# Species taxonomy for gene normalization

György Móra*[1] and Richárd Farkas*[2]

[1]University of Szeged, Department of Informatics, Szeged, Hungary
[2]Hungarian Academy of Sciences, Research Group on Artificial Intelligence, Szeged, Hungary

Email: *gymora@inf.u-szeged.hu; *rfarkas@inf.u-szeged.hu;

*Corresponding author

## Abstract

**Background:** The task of gene normalization is to assign a unique identifier from a database to the gene mentions. Using these identifiers a great deal of information can be gathered from external databases such as interactions, pathways, sequences and protein structures. Normalizing gene mentions in articles is a difficult task as the inter-species ambiguity of the gene mentions in biomedical publications is high. The experiences gained from the BioCreative II Gene Normalization Task indicate that the biggest challenge in gene normalization is the recognition of the species that a specific gene mention belongs to. In biomedical scientific articles the authors often use taxonomical entities besides concrete species mentions as references to different group of organisms. Species taxonomies are hierarchical systems (trees) of living creatures and therefore provide a classification of species. Here we investigate the added value of the utilization of taxonomic entity mentions in the inter-species gene normalization task.

**Results:** We present a method which marks those words mentioning all taxonomic entities (genus, family, etc.) and applies filtering heuristics to select the taxonomic entities referring to species mentioned in the document. These entities are then treated as species mentions together with standard species annotations and we employ them in gene normalization.

**Conclusion:** After experiments were carried out on the BioCreative III Gene Normalization Task's data-set to investigate the contribution of the additional species mentions to the gene disambiguation task, we found that our approach improves the performance of the inter-species gene mention disambiguator, both in terms of precision and recall.

## Background

A vast amount of information is present in biological scientific publications. Even a complete subset of these documents in a particular scientific topic is too large for any scientist to read nowadays. This is why search engines and information extraction systems have been developed to support the life scientist in finding the information needed. The key building brick of an information extraction system is a named entity recognizer which can identify biological entities such as genes and gene products, cell lines and organism names in a text.

Besides the identification of entity mentions it is important to normalize them. Gene normalization (GN) is a process where unique database identifiers are assigned to gene mentions, where these mentions

refer to a specific gene entry. Gene databases may contain information related to these genes such as sequences, gene products and interaction and pathway information. For instance, a system applying entity normalization can assist automatic pathway finding systems and support pharmacological investigations.

Recent studies [1, 2] indicate that the intra-species ambiguity of gene symbols is much lower than the ambiguity between species, so it is important to determine which species the gene mention belongs to. The use of synonyms instead of official gene symbols also increases ambiguity and some authors prefer to use these synonyms instead of official symbols.

Current inter-species GN approaches focus on species words and employ species mention detectors to recognize them [2, 3]. Then normalization systems use machine learnt model or hand-crafted rules to determine the species associated with a particular gene mention. The NACTEM's Species Disambiguator we applied makes use of a natural language parser to exploit the linguistical relations between the gene and species mentions [3].

There are species mention detectors available with suitable precision and recall, but these systems focus on identifying exact species mentions, such as the scientific and common names of living organisms but not the names of groups, classes, genus or other taxonomic categories. The authors can use these taxonomic names in a general way for a group of species or as references to a finite set of species mentioned earlier in the document. If the taxonomic name refers to an exact species, then an inter-species gene disambiguation system can exploit this information.

We used the BioCreative III full-text, document-level corpus for our evaluation because we found no suitable mention-level gold-standard dataset with inter-species gene normalization. Current trends in biomedical text mining are directed towards systems that work on full-text articles rather than just abstracts. The two other corpora [2, 3] available for inter-species GN evaluation are based on the Biocreative II Gene Normalization Task's dataset and consist of biomedical article abstracts [4].

The corpus used by Hakenberg et al. [2] contains all of the gene identifiers mentioned in the abstract, but it is annotated only at the document level. Although the corpus introduced by Wang et al. [3] is annotated at the mention level, every entity was annotated with only one gene id and in cases like *'rat and human BMP4'*, multiple identifiers should be assigned. We compared the annotations of the two corpora – they consist of the same document set – and there were significant differences in the genes annotated for a given document. We decided to use the BioCreativeIII corpus and not the document-level corpus used by Hakenberg et al. [2] because the latter contain only abstracts and we think that full text documents contain more taxonomical entity mentions because of their writing style.

We implemented a taxonomic name identifying system that tags expressions in biomedical scientific texts mentioning taxonomic entities (TE), and with heuristical rules determines the exact species that they refer to. Our approach was extrinsically evaluated in an inter-species gene mention normalization setting. Our results show that the annotation of TEs does indeed improve the performance of a state-of-the-art GN system.

## Methods

### Dataset

We used the BioCreative III Shared Task Gene Normalization dataset for the evaluation [5]. The dataset consists of manually annotated full-text articles. A subset of the documents was fully annotated and in the remaining part only the important genes were recognized. Here we used the fully annotated subset of documents for the evaluation of our approach.

Just genes and gene products from the Entrez Gene database [6] that were clearly related to a species were annotated. Genes which had no Entrez gene identifier and gene mentions that refer to a group of genes were not annotated at all. The annotation also does not contain a gene mention when the species associated with the gene cannot be determined – even with domain knowledge. Entrez Gene identifiers of genes contained in each document were provided without the given gene mention being marked.

We annotated the species names in the documents by the LINNAEUS species name identification system [7] for biomedical literature, which assigns NCBI Taxonomy [6] identifiers to species mentions.

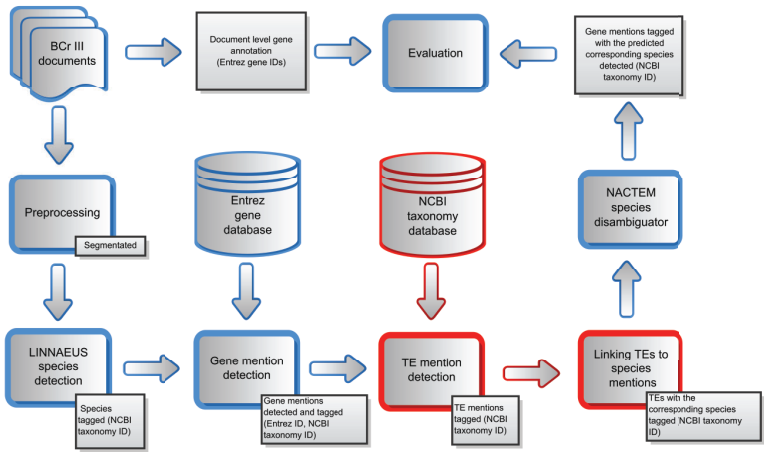


Figure 1: Flowchart of the experimental set-up (the TE mention recognition and mapping subsystems are marked with red)

### Gene mention tagging

The gene mentions were tagged in the document by our dictionary-based gene mention tagger, which assigned all of the possible Entrez Gene identifiers to the gene mentions. The dictionary mapping is based on the NLM's string normalizing method. The normalized substrings of each sentence are matched against the normalized synonyms of Entrez Gene names in our database. Then hand-crafted rules are applied to filter out false positive entity mentions and eliminate overlapping annotations of the same gene mention. One-token long entities are only accepted when they contain numerals or non-standard capitalization and if they are at least two characters long. Mentions longer than one token are accepted without restriction. The gene mentions had the possible Entrez Gene identifiers with the gene's NCBI Taxonomy species id assigned.

### Experimental set-up

We used the NACTEM's Species Disambiguator component from the uCompare system to provide inter-species gene-normalization [3, 8]. This component assigns NCBI Taxonomy identifiers to each gene mention. The module applies the species annotations in the document to determine the species associated with the gene mentions. Two different types of analysis were carried out. One was with just the species mentions tagged by LINNAEUS (baseline) and the other used an extended set of species mentions containing TEs mapped to species by our system. The only difference was that the second set-up included our TE mention mapping module and the TE mentions were mapped to species mentions before gene mention normalization. The gene-mention normalization was then evaluated at the document level. A flowchart of the experimental set-up can be seen in Figure 1. With this we investigated the added value of TEs inside a state-of-the art gene mention recognizer (leaving the other component of the system unchanged).

### Recognizing alternative species mentions

The annotation of taxonomic entities (TE) was done using the same method as that for gene tagging. The synonyms of the NCBI Taxonomy entries were matched against the text and taxonomy identifiers were assigned to the mentions. TE mentions referring to taxonomic groups that had no members annotated in the text were filtered out.

The references between TE and species mentions were identified by using the following set of heuristical rules:

- Only species descending from the taxonomic category of the TE in the NCBI Taxonomy were regarded as possibly referred species.
- If the sentence containing the TE mention also contained a candidate species mention like

| | TE - | TE + | Tagged TE - | Tagged TE + |
|---|---|---|---|---|
| Precision | 0.668 | 0.695 | 0.668 | 0.695 |
| Recall | 0.571 | 0.610 | 0.798 | 0.853 |
| F-measure | 0.616 | 0.650 | 0.727 | 0.766 |

Table 1: Performance values of the GN setting without (TE -) and with alternative species (TE +) utilized in the normalization. Results marked with "tagged" were produced by an evaluation applied only on a subset of the genes taken from the evaluation set, which were then successfully mapped to the documents by the dictionary mapper.

this, then the TE was considered to refer to this species.

- If multiple species satisfied the descendant criteria then the taxonomic entry was considered to refer to multiple species or refer to the general taxonomic class and both were removed.
- If there was no species annotated in the sentence, the search was continued at the paragraph, section, and document level, respectively.
- At the end only TE mentions annotated with one species were kept and used as alternative species mentions in our experiments.

## Results

The NACTEM's inter-species gene normalization system tagged the gene mentions with a species identifier, but the datasets available consist of Entrez Gene identifiers assigned to the documents. To evaluate the performance of our approach we mapped the species identifiers assigned by the normalizer to gene identifiers and evaluated the resulting set of Entrez identifiers at the document level with the standard F-measure metric (see Table 1).

The dictionary mapper does not provide a mapping for each gene identifier of the evaluation data set. Therefore we provide additional scores –focusing on the performance of the inter-species normalization instead of the performance of the dictionary lookup– by removing false negatives which were not annotated by the dictionary lookup ("tagged" in Table 1).

Both the precision and the recall of the inter-species gene mention normalization rose by 4-5 percentage points when utilizing TE mentions present in biomedical articles.

## Discussion

The performance of our approach compared to the state-of-the-art baseline method has an interesting distribution. The gene normalization with alternative species mentions outperformed the baseline system in 7 out of the 32 documents and there were only two cases where our approach achieved a lower F-measure. In these two cases our method added only 1-1 false positives and hence it did not affect the overall results significantly.

There were 5 documents where there were no alternative species mentions tagged by our system, so the performance of the disambiguation was the same. In the remaining 17 documents – where the TE + and TE - achieved the same results – only a few TEs were recognized. A manual inspection of the document-set showed that these differences were caused by the different writing styles of the authors. Some authors exclusively use concrete species names when referring to an organism and also use TE names to refer to species.

We evaluated the 10 documents containing a significant amount of TE mentions and the overall F-score rose from 0.40 to 0.65. The precision and recall values went up from 0.37 to 0.56 and from 0.44 to 0.77, respectively. This subset of the BioCreative III documents represents those biomedical articles where the authors often refer to organisms using broader terms instead of using exact organism names.

The following examples show how the TE mentions can aid gene normalization.

> "Indeed, elevated expression of **Drosophila** *MOF*, which counteracts ISWI activity ..."

Here the exact organism name (*D. melanogaster*) was mentioned elsewhere in the document, so the TE (in bold type) was successfully mapped to *Drosophila*

*melanogaster* because no other species belonging to the *Drosophila* subgenus was found in the given context. The species identifier of the gene mention (in Italics) was correctly determined by utilizing the identified alternative gene mention.

Wider TEs terms (like *plants*) were also successfully mapped to the corresponding species mentions in the text and produced correct gene normalization.

> "By studying **plants** with mutations in this gene, we found that *CBP60g* contributes to the increases ..."

When no plants other than *Arabidopsis thaliana* were mentioned in the given context it was possible to identify the TE *plant* by the label *A. thaliana*.

There were some documents where both of the procedures achieved low scores. An analysis later revealed that the LINNAEUS species detector was unable to identify species mentions in some cases where the authors used only short and ambiguous variants of the organism name, like *Drosophila* instead of *D. Melanogaster* or *Drosophila Melanogaster*. Even when the TEs were identified in the document and no species mentions were annotated, the TEs were filtered out. If there was no species identified in a document the NACTEM's gene disambiguator chose *Homo sapiens (human)* as the default organism for gene normalization.

If a TE covers a large number of organisms (like the TE *animal*), then false positive species associations can occur. For example, if the author uses a TE as a general term rather than as a taxonomic category. In the next negative example the word animal was referenced to *C. elegans* by mistake, but the word was used in the sense of other mammals like human rather than a worm like *C. elegans*. As a result *HCF-1 protein* was incorrectly identified as a gene product belonging to *C. elegans* instead of a human protein.

> "...we have undertaken a genetic analysis in ***C. elegans*** to study *HCF-1-protein* function in **animal** development. The C. elegans HCF-1-related protein is an amino acid protein encoded by the hcf-1 gene and referred to here as Ce HCF-1."

Another source of incorrect normalization is when the author refers globally to the group of organisms, but our heuristics link the TE to an exact species mention in the document. In the next example the TE *vertebrate* was used only to name the vertebrates in general, but it was incorrectly referenced to *D. simulans* – the only vertebrate species identified in the document. Also, *D. simulans* was incorrectly identified by LINNAEUS as a rodent (*Dipodomys simulans*) and not as an insect (*Drosophila simulans*).

> "In spite of the similar global function of **insect** and **vertebrate** *OBPs* ..."

## Conclusions

By utilizing the TE mentions as alternative species mentions, the approach we presented here improves the performance of a state-of-the-art inter-species gene normalization tool. The overall F-scores measured on the BioCreative III GN dataset rose from 0.61 to 0.65. On a subset of the dataset – where the writing style of the authors causes the classical approach to achieve poorer results than on the rest of the testset – our method increased the F-score on this set from 0.40 to 0.64.

A subsequent error analysis indicated that more sophisticated methods are required to resolve the references between TE mentions and species mentions. We plan to develop an integrated species mention and alternative organism mention system in the near future.

## Authors contributions

György Móra developed the software tools used for mention detection, taxonomy browsing, TE-to-species linking and the evaluation of the results. He was responsible for the statistical analysis done in this study. Richárd Farkas supervised the work and participated in the writing of the manuscript. The authors would like to thank those who maintain the Entrez Gene and NCBI Taxonomy databases [6], the authors of the NACTEM's Species Disambiguator [3] and the authors of the LINNAEUS species name recognizer [7] for making these tools available.

## Acknowledgements

This work was supported in part by the NKTH grant (project codename TEXTREND) of the Hungarian government.

## References


1. Chen L, Liu H, Friedman C: **Gene name ambiguity of eukaryotic nomenclatures**. *Bioinformatics* 2005, **21**(2):248–256.
2. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G: **Inter-species normalization of gene mentions with GNAT**. *Bioinformatics* 2008, **24**(16):i126–i132.
3. Wang X, Tsujii J, Ananiadou S: **Disambiguating the species of biomedical named entities using natural language parsers**. *Bioinformatics* 2010, **26**(5):661–667.
4. Morgan A, Lu Z, Wang X, Cohen A, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J, Sun C, Liu Hh, Torres R, Krauthammer M, Lau W, Liu H, Hsu CN, Schuemie M, Cohen KB, Hirschman L: **Overview of BioCreative II gene normalization**. *Genome Biology* 2008, **9**(Suppl 2):S3, [http://genomebiology.com/2008/9/S2/S3].
5. **BioCreative III Gene Normalization Task** [http://www.biocreative.org/tasks/biocreative-iii/gn/].
6. *The NCBI handbook*. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information 2002, [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books].
7. Gerner M, Nenadic G, Bergman CM: **LINNAEUS: a species name identification system for biomedical literature.** *BMC bioinformatics* 2010, **11**:85+, [http://dx.doi.org/10.1186/1471-2105-11-85].
8. Kano Y, Baumgartner WA, McCrohon L, Ananiadou S, Cohen KB, Hunter L, Tsujii J: **U-Compare: share and compare text mining tools with UIMA**. *Bioinformatics* 2009, **25**(15):1997–1998, [http://dx.doi.org/10.1093/bioinformatics/btp289].