Compounds In Literature (CIL): A web server for fast and efficient assessment of the cellular effect of compounds

Grüning B, Senger C, Günther S*

Department of Pharmaceutical Bioinformatics, Institute for Pharmaceutical Sciences, University of Freiburg, Germany;
**e-mail: stefan.guenther@pharmazie.uni-freiburg.de*

Introduction

Effects of novel compounds which are newly synthesized or isolated are difficult to assess. Gathering information about the compounds from literature and databases is an essential step preceding experimental tests. Similar molecules previously characterized may provide a clue of potential molecular functions, characteristics, and protein interactions. However, the desired information resides in huge archives like PubMed (>16×10^6 articles) [1] and PubChem (>24×10^6 compounds) [2]. Hence, data inspection is a time-consuming challenge and the risk to miss relevant information is high.

Aims

Based on literature screening with similar molecules we provide a one-step solution for a comprehensive report of potential targets and molecular functions of novel compounds .

Methods

All compounds and associated synonyms available at PubChem serve as a resource for structural similarity searches. Compounds can be uploaded to CIL in various file formats, edited via molecule editor or searched by compound name. Molecules are translated to structural fingerprints (OpenBabel). Similar compounds are identified by calculation of Tanimoto coefficients. Synonyms of those compounds are searched in all abstracts as well as MeSH terms of the PubMed database. A local version of PubMed supplied with full text indices is stored on a PostgreSQL database server. Analogously, 30,000 distinct genes/proteins (~130,000 synonyms) of the Human Genome Ontology (HUGO) were identified in PubMed abstracts. A text-mining software [3] was applied to reduce false positive hits resulting from ambiguous protein names. Thus, context information is considered to classify ambiguous words correctly as gene names. An overall evaluation represents the number of abstracts that mention certain compound-protein relations as a basis for data visualization.

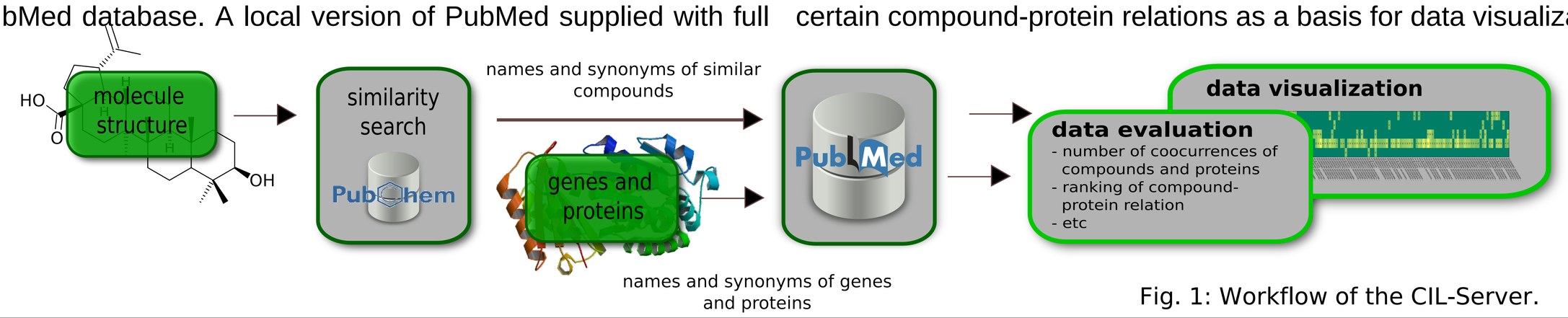

Fig. 1: Workflow of the CIL-Server.

Results

CIL identifies the most similar structures found in the database to a given query structure. Potential target proteins clustered by functional classes/processes (Gene Ontology terms) are provided for each compound. Frequencies of co-occurrences in literature are shown in the resulting plot (s. figure 2). Any compound-target relation detected with CIL is linked to the corresponding PubMed articles. Found proteins and compounds are highlighted in the associated abstracts. Filter and analysis options for detailed reports of compounds, proteins, and functions are available in the CIL web interface. CIL will soon be available at **http://cil.pharmaceutical-bioinformatics.de**.

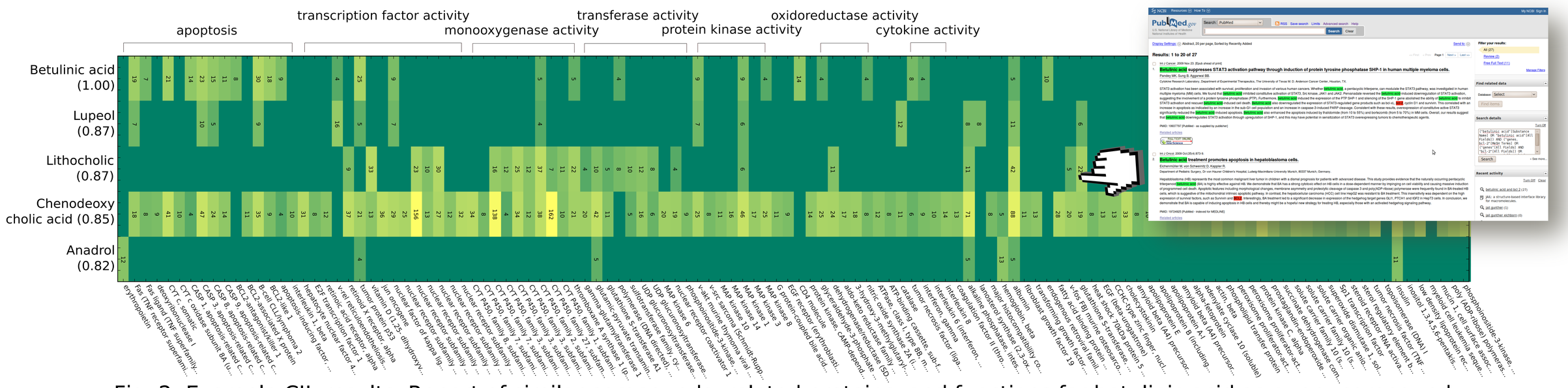

Fig. 2: Example CIL results. Report of similar compounds, related proteins, and functions for betulinic acid as query compound.

Conclusion

CIL is an efficient instrument to obtain a comprehensive survey of target relations of molecules described in literature. Results of similarity searches suggest potential cellular effects of molecules that are uncharacterized in literature, yet.

References

[1] Hunter L, Cohen KB. Biomedical language processing: What's beyond PubMed? Mol Cell, 2006. 21:589-94.
[2] Wang Y *et al.* PubChem: A public information system for analyzing bioactivities of small molecules. Nucleic Acids Res, 2009. 37:W623-33.
[3] Hur J *et al.* SciMiner: Web-based literature mining tool for target identification and functional enrichment analysis. Bioinformatics, 2009. 25:838-840.

Albert-Ludwigs-Universität Freiburg

Black Forest Grid

Bundesministerium für Bildung und Forschung

UNI FREIBURG