GeneView – Retrieval and Visualisation of Biomedical Text Resources

Philippe Thomas\textsuperscript{1}, Johannes Starlinger\textsuperscript{1}, Jörg Hakenberg\textsuperscript{2}, Ulf Leser\textsuperscript{1}

\textsuperscript{1} Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany,
\textsuperscript{2} Computer Science Department, Arizona State University, Tempe, AZ 85287, USA

Scientific literature is one of the main resources for functional knowledge in the life sciences. For instance, MEDLINE covers currently more than 19 million citations. Life scientists spend a great amount of time searching literature for gene specific information. To this end, we introduce GeneView, a text mining application capable of searching, ranking and visualizing biomedical publications in an individual way.

Searching biological concepts like genes is difficult as they often have synonyms and an ordinary full text query returns only a subset of relevant articles. On the other hand gene names may be shared across several genes, resulting in false results. Identification and normalization of named entities enables a more precise and complete search than an ordinary full text query. Our system currently identifies and grounds four different named entities: 1.) gene/protein-, 2.) single nucleotide polymorphism-, 3.) disease- and 4.) drug-terms.

GeneView has been designed to address both BioCreative3-IAT subtasks, indexing and retrieval. Thus, a potential use case of the system is the identification and ranking of articles discussing a gene of specific interest. Our default ranker uses a TF/IDF-ranking scheme and weights important sections like title, abstract, results and discussion higher. We further include custom ranking capabilities, like prioritizing articles by the total number of genes or mutations. Additionally, we compute for each query a list of significantly associated entities using $\chi^2$ test. For example, for the query "colon cancer" well-known disease relevant genes, like ERBB2 or EGFR are identified.

For a single full text, entities are highlighted and if available, links to specialized databases like EntrezGene, UniProt and KEGG are presented. For each entity type a short synopsis of all found entities is provided. The order of entities can be adjusted by the users according to their need.

We believe that GeneView helps researchers to decrease the amount of time needed to keep pace with the increasing biomedical literature. We plan to evaluate our rankings using the gene2pubmed and GeneRif information provided by NCBI.