Chemical Data Retrieval and Management
ChEMBL, ChEBI, and the Chemistry Development Kit

Stephan A. Beisken
What is EMBL-EBI?

- Part of the European Molecular Biology Laboratory
- International, non-profit research institute
- Europe’s hub for biological data services and research
EMBL-EBI’s mission

• Provide **freely available data** and bioinformatics services to all facets of the scientific community in ways that promote scientific progress

• Contribute to the advancement of biology through basic investigator-driven **research** in bioinformatics

• Provide advanced bioinformatics **training** to scientists at all levels, from PhD students to independent investigators

• Help disseminate cutting-edge technologies to **industry**

• Coordinate biological data provision throughout **Europe**
Data resources at EMBL-EBI

Genomes & variation
- Ensembl
- Ensembl Genomes
- Genome-phenome archive
- Metagenomics

Nucleotide sequences
- European Nucleotide Archive (ENA)

Expression
- Array Express
- Expression Atlas
- PRIDE
- R-Workbench

Molecular structures
- Protein Data Bank in Europe
- PDBsum
- ProFunc

Proteins
- The Universal Protein Resource (UniProt)
- InterPro

Chemical biology
- ChEMBL
- ChEBI

Pathways
- IntAct
- Reactome
- Metabolights

Literature & ontology
- Europe PubMed Central
- Gene Ontology

Patent sequences
- Non-redundant patent sequence dbs
- Patent compounds

Systems
- BioModels
- Enzyme Portal
- BioSamples
Introduction to ChEBI

- **Chemical Entities of Biological Interest**
- Focused on ‘small’ chemical entities: *no proteins or nucleic acids*
- **Scope (material entities)**
  - mainly molecules
  - but also salts, polymers, atoms, electrons, groups, residues
- **Scope (dependent entities)**
  - chemical role
  - biological role
  - application

Access ChEBI at [http://www.ebi.ac.uk/chebi/](http://www.ebi.ac.uk/chebi/)
Chemicals - ChEBI

Nomenclature
caffeine
1,3,7-trimethylxanthine
methyltheobromine

Chemical data
Formula: C8H10N4O2
Charge: 0
Mass: 194.19

Ontology
metabolite
CNS stimulant
trimethylxanthines

Database Xrefs
MSDchem: CFF
KEGG DRUG: D00528

Chemical Informatics
InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
SMILES CN1C(=O)N(C)c2ncn(C)c2C1=O
ChEBI ontology

Organised into three sub-ontologies, namely

- Molecular structure ontology
- Subatomic particle ontology
- Role ontology

(R)-adrenaline
ChEBI Ontology

(R)-2-hydroxystearate 3-hydroxybenzoate (+)-epigallocatechin 4-guanidinobutanoic acid zwiteron adipate(1-) 3-hydroxy carboxylic acid alpha,omega-dicarboxylic acid catechin adenosine 3'-phosphate 3'-coumarate chlorophenoxyacetic acid flavan-3,3',4',5,5',7-hexol carboxamidine D-fructose flavouring agent monohydroxybenzoic acid ecotine zwiteron butyric acid glycosylglucose purine ribonucleoside 3'-monophosphate methoxybenzoic acid pyrimidinemonocarboxylic acid uridines purine nucleobase laxative esoclinate guaninides nutraceutical vanillate p-block molecular entity

antioxidant secondary metabolite valeric acid cytosine biotins osmolyte succinimide hydroxybenzoic acid tyrosine kinase inhibitor hydroxyanthraquinones oxopurine synthetic auxin phenoxyacetic herbicide cytidines citramalate(2-) hydroxybenzaldehyde phenols ketohexose monophosphate biotinate altrose coenzyme hydroxycoumarin cystathionines benzaldehydes anticonvulsant coumaric acid arabinofl D-aldohexose 3'-AMP(2-) 4-hydroxybenzoate antineoplastic agent 9H-purine 2-hydroxyxystearic acid 2-aminopurines 4-methyl-2-oxopentanoate 2-hydroxydicarboxylic acid
Structure Search

- XLS Reader
- Molecule to CDK
- CDK to Molecule
- ChEBI Searcher
- Column Rename (Regex)
- ChEBI Retriever
- Missing Value
- Column Filter
- Interactive Table

Filtered table - Column Filter

Dialog - ChEBI Searcher

Table View - Interactive Table

File

Properties
- Table "default" - Rows: 42
- Spec - Columns: 2

Properties
- Row ID
- Name
- InChI

Parameters
- CHEBI Service
- Get hit entity
- Max results
- Tanimoto cutoff
- Search category structure
- Search type
- CHEBI Stars
- CHEBI entity

Table View
- Row ID
- Name
- InChI
- CHEBI Name
- CHEBI

Row 0063
1,8-diaminooctane
CHEBI:31112

Row 1000
1H-indole
CHEBI:16881

Row 1001
1H-indole-1-acetic acid
CHEBI:12214

Row 1007
1-methyl-2-hydroxoadenine
CHEBI:16203

Row 1020
1-methyltrytophan
CHEBI:72921

Row 1224
2-aminoacidic acid
CHEBI:10020
Introduction to ChEMBL?

- Open access database for drug discovery
- Freely available – searchable and downloadable
- Contents:
  - Bioactivity data manually extracted from the primary medicinal chemistry literature
  - Deposited data from neglected disease screening (e.g. Malaria)
  - Subset of data from PubChem
- Bioactivity data is associated with a biological target and a chemical structure
- Updated regularly with new data

Access ChEMBL at [http://www.ebi.ac.uk/chembl/](http://www.ebi.ac.uk/chembl/)
Chemicals ChEMBL?

- ChEMBL_15
- Compounds: 1,254,575
- Assays: 679,259
- Targets: 9,570
- Publications: 48,735
- Activities: 10,509,572
- Data sources: 16

Increase of >230,000 compounds from literature since ChEMBL01
### Bioactivities query for hERG

#### ChEMBLdb Connector

- **Compound Search**
- **Flow Variables**

#### Activity value, assay description, compound, reference

<table>
<thead>
<tr>
<th>Row</th>
<th>ChEMBL ID</th>
<th>Protein Accession</th>
<th>Search Type</th>
<th>Bioactivities</th>
<th>Activity, assay description, compound, reference</th>
</tr>
</thead>
</table>
Compound Search

Query

Config Dialog

List of NNs
Polypharmacology Profile

Find NNs

Filter, summarise & pivot

Retrieve bioactivities

Query

- Adenosine A1 receptor
- Adenosine A2a receptor
- Adenosine A3 receptor
- Cyclin A2
- Cyclin T1
- Cyclin-dependent kinase 2
- Cyclin-dependent kinase 5/CDK5 activator 1
- Cyclin-dependent kinase 9
- DNA-dependent protein kinase
- GBM
- Glycogen synthase kinase-3 beta
- HEK293 (Embryonic kidney fibroblasts)
- LS174T (Colon adenocarcinoma cells)
- PI3-kinase p110-alpha subunit
- PI3-kinase p110-beta subunit
- PI3-kinase p110-delta subunit
- PI3-kinase p110-gamma subunit
- Serine/threonine-protein kinase AKT2
- Tyrosine-protein kinase JAK2
- Tyrosine-protein kinase JAK3
- cAMP-dependent protein kinase alpha-catalytic subunit
- cAMP-dependent protein kinase beta-1 catalytic subunit
- cAMP-dependent protein kinase, gamma catalytic subunit

Compounds

- CHEMBL220335
- CHEMBL402493
- CHEMBL241134
- CHEMBL423968
- CHEMBL164596
- CHEMBL227695
- CHEMBL227727
- CHEMBL227866
- CHEMBL228042
- CHEMBL375659
- CHEMBL375696
- CHEMBL1603976
- CHEMBL1603016
Downstream: What next?

• Chemical space clustering & visualisation
• (Q)SAR analysis
  • Data modeling, activity cliffs, FW, MMP analysis
  • Bioisosteric replacements mining
• De novo design
  • Evolutionary compound optimisation
• Target fishing
• (off-)target prediction and ADR analysis
• Polypharmacology networks
• Druggability / Drug-likeness
An Open Source Java™ Library for Structural Chem- and Bioinformatics

- version 1.5.x → KNIME-CDK

Features

- reads and writes chemical data formats
- renders chemical structures
- substructure search and highlighting
- signature and fingerprint generation
- qsar descriptors
- ...
- ...
- ...
- chemical graph theory

http://tech.knime.org/community/cdk
http://tech.knime.org/forum/cdk
Common Features

SDF Reader → Molecule to CDK → Substructure Search → Molecular Properties → M-methylstyrene → sp3 char, sum formula, major isotopic mass → Symmetry

Molecules with fragment: 25 - Substructure Search/M-methylstyrene

Row 40

Row 47

Row 48

Out-Port name: 26 - Symmetry

Table "default" - Rows: 215 - Spec - Columns: 2 - Properties - Flow Variables

Row ID  Molecule

Row 140

EMBL-EBI
Similarity Search

After generating the similarity matrix, we replace values that are greater than our threshold (Tanimoto cutoff) with reference molecules. These are the 'most similar' molecules, i.e., the result of our search.

The Loop End collates the results of our column-wise operations. The output table contains only molecules exceeding the Tanimoto threshold.

Reformat result matrix to row-centered representation. The first column contains the query molecules, subsequent columns contain the similar reference molecule.
Workflows are Great!
Acknowledgements

• Cheminformatics and Metabolism Team
  • Dr. Christoph Steinbeck
  • the whole team

• Computational Chemical Biology Team
  • Mr. Edmund Duesbury
  • Dr. George Papadatos

• Syngenta
  • Mr. Mark Earll
  • Dr. Mark Seymour
  • Dr. David Portwood

Thank you for your attention.