# Text Mining with CoRS – a Compound Research System

**Döring K[1,*], Grüning B[1], Flemming S[1], Senger C[1], Günther S[1]**

[1] *Department of Pharmaceutical Bioinformatics, Institute for Pharmaceutical Sciences, University of Freiburg, Germany*
**e-mail: kersten.doering@pharmazie.uni-freiburg.de*

## Recent Developments

Searching for interacting compounds for a given protein and vice versa can be an elaborative task. *Compounds in Literature (CIL)* [1] is a web service searching proteins co-occuring with a given compound identifier in all PubMed abstracts. If non-scientific compound names do not exist, similar compounds are identified by structural descriptors for chemical molecules [2].
While CIL has a compound-centric view, the recently published system *Protein-Literature Investigation for Interacting Compounds (prolific)* [3] searches co-occurrences of compounds with a given protein and similar sequences. An example search for the *dopamine receptor* with prolific is shown in figure 1. Results are displayed as heat map to facilitate the selection on pairs of biomolecules and related PubMed abstracts. A large number indicates a well-known relationship. It is possible to restrict results to compound-protein pairs in one sentence. Furthermore, a preliminary step for the characterisation of relationships is to filter the results for sentences with biomolecules enclosing curated 'relationship' verbs or GO terms [4].

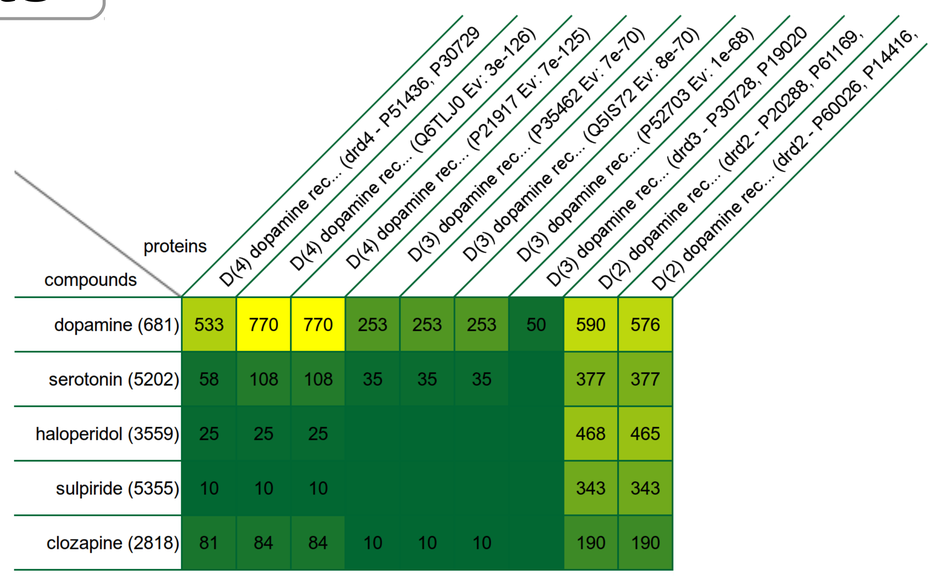


**Fig. 1: prolific heat map: Search for co-occurring compounds with the *dopamine receptor* as well as for similar proteins.**
**The boxes in different colours show the amount of abstracts containing the compound and the protein name. The compounds are ranked by row sum. The database identifiers for PubChem and UniProt are given in parentheses.**

## Future Prospects

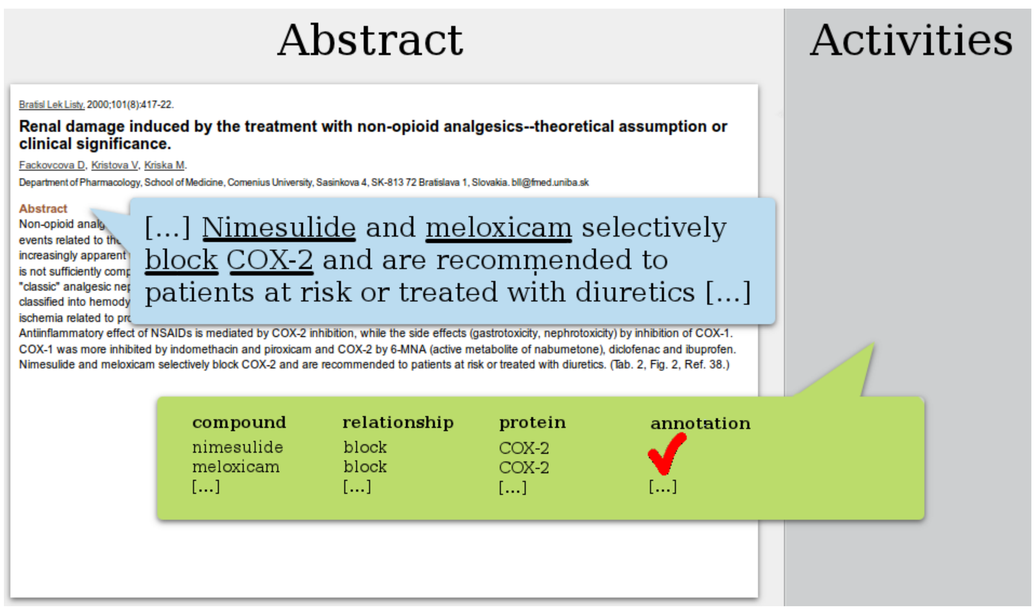


**Fig. 2: This is a first draft of the CoRS interface. It will work as a browser plugin and provide an activity-sidebar. Any biomolecule in the text will be highlighted.**
**In the blue box, two different compounds, the prote*in cyclooxygenase-2 (COX*-2), and the 'relationship' verb *block* are identified. To increase tagging quality and find new relationships, CoRS will support a curation interface for the user as illustrated in the green box.**
**Furthermore, possible features will be structure depiction by 'mouse over' or directly starting searches in CIL and prolific.**

The CoRS web project will support the screening of scientific articles to extract information about interactions of biomolecules like compounds and proteins. Its interface will be implemented as a browser plugin to assist the researcher in studying compound-protein relationships in any kind of browser text by highlighting biomolecules. Furthermore, the user can annotate newly identified interactions while studying a number of abstracts. The tagging of compounds and proteins will be supported by well-known web services and programmes such as OSCAR [5] or Whatizit [6]. A first draft of the web interface is shown in figure 2.
The back end of the CoRS web project will be trained to identify phrases including relationship-verbs, -nouns, -bigrams, or -phrases connected to all found biomolecules. Different machine learning algorithms will be investigated for their usability in this natural language processing tasks such as Bayesian classifiers and random forests [7].
The classified and restructured interaction types will be used to extract and visualise compound-protein relationships in an automated way.

## Conclusion

The CoRS search engine will reveal compound-protein relationships out of millions of abstracts by classifiying their type of interaction like induction, activation, inhibition, etc. This is possible by applying machine learning approaches adapted to the field of computer linguistics on sentences with tagged compounds, proteins, and interaction-phrases.
The system will support scientists working in the field of drug discovery, e.g. with suggestions for new interactions based on similarity screenings as shown in figure 3.

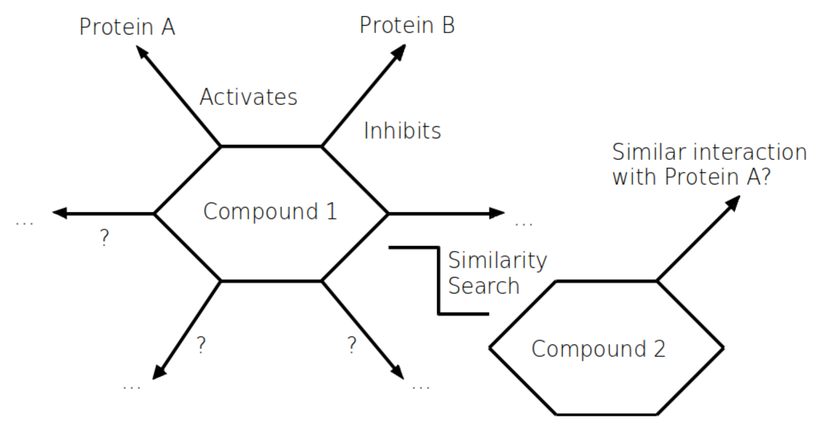


**Fig. 3: Find new druglike compounds by classifying compound-protein relationships in literature and draw a conclusion for similar compounds.**

**References**
[1] http://www.pharmaceutical-bioinformatics.com/cil
[2] Grüning *et al.,* 2011. Compounds In Literature (CIL): screening for compounds and relatives in PubMed. Bioinformatics 27:1341-2.
[3] http://www.pharmaceutical-bioinformatics.com/prolific
[4] Senger *et al.,* 2012. Mining and Evaluation of Molecular Relationships in Literature. Bioinformatics 28:709-14
[5] Corbett *et al.,* 2006.High-Throughput Identification of Chemistry in Life Science Texts. Computational Life Sciences II:107–118
[6] Rebholz-Schumann *et al.,* 2008. Text processing through Web services: calling Whatizit. Bioinformatics 24:296-298
[7] Weiss *et al.,* 2010. Fundamentals of Predictive Text Mining. Springer, Texts in Computer Science 41

Pharmaceutical Bioinformatics
University of Freiburg

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/

DFG Deutsche Forschungsgemeinschaft

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

UNI FREIBURG