Large-scale Extraction of Protein/Gene Relations for Model Organisms

Jasmin Šarić [a] **Lars Juhl Jensen** [b] **Rossitza Ouzounova** [b] **Isabel Rojas** [a] **Peer Bork** [b]
[a] EML Research gGmbH, Heidelberg, Germany
[b] European Molecular Biology Laboratory, Heidelberg, Germany

Motivation

We have previously developed a rule based approach for extracting information on the regulation of gene expression in yeast. The biomedical literature, however, contains information on several other equally important regulatory mechanisms, in particular phosphorylation, which we now expanded our rule based system to also extract.

Results

This paper presents new results for extraction of relational information from biomedical text. We have improved our system to both capture new types of linguistic constructions as well as new types of biological information (i.e.(de-)phosphorylation). The precision of our system is stable with a slight increase in recall. From almost one million PubMed abstracts related to four model organisms, we manage to extract regulatory networks and binary phosphorylations comprising **3319** relation chunks. The accuracy is **83–90%** and **86–95%** for gene expression and (de-)phosphorylation relations, respectively. To achieve this, we made use of an organism-specific resource of gene/protein names considerably larger than those used in most other biology related information extraction approaches. These names were included in the lexicon when retraining the part-of-speech tagger on the GENIA corpus. For the domain in question an accuracy of **96.4%** was attained on POS-tags. It should be noted that the rules were developed for yeast and successfully applied to both abstracts and full-text articles related to other organisms with comparable accuracy.

Availability

The revised GENIA corpus, the POS-tagger, and the full sets of extracted relations are available upon request.

Full article forthcoming in *Bioinformatics*.

Address for Correspondence:
Jasmin Šarić
EML Research gGmbH
Villa Bosch
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg
Germany
phone: +49-6221-533-256
fax: +49-6221-533-298
Jasmin.Saric@eml-r.villa-bosch.de