ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics (Extended Abstract[1])

Jiangming Sun[1], Nina Jeliazkova[2], Vladimir Chupakhin[3], Jose-Felipe Golib-Dzib[4], Lars Carlsson[1], Jörg Wegner[3], Hugo Ceulemans[3], Ivan Georgiev[2], Vedrin Jeliazkov[2], Nikolay Kochev[2,5], Thomas J. Ashby[6], and Hongming Chen[1]

[1] Discovery Sciences, AstraZeneca R&D Gothenburg, Sweden
[2] Ideaconsult Ltd., 4. Angel Kanchev Str., 1000 Sofia, Bulgaria
[3] Computational Biology, Discovery Sciences, Janssen Pharmaceutica NV, Beerse, Belgium
[4] Computational Biology, Discovery Sciences, Janssen Cilag SA, Toledo, Spain
[5] Department of Analytical Chemistry and Computer Chemistry, U. Plovdiv, Plovdiv, Bulgaria
[6] Imec vzw, Kappeldreef 75, 3001 Leuven, Belgium.

Abstract. Chemogenomics data generally refers to the activity data of chemical compounds on an array of protein targets and represents an important source of information for building *in silico* target prediction models. The increasing volume of chemogenomics data offers exciting opportunities to build models based on Big Data. Preparing a high quality data set is a vital step in realizing this goal and this work aims to compile such a comprehensive chemogenomics dataset. This dataset comprises over 70 million SAR data points from publicly available databases (PubChem and ChEMBL) including structure, target information and activity annotations. Our aspiration is to create a useful chemogenomics resource reflecting industry-scale data not only for building predictive models of in silico polypharmacology and off-target effects but also for the validation of cheminformatics approaches in general.

1 Background

In pharmacology, there has been a remarkable increase in the amount of available compound structure and activity relation (SAR) data, contributed mainly by the development of high throughput screening (HTS) technologies and combinatorial chemistry for compound synthesis. These SAR data points represent an important resource for chemogenomics modelling, a computational strategy in drug discovery that investigates an interaction of a large set of compounds (one or more libraries) against families of functionally related proteins.

Databases such as PubChem and ChEMBL are examples of large public domain repositories of this kind of information. The aforementioned publicly available data-bases have been widely used in numerous cheminformatics studies. However, the curated data are quite heterogeneous and lack a standard way for annotating biological endpoints, mode of action and target identifier. There is an urgent need to create an integrated data source with a standardized form for chemical structure, activity annotation and target identifier, covering as large a

chemical and target space as possible. In this work, by combining active and inactive compounds from both PubChem and ChEMBL, we created an integrated dataset for cheminformatics modeling purposes.

2 Dataset curation

Data cleaning and standardisation procedures were applied in preparing both chemical structures and bioactivity data. Bioassays were restricted to only those comprising a single target. 58,235 and 92,147 single targets containing concentration response (CR) type assays (confirmatory type in PubChem) remained in PubChem and ChEMBL, respectively. Inactive compounds in CR assays were kept as inactive entries. Compounds that were labelled as inactive in PubChem screening assays (assays run with a single concentration) were also kept as inactive records. Finally, targets which have < 20 active compounds were removed from the final dataset. Entrez ID , gene symbol and gene orthologue were collected as information for the target.

Address `https://solr.idea-consult.net/search/excape/` can be used to download the ExCAPE-DB, and it is also uploaded to the `Zenodo.org` repository and available from there as https://doi.org/10.5281/zenodo.173258.

In total there are 998,131 unique compounds and 70,850,163 SAR data points. These SAR data points cover 1667 targets. It constitutes a curated large scale chemogenomics set freely available in the public domain under the Creative Commons Attribution Share-Alike 4.0 license. The dataset is useful for building QSAR models for predicting activity against one or more specific targets for novel compounds and will also serve as a benchmark dataset for evaluating the performance of various machine-learning algorithms, especially multi-target learning algorithms. Inclusion of inactive compounds from PubChem better mimics chemogenomics datasets available in the pharmaceutical industry.

3 Conclusion

ExCAPE-DB is a large public chemogenomics dataset based on the PubChem and ChEMBL databases, and large scale standardisation (including tautomerization) of chemical structures using open source cheminformatics software was performed in data curation. Comprehensive compound related information such as target activity label, fingerprint based descriptors and InChIKey, and target related information such as Entrez IDs and official gene symbols were collected and are easily accessible in the publicly available database. The active labels were determined based on their dose–response data to make sure the data quality is as high as possible. This 'Big Data' set covers large number of targets reported in the literature and can be used for building holistic multi-target QSAR models for target prediction. To the best of our knowledge, this is first attempt to build such a large scale and searchable open access database for QSAR modelling.

Address https://doi.org/10.1186/s13321-017-0203-5 is the location of the original full length article.