Text-mining-based retrieval of protein networks

Lars Juhl Jensen[1]* and John H. Morris[2]

[1] Disease Systems Biology Program, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
[2] Resource for Biocomputing, Visualization, and Informatics, University of California, San Francisco, USA.
* Contact lars.juhl.jensen@cpr.ku.dk

Protein networks can provide an overview of the often hundreds of proteins involved in a given process. However, the starting point — a comprehensive list of the proteins — is usually not readily available. We have developed a text-mining-based system that allows researchers to quickly and easily retrieve a network of proteins, starting from a PubMed query, and plan to use the Open Annotation standard to track provenance of the results.

The text-mining-based network retrieval is made available to end users via the new Cytoscape app for the STRING database (http://apps.cytoscape.org/apps/stringapp). Once installed, the user can import a network by specifying a PubMed query that defines the topic, the organism of interest, the desired number of proteins, and the confidence cutoff for the interactions in the network. The app then queries the NCBI E-utilities to obtain the PMIDs of abstracts that match the query (at most 40,000), submits these to one REST API to obtain a ranked list of proteins mentioned within the abstracts, and finally submits this list of proteins to another API to obtain an association network from the STRING database (1).

On the server side, proteins from the specified organism are ranked based on how often they appear in abstracts matching the PubMed query relative to in PubMed as a whole. To facilitate this, we on a weekly basis run named entity recognition (NER) on a local copy of PubMed using a highly efficient dictionary-based tagging engine (2). The NER results and precalculated background counts are stored and indexed in a PostgreSQL database, allowing the scoring and ranking of proteins to be performed entirely within database engine as a single SQL query. This makes it possible to rank all proteins with respect to 40,000 abstracts in a just few seconds.

Whereas the system is already fully functional, there is currently no mechanism for users to track the provenance of the underlying text mining. We plan to embed this in the existing JSON response, using of the Open Annotation standard (3) to represent the NER results.

This work was in part funded by the Novo Nordisk Foundation (NNF14CC0001).

1. Szklarczyk D, Franceschini A, Wyder S, *et al.* (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**:D447–D452.
2. Pafilis E, Pletscher-Frankild S, Fanini L, et al. (2013). The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. PLoS One, 8:e65390.
3. Pyysalo S, Campos J, Cejuela JM, *et al.* (2015). Sharing annotations better: RESTful Open Annotation. *Proceedings of ACL-IJCNLP 2015*, 91–96.