MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190227

Identifying Phytochemicals from Biomedical Literature Utilizing Semantic Knowledge Sources

Indra Neil Sarkar, PhD, MLIS[a], Wayne Law, PhD[b], Michael J. Balick, PhD[b]

[a]*Center for Biomedical Informatics, Brown University, Providence, RI, USA*
[b]*Institute for Economic Botany, New York Botanical Garden, Bronx, NY, USA*

Abstract

Chemicals derived from plants (phytochemicals) are major concepts of interest in the study of medicinal plants. To date, efforts to catalogue and organize phytochemical knowledge have resorted to manual approaches. This study explored the potential to leverage publicly accessible semantic knowledge sources for identifying possible phytochemicals. Within the context of this feasibility study, putative phytochemicals were identified for more than 4,000 plants from the Medical Subject Headings Supplementary Concept Records and the Semantic MEDLINE Database. An examination of phytochemicals identified for five selected plant species using the method developed here reveals that there is a disparity in electronically catalogued phytochemical knowledge compared to information from Dr. Duke's Phytochemical and Ethnobotanical Databases maintained by the United States Department of Agriculture. The results therefore suggest that semantic knowledge sources for biomedicine can be utilized as a source for identifying potential phytochemicals and thus contribute to the overall curation of plant phytochemical knowledge.

Keywords:
Plants, Medicinal; Knowledge Bases; Phytochemicals

Introduction

A major facet of ethnobotany, the study of human uses of plant species, is the identification of chemicals that may have an active role in potential medicinal effects [1]. Such chemicals, referred to as "phytochemicals," are identified through a range of extraction and analysis techniques [2]. Reports of phytochemicals associated with a given plant species are then catalogued in monographs or articles that provide description of their actions and constituency. A major foundational step in evaluating the potential medicinal utility of a given plant species therefore requires a listing of associated phytochemicals. The process for identifying and recording phytochemical information is mostly manual, labor intensive, and costly.

A limited number of electronic databases exist, including Natural Products Alert (NAPRALERT) [3] and Dr. Duke's Phytochemical and Ethnobotanical Databases (Dr. Duke's) [4], and are artifacts of manually curated resources (e.g., more than 200,000 articles for NAPRALERT and a limited number of monographs for Dr. Duke's). Maintenance and updating information within such databases can be difficult due to challenges in available resources that are not able to keep up with the growing volume of knowledge.

Within biomedicine, there have been significant advances in developing computational approaches for identifying relevant entities from electronically accessible literature resources. Such approaches commonly utilize publicly accessible biomedical knowledge sources, which are enriched with semantic information that facilitates the inference of putative relationships. Those resources of note include the Medical Subject Headings (MeSH) Supplemental Concept Records (SCR) [5], the Unified Medical Language System [6] Metathesaurus (UMLS Meta), the UMLS Semantic Network (UMLS SN), and the Semantic MEDLINE Database (SemMedDB) [7].

MeSH SCR provides an index of chemicals, drugs, and other concepts of interest to MeSH descriptors (which are, in turn, used for cataloguing biomedical artifacts such as publications in MEDLINE). UMLS Meta is a collection of more than two million biomedical concepts collected from more than 200 classifications, codings, thesauri, and controlled vocabularies organized based on synonymy and discernable relationships between concepts. UMLS SN is a set of broad subject categories that organize concepts from UMLS Meta and relationships among them [8]. Finally, SemMedDB is a database of more than 80 million semantic predications (subject-predicate-object triples), which have been extracted by the SemRep [9] natural language processing tool and underpin the Semantic MEDLINE system [10].

Biomedical knowledge resources are thus designed with at least two purposes, to:

1. Facilitate information retrieval tasks; and
2. Support identification of putative relationships between biomedical concepts.

With regards to the latter, a number of studies have demonstrated how the aforementioned resources can be leveraged for the identification of disease risk factors [11], clinical adverse events [12], disease relationships based on genetic knowledge [13], gene-disease relationships [14], as well as many others. To date, there have been no studies exploring the potential to leverage biomedical knowledge sources for the identification of phytochemicals.

The purpose of this feasibility study was to develop an approach to identify phytochemicals from MeSH SCR and SemMedDB using biomedical concepts indexed in UMLS Meta. In addition to identifying putative phytochemicals, a detailed manual comparison of predicted phytochemicals was done for five selected plant species.

Methods

Recent versions of MeSH, UMLS Meta, UMLS SN, and SemMedDB were accessed from a local MySQL database. The list of plant species that were analyzed for this study originated from NCBI Taxonomy. Processing and analysis of data were done through programs written in Julia [15]. A graphical overview of the process for identifying chemicals associated with plants is shown in Figure 1.

Identification of Chemicals from MeSH SCR

MeSH was queried for each plant species name using an exact match of all plant names from NCBI Taxonomy, resulting in a set of entry terms. For each entry term, the associated MeSH descriptor was identified. When available, additional relevant MeSH descriptors and associated entry terms were identified through entries in the "See Also" (FX) field. Only FX field entries of UMLS semantic type "Organic Chemical" (T109; 'orch') or "Pharmacological Substance" (T121; 'phsu') were included. For the final set of MeSH descriptors, chemical names and corresponding UMLS Meta concept unique identifiers (CUIs) were retrieved by querying the MeSH SCR. The overall process used to identify chemicals from MeSH SCR is graphically depicted in Figure 1 as A1-A7 (green arrows; top half of figure).

Identification of Chemicals from SemMedDB

For each plant species name from NCBI Taxonomy, three queries were made of strings indexed in UMLS Meta to identify a query set of UMLS Meta CUIs:

1. The plant species name itself;
2. The plant species name plus the word "extract"; and
3. The set of MeSH entry terms determined as an artifact of the previous identification of chemicals from the MeSH SCR chemical identification process.

All UMLS Meta CUIs for the resulting query set were required to be of semantic type "Plant" (T002; 'plnt'). SemMedDB was then queried with the query CUIs for the given plant for predications that included one of the following predicate types: "ISA", "LOCATION_OF", or "CONVERTS_TO."

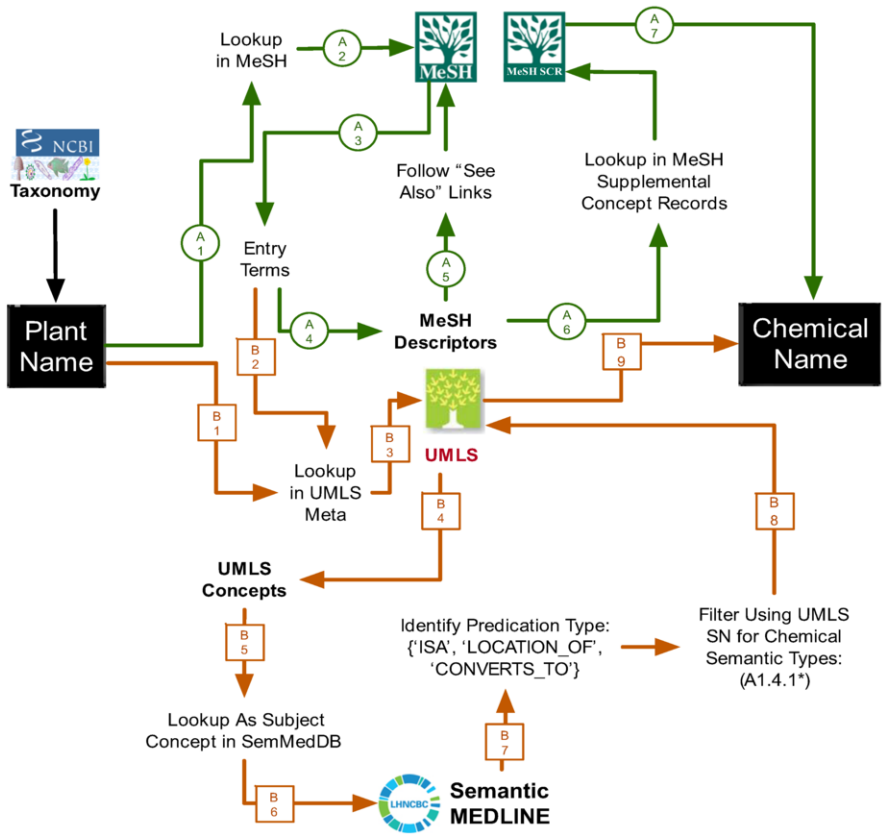

Figure 1. Overview of Approach for Identifying Chemicals Associated with Plants. Two types of knowledge sources were used to identify chemicals associated with plant species as listed in NCBI Taxonomy: (A) MeSH Supplemental Concept Records (SCR); and (B) Semantic MEDLINE. For MeSH SCR, first the full set of MeSH descriptors and entry terms were determined (A1-A5) and then the MeSH SCR was queried to identify associated chemicals (A6-A7). In addition to the NCBI Taxonomy plant species name (B1), entry terms from the MeSH SCR search (B2) were used to identify relevant concepts in the UMLS Meta (B3-B4). These concepts were then searched in Semantic MEDLINE as subject concepts (B6) and restricted to specific predicate types of interest (B7) and filtered for chemical related semantic types (B8-B9).

Table 1: Summary of agreement between chemicals associated with selected plants based on information from Dr. Duke's versus the developed approach (phyotkb_chem). For each of the five plant species examined in detail, the number of chemicals identified by the developed approach from Medical Subject Headings Supplemental Concept Records (SCR) or SemMedDB (SM), as unique to phytokb_chem (u_p), shared by both phytokb_chem and Dr. Duke's (c), unique to Dr. Duke's (u_d), or chemicals that were catalogued in both phytokb_chem and Dr. Duke's but were not identified as either as being associated with a given plant species ($u_{\neg p \neg d}$). For each plant species, respective values for the F-measure (fm), Matthews Correlation Coefficient (mcc), and Cohen's Kappa (κ) are shown.

Plant species (common name)	SCR	SM	u_p	c	u_d	$u_{\neg p \neg d}$	*fm*	*mcc*	κ
Calendula officinalis L. (Marigold)	2	24	24	1	61	5481	0.023	0.103	0.017
Cannabis sativa L. (Marijuana)	42	46	77	9	63	5476	0.011	0.204	0.101
Papaver somniferum L. (Poppy)	5	64	54	14	156	5375	0.118	0.279	0.102
Senna alexandrina Mill. (Senna)	19	10	25	2	29	5518	0.069	0.179	0.064
Solanum lycopersicum L. (Tomato)	0	896	810	85	125	5177	0.154	0.154	0.105

The UMLS Meta CUIs for the objects were retrieved. The candidate objects were filtered for concepts contained within the "Chemical" (T003; 'chem') hierarchy of the UMLS SN (A1.4.1*). The process used to identify chemicals from SemMedDB is graphically depicted in Figure 1 as B1-B11 (orange arrows; bottom half of figure).

Evaluation for Selected Plant Species

For a chosen set of five plant species with known phytochemical properties (*Calendula officinalis* L. [Marigold], *Cannabis sativa* L. [Marijuana], *Papaver somniferum* L. [Poppy], *Senna alexandrina* Mill. [Senna], *Solanum lycopersicum* L. [Tomato]), associated chemicals were retrieved from Dr. Duke's. A complete list of chemicals listed in all plants catalogued in Dr. Duke's was also retrieved.

The full set of Dr. Duke's chemicals were mapped to UMLS Meta CUIs by direct lookup; those chemicals that could not be mapped to a UMLS Meta CUI were not included in the evaluation. The set of chemicals retrieved through the process developed in this study (phytokb_chem) were compared to the set of chemicals from Dr. Duke's for each plant species of interest. In addition to a proportional analysis of the chemicals suggested to be associated with each plant species, three agreement statistics were calculated with Dr. Duke's and phytokb_chem, serving as reference standards to each other:

1. F-Measure (*fm*), which is the harmonic mean of precision and recall for a set of classifications:

$$fm = \frac{2 \cdot c}{(2 \cdot c) + u_p + u_d}$$

2. Matthews Correlation Coefficient (*mcc*), which quantifies the correlation between two classification systems:

$$mcc = \frac{(c \cdot u_{\neg p \neg d}) - (u_p \cdot u_d)}{\sqrt{(2 \cdot c) \cdot (c + u_p) \cdot (u_{\neg p \neg d} + u_p) \cdot (u_{\neg p \neg d} + u_d)}}$$

3. Cohen's Kappa (κ), which ascertains the quality of the relative accuracy of a given classification system as a function of true accuracy compared to random accuracy:

$$\kappa = \frac{\left(\frac{c + u_{\neg p \neg d}}{c + u_{\neg p \neg d} + u_p + u_d}\right) - \left(\frac{\left((u_{\neg p \neg d} + u_p) \cdot (u_{\neg p \neg d} + u_d)\right) + \left((u_d + c) \cdot (u_p + c)\right)}{(c + u_p + u_d + u_{\neg p \neg d})^2}\right)}{1 - \left(\frac{\left((u_{\neg p \neg d} + u_p) \cdot (u_{\neg p \neg d} + u_d)\right) + \left((u_d + c) \cdot (u_p + c)\right)}{(c + u_p + u_d + u_{\neg p \neg d})^2}\right)}$$

Where c represents the number of chemicals in common between Dr. Duke's and phytokb_chem; u_p represents the number of chemicals unique to phytokb_chem; u_d represents the number of chemicals unique to Dr. Duke's; and $u_{\neg p \neg d}$ represents the number of chemicals that are catalogued in either phytokb_chem or Dr. Duke's but not accounted for by u_p or u_d.

Results

A total of 127,597 plant species were searched in MeSH SCR and SemMedDB, resulting in 4,361 plants having at least one chemical. The mean number of chemicals per plant species was 9±100 [95%CI 6-12], with an inclusive range of 1 to 5,589 chemicals per plant species. For the five plant species manually examined relative to Dr. Duke's, there was little overall agreement (summary of agreement statistics shown in Table 1), suggesting that the approach developed here offers potentially synergistic information about chemicals associated with plant species. Detailed results are available as supplementary information at:

https://sites.google.com/a/brown.edu/phytokb/medinfo2019

Discussion

Identification of phytochemicals is an essential aspect in the study of medicinal plants. In advance of the laborious process of developing and implementing screening programs for ascertaining the phytochemicals for a given plant species, it is paramount to have an understanding of what is known about the plant species. To date, the process of cataloguing known phytochemicals relies on a manual curation process. Screening programs, such as those previously led by the National Cancer Institute [16], have led to the identification of active chemicals that have led to significant drugs (e.g., taxol [17]). A major challenge with such screening programs is the difficulty associated with developing appropriate protocols for reliable extraction of phytochemicals [18].

With recent interest in exploring plants as a potential source for new or complementary therapies [19,20], there is a great need to develop robust methodologies for identifying knowledge about plants lest effort be wasted in identifying already known (but perhaps lost) phytochemical knowledge. As with other sectors of biomedicine that have leveraged biomedical knowledge sources, there is a significant opportunity to utilize automated approaches to determine the latest recorded information. This feasibility study demonstrates that there is a strong discordance between existing resources (e.g., Dr. Duke's) and phytochemicals that can be identified using the MeSH Supplemental Concept Records and SemMedDB.

In considering biomedical knowledge sources that may be of utility for identifying phytochemicals, this study focused on using MeSH SCR and SemMedDB. The results of this study

suggest that MeSH SCR and SemMedDB offer complementary knowledge about chemicals. MeSH SCR data are based on drug and chemical information that appear in biomedical literature and can be mapped to MeSH descriptors. As with the MeSH thesaurus, this well-curated list of associations can generally be trusted as *bona fide* associations. However, like the MeSH thesaurus, it is limited in the scope and range of chemicals listed based principles for indexing MEDLINE. Resources like the MeSH thesaurus and MeSH SCR are primarily designed to facilitate information retrieval tasks (i.e., to identify relevant literature for a given query). Nonetheless, the results of this study demonstrate that there is utility in leveraging SCR information to identify potential phytochemicals.

In contrast to MeSH SCR, SemMedDB consists of a detailed set of predications that have been determined using a natural language processing (NLP) system (SemRep). The predications underpin the Semantic MEDLINE system, facilitating the process of retrieving relevant biomedical literature based on semantic relationships between concepts. SemRep based data have been used in a number of text mining studies (e.g., to identify interactions between drugs [21] or proteins [22] from MEDLINE). There have been previous studies in leveraging NLP approaches for studying medicinal plants [23-25]). The present study is, to the best of our knowledge, the first to leverage a predicate extraction system like SemRep for the identification of plant-chemical associations.

A major limitation in the evaluation of the quality of extracted relationships is the limited availability of suitable reference standards. In this study, the phytochemical candidates identified by the developed system were compared to a popular electronic catalogue of phytotherapy knowledge (Dr. Duke's). However, it is important to note that such resources are not "true" reference standards, the information catalogued is incomplete and also may contain inconsistent errors (e.g., chemicals like alcohol are often listed as chemical components, but these are likely artifacts of the phytochemical extraction process [26]). There is some information about putative medicinal uses, but these are supported by limited primary literature. Other resources, like NAPRALERT, have references to primary literature but also may suffer from incorrectness or incompleteness. In this study, the comparison was done with Dr. Duke's because it is a freely accessible resource. Future work will expand to include comparisons to NAPRALERT (which requires a fee for complete access to data on par with Dr. Duke's).

Ultimately, it is envisioned that the work presented in this study will be seen as complementary to resources like Dr. Duke's and NAPRALERT. However, there are still some areas of potential improvement that we consider essential before the results of the developed system could be seen as a reliable catalogue. The first is addressing the issue of identifying chemicals mentioned in literature that are associated with plants only because of the extraction process (e.g., the aforementioned alcohol). While this only occurs in a small percentage of the extractions, it may be easily addressed through a combination of stop words (which would consist of chemicals commonly used for extractions) and information theoretic approaches to identify chemicals that occur commonly and may not be of interest (e.g., water).

Another area of improvement that is planned includes developing a confidence score for the chemicals identified for each plant species. For example, there may be utility in weighting certain predicates in combination with their relative frequency in order to identify chemicals that are of interest for a particular plant species. A more detailed evaluation is needed for each of the data sources for more than five plant species; five plants was chosen to support a manual assessment to verify potential for species that are known to have phytochemicals. Nonetheless, the results of this initial study suggest that there is promise in leveraging biomedical knowledge sources for identifying phytochemicals that are not currently available in electronic phytochemical resources such as Dr. Duke's or NAPRALERT.

Acknowledging that a given plant species has at least one thousand phytochemicals, another important contribution of automated systems such as the one developed here is a tabulation of reported phytochemicals that may supplement data which are in resources like Dr. Duke's or NAPRALERT. In doing so, one can identify plant species that have been heavily studied (e.g., those with more than 1,000 phytochemicals, such as *Solanum lycopersicum* L. [Tomato] as identified either by the process developed here or in Dr. Duke's) versus those that have not had extensive phytochemical analyses shared in accessible resources (e.g., *Calendula officinalis* L. [Marigold] which had only 86 phytochemicals identified either by the process developed here or in Dr. Duke's amidst its long history of medicinal use [27]). The results of this study may thus be used to continually curate available electronic literature sources and, in combination with electronic resources like Dr. Duke's or NAPRALERT, provide a metric for identifying plant species that have been understudied. This would be especially valuable in the context of studying plants that may have therapeutic indications (e.g., based on ethnobotanical survey knowledge).

Biomedical knowledge sources are principally aimed at cataloguing and retrieving information mostly relative to disease knowledge. This can make it challenging to identify non-traditional biomedical concepts, such as plant species. In this study, a combination of MeSH entry terms, MeSH "see also" entries, and UMLS string lookups were used to identify the array of relevant concepts for a given plant species. The source used for taxonomic plant species name in this study was NCBI Taxonomy, which is limited to organisms that have associated data in other databases within NCBI (e.g., GenBank), and does not necessarily reflect the full set of plant species that are catalogued in more comprehensive resources like the International Plant Name Index [28]. Similarly, the set of chemical names indexed within UMLS Meta is not necessarily the full set of known chemicals (i.e., the 5405±138 catalogued chemicals that could not be mapped to UMLS; these were excluded from the evaluation of this study). Future work will therefore need to utilize a more robust approach for identifying plant species names and chemical names that can be added as source concepts for SemRep. Finally, the three target predicate types ("ISA", "LOCATION_OF", or "CONVERTS_TO") used for this study were determined through manual examination of predications associated with plant concepts within SemMedDB; there may additional predications that could be considered (including those that may be added to SemRep) in future work.

This study focused on analyzing biomedical literature indexed in MEDLINE. It is our expectation that the techniques described here may be used to develop approaches to identify phytochemical knowledge from other sources of electronic knowledge (e.g., PubMedCentral or the Biodiversity Heritage Library).

Conclusions

Knowledge about phytochemicals is embedded across a number of resources, including biomedical literature. This

study demonstrates how semantic biomedical knowledge sources can be leveraged to identify potential phytochemicals from literature based resources, focusing on the MeSH Supplement Concept Records and the Semantic MEDLINE Database. The results suggest that automated approaches, such as developed here, can identify a largely non-overlapping, complementary set of potential phytochemicals compared to an existing manually curated resource (Dr. Duke's Phytochemical and Ethnobotanical Databases).

Acknowledgements

This study was supported in part by the the National Institutes of Health grants R01LM011963 and U54GM115677. The content is solely the responsibility of the authors and does not reflect the official views of the National Institutes of Health.

References

[1] D.P. Briskin, Medicinal plants and phytomedicines. Linking plant biochemistry and physiology to human health, *Plant Physiol* **124** (2000), 507-514.

[2] J.B. Harborne, *Phytochemical Methods: A Guide to Modern Techniques of Plant Analysis*, Chapman and Hall, London ; New York, 1980.

[3] W.D. Loub, N.R. Farnsworth, D.D. Soejarto, and M.L. Quinn, NAPRALERT: Computer Handling of Natural Product Research Data, *J Chem Inf Comput Sci* **25** (1985), 99-103.

[4] U.S. Department of Agriculture Agricultural Research Service, Dr. Duke's Phytochemical and Ethnobotanical Databases, 1992-2016.

[5] U.S. National Library of Medicine, MeSH Record Types, 2014.

[6] O. Bodenreider, The Unified Medical Language System (UMLS): Integrating biomedical terminology, *Nucleic Acids Res* **32** (2004), D267-270.

[7] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T.C. Rindflesch, SemMedDB: A PubMed-scale repository of biomedical semantic predications, *Bioinformatics* **28** (2012), 3158-3160.

[8] A.T. McCray, An upper-level ontology for the biomedical domain, *Comp Funct Genomics* **4** (2003), 80-84.

[9] T.C. Rindflesch and M. Fiszman, The Interaction of Domain Knowledge and Linguistic Structure in natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text, *J Biomed Inform* **36** (2003), 462-477.

[10] T.C. Rindflesch, H. Kilicoglu, M. Fiszman, G. Rosemblat, and D. Shin, Semantic MEDLINE: An Advanced Information Management Application for Biomedicine, *Information Services & Use* **2011** (2011), 15-21.

[11] M. Fiszman, G. Rosemblat, C.B. Ahlers, and T.C. Rindflesch, Identifying Risk Factors for Metabolic Syndrome in Biomedical Text, *AMIA Annu Symp Proc* (2007), 249-253.

[12] G.B. Melton and G. Hripcsak, Automated detection of adverse events using natural language processing of discharge summaries, *J Am Med Inform Assoc* **12** (2005), 448-457.

[13] I.N. Sarkar, A Vector Space Model Approach to Identify Genetically Related Diseases, *J Am Med Inform Assoc* **19** (2012), 249-254.

[14] C.B. Ahlers, M. Fiszman, D. Demner-Fushman, F.M. Lang, and T.C. Rindflesch, Extracting semantic predications from Medline citations for pharmacogenomics, *Pac Symp Biocomput* (2007), 209-220.

[15] The Julia Programming Language (2019). https://julialang.org (Accessed April 1, 2019).

[16] M. Suffness and J. Douros, Current Status of the NCI Plant and Animal Product Program, *J Nat Prod* **45** (1982), 1-14.

[17] M.E. Wall and M.C. Wani, Camptothecin and Taxol: from Discovery to Clinic, *J Ethnopharmacol* **51** (1996), 239-253.

[18] N.R. Farnsworth, Biological and phytochemical screening of plants, *J Pharm Sci* **55** (1966), 225-276.

[19] A.S. Bouldin, M.C. Smith, D.D. Garner, S.L. Szeinbach, D.A. Frate, and E.M. Croom, Pharmacy and herbal medicine in the US, *Social Science & Medicine* **49** (1999), 279-289.

[20] J.P. Kelly, D.W. Kaufman, K. Kelley, L. Rosenberg, T.E. Anderson, and A.A. Mitchell, Recent trends in use of herbal and other natural products, *Arch Intern Med* **165** (2005), 281-286.

[21] R. Zhang, M.J. Cairelli, M. Fiszman, G. Rosemblat, H. Kilicoglu, T.C. Rindflesch, S.V. Pakhomov, and G.B. Melton, Using semantic predications to uncover drug-drug interactions in clinical data, *J Biomed Inform* **49** (2014), 134-147.

[22] I.N. Sarkar and T.C. Rindflesch, Discovering protein similarity using natural language processing, *Proc AMIA Symp* (2002), 677-681.

[23] V. Sharma and I.N. Sarkar, Leveraging Biodiversity Knowledge for Potential Phyto-Therapeutic Applications, *J Am Med Inform Assoc* **20** (2013), 668-679.

[24] V. Sharma and I.N. Sarkar, Leveraging concept-based approaches to identify potential phyto-therapies, *J Biomed Inform* **46** (2013), 602-614.

[25] P. Srinivasan and B. Libbus, Mining MEDLINE for implicit links between dietary substances and diseases, *Bioinformatics* **20 Suppl 1** (2004), i290-296.

[26] S. Sasidharan, Y. Chen, D. Saravanan, K.M. Sundram, and L. Yoga Latha, Extraction, Isolation and Characterization of Bioactive Compounds from Plants' Extracts, *Afr J Tradit Complement Altern Med* **8** (2011), 1-10.

[27] E. Basch, S. Bent, I. Foppa, S. Haskmi, D. Kroll, M. Mele, P. Szapary, C. Ulbricht, M. Vora, and S. Yong, Marigold (Calendula officinalis L.): An evidence-based systematic review by the Natural Standard Research Collaboration, *J Herb Pharmacother* **6** (2006), 135-159.

[28] J. Croft, N. Cross, S. Hinchcliffe, E.N. Lughadha, P.F. Stevens, J.G. West, and G. Whitbread, Plant Names for the 21st Century: The International Plant Names Index, a Distributed Data Source of General Accessibility, *Taxon* **48** (1999), 317.

Address for correspondence

Indra Neil Sarkar, PhD, MLIS, FACMI

Brown University

Box G-R

Providence, RI 02912 USA

neil_sarkar@brown.edu