Workshop 4: Pathways & Statistics

5th International Metabolomics Society Conference
Aug. 30, 2009 Edmonton, AB
Today’s Agenda

• Statistics, Informatics and Pathway Databases in Metabolomics - D. Wishart
• Cheminformatics and Metabolism at the EBI - C. Steinbeck
• Reactome Knowledgebase of Human Pathways & Processes - R. Haw
• Metabolomic Data Analysis using MetaboAnalyst - J. Xia
• Malaria Infections: From LC/MS to Pathway Analysis - T. Sana
Slides (to be) Located At:

http://www.metabolomicsssociety.org  --> Resources --> Tutorials
Statistics, Informatics and Pathway Databases in Metabolomics: An Introduction

David Wishart, University of Alberta
5th International Metabolomics Society Conference
Aug. 30, 2009 Edmonton, AB
2 Routes to Metabolonomics

Quantitative (Targeted) Methods

Chemometric (Profiling) Methods

- hippurate
- fumarate
- urea
- allantoin
- creatinine
- TMAO
- citrate
- 2-oxoglutarate
- succinate
- water

- PC1
- PC2

- ANIT
- Control
- PAP
Profiling (Untargeted)

Sample Prep

Data Collection

Metabolite Identification

Data Reduction

ANIT

Control

PAP
Quantitative (Targeted)

Sample Prep

Biological Interpretation

Metabolite Identification & Quantification

Data Reduction

ANIT

PAP

Control

PC1

PC2
Key Informatics Challenges in Metabolomics

- Data integrity and quality
- Data alignment and normalization
- Data reduction and classification
- Assessment of significance
- Metabolite identification/quantification
- Pathway mapping and identification
- Biological interpretation
Data Integrity/Quality

• LC-MS and GC-MS have high number of false positive peaks
• Problems with adducts (LC), extra derivatization products (GC), isotopes, breakdown products (ionization issues), etc.
• Not usually a problem with NMR
• Check using replicates and adduct calculators

MZedDB http://maltese.dbs.aber.ac.uk:8888/hrmet/index.html
HMDB http://www.hmdb.ca/search/spectra?type=ms_search
Data/Spectral Alignment

- Important for LC-MS and GC-MS studies
- Not so important for NMR (pH variation)
- Many programs available (XCMS, ChromA, Mzmine)
- Most based on time warping algorithms

http://mzmine.sourceforge.net/
http://bibiserv.techfak.uni-bielefeld.de/chroma
http://metlin.scripps.edu/download/
Data Normalization/Scaling

- Can scale to sample or scale to feature
- Scaling to whole sample controls for dilution
- Normalize to integrated area, probabilistic quotient method, internal standard, sample specific (weight or volume of sample)
- Choice depends on sample & circumstances

Same or different?
Data Normalization/Scaling

- Can scale to sample or scale to feature
- Scaling to feature(s) helps manage outliers
- Several feature scaling options available: log transformation, auto-scaling, Pareto scaling, probabilistic quotient, and range scaling

MetaboAnalyst http://www.metaboanalyst.ca
Data Reduction

• Data filtering (remove solvent peaks, noise filtering, false positives, outlier removal -- needs justification)

• Data binning (simplifies spectrum or chromatogram for further analysis)

• Dimensional reduction or feature selection to reduce number of features or factors to consider (PCA)

• Clustering to find similarity
Binning (3000 pts to 14 bins)

$x_i, y_i$

$x = 232.1$ (AOC)

$y = 10$ (bin #)
Dimension Reduction & PCA

- **PCA – Principal Component Analysis**
- **Process that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components**
- **Reduces 1000’s of variables to 2-3 key features**
Principal Component Analysis

Hundreds of peaks → 2 components

PCA captures what should be visually detectable
If you can’t see it, PCA probably won’t help
Visualizing PCA

- PCA of a “bagel”
- One projection produces a weiner
- Another projection produces an “O”
- The “O” projection captures most of the variation and has the largest eigenvector (PC1)
- The weiner projection is PC2 and gives depth info
PCA - The Details

- PCA involves the calculation of the eigenvalue (singular value) decomposition of a data covariance matrix.
- PCA is an orthogonal linear transformation.
- PCA transforms data to a new coordinate system so that the greatest variance of the data comes to lie on the first coordinate (1st PC), the second greatest variance on the 2nd PC etc.

\[
\begin{align*}
\text{scores} & = \text{loadings} \times \text{data} \\
t_1 & = p_1x_1 + p_2x_2 + p_3x_3 + \ldots + p_nx_n
\end{align*}
\]
Visualizing PCA

- Airport data from USA
- 5000 “samples”
- $X_1$ - latitude
- $X_2$ - longitude
- $X_3$ - altitude
- What should you expect?

Data from Roy Goodacre (U of Manchester)
PCA is equivalent to K-means clustering
K-means Clustering

Rule: $\lambda_T = \lambda_{\text{centroid}} \pm 50 \text{ nm}$
Breaching the Data Barrier

Unsupervised Methods
- PCA
- K-means clustering
- Factor Analysis

Supervised Methods
- PLS-DA
- LDA
- PLS-Regression

Machine Learning
- Neural Networks
- Support Vector Machines
- Bayesian Belief Net
Data Analysis Progression

- **Unsupervised Methods**
  - PCA or cluster to see if natural clusters form or if data separates well
  - Data is “unlabeled” (no prior knowledge)

- **Supervised Methods/Machine Learning**
  - Data is labeled (prior knowledge)
  - Used to see if data can be classified
  - Helps separate less obvious clusters or features

- **Statistical Significance**
  - Supervised methods always generate clusters -- this can be very misleading
  - Check if clusters are real by label permutation
Testing Significance

PCA

Labelled data

PLS-DA/SVM

Permutated data

Separation score

PLS-DA/SVM
Note of Caution

• Supervised classification methods are powerful
  – Learn from experience
  – Generalize from previous examples
  – Perform pattern recognition

• Too many people skip the PCA or clustering steps and jump straight to supervised methods

• Some get great separation and think the job is done - *this is where the errors begin*…

• Too many don’t assess significance using permutation testing or n-fold cross validation

• If separation isn’t partially obvious by eye-balling your data, you may be treading on thin ice

Still confused? - Jeff Xia (MetaboAnalyst)
So you’ve got your key features identified…. What next?

Metabolite Identification
2 Routes to Metabolomics

Quantitative (Targeted) Methods

Chemometric (Profiling) Methods

- hippurate
- fumarate
- allantoin
- urea
- water
- TMAO
- creatinine
- taurine
- citrate
- 2-oxoglutarate
- succinate

- ANIT
- Control
- PAP
Metabolite Identification

- Two scenarios identification of “known unknowns” and “unknown unknowns”
- For “known unknowns” use spectral or metabolite libraries to ID and quantify via spectral deconvolution
- For “unknown unknowns” (truly novel) use computer-aided structure elucidation methods (CASE)

“...there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns -- the ones we don't know we don't know.”
Technology & Sensitivity

# Metabolites detected (Log_{10})

- Known unknowns
- Unknown unknowns

Sensitivity or LDL

- NMR
- GC-MS TOF
- GC-MS Quad
- LC-MS or DI-MS
Metabolite ID by Spectral Deconvolution (NMR)
Chenomx & AMIX

www.chenomx.com

www bruker-biospin.com
NMR Spectral DBs

SBDS (http://riodb01.ibase.aist.go.jp)

NMRShiftDB (www.ebi.ac.uk/nrshiftdb/)

HMDB (www.hmdb.ca)

BMRB (www.bmrb.wisc.edu)
NMR Compound ID - HMDB

Phenyllactate
Phenylpyruvate
Phenylacetic acid
Tropic acid
Benzyl alcohol
...

Peak list to HMDB

High scoring matches
Metabolite ID by GC/LC-MS

GC or LC total Ion chromatogram

Spectral deconvolution

Component MS spectra

MS database

match

match

match
MS Spectral DBs

NIST/AMDIS (http://chemdata.nist.gov)

Metlin (http://metlin.scripps.edu/)

HMDB (www.hmdb.ca)

MassBank (www.massbank.jp)
Mass Matching DBs

ChEBI (www.ebi.ac.uk/chebi/)


ChemSpider (www.chemspider.com)

HMDB (www.hmdb.ca)
MS Compound ID - HMDB

Phenyllactate
Phenylpyruvate
Atrolactic acid
Homovanillin
Coumaric acid
...

LC-MS Spectrum

Peak list to HMDB

High scoring matches
What about the “Unknown Unknowns”? 

CASE - Christoph Steinbeck
Top - Down CASE Methods

Known metabolites (20,000)

Match observed spectra to predicted spectra to ID

Predicted biotransformations (20,000 --> 200,000)

Predicted MS, MS/MS, NMR, GC-MS Spectra
Bottom-Up (Traditional) CASE

Known metabolite substructures or metabolite EI or CID fragments

Predicted (or known) MS, MS/MS, NMR, GC-MS fragment spectra

Match observed spectra to predicted spectra to ID

Neural Network or GA driven fragment assembly
Key Informatics Challenges in Metabolomics

• Data integrity and quality
• Data alignment and normalization
• Data reduction and classification
• Assessment of significance
• Metabolite identification/quantification
• Pathway mapping and identification
• Biological interpretation
Pathway DBs

KEGG (www.genome.jp/kegg/)

HMDB (www.hmdb.ca)

BioCyc/MetaCyc

Reactome (www.reactome.org)
The Small Molecule Pathway Database (SMPDB)

http://www.smpdb.ca
SMPDB

- 350 hand-drawn pathways relevant to human/mammalian metabolism
  - 155 drug pathways
  - 72 disease pathways
  - 12 signalling pathways
  - 70 standard metabolic pathways
- Searching and browsing capabilities
- Metabolite mapping capabilities
- Captures structure, organelle, cellular and organ information
Exploring Pathways with SMPDB
Mapping Metabolites with SMPDB
Mapping Metabolite Concentrations with SMPDB
Conclusion

- Statistics, Informatics and Pathway Databases in Metabolomics - D. Wishart
- Cheminformatics and Metabolism at the EBI - C. Steinbeck
- Reactome Knowledgebase of Human Pathways & Processes - R. Haw
- Metabolomic Data Analysis using MetaboAnalyst - J. Xia
- Malaria Infections: From LC/MS to Pathway Analysis - T. Sana