Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-210

Classification and Prioritization of Biomedical Literature for the Comparative Toxicogenomics Database

Dina VISHNYAKOVA[a,b,d,1] , Emilie PASCHE[a,b,d], Julien GOBEILL[a,c,d], Arnaud GAUDINAT[a,c,d] , Christian LOVIS[b,d] and Patrick RUCH[a,c,d]
[a] *BiTeM Group*
[b] *Division of Medical Information Sciences, University and University Hospitals of Geneva, Switzerland*
[c] *Information Science Department, University of Applied Science*
[d] *Geneva, Switzerland*

Abstract. We present a new approach to perform biomedical documents classification and prioritization for the Comparative Toxicogenomics Database (CTD). This approach is motivated by needs such as literature curation, in particular applied to the human health environment domain. The unique integration of chemical, genes/proteins and disease data in the biomedical literature may advance the identification of exposure and disease biomarkers, mechanisms of chemical actions, and the complex aetiologies of chronic diseases. Our approach aims to assist biomedical researchers when searching for relevant articles for CTD. The task is functionally defined as a binary classification task, where selected articles must also be ranked by order of relevance. We design a SVM classifier, which combines three main feature sets: an information retrieval system (EAGLi), a biomedical named-entity recognizer (MeSH term extraction), a gene normalization (GN) service (NormaGene) and an ad-hoc keyword recognizer for diseases and chemicals. The evaluation of the gene identification module was done on BioCreativeIII test data. Disease normalization is achieved with 95% precision and 93% of recall. The evaluation of the classification was done on the corpus provided by BioCreative organizers in 2012. The approach showed promising performance on the test data.

Keywords. Information Retrieval, literature curation, ontology look-up services.

Introduction

Since last 10 years the interest in information retrieval and text mining applied to the biomedical literature is rapidly increasing. The biggest database of abstracts on life science and biomedical topics PubMed, in the beginning of 2012, has over 21.47 millions records; around 12 millions of these articles are listed with their abstracts; about 500.000 new records are added each year [1]. Information about entities such as genes, diseases, chemicals and etc, is available in articles in a free textual format, which is comprehensive for humans. Biocurators of the Comparative Toxicogenomic

[1] Corresponding Author: Dina Vishnyakova; SSIM; University Hospitals of Geneva; 4, rue Gabrielle-Perret-Gentil; CH-1211 Geneva 14; Tel: +41 22 372 61 99; email: dina.vishnyakova@hcuge.ch

Database (CTD) read a scientific article and convert free-text information into a structured format using official nomenclatures, incorporating external control vocabularies for chemicals, genes, diseases and organisms [2]. Free-text information is difficult to interpret for information retrieval systems and manual curation of this information is a routine and costly task. As a consequence, there have been developed many methods, which address such tasks as identification of gene/protein mentions or extraction of protein-protein interactions and etc, performing with good results. Currently, the focus of text mining community is shifting towards the interest to employ computers to assist human curators, more specifically to prioritize articles to be curated.

Our approach is based on the combination of Information Retrieval approaches. The feature selection method incorporates meta-data of documents and results retrieved with EAGLi [3][4] and NormaGene [5][6] systems.

1. Data and Methods

1.1. Data overview

BioCreative Workshop 2012 Triage-I provides the benchmark in the biomedical domain; we used a subset of 1059 articles. These articles have been manually curated and contain information on found entities such as genes, chemicals, diseases and their interactions. There were given four main chemicals and a subset of 1059 articles curated to these four chemicals.

Table 1. Distribution of curated articles for each chemical

Chemical Name	Number of articles per chemical	Curated articles per chemical in %
Raloxifene	270	60
2-Acetylaminofluorene	178	45,5
Amsacrine	69	53
Quercetimin	542	77

The decision about curation is done on the observation of curated chemicals, diseases and genes and their interactions in the article. Distributions of curated articles referring to selected chemicals are presented in Table 1.

Approximately half of the articles in the benchmark contain no information about chemicals or genes; only half of the articles have information about both genes and chemicals, and only few have information about diseases. The distribution of entities in the benchmark is shown in Table 2.

Relevant data from the abstract of an article is coded using controlled vocabularies. Chemicals vocabulary is trimmed in order to remove terms that are not considered to be chemicals of interest to CTD (e.g., the “Amino Acids, Peptides, and Proteins” branch or the “Nucleic Acids, Nucleotides, and Nucleosides” branch, etc.) [7]. Curated genes are based on the NCBI identifiers and unlike EntrezGene, a gene in CTD represents the gene for all species [8]. The diseases vocabulary is a mix of the Online Mendelian Inheritance in Man (OMIM) terms and the MeSH “Disease” [C] and “Mental Disorders” [F03] hierarchies [9].

Table 2. Distribution of entities in the benchmark

Entity Name	Number of articles
Chemicals	654
Genes	643
Diseases	28
Genes and Chemicals	602
Main Chemical in titles	381

1.2. Methods

We designed a SVN classifier for the binary classification of articles (with curated and not curated classes). This classifier combines three main feature sets: an information retrieval system (EAGLi), a biomedical named-entity recognizer (MeSH term extraction); a gene normalization service (NormaGene) and an ad-hoc keyword recognizer for diseases and chemicals. Selected features for SVN classifier are presented in Table 3.

The first feature set contains information about MeSH terms of articles extracted from the PubMed. We extract this information with the information retrieval system EAGLi. According to our observations, curated articles usually have as one of the main MeSH terms such term as pharmacology, toxicity, drug therapy, metabolism, drug effects, chemistry and chemical synthesis. Another observation justified that extracted (from PubMed) MeSH terms of a curated article often contain the name of the main chemical. The main chemical is a chemical according to which the classification is done, in our case, for example raloxifene, amsacrine and etc.

In the second feature set we incorporate meta-data received from the gene normalization system NormaGene. On gene name detection step we face ambiguity of gene names, e.g. homonyms and synonyms. NormaGene approves all gene candidates by Gene Protein Synonyms Database (GPSDB) [10]. Returned results from NormaGene are compared to the CTD genes control vocabulary. Genes from this vocabulary are based upon imported gene pages from EntrezGene; however, unlike EntrezGene, a gene page in CTD represents the gene for all species. This representation of the gene eases constrains of NormaGene on a cross-species normalization.

The third feature set is an ad-hoc keyword recognizer for diseases and chemicals. This keyword recognizer is based on the control vocabularies provided by CTD. The chemical vocabulary is a modified subset of descriptors from the “Chemicals and Drugs” category and Supplementary Concept Records from MeSH. Compare to MeSH, CTD merged the descriptors and accompanying concepts into a single hierarchy. Several branches of the original MeSH vocabulary were excluded from CTD's chemical vocabulary because they are not molecular reagents, environmental chemicals or clinical drugs (e.g., “Nucleic Acids, Nucleotides, and Nucleosides” and “Purines”). Other branches were excluded because they are simply broad chemical classes that do not contain more specific terms (e.g., “Solutions” and “Poisons”) [11]. The provided disease vocabulary is a modified subset of descriptors from the “Diseases” category of MeSH combined with genetic disorders from the OMIM database. OMIM contains

textual information, references related to diseases, links to MEDLINE and sequence records in the Entrez system, and links to additional related resources at NCBI[11].

Table 3. Selected Features for the SVN Classifier

Features	Values
Given chemical name in the abstract	binary
Given chemical name in the title	binary
Given chemical in MeSH terms of the article	binary
Appearance of chemicals in the abstract	binary
Quantity of found chemicals in the abstract	integer(1..n)
Appearance of genes in the abstract	binary
Quantity of detected genes	integer (1..n)
Appearance of chemicals and genes in the abstract	binary
Appearance of diseases in the abstract	binary
Quantity of disease names	integer (1..n)
Appearance of "pharmacology", "toxicity", "drug therapy", "metabolism", "drug effects", "chemistry" and "chemical synthesis" as the main MeSH terms of the article	binary
Quantity of main MeSH terms containing "pharmacology", "toxicity", "drug therapy", "metabolism", "drug effects", "chemistry" and "chemical synthesis"	integer(1..n)

2. Results and Conclusion

From the BioCreative 2012 data, we evaluated the effectiveness of our ad hoc terms recognizer for diseases. Our methods achieved 95% of precision and 92% of recall when tagging diseases in the training sample.

Table 4. Results of Our approach for the task-I of BioCreative 2012.

Chemical/Quantity of articles	Intermediate MAP Score	Curated Gene	Curated Chemical	Curated Disease
Urethane/204	0.637	0.08	0.705	0.3
Phenacetin/86	0.831	0.203	0.676	0.5
Cyclophosphamide/154	0.716	0.117	0.747	0.582

Table 5. Results of evaluation performed by our approach with different input for gene detection on the test data of task-I of BioCreative 2012.

Chemical/Quantity of articles	Intermediate MAP Score	Curated Gene	Curated Chemical	Curated Disease
Urethane/204	0.632	0.131	0.705	0.3
Phenacetin/86	0.830	0.295	0.676	0.5
Carcinoma/154	0.710	0.191	0.747	0.582

In order to tune our binary classifier, we performed ten folders cross-validation and achieved accuracy of 80.5%. We applied the optimal model on the test data and obtained an accuracy of 77%, which suggests some moderate overfitting phenomena. The results of our approach of test data are provided in Table 4.

According to results in Table 4, the curated gene score relatively low compared to chemicals and disease scores, which confirms that gene and gene product recognition seems more challenging than recognition of other biomedical entity recognition tasks such as chemicals. We recomputed results, applying some changes to the parameters for the gene detection task, and removing the restriction of the overlapping gene names. The recomputed results for the gene detection subtask are found in Table 5. While the Curated Entities scores in final results of BioCreative 2012 were based on recall scores it is obvious that overlapping gene candidates in the final results do improve the "Curated Gene" score. At the same time they decrease the "Intermediate MAP" score. This can be explained by the changes in the classification model, which reduces the ranking score of input articles, when a large amount of gene candidates is found.

In conclusion, our approach showed competitive performance, in particular for the recognition of chemical compounds. Intermediate MAP score showed that the selected SVM model produced promising results on the test data. In contrast, the identification of pathologies seems nearly as difficult as the recognition of genes and gene products. Further experiments are needed to explain where is the power of the approach as well as to start explaining the differences observed regarding the recognition power of some of the entity types.

Acknowledgements. The DebugIT project (http://www.debugit.eu) is receiving funding from the European Community's Seventh Framework Programme under grant agreement n°FP7-217139, which is gratefully acknowledged. The information in this document reflects solely the views of the authors and no guarantee or warranty is given that it is fit for any particular purpose. The European Commission, Directorate General Information Society and Media, Brussels, is not liable for any use that may be made of the information contained therein.

References

[1] PubMed [http://www.ncbi.nlm.nih.gov/pubmed?term=1800%3A2100%5Bdp%5D] (23.01.2012)
[2] Davis AP, Wiegers TC, Murphy CG, and Mattingly CJ. 2011. The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database. *Database.*2011, Oxford.
[3] EAGLi [http://eagl.unige.ch/EAGLi/] (23.01.2012)
[4] Gobeill J, Pasche E, Teodoro D, Veuthey AL, Lovis C, Ruch P. Question answering for biology and medicine. *Information Technology and Application in Biomedicine (ITAB 2009)*. 2009.
[5] NormaGene [http://pingu.unige.ch:8080/NormaGene](23.01.2012)
[6] Lu Z, Kao HY, Wei CH, Huang M, Liu J, Kuo CJ, Hsu CN, Tsai RT, Dai HJ, Okazaki N, Cho HC, Gerner M, Solt I, Agarwal S, Liu F, Vishnyakova D, Ruch P, Romacker M, Rinaldi F, Bhattacharya S, Srinivasan P, Liu H, Torii M, Matos S, Campos D, Verspoor K, Livingston KM, Wilbur WJ. The gene normalization task in BioCreative III. *BMC Bioinformatics*. 2011 Oct 3; 12 Suppl 8:S2.
[7] CTD Chemicals data [http://ctdbase.org/downloads/#allchems] (23.01.2012)
[8] CTD Genes data [http://ctdbase.org/downloads/#allgenes](23.01.2012)
[9] CTD Disease data [http://ctdbase.org/downloads/#alldiseases](23.01.2012)
[10] Pillet V, Zehnder M, Seewald AK, Veuthey AL, Petrak J. GPSDB: a new database for synonyms expansion of gene and protein names. *Bioinformatics*. 2005 Apr 15; 21(**8**):1743-4. Epub 2004 Dec 21.
[11] Wiegers TC, Davis AP, Cohen KB, Hirschman L, Mattingly CJ. Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD). *BMC Bioinformatics.* 2009 Oct 8;10:326.