BIOCHIP JOURNAL, Vol. 2, No. 4, 274-279, December 2008

THINKherb: *T*he *H*erb *IN*formation *K*nowledge base
– The Chip Content Database for Herbal Medicine

Banggul Oh[1], ChanHwi Um[2], MiKyung Lee[1], Hyunsu Bae[1], MooChang Hong[1], MinKyu Shin[1] & YangSeok Kim[1,2]

[1]Department of Physiology, College of Oriental Medicine, Kyung Hee University, #1 Hoeki-dong Dongdaemoon-gu, Seoul 130-701, Korea
[2]Bioinformatics Unit, ISTECH Inc. No. 506, Woongshin Art Plaza, 847 Janghang2-dong, Ilsan-gu, Goyang-si, Gyeonggi-do 411-837, Korea
Correspondence and requests for materials should be addressed to Y.S. Kim (yskim1158@khu.ac.kr)

Accepted 27 October 2008

Abstract

Motivation: Herbal products are widely used in the field of medicine. The effect of herb is well-known in the field of medicine[1]. It is applied widely from simple pain controls to many severe diseases. Yet, due to shortage of scientific evidence on molecular mechanism of herbs, it had remained as only an alternative choice to conventional drugs. With increase in clinical use, researchers are growing interest in scientific proof and molecular analysis of herbal effects, and the number of herb-related articles is increasing rapidly. In this information deluge, an efficient information system is essential. Although there are many herb databases, molecular information is more of an unexplored area. Therefore, we launched a novel scheme to construct an intelligent herb information system covering extensively from clinical applications to molecular mechanisms. The constructed database can be used for content generation of expression chip for the mechanism study of herbal medicine.

Results: THINKherb is an innovative herb database to provide integrated, knowledge-based information. It is the first in herb information system to combine clinical herb data with molecular information. In addition, another characteristic of THINKherb is expansion of data to provide potential herb interaction information. The DB contains: 499 herbs, 1,238 genes (human. mouse. rat included), 825 diseases, 245 pharmacological activity, and 373 signaling pathways (Human 148.Mouse 121.Rat 104). Entire data were manually annotated by experts for accuracy. The database is available on website http://220.127.168.145:8888/herb/.

Keywords: Database, Bioinformatics, Evidence based medicine, Medical informatics, Knowledge representation

Introduction

In the field of medicine, scientists and practitioners routinely confront extensive number of information and data scattered throughout the published literature. Currently there are more than 120,000 scientific journals, and at least 500,000 medical articles and 4,000,000 scientific articles are being published every year[2]. In fact, PubMed holds more than 9giga bytes of information and 15,575,607 numbers of abstracts based on the tally updated to 2006[3].

In this information deluge, a researcher would have to scan 130 different journals and read 27 papers per day to follow a single disease, such as breast cancer[4].

The situation is same in herbal medicine. Currently, with growing interest in scientific proof and molecular analysis of herbs, the number of herb-related articles is increasing rapidly. Already, more than 20,000 herb-related articles have been published, and there are more than 600 herb-related journals[5].

With such abundant information scattered throughout published literature, and not in any form of database for easy access, delay in any information retrieval is inevitable. This brings inaccuracy in clinical use of herbs, and delay in scientific research.

Therefore, the need of system development to manage this surge of information is crucial. There are several herb databases already in use. However, contents in most databases are limited to clinical applications. Molecular information is not yet commonly approached in the area of herbal medicine. Also, they do not provide integrated information, but rather scattered information[6].

We have constructed a new herb information system to overcome the limitation of existing systems. The aims of this project are: 1) to establish a knowledge-based herb information system covering entire currently published literature based on PubMed. 2) to make an attempt in providing molecular information of herbs for scientific approach, and provide integrated herb information in correlation with diseases, genes and pharmacological activity. 3) to maximize the accuracy of information by thorough manual annotation

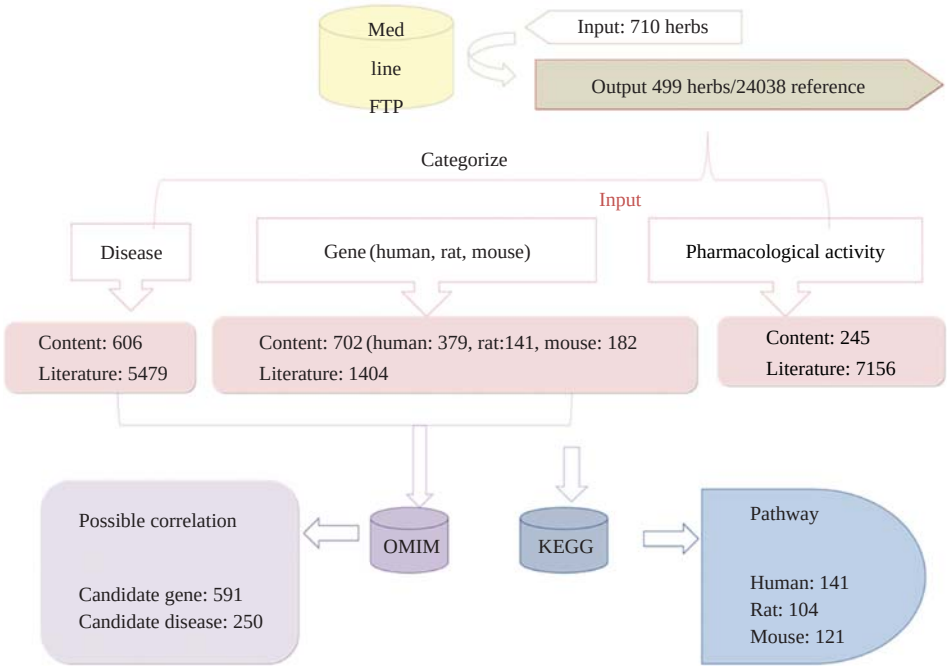

Figure 1. THINKherb establishment flowchart.

and validation.

Also the database has expanded its' information to prediction level based on knowledge and logic (details described in the article) to meet the researchers' needs.

Many up-to-date information technologies and extensive manual annotation were applied to the database to provide users with accurate and scientific knowledge-based information of herbs.

Results

Total number of 499 herbs is covered in the database. Initially, 710 herbs were each listed in its' scientific name. Out of 710 herbs, 503 were found in Medline abstracts. For each of 499 herbs, information was extracted to maximum extent from Medline, and 24038 abstracts were retrieved. The abstracts are the initial set of THINKherb database. Information was categorized resulted as below.

a. Gene: Out of 24,038 initial references, 2,488 articles were found to include information on genes. After redundancy check, final number of 1,404 references remained. Total of 702 genes are covered in the references. For extraction of accurate gene information, manual revision of first extraction result was performed. Also, we inserted additional gene contents that are frequently seen in herb-related articles, but were not extracted the first time due to technical limitations.

b. Disease: Out of the initial data set, 8,987 herb-related articles were found to withhold information on diseases. After redundancy check, final number of 5,479 references was remained. Number of diseases covered in the references is 606.

c. Pharmacological Activity: Out of the initial data set, 7,156 herb-related articles were found to have information on pharmacological activity of herbs. After redundancy check, final number of 5,479 references was remained. Number of pharmacological activities covered in the references is 245.

Figure 1 is an overall flowchart of THINKherb to illustrate the resource and information result of the database.

The THINKherb database is now available on website (http://220.127.168.145:8888/herb/). Figure 2 is a capture image of the main page of THINKherb.

Validation

To check up the confidence of our contents, validation was performed. Validation of data included both ruling out false positives factors, and confirming true positive set to estimate reliability of the whole data set.

"Gene2pubmed" from NCBI was used as a true positive set. Gene2pubmed is a file providing authorized linkage between ncbigene ID and PubMed ID. This true positive set was compared with the data set in the DB gene information, and the contents matched up to 95% giving considerable reliability to contents in the DB. The missing factors were manually search-

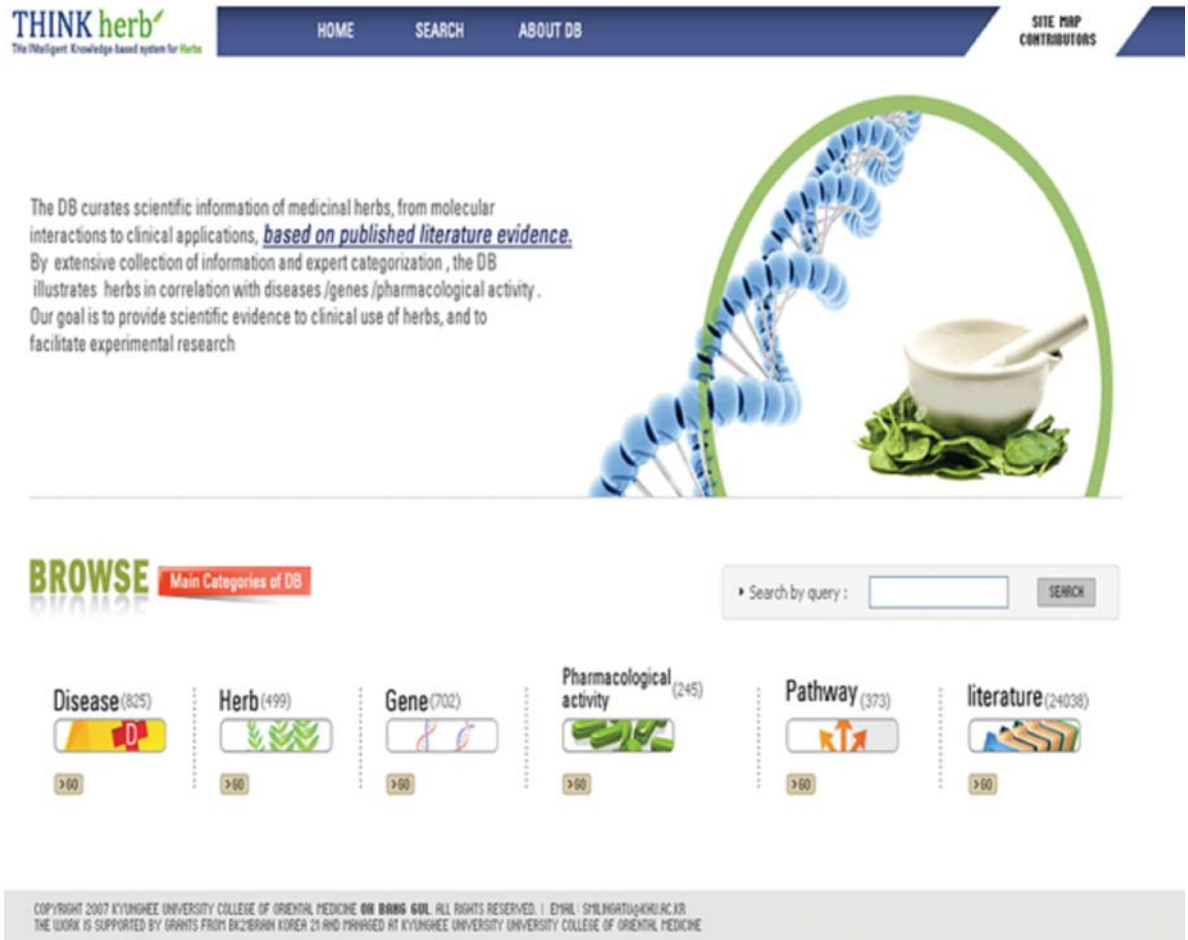

Figure 2. Main page of the constructed database.

ed to analyze reason for omission. One reason was that the gene name was not mentioned in abstracts, whereas my data extraction was based on abstract texts. Another was that the gene name/symbol was too broad for information extraction that it was ruled out due to technical limitations.

Ruling out false positive factors is critical in construction of a reliable database. Many computational methods have been developed for precise data extraction. Yet manual annotation, although the work requires tedious effort, holds its' advantage in elaborateness. The database contents were manually annotated, in repetition, adding credibility to the information.

Application Case

As the object of THINKherb is to provide information of 'herb', the most ideal DB search starts from the scientific name of an herb.

Alphabetically listed in its' scientific name, users can easily carry out a search. Direct input of query or selection by browsing the content list is both possible. For each herb contents, icons are marked that stand for provided information. Click on the icon links to corresponding information.

For example, available information on Schizandra chinensis are all five categories; literature, disease, gene, pharmacological activity and pathway. Users can easily link to detailed information of each category. Furthermore, direct links to websites, such as PubMed or KEGG are possible. In viewing each category, list other herbs that include the same information are provided.

For efficient retrieval of useful results, other search options are also available. For example, by searching a disease, users will be able to obtain a list of herbs that regulates or have potential to regulate (candidate) the query disease. Selection of an herb from the given list equals the result of herb search described above.

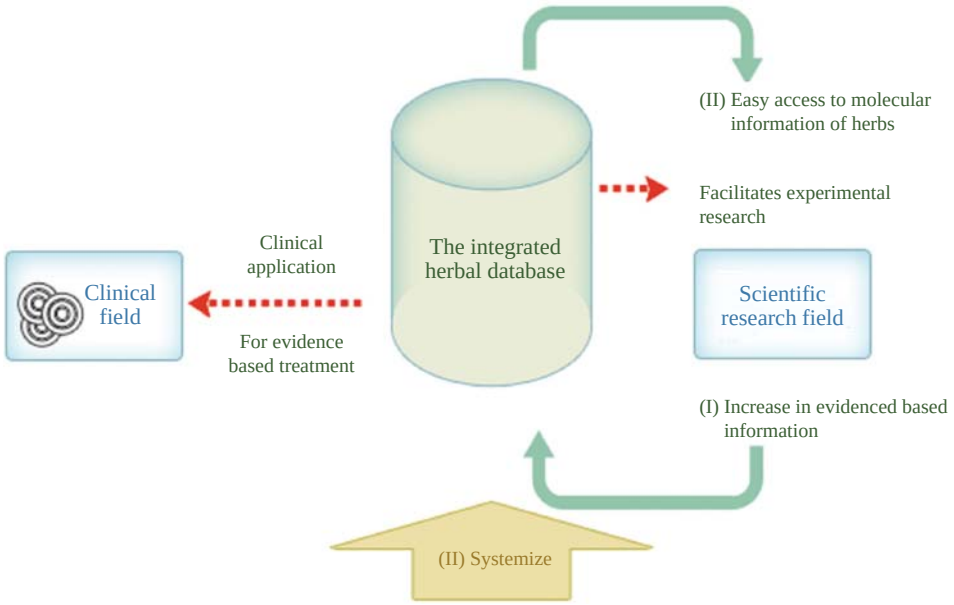

Figure 3. Application of the database in clinical and research fields.

Discussion

This database is an integrated herbal database providing a wide range of herb related information. The database covers from information applicable in clinical use, such as, disease and pharmacological activity, to information on genes and pathways for benefit of researchers. This database is the first in the field of oriental medicine to provide extensive information of herbs including molecular interaction.

Distinguishing features of THINKherb can be summarized as the followings.

First, it was an innovative approach to construct an herb database, with information focused on molecular mechanism of herbs. This is to aid in accelerating experimental research for herbs. Currently, although there are several herb databases, most information is limited to clinical applications. Therefore, in order to carry out a scientific experiment for analysis of herbal effects, a researcher has to go through tedious and time-consuming work in search for right information. With THINKherb allowing easy access to information concerning molecular mechanism of herbs, researchers will be able to save time and effort. Hopefully with acceleration of herb research, scientific evidence of herb effects will be increased. This will be an important contribution in promoting evidence-based practice in oriental medicine[7].

Second, information in THINKherb is not limited to only experimentally proven facts. Logical system was designed to expand the providing information to a certain prediction level. Based on the logical system and technical programming performance, we were able to draw out potential herb effects. Disease with possibility of herb curing effect is named candidate disease. Gene predicted to interact with mechanism of herbal effect is named candidate genes. This expansion in scope of information is a venture in herb information system. But it is expected to play a significant role in promoting experimental research in the field of herbal medicine.

Lastly, the credibility and quality of contents in THINKherb is highly remarkable. Herb contents were selected and categorized by a licensed expert in herbal medicine. Disease contents cover extensively from general diseases to genetic disorders. And most importantly, to maximize the reliability of the database, contents were followed up with thorough and repetitive manual annotation by all participants of this project. This manual work proved to be a dependable complement to limitations of technical procedures.

All in all, THINKherb is expected to benefit both clinicians who use herbal treatment for curing diseases and scientists who study molecular mechanism of herbs. With THINKherb, scientific evidence to use of herbs is just a ‘click away’. This will guide clinicians to enhance accuracy in herbal treatments. As for the research field, easier access to molecular information of herbs will promote experimental researches, thus increasing knowledge of the genes in association with mechanism of herbal effects. This, in the future, will allow researchers to address more complicated issues, such as, herb interaction in level of regulatory network

and systems biology.

The database is a non-profit database, established solely for the purpose of research and development of herbal medicine.

```
unless (open(A, "genenameI-geneID.txt")){
        die ("cannot open file1\n");
}
@array=<A>;
unless (open (B,"geneID-omimID.txt")){
        die ("cannot open file2\n");}
@array2=<B>;
$count1=0;count2=0;
#$temp3_re="";
$length1=@array;
$length2=@array2;
while($count1<$length1){
        $temp1=@array[$count1];
        $temp3=@array[$count1+1];
  $temp2=@array2[$count2];
        $temp4=@array2[$count2+1];
        while($count2<$length2){
                $temp1=@array[$count1];
        $temp3=@array[$count1+1];
  $temp2=@array2[$count2];
        $temp4=@array2[$count2+1];
                if($temp2=~/$temp3/){
          #$temp3=~ s/\n//;
                        $temp2=@array[$count1];
        $temp3=@array[$count1+1];
$temp2=@array2[$count2];
        $temp4=@array2[$count2+1];
                }
                $count2++;                }
        $count1++;$count2=0;}
print ("end of loop");
```

Figure 4. Example of Perl script used in information extraction process.

Methods

Content Selection

a. Herb: In order for herb contents to be extensive but at the same time contain reliable sources, content selection was based on the textbook of herbal medicine used in every Oriental Medical school in Korea[8]. To avoid redundancy in extensive synonymy of herb nomenclature, 710 herbs were each listed in its' scientific name for the initial set of the database[9].

b. Gene: Gene contents were derived from HGNC (HUGO Gene Nomenclature Committee) database[10]. To prevent false positive results in information extraction, we ruled out gene symbols of the following factors: i) one-lettered gene symbol, ii) gene symbols in same letters as a common word (i.e. CAT, Sit)[11], iii) gene symbols that are abbreviation of other common meanings (i.e. LD: Linkage disequilibrium ACD: allergic contact dermatitis). Also, we manually added gene contents that are frequently seen in herb-related articles, but were not extracted the first time due to technical limitations.

c. Disease: Disease contents include both general disorders and genetic diseases. General disorder contents were derived from web database of Karolinska Institute[12]. Genetic disease contents were derived from OMIM, Online Mendelian Inheritance in Man, a catalog of human genes and genetic disorders developed for the World Wide Web by NCBI.

d. Pharmacological activity: List of pharmacological activities was derived from NAPRALERT, a natural products database developed by UCI (University of Illinois at Chicago)[13].

Information Extraction

For manipulating large text data, Perl (ver.5.8.8) was used as a programming language[14]. Basic algorithm was to use pattern matching in Perl programming. Unique Perl scripts were generated for performing

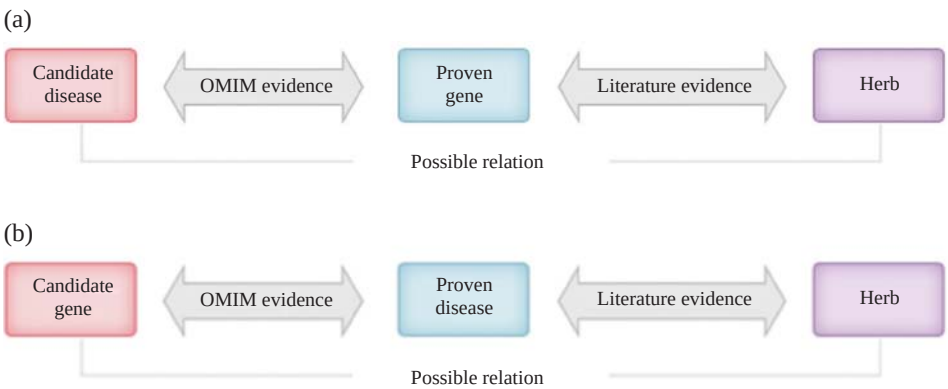

Figure 5. (a) Logic design for prediction of candidate gene, (b) Logic design for prediction of candidate disease.

every kinds of data processing such as data extracting, filtering and changing. Figure 4 is an example of Perl script used in information extraction process.

Literature Collection

The initial data set of THINKherb is set of articles related to the herb contents. It was extracted from Medline files. With acquired license of NLM's database (license code: JCL), we were able to obtain entire Medline information in FTP files updated to September, 2006.

Categorization

Then the initial data set of 24,038 references was categorized according to comprehending information of disease, gene, and pharmacological activity in the references.

Integration & Prediction

Whereas procedures stated above are proven information based on literature, we expanded our data to prediction level using OMIM (Online Mendelian Inheritance in Men) database. From OMIM, we extracted diseases correlating to the genes that were proven to be related to herbs based on Medline. Therefore those diseases are predicted to be possibly regulated by the corresponding herb. Same logic is applied for extracting gene information from OMIM. In order to prevent confusion between gene/disease information from Medline and OMIM, contents from Medline is titled "Proven" gene/disease, those from OMIM is titled "Candidate" gene/disease.

The "proven genes" are once again used to extract pathway information from KEGG. Given that the proven genes interact with herbs, by mapping the genes in pathways we can see herb interaction in regulatory network level.

Development Environment

Oracle 10 g/Linux was selected as a database management system for easy handling and safe storage of created data and Java was used for constructing database-web interface to provide Graphic User Interface.

Acknowledgements

We thank Bioinformatics Unit of ISTECH Inc. for assistance in computational programming. We also thank the College of Oriental Medicine, Physiology Department at KyungHee University for supporting the research.

Funding: This research is funded by Brain Korea (BK) 21 project.

References

1. Chan, E. Quality of efficacy research in complementary and alternative medicine. *JAMA* **299**, 2685-2686 (2008).
2. Ananiadou, S. *et al*. Text mining and its potential applications in systems biology. *Trends Biotechnol*. **24**, 571-579 (2006).
3. http://www.nlm.nih.gov/bsd/Medline.
4. Baasiri, R.A., Glasser, S.R., Steffen, D.L. & Wheeler, D.A. The Breast Cancer Gene Database: a collaborative information resource. *Oncogene* **18**, 7958-7965 (1999).
5. Chen, Y.Z. *et al*. Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation. *Br J Pharmacol*. **149**, 1092-1103 (2006).
6. Wishart, D.S. *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **1**, 34 (2006).
7. Tomoko, O. & Toshihisa, T. Development information extraction tool: Toward construction of protein-protein interaction database, *Genome Informatics*, 292-293 (1998).
8. Kwon, S.B. *et al*. Herbal Medicine, Herbal medicine of Oriental Medical School Compilation Committee, Young Lim Publishing (2004).
9. Farah, M.H. *et al*. Botanical nomenclature in pharmacovigilance and a recommendation for standardisation. *Phytomedicine* **29**, 1023-1029 (2006).
10. Eyre, T.A. Povey, S., Bruford, E.A. & Lush, M.J. The HUGO gene nomenclature database, 2006 updates. *Nucleic Acids Res*. **Jan 1; 34(Database issue)**, D319-321 (2006).
11. Hong, Yu. *et al*. Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. *Journal of Biomedical Informatics* **40**, 150-159 (2007).
12. http://www.mic.ki.se/Diseases.
13. http://www.napralert.org.
14. http://www.perl.org/.