Canadian Bioinformatics Workshops

www.bioinformatics.ca
Learning Objectives

• To learn about different database models
• To gain awareness of different kinds of metabolomic databases
• To learn about public NMR and MS spectral databases
• To learn about pathway databases
• To learn about comprehensive public metabolomic databases
Two Solitudes

Cheminformatics vs. Bioinformatics

- Established in the 1960’s
- Designed for the needs of organic chemists
- User-pay, limited public access
- Funded by large companies (MDL, Bielstein, Sigma, CAS)

- Established in the 1990’s
- Designed for needs of molecular biologists
- Web-based, open access model
- Funded by large gov’t agencies (NCBI, EBI, NIH, GC)
What’s A Database For?

- Information consolidation & linkage
- Information retrieval (query matching)
- Reference values, reference data, reference sequences, reference images
- Data for training/testing algorithms
- Similarity searching (image, spectra, structure, sequence, text)
- Prediction (structure, function, property, phylogeny, activity, relationship)

Database Evolution

- Hobby database (flatfile)
  - Limited coverage
  - Limited depth

- Curated, non-redundant (relational, warehouse)
  - Greater coverage
  - Greater depth

- Archived, open deposit, redundant (relational, distributed)
  - Extensive coverage
  - Modest depth

Increasing Cost + Resources

Size of user community
Database Evolution

Hobby database (flatfile)

Curated, non-redundant (relational, warehouse)

Archived, open deposit, redundant (relational, distributed)

Need for standardization
Dependence on automation
Querying capabilities

The Problem with Metabolomics

Genomics
Gene IDs + Transcript Abundance

Proteomics
Protein IDs + Concentrations

Metabolomics
Metabolite IDs + Concentrations
Metabolomics Databases

- Most data for metabolomics is still in textbooks or print journals (100+ years of clinical chemistry, 75 years of classic biochemistry)
- Field lags behind genomics/proteomics by about 20 years
- Challenge is to appeal to different user communities (metabolomics researchers, analytical chemists, plant chemists, clinical chemists, physicians, drug researchers, NMR specialists, MS specialists, bioinformaticians, standards setters, etc.)

Databases for Metabolomics

- **NMR spectral databases**
  - Primarily small molecule spectra, not all metabolites
- **MS or MS/MS spectral databases**
  - Primarily small molecule spectra, not all metabolites
- **Compound databases**
  - Mostly compound names, structures, IDs, physprops
- **Pathway databases**
  - Mix of metabolite, drug, protein, signaling pathways
- **Comprehensive metabolomic databases**
  - Combines most/all of the above, focus on metabolites
NMR Spectral DBs

SBDS

NMRShiftDB

MMCD

BMRB

SDBS

http://sdbbs.db.aist.go.jp
SDBS

- Maintained in Japan by AIST (since 1970’s)
- Includes 24,700 MS spectra, 15,400 $^1$H NMR spectra, 13,600 $^{13}$C NMR spectra, 52,500 FT-IR spectra on 34,000 cmpds
- Extensive suite of spectral search tools
- Most compounds are not metabolites, but still very useful for many researchers

BioMagResBank

http://www.bmrblabs.wisc.edu/metabolomics/
BioMagResBank

- 868 reference metabolites
- 5-6 NMR spectra ($^1$H, $^{13}$C, 1D, 2D) per compound
- Search by name, synonyms, InChI, formula, SMILES
- Focus initially on plant metabolites (Arabidopsis) although now includes other mammalian metabolites
- Assignments for many are available

NMRShiftDB 2

- 50,515 1H/13C NMR Spectra
- 42,444 structures

http://nmrshiftdb.nmr.uni-koeln.de/
NMRShiftDB

- Database originally developed by Christoph Steinbeck (also leads ChEBI)
- Not restricted to metabolites, includes many organic compounds
- Supports chemical shift prediction
- Can search by name, structure or chemical shifts (peaks and Jcamp file)
- Includes chemical shift assignments (but in organic solvents)

MMCD

20,306 cmpds
794 1H NMR
794 13C NMR
794 TOCSY
794 13C HSQC
300 1H NMR (Lit)
907 13C NMR (Lit)
525 HSQC (Lit)
2021 MS (Lit)

http://mmcd.nmr.fam.wisc.edu/
MMCD

- Supports structure, name, NMR (shifts), MS (peaks) searches
- Data includes chemical formula, names and synonyms, structure, physical and chemical properties, NMR and MS data, NMR chemical shifts, species associations and extensive links to images, references, and other public databases

MS Spectral DBs

- NIST/AMDIS
- Metlin
- GolmDB
- MassBank
The Golm Database

- GC-MS (Quad and TOF) database
- Contains MSRI (MS + retention index) or MST data for 1450 identified metabolites
- Includes 10,336 spectra linked to analytes
- Downloadable libraries compatible with NIST08 and AMDIS software
- Primary focus on plant metabolites
- Supports compound name and MS queries
- MS submissions via NIST08 or AMDIS format

Golm Database

http://gmd.mpimp-golm.mpg.de/
Searching The Golm Database

This page facilitates the search of metabolites within the Golm by means of user-submitted MS-IE spectra consisting of retention index (R values) and mass/charge ratios. In addition, a functional group prediction will help to characterize those metabolites without available reference mass spectra included in the Golm so far. Instead, the unknown metabolite is characterized by predicted presence or absence of functional groups. For more uses this functionality presented here is exposed as 'Use this web service'.

We kindly ask users to cite the following paper when publishing results derived from this service:
LC-MS Spectral DBs

- MoNA – 236,604 spectra, 69,946 cmpds** (12,000)
- METLIN – 68,124 spectra, 13,048 cmpds
- mzCloud – 422,349 spectra, 2975 cmpds
- NIST14 MS/MS – 234,284 spectra, 9344 cmpds
- MassBank – 28,185 spectra, 11,500 cmpds
- Wiley LC-MS^n – >10,000 spectra, 4500 poisons
- ReSpect – 9107 spectra, 3595 cmpds
- GNPS – 9000 spectra, 4200 natural products

Total #compounds with exp. MS/MS spectra ~20,000

Metlin Database

- LC-MS database maintained at the Scripps Center for Metabolomics
- Currently lists 240,588 metabolites
- 65,776 high resolution MS/MS spectra
- Metlin has 12,834 compounds with high resolution MS/MS spectra (but about 8000 of these are peptides)
- 4800 MS/MS spectra of non-peptide metabolites
Metlin MS Search

Step 1: Enter Mass
Step 2: Select Charge
Step 3: Select “all”
Step 4: “Find Metabolites”

http://metlin.scripps.edu/metabo_search_alt2.php

Metlin Results
Metlin MS/MS Search

- mzXML
- mzML
- mzData

http://metlin.scripps.edu/upload.php

MassBank

- Very nicely maintained and easily searchable collection of mostly metabolite MS spectra
- Includes ESI-QTOF, ESI-QqQ, GC-EI-TOF, EI, ESI-FTICR, Ion-trap, etc.
- Covers 41,092 MS spectra from approximately 15,300 compounds
- Archives data from ~20 different sources (Japan, Germany, US, etc)
MassBank

Peak Search (MassBank)
Compound DBs

ChEBI

ChEBI is pronounced “KEBEE”

Chemical Entities of Biological Interest

Contains 44,263 “3 star” compounds

Most compounds are from KEGG, LipidMaps, DrugBank, Patents

Most data is on names, ontology, synonyms, MW, formula and structure

Searchable by name, formula, structure

PubChem

ChemSpider

Ligand Expo
PubChem

- NIH database of 32 million compounds and 75 million substances, 5000+ HT screens of compounds
- Compound must have <1000 atoms
- 80 Database vendors/depositors
- Substances are “impure/duplicates” compounds are single entities (CID is gives a unique compound)
- Entries include synonyms, chemical properties, chemical structure including SMILES and InChI strings, bioactivity, and links to structurally related compounds and other NCBI databases like PubMed
- Searchable by name, formula, MW range, structure, hbond donor/acceptor count, XlogP

ChemSpider

- Contains 28 million compounds from 400 data sources
- Searchable by name, synonym, InChI, structure, registry #, SMILES, calculated properties (but not by formula or mass)
- Data includes names, synonyms, wikipedia articles, descriptions, data sources, suppliers, patents, articles, properties, MESH headings, pharmacology links, spectra (UV, IR, NMR, MS) sourced from other sites
Ligand Expo

- Contains the small molecules in the PDB
- Useful because it links chemicals/metabolites/drugs to their targets
- Also provides 3D structure coordinates
- Searchable via 3-letter chemical identifier code, molecular name, molecular formula, SMILES description, InChi, 3D structure, MOL/SDF sketch

Other Compound DBs

- 3DMet
- KNAPSAcK
- MyCompoundID
- LipidMaps
Other DBs

- **3DMet**
  - 3D structure database of 8581 natural metabolites

- **KNApSAcK**
  - Database of 50,000 plant metabolites linked to species information

- **MyCompoundID**
  - Database of 11 million metabolically transformed metabolites

- **LipidMaps**
  - Database with 30,000 lipids (Fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, sterols, prenols, saccharolipids, polyketides)

MyCompoundID

- Database of predicted MS (MW) data for “metabolized metabolites”
- 76 metabolic transformations, modifications or fragmentations
- 8021 starting metabolites
- 375,809 MWs for first pass metabolism
- 10,584,000 MWs for second pass metabolism
MyCompoundID

http://mycompoundid.org

Module 3 bioinformatics.ca

Possible reactions

http://mycompoundid.org

Module 3 bioinformatics.ca
MyCompoundID Output

<table>
<thead>
<tr>
<th>#</th>
<th>Name ID</th>
<th>Common Name</th>
<th>Exact Mass</th>
<th>Formula</th>
<th>Charge State</th>
<th>PPM Error</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>33575506</td>
<td>2,4-Dihydroxybenzoic acid</td>
<td>192.09580</td>
<td>C_{7}H_{6}O_{3}</td>
<td>1+</td>
<td>0.22</td>
<td>4.798E-04</td>
</tr>
<tr>
<td>2</td>
<td>33575507</td>
<td>2,5-Dihydroxybenzoic acid</td>
<td>192.09580</td>
<td>C_{7}H_{6}O_{3}</td>
<td>1+</td>
<td>0.22</td>
<td>4.798E-04</td>
</tr>
<tr>
<td>3</td>
<td>33575510</td>
<td>Phloroglucinol</td>
<td>192.09580</td>
<td>C_{7}H_{6}O_{3}</td>
<td>1+</td>
<td>0.22</td>
<td>4.798E-04</td>
</tr>
<tr>
<td>4</td>
<td>33575511</td>
<td>Nicotine</td>
<td>207.133580</td>
<td>C_{8}H_{12}N_{2}</td>
<td>1+</td>
<td>0.22</td>
<td>4.798E-04</td>
</tr>
</tbody>
</table>

Pathway DBs

- KEGG (www.genome.jp/kegg/)
- SMPDB (www.smpdb.ca)
- BioCyc/MetaCyc
- Reactome (www.reactome.org)
Pathway DBs

- Rich source of biological data that relates metabolites to genes, proteins, diseases, signaling events and processes
- Provide various tools to permit visualization and gene/metabolite mapping
- Often cover multiple species

KEGG – Kyoto Encyclopedia of Genes and Genomes

http://www.genome.jp/kegg/
KEGG

http://www.genome.jp/kegg/compound/

The Small Molecule Pathway Database (SMPDB)

http://www.smpdb.ca
SMPDB

- Nearly 900 hand-drawn small molecule pathways
  - 384 drug pathways
  - 232 disease pathways
  - 220 metabolic pathways
  - 40+ other pathways
- Depicts cell compartments, organelles, protein locations, 4° structures
- Maps gene chip & metabolomic data
- Converts gene, protein or chemical lists to pathways or disease diagnoses
Exploring Pathways with SMPDB

Mapping Metabolites with SMPDB
Mapping Metabolite/Gene Concentrations with SMPDB

PathWhiz

- Webserver designed to permit creation of colourful, biologically accurate pathway diagrams that are machine readable and interactive
- Supports BioPAX, SBML and SBGN conversion as well as SVG and PNG image generation
- Google Maps-style viewer

http://smpdb.ca/pathwhiz
Building Pathways with PathWhiz

In the pathway drawing view, components can be rendered and joined together to illustrate different biological processes.

Different types of components to render

Options to customize components

Rendered components are draggable and snap to other components
Component visualizations include generic and specialized compounds, drugs, nucleic acids, enzymes, receptors, transporters, organs, organelles, membranes, and interaction arrow types.
Comprehensive MetDBs

- Must contain >1000 metabolites
- Usually are organism specific
- Continuously updated
- Must contain
  - chemical + pathway data
  - or chemical + spectral + biological data
  - or chemical + pathway + spectral data
  - or chemical + pathway + spectral + biological data
MetaboLights

- The “GenBank” for Metabolomics
- Operated by the EBI
- Supports data uploads of metabolomics experiments (spectra, compounds, lists, etc.)
- Has useful metabolomic data for searching, querying and download (linked to ChEBI)
- Complies with MSI
- Deposit your data here

UofA Metabolomics Databases

- www.hmdb.ca
- www.drugbank.ca
- www.ymdb.ca
- www.phenol-explorer.eu
- www.ecmdb.ca
- www.foodb.ca
- www.cowmetdb.ca
- www.t3db.ca
- www.smpdb.ca
- www.csfmetabolome.ca
- www.serummetabolome.ca
- www.urinemetabolome.ca
The Human Metabolome Project

- $7.5 million Genome Canada Project launched in Jan. 2005 - Based at the University of Alberta
- Mandate to quantify and identify all metabolites in biofluids such as urine, CSF and blood as well as tissues using HT experiments and text analysis
- Associate metabolite concentrations to human diseases or conditions
- Make all data freely and electronically accessible (HMDB, DrugBank, FooDB, T3DB)
- Develop novel technologies and software to improve metabolome coverage and metabolomic throughput

History of the Human Metabolome

- 2004 – 690 known human metabolites listed in KEGG & HumanCyc
- 2006 – First release of human metabolome database (HMDB) contains 2180 metabolites
- 2009 – HMDB 2.0 lists 6408 metabolites
- 2013 – 37,170 metabolites in HMDB 3.0
- 2016 – 41,993 metabolites in HMDB today
- 20?? – >100,000 metabolites thought to be detectable in the human body
Human Metabolomes (2015)

- 3670 (T3DB) - Toxins/Env. Chemicals
- 1240 (DrugBank) - Drug metabolites
- 28500 (FooDB) - Food additives/Phytochemicals
- 1550 (DrugBank) - Drugs
- 19700 (HMDB) - Endogenous metabolites

Meet the Metabolomes...

- http://www.hmdb.ca
- http://www.T3DB.ca
- http://www.foodb.ca
- http://www.drugbank.ca
HMDB Features/Content

- 41,993 metabolites
- 170 bacterial (gut microbe) metabolites
- Normal/abnormal concentrations
- 700+ disease links
- 2200 NMR spectra
- 7600 MS spectra
- 2000 GC-MS spectra

- Sequence search tools
- Spectral search tools
- Extensive browsing tools
- Pathway search tools
- Structure searches
- Biofluid browsing
- Text search tools
- Full data downloads

Inside the HMDB
Inside the HMDB

102 data fields

MS Spectral Searching
Module 3 bioinformatics.ca

MS/MS Spectral Searching

Module 3 bioinformatics.ca

NMR Spectral Searching
The HMDB Biofluids Database

- Reference metabolite concentrations for >650 different diseases & conditions
- Abnormal and normal metabolite concentrations for >15 biofluids and >5000 different metabolites
- Designed for clinical chemists & physicians
- Largest & most complete resource of its kind
The Drug Database (DrugBank)

- Database about drugs, drug targets & mechanisms
- 1552 small molecule drugs
- Detailed ADMET, MOA and pharmacokinetic data
- >1000 drugs with metabolizing enzyme data
- >1200 drug metabolites
- >600 drugs with transporters
- >600 MS+NMR spectra
- >4200 unique drug targets
- >14,000 drug-drug interactions
- Supports sequence, spectral, structure and text searches as well as compound browsing
- Full data downloads

Inside DrugBank

http://www.drugbank.ca
DrugBank Query Tools

Category Browse

ChemQuery

Module 3 bioinformatics.ca

DrugBank Query Tools

SeqSearch

DataExtractor

Module 3 bioinformatics.ca
The Toxic Exposome Database (T3DB)

- Comprehensive data on toxic compounds (drugs, pesticides, herbicides, endocrine disruptors, drugs, solvents, PCBs, furans, carcinogens, etc.)
- Detailed mechanisms, binding constants, target info
- >3600 toxic compounds
- ~2100 toxic targets
- >15,800 gene-chemical links
- >1900 reference spectra
- Full data downloads

http://www.t3db.ca

The Food Constituent Database (FooDB)

- Database of 30,000+ compounds found in foods and their effects on flavour, aroma, colour and human health
- Average plant food contains >3000 different compounds
- Many times more sophisticated and more comprehensive than what you find on your cereal box
- You are what you eat...

http://www.foodb.ca
The Yeast Metabolome Database (YMDB)

- >2000 yeast metabolites from 39 different growth substrates
- >180 wine compounds
- 1104 protein/enzyme metabolite associations
- 3774 NMR or GC-MS or LC-MS reference spectra
- 916 reactions, 66 pathways
- Supports sequence, spectral, structure and text searches as well as compound browsing
- 78 data fields per compound

http://www.ymdb.ca

The E. coli Metabolome Database (ECMDB)

- 2717 E. coli metabolites
- 1573 genes (1205 enzymes, 299 transporters)
- 3145 chemical reactions
- 4300 references
- 125 pathways
- Supports sequence, structure & text searches as well as compound browsing
- 80 data fields per compound
- 4965 NMR and MS spectra
- Corrects many errors and erroneous entries in EcoCyc

http://www.ecmdb.ca
Database Comparison

<table>
<thead>
<tr>
<th>Database</th>
<th>Nomenclat.</th>
<th>Links + Refs</th>
<th>MS Spectra</th>
<th>NMR Spectra</th>
<th>Pathways</th>
<th>Structures</th>
<th>Descriptions</th>
<th>Chem Props</th>
<th>Physiol. data</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMDB</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>KEGG</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PubChem</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MMCD</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ChEBI</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SDBS</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Reactome</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Metlin</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cyc DBs</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Exercises - Options

- Finish the spectral identification project you started this afternoon
- Try identifying and/or annotating your metabolites using the databases we just described
- Explore the software tools and databases we just covered

*The best way of learning is by doing*