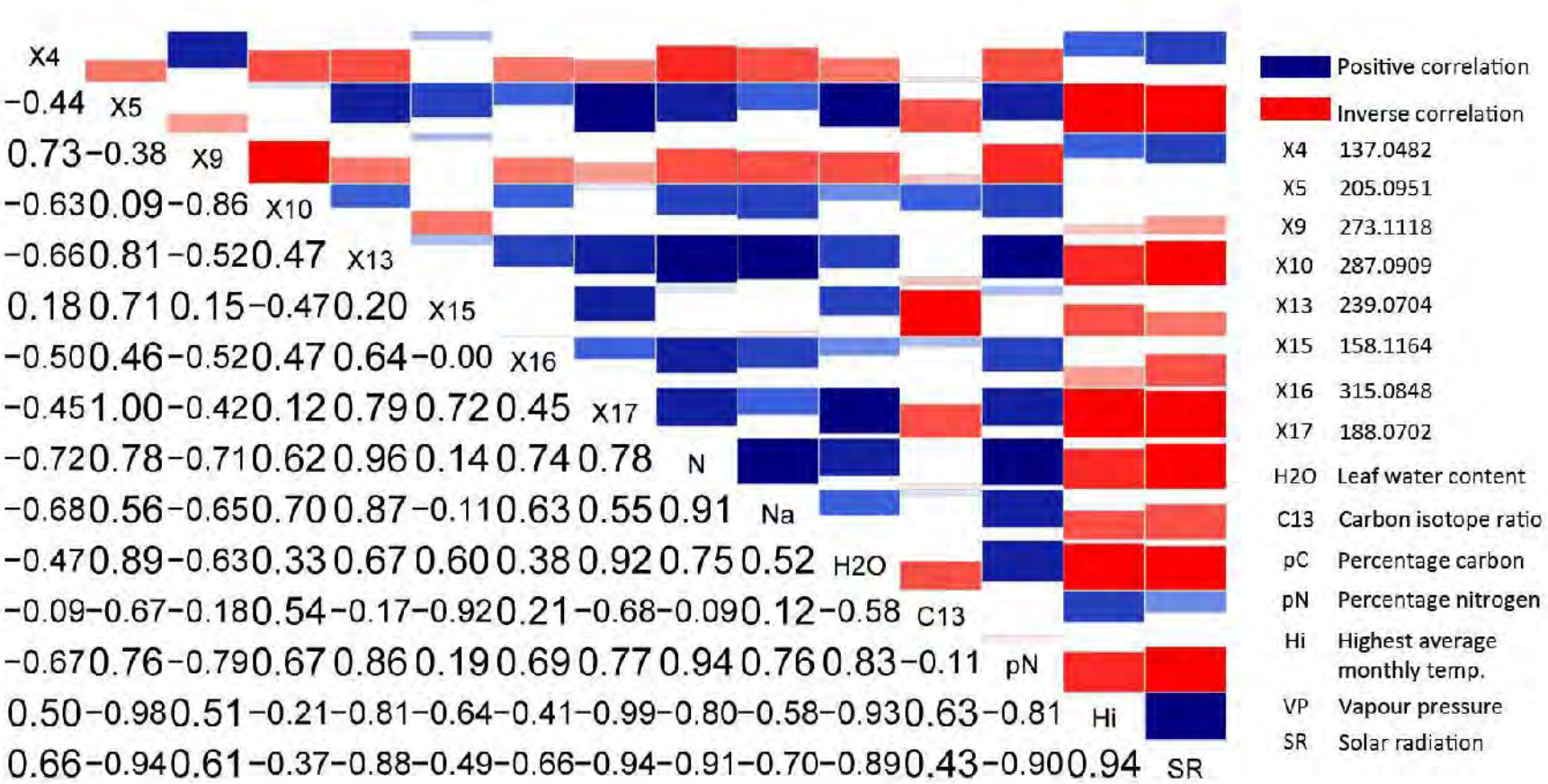


Figure 5.13: **Cross correlation matrices with informative ion intensities and relevant factor data.** Only ions which cover 60% of the variance were considered and those with intensities significantly correlated with various climate, nutrient, and physiological data are presented. Relevant ions are represented indicated in the key to the right. Factors determined as significant in Chapter 4 were also considered: "N" and "pN" are independent measures of nitrogen content, "Na" is sodium content, "H2O" is leaf water content, "C13" is $\delta^{13}C/^{12}C$ ratio, "Hi" is temperature highs, and "SR" is solar radiation.

The strongest correlations are seen between N and X13 (0.96), Na and X13 (0.87), leaf water content and X5 (0.89) and X17(0.92). The strongest inverse correlations are high temperature and solar radiation with X5 (-0.98 and -0.94), X13 (-0.81, -0.88), and X17 (-0.99, -0.94) as well as $\delta^{13}C/^{12}C$ ratio and X15(-0.92). Indicating at least seasonal change in concentration of these ions in *G. africana*.

Figure 5.12 indicated that there were at least two ions in the barcodes which represented potential outliers and the seasonal variation in the ion intensities explored in Figure 5.13 indicate that at least for some of the ions, there is a definite seasonal trend in their intensities. however the clustering in the dendrograms in Figure 5.12, as well as in Figure 5.11 indicate that there is enough remaining variation in the barcodes for species specific clustering. Collectively, these data suggest that the barcodes should be stable enough to build a classification model.

5.5 Random Forest barcode classification model

After the stability of the metabolic barcodes had been established, the other parameters (climate, nutrient, and physiological) analysed in Chapter 4 could then be selectively employed to build a classification model. The seasons defined in Chapter 2 were confirmed in Chapter 4 and the model was trained on each sample class (winter and summer) and tested on the other (summer and winter). In this way, the stability of the model was selectively tested based on biological analyses which were determined independently of the LC-MS analysis of the metabolites. The model was further validated using 10-fold K-means cross-validation.

To accomplish this, the machine learning technique Random Forest was used to build the classification model. Random Forest builds a model based on a number of decision trees generated by sampling a number of variables at each node between sample classes. Each tree forms its own classification route which the Random Forest algorithm then blends together to form a joint classification model. There is no pruning in this system, and each tree is allowed to grow as much as possible. When the trees are blended together, the variable selection that is most common between all of the samplings is voted on and becomes the collective forest model (Breiman, 2001).

This algorithm was selected due to its ability to handle missing data and to balance unbalanced data sets. These features allow a model to be built based on a reduced dataset and then to be applied to samples from different species which will inevitably have different numbers of compounds as well as generally different compound compositions (Scott et al., 2010).

5.5.1 Random Forest

As with other machine learning techniques, the Random Forest classification algorithm builds a classification model based on a subset of data (training set). Cross validation is handled internally through the out of bag error rate (OOB) which is returned after the model is generated. The Random Forest algorithm generates a large number of decision trees based on a bootstrap re-sample of observations (plant samples) presented to the model and then the tree with the majority of votes defines the aggregate model (Liland, 2011).

Each tree is generated using a different bootstrap sample of all of the barcode ions from the training set of defined Aizoaceae leaf sample classes. In the creation of each tree, about one third of the bootstrap observations are left out of the construction of the model. The data that was left out of the model generation is then tested on the model and the resulting differences are reported as the misclassification rate and the OOB error. Mitchell (2011) explored the use of this method for cross-validation in cases where there are many more variables than there are samples and found that the predictive power is actually higher than what is suggested by the OOB or that it is a pessimistic representation of model accuracy. This is a useful starting point for a study with about 23,000 variables distributed over about 50 samples. Thus, assuming that a low OOB error is achieved, the model should successfully predict the classes of unknown samples, or a testing set (Breiman, 2001).

As an additional cross validation step, K-means cross validation can also be employed. In K-means, K% of the data are left out of the model and the model is then generated on the remainder of the data. This is rotated across the entire dataset until all of the data has been either trained or tested and the resulting difference represents the model error (Xia et al., 2009). To test the predictive robustness of the model the logarithmic loss function was employed.

5.5.2 Using the parameters tested to determine model robustness

Because climate played a considerable role in the physiological and nutrient response of the plants (see Chapter 4) and because at least some of the informative ion intensities were correlated with changes in high temperature and solar radiation (see Figure 5.13), the samples were divided strictly by seasons. Two seasons were defined in Chapter 2, the first being winter, which lasted from April 2011 to August 2011 and the second, summer, which lasted from October 2011 to March 2012. The September samples were removed from consideration as this was considered a transition month. As is shown in Table 5.1, a slightly uneven number of samples from each species is distributed to each season.

(a) Winter

Species	Number of samples
A. noctiflora	4
C. edulis	5
G. africana	5
R. robusta	5
T. fruticosa	4

(b) Summer

Species	Number of samples
A. noctiflora	4
C. edulis	6
G. africana	6
R. robusta	6
T. fruticosa	3

Table 5.1: Distribution of species samples across summer and winter months used in models

To distinguish the differences in various models using seasons to divide the samples, the same sample populations must be consistent each time. Because random forest is an ensemble method where many bootstrapped models are combined together to generate a final model, it is important to denote if the model is capable of fitting the entire training set accurately. To represent the probabilities of each of the sample to be properly classified, a visualisation of the scaling matrix was generated using a multidimensional scaling plot (MDS). As was true in the PCA and hierarchical clustering methods presented earlier, in MDS, the closer the representations of the sample classes are to each other and the less total overlap of samples, the better the classification. Overlap and sample nearness indicate an increased probability of misclassification error (Breiman, 2001).

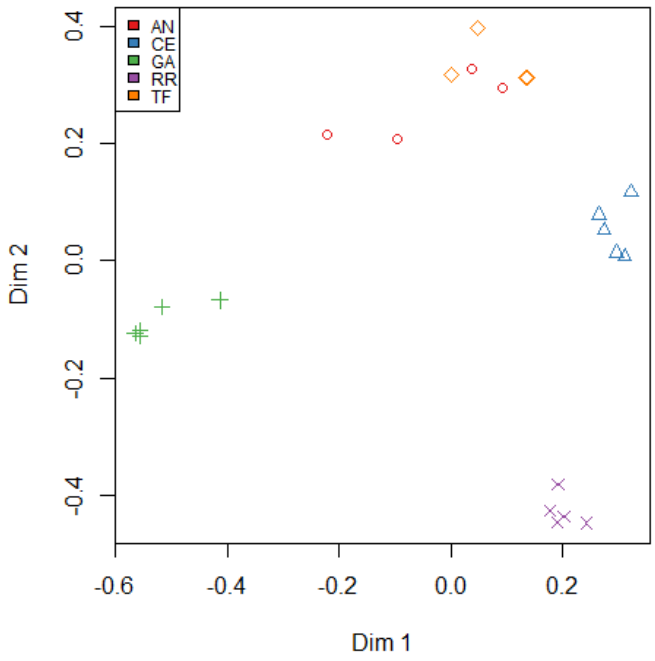
The ability of the model to accurately classify the sample classes in the training data is then represented in a confusion matrix where the misclassification error for each species class for the training set is presented along with the OOB and an additional cross-validation analysis using 10-fold K-means clustering.

To understand which variables contributed most to generating the model, the importance of the variables was determined from the OOB. The importance of each variable was assessed by removing it from the model and determining the resulting mean decrease in accuracy of the model as a whole as change in the OOB. This is plotted in accompanying dot charts where the higher the mean decrease in accuracy, the more important the variable was in the construction of the model (Shaik and Ramakrishna, 2014; Breiman, 2001).

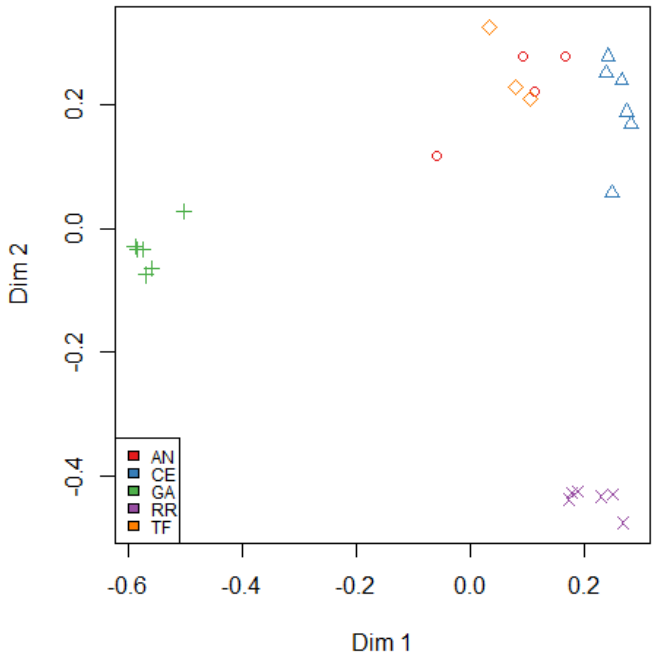
The robustness of the predictive power for each model for each season was then assessed by testing the model on the data from the opposite season. The predictive power of each model is presented as a log loss estimate where estimates closer to 0 represent more robust models.

5.6 Model based on barcode variables only

Samples were separated into winter only and summer only collections as described above. A classification model was then generated for each season based only on the barcode ions using the pre-set settings of R's "randomForest" package which automatically generated 500 decision trees with which to build a blended forest model. The first two dimension of the MDS plots presented in Figure 5.14 show how well the generated winter(Figure 5.14a) and summer (Figure 5.14b) models were respectively able to cluster the sample classes. In both cases, the first two dimensions of the MDS plot suggest strong classification of most of the sample classes except for *A. noctiflora* and *T. fruticosa*.



(a) *Winter*



(b) *Summer*

Figure 5.14: **MDS plots of barcode ion-based models.** Figure 5.14a represents the model based on the winter samples and Figure 5.14b represents the model based on the summer samples.

A confusion matrix was then generated to determine if the model was unable to correctly classify any of the samples used to build the model. The confusion matrices 5.2 indicates a perfect classification of the training data with 0 misclassifications in either the winter (Table 5.2a) or summer (Table 5.2b) models. Additionally, both models resulted in 0% OOB error. Further cross

validation using 10-fold K-means cross validation resulted in 0% error for the winter model and 2.4% error for the summer model. It was expected that there would be greater error in the summer model than in the winter model due to leaf abscission.

(a) Winter

Species	AN	CE	GA	RR	TF	Error
AN	4	0	0	0	0	0
CE	0	5	0	0	0	0
GA	0	0	5	0	0	0
RR	0	0	0	5	0	0
TF	0	0	0	0	4	0

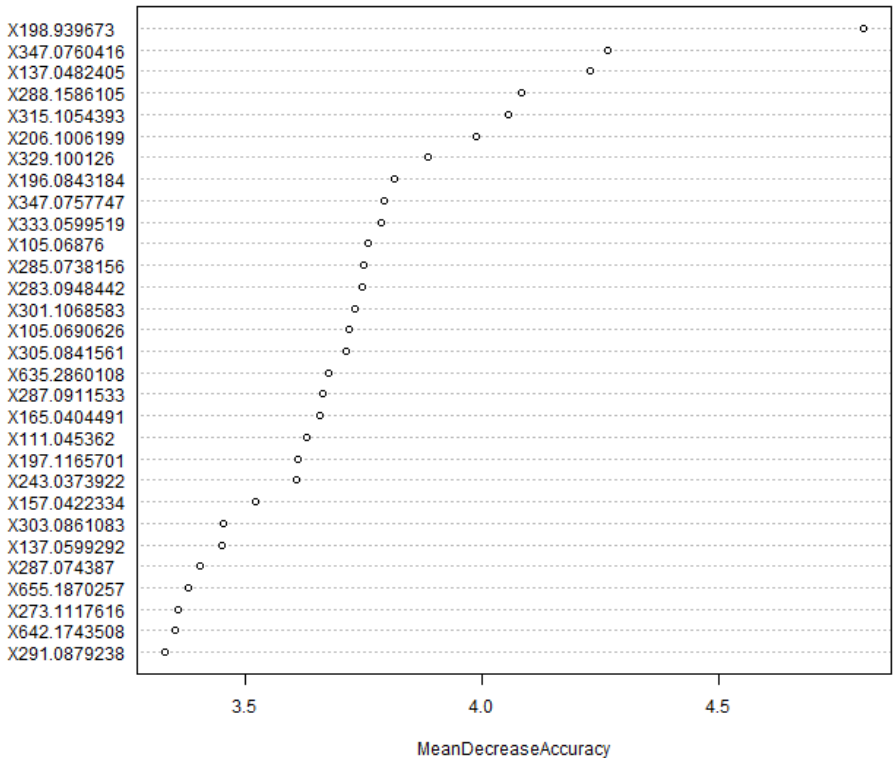
(b) Summer

Species	AN	CE	GA	RR	TF	Error
AN	4	0	0	0	0	0
CE	0	6	0	0	0	0
GA	0	0	6	0	0	0
RR	0	0	0	6	0	0
TF	0	0	0	0	3	0

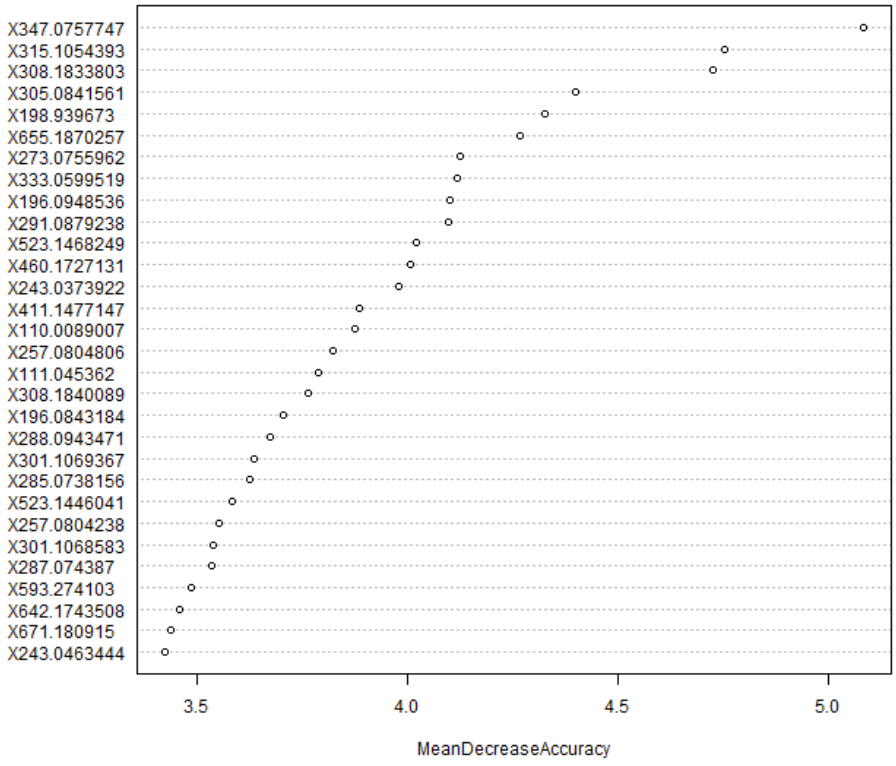
Table 5.2: **Confusion matrix from model based on leaf sample barcodes.** "AN" represents *A. noctiflora*, "CE" represents *C. edulis*, "GA" represents *G. africana*, "RR" represents *R. robusta*, "TF" represent *T. fruticosa*. Correct identifications can be seen in the matrix diagonal. Error represents the percentage of misclassifications for a particular species.

While the MDS plots of the winter and summer models (Figure 5.14) indicated that there was some overlap in sample classes and in particular in the samples of *A. noctiflora* and *T. fruticosa*, the models from both season were able to identify the samples of each class in the training set with a high very degree of accuracy.

The most important ions for model clustering were identified by ranking the mean decrease in accuracy of the ions used to generate the models as shown in Figure 5.15.



(a) *Winter*



(b) *Summer*

Figure 5.15: **Variable importance of models generated from barcode ions as indicated by mean decrease in importance.** Figure 5.15a displays the ions important in the winter model and Figure 5.15b displays the ions important in the summer model.

While the order of the most important ions changes slightly between the models, they are ultimately quite similar in the selection of which ions were used. There are fewer highly important ions in the winter model than there are in the summer model

(defined as mean decrease in accuracy > 4.0) which indicates that the summer barcodes are more stable than the winter barcodes. This makes sense as the onset of water deficit stress of the plants in the summer months would lead to less diverse metabolic activity.

5.6.1 Testing the models on the opposite seasons

The summer and winter models were then tested on the data remaining from the opposite seasons. The models were both able to predict the sample classes with 100% accuracy. The log loss estimate of prediction robustness resulted in a score of 0.39 for the winter model and 0.36 for the summer model which are highly significant scores. The slightly greater significance in the summer model log loss again is likely a reflection of the reduction in metabolic activity in the plants in the summer months.

From this, it can be concluded that despite the significant role that climate played on ion intensities over the year, the difference in ion intensities of barcode ions between the summer and winter samples was not sufficient to confuse the model.

5.6.2 Model stability considering the entire fingerprint

To determine how the models would perform on additional fingerprint data the entire fingerprints from the opposite seasons were tested to see if a stable model could be applied to full metabolic fingerprints rather than just isolated bar codes. As expected, when the entire fingerprints were applied with over 23,000 variables, there was a 0% OOB error rate. The lag in computation and the space requirements for RAM storage on whole metabolic fingerprints was significantly reduced when searching for a reduced number of ions. When attempts were made to fit a Random Forest model to complete metabolic fingerprints, it took two hours to compile. Fitting entire metabolic finger prints to a metabolic barcode model took fewer than 10 seconds.

5.7 Analysis of putatively identified compounds as chemotaxonomic markers

In keeping with more directed metabolic fingerprinting techniques, hierarchical clustering was then performed on the 74 compounds which were identified within 5-10 ppm of the calculated mass values from the internal database generated from literature reports of previously discovered Aizoaceae compounds and primary metabolites.

As is shown in Figure 5.16, species-specific clustering mostly occurs but does not agree with what has been previously determined from phylogenetics data where *C. edulis* is shown to be more closely related to *A. noctiflora* and *T. fruticosa* than it is to *R. robusta* (Klak et al., 2003, 2013).

G. africana samples A4 and A7 are separated from the rest of the main *G. africana* cluster by their high ion intensity for putatively identified (*E*)2',4'-dihydroxychalcone (ion 241.0865). The *G. africana* samples generally seem to cluster around (*E*)2',4'-dihydroxychalcone (ion 241.0865), 7,8-Dimethoxyflavanone (ion 285.1127), and Pinostrobin (ion 271.0970).

This dendrogram also does not separate all of the *T. fruticosa* and *A. noctiflora* samples. Interestingly, *R. robusta* seems to separate decisively from the other species, suggesting the presence of high concentrations of putatively identified procyanidin B2 (ion 579.1503) and catechin/epicatechin (ion 291.0869) which appear to separate it's samples from those of its close relative *C. edulis*.

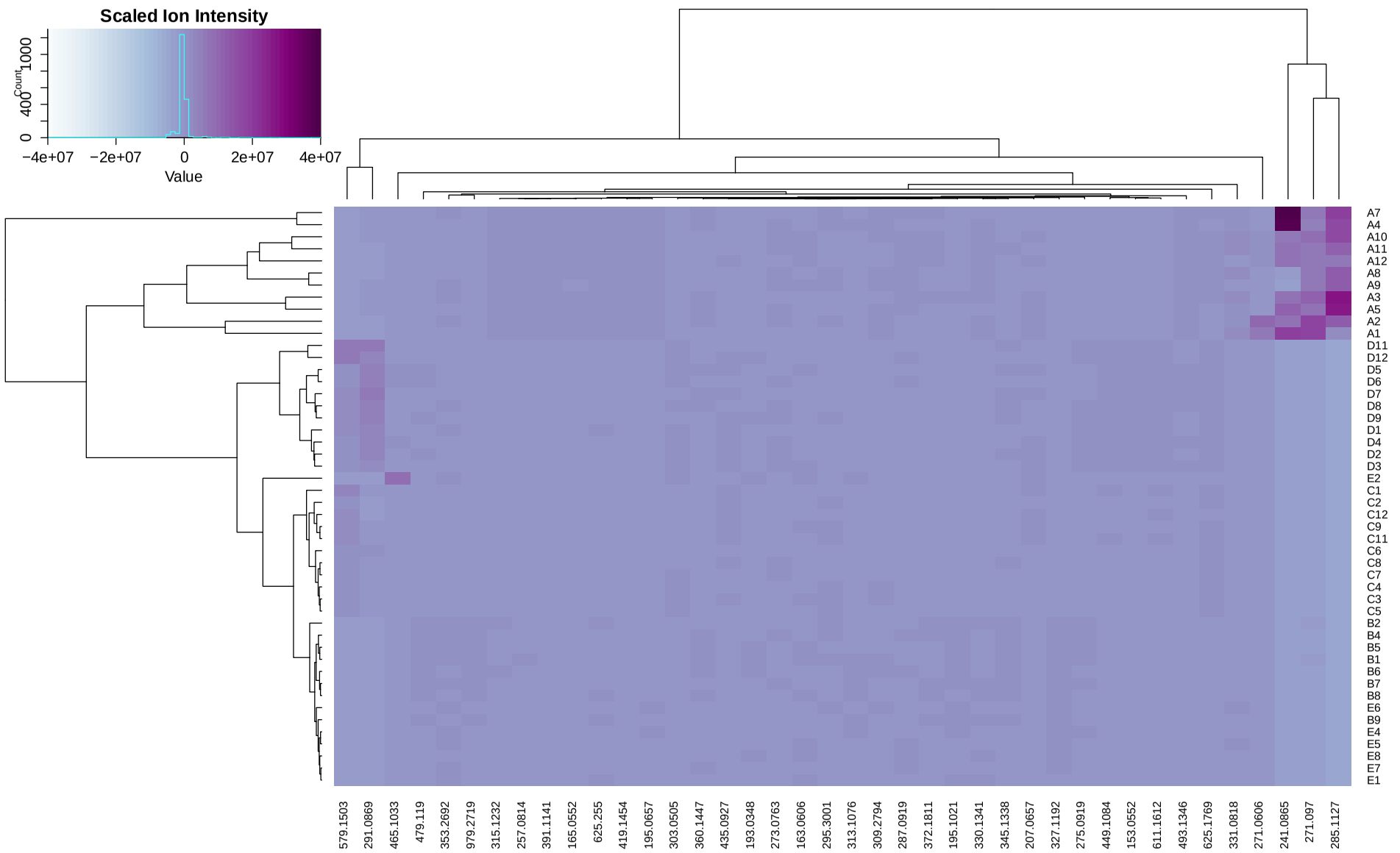


Figure 5.16: **Hierarchical clustering of ions representing putatively identified compounds from Aizoaceae literature.** Here the species are represented as 'A' for *G. africana*, 'B' for *A. noctiflora*, 'C' for *C. edulis*, 'D' for *R. robusta*, and 'E' for *T. fruticosa*.

The lack of horizontal distance between the individual samples or individual ions is mainly due to the low ion intensities of the vast majority of the ions considered. Thus, restricting the data to only these putatively identified compounds is not a basis for representing the distinctiveness of the species from each other.

This is corroborated further by the PCA analysis of these compounds, shown in Figure 5.17, where approximately 90% of the variance ($PC1 + PC2$) separates the individual *G. africana* samples farther from each other than any of the other species from each other. In fact, the other species are clustered so tightly on top of each other as to make them indistinguishable. As more secondary metabolites have been identified from *G. africana* than any of the other species in this study, this is not entirely unexpected. As was indicated in the dendrogram Figure 5.17, the variables with the greatest variance are most significant in *G. africana*. The variability described by PC2 serves to further separate the *G. africana* samples, thus the remaining variables are not distinctly different enough between the remaining Aizoaceae leaf samples to separate them from each other. This is further demonstrated by the fact that collectively PC1 and PC2 cover 90% of the total variance in this data set and only *G. africana* is distinctively separated.

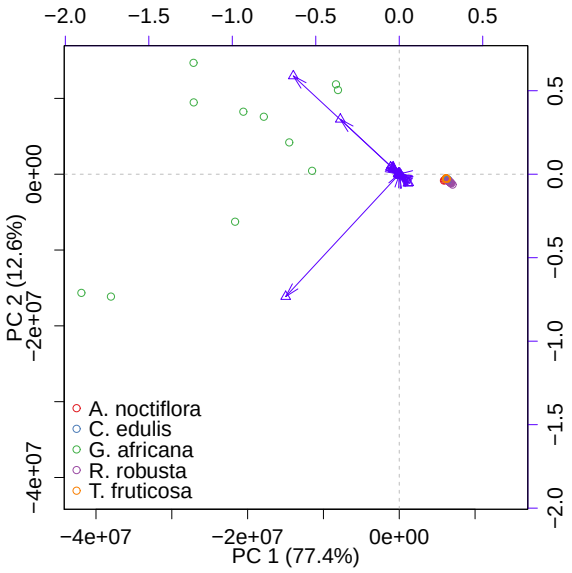


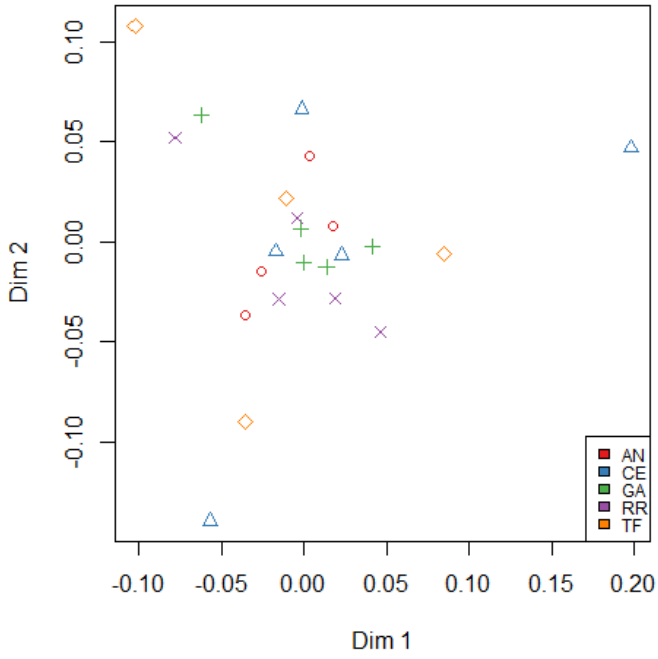
Figure 5.17: **PCA biplot of the tentatively identified compounds from all Aizoaceae leaf metabolite fingerprints.**

Both undirected clustering methods resulted in only a minor degree of taxon separation and were inconsistent with phylogenetic evidence. This indicates a general lack of robustness in the clustering using only this subset of ion data.

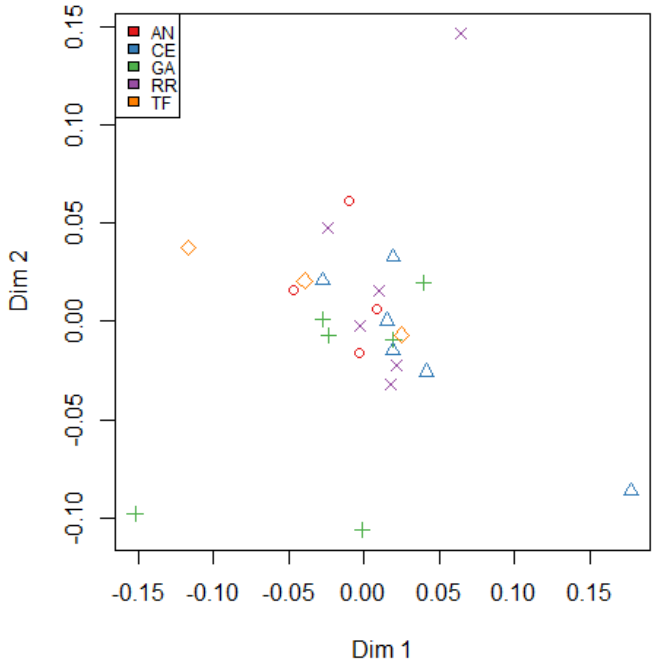
5.7.1 Model based on putatively identified compounds only

Because there was some separation in the hierarchical clustering method, an attempt was made at generating model for species classification using only these variables. To this end, samples were separated into winter only and summer only collections as described above and a classification model was generated for each season based only on the ions representing putatively identified compounds. The model was generated using the pre-set settings of R's "randomForest" package which automatically generated 500 decision trees with which to build a blended forest model.

The MDS plots in Figure 5.18 indicate that there is essentially zero separation between species in either the winter (Figure 5.18a)or the summer (Figure 5.18b) models generated using the putatively identified compounds. High levels of overlap and poor sample class separation in the MDS plot indicate that there is a very high probability of misclassification in the model.



(a) *Winter*



(b) *Summer*

Figure 5.18: **MDS plots of putatively identified compound models.** Figure 5.18a represents the model based on the winter samples and Figure 5.18b represents the model based on the summer samples.

This was subsequently demonstrated by the confusion matrices for both models (Figure 5.3) where the winter model (Table 5.3a) shows a 100% misclassification rate for all but the *R. robusta* samples (with an 80% misclassification rate) and the summer model (Table 5.3b) shows 100% misclassification rate for two species and $> 50\%$ misclassification for the other three.

(a) Winter

Species	AN	CE	GA	RR	TF	Error
AN	0	0	3	0	1	1.00
CE	0	0	1	2	2	1.00
GA	0	0	0	3	2	1.00
RR	0	0	3	1	1	0.80
TF	0	0	2	2	0	1.00

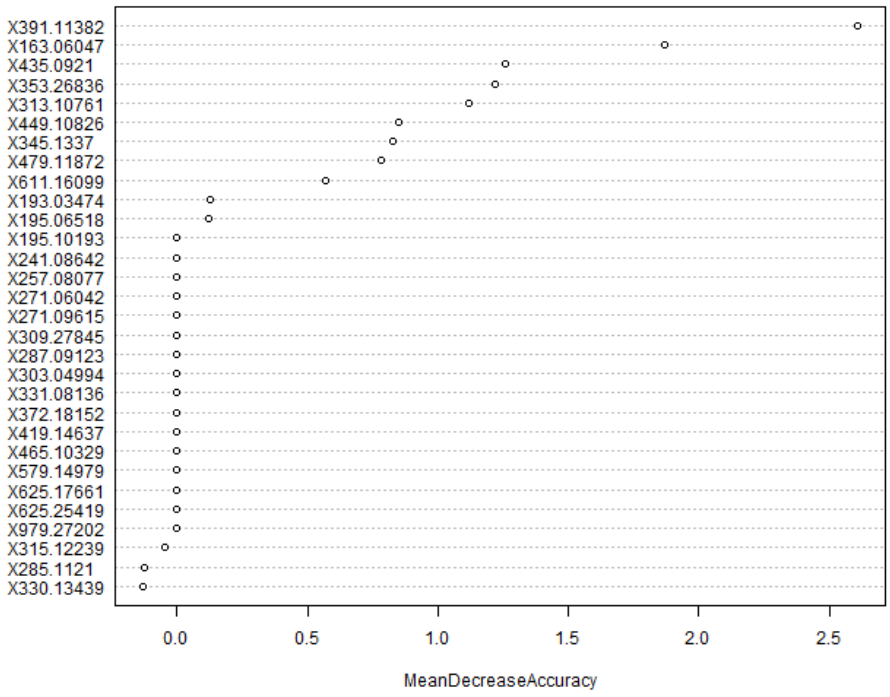
(b) Summer

Species	AN	CE	GA	RR	TF	Error
AN	0	1	2	1	0	1.00
CE	0	2	0	4	0	0.67
GA	0	2	2	2	0	0.67
RR	0	3	2	1	0	0.83
TF	0	0	1	2	0	1.00

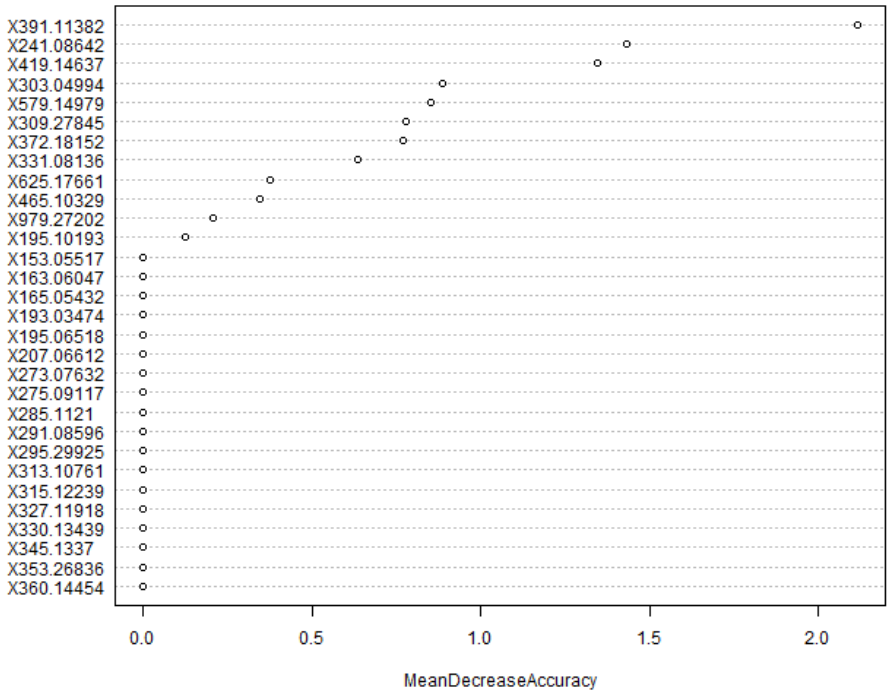
Table 5.3: **Confusion matrix from model based on putatively identified compounds from leaf samples.** "AN" represents *A. noctiflora*, "CE" represents *C. edulis*, "GA" represents *G. africana*, "RR" represents *R. robusta*, "TF" represent *T. fruticosa*. Correct identifications can be seen in the matrix diagonal. Error represents the percentage of misclassifications for a particular species.

The OOB for the winter model was 95.65% and for the summer model, 80.00%. The 10-Fold K-means cross validation error was 98.26% for the winter model and 87.60% for the summer model.

The variables important for the model classification were determined as shown in Figure 5.19a for the winter model and Figure 5.19b for the summer model. In both cases the majority of putatively identified compounds contribute nothing to the OOB, while the rest minimally increase the accuracy of the model. In the winter model, three compounds appear to negatively impact the OOB.



(a) *Winter*



(b) *Summer*

Figure 5.19: **Variable importance of models based on putatively identified compounds from leaf samples.** Figure 5.19a displays the ions important in the winter model and Figure 5.19b displays the ions important in the summer model.

Finally, the models were tested on the data from the opposite seasons resulting in a log loss of 1.59 for the winter model and 1.57 for the summer model.

While it is probable that at least some of the putative identifications made represent the compounds reported in the literature

for Aizoaceae species, at worst these are a random sampling of a relatively small number of ions. Here we showed that even this random sampling resulted in almost species-specific clustering in HCA analysis. Unfortunately the clustering of all but the *G. africana* samples was quite weak and, when comparing OOB, cross validation, and log loss, a classification model based on these ions was much weaker than a model generated from the ions which constitute unique chemical signatures for these species. This suggests that care should be taken when considering feature selection based on available literature only.

5.8 Monte Carlo of 500 random samplings of 125 ions

To determine how well the bar code model compared to comparable models based on variables randomly selected from the 23,307 ions from the original data matrix, a Monte Carlo method was employed with 500 sets of randomly selected groups of 125 ions sampled without replacement. The models were again generated using the pre-set settings of R's "randomForest" package.

For illustrative purposes, the confusion matrices and OOB for 5 models from each season are herein displayed in Table 5.4. For each model, the OOB is between 65% and 100% which indicates that the ability of the models to identify species from the training data set is very low when considering 125 randomly selected variables. The distribution of OOB for the 500 winter and summer models respectively is shown in Figure 5.20 where the average OOB is 88.94% and 89.13% for the winter and summer models respectively.

(a) Ions A Winter Model- 91.3% OOB

Species	AN	CE	GA	RR	TF	Error
AN	1	1	0	2	0	0.75
CE	0	0	0	5	0	1.00
GA	1	0	0	3	1	1.00
RR	1	1	2	1	0	0.80
TF	0	2	1	1	0	1.00

(b) Ions A Summer Model- 80.0% OOB

Species	AN	CE	GA	RR	TF	Error
AN	0	0	3	1	0	1.00
CE	0	1	0	5	0	0.83
GA	0	1	3	2	0	0.50
RR	0	5	0	1	0	0.83
TF	0	0	1	2	0	1.00

(c) Ions B Winter Model- 95.6% OOB

Species	AN	CE	GA	RR	TF	Error
AN	0	2	0	2	0	1.00
CE	0	1	0	4	0	0.80
GA	0	1	0	4	0	1.00
RR	1	4	0	0	0	1.00
TF	0	4	0	0	0	1.00

(d) Ions B Summer Model- 76% OOB

Species	AN	CE	GA	RR	TF	Error
AN	0	1	1	2	0	1.00
CE	0	0	0	6	0	1.00
GA	0	1	3	2	0	0.50
RR	0	3	0	3	0	0.50
TF	0	1	2	0	0	1.00

(e) Ions C Winter Model- 91.3% OOB

Species	AN	CE	GA	RR	TF	Error
AN	0	0	1	3	0	1.00
CE	0	0	0	5	0	1.00
GA	0	2	0	2	1	1.00
RR	0	2	1	2	0	0.60
TF	1	0	2	1	0	1.00

(f) Ions C Summer Model- 88% OOB

Species	AN	CE	GA	RR	TF	Error
AN	0	0	3	1	0	1.00
CE	0	1	0	5	0	0.83
GA	0	2	2	2	0	0.67
RR	0	6	0	0	0	1.00
TF	0	3	0	0	0	1.00

(g) Ions D Winter Model- 86.96% OOB

Species	AN	CE	GA	RR	TF	Error
AN	0	2	1	1	0	1.00
CE	0	2	0	3	0	0.60
GA	0	4	1	0	0	0.80
RR	0	5	0	0	0	1.00
TF	0	3	1	0	0	1.00

(h) Ions D Summer Model- 68.00% OOB

Species	AN	CE	GA	RR	TF	Error
AN	0	2	2	0	0	1.00
CE	0	3	0	3	0	0.50
GA	0	1	5	0	0	0.17
RR	0	6	0	0	0	1.00
TF	0	2	1	0	0	1.00

(i) Ions E Winter Model- 95.65% OOB

Species	AN	CE	GA	RR	TF	Error
AN	0	0	4	0	0	1.00
CE	0	0	2	3	0	1.00
GA	1	0	1	3	0	0.80
RR	0	3	2	0	0	1.00
TF	0	1	2	1	0	1.00

(j) Ions E Summer Model- 80.00% OOB

Species	AN	CE	GA	RR	TF	Error
AN	0	0	3	1	0	1.00
CE	0	2	0	4	0	0.67
GA	0	4	2	0	0	0.67
RR	0	5	0	1	0	0.83
TF	0	3	0	0	0	1.00

Table 5.4: **Confusion matrix from model based on summer leaf sample barcodes.** "AN" represents *A. noctiflora*, "CE" represents *C. edulis*, "GA" represents *G. africana*, "RR" represents *R. robusta*, "TF" represent *T. fruticosa*. Correct identifications can be seen in the matrix diagonal. Error represents the percentage of misclassifications for a particular species. In models A-E, summer and winter models are based on the same subset of ions.

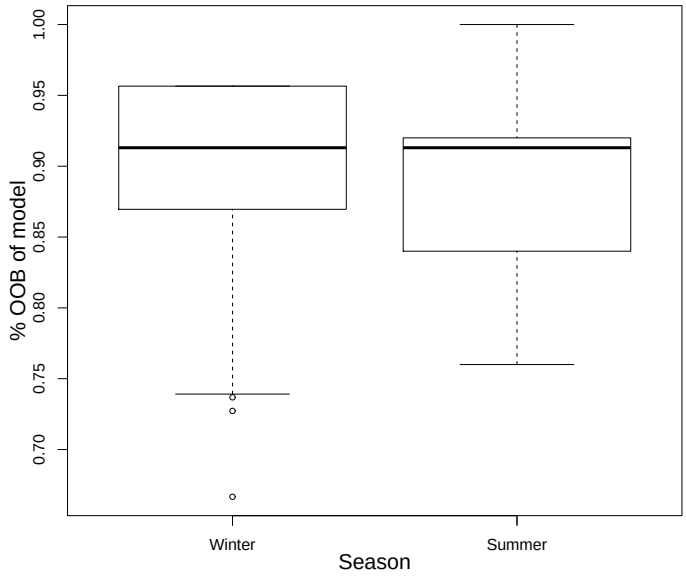


Figure 5.20: **Boxplot of OOB of models generated from random selection of ions.** Bars represent the average log loss of models A-E for Winter samples and Summer samples respectively.

After the models were generated, they were tested on the data from the opposite seasons. This resulted in very high average log loss of 1.53 for the winter models and 1.50 for the summer models. As is shown in Figure 5.21, the average log loss of winter and summer models was very close, as was the dispersion.

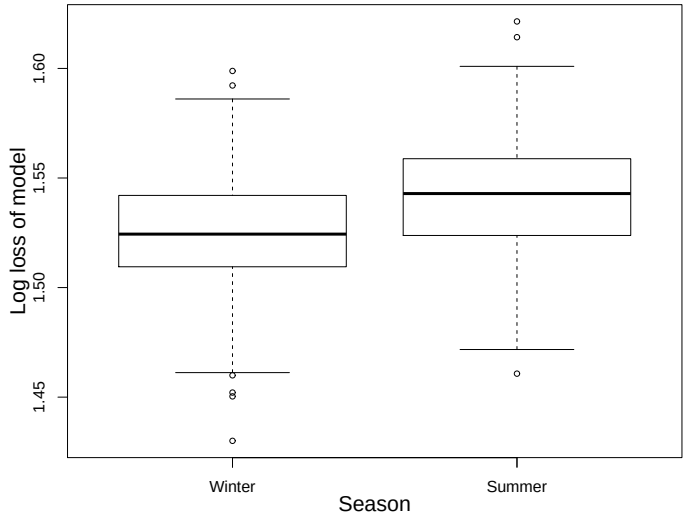


Figure 5.21: **Average log loss of models generated from random selection of ions.** Bars represent the average log loss of models A-E for Winter samples and Summer samples respectively.

5.9 Summary of findings from various classification models

The reduction of the metabolic fingerprints of the 5 Aizoaceae species into barcodes resulted in a reduced ion pool which was used to generate a highly specific classification model for those five species. Where previous phylogenetic comparisons have struggled to separate the morphologically distinct *C. edulis* and *R. robusta* species, the barcode clustering methods did so perfectly for the training set and the resulting log loss for the winter and summer models were both highly significant at 0.39 and 0.36 respectively, which additionally suggests high predictive robustness for models based on either season.

Further, it was also shown that the barcode model could be applied to an entire fingerprint such that minimal additional processing would be needed for high-throughput identification techniques to be employed in the identification of future samples from these species.

Modelling based only on putatively identified compounds showed little to no ability to appropriately classify species as was indicated by log loss values of 1.59 and 1.57 for the winter and summer models respectively. This is probably due to a variety of

factors, including the low total number of variables considered (74), no work having been been previously done on the metabolism of three of the five species being considered, but mainly that the ions selected showed very little change in intensity between species.

The average modelling based on 500 collections of 125 randomly selected ions also showed little predictive use with an average log loss of 1.53 and 1.55 for the winter and summer models collectively. While this the slight increase in predictive power over the putatively identified compounds was seen, this is more likely due to the low number of variables considered in the putative identification model rather than anything else.

Ultimately, the barcode models based on the winter and summer data faired about 5 time better in log loss than the models based on the putative compound identifications or randomly selected ions.

5.9.1 Final model with all of the samples considered

As the models were stable when only the winter data and only the summer data were considered, and were able to accurately predict the opposite season's samples respectively, the potential for over-fitting when considering the entire sample pool was considered unlikely. Thus, the final model could be built on a combination of all of the samples for use in future species predictions. To do this, a model was then generated using all of the leaf samples from all of the Aizoaceae species from the entire year considering the barcodes ions only. This resulted in a model with a 0% out of bag error rate (OOB) as can be seen in the confusion matrix in Table 5.5.

Table 5.5: **Confusion matrix from model based on leaf sample barcodes.** "AN" represents *A. noctiflora*, "CE" represents *C. edulis*, "GA" represents *G. africana*, "RR" represents *R. robusta*, "TF" represent *T. fruticosa*. Correct identifications can be seen in the matrix diagonal. Error represents the number of misclassifications for a particular species.

Species	AN	CE	GA	RR	TF	Error
AN	8	0	0	0	0	0
CE	0	11	0	0	0	0
GA	0	0	11	0	0	0
RR	0	0	0	11	0	0
TF	0	0	0	0	7	0

Sample identifications occurred with 100% accuracy resulting in 0% classification error. The OOB error rate is determined as the percentage of the sum of the misclassification errors with respect to the total number of samples tested. As is shown in Table 5.5, there were 0 misclassification errors, so the OOB error was 0%.

10-fold K-means cross validation was then employed as an additional cross validation step which also resulted in 0% error. Thus a strong model was possible with random testing.

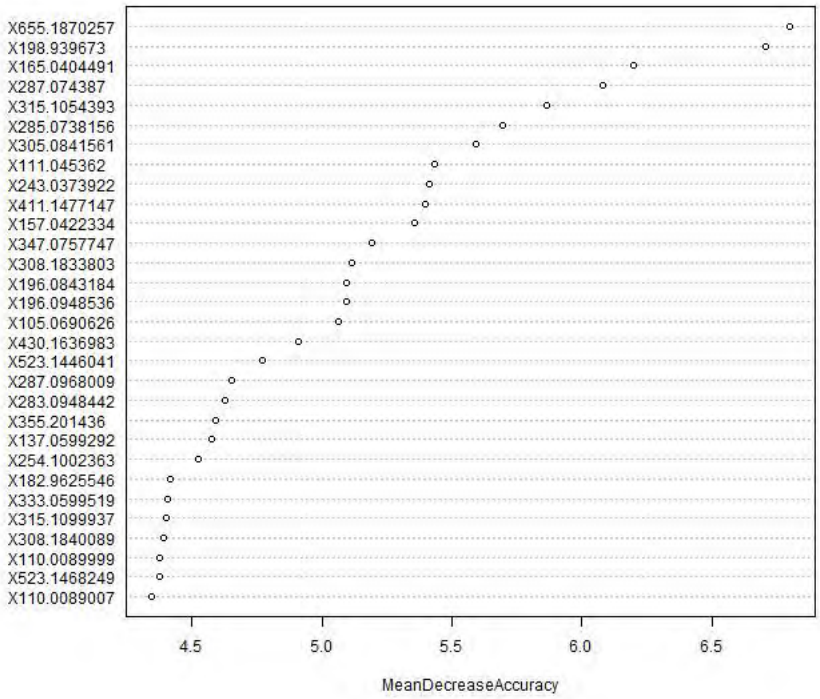


Figure 5.22: **Variable importance in model based on all samples.**

The total importance that the ions play in a mixed season model (6.8%) is significantly higher than in either the winter or the summer models (see Figure 5.15a (4.9%) and Figure 5.15a (5.1%)) and more ions play a more significant role in the model generation. Of the top 10 ions in the mixed model, all are in the top 10 important ions in either the winter or the summer models.

Unfortunately, as the dataset was sparse to begin with, it is not possible to calculate a log loss from this final model.

Chapter 6

Conclusions and the next steps

This study was performed to test the hypothesis that a specific subset of LC-MS ions were consistent and stable enough features to be used to identify closely related plant species from each other. Because general metabolism is known to change with environmental conditions, samples were collected over a year in order to accommodate typical annual metabolic flux. Various climate, nutrient, and physiological parameters were also analysed to identify time points which demonstrated the greatest variation in overall metabolism and to aid in model validation. Ultimately, a feature selection pipeline was developed to identify ion candidates from LC-MS metabolic fingerprints containing over 23,000 ions from plant extracts of five different Aizoaceae species, or to make species' barcodes.

It was then demonstrated that the various climate and nutrient factors analysed were highly correlated with the intensities of many of the barcode ions. The next step was to generate a classification model. By dividing the metabolic barcodes of the individual plant samples into testing and training sets around the sampling times where various measurements suggested that metabolic transitions had occurred, it was shown that despite sometimes drastic changes in ion intensities, the classification models could still distinguish the species with a high degree of accuracy. This suggests that the fingerprints were stable when all of the conditions were held equal - the same plant populations, the same person processing the material, the same LC-MS used to analyse the samples. Using this platform, it was also shown that it is possible to input the entire metabolic fingerprint of a sample into the model based on barcodes and still successfully identify species, which greatly reduced computational needs for identification and further suggested model robustness.

When clustering analyses were applied to the compounds putatively identified from the literature together with various primary metabolites, there was generally species-specific clustering, but the model generated from this data had very little predictive power, indicating a lack of robustness in using such a targeted subset of the data for classification model generation. These results are at least somewhat biased by the number of studies which have examined *G. africana* and *C. edulis* as the other three study species have not been chemically profiled at any level.

The Monte Carlo model averages of 500 models of 125 randomly selected ions faired slightly better than the models generated from the putatively identified compounds but their average log loss was still five times higher than the barcode models. The results of all of these experiments further suggest that the barcode model was robust and that the feature selection pipeline choose highly specific ions for the species of interest.

Considering the outcome of the statistical analyses, the barcode method appears to represent a new way of approaching metabolic fingerprinting and species identification which could be employed on a greater scale for understanding chemobiodiversity in a rapid, unique, and highly specific way. In addition, by first comparing species which are more distantly related to generate a metabolic barcode, a database of unique chemical signatures would also be produced which could potentially be used for the identification of novel secondary structures.

The use of barcode model generated in this study allowed successful distinction of the taxonomically difficult *C. edulis* and *R. robusta*. However, even with the many ways that metabolic fingerprints were assessed, the *T. fruticosa* samples always grouped more closely with the *A. noctiflora* samples than with *G. africana* samples with which it is supposed to share a subfamily (Klak et al., 2003). Further analyses will have to be conducted to ascertain where this discrepancy arises. While metabolic barcoding should not replace phylogenetic or morphological studies for biodiversity and taxonomic assessment, this study presents significant evidence that metabolic barcode analysis can be used as a complementary technique, and in the future, perhaps as a preliminary assessment of species classes.

Experiments considering climate, nutrients, and physiology revealed the importance of high temperatures and solar radiation on the nutrient uptake and the physiology of the species studied. Interestingly, Na appears to be selectively concentrated at extraordinary levels, at times correlating strongly with N uptake across the Aizoaceae species. This suggests that these species may commonly use Na accumulation as a counterion to enhance nitrate uptake.

The δ^{13}C ratios of *A. noctiflora* suggest that this species transitioned from C3 to CAM over the course of the study period, indicating that at least this species can utilise facultative CAM. While none of the rest of the species appear to have transitioned

from C3 to CAM, all of the other Aizoaceae species studied (except for *T. fruticosa* where there was not enough data for analysis), have a significant inverse correlation between $\delta^{13}C$ ratios and temperature and solar radiation and significant correlations with leaf water content which at least suggests reduced carbon uptake during that time and potential CAM idling. Previous reports have suggested that *C. edulis* is also capable of facultative CAM (Winter et al., 1976), although this was not seen in this study.

Of the 108 compounds described in the literature from Aizoaceae species, 72 were sufficiently polar to have been solubilised in ethanol, the extraction solvent. Of those, 66% were putatively detected across all of the processed samples. While true positive hits can only really be determined with the use of analytical standards or through further physiochemical analyses, this is a good starting place to determine which standards should be considered. The major challenge in metabolomics work flows is the identification of unknown compounds due to the vast number of metabolites that living organisms are composed of. This is further complicated by the presence of human, fungal, and bacterial metabolites also stored in most LC-MS databases as well as many pharmaceuticals and other industrial compounds. The results here presented indicate that plant databases where metabolites can be selected on the basis of genetic lineages, such as from the Dictionary of Natural Products and KNApSAcK, are essential platforms to initiate future studies. As various secondary metabolites were putatively identified in the study species which have never been chemically profiled before these also offer a starting point to explore the evolutionary lineage of secondary metabolic pathways in the Aizoaceae family if some analytical standards can be selectively employed.

6.1 The next steps

While this study outlines a preliminary investigation into potentially useful new tools for the examination of biodiversity in plants using metabolic barcodes, a few additional experiments would be useful to further support the methods discussed.

6.1.1 Consideration of barcode stability over geographical distance

To improve the robustness of this system generally, it would be ideal to study a number of plants from the same species sampled from more geographically distant populations as greater ecological diversity will further impact metabolite production. Studies such as that of the volatile secondary metabolites in *Myrothamnus moschatus* (Baillon) Niedenzu by Randrianarivo et al. (2013) show distinct chemotypes which are regionally specific in this species.

This is especially important for species in families such as Aizoaceae with the recent emergence of many of its species as barcode ions may be more difficult to identify. Future studies should also focus more on the profiles of individual plants rather than pooled samples to get a better idea metabolite variation across individuals and will hopefully include a greater number of total samples.

6.1.2 Consideration of barcode stability over multiple LC-MS platforms

In order for the model described to be used practically, the robustness of the LC-MS method also needs to be ascertained across multiple LC-MS systems and with other detector types to establish the feasibility of its use outside of the platform herein utilised. Issues such as machine age, column integrity, and further analysis of preprocessing across multiple software platforms, are also critical for a true understanding of first fingerprint stability and then of barcode stability.

The ultimate strength of using a LC-MS system is the number of compounds which can be explored. The greatest potential of this project would be the identification of unique chemical signals for the discovery of novel chemical constructs. Thus, the next logical step for someone trying to apply this exact method who might be employing it on a regular basis, would be to isolate and characterise the informative ions.

Appendix A

Compound lists

Appendix A consists of various tables relating to ion data, including the primary and secondary metabolites considered for putative identification in Table A.2 and Table A.1 respectively. This appendix also contains a list of the accurate masses of the 125 barcode ions and their retention times in Table A.3.

A.1 Compounds considered for putative identification

108 compounds identified from various Aizoaceae species in previous studies from more than 30 journal articles and books were compiled in Table A.1. The majority of the compounds listed are secondary metabolites with some fatty acids. While we would expect to see these in various quantities in most plants, they are here listed as a way of further validating the LC-MS data pre-processing methodology as our extraction method and LC-MS parameters should select against the presence of fatty acids in our samples. A table of common plant primary metabolites follows in Table A.2.

Table A.1: **Compounds previously identified from Aizoaceae species.** "M" represents the accurate mass, "M+H" represents the protonated mass, and "M+EtOH" represents the mass of the protonated ester or ethyl glycoside of the various compounds being considered. Confidence of the identification is indicated in blue for ion masses found within 5 ppm of the expected mass and red for ion masses found within 10 ppm of the expected mass in at least one sample in the processed ion list. In cases where neither the "M+H" or the "M+EtOH" masses are highlighted, no ion was detected within 10 ppm of the expected mass in the final processed data.

Compound	Formula	M	M+H	M+EtOH	Reference	Species
4-Methoxybenzoic acid	C8H8O3	152.0473	153.0552		36	*T. portulacastrum*
5-Hydroxy-2-methoxy benzaldehyde	C8H8O3	152.0473	153.0552		36	*T. portulacastrum*
p-Coumaric acid	C9H8O3	164.0473	165.0552	193.0864	19	*C. edulis*
Cysteic Acid	C3H7NO5S	169.0045	170.0123	197.0358	n/a	n/a
Capric acid 10:0	C10H20O2	172.14630	173.15420	201.18550	8	*G. lotoides*
p-Propoxybenzoic acid	C10H12O3	180.0786	181.0865		36	*T. portulacastrum*
Citric acid+A13:A31	C6H8O7	192.027	193.0348	221.0661	17	*C. edulis*
Ferulic acid	C10H10O4	194.0579	195.0657	223.097	14	*C. edulis*
Benzoic acid, 4-ethoxy-, ethyl ester	C11H14O3	194.09430	195.10210		21	*S. portulacastrum*
Lauric acid 12:0	C12H24O2	200.1776	201.1855	229.2167	8, 18	*C. edulis G. lotoides*
Leptorumol	C11H10O4	206.0579	207.0657		10,13	*T. portulacastrum*
Myristic acid 14:0	C14H28O2	228.2089	229.2168	257.248	8, 18	*C. edulis G. lotoides*

(E)2',4'-dihydroxychalcone	C15H12O3	240.0786	241.0865		1, 4	*G. africana*
2',4'-dihydroxydihydrochalcone	C15H14O3	242.0943	243.1021		1, 2, 3, 5	*G. africana*
Pentadecanoic acid 15:0	C15H30O2	242.22460	243.23250	271.26380	8, 18	*C. edulis G. lotoides*
Palmitoleic acid 16:1	C16H30O2	254.22460	255.23250	283.26380	8, 18	*C. edulis G. lotoides*
Pinocembrin	C15H12O4	256.0736	257.0814		1	*G. africana*
(*E*)-3,2,4-trihydroxychalcone	C15H12O4	256.0736	257.0814		1, 3, 5	G. africana
Palmitic acid 16:0	C16H32O	256.2402	257.2481	285.2793	8, 18	*C. edulis G. lotoides*
5,7,2-trihydroxyflavone	C15H10O5	270.0528	271.0606		1, 2, 3	*G. africana*
Pinostrobin	C16H14O4	270.0892	271.097		1, 4	*G. africana*
Margaric acid	C17H34O2	270.2559	271.2637	299.295	18	*C. edulis*
Heptadecanoic acid 17:0	C17H34O2	270.25590	271.26380	299.29510	8	*G. lotoides*
(*2S*)-5,7,2'-Trihydroxyflavanone	C15H12O5	272.0685	273.0763		3	*G. africana*
Phloretin	C15H14O5	274.0841	275.0919		19	*C. edulis*
α-Linolenic acid 18:3 (w-3)	C18H30O2	278.22460	279.23250	307.26380	8, 18	*C. edulis G. lotoides*
Linoleic acid 18:2 (w-6)	C18H32O2	280.24020	281.24810	309.27940	8, 18	*C. edulis G. lotoides*
6,10,14-Trimethyl-2-methylenepentadecanal	C19H36O	280.27660	281.28440		32	*T. tetragonioides*
Oleic acid 18:1	C18H34O2	282.25590	283.26380	311.29510	8, 21	*G. lotoides S. portulacastrum*
7,8-Dimethoxyflavanone	C17H16O4	284.10490	285.11270		29, 33	T. expansa
Hexadecanoic acid, ethyl ester	C18H36O2	284.27150	285.27940		21	*S. portulacastrum*
Stearic acid 18:0	C18H36O2	284.27150	285.27940	313.31070	8, 18	*C. edulis G. lotoides*
dihydroechinoidinin	C16H14O5	286.0841	287.0919		1, 2, 3, 4, 5	*G. africana*
(*E*)-3,2,4-trihydroxy-3-methoxychalcone	C16H14O5	286.0841	287.0919		1, 3, 5	*G. africana*
(-)-Mesembrine	C17H23NO3	289.16780	290.17560		26	*M. anatomicum M. expansum M. tortuosum S. anatomicum S. expansum S. namaquense S. stricum S. tortuosum S. namaquense*

Mesembrenone	C17H21NO3	287.15210	288.16000		27, 28	*M. anatomicum M. expansum M. tortuosum S. anatomicum S. expansum S. tortuosum*
Mesembrine	C17H23NO3	289.16780	290.17560		21	*S. portulacastrum*
Catechin	C15H14O6	290.079	291.0869		1, 7, 14, 15, 22	*C. edulis*
()-Epicatechin	C15H14O6	290.079	291.0869		15, 16, 19, and 22	*C. edulis*
Mesembranol	C17H25NO3	291.18340	292.19130		20	*M. oppositifolia*
Mesembrinol	C17H25NO3	291.18340	292.19130		27, 28	*M. anatomicum M. expansum M. tortuosum S. anatomicum S. expansum S. tortuosum*
Phytal	C20H38O	294.29230	295.30010		32	*T. tetragonioides*
(*C*)-Methylflavone	C18H16O4	296.1049	297.1127		10	*T. portulacastrum*
Phytol	C20H40O	296.30790	297.31570		21	*S. portulacastrum*
Quercetin	C15H10O7	302.0427	303.0505		19	*C. edulis*
Arachidonic acid 20:4 (w-6)	C20H32O2	304.24020	305.24810	333.27940	8	*G. lotoides*
Eicosenoic acid 20:1	C20H38O2	310.28720	311.29510	339.32640	8	*G. lotoides*
5,2'-Dihydroxy-7-methoxy-6,8-dimethylflavone	C18H16O5	312.0998	313.1076		35, 13	*T. portulacastrum*
Arachidic acid 20:0	C20H40O2	312.3028	313.3107	341.3419	8, 18	*C. edulis G. lotoides*
4,4'-Oxyneolign-9,9'-dioic acid	C18H18O5	314.11540	315.12320		9	*A. cordifolia*
(+)-Sceletium A4	C20H24N2O2	324.18380	325.19160		25	*S. namaquense S. stricum*
Humilixanthin	C14H18N2O7	326.11140	327.11920	355.15050	23	*D. luteum L. aurantiacus*
(-)-Tortuosamine	C20H26N2O2	326.19940	327.20730		25	*S. namaquense S. stricum*
Heneicosylic acid	C21H42O2	326.3185	327.3263	355.3576	18	*C. edulis*
1-Docosanol	C22H46O	326.35490	327.36270		21	*S. portulacastrum*
Docosahexaenoic acid 22:6 (w-3)	C22H32O2	328.24020	329.24810	357.27940	8	*G. lotoides*
(*2S,E*)-*N*-[2-Hydroxy-2-(4-hydroxyphenyl)ethyl] ferulamide	C18H19NO5	329.12630	330.13410		12	*A. cordifolia*

Eupalitin	C17H14O7	330.07400	331.08180		29, 30	*S. portulacastrum*
Behenic acid 22:0	C22H44O2	340.33410	341.34200	369.37330	8, 18	*C. edulis G. lotoides*
3-Methoxy-4,4-oxyneolign-9,9-dioic acid	C19H20O6	344.12600	345.13380		9	*A. cordifolia*
3-Methoxy-2,4-oxyneolign-9,9-dioic acid	C19H20O6	344.12600	345.13380		9	*A. cordifolia*
9,12,15- Octadecatrienoic acid, 2,3-dihydroxypropyl ester, (Z,Z,Z)-	C21H36O4	352.26140	353.26920		21	*S. portulacastrum*
1-Monolinoleoylglycerol	C21H38O4	354.27700	355.28480		21	*S. portulacastrum*
Tricosylic acid	C23H46O2	354.3498	355.3576	383.3889	18	*C. edulis*
(*E*)-*N*-[2-Hydroxy-2-(4-hydroxy-3-methoxyphenyl)- ethyl] ferulamide	C19H21NO6	359.13690	360.14470		12	*A. cordifolia*
Lignoceric acid 24:0	C24H48O2	368.36540	369.37330	397.40460	8, 18	*C. edulis G. lotoides*
(*E*)-*N*-[2-(4-Hydroxyphenyl)-2-propoxyethyl] ferulamide	C21H25NO5	371.17330	372.18110		12	*A. cordifolia*
Dimethyl 3-methoxy-4,4-oxyneolign-9,9-dioate	C21H24O6	372.15730	373.16510		9	*A. cordifolia*
Betanidin	C18H16N2O8	389.09790	390.10580	418.13700	23	*M. edule*
Dopaxanthin	C18H18N2O8	390.1063	391.1141	419.1454	11, 23	*G. longum*
Cerotic acid	C26H52O2	396.3967	397.4046	425.4358	18	*C. edulis*
3,3,5-Trimethoxy-4,4-oxyneolign-9,9-dioic acid	C21H24O8	404.14710	405.15490		9	*A. cordifolia*
Squalene	C30H50	410.39130	411.39920		21	*S. portulacastrum*
Prodelphinidin B6	C21H18O9	414.09510	415.10290		29, 34	*N. meyeri*
Lupeone	C30H48O	424.3705	425.3783		4	*G. africana*
Montanic acid	C28H56O2	424.428	425.4359	453.4671	18	*C. edulis*
β-amyrin	C30H50O	426.3862	427.394		7,16,22	*C. edulis*
Vitamin E	C29H50O2	430.38110	431.38890		21	*S. portulacastrum*
Avicularin	C20H18O11	434.0849	435.0927		19	*C. edulis*
3,3,5,5-Tetramethoxy-4,4-oxyneolign-9,9-dioic acid	C22H26O9	434.15770	435.16550		9	*A. cordifolia*
Ethyl iso-allocholate	C26H44O5	436.3189	437.3267		21	*M. crystallinum*
Uvaol	C30H50O2	442.3811	443.3889		7,16,22	*C. edulis*
Quercitrin	C21H20O11	448.1006	449.1084		19	*C. edulis*
Oleanolic acid	C30H48O3	456.3603	457.3682	485.3994	7,16, 22	*C. edulis*

Hyperoside	C21H20O12	464.0955	465.1033		14	*C. edulis*
Isoquercitin	C21H20O12	464.0955	465.1033		19	*C. edulis*
Isorhamnetin 3-*O*-glucoside	C22H22O12	478.1111	479.119		19 and 15	*C. edulis*
Eupalitin 3-glucoside	C23H24O12	492.12680	493.13460		29, 31	*S. portulacastrum*
3-Acetyl aleuritolic acid	C32H50O4	498.3709	499.3787		36	*T. portulacastrum*
Betanin	C24H26N2O13	551.15080	552.15860	580.18990	23	*Mesembryanthemum spp. D. floribundum C. acinaciformis*
Isobetanin	C24H26N2O13	551.15080	552.15860	580.18990	23	*M. conspicuum M. edule*
Rhodopin	C40H58O	554.4488	555.4566		21	*M. crystallinum*
Oleic acid, eicosyl ester	C38H74O2	562.56890	563.57670		21	*S. portulacastrum*
Trianthenol	C40H78O	574.6053	575.6131		36	*T. portulacastrum*
Procyanidin B2	C30H26O12	578.1424	579.1503		15, 19, and 22	*C. edulis*
Rutin	C27H30O16	610.1534	611.1612		14	*C. edulis*
Neohesperidin	C28H34O15	610.1898	611.1976		14	*C. edulis*
Isorhamnetin glucosyl-rhamnoside	C28H32O16	624.169	625.1769		15	*C. edulis*
(*E*,*E*)-*N*,*N*-Dityramin-4,4'-dihydroxy-3,5'-dimethoxy-b,3'-bicinnamamide	C36H36N2O8	624.24720	625.25500		12	*A. cordifolia*
7-Hydroxy-1-(4-hydroxy-3-methoxyphenyl)-*N*2,*N*3-bis(4-hydroxyphenethyl)-6-methoxy-1,2-dihydro-naphthalene-2,3-dicarboxamide	C36H36N2O8	624.24720	625.25500		12	*A. cordifolia*
Lampranthin II	C34H34N2O16	726.19080	727.19870	755.22990	23	*L. peersii L. sociorum*
Monogalactosyldiacylglycerol	C47H78O10	802.5595	803.5673		7, 22	*C. edulis*
Procyanidin C1	C45H38O18	866.20580	867.21360		29, 34	*N. meyeri*
Mesembryanthin	C44H50O25	978.2641	979.2719		24	*M. crystallinum*

Citation	Author
1	Mativandlela 2006
2	Mativandlela et al. (2008)
3	Mativandlela et al. (2009)
4	Vries et al. (2005)
6	Khajuria et al. (1982)
7	Martins et al. (2011)
8	Mengesha and Youan (2010)
9	Dellagreca et al. (2005)
10	Kavitha et al. (2014)
11	Gandía-herrero et al. (2005)
12	Dellagreca et al. (2006)
13	Kokpol et al. (1997)
14	van der Watt and Pretorius (2001)
15	Falleh et al. (2011b)
16	Martins et al. (2011)
17	Wheat (2014)
18	Custódio et al. (2012)
19	Falleh et al. (2011a)
20	Chopin et al. (1984)
21	Sheela and Uthayakumari (2013)
22	Martins et al. (2010)
23	Harborne (1999a)
24	Vogt et al. (1999)
25	Hayashi et al. (2002)
26	Wang (1986)
27	Sun et al. (1998)
28	Harborne (1999a)
29	Harborne (1999b)
30	Quijano et al. (1970)
31	Khajuria et al. (1982)
32	Aoki et al. (1982)
33	Kemp et al. (1979)
34	Kolodziej (1983)
35	Kokpol et al. (1997)
36	Nawaz et al. (2001)

Table A.2: **Common plant primary metabolites.** "M" represents the accurate mass, "M+H" represents the protonated mass, and "M+EtOH" represents the mass of the protonated ester or ethyl glycoside of the various compounds being considered. Confidence of the identification is indicated in blue for ion masses found within 5 ppm of the expected mass and red for ion masses found within 10 ppm of the expected mass in at least one sample in the processed ion list. In cases where neither the "M+H" or the "M+EtOH" masses are highlighted, no ion was detected within 10 ppm of the expected mass in the final processed data.

Name	M	M+H	M+EtOH

Ribitol	152.06847	153.0763	n/a
L-Valine	117.07898	118.08681	146.11811
L-Serine	105.04259	106.05042	134.08172
L-Leucine	131.09463	132.10246	160.13376
L-Threonine	119.05824	120.06607	148.09737
L-Glycine	131.09463	132.10246	160.13376
L-Alanine	89.04768	90.05551	118.08681
L-Methionine	149.05105	150.05888	178.09018
L-Proline	115.06333	116.07116	144.10246
L-4-hydroxyproline	131.05824	132.06607	160.09737
L-Phenylalanine	165.07898	166.08681	194.11811
L-Glutamic Acid	147.05316	148.06099	176.09229
L-Asparagine	132.05349	133.06132	161.09262
L-Aspartic Acid	133.03751	134.04534	162.07664
L-Glutamine	146.06914	147.07697	175.10827
L-Histidine	155.06948	156.07731	184.10861
L-Lysine	146.10553	147.11336	175.14466
L-Tyrosine	181.07389	182.08172	210.11302
L-Tryptophan	204.08988	205.09771	233.12901
L-Arginine	174.11168	175.11951	203.15081
L-Cysteine	121.01975	122.02758	150.05888
L-Isoleucine	131.09463	132.10246	160.13376
Citric Acid	192.027	193.03483	221.06613
Cysteic Acid	169.00449	170.01232	198.04362
Fumaric Acid	116.01096	117.01879	145.05009
DL-Isocitric Acid trisodium salt	257.97284	258.98067	287.01197
DL-Malic Acid	134.0215	135.02933	163.06063
Oxalic Acid	134.02152	135.02935	163.06065
Oxaloacetic Acid	132.00587	133.0137	161.045
Succinic Acid	118.02661	119.03444	147.06574
L(-)-Lactic acid	90.03169	91.03952	119.07082
L-(+)-Arabinose	150.05282	151.06065	179.09195
D-Arabinose	150.05282	151.06065	179.09195

Cellobiose	342.11621	343.12404	371.15534
Fructose	180.06339	181.07122	209.10252
L-(-)-Fucose	164.06847	165.0763	193.1076
D-(+)-Galactose	180.06339	181.07122	209.10252
Glucose	180.06339	181.07122	209.10252
1-Kestose	504.16903	505.17686	533.20816
1,1,1-Kestopentaose	828.27468	829.28251	857.31381
1,1-Kestotetraose	666.2219	667.22973	695.26103
D-(+)-Maltose	666.22186	667.22969	695.26099
D-(+)-Mannose	180.06339	181.07122	209.10252
Raffinose	504.16903	505.17686	533.20816
Ribose	150.05282	151.06065	179.09195
Stachyose	666.22186	667.22969	695.26099
Sucrose	342.11621	343.12404	371.15534
D-(+)-Trehalose	342.11621	343.12404	371.15534
D-(+)-Xylose	150.05282	151.06065	179.09195
2-ketoglutarate	189.98541	190.99324	219.02454
Fructose-6-phosphate	260.02972	261.03755	289.06885
D-arabitol	152.06847	153.0763	n/a
D-Erythrose	120.04226	121.05009	149.08139
Adonitol	152.06847	153.0763	n/a
D-mannitol	182.079	183.08683	n/a
D-sorbitol	182.07904	183.08687	n/a
Galactitol	182.07904	183.08687	n/a
Glycerol	92.04734	93.05517	n/a
Iso-erythritol	122.05791	123.06574	n/a
Maltitol	344.13186	345.13969	n/a
Xylitol	152.06847	153.0763	n/a

A.2 Barcode ions

The identification of a subset of ions from the metabolic fingerprints of the study species as described in Chapter 5 are presented in the following table.

Table A.3: **List of barcode ions.** Ions are arranged in order from most informative to least informative.

Ion	m/z	Rt
1	241.08576	30.02
2	121.05087	11.79
3	188.07016	14.38
4	315.08481	26.06
5	158.11644	12.22
6	317.09935	28.14
7	239.07037	24.86
8	329.09835	29.13
9	308.18338	15.61
10	287.09094	29.26
11	273.11176	26.77
12	621.30658	33.02
13	411.14771	5.37
14	179.03323	19.23
15	205.09511	14.37
16	137.04824	11.35
17	337.06997	26.13
18	347.07577	17.08
19	239.07396	23.57
20	315.10999	31.93
21	137.05993	6.82
22	165.04045	10.82
23	217.06830	5.39
24	116.07062	5.68
25	621.30681	32.63
26	121.05086	11.27
27	105.06876	14.76
28	287.07439	16.88
29	315.10544	17.54
30	138.05204	5.56
31	352.33867	30.92
32	355.20144	17.49

33	255.06409	26.93
34	303.08611	25.16
35	257.08042	28.63
36	104.10770	5.25
37	177.05318	15.13
38	321.21674	18.47
39	167.03352	29.43
40	111.04536	6.08
41	495.11380	16.70
42	121.05086	11.21
43	257.08048	28.20
44	241.08560	30.03
45	177.05332	15.81
46	166.08598	12.49
47	153.01807	25.17
48	460.17271	24.73
49	159.09090	14.38
50	333.05995	16.39
51	261.14337	13.97
52	105.06906	15.44
53	288.09435	29.46
54	254.10024	13.94
55	215.00652	7.19
56	291.08792	16.76
57	350.17561	27.63
58	347.07604	17.28
59	179.03323	19.01
60	525.30462	17.39
61	157.04223	12.81
62	110.00900	4.70
63	642.17435	16.36
64	255.06484	27.55
65	146.05862	14.22
66	283.09484	31.20

67	144.07941	14.52
68	215.01492	7.66
69	333.09657	24.00
70	203.05259	5.35
71	243.03739	12.89
72	151.06446	12.26
73	297.10912	24.96
74	641.17059	15.73
75	546.39826	29.07
76	621.30639	32.94
77	243.04634	13.28
78	523.14682	18.17
79	196.08432	12.23
80	177.08950	31.06
81	323.08821	27.11
82	523.14460	18.22
83	182.96255	4.64
84	593.27410	29.23
85	655.18703	17.05
86	273.07560	25.14
87	287.09680	11.16
88	196.09485	6.51
89	339.10748	17.42
90	271.09556	24.67
91	243.04696	13.50
92	130.04977	6.36
93	197.11657	20.03
94	198.93967	4.62
95	430.16370	25.31
96	119.04775	6.72
97	205.09647	14.38
98	179.03190	30.70
99	285.11115	32.52
100	206.10062	13.25

101	240.07283	25.65
102	317.10107	26.89
103	121.05087	12.01
104	671.18092	15.04
105	621.30664	32.86
106	110.00890	4.54
107	301.10694	29.15
108	657.16432	14.15
109	288.15861	17.08
110	301.10686	27.19
111	180.09249	6.47
112	158.11622	13.05
113	287.09111	26.63
114	287.09115	29.55
115	239.07080	24.03
116	285.07382	28.76
117	179.03326	18.18
118	292.18684	14.92
119	329.10013	31.53
120	411.19877	15.90
121	635.28601	29.95
122	233.06246	5.80
123	308.18401	15.50
124	379.26873	15.17
125	305.08416	6.30

Appendix B

R scripts

R code was written as a plain text formatting syntax or markdown file so that during development, each section could be individually tested. HTMLs are available upon request. Datasets and modifications thereof are referred to as "squid" for ease of identification.

B.1 Data normalisation

```
Normalise data- http://www.inside-r.org/packages/cran/MetabolAnalyze/docs/scaling
```{r}
centerSquid=scale(squid, center=T, scale=F)
library(MetabolAnalyze)
squid.sc=scaling(centerSquid, type = ''Pareto'')#scale
table(is.na(squid.sc))#Are there any NAs?
```
```

B.2 Hierarchical clustering

```
Hierarchical Clustering
```{r fig.width=10, fig.height=5}
par(mar=c(0,4,2,2))
par(oma=c(0,0,0,0))
d <- dist(squid.sc) # distance matrix
fit <- hclust(d, method=''ward.D'')
plot(fit, cex= 0.75, main=NULL) # display dendogram
```
```

B.3 PCA

```
PCA Page 46 ChemometricsWithR
```{r}
squid.svd = svd(squid.sc) #Singular value decomposition
squid.scores = squid.svd$u %*% diag(squid.svd$d) #Defining scores from SVD
squid.loadings = squid.svd$v #Defining loadings from SVD
squid.vars = squid.svd$d^2 / (nrow(squid) - 1) #The variation of each PC
squid.totalvar = sum(squid.vars)#The variation all PCs added together
squid.relvars = squid.vars / squid.totalvar #Fraction of var for each PC
variances = 100 * round(squid.relvars, digits = 3) #Final %variation of PCs
rownames(squid.scores) = row.names(squid) #Naming the score rows
rownames(squid.loadings) = (1:23307) #Naming the loading rows
```
```

```
Biplot visualization
```{r}
```
```

```
par(mfrow= c(1,1))
par(mar= c(4,4,2,2))
par(oma= c(0,0,0,0))
letter= c(rep(c(''G. africana'', ''A. noctiflora'', ''C. edulis'', ''R. robusta'', ''T. fruticosa''), time
squid.fin= cbind.data.frame(letter, squid.sc)
mslevels= squid.fin$letter
palette= brewer.pal(9, ''Set1'')
msdata.PCA= PCA(squid.fin[,2:23308])
biplot(msdata.PCA, pc=c(1,2), score.col=palette[(mslevels)], show.names=c(''none''))
legend(''bottomleft'', levels(mslevels), col=palette, pch=1, bty=''n'', y.intersp=1.25)
'''
```

B.4 How many PCs to use?

How many PCs to use? - relative variance

```
'''{r}
par(mfrow = c(1,2))
par(mar=c(4,4,1,1))
par(oma=c(0,0,0,0))
par(cex=.7,font=1)
squid.variances= 100*round(squid.relvars, digits = 3)
barplot(squid.variances[1:48], names.arg = paste(''PC'', 1:48), ylab=''Percent of total variance'',
xlab=''Principal Components'', ylim=c(0,55))
box()#All PCs
barplot(squid.variances[1:10], names.arg = paste(''PC'', 1:10), ylab=''Percent of total variance'',
xlab=''Top 10 Principal Components'', ylim=c(0,55))
box()#Top 10 PCs
relCumSum= cumsum(squid.variances)#cumulative summation of values
plot(relCumSum,
xlab=''All PCs'', ylab=''Percent of total variance'')
relWhich=which(relCumSum <90)#PCs reponsible for 90% of the variation
plot(relCumSum[relWhich], xlab=''PCs responsible for 90% of Variation'',
ylab=''Fraction of total variance'', ylim=c(50,100))
relWhich#Which PCs are responsible for 90% of the total variance?
'''
```

B.5 Weighing PCs

Weight PCs

```
'''{r}
par(mfrow = c(1,2))
par(mar=c(4,4,2,2))
par(oma=c(0,0,0,0))
squidHi=squid.loadings[,1:12] #Plot the values from the first 12 PCs
plot(squidHi, xlab = paste(''PC 1 ('', variances[1], ''%)'', sep = ''''), cex.lab=.75,
ylab = paste(''PC 2 ('', variances[2], ''%)'', sep = ''''), cex.lab=.75, main=''A'') #Plot first two
PCs
squidRel=squid.relvars[1:12]
heavySquid=squidHi*squidRel #Weight PCs by multiplying values by their relative variance
explained
plot(heavySquid, xlab = paste(''PC 1 ('', variances[1], ''%)'', sep = ''''), cex.lab=.75,
ylab = paste(''PC 2 ('', variances[2], ''%)'', sep = ''''), cex.lab=.75, main=''B'') #Plot weighted
first two PCs
'''
```

B.6 Determining leverage scores

```
Leverage scores determined for each ion from the cumulative variance of PCs
```{r}
varySquid=rowSums((heavySquid[,1:12])^2) #A leverage score is determined for each Ion from the
PCs covering 90% of the variance
perSquid= varySquid/(sum(varySquid))#Leverage score as percent of variance
plot(perSquid, xlab=''Ions Arranged by m/z'', ylab=''Leverage Score'')
quantile(perSquid) #Leverage score distribution
sortSquid=sort(perSquid, index.return=T, decreasing = F) #Sorting scores
squidCumSum= cumsum(sortSquid$x)#cumulative summation of values
plot(squidCumSum, ylab=''Cumulative sum of Ion Variance'', xlab=''Ions Arranged by Increasing
Variance'')
squidWhich=which(squidCumSum >.1)#Ions with variation 90%
plot(squidCumSum[squidWhich], xlab=''Number of Ions that make up 90% Variance'', ylab=''Leverage
Score (%)'')
```

```{r}
squidWhich=which(squidCumSum >.1)#Ions with variation 90%
plot(squidCumSum[squidWhich], xlab=''Number of Ions that make up 90% Variance'', ylab=''Leverage
Score (%)'')
```

Make matrix from informative ions
```{r}
squidIons= sortSquid$ix[squidWhich]#Create vector of informative ion positions
squidInformative= squid[,squidIons]#Create matrix of informative ions with their original
intensities
ncol(squidInformative)#How many informative ions are there?
print(colnames(squidInformative))#Ion names
write.table(squidInformative, ''c:/Users/User/Desktop/sw.csv'', sep='','')
quantile(rowSums(squidInformative))
```
```

B.7 Heatmap

```
Make heatmap
```{r, fig.width=10, fig.height=10}
par(mfrow = c(1,1))
par(mar=c(4,4,2,2))
par(oma=c(0,0,0,0))
library(RColorBrewer)
library(grDevices)
library(gplots)
cols=brewer.pal(5, ''Blues'')
pal=colorRampPalette(cols)

squidInformative.sc= squid.sc[,squidIons] #Create matrix of informative ions with their scaled
intensities
ncol(squidInformative.sc) #Confirm ion number
print(colnames(squidInformative.sc)) #Confirm ion names

squidInHeat= as.matrix(squidInformative) #Convert to matrix
heatmap.2(squidInHeat, tracecol=''white'', col=pal(50), sepcolor=Null, keysize = 1) #Generate
heatmap
```
```

B.8 Generating the various random forest classification models

B.8.1 Barcode models and putatively identified compound models

```
M4. Ensure that data has fewer than 32 levels
```{r}
which(sapply(squidBound, function(y) nlevels(y) > 32))
```
* There must be 0!!!

Create and combine a multitude of decision trees using 'randomForest' with 500 trees
```{r}
set.seed(456)
library(randomForest)#Load package
library(MASS)#Load package
library(ipred)
par(mfrow = c(1,1))
par(mar=c(4,4,2,2))
par(oma=c(0,0,0,0))

#10-fold cross-validation all barcode data
error.RF <- numeric(10)
for(i in 1:10) error.RF[i]= errorest(species~., data=squidBound, model=randomForest, mtry=2)$error
summary(error.RF)
```

Model building based on winter samples
```{r}
par(mfrow = c(1,1))
par(mar=c(4,4,2,2))
par(oma=c(0,0,0,0))
set.seed(125)
squidWinter=squidBound[c(1:5, 12:15, 20:24, 31:35, 42:45),]
squidFTrain= randomForest(species~., data=squidWinter, mtry=10, importance=T, prox=T,
do.trace=25)#Winter samples
print(squidFTrain)#Stats
plot(squidFTrain)#How many trees really needed for model?

MDSplot(squidFTrain, squidBound$species)
legend(``topleft'', legend=levels(squidBound$species), fill=brewer.pal(5, ``Set1''))

#10-fold cross-validation winter data
set.seed(131)
error.RF <- numeric(10)
for(i in 1:10) error.RF[i]= errorest(species~., data=squidBound[c(1:5,12:15, 20:24, 31:35,
42:45),], model=randomForest, mtry=10)$error
summary(error.RF)
```

Summer predictions
```{r}
squidSummer=squidBound[c(6:11, 16:19, 25:30, 36:41, 46:48),] #Define summer samples
squidPred= predict(squidFTrain, squidSummer)
print(squidPred)
table(squidTest$species)#How many observations of each species in test set
plot(squidPred, xlab=``Species'', ylab=``Number of Observations'', ylim=c(0,7))#How many samples
were identified as each species
box()
```
```

```
Flipping the testing and training data
```{r}
squidFTrain= randomForest(species~., data=squidSummer, mtry=10, importance=T, prox=T, do.trace=25)
#Winter samples
print(squidFTrain)#Stats
plot(squidFTrain)#How many trees really needed for model?

MDSplot(squidFTrain, squidBound$species)
legend(``topleft'', legend=levels(squidBound$species), fill=brewer.pal(5, ``Set1''))

#10-fold cross-validation winter data
set.seed(131)
error.RF <- numeric(10)
for(i in 1:10) error.RF[i]= errorest(species~., data=squidSummer, model=randomForest,
mtry=10)$error
summary(error.RF)

squidPred= predict(squidFTrain, squidWinter)
print(squidPred)
table(squidTest$species)#How many observations of each species in test set
plot(squidPred, xlab=``Species'', ylab=``Number of Observations'', ylim=c(0,7))#How many samples
were identified as each species
box()
```

What happens when the full metabolic fingerprint is applied to the model instead of just the
barcodes?
```{r}
squidBig= droplevels(data.frame(cbind(species, squid)))
squidPredict= predict(squidFTrain, squidBig)
print(squidPredict)
plot(squidPredict, xlab=``Species'', ylab=``Number of Observations'', ylim=c(0,12))#How many samples
were identified as each species
box()
```
```

This was also done for the model based on the internal database entries, the only modification being, the swapping of data matrices.

B.8.2 Monte Carlo model

```
Generate 500 models based on a random sampling of ions.
```{r}
resultW= vector(``list'',500)
resultS= vector(``list''500)
for (i in 1:500) {
#Generate Winter and Summer data sets
A=sample.int(23064, size =125, replace = FALSE)
SquidA=allSquid[,c(1,A)]
squidAW=SquidA[c(winter),]
squidAS=SquidA[c(summer),]

#Generating dummy probs
DummySProb=dummy(squidAS$species)
DummyWProb=dummy(squidAW$species)

#Generate predictive models
squidFAW= randomForest(species~.,squidAW, mtry=10, importance=T, prox=T)#Winter A
squidFAS= randomForest(species~.,squidAS, mtry=10, importance=T, prox=T)#Summer A
```
```

```
#Predicting opp Season
squidProbAS= predict(squidFAW, squidAS, type=''prob'')
squidProbAW= predict(squidFAS, squidAW, type=''prob'')

#Running log loss:
LLAS=MultiLogLoss(DummySProb,squidProbAS)
LLAW=MultiLogLoss(DummyWProb,squidProbAW)

#Save results
resultW[[i]]= LLAS
resultS[[i]]= LLAW
}
'''

'''{r}
par(mfrow = c(1,1))
#Log loss results:
W=as.numeric(resultW)
hist(W)
S=as.numeric(resultS)
hist(S)
logloss=c(W, S)

meanlossW=mean(W)
SDlossW=sd(W)
meanlossS=mean(S)
SDlossS=sd(S)

boxplot(W, S, xlab=''Season'', ylab=''Log loss of model'', cex=0.75, cex.axis=0.75)
axis(1, at=c(1,2), lab=c(''Winter'', ''Summer''), cex.axis=0.75)
'''

'''{r}
Winter Model
A=data.frame(resultW[[1]]$err.rate)
B=A[,1]
C=mean(B)#88.94%
D=sd(B)#5.95%

E=data.frame(resultS[[1]]$err.rate)
G=E[,1]
H=mean(G)#89.13%
I=sd(G)#5.23

boxplot(B, G, xlab=''Season'', ylab=''% OOB of model'', cex=0.75, cex.axis=0.75)
axis(1, at=c(1,2), lab=c(''Winter'', ''Summer''), cex.axis=0.75)
'''
```

Bibliography

Agam, N. and Berliner, P. (2006). Dew formation and water vapor adsorption in semi-arid environments- A review. *J. Arid Environ.*, 65(4):572–590.

Ahmad, M., Khan, M. A., Zafar, M., Arshad, M., Sultana, S., and Abbasi, B. H. (2010). Use of chemotaxonomic markers for misidentified medicinal plants used in traditional medicines. *J. Med. Plants Res.*, 4(13):1244–1252.

Aliferis, K. A. and Cubeta, M. A. (2013). Chemotaxonomy of fungi in the *Rhizoctonia solani* species complex performing GC/MS metabolite profiling. *Metabolomics*, 9:159–169.

Allsopp, N. (1999). Effects of grazing and cultivation on soil patterns and processes in the Paulshoek area of Namaqualand. *Plant Ecol.*, 142(1):179–187.

Allsopp, N., Laurent, C., Debeaudoin, L. M., and Igshaan Samuels, M. (2007). Environmental perceptions and practices of livestock keepers on the Namaqualand Commons challenge conventional rangeland management. *J. Arid Environ.*, 70(4):740–754.

Allwood, J. W. and Goodacre, R. (2010). An introduction to liquid chromatography-mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochem. Anal. PCA*, 21(1):33–47.

Aoki, T., Takagi, K., Hirata, T., and Suga, T. (1982). Two naturally occurring acyclic diterpene and norditerpene aldehydes from *Tetragonia tetragonoides*. *Phytochemistry*, 21(6):1361–1363.

Bath, G., Wyk, J. V., and Pettey, K. (2005). Control measures for some important and unusual goat diseases in southern Africa. *Small Rumin. Res.*, 60:127–140.

Berkov, S., Viladomat, F., Codina, C., Suárez, S., Ravelo, A., and Bastida, J. (2012). GC-MS of amaryllidaceous galanthamine-type alkaloids. *J. Mass Spectrom.*, 47(8):1065–1073.

Boccard, J., Kalousis, A., Hilario, M., Lantéri, P., Hanafi, M., Mazerolles, G., Wolfender, J.-L., Carrupt, P.-A., and Rudaz, S. (2010). Standard machine learning algorithms applied to UPLC-TOF/MS metabolic fingerprinting for the discovery of wound biomarkers in *Arabidopsis thaliana*. *Chemom. Intell. Lab. Syst.*, 104(1):20–27.

Borchardt, J. K. (2002). The beginnings of drug therapy: Ancient Mesopotamian medicine. *Drug News Perspect*, 15(3):187–192.

Botha, C. J. and Penrith, M.-L. (2008). Poisonous plants of veterinary and human importance in southern Africa. *J. Ethnopharmacol.*, 119(3):549–58.

Botha, C. J., Rundberget, T., Swan, G. E., Mülders, M. S. G., and Flåø yen, A. (2003). Toxicokinetics of cotyledoside following intravenous administration to sheep. *J. S. Afr. Vet. Assoc.*, 74(1):7–10.

Botha, C. J., Rundberget, T., Wilkins, L., Mülders, M. S., Flåø yen, A., and van Aardt, M. P. (2001). Seasonal variation in cotyledoside concentration of *Tylecodon wallichii* (Harv.) Tolken subsp. wallichii sampled in a krimpsiekte-prevalent region. *Onderstepoort J. Vet. Res.*, 68:1–9.

Bourgaud, F., Gravot, A., Milesi, S., and Gontier, E. (2001). Production of plant secondary metabolites: A historical perspective. *Plant Sci.*, 161(5):839–851.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45:5–32.

Calderón, A. I., Hodel, A., Craig, E., and Gupta, M. P. (2013). Triterpenes and fatty acids from Discophora guianensis identified by GCMS. *Biochem. Syst. Ecol.*, 50:16–18.

Carrick, P. J. (2003). The establishment ecology of two widespread Succulent Karoo plant species as a means of understanding change in rangelands. *Trans. R. Soc. South Africa*, 59(2):39–40.

Chopin, J., Dellamonica, G., Markham, K. R., Nair, A., and Gunasegaran, R. (1984). 2-p-coumaroylvitexin 7-glucoside from *Mollugo oppositifolia*. *Phytochemistry*, 23(9):2106–2108.

Custódio, L., Ferreira, A. C., Pereira, H., Silvestre, L., Vizetto-Duarte, C., Barreira, L., Rauter, A. P., Alberício, F., and Varela, J. a. (2012). The marine halophytes *Carpobrotus edulis* L. and *Arthrocnemum macrostachyum* L. are potential sources of nutritionally important PUFAs and metabolites with antioxidant, metal chelating and anticholinesterase inhibitory activities. *Bot. Mar.*, 55:281–288.

Dace, H. (2014). *The metabolomics of desiccation tolerance in Xerophyta humilis*. Masters, University of Cape Town.

D'Auria, J. C. and Gershenzon, J. (2005). The secondary metabolism of *Arabidopsis thaliana*: Growing like a weed. *Curr. Opin. Plant Biol.*, 8(3):308–16.

De Bruyne, K., Slabbinck, B., Waegeman, W., Vauterin, P., De Baets, B., and Vandamme, P. (2011). Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. *Syst. Appl. Microbiol.*, 34(1):20–29.

Dellagreca, M., Marino, C. D., Previtera, L., Purcaro, R., Zarrelli, A., Ii, F., Universitario, C., Sant, M., Cinthia, V., and Organica, C. (2005). Apteniols A-F, oxyneolignans from the leaves of *Aptenia cordifolia*. *Tetrahedron 61*, 61:11924–11929.

Dellagreca, M., Previtera, L., Purcaro, R., Zarrelli, A., Ii, F., and Universitario, C. (2006). Cinnamic acid amides and lignanamides from *Aptenia cordifolia*. *Tetrahedron*, 62:2877–2882.

Desmet, P. (2007). Namaqualand- A brief overview of the physical and floristic environment. *J. Arid Environ.*, 70(4):570–587.

Dieterle, F., Riefke, B., Schlotterbeck, G., Ross, A., Senn, H., and Amberg, A. (2011). NMR and MS methods for metabonomics. In Gautier, J.-C., editor, *Drug safety evaluation: Methods and protocols, methods in molecular biology*, volume 691 of *Methods in Molecular Biology*, pages 385–415. Humana Press, Totowa, NJ.

Driver, M., Raimondo, D., Maze, K., Pfab, M., and Helme, N. (2009). *Red list of South African plants*. Strelitzia, South African National Biodiversity Institute, Pretoria, South Africa, 25th edition.

Dunn, W. B., Erban, A., Weber, R. J. M., Creek, D. J., Neumann, S., Kopka, J., and Viant, M. R. (2013). Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, 9:44–66.

Dyubeni, L. and Buwa, L. V. (2012). An ethnobotanical study of plants used for the treatment of ear, nose and throat (ENT) infections in Nkonkobe Municipality, South Africa. *J. Med. Plants Res.*, 6(14):2721–2726.

El Bazaoui, A., Bellimam, M. A., and Soulaymani, A. (2011). Nine new tropane alkaloids from *Datura stramonium* L. identified by GC/MS. *Fitoterapia*, 82(2):193–197.

Elvira, M., Martucci, P., Vos, R. C. H. D., Carollo, C. A., and Gobbo-neto, L. (2014). Metabolomics as a potential chemotaxonomical tool: Application in the genus *Vernonia* Schreb. *PLoS One*, 9(4):1–8.

Ernst, M., Silva, D. B., Silva, R. R., Vêncio, R. Z. N., and Lopes, N. P. (2014). Mass spectrometry in plant metabolomics strategies: From analytical platforms to data acquisition and processing. *Nat. Prod. Rep.*, 31(6):784–806.

Falleh, H., Ksouri, R., Medini, F., Guyot, S., Abdelly, C., and Magné, C. (2011a). Antioxidant activity and phenolic composition of the medicinal and edible halophyte *Mesembryanthemum edule* L. *Ind. Crops Prod.*, 34:1066–1071.

Falleh, H., Oueslati, S., Guyot, S., Dali, A. B., Magné, C., Abdelly, C., and Ksouri, R. (2011b). LC/ESI-MS/MS characterisation of procyanidins and propelargonidins responsible for the strong antioxidant activity of the edible halophyte *Mesembryanthemum edule* L. *Food Chem.*, 127:1732–1738.

Farag, M. A., Porzel, A., and Wessjohann, L. A. (2012). Comparative metabolite profiling and fingerprinting of medicinal licorice roots using a multiplex approach of GC-MS, LC-MS and 1D NMR techniques. *Phytochemistry*, 76:60–72.

Field, B., Jordán, F., and Osbourn, A. (2006). First encounters- Deployment of defence-related natural products by plants. *New Phytol.*, 172(2):193–207.

Gandía-herrero, A. F., Escribano, J., García-Carmona, F., Planta, S., November, N., Gandia-herrero, F., and Garcia-carmona, F. (2005). Betaxanthins as pigments responsible for visible fluorescence in flowers. *Planta*, 222(4):586–593.

Gao, W., Yang, H., Qi, L.-W., Liu, E.-H., Ren, M.-T., Yan, Y.-T., Chen, J., and Li, P. (2012). Unbiased metabolite profiling by liquid chromatography-quadrupole time-of-flight mass spectrometry and multivariate data analysis for herbal authentication: classification of seven Lonicera species flower buds. *J. Chromatogr. A*, 1245:109–116.

Goldstein, P. Z. and Desalle, R. (2010). Integrating DNA barcode data and taxonomic practice: Determination, discovery, and description. *Bioessays*, 33:135–147.

Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., and Kell, D. B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.*, 22(5):245–252.

Govindaraghavan, S., Hennell, J. R., and Sucher, N. J. (2012). From classical taxonomy to genome and metabolome: Towards comprehensive quality standards for medicinal herb raw materials and extracts. *Fitoterapia*, 83(6):979–988.

Greshoff, M. (1909). Phytochemical investigations at Kew. *Bull. Misc. Inf.*, 10:397–418.

Halabalaki, M., Vougogiannopoulou, K., Mikros, E., and Skaltsounis, A. L. (2014). Recent advances and new strategies in the NMR-based identification of natural products. *Curr. Opin. Biotechnol.*, 25:1–7.

Hänsch, R. and Mendel, R. R. (2009). Physiological functions of mineral micronutrients (Cu, Zn, Mn, Fe, Ni, Mo, B, Cl). *Curr. Opin. Plant Biol.*, 12(3):259–266.

Harborne, J. B. (1999a). *Phytochemical dictionary: A handbook of bioactive compounds*. Taylor and Francis, Padstow, UK, 2nd edition.

Harborne, J. B. (1999b). *The handbook of natural flavonoids*. Wiley, Chichester.

Hartmann, T. (2007). From waste products to ecochemicals: Fifty years research of plant secondary metabolism. *Phytochemistry*, 68:2831–2846.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *Unsurpervised learning*. Springer New York, New York, NY, 2nd edition.

Hayashi, M., Unno, T., Takahashi, M., and Ogasawara, K. (2002). A new enantioselective route to the *Sceletium* alkaloids via a cyclopentanonecyclohexenone transformation. *Tetrahedron Lett.*, 43(8):1461–1464.

Heaton, T. H. E. (1987). The 15N/14N ratios of plants in South Africa and Namibia: Relationship to climate and coastal/saline environments. *Int. Assoc. Ecol.*, 74(2):236–246.

Hendricks, H. H., Novellie, P. A., Bond, W. J., and Midgley, J. J. (2002). Diet selection of goats in the communally grazed Richtersveld National Park Diet selection of goats in the communally grazed Richtersveld National Park. *African J. Range Forage Sci.*, 19(1):1–11.

Herrera, A. (2008). Crassulacean acid metabolism and fitness under water deficit stress: If not for carbon gain, what is facultative CAM good for? *Ann. Bot.*, 103(4):645–653.

Hoffman, M., Allsopp, N., and Rohde, R. (2007). Sustainable land use in Namaqualand, South Africa: Key issues in an interdisciplinary debate. *J. Arid Environ.*, 70(4):561–569.

Hoffman, T. (2005). Weather data for Paulshoek. Technical report, University of Cape Town, Cape Town.

Howley, T., Madden, M. G., Connell, M.-l. O., and Ryder, A. G. (2006). The effect of principal component analysis on machine learning accuracy with high dimensional spectral data. *Knowledge-Based Syst.*

Incerti, G., Romano, A., Termolino, P., and Lanzotti, V. (2013). Metabolomic fingerprinting using nuclear magnetic resonance and multivariate data analysis as a tool for biodiversity. *Plant Biosyst.*, 147(4):947–954.

Ji, H., Li, X., and Zhang, H. (2009). Natural products and drug discovery- Can thousands of years of ancient medical knowledge lead us to new and powerful drug combinations in the fight against cance and dementia? *EMBO Rep.*, 10(3):194–200.

Jones, H. G. (2013). Radiation. In Jones, H. G., editor, *Plants and microclimate: A quantitative approach to environmental plant physiology.*, chapter 2, pages 10–21. Cambridge University Press, Cambridge, UK, 3rd edition.

Jones, J. D. G. and Dangl, J. L. (2006). The plant immune system. *Nature*, 444(7117):323–329.

Kalra, Y. P. (1998). *Handbook of reference methods for plant analysis*. CRC Press, Boca Raton.

Kanfer, I. and Patnala, S. (2013). Chemotaxonomic studies of mesembrine-type alkaloids in *Sceletium* plant species. *African J. Sci.*, 109(3):5–9.

Kanniah, K. D., Beringer, J., North, P., and Hutley, L. (2012). Control of atmospheric particles on diffuse radiation and terrestrial plant productivity: A review. *Prog. Phys. Geogr.*, 36(2):209–237.

Karsai, I. and Kampis, G. (2010). The crossroads between biology and mathematics: The scientific method as the basics of scientific literacy. *Bioscience*, 60(8):632–638.

Katajamaa, M., Miettinen, J., and Oresic, M. (2006). MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, 22(5):634–636.

Katajamaa, M. and Oresic, M. (2005). Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics*, 6:179.

Kavitha, D., Parvatham, R., and Padma, P. R. (2014). Assessment of *Trianthema portulacastrum* for its antimicrobial potential and investigation of their phytochemicals using HPTLC, GC-MS, and IR. *Int. J. Pharm. Pharm. Sci.*, 6(1):675–686.

Kehoe, B. D. (1912). Preliminary note on the poisonous properties of *Cotyledon orbiculata*.

Kellerman, T. S. (2009). Poisonous plants. *Onderstepoort J. Vet. Res.*, 76(1):19–23.

Kellerman, T. S., Coetzer, J. A. W., and Naude, T. W. (1988). *Plant poisonings and mycotoxicoses of livestock in Southern Africa*. Oxford University Press, USA, Cape Town.

Kemp, M. S., Burden, R. S., and Brown, C. (1979). A new naturally occurring flavanone from *Tetragonia expansa*. *Phytochemistry*, 18(10):1765–1766.

Khajuria, R. K., Suri, K. A., Suri, O. P., and Atal, C. K. (1982). 3,5,4-trihydroxy-6,7-dimethoxyflavone 3-O-glucoside from *Sesuvium portulacastrum*. *Phytochemistry*, 21(5):1179–1180.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput. Biol.*, 8(2):1–10.

Kim, H., Choi, Y. H., Choi, K., Park, J. S., Kim, H. S., Jeon, J. H., Heu, M. S., Shin, D., and Lee, J. (2012). Metabolic classification of herb plants by NMR-based metabolomics. *J. Korean Magn. Reson. Soc.*, 16(2):91–102.

Kim, H. K. and Choi, Y. H. (2010). NMR-based metabolomic analysis of plants. *Nat. Protoc.*, 5(3):536–549.

Kim, H. K., Choi, Y. H., and Verpoorte, R. (2011). NMR-based plant metabolomics: where do we stand, where do we go? *Trends Biotechnol.*, 29(6):267–275.

Kingston, D. G. I. (2011). Modern natural products drug discovery and its relevance to biodiversity conservation. *J. Nat. Prod.*, 74(3):496–511.

Klak, C., Bruyns, P. V., and Hanáček, P. (2013). A phylogenetic hypothesis for the recently diversified *Ruschieae* (Aizoaceae) in southern Africa. *Mol. Phylogenet. Evol.*, 69(3):1005–1020.

Klak, C., Bruyns, P. V., and Hedderson, T. A. J. (2007). A phylogeny and new classification for *Mesembryanthemoideae* (Aizoaceae). *Taxon*, 56(3):737–756.

Klak, C., Khunou, A., Reeves, G., and Hedderson, T. (2003). A pylogenetic hypothesis for the Aizoaceae(Caryophyllales) based on four plastid DNA regions. *Am. J. Bot.*, 90(10):1433–1445.

Knowles, C.-l. (2005). *Synergistic effects of mixtures of the kresoxim-methyl fungicide and medicinal plant extracts in vitro and in vivo against Botrytis cinerea*. Magister scientiae, University of the Western Cape.

Kokpol, U., Wannachet-Isara, N., Tip-Pyang, S., Chavasiri, W., Veerachato, G., Simpson, J., and Weavers, R. T. (1997). A C-methylflavone from *Trianthema portulacastrum*. *Phytochemistry*, 44(4):719–722.

Kolodziej, H. (1983). The first naturally occurring 4-aryl flavan-3-ol. *Tetrahedron Lett.*, 24(17):1825–1828.

Lange, E., Tautenhahn, R., Neumann, S., and Gröpl, C. (2008). Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, 9(1):375.

Lewinsohn, E. and Gijzen, M. (2009). Phytochemical diversity: The sounds of silent metabolism. *Plant Sci.*, 176(2):161–169.

Li, J. W. and Vederas, J. C. (2009). Drug discovery and natural products: End of an era or an endless frontier? *Science.*, 325(5937):161–165.

Li, Y., Chen, J., Li, Y., Li, Q., Zheng, Y., Fu, Y., and Li, P. (2011). Screening and characterization of natural antioxidants in four *Glycyrrhiza* species by liquid chromatography coupled with electrospray ionization quadrupole time-of-flight tandem mass spectrometry. *J. Chromatogr. A*, 1218(45):8181–8191.

Libik, M., Konieczny, R., Surowka, E., and Miszalski, Z. (2005). Superoxide dismutase activity in organs of *Mesebryanthemum crystallinum* L. at different stages of CAM development. *ACTA Biol. Cracoviensia*, 47(1):199–204.

Liland, K. H. (2011). Multivariate methods in metabolomics From pre-processing to dimension reduction and statistical analysis. *Trends Anal. Chem.*, 30(6):827–841.

Liu, Z., Liu, Y., Liu, C., Song, Z., Li, Q., Zha, Q., Lu, C., Wang, C., Ning, Z., Zhang, Y., Tian, C., and Lu, A. (2013). The chemotaxonomic classification of *Rhodiola* plants and its correlation with morphological characteristics and genetic taxonomy. *Chem. Cent. J.*, 7(1):118.

Lorenz, P., Duckstein, S., Conrad, J., Knödler, M., Meyer, U., and Stintzing, F. C. (2012). An approach to the chemotaxonomic differentiation of two European dog's mercury species: *Mercurialis annua* L. and *M. perennis* L. *Chem. Biodivers.*, 9(2):282–297.

Maathuis, F. J. M. (2009). Physiological functions of mineral macronutrients. *Curr. Opin. Plant Biol.*, 12(3):250–258.

Mabona, U. and Van Vuuren, S. (2013). Southern African medicinal plants used to treat skin diseases. *South African J. Bot.*, 87:175–193.

MacKay, D. J. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press, 7th edition.

Mankga, L. T., Yessoufou, K., and Moteetee, A. M. (2013). Efficacy of the core DNA barcodes in identifying poorly conserved and processed plant materials commonly used in South African traditional medicine. *Zookeys*, 365:215–233.

Martins, A., Vasas, A., Schelz, Z. S., Viveiros, M., Molnár, J., Hohmann, J., and Amaral, L. (2010). Constituents of *Carpobrotus edulis* inhibit P-glycoprotein of MDR1-transfected mouse lymphoma cells. *Anticancer Res.*, 30:829–836.

Martins, A., Vasas, A., Viveiros, M., Molnár, J., Hohmann, J., and Amaral, L. (2011). Antibacterial properties of compounds isolated from *Carpobrotus edulis*. *Int. J. Antimicrob. Agents*, 37(5):438–444.

Mativandlela, S. P. N., Meyer, J. J. M., Hussein, A. A., Houghton, P. J., Hamilton, C. J., and Lall, N. (2008). Activity against *Mycobacterium smegmatis* and *M. tuberculosis* by extract of South African medicinal plants. *Phyther. Res.*, 22:841–845.

Mativandlela, S. P. N., Muthivhi, T., Kikuchi, H., Oshima, Y., Hamilton, C., Hussein, A. A., van der Walt, M. L., Houghton, P. J., and Lall, N. (2009). Antimycobacterial flavonoids from the leaf extract of *Galenia africana*. *J. Nat. Prod.*, 72(12):2169–2171.

Matsuda, F., Shinbo, Y., Oikawa, A., Hirai, M. Y., Fiehn, O., Kanaya, S., and Saito, K. (2009). Assessment of Metabolome Annotation Quality: A Method for Evaluating the False Discovery Rate of Elemental Composition Searches. *PLoS One*, 4(10):e7490.

Mattson, W. J. (1980). Herbivory in relation to plant nitrogen content. *Annu. Rev. Ecol. Syst.*, 11:119–161.

Mengesha, A. E. and Youan, B. C. (2010). Anticancer activity and nutritional value of extracts of the seed of *Glinus lotoides*. *J. Nutr. Sci. Vitaminol*, 56:311–318.

Messina, A., Callahan, D. L., Walsh, N. G., Hoebee, S. E., and Green, P. T. (2014). Testing the boundaries of closely related daisy taxa using metabolomic profiling. *Taxon*, 63(2):367–376.

Miller, R. O. (1998). High-temperature oxidation: Dry ashing. In Kalra, Y. P., editor, *Handbook of reference methods for plant analysis.*, chapter 5, pages 51–89. CRC Press.

Mitchell, M. W. (2011). Bias of the random forest out-of-bag (OOB) error for certain input parameters. *Open J. Stat.*, 2011(October):205–211.

Monteiro, M. S., Carvalho, M., Bastos, M. L., and Guedes de Pinho, P. (2013). Metabolomics analysis for biomarker discovery: Advances and challenges. *Curr. Med. Chem.*, 20(2):257–271.

Moore, B. D., Andrew, R. L., Carsten, K., and Foley, W. J. (2013). Explaining intraspecific diversity in plant secondary metabolites in an ecological context. *New Phytol.*, 10:1–18.

Mushtaq, M. Y., Choi, Y. H., Verpoorte, R., and Wilson, E. G. (2014). Extraction for metabolomics: Access to the metabolome. *Phytochem. Anal.*, 25(4):291–306.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403(6772):853–858.

Nawaz, H. R., Malik, A., and Ali, M. S. (2001). Trianthenol: an antifungal tetraterpenoid from *Trianthema portulacastrum* (Aizoaceae). *Phytochemistry*, 56(1):99–102.

Newman, D. J., Cragg, G. M., and Snader, K. M. (2000). The influence of natural products upon drug discovery. *Nat. Prod. Rep.*, 17(3):215–234.

Niewiadomska, E., Bilger, W., Gruca, M., Mulisch, M., Miszalski, Z., and Krupinska, K. (2011). CAM-related changes in chloroplastic metabolism of *Mesembryanthemum crystallinum* L. *Planta*, 233(2):275–285.

Nobeli, I., Ponstingl, H., Krissinel, E. B., and Thornton, J. M. (2003). A structure-based anatomy of the *E.coli* metabolome. *J. Mol. Biol.*, 334(4):697–719.

Non-Affiliated Soil Analysis Work Committee (1990). *Handbook of standard soil testing methods for advisory purposes*. Soil Science Society of South Africa, Pretoria, South Africa.

Nortje, J. M. (2011). *Medicinal ethnobotany of the Kamiesberg, Namaqualand, Northern Cape Province, South Africa*. Msc, University of Johannesburg, Johannesburg.

Novoa, A. and González, L. (2014). Impacts of *Carpobrotus edulis* (L.) N.E.Br. on the germination, establishment and survival of native plants: A clue for assessing its competitive strength. *PLoS One*, 9(9):1–12.

Nugent, R. and Meila, M. (2010). An overview of clustering applied to molecular biology. In Bang, H., Zhou, X. K., Mazumdar, M., and Van Epps, H. L., editors, *Statistical methods in molecular biology*, chapter 12, pages 369–404. Humana Press, Totowa, NJ, 1st edition.

Padial, J. M., Miralles, A., la Riva, I. D., and Vences, M. (2010). The integrative future of taxonomy- Review. *Front. Zool.*, 7(16):1–14.

Peterson, B. J. and Fry, B. (1987). Stable isotopes in ecosystem studies. *Annu. Rev. Ecol. Syst.*, 18(1):293–320.

Pichersky, E. and Gang, D. R. (2000). Genetics and biochemistry of secondary metabolites in plants: An evolutionary perspective. *Trends Plant Sci.*, 5(10):439–445.

Pilon-Smits, E. A., Quinn, C. F., Tapken, W., Malagoli, M., and Schiavon, M. (2009). Physiological functions of beneficial elements. *Curr. Opin. Plant Biol.*, 12(3):267–274.

Pluskal, T., Castillo, S., Villar-Briones, A., and Oresic, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11(1):395.

Pluskal, T., Uehara, T., and Yanagida, M. (2012). Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching. *Anal. Chem.*, 84(10):4396–4403.

Pool, E. J., Klaasen, J. A., and Shoko, Y. P. (2009). The immunotoxicity of *Dicerothamnus rhinocerotis* and *Galenia africana*. *African J. Biotechnol.*, 8(16):3846–3850.

Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., and Stolovitzky, G. (2010). Towards a rigorous assessment of systems biology models: The DREAM3 challenges. *PLoS One*, 5(2):1–18.

Quijano, L., Malanco, F., and Ríos, T. (1970). The structures of eupalin and eupatolin: Two new flavonol rhamnosides isolated from *Eupatorium ligustrinum* D.C. *Tetrahedron*, 26(12):2851–2859.

Radulović, N. S. and Dekić, M. S. (2013). Volatiles of *Geranium purpureum* Vill. and *Geranium phaeum* L.: chemotaxonomy of balkan *Geranium* and *Erodium* species (Geraniaceae). *Chem. Biodivers.*, 10(11):2042–2052.

Randrianarivo, E., Rasoanaivo, P., Nicoletti, M., Razafimahefa, S., Lefebvre, M., Papa, F., Vittori, S., and Maggi, F. (2013). Essential-oil polymorphism in the 'resurrection plant' *Myrothamnus moschatus* and associated ethnobotanical knowledge. *Chem. Biodivers.*, 10(11):1987–98.

Rasmann, S. and Agrawal, A. A. (2009). Plant defense against herbivory: Progress in identifying synergism, redundancy, and antagonism between resistance traits. *Curr. Opin. Plant Biol.*, 12:473–478.

Riginos, C. and Hoffman, M. T. (2003). Changes in population biology of two succulent shrubs along a grazing gradient. *J. Appl. Ecol.*, 40(4):615–625.

Rohde, R. F. and Hoffman, M. T. (2008). One hundred years of separation: The historical ecology of a South African Coloured Reserve. *Africa (Lond).*, 78(2):189–222.

Rubingh, C. M., Bijlsma, S., Derks, E. P. P. a., Bobeldijk, I., Verheij, E. R., Kochhar, S., and Smilde, A. K. (2006). Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics*, 2(2):53–61.

Safer, S., Cicek, S. S., Pieri, V., Schwaiger, S., Schneider, P., Wissemann, V., and Stuppner, H. (2011). Metabolic fingerprinting of *Leontopodium* species (Asteraceae) by means of 1H NMR and HPLC-ESI-MS. *Phytochemistry*, 72(11-12):1379–1389.

Sandasi, M., Kamatou, G. P., Combrinck, S., and Viljoen, A. M. (2013). A chemotaxonomic assessment of four indigenous South African *Lippia* species using GCMS and vibrational spectroscopy of the essential oils. *Biochem. Syst. Ecol.*, 51:142–152.

Sandasi, M., Kamatou, G. P. P., and Viljoen, A. M. (2012). An untargeted metabolomic approach in the chemotaxonomic assessment of two *Salvia* species as a potential source of α-bisabolol. *Phytochemistry*, 84:94–101.

Santos Pimenta, L. P., Kim, H. K., Verpoorte, R., and Choi, Y. H. (2013). NMR-based metabolomics: A probe to utilize biodiversity. In Roesssner, U. and Dias, D. A., editors, *Metabolomics tools for natural product discovery- Methods and protocols*, chapter 9, pages 117–128. Humana Press, Totowa, NJ.

Sarker, S. D. and Nahar, L. (2012). An introduction to natural products isolation. In Walker, J. M., editor, *Natural products isolation: Methods in molecular biology*, volume 864 of *Methods in Molecular Biology*, pages 1–25. Humana Press, Totowa, NJ.

Sato, S., Arita, M., Soga, T., Nishioka, T., and Tomita, M. (2008). Time-resolved metabolomics reveals metabolic modulation in rice foliage. *BMC Syst. Biol.*, 2:51.

Saxton, K. E. and Rawls, W. J. (2006). Soil water characteristic estimates by texture and organic matter for hydrologic solutions. *Soil Sci. Soc. Am. J.*, 70(5):1569–1578.

Scott, I. M., Vermeer, C. P., Liakata, M., Corol, D. I., Ward, J. L., Lin, W., Johnson, H. E., Whitehead, L., Kular, B., Baker, J. M., Walsh, S., Dave, A., Larson, T. R., Graham, I. a., Wang, T. L., King, R. D., Draper, J., and Beale, M. H. (2010). Enhancement of plant metabolite fingerprinting by machine learning. *Plant Physiol.*, 153(4):1506–1520.

Semenya, S., Potgieter, M., and Erasmus, L. (2012). Ethnobotanical survey of medicinal plants used by Bapedi healers to treat diabetes mellitus in the Limpopo Province, South Africa. *J. Ethnopharmacol.*, 141(1):440–445.

Shaik, R. and Ramakrishna, W. (2014). Machine learning approaches distinguish multiple stress conditions using stress-responsive genes and identify candidate genes for broad resistance in rice. *Plant Physiol.*, 164(1):481–95.

Sheela, D. and Uthayakumari, F. (2013). GC-MS analysis of bioactive constituents from coastal sand dune taxon- *Sesuvium portulacastrum* (L.). *Biosci. Discov.*, 4(1):47–53.

Silvera, K., Neubig, K. M., Whitten, W. M., Williams, N. H., Winter, K., and Cushman, J. C. (2010). Evolution along the crassulacean acid metabolism continuum. *Funct. Plant Biol.*, 37(11):995–1010.

Simons, L. and Allsopp, N. (2007). Rehabilitation of rangelands in Paulshoek, Namaqualand: Understanding vegetation change using biophysical manipulations. *J. Arid Environ.*, 70(4):755–766.

Smith, B. N. (1972). Natural abundance of the stable iostopes of carbon in biological systems. *Bioscience*, 22(4):226–231.

Soltis, D. E., Gitzendanner, M. A., Stull, G., Chester, M., Chanderbali, A., Chamala, S., Jordon-Thaden, I., Soltis, P. S., Schnable, P. S., and Barbazuk, W. B. (2013). The potential of genomics in plant systematics. *Taxon*, 63(5):886–898.

Steyn, P. S., van Heerden, F. R., Vleggaar, R., and Anderson, L. A. P. (1986). Bufadienolide glycosides of the Crassulaceae. Structure and stereochemistry of orbicusides A & C, novel toxic metabolites of *Cotyledon orbiculata*. *J. Chem. Soc.*, 2(3):1633–1636.

Sumner, L. W. (2006). Current status and forward looking thoughts on LC/MS metabolomics. In K.Saito, Dixon, R., and L.Willmitzer, editors, *Biotechnology in agriculture and forestry*, volume 57, chapter 12, pages 21–32. Springer, Verlag Berlin Heidelberg.

Sumner, L. W., Samuel, T., Noble, R., Gmbh, S.-a. D., Barrett, D., Beale, M. H., and Hardy, N. (2007). Proposed minimum reporting standards for chemical analysis, Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*, 3(3):211–221.

Sun, B., Ricardo, J. M., and Spranger, I. (1998). Separation of grape and wine proanthocyanidins according to their degree of polymerization. *J. Food Agric. Chem.*, 46(97):1390–1396.

Tautenhahn, R., Böttcher, C., and Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, 9:504.

Theis, N. and Lerdau, M. (2003). The evolution of function in plant secondary metabolites. *Int. J. Plant Sci.*, 164(3):93–102.

Todd, S. W. and Hoffman, M. T. (1999). A fence-line contrast reveals effects of heavy grazing on plant diversity and community composition in Namaqualand, South Africa. *Plant Ecol.*, 142:169–178.

Tolstikov, V. V., Fiehn, O., and Tanaka, N. (2007). Application of liquid chromatography-mass spectrometry analysis in metabolomics. In Weckwerth, W., editor, *Metabolomics: Methods and protocols*, chapter 9, pages 141–155. Humana Press, Totowa, NJ.

Trudell, S. a., Rygiewicz, P. T., and Edmonds, R. L. (2004). Patterns of nitrogen and carbon stable isotope ratios in macrofungi, plants and soils in two old-growth conifer forests. *New Phytol.*, 164(2):317–335.

van den Berg, R. a., Hoefsloot, H. C. J., Westerhuis, J. a., Smilde, A. K., and van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, 7(1):142.

van den Ouweland, J. M. W. and Kema, I. P. (2012). The role of liquid chromatography-tandem mass spectrometry in the clinical laboratory. *J. Chromatogr. B*, 883-884:18–32.

van der Watt, E. and Pretorius, J. C. (2001). Purification and identification of active antibacterial components in *Carpobrotus edulis* L. *J. Ehnopharmacology*, 76(1):87–91.

Vogeser, M. and Seger, C. (2010). Pitfalls associated with the use of liquid chromatography-tandem mass spectrometry in the clinical laboratory. *Clin. Chem.*, 56(8):1234–1244.

Vogt, T., Ibdah, M., Schmidt, J., Wray, V., Nimtz, M., and Strack, D. (1999). Light-induced betacyanin and flavonol accumulation in bladder cells of *Mesembryanthemum crystallinum*. *Phytochemistry*, 52(4):583–592.

Vries, F. A., Bitar, H. E., Green, I. R., Klaasen, J. A., Mabusela, W. T., Bodo, B., and Johnson, Q. (2005). An antifungal active extract from the aerial parts of *Galenia africana*. In *NAPRECA symposium- Book of proceedings*, pages 123–131, Antananarivo, Madagascar.

Wang (1986). *Handbook of effective components in vegetable medicines*. Peaple Health Press, Bejing, China, 1st edition.

Want, E. and Masson, P. (2011). Processing and analysis of GC/LC-MS-based metabolomics data. In Metz, T. O., editor, *Metabolic profiling*, volume 708 of *Methods in Molecular Biology*, pages 277–298. Humana Press, Totowa, NJ.

Wehrens, R. (2011). *Chemometrics with R- Multivariate data analysis in the natural sciences and life sciences*. Springer New York, 1st edition.

Wheat, N. (2014). *An ethnobotanical, phytochemical and metabolomics investigation of plants from the Paulshoek Communal Area, Namaqualand*. Phd, University of Cape Town, Cape Town.

Winter, K., Lüttge, U., Winter, E., and Troughton, J. H. (1978). Seasonal shift from C3 photosynthesis to Crassulacean acid metabolism in *Mesembryanthemum crystallinum* growing in its natural environment. *Int. Assoc. Ecol.*, 34(2):225–237.

Winter, K., Troughton, J. H., Evenari, M., Läuchli, A., Lüttge, U., and Url, S. (1976). Mineral ion composition and occurrence of CAM-like diurnal malate fluctuations in plants of coastal and desert habitats of Israel and the Sinai. *Oecologia*, 25(2):125–143.

Xia, J., Psychogios, N., Young, N., and Wishart, D. S. (2009). MetaboAnalyst: A web server for metabolomic data analysis and interpretation. *Nucleic Acids Res.*, 37:652–660.

Xue, S.-Y., Li, Z.-Y., Zhi, H.-J., Sun, H.-F., Zhang, L.-Z., Guo, X.-Q., and Qin, X.-M. (2012). Metabolic fingerprinting investigation of *Tussilago farfara* L. by GCMS and multivariate data analysis. *Biochem. Syst. Ecol.*, 41:6–12.

Yi, L., Shi, S., Yi, Z., He, R., Lu, H., and Liang, Y. (2014). MeOx-TMS derivatization for GC-MS metabolic profiling of urine and application in the discrimination between normal C57BL/6J and type 2 diabetic KK-Ay mice. *Anal. Methods*, 6(12):43804387.

Yip, C., Mahoney, M. W., Szalay, A. S., Csabai, I., Budavári, T., Wyse, R. F. G., and Dobos, L. (2014). Objective identification of informative wavelength regions in galaxy spectra. *Astron. J.*, 147(5):110–125.

Yun, B., Yan, Z., Amir, R., Hong, S., Jin, Y., Lee, E., and Loake, G. J. (2012). Plant natural products: History, limitations and the potential of cambial meristematic cells. *Biotechnol. Genet. Eng. Rev.*, 28(1):47–60.

Zafar, M., Khan, M. A., Ahmad, M., Sultana, S., Qureshi, R., and Tareen, R. B. (2010). Authentication of misidentified crude herbal drugs marketed in Pakistan. *J. Med. Plants Res.*, 4(15):1584–1593.

Zhang, Q., Manzoni, S., Katul, G., Porporato, A., and Yang, D. (2014). The hysteretic evapotranspiration- Vapor pressure deficit relation. *J. Geophys. Res. Biogeosciences*, 119:125–140.

Zhou, B., Xiao, J. F., Tuli, L., and Ressom, H. W. (2012). LC-MS-based metabolomics. *Mol. Biosyst.*, 8(2):470–481.