Mining MEDLINE for Implicit Links between Dietary Substances and Diseases*

Padmini Srinivasan[@] & Bisharah Libbus[+]

[@]School of Library and Information Science, University of Iowa, Iowa City, IA, 52242
padmini-srinivasan@uiowa.edu, 319-335-5708(W), 319-354-7553(F)
[+]Lister Hill Research Center, National Library of Medicine,Bethesda, MD, 20852
libbus@nlm.nih.gov, 301-594-8205

Abstract

This research presents our open discovery algorithm which is a text mining algorithm. We demonstrate that this algorithm may be used to uncover information that could form the basis of new hypotheses. In particular, we use it to discover novel uses for Curcumin Longa, a dietary substance, highly regarded for its therapeutic properties in Asia. Several diseases are identified as offering novel research contexts for curcumin. We analyze select suggestions: retinal diseases, Crohn's disease and disorders related to the spinal cord. Our analysis suggests that there is strong evidence in favour 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 open discovery algorithm may be used to find novel uses for dietary and pharmacologic substances. More generally, open discovery may be used to uncover information that potentially sheds new light on a given topic of interest.

Keywords: text mining, knowledge discovery, hypothesis generation, curcumin, turmeric.

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 labs. Serendipity may be influenced by several intangibles including researcher intuition, prior experience and knowledge, and the ability to creatively span multiple disciplines.

But what if there were 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 simultaneously analyze the published literature from multiple disciplines, sift through the evidence, identify implicit connections

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

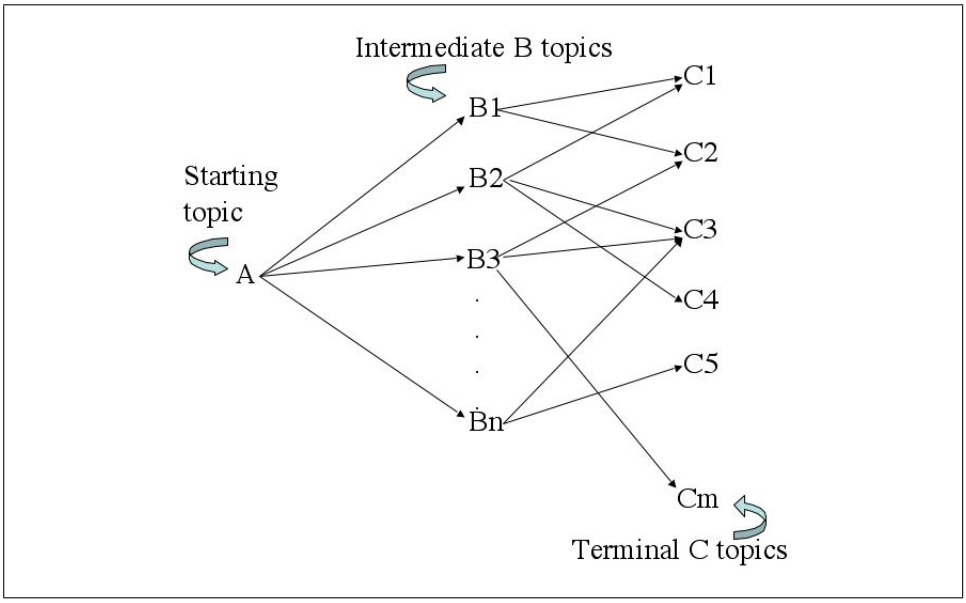

Figure 1: Open Discovery: General Model for Discovering Implicit Links between Topics.

that have potential, and then ensure that these are novel in 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 described [1].

Several successes may be attributed to text mining research. The pioneering work of Swanson and Smalheiser starting in the mid 80s resulted in several hypotheses that were later verified in clinical studies. For example, Swanson proposed that fish oils may be used to treat Raynauds disease [10] and this was later corroborated [3]. More recently they used their text mining approach to identify viruses that may be used as bioweapons [12]. Others have proposed for example, therapeutic uses for thalidomide [14] and found functional connections between genes [2, 5].

Our text mining algorithms follow the discovery framework set by Swanson and Smalheiser. In essence they proposed an *open discovery* approach which is depicted in Figure 1. This process is initiated with a single topic (A) of any type such as a disease, a pharmacologic substance or a gene. Starting with the A topic and navigating through intermediate topics (B1, B2, ...) the goal is to reach terminal topics (C1, C2 etc.) that shed new light on A. Swanson used this discovery processes to find the Raynauds 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, ...) between them [6, 7, 8, 11] . 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 [4, 13, 9]. In our replication of their eight open and closed discoveries, algorithms we designed were the most effective while requiring the least amount of manual input and analyses [9].

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 (*Curcumin Longa*) a dietary substance commonly used in Asia. Our algorithm identifies several diseases or disorders that could be 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.

SYSTEM AND METHODS

Open Discovery Algorithm

We implement Swanson's open discovery process using *topic profiles*. Topics may be simple (eg. Tylenol) or more complex such as Calcium channel blockers and Alzheimers disease. A topic profile is a set of terms (single words and phrases) representing the topic. We first retrieve a set of relevant documents for the topic and then extract terms from them. Term weights are calculated to indicate relative importance in representing the topic. When applying open discovery to MEDLINE, terms extracted are MeSH (Medical Subject Headings) metadata terms which 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 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 may restrict the discovery process to certain semantic types. The topic profiles are then restricted to MeSH terms belonging to those 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 $log(N/TF_i)$. 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.

$$weight(t_i) = v_i / \sqrt{v_1^2 + v_1^2 + ... + v_r^2}, \quad (1)$$

where $v_i = TF_i * log(N/TF_i)$ and there are r terms in the profile.

We outline our open discovery algorithm below. The algorithm is shaped by semantic type constraints. Also, larger M values for example, broaden the pipeline through which C topics are selected. 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.

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.

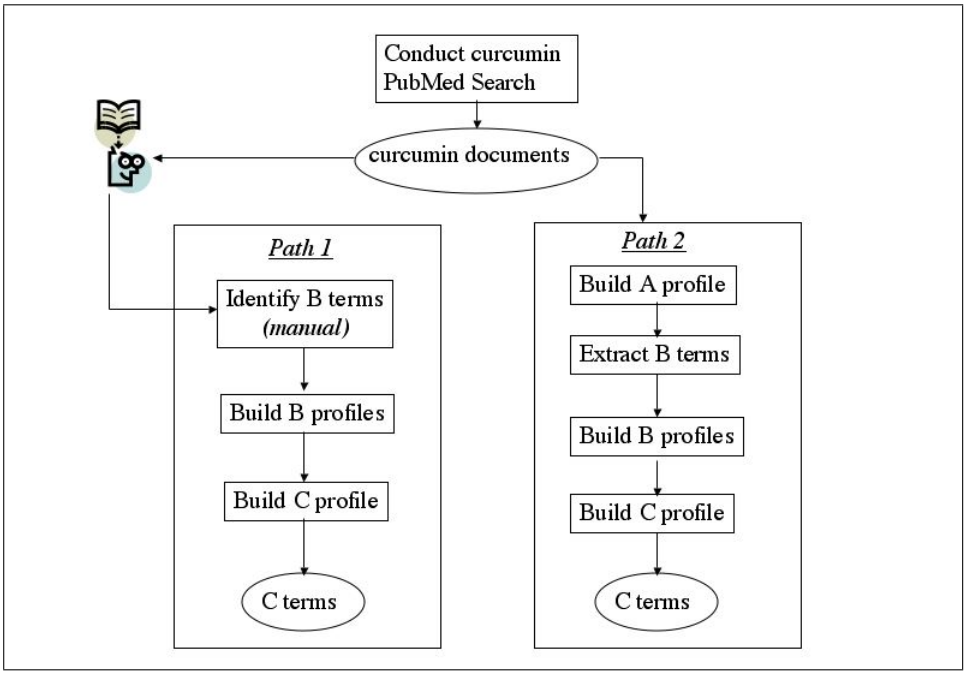

Figure 2: Parallel Experiments involving (1) Automatically and (2) Manually Selected B Terms.

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 *Disease or Syndrome* and *Neoplastic Process*. *Neoplastic Process* includes MeSH terms referring to cancers. We experimented with M values set to 5, 10 and 15. However as seen later, 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 the two paths differ.

RESULTS

A total of 1,175 PubMed documents were retrieved from the curcumin search. The majority of these publications (1,043, 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.

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

Table 1: Automatically Identified B Terms. (AAPP: Amino Acid, Peptide or Protein semantic type).

B Terms

Table 1 shows the top 10 MeSH terms that were automatically selected (step 2, M=10). 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 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, that may be accessed directly at NLM's pubmed website. Curcumin inhibits NF-kappaB (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. Interestingly, ten 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. If the comparison were relaxed to consider for example, the fact that JNK and ERK are MAPK members, the scores would increase.

C Terms

Table 3 presents the top ten suggested disease C terms. Columns labeled A10 and M indicate C term ranks using automatically (with M=10) and manually identified B terms respectively. For example, *Crohn Disease* is ranked 5 in the automatic method and 7 in the manual one. Observe that in step 5 of the algorithm we check for novelty of the C term w.r.t 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 parenthesis indicate ranks before manual refinement. For example, *Ischemic Attack Transient* was initially ranked 5 in the automatic method (not shown in table). However, a search

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

Table 2: Manually Identified B MeSH Terms. Italicized terms were also identified automatically.

on curcumin AND ischemia retrieved many documents and hence this was removed from the list of novel C terms for A. Interestingly the top 3 suggestions after refinement are identical for automatic and manual methods. Also there is an overlap of 43% in top 10 C terms from both methods. Recall and precision for the automatic method judged against the manual results for these 10 top listings are both at 60%. Interestingly the automatic algorithm run with M=5 or M=15 does not change these recall and precision scores. We only observe slight changes in the relative rankings of C terms with the top 3 suggestions remaining the same.

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 could be beneficial or harmful. Next we present such an analysis (conducted by the second author) for entries 'Retina', 'Spinal Cord', and 'Crohn Disease', i.e., for associated diseases/disorders. The first two are the top two entries while Crohn's is analyzed as it is the last ranked manual entry that also appears in the automatic list. In each case the goal in analysis is to identify biochemical pathways potentially connecting the disease and curcumin.

ANALYSIS

Retinal Diseases

Retinal diseases could result from complications due 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). 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-1beta (IL-1 beta) and cyclooxygenase 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).

Disease	A10	M	Disease	A10	M
Retina	1 (1)	1 (1)	Cystic Fibrosis	8 (33)	
Spinal Cord	2(2)	2 (8)	Epilepsy	9 (35)	
Cytomegalovirus	3 (18)	3 (10)	Uremia	10 (36)	
Amyotrophic Lateral Sclerosis	4 (25)		Choriocarcinoma		6 (22)
Crohn Disease	5 (26)	7 (32)	Sarcoma Kaposi		8 (34)
Lupus Erythematosus Systemic	6 (27)	5 (19)	Graves Disease		9 (39)
Hodgkin Disease	7 (29)	4 (17)	Sjorgens Syndrome		10 (42)

Table 3: Novel C diseases terms. A10: Automatic with parameter $M = 10$, M: Manual.

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

The family of mitogen-activated protein kinases (MAPK) is another group of genes that has an important role in retinal disease. These include extracellular signal-regulated kinases (ERK), c-Jun amino(N)-terminal kinase (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-alpha (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-1beta activation, which is induced by the presence of retinal holes, a key feature of diabetic retinopathy (12824248). NF-kappaB, whose level changes as an early response to inflammation, also stimulates these MAPK genes (12878172). Moreover, activation of TNF-alpha was followed by increased transcription of NF-kappaB (12878172). Also activation of NF-kappaB 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-1 beta, 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-kappaB activation (11753638, 11506818, 12878172, 12825130). Having shown that these genes, in particular, IL-1beta, COX-2, TNF-alpha, JNK, ERK, NF-kappaB, 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-kappaB, a pro-inflammatory molecule that confers susceptibility to Crohn's disease (12840668, 12676561,12626759, 12527755, 12673278).

Additionally, Crohn disease and ulcerative colitis, both IBDs (Irritable Bowel Disease) 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 IFN-gamma (IFN-gamma) on the one hand and the anti-inflammatories IL-4 (interleukin-4) and transforming growth factor (TGF-beta) activity, is believed to control chronic intestinal inflammation (11994418). Restoring TGF-beta1 signaling in chronic IBD, by inhibition of Smad7, results in the TGF-beta1 induced inhibition of cytokine production (11518734). Intestinal cells from patients with Crohn's disease produced IL-18, a pleiotropic cytokine that augments IFN-gamma 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-alpha RNA (11570528). The regulation of TNF-alpha, a key mediator in the inflammatory process, is interconnected with MAPK pathways in IBD: the p38alpha, 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-kappaB activation (9616307) as also inhibition of IL-1beta, IL-6, IL-8 and TNF-alpha production (9468102, 12005259) are shown to be beneficial.

Curcumin inhibits a number of these genes and cytokines. For example it inhibits NF-kappaB activation and in turn suppresses TNF-alpha 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-beta: in wounds, which involve inflammation, the observed beneficial effect of curcumin treatment was attributed to an increase in TGF-beta (9776860). Curcumin significantly inhibited IL-12 leading to decreased IFN-gamma 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-beta and LT-beta production increased and was followed 18 h later by TGF-beta1 upregulation (12127673, 12828562). Once again TNF-alpha 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 upregulated 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-alpha, (14593216, 12165135). Similar patient observations were made for IL-1alpha, IL-1beta, in addition to TNF-alpha and IL-6 levels while blocking of IL-1 and TNF-alpha receptors significantly reduced their expression (12111861).

After nerve injury, COX-2 is upregulated 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 postoperative 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 autoimmune encephalomyelitis (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-gamma and TNF-alpha 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-gamma and TNF-alpha, and upregulated IL-4, IL-10 and TGF-beta 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-gamma and TNF-alpha, to GPBP/IFA-immunized animals (11385625). Interestingly it has been observed that, CNS-confined inflammation induced by IFN-gamma may induce protective immunological countermechanisms in EAE/multiple sclerosis (11466408). The profile of TNF-alpha 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-alpha 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-alpha, IL-12, IFN-gamma, 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.

CONCLUSION

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. The connections are based primarily on genes such as TNF-alpha, MAPK, NF-kappaB, COX-2 and other cytokines and interleukins. We also compared two versions of the open discovery process. One where B terms were automatically selected and the other where 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 10 C terms finally suggested by the automatic method when judged against the manual method's output were both 60%. Overlap between the two C sets was 43%. 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 pharmacologic 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.

References

[1] Blagosklonny, M. V., & Pardee, A. B. Unearthing the gems. (2002). Nature, 41 6, 373.

[2] Chaussabel D. & Sher A. (2002). Mining microarray expression data by literature profiling. Genome Biology, 3(10):research0055.1-0055.16.

[3] DiGiacomo R.A, Kremer J.M, & Shah D.M. (1989). Fish oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study. American Journal of Medicine, 8, 158-164.

[4] Lindsay, R.K, & Gordon, M.D. (1999). Literature-based discovery by lexical statistics. JASIS, 50(7), 574-587.

[5] Shatkay, H., Edwards, S., Wilbur, W.J., & Boguski, M. (2000). Genes, Themes and Microarrays. Using information retrieval for large-scale gene analysis. ISMB, La Jolla, CA, 317-328.

[6] Smalheiser, N.R., & Swanson, D.R. (1996a). Indomethacin and Alzheimer's disease. Neurology, 46:583.

[7] Smalheiser, N.R., & Swanson, D.R. (1996b). Linking estrogen to Alzheimer's disease: An informatics approach. Neurology, 47, 809-810.

[8] Smalheiser, N.R, & Swanson, D.R. (1998). Calcium-independent phospholipase A2 and Schizophrenia. Archives of General Psychiatry. 55(8), 752-753.

[9] Srinivasan, P. Text Mining: Generating Hypotheses from MEDLINE. JASIST. To appear.

[10] Swanson, DR. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30, 7-18.

[11] Swanson, D.R. (1988). Migraine and Magnesium: Eleven neglected connections. Perspectives in Biology and Medicine, 31, 526-557.

[12] Swanson, D.R., Smalheiser, N.R., & Bookstein, A. (2001). Information discovery from complementary literatures: categorizing viruses as potential weapons. JASIS, 52(10), 797-812.

[13] Weeber, M., Klein, H., Berg, L., & Vos, R. (2001). Using concepts in literature-based discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium discoveries. JASIST, 52(7), 548-557.

[14] Weeber, M., Vos, R., Klein, H., de Jong-Van den Berg, L.T.W., Aronson, A & Molema, G. (2003). Generating hypotheses by discovering implicit associations in the literature: A case report for new potential therapeutic uses for Thalidomide. JAMIA, 10(3), 252-259.