NAIST-IS-DD1561027

Doctoral Dissertation

Data-Intensive Science of Plant Classification based on Metabolite-Content Similarity

Liu Kang

March 12, 2018

Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to the Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Engineering.

Thesis Committee:

Professor Shigehiko Kanaya	(Supervisor)
Professor Shoji Kasahara	(Co-supervisor)
Associate Professor Md. Altaf-Ul-Amin	(Co-supervisor)
Associate Professor Naoaki Ono	(Co-supervisor)

Data-Intensive Science of Plant Classification based on Metabolite-Content Similarity [1]

Liu Kang

Abstract

Metabolite-content (MC) refers to all small molecules which are the products or intermediates of metabolism within an organism. The secondary-metabolite-content of a plants is highly related to its pathways which are constrained to evolutionary phylogeny, and are also related to the bioactive compounds of the plant which determine the medicinal and nutritional features of the plant. In this study, we consider the metabolite features of plants as a new taxonomic marker and classify plants based on the MC-similarity of them using KNApSAcK Core DB. For reducing the effect of missing plant-metabolite relation data, we propose two approaches to compensate for the limitations of missing data: (1) Classification of Plants based on Chemical Structure Similarity of Metabolite-Content. By this approach, we calculated the structural similarity of all metabolite pairs by Tanimoto coefficients (TCs), and determined the MC-similarity of plants based on the background population of TCs. (2) Clustering Plants based on Structural-Similarity Network of Metabolites. By this approach, we applied a network based approach to abstract structurally similar metabolite groups as features, and measured the phylogenetic distance by a binary method. Then we classified the plants by hierarchical clustering method and compared the resulted classification of plants with NCBI taxonomy. The results prove that the MC-similarity of plants is associated with the pathway and bioactive similarity, and can be regarded as a taxonomy marker which takes into account both general phylogenetic relations and the relations between plants based on bioactive features. We also extended our finding by using phylogenetic statistic method to investigate the predictive power of MC-similarity in exploration of edible and medicinal plants for bioprospecting. We reconstructed the phylogenetic trees for the same set of plants based on MC-similarity and sequence-similarity. We then applied D statistic to test phylogenetic signal of medicinal/edible plants for the obtained phylogenetic trees and identified the hot nodes that were significantly overrepresented by plants of medicinal/edible uses. The result shows that comparing with sequence-based approach, plants with medicinal/edible uses are more significantly clustered in MC-based phylogenetic trees. The hot nodes in MC-based phylogenetic trees tend to encompass more medicinal/edible plants, and could highlighted different groups of medicinal plants.

Keywords:

Metabolite-content, secondary metabolite, chemosystematics, phylogeny, prediction

[1] Doctoral Dissertation, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1561027, March 12, 2018

Acknowledgements

First and foremost, I owe special thanks to my mentor Shigehiko Kanaya, for his continuous support for both my research work and life in Japan. Before I came to Japan, I only had software development experience and knew few about bioinformatics and systems biology. He accepted me as his student and gave me the opportunity to study in the CSB Lab. He led me into the world of bioinformatics and biological data science, and gave me many patient guidance and precious suggestions.

I would like to thank the thesis committee, Professor Shigehiko Kanaya, Professor Shoji Kasahara, Associate Professor Md. Altaf-Ul-Amin and Associate Professor Naoaki Ono. Thank for their valuable comments and supervision of my work. Their valuable advice, guidance, critics made the dissertation improved greatly.

I want to give my great appreciation to Associate Professor Md. Altaf-Ul-Amin who directly supervised my research work. He taught me how to compose a scientific paper and how to become a bioinformatics researcher. He gave me many suggestive idea and made significant contributions towards shaping the general direction of my research work in the period of my doctoral course.

I would like to thank all the other staff in CSB Lab: Assistant Professor Tetsuo Sato, Assistant Professor Alex Ming Huang, Aki H. Morita, Minako Ohashi. It is very kind of you to assist me in daily work and life. I also want to thank Professor Takaaki Nishioka for help me to modify my paper. I wish to thank all my fellow lab member in CSB Lab especially Dr. Donghan Li, Dr. Yu Bai, Dr. Sony Hortono Wijaya, Dr. Kibinge Nelson Kipchirchir and Dr. Azian Azamimi Abdullah for their support in my studies.

I also would like to express my appreciation to all staff in NAIST International Student Affairs Section for their support throughout my studies. I also wish to thank MEXT and CSC for their financial support during my doctoral course in Japan.

Last but not lease, my appreciation also extends to my parents and fiancee for their unconditional support. With their companion and encouragement, I could achieve my doctoral degree. I dedicate this thesis to them.

March 12, 2018
Liu Kang

List of Abbreviations

MC	Metabolite-Content
PCR	Polymerase Chain Reaction
NGS	Next Generation Sequencing
DNA	Deoxyribonucleic Acid
m-RNA	Messenger RNA (Ribonucleic Acid)
KEGG	Kyoto Encyclopedia of Genes and Genomes
NCBI	National Center for Biotechnology Information
WHO	World Health Organization
DB	Database
TCs	Tanimoto Coefficients
SVM	Support Vector Machine
MOL	Molecular data file created in the MDL Molfile format
SDF	Structure Definition File
rbcL	Ribulose-1,5-bisphosphate carboxylase/oxygenase
matK	Maturase K
ITS2	Internal Transcribed Spacer 2

Contents

Chapter 1 Introduction 1

1.1. Background 1

1.2. Research Problem and Objectives... 10

1.3. Dissertation Outline 11

Chapter 2 Classification of Plants based on Chemical Structure Similarity of Metabolite-Content 13

2.1. Background 13

2.2. Datasets and Preliminaries 15

2.3. Methods 17

2.3.1. Background Population of Tanimoto Coefficients 17

2.3.2. Plant-Plant Similarity Score 18

2.4. Results and Discussion 20

2.4.1. Dataset and Selection of Thresholds 20

2.4.2. Classification of Plants 23

2.5. Summary 31

Chapter 3 Clustering Plants based on Structural-Similarity Network of Metabolites 33

3.1. Background 33

3.2. Datasets and Preliminaries 34

3.3. Methods 35

3.3.1. Network Construction of Metabolites Based on Chemical Structure Similarity. 35

3.3.2. Clustering of Metabolites Based on DPClus. 36

3.3.3. Clustering of Plants Based on Metabolite Groups. 37

3.3.4. Classification of Plants by SVMs 38

3.4. Results and Discussion 39

3.4.1. Plant Representation Based on Metabolite-Content Similarity. 39

3.4.2. Clustering of Plants Based on Metabolite-Content Similarity. 43

3.4.3. Predicting Currently Unknown Plant-Metabolite Relations. 57

3.4.4. Relationship between Metabolite-Content and Uses of Plants. 61

3.5. Summary 63

Chapter 4 Metabolite-Content-Guided Prediction of Medicinal/Edible Properties in Plants for Bioprospecting 65

4.1. Background 65
4.2. Datasets and Preliminaries 67
4.3. Methods 68
4.3.1. Phylogenetic Reconstruction...68
4.3.2. Phylogenetic and Statistical Analyses 69
4.3.3. Evolutionary Patterns of Medicinal/Edible Properties 69
4.4 Results and Discussion 70
4.4.1. Phylogenetic Signal of Medicinal and Edible Plants............................73
4.4.2. Hot Nodes of Medicinal and Edible Plants 75
4.4.3. Predicting of Medicinal Properties by Metabolite-Content-Similarity 77
4.5. Summary 78

Chapter 5 Conclusion...80

Bibliography **82**

Appendix A The predicted unrecorded plant-metabolite relations **93**

Appendix B GenBank ID and use information of sample plants **112**

Appendix C The 90 plants with high priority for future screening for overall medicinal bioactivity **120**

List of Figures

1.1. Six secondary metabolites and source plants. (Facchini et al., 2012)............................... 3

1.2. Distribution of iridoids in *Lamiaceae* (Wink, 2003)... 4

1.3. Subjects arising from the systematic study of plant chemical constituents. (Reynolds, 2007) 9

2.1. The degree distribution of plants in the plant-metabolite bipartite graph. 16

2.2. The distributions of TCs in our dataset.. 19

2.3. Baker's Gamma correlation coefficient corresponding to different threshold levels. 22

2.4. Hierarchical dendrogram plot of 102 plants based on their MC-similarity, using 1% TCs as threshold... 23

3.1. Plant-metabolite relations and plant versus metabolite-group relations....................... 35

3.2. The degree distribution of species in the species-metabolite bipartite graph. 39

3.3. Structural-similarity-based network of metabolites.. 41

3.4. Frequency of metabolite groups with respect to group size. 43

3.5. Comparison of Simpson- and Jaccard-based dendrograms. 45

3.6. The family or genus level phylogeny patterns in proposed classification of 216 plants. 47

4.1. Overview of 190 plants included in *rbcL*, *matK* and ITS2 sample groups. 68

4.2. Phylogenetic trees and the hot nodes of medicinal/edible features for sequence- and MC-based approaches... 72

4.3. MC-based phylogenetic tree for 1047 plants, with the hot nodes of medicinal/edible plants.. 78

List of Tables

2.1. Classification of 102 plants. 27

3.1. Classification of 216 plants. 48

3.2. Reported plant-metabolite relations of 6 plants of genus *Citrus* with a given metabolite group 60

3.3. Predicted unrecorded metabolites for 6 *Citrus* plants. 60

3.4. Resulting confusion matrix from support vector machine (SVM) algorithm. 62

4.1. The amount of complete and partial sequences data of *rbcL*, *matK* and ITS2 sample groups. 73

4.2. The amount of plants in each category of uses 73

4.3. Phylogenetic signal of medicinal/edible features in sequence-based and metabolite-content-based trees for the *rbcL*, *matK* and ITS2 sample groups. 74

4.4. The number and proportion of medicinal/edible plants within the clades of hot nodes. 76

Chapter 1

Introduction

This dissertation summarizes the author's research experience in integrating the metabolite-content (MC) data of plants and exploring the systematic and phylogenetic value of metabolite-content data, including classifying plants by MC-similarity and analyzing the relationships among plants, secondary metabolite-content and biological activities. This study also investigates the predictive power of MC-similarity in exploring nutritional and medicinal properties in plants, by comparing with molecular phylogenetic based approach. This chapter describes the general background, the research problem and objectives, and also explains what are to be expected from the rest of the dissertation.

1.1. Background

(a) Metabolite-Content of Plants

One of the outstanding features of plants in addition to their form is their chemistry. The chemical observations of plants, from color, scent and taste, to their nutritional value or their poisonous nature, lead the inquisitive person to a study of the chemical components and this in turn leads to uses of these chemicals for practical economic and health purposes (Reynolds, 2007). Metabolomics is the scientific study of quantification of low mass compounds profiles and analysis of chemical processes involving metabolites in a comprehensive fashion. In general, metabolites can be divided into two groups: primary and secondary metabolites. Primary metabolites are directly involved in the normal growth, development and reproduction. On the other hand, secondary metabolites are not directly involved in these processes, but usually have important ecological functions, such as inter- or intra-species communication, antifungal, antimicrobial activities and also as a defense against pests and pathogens (Agostini-costa et al. 2012).

The metabolome has been defined as the qualitative and the quantitative collection of all low-weight molecules(metabolites) present in the cell that are participants in general metabolic reactions and that are required for the maintenance, growth, and normal function of a cell. The number of the different molecules in the metabolome varies depending on the organism being studied. However, the metabolome is not stable and inconstantly changing due to all the chemical reactions occurring in the cell (Tauler & Walczak, 2009). For investigating the metabolite features of an organism following an evolutionary perspective, we propose the concept of metabolite-content in this thesis. Metabolite-content refers to all small molecules which are the products or intermediates of metabolism (metabolites) that are present within a biological organism. It differs from metabolome in that the metabolite-content mainly focuses on the qualitative collection of small metabolites and ignores the quantitative differences, which is instable with different parts and stage of one organism.

For plants, the metabolite-content are mainly represented as secondary metabolites, which are often similar within members of a clade (Hegnauer, 1967; Pichersky & Gang, 2000; Wink, 2003). Vascular plants contain an enormous variety of secondary metabolites, which vary according to family and species. The restricted distribution of secondary metabolites makes a major contribution to the specific odours, tastes and colors of plants (Bennett & Wallsgrove, 1994). Secondary metabolites are present in all higher plants, usually in a high structural diversity (Figure 1.1). The pattern of secondary metabolites in a given plant is complex, it changes in a tissue- and organ specific way; regularly, differences can be seen between different developmental stages between individuals and populations. Decades ago botanists preferred the simpler interpretation that secondary metabolites were waste products of primary metabolism and that structural diversity would only reflect a play of nature. But now, adaptive explanations are more favoured again to explain the existence and diversity of secondary metabolites (Wink, 2003).

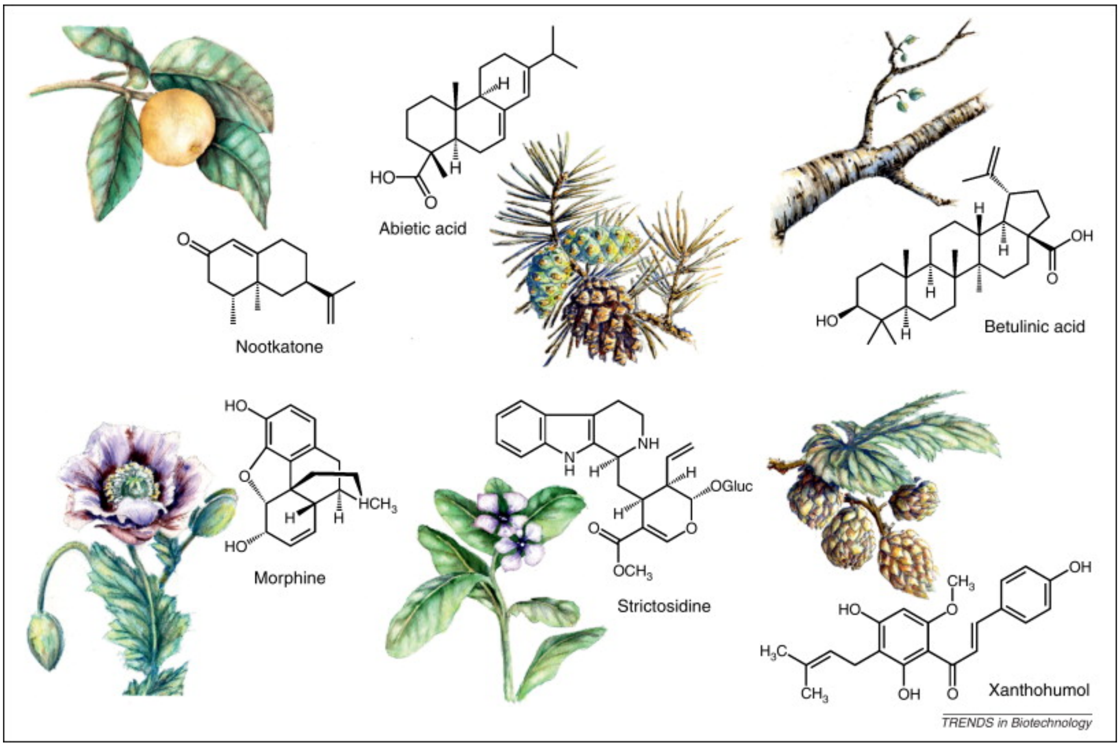


Figure 1.1. *Six secondary metabolites and source plants* (Facchini et al., 2012).

Secondary metabolites, at least the major ones present in a plant, apparently function as defence (against herbivores, microbes, viruses or competing plants) and signal compounds (to attract pollinating or seed dispersing animals). They are thus important for the plant's survival and reproductive fitness. Absence of secondary metabolites dose not result in immediate death, but rather in long-term impairment of the organism's survivability, fecundity, or aesthetics, or perhaps in no significant change at all. Secondary metabolites often play an important role in plant defense against herbivory and other interspecies defenses. Secondary metabolites therefore represent adaptive characters that have been subjected to natural selection during evolution, and are often restricted to a narrow set of plants within a phylogenetic group (Figure 1.2) (Wink, 2003).

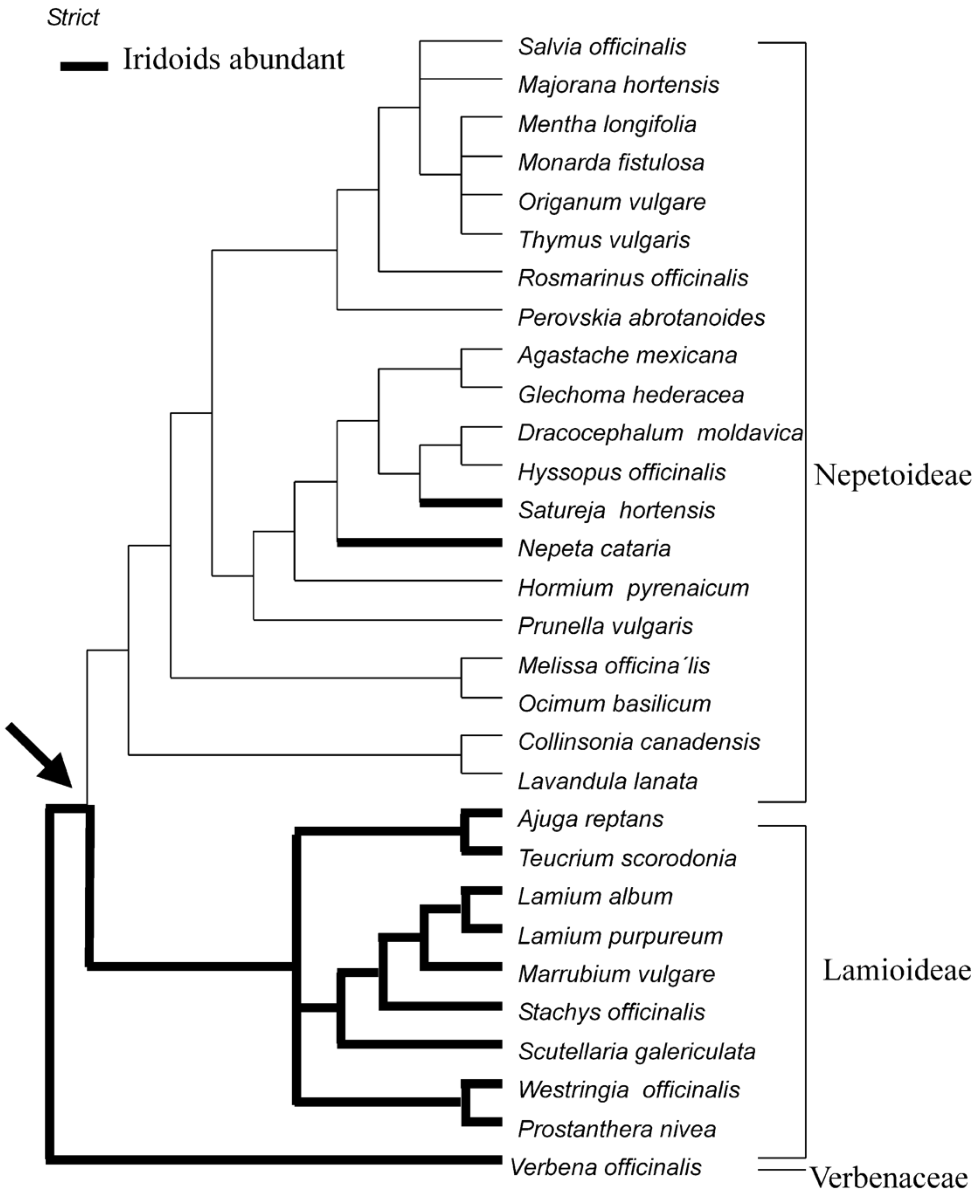


Figure 1.2. *Distribution of iridoids in Lamiaceae. The phylogenetic tree was constructed from complete rbcL sequences. The iridoid glycosides are common in members of the subfamily Lamioideae* (Wink, 2003).

(b) Plant Phylogeny and Chemosystematics

Phylogenetic is the study of the evolutionary history and relationships among individuals of a group of related species. The result of these analyses is called a phylogeny, which is often represented by a tree diagram called phylogenetic tree (Davis & Jerrold, 2014). Plant phylogeny is a phylogeny of plants. Plant phylogeny is a useful and essential tool for plant taxonomy, the science that explores, describes, names, and classifies plants. Morphology, anatomy, and genetics are the main sources of characters used in today's plants taxonomy (Rouhan & Gaudeul, 2014).

The systematic and phylogenetic analysis of plants is traditionally based on macroscopic and microscopic morphological characteristics and is known to be turbulent (Besse, 2014). With the advent of phylogeny, taxonomy has been evolved into systematics which allowed classification based on the evolutional relationship among organisms. The study of DNA and to a certain extent m-RNA and proteins has led to the immense subject of molecular biology, which has been increasingly applied to reconstruct the phylogeny of higher and lower plants (Soltis et al., 1992; Reynolds, 2007). Molecular phylogenetics is the study of phylogeny that analyses hereditary molecular differences, mainly in DNA sequences, to gain information on an organism's evolutionary relationships. This powerful approach, which provides the best phylogenetic resolution so far, is facilitated by rapid DNA amplification techniques such as polymerase chain reaction (PCR), by rapid DNA sequencing methods such as Next Generation Sequencing (NGS) (Metzker, 2010), and by powerful computation with adequate software programs (such as PAUP, MEGA, MrBayes, RAXML) (Swofford, 2003; Kumar et al., 1993; Ronquist et al., 2012; Stamatakis, 2014).

The use of molecular data in plant phylogeny has been highly successful in many instances, but also has faced some limitations and cautions to consider. First, this approach does not describe the chemical make-up of the plants, although it may comment on the inter-relationship of the chemicals and their biosynthetic systems. The molecular biology, which analysis DNA, m-RNA, and proteins and the interaction between them, does not describe the small molecules (secondary metabolites) in plants, or how they relate to each other, to the plants containing them or to the environment. Second, current technologies that use genomic compartments instead of the entire genome data usually only partially reveal the evolutional relations among plants (Hinchliff & Smith, 2014). The number of organisms with completely known genomes in Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa et al., 2017) has now reached 5093 but includes only 80 plants (October 2017). This indicates that it is still impractical to reconstruct plant phylogeny using the entire genome information. Third, recent research has indicated that horizontal gene transfer occurs in multicellular eukaryotes, especially in plants, and has an important role in their eukaryotic evolution. This suggests that phylogenetic reconstruction cannot be determined conclusively from sequence data (Keeling & Palmer, 2008; Gao et al., 2014).

As a rule, a single group of secondary metabolite dominates within a given taxon. Therefore, the metabolite characters could be regarded as a taxonomic marker and become a matter of interpretation the evolutionary relation among species in the same way as traditional morphological markers. Classifying plants based on the basis of their chemical constituents could be helpful in discovering new edible and medicinal plants and solving selected taxonomical problems (Wink, 2003; Reynolds, 2007; Singh, 2016). The systematics of plants based on their chemical constituents is known as plant chemosystematics. Different from molecular phylogeny which focuses on macromolecules such as DNA and proteins, plant chemosystematics is the molecular systematics of plants using secondary metabolites or micromoleculars.

Plant chemosystematics has initially been used to distinguish plants and other organisms that are useful for food and those best avoided. This knowledge has been progressively formalized with useful, harmful and inactive chemical constituents from relevant taxa now identified and recorded. Plants chemosystematics could reveal the general natural history of the plant with reference to its relationship to similar plants and its interaction with its environment (Reynolds, 2007; Heywood, 2013; Singh, 2016).

Traditional chemosystematics of plants is based on the presence or absence of selected secondary metabolites (Singh, 2016; Wink, 2003). This approach is based on the hypothesis that the selected secondary metabolites dominate within a given taxon. However, an inevitable conclusion drawn from the observations is that the expression of secondary metabolites of a given structural type has invariably arisen in a number of occasions in different parts of the plant kingdom. In addition, the same compounds are frequently produced by quite different biosynthetic pathways in unrelated plants. For example, in the family *Lamiaceae*, the iridoid glycosides are commonly found in members of the subfamily *Lamioideae* but occasions found in other subfamilies of *Lamiaceae* (Figure 1.2). This discrepancy could be due, either to convergent evolution or differential gene expression strategy. It is likely that in some cases the gene that encode the enzymes for the production of a given structure or structural skeleton have evolved early during evolution, and might be "switched off" for some plants and "switched on" again at some later point (Wink, 2003; Pichersky & Gang, 2000;). The inevitable irregular distribution of selected secondary metabolites reveals the limitation of traditional chemosystematics approach, and the necessity of applying a holistic approach involving metabolite-content data when conduct systematics classification of plants.

A great variety of secondary metabolites constitute the main body of metabolite-content of plants. These secondary metabolite-content data, which describes the small molecules in plants, are highly associated with the expression of bioactive compounds thus related with the nutrition and medicinal value of plant which is usually neglected by the molecular phylogenetic analysis. The incorporation of phylogenetic into chemosystematics studies suggests that the compilation of all these data on chemical composition (metabolite-content) of plants is not only an auxiliary method to molecular biology based phylogenetic analysis, but also provide a unique perspective in plant phylogeny and taxonomy. First, the inconsistent secondary metabolite profile means that the systematic value of metabolite-content characters becomes a matter of interpretation in the same way as traditional morphological markers. Second, the metabolite-content of a plants also reveal the special chemical features which is difficult to analyzed by molecular phylogenetic.

(c) Natural Products and Bioprospecting

Natural products are chemical compounds or substances produced naturally by living organisms. In the broadest sense, natural products include any substance produced by life. Bioprospecting is the process of discovery and commercialization of new products based on biological resources (Strobel & Daisy, 2003). A growing need for new bioactive compounds in the pharmaceutical and the food industries stresses the importance of prospecting for novel bio-resource (Berg et al., 2013; Woolhouse & Farrar, 2014). Since the chemical diversity of compounds as comprised in biological resources is higher than synthetic chemistry achieves, bio-resources have great potential to hold a manifold of promising compounds for biotechnological application (Bérdy, 2012; Nováková & Farkašovský, 2013).

Plants are the major contributors of natural products and are usually rich in nutritional or medicinal properties, which is attributed to the complex secondary metabolite constituents of them. (Dahanukar et al., 2000; Veeresham, 2012; Cseke et al., 2016). Secondary metabolites often contain more than one functional group; they therefore often exhibit multiple functionalities and bioactivities. Many natural products related subjects arise from the systematic study of plant chemical constituents (Figure 1.3). Generally, the natural products of plants include food products (edible plants), medicine products (medicinal plants), and other products such as structure and decoration. The edible plants, including fruits, vegetable, nuts and cereals etc., occupy an important position in people's daily diet and is directly related to human nutrition. The medicinal uses of plants include traditional herbal medicinal and pharmaceutical drugs derived from plants. In many cases there is no clear demarcation of edible plants and medicinal plants because many edible plants possess medicinal value.

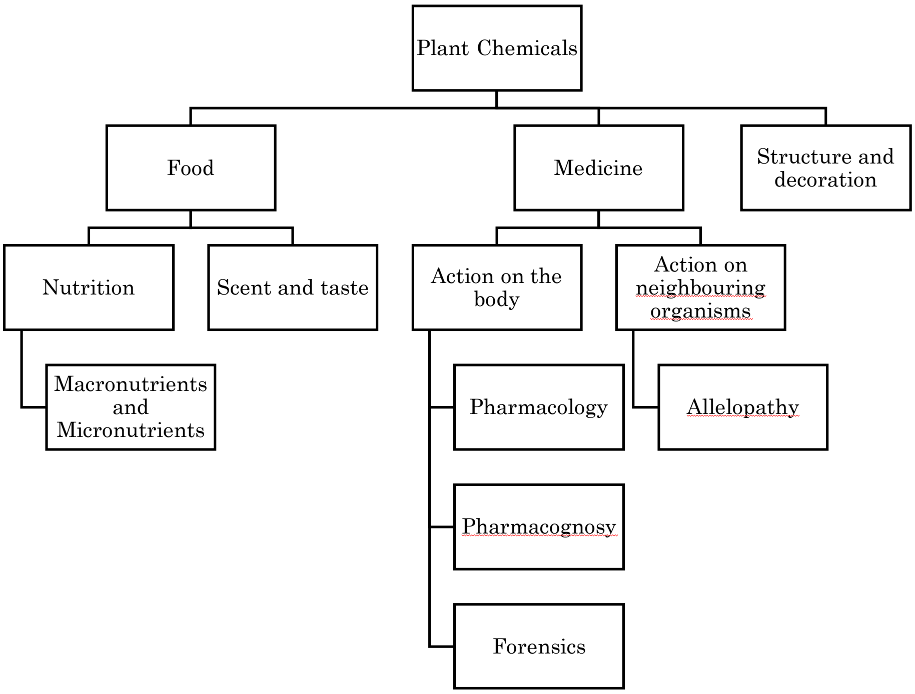


Figure 1.3. *Subjects arising from the systematic study of plant chemical constituents* (Reynolds, 2007).

Plants are an important source of novel pharmacologically active compounds with many pharmaceutical drugs have been derived directly or indirectly from plants, and have played a central role in human health-care since ancient times. Many natural products from plants are biologically active and have been used for thousands of years as traditional medicines (Newman et al., 2000; Cragg & Newman, 2013; Fabricant & Farnsworth, 2001). Even at the dawn of 21st century the importance of medicinal plants is enormous – about 25% of the drugs prescribed worldwide come from plants, and 11% of the 252 drugs considered as basic and essential by World Health Organization (WHO) were exclusively of flowering plant origin (Veeresham, 2012). With the development of modern technology, more and more plant extracts have been found to be useful to medical practice (Ouyang et al., 2014). However, the potential of plants to yield new valuable drugs is under threat due to the alarming bio-diversity loss, with recent estimates indicating that one in five of plants on earth is threatened with extinction (Brummitt & Bachman, 2010). Therefore, there is an urgent need for a time-efficient and systematic approach for unlocking the potential of plants in drug discovery.

1.2. Research Problem and Objectives

Since the phylogenetic reconstruction based on comparison of DNA sequence has many limitations, many researchers have begun to explore phylogenetic distance between species based on metabolite constituents, either alone or in combination with sequence features. Clemente et al. (2007) presented a method for assessing the structural similarity of metabolic pathways for several organisms and reconstructed phylogenies that were very similar to the National Center for Biotechnology Information (NCBI) taxonomy. Borenstein et al. (2008) predicted the phylogenetic tree by comparing the "seed set" of metabolic networks. Mano et al. (2010) considered the topology of pathways as chains and used a pathway-alignment method to classify species. Chang et al. (2011) proposed an approach from the perspective of enzyme substrates and corresponding products in which each organism is represented as a vector of substrate-product pairs. The vectors were then compared to reconstruct a phylogenetic tree. Ma et al. (2013) demonstrated the usefulness of the global alignment of multiple metabolic networks to infer the phylogenetic relationships between species. A. A. Abdullah et al. (2015) classified microorganism species based on the volatile metabolites emitted by them, and the results have been well explained in terms of their pathogenicity. However, most of these studies have focused on microorganisms, such as archaea, rather than multicellular eukaryotesm.

The systematics of plants based on their chemical constituents is often limited by the lack of plant-metabolite relation data. Traditional chemosystematics of plants is based on the presence or absence of selected secondary metabolites (Wink, 2003; Singh, 2016), which is far from the holistic approach involving metabolite-content. The inconsistency of molecular phylogenetic with chemosystematics studies suggests that the metabolite-content of plants may reveal more information of the interaction and bioactive similarity of plants. Such MC-based classification not only reveals the phylogenetic relationship of plants but also can be used for studying the relationship of plants in terms of their bioactive properties, guiding prediction of medicinal properties in bioprospecting, exploring new nutritional or economic uses of plants, and solving taxonomical problems.

With the rapid development of metabolomics, metabolite-related databases (DBs) have been created, including KNApSAcK, which contains accumulated information about species-metabolite relations including information about many secondary metabolites of plants (Shinbo et al., 2006). Such information can be used in the systems-biological studies on the interactions between plants, including the activities of medicinal plants as well as interactions between plants and their environments (Afendi et al., 2012). With the development of plants metabolomics and big data biology (Marx, 2013; Altaf-Ul-Amin et al., 2014), it is now possible to investigate the systematics value of metabolite-content of plants on a cross-class level.

In this study, we attempt to classify a set of cross-family plants (plants spread over different families) based on their MC-similarity, using KNApSAcK Core DB as sources. The main objective of this study is to investigate the systematics value of metabolite-content of plants. We consider metabolite-content of a plant as a taxonomy marker like genome data, as metabolite-content is highly related to its pathways which are regulated by the related enzyme genes, and MC-similarity could explore both the evolutional and bioactivity relations between plants. The second objective of this study is to explore structurally similar metabolite groups by MC-similarity. These metabolite groups are related to specific metabolic pathways, and can be used to predict currently unknown plant-metabolite relations. The third objective is to facilitate the prediction of medicinal/edible properties in plant bioprospecting using metabolite-content data.

1.3. Dissertation Outline

This dissertation outline is organized as follows. In Chapter 2, we describe an approach of plant classification based on their MC-similarity. We calculate the structural similarity of all metabolite pairs by Tanimoto coefficients (TCs), and determine the MC-similarity of plants based on the background population of TCs. We classify 102 plants into 28 groups by hierarchical clustering method, and prove that metabolite-content of plants is consistent with both phylogenetic and bioactive characters of plants. In Chapter 3, we describe an approach of classifying plants based on structural-similarity network of metabolites. By this approach, we apply DPClus algorithm to abstract structurally similar metabolite groups as features, and measured the phylogenetic distance of plants by these metabolite groups. We classify 216 plants into 48 groups by hierarchical clustering method. We also use these structurally similar metabolite groups to predict a series of unknown plant-metabolite relations. Additionally, we extend our analysis by implementing support vector machine (SVM) algorithm to study the relationship between metabolite-content and uses of plants. In Chapter 4, We investigate the predictive power of MC-similarity in exploration of edible and medicinal plants for bioprospecting by using phylogenetic statistic method. We reconstructed the phylogenetic trees for the same set of plants based on MC-similarity based approach and sequence-similarity based approach. We then applied D statistic to test phylogenetic signal of medicinal/edible plants for the obtained phylogenetic trees and identified the hot nodes that were significantly overrepresented by plants of medicinal/edible uses. Finally, in Chapter 5 we give conclusive remarks of this dissertation.

Chapter 2

Classification of Plants based on Chemical Structure Similarity of Metabolite-Content

2.1. Background

The systematics of plants by their chemical constituents is known as “Phytotaxonomy” or “Chemosystematics”. This subject has been used to distinguish plants by their bioactive characters. Traditional chemosystematics of plants is based on the presence or absence of selected secondary metabolites (Singh, 2016; Wink, 2003). However, an inevitable conclusion drawn from the observations is that the expression of secondary metabolites of a given structural type has almost invariably arisen in a number of occasions in different parts of the plant kingdom. It indicates that it is necessary to use a holistic approach considering all of the metabolites when classifying plants by their metabolite constituents. The incomplete data of metabolite constituents of plants limits the ability of chemosystematics for solving taxonomical problems.

Biology has recently become a “big-data science” mainly supported by the advances in high-throughput experimental technologies, and has significant roles to play in versatile disciplines of scientific research. (Altaf-Ul-Amin et al., 2014). Big data biology is a data-intensive science, which has emerged because of the rapidly increasing volume of molecular biological data in omics fields such as genomics, transcriptomics, proteomics and metabolomics (Kelling et al. 2009; Patterson et al. 2010). The metabolomics of plants is developing rapidly and will become an important topic in the systems-biological studies of interaction between plants and human (Bino et al., 2004; Macel et al., 2010; Saito & Matsuda 2010; Tohge & Fernie 2010).

With the rapid development of metabolomics and the explosively growing data scale, the development of metabolite-related databases (DBs) incorporating different species has become a very important theme in big data biology. To address this need, the KNApSAcK

Family database has been developed (Afendi et al., 2012). The KNApSAcK Family DB mainly contains two types of binary relationships: the Metabolomics DB system and the Multifaceted Plant Usage DB. The KNApSAcK Family database systems have been utilized in a number of studies in metabolomics. For example, previously the KNApSAcK Family DB systems have been used to understand the medicinal usage of plants based on traditional and modern knowledge (Afendi et al. 2012; Afendi et al. 2013; Wijaya et al. 2014). The KNApSAcK Core DB is a major member of the KNApSAcK Family database systems. This system contains accumulated information about species-metabolite relations including information about many secondary metabolites of plants. Such information can be used in the systems-biological studies on the interactions between plants, including the activities of medicinal plants as well as interactions between plants and their environments (Shinbo et al., 2006; Afendi et al., 2012). The KNApSAcK Core DB contains 111,199 species-metabolite relationships that encompass 25,658 species and 50,899 metabolites, and these numbers are still growing (http://kanaya.naist.jp/knapsack_jsp/).

To perform a holistic review on the metabolite features of a species, the concept of metabolite-content has been proposed. Metabolite-content refers to all small molecules which are the products or intermediates of metabolism (metabolites) that are present within a biological organism. Metabolite-content can be used to distinguish plants and other organisms. For example, previously microorganisms have been classified based on the volatile organic compounds emitted by them (Abdullah et al., 2015). The metabolite-content of plants is dominated by secondary metabolites, which are usually in a high structural diversity and often similar within members of a clade (Hegnauer, 1967; Pichersky et al., 2000; Wink, 2003). The KNApSAcK Core DB can be considered an advanced source of metabolite-content data of plants. As a rule, secondary metabolites are often similar within members of a clade, and plants within a taxon often represent similar metabolite-content and bioactive properties. Therefore, the metabolite-content of plants can be used as a taxonomy marker to distinguish plants. Moreover, the expression of secondary metabolites of a given structural type has frequently arisen on a number of occasions in different parts of the plant kingdom. The inevitable irregular distribution of selected secondary metabolites suggests that the metabolite-content of plants may reveal more information of the interaction and bioactive similarity of plants than morphology features. As a hypothesis, we consider metabolite-content of a plant as a taxonomy marker like morphology features, as the metabolite-content of a plant is highly related to its pathways which are regulated by the related enzyme genes, and the similarity of metabolite-content could explore both the evolutional and bioactivity relations between plants. Such MC-similarity based classification not only reveals the phylogenetic relationship of plants but also can be used for studying the relationship of plants in terms of their bioactive properties, guiding prediction of medicinal properties in bioprospecting, exploring new nutritional or economic uses of plants, and solving taxonomical problems. With the development of plant metabolomics and big data biology, it is now possible to investigate the metabolite-content of plants on a cross-class level (Marx, 2013; Altaf-Ul-Amin et al., 2014).

In this chapter, we describe a novel plant classification method based on chemical stricture similarity of metabolite-content. We collect metabolite-content data of plants and structure data of compounds from the KNApSAcK Core DB, and measure the structural similarity between two compounds utilizing the concept of Tanimoto coefficients (TCs) (Godden et al., 2000; Chen et al., 2002; Willett, 2014). Next, MC-similarity between plants is calculated based on the background population of TCs. We use the ratio of Tanimoto values above a rational threshold to assess the MC-similarity between plant pairs. Finally, we classify the plants based on the MC-similarity and compare our classification with NCBI taxonomy (http://www.ncbi.nlm.nih.gov/taxonomy). The result proves that metabolite-content of plants can be regarded as a taxonomy marker which takes into account both general phylogenetic relations and relation between plants based on bioactivity and economic uses.

2.2. Datasets and Preliminaries

The major input data in our study are species–metabolite relationships that have been obtained from the KNApSAcK Core DB (Afendi et al., 2012). The KNApSAcK core DB contains most of the published information about species–metabolite relations but obviously it is far from complete information regarding plants and other living organisms. In the preprocessing step, we removed the plants that are associated to a few number of metabolites according to our data since the known MC data of these plants are not enough to reveal their connection with other plants. For the sake of balance, we also removed some plants containing many metabolites, e.g. *Arabidopsis thaliana*, which has been frequently studied as a model plant. Figure 2.1 shows the distribution of plants with respected to related metabolites.

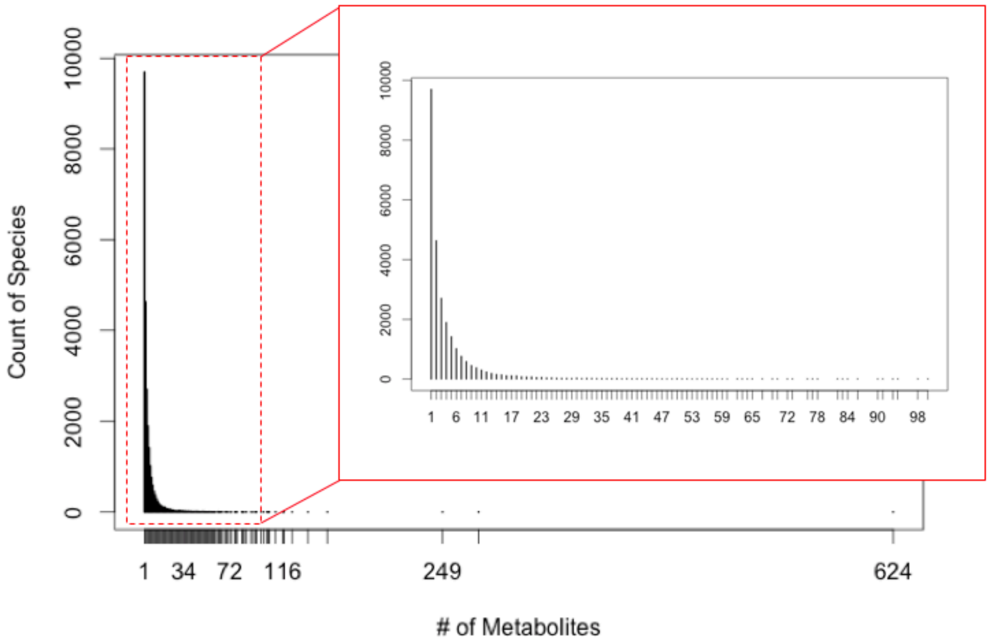


Figure 2.1. *The degree distribution of plants in the plant-metabolite bipartite graph. The x-axes represents the count of metabolites belonging to one species and the y-axes represents the frequency of such species. The initial part of the distribution is shown in the inset figure. The plants containing few metabolites (<45) and excessive metabolites (>150) have been removed from our dataset.*

We also collected the MOL file (Molecular data file created in the MDL Molfile format) which contain structure information for a single molecular compound for all the metabolites from KNApSAcK family DB, as additional input data. The MOL files can be transformed into SDF files (Structure Definition File) which can be used to generate atom pair fingerprints by ChemMine package in R (Cao et al., 2008). The atom pair fingerprints are used to assess the structure similarity between metabolite compounds.

2.3. Methods

2.3.1. Background Population of Tanimoto Coefficients

To classify the plants, we propose a method for measuring the MC-similarity between plants. A straight forward way for this is to represent the plant-metabolite relations as a binary matrix and then by means of a metric or measure to calculate similarity scores between plants and subsequently using such scores to classify the plants. However, we propose a method involving structural similarity between MCs of plants. The intuition behind this approach is to compensate the gap of incomplete data and to take into consideration the fact that structurally similar metabolites are part of the same or similar pathways. Adjacent metabolites along the metabolic pathways are often related to similar substructures and plants with highly structurally similar metabolites are likely to be within the same category.

We utilize Tanimoto coefficient to measure the structural similarity between two compounds (Godden et al., 2000). Willett (2014) investigated different structural similarity measures and concluded that chemoinformatics research on structural similarity would continue to be largely based on the use of 2D fingerprints, and the Tanimoto coefficient has been established as the standard for similarity searching. The Tanimoto coefficient between two metabolites A and B is defined as following, which is the proportion of the features shared by two compounds divided by their union:

$$Tanimoto = \frac{AB}{A + B - AB} \quad (2.1)$$

The variable AB is the number of features common in both compounds, while A and B are the number of features that are related to the respective individual compounds. The

Tanimoto coefficient has a range from 0 to 1 with higher values indicating greater similarity than lower ones. The Tanimoto coefficient can be calculated from compound fingerprints using the R package ChemMine (Cao et al., 2008).

We calculate the Tanimoto coefficients (TCs) for all the metabolite compound pairs in our dataset and this constitute the background population of TCs in the present research. Figure 2.2(A) shows the background distribution of TCs by dividing the range of 0 to 1 into 100 slots.

2.3.2. Plant-Plant Similarity Score

In this paper, we introduce a novel method for calculating a similarity score between plants based on metabolite-contents by means of Tanimoto coefficients. Different plants are reported to have different number of metabolites because some plants have been studied for longer time and by more groups compared to others. Our objective is to develop a standard method such that the similarity score falls between a range of 0 to 1.

Assume that there are two plants *A* (containing M_a metabolites) and *B* (containing M_b metabolites) and they share M_{ab} common metabolites. Let n be the number of unique metabolite pairs corresponding to plant pair *A* and *B*. It is easy to derive that n can be expressed by equation (2.2) as follows:

$$n = M_a M_b - \frac{M_{ab}(M_{ab}+1)}{2} \qquad (2.2)$$

For all the metabolite pairs we can get the Tanimoto coefficients and make a distribution of TCs corresponding to a pair of plants. In such a distribution, when two plants contain structurally similar metabolites, more and higher slots are likely to be on the right side. For example, Figure 2.2(B) shows the distribution of TCs corresponding to plants *Garcinia mangostana* and *Garcinia dulcis* which are in the same genus. On the other hand, when the metabolite-contents of two plants are not similar, the height of the slots on the left side are likely to be larger. For example, Figure 2.2(C) shows the distribution of TCs corresponding to plants *Phaseolus vulgaris* and *Curcuma zedoaria* which are taxonomically far different from each other.

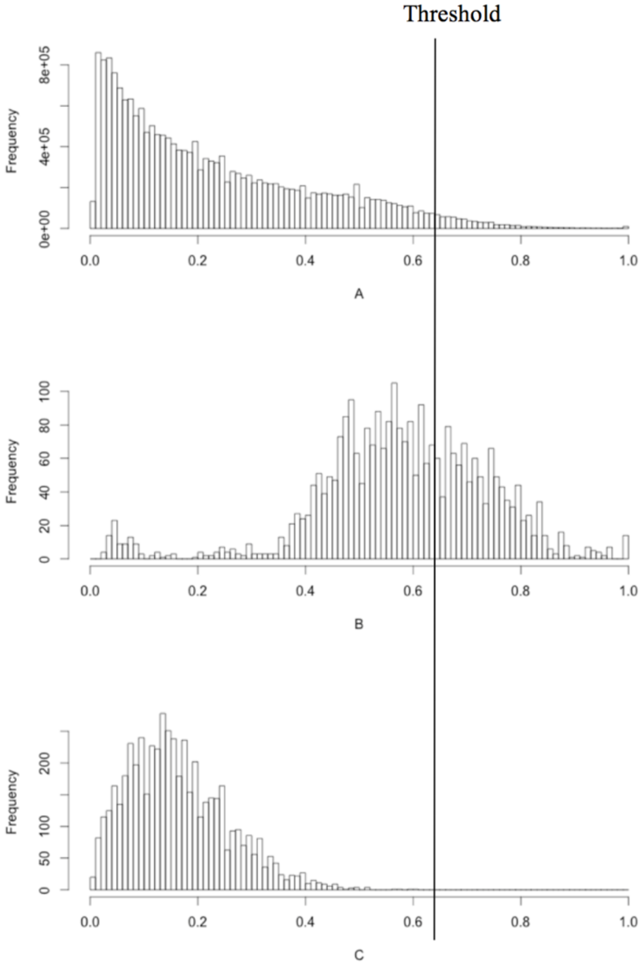


Figure 2.2. *The distributions of TCs in our dataset. (A) The distribution of TCs for all metabolite pairs which we call background population. A threshold is selected based on the background population of TCs. The MC-similarity of two plants is equivalent to the proportion of the count of TCs that are larger than the threshold. (B) The distribution of TCs for two plants (Garcinia mangostana & Garcinia dulcis) having highly similar metabolite-contents. (C) The distribution of TCs for two plants (Phaseolus vulgaris & Curcuma zedoaria) having dissimilar metabolite-contents.*

Let the TCs corresponding to the pair of plants A and B are x_1, x_2,..... x_n, where n is defined in equation (2.2) above. Now for a given threshold TC value T, we calculate the similarity score SS between plants A and B as follows:

$$SS = \frac{1}{n}\sum_{i=1}^{n} Y_i\,, \quad Y_i = \begin{cases} 1, x_i \geq T \\ 0, x_i < T \end{cases} \qquad (2.3)$$

Thus SS is equivalent to the proportion of the count of Tanimoto values that are larger than the threshold and therefore the SS ranges from 0 to 1. By this method, we assess the MC-similarity for all plant pairs reducing the influence of missing metabolite data. We transform the similarity coefficients to distance coefficients by the following transformation $d = 1 - s$ (d: distance score; s: similarity score) and classify the plants using Ward's hierarchical clustering method.

To determine an optimal solution of the threshold T, we classify the plants by various thresholds T and compare the resulted dendrograms with NCBI taxonomy based on a similarity score called Baker's Gamma correlation coefficient using R package dendextend (Baker, 1974; Galili, 2015;). Baker's Gamma is a measure of accosiation (similarity) between two trees of hierarchical clustering. Baker's Gamma correlation coefficient ranges from -1 to +1, with positive values meaning that the two trees are statistically similar. This measure is not affected by the height of a branch but only by its relative position compared with other branches.

2.4. Results and Discussion

2.4.1. Dataset and Selection of Thresholds

The KNApSAcK Core DB has been developed by collecting information on numerous metabolites of various organisms from published literatures and several databases (DBs), including PubChem. The KNApSAcK family DB represents data as binary relationships and the degree distribution of plants in the plant-metabolite bipartite graph follows a scale-free trend (Shown in Fig. 2.1). For our current research we selected 102 plants involving totally

4448 metabolites such that each plant contains no less than 45 metabolites.

There are 4448 metabolites and hence 9890128 unique metabolite pairs in our dataset. For every metabolite pair, we calculate their Tanimoto coefficients (TCs), which comprise the background population of TCs (Shown in Fig. 2.2(a)). We determine the threshold TCs based on the background population of TCs, which we use to evaluate the plant-plant similarity score. The background population of TCs ranges from 0.007 (minimum) to 1 (maximum) with the mean value of 0.231 and median value of 0.177.

In this work, we utilized several thresholds of TC values as follows: 0.129, 0.177, 0.233, 0.306, 0.4, 0.522, 0.605, 0.746 which correspond to top 60%, 50%, 40%, 30%, 20%, 10%, 5% and 1 % TCs respectively in the context of the background population. Our proposed similarity score between plants try to measure the abundance of structurally similar metabolites between plants compared to structurally dissimilar metabolites in the context of a given threshold.

Corresponding to each threshold as mentioned in the previous section, we determined plant-plant distance matrix and we performed Ward’s hierarchical clustering to reconstruct a phylogenetic tree of 102 plants. We use NCBI taxonomy generated using a web-based tool from NCBI homepage (http://www.ncbi.nlm.nih.gov/taxonomy) as the reference classification to evaluate our dendrograms. Our classifications are based on very limited number of reported metabolites and hence not very similar to the NCBI classification nevertheless the trend is promising and it implies that addition of more information would improve the results. We compared our classification trees with NCBI taxonomy by Baker's Gamma correlation coefficient(Baker, 1974; Galili, 2015;). Figure 2.3 shows the results of comparison.

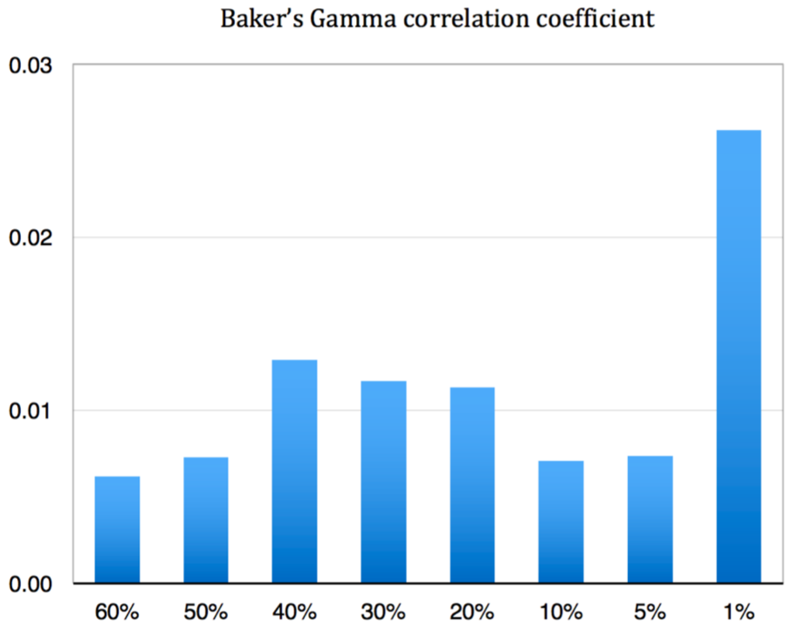


Figure 2.3. *Baker's Gamma correlation coefficient corresponding to different threshold levels. The value of Baker's Gamma correlation coefficient between two tree can range between -1 to +1, with higher values indicating greater similarity between two trees than lower ones.*

Because our metabolite-content dataset is incomplete and the metabolite-contents of plants are influenced by ecology and environment, our tree is not highly similar with NCBI taxonomy tree. But in Fig 2.3, all the trees correspond to positive values of Baker's Gamma correlation coefficient, which means that all the trees are statistically similar with the NCBI taxonomy. We can also observe two peaks, one at 40% and another at 1%. In case of 40% similarity scores between plants are determined based on much more metabolite pairs compared to 1% case. In case of 1 %, only highly structurally similar metabolite pairs are considered. This implies that plants are likely to be in the same class if they either contain highly structurally similar metabolites or many metabolites with reasonable structural similarity. The best result is obtained by using the threshold corresponding to top 1% TCs as shown in Figure 2.4. We cut the dendrogram tree corresponding to the best result (i.e. in case of 1% threshold) to 28 clusters at an empirical threshold height of 0.98, and interpret the clusters with their taxonomy category and economic uses (edible, medicinal etc.) as it is shown in Table 2.1. The order of plants in Table 2.1 is according to the dendrogram of hierarchical clustering.

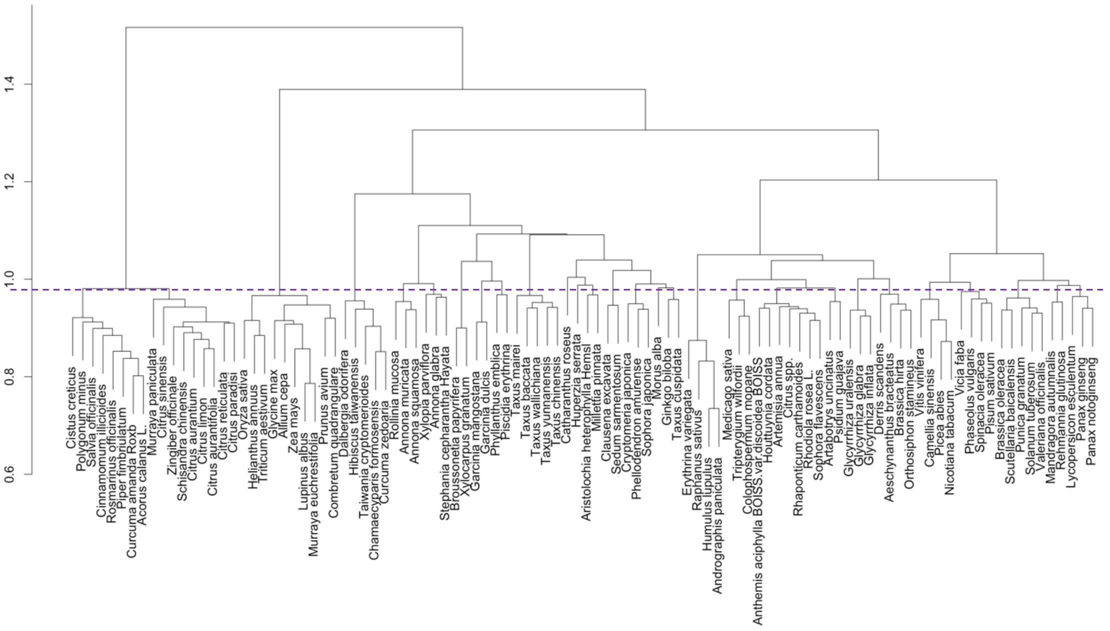


Figure 2.4. *Hierarchical dendrogram plot of 102 plants based on their MC-similarity, using 1% TCs as threshold. The 102 plants are classified in to 28 clusters by cutting the dendrogram with an empirical threshold height 0.98.*

2.4.2. Classification of Plants

The main defined ranks in NCBI taxonomic hierarchy are as follows: *superkingdom, kingdom, phylum, subclass, order, family, subfamily, tribe, genus, species* (from high to low). The highest rank in our dataset is *phylum (Streptophyta)*. There are 102 plants in our data from 45 different families. At genus level the largest group consist of 6 plants belonging to genus *Citrus* and at the family level the largest group consist of 17 plants belonging to *Fabaceae* family. Some species are isolated in our dataset i.e. they are the only candidate from a sub-class.

When more than one plants are of the same rank within a cluster we mention the rank giving priority to lower rank in column 3 of Table 2.1. From Table 2.1 we can see that many clusters are rich with plants of similar taxonomic ranks. In case of 9 clusters (5, 7, 8, 9, 10, 19, 22, 27, 28) all plants of an individual cluster can be assigned to the same genus or family or subclass. Cluster 5 contains 3 plants of family *Annonaceae.* Cluster 7 contains 2 plants of subclass *rosids.* Cluster 8 contains 2 plants of genus *Garcinia.* Cluster 9 contains 2 plants of subclass *rosids.* Cluster 10 contains 5 plants of genus *Taxus.* Cluster 19 contains 3 plants of subclass *rosids*, where 2 plants are of family *Fabaceae.* Cluster 22 contains 3 plants of genus *Glycyrrhiza.* Cluster 27 contains 2 plants of subclass *asterids.* Cluster 28 contains 3 plants of subclass *asterids*, where 2 plants are of genus *Panax.*

Plants of similar taxonomical ranks are also accumulated in other clusters. In Cluster 1, *Salvia officinalis* and *Rosmarinus officinalis* belong to tribe *Mentheae*, *Cinnamomum illicioides* and *Piper fimbriulatum* belong to subclass *Magnoliidae*, *Curcuma Amanda* and *Roxb Acorus calanus L.* belong to class *Liliopsida.* In Cluster 2, 7 plants belong to subfamily *Aurantioideae* of which 6 plants belong to genus *Citrus.* Cluster 3 contains total 10 plants, out of them 4 plants belong to class *Liliopsida* of which 3 plants belong to family *Poaceae,* and 5 other plants belong to subclass *rosids* of which 2 plants belong to the same family *Papilionoideae.* Cluster 4 contains 2 plants of family *Cupressaceae* and 2 plants of subclass *rosids.* Cluster 6 contains 2 plants of genus *Annona.* Cluster 15 contains 2 plants of subclass *rosids.* Cluster 17 contains 2 plants of *Acrogymnospermae.* Cluster 18 contains 3 plants of subclass *rosids.* Cluster 20 contains 3 plants of *Asteraceae* family and 2 plants of subclass *rosids.* Cluster 23 contains 2 plants of order *Lamiales*, and 2 plants of subclass *rosids.* Cluster 24 contains 2 plants of subclass *asterids.* Cluster 25 contains 3 plants of subfamily *Papilionoideae.* Cluster 26 contains 2 plants of subclass *rosids* and 2 plants of subclass *asterids.* Our method also successfully placed some of the isolated plants of different orders to individual single plant cluster e.g. cluster 11, 12.

Our MC-similarity based classification of plants is consistent with chemosystematics pattern of plants. Some deviations in our classification from NCBI classification can be explained in terms of ecological relationships or bioactivity similarity between plants.

Generally, medicinal properties are not randomly distributed in different classes of plants. Some plant groups are represented by more medicinal plants than others. It is suggested that there is a phylogenetic pattern in medicinal properties even within one genus (Saslis-Lagoudakis et al., 2011). A similar distribution is also observed in our classification that medicinal plants and edible plants are distributed in different groups. Interestingly, out of the two biggest plant groups (cluster 3 and cluster 1), one is dominated by medicinal plants and the other is dominated by edible plants even though these two groups are not represented by plants of the similar taxonomic ranks. This implies the importance of MC-similarity based classification of plants. The accumulation of medicinal plants in cluster 1, which are distributed across different genus is a result of rational chemosystematics pattern in those plants. Thus it can be concluded that the proposed method is suitable for searching new medicinal and edible plants. Another example is plants of the genus *Taxus*. Most of the plants of genus *Taxus* can be used in medicines. In our classification, all plants of genus *Taxus* within our dataset are assigned together to cluster 10 except *Taxus cuspidate*, which is poisonous and frequently used as timber instead of a drug. It also shows that considering metabolite-content patterns is more useful compared to phylogenetic patterns for searching medicinal plants. We also observe that plants belonging to *Fabaceae* family are classified into different clusters according to their economic use pattern. *Glycyrrhiza uralensis*, *Glycyrrhiza glabra* and *Glycyrrhiza inflata* are classified in Cluster 22 as 3 medicinal plants. *Vicia faba*, *Phaseolus vulgaris* and *Pisum sativum* are classified in Cluster 25 as 3 edible plants. *Medicago sativa* and *Colophospermum mopane* are classified in Cluster 19 as 2 un-edible plants. Other plants can also be found in other clusters with the similar economic use patterns.

Furthermore, we analyze all the clusters based on usage pattern of plants. For each cluster of size more than five, more than 70% plants belong to one of the following three categories: edible, medicinal and un-edible. Also, half of the smaller clusters are dominated by plants of certain usage. The larger clusters tend to better reflect the usage pattern based classification. Six clusters (10, 14, 21, 22, 23, 27) consist of 100% medicinal plants. Five clusters (5, 8, 12,16, 25) consist of 100% edible plants. Three clusters (1, 18, 20) consist of more than 70% medicinal plants, and 2 clusters (2, 3) consist of more than 70% edible plants.

In addition, 2 clusters (4, 11) are mostly composed of un-edible (decorative/timber) plants. 10 clusters (6, 7, 9, 13, 17, 15, 19, 24, 26, 28) are not dominated by plants of specific uses but most of them are dominated by plants of the similar taxonomic ranks.

Our classification is taking into account both general phylogenetic relations and relation between plants based on bioactive similarity. Therefore, our method can be used for searching new plants useful for medicine or other economic purposes.

The metabolite-content data we used in this study is far from complete. Also we considered a very limited number of plants due to lack of data. However, from the trends of the results we obtained it can be concluded that if more information is added then our method would be able to perform much better classification of plants.

Table 2.1. *Classification of 102 plants. Cluster ID, plant names, taxonomic ranks and economic uses are mentioned in consecutive columns. The taxonomic ranks with ' * ' means the plants above it belong to the same subclass or subfamily.*

ID	Plants	Taxonomy	Usage Pattern
1	*Salvia officinalis*	tribe(*Mentheae*)	medicinal
	Rosmarinus officinalis	tribe(*Mentheae*)	decorative
	Cinnamomum illicioides	subclass(*Magnoliidae*)	medicinal
	Piper fimbriulatum	subclass(*Magnoliidae*)	Toxicity
	Curcuma amanda Roxb	class(*Liliopsida*)	medicinal
	Acorus calanus L.	class(*Liliopsida*)	Toxicity
	Cistus creticus	phylum(*Streptophyta*)	decorative
	Polygonum minus	phylum(*Streptophyta*)	medicinal
2	*Citrus aurantium*	genus(*Citrus*)	edible
	Citrus limon	genus(*Citrus*)	edible
	Citrus aurantifolia	genus(*Citrus*)	edible
	Citrus reticulata	genus(*Citrus*)	edible
	Citrus paradisi	genus(*Citrus*)	edible
	Citrus sinensis	genus(*Citrus*)	edible
	Murraya paniculata	subfamily(*Aurantioideae*)*	decorative
	Zingiber officinale	phylum(*Streptophyta*)	medicinal/edible
	Schisandra chinensis	phylum(*Streptophyta*)	medicinal

Table 2.1. *Classification of 102 plants.* (Cont.)

ID	Plants	Taxonomy	Usage Pattern
3	*Oryza sativa*	family(*Poaceae*)	edible
	Triticum aestivum	family(*Poaceae*)	edible
	Zea mays	family(*Poaceae*)	edible
	Allium cepa	class(*Liliopsida*)*	edible
	Glycine max	subfamily(*Papilionoideae*)	edible
	Lupinus albus	subfamily(*Papilionoideae*)	decorative
	Murraya euchrestifolia	subclass(*rosids*)*	edible
	Prunus avium	subclass(*rosids*)*	edible
	Combretum quadrangulare	subclass(*rosids*)*	Timber
	Helianthus annuus	phylum(*Streptophyta*)	edible
4	*Taiwania cryptomerioides*	family(*Cupressaceae*)	decorative
	Chamaecyparis formosensis	family(*Cupressaceae*)	Timber
	Dalbergia odorifera	subclass(*rosids*)	decorative
	Hibiscus taiwanensis	subclass(*rosids*)	decorative
	Curcuma zedoaria	phylum(*Streptophyta*)	edible
5	*Rollinia mucosa*	family(*Annonaceae*)	edible
	Annona muricata	family(*Annonaceae*)	edible
	Annona squamosa	family(*Annonaceae*)	edible
6	*Xylopia parviflora*	genus(*Annona*)	medicinal
	Annona glabra	genus(*Annona*)	edible
	Stephania cepharantha Hayata	phylum(*Streptophyta*)	medicinal
7	*Broussonetia papyrifera*	subclass(*rosids*)	medicinal
	Xylocarpus granatum	subclass(*rosids*)	Timber

Table 2.1. *Classification of 102 plants.* (Cont.)

ID	Plants	Taxonomy	Usage Pattern
8	*Garcinia mangostana*	genus(*Garcinia*)	medicinal/edible
	Garcinia dulcis	genus(*Garcinia*)	edible
9	*Phyllanthus emblica*	subclass(*rosids*)	edible
	Piscidia erythrina	subclass(*rosids*)	decorative
10	*Taxus mairei*	genus(*Taxus*)	medicinal
	Taxus baccata	genus(*Taxus*)	medicinal
	Taxus wallichiana	genus(*Taxus*)	medicinal
	Taxus yunnanensis	genus(*Taxus*)	medicinal
	Taxus chinensis	genus(*Taxus*)	medicinal
11	*Catharanthus roseus*	order(*Gentianales*)	decorative
12	*Huperzia serrata*	order(*Lycopodiales*)	edible
13	*Aristolochia heterophylla Hemsl*	phylum(*Streptophyta*)	medicinal
	Millettia pinnata	phylum(*Streptophyta*)	decorative
14	*Clausena excavata*	phylum(*Streptophyta*)	medicinal
	Sedum sarmentosum	phylum(*Streptophyta*)	medicinal
15	*Phellodendron amurense*	subclass(*rosids*)	medicinal
	Sophora japonica	subclass(*rosids*)	medicinal
	Cryptomeria japonica	phylum(*Streptophyta*)	Timber
16	*Morus alba*	order(*Rosales*)	medicinal/edible
17	*Ginkgo biloba*	*Acrogymnospermae*	medicinal/edible
	Taxus cuspidata	*Acrogymnospermae*	Timber/poisonous

Table 2.1. *Classification of 102 plants.* (Cont.)

ID	Plants	Taxonomy	Usage Pattern
18	*Erythrina variegata*	subclass(*rosids*)	Toxicity/medicinal
	Raphanus sativus	subclass(*rosids*)	medicinal/edible
	Humulus lupulus	subclass(*rosids*)	edible
	Andrographis paniculata	phylum(*Streptophyta*)	medicinal
19	*Medicago sativa*	family(*Fabaceae*)	forage
	Colophospermum mopane	family(*Fabaceae*)	Timber
	Tripterygium wilfordii	subclass(*rosids*)*	medicinal
20	*Anthemis aciphylla BOISS*	family(*Asteraceae*)	medicinal
	Rhaponticum carthamoides	family(*Asteraceae*)	medicinal
	Artemisia annua	family(*Asteraceae*)	medicinal
	Citrus spp.	subclass(*rosids*)	edible
	Sophora flavescens	subclass(*rosids*)	medicinal
	Houttuynia cordata	phylum(*Streptophyta*)	medicinal/edible
	Rhodiola rosea L.	phylum(*Streptophyta*)	medicinal
21	*Artabotrys uncinatus*	phylum(*Streptophyta*)	medicinal
	Psidium guajava	phylum(*Streptophyta*)	medicinal/edible
22	*Glycyrrhiza uralensis*	genus(*Glycyrrhiza*)	medicinal
	Glycyrrhiza glabra	genus(*Glycyrrhiza*)	medicinal
	Glycyrrhiza inflata	genus(*Glycyrrhiza*)	medicinal
23	*Aeschynanthus bracteatus*	order(*Lamiales*)	not available
	Orthosiphon stamineus	order(*Lamiales*)	medicinal
	Derris scandens	subclass(*rosids*)	medicinal
	Brassica hirta	subclass(*rosids*)	medicinal

Table 2.1. *Classification of 102 plants.* (Cont.)

ID	Plants	Taxonomy	Usage Pattern
24	*Nicotiana tabacum*	subclass(*asterids*)	medicinal
	Camellia sinensis	subclass(*asterids*)	edible
	Vitis vinifera	phylum(*Streptophyta*)	edible
	Picea abies	phylum(*Streptophyta*)	decorative
25	*Vicia faba*	subfamily(*Papilionoideae*)	edible
	Phaseolus vulgaris	subfamily(*Papilionoideae*)	edible
	Pisum sativum	subfamily(*Papilionoideae*)	edible
	Spinacia oleracea	phylum(*Streptophyta*)	edible
26	*Brassica oleracea*	subclass(*rosids*)	edible
	Punica granatum	subclass(*rosids*)	edible
	Scutellaria baicalensis	subclass(*asterids*)	medicinal
	Solanum tuberosum	subclass(*asterids*)	edible
	Valeriana officinalis	subclass(*asterids*)	medicinal
27	*Mandragora autumnalis*	subclass(*asterids*)	medicinal
	Rehmannia glutinosa	subclass(*asterids*)	medicinal
28	*Panax ginseng*	genus(*Panax*)	medicinal
	Panax notoginseng	genus(*Panax*)	medicinal
	Lycopersicon esculentum	subclass(*asterids*)*	edible

2.5. Summary

In this chapter, we described an approach for classifying plants by comparing the metabolite-contents and for studying the relationship of plants in term of their bioactive properties. We show that by this approach we can produce a classification of plants similar to the traditional plant taxonomy. This method can be regarded as a novel chemosystematics method which consider the global metabolite-contents of plants instead of a group of metabolites as some previous researches.

Tanimoto coefficients (TCs) has been utilized to assess similarity between metabolite compounds and the plant-plant similarity score is calculated based on the background population of TCs. To determine the optimal threshold, we sorted the background population of TCs and utilized several thresholds of TC respectively in the context of the background population. NCBI taxonomy tree is utilized as reference tree to evaluate our dendrograms and to determine the best threshold. Finally, we classify 102 plants into 28 clusters and analyze the phylogeny and bioactivity relationship within each cluster.

The result proves that MC-similarity of plants is associated to the pathway and bioactive similarity and can be regarded as a taxonomy marker which takes into account both general phylogenetic relations and relation between plants based on bioactivity and economic uses.

Chapter 3

Clustering Plants based on Structural-Similarity Network of Metabolites

3.1. Background

In previous chapter, we described an approach of plant classification by similarity of their metabolite-contents. We calculate structural similarity scores for all of the metabolite pairs using Tanimoto coefficients, and calculate plant-plant similarity scores based on the distribution of TCs. With this approach, we consider all of the metabolites and deal with them equally without taking into account the relations between metabolites. However, the metabolites within a cell are not isolated but participant in general metabolic reactions with other metabolites, and is constantly changing due to all the chemical reactions occurring in this cell. As we described before, the metabolite-content of a plant is highly related to its pathways which are regulated by the related enzyme genes, and the adjacent metabolites along a metabolic pathway are often related to similar substructures. Moreover, for most of plants only a set of metabolites are identified, and there are many data-gaps for metabolite-contents of them. Therefore, it is necessary to find an approach of integrating the metabolite-content data to reduce the disturbance of missing metabolite-content data.

In this chapter, we utilize the plant-metabolite relation data obtained from KNApSAcK Core DB to assess the systematics value of metabolite-content data of plants, and investigate the evolutionary and bioactive similarity among plants based on their MC-similarity. The metabolite-content data of plants and structure data of compounds are mainly obtained from the KNApSAcK Core DB and partially from PubChem DB (Bolton et al., 2008; Wang et al., 2009). We measure the structural similarity between two metabolites by using the concept of the Tanimoto coefficient (Godden et al., 2000; Chen & Reynolds, 2002), construct a network by selecting highly structurally similar metabolite pairs, and determine structurally similar groups of metabolites by using the DPClus algorithm (Altaf-Ul-Amin et al., 2006). We then link plants to such metabolite groups instead of individual metabolites to represent the plants as binary vectors. Several structurally similar metabolites are generally involved in a metabolic pathway. Thus, the use of structurally similar metabolite groups in this study can help to reduce the effect of missing data. Next, the MC-similarity between plants is calculated based on binary similarity coefficients which then transformed into MC-distances. Plants are finally classified using the hierarchical clustering method, and the resulting classification is evaluated by comparing it with the NCBI taxonomy (Federhen, 2011). Our classification results reveal both the phylogeny- and bioactivity-based relations among plants. We also use a support vector machine (SVM) algorithm to classify the plants by their economic uses (Cortes & Vapnik, 1995; Hsu et al., 2003). The performance of the classification reveals the predictive power of metabolite-content in exploring nutritional and medicinal properties of plants. As a byproduct of our analysis, we can predict some currently unknown species-metabolite relations.

3.2. Datasets and Preliminaries

The major input data are species-metabolite relationships obtained from the KNApSAcK Core DB, which is a part of the KNApSAcK Family DB (Afendi et al., 2012). The KNApSAcK Core DB contains most of the published information about species-metabolite relations, but this is obviously far from complete regarding plants and other living organisms. In the preprocessing step, we removed the plants with inadequate plant-metabolite relations to guarantee that the amount of metabolite-contents of selected plants is sufficient enough to reveal their interrelations.

We collected the molecular structure description files for the metabolites in our dataset as additional input data. The KNApSAcK Core DB provides MOL molecular structure files for most of the metabolites. For metabolite compounds with structure files that cannot be obtained from the KNApSAcK Core DB, we downloaded the SDF files directly from the PubChem DB (Bolton et al., 2008; Wang et al., 2009). We used R package ChemmineR (v2.26.0) to generate atom pair fingerprints from molecular structure description files for all the metabolite compounds (Cao et al., 2008). These molecular fingerprints were used to measure the structural similarity for all the metabolite pairs. Figure 3.1(a) illustrates the binary plant-metabolite relations and corresponding molecular fingerprints.

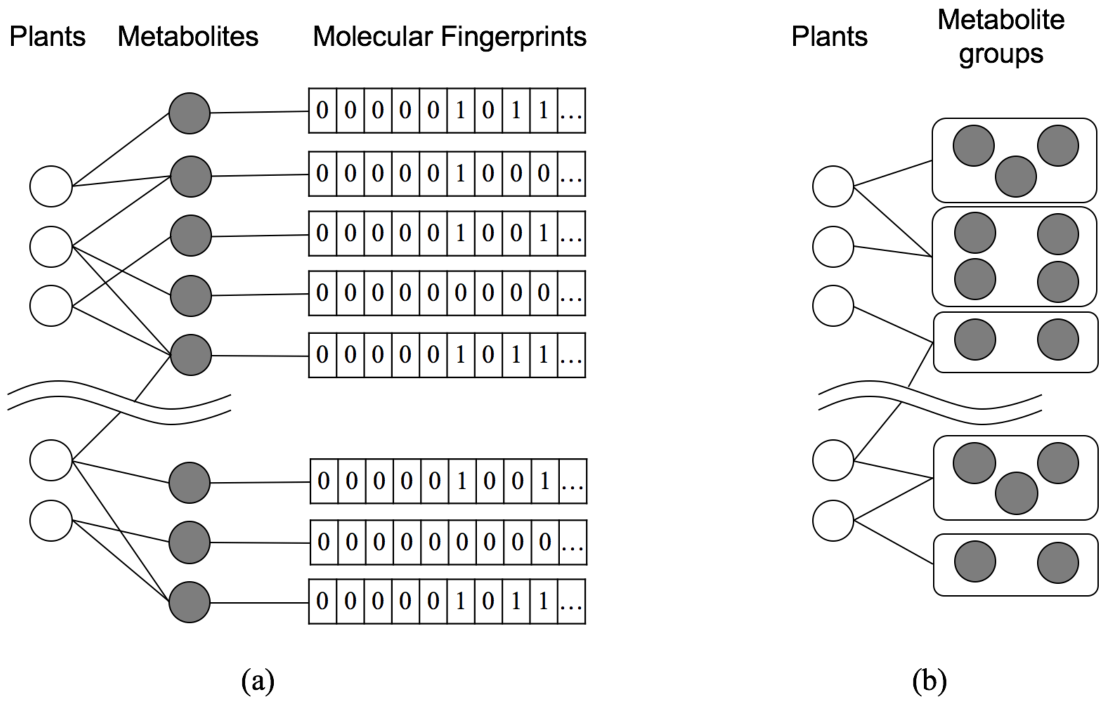


Figure 3.1. *Plant-metabolite relations and plant versus metabolite-group relations. (a) Bipartite graph of plant-metabolite relations. Molecular structures of metabolites are described by 166-bit atom pair fingerprints, which are used to calculate Tanimoto structure similarity score for each metabolite pair. (b) Bipartite graph of plant versus metabolite-group relations. Each plant has been associated with metabolite groups instead of single metabolites to reduce effect of incomplete data.*

3.3. Methods

3.3.1. Network Construction of Metabolites Based on Chemical Structure Similarity.

Very little is known of the complete set of metabolite-contents of plants. Therefore, for classifying plants based on currently available metabolite-content data, an approach that can compensate for the limitations of missing data is needed. Adjacent metabolites along a metabolic pathway are often related to similar substructures; therefore, it can be assumed that structurally similar metabolites are involved in the same or similar pathway. Therefore, plants that share highly structurally similar metabolites are likely to have common pathways; thus, likely to be within the same category and represent similar bioactivity. To compensate for the gap in missing data, we primarily linked plants to structurally similar metabolite groups instead of individual metabolites for this study.

For the purpose of determining structurally similar metabolite groups, we initially constructed a network of metabolites based on chemical structure similarity. We used the Tanimoto coefficient to measure the structural similarity between two metabolites as before (Godden et al., 2000). The Tanimoto coefficient has a range from 0 to 1 with higher values indicating greater similarity than lower ones. Empirically, a Tanimoto coefficient value larger than 0.85 indicates that the compared compounds represent highly similar bioactive features (Martin et al., 2002). We used 0.85 as the threshold to insert an edge between two metabolites and constructed a network of metabolites.

3.3.2. Clustering of Metabolites Based on DPClus

The DPClus algorithm is a graph-clustering algorithm, which has been developed based on a graph-clustering algorithm that can extract densely connected nodes as a cluster (Md Altaf-Ul-Amin et al. 2006). Initially, the algorithm was purposely developed to detect and visualize clusters of proteins in interaction networks which mostly represent molecular biological functional units. Here, we explore the possibility of this algorithm to the structural similarity network of metabolites.

This algorithm can be applied to an undirected simple graph $G = (N, E)$ that consists of a finite set of nodes N and a finite set of edges E. Two important parameters are used in this algorithm, i.e., density d and cluster property cp. Density d_k of any cluster k is the ratio of the number of edges present in the cluster ($|E|$) to the maximum possible number of edges in the cluster ($|E|_{max}$). The cluster property of a node n with respect to cluster k is represented as

$$cp_{nk} = \frac{E_{nk}}{d_k \times N_k}, \qquad (3.1)$$

where N_k is the number of nodes in k, E_{nk} is the total number of edges between n and each node of k.

In this study, we apply the DPClus algorithm to the structural similarity network of metabolites. The metabolites are divided into many groups such that each group contains structurally similar compounds and can be treated as a distinctive pattern of structure. Each metabolite group might be related to a certain pathway, which is related to the phylogeny and ecology of plants. A plant is related to a metabolite group if it is related to any metabolite in the group. Thus, the original plant-metabolite relations are transformed into plant versus metabolite-group relations, as shown in Figure 3.1(b). We use such groups to measure the similarity between plants; thus, reducing the effects of incomplete metabolite-content data.

3.3.3. Clustering of Plants Based on Metabolite Groups.

The relations between plants and structurally similar metabolite groups can be expressed with a sparse binary matrix, which is defined as *M*. Element $M_{ij} = 1$ means that plant *i* contains at least one metabolite of group *j*, and $M_{ij} = 0$ means that plant *i* contains no metabolite of group *j*. Therefore, for each plant, we obtain a binary vector such that each bit corresponds to the presence or absence of a metabolite group.

Let two plants be described by the binary vectors *x* and *y*, each comprised of p variables with values either 1 or 0 ("1" indicates presence while "0" indicates absence), and p is the total number of metabolite groups. The Simpson similarity coefficient between plants can be calculated as

$$S_{Sim} = \frac{a}{\min\{(a+b),(a+c)\}} \qquad (3.2)$$

Here, *a, b,* and *c* are the frequencies of the events $x\&y$, $x\&\bar{y}$, and $\bar{x}\&y$, respectively (Choi et al., 2010; Fallaw, 1979; Ma & Zeng, 2004).

To strengthen our finding with more support, we also used five binary coefficients which stand out from 51 coefficients in the virtual screening experiments for chemoinformatics data:

Jaccard-Tanimoto (JT), Jaccard (Ja), Sokal-Sneath (SS), Consonni-Todeschini (CT) and Gleason (Gle) (Todeschini, 2012). These binary similarity coefficients can be calculated by Equation 3.3 - 3.7.

$$S_{JT} = \frac{a}{a + b + c} \quad (3.3)$$

$$S_{Ja} = \frac{3a}{3a + b + c} \quad (3.4)$$

$$S_{SS} = \frac{a}{a + 2b + 2c} \quad (3.5)$$

$$S_{CT} = \frac{\ln(1 + a)}{\ln(1 + a + b + c)} \quad (3.6)$$

$$S_{Gle} = \frac{2a}{2a + b + c} \quad (3.7)$$

We transformed a similarity coefficient, s, to a distance coefficient, d, by the transformation $d = 1 - s$ and classified the plants by using Ward's hierarchical clustering method using R.

3.3.4. Classification of Plants by SVMs.

Support vector machines are supervised machine learning models for classification and regression analysis (Cortes & Vapnik, 1995; Hsu et al., 2003). An SVM training algorithm builds a model by constructing decision boundaries in feature space. Examples are predicted to belong to a category based on the boundaries.

To study the relationship between metabolite groups and economic uses of plants and evaluate the predictive power of metabolite-content in guiding the discovery of natural products or medicinal properties in plants, we used an SVM algorithm, which was implemented by the function *svm* in R package e1071 v1.6-7, to classify plants by using default parameters (Chang & Lin, 2011; Fan et al., 2005; Dimitriadou et al., 2005). We used economic uses as labels and corresponding metabolite groups as features. The classification performance is evaluated by using a confusion matrix. In a confusion matrix, the sum of a column represents the instances in a predicted class, while the sum of a row represents the instances in an actual class. All programs in this research were run in R v3.3.1.

3.4. Results and Discussion

3.4.1. Plant Representation Based on Metabolite-Content Similarity.

The KNApSAcK Core DB contains a total of 111,199 species-metabolite binary relations that encompass 25,658 species and 50,899 metabolites. This DB was developed by collecting information on numerous metabolites of various organisms from published literature and several DBs, including PubChem (Bolton et al., 2008; Wang et al., 2009). The species-metabolite relations in the KNApSAcK Core DB can be represented as a bipartite graph, as shown in Figure 3.1(a). Figure 3.2 shows the degree distribution of species in a species-metabolite bipartite graph. This distribution follows a power law trend (Fig. 3.2) (Jeong et al., 2001).

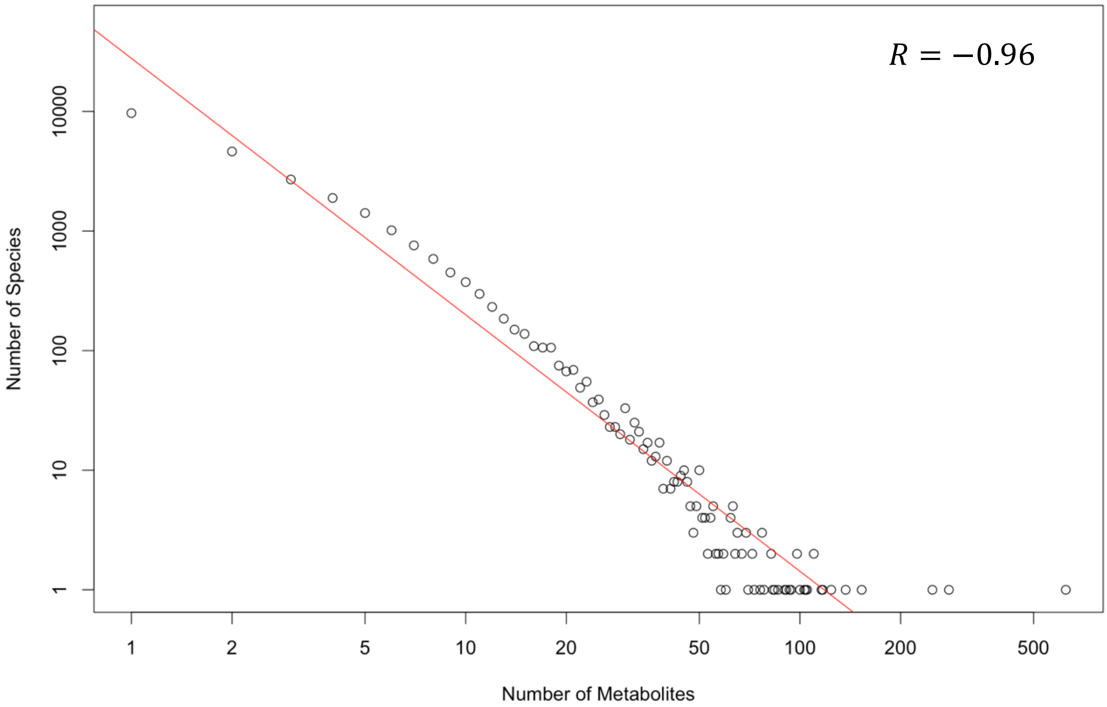
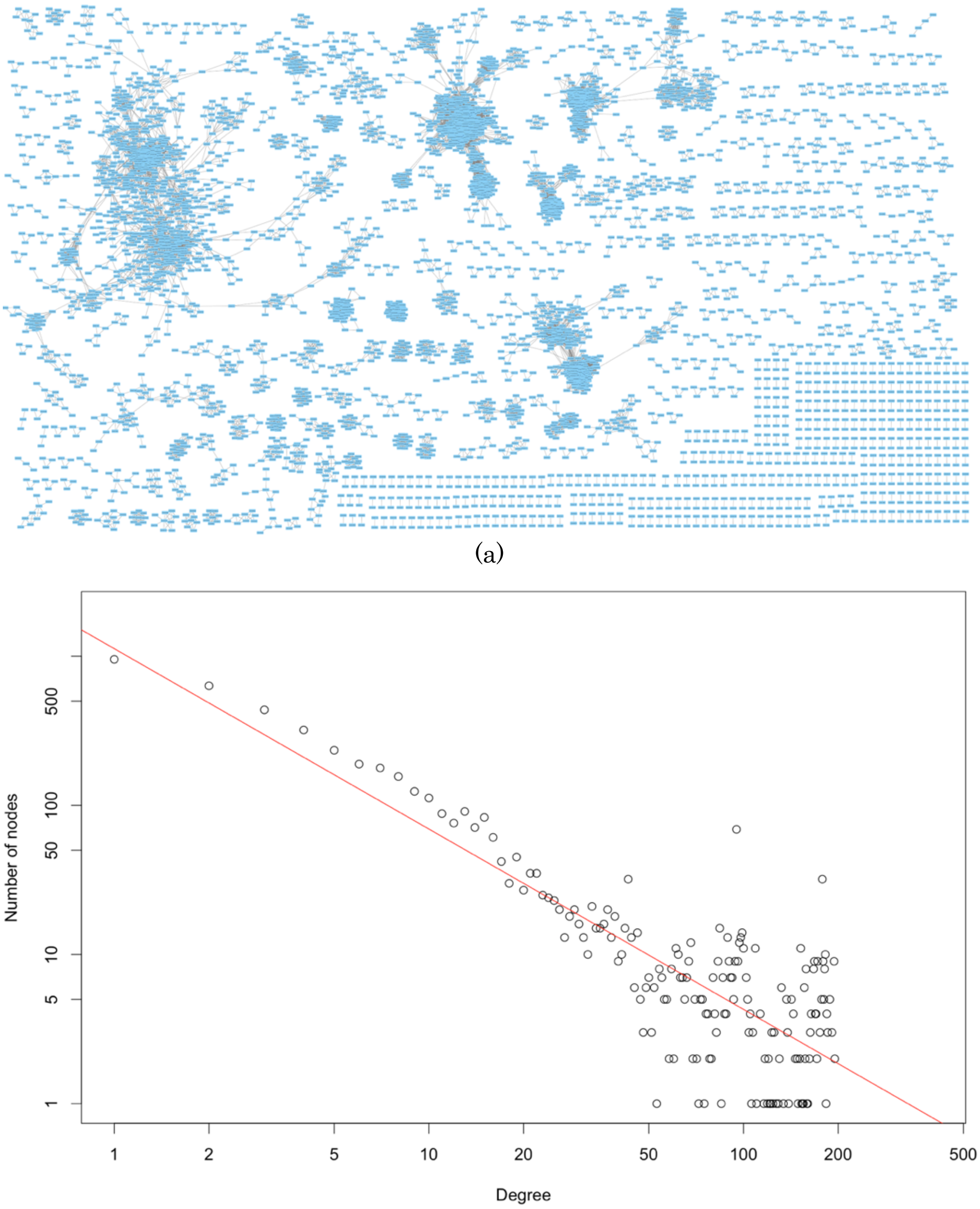


Figure 3.2. *The degree distribution of species in the species-metabolite bipartite graph. The x-axes represents the number of metabolites belonging to one species and the y-axes represents the frequency of such species.*

The metabolite-content data of plants in the KNApSAcK Core DB is unbalanced, i.e., many plants are associated with only a few metabolites and a few plants are associated with many metabolites, while other plants are in a between situation. One of the reasons behind this is that different plants have metabolic pathways of varying complexity. Medicinal plants usually contain more metabolites compared to edible plants because the former have gone through less artificial selections and preserved more secondary metabolites during evolution. Another reason is that the metabolomics of some important plants have been studied more systematically. The recorded metabolite-content of such plants are more comprehensive compared to wild plants. Therefore, in our current research, we selected 216 plants from a total of 25658 species in the KNApSAcK Core DB, such that each of the 216 plants is reported to be associated with no less than 30 metabolites, with 135 being the maximum number and 31 being the minimum. There is a total of 6522 metabolites related to the 216 plants in our input dataset.

We dealt with 6522 metabolites involving 216 plants. We determined the Tanimoto coefficients between all possible metabolite pairs (21264981 pairs). We selected 54528 metabolite pairs with Tanimoto values greater than 0.85, which are 0.25% of all the metabolite pairs. On average, each metabolite is related to about eight different metabolites. We connected all the selected metabolite pairs and constructed a network of metabolites, as shown in Figure 3.3(a). This network involves 5085 metabolites and the other 1437 metabolites are not included in the network, i.e., each of these metabolites is not structurally similar with any other metabolites. The 5085 metabolites included in the network are divided into 669 connected components. Figure 3.3(b) shows the degree distribution of the network of metabolites. This distribution also follows a power law trend (Jeong et al., 2001).



(a)

(b)

Figure 3.3. *Structural-similarity-based network of metabolites. (a) Structural-similarity-based network of metabolites (plotted using network analysis tool Cytoscape v3.3.0). This network is composed of many isolated components, and each component contains different number of nodes. (b) Degree distribution of the network in log scale.*

To compensate for the gap in incomplete data regarding species-metabolite relations, we associated plants with structurally similar metabolite groups instead of individual metabolites. To achieve this, we applied the DPClus algorithm to the network of metabolites we developed, as discussed in the previous section. We did DPClus clustering with the following settings: cluster property *cp* was set to 0.5, density value *d* was set to 0.9, minimum cluster size was set to 2, and we used the overlapping mode.

The DPClus algorithm generated 1150 clusters, i.e., metabolite groups, involving 4700 metabolites. The largest group contained 174 metabolites, and there were 510 metabolite groups containing only 2 metabolites. Figure 3.4 shows the frequency of metabolite groups with respect to size (the count of metabolites) in both normal scale and log-log scale (inset), and this distribution also follows a power law trend (Jeong et al., 2001). A total of 1822 metabolites not included in any cluster are considered as groups consisting of a single metabolite.

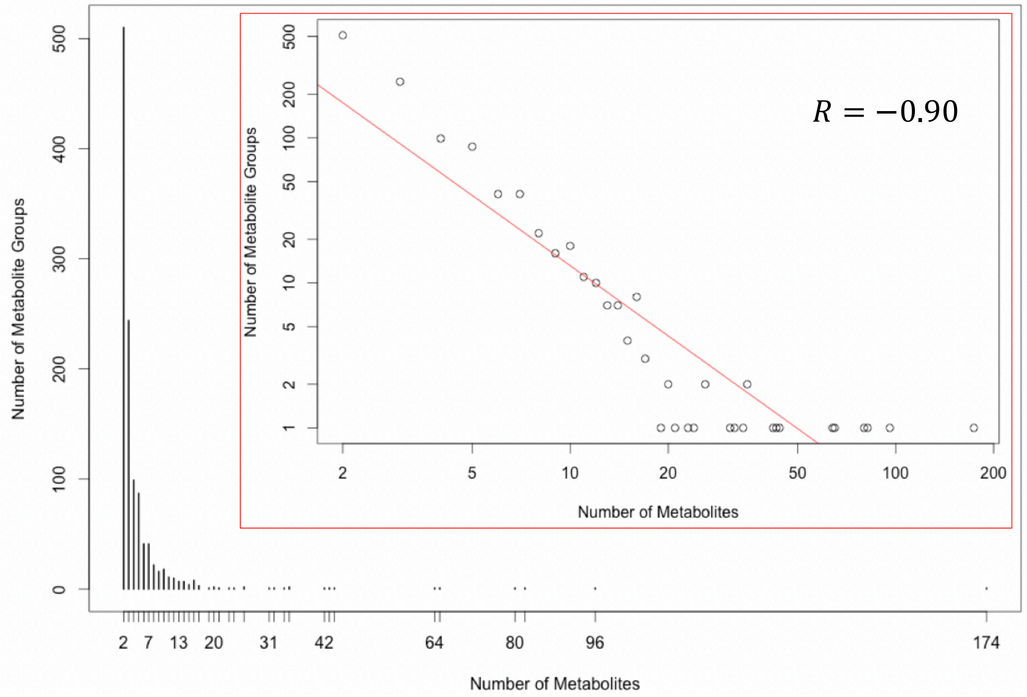


Figure 3.4. *Frequency of metabolite groups with respect to group size. X-axes represent number of metabolites belonging to one metabolite group, and y-axes represent frequency of such metabolite groups. Frequency of metabolite groups in log scale is shown in inset figure.*

All clusters, large or small, contained structurally similar metabolites. Large clusters might be related to different metabolic pathways, but small clusters are likely related to specific metabolic pathways. A plant is related to a metabolite group if it is reported to contain any metabolite in the group. A plant can be represented as a binary vector such that each bit of the vector corresponds to the presence or absence of a metabolite group.

3.4.2. Clustering of Plants Based on Metabolite-Content Similarity.

We calculated the plant-plant similarity by using Simpson coefficients and other five binary similarity coefficients which have outstanding performance for comparing chemoinformatics data (Todeschini et al., 2012). We transformed a similarity score s into a distance score d using $d = 1 - s$ then conducted Ward's hierarchical clustering analysis. Thus, we determined six dendrograms corresponding to six types of coefficients.

We used the NCBI taxonomy of the 216 plants generated using a web-based tool from the NCBI homepage (http://www.ncbi.nlm.nih.gov/taxonomy) as the reference classification (Federhen, 2011). The NCBI classification reflects the phylogenetic patterns within a plant group primarily based on morphology. According to the NCBI taxonomy, the 216 plants spread over 52 families with the largest family *Fabaceae* containing 42 plants.

We compared the dendrogram trees generated with our approach with the NCBI taxonomy based on a similarity score called Baker's Gamma correlation coefficient using R package *dendextend* v1.3.0 (Baker, 1974; Galili, 2015). Baker's Gamma correlation coefficient ranges from -1 to +1, with positive values, meaning that the two trees are statistically similar.

The results show that all the six dendrograms produced positive Baker's Gamma values (Simpson: 0.053; Jaccard-Tanimoto: 0.043; Jaccard: 0.040; Sokal-Sneath: 0.042; Consonni-Todeschini: 0.048; Gleason: 0.048), indicating that all the six trees are statistically similar with the NCBI taxonomy. Overall, the Simpson coefficient tree stands out from the remainder in our experiments with the highest Baker's Gamma value. We also illustrate this fact by comparing Simpson tree and Jaccard-Tanimoto tree by pointing out some examples in Figure 3.5. The Jarccard-Tanimoto coefficient has been used as a similarity measure to compare the enzyme content of metabolic networks in each pair of organisms (Deyasi et al., 2015). The Simpson coefficient was devised to minimize the effect of the unequal size of two faunas being compared, and having in the denominator only the number of taxa in a sample having the smaller number (Fallaw, 1979; Ma & Zeng, 2004). In the Simpson coefficient tree, more plants from the same genus or family appeared nearer to each other.

Therefore, for further explanation, we selected the Simpson-tree and classified the plants into 48 groups by cutting the dendrogram at variable threshold heights empirically chosen to enrich the clusters with plants of the same genus or family. Figure 3.6 shows the dendrogram together with group IDs produced by our classification method.

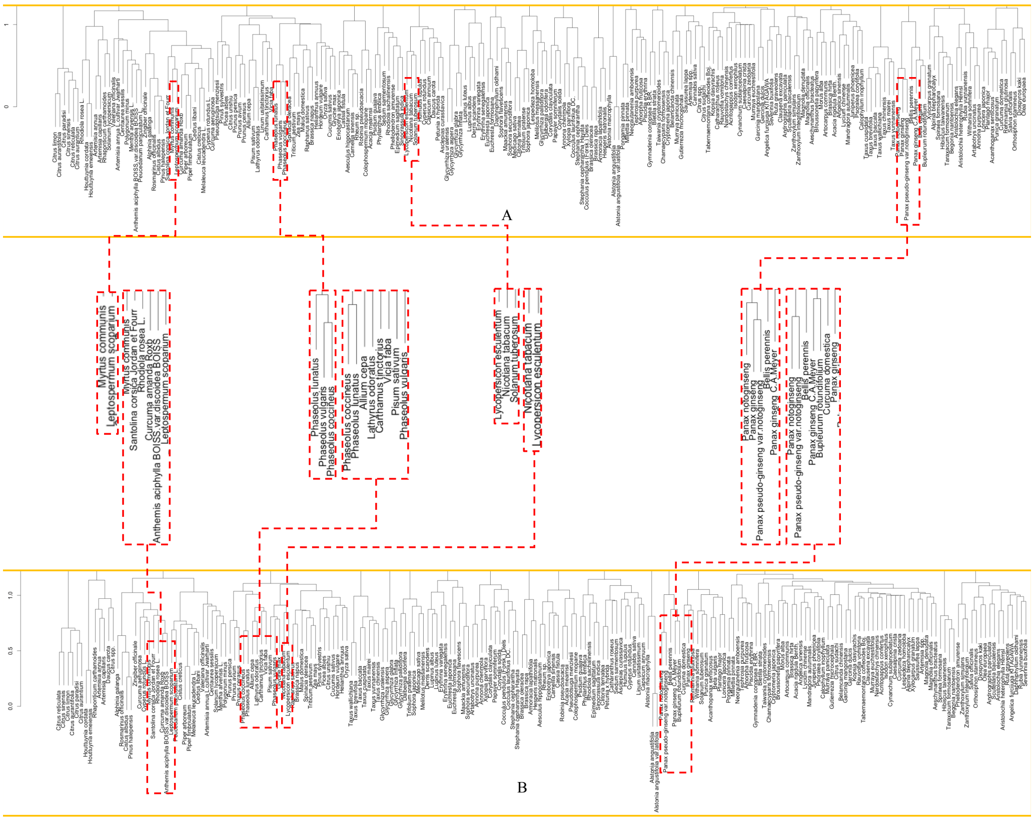


Figure 3.5. *Comparison of Simpson- and Jaccard-Tanimoto dendrograms. (A) Hierarchical dendrogram plot of classification by Simpson similarity coefficients. (B) Hierarchical dendrogram plot of classification by Jaccard-Tanimoto similarity coefficients. Myrtus communis and Leptospermum scoparium belong to family Myrtaceae. Phaseolus lunatus, Phaseolus vulgaris and Phaseolus coccineus belong to genus Phaseolus. Lycopersicon esculentum, Nicotiana tabacum and Solanum tuberosum belong to family Solanaceae. Panax notoginseng, Panax ginseng, Panax pseudo-ginseng var.notoginseng and Panax ginseng C.A.Meyer belong to genus Panax.*

The main defined ranks in the NCBI taxonomic hierarchy are as follows: *superkingdom, kingdom, phylum, subclass, order, family, subfamily, tribe, genus, species* (from high to low). We collected the taxonomy information of 216 plants that we considered in this study and annotated each plant with ranks of *family* and *genus* (we used the scientific names of plants where the first word of a plant name represents the *genus* to which the plant belongs). Table 3.1 lists the 48 groups of plants based on our clustering result with their taxonomic and use information. The plants are arranged by different groups, and for each group plants within the same *family* or *genus* are arranged together to highlight the internal phylogeny relations. In the dendrogram of Figure 3.6, neighboring plants belonging to the same *genus* or *family* are indicated by horizontal bold colored lines. Each *genus* or *family* is indicated by a specific color. It is evident that many clusters are rich with plants from the same *genus* or *family*. Thus, our results imply that plants in the same taxon correspond to similar metabolite-content. Taking into account the inadequate amount of metabolite data and limited number of plants we considered for certain families, the results from our approach are very promising. These indicate that the proposed approach was designed to compensate for the shortcomings of limited data. Some deviations in our classification from the NCBI taxonomy can be explained in terms of ecological relationships or bioactive similarity. This implies that compared to morphology-based taxonomy, MC-based classification reveals more information about the bioactive similarity among plants, which is related to the nutritional and medicinal properties of plants. Therefore, MC-based classification can be used as a time-efficient predictive tool for guiding discovery of edible and medicinal properties in wild plants.

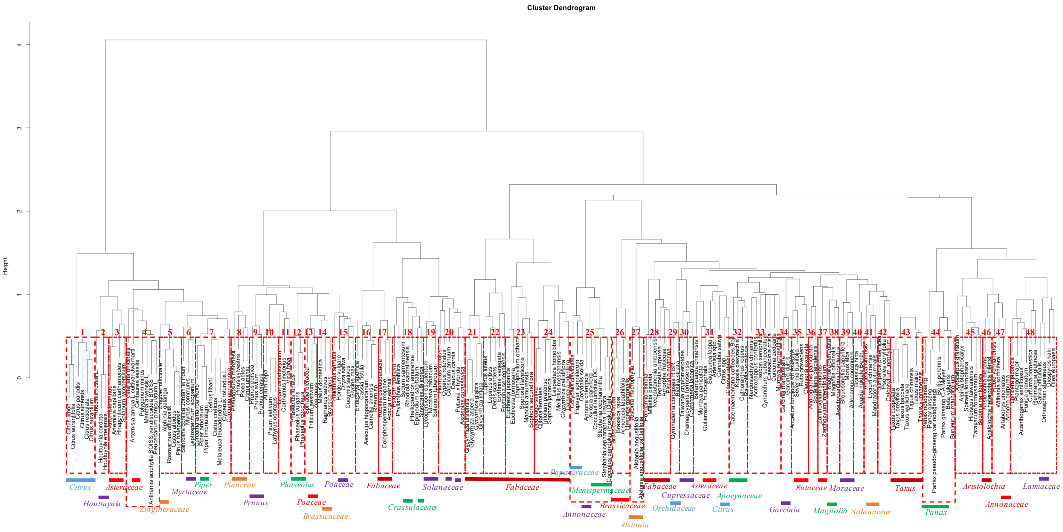


Figure 3.6. *The family or genus level phylogeny patterns in proposed classification of 216 plants. Group IDs are annotated beside the corresponding plant groups. Adjacently placed plants of the same genus or family are indicated by thick bars of specific colors.*

Table 3.1. *Classification of 216 plants. Group ID, plant names, taxonomic ranks (family), and economic uses are mentioned in consecutive columns. Economic uses of plants are represented as following abbreviations: E (edible), M (medicinal), L (landscaping,), T (timber), P (poisonous), W (wild plant). Some plants are both edible and medicinal and are annotated as M/E.*

Group	Plant	Family	Use
1	*Citrus limon*	*Rutaceae*	E
	Citrus aurantifolia	*Rutaceae*	M/E
	Citrus paradisi	*Rutaceae*	E
	Citrus sinensis	*Rutaceae*	E
	Citrus reticulata	*Rutaceae*	E
	Citrus aurantium	*Rutaceae*	E
2	*Houttuynia cordata*	*Saururaceae*	M/E
	Houttuynia emeiensis	*Saururaceae*	W
	Rhodiola rosea	*Crassulaceae*	M
3	*Artemisia annua*	*Asteraceae*	M
	Artemisia capillaris	*Asteraceae*	M
	Rhaponticum carthamoides	*Asteraceae*	W
	Solanum lycopersicum	*Solanaceae*	E
4	*Anthemis aciphylla*	*Asteraceae*	W
	Artemisia annua L	*Asteraceae*	M
	Centaurea sessilis	*Asteraceae*	W
	Valeriana officinalis	*Caprifoliaceae*	M
	Persicaria minus	*Polygonaceae*	M
	Mentha arvensis	*Lamiaceae*	M
	Peucedanum paniculatum	*Apiaceae*	W

Table 3.1. *Classification of 216 plants.* **(Cont.)**

Group	Plant	Family	Use
5	*Zingiber officinale*	*Zingiberaceae*	M/E
	Alphinia galanga	*Zingiberaceae*	M/E
	Rosmarinus officinalis	*Lamiaceae*	M
	Cistus albidus	*Cistaceae*	W
	Pinus halepensis	*Pinaceae*	L
6	*Myrtus communis*	*Myrtaceae*	M
	Leptospermum scoparium	*Myrtaceae*	M
	Santolina corsica	*Asteraceae*	W
7	*Curcuma amanda*	*Zingiberaceae*	M/E
	Curcuma aeruginosa	*Zingiberaceae*	W
	Cistus creticus	*Cistaceae*	W
	Melaleuca leucadendra	*Myrtaceae*	M
	Piper arboreum	*Piperaceae*	W
	Piper fimbriulatum	*Piperaceae*	W
	Cedrus libani	*Pinaceae*	L
	Cyperus rotundus	*Cyperaceae*	M
8	*Pseudotsuga menziesii*	*Pinaceae*	T
	Pinus sylvestris	*Pinaceae*	T
	Picea abies	*Pinaceae*	T
	Citrus unshiu	*Rutaceae*	E
9	*Prunus persica*	*Rosaceae*	E
	Prunus avium	*Rosaceae*	E
	Prunus cerasus	*Rosaceae*	E

Table 3.1. *Classification of 216 plants.* **(Cont.)**

Group	Plant	Family	Use
10	*Pisum sativum*	*Fabaceae*	E
	Lathyrus odoratus	*Fabaceae*	L
	Allium cepa	*Amaryllidaceae*	E
11	*Linum usitatissimum*	*Linaceae*	T
	Vicia faba	*Fabaceae*	E
	Carthamus tinctorius	*Asteraceae*	M
12	*Phaseolus lunatus*	*Fabaceae*	E
	Phaseolus vulgaris	*Fabaceae*	E
	Phaseolus coccineus	*Fabaceae*	E
13	*Triticum aestivum*	*Poaceae*	E
	Zea mays	*Poaceae*	E
	Spinacia oleracea	*Amaranthaceae*	E
14	*Raphanus sativus*	*Brassicaceae*	E
	Brassica napus	*Brassicaceae*	P
	Malus domestica	*Rosaceae*	E
15	*Hordeum vulgare*	*Poaceae*	E
	Oryza sativa	*Poaceae*	E
	Cucumis sativus	*Cucurbitaceae*	E
	Glycine max	*Fabaceae*	E
	Helianthus annuus	*Asteraceae*	E
16	*Eriobotrya japonica*	*Rosaceae*	E
	Cassia fistula	*Fabaceae*	M
	Aesculus hippocastanum	*Hippocastanaceae*	P
	Camellia sinensis	*Theaceae*	E
	Rheum sp.	*Polygonaceae*	W

Table 3.1. *Classification of 216 plants.* **(Cont.)**

Group	Plant	Family	Use
17	*Robinia pseudoacacia*	*Fabaceae*	L
	Colophospermum mopane	*Fabaceae*	T
	Acacia mearnsii	*Fabaceae*	W
18	*Sinocrassula indica*	*Crassulaceae*	M
	Sedum sarmentosum	*Crassulaceae*	M
	Rhodiola sachalinensis	*Crassulaceae*	M
	Phyllanthus emblica	*Phyllanthaceae*	M/E
	Psidium guajava	*Myrtaceae*	E
	Phellodendron amurense	*Rutaceae*	M
	Epimedium sagittatum	*Berberidaceae*	M
19	*Solanum lycopersicum*	*Solanaceae*	E
	Solanum tuberosum	*Solanaceae*	E
	Nicotiana tabacum	*Solanaceae*	M
20	*Capsicum annuum*	*Solanaceae*	E
	Petunia x hybrida	*Solanaceae*	L
	Daucus carota	*Apiaceae*	W
	Asclepias curassavica	*Apocynaceae*	L
	Humulus lupulus	*Cannabaceae*	M
	Cyperus rotundus	*Cyperaceae*	M
21	*Glycyrrhiza uralensis*	*Fabaceae*	M
	Glycyrrhiza aspera	*Fabaceae*	W
	Glycyrrhiza glabra	*Fabaceae*	M/E
	Glycyrrhiza inflata	*Fabaceae*	M

Table 3.1. *Classification of 216 plants.* (Cont.)

Group	Plant	Family	Use
22	*Lupinus luteus*	*Fabaceae*	W
	Lupinus albus	*Fabaceae*	E
	Derris scandens	*Fabaceae*	W
	Erythrina variegata	*Fabaceae*	L
	Erythrina senegalensis	*Fabaceae*	M
23	*Euchresta japonica*	*Fabaceae*	W
	Euchresta formosana	*Fabaceae*	W
	Sophora flavescens	*Fabaceae*	M
	Maackia amurensis	*Fabaceae*	L
	Sophora secundiflora	*Fabaceae*	W
	Daphniphyllum oldhami	*Daphniphyllaceae*	M
24	*Medicago sativa*	*Fabaceae*	E
	Clitoria ternatea	*Fabaceae*	E
	Trifolium pratense	*Fabaceae*	M
	Sophora japonica	*Fabaceae*	T
	Lespedeza homoloba	*Fabaceae*	W
	Melilotus messanensis	*Fabaceae*	W
	Glycyrrhiza pallidiflora	*Fabaceae*	W
	Dalbergia odorifera	*Fabaceae*	T

Table 3.1. *Classification of 216 plants.* (Cont.)

Group	Plant	Family	Use
25	*Corydalis claviculata*	*Papaveraceae*	W
	Papaver somniferum	*Papaveraceae*	M
	Corydalis solida	*Papaveraceae*	W
	Cocculus laurifolius	*Menispermaceae*	W
	Stephania cepharantha	*Menispermaceae*	W
	Stephania cepharantha	*Menispermaceae*	W
	Cocculus pendulus	*Menispermaceae*	W
	Annona cherimola	*Annonaceae*	E
	Xylopia parviflora	*Annonaceae*	W
26	*Brassica oleracea*	*Brassicaceae*	E
	Brassica rapa	*Brassicaceae*	E
	Armoracia lapathifolia	*Brassicaceae*	E
	Hesperis matronalis	*Brassicaceae*	L
27	*Alstonia macrophylla*	*Apocynaceae*	T
	Alstonia angustifolia	*Apocynaceae*	M
	Alstonia angustifolia var.latifolia	*Apocynaceae*	M
28	*Millettia pinnata*	*Fabaceae*	L
	Millettia pinnata	*Fabaceae*	L
	Neorautanenia amboensis	*Fabaceae*	W
	Tephrosia purpurea	*Fabaceae*	P
	Amorpha fruticosa	*Fabaceae*	L
	Piscidia erythrina	*Fabaceae*	T
29	*Gymnadenia conopsea*	*Orchidaceae*	M
	Bletilla striata	*Orchidaceae*	M

Table 3.1. *Classification of 216 plants.* (Cont.)

Group	Plant	Family	Use
30	*Taiwania cryptomerioides*	*Cupressaceae*	T
	Chamaecyparis formosensis	*Cupressaceae*	T
	Cryptomeria japonica	*Cupressaceae*	T
31	*Gutierrezia microcephala*	*Asteraceae*	P
	Saussurea lappa	*Asteraceae*	M
	Artemisia spp.	*Asteraceae*	W
	Citrus spp.	*Rutaceae*	E
	Citrus sudachi	*Rutaceae*	M
	Murraya paniculata	*Rutaceae*	M
	Cannabis sativa	*Cannabaceae*	M
	Iris domestica	*Iridaceae*	M
32	*Tabernaemontana coffeoides*	*Apocynaceae*	W
	Kopsia dasyrachis	*Apocynaceae*	W
	Catharanthus roseus	*Apocynaceae*	M
	Rauvolfia vomitoria	*Apocynaceae*	W
33	*Nardostachys chinensis*	*Caprifoliaceae*	W
	Acritopappus confertus	*Asteraceae*	W
	Isodon xerophilus	*Lamiaceae*	W
	Cynanchum sublanceolatum	*Apocynaceae*	W
	Caesalpinia crista	*Fabaceae*	T
	Murraya euchrestifolia	*Rutaceae*	W
	Curcuma zedoaria	*Zingiberaceae*	E
34	*Garcinia mangostana*	*Clusiaceae*	M/E
	Garcinia dulcis	*Clusiaceae*	W

Table 3.1. *Classification of 216 plants.* **(Cont.)**

Group	Plant	Family	Use
35	*Atalantia buxifolia*	*Rutaceae*	W
	Ruta graveolens	*Rutaceae*	M/E
	Clausena excavata	*Rutaceae*	W
	Angelica furcijuga	*Apiaceae*	M/E
36	*Andrographis paniculata*	*Acanthaceae*	M
	Scutellaria baicalensis	*Lamiaceae*	M
37	*Zanthoxylum simulans*	*Rutaceae*	M
	Zanthoxylum integrifoliolum	*Rutaceae*	W
38	*Magnolia denudata*	*Magnoliaceae*	M
	Magnolia officinalis	*Magnoliaceae*	M
	Aeschynanthus bracteatus	*Gesneriaceae*	W
39	*Broussonetia papyrifera*	*Moraceae*	E
	Morus alba	*Moraceae*	M/E
	Artocarpus communis	*Moraceae*	E
40	*Sinapis alba*	*Brassicaceae*	E
	Vachellia rigidula	*Fabaceae*	E
41	*Lycium chinense*	*Solanaceae*	M
	Mandragora autumnalis	*Solanaceae*	M
	Angelica sinensis	*Apiaceae*	M
42	*Cullen corylifolium*	*Fabaceae*	M
	Calophyllum inophyllum	*Calophyllaceae*	T
	Juniperus phoenicea	*Cupressaceae*	W

Table 3.1. *Classification of 216 plants.* (Cont.)

Group	Plant	Family	Use
43	*Taxus cuspidata*	*Taxaceae*	P
	Taxus brevifolia	*Taxaceae*	M
	Taxus baccata	*Taxaceae*	M
	Taxus wallichiana	*Taxaceae*	M
	Taxus chinensis	*Taxaceae*	M
	Taxus mairei	*Taxaceae*	M
	Taxus yunnanensis	*Taxaceae*	M
44	*Panax notoginseng*	*Araliaceae*	M
	Panax ginseng	*Araliaceae*	M
	Panax pseudo-ginseng var.notoginseng	*Araliaceae*	M
	Panax ginseng C.A.Meyer	*Araliaceae*	M
	Bupleurum rotundifolium	*Apiaceae*	M
	Beta vulgaris	*Amaranthaceae*	E
	Bellis perennis	*Asteraceae*	M/E
45	*Xylocarpus granatum*	*Meliaceae*	W
	Spiraea formosana	*Rosaceae*	W
	Hibiscus taiwanensis	*Malvaceae*	W
	Begonia nantoensis	*Begoniaceae*	W
	Alpinia blepharocalyx	*Zingiberaceae*	W
	Taraxacum formosanum	*Asteraceae*	W
46	*Aristolochia elegans*	*Aristolochiaceae*	L
	Aristolochia heterophylla	*Aristolochiaceae*	M

Table 3.1. *Classification of 216 plants.* **(Cont.)**

Group	Plant	Family	Use
47	*Artabotrys uncinatus*	*Annonaceae*	W
	Annona purpurea	*Annonaceae*	E
	Rubia yunnanensis	*Rubiaceae*	M
	Withania somnifera	*Solanaceae*	M
48	*Salvia officinalis*	*Lamiaceae*	M/E
	Orthosiphon stamineus	*Lamiaceae*	W
	Plantago major	*Plantaginaceae*	M
	Rehmannia glutinosa	*Rehmanniaceae*	M
	Olea europaea	*Oleaceae*	M/E
	Lonicera japonica	*Caprifoliaceae*	M
	Eleutherococcus senticosus	*Araliaceae*	M
	Diospyros kaki	*Ebenaceae*	E
	Punica granatum	*Lythraceae*	E
	Curcuma domestica	*Zingiberaceae*	M/E

3.4.3. Predicting Currently Unknown Plant-Metabolite Relations.

The species–metabolite relation data in the KNApSAcK Core DB were collected from previously published papers. Many more plant-metabolite relations will inevitably be discovered in the future. However, based on our study, we can predict some not yet known plant-metabolite relations. When several plants are included in the same cluster with our approach, it implies that those plants contain many metabolites that are either the same or different but structurally very similar. When several plants contain a different subset of a group of structurally similar metabolites and they are very close according to morphological taxonomy, we can assume that all those plants contain the union of the metabolites currently detected in them. The basis of this assumption is that similar metabolic pathways are expected to be active in plants within a given taxon group.

In our experiments, we found structurally similar metabolite groups of different sizes, large and small. However, the metabolites belonging to a smaller group are likely to be closely related along a certain metabolic pathway. Therefore, for predicting currently unknown plant-metabolite relations, we focused on only smaller metabolite groups and empirically considered the metabolite groups of size no more than eight.

In summary, we follow the following steps to improve prediction accuracy:

Step 1: We select a group of plants that are in the same cluster according to our approach and at the same time belong to the same genus or family. Let us call such a group *S*.

Step 2: We determine the set (*K*) of structurally similar metabolite groups of size no more than eight such that each metabolite group is associated with at least two plants in *S*.

Step 3: All the metabolites of a metabolite group in *K* are assigned to the plants in S which are associated with the group. This process is repeated for each group in *K*.

Based on known information, however, we exclude some metabolites that are mainly structure isomers from this prediction process because some isomers are usually produced by different pathways (Dewick, 2009; McMurry & Begley, 2005). We discuss this method with an example as follows.

Predicting metabolites for *Citrus* plants: Six *Citrus* plants (*Citrus limon*, *Citrus aurantifolia*, *Citrus paradisi*, *Citrus sinensis*, *Citrus reticulata* and *Citrus aurantium*) are considered an excellent group in our classification (Group 1 in Table 3.1, we call it group S), and belong to the same genus (*Citrus*). We extract the set K of metabolite groups (with size no more than eight) in which each metabolite group is associated with at least two plants in S. There is a total 58 such metabolite groups in K. For each metabolite group in K which is related to multiple plants, we can construct a plant-metabolite table. Table 3.2 is a plant-metabolite table for a given metabolite group that contains two metabolites: Limonene and Cyclohexane, and their association to six plants in S. In Table 3.2, “1” means that the metabolite is reported in the corresponding plant and “0” means that the metabolite is unreported in that plant. We treat all these unreported plant-metabolite relations as currently unknown but actual relations. We repeat this process for all 58 metabolite groups in K and obtain a list of unrecorded metabolites for the plants in S, which we show in Table

3.3. We also verify some predicted plant-metabolite relations in Table 3.3 by published literatures. The literatures which record these relations are appended behind the corresponding metabolites. It proves that following this method, we can predict some currently unrecorded metabolites and find some widespread medicinal species that can be substitutions of more endangered relatives currently being used (Saslis-Lagoudakis et al., 2011).

Not all the predicted metabolites might actually be produced in given plants because of the complexity of metabolic pathway evolution. On the contrary, many true relations could not be predicted due to the limitation of the incomplete data source. However, with developments in plant metabolomics, we may be able to add more plant-metabolite relations in our analysis in the future and produce better results. For other plant groups, we can also predict numerous of unrecorded metabolites. We list all the predicted plant-metabolite relations in Appendix A.

Table 3.2. *Reported plant-metabolite relations of 6 plants of genus Citrus with a given metabolite group (including 2 metabolites: Limonene, Cyclohexane). 1/0 indicates presence/absence of a metabolite in a plant.*

	Citrus limon	*Citrus aurantifolia*	*Citrus paradisi*	*Citrus sinensis*	*Citrus reticulata*	*Citrus aurantium*
Limonene	1	1	1	1	1	1
Cyclohexane	0	1	1	1	1	0

Table 3.3. *Predicted unrecorded metabolites for 6 Citrus plants, encompassing 38 plant-metabolite relations.*

Species	Predicted unrecorded metabolites
Citrus limon	Gibberellin A4; Methyl salicylate; Cyclohexane; o-Isopropenyl toluene; Jasmonic acid; 10'-Apoviolaxanthal; alpha-trans-Bergamotene
Citrus aurantifolia	Methyl salicylate; Citral; Benzeneacetaldehyde; o-Isopropenyl toluene; Methyl epijasmonate; Salvigenin
Citrus paradisi	Rhoifolin (Refaat et al., 2015); Isopropanol; Methyl salicylate; Citral (Njoroge et al., 2005); Benzeneacetaldehyde; o-Isopropenyl toluene
Citrus sinensis	Isoscutellarein 7,8-dimethyl ether; Isoscutellarein 7,8,4'-trimethyl ether; o-Isopropenyl toluene; Methyl epijasmonate; Salvigenin; Gibberellin A53; Violaxanthin (Roussos, 2016)
Citrus reticulata	Gibberellin A81; Gibberellin A9 ; Isopropanol; Citral; 6-Demethoxytangeritin; Tetramethylscutellarein
Citrus aurantium	Apigenin 7-rutinoside; Methyl salicylate; Salvigenin; Cyclohexane; Benzeneacetaldehyde (Najafian & Rowshan, 2012); o-Isopropenyl toluene

3.4.4. Relationship between Metabolite-Content and Uses of Plants.

Our unsupervised approach for classifying plants is based on MC-similarity using hierarchical clustering. Our results substantially match those of traditional morphology-based taxonomy. However, our results further reflect the usage patterns of plants.

The metabolite-content of plants is always related to their bioactive properties, and the similarity of the metabolite-content of plants can reveal their bioactive similarity. Generally, medicinal properties are not randomly distributed in different classes of plants. Some plant classes are represented by more medicinal plants than others. It is suggested that there is a phylogenetic pattern in medicinal properties even within one genus (Saslis-Lagoudakis et al., 2011; Rønsted et al., 2012; Ernst et al., 2016). A similar distribution could also be observed in our classification that plants with certain uses are concentrated in the same group. Many plant groups in our classification are of similar usage patterns. A plant is frequently related to multiple uses, but we only consider the most common use in this paper. We collected all the plant resource information from published literature and online sources, and annotated plants by their uses such as medicinal, edible, ornamental, forestry, poisonous, and timber. Table 3.1 lists the usage patterns of 216 plants. The economic uses of plants are represented by different letters (E: edible, M: medicinal, L: landscaping, including forestry and ornamental plants, T: timber, P: poisonous, W: wild plants that are not yet widely used by humans). Eleven groups (ID: 1, 9, 10, 12, 13, 14, 15, 19, 26, 39, 40) involving 38 plants mostly consist of edible plants, and 14 groups (ID: 2, 4, 6, 18, 21, 27, 29, 31, 36, 38, 41, 43, 44, 48) involving 69 plants mostly consist of medicinal plants. Moreover, 3 groups (ID: 8, 17, 30) involving 10 plants mostly consist of landscaping or timber plants. This implies that the proposed classification approach of plants is consistent with their economic uses.

In this section, we investigate the relations between usage patterns and metabolite-content of plants using a supervised classification technique. We considered every metabolite group as a pathway pattern such that each group can be used as a feature for classifying plants by their uses. For this analysis, we considered 48 edible plants (E), 81 medicinal plants (M), 14 timber plants (T), 14 landscaping plants (including forestry and ornamental plants), and 5 poisonous plants (P). We considered the plants that have both edible and medicinal uses (plants with “M/E” in Table 3.1) as medicinal plants. We applied an SVM algorithm to classify the plants, using economic uses of plants as labels and corresponding metabolite groups as features. Classification performance was evaluated from the resulting confusion matrix, as shown in Table 3.4. The rows of the confusion matrix indicate documented uses of plants and columns indicate the predicted uses from the SVM algorithm. *Recognition rate* is the proportion of correctly predicted plants corresponding to a class.

Table 3.4. *Resulting confusion matrix from support vector machine (SVM) algorithm.162 plants are labeled as edible (E), medicinal (M), timber (T), landscaping (L), and poisonous (P), and SVM model was constructed to classify them.*

	M	E	T	L	P	Recognition rate [%]
M	81	0	0	0	0	100
E	1	47	0	0	0	97.9
T	6	0	8	0	0	57.2
L	8	1	0	5	0	35.7
P	4	1	0	0	0	0

Total: 162 plants, Accuracy: 87.0%

We found that all the medicinal plants and all but one edible plant were classified correctly. This implies that the metabolite-content of medicinal and edible plants substantially differs. However, half the timber and landscaping plants were classified as medicinal plants. Therefore, timber and landscaping plants are somewhat related to medicinal plants in terms of metabolite-content. All the poisonous plants were classified incorrectly: four plants were classified as medicinal plants and one as edible. This implies that poisonous plants are more similar to medicinal plants. Many poisonous plants can be used in treating specific diseases if the doses are carefully controlled (Tamilselvan et al., 2014). In summary, edible plants represent exclusive metabolite-content and can be differently classified from inedible plants. Furthermore, MC-based classification also reveals the predictive power of medicinal properties in bioprospecting. This indicates that our proposed approach can be used for exploring nutritional or medicinal properties of plants.

3.5. Summary

We proposed an approach for comparing the metabolite-content of plants and classifying plants by their metabolite-content similarity. We showed that with this approach we can classify plants similar to the traditional morphology-based plant taxonomy. Naturally, this work can be generalized from various perspectives. First, our approach can be regarded as a novel chemosystematics method that can be used to consider the global metabolite-contents of plants instead of a group of metabolites as done in previous research. The resulting classification is consistent with natural phylogenetic and chemosystematics patterns of plants. Some deviations in our classification from the NCBI taxonomy can be explained in terms of bioactive similarity. Moreover, the complexity and known extent of metabolite-content varies for different plants. We found that the Simpson coefficient can minimize the effect of the unequal size of the metabolite-content of organisms and performs better in comparing metabolite-content of plants than the other coefficients.

Comparing with the classification approach of Chapter 2, this approach took into account the relation of metabolites within the same pathway (metabolite groups), and could further reduce the influence of missing data. The time efficiency was also improved due to the integration of metabolite-content data. We also described a method for predicting unrecorded metabolites by structurally similar metabolite groups and phylogenetic relation of plants. With this method, we can predict some unrecorded metabolites and find new medicinal/edible plants from wild plants that have not been used by humans. Moreover, we studied the relation between the metabolite-contents of plants and their economic uses. We found that edible and medicinal plants represent unique metabolic pathway patterns and can be classified with an SVM algorithm with our integrated metabolite-content data. Our proposed MC-based plant-classification approach reveals the predictive power of medicinal properties in bioprospecting. The performance of this approach depends on the completeness of the metabolite-content data we use because metabolite groups, which were regarded as metabolic pathway patterns in our research, have been extracted from the background network of metabolites by using the DPClus algorithm. Therefore, if we can add more plant-metabolite relations, we can classify metabolites and species more accurately. Also, metabolites along identical pathways always correspond to high structural similarity. Our approach will be useful for predicting metabolic pathways in plants.

Chapter 4

Metabolite-Content-Guided Prediction of Medicinal /Edible Properties in Plants for Bioprospecting

4.1. Background

Plants are the major contributors of natural products and are usually rich in nutritional or medicinal properties, which is attributed to the complex secondary metabolite constituents of them. (Dahanukar et al., 2000; Veeresham, 2012; Cseke et al., 2016). Plants are an important source of novel pharmacologically active compounds with many pharmaceutical drugs have been derived directly or indirectly from plants, and have played a central role in human health-care since ancient times (Newman et al., 2000; Cragg & Newman, 2013; Fabricant & Farnsworth, 2001). Many pharmaceutical drugs are derived from plants that were first used in traditional systems of medicine (Fabricant & Farnsworth, 2001). According to the World Health Organization, about 25% of medicines are plant-derived (Veeresham, 2012).

Discoveries of novel molecules and advances in production of plant-based products have revived interest in natural product research (Paradise et al., 2006; Graham et al., 2010). The number of traditionally used plant species worldwide is estimated to be between 10,000 and 53,000 (McChesney et al., 2007); however, only a small proportion have been screened for biological activity (Soejarto et al., 2005; Gurib-Fakim, 2006), and the plants from some regions are less studied than others. Moreover, the potential of plants to yield new valuable drugs is under threat due to the alarming bio-diversity loss, with recent estimates indicating that every fifth plant species on earth is threatened with extinction (Brummitt & Bachman, 2010). Therefore, there is an urgent need for a time-efficient and systematic approach for unlocking the potential of plants in drug discovery.

A correlation between phylogeny and biosynthetic pathways could offer a predictive approach enabling more efficient selection of plants for drug discovery. Following the assumption that plant-derived chemicals are constrained to evolutionary plant lineages, phylogeny-guided approaches have been seen as one of the time-efficient and informed approaches to plant-based drug discovery (Rønsted et al., 2012; Ernst et al., 2016). A series of studies have been conducted and verified that phylogeny is an efficient tool to facilitate drug discovery for diverse genera across different regions or cultures (Rønsted et al., 2008; Saslis-Lagoudakis et al., 2011; Saslis-Lagoudakis et al., 2012; Rønsted et al., 2012; Yessoufou et al., 2015; Ernst et al., 2016). However, most of these studies mainly focused on a small cluster of genera, which limits its practical application due to the limitation of incomplete sequence data. Phylogenetic distance correlated to feature similarity of species will also be invalid once beyond a certain threshold (Kelly et al., 2014). Therefore, a special perspective different from molecular biology is valuable for understanding the evolution of bioactive features and facilitating prediction and discovery of medicinal properties in plants.

Besides molecular biology which is in view of nucleotide sequence comparison, metabolite feature is also closely related to the evolution of pathways for both primary and secondary metabolites. The secondary metabolite constitutes of a plants is highly related to its pathways which are constrained to evolutionary phylogeny, and also related to the bioactive compounds of the plant which determine the medicinal and nutritional features of the plant (Wink, 2003). The systemization of plants on the basis of their chemical constituents, which is also known as plant chemosystematics, could be helpful in solving selected taxonomical problems and exploring nutritional and medicinal properties in plants. Traditional chemosystematics of plants is based on the presence of selected metabolites (Singh, 2016). The incomplete data of metabolite constituents of plants limits its ability to solve taxonomical problems and discovery of new natural products or medicinal properties of plants (Singh, 2016; Wink, 2003). Comparative classification of plants based on the MC-similarity of them could facilitate the exploration of evolution and bioactivity relations between plants.

Here, we investigate the phylogenetic value of metabolite-content data, especially the predictive power of MC-similarity in exploration of medicinal and edible plants in bioprospecting, using the KNApSAcK Core DB as metabolite-content data source. In this thesis, we reconstructed the phylogenetic tree for a set of plants which are distributed in different genera and families by their metabolite-content data obtained from KNApSAcK Core DB, using the approach described in chapter 3. We also reconstruct the phylogenetic tree based on common DNA barcodes for the same set of plants, to investigate the predictive power of these two approaches, sequence-similarity- and MC-similarity-based method, in guiding the prediction of medicinal/edible plants in bioprospecting.

4.2. Datasets and Preliminaries

The input metabolite-content data are species-metabolite relationships obtained from the KNApSAcK Core DB, which is a part of the KNApSAcK Family DB (Afendi et al., 2012). We removed the plants with inadequate plant-metabolite relations to guarantee that the amount of metabolite-content of selected plants is sufficient enough to reveal their interrelations. The KNApSAcK Core DB also provides MOL molecular structure files for the metabolite compounds. We used R package ChemmineR (v2.26.0) to generate atom pair fingerprints from molecular structure description files (Cao et al., 2008). And these molecular fingerprints were used to measure the structural similarity for all the metabolite pairs.

In this study, we also reconstructed phylogenetic tree for the same plant samples we used previously based on three common DNA barcodes: two chloroplast barcodes *rbcL* (Ribulose-1,5-bisphosphate carboxylase/oxygenase) and *matK* (Maturase K), and one nuclear barcode ITS2 (internal transcribed spacer 2). The DNA sequence data are collected from GenBank (Benson et al., 2012), and certainly there is lack of data for some plants. Here we select the plants with both abundant metabolite-content data (no less than 30 metabolites) and corresponding DNA barcode data as samples. There are 190 plants in total belong to 51 different families, with 172 plants in *rbcL* group, 165 plants in *matK* group and 160 plants in *ITS2* group. The venn diagram of these three groups is shown in Figure 4.1.

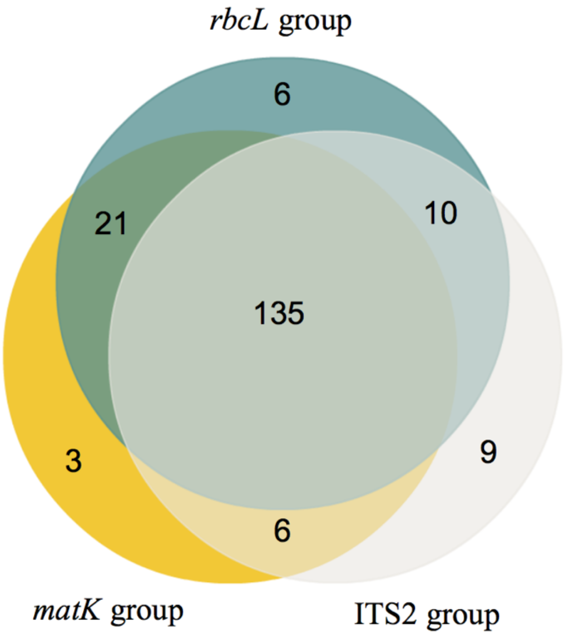


Figure 4.1. *Overview of 190 plants included in rbcL, matK and ITS2 sample groups.*

4.3. Methods

4.3.1. Phylogenetic Reconstruction

In this study, we produce phylogenetic hypothesis for each groups of samples by compiling DNA sequence data from the plastid markers *rbcL*, *matK* and nuclear marker *ITS2* respectively. Species names and corresponding GenBank accessions are listed in Appendix B. The sequence data of *rbcL*, *matK* and ITS2 are aligned by Clustal X 2.0 to compensate the missing and gapping data. Bayesian analyses of each sample groups were performed with MrBayes v3.2 (Larkin et al., 2007; Huelsenbeck & Ronquist, 2001). We produced Bayesian phylogenetic hypothesis using the $\mathrm{GTR} + \mathrm{I} + \Gamma$ model (Parameters: lset NST = 6 RATES = gamma). For each group we perform the analysis with more than 1,000,000 generations. The average standard deviation of the split frequencies (i.e., the average of all standard deviations of all observed splits between two independent analyses from different random trees) is down to <0.05 after the analysis is finished.

4.3.2. Phylogenetic and Statistical Analyses

We assess the relationship of phylogeny with the medicinal/edible properties by calculating the phylogenetic signal of medicinal/edible plants. We investigate the strength in phylogenetic signal of medicinal/edible plants using the D statistic, a measure of phylogenetic signal, implemented by the function *phylo.d* in the R package *caper* (Fritz & Purvis, 2010; Orme, 2013). D is calculated as follows,

$$D = \frac{\sum d_{obs} - mean(\sum d_b)}{mean(\sum d_r) - mean(\sum d_b)} \quad (4.1)$$

where $\sum d_{obs}$ is the observed number of changes in the binary trait (medicinal/edible properties) across the ultrametric phylogeny, $mean(\sum d_r)$ is the mean number of changes generated from 1000 random permutations of the species values at the tips of the phylogeny, and $mean(\sum d_b)$ is the mean number of changes generated from 1000 simulations of the evolution for the character by a Brownian motion model of evolution with likelihood of change being specified as that which produces the same number of tip species with each character state as the observed pattern.

The D statistic generates a value that usually lies between 0 (indicates the trait is highly correlated with phylogeny) and 1 (indicates the trait has evolved in essentially a random manner). Two p-values are calculated for the D statistic: $p(D < 1)$ indicates whether the D metric is significantly smaller than 1, meaning that the trait (medicinal/edible properties) is not randomly distributed over the phylogeny. The second p-value, $p(D > 0)$ indicates whether the D metric is significantly greater than 0, meaning that the trait (medicinal/edible properties) has a significantly different distribution on the phylogeny from the standard Brownian model of evolution. The phylogenetic signal is considered strong if $p(D < 1) < 0.05$ and $p(D > 0) > 0.05$.

4.3.3. Evolutionary Patterns of Medicinal/Edible Properties

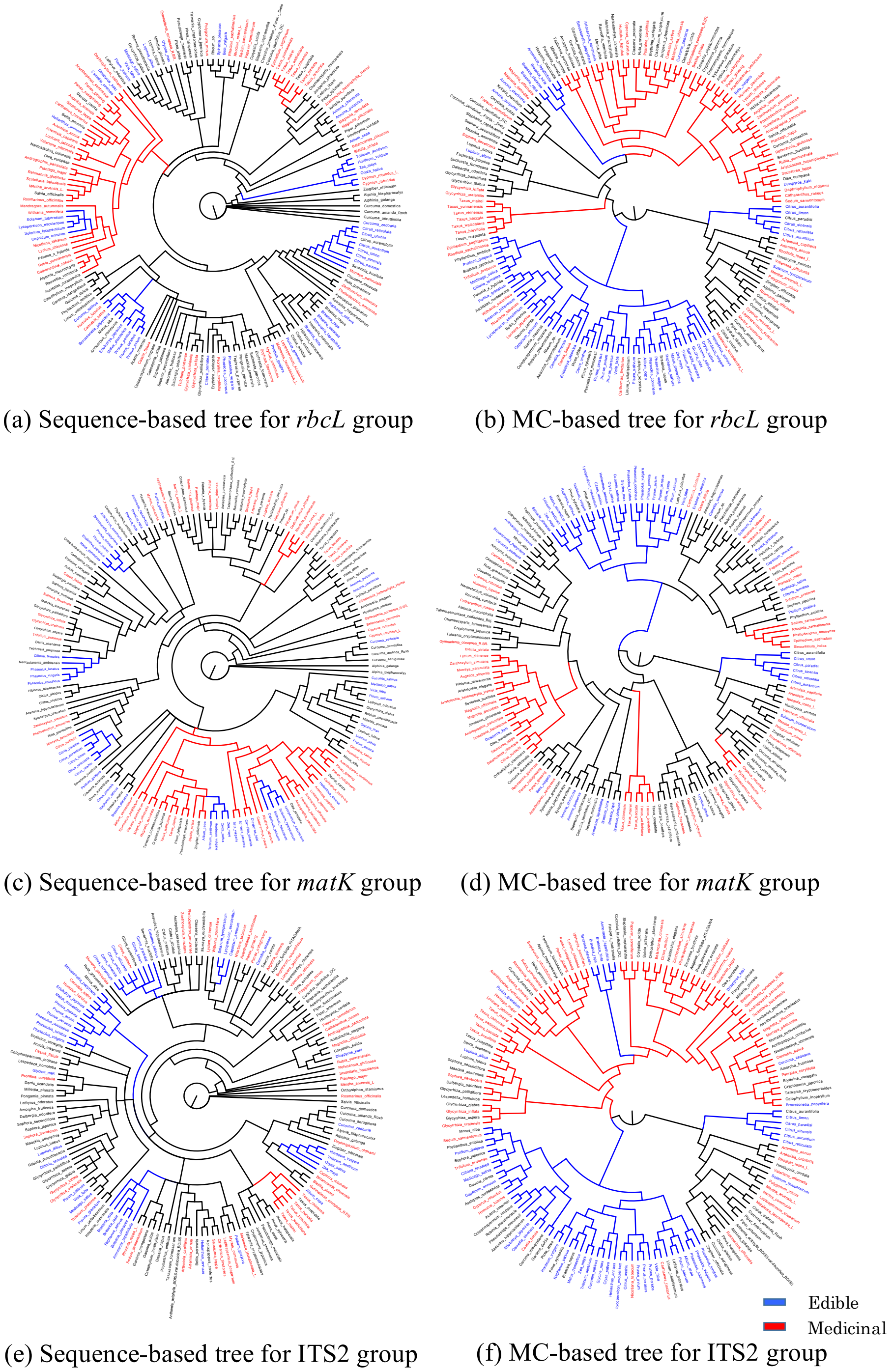
To narrow down the number of species chosen for an early stage medicinal/edible plants discovery screening, we identified the position of phylogeny clustering for medicinal/edible properties. We highlight such hot nodes (nodes that encompass significantly more medicinal/edible plants than the rest of the tree) by using the "nodesig" command in PHYLOCOM v4.2 for all of the phylogenetic trees (Webb et al., 2008). This option was used to determine the position of phylogenetic clustering in a community sample by testing each node of the phylogenetic tree for overabundance in medicinal/edible terminal taxa distal to it. Observed patterns for each phylogenetic trees were compared with those for random samples of the same size per case, drawn from the phylogeny.

For these hot nodes in each of the phylogenetic tree we obtained, we recorded the percentage of the total and medicinal properties included in them. We compared the observed number of medicinal/edible plants encompassed in the hot nodes to the one expected to be found randomly in the percentage of the plants encompassed in the hot nodes; this was the gain in percentage of medicinal hits compared with random.

4.4. Results and Discussion

All of the sequence data were downloaded from GenBank (https://www.ncbi.nlm.nih.gov/genbank/). The GenBank accession numbers and uses of species are listed in Appendix B. It should be noted that not all samples have complete sequence data. The amount of complete and partial sequences data of each groups are shown in Table 4.1. The ubiquitous missing and incomplete sequence data indicates that now the sequence data of plants included in GenBank are far from covering most of the plants, especially wild plants that not have been fully explored by people. The KNApSAcK species-metabolite relation database is also far from complete with a large amount of data fragmentation. However, the plants with abundant metabolite data included in KNApSAcK database are frequently inconsistent with plants with complete sequence data included in GenBank. The metabolite-content data of plants in KNApSAcK could be seen as a necessary supplement of gene data in GenBank for facilitating the analysis of evolutionary relation between plants and guiding the prediction of medicinal/edible plants since the plants covered by these two database are complementary to each other. The plant samples selected in our research are performing both adequate sequence and metabolite-content data with acceptable data missing. Thus we could investigate the effect of these two types of data in extracting medicinal/edible patterns from the same plant samples.

The phylogenetic trees of the three sample groups reconstructed by corresponding gene data and metabolite-content data are shown in Figure 4.2. The useful information of plants was collected from published literature and online sources, and annotated as seven categories: edible plants, medicinal plants, medicinal/edible multi-useful plants (M/E), landscaping plants, timber plants, poisonous plants and wild plants. The amount of plants with each usage category is listed in table 4.2.


(a) Sequence-based tree for *rbcL* group

(b) MC-based tree for *rbcL* group

(c) Sequence-based tree for *matK* group

(d) MC-based tree for *matK* group

(e) Sequence-based tree for ITS2 group

(f) MC-based tree for ITS2 group

Figure 4.2. *Phylogenetic trees and the hot nodes of medicinal/edible features for sequence- and MC-based approaches.*

Table 4.1. *The amount of complete and partial sequences data of rbcL, matK and ITS2 sample groups.*

	rbcL	*matK*	ITS2
Null	18	25	30
Complete sequence	73	112	131
Partial sequence	99	53	29

Table 4.2. *The amount of plants in each category of uses.*

Edible	Medicinal	M/E	Wild	Lanscaping	Timber	Poisonous
47	60	15	38	13	13	4

4.4.1. Phylogenetic Signal of Medicinal and Edible Plants

We investigated the strength in phylogenetic signal of medicinal and edible categories for each phylogenetic trees we obtained using the D statistic, which is shown in Table 4.3. We found that plants with medicinal/edible uses are significantly phylogenetically clustered in MC-based phylogenetic trees for all the three sample groups. The *rbcL*- and *matK*-based trees also show moderate phylogenetic signal for medicinal/edible plants but much weaker than that in MC-based trees. The ITS2-based tree shows weak phylogenetic signal for both medicinal and edible plants.

Table 4.3. *Phylogenetic signal of medicinal/edible features in sequence-based and metabolite-content-based trees.*

Phylogenetic tree	feature	D estimate	P(D<1)	P(D>0)
rbcL group (sequence)	Edible	0.234~0.355	0	0.026~0.126
	Medicinal	0.341~0.427	0	0.004~0.042
rbcL group (MC)	Edible	-0.053~0.002	0	0.535~0.6
	Medicinal	0.165~0.212	0	0.253~0.323
matK group (sequence)	Edible	0.197~0.274	0	0.093~0.184
	Medicinal	0.433~0.519	0	0.001~0.022
matK group (MC)	Edible	-0.206~-0.158	0	0.682~0.752
	Medicinal	-0.045~0.001	0	0.517~0.580
ITS2 group (sequence)	Edible	0.214~0.326	0	0.051~0.160
	Medicinal	0.470~0.604	0~0.002	0~0.006
ITS2 group (MC)	Edible	-0.118~-0.049	0	0.584~0.663
	Medicinal	0.354-0.391	0~0.003	0.091~0.151
MC-based tree for 1047 plants	Edible	0.768~0.773	0	0
	Medicinal	0.906~0.910	0.001~0.005	0

Generally, the edible plants are more significantly phylogenetically clustered than medicinal plants in all the three sample groups for both of the two approaches, with lower D estimate values and higher P(D>0) values. This suggests that comparing with edible plants, the distribution of medicinal plants reveal some but less phylogenetic relations. This observation could be explained as the different chemical constituent of edible and medicinal plants. The edible plants tend to contain massive primary metabolites, or the intermediate between primary and secondary metabolism (i.e., Gibberellin), which perform more stable expression during the evolution. The gene expression mechanism of medicinal plants is much subtler than edible plants and is related to the expression of small secondary metabolites which are sometimes randomly distributed along the clades. Thus we might found more phylogenetic patterns for medicinal plants by skipping gene data and comparing metabolite-content directly. Considering the genome data available from GenBank is usually incomplete, the metabolite-content data could be regarded as a novel approach for predicting medicinal properties.

4.4.2. Hot Nodes of Medicinal and Edible Plants

As a tentative approach to narrow down the number of medicinal/edible plants selected for bioactivity screening, we also identified the hot nodes that are significantly overrepresented by species of medicinal/edible uses. Phylogenetic clustering was found for edible and medicinal plants in all of the tested phylogenetic trees except ITS2 sequence based tree. The hot nodes in MC-based phylogenetic trees tend to encompass more medicinal and edible plants than sequence-based phylogenetic trees. This suggests that comparing with sequence-based approach it is more effective to explore phylogenetic patterns for medicinal and edible plants with the MC-based approach. We also compare the observed patterns for edible and medicinal plants with those for random samples of the same size drawn from the phylogeny. For these hot nodes in each of the tested phylogenetic trees, we recorded the percentage of edible and medicinal plants included in them. We compared the observed number of medicinal/edible plants encompassed in the hot nodes to the one expected to be found randomly in the percentage of the plants encompassed in the hot nodes, and this was the gain in percentage of medicinal/edible hits compared with random. (Table 4.4)

Table 4.4. *The number and proportion of medicinal/edible plants within the clades of hot nodes. Total plants included (%): The number (percentage) of the total plants included in the hot nodes of medicinal/edible uses. Medicinal/edible Hits (%): The number (percentage) of the medicinal/edible plants included in the hot nodes of medicinal/edible uses. Gain in medicinal/edible hits: the percentage of gain in medicinal/edible plants included in hot nodes, compare with what would be expected by chance. Co-included plants (hits): the number of (medicinal/edible hits) plants included in the hot nodes of medicinal/edible uses for both of the sequence- and MC-based phylogenetic trees.*

Phylogenetic tree	feature	Total plants included (%)	Medicinal/edible Hits (%)	Gain in edible/ medicinal hits	Co-included plants (hits)
rbcL group (sequence)	Edible	30 (17.4%)	20 (43.5%)	150%	Edible: 20 (18) Medicinal: 27 (20)
	Medicinal	46 (26.7%)	29 (50.9%)	90.6%	
rbcL group (MC)	Edible	64 (37.2%)	37 (80.4%)	116.1%	
	Medicinal	64 (37.2%)	32 (56.1%)	50.8%	
matK group (sequence)	Edible	23 (13.9%)	21 (44.7%)	221.6%	Edible: 16 (16) Medicinal: 12 (10)
	Medicinal	44 (26.7%)	23 (42.6%)	59.7%	
matK group (MC)	Edible	32 (19.4%)	26 (55.3%)	185.1%	
	Medicinal	34 (20.6%)	25 (46.3%)	124.7%	
ITS2 group (sequence)	Edible	35 (21.9%)	27 (65.0%)	196.8%	Edible: 30 (25) Medicinal: 5 (5)
	Medicinal	5 (3.1%)	5 (9.6%)	207.7%	
ITS2 group (MC)	Edible	61 (38.1%)	35 (85.4%)	124.1%	
	Medicinal	82 (51.2%)	35 (67.3%)	31.4 %	

The phylogenetic distribution of edible and medicinal plants encompassed by hot nodes also shows that the edible plants perform more converge trends and gains in percentage of hits. This indicates that the edible features of plants are more closely associated with the phylogeny as well as the MC-similarity of plants, and also suggests that there maybe many unexplored medicinal properties within the plant kingdoms. Moreover, we also investigated the coincidence rates of the medicinal/edible plants encompassed by hot nodes between the sequence-based and MC-based phylogenetic trees. We found that there is not significantly coincidence of medicinal/edible plants encompassed by hot nodes of these two types of phylogenetic trees. In other words, the medicinal/edible patterns identified by MC-similarity shows no significant similarity to the medicinal/edible patterns identified by sequence-based approach.

Our findings thus indicate that the MC-based approach might highlighted different group of medicinal/edible plants with phylogeny approach, and might reflect more unexplored medicinal/edible potential not associated with the genome-sequence similarity.

4.4.3. Predicting of Medicinal Properties by Metabolite-Content-Similarity

As a meaningful attempt, we imported more plant-metabolite relation data (28123 plant-metabolite relations associated with 1047 plants) and reconstructed phylogenetic tree by MC-similarity (Figure 4.3). We selected plants containing at least 14 metabolites to ensure data integrity. Plant usage information (edible or medicinal uses) was imported from KNApSAcK WorldMap DB (Afendi et al., 2012). For the total 1047 tested plants, we found medicinal or edible uses information of 605 plants from WorldMap DB, with 563 plants having medicinal values, 345 plants having edible values. There are totally 303 plants with both medicinal and edible values. The remaining 442 plants which are lack of usage information are regarded as wild plants from which we can explore new medicinal properties.

Both edible and medicinal plants in this phylogeny show weak phylogenetic signals (Table 4.3). The observed patterns of medicinal and edible plants are different from the random shuffle, but also different from the Brownian evolution model. That maybe due to the incomplete of metabolite-content data and plant uses information. However, from the hot nodes we can still find many phylogenetic patterns for medicinal plants (Figure 4.3). The hot nodes for medicinal plants encompass 288 plants in the MC-based phylogenetic tree, including 198 recorded medicinal plants. The remaining 90 wild plants encompassed by the hot nodes should be given priority for future screening for overall medicinal bioactivity because these plants perform highly MC-similarity with 198 medicinal plants. We list the 90 plants with high priority for future screening for overall medicinal bioactivity in Appendix C.

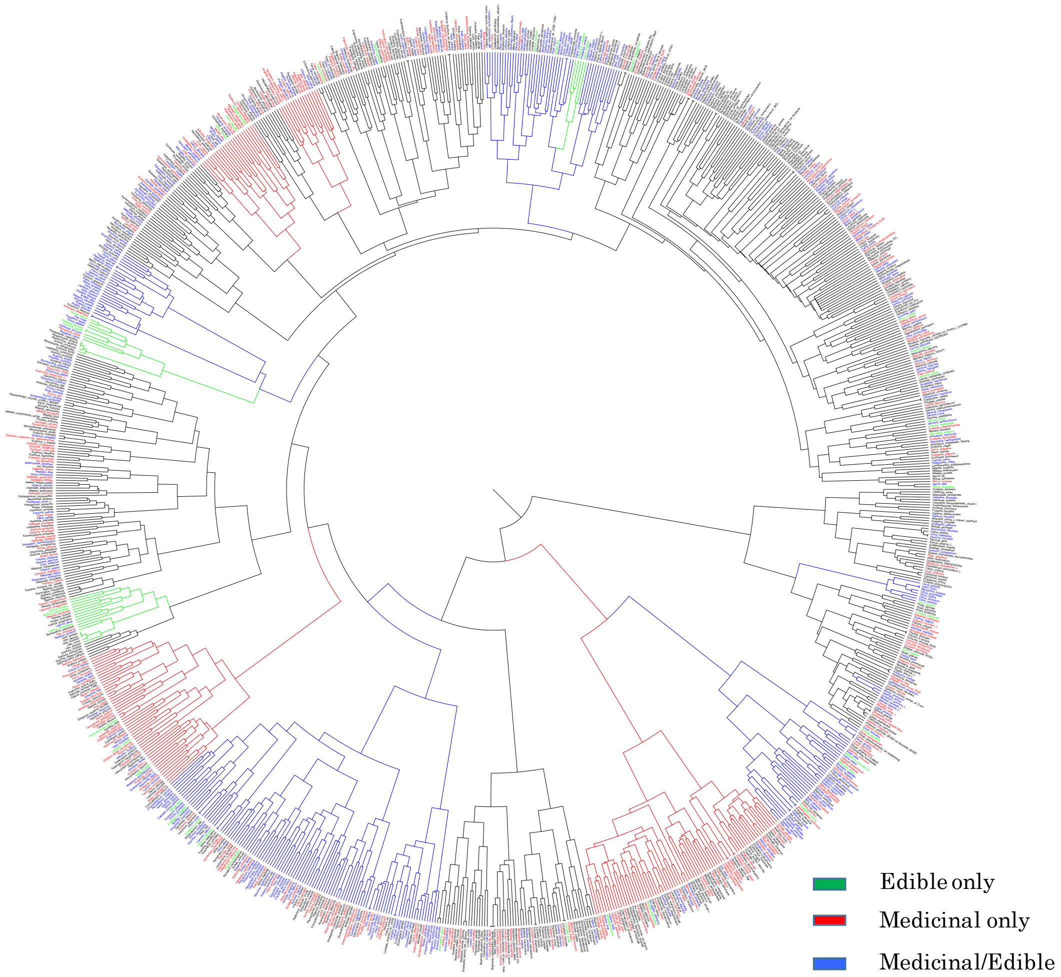


Figure 4.3. *MC-based phylogenetic tree for 1047 plants, with the hot nodes of medicinal/edible plants.*

4.5. Summary

Many researches have proved that medicinal and edible plants were derived mostly from some lineages, and tend to be clustered rather than scattered in the phylogenetic tree. Our study reveals that besides the genome sequence data, metabolite-content data is also closely associated with medicinal and edible bioactivity of plants and can be used to explore the medicinal/edible patterns in a different perspective from DNA sequence based plant phylogeny.

We found that our MC-based approach performs fair even better performance in predictive power of medicinal/edible properties comparing with DNA sequence based approach. Moreover, the hot nodes of MC-based phylogenetic tree highlight different medicinal/edible patterns comparing with DNA-sequence-based approach. This implies that MC-based approach could reflect unexplored medicinal/edible potential not recovered by the sequence-based approach.

Since phylogenetic analysis based plant bioprospecting is frequently confined to the lack of DNA sequence data, it is rational to utilize metabolite-content data to extent the limitation of phylogeny based bioprospecting. MC-based plant phylogeny reconstruction could provide a new perspective in plant bioprospecting, and the predictive power of metabolite-content data for medicinal/edible plants will also be improved with the improvement and completeness of metabolite-content database in future.

Chapter 5

Conclusion

Plants are the major contributors of natural products and are usually rich in nutritional or medicinal properties, which is attributed to the complex secondary metabolite constituent of them. The biological classification of plants, or the plant systematics, is one of the most ancient discipline which stretches from the traditional morphology based plant taxonomy to modern molecular phylogenetic. This study is conducted in order to investigate the systematic value of metabolite-content data of plants, and the relationships among metabolite-contents of plants and their biological activities.

In this dissertation, we utilized data-intensive science for conduction classification of plants based on the MC-similarity of them using KNApSAcK Core DB. We have proposed two approaches of classifying plants by their MC-similarity: (1) Classification of Plants based on Chemical Structure Similarity of Metabolite-Content. By this approach, we calculated the structural similarity of all metabolite pairs by Tanimoto coefficients (TCs), and determined the MC-similarity of plants based on the background population of TCs. We classified 102 plants into 28 groups by hierarchical clustering method. (2) Clustering Plants based on Structural-Similarity Network of Metabolites. By this approach, we applied a network based approach to abstract structurally similar metabolite groups as features, and measured the phylogenetic distance by a binary method. We classified 216 plants into 48 groups by hierarchical clustering method. We compared the resulted classifications of plants with NCBI taxonomy. The result proves that the MC-similarity of plants is associated to the pathway and bioactivity similarity, and can be regarded as a taxonomy marker which takes into account both general phylogenetic relations and the relations between plants based on bioactive features. Both these two methods have the ability to compensate for the limitations of missing data, and by the second method we can even predict some currently unknown plant-metabolite relations.

We also extended our finding by using phylogenetic statistic method to investigate the predictive power of MC-similarity in exploration of edible and medicinal plants for bioprospecting. We reconstructed the phylogenetic trees for the same set of plants based on MC-based approach and sequence-based approach. We then applied D statistic to test phylogenetic signal of medicinal/edible plants for the obtained phylogenetic trees and identified the hot nodes that were significantly overrepresented by plants of medicinal/edible uses. The result shows that comparing with sequence-based approach, plants with medicinal/edible uses are more significantly clustered in MC-based phylogenetic trees than sequence-based phylogenetic trees. The hot nodes in MC-based phylogenetic trees tend to encompass more medicinal/edible plants, and could highlighted different groups of medicinal/edible plants. We also imported plant-metabolite relation data and plant usage information from KNApSAcK Core DB and KNApSAcK WorldMap DB, and used this approach to predict some medicinal plants.

The performance of our approach depends on the completeness of the metabolite-content data we imported. In future, with the developments in plant metabolomics, we may be able to add more plant-metabolite relations in our analysis and produce better results. Moreover, MC-based plant phylogeny reconstruction could provide a new perspective in plant bioprospecting, and the predictive power of metabolite-content data for medicinal/edible plants will be improved with the improvement and completeness of metabolite-content database in future. Also, metabolites along identical pathways always correspond to high structural similarity. Our approaches will be useful for predicting metabolic pathways in plants.

Bibliography

Abdullah, A. A., Altaf-Ul-Amin, M., Ono, N., Sato, T., Sugiura, T., Morita, A. H., ... & Kanaya, S. (2015). Development and mining of a volatile organic compound database. *BioMed research international*, *2015*.

Afendi, F. M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K., ... & Saito, K. (2012). KNApSAcK family databases: integrated metabolite–plant species databases for multifaceted plant research. *Plant and Cell Physiology*, *53*(2), e1-e1.

Afendi, F. M., Ono, N., Nakamura, Y., Nakamura, K., Darusman, L. K., Kibinge, N., ... & Kanaya, S. (2013). Data mining methods for omics and knowledge of crude medicinal plants toward big data biology. *Computational and Structural Biotechnology Journal*, *4*(5), 1-14.

Agostini-costa, T.S. et al., 2012. Secondary Metabolites. Chromatography and Its Applications, 1, pp.131–164.

Altaf-Ul-Amin, M., Afendi, F. M., Kiboi, S. K., & Kanaya, S. (2014). Systems biology in the context of big data and networks. *BioMed research international*, *2014*.

Altaf-Ul-Amin, M., Tsuji, H., Kurokawa, K., Asahi, H., Shinbo, Y., & Kanaya, S. (2006). DPClus: a density-periphery based graph clustering software mainly focused on detection of protein complexes in interaction networks. *Journal of Computer Aided Chemistry*, *7*, 150-156.

Altaf-Ul-Amin, M., Wada, M., & Kanaya, S. (2012). Partitioning a PPI network into overlapping modules constrained by high-density and periphery tracking. *ISRN Biomathematics*, *2012*.

Baker, F. B. (1974). Stability of two hierarchical grouping techniques Case I: Sensitivity to data errors. *Journal of the American Statistical Association*, *69*(346), 440-445.

Bennett, R. N., & Wallsgrove, R. M. (1994). Secondary metabolites in plant defence mechanisms. *New phytologist*, *127*(4), 617-633.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic acids research*, *41*(D1), D36-D42.

Berg, G., Zachow, C., Müller, H., Philipps, J., & Tilcher, R. (2013). Next-generation bio-products sowing the seeds of success for sustainable agriculture.

Bérdy, J. (2012). Thoughts and facts about antibiotics: where we are now and where we are heading. *The Journal of antibiotics*, *65*(8), 385-395.

Besse, P. (2014). *Molecular plant taxonomy: methods and protocols* (No. 581.8 M718m). Humana Press,.

Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., ... & Trethewey, R. N. (2004). Potential of metabolomics as a functional genomics tool. *Trends in plant science*, *9*(9), 418-425.

Bolton, E. E., Wang, Y., Thiessen, P. A., & Bryant, S. H. (2008). PubChem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry*, *4*, 217-241.

Borenstein, E., Kupiec, M., Feldman, M. W., & Ruppin, E. (2008). Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proceedings of the National Academy of Sciences*, *105*(38), 14482-14487.

Brummitt, N. A., & Bachman, S. P. (2010). Plants under pressure—a global assessment: the first report of the IUCN sampled red list index for plants. *Kew, UK: Royal Botanic Gardens*.

Cao, Y., Charisi, A., Cheng, L. C., Jiang, T., & Girke, T. (2008). ChemmineR: a compound mining framework for R. *Bioinformatics*, *24*(15), 1733-1734.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), 27.

Chang, C. W., Lyu, P. C., & Arita, M. (2011). Reconstructing phylogeny from metabolic substrate-product relationships. *BMC bioinformatics*, *12*(1), S27.

Chen, X., & Reynolds, C. H. (2002). Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. Journal of chemical information and computer sciences, *42*(6), 1407-1414.

Choi, S. S., Cha, S. H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, *8*(1), 43-48.

Clemente, J. C., Satou, K., & Valiente, G. (2007). Phylogenetic reconstruction from non-genomic data. *Bioinformatics*, *23*(2), e110-e115.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Cragg, G. M., & Newman, D. J. (2013). Natural products: a continuing source of novel drug leads. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1830*(6), 3670-3695.

Cseke, L. J., Kirakosyan, A., Kaufman, P. B., Warber, S., Duke, J. A., & Brielmann, H. L. (2016). *Natural products from plants*. CRC press.

Dahanukar, S. A., Kulkarni, R. A., & Rege, N. N. (2000). Pharmacology of medicinal plants and natural products. *Indian journal of pharmacology*, *32*(4), S81-S118.

Davis, Jerrold I. (2014). Plant phylogeny. In *AccessScience*. McGraw-Hill Education. https://doi.org/10.1036/1097-8542.524400

Dewick, P. M., (2009). *Medicinal natural products: a biosynthetic approach*. John Wiley and Sons, Chichester, UK, 3rd edition, 2009.

Deyasi, K., Banerjee, A., & Deb, B. (2015). Phylogeny of metabolic networks: A spectral graph theoretical approach. *Journal of biosciences*, *40*(4), 799-808.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2005). Misc Functions of the Department of Statistics (e1071), TU Wien. *R package version*, 1-5.

Ernst, M., Saslis-Lagoudakis, C. H., Grace, O. M., Nilsson, N., Simonsen, H. T., Horn, J. W., & Rønsted, N. (2016). Evolutionary prediction of medicinal properties in the genus *Euphorbia L. Scientific reports*, *6*, 30531.

Fabricant, D. S., & Farnsworth, N. R. (2001). The value of plants used in traditional medicine for drug discovery. *Environmental health perspectives*, *109*(Suppl 1), 69.

Facchini, P. J., Bohlmann, J., Covello, P. S., De Luca, V., Mahadevan, R., Page, J. E., ... & Martin, V. J. (2012). Synthetic biosystems for the production of high-value plant metabolites. *Trends in biotechnology*, *30*(3), 127-131.

Fallaw, W. C. (1979). A test of the Simpson coefficient and other binary coefficients of faunal similarity. *Journal of Paleontology*, 1029-1034.

Fan, R. E., Chen, P. H., & Lin, C. J. (2005). Working set selection using second order information for training support vector machines. *Journal of machine learning research*, *6*(Dec), 1889-1918.

Federhen, S. (2011). The NCBI taxonomy database. *Nucleic acids research*, *40*(D1), D136-D143.

Fritz, S. A., & Purvis, A. (2010). Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conservation Biology*, *24*(4), 1042-1051.

Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, *31*(22), 3718-3720.

Gao, C., Ren, X., Mason, A. S., Liu, H., Xiao, M., Li, J., & Fu, D. (2014). Horizontal gene transfer in plants. *Functional & integrative genomics*, *14*(1), 23-29.

Godden, J. W., Xue, L., & Bajorath, J. (2000). Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *Journal of Chemical Information and Computer Sciences*, *40*(1), 163-166.

Graham, I. A., Besser, K., Blumer, S., Branigan, C. A., Czechowski, T., Elias, L., ... & Larson, T. R. (2010). The genetic map of *Artemisia annua* L. identifies loci affecting yield of the antimalarial drug artemisinin. *science*, *327*(5963), 328-331.

Gurib-Fakim, A. (2006). Medicinal plants: traditions of yesterday and drugs of tomorrow. *Molecular aspects of Medicine*, *27*(1), 1-93.

Hegnauer, R. (1967). Chemical characters in plant taxonomy: some possibilities and limitations. Pure and Applied Chemistry, *14*(1), 173-188.

Heywood, V. H. (2013). Chemosystematies—an artificial discipline. *Chemistry in Botanical Classification: Medicine and Natural Sciences: Medicine and Natural Sciences*, 41.

Hinchliff, C. E., & Smith, S. A. (2014). Some limitations of public sequence data for phylogenetic inference (in plants). *PloS one*, *9*(7), e98986.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754-755.

Ikeda, S., Abe, T., Nakamura, Y., Kibinge, N., Hirai Morita, A., Nakatani, A., ... & Kanaya, S. (2013). Systematization of the protein sequence diversity in enzymes related to secondary metabolic pathways in plants, in the context of big data biology inspired by the KNApSAcK Motorcycle database. *Plant and Cell Physiology*, *54*(5), 711-727.

Jeong, H., Mason, S. P., Barabási, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, *411*(6833), 41-42.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, *45*(D1), D353-D361.

Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, *9*(8).

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive science: a new paradigm for biodiversity studies. *BioScience*, *59*(7), 613-620.

Kelly, S., Grenyer, R., & Scotland, R. W. (2014). Phylogenetic trees do not reliably predict feature diversity. *Diversity and Distributions*, *20*(5), 600-612.

Kumar S, Tamura K, Nei M (1993) MEGA: Molecular Evolu- tionary Genetics Analysis. Version 1.0. The Pennsylvania State University, University Park, Pennsylvania

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., ... & Thompson, J. D. (2007). Clustal W and Clustal X version 2.0. *bioinformatics*, *23*(21), 2947-2948.

Ma, C. Y., Lin, S. H., Lee, C. C., Tang, C. Y., Berger, B., & Liao, C. S. (2013). Reconstruction of phyletic trees by global alignment of multiple metabolic networks. *BMC bioinformatics*, *14*(2), S12.

Ma, H. W., & Zeng, A. P. (2004). Phylogenetic comparison of metabolic capacities of organisms at genome level. *Molecular phylogenetics and evolution*, *31*(1), 204-213.

Mano, A., Tuller, T., Béjà, O., & Pinter, R. Y. (2010). Comparative classification of species and the study of pathway evolution based on the alignment of metabolic pathways. *BMC bioinformatics*, *11*(1), S38.

Macel, M., Van, D. A. M., Nicole, M., & KEURENTJES, J. J. (2010). Metabolomics: the chemistry between ecology and genetics. *Molecular ecology resources*, *10*(4), 583-593.

Martin, Y. C., Kofron, J. L., & Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity?. *Journal of medicinal chemistry*, *45*(19), 4350-4358.

Marx, V. (2013). Biology: The big challenges of big data. *Nature*, *498*(7453), 255-260.

McChesney, J. D., Venkataraman, S. K., & Henri, J. T. (2007). Plant natural products: back to the future or into extinction?. *Phytochemistry*, *68*(14), 2015-2022.

McMurry, J., & Begley, T. P. (2005). *The organic chemistry of biological pathways*, chapter 3, Roberts and Company Publishers, Engle- wood, Colo, USA.

Metzker, M. L. (2010). Sequencing technologies-the next generation. *Nature reviews genetics*, *11*(1).

Nakamura, Y., Mochamad Afendi, F., Kawsar Parvin, A., Ono, N., Tanaka, K., Hirai Morita, A., ... & Kanaya, S. (2014). KNApSAcK metabolite activity database for retrieving the relationships between metabolites and biological activities. *Plant and Cell Physiology*, *55*(1), e7-e7.

Newman, D. J., Cragg, G. M., & Snader, K. M. (2000). The influence of natural products upon drug discovery. *Natural product reports*, *17*(3), 215-234.

Nováková, J., & Farkašovský, M. (2013). Bioprospecting microbial metagenome for natural products. *Biologia*, *68*(6), 1079-1086.

Orme, D. (2013). The caper package: comparative analysis of phylogenetics and evolution in R. *R package version*, *5*(2), 1-36.

Ouyang, L., Luo, Y., Tian, M., Zhang, S. Y., Lu, R., Wang, J. H., ... & Li, X. (2014). Plant natural products: from traditional compounds to new emerging drugs in cancer therapy. *Cell proliferation*, *47*(6), 506-515.

Patterson, D. J., Cooper, J., Kirk, P. M., Pyle, R. L., & Remsen, D. P. (2010). Names are key to the big new biology. *Trends in ecology & evolution*, *25*(12), 686-691.

Pichersky, E., & Gang, D. R. (2000). Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. Trends in plant science, *5*(10), 439-445.

Reynolds, T. (2007). The evolution of chemosystematics. *Phytochemistry*, *68*(22), 2887-2895.

Ro, D. K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., ... & Chang, M. C. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, *440*(7086), 940-943.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., ... & Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, *61*(3), 539-542.

Rønsted, N., Savolainen, V., Mølgaard, P., & Jäger, A. K. (2008). Phylogenetic selection of Narcissus species for drug discovery. *Biochemical Systematics and Ecology*, *36*(5), 417-422.

Rønsted, N., Symonds, M. R., Birkholm, T., Christensen, S. B., Meerow, A. W., Molander, M., ... & Stafford, G. I. (2012). Can phylogeny predict chemical diversity and potential medicinal activity of plants? A case study of Amaryllidaceae. *BMC evolutionary biology*, *12*(1), 182.

Rouhan, G., & Gaudeul, M. (2014). Plant taxonomy: a historical perspective, current challenges, and perspectives. *Molecular Plant Taxonomy: Methods and Protocols*, 1-37.

Saito, K., & Matsuda, F. (2010). Metabolomics for functional genomics, systems biology, and biotechnology. *Annual review of plant biology*, *61*, 463-489.

Saslis-Lagoudakis, C. H., Klitgaard, B. B., Forest, F., Francis, L., Savolainen, V., Williamson, E. M., & Hawkins, J. A. (2011). The use of phylogeny to interpret cross-cultural patterns in plant use and guide medicinal plant discovery: an example from Pterocarpus (Leguminosae). *PloS one*, *6*(7), e22275.

Saslis-Lagoudakis, C. H., Savolainen, V., Williamson, E. M., Forest, F., Wagstaff, S. J., Baral, S. R., ... & Hawkins, J. A. (2012). Phylogenies reveal predictive power of traditional medicine in bioprospecting. *Proceedings of the National Academy of Sciences*, *109*(39), 15835-15840.

Shinbo, Y., Nakamura, Y., Altaf-Ul-Amin, M., Asahi, H., Kurokawa, K., Arita, M., ... & Kanaya, S. (2006). KNApSAcK: a comprehensive species-metabolite relationship database. *Plant metabolomics*, 165-181.

Singh, R. (2016). Chemotaxonomy: a tool for plant classification. *Journal of Medicinal Plants*, *4*(2), 90-93.

Soejarto, D. D., Fong, H. H. S., Tan, G. T., Zhang, H. J., Ma, C. Y., Franzblau, S. G., ... &

Xuan, L. T. (2005). Ethnobotany/ethnopharmacology and mass bioprospecting: Issues on intellectual property and benefit-sharing. *Journal of ethnopharmacology*, *100*(1), 15-22.

Soltis, P., Soltis, D.E., Doyle, J.J., 1992. Molecular Systematics of Plants. Chapman & Hall, London.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.

Strobel, G., & Daisy, B. (2003). Bioprospecting for microbial endophytes and their natural products. *Microbiology and molecular biology reviews*, *67*(4), 491-502.

Swofford, D. L. (2003). PAUP*: phylogenetic analysis using parsimony, version 4.0 b10.

Tamilselvan, N., Thirumalai, T., Shyamala, P., & David, E. (2014). A review on some poisonous plants and their medicinal values. *Journal of Acute Disease*, *3*(2), 85-89.

Tauler, R., & Walczak, B. (2009). *Comprehensive chemometrics*. S. D. Brown (Ed.). Elsevier Science.

Tohge, T., & Fernie, A. R. (2010). Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. *Nature protocols*, *5*(6), 1210-1227.

Veeresham, C. (2012). Natural products derived from plants as a source of drugs.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., & Bryant, S. H. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, *37*(suppl_2), W623-W633.

Webb, C. O., Ackerly, D. D., & Kembel, S. W. (2008). Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*, *24*(18), 2098-2100.

Wijaya, S. H., Husnawati, H., Afendi, F. M., Batubara, I., Darusman, L. K., Altaf-Ul-Amin, M., ... & Kanaya, S. (2014). Supervised clustering based on DPClusO: Prediction of plant-disease relations using Jamu formulas of KNApSAcK database. *BioMed research international*, *2014*.

Willett, P. (2014). The calculation of molecular structural similarity: principles and practice. Molecular Informatics, *33*(6-7), 403-413.

Wink, M. (2003). Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. Phytochemistry, *64*(1), 3-19.

Wink, M., & Witte, L. (1983). Evidence for a wide-spread occurrence of the genes of quinolizidine alkaloid biosynthesis: induction of alkaloid accumulation in cell suspension cultures of alkaloid-'free'species. *FEBS letters*, *159*(1-2), 196-200.

Woolhouse, M., & Farrar, J. (2014). Policy: An intergovernmental panel on antimicrobial resistance. *Nature*, *509*(7502), 555-557.

Yessoufou, K., Daru, B. H., & Muasya, A. M. (2015). Phylogenetic exploration of commonly used medicinal plants in South Africa. *Molecular ecology resources*, *15*(2), 405-413.

Njoroge, S. M., Koaze, H., Karanja, P. N., & Sawamura, M. (2005). Volatile constituents of redblush grapefruit (Citrus paradisi) and pummelo (Citrus grandis) peel essential oils from Kenya. *Journal of agricultural and food chemistry*, *53*(25), 9790-9794.

Najafian, S., & Rowshan, V. (2012). Comparative of HS SPME and HD techniques in Citrus aurantium L. *Int J Med Arom Plants*, *2*, 488-494.

Refaat, J., Desoukey, S. Y., Ramadan, M. A., & Kamel, M. S. (2015). Rhoifolin: A review of sources and biological activities. *Int J Pharmacognosy*, *2*(3), 102-09.

Roussos, P. A. (2016). Orange (Citrus sinensis (L.) Osbeck). In *Nutritional Composition of Fruit Cultivars* (pp. 469-496).

Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., & Willett, P. (2012). Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. *Journal of chemical information and modeling*, *52*(11), 2884-2901.

Achievements

Reviewed publications

1. Liu, K., Altaf-Ul-Amin, M., Abdullah, A. A., Morita, A. H., Shiraishi, M., & Kanaya, S. (2015, August). A novel plant classification method based on similarities in chemical structures of metabolite contents obtained from the KNApSAcK database. *ISHS Acta Horticulturae 1169* (pp. 139-150) (in Chapter 2).
2. Liu, K., Abdullah, A. A., Huang, M., Nishioka, T., Altaf-Ul-Amin, M., & Kanaya, S. (2017). Novel Approach to Classify Plants Based on Metabolite-Content Similarity. *BioMed research international*, *2017* (in Chapter 3).

Submited paper

1. Kang Liu, Aki H. Morita, Md. Altaf-Ul-Amin and Shigehiko Kanaya, Metabolite-content-guided prediction of medicinal/edible properties in plants for bioprospecting (in Chapter 4; submitted to *Molecular ecology resources* in January, 2018)

International conferences

1. Md. Altaf-Ul-Amin, Kang Liu, Azian Azamimi Abdullah, Aki H. Morita, Makio Shiraishi and Shigehiko Kanaya, 2016, Classification of plants based on chemical structure similarity of metabolite contents obtained from KNApSAcK database*, Applied bioinformatics in life sciences, VIB Conference Series* (17-18 March 2016, Leuven, Belgium) (in Chapter 2)
2. Kang Liu, Aki H. Morita, Shigehiko Kanaya, Md. Altaf-Ul-Amin, 2018, Metabolite-content-guided prediction of edible/medicinal properties in plant bioprospecting, *The Sixteenth Asia Pacific Bioinformatics Conference* (15-17 January 2018, Yokohama, Japan) (in Chapter 4)

Appendices

Appendix A

The predicted unrecorded plant-metabolite relations, involving 524 metabolites and 117 plants.

Metabolite	**Plant**
Gibberellin A4	*Citrus limon; Phaseolus lunatus; Nicotiana tabacum; Lupinus luteus*
Methyl salicylate	*Citrus limon; Citrus aurantifolia; Citrus paradisi; Citrus aurantium*
Cyclohexane	*Citrus limon; Citrus aurantium; Artemisia annua; Artemisia capillaris; Centaurea sessilis; Zingiber officinale*
o-Isopropenyl toluene	*Citrus limon; Citrus aurantifolia; Citrus paradisi; Citrus sinensis; Citrus aurantium*
Jasmonic acid	*Citrus limon*
10'-Apoviolaxanthal	*Citrus limon*
alpha-trans-Bergamotene	*Citrus limon*
Citral	*Citrus aurantifolia; Citrus paradisi; Citrus reticulata*
Benzeneacetaldehyde	*Citrus aurantifolia; Citrus paradisi; Citrus aurantium*
Methyl epijasmonate	*Citrus aurantifolia; Citrus sinensis*
Salvigenin	*Citrus aurantifolia; Citrus sinensis; Citrus aurantium*
Rhoifolin	*Citrus paradisi*
Isopropanol	*Citrus paradisi; Citrus reticulata*
Isoscutellarein 7,8-dimethyl ether	*Citrus sinensis*
Isoscutellarein 7,8,4'-trimethyl ether	*Citrus sinensis*
Gibberellin A53	*Citrus sinensis; Lathyrus odoratus; Phaseolus lunatus; Phaseolus coccineus; Hordeum vulgare; Nicotiana tabacum*
Violaxanthin	*Citrus sinensis*
Gibberellin A81	*Citrus reticulata*
Gibberellin A9	*Citrus reticulata*
6-Demethoxytangeritin	*Citrus reticulata*
Tetramethylscutellarein	*Citrus reticulata*

Metabolite	Plant
Apigenin 7-rutinoside	*Citrus aurantium*
Caryophyllene oxide	*Houttuynia emeiensis*
Capillaridin E	*Artemisia annua; Rhaponticum carthamoides*
Limonene	*Artemisia annua; Artemisia capillaris; Artemisia annua L.cultivar Jwarharti*
Cyclosativene	*Artemisia annua*
2-Nonanone	*Artemisia annua*
Kaempferol	*Artemisia capillaris; Sedum sarmentosum; Medicago sativa; Cryptomeria japonica*
Salicylic acid	*Artemisia capillaris*
Apigenin	*Artemisia capillaris*
Octanal	*Artemisia capillaris*
Rhamnocitrin	*Rhaponticum carthamoides*
Genkwanin	*Rhaponticum carthamoides*
p-Cymene	*Rhaponticum carthamoides*
Hexahydrofarnesyl acetone	*Artemisia annua L.cultivar Jwarharti*
Caprylic acid	*Centaurea sessilis*
Isophytol	*Centaurea sessilis; Anthemis aciphylla*
Phytol	*Centaurea sessilis; Anthemis aciphylla*
Bicyclogermacren	*Centaurea sessilis*
(2E)-Octenal	*Centaurea sessilis*
R-(+)-trans-Verbenol	*Centaurea sessilis*
2-Pentylfuran	*Anthemis aciphylla*
Isovaleraldehyde	*Alphinia galanga*
alpha-Thujene	*Leptospermum scoparium*
Neryl acetate	*Leptospermum scoparium*
6,9-Guaiadiene	*Piper arboreum*
Geraniol	*Piper arboreum*
Germacrene D	*Piper fimbriulatum*
Ampelopsin 7-glucoside	*Pseudotsuga menziesii*

Metabolite	Plant
Indole-3-carboxylic acid	*Pseudotsuga menziesii; Picea abies; Lycopersicon esculentum*
Gibberellin A9	*Pinus sylvestris; Lathyrus odoratus; Brassica napus; Hordeum vulgare*
Gibberellin A1	*Pinus sylvestris; Lycopersicon esculentum; Lupinus luteus*
Gibberellin A7	*Pinus sylvestris*
Indole-3-acetic acid	*Pinus sylvestris; Brassica rapa*
Methyl indole-3-acetate	*Pinus sylvestris*
Syringetin 3-(6''-acetylglucoside)	*Pinus sylvestris*
Laricitrin 3-(6''-acetylglucoside)	*Pinus sylvestris*
Taxifolin 3'-glucoside	*Picea abies*
Isorhamnetin 3-(6''-acetylgalactoside)	*Picea abies*
Gibberellin A20	*Prunus persica; Lycopersicon esculentum*
Afzelechin	*Prunus avium*
Gibberellin A44	*Prunus avium; Hordeum vulgare*
Gibberellin A30	*Prunus avium*
Chrysin	*Prunus avium*
Peonidin 3-rhamnoside	*Pisum sativum*
Gibberellin A97	*Lathyrus odoratus; Zea mays; Oryza sativa*
Gibberellin A51	*Lathyrus odoratus; Hordeum vulgare*
Gibberellin A12	*Lathyrus odoratus; Triticum aestivum; Brassica napus*
Gibberellin A24	*Lathyrus odoratus; Brassica napus; Hordeum vulgare*
Gibberellin A44	*Lathyrus odoratus*
Isosojagol	*Phaseolus lunatus*
Isoferreirin	*Phaseolus lunatus; Phaseolus vulgaris*
7,4'-Dihydroxy-5,2'-dimethoxyisoflavanone	*Phaseolus lunatus; Phaseolus vulgaris*
5-Deoxykievitol	*Phaseolus vulgaris*
(-)-Glycinol	*Phaseolus vulgaris*
Psoralidin	*Phaseolus coccineus*
Gibberellin A19	*Phaseolus coccineus; Lycopersicon esculentum*

Metabolite	**Plant**
Gibberellin A37	*Phaseolus coccineus*
Gibberellin A14	*Triticum aestivum*
Gibberellin A25	*Zea mays; Brassica napus*
Neoglucobrassicin	*Raphanus sativus; Armoracia lapathifolia*
Gibberellin A53	*Raphanus sativus*
Gibberellin A40	*Brassica napus*
Gibberellin A34	*Hordeum vulgare*
Gibberellin A8	*Oryza sativa*
Fisetin 3-methyl ether	*Robinia pseudoacacia*
Garbanzol	*Robinia pseudoacacia*
Butin	*Robinia pseudoacacia*
Gallocatechin	*Colophospermum mopane*
(+)-Gallocatechin	*Colophospermum mopane*
Dihydrorobinetin	*Acacia mearnsii*
Dihydroquercetin	*Acacia mearnsii*
Liquiritigenin	*Acacia mearnsii*
(-)-Epicatechin	*Acacia mearnsii*
ent-Epifisetinidol	*Acacia mearnsii*
Kaempferide	*Rhodiola sachalinensis; Sinocrassula indica*
Sinocrassoside C1	*Rhodiola sachalinensis*
Rhodionin	*Sinocrassula indica*
Rhodiosin	*Sinocrassula indica*
Nicotinic acid	*Lycopersicon esculentum*
9-Ribosyl-cis-zeatin	*Lycopersicon esculentum*
Indole-3-acetonitrile	*Nicotiana tabacum; Solanum tuberosum*
Trigonelline	*Nicotiana tabacum*
Phytuberin	*Nicotiana tabacum*
(-)-Phytuberin	*Nicotiana tabacum*
1-Caffeoyl-beta-D-glucose	*Nicotiana tabacum*

Metabolite	**Plant**
Desacetylphytuberin	*Solanum tuberosum*
9-Ribosyl-trans-zeatin	*Solanum tuberosum*
Prunetin	*Glycyrrhiza uralensis; Medicago sativa; Clitoria ternatea; Melilotus messanensis; Glycyrrhiza pallidiflora; Dalbergia odorifera*
Licoisoflavone A	*Glycyrrhiza uralensis; Glycyrrhiza inflata*
Glabridin	*Glycyrrhiza uralensis*
Topazolin	*Glycyrrhiza uralensis*
Isoderrone	*Glycyrrhiza uralensis; Glycyrrhiza aspera*
Licoflavone B	*Glycyrrhiza uralensis*
3'-Dimethylallylkievitone	*Glycyrrhiza uralensis*
Wighteone	*Glycyrrhiza uralensis; Glycyrrhiza aspera; Lupinus luteus; Derris scandens*
Hydroxywighteone	*Glycyrrhiza uralensis; Glycyrrhiza aspera; Lupinus albus; Erythrina variegata*
Licopyranocoumarin	*Glycyrrhiza aspera; Glycyrrhiza inflata*
Kanzonol W	*Glycyrrhiza aspera*
Licuroside	*Glycyrrhiza aspera; Glycyrrhiza inflata*
Neoisoliquiritigenin	*Glycyrrhiza aspera; Glycyrrhiza inflata*
1-O-Methylglycyrol	*Glycyrrhiza aspera; Glycyrrhiza glabra*
Kanzonol P	*Glycyrrhiza aspera; Glycyrrhiza glabra*
Neoliquiritin	*Glycyrrhiza aspera; Glycyrrhiza glabra; Glycyrrhiza inflata*
Licoricone	*Glycyrrhiza aspera; Glycyrrhiza glabra; Glycyrrhiza inflata*
Licoricesaponin A3	*Glycyrrhiza aspera; Glycyrrhiza glabra*
Licoricesaponin G2	*Glycyrrhiza aspera; Glycyrrhiza glabra*
Licoricesaponin C2	*Glycyrrhiza aspera; Glycyrrhiza glabra; Glycyrrhiza inflata*
Licoricesaponin E2	*Glycyrrhiza aspera; Glycyrrhiza glabra; Glycyrrhiza inflata*
Licoflavonol	*Glycyrrhiza aspera*
Gancaonin Q	*Glycyrrhiza aspera; Glycyrrhiza inflata*
Glisoflavanone	*Glycyrrhiza aspera*
Kanzonol I	*Glycyrrhiza aspera; Glycyrrhiza glabra*

Metabolite	**Plant**
Glycyrrhizol A	*Glycyrrhiza aspera*
Licoriphenone	*Glycyrrhiza aspera*
Licofuranocoumarin	*Glycyrrhiza glabra; Glycyrrhiza inflata*
Glyasperin L	*Glycyrrhiza glabra; Glycyrrhiza inflata*
Licocoumarone	*Glycyrrhiza glabra; Glycyrrhiza inflata*
Glyinflanin E	*Glycyrrhiza glabra*
Glyinflanin F	*Glycyrrhiza glabra*
Gancaonin A	*Glycyrrhiza glabra*
Semilicoisoflavone B	*Glycyrrhiza inflata*
Licoagrochalcone D	*Glycyrrhiza inflata*
Pratensol	*Lupinus luteus; Derris scandens; Clitoria ternatea; Melilotus messanensis; Glycyrrhiza pallidiflora; Dalbergia odorifera*
Indicanin E	*Lupinus luteus; Lupinus albus; Derris scandens*
Derrisisoflavone B	*Lupinus luteus; Lupinus albus*
Lupinisol A	*Lupinus luteus*
Lupalbigenin	*Lupinus luteus*
Derrisisoflavone F	*Lupinus luteus; Lupinus albus*
Lupinalbin B	*Lupinus luteus*
Lupinalbin G	*Lupinus luteus*
Laburnetin	*Lupinus luteus; Lupinus albus*
3'-Methylorobol	*Lupinus luteus; Lupinus albus*
Genistein	*Lupinus albus; Melilotus messanensis; Glycyrrhiza pallidiflora; Dalbergia odorifera*
Isoprunetin	*Lupinus albus; Derris scandens*
Scanderone	*Lupinus albus*
Chandalone	*Lupinus albus*
Ulexone A	*Lupinus albus*
Isolupalbigenin	*Lupinus albus; Derris scandens*
Gibberellin A18	*Lupinus albus*

Metabolite	**Plant**
Gibberellin A23	*Lupinus albus*
Lupinol C	*Lupinus albus*
Scandenal	*Lupinus albus; Erythrina variegata*
Lupisoflavone	*Derris scandens*
Alpinumisoflavone	*Derris scandens*
Derrone	*Derris scandens; Erythrina variegata*
Angustone B	*Derris scandens*
Angustone C	*Derris scandens*
Lupinisolone A	*Derris scandens*
Lupinisoflavone H	*Derris scandens*
Lupinifolin	*Derris scandens*
Barpisoflavone C	*Erythrina variegata*
Erysenegalensein K	*Erythrina variegata*
8-Prenylluteone	*Erythrina variegata*
Robustic acid	*Erythrina variegata; Erythrina senegalensis*
Eturunagarone	*Erythrina variegata; Erythrina senegalensis*
Bidwillon B	*Erythrina senegalensis*
Lonchocarpic acid	*Erythrina senegalensis*
Lonchocarpenin	*Erythrina senegalensis*
Scandenin	*Erythrina senegalensis*
Euchretin E	*Euchresta japonica*
Euchretin D	*Euchresta japonica*
Euchretin M	*Euchresta japonica*
Cytisine	*Euchresta japonica; Sophora flavescens; Maackia amurensis*
Formononetin	*Euchresta formosana*
Secundiflorol I	*Euchresta formosana; Sophora flavescens; Maackia amurensis*
Kushenin	*Euchresta formosana; Maackia amurensis; Sophora secundiflora; Sophora japonica; Glycyrrhiza pallidiflora*
Euchrenone a4	*Euchresta formosana*

Metabolite	**Plant**
Amorilin	*Euchresta formosana*
(-)-N-Methylcytisine	*Euchresta formosana*
Daidzein	*Sophora flavescens; Maackia amurensis; Melilotus messanensis; Glycyrrhiza pallidiflora; Dalbergia odorifera*
Thermopsine	*Sophora flavescens; Maackia amurensis*
(-)-6alpha-Hydroxylupanine	*Sophora flavescens; Sophora secundiflora*
(-)-6alpha-methoxylupanine	*Sophora flavescens; Sophora secundiflora*
Pratensein	*Maackia amurensis*
Mamanine	*Maackia amurensis; Sophora secundiflora*
5,6-Dehydrolupanine	*Maackia amurensis*
Argentine	*Maackia amurensis*
Maackiain	*Sophora secundiflora*
(-)-12,12'-Methylenedicytisine	*Sophora secundiflora*
Soyasapogenol E	*Medicago sativa; Trifolium pratense*
Xenognosin B	*Medicago sativa; Clitoria ternatea; Trifolium pratense; Sophora japonica; Melilotus messanensis*
Vestitone	*Medicago sativa*
Anhydroglycinol	*Medicago sativa; Trifolium pratense; Melilotus messanensis*
Lespedezol A1	*Medicago sativa; Trifolium pratense; Melilotus messanensis*
6a-Hydroxymedicarpin	*Medicago sativa; Glycyrrhiza pallidiflora; Dalbergia odorifera*
Sophorophenolone	*Medicago sativa*
Delphinidine	*Medicago sativa; Trifolium pratense; Sophora japonica; Melilotus messanensis*
Quercetin	*Medicago sativa; Clitoria ternatea; Melilotus messanensis*
Lespedezol G1	*Medicago sativa*
Licoagroside C	*Medicago sativa*
Pratol	*Medicago sativa; Glycyrrhiza pallidiflora*
Licoagroside E	*Medicago sativa*
7,4'-Di-O-methyldaidzein	*Medicago sativa; Trifolium pratense; Melilotus messanensis; Glycyrrhiza pallidiflora; Dalbergia odorifera*

Metabolite	**Plant**
Licodione 2'-methyl ether	*Medicago sativa*
Tricin	*Clitoria ternatea; Trifolium pratense*
Myricetin	*Clitoria ternatea; Trifolium pratense; Sophora japonica*
Irisolidone	*Trifolium pratense; Glycyrrhiza pallidiflora*
Erythrinin C	*Trifolium pratense; Sophora japonica*
Licoagroisoflavone	*Trifolium pratense; Sophora japonica*
Kushenin	*Trifolium pratense*
Tectoridin	*Trifolium pratense*
Sophojaponicin	*Trifolium pratense; Glycyrrhiza pallidiflora*
7,3'-Dimethylorobol	*Trifolium pratense*
Irisolidone 7-O-beta-D-glucoside	*Trifolium pratense*
(-)-Maackiain	*Trifolium pratense; Lespedeza homoloba*
Irilone	*Sophora japonica; Glycyrrhiza pallidiflora*
Isokaempferide	*Sophora japonica*
Trifoliol	*Sophora japonica*
Trifolirhizin	*Sophora japonica*
Irilone 4'-O-glucoside	*Sophora japonica*
(+)-6a-Hydroxymaackiain	*Sophora japonica; Lespedeza homoloba; Glycyrrhiza pallidiflora*
Coumestrol	*Lespedeza homoloba*
Bolusanthin III	*Lespedeza homoloba*
9-O-Methylcoumestrol	*Lespedeza homoloba*
5'-Methoxysativan	*Lespedeza homoloba*
Sativanone	*Melilotus messanensis*
2-Methoxymedicarpin	*Melilotus messanensis*
Coumestrin	*Glycyrrhiza pallidiflora*
Haginin D	*Glycyrrhiza pallidiflora*
Lespedezol F1	*Glycyrrhiza pallidiflora*
2'-Methoxyisoliquiritigenin	*Glycyrrhiza pallidiflora*
Echinatin	*Dalbergia odorifera*

Metabolite	**Plant**
Glypallichalcone	*Dalbergia odorifera*
Melilotocarpan B	*Dalbergia odorifera*
Cryptopine	*Corydalis claviculata; Corydalis solida*
Allocryptopine	*Corydalis claviculata; Papaver somniferum*
Rhoeadine	*Corydalis claviculata*
Norribasine	*Papaver somniferum*
Ribasine	*Papaver somniferum*
Sinoacutine	*Papaver somniferum*
alpha-Hydrastine	*Papaver somniferum*
Noscapine	*Corydalis solida*
Narcotoline	*Corydalis solida*
Salutaridine	*Corydalis solida*
Oxocularicine	*Corydalis solida*
Glaziovine	*Annona cherimola; Cocculus laurifolius; Artabotrys uncinatus*
Calycinine	*Annona cherimola*
Juzirine	*Annona cherimola*
Pronuciferine	*Xylopia parviflora; Stephania cepharantha Hayata*
Annocherine A	*Xylopia parviflora*
Annocherine B	*Xylopia parviflora*
Michelalbine	*Xylopia parviflora*
Micheline A	*Xylopia parviflora*
14-Episinomenine	*Cocculus laurifolius; Stephania cepharantha*
Cephamonine	*Stephania cepharantha*
Aknadilactam	*Stephania cepharantha*
Stephodeline	*Stephania cepharantha*
Tannagine	*Stephania cepharantha*
Aknadinine	*Stephania cepharantha*
Juziphine	*Stephania cepharantha*
N-Methylasimilobine	*Stephania cepharantha*

Metabolite	**Plant**
Coclaurine	*Stephania cepharantha Hayata*
Anonaine	*Stephania cepharantha Hayata*
Aknadicine	*Stephania cepharantha Hayata*
Cucoline	*Stephania cepharantha Hayata*
Oblongine	*Stephania cepharantha Hayata*
1-Methoxybrassitin	*Brassica oleracea*
Glucoviorylin	*Brassica oleracea; Brassica rapa; Hesperis matronalis*
Glucolepidiin	*Brassica oleracea; Brassica rapa; Hesperis matronalis*
Caulilexin C	*Brassica oleracea*
Brassicanal B	*Brassica oleracea*
Benzylglucosinolate	*Brassica oleracea; Brassica rapa; Hesperis matronalis*
Glucocheirolin	*Brassica oleracea; Brassica rapa; Hesperis matronalis*
Spirobrassinin	*Brassica oleracea*
3-Hydroxybutyl glucosinolate	*Brassica oleracea; Brassica rapa; Hesperis matronalis*
4-Methoxyglucobrassicin	*Brassica oleracea*
Sinigrin	*Brassica rapa; Hesperis matronalis*
Methyl anthranilate	*Brassica rapa*
Glucoberteroin	*Brassica rapa*
1-Methoxyspirobrassinin	*Brassica rapa*
Glucosinalbate	*Armoracia lapathifolia*
Glucoalyssin	*Armoracia lapathifolia*
Glucohesperin	*Armoracia lapathifolia*
Glucoiberin	*Hesperis matronalis*
Glucobrassicanapin	*Hesperis matronalis*
Glucolesquerellin	*Hesperis matronalis*
2-Methoxybenzyl glucosinolate	*Hesperis matronalis*
Alstolactone	*Alstonia macrophylla*
Isoalstonisine	*Alstonia macrophylla*
Normacusine B	*Alstonia macrophylla*

Metabolite	**Plant**
N(4)-Demethylalstonerinal	*Alstonia macrophylla; Alstonia angustifolia*
Cathafoline N(4)-oxide	*Alstonia macrophylla*
Alstophyllal	*Alstonia angustifolia; Alstonia angustifolia var.latifolia*
Macrocarpine A	*Alstonia angustifolia; Alstonia angustifolia var.latifolia*
Macrocarpine B	*Alstonia angustifolia; Alstonia angustifolia var.latifolia*
Alstohentine	*Alstonia angustifolia; Alstonia angustifolia var.latifolia*
Alstonisine	*Alstonia angustifolia*
Alstomaline	*Alstonia angustifolia; Alstonia angustifolia var.latifolia*
N1-Demethylalstophylline	*Alstonia angustifolia; Alstonia angustifolia var.latifolia*
(-)-Vincamajine	*Alstonia angustifolia var.latifolia*
Isopongaflavone	*Pongamia pinnata; Tephrosia purpurea*
Pongaglabol methyl ether	*Pongamia pinnata; Tephrosia purpurea*
Kanjone	*Pongamia pinnata; Tephrosia purpurea*
Lanceolatin B	*Pongamia pinnata*
Glabranin	*Pongamia pinnata*
Purpurenone	*Millettia pinnata*
O-Methylpongamol	*Millettia pinnata*
Purpuritenin B	*Millettia pinnata*
Pinnatin	*Millettia pinnata*
Piscisoflavone D	*Millettia pinnata*
Praecansone B	*Millettia pinnata*
Ovalichromene B	*Millettia pinnata*
Piscisoflavone A	*Neorautanenia amboensis*
Millettone	*Neorautanenia amboensis*
Pongachalcone I	*Tephrosia purpurea*
Pongapinnol D	*Tephrosia purpurea*
Karanjachromene	*Tephrosia purpurea*
Dihydroamorphigenin	*Tephrosia purpurea; Piscidia erythrina*
Dalpanol	*Tephrosia purpurea; Piscidia erythrina*

Metabolite	**Plant**
Amorphigenol	*Tephrosia purpurea; Piscidia erythrina*
Amorphigenin	*Tephrosia purpurea; Piscidia erythrina*
Jamaicin	*Tephrosia purpurea; Amorpha fruticosa*
Ichthynone	*Tephrosia purpurea; Amorpha fruticosa*
Dehydromillettone	*Tephrosia purpurea; Amorpha fruticosa*
3-O-Demethylamorphigenin	*Tephrosia purpurea; Piscidia erythrina*
Rotenone	*Amorpha fruticosa*
Candidone	*Piscidia erythrina*
Dehydrodeguelin	*Piscidia erythrina*
Ambofuranol	*Piscidia erythrina*
Neorautenanol	*Piscidia erythrina*
Isopongachromene	*Piscidia erythrina*
Blestrianol A	*Gymnadenia conopsea*
Blestrin A	*Gymnadenia conopsea*
Blestrin B	*Gymnadenia conopsea*
Blestrin C	*Gymnadenia conopsea*
Blestrin D	*Gymnadenia conopsea*
Isoarundinin-II	*Gymnadenia conopsea*
Bulbocodin C	*Bletilla striata*
Bulbocodin D	*Bletilla striata*
Gymconopin C	*Bletilla striata*
Bulbocol	*Bletilla striata*
Gymconopin D	*Bletilla striata*
12-Methylferruginol	*Taiwania cryptomerioides*
Pisiferal	*Taiwania cryptomerioides*
(-)-Nortrachelogenin	*Taiwania cryptomerioides; Cryptomeria japonica*
Matairesinol	*Chamaecyparis formosensis*
Cryptomeridiol	*Chamaecyparis formosensis*
Cubeb camphor	*Chamaecyparis formosensis*

Metabolite	Plant
Epicubebol	*Chamaecyparis formosensis*
Diphyllin	*Chamaecyparis formosensis*
Sandaracopimarinal	*Chamaecyparis formosensis*
Sugiol	*Chamaecyparis formosensis*
3,7,4'-Tri-O-methylkaempferol	*Chamaecyparis formosensis*
19-Hydroxyferruginol	*Chamaecyparis formosensis*
Pisiferanol	*Cryptomeria japonica*
Pisiferol	*Cryptomeria japonica*
Chaenocephalol	*Cryptomeria japonica*
(-)-Pluviatolide	*Cryptomeria japonica*
12-O-Methylpisiferanol	*Cryptomeria japonica*
alpha-Cadinol	*Cryptomeria japonica*
Sugiol methyl ether	*Cryptomeria japonica*
(+)-beta-Cyclocostunolide	*Artemisia spp.*
Arbusculin C	*Saussurea lappa*
Nortetraphyllicine	*Tabernaemontana coffeoides*
Vincamine	*Kopsia dasyrachis*
Ajmalicine	*Kopsia dasyrachis; Rauvolfia vomitoria*
Sarpagine	*Catharanthus roseus*
Alstonine	*Catharanthus roseus*
(+)-Isoeburnamine	*Catharanthus roseus*
(+)-Eburnamonine	*Catharanthus roseus*
14,15-Dihydroxyvincadifformine	*Catharanthus roseus*
Serpentine	*Rauvolfia vomitoria*
10-O-Methylsarpagine	*Rauvolfia vomitoria*
Dulcisxanthone B	*Garcinia mangostana*
Cowaxanthone D	*Garcinia mangostana*
Isonormangostin	*Garcinia mangostana*
Normangostin	*Garcinia dulcis*

Metabolite	**Plant**
Garcinone C	*Garcinia dulcis*
Mangostenone D	*Garcinia dulcis*
Mangostenone E	*Garcinia dulcis*
Tovophyllin B	*Garcinia dulcis*
Arborinine	*Severinia buxifolia*
Xanthoxyletin	*Ruta graveolens*
Clausamine A	*Ruta graveolens*
1,2,3-Trihydroxyacridone	*Ruta graveolens*
Bergapten	*Clausena excavata*
Psoralen	*Clausena excavata*
Rutacridone	*Clausena excavata*
Isoscopoletin	*Zanthoxylum simulans*
N-Methylflindersine	*Zanthoxylum simulans*
(-)-5-Methoxybalanophonin	*Zanthoxylum simulans*
beta-Sitosterone	*Zanthoxylum integrifoliolum*
Sepesteonol	*Zanthoxylum integrifoliolum*
Zanthobungeanine	*Zanthoxylum integrifoliolum*
Flindersine	*Zanthoxylum integrifoliolum*
Scopoletin	*Broussonetia papyrifera*
Sanggenon L	*Broussonetia papyrifera*
Gemichalcone B	*Broussonetia papyrifera*
Gemichalcone C	*Broussonetia papyrifera*
Artocommunol CA	*Morus alba*
Artocommunol CE	*Morus alba*
Marmesin	*Morus alba*
Dihydrocycloartomunin	*Morus alba*
Dihydroisocycloartomunin	*Morus alba*
Cycloartomunin	*Morus alba*
Artochamin D	*Morus alba*

Metabolite	**Plant**
Broussoflavonol A	*Morus alba*
(-)-Cycloartocarpin	*Morus alba*
Morusin	*Artocarpus communis*
Sanggenon M	*Artocarpus communis*
Cyclomulberrin	*Artocarpus communis*
Sanggenon A	*Artocarpus communis*
Mulberrin	*Artocarpus communis*
Demethoxyisogemichalcone C	*Artocarpus communis*
Isogemichalcone C	*Artocarpus communis*
Calystegin B2	*Lycium chinense*
Atropine	*Lycium chinense*
Calystegine B4	*Mandragora autumnalis*
Baccatin V	*Taxus cuspidata; Taxus brevifolia; Taxus wallichiana; Taxus chinensis; Taxus mairei; Taxus yunnanensis*
Taxol D	*Taxus cuspidata; Taxus chinensis; Taxus yunnanensis*
Dantaxusin D	*Taxus cuspidata; Taxus baccata; Taxus chinensis; Taxus mairei*
9-Deacetyltaxinine	*Taxus cuspidata; Taxus chinensis; Taxus mairei; Taxus yunnanensis*
Taxchin B	*Taxus cuspidata; Taxus baccata; Taxus mairei; Taxus yunnanensis*
Taxinine B	*Taxus cuspidata; Taxus chinensis*
Taxine I	*Taxus cuspidata; Taxus chinensis*
Taxumairol A	*Taxus cuspidata; Taxus baccata; Taxus chinensis*
Taxchinin H	*Taxus cuspidata; Taxus brevifolia*
Taxol B	*Taxus cuspidata*
Dantaxusin A	*Taxus cuspidata; Taxus baccata; Taxus chinensis; Taxus mairei*
Taxayuntin E	*Taxus cuspidata; Taxus chinensis*
Taxayuntin F	*Taxus cuspidata*
Taxchinin A	*Taxus cuspidata*
Taxchinin D	*Taxus cuspidata; Taxus baccata; Taxus wallichiana*
Taxchinin C	*Taxus cuspidata; Taxus brevifolia; Taxus yunnanensis*

Metabolite	**Plant**
Taxchinin K	*Taxus cuspidata; Taxus brevifolia; Taxus yunnanensis*
Taxayuntin H	*Taxus cuspidata; Taxus brevifolia; Taxus baccata; Taxus wallichiana; Taxus mairei*
Baccatin IV	*Taxus cuspidata; Taxus yunnanensis*
Taxamairin B	*Taxus cuspidata*
Dantaxusin C	*Taxus cuspidata; Taxus baccata; Taxus mairei*
Taxinine A	*Taxus cuspidata*
7-Deacetylcanadensene	*Taxus cuspidata; Taxus baccata*
Chinentaxunine	*Taxus cuspidata*
N-Methyltaxol C	*Taxus cuspidata; Taxus chinensis; Taxus yunnanensis*
14beta-benzoyloxybaccatin IV	*Taxus cuspidata; Taxus baccata; Taxus mairei*
5-Decinnamoyltaxuspine D	*Taxus cuspidata; Taxus yunnanensis*
5-Cinnamoyltaxicin I triacetate	*Taxus cuspidata; Taxus chinensis; Taxus mairei; Taxus yunnanensis*
5-epi-Canadensene	*Taxus cuspidata; Taxus chinensis; Taxus mairei*
7,2'-Bisdeacetoxyaustropicatine	*Taxus cuspidata*
Taxuspinanane I	*Taxus brevifolia; Taxus baccata; Taxus wallichiana; Taxus chinensis; Taxus mairei; Taxus yunnanensis*
Ormosin VI	*Taxus brevifolia; Taxus mairei*
Taxchinin B	*Taxus brevifolia; Taxus wallichiana*
Taxuspine V	*Taxus brevifolia; Taxus mairei*
14beta-Hydroxytaxusin	*Taxus brevifolia; Taxus baccata; Taxus wallichiana*
Epitaxol	*Taxus baccata; Taxus wallichiana; Taxus chinensis; Taxus mairei; Taxus yunnanensis*
Taxachitriene A	*Taxus baccata*
Taxachitriene B	*Taxus baccata*
Taxuspine S	*Taxus baccata; Taxus chinensis; Taxus mairei*
Taxezopidine L	*Taxus baccata; Taxus mairei*
Taxol C	*Taxus baccata*
Taxuspinanane A	*Taxus baccata; Taxus chinensis; Taxus yunnanensis*
Taxinine E	*Taxus baccata; Taxus chinensis*

Metabolite	**Plant**
Taxuspine Z	*Taxus baccata; Taxus wallichiana*
Taxezopidine E	*Taxus baccata*
Taxuspine X	*Taxus baccata*
Taxezopidine F	*Taxus baccata; Taxus yunnanensis*
Taxuyunnanine C	*Taxus baccata*
Taxuspinanane B	*Taxus baccata; Taxus wallichiana; Taxus chinensis*
Ponasterone A	*Taxus baccata; Taxus chinensis*
Decinnamoyltaxinine E	*Taxus baccata*
Taxezopidine H	*Taxus baccata; Taxus mairei*
Taxuspine D	*Taxus baccata; Taxus mairei*
Taxin B	*Taxus baccata*
Taxuspine U	*Taxus baccata; Taxus mairei*
5-Deacetyltaxachitriene B	*Taxus baccata*
2-Deacetyltaxachitriene A	*Taxus baccata*
Taxuspine A	*Taxus wallichiana; Taxus chinensis*
Taxuspine P	*Taxus wallichiana; Taxus yunnanensis*
Isolariciresinol	*Taxus wallichiana; Taxus yunnanensis*
Taxuspine T	*Taxus chinensis; Taxus mairei*
Taxine II	*Taxus chinensis*
Taxuspine B	*Taxus chinensis; Taxus mairei*
Taxuspine E	*Taxus chinensis; Taxus yunnanensis*
Taxayuntin J	*Taxus chinensis; Taxus mairei*
Taxumairol F	*Taxus chinensis; Taxus yunnanensis*
Deaminoacyltaxine A	*Taxus chinensis; Taxus mairei; Taxus yunnanensis*
Isotaxine B	*Taxus chinensis; Taxus yunnanensis*
Taxinine H	*Taxus chinensis; Taxus mairei*
Dantaxusin B	*Taxus chinensis; Taxus mairei*
Taxayuntin A	*Taxus chinensis*
2-Deacetyldecinnamoyltaxinine E	*Taxus chinensis; Taxus mairei*

Metabolite	**Plant**
Taxuyunnanine N	*Taxus mairei*
10-Deacetyltaxol A	*Taxus mairei*
13-Deacetoxybaccatin I	*Taxus mairei; Taxus yunnanensis*
Taxumairol V	*Taxus yunnanensis*
Taxchinin I	*Taxus yunnanensis*
Methyl pyroglutamate	*Panax notoginseng*
Ginsenoside Rh7	*Panax notoginseng; Panax pseudo-ginseng var.notoginseng*
Ginsenoyne E	*Panax ginseng; Panax pseudo-ginseng var.notoginseng*
Notoginsenoside	*Panax ginseng*
Panaxydol	*Panax pseudo-ginseng var.notoginseng*
Ginsenoside Re	*Panax pseudo-ginseng var.notoginseng*
Aristolactam IIIa	*Aristolochia elegans*
Aristolochate I	*Aristolochia elegans*
Aristolactam I	*Aristolochia elegans*
Methyl aristolochate	*Aristolochia elegans*
Aristolochate C	*Aristolochia elegans*
Aristolochic acid I	*Aristolochia heterophylla*
4-Hydroxybenzoic acid	*Aristolochia heterophylla*
O-Methylflavinantine	*Artabotrys uncinatus*
Pallidine	*Artabotrys uncinatus*
Norpallidine	*Artabotrys uncinatus*
Lysicamine	*Artabotrys uncinatus*
Flavinantine	*Annona purpurea*

Appendix B

GenBank ID (*rbcL*, *matK*, ITS2) and use information of sample plants. Economic uses of plants are represented as following abbreviations: E (edible), M (medicinal), L (landscaping,), T (timber), P (poisonous), W (wild plant). Some plants are both edible and medicinal and are annotated as M/E. (*Partial sequence data)

Plant name	***rbcL***	***matK***	**ITS2**	**uses**
Rosmarinus officinalis	NC_027259.1	NC_027259.1	EU796893.1	M
Anthemis aciphylla BOISS. var.discoidea BOISS			*FM957767.1	W
Acritopappus confertus			*KP454449.1	W
Nardostachys chinensis	*AF446950.1	AF446920.1	*AY236190.1	W
Valeriana officinalis	L13934.1	*AY362532.1	EU796889.1	M
Mentha arvensis L.	*HQ590183.1	*JN896123.1	AY656005.1	M
Solanum lycopersicum	NC_007898.3	NC_007898.3	AB373816.1	E
Cyperus rotundus L.	*AM999813.1	*KX369513.1		M
Zingiber officinale	KM213122.1	KM213122.1	KC582868.1	M/E
Alphinia galanga	*KY189086.1	AF478815.1	AF478715.1	M/E
Curcuma amada Roxb	*KF981156.1	*KJ872380.1	AH009165.2	M/E
Curcuma aeruginosa	*KX608611.1	AF478840.1	DQ438047.1	W
Pinus halepensis	JN854197.1	JN854197.1	AF037007.1	L
Cedrus libani	*HG765043.1			L
Cistus albidus	*FJ225860.1	*DQ092975.1	*DQ092933.1	W
Melaleuca leucadendra L.	*KX527090.1		*EU410106.1	M
Cistus creticus	*FJ225862.1	*DQ092979.1	*DQ092937.1	W
Myrtus communis	JQ730673.1	AY525136.2	GU984341.1	M

Plant name	***rbcL***	***matK***	**ITS2**	**uses**
Leptospermum scoparium	*HM850121.1	*KM065275.1	KM065050.1	M
Rhodiola rosea L.	*KM360979.1	*KP114859.1	KF454616.1	M
Piper arboreum	*GQ981830.1		EF056223.1	W
Piper fimbriulatum			EF056254.1	W
Polygonum minus	*FM883633.1	*JN896184.1	EU196895.1	M
Brassica hirta	*HM849823.1	LC064389.1	FJ609733.1	E
Saussurea lappa	*KX527328.1	*KX526536.1	KJ721545.1	M
Artemisia annua	*KJ667633.1	*HM989754.1	KX219675.1	M
Artemisia capillaris	NC_031400.1	NC_031400.1	KT965668.1	M
Olea europaea	NC_013707.2	NC_013707.2	KJ188984.1	M/E
Juniperus phoenicea	*HM024320.1	*HM024042.1	GU197870.1	W
Hesperis matronalis	*KM360815.1	*HQ593319.1	AJ628314.1	L
Citrus sinensis	DQ864733.1	DQ864733.1	AB456127.1	E
Citrus reticulata	*AB505952.1	AB626773.1	AB456115.1	E
Citrus aurantium	*AB505953.1	AB626798.1	AB456126.1	E
Citrus paradisi	*AJ238407.1	*JN315360.1	AB456065.1	E
Citrus limon	*AB505956.1	AB762353.1	AB456128.1	E
Citrus aurantifolia	KJ865401.1	KJ865401.1	AB456118.1	M/E
Houttuynia cordata	*AY572259.1	DQ212712.1	*AM777852.1	M/E
Helianthus annuus	NC_007977.1	NC_007977.1	KF767534	E
Carthamus tinctorius	KM207677.1	KM207677.1	KX108699.1	M
Hordeum vulgare	KC912687.1	KC912687.1	KM252865.1	E
Triticum aestivum	KJ592713.1	KJ592713.1	AJ301799.1	E
Zea mays	NC_001666.2	NC_001666.2	*KJ474678.1	E
Oryza sativa	KM103369.1	KM103369.1	KM252851.1	E
Allium cepa	KM088013.1	KM088013.1	AM492188.1	E

Plant name	***rbcL***	***matK***	**ITS2**	**uses**
Picea abies	*EU364777.1	AB161012.1	AJ243167.1	T
Pinus sylvestris	*JF701589.1	AB097781.1	AF037003.1	T
Brassica napus	NC_016734.1	NC_016734.1	AB496975.1	P
Cucumis sativus	DQ119058.1	DQ119058.1	AJ488213.1	E
Glycine max	NC_007942.1	NC_007942.1	AJ011337.1	E
Phaseolus lunatus		DQ445985.1	Y19456.1	E
Phaseolus vulgaris	EU196765.1	EU196765.1	GU217644.1	E
Phaseolus coccineus	*LT576851.1	DQ445966.1	Y19453.1	E
Pisum sativum	KJ806203.1	KJ806203.1	AB107208.1	E
Lathyrus odoratus	KJ850237.1	KJ850237.1	KX287478.1	L
Vicia faba	KF042344.1	KF042344.1	*EU288904.1	E
Linum usitatissimum	FJ169596.1		EU307117.1	T
Malus domestica	*KM360872.1	AM042561.1	AF186484.1	E
Prunus cerasus	*HQ235416.1	*FN668844.1	FJ899099.1	E
Prunus persica	HQ336405.1	HQ336405.1	*KX674813.1	E
Prunus avium	*HQ235394.1	*AM503828.1	HQ332169.1	E
Citrus unshiu	*AB505946.1	AB626802.1	AB456117.1	E
Spinacia oleracea	NC_002202.1	NC_002202.1		E
Camellia sinensis	KC143082.1	KC143082.1	*EU579773.1	E
Pseudotsuga menziesii	JN854170.1	JN854170.1	AF041353.1	T
Cassia fistula	*U74195.1	*JQ301870.1	JQ301830.1	M
Colophospermum mopane	*JX572425.1	AY386894.1	AY955788.1	T
Robinia pseudoacacia	KJ468102.1	KJ468102.1	GU217616.1	L
Acacia mearnsii	*KF532045.1	HM020723.1	KF048786.1	W
Garcinia mangostana	*JX664049.1		AJ509214.1	M/E
Garcinia dulcis	JF738433.1		EU128468.1	W

Plant name	***rbcL***	***matK***	**ITS2**	**uses**
Eriobotrya japonica	KT808478.1	DQ860462.1	FJ449737.1	E
Aesculus hippocastanum	*KM360616.1	EU687725.1	EU687637.1	P
Rheum sp.	*EU840308.1	EU840469.1		W
Raphanus sativus	NC_024469.1	NC_024469.1	AY746462.1	E
Armoracia lapathifolia	*KM360651.1	LC064385.1	AF078032.1	E
Brassica oleracea	KR233156.1	KR233156.1	GQ891877.1	E
Brassica rapa	AY167977.1	AY541619.1	KF454313.1	E
Daucus carota	DQ898156.1	DQ898156.1	AH003468.2	W
Asclepias curassavica	*EU916742.1	*DQ026716.1	AM396884.1	L
Nicotiana tabacum	NC_001879.2	NC_001879.2	*KP893959.1	M
Capsicum annuum	KR078313.1	KR078313.1	*KP893996.1	E
Lycopersicon esculentum	NC_007898.3	NC_007898.3	AJ300201.1	E
Cyperus rotundus	*KJ773433.1	*KX369513.1	*KX675088.1	M
Humulus lupulus	NC_028032.1	NC_028032.1	AB033891.1	M
Catharanthus roseus	KC561139.1	KC561139.1	HQ130657.2	M
Petunia x hybrida	*HM850249.1	*EF439018.1		L
Diospyros kaki	NC_030789.1	NC_030789.1	AB175009.1	E
Clitoria ternatea	*U74237.1	EU717427.1	AF467038.1	E
Sedum sarmentosum	NC_023085.1	NC_023085.1	*GQ434462.1	M
Psidium guajava	NC_033355.1	NC_033355.1	*AB354956.1	E
Phyllanthus emblica	*AY765269.1	AY936594.1	*KU508339.1	M/E
Phellodendron amurense	*AF066804.1	FJ716737.1	*KT972670.1	M
Epimedium sagittatum	NC_029428.1	NC_029428.1		M
Rhodiola sachalinensis	*KJ570585.1	*KJ570498.1		M
Sinocrassula indica		*AF115679.1		M
Amorpha fruticosa	KP126864.1	KP126864.1	GU217619.1	L

Plant name	***rbcL***	***matK***	**ITS2**	**uses**
Glycyrrhiza uralensis	*AB012126.1	AB280741.1	AB649775.1	M
Glycyrrhiza aspera		*JQ669639.1	GQ246126.1	W
Glycyrrhiza glabra	NC_024038.1	NC_024038.1	*KX675022.1	M/E
Glycyrrhiza inflata	*AB012127.1	AB280743.1	JF778868.1	M
Erythrina variegata	*KF496750.1	*KU587466.1	KJ716427.1	L
Sophora japonica	*U74230.1	*HM049517.1	JQ676976.1	T
Medicago sativa	KU321683.1	KU321683.1	Z99236.1	E
Trifolium pratense	KP126856.1	KP126856.1	AF154620.1	M
Lespedeza homoloba			KY174702.1	W
Glycyrrhiza pallidiflora	*HM142228.1	EF685997.1	GQ246130.1	W
Dalbergia odorifera	*KM510281.1	*KM521320.1	*GQ434362.1	T
Neorautanenia amboensis		*KX213174.1		W
Lupinus luteus	NC_023090.1	NC_023090.1	AF007478.1	W
Lupinus albus	KJ468099.1	KJ468099.1	AF007481.1	E
Derris scandens		JX506621.1	JX506450.1	W
Euchresta japonica	*AB127040.1			W
Euchresta formosana	*AB127039.1			W
Sophora flavescens	*AB127037.1	*HM049520.1	GU217622.1	M
Maackia amurensis	*AB127041.1	AY386944.1	Z72352.1	L
Sophora secundiflora	*Z70141.1		AF174638.1	W
Daphniphyllum oldhami	KC737396.1	KC737244.1	JN040993.1	M
Annona purpurea	*KM068886.1	*JQ586490.1		E
Annona cherimola	NC_030166.1	NC_030166.1		E
Xylopia parviflora	*JF265661.1	*JF271002.1		W
Cocculus laurifolius DC.	*JN051677.1	AF542588.2	KM092304.1	W
Stephania cepharantha	*JN051691.1	*GU373530.1	AY017400.1	W

Plant name	***rbcL***	***matK***	**ITS2**	**uses**
Cocculus pendulus (Forsk.) Diels	*FJ026478.1			W
Corydalis solida	*KM360733.1		X85464.1	W
Papaver somniferum	NC_029434.1	NC_029434.1	DQ364699.1	M
Rubia yunnanensis	*KP098291.1		*KP098123.1	M
Taraxacum formosanum			*AY862577.1	W
Alpinia blepharocalyx	*KJ871690.1	AF478809.1	AF478709.1	W
Hibiscus taiwanensis	*KX527103.1	*KX526698.1		W
Xylocarpus granatum	*KF848252.1	*KJ784619.1		W
Acanthopanax senticosus	JN637765.1	JN637765.1	*KX674996.1	M
Panax notoginseng	KR021381.1	KR021381.1	KT380921.1	M
Panax ginseng	KM067390.1	KM067390.1	*AB043872.1	M
Bupleurum rotundifolium			AF481400.1	M
Bellis perennis	*AY395530.1	KP175061.1	JN315918.1	M/E
Lonicera japonica	NC_026839.1	NC_026839.1	EU240693.1	M
Solanum tuberosum	KM489056.2	KM489056.2		E
Withania somnifera	*FJ914179.1	*KR734871.1	JQ230981.1	M
Punica granatum	*L10223.1	*JQ730680.1	*FM887008.1	E
Beta vulgaris	KR230391.1	KR230391.1		E
Taxus wallichiana	KX431996.1	KX431996.1	EF660573.1	M
Taxus cuspidata	*DQ478793.1	AF228104.1	KU904438.1	P
Taxus brevifolia	*AF249666.1	*EU078561.1	EF660600.1	M
Taxus baccata	*AF456388.1	DQ478791.1	EF660599.1	M
Taxus chinensis	*AY450855.1	AF228103.1	AF259300.1	M
Taxus mairei	KJ123824.1	KJ123824.1	KU904440.1	M
Taxus yunnanensis	*AY450857.1			M

Plant name	***rbcL***	***matK***	**ITS2**	**uses**
Tabernaemontana coffeoides Boj.		*GU973924.1		W
Rauvolfia vomitoria	*DQ660663.1	*DQ660538.1		W
Alstonia macrophylla	*GU135289.1	*GU135060.1		T
Tephrosia purpurea	*LT576862.1	*KF545845.1		P
Pongamia pinnata	*AY289676.1		AF467493.1	L
Millettia pinnata	NC_016708.2	NC_016708.2	JX506445.1	L
Psoralea corylifolia	*JN114837.1		GU217608.1	M
Calophyllum inophyllum	*HQ332016.1	*HQ331553.1	AJ312608.2	T
Broussonetia papyrifera	*AF500347.1	*AF345326.1	AB604292.1	E
Morus alba	KU981119.1	KU981119.1	AM041998.1	M/E
Artocarpus communis	*AF500345.1	*KJ767846.1		E
Gymnadenia conopsea R.BR.	*KJ451493.1	EF612530.1	Z94068.1	M
Bletilla striata	NC_028422.1	NC_028422.1	KJ405419.1	M
Curcuma zedoaria	*GU180515.1	AB047743.1	KJ803170.1	E
Taiwania cryptomerioides	NC_016065.1	NC_016065.1	*AY916831.1	T
Chamaecyparis formosensis	*AY380879.1	*FJ475234.1		T
Cryptomeria japonica	NC_010548.1	NC_010548.1	AF387522.1	T
Angelica sinensis	*JN704983.1	*GQ434227.1	JX138965.1	M
Lycium chinense	*FJ914171.1	*AB036637.1	KC832461.1	M
Mandragora autumnalis	*HQ216129.1			M
Curcuma domestica	*KX608614.1	AB551931.1	KJ803148.1	M/E
Plantago major	*KJ204386.1	*KJ593055.1	AB281165.1	M
Rehmannia glutinosa	*FJ172725.1	*GQ434277.1	EU266023.1	M
Andrographis paniculata	KF150644.2	KF150644.2	*KT898259.1	M
Scutellaria baicalensis	NC_027262.1	NC_027262.1	JN853779.1	M

Plant name	***rbcL***	***matK***	**ITS2**	**uses**
Magnolia denudata	NC_018357.1	NC_018357.1		M
Magnolia officinalis	NC_020316.1	NC_020316.1	JF755930.1	M
Aeschynanthus bracteatus			AF349283.1	W
Angelica furcijuga KITAGAWA			DQ278164.1	M/E
Zanthoxylum simulans	*KT634182.1	EF489100.1	DQ016545.1	M
Severinia buxifolia	*AF066806.1	AB762384.1	JX144180.1	W
Aristolochia elegans		*AB060790.1	KM092119.1	L
Aristolochia heterophylla Hemsl	*KU853431.1	*KU853368.1		M
Cannabis sativa	NC_027223.1	NC_027223.1	KF454086.1	M
Citrus sudachi		AB762337.1	AB456086.1	M
Salvia officinalis	*AY570431.1	*JQ934074.1	FJ883522.1	M/E
Orthosiphon stamineus		*KM658969.1	*AY506663.1	W
Murraya paniculata	*AB505906.1	AB762389.1	KM092325.1	M
Belamcanda chinensis	*AJ309694.1	AY596652.1	JF421476.1	M
Murraya euchrestifolia			*JX144210.1	W
Ruta graveolens	*U39281.2	EF489057.1	JQ230976.1	M/E
Clausena excavata	NC_032685.1	NC_032685.1	JX144189.1	W
Caesalpinia crista	*KP094390.1	*EU361900.1		T

Appendix C

The 90 plants with high priority for future screening for overall medicinal bioactivity:

Panax pseudo-ginseng var.notoginseng; Panax ginseng C.A.Meyer; Trichosanthes tricuspidata; Bupleurum rotundifolium; Dracaena draco; Tribulus pentandrus; Solanum abutiloides; Silphium perfoliatum; Dioscorea spongiosa; Astragalus trojanus; Polygala japonica; Duranta repens; Ilex kudingcha; Kandelia candel; Baikiaea plurijuga; Dicranopteris pedata; Camellia sinensis var. viridis; Cistus incanus; Rheum sp.; Vancouveria hexandra; Melicope triphylla; Chrysothamnus viscidiflorus; Hypericum sampsonii; Anaxagorea luzonensis A.GRAY; Rhamnus disperma; Podocarpus fasciculus; Chrysothamnus nauseosus; Platanus acerifolia; Pityrogramma triangularis; Grevillea robusta; Podocarpus nivalis; Hypericum erectum Thunb.; Petunia x hybrid; Solanum spp.; Acacia dealbata; Ardisia colorata; Syzygium samarangense; Eugenia jambolana; Leptarrhena pyrolifolia; Nymphaea caerulea; Abies amabilis; Hyacinthus orientalis; Eustoma grandiflorum; Salvia splendens; Lathyrus odoratus; Rosa spp.; Rhododendron spp.; Empetrum nigrum; Vaccinium padifolium; Saussurea medusa; Crataegus pinnatifida; Betula nigra; Conocephalum conicum; Tephrosia toxicaria; Syzygium samarangense; Eugenia jambolana; Leptarrhena pyrolifolia; Nymphaea caerulea; Abies amabilis; Hyacinthus orientalis; Eustoma grandiflorum; Salvia splendens; Lathyrus odoratus; Rosa spp.; Rhododendron spp.; Empetrum nigrum; Vaccinium padifolium; Saussurea medusa; Crataegus pinnatifida; Betula nigra; Conocephalum conicum; Tephrosia toxicaria; Euphorbia supina Rafin; Oricia suaveolens; Rhodobacter sphaeroides; Erwinia uredovora; Myxococcus xanthus; Streptomyces griseus; Rhodobacter capsulatus; Corbicula sandai; Corbicula japonica; Silurus asotus; Erysimum asperum; Cibotium glaucum; Gibberella fujikuroi; Marah macrocarpus; Pharbitis purpurea; Haplophyllum patavinum; Niphogeton ternate; Chloranthus japonicus