Co-funded by the Horizon 2020 Framework Programme of the European Union

Grant agreement No. 671555

ExCAPE
Exascale Compound Activity Prediction Engines

Future and Emerging Technologies (FET)

Call: H2020-FETHPC-2014
Topic: FETHPC-1-2014
Type of action: RIA

Deliverable D3.1

Other: PublicBio
Public Chemogenomic Dataset Curation

Due date of deliverable: 29.04.2016
Actual submission date: 05.05.2016

Start date of Project: 1.9.2015 Duration: 36 months

Responsible Consortium Partner: AZ
Contributing Consortium Partners: JP,IDEA
Name of author(s) and contributors: Jiangming Sun, Hongming Chen (AZ), Jose Felipe Golib Dzib (JP), Nina Jeliazkova (IDEA)

Project co-funded by the European Union within the Horizon 2020 Framework Programme (2014-2020)		
Dissemination Level		
PU	Public	**X**
PP	Restricted to other programme participants (including the Commission Services	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Revision: V3.0
Internal reviewer(s): Vladimir Chupakhin

Document revision tracking

This page is used to follow the deliverable production from its first version until it has been reviewed by the assessment team. Please give details in the table below about successive releases.

Release number	Date	Reason of this release and/or validation	Dissemination of this release (task level, WP/ST level, Project Office Manager, Industrial Steering Committee, etc)
V1.0	15.04.2016	First draft for discussion	WP3 project partner
V2.0	02.05.2016	Second draft after gathering feedback from WP3 partners	WP3 project partner
V3.0	03.07.2016	Third draft after gathering feedback from EU reviewers	Public

Glossary

HTS	high throughput screening
MW	molecular weight
PSA	polar surface area
FSC	fraction of Sp3 carbon atom
NHR	nuclear hormone receptors

Link to Tasks

Task number	Work from task carried out	Deviations from task technical content
Task 3.1.1	Development of a standardization pipeline for chemical and biological data.	The task was carried out as expected
Task 3.1.2	Collection and curation of public biological dataset with known target annotation.	The task was carried out as expected

Table of contents

1 Executive summary

A large public chemogenomics dataset was created based on public available PubChem and ChEMBL database. Compound related information such as target activity label, fingerprint based descriptors and InChi keys, and target related information such as Entrez geneID, Genesymbol were collected in text file. This dataset will be used for the following large scale modelling study in ExCAPE.

The reason for late submission was due to complexity of data curation and consultation among WP3 partners for modification.

2 Introduction – Aim

The chemogenomics data generally refers the activity data of chemical compounds against an array of protein targets and represents an important source of information for building in silico model for target prediction. In silico protein target prediction is a well-established computational technique that offers an alternative avenue to infer target-ligand interactions by utilizing known bioactivity information [1]. These methods have played an important role in the field of efficacy prediction, the prediction of toxicity and target deconvolution in phenotypic screening.[2-4] Such approaches are designed to predict targets for orphan compounds early in the drug development phase, with the predictions forming the base of an experimental confirmation afterwards. Both structure-based and ligand-based methods exist for the prediction of protein targets for small molecule ligands.[5,6]

The goal of ExCAPE project is to develop machine learning algorithms and implementations thereof to enable the use of upcoming exascale supercomputers to solve very large scale complex compound activity problems in pharmaceutical research. Building in silico model based on chemogenomics dataset is one of the goals in this project. Recent years it has been seen that large databases evolve in the public domain to archive and organize rapidly growing amounts of chemical structures and associated activity data (Similar efforts were also carried out within large pharmaceutical companies for collecting their own proprietary data). Among others, prominent data repositories currently include PubChem[7] and ChEMBL[8] databases. One important deliverable for ExCAPE project is therefore to compile a large chemogenomics dataset based on these two major public information sources forming the benchmark dataset for validating the forthcoming algorithms developed in ExCAPE WP1 and WP2.

3 Workflow to compile public chemogenomics dataset

The latest ChEMBL20 database (https://www.ebi.ac.uk/chembl/) was used for data curation and PubChem data was downloaded from PubChem website (https://pubchem.ncbi.nlm.nih.gov/) at cut-off date of 2016-01-15. Both databases are quite heterogeneous data sources. Certain data cleaning and standardization procedures are needed for dataset preparation. The workflow for data retrieval and standardization was shown in Figure 1.

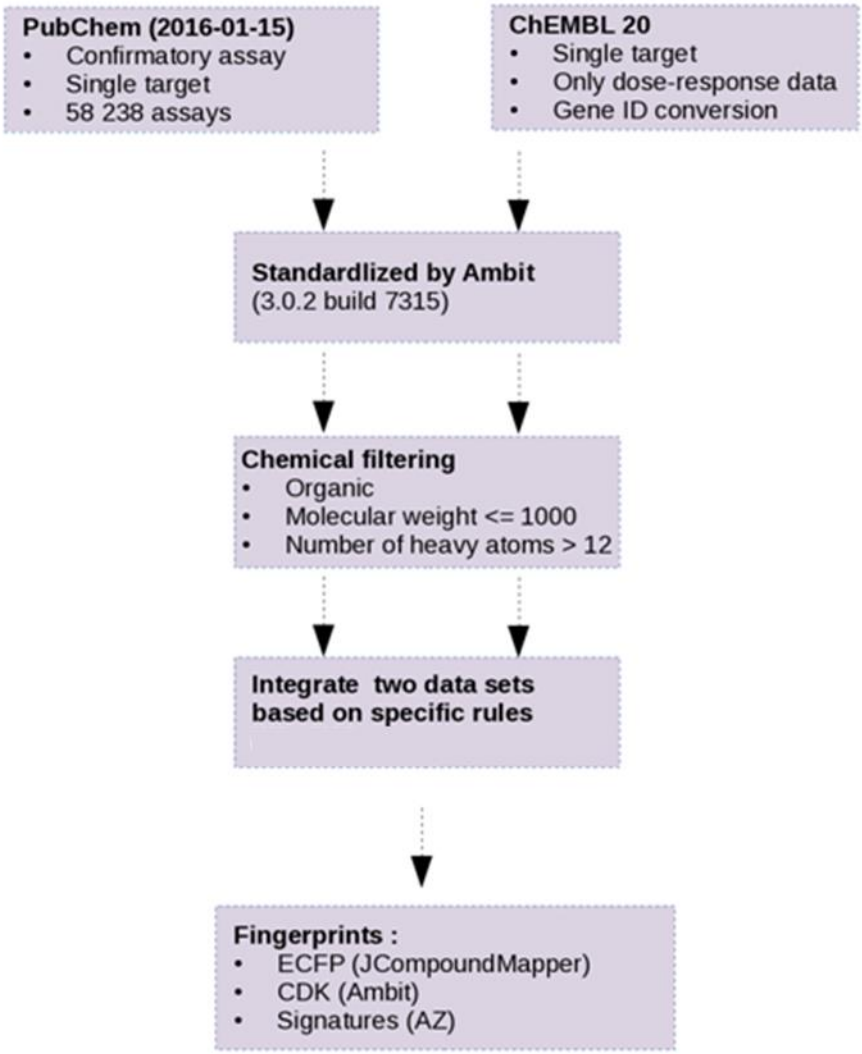

Figure 1. Workflow for data preparation

The processing protocol as shown in Figure 1 was used to extract and standardize bioactivity data. We restrain our target list to human target and only examine assays which comprise single target, i.e. blackbox (target unknown) or multi-target assays were excluded, and only targets containing more than 20 actives were considered. This is to avoid building models on targets having too few active compounds. For those filtered assays, actives whose dose-response value (In PubChem assays, only confirmatory type assays are considered) is less than

10uM (ie. potency data is higher than 10uM) were kept as active compounds. Inactives and actives whose activity value is higher than that 100uM were kept at inactives. Compounds which are labelled as inactives in PubChem single concentration assay (ie. screening assays) were also kept as inactive compounds. From each data source various attributes are being read and converted into controlled vocabularies. Those are target (Entrez Gene ID), activity value, mode of action, assay type and assay technology etc. The underlying data sources contain activity data with various result types, we tries to unify results so that results can be compared across tests (and data sources) irrespective of the original result type.

For active compounds, only dose-response results were kept for analysis. The selected compatible dose-response result types are:

- Concentration dependent activity (e.g. IC_{50}), currently defined as result in M or g/l with result type name matching
  - *C50
  - Potency
  - *GI50
  - TGI
  - MIC
  - MEC
  - KI
  - KD
  - A2

- Concentration dependent activity in log format, currently defined as results with result type name matching
  - p*C50
  - p*GI50
  - pKi or pKI
  - pKd or pKD
  - pTGI
  - pMIC
  - pMEC
  - pA2 or pa2

Activity values for actives are aggregated so that each compound in each target only has one result. For a compound-target pair the highest (max) potency is chosen as the aggregated final value. Once all the actives and inactives were collected, they were standardized by running Ambit program[9] to generate standard tautomer form and InChi key. The duplicated compounds for the same target were removed.

The compound set was further filtered based on physicochemical properties: Organic filter (no metal element appear in the structure); MW (molecular weight) < 1000; HEV (number of heavy atom) > 12. This is a much generous rule than the Lipinski rule-of-five[10], but we are trying to keep as much useful chemical information as possible while still remove some non-drug like compounds. In the last step, fingerprint descriptors were generated for all the compounds. So far JCM[11] and CDK[12] fingerprint descriptors were generated respectively.

4 Overview of the public chemogenomics dataset

Table 1 Public chemogenomics dataset

		ChEMBL	PubChem	Total
Actives	# SAR data points	538 703	315 043	853 746
	# Compounds	271 481	161 666	433 147
Inactives	# SAR data points	890 078	68 856 513	69 746 591
	# Compounds	319 645	3 120 408	3 440 054

All the data compilation and clearing was done and stored on IT4I Salomon server and the detailed processing protocols can be seen in Annex part. The final meta-data file size is around 20GB. The dataset composition can be seen in Table 1. In total there is around 3.67 million unique compounds and around 70 million SAR data points and the whole data file is around 9.4GB. These SAR data points cover 1239 human targets in total. This dataset represents a cleaned large scale chemogenomics set in public domain. The distribution of active compounds in the targets can be seen in figure 2 and 3. Overall most of targets have much fewer actives than inactives, which means that the chemogenomics dataset is highly imbalanced. This is largely due to the inactive compounds coming from the PubChem assays which contain a large number of HTS screening results.

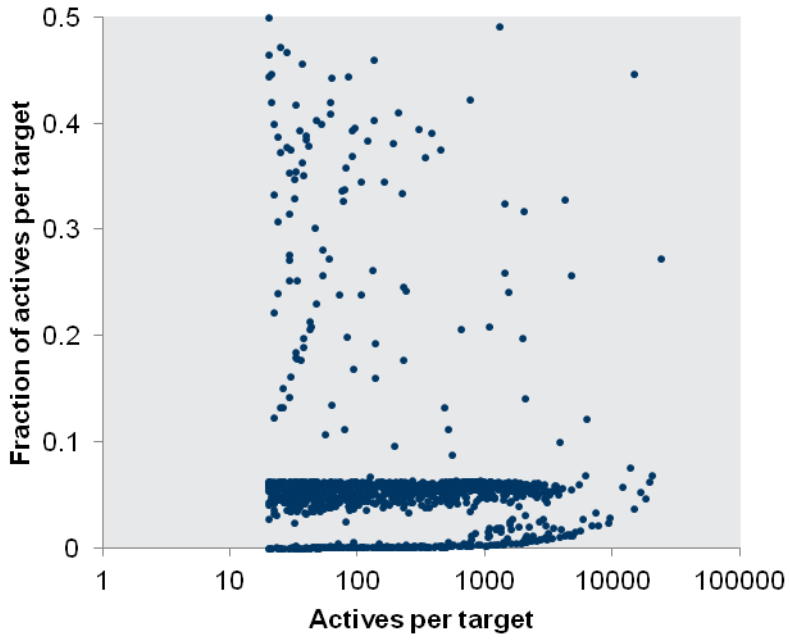

Figure 2. The distribution of active compound among the targets.

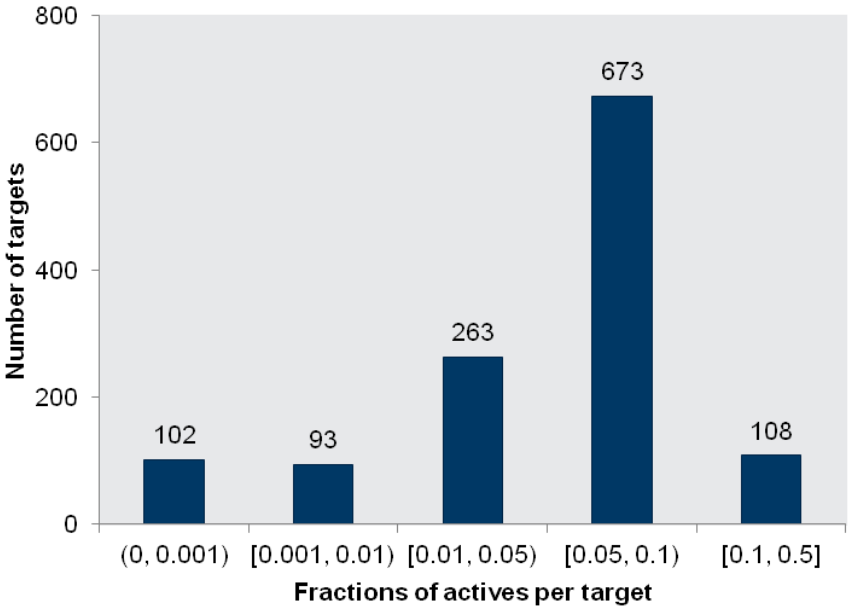

Figure 3. The fraction of actives in the dataset.

The target family distribution among the dataset was also examined. It can be seen from Figure 4 that the distribution of several major target families in the dataset. The biggest chunk of target is enzymes and second largest target family is GPCR following by ion channels and NHRs. Some further detailed target distribution can be seen in Figure 5 and 6.

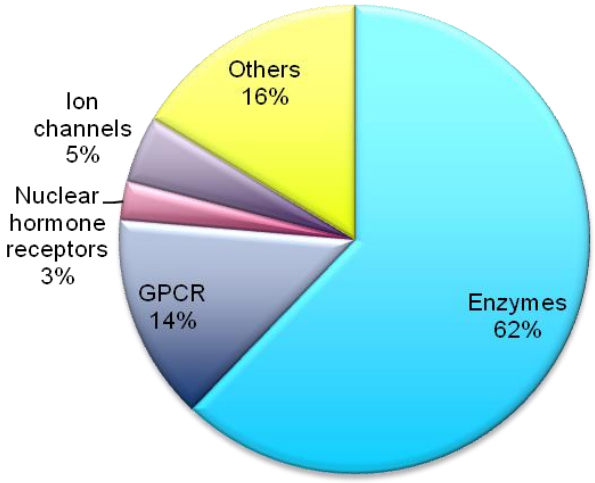

Figure 4. Target family distribution in the dataset

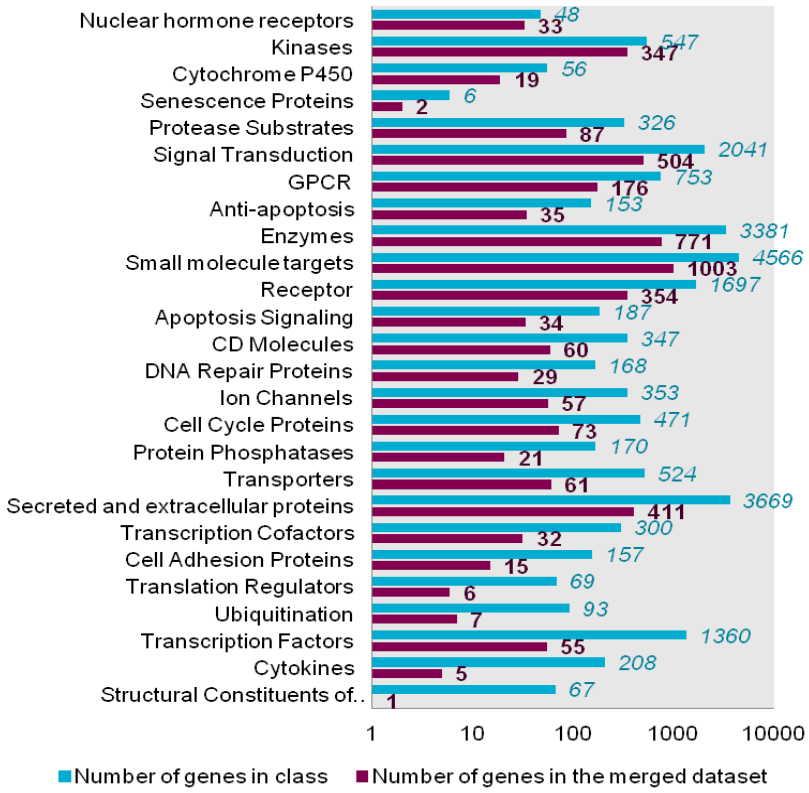

Figure 5. Detailed target subfamily distribution in the dataset.

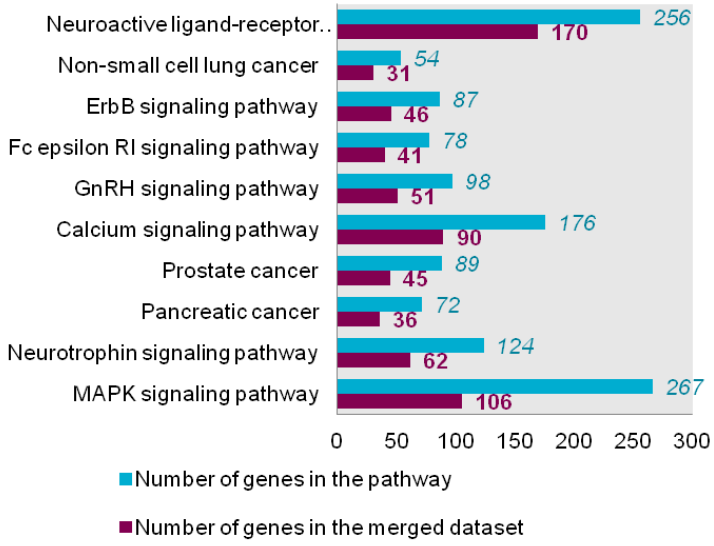

Figure 6. The target distribution according to pathways in KEGG (http://www.genome.jp/kegg/) pathway database.

The physicochemical property distribution of the dataset was also investigated. Distribution of four properties was shown in Figure 7. Figure 7a is the for molecular weight (MW), 7b is ClogP for compound lipophilicity measurement, 7c is polar surface area which represent compounds polarity and 7d is fraction of Sp3 carbon in the compound, which represent the

flatness of compounds[13]. In general, these distribution shows that most of compounds in the dataset fulfill the Lipinski rule-of-five and are drug like compounds.

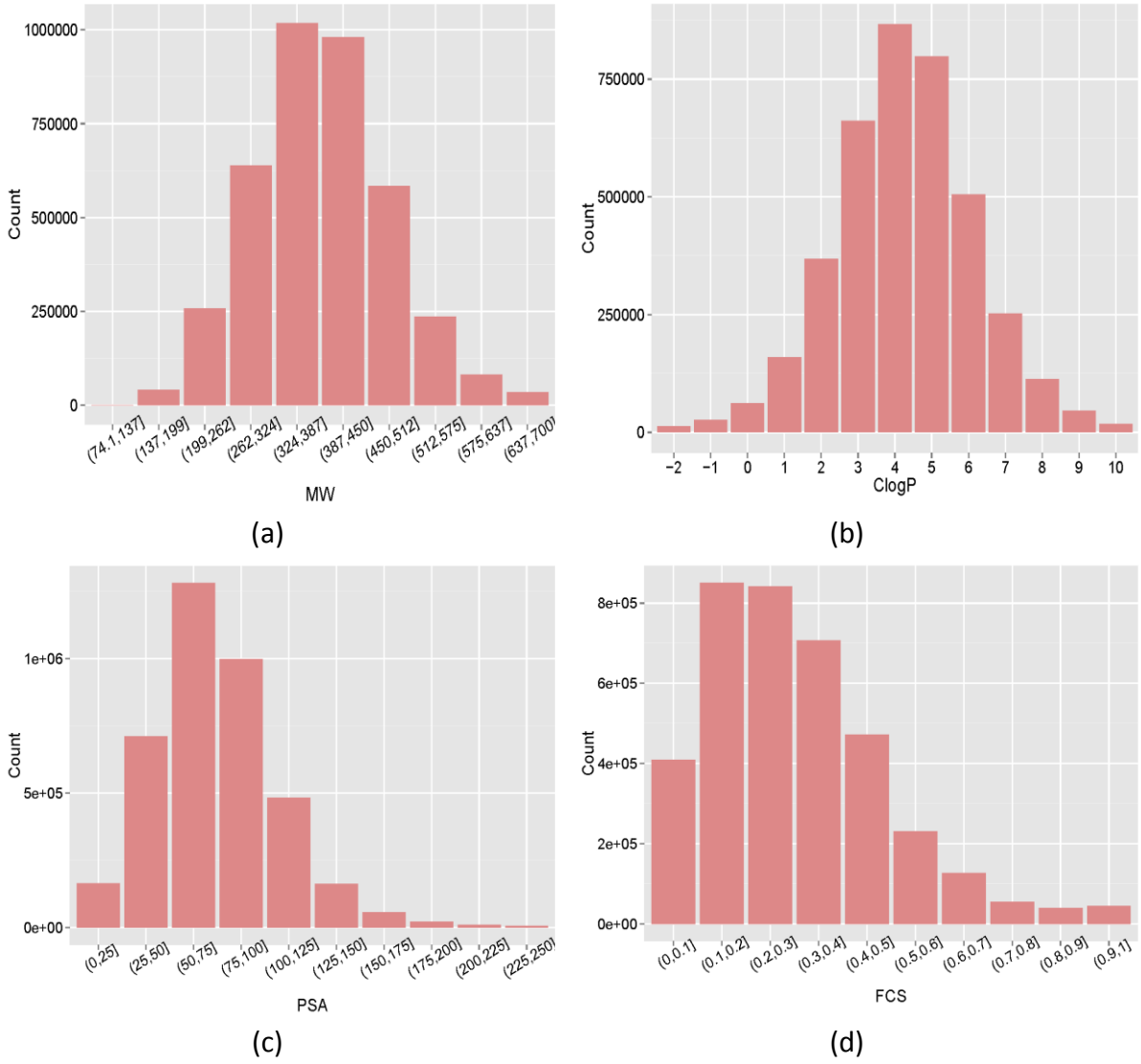

Figure 7. The distribution of (a) MW, (b) ClogP, (c) PSA and (d) FCS.

5 Conclusion

This work aims to create a large comprehensive chemogenomics dataset as a benchmark dataset for evaluating novel machine learning algorithms developed in WP1 and WP2 of ExCAPE project. This dataset comprises over 70 million SAR data points existed in public available database Pubchem and ChEMBL, structure and target information are all included in the dataset. It will serve as a useful chemogenomics data source not only for benchmark study but also for building predictive models on *in silico* polypharmacology prediction.

6 Annexes

Public chemogenomics dataset

Public chemogenomics dataset is saved at IT4I Salomon server at:

/scratch/work/project/excape-public/data/az_PubChem_ChEMBL20/Dataset_V1_2016-03-01/

- Column description in file az_Merged_Data_Pubchem_ChEMBL20.20160318.txt:

1. Compounds information related columns:

EntryID: Pubchem compound ID(CID) or ChEMBL compound ID.

AMBIT_SMILES: Generated from AMBIT(version 3.0.2, build 7385)

AMBIT_InChIKey: Generated from AMBIT(version 3.0.2, build 7385)

AMBIT_InChI: Generated from AMBIT(version 3.0.2, build 7385)

2. Targets ID columns:

GeneID: Entrez GeneID, Last modified on January 4, 2016

3. Columns of bioactivities:

ActivityFlag: A->active (pXC50>=5); N->inactive

4. Column for descriptor

JCM_ECFP_Fingerprints: Hashes of ECFP fingerprints that generated by jCompoundMapper

- Column description in file az_Merged_Data_Entrez_ID_2_HGNC_Symbol_Class.txt

Entrez_ID column : Entrez gene id. Last modified: January 4, 2016

HGNC_Symbol column : HGNC gene symbol, Homo sapiens genes GRCh38.p5

Class column: Targets are classified by http://www.silenceselect.com/listGeneClasses.do

- Column description in file az_Merged_Data_Ambit_InChIKey_2_ECFPs.txt

AMBIT_InChIKey column: Generated from AMBIT(version 3.0.2, build 7385)

JCM_ECFP_Fingerprints column: Hashes of ECFP fingerprints that generated by jCompoundMapper

- Column description in file az_Merged_Data_Pubchem_ChEMBL20.20160301.txt

EntryID column: Pubchem compound ID(CID) or ChEMBL compound ID. For all compounds: molecular weight <=1000, number of heavy atoms >12, be organic.

AMBIT_SMILES column: Generated from AMBIT(version 3.0.2, build 7315)

AMBIT_InChIKey column: Generated from AMBIT(version 3.0.2, build 7315)

AMBIT_InChI column: Generated from AMBIT(version 3.0.2, build 7315)

GeneID column: Entrez GeneID, modified on January 4, 2016

ActivityFlag column: A->active (pAct>=5); N->inactive

- Column description az_Merged_Data_Pubchem_ChEMBL_Deposit_Info.20160301.txt

EntryID column: Pubchem compound ID(CID) or ChEMBL compound ID

DB column: 20->ChEMBL; pubchem->PubChem; pubchem_unlabeled-> untested compounds in PubChem

AssayID column: Identifier of bioassay (For bioactives of actives, only single target assays were included).

Steps of public chemogenomics data curation

1 Obtain the list of PubChem assay IDs

To search assay records as confirmatory and single-target, the resulted query will be: http://www.ncbi.nlm.nih.gov/pcassay?term=%22confirmatory%22[activityoutcomemethod]%20AND%201[TargetCount]

From the result page, you can export the AID list by following:

a) From the pull-down menu following “Send to:” , select “File”

b) Select “ID List” from the “Format” menu

c) Click “Create File”

This gave the file of “confirmatory_single_target_assay.txt”. One assay ID per line.

Carried on 20160112

2 Retrieve PubChem data

2.1 Update PubChem data if necessary (option for future update)

Remove your assay IDs that already existed in the collection, i.e. confirmatory_single_target_assay.txt.

2.2 Query PubChem by assay ID

Script

```
#!/bin/bash

while read -r aid; do wget -O Output_dir/$aid.txt -o Output_dir/$aid.log "http://pubchem.ncbi.nlm.nih.gov/assay/getassay.cgi?query=bioactivity&task=download&aid=$aid"; done < confirmatory_single_target_assay.txt
```

where $aid is the assay ID and Output_dir is the path of log files.

2.3 Check the log files to find the missing assays

Script

```
#!/bin/bash

while read -r id; do if grep "saved" Output_dir/$id.log

then

    echo -e "$id\tok" >> assay_error_check.txt

else

    echo -e "$id\tfailed" >> assay_error_check.txt

fi ; done < confirmatory_single_target_assay.txt
```

where $id is the assay ID and Output_dir is the path of log files.

2.4 Header of outputs

Column 1-15

Col_1. AID

Col_2. Panel ID

Col_3. SID

Col_4. CID

Col_5. Activity

Col_6. AC Value (micromolar)

Col_7. AC Name

Col_8. BioAssay Name

Col_9. GI

Col_10. Target Name

Col_11. Outcome Method

Col_12. PubMed ID

Col_13. GeneID

Col_14. Target Count (active)

Col_15. Target Count (tested)

3 PubChem data cleaning

3.1 remove substances that have blank compound identifier (CID)

awk -F"\t" '$4!=""' input > output

where "$4" refer to column 4 -CID.

3. 2. Remove assays whose AC Names are unwanted

The list of unwanted AC Names:

a) empty

b) Active Concentration

c) AbsAC0.39_uM

d) Max_Activity_Concentration_uM

e) Max_Concentration(uM)

f) Max_Delta_Tm_Conc_uM

g) AC_50uM

h) TD50 (microM)

i) 2nd Test Concentration

j) AbsAC0.47_uM

k) AbsAC10_uM

l) Compound_Concentration

m) LD50

n) AbsAC0.54_uM

o) AbsAC35_uM

p) IC90

q) AbsAC200_uM

r) ID50

s) AbsAC40_uM

t) AbsAC1_uM

u) Km

v) ED50

w) AbsAC1000_uM

x) 64

y) ...

Again, awk 'BEGIN {FS="\t";OFS="\t"}; NR==FNR{a[$1];next} !($7 in a)' list_of_unwanted input > output.

3. 3. Only targets of human, mouse and rat are kept

Link Entrez ID to NCBI Taxonomy ID by any gene2* file under the directory of ftp://ftp.ncbi.nih.gov/gene/DATA/

Table 1 Tax IDs of three species

Tax_ID	Tax_Name
9606	Homo sapiens
10116	Rattus norvegicus
10090	Mus musculus

I already made a mapping table that included all current Entrez IDs in human, rat and mouse. - /scratch/work/project/excape-public/data/az_PubChem_ChEMBL20/Dataset_V2_2016-04-15/targets/EntrezID2ortholog_group.

You can also apply some other scripts here. Pay attention to that version of the tools and databases may cause problems.

4 ChEMBL 20

4.1 Retrieve from AZ's in-house-curate database

You can download data directly from ChEMBL if you like.

4.2 Convert discontinued Entrez IDs to current Entrez IDs if they have

According to the file “gene_history” under the directory of ftp://ftp.ncbi.nih.gov/gene/DATA/.

Header of “gene_history”:

tax_id GeneID Discontinued_GeneID Discontinued_Symbol Discontinue_Date

If GeneID is not empty, convert Discontinued_GeneID to GeneID in the ChEMBL if you mapped the Uniprot ID to Entrez ID improperly. They’re not 1-to-1 mapping, i.e. one Uniprot ID mapped to 2 Entrez IDs, 1 Entrez ID have multi Uniprot IDs.

5 Collect inactives from PubChem's screening assays

Script

wget "http://pubchem.ncbi.nlm.nih.gov/assay/getassay.cgi?query=bioactivity&task=download&geneid=$geneid&actvty=Inactive&aomethod=Screening"

where $geneid is Entrez ID from either ChEMBL 20 or the obtained PubChem dataset.

date: 20160404

6 Target annotations

Path: /scratch/work/project/excape-public/data/az_PubChem_ChEMBL20/Dataset_V2_2016-04-15/targets

EntrezID2ProteinAccession:

Covert Entrez ID to gene symbol, NCBI protein accession and UniProtKB protein accession

EntrezID2ortholog_group (option):

Covert Entrez ID to the group of gene ortholog. Group name staring with “un” like un180 means that the corresponding Entrez ID is not grouped.

Refs:

Entrez ID to Tax_ID, NCBI protein accession and gene symbol

ftp://ftp.ncbi.nih.gov/gene/DATA/gene2accession.gz

date: 20160111

NCBI protein accession to UniProtKB protein accession

ftp://ftp.ncbi.nih.gov/gene/DATA/gene_refseq_uniprotkb_collab.gz

date: 20160406

Entrez ID to Orthologs by building an ortholog table (EntrezID2ortholog_group) from

ftp://ftp.ncbi.nih.gov/gene/DATA/gene_group.gz

date: 20160404

Entrez Gene ID to discontinued ID

ftp://ftp.ncbi.nih.gov/gene/DATA/gene_history.gz

date: 20160212

7 Compounds standardization

Standardization by ambitcli-3.0.2, build 7385.

Most were retrieved from IDEA's work (/scratch/work/project/excape-public/data/standardization/standardization-current/). I performed Ambit to the missing cpds

java -jar /home/jiangming/ambitcli.jar -a standardize -i "input" -m post -d page=0 -d pagesize=-1 -d tag_smiles=AMBIT_SMILES -d tag_inchi=AMBIT_InChI -d tag_inchikey=AMBIT_InChIKey -o "output" -d tautomers=true -d splitfragments=true -d implicith=true -d smiles=true -d smilescanonical=false -d inchi=true -d neutralise=true -d isotopes=true

8 Compounds' property filtering

Filtered by property

a) Organic

b) Number of heavy atoms > 12

c) Molecular weight <= 1000.

This was actually done by a pipeline pilot protocol.

9 Compounds' descriptors

Path:

/scratch/work/project/excape-public/data/az_PubChem_ChEMBL20/Dataset_V2_2016-04-15/descriptors/

9.1 CDK fingerprints

Most were retrieved from IDEA's work (/scratch/work/project/excape-public/data/standardization/standardization-current/). I generated CDK fingerprints for missing compounds by script:

java -Xmx1536m -jar /home/jiangming/ambitcli.jar -a fingerprint -m post -d page=0 -d pagesize=-1 -d fpclass=CircularFingerprinter -d tag_tokeep=AMBIT_InChIKey -d inputtag_smiles=AMBIT_SMILES -d inputtag_inchikey=AMBIT_InChIKey -d inputtag_inchi=AMBIT_InChI -d write_count=true -i "input" -o "output" > ambit.log

9.2 JCM ECFP fingerprints

Generated by a jcm web service

Start JCM services with command:

java -Xmx4g -XX:hashCode=5 -jar .../jcm_share/jcmapper/jcompoundmapperservice/target/jcompoundmapperservice-1.0-SNAPSHOT.jar --port 8082 -a DAYLIGHT_INVARIANT_RING -d 6 --removeHydrogens false --useAromaticity true

java -Xmx4g -XX:hashCode=5 -jar ~/IT4I/scratch/work/project/excape-public/jcm_share/jcmapper/jcompoundmapperservice/target/jcompoundmapperservice-1.0-SNAPSHOT.jar --port 8082 -a DAYLIGHT_INVARIANT_RING -d 6 --removeHydrogens false --useAromaticity true

Query fingerprint by script:

while read smiles cid others; do curl --data-urlencode "smiles=$smiles" -d "fpType=ECFP" -d "atomLabelType=DAYLIGHT_INVARIANT_RING" -d "distanceCutoff=6" -s "http://localhost:8082/Fingerprints" | sed -e 's|"hash":|\nhash\n|g' | sed 's|,"value|\n|g' | grep -A 1 -B 0 -w "^hash$" | sed '/--$/d' | sed '/^hash$/d' | paste -d, -s | sed -r 's|^|'$cid'\t|g' ; done < input_smiles > output.txt

where $cid could be ambit_inchikey depend on your aim

9.3 Signatures

Generated by our in-house tool that gave two files. One is in the libsvm format – "Compounds2Signatures.libsvm.sparse.txt". Its index refer to the order of signatures in another file "Compounds2Signatures.libsvm.sparse.orderofSignatures.txt"

10 Data combination (PubChem and ChEMBL20)

Sorted by pXC50 values, this gave three datasets

I. The merged data set before aggregation

"pooled.ChEMBL_PubChem.pXC50_sorted.txt"

II. All dose-response data before aggregation

"pooled.ChEMBL_PubChem.pXC50_sorted.DoseResponsedOnly.txt"

awk –F"\t" '$5!="" && $5>-1000' pooled.ChEMBL_PubChem.pXC50_sorted.txt > pooled.ChEMBL_PubChem.pXC50_sorted.DoseResponsedOnly.txt

III. The data set after aggregation

"pooled.ChEMBL_PubChem.cleaned.Max_pXC50.atLeast20ActiveCpdsPerGeneOrthologGroup.txt"

Rules of data aggregation

Only the maximum values of Ambit_InChIKeys per Ortholog_Group were kept for actives and inactives, respectively.

a) Actives

pXC50>=5, should have at least 20 active cpds per Ortholog_Group,

.. | sort | uniq -c | sort –nr | …

b) Inactives

Remove all inactives if count of actives cpds less than 20 per Ortholog_Group.

Remove overlaps in inactives if such pairs existed in actives.

awk 'BEGIN {FS="\t";OFS="\t"}; NR==FNR…

Obtain the list of PubChem assay IDs

7 References

[1] Koutsoukas A, et al. From in silico target prediction to multi-target drug design: current databases, methods and applications. J Proteomics, 74, 2011, 2554–2574

[2] Ji Z. L., et al, In silico search of putative adverse drug reaction related proteins as a potential tool for facilitating drug adverse effect prediction. Toxicol Lett 164, 2006, 104–112.

[3] Poroikov V., et al, Top 200 medicines: can new actions be discovered through computer-aided prediction? SAR QSAR Environ Res 12, 2001, 327–344

[4] Lounkine E., et al, Largescale prediction and testing of drug activity on side-effect targets. Nature, 486, 2012, 361–367

[5] Gregori-Puigjané E., Mestres J., A ligand-based approach to mining the chemogenomic space of drugs. Comb Chem High Throughput Screen, 11, 2008, 669–676

[6] Jacob L., et al, Virtual screening of GPCRs: an in silico chemogenomics approach. BMC Bioinform, 9, 2008, 363

[7] Wang Y., et al, PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res., 37, 2009, W623–W633

[8] Bento A. P., et al, The ChEMBL bioactivity database: an update. Nucleic Acids Res., 42, 2014, 1083-1090.

[9] Kochev N., Paskaleva V., Jeliazkova N., Ambit-Tautomer: An Open Source Tool for Tautomer Generation, Molecular Informatics, 32, 2013, 481-504.

[10] Lipinski C.A., Lombardo F., Dominy B.W., Feeney P.J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev. 46, 2001, 3–26.

[11] Hinselmann G., Rosenbaum L., Jahn A., Fechner N., Zell A., jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints. J. Cheminform. 3, 2011, 3.

[12] Steinbeck C., Han Y., Kuhn S., Horlacher O., Luttmann E., Willighagen E., The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics, J. Chem. Inf. Comput. Sci., 43, 2003, 492-500.

[13] Lovering F., Bikker J., Humblet C., Escape from flatland: increasing saturation as an approach to improving clinical success. J. Med. Chem., 52, 2009, 6752-6.