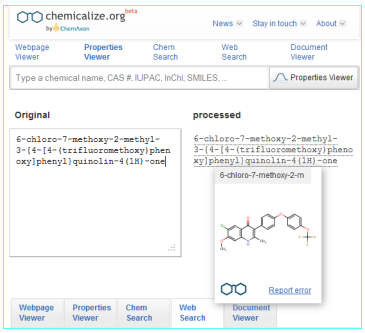
Chemicalize.org: A uniquely user-selected PubChem source of structures extracted from text

Christopher Southan, TW2Informatics, Göteborg, Sweden, Andras Stracz, Daniel Bonniot de Ruisselet and Ferenc Csizmadia, ChemAxon Kft, 1031 Budapest, Hungary. cdsouthan@hotmail.com astracz@chemaxon.com



The chemicalize.org open web application recognizes different types of chemical names in any text source and converts them into structures (the example of a new Wikipedia antimalarial lELQ-300 is shown on the right). It can thus both disinter and connect between millions of structures from their "tombs" of patents, papers, abstracts or web pages. The service has generated a searchable database of ~300,000 unique structures from user results since 2009. The Webpage Viewer and Document Viewer save visited URLs along with the extracted structures. Because the philosophy of chemicalize.org is to make chemistry more accessible this archive is deposited into PubChem and updated. It is unique in being the only source derived from user-selected content (i.e. documents and URLs are actively chosen, typically via the protein target and/or disease indication). The accumulated structures are thus "collectively crowd-sourced" and link back to chemicalize.org data pages. These include predicted properties such as pKa, logP/D, all of which can be downloaded along with SDF, SMILES, IUPAC names, and InChI. The figures below give an introduction to the functionality and exploitation synergies with PubChem.

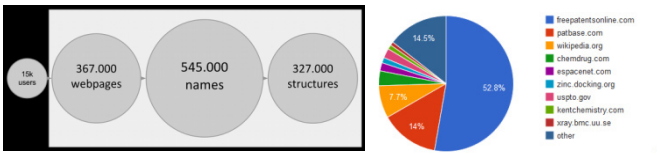


Figure 1. The left hand picture shows the chemicalize.org result statistics on the ChemAxon side. On the right you can see a breakdown of these user-selected sources by first-access to structures. Notably this is ~ 70% patent full-text sites but also 8% from Wikipedia pages.

The overall statistics associated with the PubChem content are as follows. After filtration and submission processing, 300,987 chemicalize.org structure records were merged into 297,083 CIDs, with a ranking of 27th by source size. Of these CIDs 80% are independently confirmed by at least one other submitting source in PubChem suggesting a high quality of extraction by chemicalize.org. The remaining 62,032 are unique CIDs. While 8,964 of these are mixtures, the components of which may be pre-existing CIDs, it is clear that chemicalize.org users are finding novel structures in their selected sources.

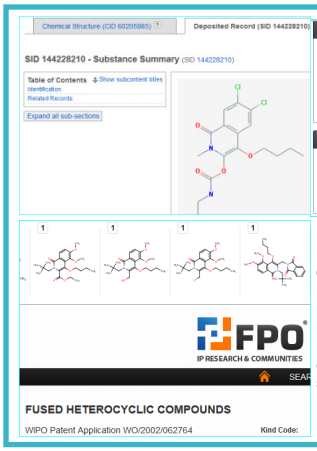


Figure 2 shows the connectivity utility via a unique CID from PubChem (top left) and part of the linked chemicalize.org record (top right). In turn, this links directly to the open patent URL selected by the original user (left). Using just one click any user can re-run the complete document extraction in chemicalize.org and download over 1,900 structures.

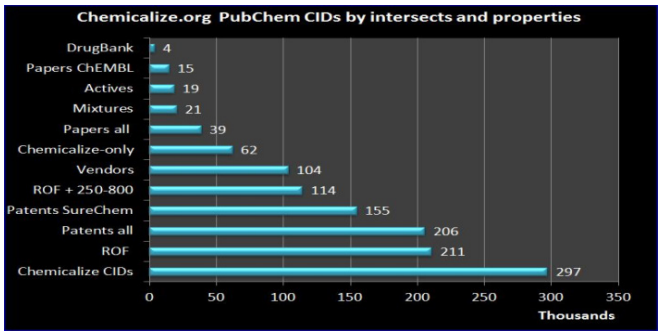


Figure 3 is a comparative analysis of the chemicalize.org compounds in PubChem. The Lipinski rule-of-five content at 71% is high (c.f. 59% for ChEMBL as a comparison to active compounds from papers). This indicates the user-focus on drug-like space. Using a more lead-like filter by restricting to a Mw range of 250-800, drops this to 38% (c.f. 58% for ChEMBL). Unsurprisingly, since patents are the major user-selected source by document origin (as shown in fig.1), the structure overlap is 69% for patent sources (SureChem, IBM,SCRIPDB and Thomson Pharma) with 52% for SureChem being the largest patent-only source. Note that coverage of DrugBank is high (60%) but for ChEMBL is low(3%). The likely reason for the latter is that, unlike patents, there are few medicinal chemistry journals with open text URLs that users can run for extractions. Using in the PubChem query interface the chemicalize.org CIDs can in intersected with any sources or property filters. Note that even if the CID was pre-existing the URL and or document links (as shown in fig. 2) may link, via chemicalize.org, to information or data that is not in the other PubChem source links

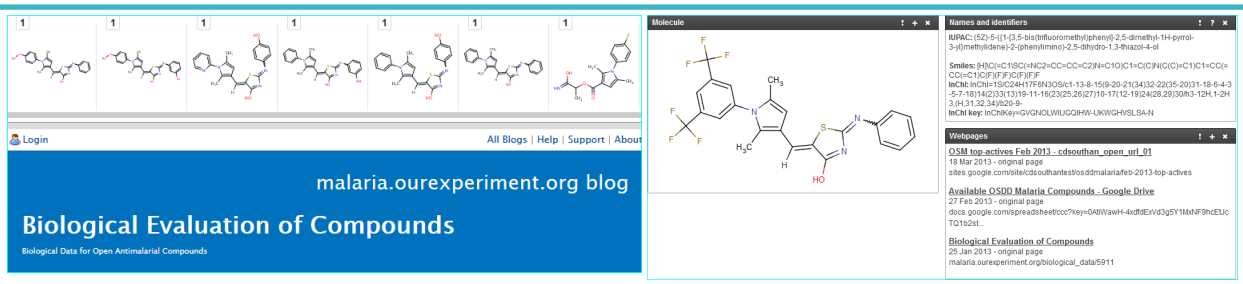


Figure 4 shows the utility both as an open sharing option and an indirect submission route to PubChem. Compounds tested in an Open Source Drug Discovery project (OSDD), for antimalarial activity in this case, have been initially extracted from the team's open web pages that are linked to assay results. Of the 15 structures, 6 in PubChem, including a low-nM lead from the project (CID 57515644 = OSM-S-38) originate uniquely from chemicalize.org, thus becoming globally searchable within the PubChem CID space many months in advance of publication. Note also that two additional links are from personal websites as another global surfacing option. The InChIKey generated by chemicalize.org is effective highly-specific for Google searching and connecting to databases and websites of all types. Quote "In NIH's view, all data should be considered for data sharing."

References and additional information

http://www.chemicalize.org/
http://www.chemaxon.com/blog/chemicalize-orgs-crowdsourced-database
http://cdsouthan.blogspot.se/2013/01/chemicalizeorg-from-chemaxon-in-pubchem.html
http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=57515644 the antimalarial lead from fig.4
Southan C. & Stracz A., 2013, **Extracting and Connecting Chemical Structures from text Sources using Chemicalize.org**. *J. Chem. Inform.*, in press (open access)
Southan C. 2013 **InChI in the wild: an assessment of InChIKey searching in Google**. *J Chem. Inform*. Feb;5(1):10. doi: 10.1186/1758-2946-5-10 (open access)
Swain M. 2012, **chemicalize.org**, *J. Chem. Inf. Model.* 52, No. 2.

chemicalize.org beta

15 Years CHEMICAL GROWTH

ChemAxon