BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i290–i296
DOI: 10.1093/bioinformatics/bth914

Mining MEDLINE for implicit links between dietary substances and diseases

Padmini Srinivasan[1,*] *and Bisharah Libbus*[2]

[1]*School of Library and Information Science, University of Iowa, Iowa City, IA 52242, USA and* [2]*Lister Hill Research Center, National Library of Medicine, Bethesda, MD 20852, USA*

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: Text mining systems aim at knowledge discovery from text collections. This work presents our text mining algorithm and demonstrates its use to uncover information that could form the basis of new hypotheses. In particular, we use it to discover novel uses for *Curcuma longa*, a dietary substance, which is highly regarded for its therapeutic properties in Asia.

Results: Several disease were identified that offer novel research contexts for curcumin. We analyze select suggestions, such as retinal diseases, Crohn's disease and disorders related to the spinal cord. Our analysis suggests that there is strong evidence in favor of a beneficial role for curcumin in these diseases. The evidence is based on curcumin's influence on several genes, such as COX-2, TNF-alpha, JNK, p38 MAPK and TGF-beta. This research suggests that our discovery algorithm may be used to suggest novel uses for dietary and pharmacological substances. More generally, our text mining algorithm may be used to uncover information that potentially sheds new light on a given topic of interest.

Availability: Contact authors.

Contact: padmini-srinivasan@uiowa.edu

INTRODUCTION

Serendipity has often shaped the pathways of science. Classic examples, include Fleming's observations of a culture of *Staphylococcus* dissolving when the plate was accidently contaminated with a blue-green mold. This discovery of penicillin eventually led to the development of antibiotics. The serendipitous discoveries of artificial sweeteners, saccharine and aspartame, although possibly not as momentous, resulted when scientists accidently tasted spills in their research laboratories. Serendipity may be influenced by several intangibles, including researcher intuition, prior experience and knowledge, and the ability to creatively span multiple disciplines.

*To whom correspondence should be addressed.

But what if there are systems designed to make such discoveries more likely and in effect make them less serendipitous? Clearly, such systems would have to separate the potentially meaningful connections from a vast, and mostly, noisy background of random associations. This, in essence, is a key goal in text mining research. Text mining systems strive to analyze simultaneously the published literature from multiple disciplines, filter through the evidence, identify implicit connections that have potential and then ensure that these are novel in the sense that they have not yet been explicitly addressed. Recently, text mining applied to the biosciences has been referred to as conceptual biology and its importance in fueling hypothesis driven biomedical explorations are described in Blagosklonny and Pardee (2002).

Several successes may be attributed to text mining research. The pioneering work of Swanson and Smalheiser beginning from the mid-1980s resulted in several hypotheses that were later verified by clinical studies. For example, Swanson (1986) proposed that fish oils may be used to treat Raynaud's disease and this was later corroborated by DiGiacomo *et al*. (1989). More recently, their text mining approach was used to identify viruses that may be used as bioweapons (Swanson *et al*., 2001) Others have proposed, e.g., therapeutic uses of thalidomide (Weeber *et al*., 2003) and found functional connections between genes (Chaussabel and Sher, 2002; Sehgal *et al*., 2003; Shatkay *et al*., 2000).

Our text mining algorithms follow the discovery framework set by Swanson and Smalheiser. In summary, they proposed an open discovery approach that is depicted in Figure 1.

This process is initiated with a single topic (A) of any type, such as a disease, a pharmacological substance or a gene. Starting with the A topic and navigating through intermediate topics (B1, B2, etc.), the goal is to reach terminal topics (C1, C2, etc.) that shed new light on A. Swanson used this discovery processes to find the Raynaud's disease (A) and fish oils (C) connection. Swanson also proposed closed discovery where two topics (A and C) form the starting point and the goal is to determine if there are novel connections (B1, B2, etc.) between them (Smalheiser and Swanson, 1996a,b,

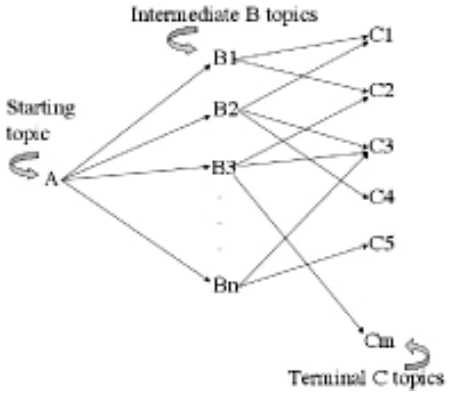

Fig. 1. Open discovery: general model for discovering implicit links between topics.

1998; Swanson, 1988). The closed discovery process supports hypothesis testing of a user's intuition regarding a relationship (a particular A–C or A–B–C combination).

The open and closed discovery framework is at the core of an active research agenda on designing alternative text mining algorithms with the Swanson and Smalheiser discoveries offering a test bed (Lindsay and Gordon, 1999; Weeber *et al*., 2001; Srinivasan, 2004). In our replication of their eight open and closed discoveries, our algorithms were the most effective and required the least amount of manual input and analyses (Srinivasan, 2004).

The goal in this research is to determine if our open discovery algorithm can uncover interesting and new information that might lead to reasonable hypotheses. In particular, we use open discovery to explore the therapeutic potential of curcumin/turmeric (*Curcuma longa*), a dietary substance, commonly used in Asia. Our algorithm identifies several diseases or disorders that could form the basis of new and testable hypotheses involving curcumin. Analysis of three suggestions, retinal diseases, Crohn's disease and problems related to the spinal cord, uncovers genetic and biochemical evidence suggesting that treatment with curcumin may indeed be beneficial.

Next, we describe our open discovery algorithm and its application to turmeric. We then present the results obtained. Following this, we present an analysis of the evidence supporting three of the diseases suggested by our algorithm. Finally, we present our conclusions.

SYSTEMS AND METHODS

Open discovery algorithm

Our implementation of Swanson's open discovery process is built on the notion of topics and their profiles. Topics may be simple such as Tylenol or more complex such as Calcium channel blockers and Alzheimer's disease. A topic's profile is a representation of the topic, which is built from a set of documents that are relevant to the topic. A topic profile is a set of terms (single words and phrases) extracted from these documents. A numerical weight is associated with each term to indicate its relative importance in representing the topic. When applying open discovery to MEDLINE, documents are retrieved via the PubMed interface (http://www.nlm.nih.gov). Also, terms extracted are MeSH (Medical Subject Headings) metadata terms that are assigned by trained indexers at NLM. Thus, our profiles are weighted vectors of MeSH terms.

We also exploit the 134 UMLS (Unified Medical Language System) semantic types. Each MeSH term is assigned to one or more semantic types. For example, interferon type II falls within both Immunologic Factor and Pharmacologic Substance semantic types. Depending on the nature of the discovery goals, we restrict the discovery process to certain semantic types. The topic profiles are then restricted to MeSH terms belonging to these semantic types.

Term weights are TF*IDF scores, where TF_i (term frequency) is the number of times the MeSH term t_i occurs in the retrieved document set and IDF_i (inverse document frequency) is $\log(N/\mathrm{TF}_i)$. N is the number of documents retrieved for the topic. Weights are normalized as shown below for term t_i. This vector of weighted MeSH terms forms the topic profile.

$$\mathrm{weight}(t_i) = v_i \Big/ \sqrt{v_1^2 + v_2^2 + \cdots + v_r^2}, \qquad (1)$$

where $v_i = \mathrm{TF}_i * \log(N/\mathrm{TF}_i)$ and there are r distinct terms in the profile.

We outline our open discovery algorithm below.

- Input: (1) an A topic, (2) ST-B and ST-C: two sets of UMLS semantic types and (3) M.
  1. Search PubMed for A, and build its topic profile (AP).
  2. For each semantic type in ST-B, select the M top ranking MeSH terms from AP. Remove duplicates. Call these (B1, B2, B3, etc.).
  3. Search PubMed for terms B1, B2, B3, etc. (independently) and build their profiles (BP1, BP2, BP3, etc.).
  4. Build a combined profile limited to ST-C semantic types where the combined weight of a MeSH term is the sum of its weights in BP1, BP2, BP3, etc. (CP).
  5. Eliminate term t in CP if a PubMed search on A AND t retrieves documents.
- Output: For each semantic type in ST-C, output MeSH terms in CP ranked by the combined weight.

The role of ST-B and ST-C in the algorithm is to apply semantic constraints to the problem and shape the path of the discovery process. Similarly, parameter M may be used to focus the discovery process. The higher this number the

bigger the scope through which one looks for novel C topics. Obviously, it takes experience to come up with reasonable values for these parameters. However, we already see some patterns emerging in the MEDLINE mining literature. For example, when looking for substances that are likely to influence a disease, several researchers have used functional semantic types, such as Cell Function and Molecular Dysfunction, for selecting intermediate pathways (e.g. Weeber *et al*., 2001). Experiments varying these semantic types have been described in our previous work (Srinivasan, 2004). The advantage in using our open discovery algorithm is that a user searching for new ideas related to A may focus on the ranked C topics identified.

Open discovery with turmeric

Turmeric/curcumin is a widely used spice in Asia and is highly regarded for its curative and analgesic properties. These include the treatment of burns, stomach ulcers and ailments, and various skin diseases. Curcumin is an antiseptic, it alleviates symptoms of the common cold and also serves as a depilatory. We initiate an open discovery process with curcumin as our A topic seeking a set of novel diseases for which the substance may be useful.

The PubMed search conducted was *Turmeric or Curcumin or Curcuma*. We limited ST-B to the three semantic types Gene or Genome; Enzyme; and Amino Acid, Peptide or Protein, since we are looking for biochemical and genetic connections between turmeric and novel diseases. ST-C was restricted to three semantic types: Body Part, Organ or Organ Component; Disease or Syndrome and Neoplastic Process. Neoplastic Process includes MeSH terms referring to cancers. Since turmeric is known to have beneficial effects for problems of the stomach, we include the first semantic type listed above to determine if turmeric may be beneficial in the context of other organs. We experimented with M values set to 5, 10 and 15. However as described below, we focus our analysis mainly on the middle value of $M = 10$.

We also experiment with a variant of our automatic open discovery process. Instead of identifying B terms automatically (step 2), these were manually identified by our user (the second author) after studying select documents retrieved by the curcumin search. Figure 2 illustrates the two parallel paths in our experiments. By varying just the mechanism for selecting B terms, we compare our automatic method with a manual method that benefits from certain decisions made by a domain specialist. We wish to determine if the C terms suggested by these two paths differ.

RESULTS

A total of 1175 PubMed documents were retrieved from the curcumin search. The majority of these publications (1043, 89%) were published in 1990 or later. This indicates a surge in interest in the health effects of this spice, which has long been valued in Asia for its medicinal properties.

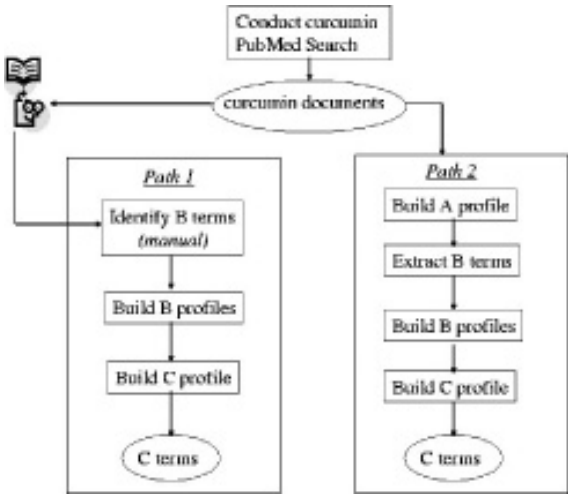

Fig. 2. Parallel experiments involving (1) automatically and (2) manually selected B terms.

B terms

Table 1 shows the top 10 MeSH terms that were automatically selected (step 2, $M = 10$) for each semantic type. After removing duplicates in step 2 there were 26 B terms. Some are very specific such as Glutathione Transferase while others represent families such as DNA-Binding Proteins and Isoenzymes. These B terms are in general relevant to curcumin. For example, curcumin strongly down-regulated c-Jun amino(N)-terminal kinase (JNK) (14627502, 12859962, 11370761, 12097302), which is a mitogen-activated protein kinase (MAPK). Numbers such as these in parantheses refer to PMIDs, i.e. identifiers of PubMed records, which may be accessed directly at NLM's pubmed website. Curcumin inhibits nuclear factor-kappa B (NF-κB) (12714587) leading to the suppression of cell proliferation and the induction of apoptosis in multiple myeloma (12393461). Also curcumin is an inhibitor of AP-1 (12853969).

Table 2 lists the 27 terms that were manually identified by the user after reviewing select turmeric documents. Note that the user is not restricted to MeSH and may select B terms from any part of the MEDLINE record. Interestingly, 10 of these terms, identified by italics, were also found by our automatic methods. The recall and precision scores for our automatic method judged against this manual set with exact match criteria are 0.37 and 0.38, respectively. This comparison is done only for our $M = 10$ run since the resulting set of 26 unique terms is the closest in size to the set of 27 selected by the user. If the comparison were relaxed to consider for example, the fact that JNK and extracellular signal-regulated kinases (ERK) are MAPK members, the scores would increase.

C terms

Table 3 presents the top three suggested MeSH C terms for each semantic type of interest. The column labeled

Table 1. Automatically identified B terms

Rank	MeSH term
Gene or genome	
1	Genes, jun
2	Genes, fos
3	Genes, APC
4	Genes, Reporter
5	Genes, Dominant
6	Genes, ras
7	Genes, rel
8	Genes, bcl-2
9	Nucleolus Organizer Region
10	Genes, myc
Enzyme	
1	MAPK
2	Glutathione Transferase
3	Protein Kinase C
4	Prostaglandin–Endoperoxide Synthase
5	Isoenzymes
6	Protein-Tyrosine Kinase
7	Caspases
8	Nitric-Oxide Synthase
9	Ornithine Decarboxylase
10	MAP Kinase Signaling System
Amino Acid, Peptide or Protein	
1	NF-kappa B
2	Transcription Factor AP-1
3	MAPK
4	Proto-Oncogene Proteins c-jun
5	Glutathione Transferase
6	Tumor Necrosis Factor
7	Glutathione
8	DNA-Binding Proteins
9	Protein Kinase C
10	Prostaglandin–Endoperoxide Synthase

Table 2. Manually identified B terms

Term
Prostaglandin–Endoperoxide Synthase
MAPK
Glutathione Transferase
NF-kappa B
Protein-Tyrosine Kinase
Protein Kinase C
Tumor Necrosis Factor
Transcription Factor AP-1
Ornithine Decarboxylase
Proto-Oncogene Proteins c-jun
TIMP-3
Tumor Growth Factor
Matrix Metalloproteinase 13
Epidermal growth factor
P53
Interferon IFN-gamma
5-lipoxygenase
JNK
Interleukins
BCL
ERK
Alkaline Phosphatase
Smad3
JAK
STAT1
HO-1
AKT

Italicized terms were also identified automatically in the $M = 10$ run.

Table 3. Novel C MeSH terms

Rank	Automatic	Manual
Body part, organ or organ component		
1	Retina (1)	Retina (1)
2	Spinal cord (2)	Pulmonary Alveoli (2)
3	Pulmonary Alveoli (4)	Spinal Cord (8)
Disease or Syndrome		
1	Cystic Fibrosis (1)	Cystic Fibrosis (5)
2	Amyotrophic Lateral Sclerosis (4)	Pleurisy (6)
3	Crohn Disease (8)	Schistosomiasis mansoni (7)
Neoplastic Process		
1	Laryngeal Neoplasms (4)	Choriocarcinoma (3)
2	Choriocarcinoma (6)	Laryngeal Neoplasms (6)
3	Meningioma (8)	Hodgkin Disease (7)

$M = 10$ for the automatic run.

'Automatic' represents the $M = 10$ run. As an example, Retina is ranked first for both methods under Body Part, Organ or Organ Component. Observe that in step 5 of the algorithm we check for novelty of the C term with respect to A. However, this step is not purely automatic since we need to consider synonymous search terms. At this point the list of novel C terms produced initially by our algorithm is further refined manually. In Table 3, the numbers in parentheses indicate ranks before manual refinement. For example, Helicobacter infections was initially ranked second in the automatic method (not shown in the table) and first in the manual method. However, a search on curcumin and helicobacter retrieved two documents of which one is about the inhibition of *Helicobacter pylori* growth by curcumin. Hence, this term was removed from the list of novel C terms for A. Interestingly, six of the suggestions after refinement are identical to both the automatic and manual methods yielding an overlap of 67%. Recall and precision for the automatic method judged against the manual results for these top listings are also both at 67%. Interestingly, the automatic algorithm run with $M = 5$ or 15 does not change these recall and precision scores. We only observe slight changes in the relative rankings of the C terms.

The user may now peruse the appropriate literature to determine the strength of the evidence and the nature of the relationship between curcumin and each suggested disease as

the substance that could be beneficial or harmful. Next, we present such an analysis (conducted manually by the second author) for entries 'Retina', 'Spinal Cord' i.e. for associated diseases/disorders and 'Crohn disease'. The first two are the top two entries for the automatic run that also appear within the top three positions for the manual run. Crohn is analyzed as it is the last ranked automatic entry in its semantic group and it does not appear in the manual list of top three. We wish to determine if Crohn may be a false positive suggestion. In each case, the goal in analysis is to identify biochemical pathways potentially connecting the disease and curcumin. We postpone analysis of entries in the Neoplastic Process category to future research.

ANALYSIS

Retinal diseases

Retinal diseases could result from complications owing to diabetes (diabetic retinopathy), or of infection and inflammation of the retina. An early sign of diabetic retinopathy, a leading cause of blindness, is the adhesion of leukocytes to the vessels of the retina, endothelial cell injury and the breakdown of the blood–retina barrier (12000720[1]). Glaucoma, the second most common cause of blindness in the world (8695555), is caused by mutations in a number of genes on chromosomes 1 and 10 as well as in other loci on chromosomes 2, 3, 8 and 7. While one or a few genetic loci control disease progression and familial transmission, it is often the case that a variety of genes may be involved in their pathophysiology. Following is a brief survey of some of the genes that may be involved in tissue injury. Such genes could provide strategies for therapeutic intervention using curcumin.

In diabetes and during inflammation, periods of hypoxia, i.e. low oxygen concentration, occur in various tissues and organs. At such times an early cellular response results in the elevated expression of interleukin-1β (IL-1β) and cyclo-oxygenase 2 (COX-2) genes (11527948, 14507857, 11821258), which in turn stimulate new blood vessel growth leading to retinopathy (12821538, 12601017). COX-2 expression was also associated with the development of glaucoma (9441697). COX-2 inhibitors suppressed blood–retinal barrier breakdown and prevented the growth of new blood vessels and thus had a protective effect on the retina (12821538, 11980873).

Another gene, tumor necrosis factor α (TNF-α), was elevated during the early stages of diabetic retinopathy and inflammation (11821258, 12706995, 11161842). Activation of TNF-α and other genes may also lead to the pathophysiology of glaucoma (10975909, 10815159). Anti-TNF-α treatment reduced leukocyte adhesion to blood vessels of the eye and vascular leakage (12714660) indicating a potential therapeutic effect for reducing ocular inflammation.

[1]MEDLINE records identified in parentheses were manually identified by the user via PubMed.

The family of MAPK is another group of genes that has an important role in retinal disease. These include ERK, JNK and p38. ERK, was induced in glaucoma (12824248). Often inflammatory responses include the induction of apoptosis or programmed cell death. Involvement of JNK in inducing apoptosis was demonstrated in retinal cells (12270637). Inhibitors of MAPK inhibited retinal pigment epithelial cell proliferation (12782163).

TNF-α (discussed above) is linked to MAPK as it activates phosphorylation of ERKs, p38 and JNK MAPK in human chondrocytes (12878172). These MAPK genes are also activated by IL-1β activation, which is induced by the presence of retinal holes, a key feature of diabetic retinopathy (12824248). NF-κB, whose level changes as an early response to inflammation, also stimulates these MAPK genes (12878172). Moreover, activation of TNF-α was followed by increased transcription of NF-κB (12878172). Also activation of NF-κB subsequently stimulated COX-2 (12807725).

Curcumin was effective in inhibiting cell proliferation of tumorigenic and non-tumorigenic breast cancer cells (12527329) and other tumor cells (12680238). It also suppressed COX-2 (12844482) and neutralized the effect of IL-1β, possibly through its effect on p38 and COX-2 and JNK (12957788). Curcumin inhibits JNK (12957788, 12854631, 12582006, 12130649, 12105223, 9674701) and also suppresses NF-κB activation (11753638, 11506818, 12878172, 12825130). Having shown that these genes, in particular, IL-1β, COX-2, TNF-α, JNK, ERK, NF-κB, etc. are involved in retinopathy and in regulating cell proliferation and leukocyte attachment and the breakdown of the blood–retina barrier, and having established that curcumin is capable of inhibiting the activity of these genes we hypothesize that curcumin may have therapeutic value in preventing or ameliorating a number of retinal pathologies, including diabetic retinopathies, ocular inflammation and glaucoma.

Crohn's disease

Crohn's disease, characterized by a chronic relapsing intestinal inflammation, has a number of genes or chromosomal loci (CARD15 or NOD2; 19p13, 16q12, 16p, 14q11–q12, 12p13.2–q24.1, 6p, 5q31, 1p36). For example, in Crohn's disease of the terminal ileum, paneth cells that are most numerous there, prominently expressed NOD2 (CARD15) (12851870). Bacterial wall protein is believed to activate the gene CARD15 as well as NF-κB, a pro-inflammatory molecule that confers susceptibility to Crohn's disease (12840668, 12676561,12626759, 12527755, 12673278).

Additionally, Crohn's disease and ulcerative colitis, both IBDs (Irritable Bowel Diseases) are commonly classified as autoimmune diseases, implicating a number of inflammatory cytokines in the process of pathogenesis. The balance between pro- and anti-inflammatory cytokines, particularly between the pro-inflammatory interferon -γ (IFN-γ) on one hand and the anti-inflammatories IL-4 and TGF-β activity, is believed to

control chronic intestinal inflammation (11994418). Restoring TGF-β1 signaling in chronic IBD, by inhibition of Smad7, results in the TGF-β1 induced inhibition of cytokine production (11518734). Intestinal cells from patients with Crohn's disease produced IL-18, a pleiotropic cytokine that augments IFN-γ production (11751987), as well as IL-12 another pro-inflammatory (11570528). In both Crohn's disease and ulcerative colitis, IL-12 and IL-17, mRNA are induced and are believed to be involved in sustaining intestinal inflammation (12678335).

Intestinal cells in Crohn's also produce TNF-α RNA (11570528). The regulation of TNF-α, a key mediator in the inflammatory process, is interconnected with MAPK pathways in IBD: the p38α, JNKs and ERK1/2 MAPKs were significantly activated (11994493). Patients with IBD can be helped by anti-TNF therapy (12047261, 12190096, 12421092). Inhibition of NF-κB activation (9616307) as also inhibition of IL-1β, IL-6, IL-8 and TNF-α production (9468102, 12005259) are shown to be beneficial.

Curcumin inhibits a number of these genes and cytokines. For example, it inhibits NF-κB activation and in turn suppresses TNF-α signaling and its target transcription factors (11753638, 11506818, 12878172, 12578124, 7786295, 12825130). Curcumin also influences MAPK-based pathways. For example, it inhibits JNK (12957788, 12854631, 12582006, 12130649, 12105223, 9674701). Curcumin also influences TGF-β: in wounds, which involve inflammation, the observed beneficial effect of curcumin treatment was attributed to an increase in TGF-β (9776860). Curcumin significantly inhibited IL-12 leading to decreased IFN-γ induction and increased induction of IL-4 (an anti-inflammatory 10510448). Given the roles of these genes in IBD in general and Crohn's and ulcerative colitis in particular, we hypothesize that curcumin may have beneficial effects on both Crohn's and ulcerative colitis.

Spinal cord

Two aspects involving the spinal cord are analyzed. First, we look at spinal cord injuries. Second, we look at experimental autoimmune encephalomyelitis (EAE), an autoimmune model resembling multiple sclerosis, the human demyelinating disorder (11043609).

Spinal cord injury leads to inflammation that once again involves the pro-inflammatory and anti-inflammatory cytokines (12165135, 14637102). For example, cytokines TNF-β and LT-β production increased and was followed 18 h later by TGF-β1 upregulation (12127673, 12828562). Once again TNF-α is observed to be a major neuroinflammatory player whose effect is synergized with other cytokines (13678668, 12471141, 12933842, 12363412). In another study IL-6 is up-regulated following injury (12932839) with its attenuation contributing to decreased inflammation (12363412). Patients with spinal cord injury had significantly higher levels of IL-2 and TNF-α, (14593216, 12165135). Similar patient observations were made for IL-1α, IL-1β, in addition to TNF-α and IL-6 levels while blocking of IL-1 and TNF-α receptors significantly reduced their expression (12111861).

After nerve injury, COX-2 is up-regulated in the spinal cord (14697327). Moreover, enhancement in spinal COX-2 expression is linked to spinal sensitization (12950462). It has also been suggested that a spinal interaction of COX-2 inhibition with opiate analgesia may allow for a reduction of post-operative pain with lower doses of opiate (12373690).

Because of its influence on these cytokines, and the inflammatory consequences of injury, curcumin is potentially beneficial in treating spinal cord injuries. Moreover, curcumin inhibition of COX-2 transcription and protein expression (11751448, 11566484) also suggests a role for curcumin in reducing post-operative spinal cord pain or pain that results from direct injury to the spinal cord.

In experimental EAE, significant levels of TNF and IL-6 were found in the spinal cord (12363412). EAE rats treated orally with Am-80, a synthetic retinoid, had transcriptional levels of the pro-inflammatory cytokines IL-6, IFN-γ and TNF-α that paralleled with the clinical symptoms (11043609). Linomide administration, which delayed the interval between immunization and onset of EAE in a dose-dependent fashion, suppressed the pro-inflammatory cytokines IFN-γ and TNF-α, and up-regulated IL-4, IL-10 and TGF-β in spinal cord sections (9630163). Suppression of the clinical signs in EAE was paralleled by reduced chemokine and cytokine expression (12112074). Clinical EAE was induced by the administration of IL-12, but not of IFN-γ and TNF-α, to GPBP/IFA-immunized animals (11385625). Interestingly it has been observed that, CNS-confined inflammation induced by IFN-γ may induce protective immunological countermechanisms in EAE/multiple sclerosis (11466408). The profile of TNF-α mRNA expression roughly paralleled the clinical signs of EAE (7593556). In the same study, IL-12 expression appeared early and before onset of clinical signs of EAE while IL-10 appeared increasingly at and after clinical recovery. Suppression of EAE by estrogen has also been postulated to occur through a hormone-dependent regulation of TNF-α production (11418693).

Once again, because of its recognized anti-inflammatory properties (14637278, 14637190) and more specifically its influence on cytokines in the spinal cord, such as TNF-α, IL-12, IFN-γ, IL-4, IL-6 (11753638, 11506818, 12878172, 12578124, 7786295, 12825130, 12594059) and their role in EAE, we hypothesize that curcumin may have beneficial effects in EAE and multiple sclerosis.

CONCLUSIONS

We presented our open discovery algorithm and results obtained when using it to look for novel therapeutic roles for turmeric. We analyzed several of the top ranked

suggestions: retinal diseases (diabetic retinopathy, inflammation and glaucoma), Crohn's disease and disorders related to the spinal cord (injuries as well as EAE). In each case, plausible connections between curcumin and the disorder were found. It should be emphasized that in each case co-occurrence in MEDLINE between curcumin and the disorder was not observed. The suggested indirect connections are based primarily on genes, such as TNF-α, MAPK, NF-κB, COX-2, and other cytokines and interleukins. We also compared two versions of the open discovery process. In one version, B terms were automatically selected and in the other these were identified manually. Recall and precision scores for the B terms identified by the automatic method when judged against the ones identified manually by the user, under a strict match condition, were 37 and 38%, respectively. Despite these relatively low numbers, the recall and precision scores for the top C terms finally suggested by the automatic method when judged against the manual method's output were both 67% as was overlap between them. These C term based scores did not change as *M* was set to 5 or 15, indicating robustness of our algorithm. One limitation in our algorithm, we have observed with this research, is the need to manually refine the C list generated by our algorithm. Thus, our immediate goal is to further automate step 5. In particular, we will explore semantic relationships expressed in the UMLS to look for synonymous and near-synonymous terms that may be used for searching. We will also continue testing our open discovery algorithm on other dietary as well as pharmacological substances. The results presented in this study suggest that our open discovery algorithm is capable of uncovering implicit information that may form the basis of new hypotheses for research.

ACKNOWLEDGEMENTS

This research was partly accomplished while the first author was a visiting faculty scholar at the National Library of Medicine, Bethesda, MD. She thanks the University of Iowa for the Faculty Scholar Award and NLM for their hospitality and acknowledges NSF grant no. IIS-0312356, which partly funded this research.

REFERENCES

Blagosklonny,M.V. and Pardee, A.B. (2002) Unearthing the gems. *Nature*, **416**, 373.

Chaussabel,D. and Sher,A. (2002) Mining microarray expression data by literature profiling. *Genome Biol.*, **3**, RESEARCH0055.1–0055.16.

DiGiacomo,R.A., Kremer,J.M. and Shah,D.M. (1989) Fish oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. *Am. J. Med.*, **8**, 158–164.

Lindsay,R.K. and Gordon,M.D. (1999) Literature-based discovery by lexical statistics. *J. Am. Soc. Inf. Sci.*, **50**, 574–587.

Sehgal,A., Qiu,X.Y. and Srinivasan,P. (2003) Mining MEDLINE metadata to explore genes and their connections. In *Proceedings of SIGIR 2003 Workshop on Text Analysis and Search for Bioinformatics*.

Shatkay,H., Edwards,S., Wilbur,W.J. and Boguski,M. (2000) Genes, themes and microarrays. Using information retrieval for large-scale gene analysis. In *Proceedings of ISMB*, La Jolla, CA, AAI Press. pp. 317–328.

Smalheiser,N.R. and Swanson,D.R. (1996a) Indomethacin and Alzheimer's disease. *Neurology*, **46**, 583.

Smalheiser,N.R. and Swanson,D.R. (1996b) Linking estrogen to Alzheimer's disease: an informatics approach. *Neurology*, **47**, 809–810.

Smalheiser,N.R. and Swanson,D.R. (1998) Calcium-independent phospholipase A2 and Schizophrenia. *Arch. Gen. Psychiatry*, **55**, 752–753.

Srinivasan,P. (2004) Text mining: generating hypotheses from MEDLINE. *J. Am. Soc. Inf. Sci. Technol.*, **55**, 396–413.

Swanson,D.R. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Persp. Biol. Med.*, **30**, 7–18.

Swanson,D.R. (1988) Migraine and magnesium: eleven neglected connections. *Persp. Biol. Med.*, **31**, 526–557.

Swanson,D.R., Smalheiser,N.R. and Bookstein,A. (2001) Information discovery from complementary literatures: categorizing viruses as potential weapons. *J. Am. Soc. Inf. Sci. Technol.*, **52**, 797–812.

Weeber,M., Klein,H., Berg,L. and Vos,R. (2001) Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine–magnesium discoveries. *J. Am. Soc. Inf. Sci. Technol.*, **52**, 548–557.

Weeber,M., Vos,R., Klein,H., de Jong-Van den Berg,L.T.W., Aronson,A. and Molema,G. (2003) Generating hypotheses by discovering implicit associations in the literature: a case report for new potential therapeutic uses for Thalidomide. *J. Am. Med. Inform. Assoc.*, **10**, 252–259.