Quantitative Analysis of Proteomics Mass Spectrometry Data

Sebastian Gibb and Korbinian Strimmer
IMISE, University of Leipzig

EMS 2013
Budapest
21 July 2013
Inhalt

1 Introduction
   Personalized Medicine
   Why Proteomics?
   Statistical Challenges

2 Preprocessing Mass Spectrometry Data
   MALDIquant Software
   Workflow
      Data import, smoothing, baseline correction, calibration, peak detection, peak alignment, peak binning, intensity matrix

3 High-Level Analysis
   Dichotomization
   Classification and Ranking with Binary Predictors
   Example from Cancer Proteomics

4 Conclusion and Outlook
I. Introduction
Predicting Disease
Personalized Medicine

Medication and tests approved for personalized medicine in Germany:

<table>
<thead>
<tr>
<th>Medikament</th>
<th>Krankheitsgebiet</th>
<th>Nebenwirkungen</th>
<th>Testauf Verordnen nach (Überempfindlichkeits-Test)</th>
<th>Konsequenz aus dem Test</th>
<th>Status</th>
<th>Quelle</th>
<th>Bemerkung</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aciclovir</td>
<td>HIV / Aids</td>
<td>Nebenwirkungen</td>
<td>Test auf Verordnen nach (Überempfindlichkeits-Test) (Aktuell: Erhöhtes Risiko für Überempfindlichkeit)</td>
<td>keine Anwendung bei positivem Test</td>
<td>Blut</td>
<td>Pille, Infektionsfall</td>
<td>positiv: Toleranz/Bindung bei ca. 5% aller Patienten; bei 48-61% dieser Patienten Überempfindlichkeitsreaktionen; negativ: Toleranz/Bindung auf die genannte Art erhöht.</td>
</tr>
<tr>
<td>Anastrazin</td>
<td>Onkologie / Anrechte Formen von Brustkrebs</td>
<td>Nebenwirkungen</td>
<td>Test auf Hormonrezeptor-positiver Brustkrebs</td>
<td>keine Anwendung bei positivem Test</td>
<td>Krebszellen</td>
<td>Pille, Infektionsfall</td>
<td>keine Anwendung bei positivem Test</td>
</tr>
<tr>
<td>Anastrozol</td>
<td>Onkologie / Anrechte Formen von Brustkrebs</td>
<td>Nebenwirkungen</td>
<td>Test auf Hormonrezeptor-positiver Brustkrebs</td>
<td>keine Anwendung bei positivem Test</td>
<td></td>
<td></td>
<td>keine Anwendung bei positivem Test</td>
</tr>
<tr>
<td>Arozuspin</td>
<td>Immununsupprimiert</td>
<td>Nebenwirkungen</td>
<td>Test auf Thrombin-Methyltransferase (TMM) - Mangel (nur bei endogen- oder exogen-(vermehrter) Hepatitis A/Alphavirus-Infektion)</td>
<td>keine Anwendung bei positivem Test</td>
<td>Blut</td>
<td>Pille</td>
<td>positiv: Testergebnis bei ca. 0,3% aller Patienten, 0% mit mittlerem Risiko, Test wird nur für diese Patienten empfohlen, da unterschiedliche Formen gegeben sein können</td>
</tr>
<tr>
<td>Carbamazepin</td>
<td>Epilepsie</td>
<td>Nebenwirkungen</td>
<td>Test auf Verordnen nach (Überempfindlichkeits-Test) (Aktuell: Erhöhtes Risiko für Nebenwirkungen)</td>
<td>keine Anwendung bei positivem Test</td>
<td>Blut</td>
<td>Pille, Infektionsfall</td>
<td>keine Anwendung bei positivem Test</td>
</tr>
<tr>
<td>Glucocorticoide</td>
<td>Onkologie / Anrechte Formen von Brustkrebs</td>
<td>Nebenwirkungen</td>
<td>Test auf nicht-exprimierte Wildtyp-KRAS-Gen</td>
<td>keine Anwendung bei positivem Test</td>
<td></td>
<td></td>
<td>keine Anwendung bei positivem Test</td>
</tr>
<tr>
<td>Prostazin</td>
<td>Onkologie / Anrechte Formen von Brustkrebs</td>
<td>Nebenwirkungen</td>
<td>Test auf nicht-exprimierte Wildtyp-KRAS-Gen</td>
<td>keine Anwendung bei positivem Test</td>
<td></td>
<td></td>
<td>keine Anwendung bei positivem Test</td>
</tr>
</tbody>
</table>

Biomarker Discovery

General aim of personalized medicine:

“the right drug for the right person”

Biomarkers are essential for:

- Classifying heterogeneous subtypes of disease
- Pinpoint precise diagnosis
- Targeted therapy
- Understanding the provenance of disease
Data from High-Throughput Platforms

Major technologies to collect genomic high-throughput data: Microarrays and Next-Generation Sequencing
Proteomics

Microarrays and NGS allow to measure concentration of RNA expressions → Transcriptomics

However, in many clinical settings (e.g. clinical diagnostics) one is often interested in expression of proteins, peptides and amino acids.

Advantage: direct biological and medical interpretation!

Systematic study of proteins and related products → Proteomics
Proteomic Mass Spectrometry

High-throughput technology for measuring proteins: Mass Spectrometry
MALDI Technology

MALDI: Matrix-Assisted Laser Desorption/Ionization

Allows the analysis of large organic molecules such as proteins. (note “matrix” is biochemical jargon referring to carrier material that helps the protein ionization).
MALDI-TOF Principle

**Ion Source:** MALDI

**Matrix-Assisted Laser Desorption/Ionization**

**Mass Analyzer:** TOF

**Time Of Flight** ($t \propto \sqrt{\frac{m}{q}}$)

**Detector**

**Quantity Measurement**

Abb. 3.14.; S. 67; "Biochemie & Pathobiochemie", Löfler G., 8. Auflage (2007), Springer Medizin Verlag
MALDI Benefits

Advantages:

- Reliable well established technology (many variants exists)
- Often combined with other experimental techniques (gels, imaging etc).
- Very cheap to run compared with other high-throughput technologies (i.e. many replications possible)
Statistical Challenges

Analysis of proteomics mass spectrometric data is more complicated than that of gene expression!

Overview of the major statistical challenges summarized by Morris et al (2007):

Preprocessing:
- Removal of systematic biases in the data
- Peak identification
- Peak alignment across spectra
- Quantification and calibration of relative peak intensities

Develop suitable methods for multivariate analysis:
- Differential protein expression
- Classification and prediction
- Feature selection

In addition, many considerations from transcriptomics also apply here (such high dimensionality, sparse models etc).
II. Preprocessing Mass Spectrometry Data
MALDIquant Software

In the form of MALDIquant we have developed a complete processing pipeline for proteomics data in R.

Motivation:

- Only relatively few open source software solutions available and very few for the R platform.
  
  http://strimmerlab.org/notes/mass-spectrometry.html

- No MALDI-TOF package fitting our needs for clinical diagnostics.

- Necessity of handling both technical and biological replicates.

- Nonlinear Peak alignment necessary for many spectra.

- Modular and easy to customize analysis routines.

- Testbed for studying new methodologies (e.g. for quantification).
## MALDIquant Family

### MALDIquant

Complete analysis pipeline for MALDI-TOF and other 2D mass spectrometry data.

### MALDIquantForeign

- Import raw data (txt, tab, csv, Bruker Daltonics fid, Ciphergen XML, mzXML, mzML).
- Export into common formats (txt, tab, csv, msd, mzML).
MALDIquant Family

MALDIquant

- single spectrum
  - smoothing
  - baseline correction
  - peak detection

- multiple spectra
  - peak alignment/peak binning
  - calibration/normalization

MALDIquantForeign

- raw data
  - fid
  - mzML
  - mzXML
  - txt, tab
  - csv

classification

- sda
- randomForest
- pamr
- ...

Sebastian Gibb and Korbinian Strimmer, Analysis of Proteomics Data, 21/7/2013
MALDIquant automatically recognizes many mass spectrometry data formats (including native Bruker files).

```r
## load MALDIquant
library("MALDIquant")
## load MALDIquantForeign
library("MALDIquantForeign")
spectra <- import("/data/ms/raw")
```
Here is a raw example spectrum from the Fiedler et al. (2009) cancer data set.

```r
plot(spectra[[1]])
```
Variance Stabilizing/Smoothing

```
spectra <- transformIntensity(spectra, sqrt)
plot(spectra[[1]])
```

---

Sebastian Gibb and Korbinian Strimmer, Analysis of Proteomics Data, 21/7/2013
Variance Stabilizing/Smoothing

Raw Data
- Data Import
- Smoothing
- Baseline Correction
- Intensity Calibration
- Peak Detection
- Peak Alignment
- Peak Binning
- Intensity Matrix
- Post Processing
- Result

```
movingAverage <- function(y) {
  return(filter(y, rep(1, 5)/5, sides = 2))
}
spectra <- transformIntensity(spectra, movingAverage)
```

![Mass Spectrum](image)
For correct quantification of peak intensities it is necessary to conduct baseline correction. This accounts for systematic bias, such as matrix effects.

MALDIquant implements a number of correction algorithms, including the SNIP approach by Ryan et al (1988) and the TopHat filter.
Baseline Correction

- Raw Data
- Data Import
- Smoothing
- Baseline Correction
- Intensity Calibration
- Peak Detection
- Peak Alignment
- Peak Binning
- Intensity Matrix
- Post Processing
- Result

Median Baseline

/data/set A - discovery leipzig/control/Pankreas_HB_L_061019_G10/0_m19/1/1SLin/fid

mass
intensity

Sebastian Gibb and Korbinian Strimmer, Analysis of Proteomics Data, 21/7/2013
Baseline Correction

1. Raw Data
2. Data Import
3. Smoothing
4. Baseline Correction
5. Intensity Calibration
6. Peak Detection
7. Peak Alignment
8. Peak Binning
9. Intensity Matrix
10. Post Processing
11. Result

Median Baseline
/data/set A - discovery leipzig/control/Pankreas_HB_L_061019_G10/0_m19/1/1SLin/fid

Mass vs. Intensity Graph

Sebastian Gibb and Korbinian Strimmer, Analysis of Proteomics Data, 21/7/2013
Baseline Correction

- Raw Data
- Data Import
- Smoothing
- Baseline Correction
- Intensity Calibration
- Peak Detection
- Peak Alignment
- Peak Binning
- Intensity Matrix
- Post Processing
- Result

Convex Hull Baseline

/data/set A – discovery leipzig/control/Pankreas_HB_L_061019_G10/0_m19/1/1SLin/fid

Sebastian Gibb and Korbinian Strimmer, Analysis of Proteomics Data, 21/7/2013
Baseline Correction

Raw Data → Data Import → Smoothing → Baseline Correction → Intensity Calibration → Peak Detection → Peak Alignment → Peak Binning → Intensity Matrix → Post Processing → Result

TopHat Baseline

mass

intensity

/data/set A - discovery leipzig/control/Pankreas_HB_L_061019_G10/0_m19/1/1SLin/fid
Baseline Correction

Raw Data → Data Import → Smoothing → Baseline Correction → Intensity Calibration → Peak Detection → Peak Alignment → Peak Binning → Intensity Matrix → Post Processing → Result

Baseline Correction

Raw Data → Data Import → Smoothing → Baseline Correction → Intensity Calibration → Peak Detection → Peak Alignment → Peak Binning → Intensity Matrix → Post Processing → Result

```r
spectra <- removeBaseline(spectra, method = "SNIP")
plot(spectra[[1]])
```

As with gene expression data, mass spectrometry data need to be calibrated. Several methods are available (TIC, Median, probabilistic quotient normalization).

```r
spectra <- standardizeTotalIonCurrent(spectra)
plot(spectra[[1]])
```

![Graph showing mass vs. intensity for Pankreas_HB_L_061019_G10.M19]
Peak Detection

Raw Data

Data Import

Smoothing

Baseline Correction

Intensity Calibration

Peak Detection

Peak Alignment

Peak Binning

Intensity Matrix

Post Processing

Result

Pankreas_HB_L_061019_G10.M19

Intensity Matrix

accepted maxima
rejected maxima
noise threshold
Peak Detection

- Raw Data
- Data Import
- Smoothing
- Baseline Correction
- Intensity Calibration
- Peak Detection
- Peak Alignment
- Peak Binning
- Intensity Matrix
- Post Processing
- Result

```r
peaks <- detectPeaks(spectra, SNR = 3)
plot(spectra[[1]])
points(peaks[[1]])
```

![Graph showing peak detection and intensity analysis](image-url)
Peak alignment across multiple spectra is a very challenging issue in mass spectrometry analysis:

- there are many peaks that occur only in few spectra
- non-linear mass shifts requires calibration along x-axis
- peak mapping needed for comparison and identification of markers
Peak Alignment/The Problem

Raw Data
- Data Import
- Smoothing
- Baseline Correction
- Intensity Calibration
- Peak Detection
- Peak Alignment
- Peak Binning
- Intensity Matrix
- Post Processing
- Result
Peak Alignment/Possible Solutions

- Raw Data
- Data Import
- Smoothing
- Baseline Correction
- Intensity Calibration
- Peak Detection
- Peak Alignment
- Peak Binning
- Intensity Matrix
- Post Processing
- Result

**DTW**
[Clifford et al. 2009]
very slow & high memory usage
$m_1 \times m_2$: \((5 \times 10^4)^2 \times 48\text{Bytes} \approx 120\text{GB})$

**COW**
[Veselkov et al. 2009]
slow, must run segment-wise to overcome nonlinear shift

**PTW**
[Bloemberg et al. 2010]
fast (\(\approx 3\text{ sec/sample on an i5@2.5 GHz, 4 GB RAM}\))

...
Peak Alignment/Our Approach

inspired by [Wang et al. 2010]
reference peaks: landmark peaks ⇒ occur in most spectra

inspired by [He et al. 2011]

- peak matching (choose highest peak in range)
- mass vs diff plot → estimate warping function

Bing Wang, Aiqin Fang, John Heim, Bogdan Bogdanov, Scott Pugh, Mark Libardoni, and Xiang Zhang. Disco: Distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics. *Analytical Chemistry*, 82(12):5069–5081, 2010

Peak Alignment/Warping Functions

warpingFunctions <- determineWarpingFunctions(peaks)
Peak Alignment/Warping Functions

spectra <- warpMassSpectra(spectra, warpingFunctions)
peaks <- warpMassPeaks(peaks, warpingFunctions)
Peak Alignment/Comparison

Raw Data
- Data Import
- Smoothing
- Baseline Correction
- Intensity Calibration
- Peak Detection
- Peak Alignment
- Peak Binning
- Intensity Matrix
- Post Processing
- Result

unwarped
- mass 4180 4190 4200 4210 4220 4230 4240
- ptw (Bloemberg et al 2010)
- MALDIquant

unwarped
- mass 9200 9250 9300 9350 9400
- ptw (Bloemberg et al 2010)
- MALDIquant

Sebastian Gibb and Korbinian Strimmer, Analysis of Proteomics Data, 21/7/2013
Peak Binning
Peak Binning

peaks <- binPeaks(peaks)

Sebastian Gibb and Korbinian Strimmer, Analysis of Proteomics Data, 21/7/2013
```r
featureMatrix <- intensityMatrix(peaks)
featureMatrix[1:12, 1:2]
```

```
## 1011.67823313295 1020.66971843048
## [1,] 3.587257e-05 0.0002467926
## [2,] 3.170674e-05 0.0002550081
## [3,] 3.517940e-05 0.0002428846
## [4,] 3.430174e-05 0.0002563571
## [5,] 3.122426e-05 0.0004472052
## [6,] NA 0.0004505502
## [7,] 2.680547e-05 0.0004240763
## [8,] 3.484823e-05 0.0003559640
## [9,] 4.327525e-05 0.0001619205
## [10,] 3.357397e-05 0.0001527801
## [11,] 4.160095e-05 0.0002912183
## [12,] 3.848561e-05 0.0002911327
```
## load libraries

```r
library("MALDIquant")
library("MALDIquantForeign")
```

## load data

```r
spectra <- import("/data/ms/raw")
```

## run spectrum based workflow

```r
spectra <- transformIntensity(spectra, sqrt)
spectra <- transformIntensity(spectra, movingAverage)
spectra <- removeBaseline(spectra)
spectra <- standardizeTotalIonCurrent(spectra)
peaks <- detectPeaks(spectra)
```

## run peak based workflow

```r
warpingFunctions <- determineWarpingFunctions(peaks)
peaks <- warpMassPeaks(peaks)
peaks <- binPeaks(peaks)
featureMatrix <- intensityMatrix(peaks)
```

## use featureMatrix for further analysis
MALDIquant GUI

MALDIquant - Workflow

Baseline Correction
Baseline Correction:
SNIP

halfWindowSize:
1 100 500

☑ Show Uncorrected Spectrum
☑ Show Corrected Spectrum

Zoom:
Mass Range:
1,000 10,000

Intensity Range:
0 111,862

Plot:
Plot Spectrum:
Pankreas_HB_L_061019_G1
Pankreas_HB_L_061019_G10
Pankreas_HB_L_061019_H7
Pankreas_HB_L_061019_H7

Pankreas_HB_L_061019_G10_M19

mass

intensity

0 50 100 150 200 250 300

2000 4000 6000 8000 10000
III. High-Level Analysis
Multivariate Analysis of Omics Data

In the last decade a multitude of statistical methods have been developed for high-dimensional gene expression data. For example:

- regularized $t$-scores (moderated $t$, shrinkage $t$) for gene ranking,
- regularized regression and classification (e.g. shrinkage discriminant analysis),
- sparse models, structured models, latent class models, and
- high-dimensional testing (e.g. Higher Criticism or FDR)

Can we use these techniques also for mass spectrometry data?
Peaks as Biomarkers

Mass spectrometric peaks contain both binary and continuous information:

1. peak may be present or absent (NA in intensity matrix)
2. intensity of peak may be up or down regulated

Both properties need to be taken into account!
Thus, usual methods from gene expression data are generally not appropriate! (But nonetheless used in practice.)
Solution: Dichotomization

Local peak thresholding (Tibshirani et al 2004) uses a peak-specific thresholding rule $I_{\text{thresh}}(m)$ to dichotomize spectral data:

1. peak intensity $I(m) > I_{\text{thresh}}(m) \rightarrow 1$
2. peak intensity $I(m) \leq I_{\text{thresh}}(m) \rightarrow 0$

This allows to take account both of absent/present as well as up/down regulated peaks.

The thresholding rule is estimated from the data by maximizing the separation between the groups (as measured by a variable importance criterion).

→ for mass spectrometry data we need categorical data analysis!
Classification and Ranking with Binary Predictors

Large-scale analysis of binary data is routine in machine learning, especially in text mining.

We suggest using similar techniques as in text mining for the analysis of mass spectrometry data, in particular:

- multivariate Bernoulli and related models for classification,
- corresponding variable importance measures (typically based on KL entropy) for peak ranking, and
- regularized inference for application in high-dimensional settings.
Multivariate Bernoulli Independence Rule

One of the most popular approaches and highly effective approach for classification of gene expression data is PAM (Tibshirani et al 2003), a variant of shrinkage diagonal discriminant analysis.

We use the same idea for mass spectrometry data:

- assume “diagonal” multivariate Bernoulli distributions for each group, $\mathbf{X}_k = (X_1, \ldots, X_d) \sim B_d(\mu_k)$ with $\Pr(x_i | \mu_i, k) = \prod_{i=1}^{d} \Pr(x_i | \mu_i, k)$

- Bayes rule gives discriminant function $d_k \propto \log \Pr(k | x)$.

- Regularized training of discriminant rule in situations with $d \geq n$.

This MVB independence rule, while very simply, has shown to be highly effective (e.g. Park 2009).
Pancreatic Cancer Proteomics Study


- Large study conducted in Leipzig and Heidelberg
- 120 participants (60 healthy vs. 60 cancer)
- 4 technical replicates per sample
- 480 MALDI spectra
Preprocessing and Peak Filtering

- Number of detected peaks per spectrum (of 480): between 134 and 257
- After merging of technical replicates, keeping only peaks that occur in all 4 measured spectra: between 52 and 146
- Keeping only peaks that occur with 75 % frequency in each group (cancer and control): between 23 and 56
Comparison of Rankings

Shrinkage $t$ vs. categorical ranking (top 10 peaks):

<table>
<thead>
<tr>
<th>Ranking</th>
<th>metric</th>
<th>discrete</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4494.71</td>
<td>4494.71</td>
</tr>
<tr>
<td>2</td>
<td>2755.41</td>
<td>8937.02</td>
</tr>
<tr>
<td>3</td>
<td>4250.76</td>
<td>4467.81</td>
</tr>
<tr>
<td>4</td>
<td>8131.39</td>
<td>5945.42</td>
</tr>
<tr>
<td>5</td>
<td>2022.73</td>
<td>2022.73</td>
</tr>
<tr>
<td>6</td>
<td>1627.86</td>
<td>1866.07</td>
</tr>
<tr>
<td>7</td>
<td>3920.23</td>
<td>4250.76</td>
</tr>
<tr>
<td>8</td>
<td>8144.25</td>
<td>2755.41</td>
</tr>
<tr>
<td>9</td>
<td>5945.42</td>
<td>5906.06</td>
</tr>
<tr>
<td>10</td>
<td>5266.03</td>
<td>2953.13</td>
</tr>
</tbody>
</table>
Clustering of Dichotomized Data

Dichotomized data exhibits near perfect split in healthy and cancer patients!
IV. Conclusion and Outlook
We have developed MALDIquant, a comprehensive R package for analysis of mass spectrometry data, with focus on clinical diagnostics.

MALDIquant is easy to use yet incorporates state-of-the art methods for preprocessing, calibration, quantification and visualization.

Multivariate analysis is best done on dichotomized peak data, using regularized categorical data analysis.
Outlook: IMS

Imaging Mass Spectrometry (IMS): combining MS data with spatial information

MALDIquant - IMS example
Many Thanks for Your Interest!
Availability

http://strimmerlab.org/software/maldiquant/

```r
## get newest MALDIquant/MALDIquantForeign directly from CRAN
install.packages(c("MALDIquant", "MALDIquantForeign"))
```