Chemistry Central Journal

ChemistryCentral

Oral presentation

Using chemical structure in open-source chemical text mining

PT Corbett* and P Murray-Rust

Address: Unilever Centre for Molecule Sciences Informatics, Lensfield Road, Cambridge, CB2 1EW, UK
* Corresponding author

from 3rd German Conference on Chemoinformatics
Goslar, Germany. 11-13 November 2007

Published: 26 March 2008

Chemistry Central Journal 2008, **2**(Suppl 1):S10 doi:10.1186/1752-153X-2-S1-S10

This abstract is available from: http://www.journal.chemistrycentral.com/content/2/S1/S10

A great wealth of chemical information is to be found in the literature. For example, PubMed contains of the order of 15 million abstracts, a significant proportion of which contain information about chemicals, their biological activity and reactivity. In order to analyse this information, it must first be extracted from the literature – a task that can be performed by computers as well as by humans. OSCAR3 is an open source chemistry text mining tool, which can find chemical names, ontology terms and experimental data in chemistry papers, biomedical abstracts and other texts [1,2]. Recent advances in recognition techniques enable recognition with precision and recall of 80% or better, adjustable to higher recall or higher precision, at a rate of about one abstract per second. A key feature of OSCAR3 is that it can produce molecular structures for the names it finds. This enables structure-based chemical informatics techniques such as substructure search and molecular similarity to be added to the repertoire of text mining methodologies, increasing the range of information that can be extracted, and the range of analyses that can be performed.

Preliminary results show that it is possible to mine metabolic reactions from a corpus of cytochrome P450 abstracts. The names of substrates and reactions can be spotted using OSCAR3, and related to each other *via* pattern matching. In many cases, it is possible to infer the products of the reactions, even though they are not stated explicitly. For example, a paper can mention "the O-demethylation of codeine". From this, it is possible to use the chemical structure of codeine – by finding the O-methyl group and removing it - to infer that the product of the reaction is morphine, even though this is not explicitly stated in the paper. Another attractive use for chemical structure in text mining makes use of molecular similarity. Given the structure of a chemical, it is possible to find structurally-related compounds *via* fingerprint-based techniques, and to correlate the occurrence of those related compounds in a corpus with the occurrence of names of cytochrome P450s to make predictions about which P450s interact with the target molecule.

References

1. Corbett P, *et al.*: *LNCS* **4216:**107-118.
2. Batchelor C, *et al.*: *Proceedings of the ACL 2007 Demo and Poster Sessions* :45-48.