Poster K-14

BioMint: a database curator's assistant for biomedical text processing

Authors:
Anne-Lise Veuthey (*Swiss Institute of Bioinformatics*)
Violaine Pillet (*Swiss Institute of Bioinformatics*)
Marc Zehnder (*Swiss Institute of Bioinformatics*)

Short Abstract: BioMinT is an information retrieval system which helps literature screening for Swiss-Prot curation. It provides a graphical user interface which expands a query on a gene/protein with synonyms, filters and ranks the retrieved abstracts according to their relevance to the query, and extracts information specific to database annotation topics.

Long Abstract:

We developed an information retrieval system to help literature screening for the annotation of the Swiss-Prot and PRINTS databases. The BioMinT prototype provides a graphical user interface which (i) expands a query on a gene/protein with synonyms, (ii) filters and ranks the abstracts retrieved from PubMed according to their relevance to the query, and (iii) retrieve information specific to topics such as protein function or subcellular location. The core of the system is composed of an information retrieval module consisting in a meta-query engine wrapped around the PubMed server. The strategy we use ensures a high recall of documents from Medline by expanding the query with related terms. For gene and protein names, such an expansion is done using a synonym database, GPSDB [1], constructed from 14 existing resources of model organisms. The database contains 559'294 synonyms referring to 292'472 proteins and is periodically updated. In this way, the user can choose the query terms from a comprehensive list of gene/protein name synonyms, sorted by species, and even distinguish between possible homonyms. The documents retrieved from PubMed are then filtered and ranked according to their relevance with regard to the query. We use the Lucene ranker for this purpose, with the possibility for the user to influence the ranking by selecting preferred and/or avoided terms. The selected documents are then processed to extract sentences containing relevant information on various topics for database annotation. Currently, classifiers using support vector machine (SVM) have been trained for extraction of sentences on eight different topics, namely protein function, domain composition, information on possible isoforms, subcellular location, tissue specificity of protein expression, associated diseases, information on protein glycosylation and phosphorylation. The algorithms were trained on corpora composed of Medline abstracts with tagged sentences. A total of 18'827 sentences on 21 biological topics relevant for Swiss-Prot annotation have been tagged. These corpora have been made available to the text-mining community (http://biomint.isb-sib.ch/). The BioMinT project was funded by the European Commission, contract-no. QLRI-CT-2002-02770 under the RTD programme