Textpresso

Information Retrieval and Extraction System for Biological Literature

Hans-Michael Müller, Arun Rangarajan, Tracy K. Teal, Kimberly van Auken, Juancarlos Chan and Paul W. Sternberg.

California Institute of Technology
Pasadena, CA 91125, USA

We developed an information retrieval and extraction system that processes the full text of biological papers. The system, called Textpresso, separates text into sentences, labels words and phrases according to an ontology (an organized lexicon), and allows queries to be performed on a database of labeled sentences. The current ontology comprises approximately one hundred categories of terms, such as “gene”, “regulation”, “human disease”, “brain area” etc., and also contains main Gene Ontology (GO) categories. Extraction of particular biological facts, such as gene-gene interactions, or the curation of GO cellular components, can be accelerated significantly by ontologies, with Textpresso automatically performing nearly as well as expert curators to identify sentences. Search engine for four literatures, *C. elegans, Drosophila, Arabidopsis* and Neuroscience have been established by us, and thirteen systems for other literatures have been developed by other groups around the world. Currently, our four systems contain 112,000 papers with 40 million sentences, all systems worldwide contain 190,000 papers with approximately 65 million sentences.

Dagstuhl Seminar Proceedings 08131
Ontologies and Text Mining for Life Sciences : Current Status and Future Perspectives
http://drops.dagstuhl.de/opus/volltexte/2008/1516