Abstract

**Background**: In the field of traditional medicine, there has been a great deal of research on the effects exhibited by medicinal material components and their mechanisms of action. However, since there are no systematic resources of components of medicinal materials, effort to investigate these components on a study-by-study basis is required. Several medicine databases have already been built using classical literature and published books. However, each database provides different information, and the origin of each chemical compound is not clearly stated. The aim of this study was to systematically construct a database of medicinal material components based on articles listed in PubMed Central.

**Description**: Journal articles listed in PubMed Central were searched for sentences containing the names of 799 medicines as well as the names of compounds in the PubChem Substance database; those relating to minerals and animals were excluded. After confirmation that the compound names mentioned were medicinal material components, a compilation of information for 676 compounds was built for 154 different medicines. The database also provides a clear indication of the source articles from which each material component was derived.

**Conclusions**: A medicinal compound database clearly referencing and integrating different sources was constructed. The results of this study, though directly constructed, will serve as a foundation for people to conduct research with automated annotations for medicinal compound names. This database of information on medicinal compounds also enables the discovery of new knowledge contained in the classic literature of traditional medicine and the scientific evidence of modern medicine.

**Keywords**

medicinal materials, compound, chemical names, annotation, literature, traditional medicine
**Background**

In the field of traditional medicine, many texts have addressed the effects of medicinal materials used to treat disease, and the number is continually rising. Such texts generally describe a mechanism by which the medicinal material components have a given effect on a gene or enzyme. Most texts, however, do not describe the effects of every component in a medicine, but introduce the effects of a few selected components. Moreover, while chemical compounds are introduced as having been extracted from specific medicinal materials in some texts, the same chemical compounds can also be extracted from different medicinal materials.

Elsewhere, knowledge of the effects and use of medicinal materials in traditional medicine has been described in classical texts since ancient times. The knowledge imparted in this literature is based on historical, long-term clinical experience rather than on modern, scientific evaluations in academic papers. There is thus no content regarding the components of the medicinal materials, and no description of which diseases the medicinal materials are effective against in clinical practice.

Recently, several medicinal material databases comprising medicinal material components have been constructed based on classical literature in the field of traditional medicine or modern research results; Table 1 presents representative examples thereof.

The “Traditional Chinese Medical Literature Analysis and Retrieval System” (TCMLARS) provides search and analysis functions for biomedical journal articles published in China. It has fields for paper topics, authors, and abstracts, as in MEDLINE [4], and includes fields like TCM drug pharmacology and drug compatibility.

TCM Online is a database constructed by the Information Institute of TCM Academy, which provides integrated access to databases of patient records, TCM, traditional Chinese drugs, TCM literature, and traditional Tibetan drugs. TCMLARS is one example of the databases in TCM Online. TCM Online and TCMLARS provide information on chemical compounds for all medicinal materials, but the sources of this information are not clear.

The Chemical Database of Traditional Chinese Medicine (CHEM-TCM) is comprised of 2 databases. One is a database for 8,264 chemical constituents from 240 medicinal materials most commonly used in China; the other is a database on the target specificity of bioactive plant compounds. The former, in particular, extracts data from the books “The Chemical Constituents of Oriental Herbs” [25], “Chinese
Materia Medica: Chemistry, Pharmacology & Applications” [26], “Traditional Chinese Medicines: Molecular Structures, Natural Sources & Applications” [27], “Handbook of Phytochemical Constituents of GRAS Herbs and Other Economic Plants” [28], and “Dictionary of Natural Products”[29].

The Traditional Chinese Medicine Database (TCMD) was constructed from a book called “Traditional Chinese Medicines: Molecular Structures, Natural Sources and Applications” [27]. The database includes information on Chinese medicines, original plants, and bioactive compounds, and contains 6,800 molecular compounds isolated from more than 1,540 species of natural plants used as TCM as well as plants, herbs, animals, and fungi.

The 3D Structural Database includes biochemical components extracted from medicinal materials in TCM. The database contains 2,083 TCM herb records from 298 families and 10,564 herbal component records. It also includes an optimized 3D structure for at least 90% of the components. Approximately 80% of the records are from a systematic survey of the literature published in Chinese since 1980.

The Traditional Oriental Medicine Database (TradiMed) contains drug formulations from traditional Oriental medicine, constituent chemical compound components of medicinal materials, botanical and taxonomical information on medicinal materials, processing methods, clinical case studies for medicinal materials, and disease classifications. The site is presently offered in Korean and English, but does not specify sources.

There is thus a great deal of data on medicinal materials and their constituent components in the field of traditional medicine. The textual content relating to the constituent components, however, is fragmented and cannot be integrated with information on chemical compounds. Databases on medicinal materials primarily contain data systematically compiled from the classical literature and published books, and Chinese-language articles, rather than English-language articles; however, the sources for the individual chemical compounds are not clear. Moreover, because of the differences in the chemical compound information provided by these databases, it is difficult to determine which database to use to obtain information on chemical compounds. Beyond such medicinal material databases, government ministries of some nations have published pharmacopoeias [1-3], but these only contain major chemical compounds, and only a small number of constituent components of medicinal materials are searchable.
The aim of this study was to construct a traditional medicinal material database whereby sources could be clearly stated and systematically integrated. To this end, research focused on searching journal articles in PubMed Central [19] for medicine and chemical names and detecting relationships between the names of medicines and chemical compounds.

The scientific community provided significant annotations of gene and protein names in the literature [15], and several studies have proposed methods for detecting chemical compound names in the literature.

Zhang et al. [16] proposed an automated chemical name annotation system. They used MeSH vocabulary to annotate chemical compound names for 26,000 MEDLINE articles, and obtained up to 96% precision with 28% recall versus the results derived from manual annotations by National Library of Medicine (NLM) indexers.

Klinger et al. [17] proposed a machine-learning approach based on Conditional Random Fields (CRF) in order to detect International Union of Pure and Applied Chemistry (IUPAC) and IUPAC-like chemical names. In that study, an F-measure of 85.6% was obtained for the MEDLINE corpus.

Hettne et al. [18] proposed a chemical dictionary for identifying small molecules and drugs in text. This dictionary was built through a normalization process by gathering data from multiple chemical resources and provided 67% precision and 40% recall.

Such studies have proposed methods for detecting chemical compound names within the entire host of biomedical literature. However, there has been no proposal for a method to discover the relationships between different elements such as between chemical compound names and medicinal materials.

**Construction and content**

**Data sources**

The scientific literature from which information on medicinal materials and their constituent components was extracted made it possible to use the online archive of biomedical journal articles provided by PubMed Central (PMC). PMC is a free digital database of full-text scientific articles in the biomedical and life science fields, developed by NLM. Journal articles are provided in XML format and can be downloaded from an ftp site [19].
A single medicinal material can be called by various different names. Despite the existence of binomial nomenclature, which is a formal system for giving names to species of living things, each country or community has its own collection of common names. Therefore, we used common names, Latin names, and scientific names (binomen) of medicinal materials mentioned in the Korean Pharmacopoeia [1], Chinese Pharmacopoeia [2], and Japanese Pharmacopoeia [3] published by the Korean, Chinese, and Japanese governments in order to provide the names of objectively proven medicinal materials. After excluding minerals and animals from the medicinal materials used in the field of traditional medicine, 799 plants were used for searches.

Although the nomenclature for chemical names is regulated by IUPAC, chemical compound names have a variety of synonyms. Synonymous terms for chemical compounds in the PubChem database, which provides information on the biological activities of small molecules, were used. PubChem is connected to the PubChem Substance, PubChem Compound, and PubChem BioAssay databases, each of which has substance information, compound structures, and bioactivity data. The PubChem Substance database contains chemical substance information including chemical structures, synonyms, and descriptions; synonyms for chemical compound names were extracted from the synonym information housed here. The PubChem Substance database permits downloading from the PubChem ftp site [20] in a compressed XML file format. In the XML files, synonym information is provided as an element tag “PC-Substance_synonyms_E”; about 73 million chemical compound names are assembled.

Construction

The medicinal material chemical compound database was constructed with PMC articles as follows. First, XML parsing was used to extract article abstracts and main bodies, and the Lingpipe Sentence Model [21] was employed to create segments with constituent sentences. Segmented sentences were indexed using a StandardAnalyzer, which is a basic analyzer by Apache Lucene [22]. Sentences containing medicinal material names were searched from the index, and results were stored as a separate index file using a WhitespaceAnalyzer to which a lower-case filter had been added. During the medicinal material name search, the slop value was set to 10 such that the search would be carried out even when the word order was switched in medicinal material names present in two or more places. In the following step, no search was conducted during the analysis of the chemical
compound names and indexing was performed using the WhitespaceAnalyzer to prevent analysis of chemical compound names. Then, sentences containing chemical compound names were searched from the index. In the sentences thus searched, the fact that one sentence might contain both a medicinal material name and a chemical compound name does not mean the sentence describes the chemical compound as a constituent component of that medicinal material. A Doctor of Korean Medicine and a biology major read the statements directly and determined whether the chemical compound was a constituent component of the medicinal material.

**Content**

When information on medicinal material constituent components was extracted from journal articles in PubMed central, 676 chemical compounds were detected for 154 medicinal materials. Chemical compound names do not have one representative name among their synonyms, so synonyms for chemical compound names were not processed, to ensure they would be used as written in the articles.

Table 2 shows the top 10 medicinal materials in terms of the number of constituent components from the 154 medicinal materials, and shows the number of chemical compounds therein. The medicinal materials in this table clearly have a large number of constituent components, and many studies relating to the constituent components of these medicinal materials are underway.

Table 3 shows the top 10 constituent components from 676 constituent chemical compound components contained in a large number of medicinal materials; it also shows the number of medicinal materials in which the chemical compounds are contained. Based purely on the results in this table, it is clear these compounds are important by virtue of the large number of medicinal materials in which they are contained. These chemical compounds, however, are names representing different kinds of widely used compounds; therefore, they would be detected at a considerable frequency in the literature. As an example, saponins are a class of glycoside; saponins that are extracted from ginseng are called ginsenosides. Each medicinal material also contains the chemical compounds at different constituent ratios. As such, the fact that a chemical compound is included in a medicinal material does not mean the effects of the chemical compound will be exhibited in all such medicinal materials.

Information on the chemical compounds of medicinal materials can be searched on
Figure 1 shows the results of a search for the keyword “Ephedr” on this site. When a medicinal material name or chemical compound name is inputted into the search field above and the “Search” button is clicked, a list of medicinal materials and chemical compounds having names that include the search query is shown in different respective sections. For the name selection, a medicinal material name is shown in the format “Latin Name (common name), [Korean name (Chinese name)]”, and below the medicinal material name are displayed the scientific names and a list of constituent compounds of the medicinal material. When a constituent compound is clicked, the name of that compound is searched for again. Next to any compound are displayed links to PubMed articles in which the medicinal material and chemical compound have been discussed. Because the medicinal material name used in each article varies from article to article, in reality, the medicinal material name retrieved using Lucene is displayed.

The chemical compound section shows a compound list, and below each compound name, a list of medicinal materials having the compound as a constituent ingredient is shown. As in the medicinal material section, clicking on a medicinal material prompts another search for the name of that medicinal material; next to the medicinal material are displayed links to PubMed articles in which the medicinal material and chemical compound have been discussed, as in the case of the medicinal material names.

All medicinal material and chemical compound information can be downloaded as an ontology file, written in OWL, from the abovementioned search site. The medicinal materials in the present study are the same as the medicinal materials included in the medicinal material ontology built by Jang et al. As such, the medicinal materials of the present study and those in the study by Jang et al. are connected by corresponding sameAs relationships. Figure 2 is a representative example of a Resource Description Framework (RDF) graph for Ephedrae Herba from the ontology. Blue oval nodes represent objects, and green oval nodes denote data values. The rectangular shapes show the class names of the nodes.

**Utility and Discussion**

The database of constituent components in medicinal materials constructed here includes medicinal material names, constituent component names, and links to articles where constituent components are mentioned. Use of this database makes it possible for researchers to check texts for what
constituent chemical compounds a medicinal material contains and allows them to easily check relevant information through reference links to an article.

Much research has addressed Literature Based Discovery (LBD), in which hidden and important connections are sought between items of information embedded in published literature. LBD was first introduced in the Swanson study [30] published in 1986. Swanson found a connection between disjointed literature areas linking Fish Oil and Raynaud’s syndrome. Ever since then, there has been research in the LBD field searching for unknown links in the published literature primarily in the medical domain. The field of traditional medicine in particular has a great deal of knowledge that is based on classic literature. Classic literature, however, lacks information on chemical compounds, and relationships exist for a medicinal material and the illnesses to be treated by the medicinal material. Therefore, the use of information on medicinal materials and chemical compounds, of the relationships between medicinal materials and illnesses in the classic literature, and the relationships between chemical compounds and illnesses as per modern experimentation, enables the discovery of new scientific evidence.

This study involved people directly reading texts to detect the entire corpus of information on medicinal materials and constituent components. However, the fact that articles are continually being published means that methods of automation are needed. Use of the database constructed here could allow for the construction of Named Entity Recognition (NER) for medicinal materials and constituent components. It would also be possible to automate the search for medicinal material names and chemical compound names in order to analyze patterns in the relationships between them. For example, phrases like “derived from,” “isolated from,” “comprise,” and “contain” are verbs or verb phrases primarily present between a medicinal material and its constituent components. Although such clearly stated verbs may be apparent, they may also be linked simply by a comma, and sentences that are more complex may link multiple medicinal material names and chemical compound names. Data mining between such medicinal materials and chemical compounds remains a requirement for future research.

In the future, we plan to update the version of the data sources, such as articles from PubMed Central and chemicals names listed on PubChem, used in this study. Currently, the articles are distinguished by the file names, and each sentence of the manuscript has a unique sentence identifier number. Chemical information is updated daily in PubChem site and the downloading of the updated
“Substance” information is possible. However, as articles and chemical information are continuously added with time, an automated system is necessary to efficiently manage our database. Therefore, we will study a method and a system for updating the database based on data mining. Furthermore, using this system, we will analyze additional articles and compounds, and the results will be updated in our database for the relevant medicinal materials.

Conclusions

In the field of traditional medicine, a large number of researchers work by exploring the effects of extracts of medicinal materials. Exploring the effects of medicinal materials primarily requires knowledge of what components constitute those materials. Researchers manually search for constituent components by querying articles in PubMed. However, not all of the chemical compounds in a medicinal material are stated within an article, and there are fragmented mentions only of several main chemical compounds. Recently, it has become possible to make use of medicinal material databases that contain the constituent chemical compounds of medicinal materials. However, not only does each database provide different information on such chemical compounds, it is difficult to identify the sources of this information; therefore, it is challenging to determine which database to use for chemical compound information. Therefore, we constructed a medicinal compound database that clearly references and integrates different sources. To that end, we constructed a medicinal material constituent component database by searching for medicinal material and chemical component names in articles listed in PubMed Central and by directly identifying whether the compounds were constituent components of the medicinal materials. A clear indication of the source article from which each material component derived was provided.
Availability and requirements

Web page: http://tkm.kiom.re.kr/compound/index.do

License: None

Any restrictions to use by non-academics: None

Acknowledgements

This research was supported by the Ministry of Education, Science and Technology (MEST), under the “Ontology-based Traditional Korean Medicine Knowledge Framework” project, grant K12092.

References

6. The TCM online database searching website [http://cowork.cintcm.com/engine/windex.jsp]
8. TCM Online [http://www.cintcm.com]
### Table 1 Databases of medicinal materials and their chemical compounds

<table>
<thead>
<tr>
<th>Databases</th>
<th>Descriptions</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional Chinese Medical Literature Analysis and Retrieval System (TCMLARS)</td>
<td>TCM bibliographic literature database containing TCM journal articles; includes information on pharmacology, drug compatibility, etc. of Chinese herbs</td>
<td>Feng et al. [5, 6]</td>
</tr>
<tr>
<td>TCM online</td>
<td>Provides multiples databases like TCM medicine database, TCM drug database, TCM literature database</td>
<td>Lukman et al. [7, 8]</td>
</tr>
<tr>
<td>Chemical Database of Traditional Chinese Medicine (CHEM-TCM)</td>
<td>Chemical constituents database for TCM medicinal materials and target specificities database construct for bioactive plant compounds</td>
<td>Ehrman et al. [9, 10]</td>
</tr>
<tr>
<td>Traditional Chinese Medicine Database (TCMD)</td>
<td>Provides information on Chinese medicines, original plants, bioactive compounds</td>
<td>He et al. [11]</td>
</tr>
<tr>
<td>3D structural database</td>
<td>3D structure of medicinal material components for searching TCM medicines and molecular information</td>
<td>Qiao et al. [12, 13]</td>
</tr>
<tr>
<td>Traditional Oriental Medicine Database (TradiMed)</td>
<td>Database containing drug formulations, medicinal material constituent chemical compounds, biological information, and disease classifications</td>
<td>Chang et al. [14]</td>
</tr>
</tbody>
</table>

### Table 2 List of medicinal materials containing a large number of constituent components

<table>
<thead>
<tr>
<th>Medicinal material name</th>
<th>Number of chemical compounds per medicinal material</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycyrrhizae Radix et Rhizoma</td>
<td>69</td>
</tr>
<tr>
<td>Ginseng Radix</td>
<td>53</td>
</tr>
<tr>
<td>Chemical compound name</td>
<td>Number of medicinal materials in which the chemical compound is contained</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>saponin</td>
<td>26</td>
</tr>
<tr>
<td>glycoside</td>
<td>18</td>
</tr>
<tr>
<td>polysaccharide</td>
<td>18</td>
</tr>
<tr>
<td>polyphenol</td>
<td>12</td>
</tr>
<tr>
<td>tannin</td>
<td>10</td>
</tr>
<tr>
<td>quercetin</td>
<td>10</td>
</tr>
<tr>
<td>triterpenoid</td>
<td>10</td>
</tr>
<tr>
<td>polyphenolic</td>
<td>8</td>
</tr>
<tr>
<td>glucoside</td>
<td>8</td>
</tr>
<tr>
<td>sterol</td>
<td>8</td>
</tr>
</tbody>
</table>
Figure 1 Results of a search for the keyword “Ephedr”

Figure 2 RDF graph for Ephedrae Herba in the ontology
Medicinal Material

**Ephedrae Herba (Ephedra Herb)**, [마황 (麻黃)]

Scientific Names: Ephedra sinica Stapf.
- Ephedra intermedia Schrenk et C. A. Mey
- Ephedra intermedia Schr et C. A. Mey
- Ephedra equisetina Bunge
- Ephedra equisetina Bge.
- Ephedra sinica Stapf
- Ephedra equisetina Bunge.
- Ephedra intermedia Schrenk et C. A. Mey.
- Ephedra intermedia Schr et C. A. Meyer

Compounds (with the name of the medicinal material in an article) [8]
- norepinephrine: Ephedrae Herba [1]
- epinephrine: Ephedra Herb [1]
- phenylpropanolamine: Ephedra Herb [1]
- catechin: Ephedrae Herba [1]
- methylephedrine: Ephedra Herb [1]
- ephedrine: Ephedrae Herba [1][2][3], Ephedra sinica Stapf [1], Ephedra Herb [1][2][3]
- sympathomimetic agent: Ephedra Herb [1]
- pseudoephedrine: Ephedra Herb [1]

**Compound**

- ephedrine
  
  Medicinal Materials [1]
  - Ephedrae Herba (Ephedra Herb), [마황 (麻黃)]: Ephedrae Herba [1][2][3], Ephedra sinica Stapf [4], Ephedra Herb [5][6][7]

- methylephedrine
  
  Medicinal Materials [1]
  - Ephedrae Herba (Ephedra Herb), [마황 (麻黃)]: Ephedra Herb [1]