

Mining MEDLINE Metadata to Explore Genes and their Connections

Aditya Sehgal⁺, Xin Ying Qiu[®], Padmini Srinivasan^{®!}

⁺Computer Science Department

[®]Department of Management Sciences

[!]School of Library and Information Science

The University of Iowa

Iowa City, IA 52242

asehgal@cs.uiowa.edu, xiqu@blue.weeg.uiowa.edu, padmini-srinivasan@uiowa.edu

ABSTRACT

Biomedical scientists are faced with enormous challenges when tracking new discoveries made and research produced in their domain of interest. These challenges are exacerbated by the need to track research in other domains that might possibly be relevant to one's own research. Tools that support researchers by mining appropriate information from text collections would be of immense value. In this context we present a system that may be used by scientists to explore topics and their relationships using a collection such as MEDLINE. Several kinds of analyses are supported. One may study the features of an individual topic or those of a group of topics where these features are identified from the text collection. Functions are provided to examine topic co-occurrence and topic similarity based associations. The user may also explore indirect connections between topics by using a closed discovery process that is part of the system. In this paper we describe our system which offers a common framework for these functions. We illustrate the design of the system with a dataset of topics from the pediatrics genetics domain.

1. INTRODUCTION

The growth rate of biomedical knowledge creates an enormous challenge for scientists trying to keep pace with developments in their field. At the same time the vast collections of biomedical publications offer an excellent opportunity for text mining, i.e., the automatic discovery of knowledge. Text mining is similar to data mining [4] in its goal. But instead of mining a collection of well structured data, text mining operates off text collections that are at best semi-structured. In both cases the *knowledge* discovered are essentially propositions or hypotheses, that require further study and verification. Text mining has attracted the attention of many researchers e.g., [1, 5, 6, 7, 9, 13, 20, 15, 24],

including those in biomedicine. A recent article in Nature [2] referring to text mining as conceptual biology speaks to its legitimacy as a field that fuels hypothesis driven biomedical explorations. Examples of recent text mining applications include automatically identifying viruses that may be used as bioweapons [20], proposing therapeutic uses for thalidomide [23] and finding functional connections between genes [11].

We recently proposed and evaluated algorithms that mine metadata assigned to text for hypothesis discovery (text metadata mining). Our algorithms follow the discovery framework initiated by Swanson in the mid 1980s. His aim was to process MEDLINE in particular ways and generate interesting hypotheses concerning specific diseases and health problems. Given two topics that are bibliographically disconnected areas of specialization, Swanson explored potential linkages via intermediate topics or specializations. Over the past two decades and in collaboration with Smalheiser, Swanson proposed several interesting hypotheses [22, 21, 20, 13], that were later validated by bioscientists. Since their pioneering contributions this kind of knowledge discovery work has attracted the attention of other researchers [10, 24] besides us. In earlier research we were able to successfully replicate the Swanson and Smalheiser discoveries using a text metadata mining procedure [16]. We have also used our metadata mining approach to explore the global distribution of disease research represented in MEDLINE [17].

In this paper our aim is to present a system that offers our text metadata mining procedures within a larger context of mining functions. Users may employ our system to explore individual topics as well as their relationships. Aspects that may be explored include topic co-occurrence, topic similarity and indirect relationships exposed using our metadata mining procedures.

To illustrate our ideas, consider a user who is interested in the gene 'myosin light chain 2'. The first question she asks is: what does the text collection, in this case MEDLINE, say about this gene? One may of course retrieve the relevant documents from the text collection and stop after displaying these to our user. However, our goal is to go beyond document retrieval and provide her with an automatically derived summary description of the gene. A useful description is one that identifies for her the gene's key features or properties. Thus we display topic profiles generated from metadata assigned to the retrieved documents (in this case

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR'03 Workshop on Bioinformatics, July 28 -August 1, 2003, Toronto, Canada.

Copyright 2003 ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

MeSH terms). Consider now an extension of this scenario, where the user specifies a group of topics. For example, our user is interested in a set of genes that cluster together in a microarray experiment. Can their individual descriptions be analyzed such that features in common or features that are unique are identified? In response we present information summarizing the profiles of the gene topics. Next our user is interested in exploring ‘similarity’ between genes. Moreover, she seeks to assess similarity relative to a subset of the genes’ features. These can be handled, again using our topic profiles. Finally she would like to explore potential connections between pairs of genes even if the pairs have no retrieved documents in common. We support such enquiry using our metadata-based closed discovery process. In general, our system presents a metadata-based text mining system that offers the capacity to explore topics along these different dimensions. In this paper we describe this system and also demonstrate it with a group of genes identified by a researcher in the pediatrics genetics department at the University of Iowa.

2. METADATA MINING

Figure 1 provides an overview of our metadata-based text mining system. It has 3 major modules for: dataset specification, dataset processing and dataset analysis. The user is involved in the first and third modules. Completion of the first module triggers the second module which runs independent of the user.

Our user first defines a dataset of interest consisting of any number of topics. A dataset may include topics that are both already known to the system and those that are new. Topic selection/specification functions enable this step of dataset specification. The system first performs some data management chores such as providing each topic with a unique identifier and establishing links between the topics and the dataset which also has a unique identifier. These details are stored in a backend database. After this, dataset processing begins. In this module documents are first retrieved for each topic from the text collection. Next topic profiles are build. A variety of association scores are then computed based on co-occurrence and profile similarity. Profiles and association scores are stored in the backend database for persistent storage. The last module supports analysis of the dataset by the user. The user may exercise various options such as explore individual topic profiles; view co-occurrence based graphs of the topics; view the similarity scores between topics and also conduct closed discovery procedures between topics.

3. DATASET SPECIFICATION

3.1 Topics

A topic is essentially any search specification supported by an external retrieval system (web accessible) that interfaces with the text collection being mined. In our case, topics are MEDLINE searches supported by the PubMed interface. Each topic is given a unique id and is added to the persistent database. Topics profiled may be as simple as those described by single word searches (eg. *Tylenol*) or they may be more complex such as *Calcium channel blockers AND Alzheimers disease*.

3.2 Datasets

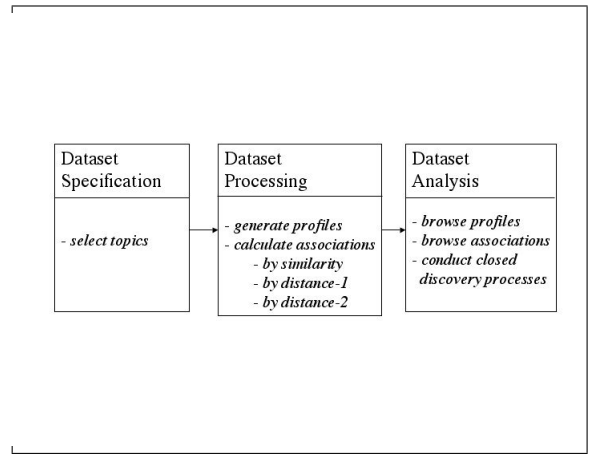


Figure 1: Three Major Modules of the System

A dataset is a group of topics that the user wants to jointly analyze. Datasets are also persistent in our system thus allowing the same or other users to analyse them at any point after their creation. Datasets need not be mutually exclusive in their topic composition. In fact datasets may overlap or one may be a true subset of another. Finally, at one extreme datasets may contain only a single topic just as a dataset may contain all topics thus far known to the system.

In this paper we illustrate our system with a dataset consisting of 26 genes. These genes were specified by a pediatrics geneticist working with the output of a microarray experiment. The genes are of interest because they cluster together in terms of their expression patterns. The reason for their similarity in expression is unknown. Thus the user is interested in determining what the MEDLINE database has to say about these 26 genes. Table 1 lists these genes and provides some details about each gene. Topics are formed from this information. For example, the topic derived for the gene in the first row is: ‘*MLC2a OR myosin light chain 2 OR Mylc2a OR RLC-A OR MLC-2alpha*’.

4. DATASET PROCESSING

4.1 Topic Profiles

Our analysis algorithms operate on topic profiles. A profile is essentially a set of characteristic terms representing the topic¹. In this work, we build MeSH topic profiles from MEDLINE.

First we retrieve a relevant subset of documents from the text collection. We then identify the MeSH terms assigned to these documents and generate weighted vectors from this information as shown below for a topic T_i .

$$Profile(T_i) = \{w_{i,1}m_1, w_{i,2}m_2, \dots, w_{i,n}m_n\} \quad (1)$$

where m_j represents a MeSH term, $w_{i,j}$ its weight and there are totally n terms in the MeSH vocabulary. (We discuss weights shortly).

¹Note that these may in general be extracted from the free-text and/or metadata portions of the documents.

Gene Symbol	Gene Name	Sample Aliases
MLC2a	myosin light chain 2	Mylc2a; RLC-A; MLC-2alpha
GATA4	GATA-binding protein 4	Gata-4
ANF	atrial natriuretic factor	NPPA; ANP; PND; CDD-ANF; Pnd; RATANF
MYL1	myosin light chain 1	MLC1F; MLC3F; A1 catalytic; A2 catalytic
GJA5	gap junction protein	alpha 5; 40kDa (connexin 40); Cx40; Cnx40; Gja-5
TPM1	tropomyosin 1 (alpha)	CMH3; TMS; Tmpa; Tpm-1; alpha-TM; Tma2
HSPB2	heat shock 27kDa protein 2	MKBP; HSP27; HSP72; HS.78846; MKBP; 27kDa
Cryab	crystallin	alpha B; Crya; Crya-2; CTPP2; AACRYA
Msg1	melanocyte-specific gene 1 protein	
Ldhb	lactate dehydrogenase B	Ldh2
TIMP4	tissue inhibitor of metalloproteinase 4	TIMP-4
RPL11	ribosomal protein L11	DL11; CG7726; CT23337; l(2)k16914
APOBEC-2	apolipoprotein B mRNA editing enzyme	catalytic polypeptide-like 2; APOBEC2; ARP1; ARCD1
TnI	tropoin I	wupA; HDP; HLI; hdp; HL-I; wup-A; CG7178; heldup
TnC	Troponin C	TpnC41C; TPNC; CG2981; TnC41C; CT10051; TNNC
MMP14	matrix metalloproteinase protein 14 (membrane inserted)	MMP-X1; MTMMP1; MT1-MMP
Spna2	alpha-fodrin	spna-2 brain; 2610027H02Rik; A2a; SPTAN1
Tgm2	tissue-type transglutaminase	TgaseII; TGC; G[a]h; tissue transglutaminase
NADP+ -specific ICDH	NADP+-specific isocitrate dehydrogenase	IDH; IDP; PICD; IDH1; IDHM; ICD-M; mNADP-IDH
RAD	ras associated with diabetes gene	RAD1; RRAD; Ras-like GTP-binding protein rad
SP120	SP120	MGI; SAFA; hnRNPU; AA408410; AI256620; AL024437
MYBPC3	myosin binding protein C	FHC; CMH4; MYBP-C; myosin binding protein C cardiac
ATP2A2	brain Ca2+ ATPase	DD; DAR
ATP2A2	Ca2+/Mg2+ ATPase	Ca2+/Mg2+ ecto-ATPase; (Ca2+ + Mg2+) ATPase
SPTR	beta tyrosine phosphatase	
FABP3	fatty acid binding protein	MDGI; H-FABP; Fabph1; Fabph4; Fabph-1; Fabph-4

Table 1: Dataset of 26 genes from a microarray experiment in pediatrics genetics.

Profiles may be as current as the text collection or generated from subsets corresponding to particular time periods. Such temporal profiles may support trend analysis as in [9, 16].

4.2 Employing Semantic Types in Profiles

The MEDLINE MeSH vocabulary has been classified into 134 groups within the UMLS (Unified Medical Language System)². These groups are called semantic types (NLM, 2002). *Cell Function*, *Sign or Symptom* are two examples of 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. More generally, semantic types represent ‘categories’ that have been used to classify the MeSH metadata. We make use of these semantic types to differentiate between the MeSH terms in our topic profiles. In particular semantic types supports the notion of *views* in our system. Henceforth, MeSH terms will be in small case while semantic types will have the first character of each word capitalized. Both will appear in italics.

Figure 2 which outlines our procedure for building profiles shows how we involve these semantic types. Basically MeSH terms are separated by semantic type and term weights are computed within the context of a semantic type. This results in a vector of MeSH term vectors, one for each of the 134 UMLS semantic types. Thus,

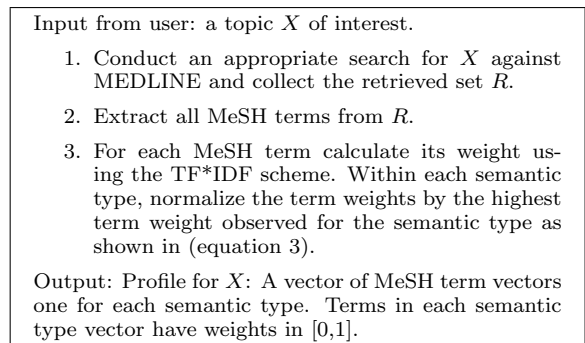


Figure 2: Procedure for generating TF*IDF Weighted MeSH Profiles.

²<http://umlsks.nlm.nih.gov>

$$Profile(T_i) = \{ \{ w_{i,1,1}m_{1,1}, w_{i,1,2}m_{1,2}, \dots \}, \dots, \{ w_{i,134,1}m_{134,1}, w_{i,134,2}m_{134,2}, \dots \} \} \quad (2)$$

where $m_{x,y}$ represents the MeSH term m_y that belongs to the semantic type x and $w_{i,x,y}$ is the computed weight for $m_{x,y}$. Weights may be computed using any appropriate weighting scheme (such as mutual information and log likelihood). Below we use the TF*IDF (term frequency * inverse document frequency) [14] weighting scheme and then normalize the weights:

$$w_{i,x,y} = v_{i,x,y} / highest(v_{i,x,l}), \quad (3)$$

where $l = 1, \dots, r$ and $v_{i,x,y} = n_{i,x,y} * \log(N/n_{x,y})$. Here N is the number of documents in the database, $n_{x,y}$ is the number of documents in which $m_{x,y}$ occurs and $n_{i,x,y}$ is the number of retrieved documents for T_i in which $m_{x,y}$ occurs. Normalization by $highest(v_{i,x,l})$, the highest value for $v_{i,x,y}$ observed for the MeSH terms with semantic type x , yields weights that are in $[0,1]$ within each semantic type. (Note that there are r terms in the domain for semantic type x).

Thus a profile represents the relative importance, within semantic types, of the different MeSH terms associated with the topic's document set. When appropriate, this profile may be focussed or limited to a specific *view* i.e., terms with particular semantic types. For example, profiles of genes may be limited to functional semantic types such as *Cell Function* and *Pathologic Function*. In a recent paper [17] we used profiles of diseases limited to *Geographic Area* to explore the global distribution of research on various diseases. We then compared such distributions with disease prevalence (distribution) data.

4.2.1 Example Profile

We use the gene 'TnI' as an example topic to illustrate profiles. Table 2 presents details of the counts pertaining to the set of retrieved documents and the profile. The search included TnI and all alias terms as well as general terms such as 'genes' and 'genetics' to limit the retrieved documents to those that are most likely to be about genes. Five top ranked MeSH terms and their weights (equation 3) are shown for a sample set of semantic types. Semantic types with the most terms are also identified. For example, *Amino Acid, Peptide, or Protein* with 363 terms is the top ranked type. With respect to the distribution of weights, a threshold of 0.5 yields a profile of 206 terms which is only 12% of the original 1,720 terms. A threshold of 0.3 gives 336 (20%) while a threshold of 0.7 gives 157 terms which is less than 9% of the terms.

4.3 Topic Co-Occurrence

Co-occurrence as a phenomenon suggesting relatedness has been explored in several papers [8, 18, 19]. Although co-occurrence may capture meaningful *known* relationships, our emphasis is on identifying novel relations. Therefore, we extend the standard notion of co-occurrence.

4.3.1 Distance-1 co-occurrence links:

If D_i and D_j represent the documents retrieved for two topics T_i and T_j then if D_i and D_j overlap we say that the two topics exhibit distance-1 co-occurrence. We measure the association strength, $Association_{distance-1}(T_i, T_j) =$

$$|D_i \cap D_j| / |D_i \cup D_j| \quad (4)$$

4.3.2 Distance-2 co-occurrence links:

Consider topics T_i and T_j that are at distance-1 and T_j and T_k that are at distance-1. Thus it may be assumed that there is some meaningful connection between the topics of each co-occurring pair. Assume now that T_i and T_k do not co-occur. It may be concluded that a direct connection between these two topics has not been identified. However, given the co-occurrence patterns it may be inferred that a meaningful connection between T_i and T_k is likely, via the intermediate topic T_j . Thus T_j forms a bridging topic between the other two. We then say that T_i and T_k are at distance-2 from each other. Observe that distance-1 pairs cannot be at distance-2. Moreover, all pairs of topics that are at distance-2 are not necessarily equal. Some pairs may have just a single bridging topic while others may have several. We suggest that the greater the number of bridging topics the more likely the two topics at distance-2 are related in a meaningful way.

Let T_i and T_j be two topics at distance-2. Let P_i and P_j represent the set of topics connecting at distance-1 with T_i and T_j respectively. $Association_{distance-2}(T_i, T_j) =$

$$|P_i \cap P_j| / |P_i \cup P_j| \quad (5)$$

4.4 Profile Similarity

The similarity between two topics T_i and T_j is calculated as the cosine score between their profile vectors.

$Association_{similarity}(T_i, T_j) =$

$$\sum_{r=1}^p (w_{ir} * w_{jr}) / \sqrt{\sum_{r=1}^p w_{ir}^2 * \sum_{r=1}^p w_{jr}^2} \quad (6)$$

where w_{ir} , for example, is the weight of r , a metadata term - metadata category combination, in $Profile(T_i)$ and there are p metadata term - metadata category combinations. Note that this association measure when based on the profiles is not dependent on the co-occurrence of T_i and T_j . A pair may never co-occur and yet have a high association score. Again profiles may be limited to certain categories of metadata, thus supporting more specialized analysis.

In summary, during the dataset processing phase profiles are built for individual topics and added to persistent storage. Association scores are also calculated and added to persistent storage for future display to the user.

5. DATASET ANALYSIS

Having processed the dataset, it is now available for analysis by the user. Besides browsing profiles and association scores, the key analysis step available at this point is our closed discovery process which is described next.

5.1 Closed Discovery Analysis: Swanson and Smalheiser Algorithm

In previous research we explored Swanson and Smalheiser's closed discovery algorithm for discovering novel relationships between topics. Given two topics T_i and T_j that do not co-occur, the aim is to identify key concepts that might represent connections. For example, in 1986, Swanson [22] had the intuition that fish oils (called the C concept) may be useful in treating Raynauds disease (called the A concept). Their procedure started with independent literature searches on the initiating A and C concepts. Titles of retrieved documents were scanned manually for interesting po-

<p>Gene: TnI</p> <p>Number of documents retrieved: 1,247</p> <p>Number of MeSH term instances in the document set: 27,355</p> <p>Number of unique MeSH terms in the document set: 1,720</p> <p>Profile: (top 5 terms for a few semantic types are shown below)</p> <p>Semantic Type: <i>Body Part, Organ, or Organ Component:</i> <i>muscle, skeletal (1.00), muscles (0.95), heart (0.47), heart ventricle (0.16), heart atrium (0.06),</i></p> <p>Semantic Type: <i>Cell:</i> <i>cells, cultured (1.00), muscle fibers, fast-twitch (0.79), cell line (0.79), muscle fibers (0.70), muscle fibers, slow-twitch (0.64)</i></p> <p>Semantic Type: <i>Cell Function:</i> <i>cell differentiation (1.00), calcium signaling (0.18), cell division (0.17), cell fusion (0.11), cell movement (0.09)</i></p> <p>Semantic Type: <i>Molecular Function:</i> <i>enzyme activation (1.00), mutagenesis (0.99), linkage (genetics) (0.91), signal transduction (0.63), binding, competitive (0.42)</i></p> <p>Semantic Type: <i>Disease or Syndrome:</i> <i>cardiomyopathy, hypertrophic (1.00), cardiomyopathy, congestive (0.19), myocardial diseases (0.15), myocardial ischemia (0.13), heart failure, congestive (0.13)</i></p> <p>Semantic Type: <i>Pathologic Function:</i> <i>cardiomegaly (1.00), death, sudden, cardiac (0.66), hypertrophy, left ventricular (0.38), ventricular dysfunction, left (0.38), death, sudden (0.20)</i></p> <p>Semantic Type: <i>Enzyme:</i> <i>myosins (1.00), adenosinetriphosphatase (0.21), cyclic amp-dependent protein kinases (0.21), creatine kinase (0.19), chloramphenicol O-acetyltransferase (0.19),</i></p> <p>Semantic Type: <i>Physiologic Function:</i> <i>regeneration (1.00), fetal development (0.91), down-regulation (0.54), energy metabolism (0.45), up-regulation (0.27)</i></p> <p>Number of unique MeSH terms in profile: 1,720</p> <p>Total number of MeSH term entries in profile: 2,511 (a term can be in multiple semantic types)</p> <p>Top 5 Semantic types ranked by number of terms: <i>Amino Acid, Peptide, or Protein (363), Biologically Active Substance (224), Laboratory Procedure (128), Enzyme (121), Pharmacologic Substance (109)</i></p> <p>Number of semantic types with at least 1 term in profile: 111 (out of 134 possible)</p>
--

Table 2: Example Profile. Topic Gene: TnI. (Only the top 5 MeSH terms and weights are shown within select semantic types.)

tential connections between A and C. In particular Swanson observed from the A literature that Raynaud’s syndrome is exacerbated by high platelet aggregability, high blood viscosity, and vasoconstriction. He also observed from the C literature that these same phenomena are reduced by dietary fish oils. Moreover, the A and C literatures had no overlap. Thus by reading the titles in these two literatures, Swanson was able to suggest that fish oils may be beneficial to patients with Raynauds, which was later validated in [3].

Swanson and Smalheiser made several other predictions using this closed process. For example, they also identified 11 different pathways between Migraine and Magnesium [21]. The hurdle in their procedures is the need to manually examine, comprehend and select phrases at different steps. Their ARROWSMITH program (which includes a stoplist of about 8000 words) is designed to assist a user with these discovery processes.

In previous research [16] we described and tested a metadata based algorithm that successfully replicated the various Swanson and Smalheiser discoveries. Figure 3 outlines the key steps in our algorithm. Profiles are built for the two topics. P top ranking terms are retained for each semantic type. A profile of MeSH terms in common between AP and CP is then built which represents potentially novel connections. Also the constraint in step 4 may be relaxed to offer the user the opportunity to explore connections between A and C that may not be totally novel in the literature.

For example in the Raynauds - fish oils discovery our algorithm identified the key concepts at the topic ranks as shown in Table 3. Parameters P was set to 20 for the combined profile. The ranks of the key connecting terms indicates that all three pathways are discovered within the top 3 ranks.

Metadata Term (MeSH Term)	Metadata Category (Semantic Type)	Rank
Platelet Aggregation	Cell Function	1
Platelet Adhesiveness	Cell Function	3
Blood Viscosity	Physiologic Function	1
Thrombosis	Finding	1
Vasoconstriction	Finding	3

Table 3: Pathways between Raynauds and Fish Oils

Thus we offer our closed discovery algorithm as one of the functions in our system for analysing datasets.

5.2 User Interaction

During the analysis phase, the user starts by identifying the dataset that is to be analysed. At this point all the pre-computed information (profiles, distance-1, distance-2 and similarity association scores) associated with the dataset are available for display. The user is first presented with a graph where nodes are topics and links represent distance-1 and distance-2 connections. The display can also be designed to focus on a single topic. The screenshot of Figure 1 is an example of such a display. In it ‘GATA4’ is the gene topic that is the focus. If we start at the GATA4 node, brown arcs from here lead to topics that are at distance-1. The green links from these nodes in turn lead to their distance-1 connected topic nodes. Blue links represent distance-2 links. For example, GATA4 and TIMP4 are at distance-2 from each other with 3 bridging topics between them: Spna2, RPL11 and

Tam2. $Association_{similarity}$ scores are also shown for the distance-2 links. The 5 distance-2 links are also the ones with the highest $Association_{similarity}$ scores. Out of the 5 distance-2 links, the one connecting GATA4 with MLC2a has the highest number (9) of bridging topics. It seems reasonable for the user to first explore potential links between these two genes. However, if one considers the similarity score, then the distance-2 link between GATA4 and Cryab seems potentially the most interesting.

Anytime the user is presented with a topic graph he/she may click on any topic node and view the corresponding metadata profile (an example is shown in Table 2). If instead our user is interested in an average profile for the dataset as a whole, output as shown in Table 4 for our 26 genes is available.

Selecting an edge displays the results of applying our closed discovery function between the two connected topics. A sample of this output is shown in Table 5. Thus after identifying a distance-2 link of interest our user may look deeper into the literature for evidence supporting the link. In this way our user may move seamlessly between the different analysis techniques.

6. RELATED RESEARCH

Co-occurrence is the basis of many text mining applications exploring concept similarity or relatedness in the biomedical domain [8, 18, 19]. Jenssen et al., [8] generate a co-occurrence based gene-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. Stapley and Benoit [18] also exploit co-occurrence to generate a gene-gene map from MEDLINE documents containing the term *Saccharomyces cerevisiae*. While Stephens et al., [19] also use co-occurrence data to postulate gene interactions, they go beyond simple frequency based counts. They also consider how frequently the gene term (or its synonyms) occurs in the document. Associations above a particular threshold are analysed further using a list of relationship words such as *activates* and *cleaves*.

Several researchers besides us are looking beyond co-occurrence based concept associations. Shatkay et al., [11] first identify for each gene a *kernel* document describing the gene’s function. This document is then used to seek out similar documents from MEDLINE. Overlap in document sets provide estimates of the functional similarities between the source genes. In addition to the discoveries regarding Raynauds disease and Alzheimers disease, Swanson and Smalheiser discovered several other connections such as between estrogen and Alzheimers [13]. In each case they identified possible mechanisms of interaction. Lindsay and Gordon [10] and Weeber et al., [24] have explored several alternative approaches emulating the Swanson and Smalheiser discoveries. Other text mining based discoveries that are equally intriguing include the more recent work by Swanson et al., on categorizing viruses [20].

7. CONCLUSION

In this paper we have presented preliminary work on the design of a metadata-based text mining system. The design builds on our earlier research exploring a metadata-based algorithm for the Swanson and Smalheiser discovery proce-

Input from user: Two topics of interest designated, A and C.
 Parameter: P .

- Step 1: Conduct independent PubMed searches for A and C. Build the A and C MeSH profiles (equation 3). Call these profiles AP and CP respectively.
- Step 2: For each semantic type retain only the highest weighted P MeSH terms within AP and CP.
- Step 3: Compute a B profile (BP) composed of terms in common between AP and CP. The weight of a MeSH term in BP is the sum of its weights in AP and CP.
- Step 4: Eliminate terms from BP that do not represent novel associations. That is, if a search for A AND t AND C returns non zero results, then eliminate t from BP. The remaining MeSH terms are the potential B concepts of interest. (This constraint may be relaxed if the user is also interested in connections that are not necessarily novel).

Output: For each semantic type, display a ranked list of MeSH terms to the user. Each term represents a potential conceptual connection between A and C.

Figure 3: Closed Discovery Algorithm: Outline of Steps.

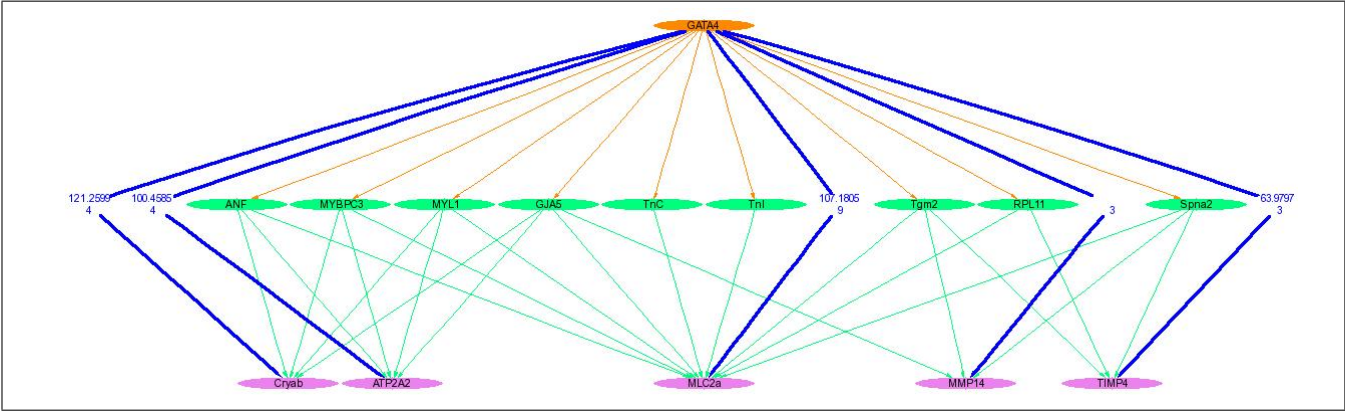


Figure 4: Screenshot: GATA4 is the focus gene topic. Distance-1 and distance-2 links are displayed.

Topic: Average profile for the whole dataset of 26 genes
PubMed Search:
Number of documents retrieved: 213,025
Number of MeSH term instances in the document set: 3,424,962
Number of unique MeSH terms in the document set: 15,671
Profile: (top 5 terms for a few semantic types are shown below)
 Semantic Type: *Body Part, Organ, or Organ Component:*
liver (0.40), muscle, skeletal (0.38), heart (0.33), muscles (0.31), brain (0.26),
 Semantic Type: *Cell:*
cells, cultured (0.70), cell line (0.61), tumor cells, cultured (0.51), fibroblasts (0.23), 3t3 cells (0.19)
 Semantic Type: *Cell Function:*
cell differentiation (0.60), cell division (0.50), apoptosis (0.33),
lymphocyte transformation (0.21), cell adhesion (0.19)
 Semantic Type: *Molecular Function:*
signal transduction (0.60), enzyme activation (0.49), linkage (genetics) (0.44), mutagenesis (0.38),
binding, competitive (0.27),
 Semantic Type: *Disease or Syndrome:*
liver (0.49), brain (0.32), cardiomyopathy, hypertrophic (0.22), lung (0.21),
hypertension (0.16),
 Semantic Type: *Pathologic Function:*
cardiomegaly (0.39), neoplasm invasiveness (0.26), inflammation (0.20), insulin resistance (0.20),
hypertrophy, left ventricular (0.20),
 Semantic Type: *Enzyme:*
myosins (0.29), isoenzymes (0.20), protein kinase c (0.18), gtp-binding proteins (0.16),
protein-serine-threonine kinases (0.13),
 Semantic Type: *Physiologic Function:*
down-regulation (0.59), up-regulation (0.57), fetal development (0.40), electrophysiology (0.21),
energy metabolism (0.15)
Number of unique MeSH terms in profile: 15,671
Total number of MeSH term entries in profile: 111,930 (a term can be in multiple semantic types)
Top 5 Semantic types ranked by number of terms: *Amino Acid, Peptide, or Protein (2185),*
Disease or Syndrome (1962), Organic Chemical (1762), Pharmacologic Substance (1677),
Biologically Active Substance (1223)
Number of semantic types with at least 1 term in profile: 130 (out of 134 possible)

Table 4: Average Profile. Dataset of 26 Genes. (Only the top 5 MeSH terms and weights are shown within select semantic types.)

MeSH Term	Semantic Type	Rank
Fetal Development	Physiologic Function	1
Down-Regulation	Physiologic Function	2
Up-Regulation	Physiologic Function	3
Neovascularization, Physiologic	Physiologic Function	4
Transcription Factors	Amino Acid, Peptide, or Protein	1
Atrial Natriuretic Factor	Amino Acid, Peptide, or Protein	2
Signal Transduction	Molecular Function	1
Enzyme Activation	Molecular Function	2
Binding, Competitive	Molecular Function	3
Sequence Deletion	Cell or Molecular Dysfunction	1
Chromosome Deletion	Cell or Molecular Dysfunction	2
Oxidative Stress	Cell or Molecular Dysfunction	3
Cell Differentiation	Cell Function	1
Embryonic Induction	Cell Function	2
Cardiomegaly	Pathologic Function	1
Hypertrophy	Pathologic Function	2
Constriction, Pathologic	Pathologic Function	3
Heart Defects, Congenital	Disease or Syndrome	1

Table 5: Output from a Closed Discovery Process between Gene MLC2a and Gene GATA4 (Only select terms and semantic types are shown. Parameter P was set at 5. Rank is within semantic type.)

dure. The system embeds the closed discovery process in a framework that supports co-occurrence and similarity based analysis as well. These various options jointly offer the user a family of operations for exploring topics and their relationships. At the foundation of the different functions is the notion of topic profiles that are built using the metadata assigned to texts. We illustrate the capabilities of our system with examples of genes selected from a dataset of genes identified by a user in the area of pediatrics genetics. In addition we have a second dataset of about 100 genes identified by another user interested in the disease domain of ‘Chondrosarcoma’. Our users are eager to explore their datasets using our system. We plan to have the user-directed analysis completed, for the collection of 26 genes, in time for the workshop.

8. REFERENCES

- [1] Andrade A, Valencia A. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, 14(7):600-607, 1998.
- [2] Blagosklonny, M. V., & Pardee, A. B. Unearthing the gems. (2002). *Nature*, 416, 373.
- [3] DiGiacomo R.A, Kremer J.M and Shah D.M. Fish oil dietary supplementation in patients with Raynaud’s phenomenon: A double-blind, controlled, prospective study. *American Journal of Medicine*, 8, 158-164, 1989.
- [4] Fayyad U.M. and Uthurusamy R. Data mining and knowledge discovery in databases (Introduction to the special section). *CACM*, 39(11):24-26, 1996.
- [5] Feldman R, Aumann Y, Amir A, Klosgen W, and Zilberstien A. Maximal association rules: A new tool for mining for keyword co-occurrences in document collections. *KDD-97*, Newport Beach, CA, 1997.
- [6] Hahn U, and Schnattinger K. Deep knowledge discovery from natural language texts. *KDD-97*, 175-178, 1997.
- [7] Hearst M. Untangling text data mining. *Proceedings of the 37th ACL Conference*, 1999.
- [8] Jenssen T-K, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28:21-28, 2001.
- [9] Lent B, Agrawal R. and Srikant R. Discovering Trends in Text Databases. *Proceedings of the 3rd International Conference on Knowledge Discovery*, *KDD-97*, Newport Beach, CA, 1997.
- [10] Lindsay R.K and Gordon M.D. Literature-based discovery by lexical statistics. *JASIS*, 50(7):574-587, 1999.
- [11] Shatkay H, Edwards S and Wilbur WJ, Boguski M. Genes, Themes and Microarrays. Using information retrieval for large-scale gene analysis. *ISMB*, La Jolla, California, USA, 317-328, 2000.
- [12] Smalheiser N.R. and Swanson D.R. Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. *Computing Methods Programs in Biomedicine*. 57(3),149-153, 1998.
- [13] Smalheiser N.R. and Swanson D.R. Linking estrogen to Alzheimer’s disease: An informatics approach. *Neurology*, 47:809-810, 1996.
- [14] Sparck Jones K. *Automatic keyword classification for information retrieval*. Butterworths, London, UK, 1971.
- [15] Srinivasan P. MeSHmap: A text mining tool for MEDLINE. *Proceedings of the AMIA Symposium*, 642-646, 2001.
- [16] Srinivasan P. *Text Mining: Generating Hypotheses from MEDLINE Submitted to: JASIST*, May 2003.
- [17] Srinivasan P. and Wedemeyer M. 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, June 2003.
- [18] Stapley B.J. and Benoit G. Bibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *PSB*, 5:526-537, 2000.
- [19] Stephens M, Palakal M, Mukhopadhaya S, Raje R., and Mostafa J. Detecting gene relations from MEDLINE abstracts. *PSB*, 483-496, 2001.
- [20] Swanson D.R, Smalheiser N.R and Bookstein A. Information discovery from complementary literatures: categorizing viruses as potential weapons. *JASIST* 52(10), 797-812. August 2001.
- [21] Swanson D.R. Migraine and Magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, 31:526-557, 1988.
- [22] Swanson DR. Fish oil, Raynaud’s syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30:7-18, 1986.
- [23] 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.
- [24] Weeber M, Klein H, Berg L, Vos R. Using concepts in literature-based discovery: Simulating Swanson’s Raynaud-Fish Oil and Migraine-Magnesium discoveries. *JASIST*, 52(7):548-557, 2001.