

## BMI 209, Fall 2006

# Challenges in Mass Spectrometry Data Analysis

Yuanyuan Xiao

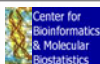
Center for Bioinformatics & Molecular Biostatistics  
UCSF Division of Biostatistics

<http://www.biostat.ucsf.edu/cbmb>



## Lecture Road Map

- ❑ **Introduction**
  - MALDI SELDI
  
- ❑ **Preprocessing of mass spectrometry data**
  - Denoising
  - Baseline subtraction
  - Peak identification
  - Normalization
  - Peak alignment
  
- ❑ **Potential limitations of proteomic profiling using mass spectrometry**



## Mass Spectrometry Proteomic Profiling

- ❑ **Surface Enhanced Laser Desorption/Ionization (SELDI-TOF-MS)**
  - Developed by Ciphergen Biosystems
  - Mass profile of a sample (serum, urine, cell lysates) in the range of 0-200K Daltons
  
- ❑ **A profile is believed to provide a rich source of information. Potentially useful for:**
  - Detection of a disease
  - Discovery of disease biomarkers
    - ✓ Identify species (protein/peptides) responsible for the differences



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

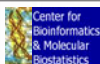
## Mass Spectrometry Proteomic Profiling

### *Specimen Types*

Blood Serum
Tissue Biopsy
Nipple Aspirate Fluid
Pancreatic Juice

### *Disease Types*

Heart disease
Ovarian cancer
Prostate cancer
Renal cancer
Breast cancer
Head & Neck cancer
Lung cancer
Pancreatic cancer



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## Advantages of Proteomic Profiling using Mass Spectrometry

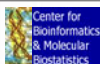
- Technology developed in the early 20<sup>th</sup> century and widely used by chemists to characterize small organic molecules
- Breakthrough: the development of methods for ionizing proteins and peptides that did not destroy them in the process
- Advantages:
  - Relatively little sample purification is required
  - Direct measurement of proteins from serum, tissue, other biological samples
  - Relatively rapid analysis time
  - Detection of proteins with m/z ranging from 0-200,000 daltons
  - Collection of useful spectra from complex mixtures
  - Mass accuracies ~0.1%



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

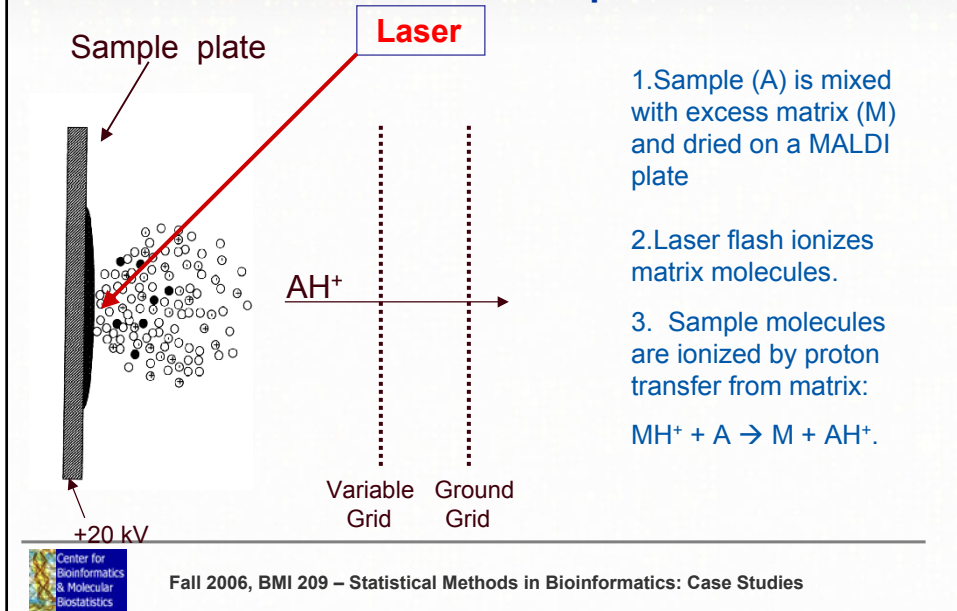
## Requirements Using Mass Spectrometry

- Substances have to be ionized to be analyzed by mass spectrometry**
- Mass-to-charge (m/z) is the fundamental measurement.**
- Sample preparation is key**



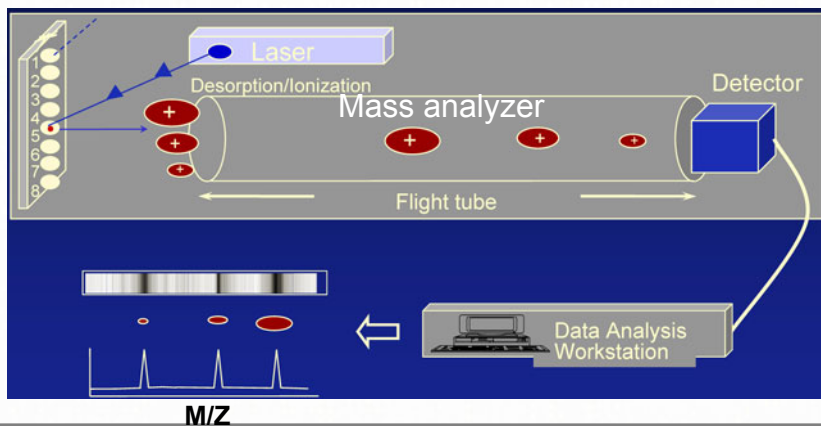
Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## Instrumentation – MALDI matrix-assisted laser desorption ionization



## SELDI ProteinChip Technology Surface Enhanced Laser Desorption/Ionization

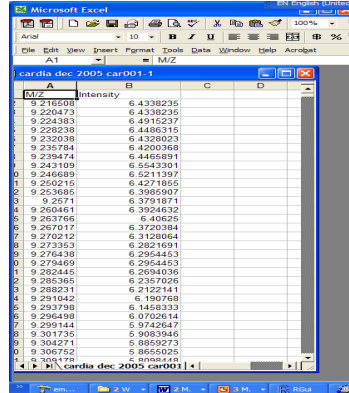
### SELDI Protein Chips



# Mass Spectra Data

## □ Spectra produced

- Mass/charge ratio (m/z) plotted against intensity
- $10^4 - 10^6$  data points per spectrum
- Sample SELDI-TOF Data

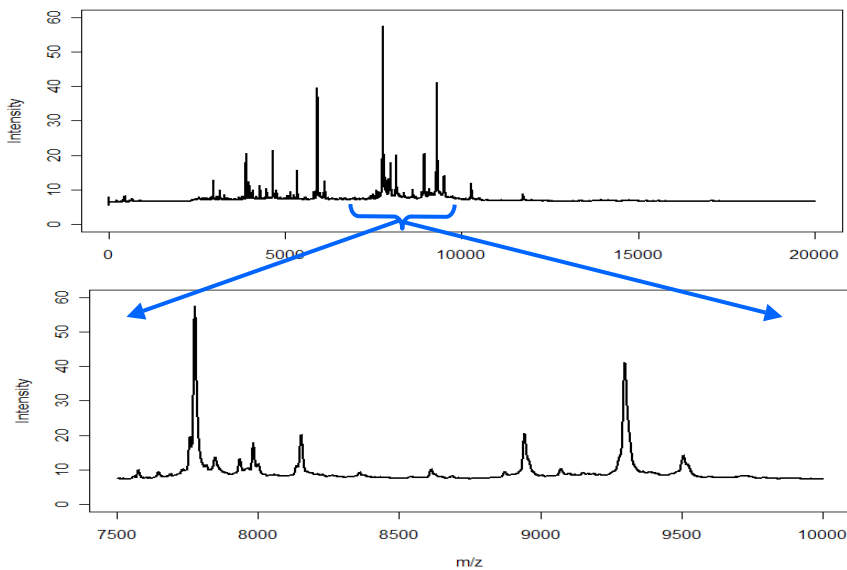


M/Z	Intensity
9.216508	6.4338236
9.220473	6.4338236
9.224383	6.4915237
9.228238	6.4489316
9.232038	6.4328023
9.235784	6.4290368
9.239474	6.4465891
9.243109	6.5543361
9.246689	6.5211397
9.250215	6.4271865
9.253685	6.3985907
9.2571	6.3791871
9.260481	6.3924632
9.263766	6.40625
9.267017	6.3720384
9.270212	6.3128064
9.273353	6.2821691
9.276438	6.2954453
9.279489	6.2954453
9.282445	6.2894036
9.285365	6.2357026
9.288231	6.2122441
9.291042	6.190768
9.293798	6.1458333
9.296498	6.0702614
9.299144	5.9742647
9.301735	5.9083946
9.304271	5.8859273
9.306752	5.8656025
9.309179	5.8464444
9.311554	5.8284444



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

# Sample SELDI Spectrum



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

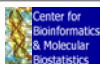
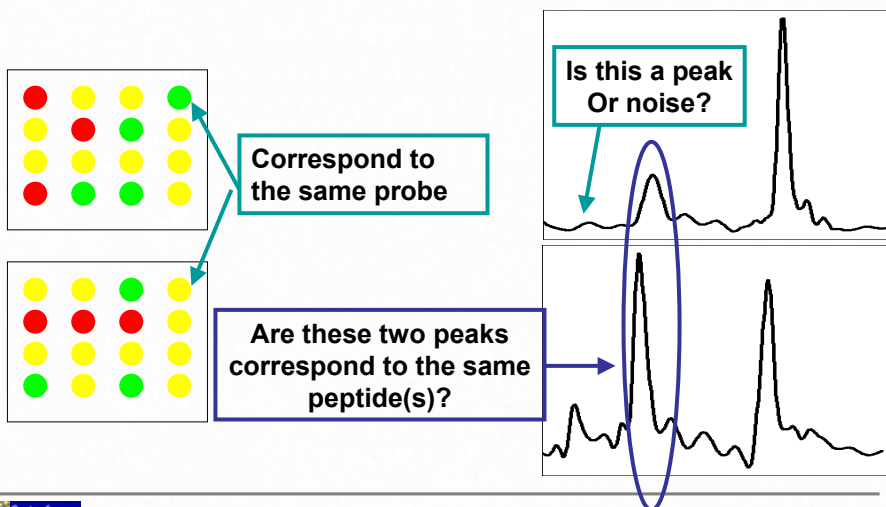
## Challenges in Mass Spectrometry Data Analysis

- ❑ Peptides represented by peaks are not known *a priori*
  - A peak may represent: noise, single peptide (known or unknown), peptide amalgamation
- ❑ M/Z values are not aligned from sample to sample
- ❑ Peak alignment is not straight-forward
- ❑ Spectra may represent tens to hundreds of thousands of data points
- ❑ Ad-hoc decision in preprocessing



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## Microarray vs Mass Spectrometry



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## Data Analysis: Preprocessing

### Aims:

- to remove noise and systematic biases in the signals while preserving useful information
- Dimension reduction step, entire collection of  $m/z$  -> a subset of  $m/z$  that represent aligned peak locations across spectra

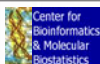
1. **Denoising**
2. **Baseline subtraction**
3. **Peak identification**
4. **Normalization**
5. **Peak alignment**



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## Preprocessing I: Denoising

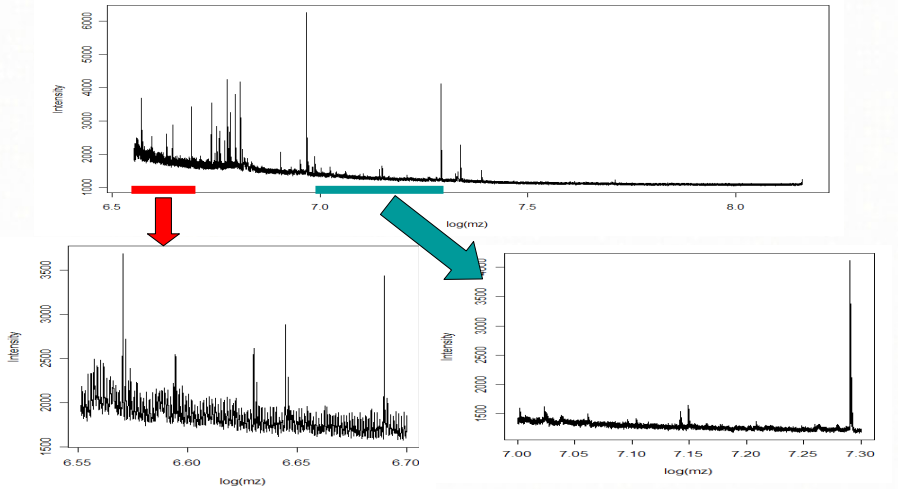
- Noise in mass spectrometry data appears as normal data points; appropriate denoising improves peak detection and downstream analysis.**
- Sources of noise**
  - Electronic noise: sensors, light sources, etc
  - Chemical noise: ions, matrix, etc
- Danger**
  - Hard to differentiate signals and noises
  - In danger of information loss



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## Preprocessing I: Denoising

An example:



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## Preprocessing I: Denoising

Wavelet Transformation for denoising

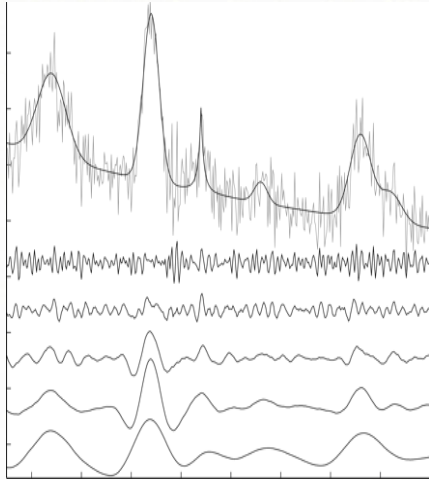
- Wavelets have been applied to many areas for removing noise
- Wavelets transformation applied to “Intensities” – one-dimensional
- A typical wavelet family chosen are the Daubechies wavelets with eight vanishing moments to filter the intensities.

Reference: Morris and Coombes, et al (University of Texas M.D. Anderson Cancer Center)  
<http://bioinformatics.mdanderson.org/cromwell.html>



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

# Preprocessing I: Denoising

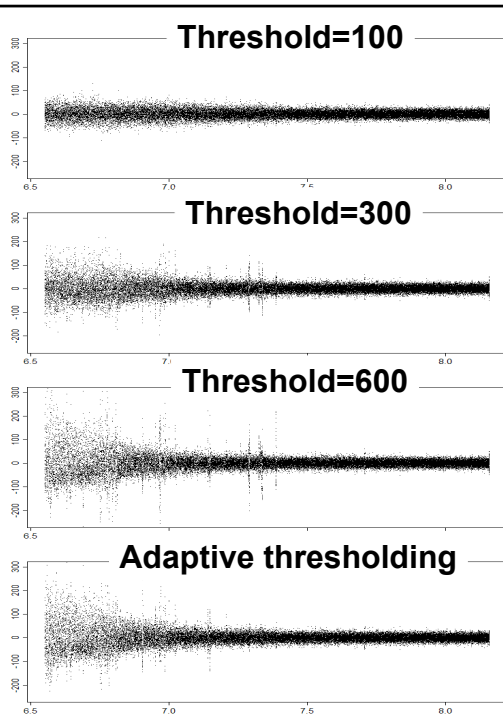


- Simulation study**
  - Signal + noise
- Ascend in the level of details, coefficients becomes smaller**
- After the transformation, a hard threshold is chosen to filter unwanted intensities, attenuating those below the threshold to zero. The denoised intensities are reconstructed when the inverse transform is applied to the filtered data.**

Randolph et al, Molecular & Cellular Proteomics 2005 4:1990

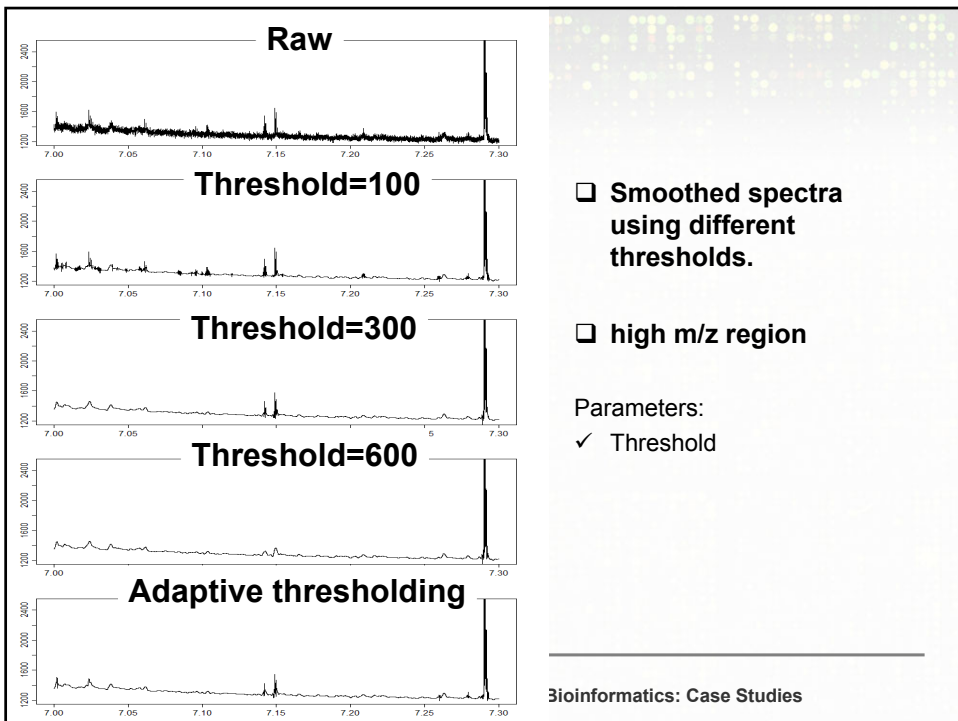
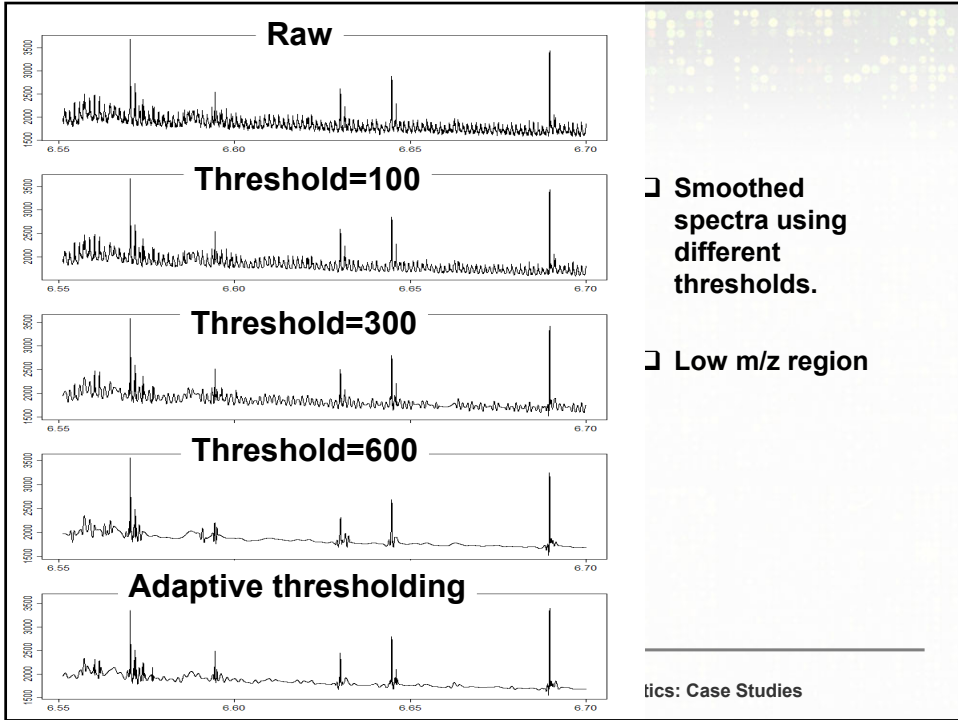


Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies



- Residual plot**  
Residuals (Raw – denoised) plotted along  $\log(m/z)$
- Low threshold**  
Danger of under-smoothing
- High threshold**  
Danger of over-smoothing
- Noise is not  $m/z$  invariant**

Bioinformatics: Case Studies

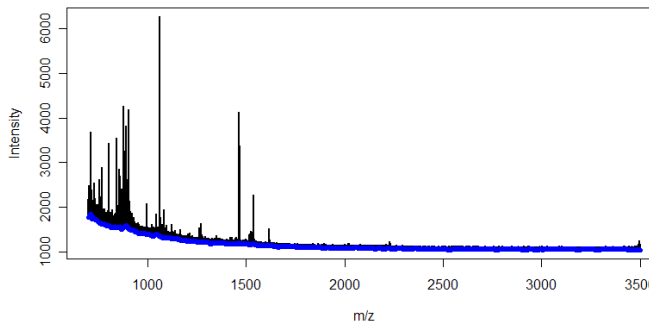
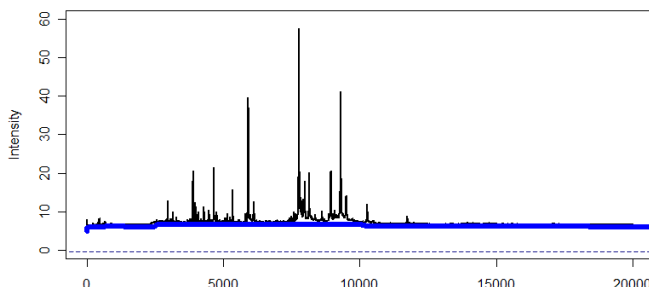


## Preprocessing II: Baseline subtraction

- Ideally a spectrum should rest more or less on the zero horizontal (intensity=0) line
- In reality, baselines are usually inflated by matrix noise.



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies



Baseline estimation

- Local minimum
- Smoothing

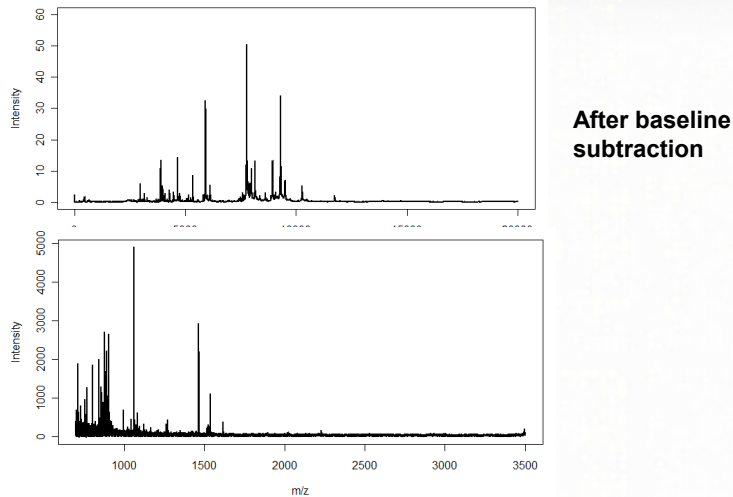
Parameters:

- Bin size



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## Preprocessing II: Baseline Subtraction



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## Preprocessing III: Peak Identification

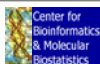
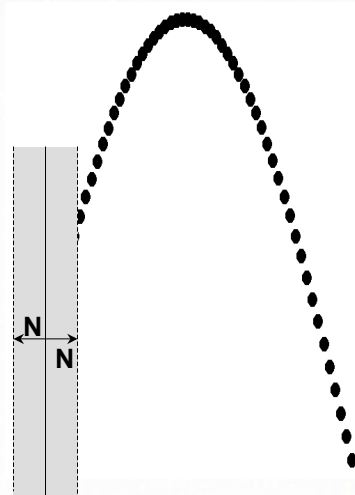
Essential step in dimension reduction

Local Maximum estimation

At each  $m/z$  point, ask “Is it the highest point in the neighborhood of  $\pm N$  points?”

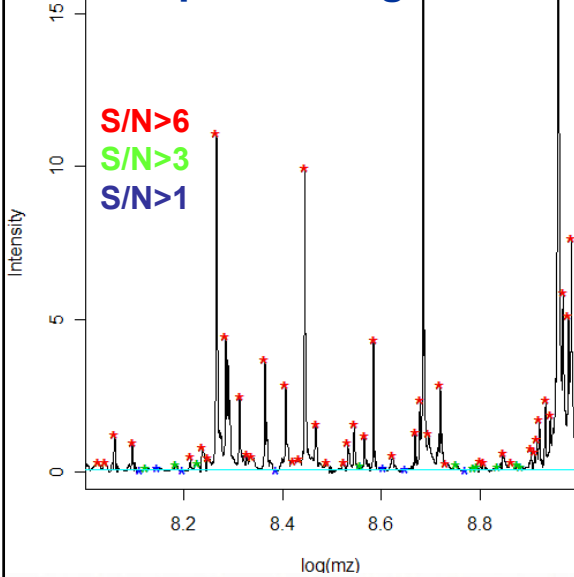
Yes = a peak

No = not a peak



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## Preprocessing III: Peak Identification



Usually refined to satisfied certain S/N criterion

Parameters:

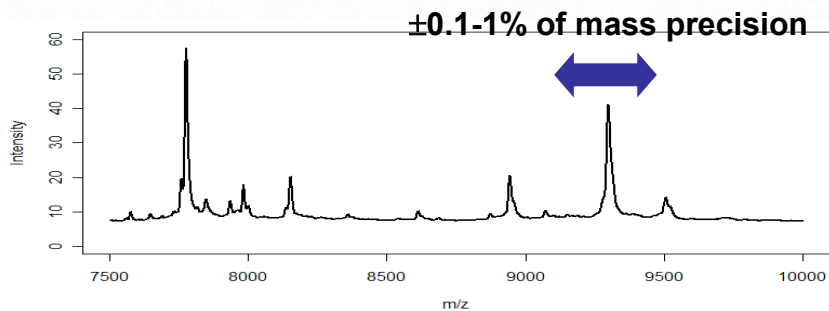
- ✓ Bin size
- ✓ S/N



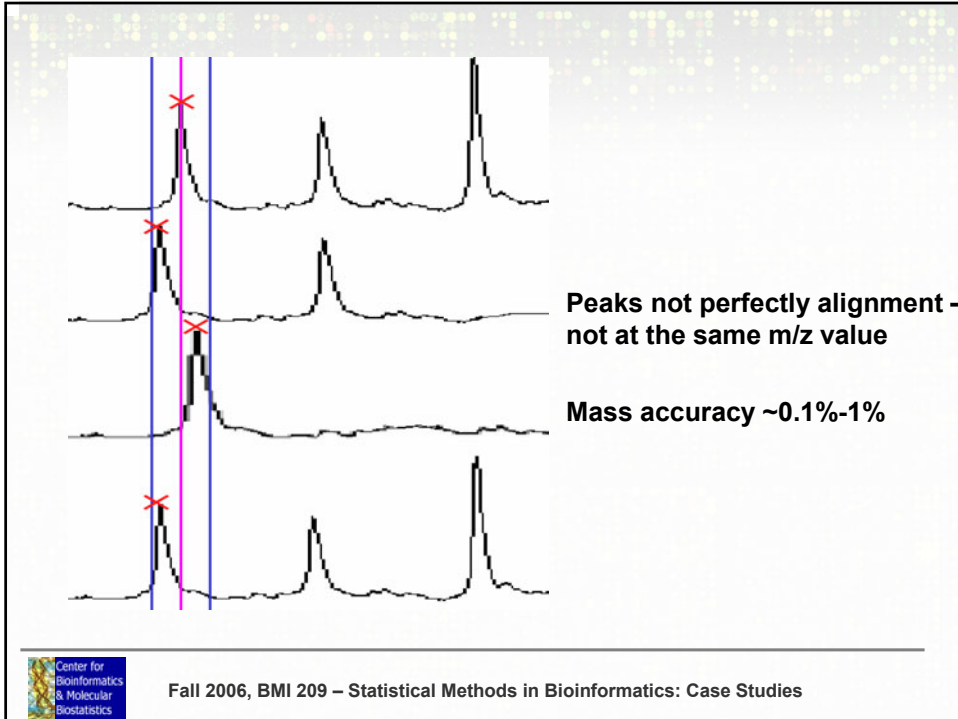
Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## Preprocessing IV: Peak Alignment

- ❑ Mass spectrometry data exhibit 2-dimensional variations
  - Vertical: Intensity measurement imprecision
  - Horizontal: Mass/Charge machine measurement imprecision



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies



## Preprocessing IV: Peak Alignment

### Binning

- R package: PROcess

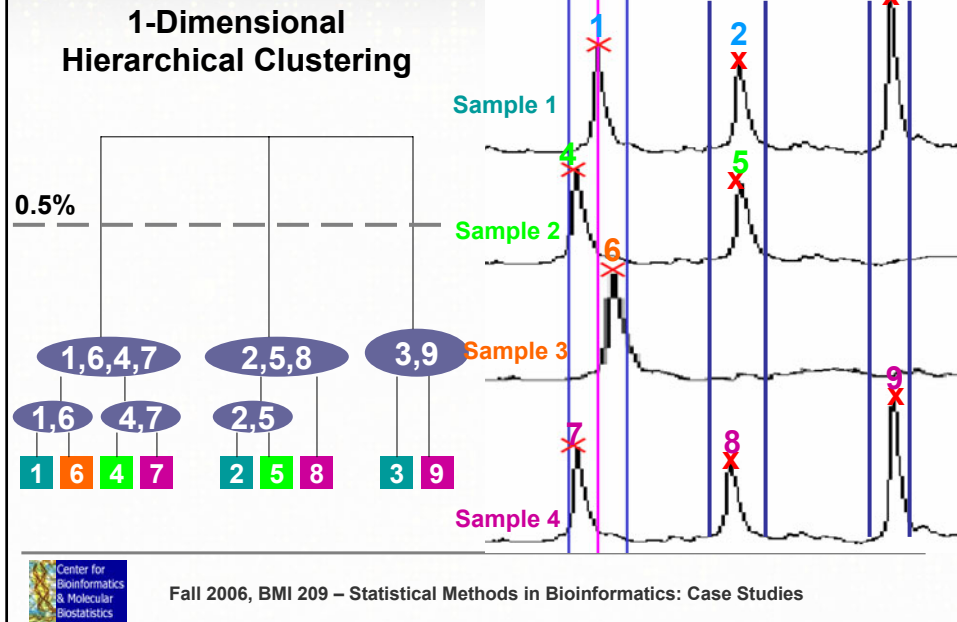
### Hierarchical clustering

- ppc, Tibshirani, *et. al*, *Bioinformatics* (2004) June 29

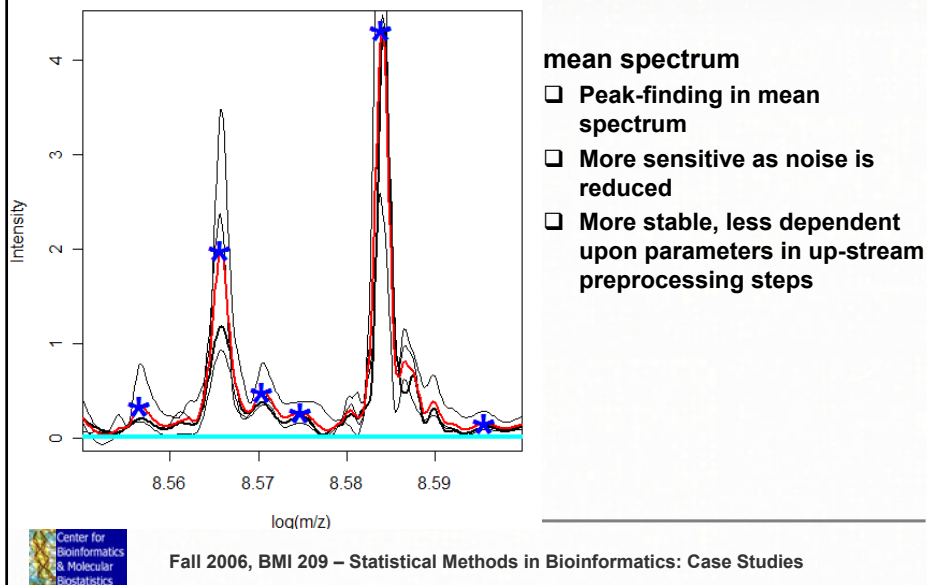
### Mean spectrum peak finding

- Morris *et al*, *Bioinformatics* (2005) 21: 1764

## Preprocessing IV: Peak Alignment



## Preprocessing IV: Peak Alignment



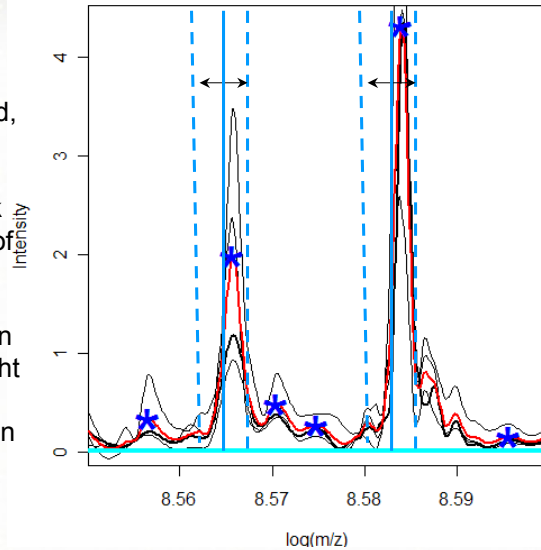
## Preprocessing IV: Peak Alignment

For each spectrum, each centroid,  
Ask,

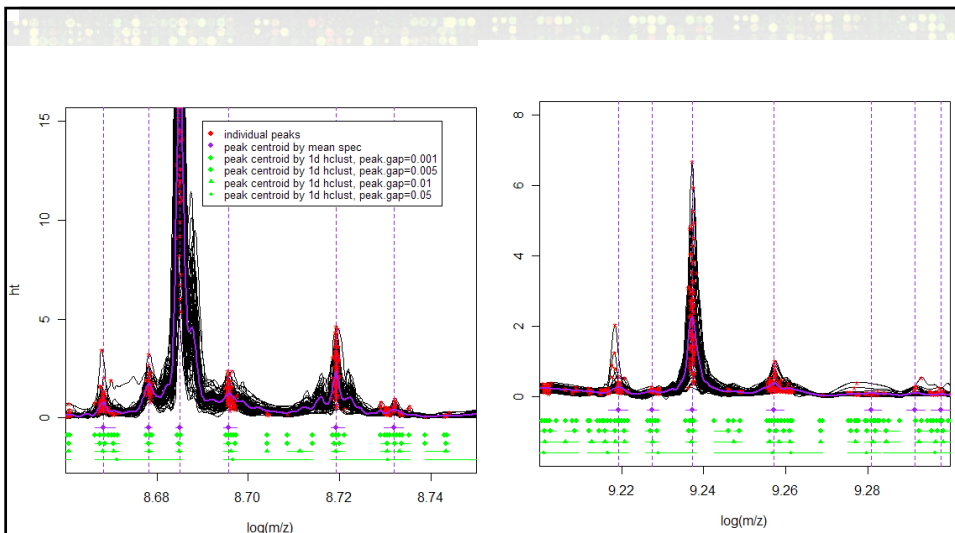
“Does this spectrum have a peak  
within the  $\pm\%m/z$  neighborhood of  
the centroid?”

Yes – this spectrum has a peak in  
this peak cluster; record the height

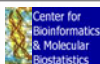
No – this spectrum has no peak in  
this peak cluster



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies



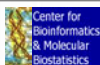
❑ 1-d hclust, locations of peak centroids sensitive to the magnitudes of mass accuracy, and other parameters used preprocessing



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

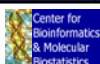
## Number of parameters involved in preprocessing

- Smoothing**
    - Thresholds
  - Baseline subtraction**
    - Bin size
  - Peak identification**
    - Bin size
    - S/N
  - Peak alignment**
    - Mass accuracy (%m/z shift)
- Parameter selection usually done by visual inspection**
  - Number of peaks highly dependent on algorithm chosen and parameters for each step**



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

	Algorithm	Any	<10%	10-50%	50-90%	>90%
<b>Peak Finding</b> S2N>1	Mean Spectrum	197	32	51	51	63
	1-d hclust, 0.001	2457	1848	492	94	23
	1-d hclust, 0.01	705	220	223	146	116
	1-d hclust, 0.05	282	25	56	56	145
<b>Peak Finding</b> S2N>2	Mean Spectrum	190	29	51	48	62
	1-d hclust, 0.001	2074	1524	439	90	21
	1-d hclust, 0.01	591	162	198	121	110
	1-d hclust, 0.05	243	21	44	44	145



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## After Preprocessing

	M/Z	Sample 1	Sample 2	Sample 3
Peak 1	1053.33	20	0	22
Peak 2	1100.20	2	3	2.5
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Peak N	8055.13	0	1	1.5

The aligned peak matrix can be used for classification, biomarker selection

Preprocessing completed

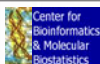


Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## Potential Limitations

### ❑ Bias of SELDI-TOF towards high abundance molecules

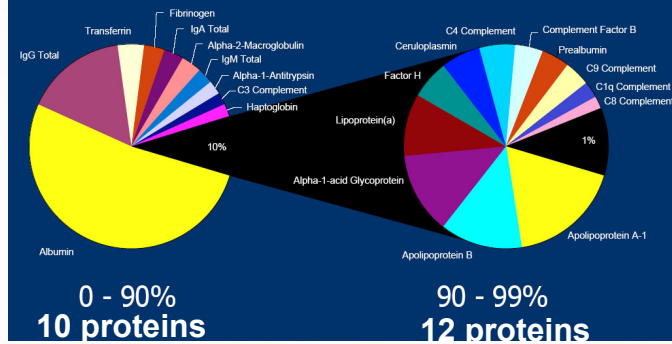
- Dynamic range of protein concentration in serum,  $\sim 10^{10}$
- Unlikely that low abundance molecules can compete with high abundance, non-informative molecules for binding
- 22 proteins constitutes 99% of proteins in serum



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

## Major Plasma Proteins

99% of plasma protein mass

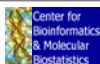


Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies

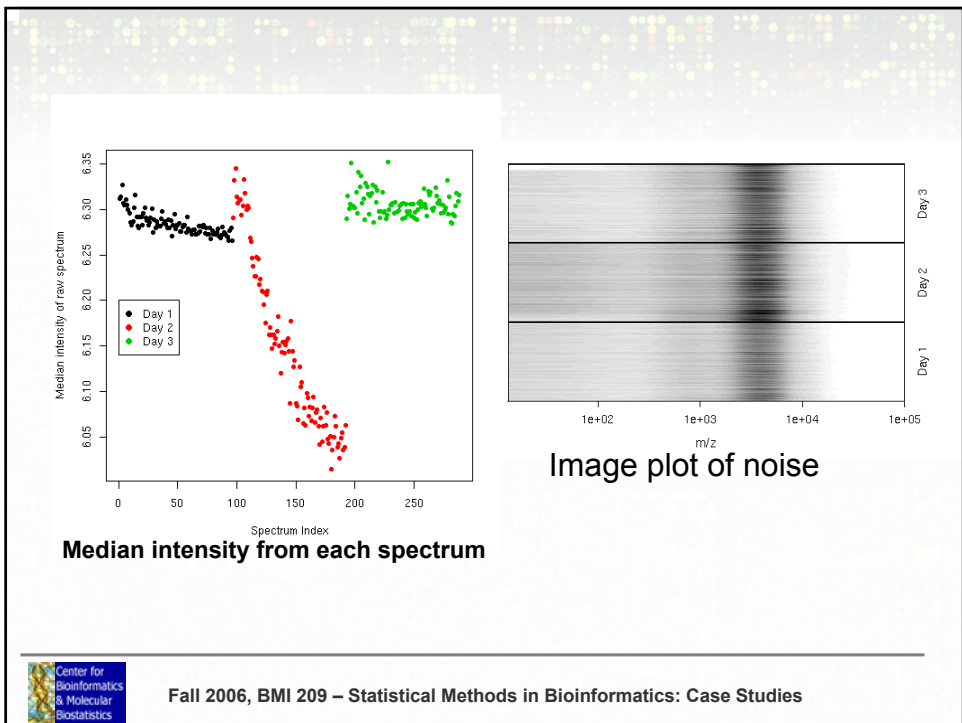
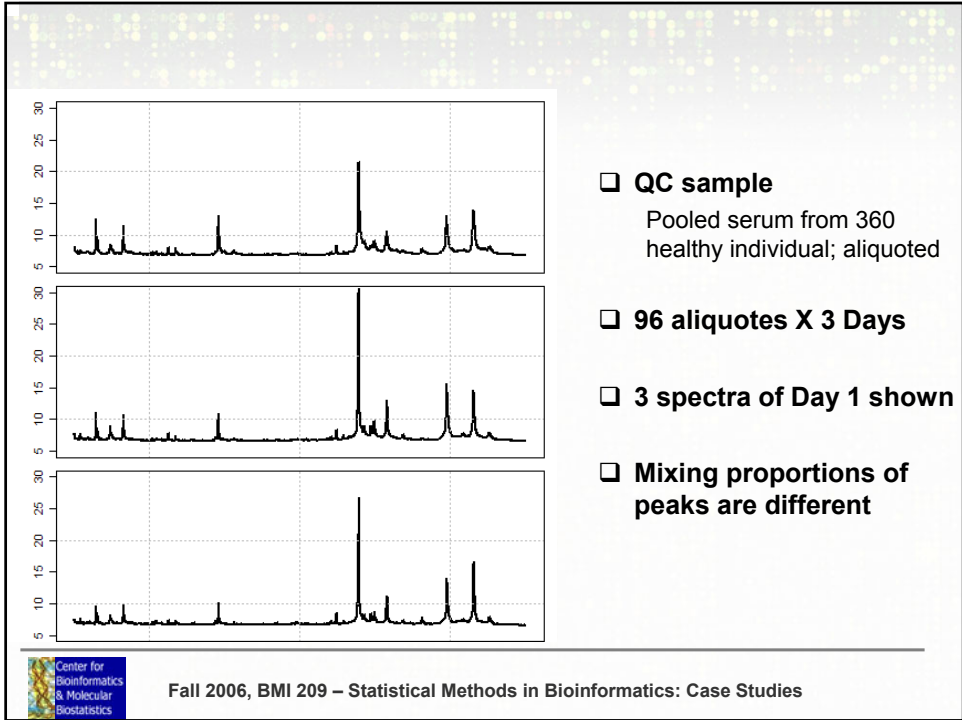
## Potential Limitations

### ❑ Lack of analytical reproducibility

- Sensitive to sample processing variables
  - ✓ Plasma or serum?
  - ✓ Aliquot storage materials
  - ✓ Number of freeze/thaw cycles
  - ✓ Time from blood collection to analysis
- Sensitive to different bioinformatic processing methods



Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies



**Table 2:** Comparison of 5 reports for prostate cancer diagnosis based on SELDI-TOF<sup>a</sup> technology

Study	Chip Type	Distinguishing Peaks M/Z <sup>b</sup>	Diagnostic Sensitivity and Specificity
Petricoin et al. (39)	Hydrophobic C16	2092, 2367, 2582, 3080, 4819, 5439, 18220	95%; 78-83%
Adam et al. (40)	IMAC-Cu	4475, 5074, 5382, <b>7024</b> , 7820, 8141, 9149, 9507, <b>9656</b>	83%; 97%
Qu et al. (41)	IMAC-Cu	<u>Non-cancer vs cancer</u> 3963, 4080, 6542, 6797, 6949, 6991, <b>7024</b> , 7885, 8067, 8356, <b>9656</b> , 9720, <u>Healthy vs BPH</u> 3486, 4071, 4580, 5298, 6099, 7054, <b>7820</b> , 7844, 8943	97-100%; 97-100%
Banez et al. (42)	WCX2	3972, 8226, 13952, 16087, 25167, 33270	63%; 77%
	IMAC-Cu	3960, 4469, 9713, 10266, 22832	66%; 38%
Lehrer et al. (43)	Hydrophobic H4	<u>Cancer and BPH vs Controls</u> 15200, 15900, 17500 <u>Cancer vs BPH</u> 15200 15900 17500	100% (specificity) 82%; 67% 82%; 100% 64%; 67%

- Same patients
- Same lab
- Same data
- Two different bioinformatic methods

Diamandis, E. MCP 2004 February 28



## Acknowledgements

**Mark Segal**  
**Mark Pletcher**  
**Jeff Tice**



**Next Week:**

**Dr. Haiyan Huang (Berkeley)**  
**A statistical framework to infer functional gene  
associations from multiple biologically  
interrelated microarray experiments: application to  
yeast and human data**



Center for  
Bioinformatics  
& Molecular  
Biostatistics

Fall 2006, BMI 209 – Statistical Methods in Bioinformatics: Case Studies