AFRL-AFOSR-JP-TR-2018-0003

Building Vietnamese herbal database towards big data science in nature-based medicine

Ly Le
INTERNATIONAL UNIVERSITY VIETNAM NATIONAL UNIVERSITY-HCM

01/04/2018
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOA
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved

OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
17-09-2017	Final	09-22-2015 – 09-21-2017

4. TITLE AND SUBTITLE

Building Vietnamese herbal database towards big data science in nature-based medicine

5a. CONTRACT NUMBER

FA2386-15-1-4119

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Ly Thi Le

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

INTERNATIONAL UNIVERSITY - VIETNAM NATIONAL UNIVERSITY
1-1 Quarter 6, Linh Trung Ward, Thu Duc District
Ho Chi Minh City, Vietnam

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AOARD
UNIT 45002
APO AP 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)

AOARD

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

15IOA006_154119

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Vietnam carries a highly diverse data of traditional medicine, in which various combinations of herbs were widely used as a remedy for many types of diseases. Poor hand-writing records and current text-based databases, however, perplex the conventionalizing and evaluating process of the canonical therapeutic effects. In efforts to reorganize this valuable information, we provide VietHerb Ontology (VHO) for herbs documented in Vietnamese traditional medicines. The ontology was constructed with temerity to provoke data communication with available ontologies for plants, metabolites, diseases, and geography in order to convey a composite description of each individual species. VHO consists of 2881 species, 10887 metabolites, 458 geographical locations, and 8046 Oriental therapeutic effects. The number of binary relationships of species-metabolite, species-therapeutic effect, species-morphology, and species-distribution are 17602, 2718, 11943, and 16089 respectively. The ontology primarily serves as open sources facilitating users in studies on structure-based drug design and simulation and develops knowledge-based prediction models for deep-learning in future. Our newly built database will be used to explore active ingredients in the effective Vietnamese herbal medicine formulations for individual diseases and to understand therapeutic effects under scientific viewpoint.

15. SUBJECT TERMS

Ontology, Vietnames herbs, metabolites, diseases, species, therapeutic effect, morphology, knowledge-base

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Hiroshi Motoda, Ph. D.
U	U	U	UU	20	19b. TELEPHONE NUMBER *(Include area code)* +81-42-511-2011

Final Report for AOARD Grant FA2386-15-1-4119

Building Vietnamese Herbal Database
Towards Big Data Science In Nature-Based Medicine

Date: 21st September, 2017

Name of Principal Investigators (PI and Co-PIs): Ly Le
- e-mail address: ly.le@hcmiu.edu.vn
- Institution: School of Biotechnology, International University, Vietnam National University, Ho Chi Minh City.
- Mailing Address: Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam
- Phone: (08) 37244270
- Fax: (08) 37244271

Period of Performance: 9/30/2015 – 9/30/2017

Abstract:

Vietnam carries a highly diverse data of traditional medicine, in which various combinations of herbs were widely used as a remedy for many types of diseases. Poor hand-writing records and current text-based databases, however, perplex the conventionalizing and evaluating process of the canonical therapeutic effects. In efforts to reorganize these valuable information, we provide VietHerb Ontology (VHO) for herbs documented in Vietnamese traditional medicines. The ontology was constructed with temerity to provoke data communication with available ontologies for plants, metabolites, diseases, and geography in order to convey a composite description of each individual species. VHO consists of 2881 species, 10887 metabolites, 458 geographical locations, and 8046 Oriental therapeutic effects. The number of binary relationships of species-metabolite, species-therapeutic effect, species-morphology, and species-distribution are 17602, 2718, 11943, and 16089 respectively. The ontology primarily serves as open sources facilitating users in studies on structure-based drug design and simulation and develops knowledge-based prediction models for deep-learning in future. Our newly built database will be used to explore active ingredients in the effective Vietnamese herbal medicine formulations for individual diseases and to understand therapeutic effects under scientific viewpoint, this project predicts therapeutic effects based on metabolite profiles. Since this project can reveal the main predictors of specific therapeutic effect, they are valuable information for further research of herbal medicine and drug development.

1. Introduction:

Aiming to a comprehensive analysis of metabolites in various level of characterized biological units, metabolomics has contributed to the holistic understanding of complex molecular interactions in biological systems by great efforts on investigating unknown impacts of metabolites on phenotypes and biological pathways [1]. In recent years, metabolomics has played an essential role in functional genomic research [2]–[7]. According to a scientific literature review [8], the number of known metabolite varies in a range from 200,000 to 1,000,000 of about 223,000 plant species [9]. To deal with extremely large data, a database (DB) of many systematically addressed relationships between metabolites and their biological origins is highly needed not only for metabolomics research but also for the other disciplines.

We introduce an ontology of herbal metabolomics to community as an official information source for both experts and non-experts who share common interests in herbal species. We constructed VietHerb Ontology (VHO), containing numerous binary relationships of species-metabolite, species-therapeutic effect, species morphology, and species-distribution. The information of herbal species, metabolites, geographical distributions, WHO-standardized diseases and orientally-viewed diseases was semi-automatically and manually collected mainly from well-known online databases and officially published hard-copied documents to achieve high reliability. Repetitive screening for unmatched term; linguistically translated mistakes and random faults in manual work has been ceaselessly and thoroughly processed by our VHO core team.

In this study, we focus on analyzing data structure, performing statistics and applying consensus algorithm on VHO. Our current statistics suggests the ontology for species-metabolite, species-therapeutic effect, species-morphology, and species-distribution contains 17602, 2718, 11943, and 16089 binary relationships respectively encompassing 2881 species, 10887 metabolites, 458 geographical locations, and 8046 Oriental therapeutic effects. The binary relationships between WHO standardized diseases and Oriental diseases currently has not been finished due to unsolved difficulties in data matching and homologous term linking. As of now, this database serves as the largest ontology-based knowledge for Vietnamese herbs.

An initial attempt has been made to classify and predict therapeutic effects from metabolites data using a few machine learning methods, e.g., Random Forest classifier, Generalized Boosted Model and Support Vector Machine classifier. Although we obtained some promising results we also identified various problems that need be solved. Since the results are still premature, we don't include them in this report. This is the main subjects for the follow-up project FA2386-17-1-4032.

2. Method/Theory/Experiment:

2.1 Main principles and data sources

All the information in VHO is non-proprietary and freely assessable to anyone without any conflict of interest. All the data item in VHO is traceable back to the original source. Data mining was processed only with open-source databases. To construct VHO, data from different sources were retrieved and incorporated. The data sources are listed below.

- Cay Thuoc (Medicinal plants): A database contains about 3000 herbal species with morphological features, therapeutic effects and medicinal formula. However, its information is uncategorized and described in plain-text with numerous text errors [10]. Information about morphological feature, distrubution, and therapeutic effect was retrieved from this database.

- Tropicos: A well-known database which has been maintained for more than 30 years. This database contains extremely large information, including about 1.3 million scientific names, 60,000 common 2 names, 4.4 million specimen records, 500,000 images and 140,000 references [11]. Information about scientific and common

name was retrieved from this database to cross check with the others.

- Theplantlist: A database which was developed and maintained under collaboration between the Royal Botanic Gardens, Kew and Missouri Botanical Garden. This database provides users with the accepted Latin names and synonyms for plant species with ranking in reliability. This database contains about 350,000 accepted Latin names, 470,000 synonyms and 240,000 unsolved cases [12]. Information about Latin name and synonym was retrieved from this database to cross check with the others.

- KNapSAcK: A comprehensive Species-Metabolite Relationship database which was developed by NARA Institute of Science and Technology, Japan. This database contain more than 100,000 binary linkages of 20,741 herbal species and 50,048 metabolites. As a database incorporated with systematic analysis, KNapSAcK facilitates metabolomics research with enormous numbers of organic compounds of defined and undefined structures [13]. Information about speciesmetabolite binary relationship was retrieved from this database.

- ChEBI: A freely accessed database containing molecular entities of small chemical compounds, including both natural and synthetic products, which targets on physiological pathways of human-being [14]. Information about molecular structure, biological role and application was retrieved from this database.

For the hard-copied documents

- *Nhung cay thuoc va vi thuoc Viet Nam (The medicinal plants and herbal formulae in Vietnam):* The book written by Professor DO, Tat Loi is a collection of many Vietnamese medicinal plants and herbal formula after decades of ceaseless research. The author first published the book in 1962 and it has been reprinted 14 times up to present. Novel information is added and modified carefully in each reprinted version to provide readers with the most reliable and precise information [15].
- *Tu dien cay thuoc Viet Nam (The dictionary for medicinal plants in Vietnam)*: The book "*The dictionary for medicinal plants in Vietnam*" by Professor VO, Van Chi is a valuable resource for re- searchers in botany, medicinal chemistry and traditional medicine. The author first published the book in 1997 and the latest version of this book was reprinted in 2013 with modified and up-to- date information of about 4700 medicinal plants and 1500 images. The book introduces identification method, classification, and the general use of medicinal plants. In addition, the author also describes pharmacological effect and main active ingredients of medicinal plants [16].

All listed hard-copied documents were used as references for morphological features, distributions, and herbal formulae. Every hard-copied document version in use is the latest version.

2.2 Database construction

VHO was built following METHONTOLOGY. The relational database is implemented in Java and MySQL DB for homogenous querying purposes specifies each attribute for a specific concept. The data were automatically standardized and stored in MySQL with RDF formats before converting into OWL formats [17].

2.3 Database content

Currently, VHO is a family database of 4 sub-databases and 1 main database. The sub-databases include effect, distribution, metabolite and morphology sub-ontology. The main database is herb main-ontology.

- Herb main ontology contains information of plant families, genus, and species (Figure 1.a). Plantae family (e.g. *euphorbiaceae*), genus (e.g. *euphorbia*) and species (e.g. euphorbia hirta) are considered as the primary, secondary and end node respectively. The herbal species are characterized by 4 nominal anotations, including Vietnamese name (herb:vn name), scientific name (herb:herb name), Vietnamese synonym (herb:vn synonym) and Latin synonym (herb:latin synonym) and 3 connective anotations which are linked to effect (herb:has effect), metabolite (herb:has metabolite), and distribution (herb:distribute in).

- Effect sub-ontology contains information of effects caused by herbal species and metabolites (Figure 1.b). The effects are divided into 4 primary nodes, including biological, therapeutic, nontherapeutic and toxical/tumorigenetic/disease- causing effect. The secondary nodes are de- fined as groups of effects (e.g. anticancer) sharing common features in physiological pathways. The end nodes are specific effects which are characterized by 4 nominal anotations, consist- ing of Vietnamese name (effect:vn name), English name (effect:herb name), Vietnamese synonym (effect:vn synonym) and English synonym (effect:latin synonym) and 1 connectve anotation linking to herbal species (effect:from herb). The hierarchy in Effect sub-ontology followed that of Metabolite Activity database of KNApSAcK families database [13].

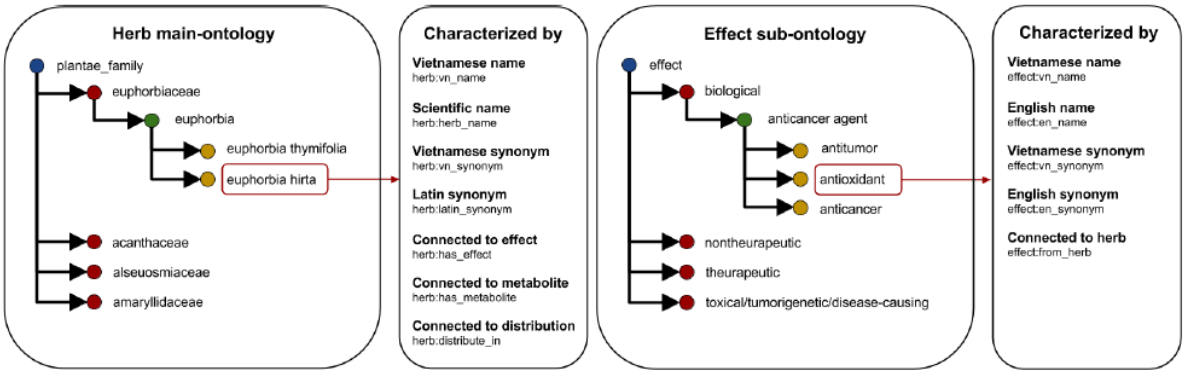
- Distribution sub-ontology contains information of geographical locations having records of plant availability (Figure 1.c). The distribution is categorized into 6 primary nodes which are 6 continents (e.g. africa continent, asia continent) and each continent is then divided into many subregions (e.g. central america, northern europe). The end nodes are nations which are characterized by 3 nominal anotations, including Vietnamese name (distri:vn name), English name (distri: herb name), and Geographical coordinates (distri: geo coordinate) and 1 connective node targeting to herbal species (distri:has herb). The hierarchy in Distribution sub-ontology with geographical coordinates followed up-to-date version of Geoname ontology [18].

- Metabolite sub-ontology contains information of phytochemicals having records of being possessed by herbs. There are 2 options of hierarchy and 1 option of non-hierachy (Figure 1.d). The first option of hierarchy followed phytochemical classification in book" Phytochemical dictionary: a handbook of bioactive compounds from plants" (e.g. alkaloids, terpenoids) [19] and the second option of hierarchy purely followed funtional groups of organic compounds (e.g. alcohol, ester). The secondary nodes in phytochemical option includes main families (e.g. alkaloid, carbohydrates and lipids). Each phytochemical family contains many sub-groups as end nodes (e.g. indole alkaloids, steroidal alkaloids). Functional group node has only 1 sub-node while the unclassed node compounds has no sub-node.

- Morphology sub-ontology contains information of all morphological features of plants, including root style, stem style, leaf style, seed style, flower style and other (Figure 1.e). Currently, they are the only nodes of this database due to incomplete refining process. The terms used for morphological feature description are extremely diverse and highly redundant. Besides the original words or the key words for description, there are millions of inaccurate and clumsy words created by non-experts and non-English-speaking people. Unfortunately, these words have been widespread in many documents (both online and hard-copied references). Text mining is planned before

hand in the next version to correct and modify them.

- There is a pressing need to develop powerful data storing methods. XML, DTDs and XML Schemas, although these are sufficient and good enough for data sharing and connecting different entities, are yet unable to reliably perform tasks with new XML vocabularies due to their lack of semantics. Like other ontologies, VHO is meant to be the fundamental vocabularies for use in applications searching for or integrating diverse information in communities. VHO's ontoboty is to create the interconnective linkages among those individual entities in each separate database (Fig 1). Every input of each sub-ontology is capable of returning output of Herb main ontology and vice versa. Effect and Metabolite sub-ontology can also interconnect to each other's, without going through main ontology (e.g. input of effect keywords can return outputs of metabolite and vice versa). Morphology and Distribution sub-ontology only interconnect to Herb main ontology, but not the others.



(a) Herb main ontology (b) Effect sub-ontology

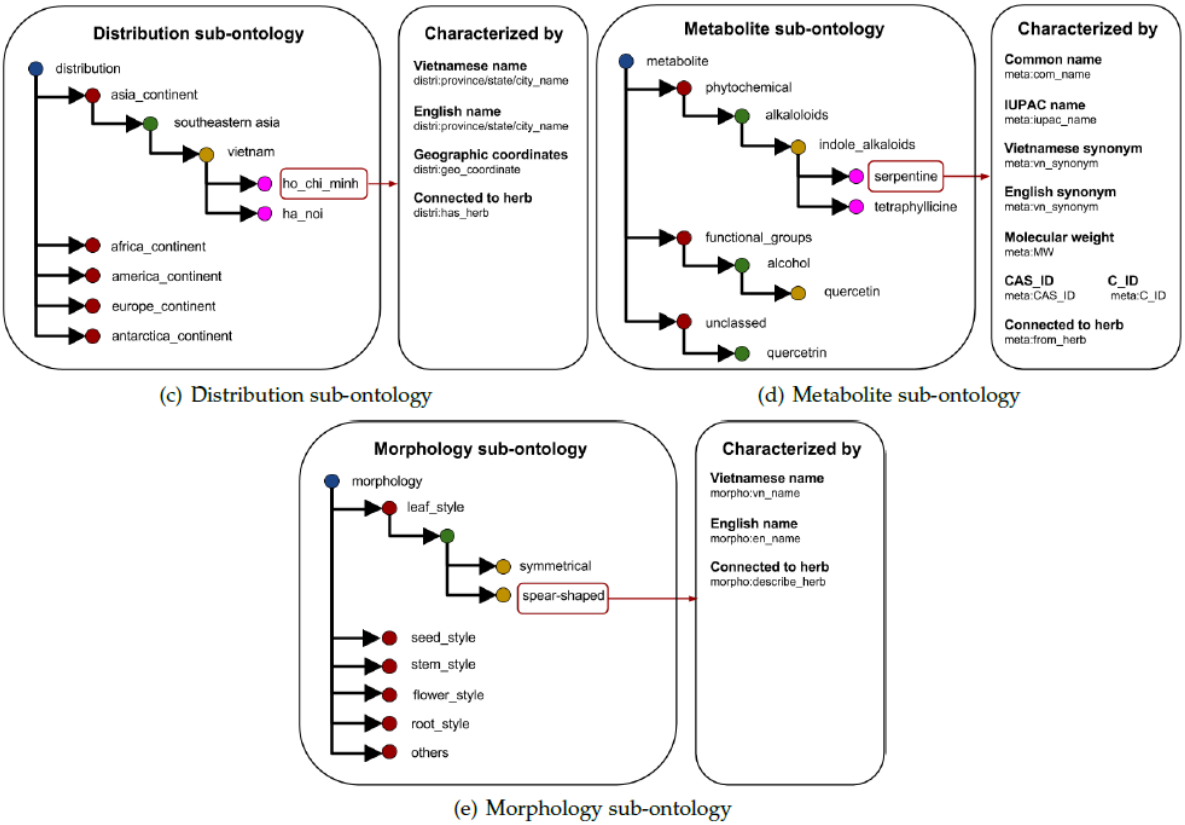


Fig. 1. **VHO structure and characterization**

2.4 Feedback and consensus

VHO allows any users to create their own version of a particular object. To solve conflicts between users, the collaborative algorithm using consensus quality has been employed [20]. The method is described in 4 phases, including Preparatory, Contribution, Consensus improvement and Controlled feedback (Fig 2).

- Phase 1 Preparatory: VHO provides users with criteria for constructing new ontology instead of using questionnaires in Delphi.
- Phase 2 Contribution: a version of the ontology will be created without interfering to the original one when a user modifies the original version. Therefore, each class in VHO can contain multiple versions depending on the users. At the moment, only administrators are allowed to inspect other users' modification.
- Phase 3 Consensus improvement: We measured the distance between a version and other versions, including the original one, as previously described in [20]. Thereby, each version has its own total distance and average. The consensus is reached only if the following inequality holds: dt mean(X) ≥ dmin(X), where dt mean(X) is the total of average distances, and dmin(X) is the minimal total distance among all versions. Else, VHO will require more modifications.
- Phase 4 Controlled feedback: In this phase, the quality of the new consensus is assessed [21] to estimate the improvement. If the consensus creates no significant

changes to the quality, the information will be updated. This consensus is considered as a new original version of a class in VHO. The process then moves back to Phase 3. In case the quality significantly changed, the administrators will determine whether the consensus is appropriate before update the original version. The algorithm is terminating only when the administrators decide that there is no need to improve.

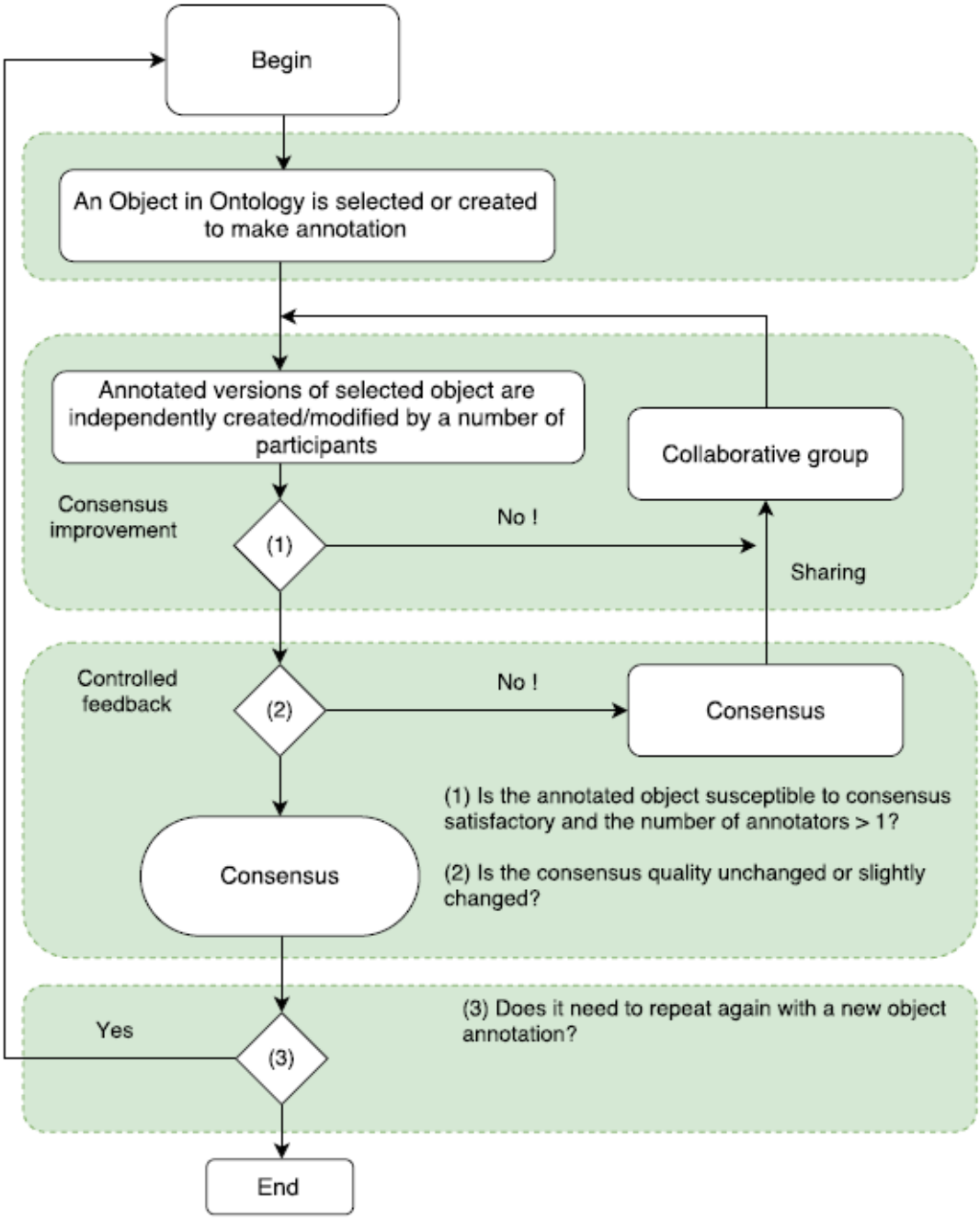


Fig. 2. **Feedback and consensus**

3 . Results:

3.1 Herbal species

VHO comprises of 3,019 species, which were notably recorded in Vietnamese documents as herbs. The number of our entries is undoubtedly incomparable to publicly available DB of herbal species. However, there are 1,698 species, occupying 56.24% of VHO, which are not available in any current plant-metabolite databases. This shows that there are many more species which are unique to Vietnam, but very few is known about their metabolomics. VHO reveals more opportunity for further investigation. Data-statistics shows that nearly 60% of herbal species is likely to be found in the Northeast Asia, e.g. Northeast of China, Taiwan, Korean, and Japan. Besides, smaller portions of 14.4%, 4.6%, 2.8%, and 0.8% of herbal species can be found in the South Asia, Australia- New Zealand, South American, and other regions, respectively. The data-statistics reveals that Cassia and Ipomoea are genus having the most widely distributed herbs compared to the others. High diversity in phenotypes partially explains the strong adaptability to various habitats. With respect to Vietnam, in terms of popularity, Leguminosae is considered as the most common familia in Vietnam, bestrewing all over the nation with 253 species, followed by Compositae, Lamiaceae, Apocynaceae and Rubiaceae with 156, 115, 109, and 97 species respectively.

3.2 Distribution density

To illustrate the distribution density of herbal species in each particular region, geographical heatmaps were plotted in both global and national (Vietnam) scales. Global heatmap of 16 sub-regions was plotted by Rpackage ”rworldmap” [22] while national heatmap of 8 sub-regions was plotted by R-packages ”rgeos” [23], ”maptools” [24] and ”ggplot2” [25]. Vietnam map was used from sources of Global Administrative Areas (GADM) [26].

Global heat map (Figure 3) shows that the South East Asia and North East Asia are the most populated in herbal distribution, followed by the Central Asia, Australia, South America and Central America.

Vietnam heat map (Figure 4) is divided into 8 major regions of Vietnam, including the Northern area (Northeast, Northwest, Red River Delta), the Central area (North Central, South Central, Central Highlands), and Southern area (Mekong River Delta, Southeast). Comparing these regions, Northeast and Central Highlands are considered to have significantly greater numbers of herbal species with 1829 and 1235 species respectively. Meanwhile, South Central is a region of the least number of species. The highest annual aridity can explain for the low biodiversity in this region. Mekong River Delta and Red River Delta are two largest deltas with high biodiversity. However, according to the database, Mekong River Delta is about 2 times less diverse than Red River Delta. Notably, many species were putatively documented as inhabitants of the whole Northern, Southern and Central areas which consist of several regions e.g. Southern area consists of Mekong River Delta, South Central, Southwest and partial Tay Nguyen region. 42.7% of herbs were annotated with at least one ambiguous geographical terms, such as Southern/Northern of Vietnam. If a species carries only those terms, it is referred as putative species and are not visualized in the heatmap.

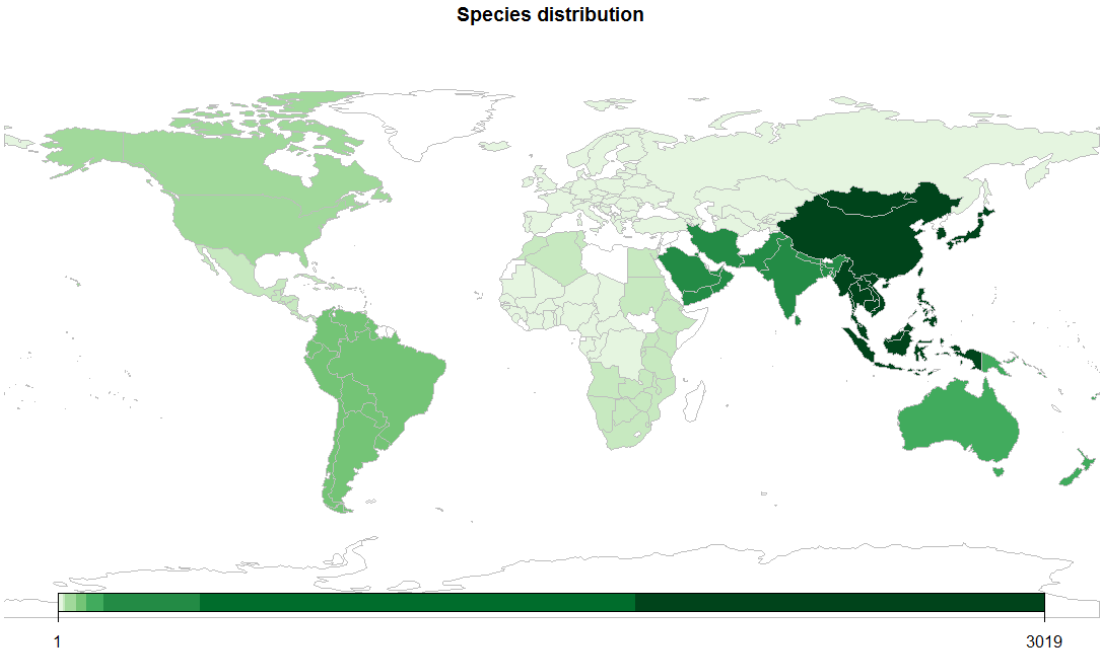


Fig. 3. **Heat map illustrates the distribution of herbal species in VietHerb at world scale.** The coloring scale was generated by mapping the geological information of each species to theWorld Map. Number of species was used as scaling unit.

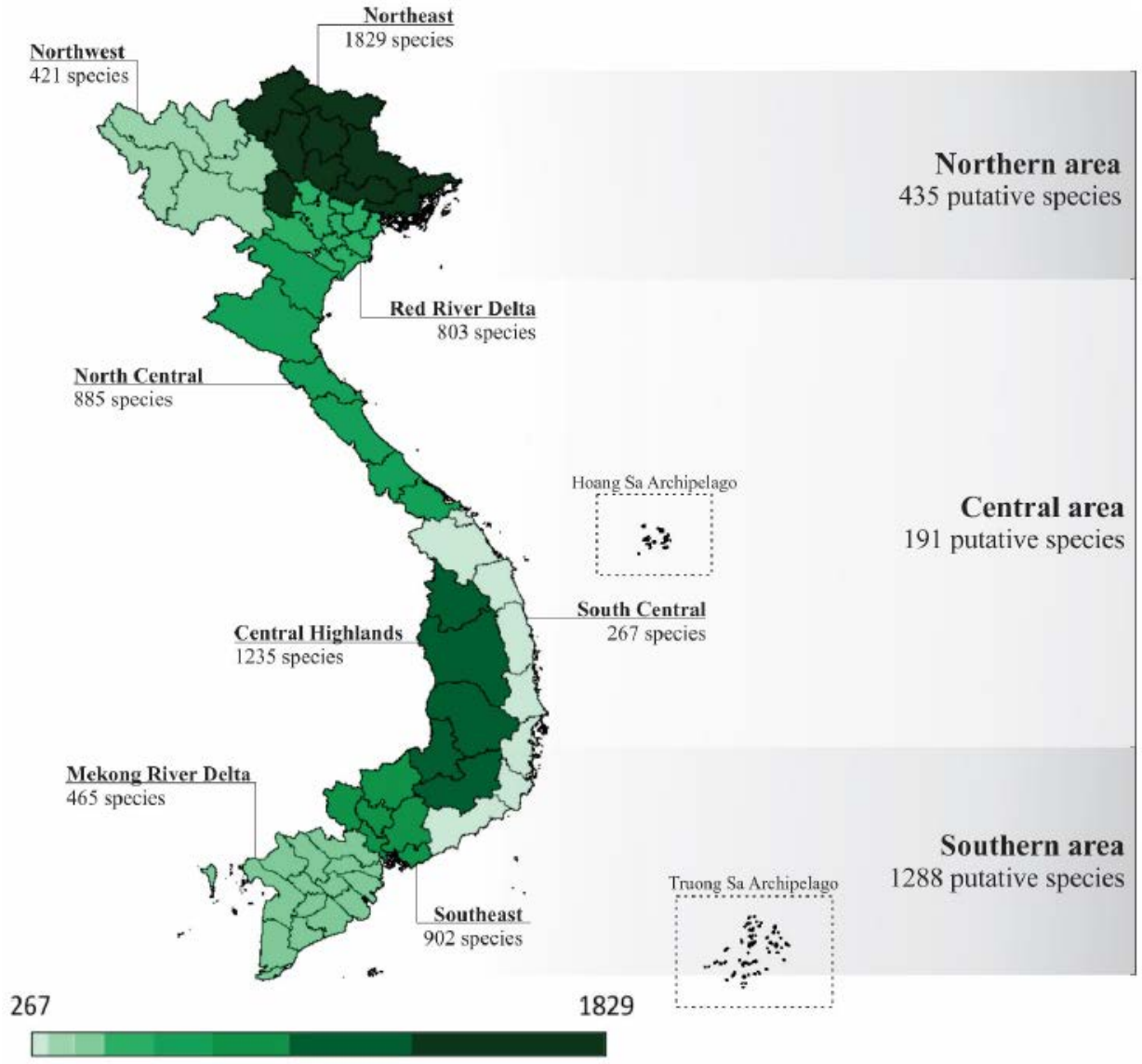


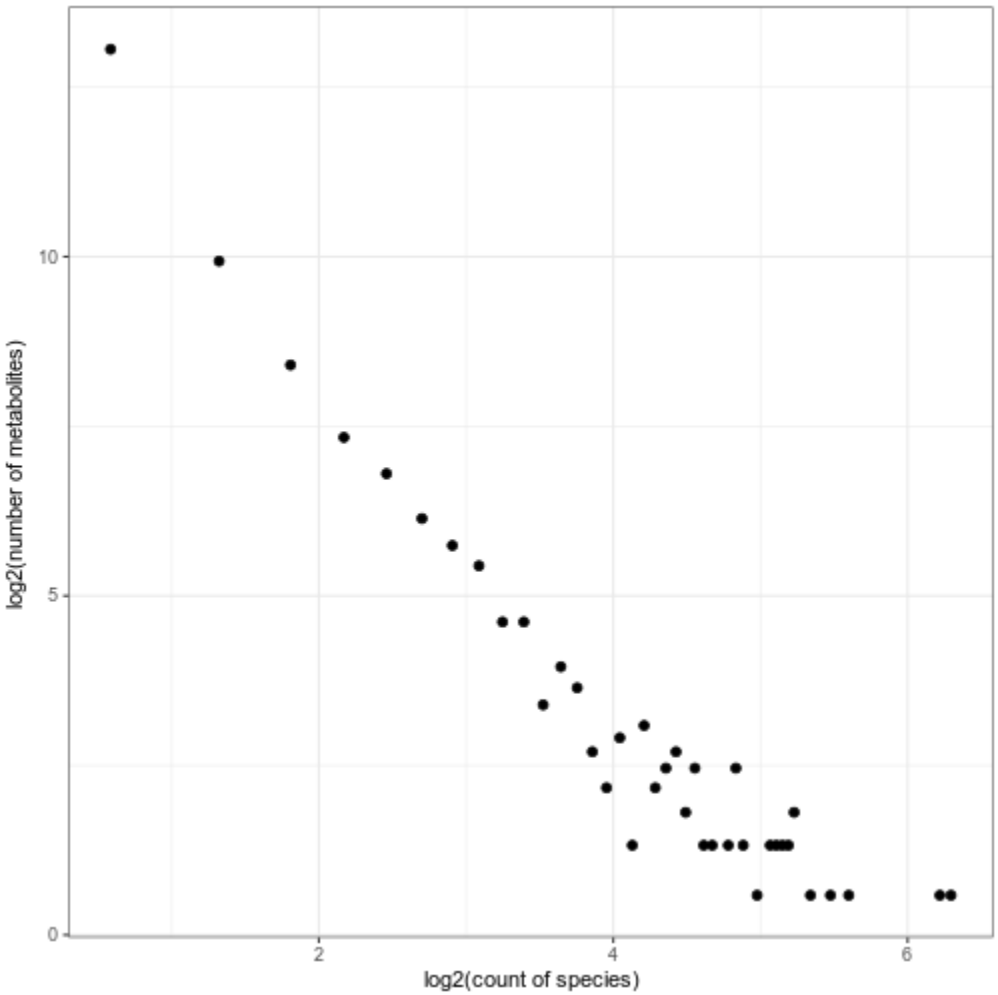
Fig. 4. **Heat map illustrates the distribution of herbal species in VietHerb at national scale.** The calculation focused on 8 core economic regions of Vietnam, including the Northern area (Northeast, Northwest, Red River Delta), the Central area (North Central, South Central, Central Highlands), and Southern area (Mekong River Delta, Southeast). Putative species refers to the herb having unspecific habitats)

3.3 Binary relationship of herbal species and metabolites

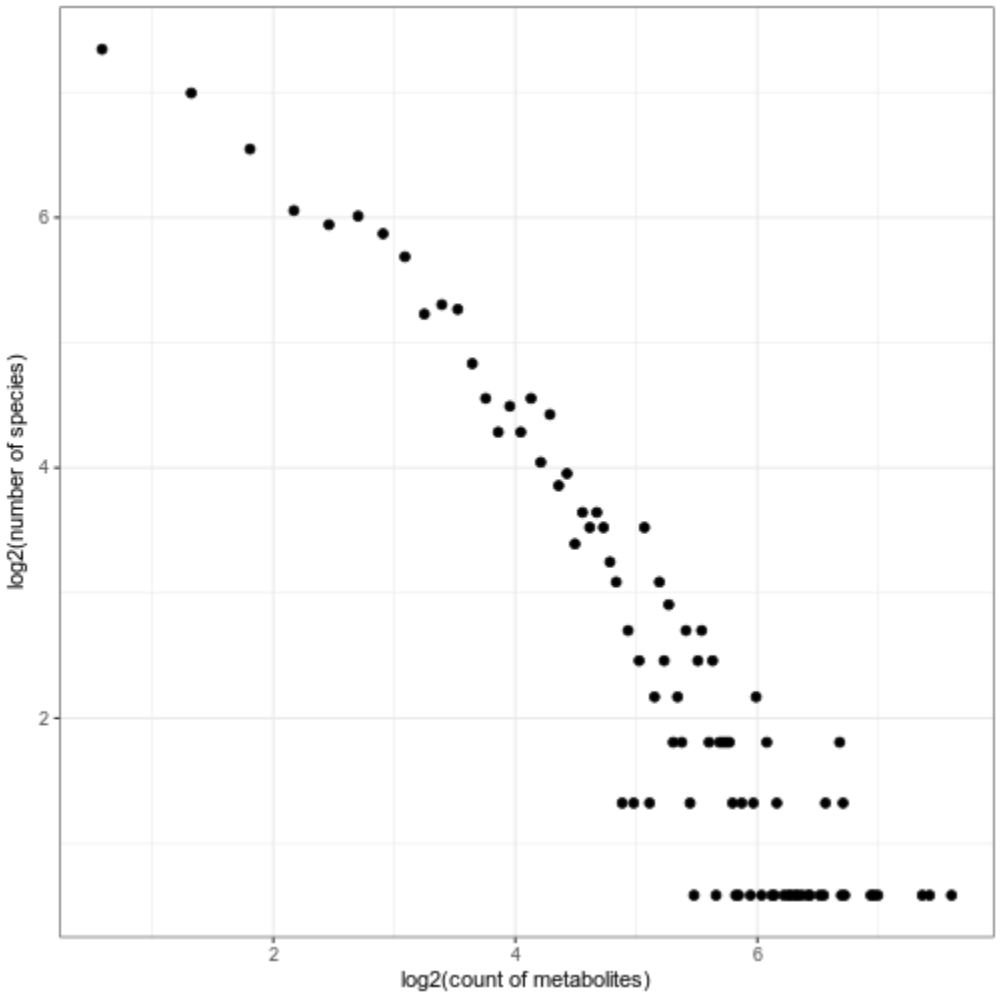
Regression analysis was used to assess the relationship between the number of herbal species and a given metabolite. Their correlation was visualized by scatter plots in log-log scale and with Spearman method (Figure 5) The plot on the left shows the relationship between a given number of metabolites and the corresponding count of herbal species containing the given set of metabolites and the plot on the right shows the relationship between a given number of herbal species and the corresponding count of metabolites contained in the given set of herbal species. The Spearman Correlation coefficient of the left-hand side and the right-hand side are -0.9300872 and -0.9109984 respectively. These

numbers confirm strong negative correlation and fit to power law with the steep slope. The analysis shows there is a very large number of herbal species containing a few metabolites and only a very small number of herbal species containing many metabolites. Although these models are likely to be used to predict the expected values of interest, other factors also need to be examined. This result revealed the fact that an extremely great number of herbal species have not been well studied on pharmacology and phytochemistry.

Principle component analysis (PCA) was based on herbs and their metabolite content. To estimate the largest possible variance, we constructed a binary matrix where each cell reports a presence of a metabolite (row) in a herb (column). Rows and columns having a total lower than 3 were excluded. As there is a majority of herbs having low metabolite counts, these herbs share a prominent amount of 0-rows which contribute to minimize their variance and promote the formation of a single cluster in PCA. Accordingly, the outliers are the herbs that consist of high metabolite counts. As illustrated in Figure 6, most species in Citrus genus belong to these outliers which reflects their popularity in research due to their huge worldwide consumption. Other popular grains such as Zea mays (maize), Pisum sativum (pea), Phaseolus vulgaris (green bean), Oryza sativa (rice), and Raphanus sativus (raddish) were found distinct to both Genus species and the main cluster. Artemisia annua is another discordant herb. Due to extensive research against malaria for decades [27], fully annotated metabolomics makes it unique. Except Citrus, no other genus-specific cluster was observed. In metabolites, gibberellins and several common aromatics were shared by most of the herbs in VHO, thus they are distantly related to the single cluster of low-herb-count metabolites. This result indicates that a large number of herbs in the Vietnamese traditional medicines have vague metabolite profiles.

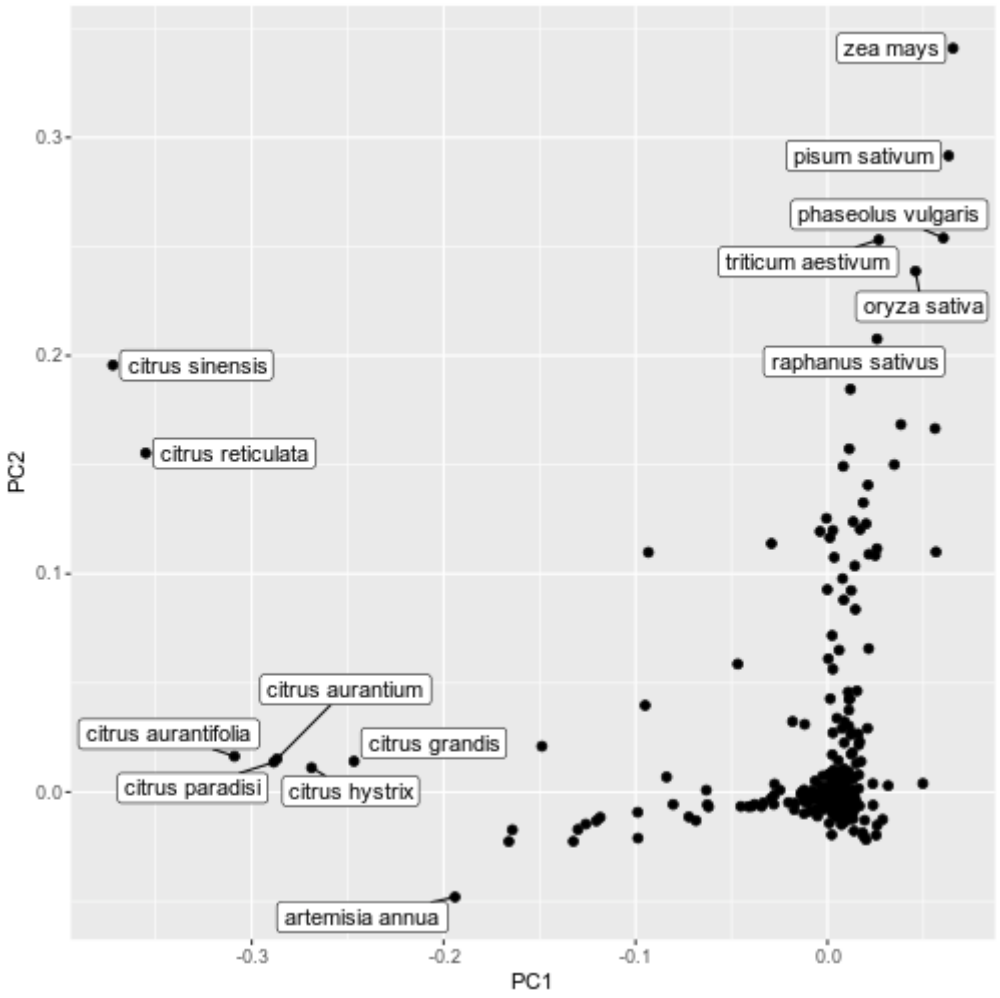


(a) Function y = -0.6392x + 4.5726

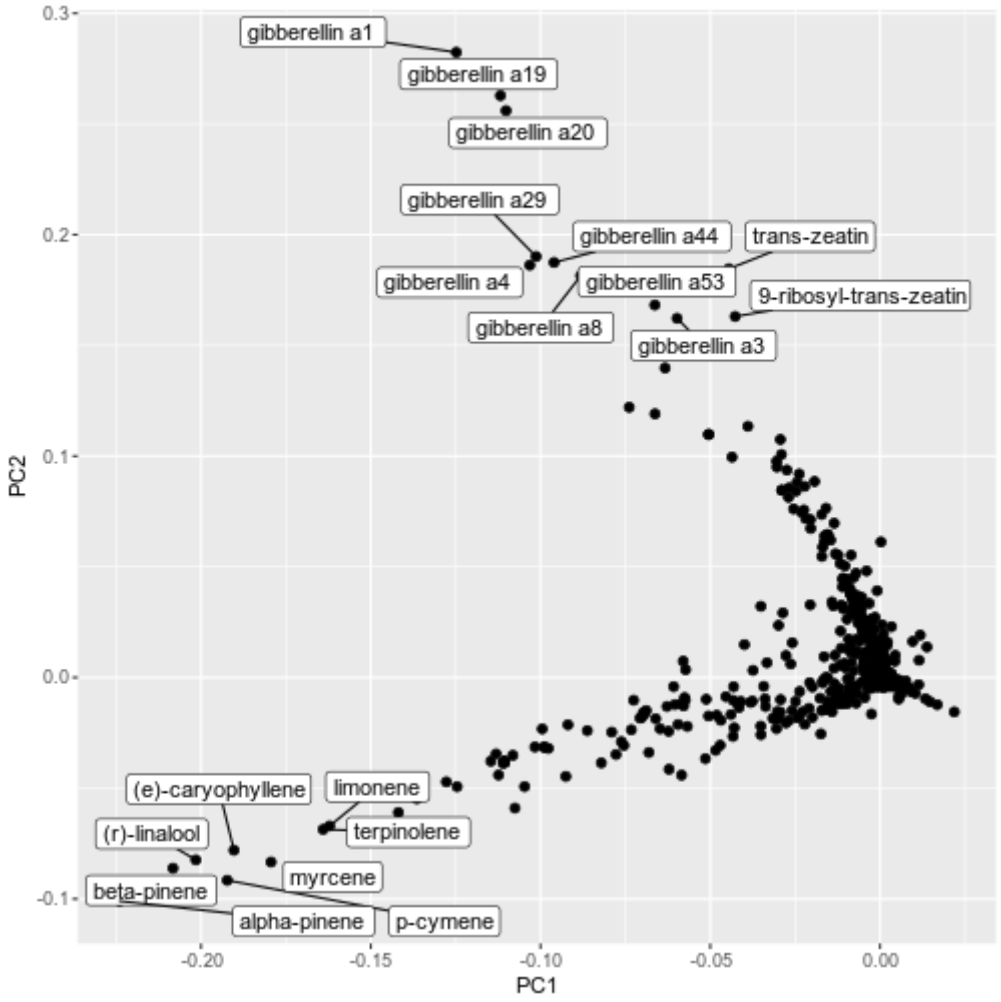


(b) Function y = −0.4896x + 4.0099

Fig. 5. **Degree distributions shows relationship of species and metabolite.** For the left-sided plot, x = ”Number of given set of metabolite” and y = “Count of species” with Spearman Correlation coefficient r1 = -0.93. For the right-sided plot, x = “Number of given set of herbal species” and y = “Count of metabolite” with Spearman Correlation coefficient r2 = -0.91



(a)



(b)

Fig. 6. **PCA of herbs and metabolites.** Outliers are nominally labeled

3. CONCLUSION

We have designed a ontology-based database which systematizes and hierarchically organizes the Vietnamese herbal information, such as traditional medicines, herbal species, therapeutic effects, and geographical distribution. Most of these data are unique and have not been recorded in traditional databases such as TCM [28] or KNApSAcK [13]. As of now, VHO contains 2881 species, 10887 metabolites, 458 geographical locations and 8046 oriental therapeutic effects, and binary relations between them: 17602 species-metabolite linkages, 2718 species-therapeutic effect linkages, 11943 speciesmorphology linkages, and 16089 species-distribution linkages respectively. Preliminary statistical analysis reveals that both the number of species per metabolite and the number of metabolites per species follow a power law. It is interesting to note that their behavior is very similar to that of social systems. The ceaseless globalization has brought traditional medicals and conventional one closer to serve human life in medicare and health science. The need of herb-based drugs whose formulae are standardized, ingredients are clarified and origins are traceable is continually expanding with a potential market segment. This trend, therefore, requires modernization of traditional medicines. To uniformize the quality of herb-based drugs, the assessment in traditional folk medicines must be verified and validated with scientific evidence. Understanding the role of modernization of traditional medicines, VHO provides not only sources of separate herbal instances in form of sub-ontologies but also information of thousand binary connections among these instances which can be used as inputs for searching the others. The multifaceted search tool of our database using semantic analysis

algorithm to efficiently deal with fragmented, incomplete and imprecise input query is currently in progress. This will enable the system to separate the input information into the shortest term in sense, then proceed semantical analysis, requisite analysis, information retrieve and result suggestion in succession. VHO also facilitates building and integrating with similar databases, given that the collection, analysis and category processes require huge amount of effort to achieve high quality, in order to enlarge the database easily. In the light of evidence, we confirm the essential contribution of VHO in phytochemistry, pharmaceutical science, and botany.

In our follow-up research, we proposed a traditional herbal medicine of Vietnamese knowledge discovery based on machine learning approach. We plan to solve many problems of our practical data such as multiclass classification, imbalanced class data,... If the data is retrieved from more diverse knowledge of herbal medicine and relative biopharmaceutical information, we can reach the better model to clarify the ability treatment behind of traditional herbal medicine. In the future, we will collect more data and continue with this study in order to improve the performance. Moreover, we will develop the new method in algorithm and methodology to mining the specific knowledge of herbal medicine.

ADDTIONAL INFORMATION

Web access and service

VHO can be accessed via the web at http://www.vietherb.com.vn. The first released version only assists users with simple textual search. For instance, scientific name of a herb or IUPAC name of a molecule can be used as a query. Semantic search tool with suggestion is planned for immediate near future. The VHO home page and manual are also available in English for international cooperation as well as in Vietnamese for non-English speaking users.

REFERENCES

[1] Bino, R. J. et al. Potential of metabolomics as a functional genomics tool. *Trends in plant science* 9, 418–425 (2004).

[2] Fiehn, O. Metabolomics–the link between genotypes and phenotypes. *Plant molecular biology* 48, 155–171 (2002).

[3] Stitt, M. & Fernie, A. R. From measurements of metabolites to metabolomics: an 'on the fly'perspective illustrated by recent studies of carbon–nitrogen interactions. *Current opinion in biotechnology* 14, 136–144 (2003).

[4] Trethewey, R. N. Metabolite profiling as an aid to metabolic engineering in plants. *Current opinion in plant biology* 7, 196–201 (2004).

[5] Pedro, M. Emerging bioinformatics for the metabolome. *Briefings in bioinformatics* 3, 134–145 (2002).

[6] Sumner, L. W., Mendes, P. & Dixon, R. A. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* 62, 817–836 (2003).

[7] Kopka, J., Fernie, A., Weckwerth, W., Gibon, Y. & Stitt, M. Metabolite profiling in plant biology: platforms and destinations. *Genome biology* 5, 109 (2004).

[8] Saito, K. & Matsuda, F. Metabolomics for functional genomics, systems biology, and biotechnology. *Annual review of plant biology* 61, 463–489 (2010).

[9] Scotland, R. W. & Wortley, A. H. How many species of seed plants are there? *Taxon* 52, 101–104 (2003).

[10] Cay Thuoc. URL http://www.caythuoc.net/.

[11] Tropicos. URL http://www.tropicos.org/.

[12] The Plant List. URL http://www.theplantlist.org/.

[13] Nakamura, Y. et al. Knapsack metabolite activity database for retrieving the relationships between metabolites and biological activities. *Plant and Cell Physiology* 55, e7–e7 (2014).

[14] Degtyarenko, K. et al. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research* 36, D344–D350 (2008).

[15] Do, L. T. *Nhung cay thuoc va vi thuoc Viet Nam (The medicinal plants and herbal formulation in Vietnam)* (Y Hoc & Thoi Dai, 2015).

[16] Van, C. V. Tu dien cay thuoc viet nam (*The dictionary for medicinal plants in Vietnam*)(2013).
[17] Fern´andez-L´opez, M., G´omez-P´erez, A. & Juristo, N. Methontology: from ontological art towards ontological engineering (1997).
[18] Vatant, B. & Wick, M. Geonames ontology (2012).
[19] Baxter, H., Harborne, J. B. & Moss, G. P. *Phytochemical dictionary: a handbook of bioactive compounds from plants* (CRC press, 1998).
[20] Duong, T. H., Tran, M. Q. & Nguyen, T. P. T. Collaborative vietnamese wordnet building using consensus quality. *Vietnam Journal of Computer Science* 1–12 (2016).
[21] Nguyen, N. T. *Advanced methods for inconsistent knowledge management* (Springer Science & Business Media, 2007).
[22] South, A. rworldmap: a new r package for mapping global data. *The R Journal* 3, 35–43 (2011).
[23] Bivand, R. & Rundel, C. rgeos: interface to geometry engineopen source (geos). *R package version* 0.3-2 (2013).
[24] Bivand, R. & Lewin-Koh, N. maptools: Tools for reading and handling spatial objects. *R package version* 0.8–27 (2013).
[25] Ginestet, C. ggplot2: elegant graphics for data analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174, 245–246 (2011).
[26] Areas, G. A. Gadm database of global administrative areas (2012).
[27] Miller, L. H. & Su, X. Artemisinin: discovery from the chinese herbal garden. Cell 146, 855–858 (2011).
[28] Chen, C. Y.-C. Tcm database@ taiwan: the world's largest traditional chinese medicine database for drug screening in silico. *PloS one* 6, e15939 (2011).
[29] Wishart, D. S. et al. HMDB: The human metabolome database. Nucleic Acids Res. 35, (2007).
[30] Ertl, P. Cheminformatics analysis of organic substituents: Identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. in Journal of Chemical Information and Computer Sciences 43, 374–380 (2003).
[31] Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. Conditional variable importance for random forests. BMC Bioinformatics 9, 307 (2008).
[32] He, H. & Garcia, E. A. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284 (2009).
[33] Anand, A., Pugalenthi, G., Fogel, G. B. & Suganthan, P. N. An approach for classification of highly imbalanced data using weighting and undersampling. Amino Acids 39, 1385–1391 (2010).

Results and Discussion: Describe significant theoretical and/or experimental research advances or findings and their significance to the field and what work may be performed in the future as a follow on project. Fellow researchers will be interested to know what impact this research has on your particular field of science.

Note: Final report is for the entire project period, not just for the last one year. Section structure need not be the same as above. **You can structure the report in the same way as you normally write a journal paper.**

List of Publications and Significant Collaborations that resulted from your AOARD supported project: In standard format showing authors, title, journal, issue, pages, and date, for each category list the following:
a) papers published in peer-reviewed journals,

Quang Ong, Phuc Tran, Thao Nguyen, Ly Le. Bioinformatics Approach in Plant Genomic Research. Current Genomics.Volume 7, Number 4, 2016

b) papers published in peer-reviewed conference proceedings,

c) papers published in non-peer-reviewed journals and conference proceedings,
d) conference presentations without papers,
e) manuscripts submitted but not yet published

VietHerb: an ontology for herbal species in Vietnamese traditional medicine. Journal of Chemical Information and Modeling.

f) provide a list any interactions with industry or with Air Force Research Laboratory scientists or significant collaborations that resulted from this work.

Attachments: Publications a), b) and c) listed above if possible.

DD882, SF425, SF298: As a separate document, please complete the invention disclosure form (**DD882)**, Federal financial report (**SF425**) and Project summary report (**SF298**), and **sign** DD882 and SF425. **SF425 must be signed by a personnel in your business office** who is authorized to sign a financial document **or the same person who signed SF424** (Application package). **If the PI signs, we need a separate e-mail from your business office** stating that the PI is authorized to sign SF425**. DD882** can be signed by the PI.

Important Note: **Attached publications are used only for internal use.** They do not go outside of AFRL because of the copyright issues of the published papers. However, **the main text goes to DTIC which can be accessible to public.** Thus, **a final report must be self-contained without reference to other documents. Submission of a report that is very similar to a full length journal article will be sufficient in most cases. The final report should give a fair account of the work performed during the period of performance.** There will be variations depending on the scope of the work. As such, there is no length or formatting constraints for the final report. Keep in mind the amount of funding you received relative to the amount of effort you put into the report. For example, do not submit a $300k report for $50k worth of funding; likewise, do not submit a $50k report for $300k worth of funding. **Include as many charts and figures as required to explain the work.**