Statistical Analysis of Mass Spectrometry Data

Kimberly F. Sellers
Department of Mathematics and Statistics
Georgetown University

Presentation Outline

- Background
  - What is Proteomics?
  - What is Mass Spectrometry?
  - Significance, benefits and drawbacks
- Uploading and plotting data
- Low-level analysis
  - One spectrum
  - Several spectra
- Quality assessment
- Finding proto-biomarkers
- Discussion

First Things First

- Load the R package, **PROcess**
  - Functions include baseline correction, normalization of spectra, peak alignment and detection

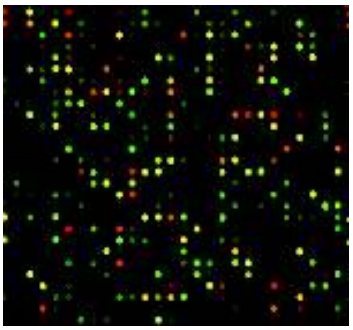
```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite("PROcess")
> library("PROcess")
```

Motivation

www.labtestsonline.org/lab/photo/images/needle_in.jpg

What is Microarray Data Analysis?

- matrix of spots
- intensity = relative abundance of mRNA
- Goal: determine differentially expressed genes



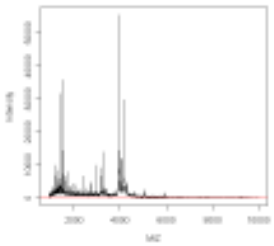
http://www.genomic.ch/pict/array/chip_exp.png

What is Proteomic Data Analysis?

- "Proteomics" = direct study of proteins
- Two popular approaches:

Two-dimensional gel electrophoresis

Mass spectrometry

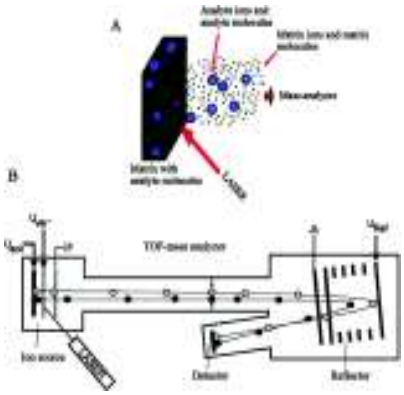


Microarray vs. Proteomic Data Analysis

- Similarities:
  - Research Goal: to detect changes
  - Technology and image processing
- Differences:
  - Poor correlation btwn protein & mRNA abundance
    - 30,000 – 40,000 genes
    - 200,000 – 2,000,000 proteins
  - Problem set-up
  - Detection
    - Microarray: Differential expression
    - Proteomics: Differential expression + differential modification

What is Mass Spectrometry?

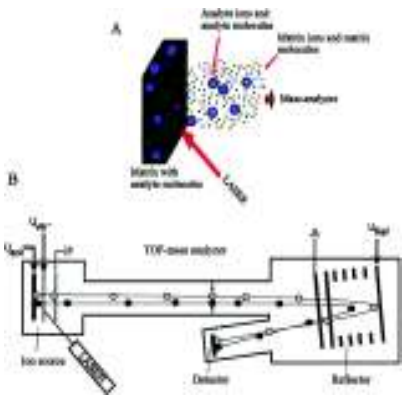
- Analytic tool to identify proteins
- Mass spectrometer: instrument that measures masses of molecules converted into ions via mass-to-charge (m/z) ratio



http://www.biochem.arizona.edu/classes/bioc471/pages/Lecture23/bi70_0437_1.gif

Experimental Procedure

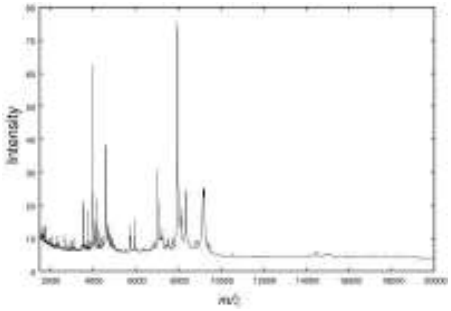
1. Mix sample with matrix (causes mixture to crystallize while drying)
2. Place metal plate containing sample into vacuum chamber
3. Nitrogen laser strikes crystal with light pulses
4. Matrix molecules absorb energy from laser and transfer it to the proteins, causing them to desorb and ionize, producing cloud of ionized protein molecules
5. Electric field accelerates ionized proteins into flight tube, where they "fly" until they strike detector that records TOF



http://www.biochem.arizona.edu/classes/bioc471/pages/Lecture23/bi70_0437_1.gif

Significance

- Spectral data consist of sequentially recorded numbers of ions arriving at detector coupled with corresponding m/z values



- Dataset contains tens or hundreds of spectra
- Peaks in intensity plot represent proteins present in sample

Benefits and Drawbacks

Benefits:

- Generates large amounts of spectral data
- Detects protein differential expression and modification in different samples
- Provides high-resolution measurements

Drawbacks:

- Nonlinearity in detector response
- Noisy data lead to high false positive peak identification rate
- systematic errors in experimentation, sample preparation, and instrumentation
- Amplitude and phase variation

Upload Data

- Here, we access example dataset provided in PROcess. Generally, you can read-in data, e.g. using "read.table" command

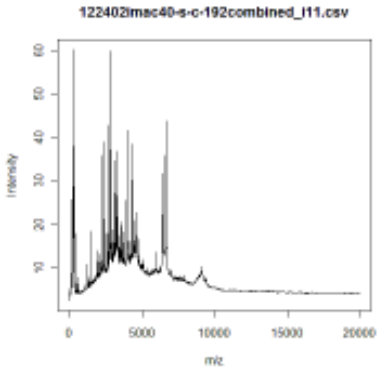
```
> fdat <- system.file("Test",package="PROcess")
> fs <- list.files(fdat,pattern="\\.*csv\\.*",full.names=TRUE)
> length(fs)
[1] 2

> f1 <- read.files(fs[1])
> dim(f1)
[1] 13482     2
> colnames(f1)
[1] "M.Z"       "Intensity"
```

This shows that there are actually two files available here to analyze. Let's consider the first of the two.

Plot Data

> plot(f1,type="l",xlab="m/z", ylab="Intensity")

> title(basename(fs[1]))



Low-level Analysis

- Goal: To automatically identify differentially expressed and modified proteins.
  - Background (i.e. baseline) correction
  - Data denoising and normalization
  - Align group replicate profiles (i.e. spectral or peak alignment)
  - Feature (i.e. peak) detection

Baseline Correction

bslnoff {PROcess} R Documentation

Baseline Subtraction

Description

This function estimates the baseline and then removes baseline from the raw spectrum

Usage

bslnoff(f, breaks = 200, qntl = 0, method = c("loess", "approx"), bw = 0.005, plot = FALSE, ...)

Arguments

f a matrix with M/Z values in the first column and intensities in the second column

breaks number of breaks set to M/Z values for finding the local minima or points below a certain quantile of intensities; breaks-1 equally spaced intervals on the log M/Z scale

qntl if 0, find local minima; if >0 find intensities < qntl*100th quantile locally

method "loess" or "approx" (linear interpolation)

bw the bandwidth to be passed to loess

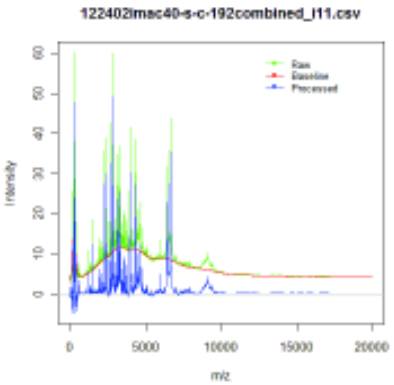
plot TRUE or FALSE, if true, it will plot the raw spectrum, theestimated baseline and the baseline substracted spectrum

... Further parameters that get passed on to plot

Baseline Correction (cont.)

> bseoff <- bslnoff(f1, method="loess", bw=0.1, xlab="m/z", plot = TRUE)

> title(basename(fs[1]))



Exercise: try changing bandwidth, quantile, etc. to see impact on baseline correction.

isPeak {PROcess} R Documentation

Locate Peaks in a Spectrum

Description

Find local maxima of a spectrum

Usage

```
isPeak(x, SNR = 2, span = 51, sm.span=11,plot=FALSE,add = FALSE,
zerothrsh=2, area.w = 0.003, ratio = 0.2, ...)
```

Arguments

to calculate area of the peak.

x	a matrix of two columns representing a spectrum, with the first column the m/z value and second the intensity
SNR	signal to noise ratio criterion for peak detection
span	parameter for estimating local variance before peak detection; default is 51 points, that is, 40 points to the left and right of a point of which the variance is being estimated.
sm.span	parameter for smoothing the spectrum before peak detection; default is 11 points, that is, 5 points to the left and right of a point being smoothed
plot	logical, plot the smoothed spectrum and peaks?
add	add to the existing raw and baseline-subtracted plot?
zerothrsh	ignore peaks whose intensity values are below zerothrsh.
area.w	the neighbourhood of a peak: max(1-area.w, 1+area.w).
ratio	if min(area) > ratio, the peak is retrained.
...	further arguments that get passed on to plot.

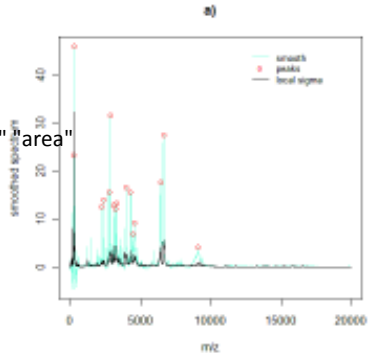
Peak detection (cont.)

> pkgobj <- isPeak(bseoff, span = 81, sm.span = 11, plot = TRUE, zerothrsh = 2, area.w = 0.003, ratio = 0.2, main="a)")

> dim(pkgobj)
[1] 13482 5

> colnames(pkgobj)
[1] "peak" "smooth" "mz" "sigmas" "area"

- peak: logical (peak or not)
- smooth: smooth of the spectrum
- mz: m/z value
- sigmas: local variation estimates
- area: peak area under curve



Peak detection (cont.)

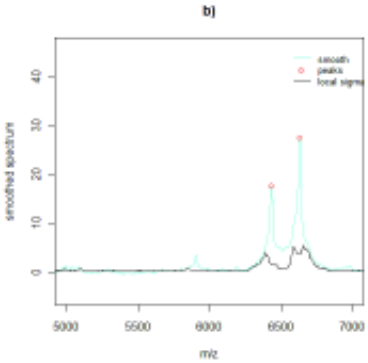
> pkgobj[pkgobj$peak == TRUE,]

	peak	smooth	mz	sigmas	area
1536	TRUE	23.239521	249.8972	5.7694057	92.17202
1653	TRUE	46.005951	290.5522	20.7377551	182.97196
4501	TRUE	12.437345	2212.0081	0.8916012	120.10662
4656	TRUE	13.999812	2367.9463	0.9767109	137.05628
5010	TRUE	15.586117	2743.9725	1.5338407	158.07309
5117	TRUE	31.667085	2863.0734	2.1175916	326.64155
5295	TRUE	12.752275	3066.8021	1.6790805	136.04358
5427	TRUE	12.061064	3222.3973	2.5204124	158.94419
5506	TRUE	13.292926	3317.3581	1.3291588	161.70659
6024	TRUE	16.649756	3974.1382	2.2875057	214.66003
6254	TRUE	15.606809	4284.7434	2.0811745	198.22859
6391	TRUE	6.880581	4475.3045	0.8058924	96.68459
6462	TRUE	9.105213	4575.6920	1.4157855	134.79582
7658	TRUE	17.541217	6433.9461	1.3702803	274.32702
7774	TRUE	27.568420	6630.9726	3.7696683	427.55682
9098	TRUE	3.990684	9090.1834	0.4322778	90.92869

Peak detection (cont.)

- To inspect peaks more closely:

> specZoom(pkgobj, xlim=c(5000,7000), main="b)")



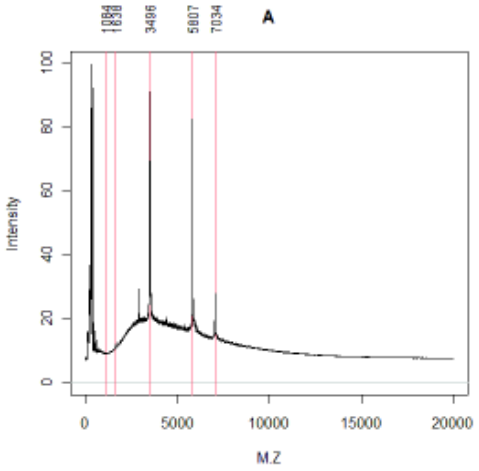
Processing Set of Calibration Spectra

- Given protein locations in the samples, **plotCali** identifies protein locations within each spectrum plot

```
amu.cali <- c(1084, 1638, 3496, 5807, 7034)
plotCali <- function(f, main, lab.cali){
    x <- read.files(f)
    plot(x, main=main, ylim=c(0, max(x[,2])), type="n")
    abline(h=0, col="gray")
    abline(v= amu.cali, col="salmon")
    if(lab.cali) axis(3, at=amu.cali, labels=amu.cali, las =3, tick =
        FALSE, col = "salmon", cex.axis=0.94)
    lines(x)
    return(invisible(x))
}
```

Processing Set of Calibration Spectra

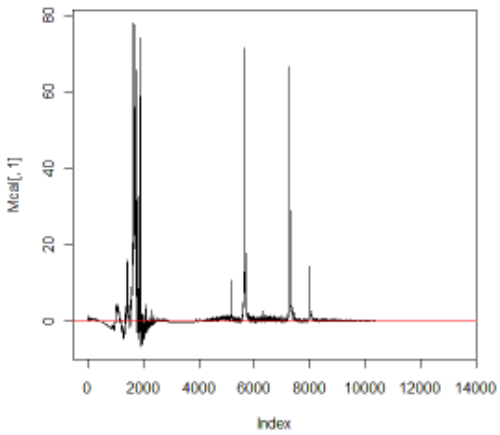
```
> dir.cali <-
system.file("calibration",
package="PROcess")
> files <- dir(dir.cali,
full.names = TRUE)
> i <- seq(along = files)
> par(mfrow=c(4,2))
> mapply(plotCali, files,
LETTERS[i], i<=2)
```



Background Subtraction on Set of Spectra

- **rmBaseline** corrects for baseline across several spectra, calling **bslnoff** for each spectrum

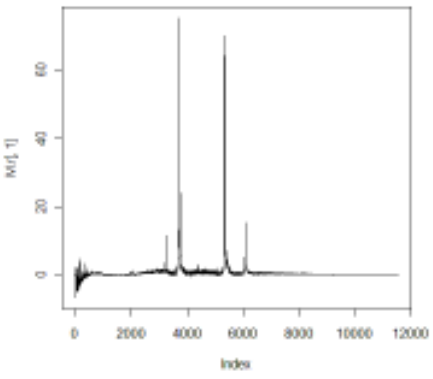
```
> Mcal <-
rmBaseline(dir.cali)
```



Normalizing Spectra

- Based on total ion normalization
  - Compute area under curve (AUC) for peaks
  - Scale all spectra to median AUC
  - Method relies on two assumptions:
    - number of over- and under-expressed proteins approximately equal (on average)
    - Number of proteins with changed expression levels is small

```
> M.r <- renorm(Mcal, cutoff = 400)
```

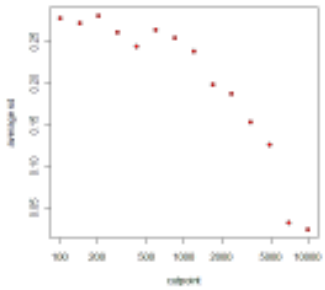


Cutoff Selection

- Noise is large for small m/z values
- Causes for elevated baseline:
  - Chemical noise
  - Ion overloading
- Find cutoff point for stable noise
- Use same cutoff for all spectra
- Algorithm:
  1. Normalize baseline-subtracted spectra to median of sums of spectra intensities
  2. Calculate standard deviation of intensities of each m/z value
  3. Compute mean of standard deviations from Step 2.

Cutoff Selection (cont.)

```
> cts <- round(10^(seq(2,4, length=14)))
> sdsFirst <- sapply(cts, avesd, Ma=Mcal)
> plot(cts, sdsFirst, xlab="cutpoint", pch=21, bg="red",  log="x",
ylab="average sd")
```



Peak Identification Among Spectra

- **getPeaks** uses **isPeak** on spectra batch
- Produces list of identified peaks among spectra, including location and relative intensity

```
> peakfile <- "C:/conferences/UAB workshop/calipeak.csv"
> getPeaks(M.r, peakfile, ratio=0.1)
```

Quality Assessment

- **quality** computes three results:
  - Quality: measures separation between signal and noise
  - Retain: relative measure of high peaks in spectrum
  - Peak: compares number of peaks in spectrum to those among all spectra
- Spectrum excluded if all are true:
  - Quality < 0.4
  - Retain < 0.1
  - Peak < 0.5

Quality Assessment (cont.)

```
> qualRes <- quality(M.r, peakfile, cutoff=400)
> qualRes
```

	Quality	Retain	peak
060503peptidecalib_1_128.csv	0.4144097	0.1710994	0.9696970
060503peptidecalib_1_16.csv	0.4558291	0.1406047	0.9696970
060503peptidecalib_1_2.csv	0.4971924	0.1178203	0.9696970
060503peptidecalib_1_256.csv	0.4095183	0.1778567	0.7272727
060503peptidecalib_1_32.csv	0.3556941	0.1297756	0.9696970
060503peptidecalib_1_4.csv	0.5220855	0.1432037	1.2121212
060503peptidecalib_1_64.csv	0.4790311	0.1430304	1.2121212
060503peptidecalib_1_8.csv	0.4174724	0.1201594	0.9696970

Finding Proto-biomarkers

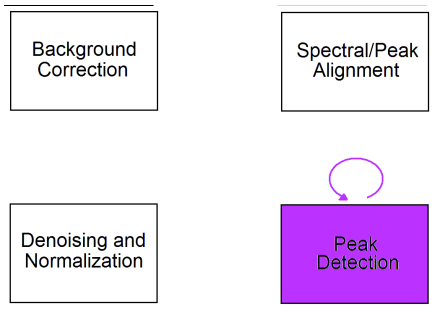
- **pk2bmkr** aligns peaks across spectra and obtains proto-biomarkers

```
> bmkfile <- "C:/conferences/UAB workshop/calibmk.csv"
> bmk1 <- pk2bmkr(peakfile, M.r, bmkfile, p.fltr = 0.5)
> mk1 <- round(as.numeric(gsub("M", "", names(bmk1))))
> mk1
[1] 2906 3498 5812 7036
```

Compare with
amu.cali <- c(1084, 1638, 3496, 5807, 7034)

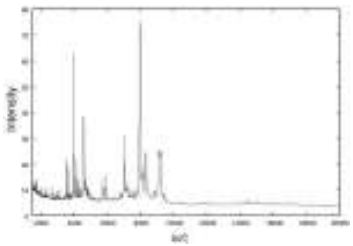
Discussion: Order of Operations

- Algorithm results are inconsistent, generally unrecoverable
- Different preprocessing algorithms could severely affect downstream analysis (Baggerly et al., 2004)



Discussion: Which approach is "best"?

- Accuracy
- False discovery rate
- Ability to detect low-lying protein spots?
- Order matters!
- Ability to automate procedure



For Added Assistance….

- Reference:
  - X Li, R Gentleman, X Lu, Q Shi, JD Iglehart, L Harris, A Miron (2005) Chapter 6: SELDI-TOF Mass Spectrometry Protein Data, in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*; R. Gentleman, VJ Carey, W Huber, RA Irizarry, S Dudoit, eds., Springer. [Note: This chapter focuses on SELDI-TOF data, but programming codes are generally applicable]

- Use the commands, “help(function-name)” and “help.start()”