Poster K-22

EBIMed and Protein Corral: EBI's information retrieval and information extraction engines for realtime analysis of Medline abstracts

Authors:
Arregui (*EBI*)
Gaudan (*EBI*)
Kirsch (*EBI*)
Rebholz-Schuhmann (*EBI*)

Short Abstract: EBIMed and Protein Corral are EBI's new Web services based on Medline abstracts. Upon keyword search the retrieved abstracts are analyzed for associations between biomedical terms (EBIMed) and for relations between proteins (Protein Corral). The identified concepts are automatically linked EBI's and NCBI's biomedical databases.

Long Abstract:

EBIMed and Protein Corral: EBI's information retrieval and information extraction engines for realtime analysis of Medline abstracts

EBIMed is EBI's new service that combines document retrieval with co-occurrence based summarization of Medline abstracts [1, 2]. EBIMed uses EBI's inhouse Medline installation, relies on keyword query to retrieve abstracts and filters for biomedical terminology maintained in different resources [3]. The extracted terminology and sentences give an overview on associations of UniProtKb proteins with other proteins, GO annotations, drugs and species [4, 5]. The co-occurrences can be interpreted as identification of protein-protein interactions, of functional annotations, of drug targets and of model organisms.

The retrieved abstracts are processed in a set of cascaded Finite State Automata. They are set up in a pipeline of modules, each of which reads and writes XML code. Identification of UniProtKb proteins is based on the content of the UniProtKb XML file, which provides the current set of protein names and synonyms [6]. Identified proteins are marked up and are linked to the corresponding UniProtKb database server entry. In the case of ambiguous acronyms, e.g. ESC for 'embryonic stem cells', then the expanded form found in conjunction with the acronym is used for disambiguation [7] . Other ambiguous protein names that are stated without their expanded version and that have a high frequency in the British National Corpus are excluded as well, such as 'BY', 'AND' and others.

Identification of GO terms is based on matching of uppercase and lowercase GO term variants, which currently suffices the demands expressed by curators. Drug names are provided from MedlinePlus . The NCBI taxonomy serves as terminological resource for the identification of species. The priority rules for the matching of the terminology is UniProtKb first, then GO:biological process, GO:cellular location, GO:molecular function, drugs and species.

It is the main design principle of EBIMed to offer to the user free selection of Medline abstracts that are then processed, he can even submit a list of PMIDs to process the abstracts thereafter. The only dependence between query and analysis is the set of retrieved documents. The query terms are not used as parameters for the analysis and the query terms may be anything else apart from protein names, drugs, species or GO terms.

An initial keyword query to EBIMed leads to the retrieval of Medline abstracts, which are then analysed for sentences with key terminology. The keyword query gears the retrieval of documents and the analysis modules function as a filter. At present EBIMed provides annotations for any UniProtKb protein in the text, if it co-occurs with another UniProtKb protein, with a GO term, a drug or a mention of a species. Such co-occurrences can be interpreted as identification of protein-protein interactions, of functional annotations, of drug targets and of model organisms.

All findings from all retrieved Medline abstracts are presented in a table. Initially, the leftmost column lists the UniProtKb protein and all other columns list the co-occurring concepts that form a hit-pair with the protein. Due to EBIMed's capability to reshuffle the hit-pairs the left-most column can as well be the list of encountered drugs or species, which leads to a different prioritization of the content extracted by EBIMed. Above this table a display shows the total number abstracts analysed and the number hit-pairs encountered.

The number of hit-pairs increases with the number of retrieved documents and the number of contained terms from the domain of molecular biology.

An evaluation based on this service lead to the following results: Precision and recall for proteins in a random sample is 53.9% and 55.4%, respectively. Precision was 86.1%, if the identification did not require to detect the correct boundaries, e.g. 'Wnt pathway' was considered to be correct instead of considering 'Wnt' as an error. This is a valid approach, since all facts are provided with its context. 33.1% of the protein-protein co-occurrences report a protein-protein interaction. In the case of drugs co-occurring with proteins, 45% of the sentences state an effect of the drug on the protein. The analysis of 4 protein-protein pairs from the Wnt/beta-catenin pathway retrieved 39% statements of interactions and provided details concerning the Dkk/LRP interaction, which are not provided from Kegg.

Protein Corral serves the identification of protein-protein interactions only [8, 9]. Protein Corral combines co-occurrence, tri-cooccurrence and chunk parsing to provide data at different precision and recall levels. Similar to EBIMed all analyses are performed in real time after the retrieval of the Medline abstracts, in contrast to EBIMed it provides syntactical analysis of sentences.

The phrase pattern denoting an interaction makes use of verbs as well as of nominalized forms of verbs. The following general verbs are used: "dissociate, assemble, attach, bind, complex, contact, couple, dimerize, link, interact, precipitate, regulate, inhibit". In addition, the following verbs indicating post translation modifications are used: "acetylate, acylate, amidate, brominate, biotinylate, carboxylate, cysteinylate, farnesylate, formylate, hydrox[iy]late, methylate, myristo?ylate, palmito?ylate, phosphorylate, pyruvate, nitrosylate, sumoylate, ubiquitin(yl)?ate".

1. Stapley,B.J., and Benoit,G. (2000) Bibliometrics: information retrieval and visualization from co-occurrence of gene names in Medline abstracts. Pac. Symp. Biocomput. 5: 529-540.
2. Jelier,R. Jenster,G., Dorssers,L.C.J., van der Eijk,C.C., van Mulligen, E.M., Mons, B., and Kors, J.A. (2005). Co-occurrence based meta-analysis of scientific texts: Retrieving functional relationships between genes. Bioinformatics, 21(9): 2049-2058.
3. Nenadic,G., Mima,H., Spasic,I., Ananiadou,S. and Tsuji,J.-I. (2002) Terminology-driven literature mining and knowledge acquisition in biomedicine. In Int. J. Med. Informatics, 67, 33--48.
4. Apweiler,R., et al. (2004) UniProt: the Universal Protein Knowledgebase. Nucleic Acids Research, Database Issue, 32, D115–D119.
5. Gene-Ontology-Consortium (2001) Creating the gene ontology resource: design and implementation. Genome Res, 11, 1425-33.
6. Kirsch,H., Gaudan,S., Rebholz-Schuhmann,D. (2005) Distributed modules for text annotation and IE applied to the biomedical domain. Int J Med Inform. 2005 Aug 4;
7. Gaudan, S., Kirsch, H., and Rebholz-Schuhmann, D. (2005) Resolving abbreviations to their senses in Medline. Bioinformatics 21(18):3658-64
8. Blaschke,C., Hirschman,L. and Valencia,A. (2003) BioCreAtIve : Critical Assessment of Information Extraction systems in Biology. http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html (Preliminary analysis)
9. Marcotte,E.M., Xenarios,I., and Eisenberg,D. (2001) Mining literature for protein-protein interactions. Bioinformatics 17, 359-63.