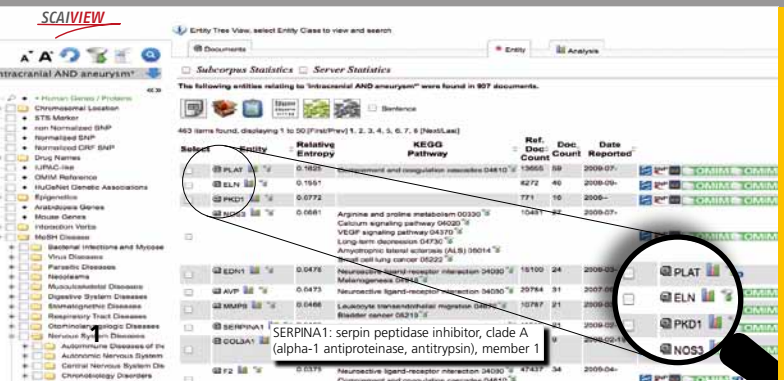
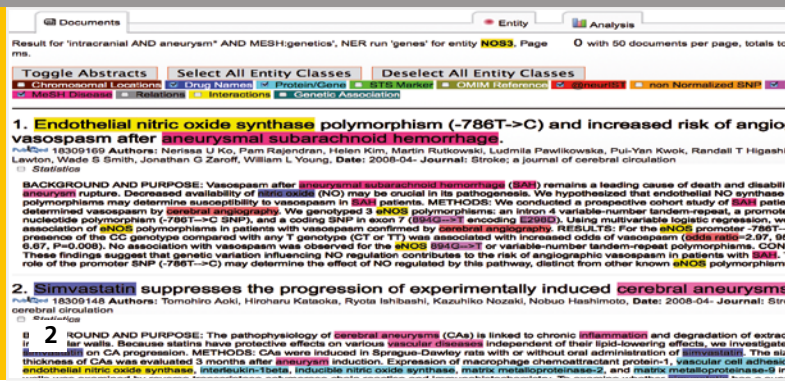
Fraunhofer
SCAI

FRAUNHOFER INSTITUTE FOR ALGORITHMS AND SCIENTIFIC COMPUTING SCAI





1 *Entity View with aggregated resultsets and linkout possibilities.*
2 *Document View: entity types are highligted in different colours, linkouts possible.*

KNOWLEDGE DISCOVERY AND SEMANTIC SEARCH IN SCIENTIFIC LITERATURE

Fraunhofer Institute for Algorithms and Scientific Computing SCAI

Schloss Birlinghoven
53754 Sankt Augustin
Germany

Contact
Dr.-Ing. Christoph M. Friedrich
phone +49 2241 14-2502
christoph.friedrich@scai.fraunhofer.de
www.scai.fraunhofer.de/scaiview

Distribution
scapos AG
phone +49 2241 14-2820
www.scapos.com

SCAIVIEW

Situation

SCAIView is an advanced search engine and addresses questions of interest to general biomedical researchers. Most of the information is given as unstructured text (publications, text fields in databases).

Advanced retrieval technology allows answering complex queries such as:

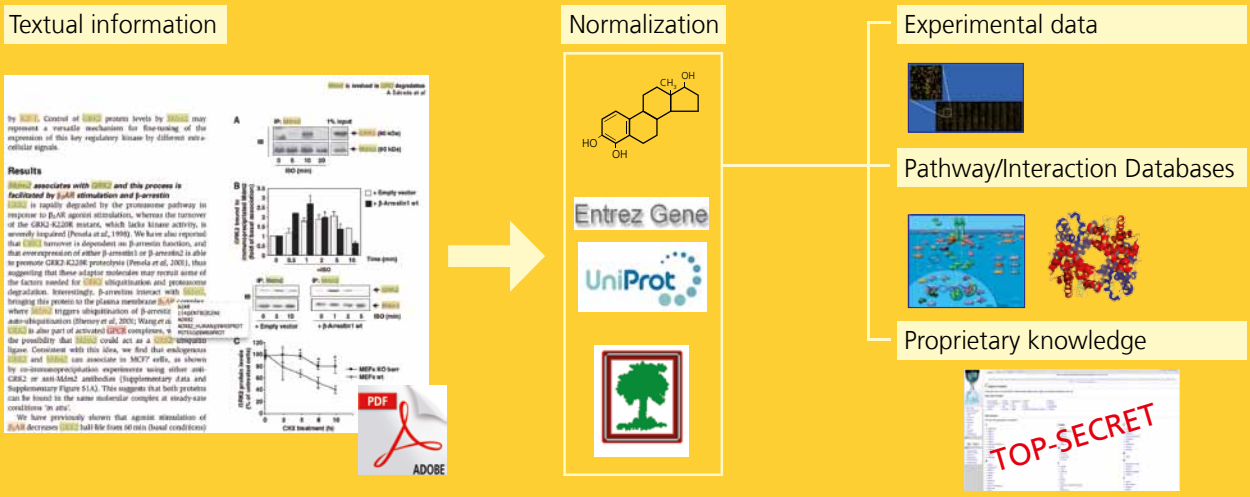
- Which genes/proteins are related to a certain context (e.g. disease/pathway/epigenetics)?
- Give me an overview of relevant biomedical concepts in my subcorpus.
- Which drugs are relevant for this context?
- To which diseases is my gene associated?
- Which chromosomes show linkage to the disease?
- Which variations are mentioned in the context of the disease and could they be found in dbSNP?
- What other diseases are possibly co-occurring with my relevant disease?

SCAIView allows for full text and biomedical concept searches with large biomedical terminologies and outstanding text mining technologies.

Document Retrieval

The documents are retrieved via free text queries in combination with semantic or ontological search of biomedical entities of interest. The biomedical entities are embedded in searchable hierarchies and span from genes, proteins, accompanied SNPs to chemical compounds and medical terminology.

With ontological Filtering, it is possible to restrict the result to a subset; e.g. genes on a KEGG pathway or in a Cytoband region.



Knowledge Discovery

The most important feature of any Knowledge Discovery tool or search engine is the ranking according to relevance of the results. For this we use a technique named relative entropy. Even if some proteins like insulin are mentioned quite often in the context of a search, it will be ranked low if it is not mentioned over-proportional in your specific query result set.

The other property of real Knowledge Discovery, novelty detection, could be shown in several biomedical applications.

Example

In a review on the "Genetics of intracranial aneurysms"[1], 18 associated genes are mentioned. A query with SCAIView for "intracranial aneurysm AND MESH: genetics" with the selection of human genes/proteins leads to the retrieval of 122 genes with top ranked hits contained in the experts review. Even new associated genes not described in the review have been found.

Successful uptake of the SCAIView information via an Application Programmers Interface for pediatric diseases in the EU project Health-e-Child could be demonstrated at HealthGrid 2009.

Advantages

- Superior text mining technology based on approximative search and machine learning
- Support for confidence information (adjustment of precision/recall)
- Combination of full text, semantic and ontology search
- Very fast retrieval from large corpora and relevance ranking of retrieved results
- Support of large resultsets (e.g. 1 Mio hits)
- Relevance ranking on aggregated entity search results
- Overview of found terminology in defined sub corpora
- Links to relevant biomedical databases (e.g. EntrezGene, dbSNP, KEGG, GO, DrugBank)
- Document visualization with user defined highlighting
- Export to Excel or Cytoscape
- Programmatic access via an Application Programmers Interface (API)

Technology

The selected biomedical entities are found by an approximate search algorithm implemented in the Fraunhofer-Gesellschaft information extraction tool ProMiner®, which additionally disambiguates synonyms of entities to unique identifiers in public available entity databases.

ProMiner has been evaluated as one of the best tools regarding protein and gene detection at the 2004 and 2007 BioCreAtIvE contest.

Additionally, non-enumerable entities like IUPAC names are found by a machine learning-based ProMiner plugin[2].

For improved performance of the retrieval engine, a highly optimized multi-threaded search engine is used.

Requirements

Client

- Firefox > 2.0
- Internet Explorer > 6.0

Server

- Minimum RAM: ≥ 2GB
- Operating System: Linux, Windows XP, Windows 7, Solaris
- Application Server: Tomcat > 5.5
- Multi-Core processors: Recommended for near linear performance scale-up

3 *Relation to experimental data, interaction data bases or propriatory data through the provided mapping.*

1 Krischek, B.; Inoue, I.: The genetics of intracranial aneurysms. Journal of Human Genetics, 2006, 51, 587-594.

2 Klinger, R.; Kolárik, C.; Fluck, J.; Hofmann-Apitius, M.; Friedrich, C. M.: Detection of IUPAC and IUPAC-like chemical names. Bioinformatics, 2008, 24, i268-i276.