

A rule-based approach for automatically identifying gene and protein names in MEDLINE abstracts

Hong Yu,¹ Vasileios Hatzivassiloglou,²
Carol Friedman,^{1,3} Ivan H. Iossifov,⁴ Andrey Rzhetsky,^{1,4} and W. John Wilbur⁵

¹Department of Medical Informatics, ²Department of Computer Science, Columbia University, New York

³Department of Computer Science, Queens College, City University of New York, New York

⁴Columbia Genome Center, Columbia University, New York

⁵National Center for Biotechnology Information, Bethesda, USA

ABSTRACT

Identifying gene and protein terms is important for obtaining biological knowledge from literature. We have developed GPmarkup (for gene and protein-name mark up), a system that automatically identifies gene and protein terms and maps gene and protein symbols (e.g., *DR3*) to names (e.g., *Death Receptor 3*) in MEDLINE abstracts.

INTRODUCTION

MEDLINE is a rich resource for biological knowledge. Natural language processing techniques are promising in extracting biological knowledge from the literature (Friedman et al. 2001). Identifying gene and protein terms in MEDLINE is a first step in applying a natural language processor to extract automatically the knowledge related to genes and proteins from the abstracts.

We developed GPmarkup (for Gene and Protein-name mark up), a software system that maps gene and protein symbols (e.g., *LARD*) to names (e.g., *lymphocyte associated receptor of death*) in MEDLINE abstracts. We also applied GPmarkup to 11 million MEDLINE abstracts (year 1966-2001) to automatically generate a knowledge source for paired gene and protein symbols and names. GPmarkup then applies the knowledge source to identify the remaining gene and protein terms in the abstracts. Note that GPmarkup does not differentiate gene terms from protein terms. Hatzivassiloglou et al (2001) has applied machine-learning approaches to disambiguate gene and protein names with 85% precision.

METHODS

We manually examined the published guidelines of the nomenclature for genes and proteins and developed a set of pattern-matching rules that map gene and protein symbols to names. The pattern-matching rules include some special abbreviations that represent amino acids (e.g., *Y* for *tyrosine*) and common conventions author apply for an abbreviation (e.g., The abbreviation matches the first letter of each word in the full form) (Yu et al. 2002). We

implemented the pattern-matching rules in GPmarkup and applied it to 11 million MEDLINE abstracts to generate a knowledge source of paired abbreviations and full forms. We then developed a rule-based approach to determine gene and protein symbols and names from the knowledge source of paired abbreviations and full forms. Our rules include "if an abbreviation contains a number, the abbreviation and full form is a pair of gene and protein symbol and name only if the full form contains one and more of the keywords including *protein(s)*, *gene(s)*, *peptide(s)*, *molecule(s)*, and *enzyme(s)*." We then applied the knowledge source of paired gene and protein symbols and names to mark up the remaining gene and protein terms in MEDLINE abstracts.

EVALUATION RESULTS

We applied GPmarkup to MEDLINE abstracts. We randomly selected (by publication time) 25 MEDLINE abstracts and evaluated the recall and precision of GPmarkup in identifying gene and protein terms and mapping gene and protein symbols to names. GPmarkup identified a total of 105 gene and protein terms, where 95 of them are correct. GPmarkup missed 47 gene and protein terms. The recall and precision of GPmarkup were 72.0% and 90.5%.

ACKNOWLEDGEMENT

Hong Yu is supported by LM07079 "Research Training Grant".

REFERENCES

- Friedman, C., P.Kra, H.Yu, M.Krauthammer and A.Rzhetsky. 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* 17: S74-82.
- Hatzivassiloglou, V., P.A. Duboue, and A.Rzhetsky. 2001. Disambiguating proteins, genes and RNA in text: a machine learning approach. *Bioinformatics* 17: S97-106.
- Yu, H., G. Hripcsak, and C. Friedman. 2002. Mapping abbreviations to full forms in biomedical articles. *J Am Med Inform Assoc* 9: 262-72.