Appeared in:

Lynette Hirschmann, Martin Krallinger, and Alfonso Valencia, editors. Proceedings of the Second BioCreative Challenge Evaluation Workshop.

Centro Nacional de Investigaciones Oncologicas, CNIO, Madrid, Spain, 2007. pp. 149–151

ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries

Juliane Fluck
juliane.fluck@scai.fhg.de

Heinz Theodor Mevissen
theo.mevissen@scai.fhg.de

Holger Dach
Holger.dach@scai.fhg.de

Marius Oster
marius.oster@scai.fhg.de

Martin Hofmann-Apitius
martin.hofmann-apitius@scai.fraunhofer.de

Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)
Department of Bioinformatics, Schloss Birlinghoven, Sankt Augustin, Germany

Abstract

For the recognition of gene and protein names and their normalization to gene and protein centered databases (Entrez Gene and UniProt) regularly updated dictionaries generated from these sources are used by the ProMiner system to search gene and protein names in scientific publications. A multistep curation process and inclusion of different biomedical dictionaries in the curation process leads to an increase of precision and recall. The recognition of names containing special parenthesis expressions augments the recall further. Human gene and protein names in the test corpus provided in BioCreAtIvE II could be recognized with the adapted ProMiner system and a regularly updated dictionary with a final F-measure of 80 %.

Keywords: named entity recognition, text-mining, gene normalization

1 Introduction

The ProMiner system was developed for the automatic generation of gene and protein name dictionaries and their recognition in scientific texts. The performance of the approach taken with ProMiner was already demonstrated in BioCreAtIvE I where an F-score of 0.8 could be reached for fly and mouse and an F-score of 0.9 for the yeast organism [3]. For BioCreAtIvE II two different training corpora, an automatic generated noisy training set (5000 Medline abstracts) and a manual annotated corpus (282 abstracts) including Entrez Gene identifier of the occurring human genes were provided. The performance of the entity recognition procedure was estimated on an independent set of 262 abstracts. Annotations for these sets were not available during the competition. On basis of a gold standard provided by human experts, submitted results were assessed by the organizers.

The recognition of human gene and protein names with ProMiner has already been used in different application scenarios like the generation of disease centric databases, e.g. the Auto Immune Data Base (AIDB) [5] or an intracranial aneurysm knowledge base in the European project @neurIST[1]. The ProMiner system includes an updating and dictionary curation process to generate gene and protein name dictionaries from the databases Entrez Gene [7] and UniProt [1]. Here, we describe the standard updating process for the human dictionary and adaptations made to the BioCreAtIvE sets. Furthermore, in the recognition module an extension for the recognition of names containing special parenthesis expressions is integrated.

2 The ProMiner software for recognition of gene and protein names

The ProMiner system has already been described in detail in ([2,3]). In this paper we give a short overview (cf. figure 1) over the sources used for the generation of the dictionaries, the different ProMiner modules, and the adaptations made for the BioCreAtIvE II gene normalization assessment.

The human dictionary is extracted from the gene description fields of human Entrez Gene entries and the protein description fields of human UniProt entries. All entries that are transitively mapped to each other in

[1] http://www.aneurist.org/

the International Protein Index (IPI) [6] are merged to one dictionary entry. For the BioCreAtIvE assessment we separate all entities containing more than one Entrez Gene entry. The dictionary used for the BioCreAtIvE assessment is based on an extraction of all files from release date 1st August 2006.

In the automatic dictionary curation, several functionalities such as acronym expansion, addition of spelling variants or filtering synonyms on the basis of regular expressions are covered. Its tasks are to add certain terms like long-forms of acronyms or spelling variant like IL1 (in addition to IL 1) to the dictionary in order to gain recall or to detect unspecific synonyms to either prune them from the dictionary (i.e. 35 kDa protein) or mark them for later processing (i.e. ambiguous synonyms). In the human dictionary, one-word synonyms are expanded with a leading „h“ (e.g. hSMRP). The new name is included (in addition to the original one) only if it is unique in the dictionary.

Additionally, a manually curated list generated through inspection of various training corpora in different former and ongoing projects (independent from BioCreAtIvE e.g. in the context of [5]) is used for curation of the human dictionary. Furthermore, we extract from the Open Biomedical Ontology (OBO) site different ontologies[2] for disease, tissue, organism and protein family names (BioMed terminology). In order to prune such unspecific gene and protein names all human dictionary synonyms matching in a ProMiner search to names from the BioMed terminology dictionary are removed. For BioCreAtIvE II false positive hit lists generated by ProMiner runs on the training set and the noisy training set are inspected by a curator and added to the curation list.

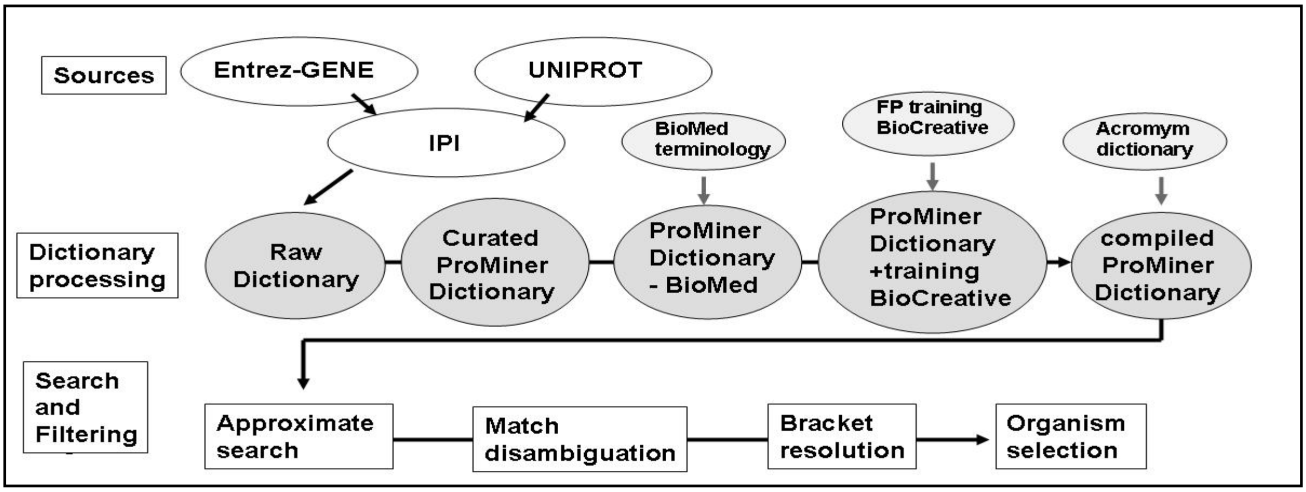

Figure 1: The ProMiner system used in BioCreAtIvE gene normalization

Finally an algorithm similar to [4] is used to extract acronyms and their long forms from all MEDLINE abstracts, generating an *acronym dictionary*. A *gene search specific acronym dictionary* is generated through the reduction to acronyms similar to gene names and removal of long forms containing gene or protein names. In the compilation step all synonyms (also acronyms and their long forms) are classified into one of several classes, which are searched with specific parameter settings like *case sensitive*, *exact* or *permuted* search in the subsequent search queries.

The search system is based on an approximate string matching algorithm enabling not only exact matches but also small variations in spelling. Synonyms which are contained in more than one Entrez Gene entry or additionally found in the acronym dictionary are labeled as ambiguous and the number of different Entrez Gene entries are memorized (D_{occur}). Hits of ambiguous synonyms are only accepted if another unique match (not labeled as ambiguous) of the same entry is found or if the user assigned disambiguation threshold (D#) is higher than the number of different Entrez Gene entries (D_{occur}).

In the training set several protein names are split by the insertion of acronyms put in brackets. As result the full name is not found (coenzyme A (HMG-CoA) synthase).To solve this problem, three runs were made. In the first run the original text was used. For the second run the full bracketed expression was removed and in the third run only the brackets are deleted. The runs are merged and the ProMiner ambiguous filter selects the appropriate matches. In order to disambiguate genes between different organisms the NCBI taxonomy

[2] Sources: http://obo.sourceforge.net/: UniProt taxonomy, Brenda tissue, Mouse adult gross anatomy, Mouse pathology, cellular component, Cell type, DiseaseOntologyV2_1

database [8] is integrated in our system and a simple co-occurrence approach is applied. A gene is rejected from the result set when it is mentioned together with any other organism or different ancestor in the phylogentic tree than Homo sapiens. A relational database system which will be described elsewhere (Dach et al., in preparation) and recursive SQL was used to accomplished this step.

3 Results

Three different runs are computed and submitted, intended to meet highest F-score, precision or recall of the ProMiner system (cf. table 1, bold). Overall, all three runs generate results which position our approach in the first quartile of all participants. The runs differ in the setting for the disambiguation threshold (controlling the result set for matches of gene names which are not unique in the dictionary) and the organism filter. In the first run this threshold was set to one (D1) allowing no matches of non-unique dictionary names. These conditions result in the highest F-measure (0.799). The second run accepts ambiguous matches (D3), increasing recall (+ 0.035) but this is accompanied by a high loss in precision (- 0.054). In the last run we also take the recognition of organism names into account and remove matches in abstracts/sentences only talking about other organisms. This approach leads to a slightly better precision (+ 0.002) but is accompanied by a high loss of recall (- 0.038) and an overall loss in F-measure (-0.02). The original dictionary not adapted to the BioCreAtIvE training corpora demonstrates a loss in precision of 0.024 compared to the final dictionary used in the submitted runs. To show the impact of ambiguity within the dictionary and maximum reachable hits with our dictionary we used a dictionary containing only the gold standard genes. Here precision as well as recall were increased by 0.05. The inclusion of bracket resolution on the training corpus results in an increase of 0.02 in recall but can not be reproduced on the test set. In this case, no differences can be observed between the different runs.

Table 1 ProMiner results

ProMiner runs on the test corpus (Test) with different user assigned disambiguation thresholds (D1, D3), organism selection (O+, O-) were submitted (Run 1-3). The next two columns present results on the test corpus with the originally dictionary without any BioCreAtIvE training (DictOrig) or a dictionary subset containing only gene entities from the gold standard (DictSub). The result on the training corpus is shown in the Train column. The last two columns provide results with a reduced ProMiner run containing no bracket resolution (-brackets) on the training and test corpus.

	Test Run 1	**Test Run 2**	**Test Run 3**	Test DictOrig	Test DictSub	Train	Train-brackets	Test-brackets
	D1, O-	**D3, O-**	**D1, O+**	D1, O-	D1, O-	D1, O-	D1, O-	D1, O-
F-measure	**0.799**	**0.790**	**0.779**	0.792	0.847	0.784	0.776	0.799
Recall	**0.768**	**0.803**	**0.730**	0.777	0.811	0.755	0.736	0.768
Precision	**0.833**	**0.779**	**0.835**	0.809	0.885	0.819	0.820	0.833
Quartile	**1**	**1**	**1**					

References

[1] Bairoch, A., Apweiler, R., Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS.: The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 33: D154-159, 2005.

[2] Hanisch D, Fluck J, Mevissen H, Zimmer R: Playing Biology's Name Game: Identifying Protein Names in Scientific Text. *Pacific Symposium on Biocomputing*, 8:403–414 2003.

[3] Hanisch, D., Fundel, K., Mevissen, H., Zimmer, R., Fluck, J.: ProMiner: Rule based protein and gene entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S14, 2005.

[4] Schwartz AS, Hearst MA: Identifying Abbreviation Definitions in Biomedical Text. *Pacific Symposium on Biocomputing*, 451-462, 2003.

[5] Karopka T, Fluck J, Mevissen HT, Glass A.: The Autoimmune Disease Database: a dynamically compiled literature-derived database. *BMC Bioinformatics*, 7:325**,** 2006.

[6] Kersey P. J., Duarte J., Williams A., Karavidopoulou Y., Birney E., Apweiler R.: The International Protein Index: An integrated database for proteomics experiments. *Proteomics,* 4(7): 1985-1988, 2004.

[7] Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T.: Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, 1;33:D54-8, 2005.

[8] Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA (2000). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res., 1;28(1):10-4, 2000*.