Introduction

As new chemical entities are often first published in patents, and some new compounds may not even be featured elsewhere, patents have become an important source of information for researchers.

Methods

When stereochemistry was removed, the results somewhat improved, with SureChEMBL returning 64.9% of the SciFinder molecules. This number was 67.1% for the "Biologically annotated" subset.

The first comparison we performed was of a binary nature – we looked at whether a compound was found at all in the SureChEMBL and IBM SIIP databases.

From the 438 compounds, 67.1% was found in at least one of the two databases, 52.7% were found in both, 29% was found only in IBM SIIP and 11.4% was found only in SureChEMBL.

Results

Discussion

The about 60% efficiency of SureChEMBL would most likely seem low to the researcher who expects every single compound of interest to be extracted from each patent. This is the reason it is surprising that the coverage was not greatly increased for "Biologically annotated" molecules. But what compounds are of interest?

SureChEMBL returned nearly 5 times more compounds than SciFinder.

Conclusion

On average, 50-68% of the “gold standard” manually curated patent chemistry database content can also be found in automatically generated databases. These latter databases are also freely available, for example, SureChEMBL will soon be available through the Open PHACTS api (http://dev.openphacts.org). IBM SIIP is also freely available, however it is a static database covering patents until 2010, whereas SureChEMBL is updated daily.

References

1. Senger et al., J. Cheminf, 2015, 7:49
3. http://www.uspto.gov/* luca.bartek1@gmail.com

Authors: Luca Bartek*, Stefan Senger, George Papadatos, Anna Gaulton

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115191, resources of which are composed of financial contribution from the European Union’s Seventh Framework Programme (FP7/2007-2013) and EFPIA companies’ kind contribution.

Comparison of automated and manual patent chemistry extraction methods