

A new methodology for gene normalization using a mix of taggers, global alignment matching and document similarity disambiguation

Mariana Neves¹, Monica Chagoyen¹, José M Carazo¹, Alberto Pascual-Montano²

¹Biocomputing Unit, Centro Nacional de Biotecnología – CSIC, C/ Darwin 3, Campus de Cantoblanco, 28049, Madrid, Spain.

²Departamento de Arquitectura de Computadores, Universidad Complutense de Madrid, Facultad de Ciencias Físicas, 28040, Madrid, Spain.

Abstract: This work proposes a new solution to the gene normalization problem by using simple gene/protein dictionaries, global alignment with predefined costs and document similarity for the disambiguation step. A mix of taggers was used to extract the mentions that are further processed in the search of the candidate gene/protein identifiers and a disambiguation of them is performed for the decision of a single solution for each mention. The matching procedure is provided by a global alignment with predefined costs for the inclusion, deletion and substitution operations. The synonyms in the dictionary are represented according to a tree retrieval structure (trie) to provide a faster approximate search. If necessary, a disambiguation step is performed based on the cosine similarity between the article and a document representative of each of the candidates, which is composed by terms related to each of the candidate genes. The evaluation of the methodology proposed resulted in a 94.92% of recall (F-Measure of 91.55%) for the yeast.

1. Motivation

The great number of scientific publications in the area of molecular biology associated to the necessity of biologists to interpret the contents of many of these articles has resulted in advances in biological text mining in order to extract information that could be helpful the biologists. The gene/protein mention and normalization are preceding tasks to some other text mining tasks, such as the identification of protein-protein interactions. The named entity recognition problem consists of identifying an entity (that may be a gene or a protein) in a text (scientific article) while the normalization task tries to map the extracted mention to its identifier in a specific database.

The two editions of the BioCreative evaluation [Hirschman et al. 2005; Morgan and Hirschman 2007] have allowed a coherent comparison among the many proposals to the solution of these problems. Furthermore, the workshops organized by BioCreative resulted in an opportunity for sharing the knowledge and experience obtained during the development of the systems and the discussion of the changes proposed by the community for the next editions of the competition.

This paper propose a new methodology to the gene normalization problem using a mix of taggers to extract the genes and protein from the text, an approximated matching based on global alignment with predefined costs and a disambiguation step that uses the cosine similarity between the text and a document representative of each of the candidate genes. The next section of this paper gives an overview of the methods already presented for the gene normalization problem, followed by the description of the methodology proposed here, results and future works.

2. Background

Many different solutions have already been proposed to the gene normalization problem and most of them shares the usual steps that consists in first extracting the gene and protein mentions from the text, followed by a matching procedure against a pre-processed dictionary of synonyms for the organisms related and an optional last step of filtering the results and/or the disambiguation of the candidates in the case that more than one identifier is found for the same mention.

The first step is usually the extraction of the mentions from the text and it may be performed by the same system responsible for the normalization [Fundel et al. 2005], or by one (or more) of the reliable freely available systems, such as Abner [Settles 2005] and Banner [Leaman and Gonzalez 2008] taggers, that claims an F-Measure of almost 70% and 82% on the BioCreative dataset, respectively.

The matching procedure may be an exact or rule-based approximated one against an extensive curated dictionary of synonyms [Fundel et al. 2005] or a string similarity approach [Crim et al. 2005] as the Jaro-Winkler metric [Cohen et al. 2003]. The dictionary of synonyms may be constructed by mixing many gene and protein Web databases [Liu et al. 2004] and synonyms may be pruned manually by experts [Fundel et al. 2005] or automatically [Crim et al. 2005].

The last and optional step (although often present) of the gene normalization problem is the disambiguation and filtering tasks. Many are the different approaches that have been proposed by the authors, as for example Support Vector Machine filters [Fundel et al. 2005] or a similarity measure between the abstract and some disambiguation vectors for each gene [Liu et al. 2004].

The evaluation of the gene normalization task is performed by concepts of recall (percentage of correctly extracted identifiers in comparison to the really correct ones), precision (percentage of correctly extracted identifiers in comparison to all of those obtained by the system) and F-Measure (harmonic average of the two preceding concepts). These concepts are calculated based on the concepts of true positive (correct extracted identifiers), false positive (false extracted identifiers) and false negative (correct not extracted identifiers).

3. Methods

The dataset of documents used for the training and testing of the system were the ones provided by BioCreative task 1B [Hirschman et al. 2005] for yeast, mouse and fly and by BioCreative 2 Gene Normalization task [Morgan and Hirschman 2007] for the human. The number of documents in each dataset (training, development and test) for each organism is presented in Table 1. The use of the training and the development corpora to train and test the system during development phase and a blind test set allows a correct comparison of the results obtained here with those from the participants of both competitions. The dataset also includes a dictionary of synonyms for each organism (Table 1) and these were the initial lists that were used for the training and testing of the system, with some few changes that are described in the next section.

3.1. Extraction of Mentions

For the extraction of gene and protein mentions the system makes use of an ensemble of taggers, including the one developed by the authors and based on Case-Based Reasoning (CBR) [Neves 2007] and the freely available Abner [Settles 2005] and Banner [Leaman and Gonzalez 2008] taggers. Many tests were performed in order to decide which combination of the three taggers produced the best results and a combination of all taggers was the final decision in order to obtain the higher recall.

The mentions obtained from each of the taggers are unified (no repetition) and those that match with stopwords (<http://www.unine.ch/info/clef/englishST.txt>) or the BioThesaurus (<http://pir.georgetown.edu/pirwww/iprolink/biothesaurus.shtml>), a freely available lexicon of biomedical terms and common English, are not considered for normalization. It were also removed mentions that coincide with organism names (such as “human” or “s. cerevisiae”, of one-character length and the one composed by Greek letters, numbers or Roman numerals only.

After all theses preprocessing tasks, the mentions are ready to be presented to the matching procedure in order to decide the corresponding gene/protein identifiers, if any. In order to improve recall and considering that the mentions extracted from the taggers may include more

than one entity, the mentions are tokenized by space and hyphen separators. Also, the obtained tokens were ordered alphabetically so as to avoid mismatching between mention and synonym due to different ordering of the same words, as proposed in [Liu et al. 2004].

	Yeast	Mouse	Fly	Human
Training	5000	5000	5000	281
Development	110	250	108	-
Test	250	250	250	262
Original list	14,633	130,208	92,540	176,541
Final list	15,815	113,190	92,538	176,541

Table 1: Corpora and dictionaries of BioCreative 1 and 2.

3.2. Dictionary of Synonyms

The dictionary of synonyms considered here is the one provided by the BioCreative competition and the number of synonyms for each organism is presented in Table 1. No operations such as exclusion of punctuations, conversion of number and Greek letters are performed on the synonyms so as they may be the most similar to the original dictionary provided by BioCreative.

The approximated matching procedure requires that each extracted mention should be compared to each synonym in the dictionary what may result in a long time processing. In order to improve the performance of the system, the synonyms have been stored in the database as a trie (a retrieval tree) [Shang and Merrett 1996] in which each word is represented by nodes of single characters according to a tree so as that words with the same prefix are located in common branches of the tree. The advantage of using a trie is that there is no need of performing repeated alignment operations when comparing a mention with synonyms that share the same prefix. Also, the search through a branch is aborted if the minimum actual cost of the alignment is higher than a predefined threshold. The result of this strategy is a clear reduction in the time of processing with no loss in the quality of the comparison.

3.3. Matching of Mentions and Synonyms

The matching strategy is an initial exact matching and if it fails, an approximated matching based on global alignments with predefined costs is used as proposed in [Tsuruoka and Tsujii 2003] for the gene mention problem. The initial costs of substitution, inclusion and deletion of characters were the ones proposed in the mentioned paper and were afterwards tuned according to experiments carried on with the development corpus for the four organisms in consideration. The final costs for each operation are presented in the Table 2.

Characters	Inclusion Deletion	Characters	Substitution
Numeral	50	Numeral by numeral	50
Punctuation/Space	1	Numeral by letter	100
Letter “s” (plurals)	10	Punctuation/Space	1
First letter “h” (human)	10	Letter by letter	50
Last letter “p” or “c” (yeast)	10	else	50
else	50		

Table 2: Predefined costs for the global alignment algorithm.

When a mention is compared against a synonym, the lower-most and right-most cell of the dynamic programming matrix is the final result of the matching to which is added 0.4 (a constant value proposed by [Tsuruoka and Tsujii 2003]) and normalized by the length of the synonym in consideration. If this normalized cost is lower than a parameterized threshold (defined as 3.0), the matching is considered valid. The comparison is performed for the synonyms in the trie through

the branches and the exploration of a branch is interrupted if none of the values in the dynamic programming matrix is lower than the threshold value, i.e., there is no more possibility of finding a matching with the synonyms under the considered branch of the tree.

3.4. Disambiguation

In case of more than one identifier is matched for the same mention, a disambiguation procedure is executed in order to decide which of the candidates is the most likely to be correct. The decision among the candidates is performed by a document similarity between the text in consideration and a document representative (referred here as gene-document) of each of the genes/proteins in consideration. A gene-document is constructed for each gene/protein based on text extracted from some fields of the freely available Web databases, such as SGD (<http://www.yeastgenome.org/>), MGI (<http://www.informatics.jax.org/>), FlyBase (<http://flybase.org/>) and Entrez Gene (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>). These were the database identifiers used in BioCreative datasets (list of synonym and gold standard). The fields of the databases considered for the construction of the gene-documents were the ones for symbols, aliases, descriptions, summaries, products, phenotypes, relationships, interactions, etc. Also, it was also taken in account information related to Gene Ontology (<http://www.geneontology.org/>) terms associated to each entity.

The text contained in these fields was tokenized and the resulting tokens were stored as a bag of words. A vector space model composed by these tokens was constructed to each document, with the exception of the words that match ordinal and cardinal numbers, predefined unit measures (such as “10-kb”) and stopwords. The resulting words are reduced to their stem with the Porter stemmer (<http://tartarus.org/~martin/PorterStemmer/>) and finally weighted in the document according to the TF-IDF measure [Shatkay and Feldman 2003]. This procedure is performed to each candidate gene-document and the article in consideration and the final decision is given by the highest cosine similarity [Shatkay and Feldman 2003] between the article and the gene-document.

4. Results

The results obtained with the BioCreative datasets are summarized in the Table 3. The table presents the precision, recall and F-Measure values and compares the last of them to the best results of the two editions of the BioCreative evaluation.

	Yeast	Mouse	Fly	Human
Precision	94.92	64.30	44.24	54.83
Recall	88.42	76.47	56.41	81.66
F-Measure	91.55	69.86	49.59	65.61
Best BC F-M	92.10	79.10	81.50	79.00

Table 3: Results for the BioCreative dataset.

The results are promising as it may be noticed that the yeast F-Measure is almost as high as the best BioCreative results. The human recall is also high enough while the remaining values need still to be improved. The fly recall is particularly low due to the fact that the taggers were not able to extract all the correct mentions from and documents so as these mentions could have been tried to be normalized by the system.

Also, the disambiguation step needs also to be improved as some of the false positive mistakes that decreases the precision values for the mouse, fly and human are due to the disambiguation decision for the wrong candidate. The same problem does not happen to the yeast dataset as ambiguity is not really a concern to this organism, as discussed in [Hirschman et al. 2005].

5. Conclusions and future works

The experiments related here have confirmed that it is feasible to obtain good results by using a simple dictionary of synonyms, available along with the corpora of BioCreative and for any user, and an approximate matching procedure along with a simple disambiguation step.

Further experiments in order to improve the matching step include some changes in the CBR tagger and trying machine learning algorithms (such as Support Vector Machines, Random Forests) that might be able to perform efficiently for different organisms with no need for a manual tuning of the methods according to the particularities of the organisms in consideration. Also, improvements in the disambiguation step might include a change in the selection of the words considered for the vector space model and the use of machine learning algorithms that might take in account some other features.

6. Acknowledgements

This work has been partially funded by the Spanish grants BIO2007-67150-C03-02, S-Gen-0166/2006, CYTED-505PI0058, TIN2005-5619, PR27/05-13964-BSCH. APM acknowledges the support of the Spanish Ramón y Cajal program. The authors acknowledge support from Integromics, S.L.

7. References

- Cohen W C, Ravikumar P and Fienberg S E. (2003) A Comparison of String Distance Metrics for Name-Matching Tasks. *IIWeb Wokshop on IJCAI-2003, International Joint Conference on Artificial Intelligence*, pp.73-78.
- Crim J, McDonald R and Pereira F. (2005) Automatically annotating documents with normalized gene lists. *BMC Bioinformatics*, 6(Suppl I): S13.
- Fundel K, Güttler D, Zimmer R and Apostolakis J. (2005) A simple approach for protein name identification: prospects and limits. *BMC Bioinformatics*, 6(Suppl I): S15.
- Hirschman L, Colosimo M, Morgan A and Yeh A. (2005) Overview of BioCreAtIvE task 1B: normalized gene list. *BMC Bioinformatics*, 6(Suppl I):S11. (<http://biocreative.sourceforge.net/>)
- Leaman R and Gonzalez G. (2008) BANNER: An executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing*, 13:652-663. (<http://banner.sourceforge.net/>)
- Liu H, Wu C and Friedman C. (2004) BioTagger: A Biological Entity Tagging System. *BioCreAtIvE Workshop Handouts*, Spain.
- Morgan A and Hirschman L. (2007) Overview of BioCreative II Gene Normalization. *Second BioCreative Challenge Evaluation Workshop*, pp. 17-27, Spain.
- Neves M. (2007) Identifying Gene Mentions by Case-Based Classification. *Second BioCreative Challenge Evaluation Workshop*, pp. 77-79, Spain.
- Settles B. (2005). ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 21(14):3191-3192. (<http://pages.cs.wisc.edu/~bsettles/abner/>)
- Shang H and Merrett T. (1996) Trie for approximate string matching. *IEEE Transactions on Knowledge and Data Engineering*, 8(4).
- Shatkay H and Feldman R. (2003) Mining the Biomedical Literature in the Genomic Era: An Overview. *Journal of Computational Biology*, vol. 10, number 6, pp. 821-855.
- Tsuruoka Y and Tsujii J. (2003) Boosting Precision and Recall of Dictionary-Based Protein Name Recognition. *ACL-03 Workshop on Natural Language Processing in Biomedicine*, pp. 41-48, Japan.