Gene mention normalization in full texts using GNAT and LINNAEUS

Illés Solt[1,2], Martin Gerner[3], Philippe Thomas[2], Goran Nenadic[4], Casey M. Bergman[3], Ulf Leser[2], Jörg Hakenberg[5§]

[1] Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, 1117 Budapest, Hungary
[2] Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
[3] Faculty of Life Sciences, University of Manchester, Manchester, M13 9PT, UK
[4] School of Computer Science, University of Manchester, Manchester, M13 9PL, UK
[5] Computer Science Department, Arizona State University, Tempe, AZ 85287, USA
[§] Corresponding author

Email addresses:
IS: solt@tmit.bme.hu
MG: martin.gerner@postgrad.manchester.ac.uk
PT: thomas@informatik.hu-berlin.de
GN: g.nenadic@manchester.ac.uk
CB: casey.bergman@manchester.ac.uk
UL: leser@informatik.hu-berlin.de
JH: joerg.hakenberg@asu.edu

Abstract

Gene mention normalization (GN) refers to the automated mapping of gene names to a unique identifier, such as an NCBI Entrez Gene ID. Such knowledge helps in indexing and retrieval, linkage to additional information (such as sequences), database curation, and data integration. We present here an ensemble system encompassing LINNAEUS for recognizing organism names and GNAT for recognition and normalization of gene mentions, taking into account the species information provided by LINNAEUS. Candidate identifiers are filtered through a series of steps that take the local context of a given mention into account. On the BioCreative III high-quality training data, our system achieves TAP-5 and TAP-20 scores of 0.36 and 0.41, respectively. On the evaluation set of 50 documents that were provided to participants, we achieve scores of 0.16 and 0.20 for TAP-5 and TAP-20, respectively. Our analysis of the evaluation results suggests that the lower scores primarily are due to significant differences in species composition, and partly due to the method for selecting the evaluation data.

Background

BioCreative is a repeated community challenge addressing various tasks in biomedical text mining, such as named entity recognition (NER) of gene and protein names, extraction of protein-protein interactions, or protein interaction detection methods. In the fourth installment in 2010, one of the tasks addressed the recognition and normalization of gene and protein names in full text publications. Participants of this task had to provide a system capable of finding all mentions of genes or proteins in a full text article and of

mapping these mentions to their respective Entrez Gene identifiers. Challenges arise from both synonymity and homonymity. Genes frequently have multiple synonyms, usage of which differs not only between authors and journals [1], but also over time. Names often also are used for several different genes (including orthologs, paralogs or unrelated genes) or even for concepts belonging to completely different semantic classes. Developing systems that overcome these challenges is critical for advancing the application of gene mention normalization in biomedical text mining.

Methods

System overview

Our processing pipeline begins by loading the collection of texts that should be annotated, after which we perform NER of species, Gene Ontology (GO) [http://www.geneontology.org/] and MeSH [http://www.nlm.nih.gov/mesh/] terms. We then use species-specific GNAT [2] gene NER modules to find gene name matches in the texts. These modules consist of combined Entrez Gene and UniProt gene name dictionaries, expanded with typical patterns of gene name variations [5]. The recognized gene mentions are assigned candidate identifiers according to the dictionary. The gene mentions are processed by a set of rule-based methods designed to filter out and score candidate identifiers, based on their syntactic and semantic context [2]. Species disambiguation of gene mentions is done by considering the local findings of species NER. Finally, gene mentions with confidence scores above a threshold are reported.

Using LINNAEUS for species NER

In order to identify the species that are discussed in a paper (which in turn determines what genes to search for), we utilize LINNAEUS [3]. LINNAEUS uses a dictionary of expanded species terms from the NCBI taxonomy, together with a variety of rule-based methods and distributional statistics to disambiguate ambiguous species mentions and reduce the number of false positives and negatives. Compared against a corpus of 100 full-text articles manually annotated for species names, LINNAEUS achieves 94% precision and 97% recall [3]. It has previously been shown that for articles linked to genes in Entrez Gene, LINNAEUS can find the species of the referenced gene in 94% (9,662/10,290) of cases where full-text was available [3].

In order to further increase the utility of LINNAEUS for detecting focus organisms of articles, even if they are not mentioned directly, we have included additional "proxy" dictionaries that link cell-lines and genera to corresponding species. The cell-line dictionary was created from the database of [4]. Genera are also tagged and linked to the member species that is most commonly mentioned in MEDLINE (for example, "Drosophila" is linked to *Drosophila melanogaster*).

Some technical re-linking of species identifiers was also necessary due to recent changes in species associations in Entrez Gene. For example, all genes that previously were linked to *Saccharomyces cerevisiae* (NCBI Taxonomy ID 4932) were instead linked to a specific strain, *S. cer. S288c* (ID 850287). This was performed for all species where we could determine that such changes had occurred in Entrez Gene.

Filtering gene names and candidate identifiers

Dictionary-based matching allows direct assignment of candidate identifiers to recognized gene mentions, based on what dictionary entries a mention matches. In a series of filtering steps, the set of mention candidate identifiers is narrowed down successively by removing false positive gene IDs and species IDs (see Table 1 for the full list, and [5] for further details). Filtering includes:

1. Use of the sentence and paragraph context surrounding the mention. The context is matched against pre-computed gene profiles and scanned for clues indicating the presence of false positives.
2. Use of species name mentions located close to the gene mention, that are used to perform cross-species disambiguation.
3. String similarity searches of the located term against the original (not expanded) terms for the candidate identifiers, which are used to determine the closest (and most distant) matches.

Table 1. List of processing filters. Filtering steps are used to expand and reduce candidate ID lists for each gene mention. Also see Figure 1.

Filter	Filtering method
MDRER	Species-dependent gene NER
REU	Joins overlapping or adjacent gene names
LRCF	Match the text surrounding the mention against context models of FPs
ICF	Filter false positives by immediate context
loadGR	Load the gene profile for each candidate gene
UNF	Filter names that refer to gene families and other un-specific mentions
NVF	Restore names removed during UNF where a synonym is used elsewhere
AF	Score mentions by string similarity against unexpanded gene synonyms
SVF	Verify ambiguous species names (“cancer”)
UMF	Mark genes that are unambiguous throughout the text as identified
MSDF	Gene mention disambiguation by context profile
ITF	Adjust mention scores based on whether the terms have been found italicized in other PubMed Central articles
SCSA	Assign relative scores to candidates per text
SCSF	Adjust scores to fit the TAP scoring scheme

Scoring candidate identifiers using context profiles for disambiguation

Gene mention disambiguation in our system is handled by an adaptation of GNAT [2]. Adjustments include: (i) more localized reliability scoring of candidate identifiers using paragraph contexts; (ii) keeping annotations consistent across paragraphs; and (iii) text-wide search for the best evidence to map a gene mention to a species.

Selecting the set of species-specific dictionaries

Due to memory constraints, the gene name dictionaries used by GNAT are restricted to a set of model organisms. The selection of what species to include is critical since it determines the species for which GNAT can recognize gene names. The species were chosen based on mention frequencies in MEDLINE and PubMed Central, to cover the majority of articles discussing particular species. In total, we used gene name dictionaries with genes from 32 species (see Table 2), covering 69% of all species mentions in MEDLINE and PubMed Central.

Table 2. List of species-specific dictionaries.
List of species for which we built and used gene name dictionaries. Column two and three give occurrence statistics in the training and test sets (species with no associated genes in both the training or test sets were omitted due to space constraints). The frequencies represent the number of genes associated to each species.

Species	Training frequency	Test frequency
Homo sapiens	121 (19.9%)	181 (10.8%)
Mus musculus	75 (12.3%)	235 (14%)
Rattus norvegicus	14 (2.3%)	41 (2.4%)
Gallus gallus	10 (1.6%)	4 (0.2%)
Saccharomyces cerevisiae S288c	166 (27.3%)	36 (2.1%)
Escherichia coli str. K-12 substr. MG1655	4 (0.6%)	1 (0%)
Arabidopsis thaliana	30 (4.9%)	9 (0.5%)
Drosophila melanogaster	58 (9.5%)	59 (3.5%)
Bos taurus	9 (1.4%)	3 (0.1%)
Caenorhabditis elegans	19 (3.1%)	9 (0.5%)
Xenopus laevis	17 (2.8%)	3 (0.1%)
Danio rerio	42 (6.9%)	7 (0.4%)
Hepatitis C virus	0	1 (0%)
Magnaporthe oryzae 70-15	0	68 (4%)
Neurospora crassa OR74A	0	2 (0.1%)
Schizosaccharomyces pombe	3 (0.4%)	5 (0.2%)
Zea mays	2 (0.3%)	0
Human immunodeficiency virus 1	0	1 (0%)
Sus scrofa	7 (1.1%)	76 (4.5%)
Triticum aestivum	2 (0.3%)	2 (0.1%)
Xenopus (Silurana) tropicalis	1 (0.1%)	0
Macaca mulatta	2 (0.3%)	0
Total	582 (95.1%)	743 (43.5%)

Results and discussion

TAP, F1-score, recall and precision on the training and test corpora

The TAP-5 scores [6] of our system on the training and test data are 0.363 and 0.157, respectively; the corresponding TAP-20 scores are 0.408 and 0.199. Per the construction of the test set, these data were considered more difficult than the training data (see overview paper). On the high-quality part of the training set, we achieved precision, recall, and F1-score of 0.536, 0.474, and 0.503, respectively.

Species recognition results

By applying LINNAEUS to the training and test corpora and comparing the identified species against the manually annotated gene identifiers, we evaluated to what extent LINNAEUS was able to find mentions belonging to the species that are associated with genes in particular papers. For the fully annotated subset of the training corpus (32 documents), the original version of LINNAEUS could find species mentions for 87% (528/607) of the annotated gene entries. When also incorporating the additional

dictionaries produced as part of this work (using cell-lines as proxies for species and linking genus names to commonly mentioned species), this rate increased to 94% (571/607). The species identifier "re-linking" was performed in both cases. Performance was lower for the manually annotated subset of the test set (50 documents), where the software was able to locate only 74% (1,242/1,670) and 80% (1,341/1,670) of gene-associated species using the original and extended dictionaries, respectively.

For both the training and test set, a preliminary inspection of a subset of false negatives suggests that the main reason for false negatives is that articles simply do not contain the appropriate species name. While it may be possible to reduce this problem by adding additional "proxy" dictionaries, it is probably not possible to completely solve it.

Analysis of filtering steps

To assess the impact of the individual components used by GNAT, we performed accuracy evaluations of the predicted gene mentions throughout the GNAT pipeline. We evaluated each filtering step, from initial species-dependent gene NER to the final disambiguation and scoring, on the high-quality portions of the BC III training set (see Table 1 and Figure 1). This analysis show that the pipeline methods that contributed the most to the increase in accuracy were the context-based filters (LRCF and ICF), the string similarity search filter (AF), the species disambiguation filter (SVF) and the gene re-classification filter (UMF).

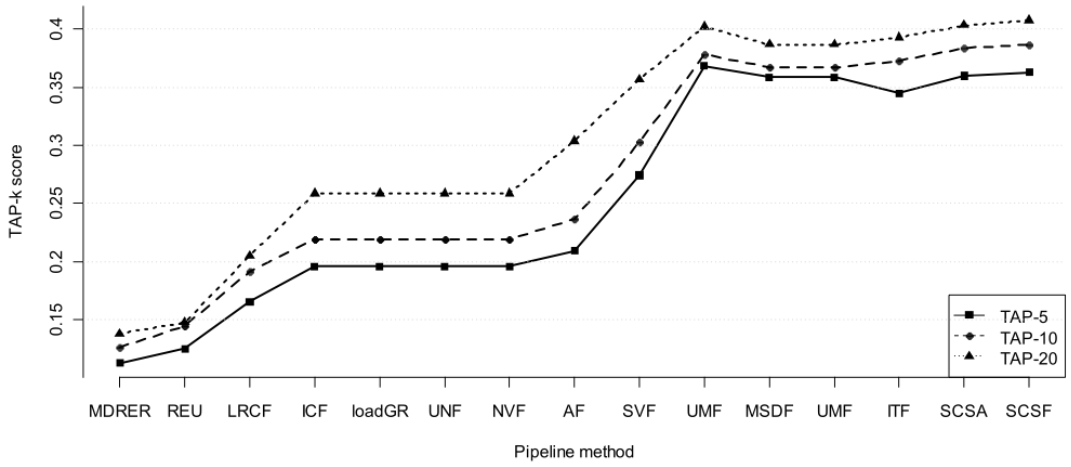

Figure 1. TAP scores after individual filtering steps on the training data. TAP-5, TAP-10, and TAP-20 scores as observed after each individual step of our processing pipeline. Table 1 describes each filtering step.

Differences between the training and test set

Our analysis of the results on the test set suggests that the primary reason for the difference in accuracy seen between the test and training set is the difference in species composition. Our species-specific gene dictionaries covered the species associated to 95% of the annotated gene entries in the training set, but only 43% of the genes in the test set (see Table 2), causing a large number of false negatives. Model organisms were much more common in the training set than in the test set, where species discussed less frequently have a more important role. For instance, 22% of the gene entries in the test

data are from *Enterobacter sp. 638*, a species mentioned extremely rarely in MEDLINE. It is clear that while the common model species are heavily over-represented in research [3], the species-specific gene dictionaries used by GNAT represent a limitation for articles that discuss less-frequently mentioned species.

Conclusions and availability

Here, we presented a system for recognizing and normalizing gene mentions in full texts used in the BioCreative III challenge's Gene Normalization (GN) task. We demonstrate the utility for species NER for guiding gene name dictionary recognition, and for gene context profiles used when performing gene normalization. Our training and test set performances differ widely, with TAP-20 scores ranging from 0.4 to 0.2. This difference can primarily be attributed to differences in species composition that could not be handled using the species-restricted approach used by our system, and to some extent the method used for the selection of test data used for evaluation (see overview article). Future work will concentrate on making the initial dictionary NER method less dependent on species-specific dictionaries in order to overcome this problem.

GNAT will be made available at http://gnat.sourceforge.net shortly after the BioCreative III workshop. LINNAEUS and the additional genus and cell-line dictionaries are available at http://linnaeus.sourceforge.net.

Author contributions

IS and MG implemented the adaptations of GNAT and LINNAEUS for BC III. MG and JH wrote the manuscript, with help from the other authors. PT tested GNAT for high-throughput applications. GN, CMB, and UL supervised the work. All authors have read and approved the final manuscript.

Acknowledgements

IS was supported by the Alexander-von-Humboldt Stiftung. MG was supported by the University of Manchester and a BBSRC CASE studentship. PT was supported by the BMBF, grant 0315417B. GN and CMB were supported in part by BBSRC, grant BB/G000093/1. JH was supported by Arizona State University.

References

1. Tamames J, Valencia A: **The success (or not) of HUGO nomenclature.** *Genome Biology* 2006, **7:**402.
2. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzales G: **Inter-species normalization of gene mentions with GNAT.** *Bioinformatics* 2008, **24:**i126-i132.
3. Gerner M, Nenadic G, Bergman CM: **LINNAEUS: a species name identification system for biomedical literature.** *BMC Bioinformatics* 2010, **11:**85.
4. Sarntivijai S, Ade AS, Athey BD, States DJ: **A bioinformatics analysis of the cell line nomenclature.** *Bioinformatics* 2008, **24:**2760-2766.
5. Hakenberg J, Plake C, Royer L, Strobelt H, Leser U, Schroeder M: **Gene mention normalization and interaction extraction with context models and sentence motifs.** *Genome Biology* 2008, **9:**S14.
6. Carroll HD, Kann MG, Sheetlin SL, Spouge JL: **Threshold Average Precision (TAP-k): a measure of retrieval designed for bioinformatics.** *Bioinformatics* 2010, **26:**1708-1713.