Metabolomic Data Processing & Statistical Analysis

Jianguo (Jeff) Xia
Dr. David Wishart Lab
University of Alberta, Canada
Outline

I. Overview of procedures for metabolomic studies

II. Introduction to different data processing & statistical methods

III. MetaboAnalyst – a web service for metabolomic data processing, analysis and annotation

IV. Conclusions & future directions
A data-centric overview of metabolomic studies

1. Data Collection
2. Data Processing
3. Data Analysis
4. Data Interpretation
Data collection

- Biological Samples → Spectra

**Separation Techniques**
- Gas Chromatography (GC)
- Liquid Chromatography (LC)
- Capillary Electrophoresis (CE)

**Detection Techniques**
- Nuclear Magnetic Resonance Spectroscopy (NMR)
- Mass Spectrometry (MS)

**Hyphenated Techniques**
- Gas Chromatography - Mass Spectrometry (GC-MS)
- Liquid Chromatography - Mass Spectrometry (LC-MS)
- Liquid Chromatography - Nuclear Magnetic Resonance (LC-NMR)
Data processing

- **Raw Spectra → Data Matrix**

<table>
<thead>
<tr>
<th>Quantitative</th>
<th>Chemometric</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Compound concentration data;</td>
<td>• Spectral bins (NMR, Direct injection–MS)</td>
</tr>
<tr>
<td>• Involving compound identification &amp; quantification;</td>
<td>• Peak lists (LC/GC – MS)</td>
</tr>
<tr>
<td>• Currently labor intensive with a lot of manual efforts</td>
<td>• Largely automated process</td>
</tr>
</tbody>
</table>
Data analysis

- Extract important features/patterns

**Exploratory Analysis**
- Data overview
- Outlier detection
- Grouping patterns

**Biomarker discovery**
- To identify metabolites that are significantly different between groups

**Classification**
- To build a model for the prediction of unlabeled new samples
Data interpretation

- Mainly a manual process
- Require domain expert knowledge
- Tools are coming:
  - Comprehensive metabolite databases
  - Network visualization
  - Pathway analysis

Features/patterns → biological knowledge
Data processing & normalization

1. Data Collection
2. Data Processing
3. Data Analysis
4. Data Interpretation
Data processing (I)

- Purposes:
  - To convert different metabolomic data into data matrices suitable for varieties of statistical analysis
  - Quality control
    - To check for inconsistencies
    - To deal with missing values
    - To remove noises
A data matrix with rows represent samples and columns represent features (concentrations/intensities/areas)

GC/LC-MS spectra
- Peak picking
- Peak alignment

Compound concentrations
- Nothing to do
Data normalization

- **Purposes:**
  - To remove systematic variation between experimental conditions unrelated to the biological differences (i.e. dilutions, mass)
    - Sample normalization (row-wise)
  - To bring variances of all features close to equal
    - Feature normalization (column-wise)
Sample normalization

- By sum or total peak area
- By a reference compound (i.e. creatinine, internal standard)
- By a reference sample
- By dry mass, volume, *etc*
Feature normalization

- Log transformation
- Scaling

<table>
<thead>
<tr>
<th>Method</th>
<th>Formula</th>
<th>Goal</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autoscaling</td>
<td>$\tilde{x}<em>{ij} = \frac{x</em>{ij} - \bar{x}_j}{s_i}$</td>
<td>Compare metabolites based on correlations</td>
<td>All metabolites become equally important</td>
<td>Inflation of the measurement errors</td>
</tr>
<tr>
<td>Range scaling</td>
<td>$\tilde{x}<em>{ij} = \frac{x</em>{ij} - \bar{x}<em>i}{(\bar{x}</em>{max} - \bar{x}_{min})}$</td>
<td>Compare metabolites relative to the biological response range</td>
<td>All metabolites become equally important. Scaling is related to biology</td>
<td>Inflation of the measurement errors and sensitive to outliers</td>
</tr>
<tr>
<td>Pareto scaling</td>
<td>$\tilde{x}<em>{ij} = \frac{x</em>{ij} - \bar{x}_j}{\sqrt{s_i}}$</td>
<td>Reduce the relative importance of large values, but keep data structure partially intact</td>
<td>Stays closer to the original measurement than autoscaling</td>
<td>Sensitive to large fold changes</td>
</tr>
</tbody>
</table>

Statistical Analysis

1. Data Collection
2. Data Processing
3. Data Analysis
4. Data Interpretation
# Data Analysis

## Univariate
- Fold change analysis,
- T-tests
- Volcano plots

## Chemometrics
- Principal component analysis (PCA)
- Partial least squares - discriminant analysis (PLS-DA)

## High-dimensional feature selection
- Significance analysis of microarrays (and metabolites) (SAM)
- Empirical Bayesian analysis of microarrays (and metabolites) (EBAM)

## Clustering
- Dendrogram & Heatmap
- K-means, Self Organizing Map (SOM)

## Classification
- Random Forests
- Support Vector Machine (SVM)
Volcano-plot

- Arrange features along dimensions of statistical (p-values from t-tests) and biological (fold changes);
- The assumption is that features with both statistical and biological significance are more likely to be true positive.
- Widely used in microarray and proteomics data analysis.
PLS-DA

- *De facto* standard for chemometric analysis
- A supervised method that uses multiple linear regression technique to find the direction of maximum covariance between a data set (X) and the class membership (Y)
- Extracted features are in the form of latent variables (LV)
PLS-DA for feature selection

Variable importance in projection or VIP score
- A weighted sum of squares of the PLS loadings. The weights are based on the amount of explained $Y$-variance in each dimension.

Based on the weighted sum of PLS-regression coefficients.
- The weights are a function of the reduction of the sums of squares across the number of PLS components.

The Fifth International Conference of Metabolomic Society
Over fitting problem

- PLS-DA tend to over fit data
  - It will try to separate classes even there is no real difference between them!

- Require more rigorous validation
  - For example, to use permutations to test the significance of class separations
1) Use the same data set with its class labels reassigned randomly.

2) Build a new model and measure its performance (B/W).

3) Repeat many times to estimate the distribution of the performance measure (not necessarily follows a normal distribution).

4) Compare the performance using the original label and the performance based on the randomly labeled data.
Multi-testing problem

- P-value appropriate to a single test situation is inappropriate to presenting evidence for a set of changed features.
  - Adjusting p-values
    - Bonferroni correction
    - Holm step-down procedure
  - Using false discovery rate (FDR)
    - A percentage indicating the expected false positives among all features predicted to be significant
    - More powerful, suitable for multiple testing
A well-established method widely used for identification of differentially expressed genes in microarray experiments is moderated $t$-tests. This method computes a statistic $d_j$ for each gene $j$, which measures the strength of the relationship between gene expression ($X$) and a response variable ($Y$). The significance of the expression of any gene is determined by repeated permutations of the data and non-parametric statistics. The SAM (Significance Analysis of Microarray) plot illustrates the observed $d(j)$ values against the expected $d(i)$ values. The cut-off values for the plot are determined by the significance level (0.05) and the False Discovery Rate (FDR) of 0.008.
Clustering

- Unsupervised learning
- Good for data overview
- Use some sort of distance measures to group samples
  - PCA
  - Heatmap & dendrogram
  - SOM & K-means
Classification

- Supervised learning
- Many traditional multivariate statistical methods are not suitable for high-dimensional data, particularly small sample size with large feature numbers
- New or improved methods, developed in the past decades for microarray data analysis
  - Support vector machine (SVM)
  - Random Forests
To develop a pipeline service for metabolomic studies
Microarray data analysis pipeline

Analytical methodology

Gene expression data set

Pre-processing

Gene selection

Marker genes

Build predictor

Cross-validation results

Model

Test data set

Test set prediction results

Pipeline representation and results

1. Heat map

Gene list significance

2. Cross-validation results

3. Test set prediction results

The Fifth International Conference of Metabolomic Society
A proposed pipeline for metabolomics studies

- Data clean-up and alignment
- Normalisation or IS correction
- Combination of duplicates
- Division into subsets
- 80% rule for missing values, imputation of missing values
- Matrix construction
- Biological question
- Mean-centering or scaling
- PLS-DA
- Cross validation
- Permutation test
- Statistical validation of model and confirmation of biomarkers
  - Possible biomarkers
  - Evaluation of possible biomarkers
  - Targeted approach
    - PLS-DA
    - Cross validation
    - Potential biomarkers
  - Artefacts
    - PLS-DA
    - Cross validation
- Data excluding possible biomarkers
  - PLS-DA
  - Cross validation
- Construction of non-informative models
- Validation of models
- Data analysis
MetaboAnalyst

-- A web service for high-throughput metabolomic data processing, analysis and annotation

-- Implementation of all the methods mentioned in the form of user-friendly web interfaces

-- www.metaboanalyst.ca
GC/LC-MS raw spectra → MS / NMR peak lists → MS / NMR spectra bins → Metabolite concentrations

Peak detection • Retention time correction
Peak alignment
Baseline filtering • Data integrity check • Missing value imputation

Data normalization
Row-wise normalization (4) • Column-wise normalization (4)

Data analysis
Univariate analysis (3) • Dimension reduction (2) • Feature selection (2) • Cluster analysis (4) • Classification (2)

Data annotation
Peak searching (3) • Pathway mapping

Data integrity check • Missing value imputation

Download
Processed data • PDF report • Images
Implementation features

MetaboAnalyst

Latest Java Server Faces (JSF) technology for web interface design

R (esp. Bioconductor packages) for backend statistical analysis & visualization

Using resources in HMDB for peak annotation, compound identification, as well as pathway mapping

Comprehensive analysis report generation & documentation
2.2 Principal Component Analysis (PCA)

PCA is an unsupervised method aiming to find the directions that best explain the variance in a data set (X) without referring to class labels (Y). The data are summarized into much fewer variables called *scores* which are weighted average of the original variables. The weighting profiles are called *loadings*. The PCA analysis is performed using the **prcomp** package. The calculation is based on singular value decomposition.

The R script `chemometrics.R` is required. Figure 6 is pairwise score plots providing an overview of the various separation patterns among the most significant PCs; Figure 7 is the scree plot showing the variances explained by the selected PCs; Figure 8 shows the 2-D score plot between selected PCs; Figure 9 shows the 3-D score plot between selected PCs; Figure 10 shows the loading plot between the selected PCs; Figure 11 shows the biplot between the selected PCs.

![Figure 6: Pairwise score plots between the selected PCs. The explained variance of each PC is shown in the corresponding diagonal cell.](image)
Some usage statistics

Over 1,200 visits since publication (~15 / day)
Current status

- Differential Analysis (Biomarker Identification)
- Class Prediction (Supervised learning)
- Class Discovery (Clustering)
- Pathway Analysis
Challenges & future directions

- Unbiased and comprehensive survey of metabolome
  - NMR only able to detect more abundant compound species (> 1 µmol)
  - MS are usually optimized to detect compounds of certain classes
- Systematic classification of compounds (ontology)
- More efficient pathway analysis & visualization
Acknowledgement

- Dr. David Wishart
- Dr. Nick Psychogios
- Nelson Young

- Alberta Ingenuity Fund (AIF)
- The Human Metabolome Project (HMP)
- University of Alberta