

Text Mining Functional Keywords Associated with Genes

Ying Liu^a, Martin Brandon^a, Shamkant Navathe^a, Ray Dingledine^c, Brian J. Ciliax^b

^aCollege of Computing, Georgia Institute of Technology, USA

^bDepartment of Neurology, ^cDepartment of Pharmacology, Emory University Medical School, USA

Abstract

Modern experimental techniques provide the ability to gather vast amounts of biological data in a single experiment (e.g. DNA microarray experiment), making it extremely difficult for the researcher to interpret the data and form conclusions about the functions of the genes. Current approaches provide useful information that organizes or relates genes, but a major shortcoming is they either do not address specific functions of the genes or are constrained by functions predefined in other databases, which can be biased, incomplete, or out-of-date. We extended Andrade and Valencia's method [1] to statistically mine functional keywords associated with genes from MEDLINE abstracts. The MEDLINE abstracts are analyzed statistically to score and rank keywords for each gene using a background set of words for baseline frequencies. We generally got very good functional keyword information about the genes we tested, which was confirmed by searching for the individual keywords in context. The keywords extracted by our algorithm reveal a wealth of potential functional concepts, which were not represented in existing public databases. We feel that this approach is general enough to apply to medical and biological literature to find other relationships: drugs vs. genes, risk-factors vs. genes, etc.

Keywords:

Text mining, MEDLINE, functional keywords, gene.

Introduction

One of the rich resources of on-line information is the scientific literature. The MEDLINE database, for example, provides bibliographic information and abstracts for more than 12 million articles that have been published in biomedical journals [2]. However, all the information contained in the database is stored as text. The rapid growth of these collections makes it increasingly difficult for humans to access the required information in a convenient and effective manner [3]. Clearly, there is a necessity of developing methods for automatic extraction of relevant information (such as keywords associated with genes) from the literature, which is written in natural language.

A number of groups are developing algorithms that link information from medical literature with gene names. Andrade and Bork [3] developed a program that links the OMIM database of human inherited diseases to keywords derived from MEDLINE with their statistical profiling algorithm. A variety of nonstatistical

approaches have also been used to organize genes. The web tool, PubGene, finds links between pairs of genes based on their co-occurrence in MEDLINE abstracts [4]. Another approach [5], the basis of the HAPI web tool, organizes gene names according to predefined hierarchical classification systems of enzymes and diseases, and includes hyperlinks to specific MEDLINE citations responsible for the individual classifications. Still another approach [6], used by the MedMiner system, automatically retrieves functional information (both keywords and gene names related to a user-defined function) from the GeneCards database, and configures it for a PubMed search. The algorithm presents the results by the specific sentence containing the information rather than by the title, speeding review of the results if the user prefers to extract relevant sentences rather than scan through the whole abstract text. Databases, such as Gene Ontology (GO) Consortium, Swiss-Prot, GenBank, and GeneCards, reduce vast literature into a few functional concepts.

Andrade and Valencia [1] introduced a statistical profiling strategy that accepts user-supplied abstracts related to a protein of interest and returns an ordered set of keywords that occur in those abstracts more often than would be expected by chance. We have extended this approach to build a system for mining functional relationships of genes by a) testing new background sets, an alternative stemming algorithm and a new, extensive stop list customized for use with the biological literature; b) using the gene-associated keywords to cluster genes with similar functions; and c) providing a list of the top-ranked keywords of each cluster as an aid to hypothesis building [7]. In this paper, we use this system to create a repository of functional keywords from MEDLINE abstracts for genes. We also compared our results with information found in public databases.

Materials and Methods

We used statistical methods to extract keywords from MEDLINE citations [1, 7]. This method estimates the significance of words by comparing the frequency of words in a given set (Query Set) of abstracts with their frequency and distribution in a Background Set of abstracts [1]. In general, words were considered to be more likely associated with a gene if their frequencies in relevant abstracts were higher than what would be expected by chance.

Stemming

Word stemming is used to truncate suffixes and trailing numerals so that words having the same root (e.g., activate, activates, and activated) are collapsed to the same word for frequency counting. In this paper, we used a simple stemming algorithms used by [1]. Two words are considered to have the same stem if they have the same beginnings and their endings differ in one or two characters (e.g. kinase and kinases; express and expressed). However, this rule will not be applied when the stem has less than five characters to avoid unwanted situations like mistaking “actin” with “active”.

Building the background set dictionary

The background set was created consisting of 50,000 randomly selected MEDLINE abstracts sorted into 1000 groups of 50 abstracts each. The background set is used to reduce the weight of commonly occurring words. Words in the abstracts of the background set were statistically analyzed to build a dictionary [1]. The frequency of word, the mean frequency of word, and the standard deviation of word a across all groups formed a baseline against which the frequency of a word in a query set (see below) could be compared. All words, their corresponding average frequencies and standard deviations form the background set dictionary.

Query Set

For each gene analyzed, word frequencies were calculated from a group of abstracts retrieved by a search, in the TITLE field, for the specific gene name and any known aliases. The aliases were found in a public database LocusLink [8]. The resulting set of abstracts (the Query Set) was processed to generate a specific keyword list. The frequencies of every word in the Query Set were compared to the mean frequencies (and standard deviations) of the same words contained in a Background Set of abstracts to calculate the significance (z-score) for each (see Keywords Statistics section below).

We tested two query sets. The first group of genes (query set #1) were calyculin (C), cathepsin H (H), cathepsin S (S), glutamine-oxaloacetate transaminase (GOT), nexin-1 (N), osteopontin (OPN), and uridine kinase (UK). To test if our system can extract new information from the medical literature, we designed a second query set for OPN using abstracts only from the year 2001 (query set #2), with the hope of extracting relevant keywords for several novel functional links between OPN and diseases, such as hypertension[9], autoimmune demyelinating diseases [10] and tumor metastasis [11].

Stop-word list

The stop-word list was used to filter out the non-scientific English words. We created the stop list [7] based on an online dictionary of 22,205 words [12]. Our initial tests led us to add methodological words that are unrelated to gene or protein function to the online dictionary as a stop-word list customized to biological applications.

Keyword list construction

In the Query Set of words, the frequency of each word was calculated.

Then the test statistic, z (or the z -score) was calculated for each word by the equation: $z = (F - A) / \sigma$. Where:

F is the frequency of the word in the query set;

A is the average frequency of the word in the background set;

σ is the standard deviation of the word frequency in the background set.

All the words and their corresponding z -scores form the keyword list. If a word appears in the query set, but not in the background set, the z -score of this word is designated as “New”. New words were treated as having the highest z -score, in order to bring potentially novel words to the attention of the investigator.

Keyword sharing among genes

All combinations of keyword lists (new words and words with z -scores > 2) were compared in order to calculate the number and frequency of shared keywords among the genes in the query set #1.

Results

Keyword list:

An example of a keyword list (Query set #1) for the gene name “OPN” is shown in Table 1 (Only the top 100 words are shown). One author (BJC) selected functional words (bold font) to illustrate the enrichment of informative words near the top of the list. Other gene keyword lists are available upon request.

Keyword sharing

Keywords generated were tabulated and ranked according to the number of genes with which they were associated (Table 2). Functionally meaningful keywords are highlighted (bold font).

Discussion

Data are fundamental units of information, whereas knowledge is the comprehension of fact or truth. Data can be represented by numbers, text, and graphics, which can be labeled, tabulated, annotated, organized, and archived in databases. From such collections of data, patterns can be extracted and analyzed by data mining techniques. On the other hand, knowledge is understood through language and usually archived as text. Thus, mining the text of expert literature could provide us with many specific concepts of interest. The term “knowledge-base” has been used to describe a database of concepts and expertise. Such a knowledge-base is typically generated manually by human experts and archived as text, possibly with links to other items. To maximize the accessibility of biomedical knowledge for investigators using DNA microarrays, we intend to generate a knowledge-base automatically through text mining and to create an unbiased, up-to-date database of functionally relevant keywords for every named gene in GenBank

Table 1. Keyword list for gene Osteopontin

Word	z-score	Word	z-score
integrin	48.3	polyelectrolytes	8.7
transwell	22.5	postovulatory	8.7
ctgf	19.5	reaggregate	8.7
fluorimetry	19.5	glomeruli	8.6
histotroph	19.5	upregulation	8.5
tcf	19.5	normoxia	8.3
trophectoderm	19.5	proteinaceous	8.1
vsm	19.5	metastasis	8.0
vibrissa	15.9	lithogenic	7.9
adpkd	13.8	losartan	7.9
atn	13.8	ranets	7.9
catagen	13.8	uremia	7.9
chitosan	13.8	hyperglycemic	7.7
diphenylene	13.8	mapk	7.6
enterokinase	13.8	kappab	7.6
iodonium	13.8	normoxic	7.6
ishikawa	13.8	antineutrophil	7.3
mdh	13.8	crevicular	7.3
nexin	13.8	doca	7.3
nondialysis	13.8	mek	7.3
ptf	13.8	neurotomy	7.3
pulpitis	13.8	periglomerular	7.3
renoprotective	13.8	mcp	7.3
vth	13.8	fibrotic	7.2
interstitium	13.2	autoregulatory	6.8
lsab	11.2	gcf	6.8
matrigel	11.2	gonadotropes	6.8
postovulation	11.2	jnk	6.8
pulmonal	11.2	muc	6.8
telogen	11.2	spatio	6.8
upar	11.2	hgf	6.8
upa	10.8	migratory	6.7
stone	10.6	thrombin	6.7
atheromatous	10.4	chemokine	6.5
urolithiasis	10.4	antagonises	6.4
erk	10.3	deoxynucleotidyl	6.4
tartrate	10.3	hypercholesterolaemia	6.4
hoxa	9.7	hyperphosphatemia	6.4
igan	9.7	oro	6.4
lucigenin	9.7	kidneys	6.4
lymphoprolifera- tion	9.7	ethylene	6.3
nephritic osteoclastogene- sis	9.7	autoimmunity	6.2
talin	9.7	henle	6.1
tyimpanosclerosis	9.7	morphogenic	6.1
mesangial	9.2	propidium	6.1
mmp	8.9	smad	6.1
metalloproteinase	8.7	spongiosa	6.1
anagen	8.7	stromelysin	6.1
deoxypyridinoline	8.7	Uninephrectomized	6.1
		transcriptase	6.0

A number of public databases have a similar goal of identifying gene function but approach it with a different philosophy than that of our project. Usually, these databases attempt to simplify the knowledge about a gene's function down to a single concept, so that the functions can be understood simply and quickly. The investigator then has a better chance of comprehending the function of many genes in a single session, and thereby construct

their own overview of the relationships amongst the genes in question.

Table 1: Keywords shared by two or more genes

# of genes	Keyword	associated gene names			
4	cytoplasmic	H	S	GOT	OPN
4	tumors	C	H	OPN	UK
3	endothelial	S	N	OPN	
3	monoclonal	H	S	OPN	
3	peritubular	C	OPN	UK	
2	antimesometrial	C	OPN		
2	arginine	H	OPN		
2	bulls	GOT	OPN		
2	cathepsin	H	S		
2	chorioamnion	C	OPN		
2	cosecretion	H	OPN		
2	coumarylamide	H	S		
2	deafferented	S	N		
2	decalcified	C	OPN		
2	dermis	C	N		
2	entorhinal	C	S		
2	immunogold	C	OPN		
2	intraarticular	N	OPN		
2	leupeptin	H	S		
2	lowicryl	H	OPN		
2	lysosomal	S	N		
2	mesometrial	C	OPN		
2	mitogens	N	OPN		
2	OPN	C	OPN		
2	postischemia	C	OPN		
2	pyruvate	H	GOT		
2	resected	C	OPN		
2	stefin	H	S		
2	thapsigargin	C	OPN		
2	tunel	C	OPN		
2	whitney	C	OPN		

This approach creates for the investigator a burdensome challenge, the development of a functional network of genes using knowledge that has been extensively pruned. The philosophy of our approach differs in that we do not wish to limit the rich information available, but rather take advantage of the speed and capacity of computers to represent and process the many concepts about functions and features of individual genes. We then apply algorithms to identify "threads" that connect the genes to each other and automatically generate a functional network. The investigator could then evaluate that network, use it to access information in the literature, gain insight, and plan new experiments.

Keywords extracted by the system

To find out the relevance of the keywords for a gene, one investigator (BJC) inspected a word list for osteopontin (OPN, Query Set #1) to select keywords with z-scores above 2.0 and filter out

general or methodological words (essentially all non-functional words related to methodology, e.g. cDNA, polyclonal, chromatography, escherichia, coli, histology, lysates, Sepharose, clone, biotinylated, recombinant, nmr, hybridization, densitometric, luciferase, polyacrylamide, immunogold, immunostaining, immunohistochemistry).

Table 2: Information on OPN Extracted from Various Internet Gene Resources

Resource	Information
Gene	cell adhesion
Ontology:	cell adhesion molecule ossification
Gene-Cards:	alternate/ related names: osteopontin precursor bone sialoprotein 1 urinary stone protein secreted phosphoprotein 1 SPP-1 nephropontin uropontin gene: SPPI or OPN composition: 314 amino acids molecular weight: 35 kD function: binds tightly to hydroxyapatite; appears to form an integral part of the mineralized matrix; probably important to cell-matrix interaction. subunit: ligand for integrin alpha-v/beta-3. alternative products: 3 isoforms are produced by alternative splicing: a/opn-a/op1b, b/opn-b/op1a, and c/opn-c. posttrans. modifications: extensively phosphorylated on serine residues. N- and O-glycosylated. similarity: belongs to the osteopontin family.
SwissProt:	[All of the above info from GeneCards is available in the human osteopontin entry for SwissProt, which also included the following fact:] Disease: this protein plays a principal role in urinary stone formation as the stone matrix.

GenBank (Keywords field)

bone phosphoprotein; bone sialoprotein; calcium binding protein; cell adhesion phosphoprotein; extracellular matrix Protein; hydroxyapatite-binding protein; integrin-binding protein; matrix protein; mOP; osteopontin; phosphoprotein; secreted phosphoprotein; sialoprotein; sialoprotein I; SPP1 gene; SPPI protein; structural protein; tumor-associated phosphoprotein; hOP.

The relevance of keywords for OPN function was determined by searching the query set of abstracts for their occurrence and reading the abstracts. Virtually every keyword was found to have at least one highly relevant meaning in the context of the OPN literature (results not shown).

Extraction of new information

The keyword list results using query set #2 showed that our system was able to identify keywords associated with newly discovered functions of OPN (Table 1). For example, our algorithms can identify the keywords and their associated z-scores captopril 2.3 (not shown), losartan 7.9, and atherosclerosis 4.4 (not shown) after the possible role of OPN in hypertension [9]. Similarly, a functional link between OPN and autoimmune demyelinating diseases [10] is suggested by the keywords demyelinating 2.1 (not shown), encephalomyelitis 2.9 (not shown), autoantigen 2.6 (not shown), and autoimmune 6.2, whereas a link to tumor metastasis [11] is pointed to by the keywords catenin 5.3 (not shown), cadherin 3.3 (not shown), and tumorigenic 4.4 (not shown). [Table 1 has words in descending order of Z score; hence the words “not shown” did not make the list.]

Comparison of keyword lists with existing ontologies and resources

Besides MEDLINE (PubMed), there are several other resources which are available over the Internet that contain useful information regarding the specific functions of genes, for example, the Gene Ontology (GO) Consortium, SwissProt, GenBank, and GeneCards. These databases necessarily reduce vast literature into a few functional concepts, whereas the algorithm-derived keywords often convey a much broader sense of the functions of genes. Using the osteopontin (OPN) gene as an example, we manually extracted all available functional information on OPN from these resources in January 2002. The extracted results are presented in Table 3. For OPN, the three GO keywords represent functional concepts, whereas the 19 words in GenBank are mostly aliases or biochemical descriptions for OPN. The GeneCards and SwissProt information are essentially the same and contain aliases, general characteristics and functional information. Taken together, a number of biological concepts regarding OPN are represented in these various databases; however, individually, there are certainly gaps in discrete topics for each database. For example, as of April 30, 2003, none of these databases identified the possible role of OPN in hypertension [9], tumor metastasis [11], or in autoimmune demyelinating disease [10]. Finally, we searched Gene Ontology for the other gene names that we used to generate our preliminary data and found: 9 keywords for nexin, 0 keywords for cathepsin H and cathepsin S, 3 keywords for calcyclin, and no entries for glutamate-oxaloacetate transaminase or uridine monophosphate kinase. Therefore, we conclude that these popular public databases are useful, but individually and collectively incomplete. This example indicates that our statistical algorithm can extract many relevant keywords, a number of which point to biological concepts not found in the existing public gene databases.

Keyword sharing

Keywords shared among genes appeared to be highly informative about the general function of the genes (Table 2).

The results indicate that the keyword lists could be used to cluster genes. Liu et al. [7] selected 26 genes in four well-defined functional groups consisting of 10 glutamate receptor subunits, 7 enzymes in dopamine metabolism, 5 cytoskeletal proteins and 4 enzymes in tyrosine synthesis as the query set. Keyword lists were generated for each of these 26 genes in four well-defined functional groups and the resulting word-by-gene matrix was converted to a symmetrical gene-by-gene matrix. A bond energy clustering algorithm [13, 14] correctly assigned 25 of 26 genes to the appropriate cluster based on the strength of keyword associations. The results support the rationale of using extracted keywords to characterize gene function and shared keywords to cluster a set of genes.

Conclusion

We designed a text mining system based on a statistical algorithm. The results showed that our system could extract functional keywords for the genes identified by expression profiling with DNA microarrays. Furthermore, the system can extract newly discovered information that cannot be found from other sources online. Stop-lists and background sets will need to be improved in order to remove “new” words and the non-functional, methodologically related words from the resulting keyword list. The functional keywords extracted by the system should be useful for clustering genes using the corresponding z-scores as measures of the relative strength of association [7].

Acknowledgements

This work was supported by NINDS (RD) and the Emory-Georgia Tech Research Consortium. We would like to thank Brian Revenaugh for computer administration and technical support and to Prof. Ashwin Ram, and graduate student Jorge Civera-Saiz for their contributions.

References

- [1] Andrade M, and Valencia A. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families, *Bioinformatics* 1998; 14: 600-607.
- [2] <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>
- [3] Andrade MA, and Bork P. Automated extraction of information in molecular biology. *FEBS Letters* 2000;476:12-17.
- [4] Jenssen TK, Laegreid A, Komorowski J, and Hovig E. A literature network of human genes for high-throughput analysis of gene expression *Nature Genetics* 2001;28: 21-28.
- [5] Masys DR, Welsh JB, Lynn FJ, Gribskov M, Klacansky I, and Corbeil J. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics* 2001;17: 319-26.
- [6] Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, and Weinstein JN. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques* 1999;27: 1210-1214.
- [7] Liu Y, Pivoshenko A, Civera J, Brandon M, Ram A, Navathe S, Ciliax BJ, and Dingledine R. *Clustering genes by functional keywords from biomedical literature*. Technical Report, College of Computing, Georgia Tech.
- [8] http://ftp.ncbi.nih.gov/refseq/LocusLink/LL_tmpl.gz.
- [9] Hartner A, Porst M, Gauer S, Prots F, Veelken R, and Hilgers KF. Glomerular osteopontin expression and macrophage infiltration in glomerulosclerosis of DOCA-salt rats. *Am J Kidney Dis* 2001;38: 153-164.
- [10] Chabas D, Baranzini SE, Mitchell D, Bernard CCA, Ritling SR, Denhardt DT, Sobel RA, Lock C, Karpuz M, Pedotti R, Heller R, Oksenberg JR, Steinman L. The influence of the proinflammatory cytokine, osteopontin, on autoimmune demyelinating disease. *Science* 2001;294: 1731-1735.
- [11] Furger KA, Menon RK, Tuck AB, Bramwell VH, and Chambers AF. The functional and clinical roles of osteopontin in cancer and metastasis. *Curr Mol Med* 2001;1: 621-632.
- [12] <http://ftp.std.com/obi/Dictionary/dict>
- [13] McCormic WT, Schweitzer PJ, and White TW Problem decomposition and data reorganization by a clustering technique. *Oper. Res.* 1972;20: 993-1009.
- [14] Navathe S, Ceri S, Wiederhold G, and Dou J. Vertical partitioning algorithms for database design. *ACM Trans. On Database Systems* 1984;9: 680-710.

Address for correspondence

Ying Liu
College of Computing
Georgia Institute of Technology
Atlanta, Ga 30332-0280
E-mail: yingliu@cc.gatech.edu
Tel: 404-808-0655