Journal of Ethnopharmacology 208 (2017) 117–128

Contents lists available at ScienceDirect

Journal of Ethnopharmacology

journal homepage: www.elsevier.com/locate/jep

A framework for the targeted selection of herbs with similar efficacy by exploiting drug repositioning technique and curated biomedical knowledge

Sang-Jun Yea[a,b], Bu-Yeo Kim[c], Chul Kim[b,*], Mun Yong Yi[a,*]

[a] Graduate School of Knowledge Service Engineering, Korea Advanced Institute of Science and Technology, Republic of Korea
[b] K-herb Research Center, Korea Institute of Oriental Medicine, Republic of Korea
[c] KM Convergence Research Division, Korea Institute of Oriental Medicine, Republic of Korea

ARTICLE INFO

Keywords:
Systems biology
Traditional Chinese medicine
Medicinal herb
Similar efficacy
Drug repositioning

ABSTRACT

Ethno pharmacological relevance: Plants have been the most important natural resources for traditional medicine and for the modern pharmaceutical industry. They have been in demand in regards to finding alternative medicinal herbs with similar efficacy. Due to the very low probability of discovering useful compounds by random screening, researchers have advocated for using targeted selection approaches. Furthermore, because drug repositioning can speed up the process of drug development, an integrated technique that exploits chemical, genetic, and disease information has been recently developed. Building upon these findings, in this paper, we propose a novel framework for the targeted selection of herbs with similar efficacy by exploiting drug repositioning technique and curated modern scientific biomedical knowledge, with the goal of improving the possibility of inferring the traditional empirical ethno-pharmacological knowledge.

Materials and methods: To rank candidate herbs on the basis of similarities against target herb, we proposed and evaluated a framework that is comprised of the following four layers: *links, extract, similarity*, and *model*. In the framework, multiple databases are linked to build an herb-compound-protein-disease network which was composed of one tripartite network and two bipartite networks allowing comprehensive and detailed information to be extracted. Further, various similarity scores between herbs are calculated, and then prediction models are trained and tested on the basis of theses similarity features.

Results: The proposed framework has been found to be feasible in terms of link loss. Out of the 50 similarities, the best one enhanced the performance of ranking herbs with similar efficacy by about 120–320% compared with our previous study. Also, the prediction model showed improved performance by about 180–480%. While building the prediction model, we identified the compound information as being the most important knowledge source and structural similarity as the most useful measure.

Conclusions: In the proposed framework, we took the knowledge of herbal medicine, chemistry, biology, and medicine into consideration to rank herbs with similar efficacy in candidates. The experimental results demonstrated that the performances of framework outperformed the baselines and identified the important knowledge source and useful similarity measure.

1. Introduction

Plants have been the most important natural resources for treating diseases since ancient civilizations and still remain important for the modern pharmaceutical industry for areas such as new drug development (Sharma and Sarkar, 2013). These can be ascertainable from the well-established systems of traditional medicine in several countries and the fact that one-third of drugs that are currently available come from natural resources with plant origin (Strohl, 2000). Since the Nagoya Protocol has been applied to the traditional knowledge associated with genetic resources that are covered by the Convention on Biological Diversity, it has become more difficult to acquire non-indigenous medicinal herbs. As a result, the demand for alternative medicinal herbs with similar efficacy has increased (Nagoya protocol, 2017). Recently, in vivo and in vitro studies were carried out to determine and compare the anti-inflammatory effects of *Peucedanum praeruptorum* Dunn and *Peucedanum decursivum* (Miq.) Maxim. on allergic lung inflammation (Lee et al., 2016). Whereas the development of new drug has been suffering from high cost, long time, and high risk. Due to the fact that efficient and effective applications of natural

* Corresponding authors.
E-mail addresses: chulnice@kiom.re.kr (C. Kim), munyi@kaist.ac.kr (M.Y. Yi).

http://dx.doi.org/10.1016/j.jep.2017.06.048
Received 28 February 2017; Received in revised form 27 June 2017; Accepted 27 June 2017
Available online 05 July 2017

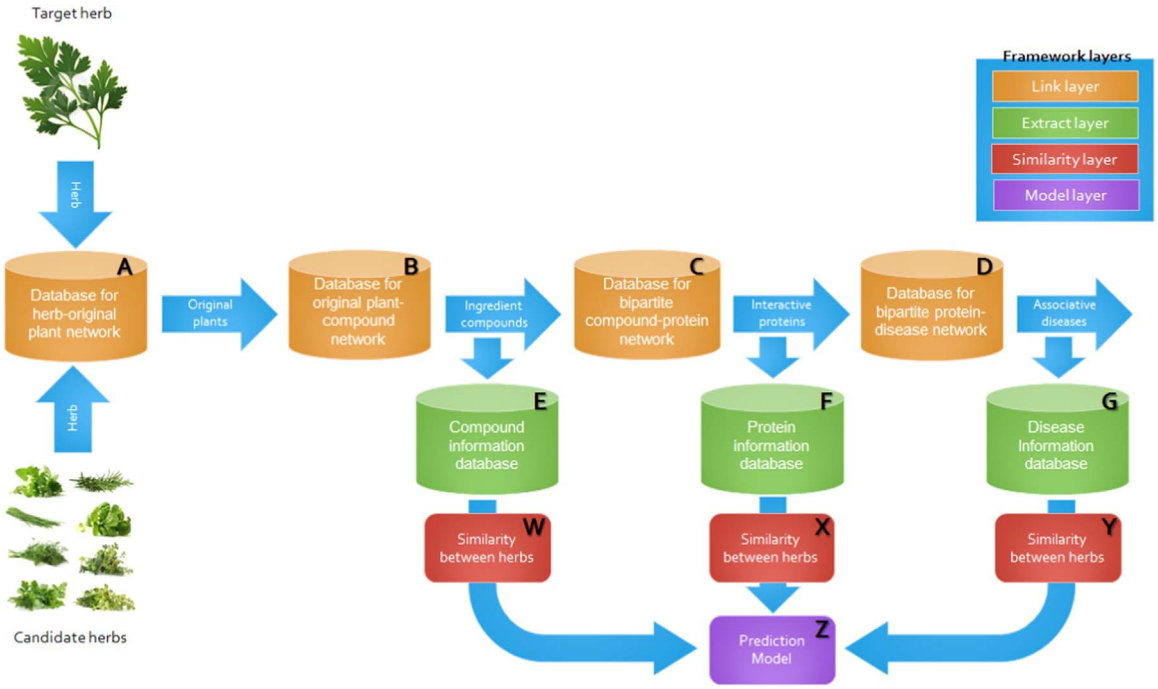

Fig. 1. The overall procedure of THED framework.

products will improve the drug discovery process and reduce the cost of drug development, various screening approaches are being developed in which natural products can be used in the drug discovery process (Harvey, 2008). One of these approaches is to adopt the same family and/or genus of plants or medicinal herbs with similar efficacy based on the assumption that they might have the same or similar bioactive ingredients and biomedical functions (Pan et al., 2013).

There is an approximate 1-in-10,000 probability of discovering useful compounds by random screening (Douwes et al., 2008), not to mention the fact that there is a huge cost involved in both time and expenditure to screen the vast number of randomly selected extracts. Thus researchers have advocated for the use of targeted selection approaches that employ phylogenetic, ecological, or ethno-pharmacological knowledge in the application of natural resources. To adopt a targeted selection approach in the discovery of new drugs, prior researchers have used Bayesian analysis (Weckerle et al., 2011), regression analysis (Douwes et al., 2008), the integration of ethnopharmacology and bioinformatics (Bernard et al., 2001), and the simple scoring system (Clark et al., 1997). To replace medicinal herbs in traditional medicine, targeted selection methods have been proposed, such as the simple mathematical and logical method (Fang et al., 2013), manual review (Zhang et al., 2012), and literature review (Medeiros et al., 2011). However, none of these studies utilized vast accumulated scientific biomedical knowledge to the fullest extent possible or they just used their knowledge of traditional medicine. In our previous study (Yea et al., 2016), a targeted selection technique was developed and evaluated over three validation datasets to rank herbs with similar efficacy by similarity scores calculated on the basis of medical subject headings (MeSH) extracted from articles in MEDLINE. It showed the possibility of inferring traditional empirical ethnopharmacological knowledge using modern scientific biomedical knowledge. However, it also had the following limitations: (1) it only used a non-curated biomedical database, (2) it did not show sufficient performance, and (3) it did not built sophisticated prediction model.

The inefficiencies connected to time and cost in new drug development has brought about the drug repositioning approach, which finds new or additional indication for existing drugs. Since drug repositioning can speed up the process of drug development, it has been in the limelight and as a result, various computational methods have been proposed (Terstappen and Reggiani, 2001; Paul et al., 2010). From amongst these methods, an integrated technique that constructs a comprehensive heterogeneous drug-molecule-disease network at distinct levels and on different scales has been recently developed providing systemic views in predicting new indications (Hurle et al., 2013; Zhang et al., 2014; Wu et al., 2013). These studies measured the similarities between the pertinent drug and disease information and combined it with a nonlinear optimization technique (Zhang et al., 2014), SVM with Kronecker product kernel (Wang et al., 2013a, 2013b), and a logistic regression classifier (Gottlieb et al., 2011), to rank the accumulated evidence for determining the connections between drug and disease. Finding herbs with similar efficacy can be thought of as being analogous to herb repositioning. However, none, as far as we know, have adopted the drug repositioning approach for selecting herbs with similar efficacy. Therefore, in this paper, we propose a novel method for the targeted selection of herbs with similar efficacy by exploiting drug repositioning technique (THED) based on curated biomedical knowledge. The curated biomedical knowledge is publicly available via an online database whose content has been collected by a number of experts via consulting, verifying, and aggregating existing sources (Buneman et al., 2008). We aimed to build a novel computational framework, called THED, for the targeted selection of herbs with similar efficacy; to evaluate the performance of THED; and to identify the relative importance of biomedical knowledge sources selected by prediction model in THED. In order to evaluate the performance of the proposed methods, we adopted the same three validation datasets and evaluation metrics employed by Yea et al. (2016).

2. Materials and methods

2.1. Overall procedure

The purpose of THED framework is to rank candidate herbs on the basis of similarities that are calculated against target herbs (i.e., to tell which candidate herb is more similar to the target herb in terms of efficacy). As depicted in Fig. 1, the THED framework is comprised four layers: *link*, *extract*, *similarity*, and *model*. The link layer is orange and composed of four databases (A, B, C, and D) to build an herb-

compound-protein-disease network which comprises one tripartite network of herb-original plants-ingredient compounds and two bipartite networks of compound-interactive proteins and protein-associative diseases. The input and output of each database are illustrated in the arrows (e.g., database C, which links compounds and proteins, takes the ingredient compounds of the original plant and gives the corresponding interactive proteins). The extract layer is green and contains three databases (E, F, and G), which provide comprehensive and detailed information about compounds, proteins, and diseases. The similarity layer is red and comprised of three similarity calculation processes (W, X, and Y), which calculate the various similarities between the target herb and candidate herbs using the information from the extract layer. The model layer is purple and predicts the similarity scores between the target herb and each candidate herb by using the trained prediction model based on the feature vector generated in the similarity layer. Finally, the ranking of each candidate herb is determined by the predicted similarity scores against the target herb. Although all four layers had some degree of differences compared to existing drug repositioning techniques, the link layer had the biggest distinctions. During drug development and clinical experiments, the drug-related information, such as chemical structures, target proteins, target diseases, and side effects, were already well known and accumulated into public databases. However, because these information about medicinal herbs are uncertain and is being gradually discovered by recent studies, the link layer is an essential component of THED, but the linking databases is not a necessary part of the drug repositioning framework.

2.2. Curated knowledge sources

2.2.1. Databases for the link layer

During the past several decades, the development of systems biology, which is the computational and mathematical modeling of complex biological systems, has heavily depended on the vast biological databases and the varieties of interconnections between them. Since various large-scale curated chemical, biological, and medical databases were developed for different objectives and are complementary to each other, one of the main problems of exploiting and interconnecting multiple databases is determining how to select suitable databases that are adequate for the purpose of studies (Yang et al., 2013). Thus, selection criteria are needed for each database in the link layer, which are listed in Table 1. Also in Table 1 the facts, including inputs and outputs, from the selected databases have been summarized. While linking databases, the type of data and link loss had to be considered between adjacent databases in the herb-compound-protein-disease network. For data type, the STITCH database generated the UniProt ID. However, the DisGeNET database cannot handle the UniProt ID and can only take the GeneID. Thus, the UniProt ID was transformed to the GeneID. On the other hand, for link loss, there was non-corresponding information between adjacent databases (e.g., some proteins didn't have the corresponding associative diseases.) As such, THED inevitably suffered from link loss. If the link loss is too big, then THED cannot be applied in biomedical or ethno-pharmacological fields. Thus, we evaluated the feasibility of linking databases in THED through experiments.

Because database A defines the scope of this study, we selected the reference dataset from our earlier study, which was built using herbs and corresponding original plants out of the *Korean Pharmacopoeia* and the *Korean Herbal Pharmacopoeia* (Yea et al., 2016). For database B, TCMID (Xue et al., 2012), which includes 8159 herbs and 25,210 compounds, was selected as it provides the compound information and the sources of information for each original plant. Tables S1–S4 provide the Supplementary material for the evaluation results of the databases in the link layer. Although TCMID is the most comprehensive database for relational information between original plants and ingredient compounds, because it has been established on the basis of

Table 1
Selected databases for the link layer.

	Name	Selection criteria	Input	Output	Fact	Reference
A	Reference dataset	None	Herbal name	Scientific name of original plants	448 herbs 610 original plants	Yea et al. (2016)
B1	Traditional Chinese Medicines Integrated Database (TCMID)	Information by original plants Comprehensiveness Information source	Scientific names of plants	Name of chemical compound PubChem CID	8159 herbs 25,210 chemicals	Xue et al. (2012)
B2	Northeast Asian Traditional Medicine Database (TM-MC)	Complementary of B_1	Scientific names of plants	Name of chemical compound PubChem CID	536 herbs14,000 chemicals	Kim et al. (2015)
C	Search Tool for Interactions of Chemicals (STITCH)	Comprehensiveness Confidence score	PubChem CID	UniProt ID	0.4 M chemicals 3.6 M proteins 1133 organisms 0.4 M associations (human)	Kuhn et al. (2013)
D	Disease Gene Network (DisGeNET)	Comprehensiveness Confidence score	Gene ID	UMLS CUI	16,000 genes13,000 diseases 0.4 M associations	Piñero et al. (2015)

traditional Chinese medicine, some herbs and original plants of this study did not exist. To complement TCMID, thus, TM-MC (Kim et al., 2015) was added to cover some herbs and original plants in traditional Korean medicine. For databases C and D, the confidence score was set for the selection criteria to prioritize and select information. This is because one chemical compound interacts with multiple proteins and one protein has associations with multiple diseases, and because the data was collected and aggregated from different sources and methods, such as manual curation, text mining, prediction, and so on. For database C, STITCH (Kuhn et al., 2013) was chosen, which includes approx. 0.4 M compounds, approx. 3.6 M proteins for 1133 organisms, and approx. 0.4 M associations for only human. Finally, for database D, DisGeNET (Piñero et al., 2015), which includes about 16,000 genes, was selected. This covers about 70% of the protein coding genes of human, 13,000 diseases, and 0.4 M associations.

2.2.2. Databases for the extract layer

The main purpose of the extract layer is to gather comprehensive and detailed information about compounds, proteins, and diseases. As such, representative online information databases that can handle input data types were chosen. The selected databases and corresponding facts, including inputs and outputs, are summarized in Table 2. First, to calculate the structural similarity between compounds, the canonical Simplified Molecular Input Line Entry System (SMILES) was extracted from PubChem. SMILES is a line notation for entering and representing molecular structures and chemical reactions. Second, to calculate structural similarities between proteins, the canonical protein sequence in FASTA was gathered from UniProt. FASTA is a text-based format for representing either nucleotide sequences or peptide sequences. Furthermore, to calculate the semantic similarities between proteins, gene ontology, which is a computational knowledge representation of how genes encode biological functions at the molecular, cellular, and tissue levels, was extracted. Gene ontology is divided into three categories: the biological process (BP), molecular function (MF), and cellular component (CC) (Gene Ontology Consortium Help, 2017). Finally, to calculate the semantic similarities between diseases, we gathered Disease Ontology (DO), which is a collection of well-established terminologies that contain disease and disease-related concepts, such as SNOMED, ICD-10, and MeSH (Disease Ontology Help, 2017), and Human Phenotype Ontology (HPO), which describes a phenotypic abnormality, such as atrial septal defect, in order to provide a standardized vocabulary of phenotypic abnormalities encountered in human diseases (Human Phenotype Ontology Help, 2017). Because many phenotypically similar diseases are caused by functionally-related genes and the concept of modularity for human diseases exists (Guzzi et al., 2012; Kann, 2010), two types of disease information, which were matched terms and broadened synonyms, were extracted while we were extracting DO and HPO. The synonyms were broadened to parent, child, or similar relationships in the semantic network of UMLS using G1, shown in Table 2. As such, four totally different kinds of ontologies were provided to calculate the semantic similarity between diseases.

2.3. Similarity calculation

In diverse science and engineering fields, how alike or unalike objects are in comparison to one another is indicated by a "similarity measure," which is usually a real-valued function and the inverse of a distance metric. And it is well defining way that "A is similar to B with respect to C." For instance, chemical compound A can be similar to compound B in respect to structure, semantics, and so on (Goodman, 1972; Nikolova and Joanna, 2003).

2.3.1. Structural similarity

Structural similarity plays an important role in modern chemistry for predicting the properties of unknown chemical compounds and screening a large chemical database in connection to the development of new drugs. This approach is based on the similar property principle, which states that structurally similar molecules exhibit similar biological activity (Johnson and Gerald, 1990; Martin et al., 2002). To calculate the structural similarity between compounds, the extracted SMILE was transformed into a fingerprint, which is an ordered list of binary bits. After that, the score was calculated with the Tanimoto coefficient (Tanimoto, 1957), which is defined as Eq. (1):

$$SIM_{tan}(c_1, c_2) = \frac{Nc_1 \,\&\, c_2}{Nc_1 + Nc_2 - Nc_1 \,\&\, c_2}, \text{where } N \text{ is the count of bits in fingerprint} \quad (1)$$

The biological function of a protein is highly dependent on its structure, which is also closely related to the sequence of amino acids. In understanding evolutionarily divergent proteins and predicting the protein functionality family, the sequence alignment is an important method for modern biology (Wright and Dyson, 1999; Marcotte et al., 1999). Thus, the representative Smith-Waterman algorithm (Smith et al., 1985) with BLOSUM62 (Henikoff and Henikoff, 1992) was adopted to calculate the structural similarity between proteins.

2.3.2. Semantic similarity

Semantic similarity is a metric defined over a set of terms, where distance or similarity between them is defined on the basis of the affinity of their meaning from the viewpoint of semantic context. There are various kinds of semantic similarity measures and different performances have been reported in different contexts (Guzzi et al., 2012). As such, two common semantic similarities, the Lin measure (Lin, 1998) and Wang measure (Wang et al., 2007), were adopted and are denoted in Eqs. (2) and (3), respectively. Theses similarity measures were calculated for every three categories of gene ontology and all four kinds of disease ontologies. The Lin semantic similarity is based on the information content (IC), which is defined as $IC(t) = -\log(p(t))$, where p(t)is the probablity of observing t. When the most informative common ancestor (MICA) is defined as $MICA(t_1, t_2) = argMAX(IC(t_j))$, where $t_j \in$ ancestors(t_1, t_2), Lin measure is defined as:

$$SIM_{lin}(t_1, t_2) = \frac{IC(MICA(t_1, t_2))}{IC(t_1) + IC(t_2)} \quad (2)$$

Whereas, the Wang semantic similarity is based on the directed acyclic

Table 2
Selected databases for the extract layer.

	Name	Input	Output	Fact	Method
E	PubChem	Name of chemical compound PubChem CID	Canonical SMILES	82.6 M compounds	OpenAPI
F	Universal Protein resource (UniProt)	UniProt ID	FASTA GeneID Gene Ontology (BP, MF, CC)	0.6 M proteins	OpenAPI
G1	Unified Medical Language System (UMLS)	UMLS CUI	UMLS CUI of parent, child, or similar relationship	1 M CUIs	Download dataset
G2	Disease Ontology (DO)	UMLS CUI	DOID	10,263 terms	Download dataset
G3	Human Phenotype Ontology (HPO)	UMLS CUI	HPOID	10,371 terms	Download dataset

graph (DAG) and each edge of the graph is weighted. When the semantic contribution of term t to term A is defined as $S_A(t) = \max\{W_e \times S_A(t')t' \in childrenof(t)\}$, if t‡A, for any term t in DAG of ancestors of A, the Wang measure is defined as:

$$SIM_{wang}(t_1,\ t_2) = \frac{\sum_{t \in T_A \cap T_B}(S_A(t) + S_B(t))}{\sum_{t_1 \in T_A} S_A(t_1) + \sum_{t_2 \in T_B} S_B(t_2)} \quad (3)$$

2.3.3. Aggregation method

The aforementioned structural and semantic similarities were calculated pairwise between two biomedical objects (e.g., the structural similarity between compounds and the semantic similarity between BP annotations of gene ontology). However, to produce similarities between herbs, the multiple pairwise similarity scores had to be combined into one herb-herb similarity score. There are several kinds of aggregation methods, and different performances have been reported in different contexts (Guzzi et al., 2012). From among them, we adopted two recent methods, called FSA and FSM (Schlicker et al., 2006). The FSA method calculates the average of the average of the maximum values of each row and column in the matrix of all pairwise similarity scores and is defined as follows:

$$AGG_{fsa} = 0.5 \times \left(\frac{\sum_{i=1}^{m} \underset{1 \le j \le n}{max} SIM_{ij}}{m} + \frac{\sum_{j=1}^{n} \underset{1 \le i \le m}{max} SIM_{ij}}{n} \right) \quad (4)$$

The FSM method calculates the maximum of the average of the maximum values of each row and column in the matrix of all pairwise similarity scores and is defined as follows:

$$AGG_{fsm} = max \left(\frac{\sum_{i=1}^{m} \underset{1 \le j \le n}{max} SIM_{ij}}{m}, \frac{\sum_{j=1}^{n} \underset{1 \le i \le m}{max} SIM_{ij}}{n} \right) \quad (5)$$

2.3.4. Set similarity

In biology and medicine, binary data, which codes presence or absence, is very common and there are a large amount of similarity or dissimilarity coefficients that have been specifically developed for binary variables to deal with this type of data. Several kinds of similarity coefficients were applied to calculate the relatedness between research objects in the computational systems of biology and medicine (Podani, 2000). From among these, we adopted two common sets of similarity methods of the Kulczynski similarity (Kulczynski, 1927) and Ochiai similarity (Ochiai, 1957). Where, a is the number of variables present in both objects, and b and c are the number of variables present in one of the two objects. The Kulczynski and Ochiai similarities are as defined below in Eqs. (6) and (7), respectively:

$$SIM_{kul} = 0.5 \times \left(\frac{a}{a+b} + \frac{a}{a+c} \right) \quad (6)$$

$$SIM_{och} = \frac{a}{\sqrt{(a+b)(a+c)}} \quad (7)$$

2.3.5. Similarity matrix

As previously explained, the structural and semantic similarities were generated between two biomedical objects and then combined using the aggregation method to create a similarity score between different herbs. For example, with protein, we extracted one kind of structural information and three kinds of semantic information, and adopted one structural and two semantic similarity measures. We also combined each of them with two aggregation methods. As such, we produced two structural and four semantic similarities for each category of gene ontology, as shown in Table 3. However, when set similarity is calculated between herbs, it doesn't need to be combined by aggregation methods. As shown in Table 3, we produced 4, 22, and 24 kinds of similarities using the information extracted from compounds, proteins, and diseases, respectively. Finally, for 50 kinds of similarities, we built matrixes with the target herbs as rows and candidate herbs as columns in order to generate a feature vector of the prediction model. Furthermore, the similarity matrix was normalized by rows with a minimum value of a row as 0 and a maximum as 1.

2.4. Prediction model

To build and evaluate the prediction model in THED, a feature vector, which is a 50-dimensional vector of the numerical data representing the similarities between herbs, had to be generated. Out of the 50 similarity matrixes explained in Subsection 2.3.5, a feature vector was created for every pair (H_T, H_C) where $H_T \in$ {target herbs} and $H_C \in$ {candidate herbs}. Each feature vector had label **Y** where H_T refers to original herb and H_C refers to alternative herb with similar efficacy and **N** where (H_T, H_C) is not the suitable relationship. The conceptual illustration for the formation of a feature vector is shown in Fig. 2. The feature vectors were then separated by the target herbs into a training set to build the prediction model and a testing set to evaluate it, which is also shown in Fig. 2.

To select classifiers in the model layer of THED, two selection criteria were set: (1) probability estimation or ranking and (2) feature importance comparison. First, because candidate herbs must be ranked on the basis of similarity, we needed a ranking capacity or probability estimation property that can be interpreted as a rank. Second, we wanted to compare the relative importance of features, which were selected by the prediction model out of 50 similarity features. This comparison might give insight into what kinds of information are critical when building a prediction model to rank herbs with similar efficacy and what kinds of data should be more comprehensively gathered. Among the various machine learning algorithms, the Support Vector Machine (SVM) as a linear classifier and Random Forest (RF) as a non-linear classifier were chosen to compare the behavior and performance of linear and non-linear prediction models. They met the two criteria listed above and have been receiving an enormous amount of attention in the systems biology and systems medicine, because they showed high performance in diverse applications, can handle a large number of features, and provide relative importance of features (Grömping, 2009). The feature importance of SVM is measured by the coefficient of the support vectors and that of RF is calculated by permuting each feature (Chang and Lin, 2008; Jiang et al., 2007). To get the generalized performance of the models, the 5-fold cross-validation approach was adopted. To achieve the best performance by each prediction model, a best feature selection process was carried out by adding features one by one, which were sorted according to the information gain scores of 50 similarity features (Guyon and Elisseeff, 2003).

2.5. Experimental setting

To evaluate the performance of THED, three of the same validation datasets corresponding to Yea et al. (2016) were adopted. The OMB dataset was built from four ancient Korean and Chinese medical books, the HEC dataset was constructed via the knowledge and consent of two traditional Korean medicinal herbal experts, and the POK dataset was made up of herbs with multiple original plants in the *Korean Pharmacopoeia* and the *Korean Herbal Pharmacopoeia*. The number of target and candidate herbs in validation datasets are shown in Table 4, and Tables S5–S7 in the Supplementary material provide samples of the three validation datasets.

Also, the same evaluation metrics, which correspond to Yea et al. (2016), were used. These are the recall, average reciprocal hit-rank (ARHR), and area under curve (AUC). The recall is the fraction of hit instances in candidates. The ARHR rewards each hit on the basis of its position in the top N list. It ranges from recall / N (worst) to recall

Table 3
The number of calculated similarity measures.

Similarity layer	Group	Subgroup[a]	Structural similarity	Semantic similarity	Set similarity	Subtotal	Total
W	Compound	SQ	2	0	2	4	4
X	Protein	SQ	2	0	2	4	22
		CC	0	4	2	6	
		MF	0	4	2	6	
		BP	0	4	2	6	
Y	Disease	DO_ID	0	4	2	6	24
		DO_SN	0	4	2	6	
		HPO_ID	0	4	2	6	
		HPO_SN	0	4	2	6	

[a] SQ: structure, CC: cellular component, MF: molecular function, BP: biological process, DO: disease ontology, HPO, human phenotype ontology, ID: matched term, SN: broadened synonym.

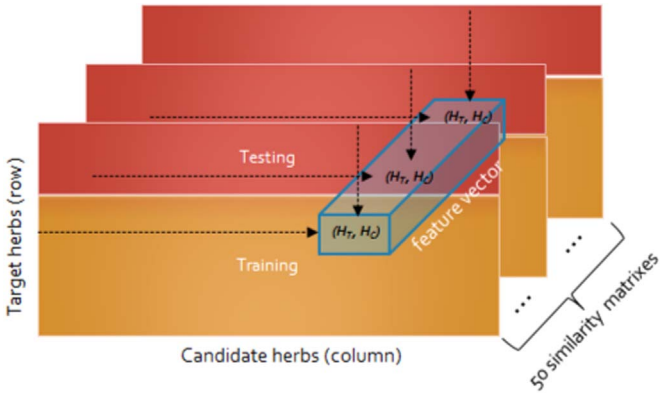

Fig. 2. The formation of feature vector, training set, and testing set.

Table 4
The number of target and candidate herbs for each validation dataset. Quoted from Yea et al. (2016).

Dataset	Target herbs	Candidate herbs
OMB	58	79
HEC	26	29
POK	267	448

(best) and is defined as given in Eq. (8) (Deshpande and Karypis, 2004). Among the candidate herbs in each validation dataset, the performance of THED was evaluated for the top10 ranks, which had a high similarity scores against the target herb. It was compared to Yea et al. (2016), which was displayed as baseline in figures and tables of the results section. The AUC is the area under the performance graph of the recall or ARHR and ranges from 0 (worst) to 1 (best).

$$ARHR = \frac{1}{N}\sum_{i=1}^{h}\frac{1}{p_i}, \text{ where } h \text{ is the number of hits and } p_i \text{ is the position within the list} \quad (8)$$

3. Results

3.1. Coverage of target and candidate herbs

As depicted in Fig. 1 and explained in Section 2.2.1, the THED framework connects four different databases in the link layer to build a biomedical network from the herb to the disease. Thus, THED inevitably suffers from link loss (i.e., because database A defined the scope of this study, it had no losses; however, the other databases lost some degree of links between adjacent databases). As depicted in Fig. 3, the coverage of target and candidate herbs displayed similar patterns. The lowest coverages occurred in the final link (i.e., database D), and they occurred in the HEC dataset in both herbs. They were 80.77% and 79.31%, respectively. However, the coverages of OMB and POK reached over 90% and 80%. As the maximum final link loss was around 20%, the THED framework was proved to be viable in terms of coverage.

3.2. Best threshold for linking databases

One of the biggest problems in exploiting large-scale biomedical databases that are collected and aggregated from different sources and methods is how to prioritize and select the information (Piñero et al., 2015), as explained in Section 2.2.1. As such, we set the information source and confidence score as the database selection criteria in the link layer. By conducting comprehensive experiments, we chose the best threshold for each link database and individual validation dataset, as shown in Table 5. (Refer to Figs. S1–S3 in the Supplementary material for the experiment results). In Table 5, the thresholds for database B denote the selected information sources from TCMID (e.g., CTM means that information combined together from HIT, TD@T, and TCM-ID provided the best result, and so it was selected for this study). The thresholds for databases C and D denote the amount of top-ranked information used according to the confidence score (e.g., Top10 means that THED performed the best when 10 top-ranked information were used.) Actually, the best thresholds for database D and the OMB dataset was top1 instead of top10. However, due to the coverage problem, we reconsidered our choice of thresholds. In regards to the actual best thresholds when building an herb-compound-protein-disease network, the OMB and POK datasets showed the best performance when a small amount of information with a high confidence score was used contrasting with the HEC datasets.

3.3. Performance evaluation of features

In THED, 50 kinds of similarities were calculated and each of them was evaluated to determine the performance of individual features in terms of recall and ARHR against the baseline. Because Yea et al. (2016) adopted the semantic similarity between MeSH terms in the articles from MEDLINE, the performances of their method are depicted in figures as baselines C, D, and G, which are the Disease, Chemical and Drugs, and Phenomena and Processes categories of MeSH, respectively. As can be seen in Fig. 4, the best feature in OMB and POK is the protein structure's similarity with the Kulczynski method. The recalls for it were 0.40 and 0.74, respectively (refer to Table S8 in the Supplementary material for the feature nomenclature). However, the best feature of HEC was the compound structure's similarity with the FSA method and its recall was 0.84 (i.e., it can predict 84% of herbs with similar efficacy within the top10 candidate herb rankings, while baseline_D only had a 45% probability rating).

The AUC of the recall graphs in Fig. 4 were calculated and summarized, as shown in Table 6. The average performance improve-

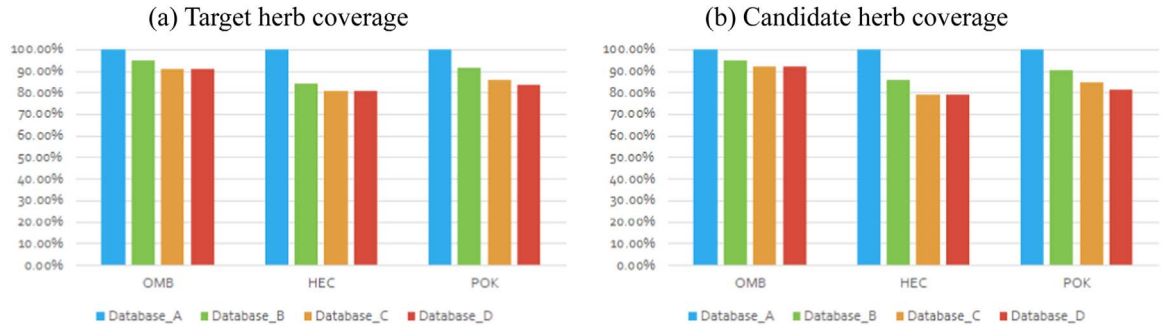

Fig. 3. The coverage of target and candidate herbs in THED framework.

Table 5
The best thresholds to link databases.

Database	OMB	HEC	POK
B	CTM[a]	CTEM[a]	CTEM[a]
C	Top1	Top10	Top1
D	Top10	Top25	Top5

[a] C: HIT (http://lifecenter.sgst.cn/hit), T: TD@T (http://tcm.cmu.edu.tw/), E: Encyclopedia of traditional Chinese medicines, M: TCM-ID (http://bidd.nus.edu.sg/group/TCMsite).

ment rate was the best in POK at 322% (top1 feature) and 312% (top10 features) compared to all three baseline performances. Next was HEC at 245% and 203%, respectively. However, the AUC showed a slight improvement in OMB. The difference in AUC between the top1 and top10 average was the smallest in POK, but the largest in HEC (i.e., the performances of the top10 features were similar in POK rather than in HEC).

Regarding ARHR, which rewards each hit position, as shown in Fig. 5, the best features in all three datasets were the same as the recall graphs above. In OMB and POK, the similarity in protein structure with the Kulczynski method was the best and were 0.17 and 0.42, respectively. The best feature of HEC was the similarity in compound structure with the FSA method and the ARHR of it was 0.49. However, baseline_D was only 0.26.

As shown in Table 7, contrary to the AUC of recall, the average improvement in performance was different in the case of top1 and the top10 average. The best improvement in the top1 feature occurred in HEC and was 319%. Regarding to the top10 average of features, the best improvement occurred in POK and was 311%. However, the AUC showed a slight improvement in OMB. The difference in the AUC between the top1 and the top10 average was the smallest in POK and the largest in HEC, which was the same as the AUC for recall.

In Table 8, the top10 features by recall are listed and some common characteristics were found. The similarity in protein structure with the Ochiai method and the similarity in compound structure with the FSA method were common features in all three datasets. There were features encompassing compounds, proteins, and diseases in OMB and HEC, but not POK (i.e., no disease-related features were found in POK). Also, no semantic similarity features existed in POK. In all three datasets, the features in structural similarity with set methods are the most frequent. However, the features of disease similarity broadened by synonyms didn't appear.

3.4. Performance evaluation of the prediction model

It is apparent that the performance of the prediction model outperformed the baselines, as depicted in Fig. 6. For example, the recalls of RF and SVM in POK were 0.99 and 0.70, respectively; whereas, that of baseline_D was only 0.39 (i.e., RF can predict 99% of herbs with similar efficacy within the top10 ranks out of 448 candidate herbs, while baseline_D only showed a 39% probability). In all three datasets, RF showed better results than SVM, which can be interpreted that a non-linear combination of features is more suitable for the prediction model to rank herbs with similar efficacy. RF predicted almost all herbs with similar efficacy within the top10 ranks in HEC and POK. However, SVM showed little difference with the baselines in OMB.

As can be seen in Table 9, the average improvement in performance was the best in POK, and it was 477% (RF) and 368% (SVM). Next was HEC, which was 265% and 261%, respectively. The difference between RF and SVM was the smallest in HEC; whereas, it was the largest in POK. Because 1.0 is considered to be the perfect score in the AUC, the performance of THED in HEC and POK proves that the drug repositioning approach adopted in THED is promising for these kinds of studies.

The ARHR graphs in Fig. 7 also proved that the prediction model outperformed baselines when rewarding the position of each hit (e.g., its performances in POK were 0.74 (RF) and 0.54 (SVM), whereas that of baseline_D was only 0.22). Like recall graphs, RF showed better performance than SVM in all three datasets.

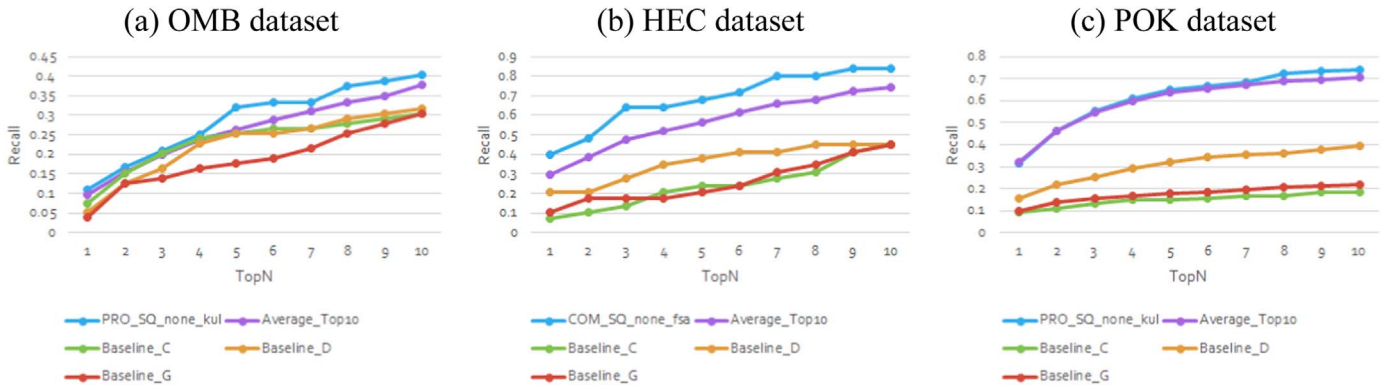

Fig. 4. The recall of top1 feature and average of top10 features against baseline. The top1 features are shown in its name.

Table 6
The AUC of recall of top1 feature and average of top10 features against baseline.

Dataset	Category	AUC	Dataset	Category	AUC	Dataset	Category	AUC
OMB	PRO_SQ_none_kul	0.2889	HEC	COM_SQ_none_fsa	0.6840	POK	PRO_SQ_none_kul	0.6137
	Average_Top10	0.2616		Average_Top10	0.5673		Average_Top10	0.5989
	Baseline_C	0.2329		Baseline_C	0.2448		Baseline_C	0.1487
	Baseline_D	0.2253		Baseline_D	0.3586		Baseline_D	0.3053
	Baseline_G	0.1886		Baseline_G	0.2586		Baseline_G	0.1740

As shown in Table 10, the average improvement of ARHR was the best in POK, and it was 527% (RF) and 373% (SVM). Next was HEC, which was 378% and 357%, respectively. The difference between RF and SVM was the smallest in HEC; whereas, it was the largest in POK.

3.5. Relative feature importance

As explained in Section 2.4, best feature selection was carried out by adding features one by one, and the number of features that achieved the best performance are listed in Table 11. RF had twice as features as SVM, and both RF and SVM had the most features in OMB. The subsequent results in this section are described on the basis of selected features, which are shown in Table 11.

The similarity features came from different knowledge source groups (i.e., compounds, proteins, and diseases). Each knowledge source group had a different number of features, as shown in Table 3 (e.g., the compound group had 4 features and protein had 22 features). As such, the distribution of features by knowledge sources was compared by percentage. Because each feature had different levels of importance when building the prediction model, the average importance of features by knowledge source was also analyzed. As shown in (a) and (b) of Fig. 8, compound information was used the most in both RF and SVM. Moreover, all of the compound features were used for all three datasets in RF. And protein information was used the second most. In SVM, there was no disease feature in HEC and POK. As depicted in (c) and (d) of Fig. 8, the compound and protein information had a higher level of importance rather than disease. Moreover, compounds had the highest importance for all three datasets in RF. Proteins had the highest level of importance in OMB and HEC in SVM. In summary, the compound information was used the most and was the most important for the prediction model, and protein was the most important for its usage. Meanwhile, knowledge of a disease proved to have little impact on the prediction model.

The similarity features were also calculated with different similarity measure groups (i.e., structural, semantic, and set). Each similarity measure group had a different number of features, as shown in Table 3 (e.g., the structural similarity had 4 features and the semantic similarity had 28 features). Thus, the distribution of features by similarity measure was analyzed by percentage. The average importance of features by similarity measure was also compared. As shown in (a) and (b) of Fig. 9, structural similarity was used the most in both RF and SVM. Moreover, in all three datasets, all structural features were used in RF, but there were no semantic features in SVM. And set features were used the second most. As depicted in (c) and (d) of Fig. 9, structural and semantic similarities had higher levels of importance than set similarity in RF, and structural similarity had the highest level of importance in OMB and HEC in SVM. In summary, structural similarity was used the most and was the most important for the prediction model. Semantic similarity in RF and set similarity in SVM were more important in regards to their usages.

4. Discussion

In our previous study (Yea et al., 2016), we developed a targeted selection method and showed the possibility of inferring traditional empirical ethno-pharmacological knowledge using non-curated biomedical knowledge. In this study, we improved the performance of our previous study, as shown in Table 12, with a novel method exploiting drug repositioning technique based on curated biomedical knowledge. When compared to random selection, our consecutive studies have enhanced the capability to predict herbs with similar efficacy by about 400–4600%. While building the prediction model for herbs with similar efficacy, we identified compound information as being the most important knowledge source and structural similarity as being the most useful measure. These methods and results prove that it might be able to shed light on finding a way of more efficient and effective application of natural products in traditional medicine and new drug development.

To analyze the internal processes of the THED framework, we adopted Peucedanum root and Angelica Dahurica root. The original plants of Peucedanum root are *Peucedanum praeruptorum* Dunn and *Angelica decursiva* Franchet et Savatier that belong to the *Umbelliferae* family. In Korea, Japan, and Taiwan, both original plants are classified as Peucedanum root. However, in China, they are identified as different herbs but are treated as having the very similar efficacy to each other. Both original plants are widely distributed through Korea and China, and are commonly harvested in winter. Their dried root-rhizome is used as Peucedanum root, which is known to possess the effects of inhibiting platelet aggregation, discharging

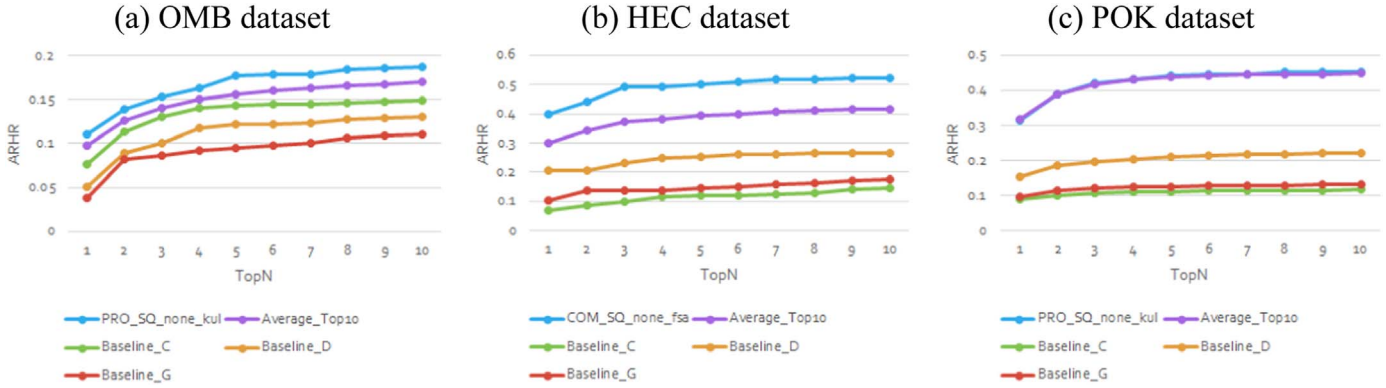

Fig. 5. The ARHR of top1 feature and average of top10 features against baseline. The top1 features are shown in its name.

Table 7
The AUC of ARHR of top1 feature and average of top10 features against baseline.

Dataset	Category	AUC	Dataset	Category	AUC	Dataset	Category	AUC
OMB	PRO_SQ_none_kul	0.1660	HEC	COM_SQ_none_fsa	0.4923	POK	PRO_SQ_none_kul	0.4249
	Average_Top10	0.1500		Average_Top10	0.3834		Average_Top10	0.4225
	Baseline_C	0.1337		Baseline_C	0.1158		Baseline_C	0.1101
	Baseline_D	0.1111		Baseline_D	0.2457		Baseline_D	0.2043
	Baseline_G	0.0919		Baseline_G	0.1486		Baseline_G	0.1238

Table 8
Top10 features of each dataset by recall.

Rank	OMB		HEC		POK	
	Feature	Recall	Feature	Recall	Feature	Recall
1	PRO_SQ_none_kul	0.402778	COM_SQ_none_fsa**	0.84	PRO_SQ_none_kul	0.739474
2	COM_SQ_none_och	0.391892	PRO_CC_none_och	0.826087	COM_SQ_none_fsa**	0.738916
3	PRO_SQ_none_fsa	0.388889	DIS_HPO-ID_wang_fsa	0.826087	COM_SQ_none_fsm	0.726601
4	PRO_SQ_none_och*	0.388889	COM_SQ_none_fsm	0.76	PRO_SQ_none_och*	0.726316
5	DIS_DO-ID_lin_fsm	0.388889	DIS_HPO-ID_lin_fsa	0.73913	COM_SQ_none_kul	0.716749
6	COM_SQ_none_kul	0.378378	DIS_DO-ID_none_kul	0.695652	COM_SQ_none_och	0.716749
7	COM_SQ_none_fsa**	0.378378	DIS_DO-ID_none_och*	0.695652	PRO_SQ_none_fsa	0.707895
8	DIS_DO-ID_none_kul	0.375	DIS_HPO-ID_none_kul	0.695652	PRO_SQ_none_fsm	0.678947
9	DIS_DO-ID_wang_fsm	0.347222	PRO_SQ_none_och	0.695652	PRO_MF_none_kul	0.655263
10	PRO_MF_lin_fsm	0.347222	PRO_BP_none_och	0.695652	PRO_BP_none_kul	0.65

*, **: common features in all three datasets.

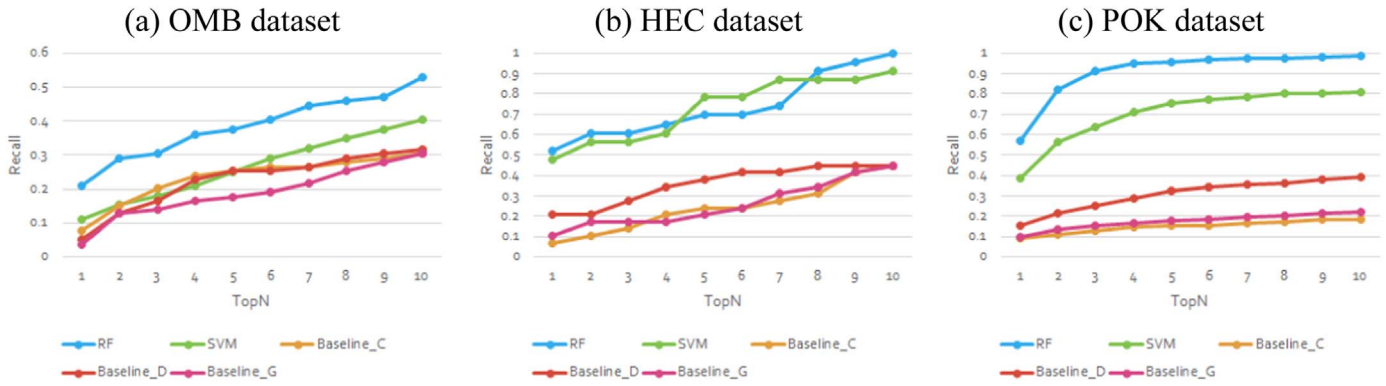

Fig. 6. The recall of prediction model against baseline.

Table 9
The AUC of recall of prediction model against baseline.

Dataset	Category	AUC	Dataset	Category	AUC	Dataset	Category	AUC
OMB	RF	0.3847	HEC	RF	0.7391	POK	RF	0.9090
	SVM	0.2639		SVM	0.7304		SVM	0.7022
	Baseline_C	0.2329		Baseline_C	0.2448		Baseline_C	0.1487
	Baseline_D	0.2253		Baseline_D	0.3586		Baseline_D	0.3053
	Baseline_G	0.1886		Baseline_G	0.2586		Baseline_G	0.1740

phlegm, inhibiting arrhythmia, etc. (Ju, 2013). While Angelica Dahurica root has the effects of acesodyne, antipyretic, and antitumor. Its original plants are *Angelica dahurica* Benth. et Hook. f. and *Angelica dahurica* Benth. et Hook. f. var *formosana* Shan et Yuan, and belong to the same genus and family as one of the original plants of the Peucedanum root. As such, two original plants of Peucedanum root and one of Angelica Dahurica root (i.e., *Angelica dahurica* Benth. et Hook. f.) were compared. The average similarity scores by knowledge source and similarity measure between *Peucedanum praeruptorum*, *Angelica decursiva*, and *Angelica dahurica* were then analyzed, as shown in Table 13. If THED works correctly, the biggest similarity score must be the one between *Peucedanum praeruptorum* and *Angelica decursiva*, because they are the original plants of the same herb. Next is the score between *Angelica decursiva* and *Angelica dahurica*, because they are in the same genus and family despite belonging to different herbs. The lowest score is between *Peucedanum praeruptorum* and *Angelica dahurica*. As noted in Table 13, when the knowledge source was a protein and a disease, slightly abnormal scores occurred, which are indicated in bold and italics. Also, the semantic similarity measure showed opposite results. However, the scores for compound information, structural similarity, and set similarity were signified in accordance with our expectations.

Next, we examined the predicted similarity scores between *Peucedanum praeruptorum, Angelica decursiva,* and *Angelica dahur-*

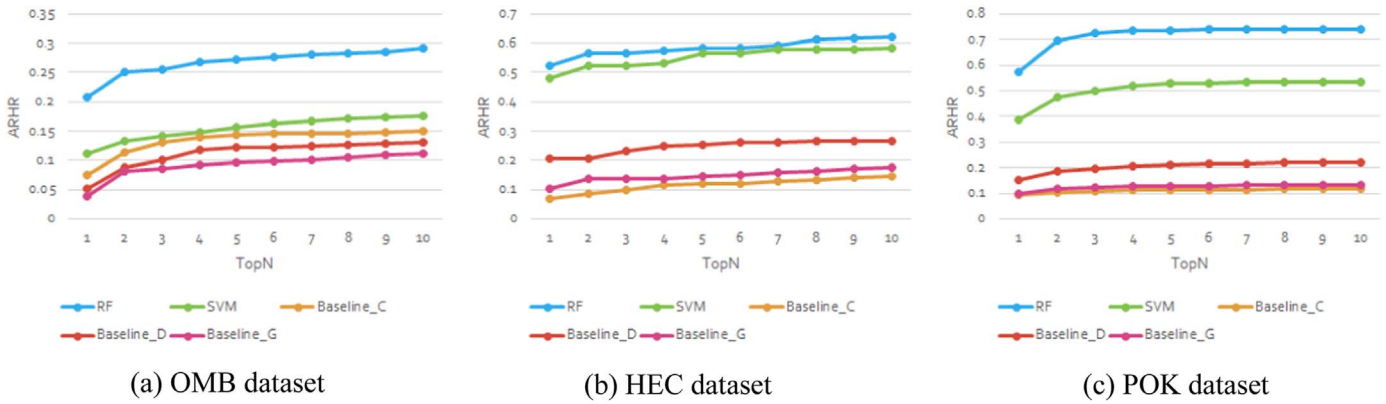

Fig. 7. The ARHR of prediction model against baseline.

Table 10
The AUC of ARHR of prediction model against baseline.

Dataset	Category	AUC	Dataset	Category	AUC	Dataset	Category	AUC
OMB	RF	0.2670	HEC	RF	0.5841	POK	RF	0.7165
	SVM	0.1541		SVM	0.5513		SVM	0.5079
	Baseline_C	0.1337		Baseline_C	0.1158		Baseline_C	0.1101
	Baseline_D	0.1111		Baseline_D	0.2457		Baseline_D	0.2043
	Baseline_G	0.0919		Baseline_G	0.1486		Baseline_G	0.1238

Table 11
The number of selected features for prediction model.

Classifier	OMB	HEC	POK
RF	27	17	14
SVM	8	3	6

ica, as shown in Table 14. These scores were generated from the prediction model of THED. Because a positive score indicates that two plants might have a similar efficacy to each other and negative score means the contrary, RF and SVM showed perfect results. For example, the biggest score of RF was 0.3916 or 0.3491, which occurred between *Peucedanum praeruptorum* and *Angelica decursiva*. The next large score was 0.0403 between *Angelica decursiva* and *Angelica dahurica*. Also, the lowest score was − 0.5207 between *Peucedanum praeruptor-*

Fig. 8. The comparison of knowledge sources in terms of distribution and importance.

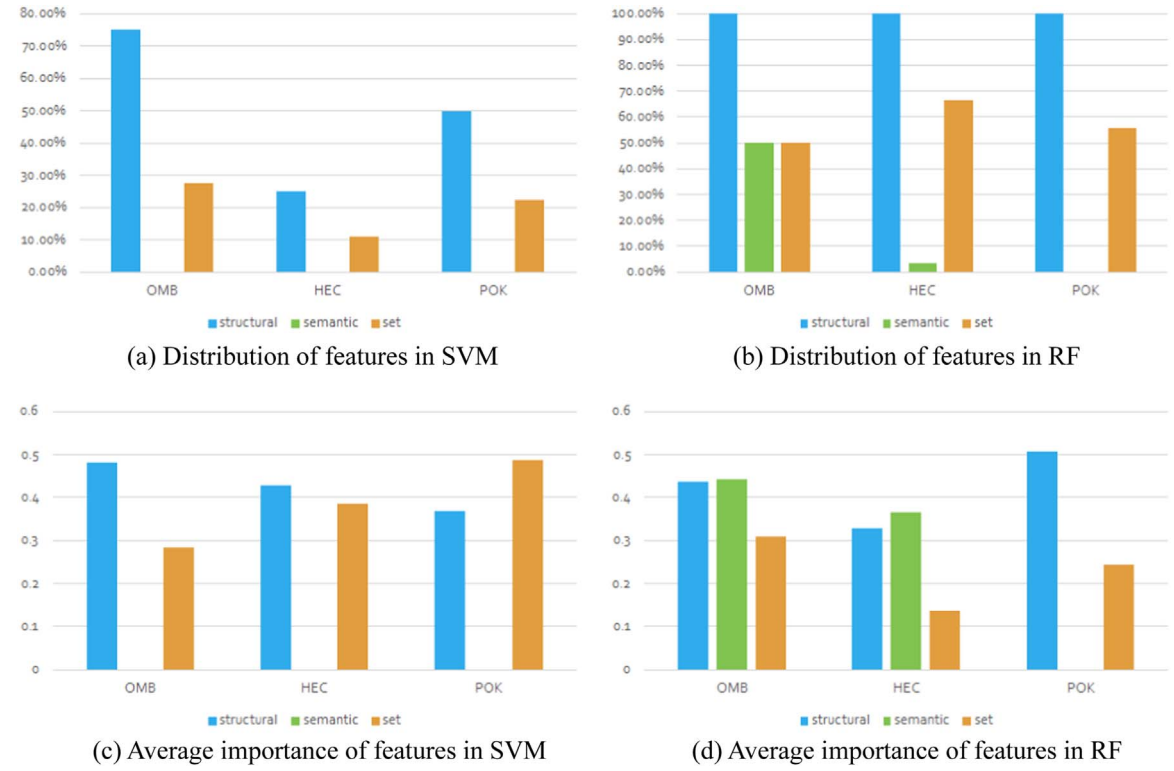

Fig. 9. The comparison of similarity measures in terms of distribution and importance.

Table 12
The comparison of best AUC of recall between studies.

Study	OMB		HEC		POK	
	Method	**AUC**	**Method**	**AUC**	**Method**	**AUC**
This	RF	0.3847	RF	0.7391	RF	0.9090
Yea et al. (2016)	Baseline_C	0.2329	Baseline_D	0.3586	Baseline_D	0.3053
None	Random	0.0833	Random	0.1897	Random	0.0197

Table 14
The predicted similarity score of RF and SVM between the original plants of Peucedanum Root and Angelica Dahurica Root.

Classifier	*Peucedanum praeruptorum*		*Angelica decursiva*	
	Angelica decursiva	*Angelica dahurica*	*Peucedanum praeruptorum*	*Angelica dahurica*
RF	0.3916	− 0.5207	0.3491	0.0403
SVM	0.4289	− 0.3493	0.3535	− 0.0563

um and *Angelica dahurica*. These corresponded with our aforementioned expectations.

5. Conclusions

In this paper, we have proposed a general computational THED framework to improve the possibility of inferring the traditional empirical ethno-pharmacological knowledge using curated modern scientific biomedical knowledge by exploiting the recently developed drug repurposing technique In the proposed framework, we took the knowledge of herbal medicine, chemistry, biology, and medicine into consideration to rank herbs with similar efficacy in candidates. The experimental results demonstrated that the performances of THED framework outperformed the baselines and identified the important knowledge source and useful similarity measure. The proposed method and experimental results support that THED can shed light on finding a way of more efficient and effective application of natural products in traditional medicine and the development of new drugs. Furthermore,

Table 13
The similarity score by knowledge source and similarity measure between the original plants of Peucedanum Root and Angelica Dahurica Root.

Category		*Peucedanum praeruptorum*		*Angelica decursiva*	
		Angelica decursiva	*Angelica dahurica*	*Peucedanum praeruptorum*	*Angelica dahurica*
Knowledge source	**Compound**	0.4534	0.2301	0.4534	0.2825
	Protein	0.5869	0.5498	0.5869	***0.5913***
	Disease	0.6774	***0.7341***	0.6774	***0.7124***
Similarity measure	**Structural**	0.3613	0.1757	0.3613	0.2701
	Semantic	0.7621	***0.8245***	0.7621	***0.8190***
	Set	0.4314	0.3543	0.4314	0.3708

as the large-scale curated databases accumulate more data with higher accurate over time, and then the THED framework can be more usefully applied in diverse ethno-pharmacology fields.

Acknowledgements

This research was supported by the project – "Maximize utilization of knowledge about herbal resource (K17404)" funded by the Korea Institute of Oriental Medicine.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jep.2017.06.048.

References

Bernard, P., et al., 2001. Ethnopharmacology and bioinformatic combination for leads discovery: application to phospholipase A 2 inhibitors. Phytochemistry 58 (6), 865–874.
Buneman, P., et al., 2008. Curated databases. In: Proceedings of the Twenty-seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 1–12.
Chang, Y.W., Lin, C.J., 2008. Feature ranking using linear SVM. WCCI Causa. Predict. Chall., 53–64.
Clark, T.E., et al., 1997. A semi-quantitative approach to the selection of appropriate candidate plant molluscicides—a South African application. J. Ethnopharmacol. 56 (1), 1–13.
Deshpande, M., Karypis, G., 2004. Item-based top-n recommendation algorithms. ACM Trans. Inf. Syst. (TOIS) 22 (1), 143–177.
Disease Ontology Help, 2017. ⟨http://disease-ontology.org/about/⟩ (Accessed 01 2017).
Douwes, E., et al., 2008. Regression analyses of southern African ethnomedicinal plants: informing the targeted selection of bioprospecting and pharmacological screening subjects. J. Ethnopharmacol. 119 (3), 356–364.
Fang, Z., et al., 2013. Replacements of rare herbs and simplifications of traditional chinese medicine formulae based on attribute similarities and pathway enrichment analysis. Evid.-Based Complement. Altern. Med. 2013, Article ID 136732, (9 pages).
Gene Ontology Consortium Help, 2017. ⟨http://geneontology.org/page/about⟩ (Accessed 01 2017).
Goodman, N., 1972. Seven Strictures on Similarity. Bobbs-Merrill, Indianapolis.
Gottlieb, A., et al., 2011. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol. Syst. Biol. 7 (1), Article ID 496, (9 pages).
Grömping, U., 2009. Variable importance assessment in regression: linear regression versus random forest. Am. Stat. 63 (4), 308–319.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.
Guzzi, P.H., et al., 2012. Semantic similarity analysis of protein data: assessment with biological features and issues. Brief. Bioinform. 13 (5), 569–585.
Harvey, A.L., 2008. Natural products in drug discovery. Drug Discov. Today 13 (19), 894–901.
Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. 89 (22), 10915–10919.
Human Phenotype Ontology Help, 2017. ⟨http://human-phenotype-ontology.github.io/about.html⟩ (Accessed 01 2017).
Hurle, M.R., et al., 2013. Computational drug repositioning: from data to therapeutics. Clin. Pharmacol. Ther. 93 (4), 335–341.
Jiang, P., et al., 2007. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res. 35 (2), 339–344.
Johnson, M.A., Gerald, M.M., 1990. Concepts and Applications of Molecular Similarity. Wiley, New York.
Ju, Y.S., 2013. Ungok Herbology. Woosuk Press, Jeonju.
Kann, M.G., 2010. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. Brief. Bioinform. 11 (1), 96–110.
Kim, S.K., et al., 2015. TM-MC: a database of medicinal materials and chemical compounds in Northeast Asian traditional medicine. BMC Complement. Altern. Med. 15 (1), Article ID 218, (8 pages).
Kuhn, M., et al., 2013. STITCH 4: integration of protein–chemical interactions with user data. Nucleic Acids Res. 42, 401–407.
Kulczynski, S., 1927. Die Pflanzenassoziationen der Pieninen. Bull. Int. Acad. Pol. Sci. Lett. Cl. Sci. Math. Nat. Bull., 57–203.
Lee, A.R., et al., 2016. Reduced allergic lung inflammation by root extracts from two species of Peucedanum through inhibition of Th2 cell activation. J. Ethnopharmacol. 196, 75–83.
Lin, D., 1998. An information-theoretic definition of similarity. ICML 98, 296–304.
Marcotte, E.M., et al., 1999. A combined algorithm for genome-wide prediction of protein function. Nature 402 (6757), 83–86.
Martin, Y.C., et al., 2002. Do structurally similar molecules have similar biological activity? J. Med. Chem. 45 (19), 4350–4358.
Medeiros, P.M.D., et al., 2011. The use of medicinal plants by migrant people: adaptation, maintenance, and replacement. Evid.-Based Complement. Altern. Med. 2012, Article ID 807452, (11 pages).
Nagoya protocol, 2017. ⟨http://en.wikipedia.org/wiki/Convention_on_Biological_Diversity⟩ (Accessed 01 2017).
Nikolova, N., Joanna, J., 2003. Approaches to measure chemical similarity–a review. QSAR Comb. Sci. 22, 1006–1026.
Ochiai, A., 1957. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. Bull. Jpn. Soc. Sci. Fish 22 (9), 526–530.
Pan, S.Y., et al., 2013. New perspectives on how to discover drugs from herbal medicines: CAM's outstanding contribution to modern therapeutics. Evid.-Based Complement. Altern. Med. 2013, Article ID 627375, (25 pages).
Paul, S.M., et al., 2010. How to improve R & D productivity: the pharmaceutical industry's grand challenge. Nat. Rev. Drug Discov. 9 (3), 203–214.
Piñero, J., et al., 2015. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database 2015, 1–17.
Podani, J., 2000. Introduction to the Exploration of Multivariate Biological Data. Backhuys, Leiden.
Schlicker, A., et al., 2006. A new measure for functional similarity of gene products based on Gene ontology. BMC Bioinform. 7 (1), Article ID 307, (16 pages).
Sharma, V., Sarkar, I.N., 2013. Bioinformatics opportunities for identification and study of medicinal plants. Brief. Bioinform. 14 (2), 238–250.
Smith, T.F., et al., 1985. The statistical distribution of nucleic acid similarities. Nucleic Acids Res. 13 (2), 645–656.
Strohl, W.R., 2000. The role of natural products in a modern drug discovery program. Drug Discov. Today 5 (2), 39–41.
Tanimoto, T.T., 1957. An Elementary Mathematical Theory of Classification and Prediction. IBM Internal Report.
Terstappen, G.C., Reggiani, A., 2001. In silico research in drug discovery. Trends Pharmacol. Sci. 22 (1), 23–26.
Wang, J.Z., et al., 2007. A new method to measure the semantic similarity of GO terms. Bioinformatics 23 (10), 1274–1281.
Wang, Y., et al., 2013a. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. PLoS One 8 (11), e78518.
Wang, Y., et al., 2013b. Computational study of drugs by integrating omics data with kernel methods. Mol. Inform. 32, 930–941.
Weckerle, C.S., et al., 2011. Quantitative methods in ethnobotany and ethnopharmacology: considering the overall flora—Hypothesis testing for over-and underused plant families with the Bayesian approach. J. Ethnopharmacol. 137 (1), 837–843.
Wright, P.E., Dyson, H.J., 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 293 (2), 321–331.
Wu, Z., et al., 2013. Network-based drug repositioning. Mol. Biosyst. 9 (6), 1268–1281.
Xue, R., et al., 2012. TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis. Nucleic Acids Res. 41, 1089–1095.
Yang, M., et al., 2013. Navigating traditional Chinese medicine network pharmacology and computational tools. Evid.-Based Complement. Altern. Med. 2013, Article ID 731969, (24 pages).
Yea, S.J., et al., 2016. A data mining approach to selecting herbs with similar efficacy: targeted selection methods based on medical subject headings (MeSH). J. Ethnopharmacol. 182, 27–34.
Zhang, L., et al., 2012. Actuality of herbs replacement in the application of recipes of past dynasties based on the data collected from National major projects of infectious diseases. Chin. J. Basic Med. Tradit. Chin. Med. 11, 45.
Zhang, P., et al., 2014. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. AMIA Annu. Symp. Proc., 1258–1267.