OPEN ACCESS

GAVIN PUBLISHERS

Current Research in Complementary & Alternative Medicine

Liu k, et al. Curr Res Complement Altern Med: CRCAM-130.

Research Article

DOI:10.29011/CRCAM-130/100030

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

Kang Liu, Aki H. Morita, Shigehiko Kanaya, Md. Altaf-Ul-Amin*

Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192, Japan

***Corresponding author:** Md. Atlaf-Ul-Amin, Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192, Japan.Tel: +81743725326; Fax: +81743725329; Email: amin-m@is.naist.jp

Citation: Liu K, Morita AH, Kanaya S, Atlaf-Ul-Amin M (2018) Metabolite-Content-Guided Prediction of Medicinal/Edible Properties in Plants for Bioprospecting. Curr Res Complement Altern Med: CRCAM-130. DOI:10.29011/CRCAM-130/100030

Received Date: 19 March, 2018; **Accepted Date:** 26 March, 2018; **Published Date:** 04 April, 2018

Abstract

Metabolite-content (MC) refers to all small molecules which are the products or intermediates of metabolism within an organism. The metabolite-contents of plants which involve numerous secondary metabolites are highly related to their nutritional and medicinal features. Previous researches have confirmed that phylogeny-guided approaches have been seen as one of the time-efficient and informative approaches to plant-based drug discovery. However, the phylogenetic reconstruction of plants is not determined conclusively from genomic sequence data. Here, we investigate the systematic value of metabolite-contents of plants, especially the predictive power of metabolite-content data in exploration of edible and medicinal properties for bioprospecting. In this study, 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. We used a network based approach to abstract structurally similar metabolites as features, and measure the phylogenetic distance by a binary method. We also reconstructed phylogenetic trees based on plastid markers *rbcL*, *matK* and ITS2 for the same set of plants, to investigate the predictive power of these two approaches, sequence- and MC-based approaches, in guiding the prediction of medicinal/edible properties.

Our results reveal that besides the genomic 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 properties in a different perspective from sequence-based approach. Our study therefore provides a new approach for plant bioprospecting, and the predictive power of metabolite-content data for medicinal/edible plants will also be improved with the improvement and completeness of the metabolite-content database.

Keywords: Chemosystematics; Metabolite-content; Phylogeny; Prediction; Secondary Metabolite

Introduction

Plants are the major contributors of natural products and are usually rich in nutritional or medicinal properties, which are attributed to the complex secondary metabolite constituents of them [1-3]. 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 [4-6]. Many pharmaceutical drugs are derived from plants that were first used in traditional systems of medicine [6]. According to the World Health Organization, about 25% of medicines are plant-derived [2].

Discoveries of novel molecules and advances in production of plant-based products have revived interest in natural product research [7,8]. The number of traditionally used plant species worldwide is estimated to be between 10,000 and 53,000; however, only a small proportion have been screened for biological activity [9-11], 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 [12]. 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 [13,14]. 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 [13-18]. However, most of these studies focused on only a small cluster of genera, which limits its practical application. This approach also faces the limitation of incomplete sequence data. Moreover, phylogenetic distance correlated to feature similarity of species will also be invalid once beyond a certain threshold [19]. Therefore, a special perspective different from sequence-based phylogeny is valuable for understanding the evolution of bioactive features and facilitating the prediction and discovery of medicinal properties in plants.

Besides molecular biology which is in the view of nucleotide sequence comparison, metabolite feature is also closely related to the evolution of pathways for both primary and secondary metabolites. Many researchers have begun to explore phylogenetic distance between species from the diversity of metabolite features, 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 [20]. Borenstein et al. (2008) predicted the phylogenetic tree by comparing the "seed set" of metabolic networks [21]. Mano et al. (2010) considered the topology of pathways as chains and used a pathway-alignment method to classify species [22]. 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 [23]. Ma et al. (2013) demonstrated the usefulness of the global alignment of multiple metabolic networks to infer the phylogenetic relationships between species [24]. 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 [25]. However, most of these studies have focused on microorganisms such as archaea, and only a few studies have involved land plants and the bioactive compounds produced by them.

The systemization of plants on the basis of their chemical constituents, which is also known as plant chemosystematics, could be helpful in solving taxonomical problems and exploring nutritional and medicinal properties from plants. Traditional chemosystematics of plants is based on the presence of selected metabolites. The incomplete data of metabolite constituents of plants limits its ability to solve taxonomical problems and discover new natural products or medicinal properties from plants [26, 27]. To perform a holistic review on the metabolite features of a species, we propose the concept of metabolite-content. Metabolite-content refers to all small molecules which are the products or intermediates of metabolism within an organism. It differs from metabolome in that the metabolite-content focuses on the qualitative collection of small metabolites and ignores the quantitative differences, which is instable with different parts and stages of one organism.

The secondary metabolite constituents of a plant are 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 it [26]. Comparative classification of plants based on their metabolite-content-similarity could be used to explore the evolutionary and bioactive relation between them [28]. Here, we investigate the phylogenetic value of metabolite-content data, especially the predictive power of metabolite-content data in exploration of medicinal and edible plants for bioprospecting, using the KNApSAcK Core DB.

The KNApSAcK Core DB is an extensive plant-metabolite relation database that can be applied in multifaceted researches of plants, such as identification of metabolites, construction of integrated databases, bioinformatics and systems biology [29-32], and can be considered an advanced source of metabolite-content data of plants. 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 [30].

In this paper, 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. We used a network based approach to abstract structurally similar metabolite groups as features, and measure the phylogenetic distance by a binary method. We also reconstructed the phylogenetic tree based on common DNA barcodes for a subset of plants, to investigate the predictive power of these two approaches, sequence- and metabolite-content-based method, in guiding the prediction of medicinal/edible plants for bioprospecting.

Material and Methods

Dataset 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 [30]. 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. 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 [33]. 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 before based on three common DNA barcodes: two chloroplast barcodes *rbcL* and *matK*, and one nuclear barcode ITS2. The DNA sequence data are collected from GenBank [34], and certainly there is lack of data for some plants. Here we select the plants with both abundant metabolite-content (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 (Figure 1).

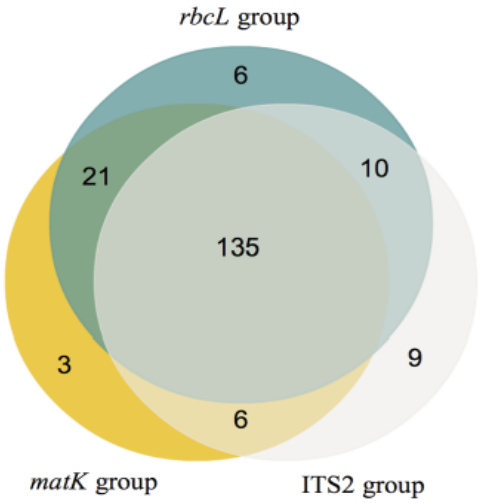


Figure 1: Overview of 190 plants included in *rbcL*, *matK* and ITS2 sample groups.

Phylogenetic Hypothesis

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. 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 [35,36]. 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.

Clustering of Plants Based on Metabolite-Content Similarity

For classifying plants based on currently available metabolite-content data, firstly we need an approach that can compensate for the limitations of missing data. Adjacent metabolites along a metabolic pathway are often related to similar substructures, and structurally similar metabolites are involved in the same or similar pathway. Therefore, plants that share highly structurally similar metabolites are likely to be within the same category and represent similar bioactivity. In this study, we linked plants to structurally similar metabolite groups instead of individual metabolites.

We used the Tanimoto coefficient to measure the structural similarity between two metabolites and constructed a network of metabolites based on chemical structure similarity [37]. The Tanimoto coefficient between two metabolites A and B is defined as follows, which is the proportion of the features shared by two compounds divided by their union:

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

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 molecular fingerprints using the R package ChemmineR [33]. Empirically, a Tanimoto coefficient value larger than 0.85 indicates that the compared compounds represent highly similar bioactive features [38]. We used 0.85 as the threshold to insert an edge between two metabolites and constructed a network of metabolites. For this network we applied a graph-clustering algorithm DPClus to generate metabolite groups that contains structurally similar compounds.

The DPClus algorithm is a graph-clustering algorithm that can be used to extract densely connected nodes as a cluster [39,40]. 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, density d and cluster property cp, are used in this algorithm. 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

$$cp_{nk} = \frac{E_{nk}}{d_k \times N_k} \quad (2.2)$$

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

By DPClus algorithm the metabolites were divided into many groups such that each group contains structurally similar compounds and can be treated as a distinctive pattern of structure. A plant is related to a metabolite group if it is related to any metabolite in the group. The original plant-metabolite relations are transformed into plant versus metabolite-group relations. We used such groups to measure the similarity between plants, thus reducing the effects of incomplete metabolite-content data. Each plant could be represented as a binary vector which is comprised of p variables with values 1 or 0 indicates presence or absence of each metabolite group. The metabolite-content-similarity of two plants was calculated by Simpson similarity coefficient as follow,

$$S_{Sim} = \frac{a}{\min\{(a+b),(a+c)\}} \quad (2.3)$$

Here, a, b, and c are the frequencies of the events x&y, $\bar{x}$&y, and x&$\bar{y}$, respectively [41-43]. 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.

Thus for each of the three sample groups (172 samples with data of gene *rbcL*, 165 samples with data of gene *matK*, 160 samples with data of gene ITS2), we can reconstruct the phylogenetic trees by sequence data and metabolite-content data. To evaluate the power of the sequence data and the metabolite-content data in predicting medicinal/edible properties in the context of these three sample groups, we reconstructed phylogenetic trees by similarity of corresponding gene sequence and similarity of metabolite-content data respectively, and performed comparative analysis for these two types of phylogenetic trees.

Phylogenetic and Statistical Analyses

We assessed the relationship of phylogeny with the medicinal/edible properties by calculating the phylogenetic signal of medicinal/edible plants. We investigated 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 [44,45]. D is calculated as follows,

$$D = \frac{\sum d_{obs} - mean(\sum d_b)}{mean(\sum d_r) - mean(\sum d_b)} \quad (2.4)$$

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) indicating 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.

Evolutionary Patterns of Medicinal/Edible Properties

To narrow down the number of species chosen for an early stage of medicinal/edible plant discovery screening, we identified the position of phylogeny clustering for medicinal/edible properties. We highlighted 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 [46]. 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 tree 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 trees we obtained, we recorded the percentage of the total and medicinal/edible 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.

Results and Discussion

All of the sequence data were downloaded from GenBank (Table 1). It should be noted that not all samples have complete sequence data (Table 2). 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 human. The KNApSAcK species-metabolite relation database is also far from complete with a large amount of data fragments. However, the plants with abundant metabolite-content 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 sequence data in GenBank for facilitating the analysis of evolutionary relations between plants and guiding the prediction of medicinal/edible plants since the plants covered by these two databases 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. We reconstructed the phylogenetic trees for the three sample groups by corresponding sequence data and metabolite-content data respectively (Figure 2).

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
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
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
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
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
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
Tabernaemontana coffeoides Boj.		*GU973924.1		W

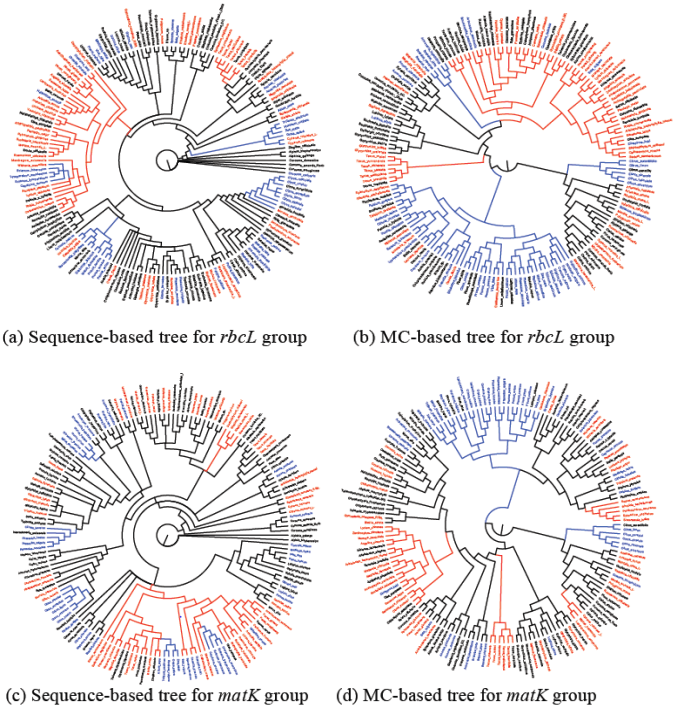
Rauvolfia vomitoria	*DQ660663.1	*DQ660538.1		W
Alstonia macrophylla	*GU135289.1	*GU135060.1		T
Tephrosia purpurea	*LT576862.1	*KF545845.1		P
Pongamia pinnata	*AY289676.1		AF467493.1	L
Millettia pinnata	NC_016708.2	NC_016708.2	JX506445.1	L
Psoralea corylifolia	*JN114837.1		GU217608.1	M
Calophyllum inophyllum	*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
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

Table 1: 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 medicinal and edible and are annotated as M/E. (*Partial sequence data).

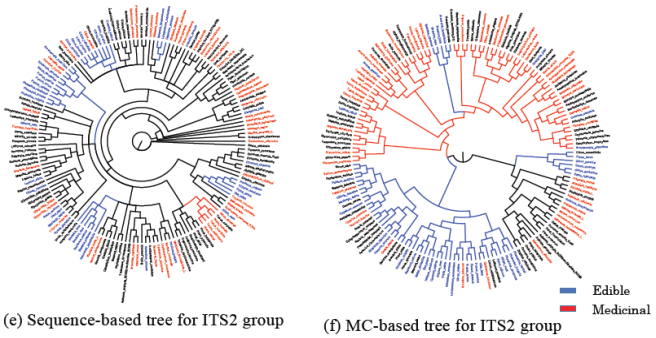
	rbcL	*matK*	ITS2
Null	18	25	30
Complete Sequence	73	112	131
Partial Sequence	99	53	29

Table 2: The amount of complete and partial sequences data of *rbcL*, *matK* and ITS2 sample groups.



(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 2: Phylogenetic trees and the hot nodes of medicinal/edible features for sequence- and MC-based approaches.

The uses 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, landscaping plants, timber plants, poisonous plants and wild plants (Table 3).

Edible	Medicinal	Medicinal/ Edible	Wild	Landscaping	Timber	Poisonous
47	60	15	38	13	13	4

Table 3: The amount of plants in each category of uses.

We investigated the strength in phylogenetic signal of medicinal and edible categories for each phylogenetic tree we obtained using the D statistic (Table 4). We found that plants with medicinal/edible uses are significantly clustered in metabolite-content-based phylogenetic trees of 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 metabolite-content-based trees. The ITS2-based tree shows weak phylogenetic signal for both medicinal and edible plants.

Phylogenetic Tree	Feature	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

Table 4: Phylogenetic signal of medicinal/edible features in sequence-based and metabolite-content (MC) based trees.

Generally, the edible plants are more 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 across the lineages reveals some but less phylogenetic relations. The 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. Moreover, the expressions of the secondary metabolites with medicinal bio-activity are more closely related to the overall metabolite features, i.e., metabolite-contents of the plants. The plants with similar metabolite-contents tend to have similar medicinal features, and such observations are more obvious comparing with sequence-based approach in our experiments. Thus we might found more phylogenetic patterns by skipping gene data and comparing metabolite-content data directly. Considering the gene data available from GenBank is usually incomplete, the metabolite-content data implies great potential applications in predicting medicinal properties.

As a tentative approach to narrow down the number of medicinal/edible plants selected for bioprospecting, we also identified the hot nodes that are significantly overrepresented by species of medicinal/edible uses (Table 5). We can observe that 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 metabolite-content 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 metabolite-content-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 phylogenies. 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.

Phylogenetic tree	Feature	Total plants included (%)	M/E Hits (%)	Gain in M/E hits (%)	Co-included plants (hits)
rbcL group (sequence)	Edible	30 (17.4%)	20 (43.5%)	150%	Edible:20 (18)
	Medicinal	46 (26.7%)	29 (50.9%)	90.60%	Medicinal:27 (20)
rbcL group (MC)	Edible	64 (37.2%)	37 (80.4%)	116.10%	-
	Medicinal	64 (37.2%)	32 (56.1%)	50.80%	-
matK group (sequence)	Edible	23 (13.9%)	21 (44.7%)	221.60%	Edible:16 (16)
	Medicinal	44 (26.7%)	23 (42.6%)	59.70%	Medicinal:12 (10)
matK group (MC)	Edible	32 (19.4%)	26 (55.3%)	185.10%	-
	Medicinal	34 (20.6%)	25 (46.3%)	124.70%	-
ITS2 group (sequence)	Edible	35 (21.9%)	27 (65.0%)	196.80%	Edible:30 (25)
	Medicinal	5 (3.1%)	5 (9.6%)	207.70%	Medicinal:5 (5)
ITS2 group (MC)	Edible	61 (38.1%)	35 (85.4%)	124.10%	-
	Medicinal	82 (51.2%)	35 (67.3%)	31.40%	-

Table 5: 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. M/E Hits (%): The number (percentage) of the medicinal/edible plants included in the hot nodes of medicinal/edible uses. Gain in M/E 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.

The phylogenetic distribution of medicinal and edible 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 metabolite-content similarity, and also suggests that there may be 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 metabolite-content-based phylogenetic trees. We found that there is not significantly coincidence of medicinal plants encompassed by hot nodes of these two types of phylogenetic trees. In other words; the medicinal patterns identified by metabolite-content-based approach shows no significant similarity to the medicinal patterns identified by sequence-based approach. Our findings thus indicate that the metabolite-content-based approach might highlighted different group of medicinal plants with sequence-based approach, and might reflect more unexplored medicinal potential not associated with the sequence-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 metabolite-content-similarity (Figure 3). We selected plants containing at least 14 metabolites to ensure data integrity. Plant uses information (edible or medicinal uses) was imported from KNApSAcK World Map DB. For the total 1047 tested plants, we found medicinal or edible uses information for 605 plants from World Map DB, with 543 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 uses information are regarded as wild plants from which we may explore new medicinal properties. The hot nodes for medicinal plants encompass 288 plants, 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 metabolite-content-similarity with other 198 medicinal plants (Table 6).

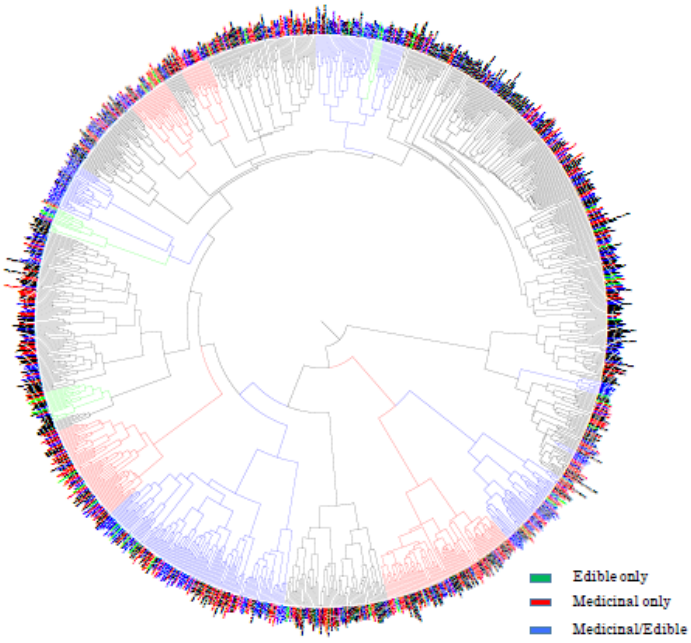


Figure 3: MC-based phylogenetic tree for 1047 plants, with the hot nodes of medicinal/edible Plants.

Panax pseudo-ginseng var.notoginseng; Panax ginseng C.A.Meyer*; Trichosanthes tricuspidata; Bupleurumrotundifolium; Dracaena draco; Tribulus pentandrus; Solanum abutiloides; Silphium perfoliatum; Dioscorea spongiosa; Astragalus trojanus; Polygala japonica; Duranta repens; Ilex kudingcha; Kandelia candel; Baikiaeaplurijuga; 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; Pityrogrammatriangularis; 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; Lathyrusodoratus; Rosa spp.; Rhododendron spp.; Empetrum nigrum; Vaccinium padifolium; Saussurea medusa; Crataegus pinnatifida; Betula nigra; Conocephalum conicum; Tephrosia toxicaria; Syzygium samarangense;Eugenia jambolana;Leptarrhenapyrolifolia;Nymphaeacaerulea;Abiesamabilis;Hyacinthusorientalis;Eustomagrandiflorum; Salvia splendens; Lathyrus odoratus; Rhododendron spp.; Empetrum nigrum; Vacciniumpadifolium; 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; Haplophyllumpatavinum; Niphogeton ternate; Chloranthus japonicus*

Table 6: The 90 plants with high priority for future screening for overall medicinal bioactivity.

Conclusion

Many researchers have proved that edible and medicinal 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 sequence data, metabolite-content data is also closely associated with medicinal and edible bioactivity of plants and can explore the medicinal/edible patterns in a different perspective from DNA sequence-based plant phylogeny.

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

Since sequence-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 sequence-based bioprospecting. Metabolite-content-based plant phylogeny reconstruction could provide a new perspective in plant bioprospecting. With the improvement of metabolite-content database and the integration of various plant pharmacopoeia, such MC-guided bioprospecting approach can be further accelerated, and the predictive power for medicinal/edible plants will also be improved with the completeness of metabolite-content database in future.

Acknowledgements:

This work was supported by the National Bioscience Database Center in Japan; the Ministry of Education, Culture, Sports, Science, and Technology of Japan (16K07223 and 17K00406), Platform project for Supporting Drug Discovery and Life Science Research funded by Japan Agency for Medical Research and Development and NAIST Big Data Project.

References

1. Dahanukar SA, Kulkarni RA, Rege N N (2000) Pharmacology of medicinal plants and natural products. Indian journal of pharmacology 32: S81-S118.
2. Ciddi Veeresham (2012) Natural products derived from plants as a source of drugs 3: 200-201.
3. Cseke LJ, Kirakosyan A, Kaufman PB, Warber S, Duke, JA, et al. (2016) Natural products from plants. CRC press.
4. Newman DJ, Cragg GM, Snader KM (2000) The influence of natural products upon drug discovery. Natural product reports 17: 215-234.
5. Cragg GM, Newman DJ (2013) Natural products: a continuing source of novel drug leads. Biochimica et Biophysica Acta (BBA)-General Subjects 1830: 3670-3695.
6. DS Fabricant, NR Farnsworth (2001) The value of plants used in traditional medicine for drug discovery. Environmental health perspectives 109: 69-75.
7. Ro DK, Paradise EM, Ouellet M, Fisher KJ, Newman KL, et al. (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 440: 940-943.
8. Graham IA, Besser K, Blumer S, Branigan CA, Czechowski T, et al. (2010) The genetic map of *Artemisia annua L.* identifies loci affecting yield of the antimalarial drug artemisinin. science 327: 328-331.
9. McChesney JD, Venkataraman SK, Henri JT (2007) Plant natural products: back to the future or into extinction?. Phytochemistry 68: 2015-2022.
10. Soejarto DD, Fong HHS, Tan GT, Zhang HJ, Ma CY, et al. (2005) Ethnobotany/ethnopharmacology and mass bioprospecting: Issues on intellectual property and benefit-sharing. Journal of ethnopharmacology 100: 15-22.
11. Gurib-Fakim A (2006) Medicinal plants: traditions of yesterday and drugs of tomorrow. Molecular aspects of Medicine 27: 1-93.
12. Brummitt NA, Bachman SP (2010) Plants under pressure-a global assessment: the first report of the IUCN sampled red list index for plants. Kew, UK: Royal Botanic Gardens.
13. Rønsted N, Symonds MR, Birkholm T, Christensen SB, Meerow AW, et al. (2012) Can phylogeny predict chemical diversity and potential medicinal activity of plants? A case study of Amaryllidaceae. BMC evolutionary biology 12: 182.
14. Ernst M, Saslis-Lagoudakis CH, Grace OM, Nilsson N, Simonsen HT, et al. (2016) Evolutionary prediction of medicinal properties in the genus *Euphorbia* L. Scientific reports 6: 30531.
15. Rønsted N, Savolainen V, Mølgaard P, Jäger AK (2008) Phylogenetic selection of Narcissus species for drug discovery. Biochemical Systematics and Ecology 36: 417-422.
16. Saslis-Lagoudakis CH, Klitgaard BB, Forest F, Francis L, Savolainen V, et al. (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: e22275.
17. Saslis-Lagoudakis CH, Savolainen V, Williamson EM, Forest F, Wagstaff SJ, et al. (2012) Phylogenies reveal predictive power of traditional medicine in bioprospecting. Proceedings of the National Academy of Sciences 109: 15835-15840.
18. Yessoufou K, Daru BH, Muasya AM (2015) Phylogenetic exploration of commonly used medicinal plants in South Africa. Molecular ecology resources 15: 405-413.
19. Kelly S, Grenyer R, Scotland RW (2014) Phylogenetic trees do not reliably predict feature diversity. Diversity and Distributions 20: 600-612.
20. Clemente JC, Satou K, Valiente G (2007) Phylogenetic reconstruction from non-genomic data. Bioinformatics 23: e110-e115.
21. Borenstein E, Kupiec M, Feldman MW, Ruppin E (2008) Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proceedings of the National Academy of Sciences 105: 14482-14487.
22. Mano A, Tuller T, Béjà O, Pinter RY (2010) Comparative classification of species and the study of pathway evolution based on the alignment

of metabolic pathways. BMC bioinformatics 11: S38.

23. Chang CW, Lyu PC, Arita M (2011) Reconstructing phylogeny from metabolic substrate-product relationships. BMC bioinformatics 12: S27.

24. Ma CY, Lin SH, Lee CC, Tang CY, Berger B, et al. (2013) Reconstruction of phyletic trees by global alignment of multiple metabolic networks. BMC bioinformatics 14: S12.

25. Abdullah AA, Altaf-Ul-Amin M, Ono N, Sato T, Sugiura T, et al. (2015) Development and mining of a volatile organic compound database. BioMed research international13: 1-13.

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

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

28. Liu K, Abdullah AA, Huang M, Nishioka T, Altaf-Ul-Amin M, et al. (2017) Novel Approach to Classify Plants Based on Metabolite-Content Similarity. BioMed research international.

29. Shinbo Y, Nakamura Y, Altaf-Ul-Amin M, Asahi H, Kurokawa K, et al. (2006) KNApSAcK: a comprehensive species-metabolite relationship database. Plant metabolomics 165-181.

30. Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, et al. (2012) KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. Plant and Cell Physiology 53: e1.

31. Ikeda S, Abe T, Nakamura Y, Kibinge N, Hirai Morita A, et al. (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: 711-727.

32. Nakamura Y, Mochamad Afendi F, Kawsar Parvin A, Ono N, Tanaka K, et al. (2014) KNApSAcK metabolite activity database for retrieving the relationships between metabolites and biological activities. Plant and Cell Physiology 55: e7.

33. Cao Y, Charisi A, Cheng LC, Jiang T, Girke T (2008) ChemmineR: a compound mining framework for R. Bioinformatics 24: 1733-1734.

34. Benson DA, Cavanaugh M, Clark K, et al. (2013) GenBank. Nucleic acids research 41: D36-D42.

35. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. bioinformatics 23: 2947-2948.

36. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754-755.

37. Godden JW, 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: 163-166.

38. Martin YC, Kofron JL, Traphagen LM (2002) Do structurally similar molecules have similar biological activity?. Journal of medicinal chemistry 45: 4350-4358.

39. Altaf-Ul-Amin M, Tsuji H, Kurokawa K, Asahi H, Shinbo Y, et al. (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.

40. 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.

41. Choi SS., Cha SH, Tappert CC (2010) A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics 8: 43-48.

42. Fallaw WC (1979) A test of the Simpson coefficient and other binary coefficients of faunal similarity. Journal of Paleontology 53: 1029-1034.

43. Ma HW, Zeng AP (2004) Phylogenetic comparison of metabolic capacities of organisms at genome level. Molecular phylogenetics and evolution 31: 204-213.

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

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

46. Webb CO, Ackerly DD, Kembel SW (2008) Phylocom: software for the analysis of phylogenetic community structure and trait evolution. Bioinformatics 24: 2098-2100.