bioRxiv preprint first posted online Feb. 23, 2017; doi: http://dx.doi.org/10.1101/111088. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY 4.0 International license.

EXTRACT 2.0: text-mining-assisted interactive annotation of biomedical named entities and ontology terms

Evangelos Pafilis[1,], Rūdolfs Bērziņš[2], and Lars Juhl Jensen[2,*]*

[1] *Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, P.O. Box 2214, 71003 Heraklion, Crete, Greece*
[2] *Cellular Network Biology Group, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen N, Denmark*

1 INTRODUCTION

Databases increasingly rely on text-mining tools to support the curation process. The BioCreative interactive annotation task recently evaluated several such tools and found our tool EXTRACT to perform favorably in terms of usability and accelerated curation by 15–25% (Wang *et al.*, 2016).

The original version of EXTRACT was designed to support annotation of metagenomic samples with semantically controlled environmental descriptors (Pafilis *et al.*, 2016). For this reason, it focused on named entity recognition of terms from the Environment Ontology (ENVO) (Buttigieg *et al.*, 2016) and ontologies relevant for describing host organisms (https://www.ncbi.nlm.nih.gov/taxonomy), tissues (Placzek *et al.*, 2017), and disease states (Kibbe *et al.*, 2015).

2 EXPANDED SCOPE OF THE TOOL

EXTRACT 2.0 expands the scope of the tool in several new directions with the aim to make it more broadly useful.

We expanded the scope from covering only diseases to covering phenotypes in general. To this end, we complemented the existing disease dictionary with terms from the Mammalian Phenotype Ontology (MPO) (Smith and Eppig, 2012). To avoid redundancy in the dictionary, we excluded MPO terms that clashed with terms already in the disease dictionary. To improve recall, we added plural and adjective endings to the names and generated variants of the form *pronoun of noun* from names of the form *noun pronoun*.

To cover also important concepts of molecular and cellular biology, we further expanded the dictionary with Gene Ontology (GO). The names from GO were processed similar to those from MPO to generate variants and improve recall.

In addition to adding more biomedical ontologies, we have expanded the tool with named entity recognition of molecular entities. To this end, we have included dictionaries of protein-coding and non-coding RNA (ncRNA) genes from STRING (Szkararczyk *et al.*, 2017) and RAIN (Junge *et al.*, 2017), respectively. We have furthermore added a dictionary of drugs and other small molecule compounds from the STITCH database (Szklarczyk *et al.*, 2016).

Together, these additional types of entities have made EXTRACT 2.0 potentially useful for many more tasks than just annotation of metagenomic samples. For example, it can be used to help annotate both proteins and ncRNAs with functions, processes, subcellular localization, tissue expression, and associated diseases. The tool, API, and documentation are freely accessible at http://extract.jensenlab.org/.

* To whom correspondence should be addressed: pafilis@hcmr.gr, lars.juhl.jensen@cpr.ku.dk

FUNDING

The Novo Nordisk Foundation (NNF14CC0001).

REFERENCES

Wang,Q., Abdul,S.S., Almeida,L. *et al.* (2016). Overview of the interactive task in BioCreative V. *Database*, **2016**, baw119.

Pafilis,E., Buttigieg,P.L., Ferrell,B. *et al.* (2016) EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation. *Database*, **2016**, baw005.

Buttigieg,P.L., Pafilis,E., Lewis,S.E. *et al.* (2016) The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *J. Biomed. Semant.*, **7**, 57.

Placzek,S., Schomburg,I., Chang,A. *et al.* (2017) BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.*, **45**, D380–D388.

Kibbe,W.A., Arze,C., Felix,V. *et al.* (2015) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.*, **43**, D1071–D1078.

Smith,C.L. and Eppig,J.T. (2012) The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm. Genome*, **23**, 654–668.

Szklarczyk,D., Morris,J.H., Cook,H. *et al.* (2017). The STRING database in 2017: quality-controlled protein-protein association. *Nucleic Acids Res.*, **45**, D362–D368.

Junge,A., Refsgaard,,J.C., Garde,C. *et al.* (2017). RAIN: RNA–protein Association and Interaction Networks. *Database*, **2017**, baw167.

Szklarczyk,D., Santos,A., von Mering,C. *et al.* (2016). STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.*, **44**, D380–D384.