LITSEEK: Public Health Literature Search by Metadata Enhancement with External Knowledge Bases

Priyanka Prabhu[1]
priyanka.prabhu@gatech.edu

Shamkant Navathe[1]
sham@cc.gatech.edu

Stephen Tyler[1]
st144@mail.gatech.edu

1 College Of Computing
Georgia Institute Of Technology
Atlanta, GA 30332-0280
Tel #: +1- 404-385-2892

ABSTRACT

Biomedical literature is an important source of information in any researcher's investigation of genes, risk factors, diseases and drugs. Often the information searched by public health researchers is distributed across multiple disparate sources that may include publications from PubMed, genomic, proteomic and pathway databases, gene expression and clinical resources and biomedical ontologies. The unstructured nature of this information makes it difficult to find relevant parts from it manually and comprehensive knowledge is further difficult to synthesize automatically. In this paper we report on LITSEEK (LITerature Search by metadata Enhancement with External Knowledgebases), a system we have developed for the benefit of researchers at the Centers for Disease Control (CDC) to enable them to search the HuGE (Human Genome for Epidemiology) database of PubMed articles, from a pharmacogenomic perspective. Besides analyzing text using TFIDF ranking and indexing of the important terms, the proposed system incorporates an automatic consultation with PharmGKB - a human-curated knowledge base about drugs, related diseases and genes, as well as with the Gene Ontology, a human-curated, well accepted ontology. We highlight the main components of our approach and illustrate how the search is enhanced by incorporating additional concepts in terms of genes/drugs/diseases (called metadata for ease of reference) from PharmGKB. Various measurements are reported with respect to the addition of these metadata terms. Preliminary results in terms of precision based on expert user feedback from CDC are encouraging. Further evaluation of the search procedure by actual researchers is under way.

Categories and Subject Descriptors

H.2.4 [Database Management Systems] –Textual Databases H.3.3 [Information Search and Retrieval] – *Information Filtering, Search Process, Selection*

General Terms

Our general terms are: Algorithms, Design.

Keywords

Metadata integration, text mining, information retrieval, pharmacogenomics, knowledge bases, search.

DTMBIO'09, November 6, 2009, Hong Kong, China.
ACM 978-1-60558-803-2/09/11.

1. ENHANCEMENT OF BIOMEDICAL LITERATURE SEARCH

1.1 The Problem

Biomedical knowledge search typically requires manual use of external knowledge bases (like ontologies and databases) in conjunction with bibliographic literature to link and relate knowledge for research purposes. Particularly, there is a specific information need about genes, diseases and drugs and their mutual relationships after one views articles from systems like PubMed. For example, a user searching for a specific gene BRCA1 may not find articles containing the alternate gene names of BRCA1. Also, he may be interested in articles containing the drugs and diseases associated with the BRCA1 gene. However, these articles may not necessarily contain the term BRCA1.

The challenges in this context are as follows:

i) Identification of genes, drugs and diseases from the unstructured natural language text.

ii) Linking the entities of interest found in step 1 with external information about them.

iii) Integration of structured (relational) and unstructured (plain text) data from multiple sources.

1.2 The Solution

We propose to solve this by integrating actual data and the specific information of interest from multiple external sources with zero manual effort. Our solution is based on metadata enhancement and query expansion where an integrated search engine aims to improve the search experience for a researcher. LITSEEK integrates itself with pharmacogenetic knowledge bases, namely the PharmGKB database used for query expansion and it provides further knowledge from Gene Ontology [Ashburner et al. 2000.]. We use the HuGE dataset [Yu et al. 2008.] as a literature database which has been compiled and curated by scientists at the CDC for epidemiological research. The HuGE database has about 20,000 articles currently. We have developed an automated classifier approach [Polavarapu et al. 2005] to assist a human expert select articles for HuGE from PubMed. The primary features of our search system are as follows:

1. For a given query term, we give an option to the user to retrieve the primary set of articles strictly based on (i) only Titles or (ii) Title and Abstracts.

2. For each article in the primary set, we expand the query with a set of related genes, drugs, and diseases (together called metadata) by consulting PharmGKB.

3. The search granularity (explained in Section 3) determines the fields of the article and metadata that should be searched for the query. Each result contains the article and its metadata. The relevant concepts from the three domains (cellular component, molecular function and biological process) from Gene Ontology are attached to the articles in the result set for further insight.

4. The system is parameterized where the thresholds for determining what is a gene from GAPSCORE (explained later) or the number of alternate genes to look up from PharmGKB etc. can be user-controlled.

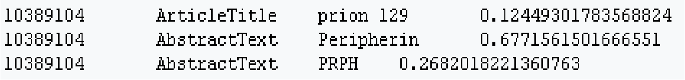
2. APPROACH AND METHODS FOR METADATA INTEGRATION

This section provides an overview of our approach to enhancing literature search with metadata. Throughout the rest of this paper, we have used the term 'article' to mean the PubMed Abstract which has a title and abstract. It does not explicitly include any metadata such as gene names, drug names, or disease names. As suggested by [Meij et al. 2005], biomedical literature search can be expanded by using thesauri-like methods. The major advantage we have is that by using a humanly curated database like PharmGKB, and ontologies constructed by domain experts, like the GO ontology, we are in a better position to assure that while recall increases, precision will not suffer..

The various steps in the working of LITSEEK are as follows:

Step 1: The dataset (given a list of PubMed Ids (PMIDs) for the HuGE database) is fetched using NCBI's EFetch utility. We produce a Lucene (http://lucene.apache.org/) document (in Lucene terminology) which is indexed on the following fields: PMID, ArticleTitle and ArticleText. Internally, Lucene also creates an inverted index on terms for fast retrieval. This data is parsed and an index is created using Lucene; it is called the primary index.

Step 2: GAPSCORE [Chang et al. 2004] uses a machine-learning based approach and returns the likely genes/proteins from an input sentence and their likelihood score (between 0-1 for increasing confidence). We prefer GAPSCORE since it is used for named entity recognition of genes and proteins as compared to more general named entity recognizers like MetaMap Transfer MMTx [Osborne et al 2007.]. GAPSCORE is run on the Lucene index. Thus, for each article we store the PMID, the location where the term was found, title (called ArticleTitle) or text of article (called ArticleText), and the term which has been identified by GAPSCORE as gene/ protein and their scores. For example, in Fig 1, Peripherin, rather than prion 129 and PRPH is more likely to be a gene. As a filtering step, we consider only those tokens with a GAPSCORE exceeding an empirical threshold of 0.5.



```
10389104      ArticleTitle    prion 129       0.12449301783568824
10389104      AbstractText    Peripherin      0.6771561501666551
10389104      AbstractText    PRPH    0.2682018221360763
```

Fig 1: Use of GAPSCORE for Gene/ Protein Recognition

Step 3: The threshold subjected entities identified by GAPSCORE are populated in the gene_protein table which contains PMID, location of the term in the article, the gene/protein term and its GAPSCORE.

Step 4: PharmGKB provides flat file data about genes, drugs and diseases. We convert this into a relational schema: drug table (PharmGKB Id, Name, Alternate Name), the gene table (PharmGKB Id, Entrez Id, Name, Symbol, Alternate Names, Alternate Symbols) and disease table (PharmGKB Id, Name, Alternate Name).

The gene/protein names from Step 3 are searched in the PharmGKB derived gene table. Since PharmGKB identifies only genes, this step makes sure that proteins are filtered from the gene_protein table using string matching. The result of the match still contains some terms like "polymerase" which are related to too many genes, drugs and diseases. This is the equivalent of a stop word for the domain. As another heuristic parameter, we remove any GAPSCORE text which references more than 10 genes. We then load these results into the pmidgene table (pmidgeneid, pmid, field, name, gapscore, geneid) that links the genes that occur in the dataset articles to the information about them in PharmGKB with their pmids (which uniquely reference the articles) and geneids (which uniquely reference the genes provided by PharmGKB).

Step 5: The relationship data in PharmGKB is unstructured as text records and requires the tables from step 4 for named entity recognition to identify the relationship types. Our schemas for relationship tables are as follows: relationship (relationshipid, pharmgkbid, relationshipType), relationshipgene (relationshipid, geneid), relationshipdisease (relationshipid, diseaseid) and relationshipdrug (relationshipid, drugid). Our LITSEEK internal design thus translates unstructured data into structured form and can incorporate data from any source (in this case PharmGKB).This relationship database gives metadata about related genes, drugs and diseases for the genes identified in Step 4.

Step 6: Data (title and abstract) from Step 1 and metadata (genes, drugs, diseases related to the contents of the title and abstract) is collected from Step 5.

.**Step 7:** The gene name/symbol from our 'pmidgene' table (from step 4) is matched with the 'gene_product' table from GO. Then a join with the 'association' table from GO gives the associated terms. This establishes a link between the pmidgene table (which contains references of articles that contain that specific gene) and the associated terms in GO. This association helps us infer which articles are associated with which GO entries. We restrict GO to the species 'Homo sapiens' and include all the three domains in GO: cellular component, biological process and molecular function. We associate the subset of articles appropriate to a concept in these three domains if a link is possible between that concept and the gene pertaining to that article. Thus for every such related concept in GO, we are able to associate exactly the articles in the retrieved set for the given query term with or without expansion.

Step 8: The XML file fetched for each given PubMed Id from NCBI's EFetch utility contains several additional fields like MESH headings, dates, authors etc in additional to the basic article we have considered for LITSEEK. Such associations are used for filtering and refining the search results further. The details are omitted here.

3. LITSEEK SEARCH ALGORITHM

The algorithm for LITSEEK as a search engine is shown in the flowchart in Fig 2. In the Figure, **Search Granularity** refers to the fields of the article that should be searched for a match with the query. The values for this are "Title", "Title and Abstract" or "All fields" (Title, abstract and metadata fields). For search granularity value of "Title", Lucene only searches the query for a match in the title of the articles present in the HuGE set. At the other extreme, for "All Fields", the Lucene Search is over the title, abstract and metadata for each article.

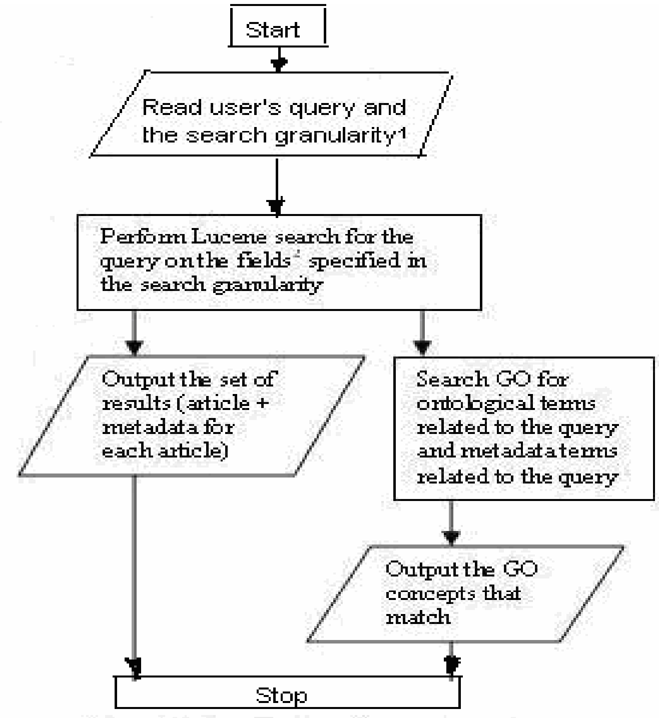
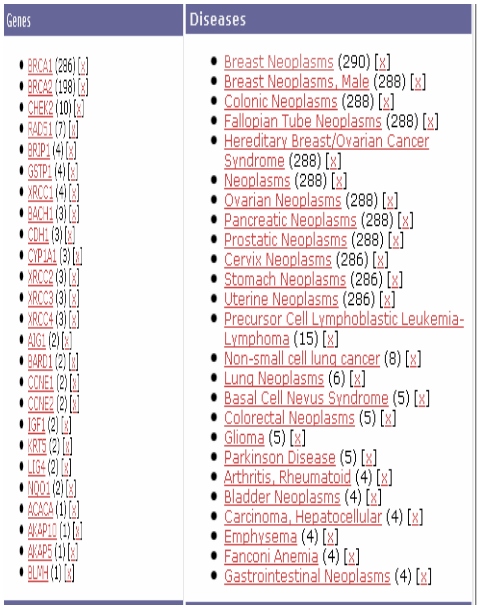

Start
Read user's query and the search granularity1
Perform Lucene search for the query on the fields1 specified in the search granularity
Output the set of results (article + metadata for each article)
Search GO for ontological terms related to the query and metadata terms related to the query
Output the GO concepts that match
Stop

Fig 2 Processing User Query with LITSEEK

This flowchart shows processing of user's query/search request with LITSEEK. The figure shows the outputs generated by the use of both PharmGKB and GO. Consider a sample query "BRCA1". The related genes we get are shown in Figure 3 and the retrieved related diseases and drugs are shown in Figures 4 and 5. Here the numbers indicate the number of articles in our dataset that contain the gene (Fig.3). The user can filter out any results by clicking on the "X" button corresponding to the listed items; that removes all the articles associated with the gene in the pmidgene table.



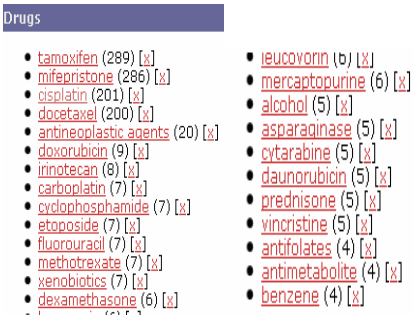
Genes

- BRCA1 (286) [x]
- BRCA2 (198) [x]
- CHEK2 (10) [x]
- RAD51 (7) [x]
- BRIP1 (4) [x]
- GSTP1 (4) [x]
- XRCC1 (4) [x]
- BACH1 (3) [x]
- CDH1 (3) [x]
- CYP1A1 (3) [x]
- XRCC2 (3) [x]
- XRCC3 (3) [x]
- XRCC4 (3) [x]
- AIG1 (2) [x]
- BARD1 (2) [x]
- CCNE1 (2) [x]
- CCNE2 (2) [x]
- IGF1 (2) [x]
- KRT5 (2) [x]
- LIG4 (2) [x]
- NQO1 (2) [x]
- ACACA (1) [x]
- AKAP10 (1) [x]
- AKAP5 (1) [x]
- BLMH (1) [x]

Diseases

- Breast Neoplasms (290) [x]
- Breast Neoplasms, Male (288) [x]
- Colonic Neoplasms (288) [x]
- Fallopian Tube Neoplasms (288) [x]
- Hereditary Breast/Ovarian Cancer Syndrome (288) [x]
- Neoplasms (288) [x]
- Ovarian Neoplasms (288) [x]
- Pancreatic Neoplasms (288) [x]
- Prostatic Neoplasms (288) [x]
- Cervix Neoplasms (286) [x]
- Stomach Neoplasms (286) [x]
- Uterine Neoplasms (286) [x]
- Precursor Cell Lymphoblastic Leukemia-Lymphoma (15) [x]
- Non-small cell lung cancer (8) [x]
- Lung Neoplasms (6) [x]
- Basal Cell Nevus Syndrome (5) [x]
- Colorectal Neoplasms (5) [x]
- Glioma (5) [x]
- Parkinson Disease (5) [x]
- Arthritis, Rheumatoid (4) [x]
- Bladder Neoplasms (4) [x]
- Carcinoma, Hepatocellular (4) [x]
- Emphysema (4) [x]
- Fanconi Anemia (4) [x]
- Gastrointestinal Neoplasms (4) [x]

Figures 3, 4 show the genes, diseases related to the query term "BRCA1"



Drugs

- tamoxifen (289) [x]
- mifepristone (286) [x]
- cisplatin (201) [x]
- docetaxel (200) [x]
- antineoplastic agents (20) [x]
- doxorubicin (9) [x]
- irinotecan (8) [x]
- carboplatin (7) [x]
- cyclophosphamide (7) [x]
- etoposide (7) [x]
- fluorouracil (7) [x]
- methotrexate (7) [x]
- xenobiotics (7) [x]
- dexamethasone (6) [x]
- leucovorin (6) [x]
- mercaptopurine (6) [x]
- alcohol (5) [x]
- asparaginase (5) [x]
- cytarabine (5) [x]
- daunorubicin (5) [x]
- prednisone (5) [x]
- vincristine (5) [x]
- antifolates (4) [x]
- antimetabolite (4) [x]
- benzene (4) [x]

Fig 5 shows the drugs related to the query term "BRCA1"

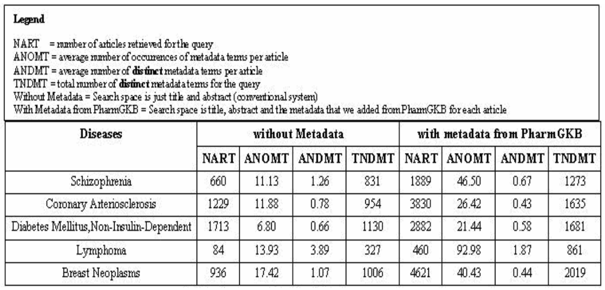
4. COMPARISON WITH RELATED WORK

GoPubMed [Doms et al. 2005] builds on PubMed and Gene Ontology and categorizes the search results to answer - *"What"*, *"Who"*, *"Where"* and *"When"* in response to user's querywith relevance ranking. EBIMed [Rebholz-Schuhmann et al 2006] analyzes Medline articles for associations between UniProt protein/gene names, Gene Ontology (GO) annotations, Drugs and Species. GoPubMed and EBIMed display relationships based on the proximity of entities in the articles, and do not focus on providing information about relationships between genes, drugs and diseases beyond what is implicit from the dataset itself, [Maojo et al. 2006] describes an ontology based approach for integrating biomedical information through mapping and unification. [Maojo et al. 2007] advocates the use of web services for linking genomic data to medical information systems. BioMOBY [Wilkinson et al. 2005], myGrid (http://www.mygrid.org.uk/) , caBIG (https://cabig.nci.nih.gov/) are major integration efforts to integrate a variety of databases to support certain applications related to problem domains like cancer. However, the goals of our work are different from them in that although we bring into our framework multiple databases such as HuGE, PharmGKB and GO, our focus here is not on the data integration aspect but on improving search using metadata.

5. EVALUATION OF USE OF METADATA

We measured the total amount of relevant information retrieved by our system with and without the query expansion process aided by retrieving related metadata (in the form of genes, drugs and diseases) from PharmGKB. This is measured in terms of the number of articles (indicative of recall) and the total number of terms that were retrieved related to the search query. The query terms were chosen in consultation with Public health experts at CDC. Table 1represents the related results for query terms that are diseases. Similar tables for query terms that are drugs and genes are not shown for space reasons. Here, we find that for a large number of cases, the proposed system retrieves a substantially higher number of articles (NART) with query expansion than without. This indicates that the number of retrieved documents related to the query is increased when metadata is added.

Table 1: Comparison for Disease names



Legend

NART = number of articles retrieved for the query
ANOMT = average number of occurences of metadata terms per article
ANDMT = average number of **distinct** metadata terms per article
TNDMT = total number of **distinct** metadata terms for the query
Without Metadata = Search space is just title and abstract (conventional system)
With Metadata from PharmGKB = Search space is title, abstract and the metadata that we added from PharmGKB for each article

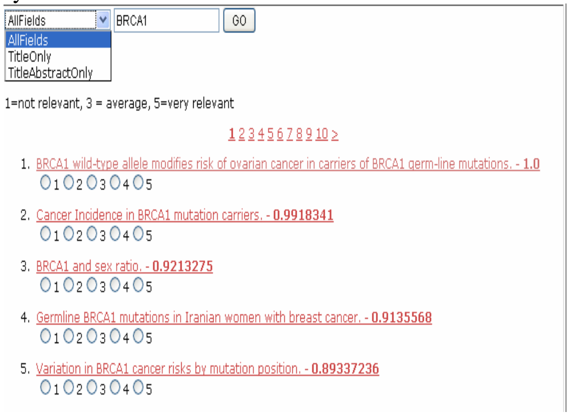
Diseases	without Metadata				with metadata from PharmGKB			
	NART	ANOMT	ANDMT	TNDMT	NART	ANOMT	ANDMT	TNDMT
Schizophrenia	660	11.13	1.26	831	1889	46.50	0.67	1273
Coronary Arteriosclerosis	1229	11.88	0.78	954	3830	26.42	0.43	1635
Diabetes Mellitus,Non-Insulin-Dependent	1713	6.80	0.66	1130	2882	21.44	0.58	1681
Lymphoma	84	13.93	3.89	327	460	92.98	1.87	861
Breast Neoplasms	936	17.42	1.07	1006	4621	40.43	0.44	2019

We also compiled similar results for query terms that are biological pathways (namely, coagulation, glucose metabolism, biotransformation, rejection, inflammatory) and clinical signs (obesity, depression, cardiometabolic risk, dyslipidemia, insulin resistance). We got modest improvements in number of articles received for most (e.g., *inflammatory* went up from 1131 to 1222) and a fairly large improvement for some (e.g., *obesity* went up from 678 to 1124). This may be attributed to the fact that the metadata used from PharmGKB contains relationships only between genes, drugs, and diseases and hence, the improvements for terms concerning clinical signs and biological pathways are not as significant as they are for genes, drugs and diseases.

The tables also represent the average number of occurrences of metadata terms per article (ANOMT). We find that LITSEEK retrieves more metadata terms per article when the search space includes the title, abstract and even the metadata section. This indicates that our approach is retrieving articles which are related to the query conceptually, without actually having the queried term matched in the title or the abstract. The relatedness is because PharmGKB has established the relationships between genes, drugs and diseases in a human-curated way and we identify how to link this to the title and abstract content by using GAPSCORE and our relational tables.

6. SEARCH QUALITY EVALUATION WITH HUMAN SUBJECTS

In order to evaluate how effectively our system returns the articles of interest and related metadata information to a typical user, we propose that the evaluators choose the type of the search (AllFields, TitleOnly or TitleAbstractOnly) and specify the query term. They can choose any terms, but typically names of drugs, genes and diseases may be of particular interest. Figure 6 shows the returned results of the query "BRCA1" – top 20 ranked results are returned on the first screen. Only the first five are shown. The score indicated after each result in this figure is calculated by using the Lucene Similarity Measure (http://lucene.apache.org/java/2_4_0/api/org/apache/lucene/search/Similarity.html). The expert user would score each result on a 5 point scale where 5 is very relevant and 1 is not relevant. The system records and tabulates these scores for each evaluator.



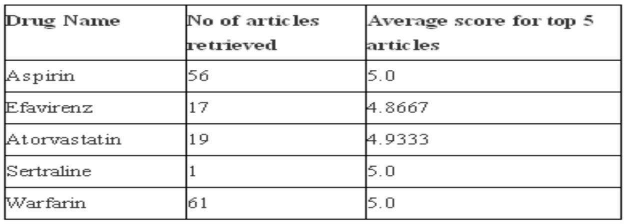
1=not relevant, 3 = average, 5=very relevant

1 2 3 4 5 6 7 8 9 10 >

1. BRCA1 wild-type allele modifies risk of ovarian cancer in carriers of BRCA1 germ-line mutations. - **1.0**
   ○1 ○2 ○3 ○4 ○5
2. Cancer Incidence in BRCA1 mutation carriers. - **0.9918341**
   ○1 ○2 ○3 ○4 ○5
3. BRCA1 and sex ratio. - **0.9213275**
   ○1 ○2 ○3 ○4 ○5
4. Germline BRCA1 mutations in Iranian women with breast cancer. - **0.9135568**
   ○1 ○2 ○3 ○4 ○5
5. Variation in BRCA1 cancer risks by mutation position. - **0.89337236**
   ○1 ○2 ○3 ○4 ○5

Fig 6 Ranked Results for Query BRCA1 using the AllFields Search Granularity

Our system has been tested by 3 expert users (a senior scientist at CDC who is M.D. M.P.H., a Ph.D. scientist and an M.P.H. scientist) using commonly prescribed drug names that yielded the results shown in Fig 7. They scored the top 5 articles in each category by retrieving them using the AllFields option which has the query expansion built in.

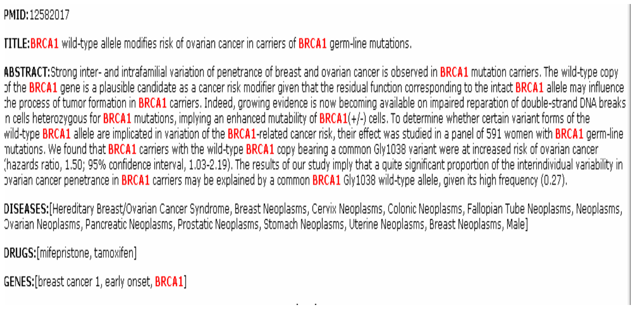


Drug Name	No of articles retrieved	Average score for top 5 articles
Aspirin	56	5.0
Efavirenz	17	4.8667
Atorvastatin	19	4.9333
Sertraline	1	5.0
Warfarin	61	5.0

Fig 7: Results of retrieval & evaluation scores by an expert user.

Thus the average score of the resulting top 5 documents for the 5 drugs is 4.96 and standard deviation is 0.0533. Future plans include a field study with a team of expert users.

Fig 8 shows the top ranked abstract and the metadata from PharmGKB that is associated with it.



PMID:12582017

TITLE:BRCA1 wild-type allele modifies risk of ovarian cancer in carriers of BRCA1 germ-line mutations.

ABSTRACT: Strong inter- and intrafamilial variation of penetrance of breast and ovarian cancer is observed in BRCA1 mutation carriers. The wild-type copy of the BRCA1 gene is a plausible candidate as a cancer risk modifier given that the residual function corresponding to the intact BRCA1 allele may influence the process of tumor formation in BRCA1 carriers. Indeed, growing evidence is now becoming available on impaired reparation of double-strand DNA breaks in cells heterozygous for BRCA1 mutations, implying an enhanced mutability of BRCA1(+/-) cells. To determine whether certain variant forms of the wild-type BRCA1 allele are implicated in variation of the BRCA1-related cancer risk, their effect was studied in a panel of 591 women with BRCA1 germ-line mutations. We found that BRCA1 carriers with the wild-type BRCA1 copy bearing a common Gly1038 variant were at increased risk of ovarian cancer (hazards ratio, 1.50; 95% confidence interval, 1.03-2.19). The results of our study imply that a quite significant proportion of the interindividual variability in ovarian cancer penetrance in BRCA1 carriers may be explained by a common BRCA1 Gly1038 wild-type allele, given its high frequency (0.27).

DISEASES:[Hereditary Breast/Ovarian Cancer Syndrome, Breast Neoplasms, Cervix Neoplasms, Colonic Neoplasms, Fallopian Tube Neoplasms, Neoplasms, Ovarian Neoplasms, Pancreatic Neoplasms, Prostatic Neoplasms, Stomach Neoplasms, Uterine Neoplasms, Breast Neoplasms, Male]

DRUGS:[mifepristone, tamoxifen]

GENES:[breast cancer 1, early onset, BRCA1]

Fig 8: Shows the article and metadata (diseases, drugs, genes related to the query) for the top ranked result for the query "BRCA1" using the AllFields Search Granularity. Note Diseases, Drugs and Genes reported to the user.

7. CONCLUSION

The focus of our work is to present users with more related results for the query input as well as enrich the information for each abstract with related genes, drugs and diseases without manual effort We also allow some user control over the behavior of the search process, e.g., the GAPSCORE relevance threshold of 0.5 for extracting gene names can be changed by the user. The maximum number of gene entries for some common words is currently limited at 10. The grouping of genes can be user-controlled as well. This approach can be used to improve search quality of literature in any domain where metadata enhancements such as those we get from PharmGKB and Gene Ontology would be an added advantage. In this paper we have attempted to show how we can integrate information from a relatively less structured XML database of literature with available metadata from other unstructured databases (e.g. PharmGKB) which are mostly textual and ontology sources (e.g., Gene Ontology) by using web services and intermediate tools (like GAPSCORE). We have used the relational model as an intermediate vehicle to have better control on searches by transforming them to SQL queries that would eventually help the end-user with their targeted goals.

8. FUTURE WORK

.In future we propose to integrate several databases which provide more exhaustive lists of genes, drugs and diseases and try to explore relationships between them and then to investigate its effects on the metadata enhancement and query expansion for bibliographic search. The schema we have used is generic so that in future it can contain data (genes / drugs / diseases and their source reference – currently PharmGKB Accession Id) from other databases. The protein names in the "gene" table are presently not used for any external database access to enhance the metadata. Some protein databases such as Swiss-Prot could be used for such enhancement.

The schema we have used is generic so that in future it can contain data (genes / drugs / diseases and their source reference – currently PharmGKB Accession Id) from other databases. This facilitates finding all the relevant information from a single database source instead of searching multiple disparate and possibly unstructured sources. Our current experiments are done with HuGE, a subset of PubMed geared for epidemiology. Similar subsets of literature for other disease categories such as different types of cancer or gene families can be combined with external knowledge sources to develop additional "enhanced search experience" systems.

9. ADDITIONAL AUTHORS

Venu Dasigi[2,] Neha Narkhede[1], Balaji Palanisamy[1],

[1]College of Computing, Georgia Institute Of Technology, Atlanta, GA 30332-0280

[2]Dept. of Computer Science & Software Engineering, Southern Polytechnic State University, Marietta, GA 30060-2855

10. ACKNOWLEDGEMENTS

We thank Drs. Marta Gwinn, Wei Yu, and Muin Khoury and Mr. Ajay Yesupriya of CDC for their inputs and encouragement. We appreciate the contributions of Lalit Kapoor, Saurav Sahay and Orlando Karam at different phases of this project. This work has been supported by a grant from CDC (Grant Number CDC-RFA-GD06-601).

11. REFERENCES

[1] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics,* 2000 May; 25(1):25-9.

[2] Chang J.T, Schütze H., Altman R.B, "GAPSCORE: finding gene and protein names one word at a time ," *Bioinformatics*, 2004 Jan 22; 20(2):216-25.

[3] Doms A. and Schroeder M., "GoPubMed: Exploring PubMed with the GeneOntology, " *Nucleic Acid Research*, 33 *(Web Server Issue)* July 1, 2005: W783—W786,

[4] Hodge A, Altman R, Klein T, "The PharmGKB: integration, aggregation, and annotation of pharmacogenomic data and knowledge," *Clinical Pharmacology & Therapeutics*, 2007 Jan; 81(1):21-4.

[5] Maojo V., Crespo J., de la Calle G., Barreiro J., Garcia-Remesal M.,"Using web services for linking genomic data to medical information systems," *Methods of information in medicine*, 2007;46(4):484-92.

[6] Maojo V., García-Remesal M., Billhardt H., Alonso-Calvo R., Pérez-Rey D., Martín-Sánchez F, "Designing new methodologies for integrating biomedical information in clinical trials," *Methods of information in medicine*, 2006; 45(2):180-5.

[7] Meij, E., Ijzereef, L., Azzopardi, L., Kamps, J., de Rijke, M., "Combining Theasauri-based Methods for Biomedical Retrieval " *Proc. 14th TREC Conference* (TREC 2005)

[8] Osborne J.D., Lin S., Zhu L., Kibbe W. A. "Mining biomedical data using MetaMap Transfer (MMtx) and the Unified Medical Language System (UMLS)," *Methods in molecular biology (Clifton, N.J.)*, Vol. 408 (2007), pp. 153-169.

[9] Polavarapu,N., Navathe, S.B., Ramnarayanan, R., Abrar ul Haque, Sahay, S. and Ying Liu: "Investigation into Biomedical Literature Screening Using Support Vector Machines," *Proc. IEEE Computational Systems Bioinformatics Conference (CSB'05)* , Stanford, Calif., Aug 2005

[10] Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Rynbeek M, Stoehr P "Protein annotation by EBIMed," *Nature biotechnology*, 2006 Aug; 24(8):902-3.

[11] Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. "A navigator for human genome epidemiology," *Nature Genetics*. 2008 Feb; 40(2):124-5. PMID: 18227866.

[12] Wilkinson M, Schoof H, Ernst R, Haase D "BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case," *Plant Physiology,* 2005 May: 138(1):5-17.