IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 14, 2018, accepted September 27, 2018, date of publication October 5, 2018, date of current version October 29, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2874089

A Data-Driven Approach for Identifying Medicinal Combinations of Natural Products

SUNYONG YOO[1,2], SUHYUN HA[1,2], MOONSHIK SHIN[1,2], KYUNGRIN NOH[3], HOJUNG NAM[4], AND DOHEON LEE[1,2]

[1]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea
[2]Bio-Synergy Research Center, Daejeon 34141, South Korea
[3]Global Business Service, IBM Korea, Seoul 07326, South Korea
[4]School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

Corresponding authors: Hojung Nam (hjnam@gist.ac.kr) and Doheon Lee (dhlee@kaist.ac.kr)

This work was supported by the Bio-Synergy Research Project of the Ministry of Science, ICT, and Future Planning through the National Research Foundation under Grant NRF-2012M3A9C4048758.

ABSTRACT Combinations of natural products have been used as important sources of disease treatments. Existing databases contain information about prescriptions, herbs, and compounds and their relationships with phenotypes, but they do not have information on the use of combinations of natural product compounds. In this paper, we identified large-scale associations between natural product combinations and phenotypes by applying an association rule mining technique to integrated information on herbal medicine, combination drugs, functional foods, molecular compounds, and target genes. The rationale behind this approach is that natural products commonly found in medicinal multicomponent mixtures have statistically significant associations with the therapeutic effects of the multicomponent mixtures. Based on a molecular network analysis and an external literature validation, we show that the inferred associations are valuable information for identifying medicinal combinations of natural products since they have statistically significant closeness proximity in the molecular layer and have much experimental evidence. All results are available through the workbench site at http://biosoft.kaist.ac.kr/coconut to facilitate the investigation of the medicinal use of natural products and their combinations.

INDEX TERMS Association rules, combination drug, combination therapy, data mining, databases, medicinal combinations, natural products, pharmaceutical technology, polypharmacology.

I. INTRODUCTION

Natural products and their mixtures have been used as a valuable source of medicinal agents, and many modern drugs are still derived from natural products [1], [2]. According to a previous review, 49% of approved cancer drugs are based on natural products, while only 25% are synthetic drugs [3]. Moreover, approximately 70-80% of the world's population depends on herbal sources for their primary health care [4]–[6].

Combination therapy is gaining attention for overcoming the critical issues of single drug treatments, such as acquired resistance and side effects [7]–[9]. Recently, several combinations of natural products with various synergies were found in herbal medicines [10]–[12], which indicates that combinations of natural products can be used as a valuable source for combination therapy. Therefore, a better understanding of natural product combinations, including their sources, efficacy and molecular mechanisms, based on accumulated knowledge is essential to identify new drug combinations.

There are several ongoing efforts to gather information on natural products. Herbal medicine databases, such as TCMID, TCM@Taiwan, HIT, TM-MC and PharmDB-K, were established to provide information about prescriptions, herbs, compounds and their relationships [13]–[17]. NPACT, NutriChem, MAPS and SuperNatural provide bioactivity and target information about natural products [18]–[21]. However, these databases rarely consider combinations of natural products. Although DCDB covers various types of information on combination drugs, it lacks information on natural products and their combinations since most compounds are small molecules [22]. Therefore, a database containing the association information between natural product combinations and phenotypes will benefit the discovery of potential therapeutic compound combinations. However, there is no

TABLE 1. License information on the 13 resources used in COCONUT.

Resource	License	Availability	Reference
KTKP	MOU	O	http://www.koreantk.com/ktkp2014/popup/en/copyright.jsp
TCMID	Custom	O	http://www.megabionet.org/tcmid/download/
KAMPO	Custom	O	https://kampo.ca/
DrugBank	Custom	O	https://scicrunch.org/resolver/SCR_002700
CTD	Custom	O	https://scicrunch.org/resolver/ SCR_006530
DCDB	Custom	O	http://www.cls.zju.edu.cn/dcdb/
BindingDB	mixed CC BY 3.0 and CC BY-SA 3.0	O	https://scicrunch.org/resolver/SCR_000390
ClinicalTrials.gov	Custom	O	https://scicrunch.org/resolver/ SCR_002309
MATADOR	CC BY-NC-SA 3.0	O	http://matador.embl.de/
TTD	none	O	https://scicrunch.org/resolver/ SCR_002309/SCR_006892
SIDER	CC BY-NC-SA 4.0	O	https://scicrunch.org/resolver/SCR_004321
STITCH	Custom	O	https://creativecommons.org/licenses/by/3.0/us/
FooDB	Custom	O	http://foodb.ca/
BFN	MOU	O	http://biofood.or.kr

database dedicated to the analysis of the potential effects of natural product combinations.

In this study, we developed an innovative database named Compound Combination-Oriented Natural Product Database with Unified Terminology (COCONUT), which contains associations between natural product combinations and phenotypes inferred from heterogeneous sources regarding herbal medicine, drug combinations, functional foods, molecular compounds and target gene information. Our fundamental hypothesis is that the natural product compounds commonly found in medicinal multicomponent mixtures for treating the phenotype are more related to the phenotype than are other compounds. Therefore, an association rule mining technique was applied to find frequent patterns between natural product combinations and phenotypes from medicinal multicomponent mixtures. Inferred associations were evaluated based on molecular network analysis and external literature. We confirmed that inferred associations have a statistically significant proximity in the molecular layer and cover the large number of results that have been reported in previous work. All integrated and inferred data of COCONUT are available through the workbench site at http://biosoft.kaist.ac.kr/coconut.

II. MATERIALS AND METHODS

A. DATA SOURCES

Korean, Chinese and Japanese herbal medicine information on prescriptions, herbs and compounds was collected from KTKP [23], TCMID [13] and Kampo [24], respectively. Food and its compound composition information was collected from FooDB [25]. Drug information was acquired from DrugBank [26], and combination drug information was extracted from DCDB [22]. Functional food information was collected from the BFN database [27]. Furthermore, compound-phenotype associations were collected from DrugBank, CTD [28], ClinicalTrials.gov [29] and SIDER [30]. Compound-gene associations were collected from DrugBank, DCDB, CTD, TTD [31], BindingDB [32], MATADOR [33] and STITCH [34]. Gene-phenotype associations were collected from CTD. Additionally, we collected protein-protein interaction (PPI) network data from BioGrid [35] and pathway data from KEGG [36], which are used to investigate the effects of natural products on the molecular layer. License and availability information of the source databases is described in Table 1.

B. DATA INTEGRATION PROCESS

The goal of this study is finding the frequent patterns between natural product combinations and phenotypes from composition and efficacy information of medicinal multicomponent mixtures. To do this, we constructed a database for integrating the comprehensive information about natural products with eight major data entities: prescription, functional food, combination drug, herb, compound, compound property, gene and phenotype. The database included 13 relationships among the data entities, such as composition and functional information of the medicinal materials or associations among compounds, phenotypes and genes. The detailed schema of COCONUT is shown in Fig. 1.

To store the data in a systematic manner, we integrated the information from heterogeneous sources into a standard format and structured the efficacy information. The construction procedure consists of following four steps (Fig. 2).

1) TERMINOLOGY UNIFICATION

In the terminology unification step, each entity instance is mapped to the corresponding international identifiers to resolve duplicate instances and to enhance interoperability

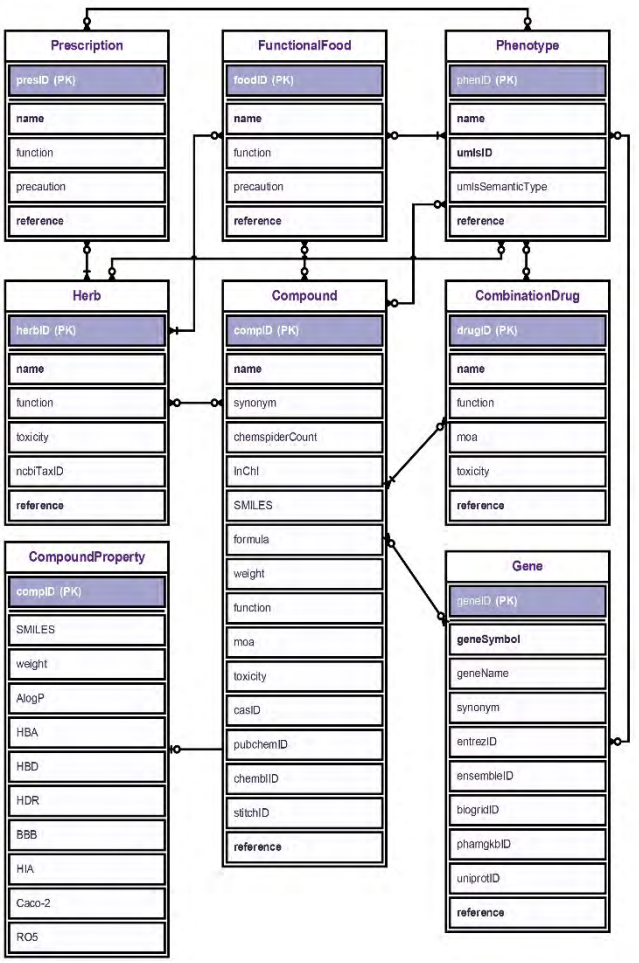

FIGURE 1. Database schema of COCONUT. There are eight major dataentities: prescription, functional food, combination drug, herb, compound, compound property, gene, and phenotype. The 13 types of relationships among the major data entities are recorded.

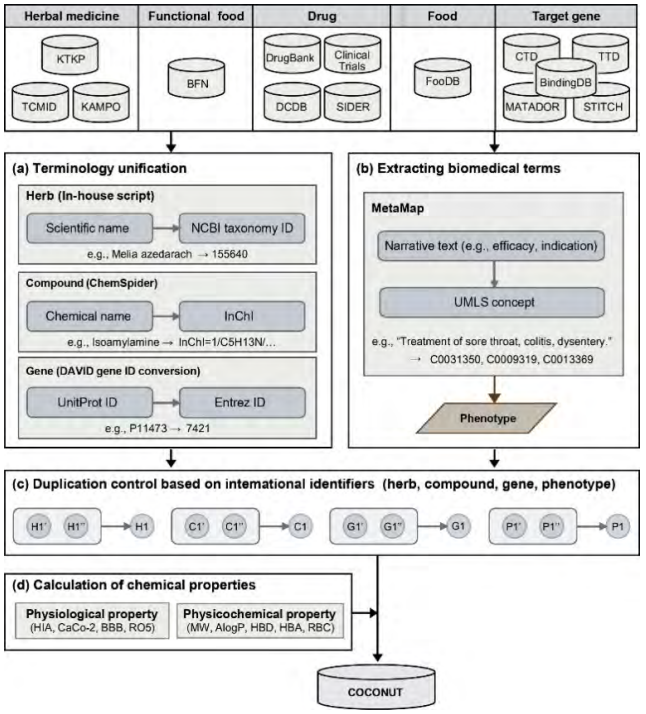

FIGURE 2. The systematic procedure for database construction (a) Terminology unification by mapping entities to international identifiers. (b) Extracting biomedical terms from narrative text using a text mining tool. (c) Performing duplication resolution on combined datasets based on international identifiers. (d) Calculating chemical properties including four physiological and five physicochemical properties.

with other public databases (Fig. 2a). Herbs were mapped to NCBI taxonomy ID, a curated classification for the organisms, based on scientific names [37]. For compounds, the ChemSpider API was used to find the international chemical identifier (InChI) [38]. Genes were mapped to Entrez gene ID with the DAVID gene ID conversion tool (DICT) [39]. Because most source databases have different types of gene information, such as gene symbols, UNIPROT ID and ENSEMBL ID, DICT was used to map the various types of gene ID to Entrez gene ID. Additionally, we identified combination drugs and functional foods with DCDB ID and brand name, respectively.

2) EXTRACTION OF BIOMEDICAL TERMS

In many databases, a large amount of efficacy and indication information is described in the narrative text. Thus, there is a problem that when a user searches for entity instances associated with a particular phenotype, all words of the narrative text should be retrieved. To extract phenotype-related terms from the narrative text, we employed MetaMap, a text-mining tool that maps biomedical text to the Unified Medical Language System (UMLS) concepts (Fig. 2b) [40]–[43]. Compared to other named entity recognition (NER) tools, MetaMap is strong in the validity and customization aspects. Many previous studies have demonstrated that MetaMap has applicability and reliability in the field of clinical and biomedical forms [44]–[46]. Additionally, MetaMap can be customized in the configuration layer [47], [48]. To avoid ambiguous results, we used MetaMap's word-sense disambiguation (WSD) module that identifies reliable words based on the context of a sentence. Depending on the characteristics of sources, we used the term processing option that recognizes input text as one phrase. Finally, UMLS concepts for biomedical text can be obtained from MetaMap.

UMLS currently integrates over 730,000 biomedical concepts from more than fifty biomedical vocabularies [41]. All concepts are categorized into 133 predefined semantic types. For each concept, UMLS editors assigned one or several semantic types. In this study, we tried to identify phenotype-related terms that could fully describe the efficacy and indications of medicinal materials. Based on the MetaMap results and definitions of semantic types, 20 semantic types were selected through manual curation to represent the phenotypes related to diseases or symptoms (Table 2). Then, we applied MetaMap to the narrative text and extracted

TABLE 2. Phenotype-related UMLS semantic types.

Abbreviation	Semantic type	Example
acab	Acquired abnormality	Liver spots
anab	Anatomical abnormality	Hernia
biof	Biologic function	Regulation
cgab	Congenital abnormality	Myelodysplasia
comd	Cell or molecular dysfunction	Metaplasia
dsyn	Disease or syndrome	Angina
emod	Experimental model of disease	Liver cirrhosis
fndg	Finding	Loss of weight
inpo	Injury or poisoning	Brain damage
lbtr	Laboratory or test result	Liver enzymes
menp	Mental process	Emotions
mobd	Mental or behavioral dysfunction	Depression
neop	Neoplastic process	Uterine fibroids
patf	Pathologic function	Fluid retention
phsf	Physiologic function	Endocrine effect
sosy	Sign or symptom	Chest pain
clna	Clinical attribute	Osmolalities
hops	Hazardous or poisonous substance	Carcinogen
bpoc	Body part, organ or organ component	Kidney
tisu	Tissue	Membranes

phenotype-related terms using the selected semantic types as filtering criteria.

3) DUPLICATION RESOLUTION

Because the data were collected from multiple sources, duplicate instances exist. Therefore, duplicate instances of the combination drug, functional food, herb, compound and gene entities are detected by comparing the international identifiers. We then merged duplicate instances into one instance and registered it with source references (Fig. 2c). For example, information about *Melia azedarach* used as an herb was collected from the KTKP and TCMID databases. Based on the NCBI taxonomy ID, we determined whether the herb information gathered from each resource is the same. If they are the same, the function and toxicity information are integrated and the reference for each information item is registered. However, prescriptions are hard to distinguish by their general name only because they contain different herb compositions, even though they have the same general name. Therefore, even if instances of a prescription entity have the same general name, they are considered as different instances.

4) CALCULATION OF CHEMICAL PROPERTIES

Chemical properties of compounds were calculated for producing physiological effects (Fig. 2d). Physiological effects include human intestinal absorption (HIA), Caco-2 permeability, blood-brain barrier (BBB), and Lipinski's rule of five (RO5). Based on the physiological effects, we can investigate whether natural products are orally bio-available, drug available or effective on certain tissues. For example, *in vivo* absorption of natural products across the gut wall can be estimated based on the Caco-2 permeability. In this study, physiological effects are calculated based on the physicochemical properties, such as molecular weight, AlogP, hydrogen-bond donors, hydrogen-bond acceptors and rotatable bond count. HIA and BBB values are calculated with Shen's method [49], while Caco-2 permeability is calculated using Pham's method [50]. RO5 are calculated with the CDK Descriptor Calculator [51]. Based on the physiological effects, we can analyze various functional activities of the natural products on the human body.

C. A DATA-DRIVEN ANALYSIS FOR IDENTIFYING MEDICINAL COMBINATIONS OF NATURAL PRODUCTS

Our fundamental hypothesis is that compounds commonly found in herbs, functional foods or combination drugs used to treat or prevent the phenotype are more related to the phenotype, compared to other compounds. To extract associations between natural product combinations and phenotypes, we applied an association rule mining analysis. The analysis procedure consists of the following three steps (Fig. 3).

1) GENERATING DATASET FOR ASSOCIATION RULE MINING

To apply association rule mining, we first constructed compound set profiles for each phenotype that contain compound composition and phenotype information of 4,370 herbs, 1,322 functional foods and 1,605 combination drugs (Fig. 3a). In the compound set profile, compounds are set of items and a phenotype is a class label. Each attribute has a value of 1 or 0; a '1' is assigned if the multicomponent mixture contains that attribute, otherwise '0' is assigned. To determine how many herbs, combination drugs and functional foods are related to phenotypes and how many compounds are contained in each entity, we examined the distribution of the number of phenotypes and the distribution of the number of compounds for each entity (Fig. 4).

2) APPLYING ASSOCIATION RULE MINING

With current biological knowledge, it is impossible to determine the dependency between natural products in medicinal materials or the linearity of relationships between combinations of natural products and phenotype. Association rule mining has the distinct advantage of being able to directly model based on conditional probabilities, avoiding the linearity assumptions underlying many classical supervised classification, regression and ranking methods [52]–[54]. Moreover, association rule mining helps avoid problems with the curse of dimensionality [55], [56]. In this study, medicinal materials contain 35,741 unique natural compounds. If we consider all possible combinations of natural compounds, then there are $2^{35,741}$ possibilities - a high curse of dimensionality. In practice, however, most of the combination drugs

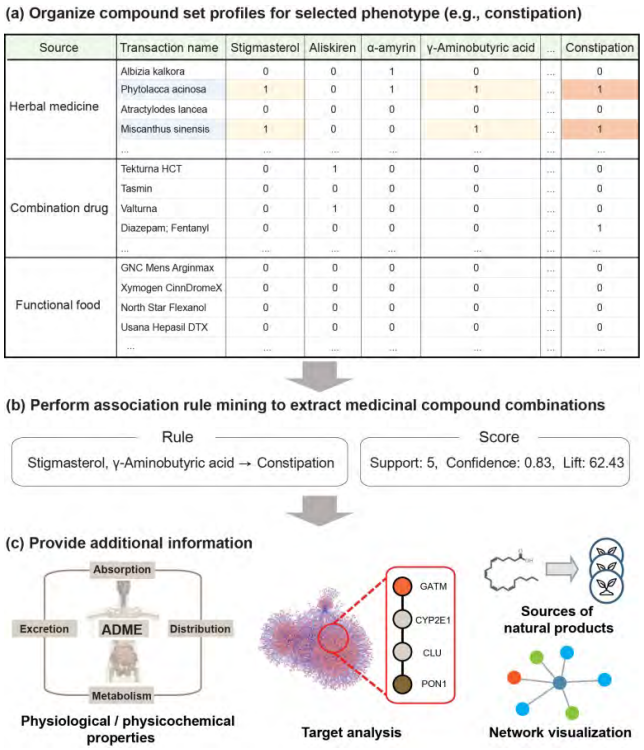

FIGURE 3. A computational framework for predicting medicinal combinations of natural product compounds. (a) Organizing compound set profiles based on compound composition and phenotypic effect information. (b) Performing association rule mining to extract frequent patterns between natural product combinations and phenotypes from medicinal multicomponent mixtures. (c) Providing additional information; physiological/physicochemical properties, target analysis, sources of natural products and network visualization.

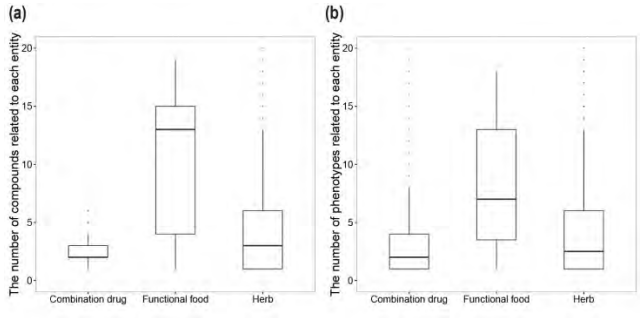

FIGURE 4. Distributions of herbs, functional food and combination drugs. (a) The distribution of the number of compounds related to each herb, functional food and combination drug. (b) The distribution of the number of phenotypes related to each herb, functional food and combination drug.

consist of two compounds [22]. Association rule mining makes predictions from subsets of coexisting items, which makes the estimation much easier. Moreover, 'lutein → fever' could be much easier to analyze than 'lutein, quercetin, biotin, spermine, etc. → fever'. As molecular mechanisms between natural compounds and phenotypes are very complex, it is important to simplify the prediction results. One other advantage of association rule mining is high-level interpretability, which enables intuitive explanations for the reasons why the inferred result occurred [57], [58]. It is useful to support further investigation of the inferred results in the functional food or drug development field.

3) STATISTICAL PARAMETERS OF ASSOCIATION RULES

A rule generated by association rule mining has the form of 'antecedent → consequent', where the antecedent is the combination of natural product compounds and the consequent is the target phenotype (Fig. 3b). For all association rules, we measured the significance with support, confidence and lift. The support value represents the ratio of instances containing both antecedent and consequent item-sets of the rule over the whole instances, which indicates the number of evidence items for the rule. Because the number of whole instances is constant, the support value is not expressed as a ratio, but rather is expressed as the number of instances containing both antecedent and consequent item-sets. For example, the support value of 10 means that there are ten medicinal materials containing the natural product combination that have effects on the target phenotype, as evidence of the rule. The confidence value is the ratio of instances containing both antecedent and consequent item-sets over the instances containing antecedent item-sets, which represents how often the rule is found to be true. For example, a confidence value of 0.5 means that the half of the medicinal materials containing the natural product combination have effects on the target phenotype. The lift value is the confidence value normalized by the number of instances containing consequent item-sets, which indicates the independence between the antecedent and consequent item-sets. For example, a lift value of 1 means that the natural product combination and target phenotype are completely independent, and a lift value of 2 means that the natural product combination is two times more dependent on target phenotype than random natural product combination.

4) ADDITIONAL INFORMATION FOR INFERRED ASSOCIATION RULES

We provide additional information of inferred associations to support further investigation on the inferred combinations of natural products (Fig. 3c). Compound properties and a list of sources of natural products are provided by searching the integrated information in COCONUT. For the target analysis, we adopted Dijkstra's algorithm to find the shortest paths between compound targets and phenotype-associated genes in the PPI network and signaling pathways [47]. All information related to inferred associations is visualized in the form of a network.

III. RESULTS

A. DATABASE CONTENTS

COCONUT contains information on 794,730 chemical compounds with calculated chemical properties. For medicinal multicomponent mixtures, data were collected for 20,259 prescriptions, 1,615 functional foods,

1,623 combination drugs and 8,492 herbs. Furthermore, we extracted 18,451 phenotype terms from the functional information of the prescription, combination drug, functional food, herb and chemical compound records. We also supplemented 39,286 genes that encode therapeutic targets or biomarkers. Each item of information was standardized based on the corresponding international identifiers (Table 3).

TABLE 3. COCONUT data entities.

Entity	International identifier	Num. of entities
Prescription	-	20,259
Combination drug	DCDB ID	1,623
Functional food	Trade/brand name	1,615
Herb	NCBI taxonomy ID	8,492
Compound	InChI	794,730
Phenotype	UMLS concept ID	18,451
Gene	Entrez gene ID	39,286

We structured the composition and functional information of the medicinal multicomponent mixtures as relationships between the entities. For example, the herbal composition of prescriptions is stored as 140,690 relationships, and the functional information of prescriptions is stored as 103,085 relationships. To support further investigation at the molecular level, we stored relationships among chemical compounds, phenotypes and genes (Table 4).

TABLE 4. Relationships of COCONUT.

Relationship (with an example)	Number of relationships
Prescription - herb (e.g., Sihogo - *Amomum villosum*)	140,690
Prescription - phenotype (e.g., Sihogo – diabetics)	103,085
Combination drug - compound (e.g., Zotrim – trimethoprim)	3,898
Combination drug - phenotype (e.g., Zotrim – urinary tract infections)	7,358
Functional food - herb (e.g., 4Life CitriShape – *garcinia*)	2,275
Functional food - phenotype (e.g., 4Life CitriShape – insulin sensitivity)	2,845
Herb - compound (e.g., *Melia azedarach* - trichilin H)	123,285
Herb - phenotype (e.g., *Melia azedarach* - abdominal pain)	103,782
Compound - phenotype (e.g., Lepirudin – thrombocytopenia)	1,372,288
Compound - gene (e.g., Lepirudin – F2)	5,298,455
Phenotype - gene (e.g., Fever – ASPG)	23,423
Inferred compound combination - phenotype (e.g., Eugenol, santalene – diarrhea)	899,476

B. ASSCIATIONS BETWEEN NATURAL PRODUCTS AND PHENOTYPES

The core data of COCONUT consists of 899,476 associations between 23,036 natural product combinations and 376 target phenotypes. For all associations, we provide support, confidence and lift values as the significance scores (Table 5). Furthermore, we examined oral bioavailability and molecular target paths to provide additional evidence, which helps us to select drug candidates among inferred associations (Table 6). For oral bioavailability, we check whether the natural products in inferred associations satisfy RO5. For all associations, 496 natural products and 266,263 natural product combinations satisfied the RO5. Next, we check whether natural product combinations are associated with the target phenotype on the molecular network. For 193,332 associations of 96 phenotypes, 26,825 associations have direct or indirect connections between compound targets and phenotype-associated genes. For example, the combination of L-(+)-arginine and biotin was inferred to be associated with diarrhea. This association is considered to be an important candidate because it is given a high confidence and lift values (conf. = 1.0 and lift = 23.63). Further analysis of COCONUT revealed that both L-(+)-arginine and biotin satisfy RO5.

TABLE 5. Examples of inferred associations.

Inferred association	Supp.	Conf.	Lift
Eugenol, Santalene→Diarrhea	6	1.00	23.63
β-sitosterol, Campesterol'→Cough	11	0.73	7.77
(+)-beta-Pinene, Piperitone→Vomiting	6	1.00	18.87
Alanine, Leucine→Hypertention	5	0.14	28.75
Leucine, DL-Lysine→Agitation	5	0.15	14.36
(+)-(R)-limonene, Caffeic acid→Cold	6	0.85	18.88
Vitamin C,Thiamine chloride→Nausea	5	0.07	11.97
Eugenol, Benzyl alcohol→Stroke	5	0.71	46.31
(+)-borneol, l-Bornyl acetate→Spleen	6	1.00	18.45
Pyrimethamine, Sulfadoxine→Malaria	6	0.75	32.98

TABLE 6. Oral bioavailability and molecular target path information for top 10 phenotypes with respect to the number of associations.

Phenotype	Num. of associations	RO [a]	PG [b]	RO & PG
Diarrhea	11042	3130	773	407
Headache	9321	2879	1259	675
Vomiting	9216	2701	560	341
Abdominal pain	9101	2585	194	110
Asthma	7982	2341	2337	1086
Fever	7715	2435	1773	1056
Bleeding	6822	2019	1055	822
Amenorrhea	6400	1893	318	530
Edema	6033	1759	1673	680
Constipation	5506	1698	699	368

[a] The number of associations satisfying RO5.
[b] The number of associations that compound targets are connected to phenotype-associated genes.

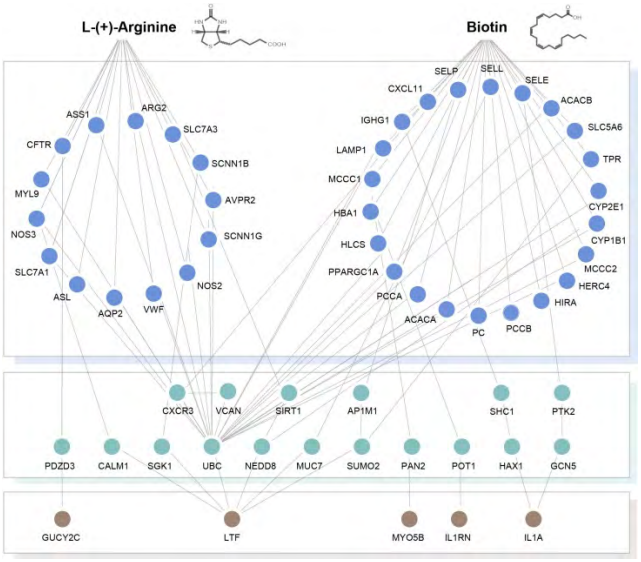

FIGURE 5. An example of target analysis for the inferred association 'L-(+)-arginine, biotin → diarrhea'. Shortest paths from compound targets (blue) to phenotype-associated genes (brown) are investigated in PPI network.

Moreover, we found that many of both targets of L-(+)-arginine and biotin are connected to the diarrhea-associated gene from the shortest path analysis (Fig. 5). Based on this approach, the COCONUT database can be used as a preliminary tool to identify the medicinal compound combination candidates from a large number of natural products.

C. PERFORMANCE EVALUATION

1) ROBUSTNESS TESTING

To verify whether inferred associations were not vulnerable with respect to variations of the dataset, we examined their robustness among different datasets with 4-fold cross-validation (Table 7). For this, the compound set profiles were divided into four folds. The instances having a therapeutic effect on the target phenotype were equally divided in each fold. Then, each three-fold was used to infer the associations between natural product combinations and target phenotype, while the remaining one was used to evaluate the robustness of the inferred associations. The robustness score of an association rule is defined as n/N, where N is the number of instances having the natural product combination in the test set, and n is the number of instances having the natural product combination with effects on the target phenotype in the test set. Consequently, the robustness score measures whether the inferred associations have a robust explanatory power for known indications of medicinal materials in the data set. For the target phenotype, associations with confidence values higher than 0.7 were selected. Then, the robustness of the target phenotype was calculated as the average robustness score of the selected associations. As a result, the average robustness ($r_{\text{avg}} = 0.733$) for 41 target phenotypes was relatively high, considering the proportion of compounds shared by both test and training sets (22%). This indicates that the inferred associations are robust in the data set.

TABLE 7. Robustness of top 10 ranked phenotypes.

Phenotype	Num. of training set	Average num. of rule	Robustness
Cardiac arrest	185	103.5	0.908
Cholera	13	67.75	0.889
Jaundice	101	149.25	0.885
Diabetes	12	4.25	0.875
Eczema	72	47.25	0.845
Hypercholesterolemia	5	4.5	0.837
Fall	170	14.25	0.826
Edema	449	5670.75	0.822
Epistaxis	59	22.25	0.820
Dysmenorrhea	51	56.5	0.817

2) EVALUATION BASED ON MOLECULAR NETWORK ANALYSIS

To evaluate the significance scores of the inferred associations, we selected the top 5% highest scoring associations and the bottom 5% lowest scoring associations based on the support, confidence and lift, respectively. In the previous study, the relative proximity (z_c) was proposed to quantify the relationship between compounds and phenotype genes in the molecular network [59]. They found that a compound tends to have a phenotype when compound targets and phenotype-associated genes are closely located on a molecular network. For each compound-phenotype pair, they compare the distance between compound targets and phenotype-associated genes to the random expectation distances, which were calculated by randomly selecting the phenotype-associated genes within the molecular network. This study also found that closest measure, which calculated the distance based on the average shortest path between the compound targets and the nearest phenotype-associated gene, showed the best performance in predicting drug efficacy, when compared with the shortest, kernel, centre and separation measures. Therefore, in this study, the relative proximity using the closest measure was used to assess the prediction results according to the significance scores. We compared the relative proximity values of the selected six association sets and the verified drug-phenotype set collected from DrugBank [26]. The molecular network was constructed by using the PPI information collected from BioGird [35] (Fig. 6).

The average relative proximity values of all sets are lower than zero, which means that compound targets and phenotype-associated genes are closer in the molecular network than randomly selected gene sets (avg.$z_c > 0$). Moreover, the average relative proximity values of associations having high support, confidence and lift values (avg. $z_c = -0.71$, -0.87 and -0.91, respectively) are comparable to the verified drug-phenotype set (avg. $z_c = -1.54$).

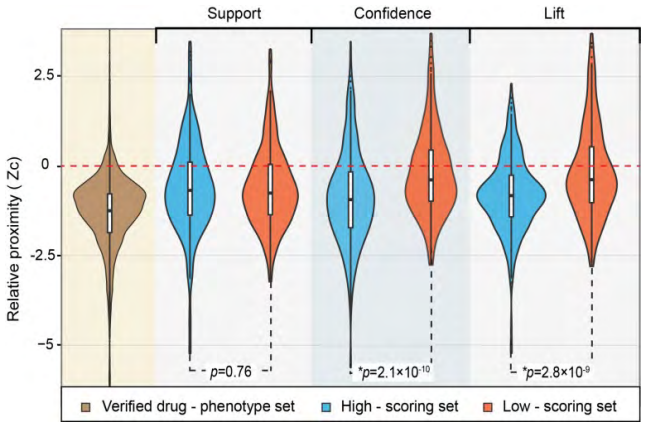

FIGURE 6. The distribution of relative proximity for inferred associations and verified drug-phenotype set. The inferred associations were ranked by the support, confidence and lift and divided into two independent sets by selecting the top 5% and bottom 5% associations, respectively. The red dotted line indicates the relative proximity of pure randomness ($z_c = 0$).

For associations selected by confidence value, the average relative proximity of a high-scoring set (avg. $z_c = -0.87$) is higher than the low-scoring set (avg. $z_c = -0.38$). Next, we performed a Mann-Whitney U test and calculated the corresponding p-values to check the significant difference between the high-scoring and low-scoring sets [60]. A p-value of the Mann-Whitney U test lower than 0.05 was considered statistically significant. From the result, we found that the difference between high- and low-scoring sets is statistically significant (p-value $= 2.1 \times 10^{-10}$). Similarly, for associations selected by the lift value, the average relative proximity of high-scoring sets (avg. $z_c = -0.91$) is higher than the low-scoring sets (avg. $z_c = -0.37$), and this difference is also significant (p-value $= 2.8 \times 10^{-9}$). These observations indicate that the associations with high confidence or lift imply significant interplay between compound targets and phenotype-associated genes. However, for associations selected by the support value, there is no significance difference (p-value $= 0.76$) between high- and low-scoring sets (avg. $z_c = -0.71$ and -0.69, respectively).

3) EVALUATION BASED ON EXTERNAL LITERATURE

In this study, the associations between natural products and phenotypes were inferred from the known efficacy of medicinal multicomponent mixtures and their chemical compositions. In other words, PubMed information for direct evidence between natural products and the phenotypes was not used during the inference process. Therefore, to evaluate the inferred associations, we employed PubMed as the independent external dataset and searched literature evidence containing the inferred associations. Using the selected six association sets described in the previous section, we checked whether the associations with high support, confidence and lift values have more evidence in the external literature than the associations with low values. To do this, co-occurrences (n_c) of compounds and a phenotype in each association were counted in 13,200,786 PubMed abstracts that were published from 1950 to 2013 [61]. For associations selected by the confidence value, the average co-occurrence count of the high-scoring sets ($n_c = 39.17$) is 6.9 times larger than the low-scoring sets ($n_c = 5.38$). Similarly, for associations selected by the lift value, the average co-occurrence count of the high-scoring sets ($n_c = 33.7$) is 10.9 times larger than the low-scoring sets ($n_c = 3.08$). These results show that confidence and lift can be used as parameters for identifying significant associations (Table 8). We also performed a Mann-Whitney U test and calculated the corresponding p-values to check the significant difference of literature evidence between the high- and low-scoring sets.

TABLE 8. External literature validation.

		Co-occurrence	Jaccard index	Fisher's exact test[a]
Support	High	7.94	1.69×10^{-4}	8
	Low	7.35	9.63×10^{-5}	5
Confidence	High	39.17	1.89×10^{-3}	20
	Low	5.38	1.15×10^{-4}	2
Lift	High	33.70	3.04×10^{-3}	22
	Low	3.08	1.03×10^{-4}	4
Mann-Whitney U test, p-value	Supp.	0.388	0.076	0.113
	Conf.	<0.001	<0.001	<0.001
	Lift	<0.001	<0.001	<0.001

[a] The number of significant associations satisfying the Fisher's exact test p-value threshold (p-value < 0.001).

Co-occurrence values do not take the frequencies of individual terms into account; they were normalized as the Jaccard index (JI) about the frequencies of individual terms. For the association sets selected by the confidence and lift values, the average Jaccard index values of high-scoring sets ($JI = 1.89 \times 10^{-3}$ and 3.04×10^{-3}, respectively) were markedly higher than those of the low-scoring sets ($JI = 1.15 \times 10^{-4}$ and 1.03×10^{-4}, respectively). Furthermore, we investigated the number of significant associations (n_f) by performing Fisher's exact test (p-value < 0.001). Fisher's exact test can assess the null hypothesis of independence by applying the hypergeometric distribution of the numbers in a contingency table [62]. To obtain a Fisher's test value of each association, the number of PubMed abstracts was counted based on whether they included the compound and whether they included the target phenotype. For association sets selected by the confidence and lift, the numbers of significant associations of the high-scoring sets ($n_f = 20$ and 22, respectively) were markedly larger than those of the low-scoring sets ($n_f = 2$ and 4, respectively). Additionally, the p-values of the Mann-Whitney U test indicated that the difference in documented evidence between the high- and low-scoring sets was significant.

However, for association sets selected by the support value, there was no significant difference between high-scoring and

low-scoring sets in the average co-occurrence count, average Jaccard index and the number of significant associations based on Fisher's exact test. This result could be due to the characteristics of support, which only indicate the appearance of the associations as mentioned in the former section. In conclusion, we suggest that inferred associations with higher significance scores can be used as potential therapeutic compound combinations for future studies.

D. WEB INTERFACE AND USE CASE

We implemented a workbench website for the COCONUT database to provide two services. First, the search service enables exploring comprehensive information about natural products (Fig. 7). The database search service can be accessed through the 'Search' tab on the main page (Fig. 7a). After selecting the entity type, users can search instances by typing query keywords or selecting an instance in the alphabetic table (Fig. 7b). Detailed information of the query instance is represented in the search result (Fig. 7c). The information on related entities is also provided in the association table with source references. Therefore, users can know the source for the information of interest. The related entities can be investigated through hyperlinks in the association table or the visualized network (Fig. 7d). As an example, *Panax ginseng* can be queried in the 'Herb' tab of the search service. From the result, users can investigate that *Panax ginseng* has been used in the treatment of diarrhea, impotence and intestinal pain, and it is composed of various natural product compounds, such as ginsenoyne I, lutein and trifolin.

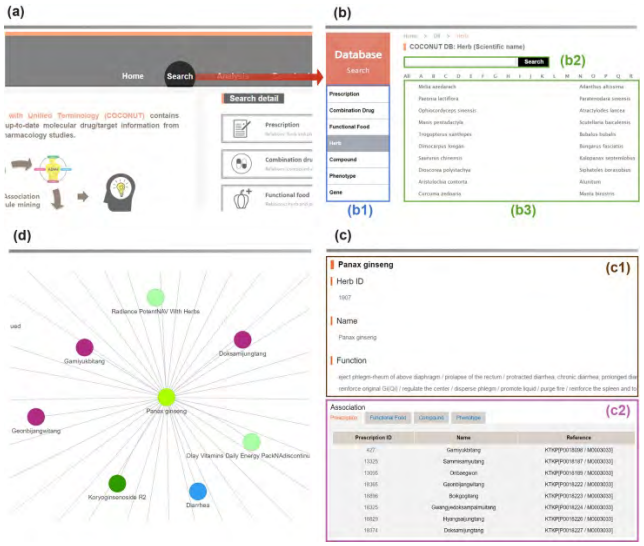

FIGURE 7. 'Search' service in the web interface. (a) The search query page can be accessed through the 'Search' tab on the main page (red circle). (b) For seven entity types (b1), users can search detailed information by keywords (b2) or the alphabetic table (b3). (c) For a query, COCONUT provides detailed information of the resulting instance (c1) and a list of associations with other entities (c2). (d) The relationships between selected instance and other entities are summarized through a network.

Second, the analysis service offers the users the ability to analyze inferred associations between natural product

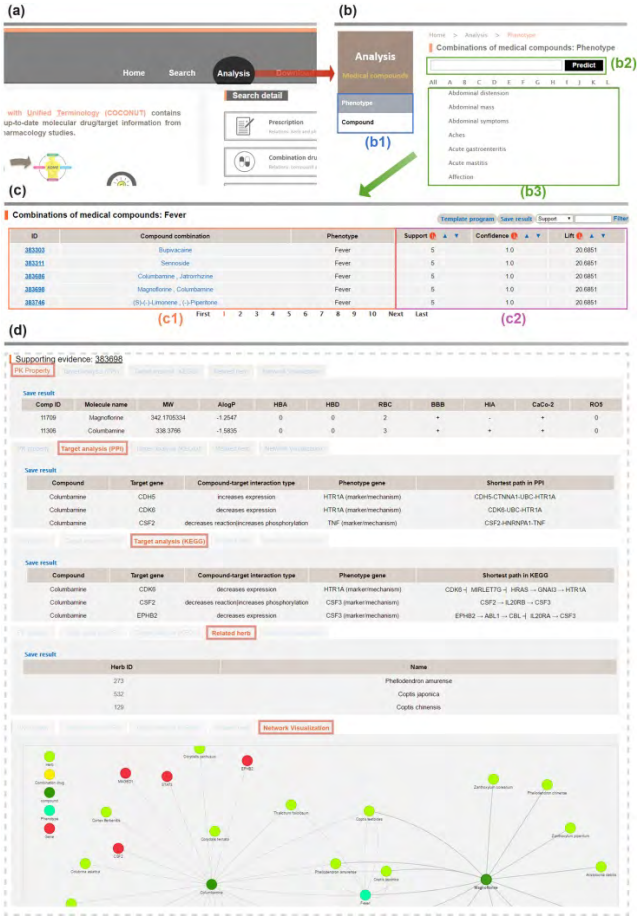

FIGURE 8. 'Analysis' service in the web interface. (a) The analysis query page can be accessed through the 'Analysis' tab on the main page (red circle). (b) For target phenotype or compound of interest (b1), users can search detailed information by keywords (b2) or the alphabetic table (b3). (c) For a query, COCONUT provides a list of associations between natural product combinations and the target phenotype (c1) with support, confidence and lift scores (c2). (d) Additional information on the selected association, such as pharmacokinetic property, target analysis, related herbs and network visualization, is provided.

combinations and phenotypes (Fig. 8). The analysis service can be accessed through the 'Analysis' tab on the main page (Fig. 8a). The user can inquire about natural product combinations for a natural product or phenotype of interest. Like the database search, users can search for information by keyword or the alphabetic table (Fig. 8b). In the analysis result, a list of associations between natural product combinations and phenotype is represented along with statistical significance parameters, such as support, confidence and lift (Fig. 8c). By adjusting the threshold of each parameter, users can focus on reliable associations. Moreover, users can analyze the association of interest more minutely in terms of pharmacokinetic property, target analysis, related herbs and network visualization (Fig. 8d). For example, fever can be queried in the 'Phenotype' tab of the analysis service. Associations between natural product combinations and fever are represented as query results. By sorting the associations with a confidence value, users can observe statistically

TABLE 9. Information on *Panax Ginseng* in COCONUT.

Known efficacy		
Phenotype term (UMLS ID)	Associated genes	Other herbs associated with the phenotype
Impotence (C0242350)	BCL2, EDN1, NOS3, PRL, VEGFA, SHBG	*Morinda officinalis* (266091), *Cistanche salsa* (161396)
Anaemia (C0002871)	AGT, ASPG, CSF2, GSR, HP, IGF2, IFNA2	*Angelica sinensis* (165353), *Prismatomeris tetrandra* (110690)
Headache (C0018681)	CSF3, GNRH1, IL6, OXT, IFNA2	*Aristolochia contorta* (266420), *Pinctada fucata* (50426)
Asthma (C0004096)	ADCY2, CAT, BCL6, AREG, HNMT, GSTP1	*Desulfovibrio desulfuricans* (876), *Citrus unshiu* (55188)
Diabetes mellitus (C0011849)	FN1, ATP6, INS, ADRB1	*Bombyx mori* (235), *Lepus mandshuricus* (392)
…		

Contained natural product information			
Natural product	Associated genes	Associated phenotypes (UMLS ID)	Other herbs containing the natural product (NCBI taxonomy ID)
Lutein	ALS2, APC, BAX, BCL2, CAT, CCND1, AQR, CCS	Acatalasia (C0268419), Asthma (C0004096), Anoxia (C0003130), Cachexia (C0006625)	*Lemna pauciostata* (89585), *Rosa rugose* (74645), *Citrus junos* (135197)
Neoxanthin	AIFM, CASP3, CASP8, CASP9	Brain Ischemia (C0007786), Edema (C0013604), Calcinosis (C0006663), Lung Tumor (C0024121)	*Brassica juncea* (3707), *Ginkgo biloba* (3311), *Hibiscus syriacus* (376), *Glycine max* (3847)
Biotin	PCCB, AVD, HLCS, SLC5A6, MCCC1, PCCA	Myositis (C0027121), Colorectal tumors (C0009404), Diabetes mellitus (C0011849), Pleurisy (C0497354)	*Allium sativum* (4682), *Glycine max* (3847) *Apis cerana* (334), *Triticum aestivum* (4565)
Adenosine	ADORA2A, ADA, ALPI, ADORA2B, ADCY10	Bronchospasm (C0006266), Backache (C0004604), Nausea (C0027497), Dyspnea (C0013404)	*Rehmannia glutinosa* (99300), *Isatis tinctoria* (161756), *Cucumis melo* (3656)
Maltol	GABRA1, GABRB1, AHR, MMP1, HECA	Autism (C0004352), Ocular Hypertension (C0028840), Breast neoplasms (C1458155), Ataxia (C0004134)	*Scutellaria baicalensis* (65409), *Macrothelypteris oligophlebia* (692138)
…			

significant associations, such as 'Magnoflorine, Columbamine → Fever'. Additional information on the association can be obtained in the bottom panel. From the PK property tab, users can find that both magnoflorine and columbamine satisfy RO5. All shortest paths between target genes and phenotype-associated genes in a molecular network, such as CDH5 (columbamine target gene) – CTNNA1 – UBC – HTR1A (fever-associated gene), can be investigated using the target analysis tab. Furthermore, users can examine the list of herbs that contain both magnoflorine and columbamine. Entities, including herbs, combination drugs, functional foods and genes, related to each compound are visualized through a network.

IV. DISCUSSION

Natural products and their combinations have distinctive advantages in drug and functional food discovery. Since they are secondary metabolites of other organisms, they are more likely to have bioactivities, and they present unique structural diversities from which we can discover novel therapeutic compounds [63], [64]. Therefore, a better understanding of the natural products through scientific analysis will provide new insights into the use of natural products as medicine.

Our study has strengths in two aspects. First, the COCONUT database provides comprehensive information about natural products in a structured and standard form. Compared to existing databases such as KTKP, TCMID and KAMPO, using a simple query, researchers can obtain the prescription, herb, compound and gene information associated with a particular phenotype because all information in COCONUT is structured. This process allows researchers to easily collect data when designing *in vitro* and *in silico* experiments for specific natural products or phenotypes. In addition, COCONUT provides standardized information; thus, it can minimize confusion or misreading of information and improve interoperability when researchers use COCONUT with other external databases. For example, when researchers identify medicinal compounds from *Panax ginseng*, they could use COCONUT in an initial stage of the experiment to collect comprehensive information about the *Panax ginseng* with standardized international identifiers, such as the known efficacy, the natural products they contain, and the associated genes (Table 9). Second, we provide promising candidates of medicinal combinations of natural products to support combination drug or functional food discovery studies. Most of the previous studies on finding medicinal agents from natural products were performed by *in vitro* assessment. However, large-scale experiments are required for a large number of natural products and their combinations. Therefore, *in silico* approaches have been proposed, primarily based on molecular analysis. However, many natural

TABLE 10. Combinations of natural product included in panax ginseng, which is expected to be effective for fever.

Inferred association	Supp.	Conf.	Lift	Medicinal materials containing inferred combinations of natural products
Stigmasterol-β-d-glucoside , serine→Fever	6	0.85	17.73	*Panax ginseng, Miscanthus sinensis, Oryza sativa, Scutellaria baicalensis, Glycine max, Rehmannia glutinosa, Angelica sinensis*
Eugenol, γ-sitosterol→Fever	6	0.85	17.73	*Panax ginseng, Lonicera japonica, Cinnamomum aromaticum, Perilla frutescens, Plantago asiatica, Scutellaria baicalensis*
(24R)-Ergost-2-en-3-ol, leucine→Fever	5	0.83	17.23	*Panax ginseng, Glycine max, Rehmannia glutinosa, Miscanthus sinensis, Oryza sativa, Scutellaria baicalensis*
β-sitosterol, β-amyrin→Fever	5	0.83	17.23	*Panax ginseng, Triticum aestivum, Glycyrrhiza uralensis, Glycine max, Plantago asiatica, Lycium chinense*
Carvacrol, phenol→Fever	5	0.83	17.23	*Panax ginseng, Morus alba, Angelica sinensis, Solanum melongena, Plantago asiatica, Hordeum vulgare*
Isofucosterol, lupeol→Fever	5	0.82	17.23	*Panax ginseng, Lycium chinense, Glycine max, Cucumis melo, Sorghum bicolor, Cucumis sativus*
Stigmasterol-β-d-glucoside, (+/-)-α-Pinene →Fever	6	0.75	15.51	*Panax ginseng, Lonicera japonica, Bupleurum chinense, Artemisia annua, Perilla frutescens, Plantago asiatica, Foeniculum vulgare, Oryza sativa*
(24R)-Ergost-2-en-3-ol, campesterol→Fever	8	0.66	13.79	*Panax ginseng, Miscanthus sinensis, Scutellaria baicalensis, Perilla frutescens, Morus alba, Eleutherococcus sessiliflorus, Aloe vera*
Stigmasterol, dauricine→Fever	5	0.62	12.92	*Panax ginseng, Akebia quinata, Sigesbeckia orientalis, Rehmannia glutinosa, Taraxacum officinale, Ziziphus jujuba*
L-(+)-Aspartic acid, γ-sitosterol→Fever	5	0.62	12.92	*Panax ginseng, Rehmannia glutinosa, Scutellaria baicalensis, Triticum aestivum, Ginkgo biloba, Matteuccia struthiopteris*
β-sitosterol, Linalool	8	0.61	12.72	*Panax ginseng, Scutellaria baicalensis, Gardenia jasminoides, Perilla frutescens, Houttuynia cordata, Plantago asiatica, Artemisia annua*

products do not have molecular structural information, and their target protein information remains mostly unknown. More importantly, the conventional *in silico* methods are aimed at investigating single agents. Hence, there is no method to predict the medicinal combinations of natural products. Our approach is different from previous *in silico* methods in that it finds combination patterns of natural products from the accumulated medicinal materials using a data-driven approach. In the performance evaluation, we performed molecular network analysis. There is no gold-standard dataset for the therapeutic effects of combinations of natural products; thus, we investigate whether there is a significant relationship between the targets of the natural products and the phenotype-associated genes in the molecular network. From the results, we confirmed that the inferred associations with high scores between natural products and phenotypes have significant associations in the molecular network. We also found that the high-scoring associations have more evidence in the external literature. This indicates that the proposed data-driven analysis enabled us to identify medicinal candidate effects of natural products. Our results can help researchers conduct further *in vitro* and in *silico* experiments by filtering natural products or herbs from a large number of candidates. For example, when researchers study the effects of *Panax ginseng* on fever, they can utilize our results to find promising combinations of natural products found in *Panax ginseng* (Table 10). Additionally, the information of chemical properties and path analysis in the COCONUT database helps to select the specific candidates before they carry out further studies.

There are some additional considerations for improving our work. First, although this study analyzed the shortest path between the known natural product targets and the phenotype associated genes, the results are insufficient due to the lack of molecular information. However, this limitation can be resolved with further experiments and improved techniques. We expect that more accurate predictions can be made in additional *in silico* studies. For example, previous studies have demonstrated that associations between compounds and targets can be predicted by investigating propagated compound effects on the disease genes in a molecular network [65], [66]. Based on this approach, researchers can estimate potential mechanism of actions of natural product combinations by calculating overlap effect in a molecular network based on COCONUT information. Second, our database contains associations between medicinal materials and phenotypes, but there is no specification about the types of associations, such as cause, treat or prevent. To overcome the limitation, we will improve the text mining method to extract the detailed association types. Finally, the database volume will be periodically expanded by integrating information from various sources. At present, COCONUT mainly

focused on herbal medicine information of northeastern Asia. Therefore, information can be biased toward a particular region or country. In the future, we will add herbal medicine information from various regions, such as India, Australia and America. Additionally, we are planning to include information not only from conventional databases but also from biomedical literature and experiment results. With the updated information, we will develop and apply various methods for more reliable prediction of the medicinal combinations of natural products. Ultimately, we believe that COCONUT will be used as a valuable resource in eliminating the bottlenecks in the current natural product research by combining with various biological information sources.

V. CONCLUSION

Natural products have been used as important sources of herbal medicine and modern drug development. COCONUT is useful for investigating natural products and their corresponding information, such as major sources, activities and efficacies. More importantly, COCONUT enables us to perform large-scale analysis on the medicinal use of natural product combinations. We believe that COCONUT will be a major bioinformatics resource for polypharmacology studies and will be of interest to pharmacologists, toxicologists and computational biologists by providing clues for the prediction of medicinal combinations from a wide range of natural products.

ACKNOWLEDGMENT

(Sunyong Yoo and Suhyun Ha are co-first authors.)

REFERENCES

[1] G. M. Cragg and D. J. Newman, "Natural products: A continuing source of novel drug leads," *Biochim. Biophys. Acta-Gen. Subjects*, vol. 1830, no. 6, pp. 3670–3695, 2013.

[2] A. Bauer and M. Brönstrup, "Industrial natural product chemistry for drug discovery and development," *Natural Product Rep.*, vol. 31, no. 1, pp. 35–60, 2014.

[3] D. J. Newman and G. M. Cragg, "Natural products as sources of new drugs from 1981 to 2014," *J. Natural Products*, vol. 79, no. 3, pp. 629–661, 2016.

[4] R. King, "Collaborating with traditional healers for HIV prevention and care in sub-Saharan Africa: Suggestions for programme managers and field workers," Joint United Nations Programme HIV/AIDS (UNAIDS), Geneva, Switzerland, Tech. Rep., Nov. 2006.

[5] Z. Qi and E. Kelley, "The WHO traditional medicine strategy 2014–2023: A perspective," *Science*, vol. 346, no. 6216, pp. S5–S6, 2014.

[6] *General Guidelines for Methodologies on Research and Evaluation of Traditional Medicine*, World Health Org., Geneva, Switzerland, 2000.

[7] J. Jia, F. Zhu, X. Ma, Z. W. Cao, Y. X. Li, and Y. Z. Chen, "Mechanisms of drug combinations: Interaction and network perspectives," *Nature Rev. Drug Discovery*, vol. 8, pp. 111–128, Jun. 2009.

[8] T.-C. Chou, "Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies," *Pharmacol. Rev.*, vol. 58, no. 3, pp. 621–681, 2006.

[9] J. Lehár *et al.*, "Synergistic drug combinations tend to improve therapeutically relevant selectivity," *Nature Biotechnol.*, vol. 27, pp. 659–666, Jul. 2009.

[10] L. Wang *et al.*, "Dissection of mechanisms of Chinese medicinal formula Realgar-*Indigo naturalis* as an effective treatment for promyelocytic leukemia," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 12, pp. 4826–4831, 2008.

[11] H. Wagner and G. Ulrich-Merzenich, "Synergy research: Approaching a new generation of phytopharmaceuticals," *Phytomedicine*, vol. 16, nos. 2–3, pp. 97–110, 2009.

[12] Y. Sun, K. Xun, Y. Wang, and X. Chen, "A systematic review of the anticancer properties of berberine, a natural product from Chinese herbs," *Anti-Cancer Drugs*, vol. 20, no. 9, pp. 757–769, 2009.

[13] R. Xue, Z. Fang, M. Zhang, Z. Yi, C. Wen, and T. Shi, "TCMID: Traditional Chinese medicine integrative database for herb molecular mechanism analysis," *Nucl. Acids Res.*, vol. 41, no. D1, pp. 1089–1095, 2012.

[14] C. Y.-C. Chen, "TCM Database@Taiwan: The world's largest traditional Chinese medicine database for drug screening *in silico*," *PLoS ONE*, vol. 6, p. e15939, Jan. 2011.

[15] H. Ye *et al.*, "HIT: Linking herbal active ingredients to targets," *Nucl. Acids Res.*, vol. 39, pp. D1055–D1059, Jan. 2011.

[16] S.-K. Kim, S. Nam, H. Jang, A. Kim, and J.-J. Lee, "TM-MC: A database of medicinal materials and chemical compounds in Northeast Asian traditional medicine," *BMC Complementary Alternative Med.*, vol. 15, p. 218, Jul. 2015.

[17] J.-H. Lee *et al.*, "PharmDB-K: Integrated bio-pharmacological network database for traditional Korean medicine," *PLoS ONE*, vol. 10, no. 11, p. e0142624, 2015.

[18] M. Mangal, P. Sagar, H. Singh, G. P. S. Raghava, and S. M. Agarwal, "NPACT: Naturally occurring plant-based anti-cancer compound-activity-target database," *Nucl. Acids Res.*, vol. 41, no. D1, pp. D1124–D1129, 2013.

[19] P. Banerjee, J. Erehman, B.-O. Gohlke, T. Wilhelm, R. Preissner, and M. Dunkel, "Super Natural II—A database of natural products," *Nucl. Acids Res.*, vol. 43, pp. D935–D939, Jan. 2015.

[20] K. Jensen, G. Panagiotou, and I. Kouskoumvekaki, "NutriChem: A systems chemical biology resource to explore the medicinal value of plant-based foods," *Nucl. Acids Res.*, vol. 43, no. D1, pp. D940–D945, 2015.

[21] U. A. Ashfaq, A. Mumtaz, T. ul Qamar, and T. Fatima, "MAPS database: Medicinal plant activities, phytochemical and structural database," *Bioinformation*, vol. 9, no. 19, pp. 993–995, 2013.

[22] Y. Liu, Q. Wei, G. Yu, W. Gai, Y. Li, and X. Chen, "DCDB 2.0: A major update of the drug combination database," *Database*, vol. 2014, p. bau124, Jan. 2014.

[23] *KTKP*. Accessed: Mar. 2, 2015. [Online]. Available: http://www.koreantk.com/

[24] *Kampo*. Accessed: Mar. 2, 2015. [Online]. Available: http://kampo.ca/

[25] *FooDB*. Accessed: Mar. 2, 2015. [Online]. Available: http://foodb.ca/

[26] V. Law *et al.*, "DrugBank 4.0: Shedding new light on drug metabolism," *Nucl. Acids Res.*, vol. 42, no. D1, pp. D1091–D1097, 2014.

[27] *BFN*. Accessed: Mar. 2, 2015. [Online]. Available: http://biofood.or.kr

[28] A. P. Davis *et al.*, "The Comparative Toxicogenomics Database's 10th year anniversary: Update 2015," *Nucl. Acids Res.*, vol. 43, no. D1, pp. D914–D920, 2015.

[29] J. E. Gillen, T. Tse, N. C. Ide, and A. T. McCray, "Design, implementation and management of a web-based data entry system for ClinicalTrials.gov," *Stud Health Technol. Inf.*, vol. 107, pp. 1466–1470, Jan. 2004.

[30] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," *Nucl. Acids Res.*, vol. 44, no. D1, pp. D1075–D1079, 2015.

[31] F. Zhu *et al.*, "Therapeutic target database update 2012: A resource for facilitating target-oriented drug discovery," *Nucl. Acids Res.*, vol. 40, no. D1, pp. D1128–D1136, 2011.

[32] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, "BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology," *Nucl. Acids Res.*, vol. 44, no. D1, pp. D1045–D1053, 2016.

[33] S. Günther *et al.*, "SuperTarget and Matador: Resources for exploring drug-target relationships," *Nucl. Acids Res.*, vol. 36, pp. D919–D922, Jan. 2008.

[34] M. Kuhn *et al.*, "STITCH 4: Integration of protein–chemical interactions with user data," *Nucl. Acids Res.*, vol. 42, no. D1, pp. D401–D407, Jan. 2014.

[35] A. Chatr-Aryamontri *et al.*, "The BioGRID interaction database: 2015 update," *Nucl. Acids Res.*, vol. 43, no. D1, pp. D470–D478, 2015.

[36] M. Kanehisa *et al.*, "From genomics to chemical genomics: New developments in KEGG," *Nucl. Acids Res.*, vol. 34, pp. D354–D357, Jan. 2006.

[37] S. Federhen, "The NCBI Taxonomy database," *Nucl. Acids Res.*, vol. 40, no. D1, pp. D136–D143, 2012.

[38] H. E. Pence and A. Williams, "ChemSpider: An online chemical information resource," *J. Chem. Edu.*, vol. 87, no. 11, pp. 1123–1124, 2010.

[39] D. W. Huang, B. T. Sherman, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki, "DAVID gene ID conversion tool," *Bioinformation*, vol. 2, no. 10, p. 428, 2008.

[40] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program," in *Proc. AMIA Symp.*, 2001, p. 17.
[41] O. Bodenreider, "The Unified Medical Language System (UMLS): Integrating biomedical terminology," *Nucl. Acids Res.*, vol. 32, pp. D267–D270, Jan. 2004.
[42] A. R. Aronson, *MetaMap: Mapping Text to the UMLS Metathesaurus*. Bethesda, MD, USA: NLM, NIH, DHHS, 2006, pp. 1–26.
[43] S. Yoo *et al.*, "*In silico* profiling of systemic effects of drugs to predict unexpected interactions," *Sci. Rep.*, vol. 8, p. 1612, Jan. 2018.
[44] J. D. Osborne, B. Gyawali, and T. Solorio. (2014). "Evaluation of YTEX and MetaMap for clinical concept recognition." [Online]. Available: https://arxiv.org/abs/1402.1668
[45] K. B. Cohen, T. Christiansen, and L. E. Hunter, "MetaMap is a superior baseline to a standard document retrieval engine for the task of finding patient cohorts in clinical free text," in *Proc. TREC*, 2011, pp. 1–2.
[46] D. A. Hanauer *et al.*, "Applying MetaMap to Medline for identifying novel associations in a large clinical dataset: A feasibility analysis," *J. Amer. Med. Informat. Assoc.*, vol. 21, pp. 925–937, Sep. 2014.
[47] N. Bhatia, N. H. Shah, D. L. Rubin, A. P. Chiang, and M. Musen, "Comparing concept recognizers for ontology-based indexing: MGREP vs. MetaMap," in *Proc. AMIA Summit Transl. Bioinform.*, San Francisco, CA, USA, 2009, pp. 1–6.
[48] K. Chard, M. Russell, Y. A. Lussier, E. A. Mendonça, and J. C. Silverstein, "A cloud-based approach to medical NLP," in *Proc. AMIA Annu. Symp.*, 2011, p. 207.
[49] J. Shen, F. Cheng, Y. Xu, W. Li, and Y. Tang, "Estimation of ADME properties with substructure pattern recognition," *J. Chem. Inf. Model.*, vol. 50, no. 6, pp. 1034–1041, 2010.
[50] H. Pham The *et al.*, "In silico prediction of Caco-2 cell permeability by a classification QSAR approach," *Mol. Inform.*, vol. 30, no. 4, pp. 376–385, 2011.
[51] *CDK Descriptor Calculator GUI*. Accessed: Aug. 3, 2017. [Online]. Available: http://www.rguha.net/code/java/cdkdesc.html
[52] N. M. Nasrabadi, "Pattern recognition and machine learning," *J. Electron. Imag.*, vol. 16, no. 4, p. 049901, 2007.
[53] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.
[54] S. Dzeroski, "Relational data mining," in *Data Mining and Knowledge Discovery Handbook*. Springer, 2009, pp. 887–911.
[55] M. Hegland, "Algorithms for association rules," in *Advanced Lectures on Machine Learning*. Springer, 2003, pp. 226–234.
[56] C. Rudin, B. Letham, and D. Madigan, "Learning theory analysis for association rules and sequential event prediction," *J. Mach. Learn. Res.*, vol. 14, pp. 3441–3492, Nov. 2013.
[57] E. Georgii, L. Richter, U. Rückert, and S. Kramer, "Analyzing microarray data using quantitative association rules," *Bioinformatics*, vol. 21, pp. ii123–ii129, Jan. 2005.
[58] E. Triantaphyllou and G. Felici, Eds., *Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques*, vol. 6. Boston, MA, USA: Springer, 2006.
[59] E. Guney, J. Menche, M. Vidal, and A.-L. Barábasi, "Network-based *in silico* drug efficacy screening," *Nature Commun.*, vol. 7, Feb. 2016, Art. no. 10331.
[60] Z. Birnbaum, "On a use of the Mann-Whitney statistic," in *Proc. 3rd Berkeley Symp. Math. Statist. Probab.*, vol. 1, 1956, pp. 13–17.
[61] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *Int. J. Artif. Intell. Tools*, vol. 13, no. 1, pp. 157–169, 2004.
[62] R. A. Fisher, "On the interpretation of X^2 from contingency tables, and the calculation of P," *J. Roy. Statist. Soc.*, vol. 85, no. 1, pp. 87–94, 1922.
[63] A. L. Harvey, "Natural products in drug discovery," *Drug Discovery Today*, vol. 13, nos. 19–20, pp. 894–901, Oct. 2008.
[64] D. A. Dias, S. Urban, and U. Roessner, "A historical overview of natural products in drug discovery," *Metabolites*, vol. 2, no. 2, pp. 303–336, 2012.
[65] J. Huang, C. Niu, C. D. Green, L. Yang, H. Mei, and J.-D. J. Han, "Systematic Prediction of Pharmacodynamic Drug-Drug Interactions through Protein-Protein-Interaction Network," *PLoS Comput. Biol.*, vol. 9, p. e1002998, Mar. 2013.
[66] K. Park, D. Kim, S. Ha, and D. Lee, "Predicting pharmacodynamic drug-drug interactions through signaling propagation interference on protein-protein interaction networks," *PloS ONE*, vol. 10, p. e0140816, Oct. 2015.

SUNYONG YOO received the B.S. and M.S. degrees in electronics and information engineering from Korea Aerospace University, South Korea, in 2007 and 2009, respectively, and the Ph.D. degree in bio and brain engineering with the Korea Advanced Institute of Science and Technology, South Korea. His research interests include bioinformatics, systems biology, network medicine, and machine learning.

SUHYUN HA received the B.S. degree in computer communication and engineering from Korea University in 2012, and the M.S. degree in bio and brain engineering from KAIST, South Korea, in 2014, where he is currently pursuing the Ph.D. degree. His research interests include bioinformatics, systems biology, database, and machine learning.

MOONSHIK SHIN received the M.S. degree in robotics program from KAIST in 2012, and the Ph.D. degree in bio and brain engineering from KAIST, South Korea, in 2018. His current research interests include bioinformatics, cheminformatics, network analysis, and machine learning.

KYUNGRIN NOH received the B.S. and M.S degrees in bio and brain engineering from KAIST, South Korea, in 2017. He is currently with IBM Korea, Global Business Service. His research interest includes bioinformatics, machine learning, and data analysis.

HOJUNG NAM received the Ph.D. degree in bio and brain engineering from KAIST, South Korea, in 2009. She is currently an Associate Professor with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, South Korea. Her research interests include bioinformatics, cheminformatics, and systems biology.

DOHEON LEE was a Visiting Professor of Stanford University, Indiana University, the Translational Genomics Research Institute, and The University of Texas at Austin, USA. He is currently a Professor in bio and brain engineering with KAIST, South Korea. He is also the Director of the Bio-Synergy Research Center, a Korean National Project, where about 30 research organizations are collaborating for bioinformatics and systems biology. He has published over 200 academic articles in bioinformatics, systems biology, and data mining. He was an Associate Editor for *ACM Transactions on Internet Technology* for nine years. He is also serving *Computers in Biology and Medicine*, the *International Journal of Data Mining in Bioinformatics*, and *Healthcare Informatics Research* as an Editorial Board Member.