Metabolomic and transcriptomic data analysis of Bioplastic-producing *Arabidopsis* using R, exploRase and GGobi

Iowa State University

Suh-Yeon Choi, Michael Lawrence, Dianne Cook, Heike Hofmann, Lauralynn Kourtz, Kristi Snell, Basil J. Nikolau and Eve Syrkin Wurtele
Outline

- Introduction to bioplastic-producing plants
- Challenges in metabolomic data analysis
- Development of R based preprocessing tool for metabolomic data analysis
- Omics data analysis using exploRase
Goal

Collaboration with METABOLIX

Bioinformatics: What limits bioplastic production in plants?

Bioplastic-producing Arabidopsis

Metabolomics

Transcriptomics

Optimize the production of bioplastic in plants
Procedure

Control plants

Bioplastic Producing plants

Extract RNA – transcriptomic analysis

Extract Metabolites – metabolomic analysis
Metabolomics data acquisition

- Plant (Treatment)
  - Extract
  - Separation (GC/LC/CE)
  - Detection (UV, MS, NMR etc...)

Data processing

- Data matrix
- Analysis

Raw data
- Baseline corrections
- Peak detection
- Peak deconvolution
- Peak matching
- Retention time correction
- Fill in missings
- Componentization
- Quantification
- Identification

exploRase

chromatoplots
Image of the Raw Data

1 sample

m/z

scan (time)
Image of the Raw Data

Scan 2747 (29.103 min): 222215.D

Ion 143.00 (142.70 to 143.70): 222215.D
Goal of preprocessing of metabolomics data

- Identify components from peaks in intensity
- Label the components as specific metabolites

Data matrix

| Metabolites | WT plant1 | PHB plant1 | WT plant2 | ...
|-------------|-----------|------------|-----------|------
| malate      | 100       | 200        | 110       |      |
| citrate     | 50        | 25         | 60        |      |
Limitation of existing tools

- Larger number of samples used with underlying experimental design
  - Most software analyze the data one by one
- Larger number of peaks of interest
  - More than ~300 metabolites detected per run
- No unified method
  - Each software uses their own algorithms
  - No comprehensive software
  - Commercial software; cannot be shared by biologists
- Some bioinformatic tools have been developed (AMDIS, XCMS, MZMine, etc), but they are lacking
  - Limited diagnostics, especially interactive visualizations
  - Do not leverage experimental design
Features/goals of new tools

1. Automated data processing tool for large set of data (over hundreds samples.)
2. Have experimental design information in data processing
3. User inspection feature during processing (over replicates, etc…)
4. User friendly GUI wizard
Proposed pipeline

1. Input as NetCDF or mzXML
2. Dealing with noise and non-detects
3. Local maximum so far above a baseline or sigma in gaussian fit
4. Matching peaks across samples, useful for RT correction
5. Retention time and m/z calibration, peak area normalization, peak filling
6. Grouping different ion peaks within a sample.
7. Usually based on a model peak in a component
8. Matrix of metabolite levels in each sample
9. Quantification
10. Identification
11. Result
12. exploRase

These steps may occur recursively, helping each other converge.
Where is the Baseline?
Background correction – existing solution

• AMDIS
  – baseline from a linear regression on all points below the median in the fitting region
  – not robust to high signal
• XCMS
  – Baseline from the second derivative of the filter translates the signal to curvature
  – subtracting linear background
• MathDAMP
  – RBE (Robust Baseline Estimation), a loess smoother that is weighted (Tukey biweight function)
  – robust to outliers (peaks)
Background correction - Loess Baseline Subtraction

• Approach used in MathDAMP
• Fit loess model to the raw profile.
• Needs to be robust to avoid fitting the peaks.
• Iterate loess fits, weighting cases with positive residuals by the Tukey biweight function (Ruckstuhl et al., 2001).
Loess Baseline Fit

LD, pCam, induced, rep 1 at 51 m/z

Loess Baseline in Red

XCMS Baseline
After Baseline Subtraction

LD, pCam, induced, rep 1 at 51 m/z

Quantile Cutoff
Metabolomics data acquisition

Plant (Treatment) → Extract → Separation (GC/LC/CE) → Detection (UV, MS, NMR etc…) → Data processing

Data processing → Data matrix → Analysis

Raw data → Baseline corrections → Peak detection → Peak deconvolution → Peak matching → Retention time correction → Fill in missings → Componentization → Quantification → Identification
Peak Detection

- Peaks are local maxima above some cutoff and exceeding adjacent minima by some threshold.
- Cutoff is a global quantile of the residuals.
- The threshold is a multiple of the standard deviation of the (residual) intensities.
- Similar approach to AMDIS.
Considering the Peak Shape

- We expect a peak to have a gaussian shape, so we fit a gaussian function to the neighborhood around each maxima.
- Neighborhoods are not allowed to overlap.
- Fits with extremely large sigma are discarded.
- About 4000 peaks detected per sample.
Example Peak Fits

Singleton

Convolution
Slicing and Dicing for the Peaks
Metabolomics data acquisition

- Plant (Treatment)
- Extract
- Separation (GC/LC/CE)
- Detection (UV, MS, NMR etc...)
- Data processing
  - Raw data
    - Baseline corrections
    - Peak detection
    - Peak deconvolution
    - Peak matching
    - Retention time correction
    - Fill in missings
    - Componentization
    - Quantification
    - Identification
- Data matrix
- Analysis
Comparing Samples

- To compare, they need to be aligned.
  - The m/z is assumed to be relatively stable.
  - Retention time likely requires correction, due to instability of the column across runs.
- Peaks between replicates should be consistent.
Retention time correction – existing solutions

- **AMDIS** – RI based (not precise)
- **METIDEA** – AMDIS + selective ion matching
- **MetAlign** – selective ion matching + back and forth...
- **XCMS** – fitting by Gaussian density estimation function
Retention time correction

• Consider the peaks in the TIC (Total Ion Count) profile, the sum over m/z (Krebs et al., 2006).
• Greedily match by the pairwise correlation between spectral intensity vectors
• Fit robust loess to ignore outliers (mismatches).
• Visually explore results using rggobi.
RT Correction Results

Unaligned (Original) Profiles

Aligned Profiles
GUI: chromatoplot (baseline correction)

- Raw image
- Profile plot
- Residual plot

Option windows; User can select

Baseline from loess fit
Baseline corrected
Next Steps

• Deconvolution of the peaks
• Matching the peaks across data set
• Identify and quantify the metabolites
  – A scriptable implementation of the methods
  – A biologist-accessible GUI
  – Plenty of interactive graphics for diagnostics
  – Integration with Bioconductor (xcms, MassSpecWavelet)
<table>
<thead>
<tr>
<th>Metabolites</th>
<th>WT plant1</th>
<th>PHB plant1</th>
<th>WT plant2</th>
</tr>
</thead>
<tbody>
<tr>
<td>malate</td>
<td>100</td>
<td>200</td>
<td>110</td>
</tr>
<tr>
<td>citrate</td>
<td>50</td>
<td>25</td>
<td>60</td>
</tr>
</tbody>
</table>

exploRase: Omics data analysis tool

Metabolic Network
• R: http://www.r-project.org/
• RGtk2: http://www.ggobi.org/RGtk2/
• rggobi: http://www.ggobi.org/rggobi/
• ggobi: http://www.ggobi.org/
• exploRase: http://www.metnetdb.org/MetNet_exploRase.htm
• chromatoplots: not available yet
Acknowledgement

◆ Department of Statistics
  Prof. Dianne Cook
  Prof. Heike Hofmann
  Michael Lawrence
  Dr. Eun-Kyung Lee

◆ Metabolix. Inc.
  Dr. Lauralynn Kourtz
  Dr. Kristi Snell

◆ Department of Genetics, Developmental and Cell Biology
  Prof. Eve Wurtele
  Suh-Yeon Choi

◆ Department of Biochemistry, Biophysics and Molecular Biology
  Prof. Basil Nikolau
  Dr. Wenxu Zhou

◆ W.M. Keck Metabolomics Research Lab
  Dr. Ann Perera