Utilization of Ontology Look-up Services in Information Retrieval for Biomedical Literature

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

Abstract. With the vast amount of biomedical data we face the necessity to improve information retrieval processes in biomedical domain. The use of biomedical ontologies facilitated the combination of various data sources (e.g. scientific literature, clinical data repository) by increasing the quality of information retrieval and reducing the maintenance efforts. In this context, we developed Ontology Look-up services (OLS), based on NEWT and MeSH vocabularies. Our services were involved in some information retrieval tasks such as gene/disease normalization. The implementation of OLS services significantly accelerated the extraction of particular biomedical facts by structuring and enriching the data context. The results of precision in normalization tasks were boosted on about 20%.

Keywords. ontology, information retrieval, literature curation.

Introduction

The problem of integrating and improving the technology to support real-world scientific problems in biomedicine, as well as its practical importance in the context of specific tasks becomes urgent. Most information of biomedical discovery is stored in detailed full text. Extraction of this information from scientific sources manually is expensive and reluctant. In many cases, the full text is not available to the researcher at all; the literature databases contain only bibliographic information and abstracts. Last ones suffer from limitations of data compression and convolution imposed by a word limit [1]. Consequently, the interest of the information retrieval community in text mining shifts to the use of machines to guide and support the user with structured and prioritized information. Therefore, there is a need to decompress information by expanding abbreviations and by mapping references between scientific terminology and terms of common language. As a result, a number of successful methods was developed for such tasks as identification and normalization of genome/proteome, pathogens identification and protein-protein interactions [2]. Gene/protein

[1] Corresponding Author: Dina Vishnyakova; SSIM; University Hospitals of Geneva; 4, rue Gabrielle-Perret-Gentil; CH-1211 Geneva 14; Tel: +41 22 372 62 32; email: dina.vishnyakova@hcuge.ch

normalization task is particularly challenging because it is not species-specific. The complexity of the task increases when there is no information provided in the text on species. Another problem is to deal with gene names that are shared among different species. Homonymy are particularly present for orthologous genes. The next challenge is the form of species in the text. It is quite common to meet implicit form of specie name in the text, for instance: *patient, woman, man, human* are referring to *Homo sapiens* and suppose to have the same identifier in taxonomy.

In order to address challenges, described above and to improve the efficiency of identification and normalization methods, we developed Ontology Look-up services (OLS), to disambiguate, to structure and to expand found information.

1. Data overview

1.1. Test Data

To assess our information retrieval systems supported by OLS, we used test data BioCreative III (BCIII) and test data provided by Biocreative Workshop'12 (BW12).

The BC III set consists of 507 articles in the field of biomedicine. This was the recent publication set, which was not manually processed by biocurators; it means that publications contain no descriptors and are not assigned to topics [3]. Review of the data showed that 70% of the articles from this set contain more than one species.

BCIII benchmark has two standards of assessment. The first standard is the so-called "golden 50" (G50), containing 50 articles of manual curation. These articles have been selected from the main collection of 507 articles. The second standard is "silver 507" (S 507). It includes best results of participants BCIII on 507 articles in combination with articles of G50 standard [3].

BC12 Track – I test data set was released in order to evaluate the performance of the participants' text-mining pipeline without their prior knowledge of the curated results. The Track I Test Dataset comprised 444 articles of manual curation, done by CTD biocurators. This set contains information about three target chemicals (urethane, phenacetin, cyclophosphamide)[4].

1.2. Data Ontology

We used following vocabularies to compile OLS:

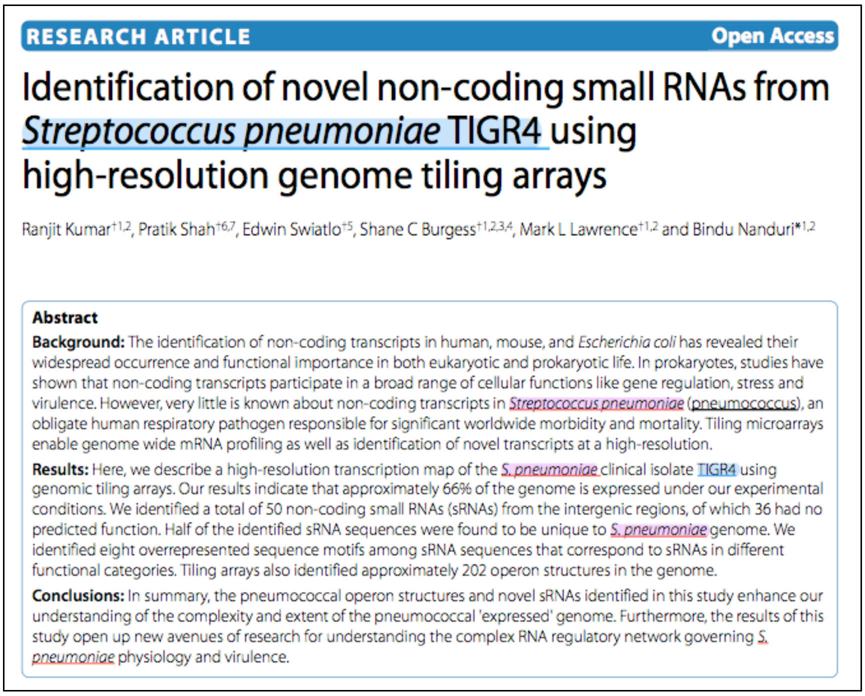
- NEWT is a database of wildlife taxonomy, maintained by UniProt. It integrates taxonomy data compiled in the NCBI database and data specific to the UniProt Knowledgebase. For most species in database, the scientific name is followed by the English common name and a synonym if available [5] Taxonomy is organized in a tree structure, which represents the taxonomic lineage.
- MeSH - Medical Subject Headings. MeSH is the controlled vocabulary lexicon of medicine. It consists of a set of terms' descriptors represented in a hierarchical structure [6]. MeSH descriptors are organized hierarchically to allow searches with various levels of specificity. The topmost level of the hierarchy includes general terms. Lower-level terms may have more than one

possible “parent” or higher-level term: For example, “Liver Neoplasms” may be reached either through “Liver Diseases” or through “Digestive System Neoplasms”. [7]

2. Methods and Results

2.1. Ontology Look-up services

Most of information retrieval methods in biomedical domain lack an approach to resolve the ambiguity by expanding or narrowing / restricting information. The ability to include or exclude relationship branches is useful in such tasks as genome normalization, where clearly identified species can significantly improve the result of genome identifier assignment [3].



RESEARCH ARTICLE Open Access

Identification of novel non-coding small RNAs from *Streptococcus pneumoniae* TIGR4 using high-resolution genome tiling arrays

Ranjit Kumar[†1,2], Pratik Shah[†6,7], Edwin Swiatlo[†5], Shane C Burgess[†1,2,3,4], Mark L Lawrence[†1,2] and Bindu Nanduri[*1,2]

Abstract

Background: The identification of non-coding transcripts in human, mouse, and *Escherichia coli* has revealed their widespread occurrence and functional importance in both eukaryotic and prokaryotic life. In prokaryotes, studies have shown that non-coding transcripts participate in a broad range of cellular functions like gene regulation, stress and virulence. However, very little is known about non-coding transcripts in *Streptococcus pneumoniae* (pneumococcus), an obligate human respiratory pathogen responsible for significant worldwide morbidity and mortality. Tiling microarrays enable genome wide mRNA profiling as well as identification of novel transcripts at a high-resolution.

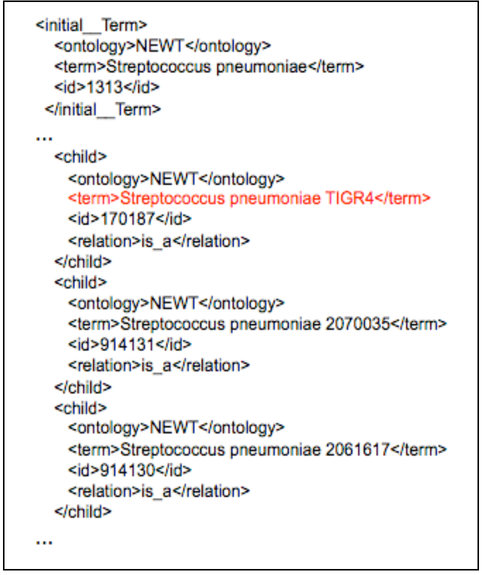
Results: Here, we describe a high-resolution transcription map of the *S. pneumoniae* clinical isolate TIGR4 using genomic tiling arrays. Our results indicate that approximately 66% of the genome is expressed under our experimental conditions. We identified a total of 50 non-coding small RNAs (sRNAs) from the intergenic regions, of which 36 had no predicted function. Half of the identified sRNA sequences were found to be unique to *S. pneumoniae* genome. We identified eight overrepresented sequence motifs among sRNA sequences that correspond to sRNAs in different functional categories. Tiling arrays also identified approximately 202 operon structures in the genome.

Conclusions: In summary, the pneumococcal operon structures and novel sRNAs identified in this study enhance our understanding of the complexity and extent of the pneumococcal 'expressed' genome. Furthermore, the results of this study open up new avenues of research for understanding the complex RNA regulatory network governing *S. pneumoniae* physiology and virulence.

Figure 1. A species names variation on an example of the article with the pmid 2887815. Different names of bacteria from the same family are marked out by color (*Streptococcus pneumoniae* and *Streptococcus pneumoniae TIGR4*).

OLS allows to search terms in the hierarchies of NEWT and MeSH by related associations of the first level or *n*-th level of such relationships as *is_a*, where the descendant belongs to a parent/parents and *has_a*, where a parent has a child / children. In addition, OLS has an access to the synonyms’ names of NEWT.

Figure 1 illustrates the ambiguity in species names normalization, where two types of bacteria belong to one family, but each has different identifier in NEWT: *Streptococcus pneumoniae* - 1313 and *Streptococcus pneumoniae TIGR4* - 170187. In the task of gene normalization, *Streptococcus pneumoniae TIGR4* would be ignored. Hence, detected gene names will be assigned (normalized) to gene identifiers with the reference to *Streptococcus pneumoniae*. It is important to eliminate the ambiguity in such cases to extend search criteria, see Fig. 2.



```
<initial__Term>
  <ontology>NEWT</ontology>
  <term>Streptococcus pneumoniae</term>
  <id>1313</id>
 </initial__Term>
...
  <child>
    <ontology>NEWT</ontology>
    <term>Streptococcus pneumoniae TIGR4</term>
    <id>170187</id>
    <relation>is_a</relation>
  </child>
  <child>
    <ontology>NEWT</ontology>
    <term>Streptococcus pneumoniae 2070035</term>
    <id>914131</id>
    <relation>is_a</relation>
  </child>
  <child>
    <ontology>NEWT</ontology>
    <term>Streptococcus pneumoniae 2061617</term>
    <id>914130</id>
    <relation>is_a</relation>
  </child>
...
```

Figure 2. OLS-NEWT results on given bacteria *Streptococcus pneumoniae*. OLS extracts all children of the given species. In red color is the child found in the text of pmid 2887815, see Fig. 1.

2.2. Results

Table 1 shows the results of NormaGene system developed for the gene normalization task of BCIII [3].

Table 1. Results of gene normalization task, produced by NormaGene.

TAP*	**G50**		**S507**	
	OLS	**Without OLS**	**OLS**	**Without OLS**
5	0.1084	0.0929	0.4268	0.3532
10	0.1581	0.1037	0.4268	0.3597
20	0.1646	0.1127	0.4268	0.3597

The system should normalize genomes and proteomes found in the text without organism names information known a priori. This complicates the normalization task where each identifier is dependent on detected specie in the text. It is not rare that the genomes name of various organisms have, often, identical name then it becomes problematic to determine what type of genome was mentioned [3]. Thus, in Table 1 are the results of the genome normalization with and without OLS.

* TAP - Threshold Average Precision (TAP-k): Evaluation metric used in BCIII campaign [9].

In table 2 are results of Toxicat system, developed for the Track-I task of BW12 [8]. Toxicat supported by OLS-MeSH expands search of chemical/disease terms and at the same time restricts detected entities, which are not relevant for the task.

Table 2. Results of Toxicat system for Track-I of BW12 where *Curated Chemical/Disease Hit Rate* is a recall of recognized entities.

Chemical/Number of articles	Curated Disease Hit Rate	Curated Chemical Hit Rate
Urethane/204	0.3	0.705
Phenacetin/86	0.5	0.676
Cyclophosphamide/154	0.58	0.747

3. Conclusions

We have shown the effectiveness of OLS approach in inter-species gene normalization task and chemical/disease detection tasks. NormaGene results showed that the correct identification of species might reduce ambiguity of orthologous genes. As for the results of Toxicat, the main task of the system was to integrate existing components for further document prioritization, see [8]; where the integration of OLS services in Toxicat system was a valuable resource to shrine the entity search and therefore to eliminate those not relevant to the main task. Although, the chemical-based approach showed competitive results [4], the disease-based approach resulted in more contrasted results. The recall for diseases seems to be satisfying and could be explained by lack of full-text information, where most of provided abstracts did not contain any information about diseases. Although current results seem suggesting that OLS can effectively help curators' tasks by providing access to more relevant contents, it is worth noticing that the effectiveness of information retrieval systems, mentioned above, is obtained by specializing some of the components.

Overall results show that our method of OLS can significantly improve the results and can generate automatically subset of biomedical knowledge, as well to provide structures of found entities for further processing, storage and retrieval of large repositories of knowledge.

References

[1] C. Blaschke, A. Valencia, Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study. *Comp Funct Genomics* **2** (2001), 196–206.

[2] C.N. Arighi, Z. Lu, M. Krallinger, K. B. Cohen, W. J. Wilbur, A. Valencia, L. Hirschman and C. H. Wu, Overview of the BioCreative III Workshop, *BMC Bioinformatics*, **12** (Suppl 8):S1 (2011)

[3] Z. Lu, W. J. Wilbur et al. The gene normalization task in BioCreative III, *BMC Bioinformatics*, **12**(Suppl 8):S2. (2011)

[4] T. Wiegers, A.P. Davis, and C.J. Mattingly, Collaborative Biocuration-Text Mining Development Task for Document Prioritization for Curation, *2012 BioCreative Workshop Proceedings* (2012)

[5] NEWT Taxonomy. http://www.uniprot.org/taxonomy/

[6] D. Trieschnigg, P. Pezik, V. Lee, F. de Jong, W. Kraaij, D. Rebholz-Schuhmann, MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics* **25 (11)** (2009): 1412-1418.

[7] C. Kahn, Multilingual Retrieval of Radiology Images, *RadioGraphics* **29** (2009), 23-29

[8] D. Vishnyakova, E. Pasche et P. Ruch, Selection of relevant articles for curation for the Comparative Toxicogenomic DataBase, *2012 BioCreative Workshop Proceedings* (2012)

[9] A.N. Carroll, D. Hyrum, Kann, G. Maricel, Sheetlin, L. Sergey, Spouge, L. John, Threshold Average Precision (TAP-k), *Bioinformatics*, Volume **26** Issue 14, (2010)