HLT-NAACL 2004 Workshop: Biolink 2004,
Linking Biological Literature, Ontologies and Databases, pp. 33-40.
Association for Computational Linguistics.

Mining MEDLINE: Postulating a Beneficial Role for *Curcumin Longa* in Retinal Diseases

Padmini Srinivasan
School of Library
& Information Science
The University of Iowa
Iowa City, IA 52242
padmini-srinivasan@uiowa.edu

Bisharah Libbus
National Library of Medicine
Bethesda, MD 20852
libbus@nlm.nih.gov

Aditya Kumar Sehgal
Computer Science
The University of Iowa
Iowa City, IA 52242
sehgal@cs.uiowa.edu

Abstract

Text mining tools are designed to assist users with the important step of hypothesis generation. In this research we apply an open discovery process to the problem of identifying novel disease or problem contexts in which a substance may have therapeutic potential. We illustrate this discovery process by executing our open discovery algorithm with turmeric (*Curcumin Longa*) as the substance being investigated. The top ranking entry suggested by the algorithm is retinal diseases. Further analysis of the literature yields evidence supporting the suggested connection between curcumin and retinal diseases. In particular, curcumin influences the activation of genes such as COX-2, TNF-alpha, JNK, ERK and NF-kappaB. These genes are in turn involved in retinal diseases such as diabetic retinopathies, ocular inflammation and glaucoma. Moreover, the evidence suggests that curcumin may have a beneficial and therapeutic role in the context of these diseases.

1 Introduction

Consider a bioscientist who is studying a particular disease. Assume that she is already well familiar with the pathophysiology and accepted therapeutic options for treating this condition and wishes to determine if there are other, yet unrecognized, substances that may have therapeutic potential. She begins by searching for documents on the disease mechanism(s) and related disorders. Very soon she finds herself immersed in a morass of pathways and possible directions that need to be further explored. It will come as no surprise if even our most determined user quickly becomes overwhelmed and discouraged. The challenge of searching for a novel therapeutic substance is at best like looking for the proverbial "needle in a haystack". However, in reality the challenge is greater since there is no assurance that there indeed is a needle in the haystack. Consequently, the goal of text mining (also known as literature mining) systems and algorithms is to assist users find such needles, if these exist at all in the literature "haystacks" (Hearst 1999).

In general, as shown in Figure 1, a user may start with any type of topic (A), be it a disease, a pharmacological substance, or a specific gene. As he navigates the literature and follows connections through appropriate intermediate topics (B1, B2 etc.), the user hopes to reach terminal topics (C1, C2 etc.) that are both relevant and novel, in the sense of shedding new information on topic (A). This text mining approach commonly referred to as 'open' discovery was pioneered by Swanson in the mid 80s. A classic example discovery is one where starting with Raynaud's disease (A) Swanson identified fish oils (C) as a substance that may have therapeutic potential (Swanson, 1986). Intermediate connections (B) such as 'blood viscosity', 'platelet aggregation' were observed. Swanson also proposed a variation called 'closed' discovery wherein starting with a pair of topics (A and C) one explores possible connections (B links) between them that are not yet recognized. In collaboration with Smalheiser, Swanson used his open and closed discovery methods on MEDLINE and proposed a number of hypotheses (eg. Swanson, 1990; Smalheiser & Swanson1996a; Smalheiser & Swanson1996b; Smalheiser & Swanson1998). The hypotheses they proposed were subsequently corroborated in clinical studies.

The text mining framework established by Swanson and Smalheiser has attracted the attention of several researchers (Gordon and Lindsay, 1996; Lindsay and Gordon, 1999; Weeber et al., 2001) besides us (Srinivasan, 2004). A key goal in these follow-up efforts has been to reduce the amount of manual effort and intervention required during the discovery process. In previous work

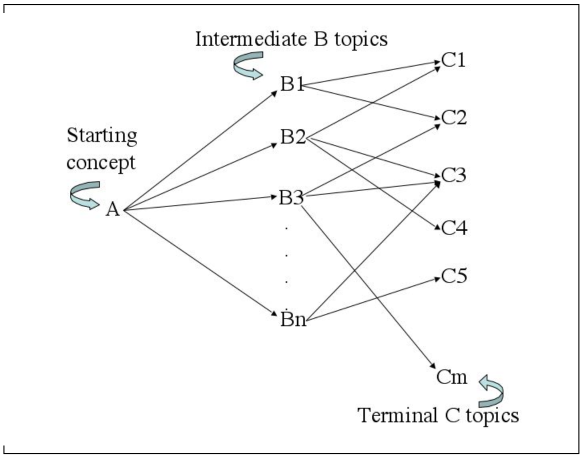

Figure 1: Open Discovery: The General Model

using algorithms for MEDLINE which we developed, we replicated the eight open and closed discoveries made by Swanson and Smalheiser. In comparison with other replication studies these algorithms were the most effective (Srinivasan, 2004). They also require the least amount of manual input and analyses. For example, in open discovery, our methods expect the user to specify only the *type* of B terms of interest. Following this our algorithm selects B terms automatically. In contrast the other methods rely more on user input for selecting B terms. Our current research demonstrates that our open discovery algorithm can be used to generate *new* hypotheses for disease treatment that could be tested. In particular, we apply our open discovery procedure to explore the therapeutic potential of curcumin/turmeric (*Curcumin Longa*) a dietary substance commonly used in Asia. We show that our automatic discovery algorithm identifies *retinal diseases* as the novel context for research on curcumin. We review genetic and biochemical evidence to indicate that curcumin may be beneficial for treating retinal diseases.

We first describe our open discovery algorithm. Next we show its application with *curcumin* as the starting point (topic A). We then present an analysis of the curcumin - retinal diseases connection. The next section is on related research. The final section presents our conclusions and plans for the next phase of this research.

2 Open Discovery

Our open discovery approach is founded on the notion of topic profiles. A topic is any subject of interest such as *treatment of hypertension* or *ATM gene*. A profile is essentially a representation of a topic that is derived from the text collection being mined. For MEDLINE our topic profiles are vectors of weighted Medical Subject Headings (MeSH). These terms belong to a controlled vocabulary and are manually assigned to each MEDLINE document by trained indexers. Given a topic of interest, our algorithm first retrieves relevant MEDLINE documents. MeSH terms are then extracted from these documents and their weights are calculated. These weighted terms form the profile vector for the topic. We discuss the method for calculating weights shortly.

We also exploit the fact that MeSH terms have been classified using 134 UMLS (Unified Medical Language System)[1] semantic types as for example *Cell Function*, *Sign or Symptom*. 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. More generally, semantic types represent ‘categories’ that have been used to classify the MeSH metadata. Semantic types are useful because depending on the nature of the discovery goals we may adopt a particular *view*, i.e., we may restrict the discovery process to consider only MeSH terms that belong to certain semantic types. In these cases the topic profiles are restricted to MeSH terms belonging to semantic types specified by the view.

We calculate term weights for the MeSH terms. Term weights are a slight modification of the commonly used TF*IDF scores. Since a MeSH term typically occurs once in a MEDLINE record, here TF_i (term frequency) equals the number of documents in which the MeSH term t_i occurs within the retrieved document set. 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.

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

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

Algorithm: Figure 2 outlines our open discovery algorithm which follows the framework shown in Figure 1. We begin by building the A topic profile restricted to ST-B semantic types. Note that all MEDLINE searches are conducted automatically via the PubMed interface[2]. We then automatically select M MeSH terms for each ST-B semantic type from this A profile and call these the B terms. Next profiles are built for each of these B terms limited to another selected set of semantic types ST-C. These B profiles are analysed in combination to select an initial pool of candidate C terms. These candidate terms are then checked for novelty in the context of the starting A topic. When the algorithm terminates the user is provided a final list of ranked, novel C terms. The higher the rank the greater

[1] *http://umlsks.nlm.nih.gov*
[2] http://www.nlm.nih.gov

the estimated confidence in the potential connection with the A topic.

At this point the rest of the process depends almost entirely on the user. (This is also the case in other implementations of the open discovery process (eg. Lindsay & Gordon 1999; Weeber et al., 2001)). It is up to the user to select A - C pairs of interest and explore the literature for supporting evidence.

The role of ST-B and ST-C in the algorithm is to apply reasonable 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. But we already see some patterns emerge in the MEDLINE mining literature. For example when looking for substances likely to influence a disease several researchers have used functional semantic types such as *Cell Function* and *Molecular Dysfunction* for selecting intermediate pathways (eg. Weeber et al., 2001). Experiments varying these semantic types have been described in our previous work (Srinivasan, 2004). Unique aspects of our algorithm in comparison to open discovery methods explored by others, include for example, the fact that our weighting scheme identifies interesting and relevant B terms at high ranks. Also, C terms are assessed by combining the evidence on their connection to the different intermediate B terms.

3 Open Discovery with Turmeric

Our interest in curcumin was sparked by the fact that this spice is widely used in Asia and is highly regarded for its curative and analgesic properties. These include the treatment of burns, stomach ulcers and ailments, and for various skin diseases. Curcumin is also used as an antiseptic, in alleviating symptoms of the common cold as well as a depilatory. A number of MEDLINE records have reported on the anti-cancer and anti-inflammatory properties of curcumin (12680238, 12678737, 12676044[3]). Our open discovery goal is aimed at determining whether there are novel disease contexts in which curcumin could prove beneficial, and to propose evidence-based hypotheses that can be experimentally verified.

Input from user: (1) an A topic of interest, (2) a set of UMLS semantic types (ST-B) for selecting B terms and a set (ST-C) for selecting C terms.
Parameter: M

- Step 1: Conduct an appropriate PubMed search for topic A, and build its MeSH profile limited to the semantic types in ST-B. Call this profile AP.

- Step 2: For each semantic type in ST-B, select the M top ranking MeSH terms from AP. Remove duplicate terms if any. These are designated the B terms (B1, B2, B3, etc.).

- Step 3: Conduct an independent PubMed search for each B term and build its profile limited to the semantic types ST-C. Call these profiles BP1, BP2, BP3, etc.

- Step 4: Compute a final combined profile where the combined weight of a MeSH term is the sum of its weights in BP1, BP2, BP3, etc. Call this initial profile CP.

- Step 5: For each term t in CP if a MEDLINE search on topic A AND t returns non zero results, eliminate t from CP.

Output: For each semantic type in ST-C, output the MeSH terms in CP ranked by combined weight. These are the C terms organized by semantic type and ranked by estimated potential.

Figure 2: Open Discovery Algorithm: Outline of Steps.

We executed our open discovery algorithm with *curcumin* as the starting topic (A). The specific PubMed search conducted was *turmeric OR curcumin OR curcuma* (done on November 15, 2003). A total of 1,175 PubMed documents were retrieved. As Figure 3 shows the majority of these publications (1,043, 89%) are rela-

[3] Numbers within parantheses such as these refer to PubMed record ids. The reader may enter these directly into the PubMed interface to retrieve the corresponding records.

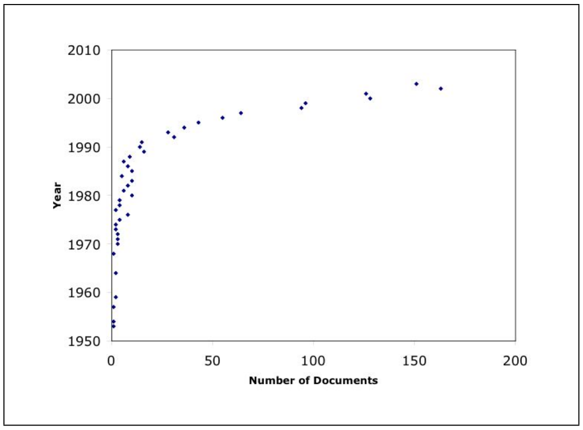

Figure 3: Temporal Distribution of Curcumin Documents

tively recent, being published in 1990 or later. This indicates a surge in scientific interest in the health effects of this spice, which has long been valued in Asia for its medicinal properties.

We limited ST-B to the three semantic types *Gene or Genome*; *Enzyme*; and *Amino Acid, Peptide or Protein*. We restricted ST-C to *Disease or Syndrome* and *Neoplastic Process*[4] and set M (the parameter specifying the number of B terms to select) to 10. These semantic types are appropriate since we are looking for biochemical and genetic connections between turmeric and novel diseases.

Table 1 shows the top 10 selected MeSH terms from each ST-B type (step 2). We can observe from the table that some of the terms appear in more than one semantic type. (This is possible since a term may be assigned to more than 1 semantic type in the UMLS). However, we remove duplicates in step 2. Also, some terms are very specific such as *Protein Kinase C* while others are broad representing families such as *DNA-Binding Proteins* and *Isoenzymes*. At present we do not distinguish between B terms using specificity. Our plan is to examine this aspect in future research.

The B terms listed in Table 1 are the top ten terms that were retrieved from a search of the literature for the semantic types Genes or Genomes, Enzymes, and Amino Acid, Peptide or Proteins. The biochemical effects of curcumin become apparent upon conducting a search of the literature for curcumin and any of these terms. Curcumin, for example, has a strong down-regulatory effect on c-Jun NH2-terminal kinase (JNK) (14627502, 12859962, 11370761, 12097302) resulting in the arrest of cell proliferation (14627502) in prostate tumor cells (12853969) and induction of apoptosis (12859962). Curcumin inhibits NF-kappaB (12714587) leading to the suppression of cell proliferation and the induction of apoptosis

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

Table 1: Intermediate B Terms. G: Gene or Genome, E: Enzyme, AAPP: Amino Acid, Peptide or Protein. Numbers indicate ranks.

[4] Neoplastic Process includes MeSH terms referring to cancers.

Terminal C MeSH Terms	Rank
Retina	1
Spinal cord	2
Testes	3
Thyroid Neoplasms	4
Ischemic Attack Transient	5

Table 2: Top Five Novel C Terms.

in multiple myeloma (12393461) and ovary cancer cells (12520734). TGF-beta1 induced IL-6 which has been implicated in the malignant progression of prostate cancers was severely impeded by curcumin through inhibition of c-Jun (matches with Genes, jun in the table) JNK (an instance of MAPK in the table) or AP-1 (12853969).

The curcumin open discovery process terminated with a ranked list of diseases. Table 2 shows the top 5 entries[5]. One observation made at this point was that the type of automated search conducted in step 5 of the algorithm to check for novelty is insufficient. At present, the search involves only the particular MeSH term intersected with the A topic. We do not yet automatically consider synonyms of the MeSH term. For example for the last entry in the table, although *Ischemic Attack Transient AND (turmeric OR curcumin OR curcuma)* retrieved 0 documents, the search *Ischemia AND (turmeric OR curcumin OR curcuma)* retrieves 17 documents. Hence this entry is unlikely to be immediately interesting to the user. However, the top two entries did not retrieve any document even after searching with different synonyms. Testes is also unlikely to be interesting since a curcumin search intersected with sperm retrieved many documents. Considering retrieval set size alone is insufficient. For instance, curcumin intersected with thyroid retrieved 5 documents. However, these appear to be peripheral to curcumin's effect on thyroid neoplasms focusing more on aspects such as hypothyroidism and toxicity. Automating query expansion using synonyms will be the subject of further research.

At this point the user may select entries and peruse the appropriate literature further to (a) determine the *nature* of the relationship between curcumin and the diseases (as the substance under study could be beneficial or harmful) and (b) assess the quality of the background knowledge that may be used to guide further study of curcumin and the disease. This manual phase may be guided by the specific B term-based pathways connecting the selected disease with curcumin. Table 3 lists the B terms that were automatically identified as connecting curcumin and 'Retina'.

MeSH Term
Genes, fos
Genes, Reporter
Genes, Dominant
MAPK
Glutathione Transferase
Protein Kinase C
Isoenzymes
Protein-Tyrosine Kinase
Caspases
Nitric-Oxide Synthase
Ornithine Decarboxylase
Proto-Oncogene Proteins c-jun
Tumor Necrosis Factor
Glutathione
DNA-Binding Proteins

Table 3: B Terms Connecting Turmeric and Retina.

In the next section we present such an analysis for 'Retina'. That is, we (the second author) examine the literature to determine if retinal diseases may be a good context in which a bioscientist may study curcumin. Our analysis indicates that indeed there is good evidence supporting the hypothesis of a beneficial role for turmeric in the context of diabetic retinopathies, ocular inflammation and glaucoma. Analysis of the other highly-ranked diseases is left for future work.

4 Turmeric - Retinal Diseases Connection

The procedure followed up to this point is 'term-centric'. That is, we automatically identify statistically interesting B terms and then generate a ranked list of C terms. We now present further analysis on the connection between retinal diseases and curcumin. In some cases reading the title and abstracts of select records provided sufficient information. In addition the full text of the document was available. Our strategy was to examine publications for biochemical or molecular biology mechanisms. In particular, we were interested in ascertaining whether any of the genes noted earlier were also involved in the pathophysiology of these retinal disorders. We focused on the genes as the critical links that connect the agent curcumin to the disorders.

Analysis: The user's goal is to identify biochemical pathways potentially connecting retinal diseases and curcumin. Retinal diseases could result from complications due to diabetes, or of infection and inflammation of the retina.

[5] Although the main semantic type for a term such as *Spinal Cord* is *Body Part, Organ, or Organ Component*, in the UMLS *Spinal Cord* is listed for at least one vocabulary as a synonym for *Spinal Cord Diseases*. It is thus also assigned the semantic type of *Disease or Syndrome*. Similar observations hold for terms *Retina* and *Testes*

Diabetic retinopathy is a leading cause of blindness. An early sign of the disease is the adhesion of leukocytes to the vessels of the retina, endothelial cell injury, and the breakdown of the blood-retina barrier (12000720). Even acute intensive insulin therapy constitutes an additional risk factor for diabetic retinopathy, due to insulin-induced hypoxia and an associated acceleration in the blood-retina barrier breakdown (11901189). Glaucoma is the second most common cause of blindness in the world (8695555) and 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 several diseases have one or a few genetic loci that 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 the process of tissue injury or inflammation and regulation of cell division. Control of the immune process and of the inflammatory response is important in combating infection and autoimmune diseases. Regulation of cell division, particularly programmed cell death, is critical in diverse diseases such as cancer and tissue regeneration, e.g. retinal injury and diseases. Regulation of the activity of 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). Similarly, the expression of COX-2 was associated with the development of glaucoma (9441697). Treatment with COX-2 inhibitors suppressed blood-retinal barrier breakdown and had an antiangiogenic effect, i.e. they 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). Anti-TNF-alpha treatment reduced leukocyte adhesion to blood vessels of the eye and vascular leakage (12714660) indicating a potential therapeutic effect for such a treatment to reduce ocular inflammation. Activation of TNF-alpha and other genes may also lead to the pathophysiology of glaucoma (10975909, 10815159).

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. One of these, ERK, was induced in glaucoma (12824248). Often inflammatory responses include the induction of apoptosis, or programmed cell death. The involvement of JNK in inducing apoptosis was demonstrated in prostate cancer (12859962, 12663665) and retinal cells (12270637). There is also a link to TNF-alpha (discussed above) which was shown to activate phosphorylation of ERKs, p38, and JNK MAPK in human chondrocytes (12878172).

IL-1beta activation, induced by the presence of retinal holes, a key feature of diabetic retinopathy, is also reported to result in the activation of a number of the MAPK genes ERK, JNK, and p38 (12824248). These conditions in turn exacerbate the disease process in that they result in proliferative and migratory cells accumulating in the wounded retina (12500176). Inhibitors of MAPK and phosphatidylinositol 3-kinase (PI3) inhibited retinal pigment epithelial cell proliferation (12782163). The breakdown in the blood-retina barrier is also suppressed by inhibitors of p38 MAPK and PI3 (11901189).

Changes in the levels of the gene NF-kappaB is an early cellular response to inflammation. Activation of TNF-alpha (discussed above) is followed by increased transcription of NF-kappaB which in turn stimulates ERK, p38, and JNK MAPK (12878172). Also activation of NF-kappaB subsequently stimulated COX-2 and matrix metalloproteinase-9 expression (12807725).

Curcumin was shown to be effective in inhibiting cell proliferation of tumorigenic and non-tumorigenic breast cancer cells (12527329) and other tumor cells (12680238). As described previously the gene COX-2 is involved in early inflammatory diabetic retinopathy (11821258). Curcumin was able to suppress COX-2 in a dose-related manner (12844482) and neutralized the effect of IL-1 beta, possibly through its effect on p38 and COX-2 and JNK (12957788). Curcumin is also a known inhibitor of JNK (12957788,12854631,12582006, 12130649, 12105223, 9674701) and a suppresser of NF-kappaB activation (11753638, 11506818, 12878172, 12825130). For example, it suppressed the induction of NF-kappaB and its dependent genes by cigarette smoke (12807725), in alcoholic liver disease (12388178) and in cultured endothelial cells (12368225).

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.

Our approach has focused on specific genes, in particular to provide clues regarding the relevant biochemical pathways. In some cases the evidence is gathered in the context of other diseases such as alcoholic liver disease with the idea that similar evidence may be found for retinal diseases. In summary it seems likely that curcumin, taken in the diet or applied topically, could prove beneficial in cases of diabetic retinopathies, retinal injury, ocular inflammation and glaucoma.

5 Related Research

Text mining, i.e., uncovering information that may lead to hypotheses, has attracted the attention of many researchers (eg. Andrade & Valencia, 1998; Gordon & Lindsay, 1996; Masys et al., 2001; Smalheiser & Swanson, 1996a; Smalheiser & Swanson, 1996b; Srinivasan & Wedemeyer 2003; Srinivasan, 2004; Swanson, 1986; Swanson, 1988; Swanson et al., 2001; Weeber, 2000). Examples of recent text mining applications include automatically identifying viruses that may be used as bioweapons (Swanson et al., 2001), proposing therapeutic uses for thalidomide (Weeber, 2003) and finding functional connections between genes (Chaussabel & Sher, 2002; Shatkay et al., 2000).

A major emphasis in text mining research has been to directly exploit co-occurrence relationships in MEDLINE. For example, Jenssen et al., (2001) generate a co-occurrence based gene network called PubGene from MEDLINE for $13,712$ named human genes. Each of PubGene's 139,756 links is weighted by the number of times the genes co-occur. Wilkinson and Huberman[6] identify communities of genes. Starting with a co-occurrence based gene network for a particular disease domain, communities are identified by repeatedly removing edges of highest betweeness (number of shortest paths traversing the edge). Applying this to the domain of colorectal cancer, they are able to identify interesting hypotheses linking genes that were for example, in the same community but had no edge between them.

Our research is based on the open discovery framework proposed by Swanson. As indicated before, Swanson and Smalheiser made several discoveries using their open and closed discovery methods (Swanson, 1986; Swanson, 1988; Swanson et al., 2001; Smalheiser & Swanson, 1996a; Smalheiser & Swanson, 1996b), that were later validated by bioscientists. These discoveries together offer a testbed of examples that are being used by other researchers to develop their own discovery algorithms (Gordon & Lindsay, 1996; Lindsay & Gordon, 1999; Srinivasan, 2004; Weeber et al., 2001).

One characteristic that may be useful in distinguishing between text mining efforts is the extent to which they are problem or sub domain specific. For example, PubGene is directly targeted towards bioinformatics researchers. In contrast, implementations such as ours that derive from the open discovery framework are not problem specific. These may be used for a variety of goals, as for example by geneticists involved in understanding the results of microarray experiments and by epidemiologists searching for links between viruses and specific populations. We believe that the next generation of text mining systems will be judged not only by their effectiveness but also by their flexibility in application.

6 Conclusions

We applied our implementation of Swanson's open discovery algorithm to the problem of identifying novel disease or problem contexts in which substances might have a therapeutic role. We used our methods to investigate the potential of turmeric or *Curcumin Longa*. Our analysis identifies a ranked list of problems for which treatment with curcumin may be beneficial with the top ranked entry pointing to retinal diseases. Guided by our algorithm, further analysis of the literature by our expert user (a geneticist) yielded good evidence in support of the hypothesis that curcumin, taken in the diet or applied topically, could prove beneficial in cases of diabetic retinopathies, ocular inflammation and glaucoma.

In future work we will analyze the other suggestions made by our open discovery methods. For example, the second suggestion is problems related to the spinal cord. The analysis will again focus on genetic mechanisms that could potentially connect curcumin with the problems. We will also explore methods to automate query expansion for the search in step 5 of the algorithm. One limitation of the discovery process concerns the evidence gathering phase when analyzing individual C terms. This process is manual and involves significant investment of time and intellect toward sifting through the literature and collecting evidence relevant to the hypothesized connections. In the next phase of our work we plan to study methods to assist in this phase.

Acknowledgments

This research was partly accomplished while the first author was a visiting faculty scholar at the National Library of Medicine, Bethesda, Maryland. She thanks the University of Iowa for the Faculty Scholar Award and NLM for their hospitality. This work was also funded in part by NSF grant no. IIS-0312356 awarded to the first author. The authors thank the anonymous reviewers for their insightful comments.

References

Andrade A, & Valencia A. 1998. Automatic extraction of keywords from scientific text: application to the

[6] Wilkinson, D., & Huberman, B. A. A method for finding communities of related genes. http://citeseer.nj.nec.com/546592.html.

knowledge domain of protein families. *Bioinformatics*, 14(7):600-607.

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

Gordon M.D & Lindsay R.K. 1996. Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *Journal of the American Society for Information Science*, 47:116-128.

Hearst M. Untangling text data mining. (1999). In: Proceedings of ACL, Annual Meeting of the Association for Computational Linguistics (invited talk), University of Maryland, Maryland, June 20-26, 1999.

Jenssen, T-K., Laegreid, A., Komorowski, J., & Hovig, E. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28:21-28.

Lindsay, R.K, & Gordon, M.D. (1999). Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science*, 50(7):574-587.

Masys, D.R., Welsh, J.B,, Fink, J.L., Gribskov, M., Klacansky, I., & Corbeil, J. 2001. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17(4):319-326.

Shatkay, H., Edwards, S., Wilbur, W.J., & Boguski, M. 2000. Genes, Themes and Microarrays. Using information retrieval for large-scale gene analysis. In: Proceedings of Intelligent Systems for Molecular Biology, La Jolla, California, 317-328.

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

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

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

Srinivasan, P. To appear 2004. Text Mining: Generating Hypotheses from MEDLINE. *Journal of the American Society for Information Science*.

Srinivasan, P., & Wedemeyer, M. (2003). Mining Concept Profiles with the Vector Model or Where on Earth are Diseases being Studied? In: Proceedings of Text Mining Workshop. Third SIAM International Conference on Data Mining. San Francisco, CA.

Swanson, DR. 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30: 7-18.

Swanson, D.R. 1988. Migraine and Magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, 31:526-557.

Swanson, D.R. (1990). Somatomedin C and Arginine: Implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine*, 33 (2):157-179.

Swanson, D.R., Smalheiser, N.R., & Bookstein, A. (2001). Information discovery from complementary literatures: categorizing viruses as potential weapons. *Journal of the American Society for Information Science*, 52(10): 797-812.

Weeber, M., Klein, H., Aronson, A.R., Mork, J.G., Jong-van den Berg, L., & Vos, R. (2000). Text-based discovery in biomedicine: the architecture of the DAD-system. In: Proceedings of AMIA, the Annual Conference of the American Medical Informatics Association, November 4-8, 2000, 903-907.

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. *Journal of the American Society for Information Science*, 52(7):548-557.

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. *Journal of the American Medical Informatics Association*, 10(3): 252-259.