Text Mining of Protein-Compound Interactions on PubMed Abstracts

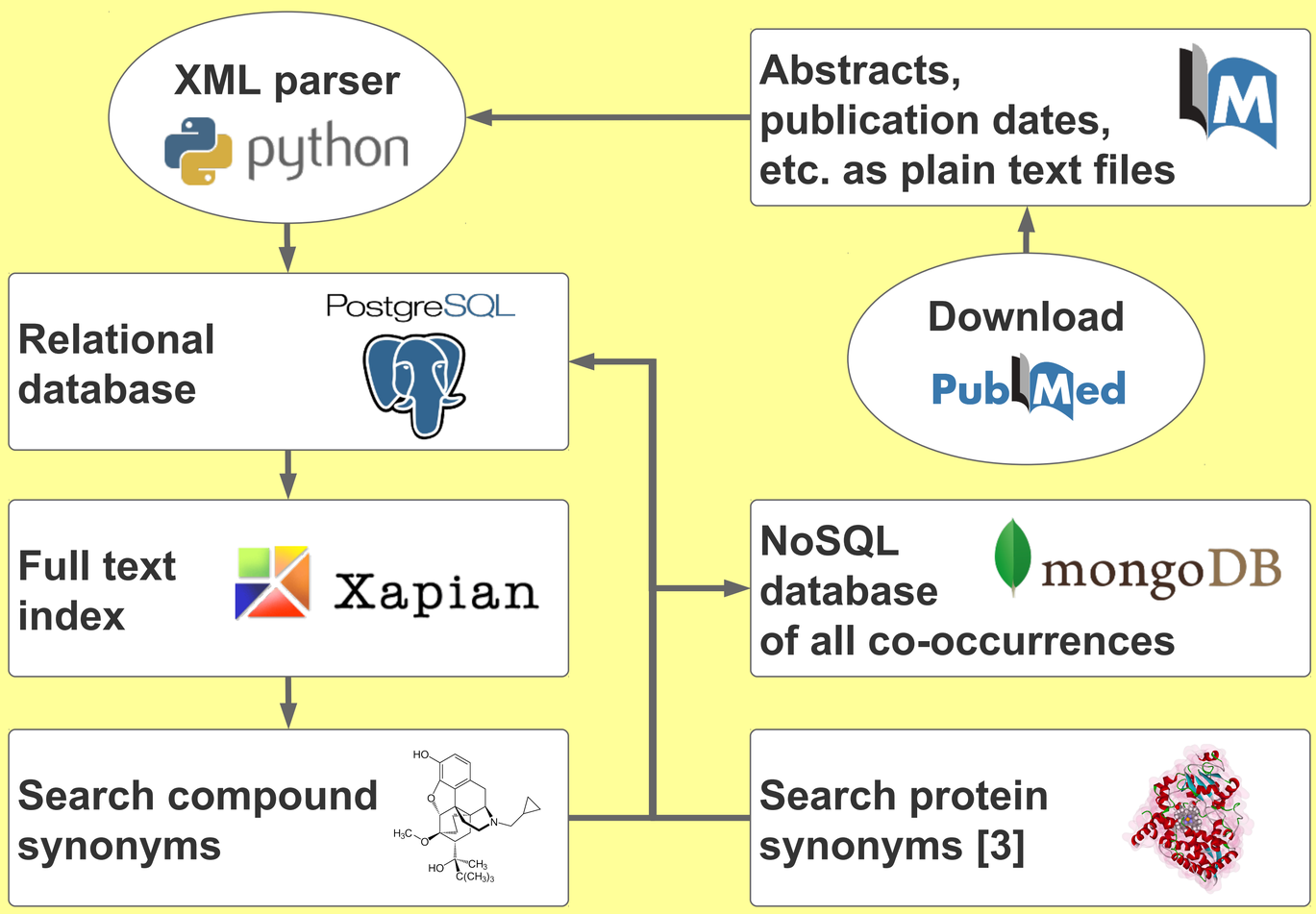
Döring K, Grüning BA, Günther S
kersten.doering@pharmazie.uni-freiburg.de
Department of Pharmaceutical Bioinformatics, Institute for Pharmaceutical Sciences,
University of Freiburg, Germany

Introduction

PubMed is a database containing around 21.5 M biomedical publication titles with around 12.5 M abstracts. Searching this continuously growing amount of literature for protein-compound interactions can be an elaborative task. We developed the web services *Compounds in Literature* (CIL) [1] and *Protein-Literature Investigation for Interacting Compounds* (prolific) [2] that search for co-occurrences of biomolecules in either a compound- or protein-centric view.

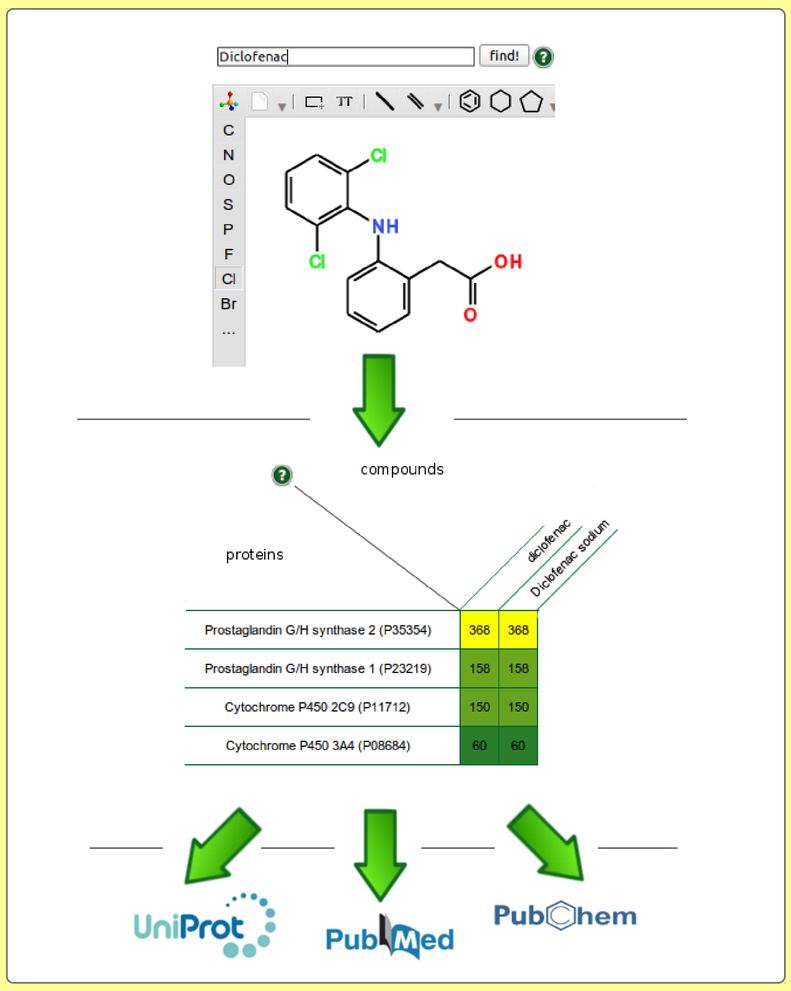
"... 21.5 M biomedical publication titles with 12.5 M abstracts."

Text Mining Workflow



The general workflow starts with parsing XML files from PubMed and loading them into a PostgreSQL relational database as well as building a full text index with Xapian. After building an "in-house" database of PubMed, it 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, we count around 1 Bn entries for 12.5 M abstracts.

Web Service Interface



In case of CIL it is possible to draw a structure or enter a name for a small molecule. Alternatively, prolific can be queried for a protein name or sequence. The result is a *heat-map* of co-occurring proteins and compounds for the query synonym as well as all similar hits.

Compounds in Literature

www.pharmaceutical-bioinformatics.de/cil

Proteins in Literature

www.pharmaceutical-bioinformatics.de/prolific

Future Prospects

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

This short abstract excerpt shows a phrase with the chemical compound *buprenorphine* (blue) and the protein *cytochrome P450 3A4* (green) enclosing the relationship word *metabolized.*
We recently started analysing such sentences to classify types of compound-protein interactions by applying machine learning methods adapted to the field of computational linguistics.

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.

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