Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Original paper

Data and Text Mining

Analyzing user-generated online content for drug discovery: Development and use of MedCrawler

Andreas Helfenstein[1], and Päivi Tammela[1,*]

[1]Centre for Drug Research, Division of Pharmaceutical Biosciences, Faculty of Pharmacy, P.O. Box 56 (Viikinkaari 5 E), FI-00014 University of Helsinki, Finland

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Ethnopharmacology, or the scientific validation of traditional medicine, is a respected starting point in drug discovery. Home remedies and traditional use of plants are still widespread, also in Western societies. Instead of perusing ancient pharmacopeias, we developed MedCrawler, which we used to analyze blog posts for mentions of home remedies and their applications. This method is free and accessible from the office computer.

Results: We developed MedCrawler, a data mining tool for analyzing user-generated blog posts aiming to find modern 'traditional' medicine or home remedies. It searches user-generated blog posts and analyzes them for correlations between medically relevant terms. We also present examples and show that this method is capable of delivering both scientifically validated uses as well as not so well documented applications, which might serve as a starting point for follow-up research.

Availability: Source code is available on GitHub at {{https://github.com/a-hel/medcrawler}}

Contact: paivi.tammela@helsinki.fi

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Ethnopharmacology, or the study of medicinal plants used by different ethnic groups, has often shown its potential as starting point for drug discovery, and follow-up studies on the ethnomedical use of plants continues to be a valuable source for plant-derived drugs (Gu *et al.*, 2014; Fabricant and Farnsworth, 2001). The traditional use, which implies the remedy being used over a prolonged timespan, vouches to some extent for the innocuousness and efficacy of the preparation. In the European Union, for example, plants that have been 'traditionally used' can qualify for a simplified registration procedure, even if safety and efficacy are not sufficiently documented by scientific literature (European Parliament et al., 2004).

Virtually all civilizations have to some extent used plants as part of their medical system, often embedded in a holistic context combining evidence-based knowledge with supernatural concepts and belief systems. The scientific validation of plants used in these folk medicines can lead to the exploration of new chemical space and can as such be a valuable asset in screening campaigns for drug discovery purposes. Especially in antibiotic research, natural products have a strong track record (Newman and Cragg, 2012), which makes the ethnopharmacological approach an interesting pillar in the containment of newly emerging resistant strains. Many, very heterogeneous, databases are specialized in ethnopharmacological knowledge, and attempts have been made to create a unified data source (Ningthoujam *et al.*, 2012, 2014). Available databases have been used for *in silico* drug discovery (Lagunin *et al.*, 2014).

From an ethical perspective, ethnopharmacology also harbors conflicts. While on one hand, it can help to preserve and distribute traditional knowledge and support the survival of local traditions (Leonti and Casu, 2013), the medicinal plants may, on the other hand, occupy a special status in these people's value system, i.e. they could be considered sacred (Posey, 2002). Furthermore, the traditional knowledge describing the medical practices constitutes intellectual property pertaining to the local

people (Soejarto et al., 2005). The financial stakes of bioprospecting can be exemplified in the story of cyclosporine, where soil samples collected on business trips to Wisconsin and Norway led to a multi-million blockbuster for the Swiss pharma giant Sandoz (now Novartis) (Borel and Kis, 1991; Svarstad *et al.*, 2000). In order to protect and reconcile the interest of all involved parties and to prevent predatory prospecting practices, bioprospecting and rights to genetic resources are now regulated in the Convention on Biological Diversity and other treaties (Efferth *et al.*, 2016; Schüklenk and Kleinsmidt, 2006).

While for a long time knowledge was transferred and shared orally or through scripture, the medium of the modern society is the Internet. The Web 2.0 makes it easy for everyone to generate and consult content and contribute actively in his new role as "prosumer", a contraction of producer-consumer coined by Toffler (1984).

An interesting product of the Web 2.0 are blogs - a modern, public form of the diary. Unlike news articles, blogs undergo little or no editing, employ various levels of writing style, often very colloquial, and optionally allow readers to comment (Elsas et al., 2008). They may contain biased and opinionated material, and information retrieval (IR) from blogs is particularly challenging (Zhang et al., 2007).

When writing a blog, the role of the authors is not limited to the creation of the content, but they are also responsible for the categorization and indexing of their content. Consequently, headings and tags (keywords) can be arbitrary, imprecise and misspelled, or use non-standardized word forms. The texts can employ authentic, recent language and incorporate neologisms (Peters and Stock, 2007). Automated IR uses different metrics to measure the quality and relevance of a blog entry, such as context and semantic analysis (Kandogan et al., 2006) or indirect quality indicators like number of readers, citations/cross-linking, spelling, use of emoticons etc. (Weerkamp et al., 2008).

In this study, we advocate the potential of the Web 2.0 - and blogs in particular - as an information vector of modern folk medicine. For that aim, we developed MedCrawler, a tool to mine and analyze data retrieved from user-generated online sources. Here, we describe the development of the tool as well as a series of example cases, which we used to demonstrate the potential of WebCrawler in searching blogs for putative plant-based remedies against common ailments such as infections and migraine. Results were then cross-examined with scientific literature. Unlike traditional ethnopharmacology, which often includes long travels and a solid local network, our approach is accessible from the office, freely available and highly automatized. It does not violate any copyright, but is still an indicator for various uses or applications of plants for medical use. Undocumented or lesser-known home remedies might, for example, be a starting point for an antimicrobial screening campaign and support the search of new antibiotics.

2 Implementation and methods

2.1 Development and features of MedCrawler

In order to analyze the representation of traditional medical knowledge on the Internet, we developed “MedCrawler”, a tool to automatically retrieve and analyze relevant blog entries. MedCrawler is written in Python and works via command line input. Given the keywords from the user, it finds relevant blog posts, extracts MeSH keywords and represents their occurrence graphically. The tool consists of three submodules (Scraper, Indexer and Grapher); their structure and interaction is shown in **Figure 1**.

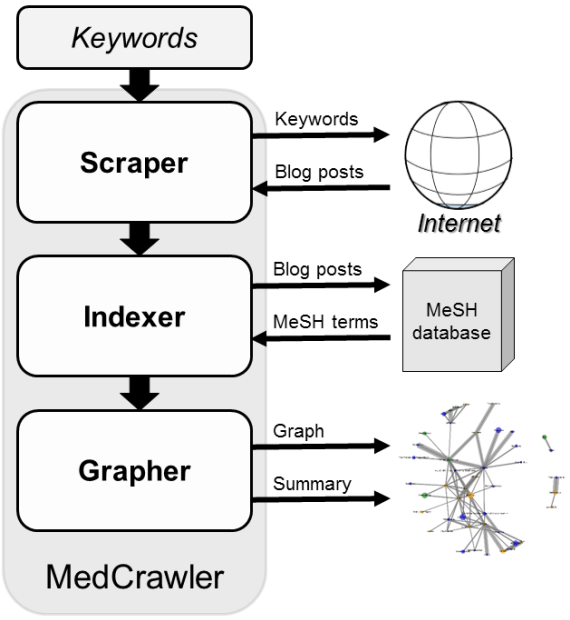

Fig 1. Structure of the MedCrawler workflow. It shows the three submodules (Scraper, Indexer and Grapher) and their interaction with various resources.

The **Scraper** crawls the web and finds blog posts from selected sources that are indexed with user-defined tags. Plugins allow to extend the Scraper’s functionality by providing an interface to different web resources. By default, MedCrawler contains the WordPress (WP) plugin. WP is a free blogging service where subscribers can publish short articles, pictures and hyperlinks. According to their website, it is the "largest self-hosted blogging tool in the world, used on millions of sites and seen by tens of millions of people every day" (WordPress.org). Individual blog posts can be indexed with tags and structured through categories. Public posts can be accessed via web API with the option to search by tag. Users can write their own plugins to include other resources and cover more online content.
The Scraper then returns the relevant blog posts to the Indexer.

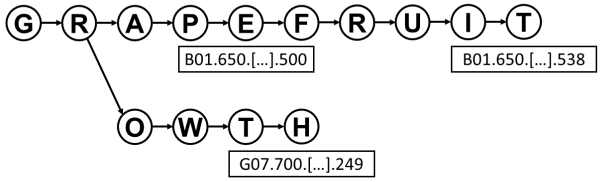

Fig 2. Scheme of the search algorithm. A word is compared character-wise with the tree. If the last character lands on a node with MeSH code, the code is returned. In any other case, the search is aborted.

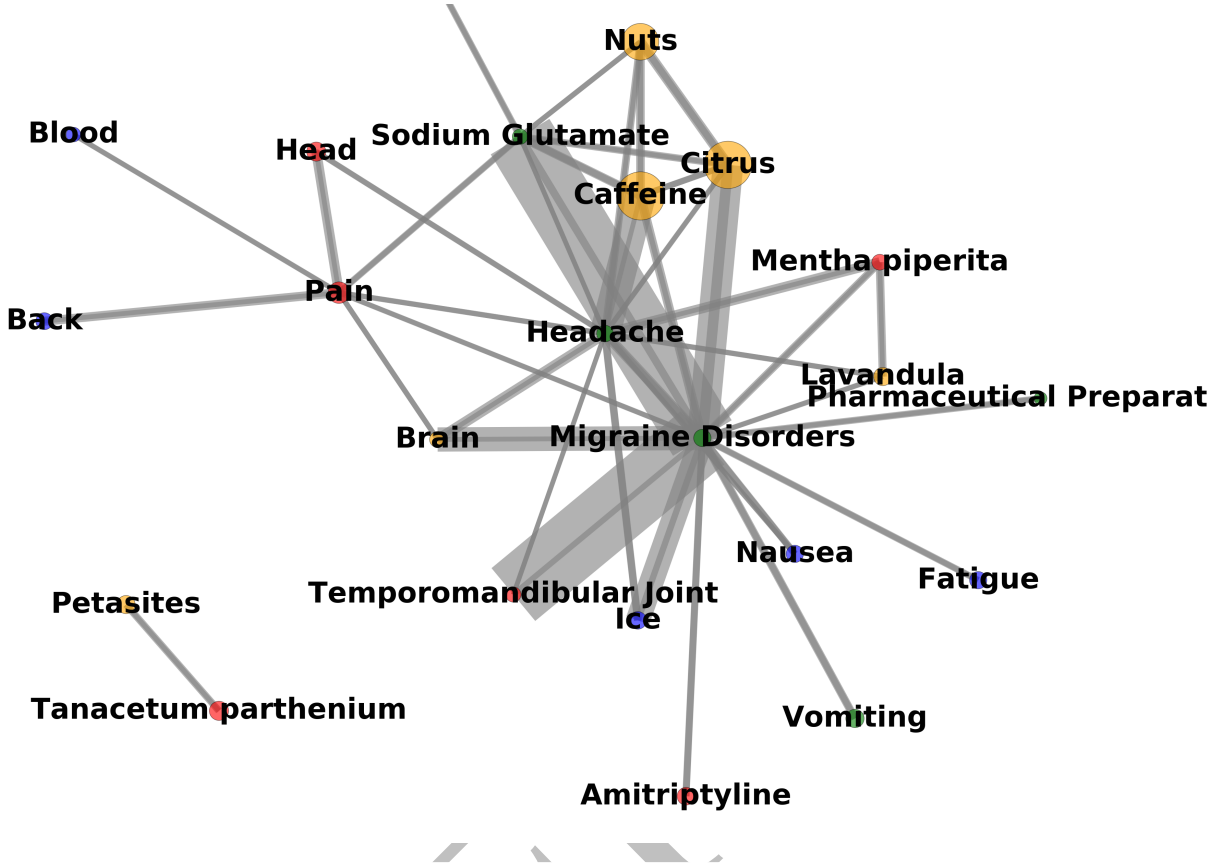

Fig 3. A graph produced with results from Example 3. Original search terms were *pain*, *headache* and *migraine*. Results were filtered to include MeSH categories A, B, C, and D. Only connections with a frequency ≥ 10 are shown. The line width is proportional to the number of counted co-occurrences.

The **Indexer** detects MeSH terms in the raw text through a list of synonyms and assigns the corresponding tree number. It is comparable to the "MeSH on demand" web service of the American National Library of Medicine (NLM). While implementation of this service into the project would be possible, its use requires a user account, and restrictions on the maximum number of requests apply. The local implementation circumvents these limitations and leads furthermore to a considerable speed-up. It is, however, dependent on the MeSH database file, which needs to be downloaded separately.

The 2016 edition of the MeSH thesaurus contains over 27,000 descriptors, and a linear search through thousands of posts would be a considerable burden on running time. The Indexer accelerates this procedure by implementing a character-based search tree, where each letter of a term is a node of the preceding letter. Word look-up occurs character-wise, and the search is aborted as soon as the following letter cannot be found in the tree (**Figure 2**). The use of shared nodes for similar terms furthermore reduces memory usage.

Finally, the **Grapher** analyzes the co-occurrences of the MeSH terms within the individual posts and builds a network graph. Node sizes are scaled according to the occurrences of the term, and edge width represents the number of co-occurrences. The Grapher makes use of the granularity of the MeSH tree numbers to filter and color-code nodes. The user can furthermore specify terms to be highlighted or excluded and indicate a minimum weight for the connections to be displayed. An example graph is shown in **Figure 3**.

Owing to its chunk-wise data retrieval and analysis strategy, MedCrawler is scalable for any arbitrary project size. Besides the initial overhead for the search tree generation, running time depends solely on the amount and length of the received posts.

The module is written in Python and can be easily launched via the provided shell script. It is platform-independent and has been tested on Mac OS X, Linux Mint and Windows 7. The tool (Python source code) is available on GitHub under *http://github.com/a-hel/medcrawler* and extensive documentation describing downloading, use and requirements, can be found at *medcrawler.readthedocs.org*. The most recent MeSH terminology can be obtained from the NLM website.

2.2 Use of MedCrawler

We demonstrated the intended use of MedCrawler by 3 examples, which also showed its strengths and limitations. We carried out a quick literature analysis of the results in order to evaluate their usefulness in the discovery of home remedies and ethnopharmacologically used plants.

The number of analyzed posts and extracted MeSH terms from examples 1-3 are listed in Table 1.

Example 1: Natural anti-infectives

Antibiotic resistance is a serious threat to public health. Natural products have repeatedly proved to be a valuable source of anti-infective compounds. In this example, we searched for the terms *virus*, *bacteria*, *infection*, *flu*, *influenza*, *common cold*, and *fever*, with the aim of finding leads for underestimated anti-infective plants. MeSH categories of interest were B01.650 (plants), B03 (bacteria), B04 (viruses), C01 (bacterial infections and mycoses), C02 (viral infections) and certain parts of D (chemicals and drugs).

Example 2: Insomnia

Self-medication with herbal teas and other preparations is common with light sleeping disorders and insomnia. Example 2 explored their use with the keywords *insomnia*, *sleepless*, and *sleeping disorder*. Results were filtered to include only the categories B (Organisms), C (Diseases), E (Analytical, Diagnostic and Therapeutic Techniques and Equipment), F (Psychiatry and Psychology), and N (Healthcare).

Example 3: Home remedies against pain and migraine

Strong pains, for example through migraines, are common yet incapacitating symptoms. Due to their frequency, we expect that many people have shared their experiences online (Ressler *et al.*, 2012). This search included the terms *pain*, *headache* and *migraine*, while looking for the categories A (Anatomy), B (Organisms), C (Diseases) and D (chemicals and drugs).

Table 1. Summary of the data retrieval

	Number of search terms	Retrieved blog posts	Found MeSH terms	Terms per post
Example 1	7	8450	36150	4.28
Example 2	3	1122	4290	3.80
Example 3	3	5730	12540	2.19

3 Results and discussion

3.1 Development

Benchmark tests

Benchmark tests were performed on a MacBook Air (Apple Corp., Cupertino, CA) with a 1.4 GHz CPU and 8 GB RAM. The test included retrieval of 5,000 posts for 3 different keywords, and performance was analyzed with cProfile. Average running time for 5,000 posts was 1134 s ± 589 s or roughly 20 min. Overhead due to the search tree construction was 32 s for 27,000 terms, and cumulative lookup time for the retrieved posts was 4 s.
Running time hence depends predominantly on the number and length of the retrieved posts and the available Internet connection bandwidth.

3.2 Use

Example 1

To elaborate the functioning of the tool, we highlighted 4 connections returned form the search in example 1: The term *influenza* was associated with *Echinacea* (9 mentions), *bacillus* with *garlic* and *solanum* (5 each), and *viruses* with *eucalyptus* (5).

Echinacea is recognized as a medicinal plant and is widely used in prophylaxis and treatment of the common cold with controversial efficacy (Karsch-Völk *et al.*, 2014). Studies have also advocated its use for influenza virus control (Pleschka *et al.*, 2009). The antibacterial effect of garlic is very well documented. Its main active compound, allicin, has shown activity against a wide range of bacteria, including multiresistant *Burkholderia sp.* and enterotoxic *Escherichia coli* (Ankri and Mirelman, 1999; Wallock-Richards *et al.*, 2014). The antiviral effect of essential oils from eucalyptus has been described, for example against *Herpes simplex* and *Influenza virus* (Astani *et al.*, 2010; Usachev *et al.*, 2013). The genus *Solanum* includes food plants such as potato (*S. tuberosum*) and tomato (*S. lycopersicum*), but also Black Nightshade (*S. nigrum*), which is used in traditional medicine, for example in Myanmar or in traditional Chinese medicine. Its anticancer activity has been shown in several studies (Aung *et al.*, 2016; Razali *et al.*, 2016; Lai *et al.*, 2016).

Example 2

Besides common relaxation techniques like yoga (11) and meditation (10), lavender (12), chamomile (8) and valerian (8) were mentioned in connection with sleep. A meta-analysis of chamomile and valerian could not confirm their efficacy, despite their widespread use against insomnia (Leach and Page, 2015). Lavender (and to a lesser extent chamomile) may have a sleep-inducing effect when inhaled (Wheatley, 2005). Despite the large amount of data considering herbal medicines and insomnia, studies are usually hard to compare due to different preparations and various endpoints (sleep quality, sleep duration, short- or long-term effect etc.).

Example 3

This example highlighted 4 interesting connections, all related to *Migraine disorders* and shown in Figure 3. The term was associated with *caffeine* (18 mentions), *Mentha piperita* (13), *Lavandula* (11), and *citrus* (11).

Caffeine is often used as an adjuvant to boost the efficacy of anti-migraine drugs (Schoenen, 2008). The use of menthol and lavender against migraine attacks have also been investigated (St Cyr *et al.*, 2015; Sasannejad *et al.*, 2012). Citrus has been documented as an anti-migraine drug in an ethnopharmacological context (Jafarpour *et al.*, 2016).

Pertinence of the MedCrawler results

For most connections, there were at least some literary records available. However, their evidence level was very heterogeneous and ranged from systematic reviews to sole mentions of reported uses. This information gap shows the potential of in-depth analyses of certain concepts.

Data retrieval was fully automated and did not undergo human supervision. This behavior is necessary on account of the given workload, but also mandates analysis being done with particular discretion. When

analyzing the results, common sense is necessary to evaluate the semantic context of the associations. Caution is advised in several cases:

- The term can be part of idiomatic expressions (e.g. 'That drives me *nuts*');
- The term can have another meaning in another language (e.g. *pain* also means *bread* in French);
- The association can be positive or negative (e.g. caffeine can help against migraine, but excessive consumption can lead to headache).

Authors of the posts have usually no scientific mission, and the subject of the post depends entirely on the writer's discretion. Current events, such as *Zika virus* infection, can lead to a temporary spike in the occurrence of certain terms. There is furthermore no guarantee for the accuracy of the content, as certain authors might be ignorant, sponsored or willingly misleading.

Scope of the search and careful keyword selection strongly correlate with the amount and usefulness of the results. While commonly encountered ailments and symptoms are a welcome subject (as in Example 3), novel or far-fetched concepts are hardly written about.

To assure the relevance of the results, the method relies on the Wisdom of Crowds through inclusion of large enough datasets. Also, positive correlations are suspected to be reported more often than negative ones. Detected correlations are considered to be pointers towards a subject that requires a closer examination and should be treated as such.

Despite all the caveats, the analysis of user-generated online content delivers comprehensible results that include scientifically validated concepts, principles from alternative medicine and speculative concepts described in ethnopharmacological contexts. Particularly the latter can serve as starting points for in-depth analysis. The results furthermore nicely reflect the uses of medical plants in nowadays' society.

To judge from the cross-analysis with scientific literature, our method seemed to nicely represent folk knowledge in the digital age, reflecting its strengths as well as its weaknesses. Unlike traditional folk medicine, blogs are a highly dynamic source of information and are subject to trends, hypes, and current events. New ideas can be picked up quickly, and unpopular concepts will soon drift into oblivion. From that point of view, our approach presents an excellent complement to 'traditional' ethnopharmacology and taps a vast resource of previously underused information.

The potential of systematic blog content analysis for research uses has been recognized, but is often limited to medical or science blogs written by physicians, patients or researchers (Kovic *et al.*, 2008; Kim, 2009; Fausto *et al.*, 2012). MedCrawler offers a method to include data from a wide variety of sources and authors, and includes automated information extraction and representation.

Being free and rapid, MedCrawler can be used as a supplementary tool to support sample library design for screening campaigns or to provide a quick overview over modern folklore. Its main advantage is the instant access to a vast pool of data, with the considerable downside that results are to be interpreted with caution. Off-label use of the source code or parts of it is equally encouraged.

Funding

This work was supported by the Academy of Finland [grant numbers 284477, 277001].

Conflict of Interest: none declared.

References

Ankri,S. and Mirelman,D. (1999) Antimicrobial properties of allicin from garlic. *Microbes Infect.*, **1**, 125–129.

Astani,A. *et al.* (2010) Comparative study on the antiviral activity of selected monoterpenes derived from essential oils. *Phyther. Res.*, **24**, 673–679.

Aung,H.T. *et al.* (2016) A Review of Traditional Medicinal Plants from Kachin State, Northern Myanmar. *Nat. Prod. Commun.*, **11**, 353–64.

Borel,J.F. and Kis,Z.L. (1991) The discovery and development of cyclosporine (Sandimmune). *Transplant. Proc.*, **23**, 1867–1874.

Efferth,T. *et al.* (2016) Biopiracy of natural products and good bioprospecting practice. *Phytomedicine*, **23**, 166–173.

Elsas,J.L. *et al.* (2008) Retrieval and Feedback Models for Blog Feed Search. In, *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08. ACM, New York, NY, USA, pp. 347–354.

EuropeanParliament (2004) Directive 2004/83/EC. *Off. J. Eur. Union*, 85–90.

Fabricant,D.S. and Farnsworth,N.R. (2001) The value of plants used in traditional medicine for drug discovery. *Environ. Health Perspect.*, **109 Suppl**, 69–75.

Fausto,S. *et al.* (2012) Research blogging: indexing and registering the change in science 2.0. *PLoS One*, **7**, e50109.

Gu,R. *et al.* (2014) Prospecting for bioactive constituents from traditional medicinal plants through ethnobotanical approaches. *Biol. Pharm. Bull.*, **37**, 903–915.

Jafarpour,M. *et al.* (2016) Effect of a traditional syrup from *Citrus medica* L. fruit juice on migraine headache: A randomized double blind placebo controlled clinical trial. *J. Ethnopharmacol.*, **179**, 170–176.

Kandogan,E. *et al.* (2006) Avatar Semantic Search: A Database Approach to Information Retrieval. In, *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06. ACM, New York, NY, USA, pp. 790–792.

Karsch-Völk,M. *et al.* (2014) Echinacea for preventing and treating the common cold. *Cochrane database Syst. Rev.*, CD000530.

Kim,S. (2009) Content analysis of cancer blog posts. *J. Med. Libr. Assoc.*, **97**, 260–6.

Kovic,I. *et al.* (2008) Examining the medical blogosphere: an online survey of medical bloggers. *J. Med. Internet Res.*, **10**, e28.

Lagunin,A.A. *et al.* (2014) Chemo- and bioinformatics resources for in silico drug discovery from medicinal plants beyond their traditional use: a critical review. *Nat. Prod. Rep.*, **31**, 1585–1611.

Lai,Y.-J. *et al.* (2016) Anti-Cancer Activity of *Solanum nigrum* (AESN) through Suppression of Mitochondrial Function and Epithelial-Mesenchymal Transition (EMT) in Breast Cancer Cells. *Molecules*, **21**.

Leach,M.J. and Page,A.T. (2015) Herbal medicine for insomnia: A systematic review and meta-analysis. *Sleep Med. Rev.*, **24**, 1–12.

Leonti,M. and Casu,L. (2013) Traditional medicines and globalization: current and future perspectives in ethnopharmacology. *Front. Pharmacol.*, **4**, 92.

Newman,D.J. and Cragg,G.M. (2012) Natural products as sources of new drugs

over the 30 years from 1981 to 2010. *J. Nat. Prod.*, **75**, 311–335.

Ningthoujam,S.S. *et al.* (2012) Challenges in developing medicinal plant databases for sharing ethnopharmacological knowledge. *J. Ethnopharmacol.*, **141**, 9–32.

Ningthoujam,S.S. *et al.* (2014) NoSQL data model for semi-automatic integration of ethnomedicinal plant data from multiple sources. **25**, 495–507.

Peters,I. and Stock,W.G. (2007) Folksonomy and information retrieval. *Proc. Am. Soc. Inf. Sci. Technol.*, **44**, 1–28.

Pleschka,S. *et al.* (2009) Anti-viral properties and mode of action of standardized *Echinacea purpurea* extract against highly pathogenic avian influenza virus (H5N1, H7N7) and swine-origin H1N1 (S-OIV). *Virol. J.*, **6**, 197.

Posey,D.A. (2002) Commodification of the sacred through intellectual property rights. *J. Ethnopharmacol.*, **83**, 3–12.

Razali,F.N. *et al.* (2016) Tumor suppression effect of *Solanum nigrum* polysaccharide fraction on Breast cancer via immunomodulation. *Int. J. Biol. Macromol.*, **92**, 185–193.

Ressler,P.K. *et al.* (2012) Communicating the experience of chronic pain and illness through blogging. *J. Med. Internet Res.*, **14**, e143.

Sasannejad,P. *et al.* (2012) Lavender essential oil in the treatment of migraine headache: a placebo-controlled clinical trial. *Eur. Neurol.*, **67**, 288–91.

Schoenen,J. (2008) Current migraine management – patient acceptability and future approaches. *Neuropsychiatr. Dis. Treat.*, **Volume 4**, 1043.

Schüklenk,U. and Kleinsmidt,A. (2006) North-south benefit sharing arrangements in bioprospecting and genetic research: A critical ethical and legal analysis. *Dev. World Bioeth.*, **6**, 122–134.

Soejarto,D.D. *et al.* (2005) Ethnobotany/ethnopharmacology and mass bioprospecting: Issues on intellectual property and benefit-sharing. *J. Ethnopharmacol.*, **100**, 15–22.

St Cyr,A. *et al.* (2015) Efficacy and Tolerability of STOPAIN for a Migraine Attack. *Front. Neurol.*, **6**, 11.

Svarstad,H. *et al.* (2000) From Norway to Novartis: cyclosporin from *Tolypocladium inflatum* in an open access bioprospecting regime. *Biodivers. Conserv.*, **9**, 1521–1541.

Toffler,A. (1984) The Third Wave Bantam.

Usachev,E. V. *et al.* (2013) Antiviral activity of tea tree and eucalyptus oil aerosol and vapour. *J. Aerosol Sci.*, **59**, 22–30.

Wallock-Richards,D. *et al.* (2014) Garlic revisited: antimicrobial activity of allicin-containing garlic extracts against *Burkholderia cepacia* complex. *PLoS One*, **9**, e112726.

Weerkamp,W. *et al.* (2008) Credibility improves topical blog post retrieval. Association for Computational Linguistics (ACL).

Wheatley,D. (2005) Medicinal plants for insomnia: a review of their pharmacology, efficacy and tolerability. *J. Psychopharmacol.*, **19**, 414–21.

WordPress.org About --- WordPress.

Zhang,W. *et al.* (2007) Opinion Retrieval from Blogs. In, *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07. ACM, New York, NY, USA, pp. 831–840.