Pathogens and Gene Product Normalization in the Biomedical Literature

Dina VISHNYAKOVA[a,b,d,1] , Emilie PASCHE[a,b,d], Douglas TEODORO[a,b,d], Christian LOVIS[b,d] and Patrick RUCH[a,c,d]

[a] *BiTeM Group*
[b] *Division of Medical Information Sciences, University Hospitals of Geneva and University of Geneva*
[c] *Information Science Department, University of Applied Science*
[d] *Geneva, Switzerland*

Abstract. We present a new approach for pathogens and gene product normalization in the biomedical literature. The idea of this approach was motivated by needs such as literature curation, in particular applied to the field of infectious diseases thus, variants of bacterial species (S. aureus, Staphyloccocus aureus…) and their gene products (protein ArsC, Arsenical pump modifier, Arsenate reductase…). The

Our approach is based on the use of an Ontology Look-up Service, a Gene Ontology Categorizer (GOCat) and Gene Normalization methods. In the pathogen detection task the use of OLS disambiguates found pathogen names. GOCat results are incorporated into overall score system to support and to confirm the decision-making in normalization process of pathogens and their genomes.

The evaluation was done on two test sets of BioCreativeIII benchmark: gold standard of manual curation (50 articles) and silver standard (507 articles) curated by collective results of BCIII participants. For the cross-species GN we achieved the precision of 46% for silver and 27% for gold sets. Pathogen normalization results showed 95% of precision and 93% of recall.

The impact of GOCat explicitly improves results of pathogen and gene normalization, basically confirming identified pathogens and boosting correct gene identifiers on the top of the results' list ranked by confidence. A correct identification of the pathogen is able to improve significantly normalization effectiveness and to solve the disambiguation problem of genes.

Keywords. Pathogen, gene normalization, Information Retrieval, infectious disease, ontology look-up service.

Introduction

Since last 10 years the interest in information retrieval and text mining applied to the biomedical literature is rapidly increasing. This interest appeared also due to the biggest public database of abstracts on life science and biomedical topics - PubMed, which, in the beginning of 2012, has over 21.47 millions records; around 12 millions of

[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

these articles are listed with their abstracts. PubMed is a free resource that is developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH). The content of PubMed such as citations and abstracts include the fields of biomedicine and health, covering portions of the life sciences, behavioral sciences, chemical sciences, and bioengineering. It also provides access to additional relevant web sites and links to the other NCBI molecular biology resources. Due to its public access it is possible to use resources of the library in scientific researches such as literature curation, novelty detection in biomedicine domain and etc. [1]

Names of pathogens and their genome have various representations in biomedicine literature. Information about infectious diseases is available in a free textual format, which is comprehensive for humans, but difficult to interpret for information retrieval systems. As a consequence, there is an increasing interest in methods, which have to detect and normalize entities such as species and genes [2] in order to provide accurate, well-structured information on demand [3].

Despite the fact that gene nomenclature is controlled by guidelines, gene normalization has to deal with highly ambiguous names. A gene entity can be described by many different terms. Moreover, the same term can be attributed to different entities. Homonymy is particularly present for orthologous genes. The complexity of the task increases if there is no information on species provided. Therefore, species identification and disambiguation may be critical in the process of finding the correct gene identifier (id).

Many systems for Gene Normalization (GN) are based on hints, such as textual structure or MeSH terms [2][4], where the abstract and the introduction are the most entity richest sections of the document. The results of our approach are not based on such hints. The estimation of results' confidence is based mainly on the meta-data of entities observed in the text and results provided by Gene Ontology Categorizer (GOCat)[4].

1. Data and Methods

1.1. Data overview

The test data provided by BioCreative III (BCIII) includes 507 articles in the biomedical domain. These articles have recently published and have not had any curated annotations yet. Overall 101 names of species have been found in the set. The overview of data shows that 70% of articles contain more than one specie name.

BCIII has three evaluation standards. The first evaluation standard is a so-called "gold 50" containing 50 articles of manual curation extracted from the entire collection of 507 articles. The second standard is a "silver 507". It consists of full collections, e.g. 507 articles, based on the combination of best submissions of BCIII participants and 50 articles from "gold standard". The third standard is "silver 50". In this standard the same articles belonging to the "gold 50" standard were included, yet the curation is done by best submissions of the BCIII participants. A gene distribution between "gold 50" and "silver 507" sets, reported in [2], shows that the "gold 50" set is not representative for the entire collection of articles. These can explain further deviation in results evaluated with both sets. The pathogen distribution in BCIII articles is shown in Table 1.

Table 1. The pathogens distribution in the BCIII testing set.

Pathogen Name	Distribution in %
Escherichia coli	9.27
Staphylococcus aureus, Enterobacter sp. 638,	4
Streptococcus pneumoniae, Bacteroides fragilis,	3.9
Mycobacterium tuberculosis	3.4
Simian immunodeficiency virus (isolate CPZ GAB1), Pseudomonas aeruginosa, Vibriocholerae, Cryptococcus neoformans, S. pneumoniae TIGR4	2.5
Sindbis virus, Staphylococcus haemolyticus, Trypanosoma congolense, Human SARS coronavirus Bacillus cereus, Human herpesvirus 1, Escherichia coli O157:H7	<1

1.2. Methods

The approach of pathogen and its genome normalization can be split into three subtasks. The first subtask is to detect entity names. In the second subtask we refine detected candidates with a dictionary. In this subtask we elaborate Gene Protein Synonyms DataBase (GPSDB) [6][7] for the gene candidate and Ontology Look-up Service (OLS) for the pathogen candidates. The third subtask filters false positives (FPs) by applying some empirical rules. In these rules the weighting scheme plays, the main role. It considers detected species, their occurrence in the text, as well as genes metadata in order to filter FPs by giving them the lowest confidence score. Approved entities are linked to unique identifiers and thus to primary sequences. The workflow of our approach is shown on the Figure 1.

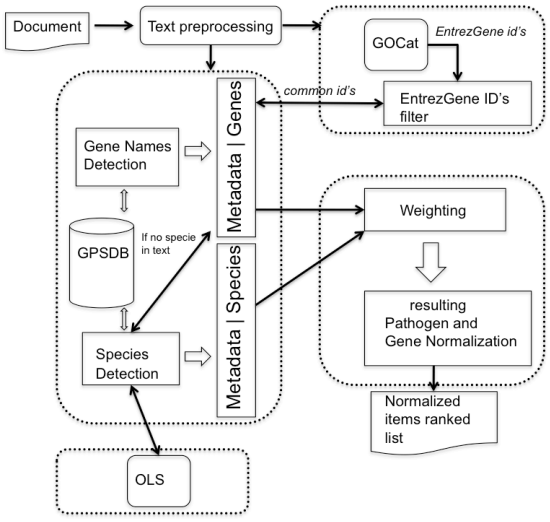

Figure 1. Workflow of Pathogens and Genes Normalization.

1.2.1. Pathogen Normalization

For Species Detection we have used simple rule-based approaches and have created specific recognition modules for a dozen of the most common pathogens, such as

Escherichia coli, *Staphylococcus aureus* and *etc*. For the found candidates validation and obtaining their ids we used GPSDB and OLS. In order to refine the scope of studied species in the text we used OLS, which provides an expanded list of entities belonging to the family of a current pathogen (Tuberculoses - a bacteria class and a group name representing the infection). The expansion of entity names is checked against the given text in order to detect implicit pathogen names related to the approved one.

1.2.2. Gene Normalization

On gene name detection step we face ambiguity of gene names, e.g. homonyms and synonyms. In our approach we use a hybrid gene name recognition module, based on Rule-Based gene/protein name detection and hidden Markov Model (HMM). All gene candidates are approved by GPSDB. A gene name frequency of occurrence and meta-data obtained from GPSDB is used for calculating the confidence score. The overall confidence score in our system is based on the species and the weight attributed to a gene name. We elaborate Gene Ontology Categorizer (GOCat) [5] in the confidence ranking. It boosts correct ids on the top of the results list [8].

2. Results

Table 2 lists results of our approach evaluated with a proposed metric for measuring retrieval efficacy called Threshold Average Precision (TAP-k) [9] on "silver 507", "gold 50" and "silver 50" standards.

Table 2. The results of evaluation performed by our system in the cross-species GN task of BC III.

TAP-k	Gold Standard/ 50 articles	Silver Standard/ 50 articles	Silver Standard/507 articles
5	0.1926	0.28	0.4368
10	0.2025	0.3157	0.4368
20	0.2097	0.3157	0.4368

These results were obtained with a specie preference weighting in order to avoid false positives results. While tuning the weighting scheme and giving more weight to gene candidates (based on NER probability, preference of the term and GOCat support) the system showed worse results, see Table 3.

From the BCIII "silver 507" standard results we extracted the species name for evaluating the efficiency of the pathogen normalization. Our approach shows 95% of precision and 92% of recall compared to species extracted from BCIII "silver 507" standard.

Table 3. The results of evaluation performed by our system in the cross-species GN task of BC III. Tuning of the weighting scheme with a preference to the observed gene names.

TAP	Gold Standard 50 With GOCat/Without GOCat		Silver Standard 50 With GOCat/Without GOCat		Silver Standard 507 With GOCat/Without GOCat	
5	0.1084	0.0329	0.2579	0.0792	0.4268	0.2332
10	0.1581	0.0437	0.2840	0.1269	0.4268	0.2397
20	0.1646	0.0527	0.2840	0.1329	0.4268	0.2397

3. Conclusion

In the section of results we have shown the performance of our approach on cross-species gene normalization. The results provided in Table 2 and 3 showed that a correct identification of the species could decrease the ambiguity of orthologous genes.

The impact of GOCat appeared effective on the data. This impact suggests that overfitting phenomena are avoided mainly because GOCat has not been originally designed for gene recognition and normalization.

While compiling statistics on gene distribution in BCIII standards we discovered that approximately 100 genes ids were not found in current version of GPSDB, which was mainly due to the late synchronization of the content with EntrezGene.

The normalization of pathogens showed effective results on the BCIII benchmark. The results of normalization of the pathogen names demonstrated that OLS was successfully used to disambiguate species entities such as genus name. The species sub-type provided by OLS is able to disambiguate the species name and genus name, which both occurred in the same text.

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]
[2] *The proceedings for the BioCreative III Workshop*, 2010, Bethesda, Maryland, USA. ISBN: 978-1-4507-3685-5
[3] D. Teodoro, R. Choquet, E. Pasche, J. Gobeill, C. Daniel, P. Ruch, C. Lovis, Biomedical Data Management: a Proposal Framework. *In Proceedings of MIE 2009* (2009)
[4] Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J et al: *Overview of BioCreative II gene normalization.* Genome Biol 2008, 9 Suppl 2:S3.
[5] GOCat – Gene Ontology Categorizer [http://eagl.unige.ch/GOCat]
[6] Gene and Protein Synonym DataBase [http://www.expasy.ch/gpsdb/]
[7] 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.
[8] Ruch P. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, **22**(6):658-664, (2006).
[9] Carroll HD, Kann MG, Sheetlin SL, Spouge JL (2010) Threshold Average Precision (TAP-k): a measure of retrieval designed for bioinformatics. *Bioinformatics* 2010, **26**(14):1708-1713.