Analysis of mass spectrometry protein data

Jenny Barrett
Section of Epidemiology and Biostatistics
Leeds Institute of Molecular Medicine
University of Leeds, UK
Outline

• Introduction
  • Proteomics, what and why?
  • Mass spectrometry

• Statistical issues
  • Experimental design
  • Preprocessing
    • calibration, baseline subtraction, normalisation
  • Peak detection and alignment
  • Multivariate exploratory data analysis
  • Statistical models to detect differences between groups
  • Classification of samples

• Concluding remarks
What is proteomics?

**Protein** - chain of amino acids in 3D structure which carry out biological functions, made from DNA via mRNA

**Proteome** - the total protein complement expressed by a cell

**Proteomics** - the study of proteomes

**Proteins are:**
- cell-specific
- time-specific
Why Proteomics?

Epigenetic factors
Feedback loops

Transcriptional control
Post-transcriptional Control e.g. alternative splicing
Translational controls
Activity controls

Closer to the disease process
Proteomics technology

• Mass spectrometry (MS) can be used to characterise the proteome when applied to the appropriate biological sample (plasma, serum, . . .).

• Many experimental pre-processing steps applied prior to MS
  • depletion of abundant proteins, fractionation, digestion, separation . . .

• Various types of MS analyzer – our focus is on MS based on time-of-flight
Mass spectrometry

- Molecules (including peptides and proteins) have different masses

- MS separates ionised molecules by their mass-to-charge (m/z) ratio

- Time-of-flight (TOF), which can be measured, differs according to m/z - can then be converted to m/z scale

- Typical range considered 2000 to 50000 Daltons (<2000 Da too much noise)
Mass Spectrometer

Accelerating Potential

Sample Chip

Grids

Detector

Acceleration phase
Example MS analysers

• **MALDI (Matrix Assisted Laser Desorption-Ionization)**
  - Apply sample to chip
  - Uses energy-absorbing matrix which facilitates ionisation without much fragmentation

• **SELDI (Surface Enhanced . . .)**
  - Like MALDI but chips are pre-coated with special surface
  - Different proteins bind to surface depending on chip type

• **iTRAQ (Isobaric Tags for Relative and Absolute Quantification)**
  - Samples are chemically labelled
Scope of this talk

• Not talking about experimental pre-processing

• MS of the proteome results in a series of data points \((x, y)\) where \(x\) is the m/z value and \(y\) is a measure of abundance

• Our focus is on the analysis of such data in a “label-free” context - we will not discuss the important step of identifying the peptides and proteins
Spectra from two duplicate samples
Spectra from two distinct samples
Experimental Design

• Quantitative measures of the proteome are influenced by many sources of variation - experimental and inter-individual

• Careful experimental design is crucial if false inferences are to be avoided, at all stages, including sample selection and sample handling

• Numerous studies have demonstrated the importance of sample handling
Peak profiles from 3 different SELDI chip types
Sample processed in 4 different ways
Banks et al, Clinical Chemistry 2005
Cluster analysis of 119 samples based on peak profiles
Serum, Citrate plasma, EDTA plasma, Heparin plasma
Essential design principles

• To avoid confounding, at sample selection stage groups to be compared should be matched wherever possible for factors such as age and sex

• To avoid bias, samples from different groups must not be treated differently (e.g. in time to processing, method of processing).

• Use randomisation wherever possible, e.g. in assignment of samples to chips
Designing efficient studies

• Attention to experimental design can also improve efficiency of studies

• More sophisticated systems of randomization may be useful to remove known sources of variation (block designs)

• Technical replicates can be used to reduce intra-individual variability

• Sample size calculations should be carried out to maximize power within constraints of the study
Analysis - pre-processing

• Calibration
  • Time-of-flight converted to m/z scale using a number of calibrants with known m/z values

• Baseline subtraction
  • Remove baseline “noise” from matrix, e.g. by Loess smoothing

• Normalisation
  • Data normalised to create comparability between spectra
Normalisation

• Aim is to remove technical sources of variation (machine performance, sample amount, etc)

• Normalisation by the (local or global) total ion current is often applied

• Concerns have been raised that this may remove relevant biological information as well as the desired experimental variability

• Optimal methods are an area of ongoing research
Peak alignment and detection

• There is imprecision on both the x (m/z) and y (abundance) axes

• Peaks in any one spectrum may be defined as points that are maximal within +/- N points on either side

• Need to also consider signal-to-noise ratio - peaks must exceed the typical local intensity values by a certain ratio

• How can peaks be aligned across spectra allowing for variability in x?
Numerous methods of peak detection - none perfect

• **Morris et al (2005):** use the mean of all spectra
  - a peak denoting a molecular feature should stand out against noise
  - circumvents problem of alignment

• **Tibshirani et al (2004):** apply one-dimensional clustering algorithm to m/z values of peaks across spectra
  - Tight clusters should represent the same biological peak possibly shifted slightly along x-axis
  - Use centroid of each cluster as consensus m/z value
Peak detection - in-house method

- Smooth each spectrum using progressively wider moving averages
- Identify peaks of size $n$ in each of these smoothed spectra (local maxima with at least $n$ lower points at each side)
- Create a frequency distribution
- Identify peak clusters
Smoothing a spectrum

window = 3
window = 39
Peak detection

• Smooth each spectrum using progressively wider moving averages

• Identify peaks of size $n$ in each of these smoothed spectra (local maxima with at least $n$ lower points at each side)

• Create a frequency distribution

• Identify peak clusters
Frequency distribution of peaks
Identifying peak clusters
Scoring peaks

• Each spectrum compared to resulting peak list to determine presence/absence of peak

• If peak is present then largest abundance measurement within certain tolerance is recorded as peak abundance

• Not all peaks can be detected and it is a subject of research to consider how to treat these (zeroes, missing, censored, . . .)
Extracting common peaks from spectra
Peak profile data

• After all this processing, raw data is reduced to ~300 to 1000 fixed peaks

• For each sample the data consist of an abundance measure at each of these peaks

• Aims
  • Identify peaks that differ between groups
  • Develop classification scheme which can be applied to new samples
Exploratory data analysis

- Proteomics data is multivariate – measurements on numerous inter-correlated peaks

- Some exploratory data analysis is useful to explore gross differences between groups – e.g. compare mean spectra

- Can carry out principal components analysis to investigate the pattern of variation more fully
Renal transplant study

• Set of 50 renal transplant recipients, 30 with clinically defined stable condition and 20 with unstable condition

• In addition 30 healthy controls and 30 patients with renal failure (prior to dialysis), similar age and gender distribution in each group

• Identical sample handling and processing protocol used

• Samples randomised to chips and all analysed in duplicate

• 3 chip types (H50, IMAC and CM10)
Which peaks differ between patients and controls?

Simple test to compare group means for each peak

• To allow for duplicated samples we use linear random effects model
• For person i, replicate k, the intensity is modelled by

\[ y_{ik} = \alpha + \beta x_i + \eta_i + \varepsilon_{ik} \]

where \( x_i \) is indicator (1/0) of case/control status,
\( \eta_i \sim N(0, \sigma_u^2) \), \( \varepsilon_{ik} \sim N(0, \sigma_e^2) \)

• Covariates can easily be included
Sample size calculations

- Sample size depends on intra- ($\sigma^2$) and inter- ($\tau^2$) sample variance, number ($m$) of technical replicates, difference in means to be detected ($\delta$)

- For a comparison of 2 groups, for a particular peak, the number of samples $n$ required in each group is

$$n = 2m \left[ \frac{Z \alpha / 2 + Z \beta}{\delta} \right]^2 \left( \tau^2 + \frac{\sigma^2}{m} \right)$$

where $\alpha$ is the significance level and $1-\beta$ the power
Estimates of variance

- Sample size requirements depend heavily on intra- and inter-sample variance ($\sigma^2$ and $\tau^2$) of the abundance measures for the peak

- $\sigma^2$ and $\tau^2$ can be estimated from pilot data

- Since there are multiple peaks, to get one estimate we may use the median, 90th percentile or maximum variance across peaks

- Generally maximum is too conservative, and median will mean power may be low for 50% of peaks
Effect of variance estimate on sample size

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>$1-\beta$</th>
<th>$m$</th>
<th>Median</th>
<th>90\textsuperscript{th} percentile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log_2(1.25)$</td>
<td>0.80</td>
<td>2</td>
<td>8</td>
<td>(16)</td>
<td>456</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>7</td>
<td>(28)</td>
<td>267</td>
</tr>
<tr>
<td></td>
<td>log$_2(1.5)$</td>
<td>2</td>
<td>11</td>
<td>(22)</td>
<td>650</td>
</tr>
<tr>
<td>0.95</td>
<td>4</td>
<td>2</td>
<td>10</td>
<td>(40)</td>
<td>380</td>
</tr>
<tr>
<td>log$_2(2)$</td>
<td>0.80</td>
<td>2</td>
<td>3</td>
<td>(6)</td>
<td>138</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>(8)</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>log$_2(2)$</td>
<td>2</td>
<td>3</td>
<td>(12)</td>
<td>197</td>
</tr>
<tr>
<td>0.95</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>(2)</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>1</td>
<td>(4)</td>
<td>28</td>
</tr>
<tr>
<td>0.95</td>
<td>0.80</td>
<td>2</td>
<td>2</td>
<td>(4)</td>
<td>68</td>
</tr>
<tr>
<td>log$_2(2)$</td>
<td>0.95</td>
<td>4</td>
<td>1</td>
<td>(4)</td>
<td>40</td>
</tr>
</tbody>
</table>

Cairns et al, Proteomics, 2009
Multiple testing

• Several hundred peaks will typically be tested. At the discovery stage the aim is not to demonstrate a difference between groups with certainty, but to ensure that a high proportion of “significant” results are true positives.

• More useful to think of false discovery rate (FDR) than false positive rate (FPR)

\[
\text{FPR} = E[\frac{\text{(# false positives)}}{\text{(# true null)}}] = \alpha
\]

\[
\text{FDR} = E[\frac{\text{(# false positives)}}{\text{(# significant)}}]
\]
False discovery rate

Let $P$ be the total number of peaks and $\pi$ the proportion of peaks that are truly differentially expressed by a particular amount, then the FDR can be estimated by:

$$\frac{\alpha(1-\pi)}{\{\alpha(1-\pi)+(1-\beta)\pi\}}$$

This can be used to inform choice of $\alpha$
FDR as a function of $\alpha$, $\beta$ and $\pi$

<table>
<thead>
<tr>
<th>$\pi$</th>
<th>$\alpha$</th>
<th>$1-\beta$</th>
<th>Number of truly differentially expressed peaks</th>
<th>Approximate expected FDR</th>
<th>Expected number of FD in 200 peaks</th>
<th>Expected number of truly differentially expressed peaks missed</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.005</td>
<td>0.001</td>
<td>0.95</td>
<td>1</td>
<td>0.173</td>
<td>0.199</td>
<td>0.05</td>
</tr>
<tr>
<td>0.005</td>
<td>0.001</td>
<td>0.90</td>
<td>1</td>
<td>0.181</td>
<td>0.199</td>
<td>0.1</td>
</tr>
<tr>
<td>0.005</td>
<td>0.001</td>
<td>0.80</td>
<td>1</td>
<td>0.199</td>
<td>0.199</td>
<td>0.2</td>
</tr>
<tr>
<td>0.05</td>
<td>0.001</td>
<td>0.95</td>
<td>10</td>
<td>0.020</td>
<td>0.19</td>
<td>0.5</td>
</tr>
<tr>
<td>0.05</td>
<td>0.001</td>
<td>0.90</td>
<td>10</td>
<td>0.021</td>
<td>0.19</td>
<td>1</td>
</tr>
<tr>
<td>0.05</td>
<td>0.001</td>
<td>0.80</td>
<td>10</td>
<td>0.023</td>
<td>0.19</td>
<td>2</td>
</tr>
<tr>
<td>0.2</td>
<td>0.001</td>
<td>0.95</td>
<td>40</td>
<td>0.004</td>
<td>0.16</td>
<td>2</td>
</tr>
<tr>
<td>0.2</td>
<td>0.001</td>
<td>0.90</td>
<td>40</td>
<td>0.004</td>
<td>0.16</td>
<td>4</td>
</tr>
<tr>
<td>0.2</td>
<td>0.001</td>
<td>0.80</td>
<td>40</td>
<td>0.005</td>
<td>0.16</td>
<td>8</td>
</tr>
<tr>
<td>0.005</td>
<td>0.01</td>
<td>0.95</td>
<td>1</td>
<td>0.677</td>
<td>1.99</td>
<td>0.05</td>
</tr>
<tr>
<td>0.005</td>
<td>0.01</td>
<td>0.90</td>
<td>1</td>
<td>0.689</td>
<td>1.99</td>
<td>0.1</td>
</tr>
<tr>
<td>0.005</td>
<td>0.01</td>
<td>0.80</td>
<td>1</td>
<td>0.713</td>
<td>1.99</td>
<td>0.2</td>
</tr>
<tr>
<td>0.05</td>
<td>0.01</td>
<td>0.95</td>
<td>10</td>
<td>0.167</td>
<td>1.9</td>
<td>0.5</td>
</tr>
<tr>
<td>0.05</td>
<td>0.01</td>
<td>0.90</td>
<td>10</td>
<td>0.174</td>
<td>1.9</td>
<td>1</td>
</tr>
<tr>
<td>0.05</td>
<td>0.01</td>
<td>0.80</td>
<td>10</td>
<td>0.192</td>
<td>1.9</td>
<td>2</td>
</tr>
<tr>
<td>0.2</td>
<td>0.01</td>
<td>0.95</td>
<td>40</td>
<td>0.040</td>
<td>1.6</td>
<td>2</td>
</tr>
<tr>
<td>0.2</td>
<td>0.01</td>
<td>0.90</td>
<td>40</td>
<td>0.043</td>
<td>1.6</td>
<td>4</td>
</tr>
<tr>
<td>0.2</td>
<td>0.01</td>
<td>0.80</td>
<td>40</td>
<td>0.048</td>
<td>1.6</td>
<td>8</td>
</tr>
<tr>
<td>0.005</td>
<td>0.05</td>
<td>0.95</td>
<td>1</td>
<td>0.913</td>
<td>9.95</td>
<td>0.05</td>
</tr>
<tr>
<td>0.005</td>
<td>0.05</td>
<td>0.90</td>
<td>1</td>
<td>0.917</td>
<td>9.95</td>
<td>0.1</td>
</tr>
<tr>
<td>0.005</td>
<td>0.05</td>
<td>0.80</td>
<td>1</td>
<td>0.926</td>
<td>9.95</td>
<td>0.2</td>
</tr>
<tr>
<td>0.05</td>
<td>0.05</td>
<td>0.95</td>
<td>10</td>
<td>0.500</td>
<td>9.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0.05</td>
<td>0.05</td>
<td>0.90</td>
<td>10</td>
<td>0.514</td>
<td>9.5</td>
<td>1</td>
</tr>
<tr>
<td>0.05</td>
<td>0.05</td>
<td>0.80</td>
<td>10</td>
<td>0.543</td>
<td>9.5</td>
<td>2</td>
</tr>
<tr>
<td>0.2</td>
<td>0.05</td>
<td>0.95</td>
<td>40</td>
<td>0.174</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>0.2</td>
<td>0.05</td>
<td>0.90</td>
<td>40</td>
<td>0.182</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>0.2</td>
<td>0.05</td>
<td>0.80</td>
<td>40</td>
<td>0.200</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>
Improving the statistical model

Various improvements have been proposed to the statistical analysis process (e.g. Oberg et al, 2008)

• Normalisation could be carried out in the same step as testing and estimating differences between groups;

• In labelled experiments a structure can be imposed on the peaks (peptides within proteins, etc.)
Classification of samples

• There are many available classification methods

• Classification trees
  • Use peak information to successively split the sample until all subjects at the same node are in the same class or a minimum node size is reached

• Key issue is to avoid over-fitting
Random Forest algorithm

- The Random Forest (RF) algorithm is a classifier based on a group of classification trees.
- In each tree, samples are successively split using the classifier that “best” discriminates between classes. Usually this is based on the Gini index, a measure of impurity of the nodes.
- Trees can be grown until each node consists entirely of samples from one class (pure node). Alternatively a minimum node size can be set as stopping rule.
- Idea of RF is to grow many trees and classify on the basis of the majority vote (Breiman, 2001).
Avoiding over-fitting

• How do we make the trees differ and be less dependent on the particular data set?
  – Each tree is based on a bootstrap sample of original data set
  – At each partition only a random subset of the variables is available to choose from to make the best split

• Each tree used to classify items NOT in the bootstrap sample (out-of-bag (OOB))
Breast cancer proteomics data

• Several groups met in Leiden, Netherlands, in 2007 to compare methods of classifying samples from breast cancer patients and controls on basis of proteomics data

• Data set:
  - Mass spectrometry (MS) proteomic profiles from 76 breast cancer cases and 77 controls formed initial data set
  - Profiles from an additional 39 cases and 39 controls used as validation data

• MS profiles:
  - The data were available after some pre-processing
  - Abundance measurements for 11,205 points corresponding to mass-to-charge (m/z) ratios along the spectrum

Mertens, SAGMB, 2008
Mean spectra from controls and cases

mean spectra (controls)

mean spectra (cases)
Two-stage analysis strategy

- **Peak detection**: reduce the 11,205 data points to 365 “robust” peaks
  - Using all 153 samples 444 peaks detected
  - Robust peak detection method reduced this to 365 peaks which appear in 7 out of 10 subsets selected

- **Random Forest**: apply RF to the 365 peaks, using 20,000 trees in the forest, and classify according to majority vote

- We have also compared this to one-stage strategy: RF applied to all 11,205 data points
Classification results

Confusion matrix of true versus predicted case-control status for the classification (a) based on all data points and (b) based on peaks

<table>
<thead>
<tr>
<th>(a) Predicted classification</th>
<th>True classification</th>
<th>(b) Predicted classification</th>
<th>True classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case</td>
<td>Control</td>
<td></td>
</tr>
<tr>
<td>Case</td>
<td>62</td>
<td>13</td>
<td>Case</td>
</tr>
<tr>
<td>Control</td>
<td>14</td>
<td>64</td>
<td>Control</td>
</tr>
</tbody>
</table>
Scatter plot of case and control samples showing votes from the two applications of the RF algorithm.
Of the groups taking part in the original project, our analysis (using RF applied to peaks) had joint lowest percentage correct from calibration data (84%) but highest from validation data (83% correct)

Barrett & Cairns, SAGMB, 2008; Hand, SAGMB, 2008
Rerunning with RF applied to raw data points, results from both calibration and validation data sets are very similar (82% and 85% correct)

Barrett & Cairns, SAGMB, 2008; Hand, SAGMB, 2008
Classification: summary

• RF has good performance as a classifier - in terms of percentage correctly classified - certainly comparable to other methods

• One especially good property is unbiased estimation of performance from initial data set

• Use of peaks has little effect on classification accuracy compared with raw (correlated) data

• Parsimony principle would favour use of summary measures where good ones can be found
Variable importance measures

• Can also see which variables have been important in classification - emphasis is on ranking of variables rather than absolute measures

• Most common VIMs are Gini index (based on impurity of nodes) and permutation-based VIM

• Have been used in numerous studies in the analysis of genetic, gene expression and proteomic data

• We have investigated VIMs but so far not found them to outperform analysis of each peak separately
General conclusions

• Careful attention to study design is essential to avoid bias and improve efficiency

• Proper understanding of (sources of) variation is needed, so that these can be minimized or removed as far as possible

• Complex analyses should be preceded by more simple descriptive analyses
Acknowledgements

Section of Epidemiology and Biostatistics and Clinical Proteomics Research Group, Leeds Institute of Molecular Medicine, University of Leeds

Especially:

David Cairns - MRC Research fellow in Statistics
Roz Banks - Professor of Biomedical proteomics
Dave Perkins - Principal research scientist in bioinformatics