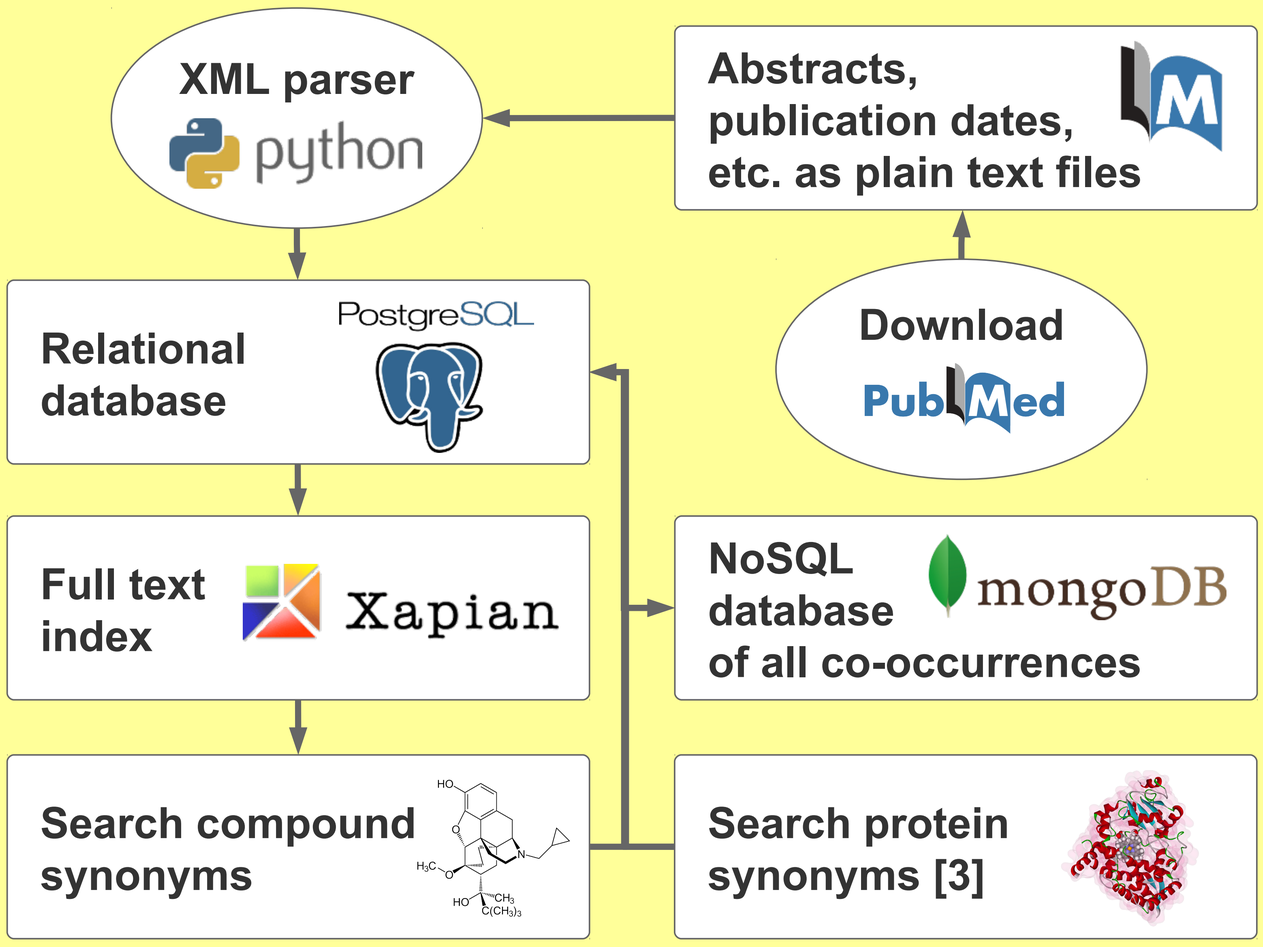
Finding Functional Interactions of Proteins and Small Molecules in Sentences of PubMed Abstracts

Döring K, Becer M, Günther S
kersten.doering@pharmazie.uni-freiburg.de
Department of Pharmaceutical Bioinformatics, Institute for Pharmaceutical Sciences,
University of Freiburg, Germany

Finding Co-occurrences



All XML files from PubMed have been downloaded, parsed and loaded into a PostgreSQL relational database. A full text index has been generated with Xapian. This "in-house" database of PubMed can be queried with different terms or synonyms, e.g. compound names. The NoSQL database MongoDB supports fast access to all co-occurring UniProt and PubChem identifiers with PubMed-IDs. Currently, the database contains around 1 Bn entries for 12.5 M abstracts.

"... 21.5 M biomedical publication titles with 12.5 M abstracts."

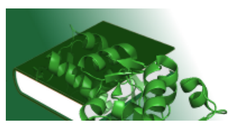
Web Services

The data can be queried by the web services *Compounds in Literature* (CIL) [1] and *Protein-Literature Investigation for Interacting Compounds* (prolific) [2] which search for co-occurrences of biomolecules in either a compound- or protein-centric view.

www.pharmaceutical-bioinformatics.de/cil
www.pharmaceutical-bioinformatics.de/prolific

"... a high frequency indicates a relationship, but what about seldom co-occurrences?."

Finding Functional Interactions



Take 10,000 abstracts from profific (year 2009)

[...] buprenorphine is metabolized through cytochrome P450 3A4 [...]

Search sentences with structure: compound – interaction word – protein [2]

Manual curation of false positive entities and annotation of true relations

1640 instances training
268 instances prediction
Class 0: no relation
Class 1: functional interaction

jSRE – java Simple Relation Extraction [4] (shallow lingusitic kernel)

Accuracy: 61.6 %
Precision: 59.6 %
Recall: 73.1 %
F1-score: 65.6 %
Specificity: 50.0 %

Next steps:
Change parameters
Cross-validation
Benchmark dataset

An arbitrary dataset of 10,000 abstracts was chosen from 2009. Sentences containing a relationship word enclosed by two biomolecules have been analysed and classified as *interaction* or *no interaction* instances. The shallow lingusitic kernel achieved a good F1-score in comparison to protein-protein or drug-drug interaction extraction results [5,6], but a low specificity, indicating that non-functional co-occurrences are more difficult to classify than functional interactions.

References

[1] Senger, Grüning *et al.,* 2012. Mining and Evaluation of Molecular Relationships in Literature. Bioinformatics 28:709-14.
[2] Grüning, Senger *et al.,* 2011. Compounds In Literature (CIL): screening for compounds and relatives in PubMed. Bioinformatics 27:1341-2.
[3] Rebholz-Schuhmann *et al.,* 2008. Text processing through Web services: calling Whatizit. Bioinformatics. 24:296-8.
[4] Giuliano *et al.,* 2006. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In: Proc. of the 11st Conf. of the European Chapter of the Association for Computational Linguistics (EACL'06).
[5] Tikk *et al.,* 2011. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. PLoS Comput. Biol. 6:e1000837.
[6] Segura-Bedmar *et al.,* 2008. Using a Shallow Linguistic Kernel for Drug-Drug Interaction Extraction. J. Biomed. Inform. 44:789-804.

The working group of Pharmaceutical Bioinformatics at the Institute for Pharmaceutical Sciences develops algorithms and software for pharmaceutical research. Our fields of research include the modeling of molecular interactions, prediction of biological effects of molecules, identification of potential new drug agents, analysis of gene expression and methylation data as well as text and data mining. The working group is part of the University of Freiburg's Research Group Program of the Excellence Initiative of the federal and state governments.

http://www.pharmaceutical-bioinformatics.com/

This work is part of the CoRS project, which is funded by the German National Research Foundation (DFG, Lis45).

DFG Deutsche Forschungsgemeinschaft

UNI FREIBURG