Finding Functional Interactions of Proteins and Small Molecules in Literature

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

Introduction

PubMed is a database containing over 23 M biomedical manuscript titles with around 13.1 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.

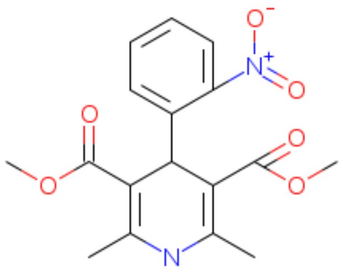
"... a high frequency suggests a relationship, but what about rare co-occurrences?."

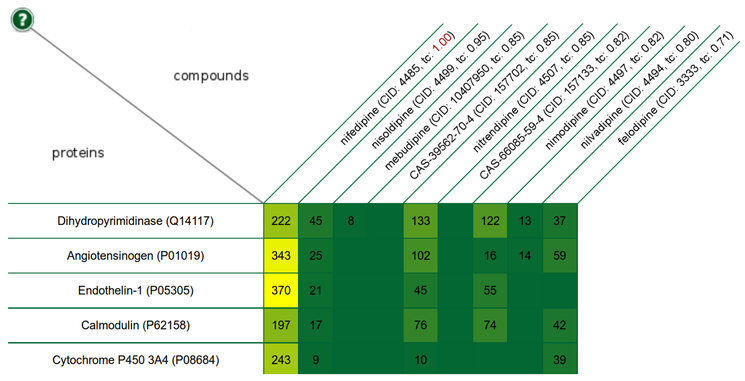
Web Service: CIL

Search for compound name, InChI or SMILES

Nifedipine find!

or draw a structure





www.pharmaceutical-bioinformatics.de/cil

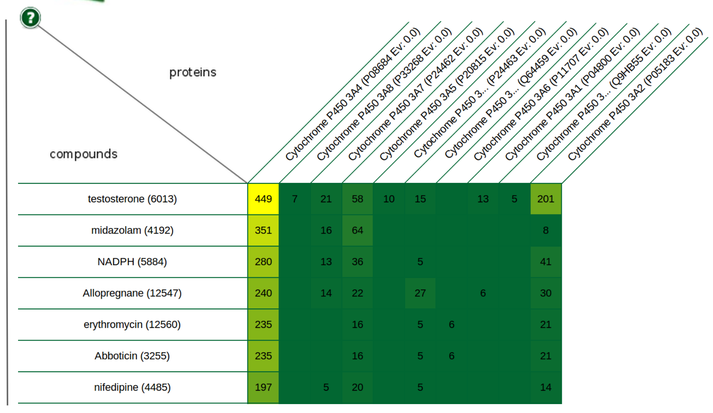
Web Service: prolific

Search for UniProt entry name, accession number, or exact protein name

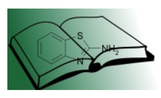
CYP3A4 find!

or specify a sequence

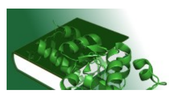
```
MALIPDLAMETWLLLAVSLVLLY
LYGTHSHGLFKKLGIPGPTPLPF
LGNILSYHKGFCMFDMECHKKYG
KVWGFYDGQQPVLAITDPDMIKT
VLVKECYSVFTNRRPFGPVGFMK
SAISIAEDEEWKRLRSLLSPTFT
SGKLKEMVPIIAQYGDVLVRNLR
REAETGKPVTLKDVFGA [...]
```



www.pharmaceutical-bioinformatics.de/prolific



UniProt PubMed PubChem



In case of CIL it is possible to draw a structure or enter a name of 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 identified similar hits.

Finding Functional Interactions in Sentences with Machine Learning

Annotation of compound-protein interactions (CPIs) in 1,289 sentences with an interaction verb enclosed (data set 1) and 1,923 sentences without it (data set 2)

DS 1: 996 sentences with interactions, 293 without

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

DS 2: 1,281 sentences with interactions, 642 without

[...] conversion of cholesterol to pregnenolone by cholesterol side-chain cleavage enzyme [...]

DS 1: 2,305 CPI pairs, 1,371 no-CPI pairs (62.7 % positives, 37.3 % negatives)
DS 2: 2,458 CPI pairs, 2,002 no-CPI pairs (55.1 % positives, 44.9 % negatives)

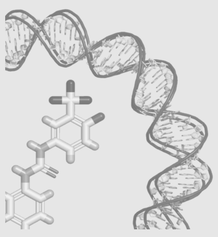
Train machine learning classifier to identify functional interactions:
jSRE – java Simple Relation Extraction (shallow linguistic kernel) [3]

Percentage of correct predictions (accuracy):
DS 1: 60.8 %
DS 2: 66.7 %

An arbitrary dataset of 40,000 abstracts was chosen from 2009. Sentences with and without a relationship verb enclosed by two biomolecules have been extracted with prolific, analysed and classified as *interaction* or *no interaction* instances. The definition of a functional interaction includes direct binding of two biomolecules as well as indirect relationships such as increase of protein expression. The jSRE machine learning model [3] for the prediction of CPIs achieved good results in comparison to protein-protein interaction extraction results [4]. The classifier of data set 2 will be included in CIL and prolific.

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] 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).
[4] Tikk *et al.,* 2011. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. PLoS Comput. Biol. 6:e1000837.



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