MS processing

Annotation & Databanks

Franck Giacomoni, Yann Guitton

11/10/2018

CC by-nc-nd :
http://creativecommons.org/licenses/by-nc-nd/2.0/fr/
Overview

Data acquisition

Extraction Data

Annotation

Biological interpretation

Identification

Statistical analysis

Extraction Data

Biological interpretation
Overview

What is a database?

“An organized collection of information”

Main goal in the “Annotation” step:

finding metabolite candidates...

...linking knowledge and experimental observations
Overview

my spectrum + my biological question
Large databank set

Chemical compounds banks

“Home based” libraries

Spectral libraries

Genome-wide metabolic reconstruction based databases

References repositories

Knowledge databases

“A bank is one way of representing information”
More than 500 references available

Commercials banks

« In House » banks
A local knowledge aggregation
Representation of one particular question
Interoperability issue or limit

Reference banks
Long term support
New submission possibilities
Curation steps

Author: McCracken, Theresa
Updated inventories

Metabolomics Society portal: http://metabolomicssociety.org/resources/metabolomics-databases

Food Biomarker Alliance portal: http://foodmetabolome.org/wpkg4

Publications:

Author: McCracken, Theresa
Overview

Quantity vs Quality

PubChem

A public repository of information on small molecules and their biological activities

157,000,000 entries

A Comprehensive Species-Metabolite Relationship Database

50,000 entries

https://fr.123rf.com
Inventory

Source: http://metabolomicsociety.org/resources/metabolomics-databases
Inventory

Chemical compounds centric resources
Name
Chemical Entities of Biological Interest
Freely available dictionary of molecular entities focused on ‘small’ chemical compounds.

Address
EMBL – EBI

Entries
41,000 entities fully annotated

www.ebi.ac.uk/chebi
Public and reference resource

Star status
Manually annotated by the ChEBI team
Manually annotated by a third party,
Marked as deleted or obsolete.

ChEBI Ontologies: Molecular structure, roles and subatomic particle information.

Manual search by mass / formula
Web services
Public and reference resource
Data
Chemical structure rules
Naming convention
Downloadable spectra
MSMS predicted data available in pos/neg mode

Lipids classification

Manual search by mass / formula / Names
Web services
Inventory

Spectral centric resources
MassBank.JP

Spectral DataBank

**Name**
Database of comprehensive, high-resolution mass spectra of metabolites

**Address**
Institute for Bioinformatics Research and Development, Japan Science and Technology Agency

**Entries**
- 41,000 MS spectra
- 47,000 compounds

www.massbank.jp/en
Public and reference resource(s)
Distributed database / Contributions
Data

- Downloadable spectra
- High precision and accurate mass spectra of primary/secondary metabolites.
- Massive work on controlled vocabulary/format: [PDF]

Detailed MS data / metadata
Spectral/structure searching utilities
Search for merged spectra
Web services

API

Tool issue on W4M
Spectral DataBank

**Name**
Metabolite database focusing on metabolomic approaches

**Entries**
- 59,000 MSMS spectra
- 64,000 structures

**Address**
Institute for Bioinformatics Research and Development, Japan Science and Technology Agency
[https://metlin.scripps.edu](https://metlin.scripps.edu)
GMD

Golm Metabolome Database

Reference mass spectra from biologically active metabolites quantified using GC-MS.

Max Planck Institute of Molecular Plant Physiology

http://gmd.mpimp-golm.mpg.de
Public and reference resource Data

Downloadable spectra
Mass spectra and retention time indices of pure reference substances
Quantitative data

Spectral searching utilities

Web services

API
Inventory

Knowledge databases
Species/Metabolomics databank

**Name**
Human Metabolome Database
Database containing detailed information about small molecule metabolites found in the human body

**Entries**
114,100 metabolite entries

**Address**
The Metabolomics Innovation Centre (TMIC)
www.hmdb.ca
Public and reference resource

Data

Downloadable spectra - MS, MSMS (and NMR)
Data from various instruments (low/high res.)
Predicted spectra
Human metabolites with chemical properties, biochemistry data and clinical data...
HMDB 4.0 propose in silico predicted spectra

MetaboCards and hudge cross-links work
Spectral/structure searching utilities
No public web service
**PHYTOHUB**

- Design for use in nutritional metabolomics studies.
- Inventories all dietary phytochemicals present in commonly consumed foods and their human metabolites.
- Data on the main dietary sources, on human metabolism, chemical features and MS spectra.
- Manually curated by an international team of scientists with expertise on phytochemicals.

**In-house tool**
Expert knowledge of food phytochemicals biotransformations

- **Chemical Structures**
- Predicted metabolites
- Physico-chemical properties
- Dietary sources
- Spectral data

**Books**: Phytochemical dictionary... Dr. Duke’s Dictionary of Food Compounds

**Databases**:
- Literature survey
- OpenBabel
- Chemaxon Marvin
- In-house tool
- Expert knowledge of food phytochemicals biotransformations

**www.phytohub.eu**

PI: Claudine Manach (INRA)
Inventory

Reference repositories
Metabolomics data repositories

MetaboLights: An Open-Access Database Repository for Metabolomics Data
Namrata S. Kale,1 Kenneth Haug,1 Pablo Conesa,1 Kalaivani Jayaseelan,1 Pablo Moreno,1 Philippe Rocca-Serra,2 Venkata Chandrasekhar Nainala,1 Rachel A. Spicer,1 Julian L. Griffin1


MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data
Kenneth Haug1, Reza M. Salek1,2, Pablo Conesa1, Janna Hastings1, Paula de Matos1, Mark Rijnbeek1, Tejasvi Mahendraker1, Mark Williams1, Steffen Neumann1, Philippe Rocca-Serra1, Eamonn Maguire2, Alejandra Gonzalez-Beltran2, Susanna-Assunta Sansone1, Julian L. Griffin2,3 and Christoph Steinbeck1,4

1European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD; 2MRC HNR, Elsie Widdowson Laboratory, Fulbourn Road, Cambridge CB1 9NL; 3Department of Biochemistry and Cambridge Systems Biology Centre, University of Cambridge, Cambridge CB2 1GA, UK; 4Department of Stress- and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany and 2Oxford e-Research Centre, University of Oxford, 7 Keble Road, Oxford OX1 3QG, UK

Received August 15, 2012; Revised September 22, 2012; Accepted October 1, 2012

ENPADASI
European Nutrition Phenotype Assessment and Data Sharing Initiative

DataCite
FIND, ACCESS, AND REUSE DATA

http://www.ebi.ac.uk/metabolights
Email: metabolights-help@ebi.ac.uk
- « Cross » species, « cross » techniques
- Accurate data collection
- « Open access »
Keynote – Thursday afternoon (6pm) :
Prof. Dr. Christoph Steinbeck

« Title still unavailable… »

Cheminformatics and Computational Metabolomics group
at the Friedrich-Schiller-University in Jena, Germany
Inventory

Genome-wide metabolic reconstruction based databases
KEGG COMPOUND is a collection of small molecules, biopolymers, and other chemical substances that are relevant to biological systems.

Entries
18,039 metabolites and small molecules

Address
Kanehisa Lab. - NPO
Bioinformatics Japan
http://www.genome.jp/kegg
Public and reference resource

Data

Compounds (not available as full resource)

Biological roles

Not adapted for metabolomics direct uses

Reference Pathway – Metabolism information

Strong link with metabolic pathways maps

WebServices

API
Tip and Trick

How to build my own bank?
Building my own bank?

Why?
- No reference banks existing
- Specific molecules or tricky forms
- Consolidation of the knowledge of the lab on chemical family annotation
- Specific analytical technics
- Sharing!

Format?
- Simple: a flat text file (notepad or excel export as CSV)
- Structured: databases

Features?
- For compounds: common name, monoisotopic mass, ids, INCHI/INCHIKEY
- For spectrum: accurate peak list (m/z, intensity, RT)
**Dbs Overlaps and Metabolome coverage**

Venn diagram showing the overlap between open mass spectral databases (HMDB, MassBank, GNPS, and ReSpect)

This table shows how PeakForest chemical library covers genome scale reconstructions of metabolic networks. This mapping is performed using MetExplore webservice.

<table>
<thead>
<tr>
<th>Metabolic network</th>
<th>Number of metabolites in the network</th>
<th>Metabolites found in PeakForest</th>
<th>Coverage of PeakForest chemical compounds in Metabolic network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>1068</td>
<td>345</td>
<td>32.3 %</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>228</td>
<td>135</td>
<td>59.21 %</td>
</tr>
<tr>
<td>Homo sapiens (Recon2)</td>
<td>1190</td>
<td>448</td>
<td>37.65 %</td>
</tr>
<tr>
<td>Homo sapiens (BioCyc)</td>
<td>850</td>
<td>371</td>
<td>43.65 %</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>1025</td>
<td>424</td>
<td>41.37 %</td>
</tr>
<tr>
<td>PlantCyc</td>
<td>2339</td>
<td>481</td>
<td>20.56 %</td>
</tr>
<tr>
<td>Rattus Norvegicus</td>
<td>1337</td>
<td>482</td>
<td>36.05 %</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>224</td>
<td>144</td>
<td>64.29 %</td>
</tr>
</tbody>
</table>

**Sources:** Vinaixa et al. 2016, TrAC - DOI: 10.1016/j.trac.2015.09.005 and PeakForest project
A case of Chinese whispers

Source: Cl. Frainay – INRA Toxalim
During annotation

*Using only one database is difficult*
*Be aware about the quality of the selected bank*
*A match in a bank is a “clue” NOT a “proof”*

Enrich your results with:
*Databanks IDs*
*INCHI (chem. structure representation)*

*And please - Make the effort to submit your experimental/annotated data to Repositories (MetaboLights, W4M – DOI, …)*
References