PubMed2Go :: Dive into Large-Scale Text Mining

Döring K, Grüning BA, Flemming S, Günther S

kersten.doering@pharmazie.uni-freiburg.de

Department of Pharmaceutical Bioinformatics, Institute for Pharmaceutical Sciences,
University of Freiburg, Germany

Introduction

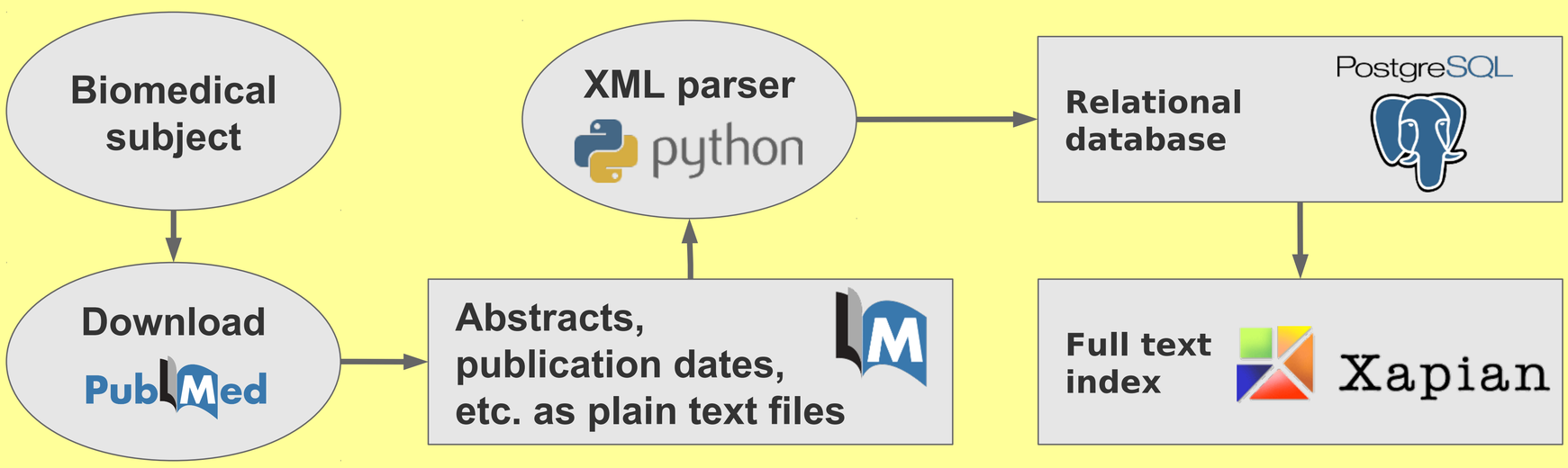
Our approach offers a one-click solution for generating a relational database in PostgreSQL [1] and a fulltext index by using Xapian [2] out of PubMed XML files [3]. Afterwards, large-scale text mining can be performed offline on the "in-house" database. The general workflow as shown in the next section was successfully applied within the webservices CIL [4] and prolific [5]. PubMed2Go was also used in the analysis of secondary metabolites of the bacteria *Streptomycetes* resulting in StreptomeDB [6].

"... development of high performance text mining."

"... be independent from a webservice."

"... easy to install on any local machine."

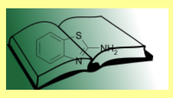
Workflow



This is our general workflow for 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 as shown in the published projects CIL, prolific, and StreptomeDB. This workflow can be used for other fields of research, too, as indicated by „space for ideas".

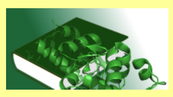
"... 21.5 M biomedical publication titles with 12.5 M abstracts."

Compounds in Literature



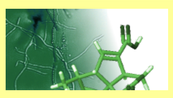
www.pharmaceutical-bioinformatics.de/cil

Proteins in Literature



www.pharmaceutical-bioinformatics.de/prolific

StreptomeDB



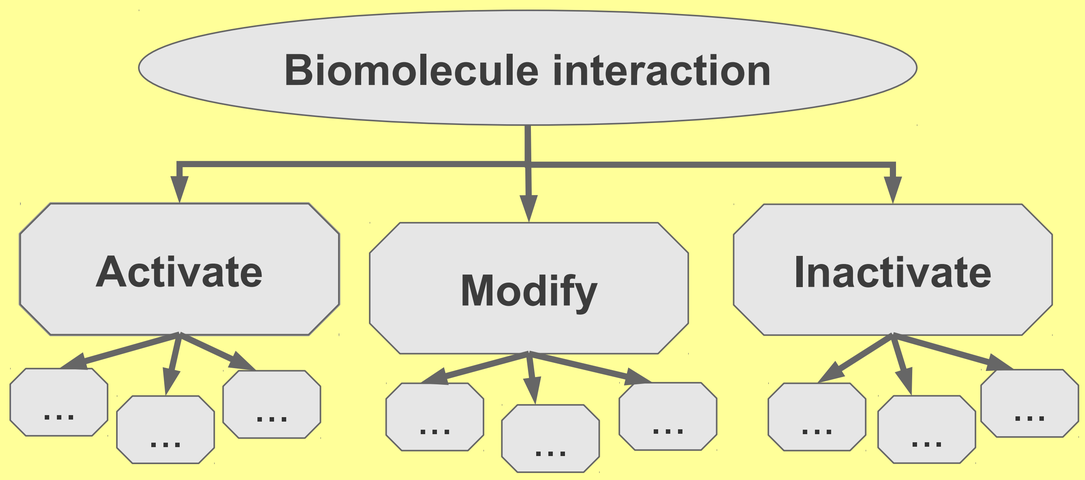
www.pharmaceutical-bioinformatics.de/streptomedb

"Space for Ideas"

?

An easily adaptable system for your own area of research!

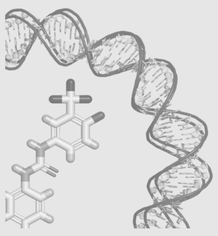
Future Prospects



The next milestone will be filtering and characterising interactions between compounds and proteins by applying an interaction ontology, e.g. BioInfer [7].

References

[1] http://www.postgresql.org
[2] http://xapian.org
[3] http://www.ncbi.nlm.nih.gov/pubmed
[4] Grüning, Senger *et al.,* 2011. Compounds In Literature (CIL): screening for compounds and relatives in PubMed. Bioinformatics 27:1341-2.
[5] Senger, Grüning *et al.,* 2012. Mining and Evaluation of Molecular Relationships in Literature. Bioinformatics 28:709-14.
[6] Lucas, Senger *et al.,* 2012. StreptomeDB: a resource for natural compounds isolated from Streptomyces species. Nucleic Acids Res. 41:D1130-6.
[7] Pyysalo *et al.,* 2010. BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics 8:50.



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/

CoRS is funded by the German National Research Foundation (DFG, Lis45).