Pathogens and Genome Normalization for Literature-based Knowledge Discovery

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

[a]*Division of Medical Information Sciences, University Hospitals of Geneva and University of Geneva, Switzerland*

[b]*Information Science Department, University of Applied Science of Geneva, Switzerland*

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

Keywords. Pathogen normalization, gene normalization.

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].

1. Data and Methods

1.1. 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.

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

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 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 have 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].

2. 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

TAP-k	50 articles of manual curation		50 articles of best submissions (same articles of manual curation)		507 articles of best submissions	
5	0.1926		0.28		0.4368	
10	0.2025		0.3157		0.4368	
20	0.2097		0.3157		0.4368	
	Evaluation with a preference on a gene identification					
	GOCat	**No GOCat**	**GOCat**	**No GOCat**	**GOCat**	**No 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

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

[1] GOCat – Gene Ontology Categorizer [http://eagl.unige.ch/GOCat]
[2] Gene and Protein Synonym DataBase [http://www.expasy.ch/gpsdb/]
[3] The proceedings for the BioCreative III Workshop, 2010, Bethesda, Maryland, USA. ISBN: 978-1-4507-3685-5