BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 9 2011, pages 1341–1342
doi:10.1093/bioinformatics/btr130

Data and text mining

Advance Access publication March 16, 2011

Compounds In Literature (CIL): screening for compounds and relatives in PubMed

Björn A. Grüning†, Christian Senger†, Anika Erxleben, Stephan Flemming and Stefan Günther*

Pharmaceutical Bioinformatics, Institute of Pharmaceutical Sciences, Albert-Ludwigs University, D-79104 Freiburg, Germany

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Searching for certain compounds in literature can be an elaborate task, with many compounds having several different synonyms. Often, only the structure is known but not its name. Furthermore, rarely investigated compounds may not be described in the available literature at all. In such cases, preceding searches for described similar compounds facilitate literature mining. Highlighted names of proteins in selected texts may further accelerate the time-consuming process of literary research. Compounds In Literature (CIL) provides a web interface to automatically find names, structures, and similar structures in over 28 million compounds of PubChem and more than 18 million citations provided by the PubMed service. CIL's pre-calculated database contains more than 56 million parent compound–abstract relations. Found compounds, relatives and abstracts are related to proteins in a concise 'heat map'-like overview. Compounds and proteins are highlighted in their respective abstracts, and are provided with links to PubChem and UniProt.

Availability: An easy-to-use web interface with detailed descriptions, help and statistics is available from http://cil.pharmaceutical-bioinformatics.de.

Contact: stefan.guenther@pharmazie.uni-freiburg.de

Received on November 29, 2010; revised on February 11, 2011; accepted on March 7, 2011

1 INTRODUCTION

The search for certain compounds in PubMed (http://www.ncbi.nlm.nih.gov/pubmed) as the largest repository of biomedical citations is an elaborate task for several reasons: (i) the vast and rapidly growing number of citations leads to time-consuming searches, particularly if researchers do not want to solely rely on MeSH (http://www.nlm.nih.gov/mesh) indexing and author-given keywords. Furthermore, information on compounds may be hidden in the context because the authors did not mention them exclusively as key words. (ii) Compounds can have a variety of different synonyms (more than 4 in 6.5% of all PubChem compounds), and each synonym should be searched to achieve a comprehensive result. (iii) If researchers do not know the exact name, a structure search has to be conducted first. (iv) A compound is newly-discovered or developed and therefore not or only rarely described in literature. Thus, preceding similarity searches can find compounds giving hints on potential biological actions. (v) To assess the biological context of the compound (i.e. interacting proteins), large parts of the texts have to be read. (vi) Considering such points, the problem of search duration arises but the result is desired instantly.

Several published approaches exist for text mining and information retrieval for compounds (Hettne *et al.*, 2009), proteins (Hur *et al.*, 2009; Rebholz-Schuhmann *et al.*, 2008) and combinations (Kuhn *et al.*, 2010; Rebholz-Schuhmann *et al.*, 2007; Zhu *et al.*, 2010). The publicly available compound resource PubChem (http://pubchem.ncbi.nlm.nih.gov) contains options for 2D similarity searches. Some of the existing tools cannot be seamlessly integrated in the desired workflow, while others lack a concise and comprehensive result overview. Thus, Compounds In Literature (CIL) was developed, a webserver that combines the related tasks of compound and literature screening. An extended computer infrastructure, specialized algorithms and data structures were used to provide a user-friendly tool for compound-related research. CIL-results provide a 'heat map'-like overview, comprising compounds, similar compounds, proteins and citations with highlighted found entities. Using this kind of structured and comprehensible information, CIL allows for analyses of obvious and hidden potential biological functions of compounds.

2 MATERIALS AND METHODS

All PubMed citations and abstracts beginning with 1975 up to 2010 were downloaded and are available in CIL (more than 18 million citations, about 10 million abstracts). Additionally, titles, abstracts, keywords, substance terms and MeSH terms were stored in a full-text index. The 28 million PubChem compounds with 55 million synonyms were downloaded and are locally available in database structures allowing for fast SMILES and InChI key searches and 2D similarity comparisons. For quick and efficient compound searches, CIL has a pre-built database using the first five synonyms of each compound for searches in every citation and abstract. Searches were conducted using the pre-processing steps described by Hettne *et al.* (2009). Additionally, we applied a self-generated stop word list for compound synonyms ('Background' page → Drug specific stop words). Furthermore, CIL allows for searches using the index and stop word lists with individually composed compound synonym sets. Proteins in all PubMed articles were retrieved via the text processing system Whatizit, which is able to find protein synonyms contained in UniProt (Rebholz-Schuhmann *et al.*, 2008). An additional protein stop word list was generated to enhance specificity ('Background' page → Protein-specific stop words). Thus, co-occurrences

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

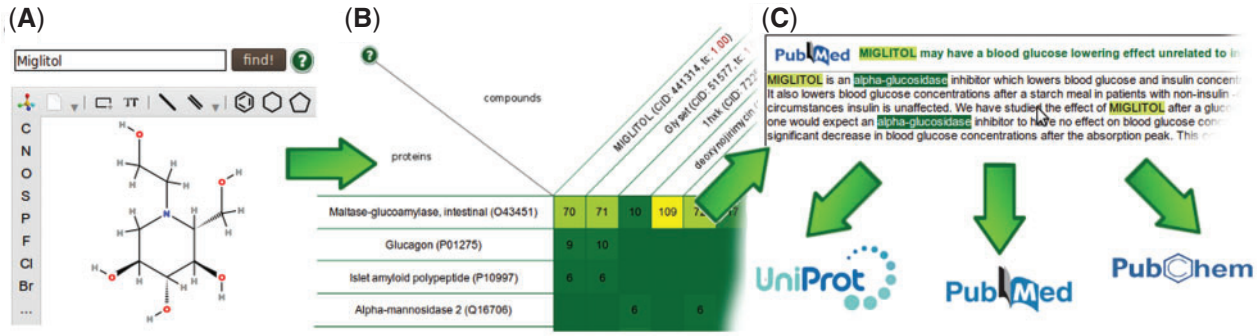

Fig. 1. CIL workflow: (**A**) search of compounds by name, SMILES, InChI or structure; (**B**) comprehensive 'heat map' overview with similar compounds, number of abstracts and occurrences of proteins; (**C**) citations with highlighted found compounds and proteins. 'Quick helps' are ubiquitously provided. Extensive links lead to further information in PubMed, PubChem and UniProt.

of compounds and proteins could be retrieved from the CIL database. Updates take place every 6 months.

3 RESULTS

The pre-calculated database contains more than 56 million parent compound–abstract and more than 133 million protein–abstract relationships. Detailed statistics are provided on the 'Statistics' page. CIL achieved a higher *F*-score (60%, precision: 52%, recall: 72%) compared to the results based on PubChem synonyms of Hettne *et al*. (2009, *F*-score: 47%, precision: 73%, recall: 35%). The CIL workflow is shown in Figure 1. CIL can be queried with compound synonyms, SMILES, InChI keys or drawn structures. Results can be obtained (i) with *one click* using the carefully pre-selected parameters and pre-calculated database, or (ii) by using the free text search and adjusting proteins' source organisms, noise filter, tanimoto coefficient for similarity searches and including or excluding certain compound synonyms. Results can be further refined by user-defined parameter settings. Search duration using the pre-calculated database primarily depends on the number of found similar compounds and limitations of the browser to quickly display large tables. Using the free text search, data retrieval from the text-index causes additional delay, thus CIL provides the possibility for the user to estimate search times. The resulting coloured table (Fig. 1b) displays the queried compound and relatives in columns. The header contains synonyms, structures and respective tanimoto coefficients as well as links to the PubChem database. Row headers contain proteins and associated synonyms and direct links to the UniProt database (http://www.uniprot.org). Table cells denote the number of found citations containing corresponding compounds and proteins. Furthermore, numbers are indicated by gradient colour shades. Cells are linked to the corresponding citations with highlighted compound and protein synonyms, which are also linked to their respective PubChem and UniProt entries (Fig. 1c). Further links to the PubMed literature service are also specified. Each component of the workflow is extensively provided with 'quick helps', providing background and additional information for each step. The CIL web site contains comprehensive help and FAQs—with detailed explanations of each step of the search, including how to use the different options. A background information section provides details on data sources, the building process and supplies e.g. utilized stop word lists.

4 CONCLUSION

CIL combines several tasks in one workflow—offering a one-click solution for otherwise elaborate, separated and time-consuming tasks. Initially searching with compound names, SMILES, InChI keys or structures, query compounds as well as similar compounds are identified in biomedical literature and are related to proteins named in the same context. Results are displayed in a concise table. CIL combines the three large datasets of PubMed, PubChem and UniProt. Given the vast number of compounds, proteins and biomedical literature, searching for information on biological functions of the query compound can be conducted in a shorter time.

Funding: This work was supported by the Excellence-Initiative (Deutsche Forschungsgemeinschaft) in the excellence cluster 'System Analysis of biogenic drugs by Pharmaceutical Bioinformatics'.

Conflict of Interest: none declared.

REFERENCES

Hettne,K.M. *et al*. (2009) A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, **25**, 2983–2991.

Hur,J. *et al*. (2009) SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics*, **25**, 838–840.

Kuhn,M. *et al*. (2010) STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.*, **38**, D552–D556.

Rebholz-Schuhmann,D. *et al*. (2007) EBIMed–text crunching to gather facts for proteins from medline. *Bioinformatics*, **23**, e237–e244.

Rebholz-Schuhmann,D. *et al*. (2008) Text processing through web services: calling whatizit. *Bioinformatics*, **24**, 296–298.

Zhu,Q. *et al*. (2010) WENDI: a tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications. *J. Cheminformat.*, **2**, 6.