Ranking and Presenting Gene-Disease Associations from Biomedical Literature

Giun Sun[1], Summer Intern
William Lau[1], Alex Wang[1], Narmada Shenoy[2], Kevin Becker[2], Huey Cheung[1]
[1]Center for Information Technology [2]National Institute on Aging

Abstract

*Pubmatrix[1] is a Web-based text mining application that allows users to search MEDLINE data using two sets of keyword terms, resulting in a frequency matrix of term co-occurrence. In this project, we have made several enhancements to improve the functionality of PubMatrix and have named it PubMatrix SE. First, ontology- and rule-based gene normalization is performed on each citation to more accurately detect the appearance of human genes and their direct products. Second, the search results can be expanded based on a custom hierarchy defined by the users to show finer details of the associations, thereby allowing them to explore relationships involving more than two entities. Lastly, we have developed a scoring metric using the term-frequency/inverse-document-frequency (TF * IDF) weighing scheme to estimate the strength and significance of the associations presented in the matrix. The system facilitates more efficient interpretation of large amounts of data in MEDLINE, allowing users to extract useful information more quickly and accurately.*

Introduction

Each citation in the Medline database is indexed by an average of 10 MeSH (Medical Subject Headings) terms. PubMed automatically matches the terms in the query against a MeSH translation table. If a match is found in this translation table, in addition to being searched as a text string, the term will also be searched within the indices as MeSH, along with all of its descendents. Currently, the search result is displayed as a flat list regardless of to which particular MeSH term an article is exactly matched. Valuable knowledge can be acquired if the terms in the matrix can become expandable to show this information. More importantly, PubMed does not automatically translate the gene symbols into their full names. Many citations are missed if only the gene symbol is queried because only the full name is used in a lot of abstracts and titles. PubMatrix, which relies solely on PubMed's search capabilities, can systematically analyze the relationships among clusters of terms in a matrix format. However, PubMatrix currently has no mechanism to rank the results nor the structure set in place.

Motivation:

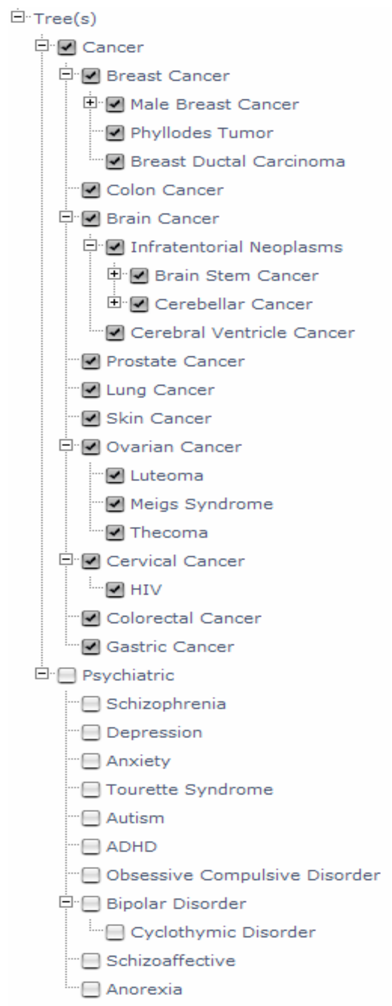
- Provide researchers in the scientific community with better access to relevant articles.
- Improve functionalities of PubMatrix to better utilize knowledge extracted from ontologies and public databases.

Goals:

- Develop a search technique and user interface for PubMatrix SE by allowing users the option to differentiate the citations in the search results based on the MeSH or a user-defined hierarchy.
- Develop insightful algorithms that assess the statistical significance of the values as well as the type of relationships.
- Implement one or more metrics for ranking the search results that will determine the diseases most related to a particular gene and vice versa.

Methods

To begin using PubMatrix SE, the user enters the search terms and gene symbols into two search boxes. A tree is created showing a relational hierarchy of the terms (Fig. 1). The user can check the nodes in the tree to be included in the result. After submission, the system creates and displays a table with all the nodes the user has selected, showing the corresponding co-occurrence frequency and association score for each pair of gene and search terms.



- ⊟ Tree(s)
  - ⊟ ☑ Cancer
    - ⊟ ☑ Breast Cancer
      - ⊞ ☑ Male Breast Cancer
      - ☑ Phyllodes Tumor
      - ☑ Breast Ductal Carcinoma
    - ☑ Colon Cancer
    - ⊟ ☑ Brain Cancer
      - ⊟ ☑ Infratentorial Neoplasms
        - ⊞ ☑ Brain Stem Cancer
        - ⊞ ☑ Cerebellar Cancer
      - ☑ Cerebral Ventricle Cancer
    - ☑ Prostate Cancer
    - ☑ Lung Cancer
    - ☑ Skin Cancer
    - ⊟ ☑ Ovarian Cancer
      - ☑ Luteoma
      - ☑ Meigs Syndrome
      - ☑ Thecoma
    - ⊟ ☑ Cervical Cancer
      - ☑ HIV
    - ☑ Colorectal Cancer
    - ☑ Gastric Cancer
  - ⊟ □ Psychiatric
    - □ Schizophrenia
    - □ Depression
    - □ Anxiety
    - □ Tourette Syndrome
    - □ Autism
    - □ ADHD
    - □ Obsessive Compulsive Disorder
    - ⊟ □ Bipolar Disorder
      - □ Cyclothymic Disorder
    - □ Schizoaffective
    - □ Anorexia

Figure 1.

An example of a user-defined hierarchy displayed in tree form.

Architecture

The MEDLINE citations are downloaded as XML files. Each citation is parsed to extract its PubMed identification, title, abstract, as well as the MeSH indexing terms (Step A in Fig. 2). The data are saved in a relational database (Step B). Identification of genes mentioned in the citations is not a trivial task. Basic keyword matching will result in substantial false positives and false negatives since the gene identifiers are often ambiguous and the same gene can be referred to in various ways. We have developed a pattern-based gene recognition algorithm to identify any human gene mentioned in the titles and abstracts (Step C). We utilize the Entrez gene knowledge-base to obtain detailed information about the genes, including official names, symbols, aliases, chromosome locations, etc. In the event of multiple matches for a term, the algorithm will calculate a score for each candidate to estimate which gene is actually being referred to by the text. The identified genes are saved in the database to establish a gene index for the citations (Step D). Currently, the database contains approximately 15,000 citations. The f-measure for the gene recognition algorithm is 0.62, using a set of 200 citations annotated by human experts as the gold standard.

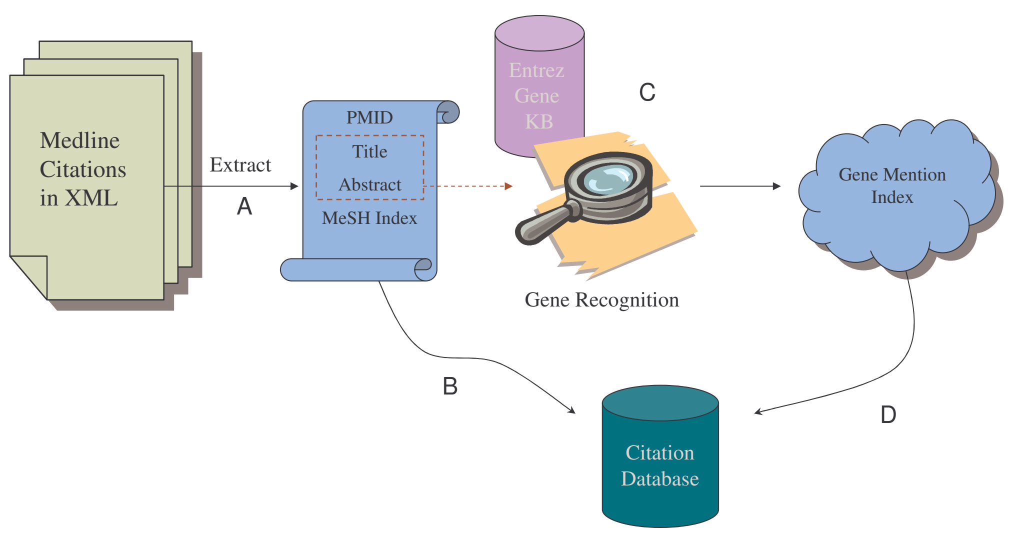


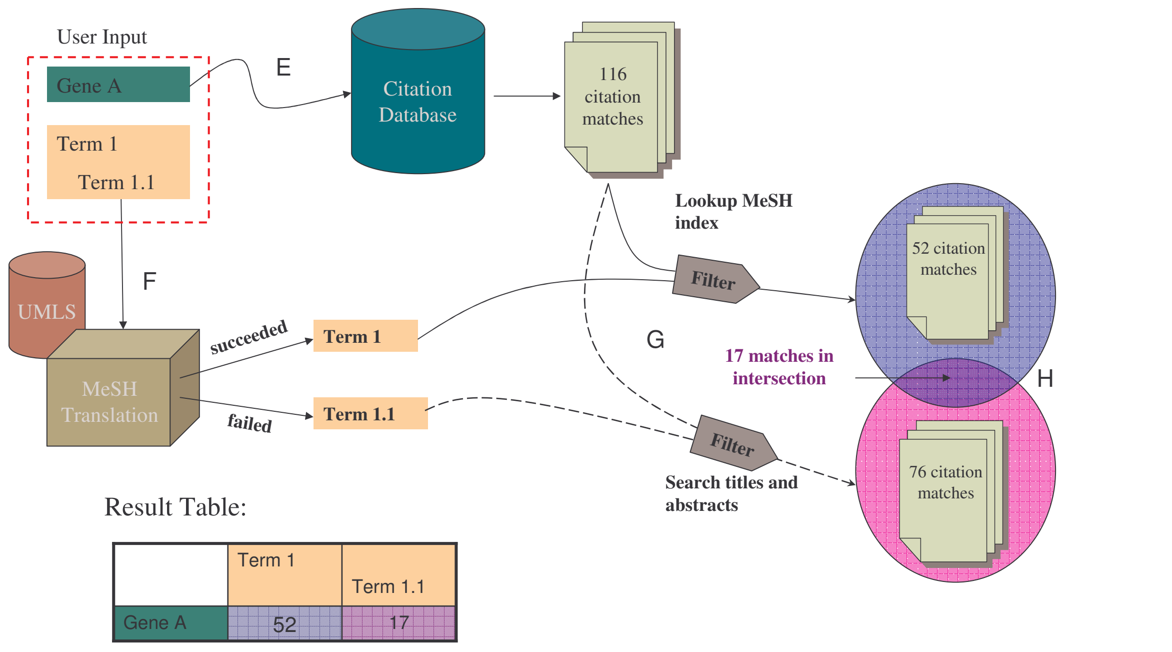
Figure 2. Schematic of MEDLINE data import into the citation database. A gene index is built for the citations using a gene recognition algorithm that is based on pattern-matching and the Entrez Gene knowledge base.

For each request, the user specifies a set of genes and a hierarchical set of search terms. The system retrieves the citations indexed by the genes of interest (Step E in Fig. 3). It will also attempt to map each search term to one or more MeSH terms via the Unified Medical Language System (UMLS) API (Step F). If the mapping succeeds, the MeSH indexing terms of the retrieved citations will be examined to compute the document frequency of co-occurrence between the gene and the search term (Step G). If the mapping fails, the system will search through the titles and the abstracts for the exact term to calculate the co-occurrence frequency. It is apparent that searching through the MeSH indexes will yield a more accurate result because natural language variations of the terms are taken into consideration in the process. Moreover, the MeSH indexes are assigned by trained experts after reading the full text of the paper. Therefore, false positives suffered by simple keyword matching are essentially eliminated. If the search term is a child of another term, as defined in the hierarchy, the system will filter out those citations that are not in its parent's list of citations (Step H).

To estimate the strength and significance of the associations, we use a scoring metric based on the term-frequency/inverse-document-frequency (TF*IDF) weighing scheme:

$$S = f(g,t)\times\log(\frac{N}{f(p(t))})$$

where *f* is a function that returns the co-occurrence frequency of the input terms, p is a function that returns an enumeration of the term and its parents, *g* is the gene, *t* is the search term, and *N* is the total number of records in the database. At present, *f*(*p*(*t*)) is computed by querying PubMed using NCBI's Entrez Utilities. The calculation is performed when the hierarchy is defined. In the result table, the scores are normalized by the maximum score of the corresponding gene.

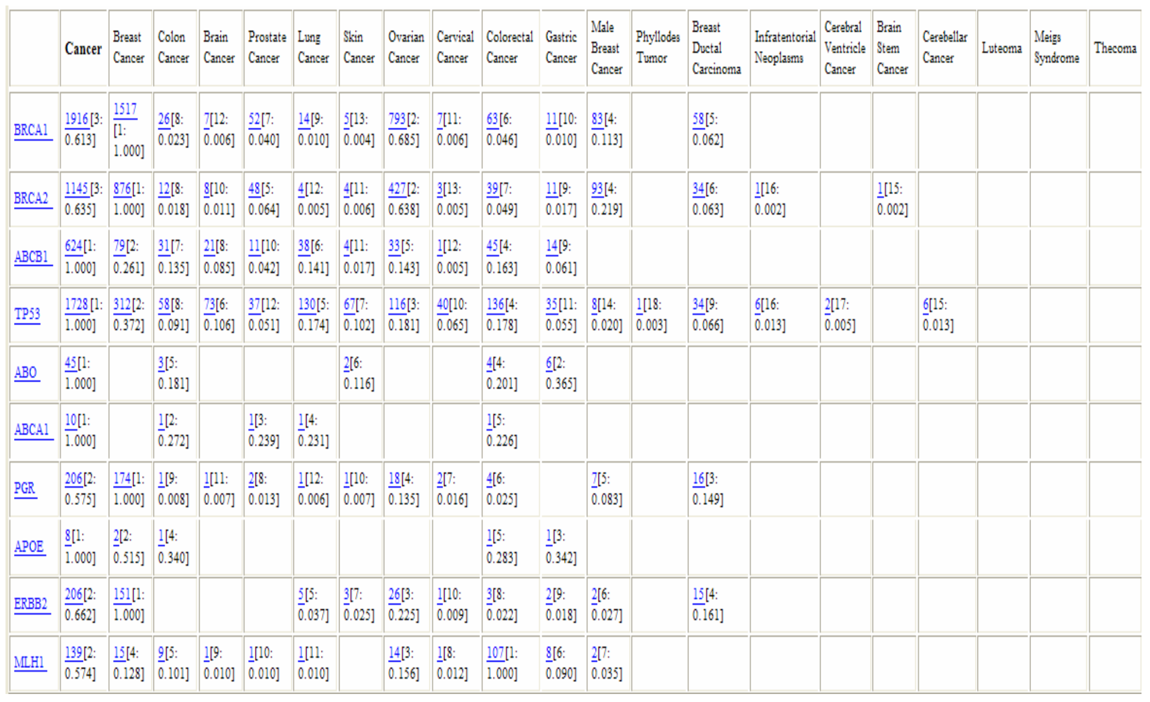


	Term 1	Term 1.1
Gene A	52	17

Figure 3. Schematic showing the steps that are carried out to generate the result table for a sample user query. The search method is slightly different for MeSH and non-MeSH terms. When a term is defined as a child of another term, only the common set is taken.

Results

We have successfully prototyped PubMatrix SE, which extends the functionality of PubMatrix. The web-based interface was designed to be simple but user-friendly. A tree hierarchy shows the relationships between terms. The table displays in a matrix format, the number of citations in the database in which both the term concept and the gene are mentioned (Fig. 4). For each gene, we compare the terms in the hierarchy to determine the strength of their association to the gene. The system has utilized a number of technologies, including Oracle 10g database, J2EE, Asynchronous JavaScript and XML (AJAX), and Apache Tomcat.



	Cancer	Breast Cancer	Colon Cancer	Brain Cancer	Prostate Cancer	Lung Cancer	Skin Cancer	Ovarian Cancer	Cervical Cancer	Colorectal Cancer	Gastric Cancer	Male Breast Cancer	Phyllodes Tumor	Breast Ductal Carcinoma	Infratentorial Neoplasms	Cerebral Ventricle Cancer	Brain Stem Cancer	Cerebellar Cancer	Luteoma	Meigs Syndrome	Thecoma
BRCA1	1916[3: 0.613]	1517 [1: 1.000]	26[8: 0.023]	7[12: 0.006]	52[7: 0.040]	14[9: 0.010]	5[13: 0.004]	793[2: 0.685]	7[11: 0.006]	63[6: 0.046]	11[10: 0.010]	83[4: 0.113]		58[5: 0.062]							
BRCA2	1145[3: 0.635]	876[1: 1.000]	12[8: 0.018]	8[10: 0.011]	48[5: 0.064]	4[12: 0.005]	4[11: 0.006]	427[2: 0.638]	3[13: 0.005]	39[7: 0.049]	11[9: 0.017]	93[4: 0.219]		34[6: 0.063]	1[16: 0.002]		1[15: 0.002]				
ABCB1	624[1: 1.000]	79[2: 0.261]	31[7: 0.135]	21[8: 0.085]	11[10: 0.042]	38[6: 0.141]	4[11: 0.017]	33[5: 0.143]	1[12: 0.005]	45[4: 0.163]	14[9: 0.061]										
TP53	1728[1: 1.000]	312[2: 0.372]	58[8: 0.091]	73[6: 0.106]	37[12: 0.051]	130[5: 0.174]	67[7: 0.102]	116[3: 0.181]	40[10: 0.065]	136[4: 0.178]	35[11: 0.055]	8[14: 0.020]	1[18: 0.003]	34[9: 0.066]	6[16: 0.013]	2[17: 0.005]		6[15: 0.013]			
ABO	45[1: 1.000]		3[5: 0.181]				2[6: 0.116]			4[4: 0.201]	6[2: 0.365]										
ABCA1	10[1: 1.000]		1[2: 0.272]		1[3: 0.239]	1[4: 0.231]				1[5: 0.226]											
PGR	206[2: 0.575]	174[1: 1.000]	1[9: 0.008]	1[11: 0.007]	2[8: 0.013]	1[12: 0.006]	1[10: 0.007]	18[4: 0.135]	2[7: 0.016]	4[6: 0.025]		7[5: 0.083]		16[3: 0.149]							
APOE	8[1: 1.000]	2[2: 0.515]	1[4: 0.340]							1[5: 0.283]	1[3: 0.342]										
ERBB2	206[2: 0.662]	151[1: 1.000]				5[5: 0.037]	3[7: 0.025]	26[3: 0.225]	1[10: 0.009]	3[8: 0.022]	2[9: 0.018]	2[6: 0.027]		15[4: 0.161]							
MLH1	139[2: 0.574]	15[4: 0.128]	9[5: 0.101]	1[9: 0.010]	1[10: 0.010]	1[11: 0.010]		14[3: 0.156]	1[8: 0.012]	107[1: 1.000]	8[6: 0.090]	2[7: 0.035]									

Figure 4. PubMatrix SE search results table. The hyperlinked number shows the frequency of articles that match the corresponding term and gene. The link opens up a new windows listing the citations within the PubMed Web site. The numbers in the bracket shows the rank (left) and score (right) of the term against all other terms that relate to a particular gene.

Future Plans

Several improvements to the current tool will be developed, including:

- Further development of other ranking metrics, such as number of citations received, z-score, impact factors of the publications.
- Automatic expansion of the hierarchy based on MeSH.
- Clustering of genes using the Gene Ontology.
- Content analysis of the text to better determine what the specific relationships between the genes and the other entities are.
- A genome browser interface for visualization.
- A more informative result matrix to show the particular relationships between terms.

References

K. G. Becker, D. A. Hosack, G. Dennis Jr, R. A. Lempicki, T. J. Bright, C. Cheadle and J. Engel, "PubMatrix: a tool for multiplex literature mining," *BMC Bioinformatics,* vol. 4, pp. 61, Dec 10 2003.