

POSTER PRESENTATION

Open Access

Species identification for gene name normalization

Illés Solt^{1,2*}, Domonkos Tikk^{1,2}, Ulf Leser¹

From Workshop on Advances in Bio Text Mining
 Ghent, Belgium. 10-11 May 2010

Background

Protein interaction networks are expensive to construct experimentally. Therefore, researchers usually refer to the literature or domain-specific databases to convey knowledge on currently known interactions. Yet the task of manual collection of knowledge from scientific papers is labor intensive, and therefore should be automated to the extent possible. For this, an important step is identifying gene and protein names (termed entities). After

identification, gene names must be mapped to database identifiers to connect them to structured knowledge. One particular problem in this step are homonymous, i.e., identical names referring to different genes in different species.

Methods

We present different approaches that aim at assigning species labels to MEDLINE abstracts. We use (1) as a

Table 1 Comparison of methods for document-level species annotation

| Species | Method | GS: MeSH terms | | | GS: UniProt references | | |
|---------------|----------------------|----------------|---------|---------|------------------------|---------|---------|
| | | P | R | F | P | R | F |
| Human | journal heuristic | 0.908 | 0.632 | 0.745 | (0.011) | 0.231 | (0.021) |
| | SVM | 0.710 | 0.775 | 0.741 | (0.024) | 0.781 | (0.046) |
| | Ali Baba [1] | 0.888 | 0.583 | 0.703 | (0.033) | 0.654 | (0.063) |
| | LINNAEUS [2] | 0.900 | 0.660 | 0.761 | (0.030) | 0.659 | (0.057) |
| | GNAT [3] (Ali Baba) | 0.878 | 0.318 | 0.467 | (0.056) | 0.618 | (0.103) |
| | GNAT (LINNAEUS) | 0.609 | 0.507 | 0.553 | (0.037) | 0.944 | (0.072) |
| | UniProt | 0.934 | 0.031 | 0.060 | (1.000) | (1.000) | (1.000) |
| <i>E.Coli</i> | journal heuristic | 0.146 | 0.310 | 0.198 | (0.008) | 0.468 | (0.015) |
| | SVM | 0.217 | 0.289 | 0.248 | (0.010) | 0.387 | (0.019) |
| | Ali Baba | 0.654 | 0.605 | 0.628 | (0.031) | 0.829 | (0.059) |
| | LINNAEUS | 0.665 | 0.602 | 0.632 | (0.032) | 0.838 | (0.061) |
| | GNAT (Ali Baba) | 0.771 | 0.301 | 0.434 | (0.064) | 0.730 | (0.118) |
| | GNAT (LINNAEUS) | 0.058 | 0.415 | 0.102 | (0.004) | 0.847 | (0.008) |
| | UniProt | 0.946 | (0.032) | (0.063) | (1.000) | (1.000) | (1.000) |
| RegulonDB [4] | 0.857 | (0.107) | (0.191) | (0.175) | 0.640 | (0.275) | |

Legend: GS - gold standard species labeling. Only human and *E. coli* shown for brevity. For comparison, we also provide inter-gold standard agreement between MeSH, UniProt and RegulonDB. Using UniProt as gold standard, only recall can be compared in a cross-corpus sense as UniProt does not reference all papers mentioning a protein. For the same reason, when using databases for prediction, only precision is comparable.

* Correspondence: solt@informatik.hu-berlin.de

¹Knowledge Management in Bioinformatics, Institute for Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
 Full list of author information is available at the end of the article

baseline, the most frequent species MeSH term of the corresponding journal represented as MeSH terms; (2) the prediction of a binary classifier (SVM) for each species; (3) species names found by the tools Ali Baba [1] or LINNAEUS [2]; (4) the species of a normalized protein mention found by GNAT [3]. For evaluation, we use two sources as gold standard document-level annotations: The MeSH terms from MEDLINE and the species from UniProt and the *E. coli*-specific RegulonDB via protein- MEDLINE references.

Results

Measurements on a random set of 200 k abstracts from MEDLINE are summarized in Table 1. For MeSH term prediction, the text based methods (Ali Baba, LINNAEUS, GNAT) show stable performance across species, while the classification methods, as they rely on training data, suffer for species with lower *prior* probability. For the most frequent species human, the bag-of-word based SVM overcomes the difficulty of missing explicit species mention by learning other clues. Using UniProt as gold standard, learning methods produce substantially higher recall, indicating that molecular biology papers are more explicitly mentioning their focus organisms. There is a considerable disagreement between gold standard databases, e.g., only 85.7 % of the papers referenced from a comprehensive *E. coli*-specific database are annotated as *E. coli* by MeSH. Reasons for this could be, i.e., incompleteness of MeSH annotations or consideration of orthologs in RegulonDB.

Conclusion

We conclude that there is no one-size-fits-all method for identifying species in abstracts. For less frequent species, direct species mention identification methods work best. The advantage of using indirect clues could only be realized for the most frequent species human, suggesting that machine learning methods should be applied after better balancing the training data. We also showed that using MeSH term queries to filter papers poses considerable limitations on recall.

Acknowledgements

Domonkos Tikk was supported by the Alexander-von-Humboldt Foundation.

Author details

¹Knowledge Management in Bioinformatics, Institute for Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.

²Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, H-1117 Budapest, Magyar Tudósok krt 2., Hungary.

Published: 6 October 2010

References

1. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U: **AliBaba: PubMed as a graph.** *Bioinformatics* 2006, **22**(19):2444-2445.
2. Gerner M, Nenadic G, Bergman C: **LINNAEUS: A species name identification system for biomedical literature.** *BMC Bioinformatics* 2010, **11**:85.
3. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G: **Inter-species normalization of gene mentions with GNAT.** *Bioinformatics* 2008, **24**(16):126-132.
4. Salgado H, Santos-Zavaleta A, Gama-Castro S, Peralta-Gil M, Penaloza-Spinola M, Martinez-Antonio A, Karp P, Collado-Vides J: **The comprehensive updated regulatory network of Escherichia coli K-12.** *BMC Bioinformatics* 2006, **7**:5.

doi:10.1186/1471-2105-11-S5-P5

Cite this article as: Solt et al.: Species identification for gene name normalization. *BMC Bioinformatics* 2010 **11**(Suppl 5):P5.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

