AFRL-AFOSR-JP-TR-2019-0045

Knowledge discovery in Vietnamese herbal medicine by use of VietHerb

Ly Le
INTERNATIONAL UNIVERSITY VIETNAM NATIONAL UNIVERSITY-HCM
QUARTER 6, LINH TRUNG, THU DUC DIST.
HO CHI MINH, 700000
VN

07/16/2019
Final Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Asian Office of Aerospace Research and Development
Unit 45002, APO AP 96338-5002

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
16-07-2019	Final	21 Jul 2017 to 20 Jul 2019

4. TITLE AND SUBTITLE
Knowledge discovery in Vietnamese herbal medicine by use of VietHerb

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA2386-17-1-4032

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Ly Le

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
INTERNATIONAL UNIVERSITY VIETNAM NATIONAL UNIVERSITY-HCM
QUARTER 6, LINH TRUNG, THU DUC DIST.
HO CHI MINH, 700000 VN

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AOARD
UNIT 45002
APO AP 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-JP-TR-2019-0045

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
'Knowledge discovery in Vietnamese herbal medicine by use of VietHerb' effort focused on examining various AI techniques in the medical domain. The first work used graph convolution and extended connectivity fingerprinting to predict metabolite efficacy against cancer. The second work applied text mining techniques to discover drug interactions between herbal and traditional medicines. For pragmatics of scale, this required an unsupervised learning model to perform relation extraction. This effort produced 2 journal publications and 3 conference publications. The database is hosted at http://vietherb.com.vn/

15. SUBJECT TERMS
ontologies, knowledge discovery, machine learning, data mining

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			LIN, ALAN
Unclassified	Unclassified	Unclassified	SAR		19b. TELEPHONE NUMBER *(Include area code)* 315-227-7009

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Final Report for AOARD Grant FA2386-17-1-4032

"Knowledge discovery in Vietnamese herbal medicine by use of VietHerb"

July 16, 2019

Name of Principal Investigators (PI and Co-PIs): Ly Le

- e-mail address : ly.le@hcmiu.edu.vn
- Institution : Computational Biology Center, International University, Vietnam National University
- Mailing Address : Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, 700000 Vietnam
- Phone : +84906578836
- Fax : +84837244271

Period of Performance: 07/21/2017 - 07/20/2019

Abstract:
As claimed by a chemo-informatics-related principle, structurally similar chemical compounds will very likely have similar biological activity. A metabolite is a chemical compound. Based on this observation, we challenged two kinds of knowledge discovery from herbal database during the first year: modeling the relationship between metabolites and therapeutic effects, and predicting whether a metabolite fights against a cancer or not, which resulted in two conference papers. Herbal medicine is often utilized in combination with conventional drugs. Herb-drug interactions (HDIs) related adverse drug reactions are becoming widely recognized because of the increasing number of clinical cases. During the second year we challenged another kind of knowledge discovery which is extracting HDI knowledge from literature, which resulted in one conference paper. During the project period, two journal papers on VietHerb were published which is the major output of the previous AOARD project and forms a basis of this project.

In this final report we focus on a graph convolution-based classification model for identifying anticancer metabolites from traditional Vietnamese herbal medicine database and semantic relation extraction for herb-drug interactions from the biomedical literature. Details of other studies as well as the reported studies are described in the papers in List of Publications.

The first work is to predict whether a metabolite fights against a cancer or not. A metabolite is a chemical compound. The challenge is how to encode this to the input of any classifier. Two approaches were attempted. One is to use Extended Connectivity Fingerprint, which extracts all the nested substructures starting from every single atom recursively and hash the resulting lists to a fixed length list of numerals. Any classifier can be used but a neural network was chosen in this study. The other is to use graph convolution, which is a deep neural network specifically designed to handle chemical compound. This selectively learns important features just like a convolutional neural network does. Both approaches give reasonably good results, but we found that the graph convolution approach generalizes better.

The second work is to investigate the impacts of herb-based products on activities of other conventional drugs when combining them in certain medical treatments. For years, patients using herb-based medications have built a misconception about the absolute safety of products derived from natural sources. In this study, text mining techniques are applied to provide users with a novel approach to save time when looking for information of HDIs. Since constructing an annotated corpus for herb-based products in traditional manner

requires a high demand for human resources and financial support, an unsupervised learning model for relation extraction was used as an alternative which eliminates to the crucial role of an annotated training set. The obtained result proposes a promising method for the HDIs extraction challenge.

Introduction:

a. Traditional Vietnamese Herbal Medicine Database

Vietnam is located within the tropical zone, has many good conditions for creature's growth and created a variety of animals, plants and other Eco-living systems [1- 4]. We introduce a DB of herbal metabolomics to community as a liable information source for both experts and non-experts who share common interests in herbal species. We constructed VietHerb Ontology (VHO), an ontology-based DB of four types of binary relationships: herbal species - metabolites, herbal species - geographical distributions, herbal species - Chinese therapeutic effects, and standardized international diseases -Chinese medical diseases. VHO will work as a core in developing VietHerb Families Database (VHFDB) in the future. According to the statistical analysis of VietHerb database: herbal species - metabolites DB, herbal species – geographical DB and herbal species - therapeutic effects DB contain 17,602; 2,695 and 2,916 binary relationships respectively encompassing 3,019 species, 10,888 metabolites, 512 geographical locations, and 4,194 therapeutic effects. It gives us a question about how to use and understand the highly content rich data. Computational methods [5-7] for herbal medicine allow us to identify required information more efficiently [8-9].

b. Modern bioinformatics to explore Traditional Vietnamese Herbal Medicine Database

Traditional Vietnamese Herbal Medicine and Western Medicine are obviously different medical systems, both of which provide support for disease diagnosis, treatment and health prevention. Different from Western herbal medicine, in which herbs are often delivered singly or combined into small formulas of herbs with the same function, Vietnamese herbalists usually prescribe combined drug therapies of multiple herbs to treat a disease. For some fatal diseases, Chinese herbal medicine has slower, but better, effects. These diseases are assumed to have complex causes and multiple targets. Investigating multicomponent herbal drugs can be a future direction of multi-target drug development for effective personalized medicine. Data mining and machine learning promise a solution for dealing with herbal information overloading [10-16]. Moreover, by creating a forecasting model, we will deeply understand these VietHerb data and their application in a higher level of knowledge.

c. Relational extraction from biomedical literature

In the growing of biomedical literature, HDIs are regularly published in journals, advancing medical literature the most practical resource for its detection [17-21]. However, a standard methodology or system to extract HDIs from textual data is still nonexistent. Raised in natural language processing and machine learning technique, Relation Extraction (RE) task can be a great advantage in the healthcare observation system [22]. This method enables recognition and extraction of appropriate information on HDIs and affording an interesting way of diminishing the time spent on examining the literature. Currently, in biomedical relation extraction field, database-based and text mining methods are popularly used [23]. Despite, HDIs information remains insufficient in digital structured references like relational database or ontology [16]. Consequently, a supervised relation extraction method is interrupted by several problems (label, structural data, ...) [24-25]. Unsupervised relation extraction is an alternative strategy to obtain a string of words which contain purposed relational entity [26]. In this approach, large amounts of data could be employed, and lots of relational sentences could be detected [27]. However, the resulting relations will be certainly difficult to map to a standard relation and it requires a particular knowledge base.

For improving the performance, word embedding which is an distributed word representation approach is adopted in the preprocess phase [28]. They have recently been shown to capture both semantic and syntactic information about words very well, setting performance records in several word similarity tasks [29]. The word embedding is used to transform entity words (drug or herb name) or relational trigger words then parse sentences in low dimensional space to increasing identification of semantic clusters among the large number of documents.

Contribution:

1.	Provided a machine learning approach to exploit the relationship between different attributes of Herbal Medicince
2.	Provided a deep learning method, which is Graph Convolution Neural Network for discovering cancer potential bioactivity of Metabolites in Vietnamese Herbal Medicine.
3.	Provided an unsupervised ML approach to retrieve the relationship between herbal medicine and drug by finding their interaction from large scale biomedical literature.

A graph convolution-based classification model for identifying anticancer metabolites from traditional vietnamese herbal medicine database

Material and methods

a. Vietherb dataset

With the support of AOARD (previous project), Vietherb, an ontology-based database for herbal sources in Vietnam, was constructed based on officially published documentations and widely known database as input for herbal-based drug development [a-2]. Vietherb Ontology (VHO) provides a reliable information resource regarding the following binary associations: herbal species – metabolites, herbal species – geographical distributions, herbal species – Chinese medical diseases. Statistically, VHO possesses 458 geographical locations, 8046 Oriental Therapeutic Effects, 10239 metabolites and 2881 species [4]. VHO is also currently the largest ontology-based resource for Vietnamese herbal medicines [4].

b. NCI-60 Developmental Therapeutics Program dataset

The NCI-60 dataset was constructed by the National Cancer Institute Developmental Therapeutics Program (DTP) in 1990 [6], which serves as a resource of more than 50 000 compounds and a variety of biology-related data such as gene expression, protein expression, miRNA expression, genetic variation and DNA copy number. The dataset was obtained by screening thousands of chemical compounds on 60 distinct human tumor cell lines to perform characterization and identification of new compounds involved in tumor growth inhibition [6]. Moreover, the collected data is also used to construct predictive models for drug sensitivity and identify molecular targets [6]. In order to predict the anticancer activity of the plant-derived metabolites, we collected two major sub-datasets from NCI-60, which are DTP One Dose (inactive) and DTP Dose Response (active) groups, respectively. Specifically, DTP One dose dataset indicates no significant growth inhibition of cancer cells and all compounds in this group are tested at only one concentration, 10-5 M. In contrast, compounds which indicate significant growth inhibition of greater than 50% are tested at 5 levels of concentration.

c. Extended Connectivity Fingerprint and Molecular Graph Convolution

In the field of chemo-informatics, ECFP is a representation of chemical structure in the form

of a fixed-length vector of binary values that are widely used for similarity searching, clustering, classification, virtual screening and structure-activity modeling [7]. Specifically, ECFP extracts properties from a chemical structure such as stereochemical information, atomic number, atomic mass, atomic charge and compresses those features into a fixed-length bit string using hash function which is used to feed into classification machine learning models [7]. However, such a fixed-length vector will very likely result in missing features or collision (different features overlap in the same position in the vector). Besides, extending the vector length to cover all features also leads to expensive computation [8].
Because of the expensively computational disadvantage of fixed fingerprint, we apply molecular graph convolution to encode only relevant features and to reduce the number of parameters, which results in more efficient computation and better predictive performance [5]. Specifically, the chemical structure is fed through a series of 3 neural layers which are graph convolution, graph gather and max pooling to perform simultaneously feature extraction and parameter reduction [5]. As a result, one continuous vector integrated with learnable parameters is generated for classification [5].
To facilitate the programming work of both ECFP and graph convolution, we employ DeepChem [9], a deep learning framework specifically designed for chemoinformatics. The framework is integrated with ECFP, graph convolution architecture and Tensorflow-based classification models, which handles most of the prediction procedure.

d. Experimental design

1. Firstly, both DTP One Dose and DTP Dose Response's compounds are filtered only at the dose of 10-5 M. For One Dose dataset, compounds which exhibit less than 5% growth inhibition are obtained as inactive compounds. For the other dataset, compounds that yield more than 50% growth inhibition are collected as active compounds.

2. Input data for the deep learning network are in the form of SMILES notation, which is a series of linear characters representing chemical structure [smiles ref]. Therefore, SMILES strings corresponding to every compound in the active and inactive dataset must be obtained by using BeautifulSoup, a Python-based web scraping package [10].

3. The datasets are filtered one more time to remove compounds of unknown SMILES. Finally, the binary labels are attached to equivalent datasets, which results in the final input data.

4. The input data is fed into ECFP-featured neural network and Molecular Graph Convolutional neural network respectively in order to make performance comparison.

5. The best predictive model from Graph Convolution Network is selected, which is then applied to the Vietherb dataset to predict anticancer activity of herbal plants based on their metabolites as input.

The detailed amount of preparation data is as follows, suggesting that our data is highly imbalanced:

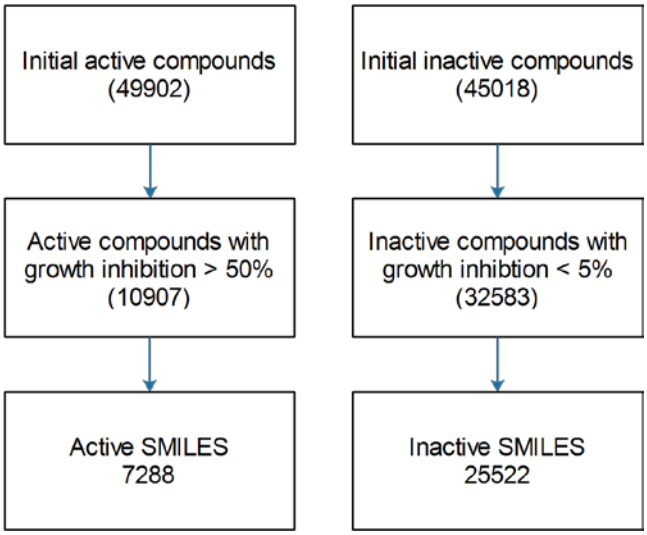

Figure 1: *The establishment of dataset for training and predicting anticancer activity from NCI-60 dataset*

e. Classification technique and featurizer

Softmax classifier: The softmax function, which is also known as normalized exponential function [11], is utilized as classifier in the final layer of graph convolutional network for classification of anticancer activity, coupled with cross-entropy as cost function [12]. Specifically, Softmax function returns probability score for each class label with probability sum of 1, thereby minimizing the output range and facilitating result interpretation.

ECFP featurizer: as the graph convolution featurizer is developed based on ECFP, the illustration of ECFP process is a necessity. In general, the generation of ECFP consists of three steps: Atom identifiers assignment, identifier iterative update and duplicate identifier removal [7].

For the first phase, atom identifiers assignment is implemented by applying Daylight atomic invariants rule [15]. Specifically, every single atom in a molecule is assigned a number which is calculated based on Daylight atomic invariants and is independent of atom numbering [7]. In addition, hydrogen atoms and hydrogen bonds are ignored when assigning identifier. The specific atom information used to calculate the identifier are as follows: number of neighboring atoms (also called heavy atoms) which are not hydrogen atoms, the deduction of valence and the number of hydrogen bonds, atomic charge, atomic number, atomic mass, hydrogen numbers and the atom containing one ring at least [7]. Those atomic information is packed or hased into a value as an atom identifier. The identifiers' value could be either negative or positive depending on the hash function.

Once the atom identifier assignment is finished, the identifier iterative update, which is a phase detecting features of a whole molecule, is performed. Each iteration corresponds to each radius. As the radius or iteration increases, the more features are to be generated, which represent the fragments of a molecule. In this phase, every single identifier is newly updated after each iteration, which exhibits the relationship or local operation between an examined atom and its neighboring atom using the following information: the iteration number and the identifier of the examined atom, the bond orders of neighboring atoms connected to the examined atom in which 1,2,3 and 4 indicate single, double, triple and aromatic, respectively. For example, considering the following part of a molecule to illustrate the iterative update idea:

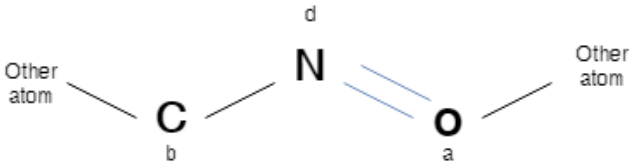

Figure 2: *Illustration of a part of molecule at iteration 0. N indicates Nitrogen atom, which is considered as the examined or core atom, has an identifier of d. C and O indicate Carbon atom with identifier of b and Oxygen atom with identifier of a, respectively. The single line represents single bond while the double line represents double bond.*

To update the identifier d of the nitrogen atom (N), an array is created to perform local operation of N to C and O. Specifically, the first two elements of the array will be the iteration number and the identifier of core atom, for example (1,d). Next, the algorithm of ECFP considers the bond of neighboring atoms with the core atom in the form of (bond order, neighboring atom identifier). For example, the C atom is connected with the N atom through a single bond, thereby generating a pair (1,b) in which 1 indicate single bond. For the O atom's connection with the N, the pair will be (2,a) in which 2 represents double bond. All above information is put into the final array as follows: [1,d,1,b,2,a]. This array is hashed into a new single value which is also a new identifier of the core atom, which represents a larger structure fragment of the molecule (illustrated in Figure 3).
The iterative updating continues until no new features are yielded. After all iterations, the results are a series of identifiers from the initial identifier assignment to the final iteration. Each identifier represents a different substructure of the whole molecule to ensure all molecular features are encoded. The act of updating identifiers leads to a problem which is that an iteration generates the same identifiers from different core atom. This means that the final list of identifiers is not always unique but likely to be redundant. Therefore, the final phase of the ECFP is to remove duplicate identifiers, which results in unique fingerprint for a molecule. In the final step, all unique identifiers of a molecule are hashed into a fixed-length vector of 1024 bits or more to obtain the digital fingerprint of a molecule [7].

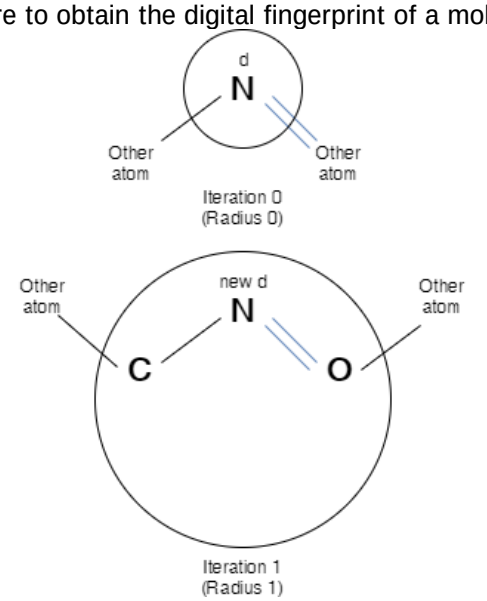

Figure 3*: Illustration of identifier d of N atom at iteration 0 and iteration 1. After each iteration, the substructure representation becomes more informative in terms of chemical structure insight*

Graph Convolution featurizer: there are three major layers serving as feature extractor in graph convolution neural network: Graph Convolution, Graph Pool and Graph Gather [13]. In analogy to ECFP featurizer, Graph Convolution also employs the same decomposition

principles of ECFP, but rather than compressing the identifiers of molecule into binary vectors using fixed hash functions, graph convolution applies adaptive learning to extract only useful features of molecules. Specifically, the graph convolution applies hidden weights in convolutional layers and the output weights in fully connected layers to the identifiers of the atom [5], thereby monitoring the parameters in accordance with supervised learning to achieve better result.
Firstly, the Graph Convolution layers perform convolution output for each node or the core atom in the molecular graph in an approach similar to convolutional network in image recognition [15]. Specifically, the neighboring atoms' feature vector are multiplied by hidden weights and the core atom's feature vector is also multiplied by a distinct weight. Then the summation of both is performed, which is put into the activation function to introduce the non-linearity to the neural network and form a new vector for the core atom [15]. (illustrated in Figures 4a and 4b).

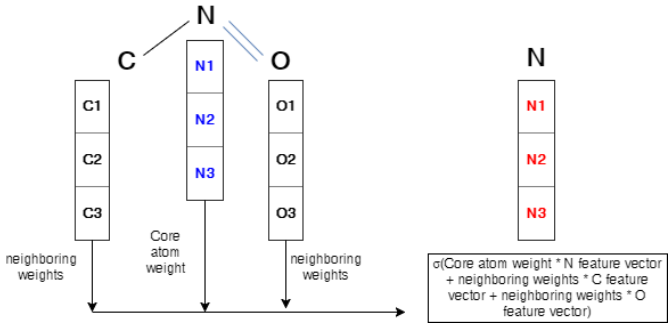

Figure 4a: *Illustration of the operation in graph convolution layer. The array of C1,C2 and C3 indicate the feature vector the C atom. The same notation is used for N and O atom, respectively. The core atom, Nitrogen, is marked as blue. To perform convolution to form the new feature of the core atom associated with 3 neighboring atoms, the summation is conducted and wrapped by an activation function. The new feature vector is marked as red. Similarly, the same process is applied to all the node in a molecular structure, which results in entirely new feature vectors for each node.*

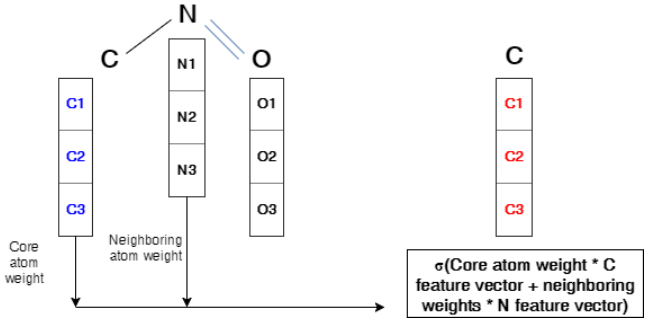

Figure 4b: *Illustration of graph convolution layer on a core atom of 2 neighboring atoms.*

Secondly, the Graph Pool performs max pool operator on each node in the molecule, which returns the maximal value chosen from neighboring atoms and the core atom [13]. In particular, each node's feature vector is updated once again to form a new one.
Finally, Graph Gather layer sums all atoms' feature vectors in Graph Layer into a continuous vector, which represents the whole molecule information [13]. The major purpose of graph convolution is to compress all useful information into a continuous vector while the parameters are reduced when undergoing fully connected layer, thereby lessening computational workload but still gaining competitive result.

Results and Discussion:

a. Result

For classification problem and highly imbalanced data, we use two widely popular metrics, which are AUC-ROC (Area Under Receiving Operating Charateristic Curve) and AUC-PRC (Area Under Precision Recall Curve), to evaluate the result.

Table 1. Classification performance based on AU-ROC

Featurizer		AU-ROC	AU-PRC
ECFP	Training set	0.9752	0.9254
	Validation set	0.9159	0.8532
Graph Convolution	Training set	0.9589	0.8616
	Validation set	0.9390	0.8309

As can be seen, both ECFP and Graph Convolution return very high AU-ROC and AU-PRC scores in training and validation set, suggesting that the neural network performed decent generalization. The ECFP returns the highest AU-ROC and AU-PRC scores in the training set but tends to be more overfitted than Graph Convolution's as the gap between ECFP's training set and validation set is 0.0593. As for Graph Convolution, the training set and validation set are close to each other with the difference of 0.011. We, thus, choose Graph Convolution model as a reliable model to screening through the Vietherb database to uncover potential anticancer candidates.

b. Prediction

Due to Vietherb database being still at the preliminary stage of development, only 3233 out of 10239 SMILES (one SMILES corresponds to one metabolite) are available for prediction. These 3233 SMILES are transformed into real-value vectors before being fed into the built graph convolution network. As a result, 1890 out of 3233 metabolites in the Vietherb database are predicted as potential anticancer candidates.

Semantic Relation Extraction for Herb-Drug Interactions from the Biomedical Literature Using an Unsupervised Learning Approach

Materials and methods

Due to the lack of annotated herbal corpus, we have to use data from several sources to build the suitable one. Herbal medicine and drug profiles are necessary to be used as entities which will be exploited in the relation extraction task while the biomedical literatures would be the resources.

a. Biomedical Literature

We extracted the biomedical data needed to build the word embedding model from PubMed Central (PMC) and PubMed [31]. PMC is one of the most popular source for free full-text articles and there are 1,884,684 full-text articles available as of March 2018 were downloaded. In addition, we also collected the abstracts of 28,337,042 articles which available on PubMed.

b. Herbal Medicine

We dowloaded and extracted 10,061 herbal medicine profiles that available at TCM Database@Taiwan and VietHerb with 48,584 synonyms from the Global Biodiversity Information Facility (GBIF).

TCM Database@Taiwan (http://tcm.cmu.edu.tw/) is currently the largest non-commercial TCM database available for download and based on information collected from Chinese medical texts and scientific publications. Information about herbal medicine was retrieved from this database [32].

GBIF: An international network and research infrastructure which was funded by the world's governments and aimed at providing digitized biological data from different sources (e.g. museum collections, survey programs, etc.) as a result of collaborative endeavours between data providers and taxonomists across many institutions. There are over 686 datasets and 85,899 published occurrences from Viet Nam. Information about synonyms of species was retrieved from this database [33].

VietHerb (http://vietherb.com.vn/) is a relational database that was built on the data of Vietnamese traditional medicine with reliable scientific reference [34]. The database provides a commercial value for herb vendors, which will not only promote the usage of traditional medicine, but also create a seamless network between research and application. As of now, VietHerb 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 species morphology linkages, and 16089 species-distribution linkages respectively. Information about Vietnamese traditional medicine was retrieved from this database.

c. Drug

The DrugBank database is a freely available resource providing data about drugs and drug targets [35]. We downloaded and extracted name of 10,562 drugs which are associated with 16,561 synonyms and published on DrugBank (version 5.1.0, released 2018-04-02).

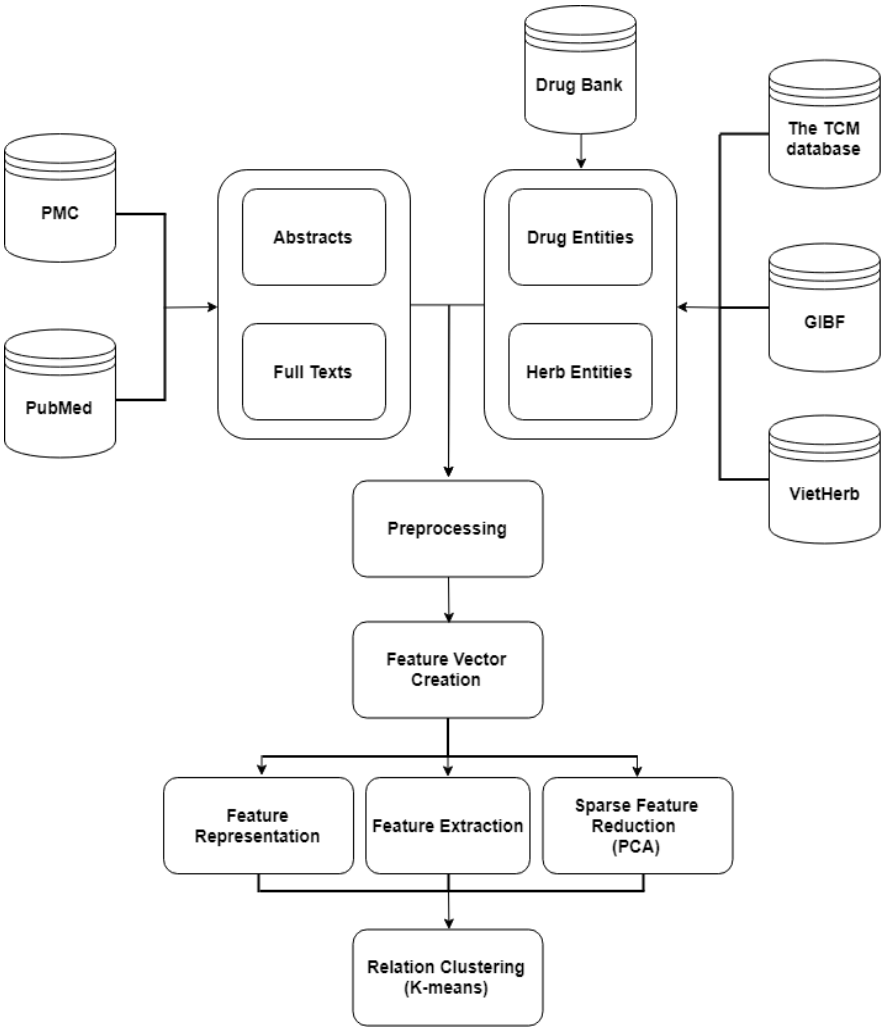

Figure 1. Over view of the system architecture

WORKFLOW

a. Data Preprocessing

The plain text which is encoded in UTF-8 was initially preprocessed for extracting textual parts from the articles and removing all special characters, used as delimiters or formatters. We also applied several different preprocessing techniques for text mining. The process contains these following steps:

Topic related identification: For increasing the specificity as well as reducing the computing resources, only documents which are related to relevant topics such as “herb”, ”traditional medicine” or “plant” were used as the input corpora. We propose to apply rule-based method in this part by using a list of trigger words for the detection. Trigger words list: [“medicine”, “combination”, “concurrent”, “herb”, “traditional”, “adverse”, “side effect”, “plant”, “natural”, “interaction”, “tcm”].

Nomarlization: The sentences were normalized by simply transforming them to lower case. Casing might be valuable in tasks such as sentiment or emotion analysis whereas it does not play a role in relation extraction. Thus, we are interested in the information and the structure of the text and not on the emphasis of the different words in the sentence.

Tokenization: Tokenization is the process of splitting up the input into tokens which

constitute the units for language processing. In our case, splitting up the paragraph into sentences then sentence into words.

Stop words / digits removal: Stop words are words that occur very often, such as" to", "for", "is" or "are". Since they are high-frequency word, so they might not add much information regarding the relation between the entities in the sentence. For the same reason, digits and short token which had no more than 2 characters were also removed.

Entity Replacement: We considered two kinds of entities in this research: herb and drug. For each sentence in the dataset, we replaced all the entities name and their synonyms by a specific format for each type of entity. The idea is that we want to increase the specificity by transforming all the synonyms into one name.

For example:

- In the literature, St John's wort decreases the INR through induction of cytochrome P450 (CYP)-mediated metabolism of warfarin and increases warfarin clearance.

- In the literature, ***herb_1*** decreases the INR through induction of cytochrome P450 (CYP)-mediated metabolism of ***drug_1*** and increases ***drug_1*** clearance.

Part-of-Speech (PoS) Tagging: Part-of-speech tagging is the process of assigning a description of eight parts-of-speech marker (including noun, verb, pronoun, preposition, adverb, conjunction, participle, and article) to each token in the corpus [36]. These tags are used to extract the domain entities and describe the syntactical role of word in the sentence. In our case, we focus only on noun or verb and remove others because the textual relation bewteen entities pairs alsmost is in verb form.

Stemming: The idea of the stemming process is transforming a given word into its root form. As a result, it reduces the size of the word space, and let us focus on the actual meaning of the word.

b. Feature Vector Creation

Word Embeddings Construction

We constructed the feature vector for words by using word2vec [37]. Word2vec is a semantic learning framework employing neural network implementation that learns distributed representations for words by deep learning [30][38]. A corpus is used as an input to estimate continuous vector representations of words as well as produces a vector space. Each vector represents for a unique word in dataset and that word's relationship to others. In terms of hyper-parameter settings, we set the vector dimension, window size, and the mininum word count equal 300, 5, and 5 respectively.

Feature Extraction

For every entity pair, we constructed a Feature Vector from a sentence that contains the entity pair to produce a vector presentation of the textual realtionship between them. Since the expression of the relation between two entities does not depend on every word in the sentence, we prefer giving a higher weight for the lexicalized dependency path between the two named entities. Therefore, we applied a novel method to re-weight the word embeddings which was trained from a previous step. The vector representation s (W, D) of each sentence is defined:

$$s(W,D) = \sum f(w_i, W, D) \cdot v(w_i), \qquad f(w_i, W, D) = \begin{cases} \frac{C_{in} \cdot |W|}{|D|} \\ C_{out}, \end{cases}$$

where W = {w1, ..., wn} is the set of terms in the sentence, D ⊂ W is the set of terms in the lexicalized dependency path between the named entities in the sentence, and v(wi) is the pre-trained word embedding vector for wi, Cin ≥ 1 and Cout are constant values experimentally set to 1.85 and 0.02 respectively. Moreover, length of lexicalized dependency path will also be considered in the calculation since shorter path represents true relationships between pair entities [39].

Sparse Feature Reduction

In unsupervised relation extraction, the lack of training data and feature selection step could be an issue when the concatenation of sparse features can skew the clustering. We suppose that using feature reduction method like Principal Component Analysis (PCA) can circumvent these sparse features and reduce bias [40].

Relation Clustering

As mentioned previously, an unsupervised approach was applied to group all of the relations that have the same type in the corpus. Since each pair of entities appearing in the same sentence contains a potential relation, we applied K-means which is a partitioning method (non-hierarchical method) to cluster all the potential relations identified. The basic idea of this method is the entities pairs which have the same type of relation should belong to the same cluster, while the different clusters should represent other kind of relations. Euclidean distance is used as distance measure.

In terms of choosing number of clusters, we aim to a balance between size of cluster and number of clusters. The smallest K should be chosen to avoid the fragmentation, but the one-third of the clusters also must be significantly large in size simply because we assume that significantly large clusters have higher chance to indicate some semantic relations [41].

Results and Discussion:

a. Result

Quantitative assessment

Firstly, we assess the results base on internal measures. We considered silhouette due to two relevant aspects in appraising a clustering: how similar an element is to its own cluster, and how dissimilar elements which belong to two different clusters are [42][43]. This mean we can judge that how consolidated a cluster is and how distinct it is from another cluster. The silhouette value is defined as:

$$s(i) = \frac{b(i) - w(i)}{max\{b(i), w(i)\}}$$

With

$$b(i) = min\{B(i,k)\}$$

where w(i) is the average distance from the ith point to the other points in its own cluster, and B(i, k) is the average distance from the ith point to points in another cluster k. This value ranges from -1 to +1. The higher score indicates that this point is very distant from other clusters while the negative value mean that point is highly assigned to the wrong cluster.

Moreover, we can average the silhouttes of all points to assess the global measures. The average silhoutte width of a cluster (ASW) and for the entire data set can be determined by the mean value of s(i) for all i in a given cluster and the mean of all the individual silhouettes - $\bar{s}(k)$ respectively:

$$\bar{s}(k) = \frac{\sum_{i=1}^{m} s(i)}{m} \quad with\ m\ is\ the\ number\ of\ objects\ in\ the\ data\ set$$

The negative value of ASW indicates that the distance from the closest point from the other cluster is smaller than the width of this cluster. This can be thought of as an appearance of an oversize cluster in the data set.

In Fig. 2 we report the average Sihouette for the value of K that lie range 2 – 50. All the cases give a positive average Sihoutte value. With K from 2 to 10, the values arrive at 0.1.

In Fig. 3 we report the Sihouette value for all the clusters with K = 5. Obviously, the cluster 4 has a very large silhouette which indicates that it quite separated from the others. The Sihoutte value of the remainings vary in [-0.1, 0.2].

In Fig. 4 we report the size of each cluster with K = 5. The cluster sizes are fairly homogeneous with small K. This balance is needed to avoid the fragmentation.

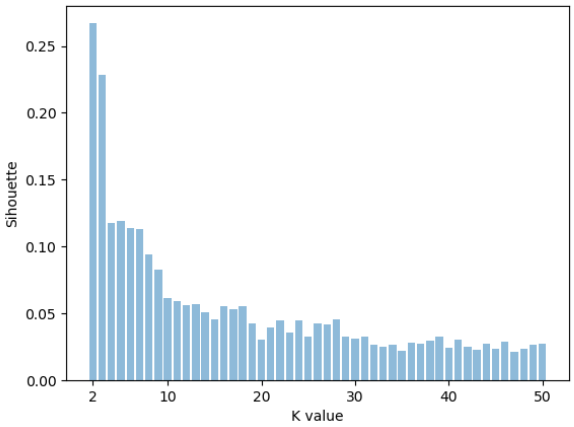

Figure 2. Average Sihouette value for different value of K

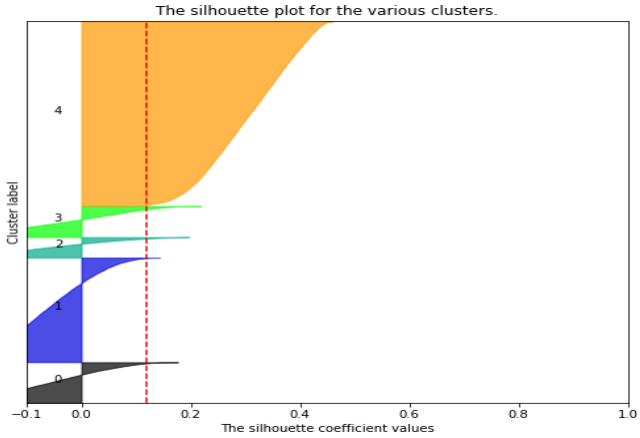

Figure 3. Sihoutte value for each cluster with K = 5

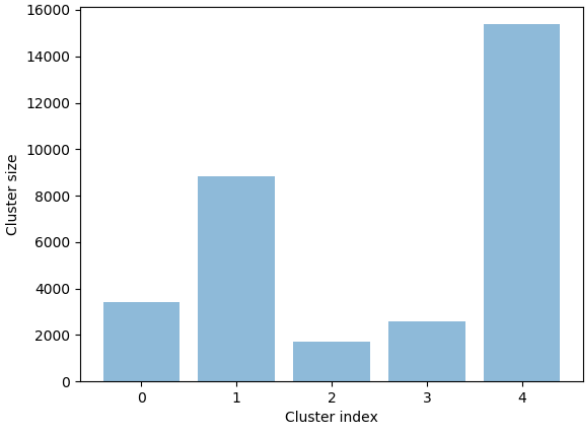

Figure 4. Size of the clusters with K = 5

Qualitative analysis

In terms of the semantic interpretation of the cluster, we try to identify the features which can be represented for the cluster to characterize the type of the relation associated with the cluster. We propose utilizing the cosine similarity between the cluster centroid and the features since they might be well suited for exploiting of the semantic meaning. In general, we consider the combination of the top four features in cosine similarity score to label each cluster. For instance, Cluster 4 is characterized by the four nearest features to the centroid:" inhibit ", "increase ", "produce ", "use ". The suitable label for the relation type of the cluster could be considered as the union of all these features. These features are the trigger words being extracted from the lexicalized dependency path between the entity pairs.

For example:

- Cluster 4: ***Ethanol*** also ***increases*** the toxicity of ***kava*** markedly

- Cluster 4: ***Boesenbergia rotunda*** gave the highest pde ***inhibition*** activity of

4.36-fold sildenafil

- Cluster 4: ***Prismatomeris tetrandra***, was identified as having the potential to develop ***inhibitors*** of ***hyaluronidase***

TABLE 1. The cluster discriminative features.

#Cluster	Discriminative features	Latent label
0	“grown”,” root”	Plant growth regulators
1	“express”, “combine”,”protein”,”cell”,”activity”	#unknown
2	“patient”,”hospital”,”trial”	#unknown
3	“extract”	Plant extraction
4	“produce”, “inhibit”, “increase “, “use “	Bioavailability of drug

Evaluation

To evaluate this system, we use 100 pairs of herb drug which are collected from articles and case reports. These pairs of entities are considered as a standard reference for potential HDIs which relate to adverse reaction (70 pairs) and non-relate to adverse reaction (30 pairs). The results from the qualitative analysis indicate that cluster #4 is the most suitable cluster to represent for potential adverse drug reaction inducing by herb-drug interaction. This mean other clusters are considered as non-related. From these parameters, the system archived 54.45 % of precision with 75.71 % of recall and 0.63 of F-score.

Conclusion

We show different kinds of knowledge discovery are possible by use of state-of-the-art machine learning approaches. All the methods developed in this project are implemented and embedded in the VietHerb database website. Internal attributes in VietHerb are confirmed to provide rich knowledge about herbal medicine. Therapeutic effects or Bioactivities of Herbal Medicine were exploited by the information of Metabolites such as chemical structure or metadata. VietHerb was combined with data in Western medicine such as FDA approved drugs to find the interaction and bridge between Western medicine and

Vietnamese Herbal medicine. There are many more opportunities that VietHerb can contribute to deepen our understanding of traditional herbal medicine. Various application tools embedded to VietHerb will continue to grow.

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:

Five papers [a-1, a-2, b-1, b-2, b-3] are listed. [b-1, b-2, b-3] are the papers that describe the main results of this report. [a-2] is the paper about VietHerb (main result of the previous AOARD project) from which the discovery work is conducted. [a-1] is a related paper authored by the same team. All papers are provided as a separate package to this report.

a) papers published in peer-reviewed journals,

1. Huyen-Trang Vu, Phuong Huynh, Hoang-Dung Tran, Ly Le.In Silico Study on Molecular Sequences for Identification of Paphiopedilum Species. Evolutional Bioinformatics. Volume 14: 1–9. https://doi.org/10.1177/117698774542
2. Nguyen-Vo, T. H., Le, T., Pham, D., Nguyen, T., Le, P., Nguyen, A., ... & Trinh, K. (2018). VIETHERB: A database for vietnamese herbal species. Journal of chemical information and modeling, 59(1), 1-9.

b) papers published in peer-reviewed conference proceedings:

1. Duy, P. T., Thanh, N. M., Vu, N. A., & Le, L. (2017, December). A machine learning approach for drug discovery from herbal medicine: Metabolite profiles to Therapeutic effects. In Proceedings of the 8th International Conference on Computational Systems-Biology and Bioinformatics (pp. 28-33). ACM.

2. Vu, N. A., Duy, P. T., & Ly, L. T. (2018, February). A graph convolution-based classification model for identifying anticancer metabolites from traditional vietnamese herbal medicine database. In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing (pp. 122-126). ACM.

3. Trinh, K., Pham, D., & Le, L. (2018, October). Semantic Relation Extraction for Herb-Drug Interactions from the Biomedical Literature Using an Unsupervised Learning Approach. In 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 334-337). IEEE.

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

Vietherb website: http://vietherb.com.vn/

Attachments: Publications 5 files.

REFERENCES

[1] Firenzuoli, F. & Gori, L. Herbal medicine today: Clinical and research issues. in Evidence-based Complementary and Alternative Medicine 4, 37–40 (2007).Ding, W. and Marchionini, G. 1997. A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[2] Nirmal, S. A., Pal, S. C., Mandal, S. C. & Mandal, S. C. Pharmacovigilance of herbal

medicines. Pharma Times 46, 19–22 (2014).Tavel, P. 2007. Modeling and Simulation Design. AK Peters Ltd., Natick, MA.
[3] Scotland, R. W. & Wortley, A. H. How many species of seed plants are there? Taxon 52, 101–104 (2003).
[4] Ly, L. VietHerb database. (2017). Available at: http://vietherb.com.vn/
[5] Lukman, S., Y. He, and S.C. Hui, Computational methods for Traditional Chinese Medicine: a survey. Comput Methods Programs Biomed, 2007. 88(3): p. 283-94.
[6] Shoemaker R.H.. The NCI60 human tumour cell line anticancer drug screen, Nat Rev. Cancer , 2006, vol. 6 (pg. 813-823)
[7] Sun, Y., et al., Towards a bioinformatics analysis of anti-Alzheimer's herbal medicines from a target network perspective. Brief Bioinform, 2013. 14(3): p. 327-43.
[8] Rogers, D. and Hahn, M., 2010. Extended-connectivity fingerprints. Journal of chemical information and modeling, 50(5), pp.742-754.
[9] Wang, X., et al., Pattern recognition approaches and computational systems tools for ultra performance liquid chromatography-mass spectrometry-based comprehensive metabolomic profiling and pathways analysis of biological data sets. Anal Chem, 2012. 84(1): p. 428-39.
[10] Richardson, L., 2007. Beautiful soup documentation.
[11] Bishop, C.M., 2006. Pattern recognition and machine learning. springer.
[12] Barbakh, W.A., Wu, Y. and Fyfe, C., 2009. Non-standard parameter adaptation for exploratory data analysis (Vol. 249). Berlin: Springer.
[13] Altae-Tran, H., Ramsundar, B., Pappu, A.S. and Pande, V., 2017. Low Data Drug Discovery with One-Shot Learning. ACS central science, 3(4), pp.283-293.
[14] Convolutional Neural Networks. http://cs231n.github.io/convolutional-networks/, Accessed: 2016-11-06
[15] Weininger, D., Weininger, A. and Weininger, J.L., 1989. SMILES. 2. Algorithm for generation of unique SMILES notation. Journal of Chemical Information and Computer Sciences, 29(2), pp.97-101.
[16] P. Posadzki, L. Watson, and E. Ernst, “Herb-drug interactions: An overview of systematic reviews,” Br. J. Clin. Pharmacol., vol. 75, no. 3, pp. 603–618, 2013.
[17] C. Tarirai, A. M. Viljoen, and J. H. Hamman, “Herb–drug pharmacokinetic interactions reviewed,” Expert Opin. Drug Metab. Toxicol., vol. 6, no. 12, pp. 1515–1538, 2010.
[18] B. J. Gurley, E. K. Fifer, and Z. Gardner, “Pharmacokinetic herb-drug interactions (Part 2): Drug interactions involving popular botanical dietary supplements and their clinical relevance,” Planta Med., vol. 78, no. 13, pp. 1490–1514, 2012.
[19] A. A. Izzo, S. Hoon-Kim, R. Radhakrishnan, and E. M. Williamson, “A Critical Approach to Evaluating Clinical Efficacy, Adverse Events and Drug Interactions of Herbal Remedies.,” Phytother. Res., 2016.
[20] X. Li et al., “Inhibitory effects of herbal constituents on P-glycoprotein in vitro and in vivo: Herb-drug interactions mediated via P-gp,” Toxicol. Appl. Pharmacol., 2014.
[21] M. Z. Liu et al., “Pharmacogenomics and herb-drug interactions: Merge of future and tradition,” Evidence-based Complementary and Alternative Medicine. 2015.
[22] C. C. Aggarwal and C. Zhai, Mining text data. 2013.
[23] O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, and B. R. South, “Evaluating the state of the art in coreference resolution for electronic medical records.,” J. Am. Med. Inform. Assoc., 2012.
[24] P. Posadzki, L. Watson, and E. Ernst, “Herb-drug interactions: An overview of systematic reviews,” Br. J. Clin. Pharmacol., 2013.
[25] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning, “Multi-instance Multi-label Learning for Relation Extraction,” Proc. 2012 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn. EMNLP ’12, 2012.
[26] S. Takamatsu, I. Sato, and H. Nakagawa, “Reducing Wrong Labels in Distant Supervision for Relation Extraction,” Jeju, Repub. Korea, 2012.
[27] D. Zeng, K. Liu, Y. Chen, and J. Zhao, “Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks,” in Proceedings of the 2015 Conference on

Empirical Methods in Natural Language Processing, 2015.
[28] C. Quan, M. Wang, and F. Ren, “An unsupervised text mining method for relation extraction from biomedical literature,” PLoS One, 2014.
[29] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, “Relation Classification via Convolutional Deep Neural Network,” Coling, 2014.
[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and hrases and their compositionality,” in NIPS, 2013.
[31] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas, “Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses,” FASEB J., 2007.
[32] C. Y. C. Chen, “TCM Database@Taiwan: The world’s largest traditional Chinese medicine database for drug screening In Silico,” PLoS One, 2011.
[33] T. Robertson et al., “The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet,” PLoS One, 2014.
[34] L. Nguyen-Vo, Hoang; Le, Tri; Pham, Duy; Nguyen, Tri; Le, Phuc; Nguyen, An; Nguyen, Thanh; Nguyen, Thien-Ngan; Nguyen, Vu; Do, Hai; Trinh, Khang; Duong, Hai; Le, “VIETHERB: A Database for Vietnamese Herbal Species,” 2016. [Online]. Available: http://vietherb.com.vn.
[35] V. Law et al., “DrugBank 4.0: Shedding new light on drug metabolism,” Nucleic Acids Res., 2014.
[36] D. Jurafsky and J. H. Martin, “Part-of-speech tagging,” Speech Lang. Process., 2016.
[37] T. Mikolov, G. Corrado, K. Chen, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Proc. Int. Conf. Learn. Represent. (ICLR 2013), 2013.
[38] T. Mikolov, W. Yih, and G. Zweig, “Linguistic regularities in continuous space word representations,” Proc. NAACL-HLT, 2013.
[39] H. Elsahar, E. Demidova, S. Gottschalk, C. Gravier, and F. Laforest, “Unsupervised Open Relation Extraction,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017.
[40] H. Abdi and L. J. Williams, “Principal component analysis,” Wiley Interdisciplinary Reviews: Computational Statistics. 2010.
[41] C. D. Manning, P. Ragahvan, and H. Schutze, An Introduction to Information Retrieval. 2009.
[42] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics). 1990.
[43] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., 1987.

DD882, SF425: As a separate document, please complete the invention disclosure form (**DD882)**, Federal financial report (**SF425**) and sign both. **SF425 must be signed by an authorized person in your business office.**

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.**