Pathogens and gene product normalization in the biomedical literature

VISHNYAKOVA, Dina, et al.

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, ...). 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 decisionmaking 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 […]

Reference

PMID : 22491118

Available at:
http://archive-ouverte.unige.ch/unige:21579

Disclaimer: layout of this document may differ from the published version.
Pathogens and Genome Normalization for Literature-based Knowledge Discovery

Dina VISHNYAKOVA\textsuperscript{a,1}, Emilie PASCHE\textsuperscript{a}, Douglas TEODORO\textsuperscript{a}, Patrick RUCH\textsuperscript{b} and Christian LOVIS\textsuperscript{a}

\textsuperscript{a}Division of Medical Information Sciences, University Hospitals of Geneva and University of Geneva, Switzerland

\textsuperscript{b}Information Science Department, University of Applied Science of Geneva, Switzerland

Abstract. We present a new approach of pathogens and genome normalization in the biomedical literature. It was motivated by needs such as literature curation, in particular applied to infectious diseases. Our approach is based on use of an Ontology Look-up Service, a Gene Ontology Categorizer and Gene Normalization methods. Gene normalization precision is $\approx 0.43\%$. Pathogen normalization results showed 95\% of precision and 93\% of recall. The results showed that a correct identification of the species can improve significantly normalization effectiveness of gene products.

Keywords. Pathogen normalization, gene normalization.

1. Introduction

Information about infectious diseases is available in a free textual format, which is difficult to interpret for information retrieval systems. Despite the fact that gene nomenclature is controlled by guidelines, gene normalization has to deal with highly ambiguous names. The species identification and disambiguation may be critical in the process of finding the correct gene identifier (id). In our approach we base results’ confidence on the meta-data of entities observed in the text and results provided by Gene Ontology Categorizer (GOCat)[1].

2. Data and Methods

2.1. Data overview

The test data provided by BioCreative III (BCIII) includes 507 articles in the biomedical domain. 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.

\textsuperscript{1} Corresponding Author: Dina Vishnyakova; SIMED; University Hospitals of Geneva; 4, rue Gabrielle-Perré-Gentil; CH-1211 Geneva 14; Tel: +41 22 372 61 99; E-mail: dina.vishnyakova@hcuge.ch
2.2. 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 validate detected candidates with a dictionary. The third subtask filters false positives (FPs) by applying some empirical rules.

For Species Detection we have used simple rule-based approaches and created recognition modules for a dozen of the most common pathogens. In order to refine the scope of studied species in the text we used Ontology-Lookup Service (OLS), which provides an expanded list of entities belonging to the family of a current pathogen.

For solving the ambiguity problem of gene names (e.g. homonyms and synonyms) we use a hybrid gene name recognition module. All gene candidates are approved by GPSDB [2]. Elaboration of GOCat boosts correct ids on the top of the results list [3].

3. Results and Conclusion

We extracted species names from BCIII test data for evaluating the efficiency of the Pathogen Normalization (PN). The result of PN shows 95% of precision and 93% of recall. The species sub-type provided by OLS is able to disambiguate the species name and genus name, which both occurred in the same text. The results of Gene Normalization (GN) of our approach are evaluated with a proposed metric called Threshold Average Precision (TAP-k) [3].

Table 1. The results of cross-species Gene Normalization

<table>
<thead>
<tr>
<th>TAP-k</th>
<th>50 articles of manual curation</th>
<th>50 articles of best submissions (same articles of manual curation)</th>
<th>507 articles of best submissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.1926</td>
<td>0.24</td>
<td>0.4368</td>
</tr>
<tr>
<td>10</td>
<td>0.2025</td>
<td>0.3157</td>
<td>0.4368</td>
</tr>
<tr>
<td>20</td>
<td>0.2097</td>
<td>0.3157</td>
<td>0.4368</td>
</tr>
</tbody>
</table>

Evaluation with a preference on a gene identification

<table>
<thead>
<tr>
<th></th>
<th>GOCat</th>
<th>No GOCat</th>
<th>GOCat</th>
<th>No GOCat</th>
<th>GOCat</th>
<th>No GOCat</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.1084</td>
<td>0.0329</td>
<td>0.2579</td>
<td>0.0792</td>
<td>0.4268</td>
<td>0.2332</td>
</tr>
<tr>
<td>10</td>
<td>0.1581</td>
<td>0.0437</td>
<td>0.2840</td>
<td>0.1269</td>
<td>0.4268</td>
<td>0.2397</td>
</tr>
<tr>
<td>20</td>
<td>0.1646</td>
<td>0.0527</td>
<td>0.2840</td>
<td>0.1329</td>
<td>0.4268</td>
<td>0.2397</td>
</tr>
</tbody>
</table>

The results provided in Table 1 showed that a correct identification of the species could decrease the ambiguity of orthologous genes. The impact of GOCat is appeared to be effective. The overfitting phenomena are avoided mainly because GOCat has not been originally designed for gene recognition and normalization.

Acknowledgements. This work is supported by grant of DebugIT. The DebugIT project (http://www.debugit.eu) is receiving funding from the European Community’s Seventh Framework Programme under grant agreement n°FP7-217139.

References