

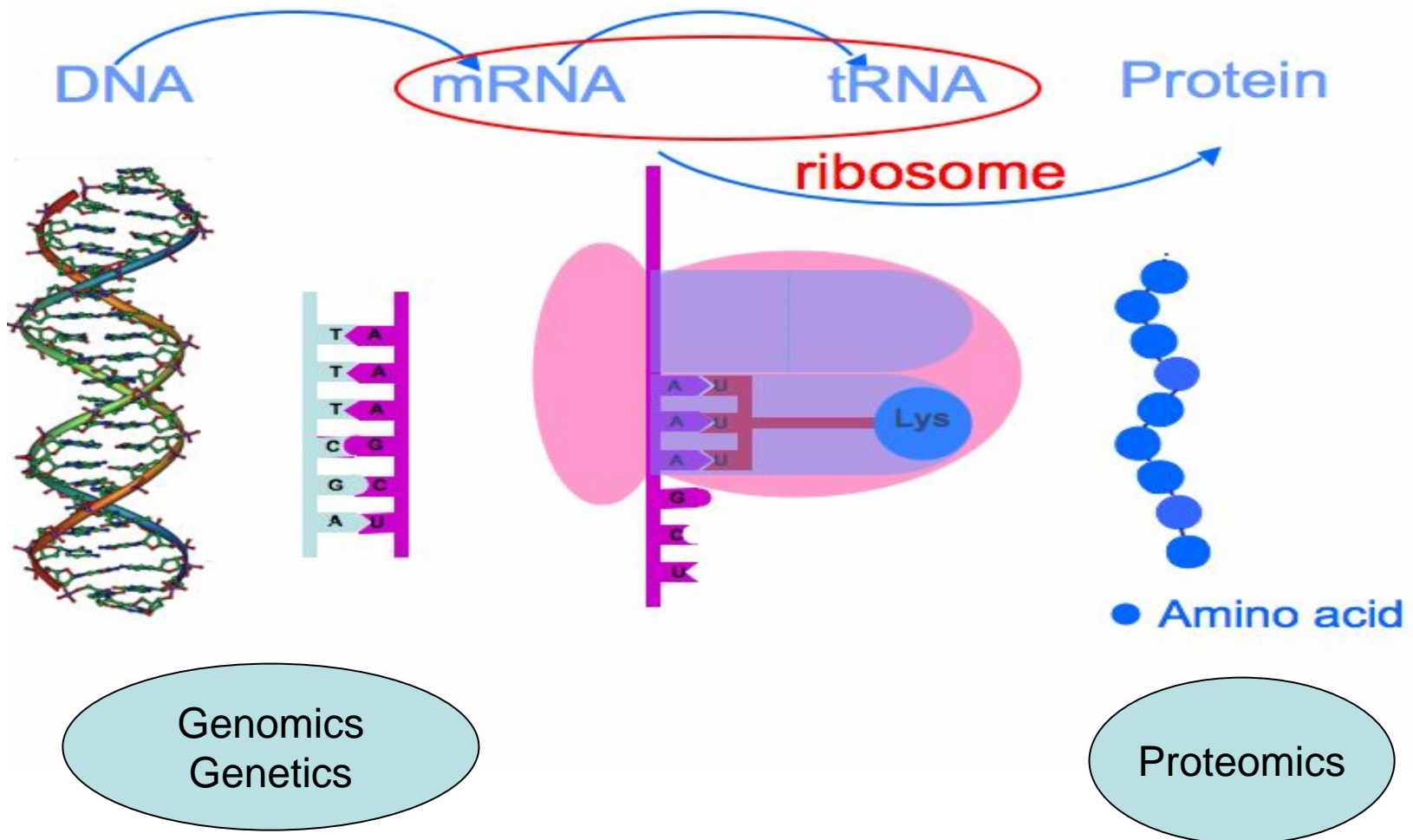
Analysis of MALDI-TOF Data: from Data Preprocessing to Model Validation for Survival Outcome

Heidi Chen, Ph.D.
Cancer Biostatistics Center
Vanderbilt University School of Medicine
March 20, 2009

Outline

- MALDI-TOF
- Data preprocessing for raw spectra
- Build a prediction model from training set
- Model validation

Biology 101

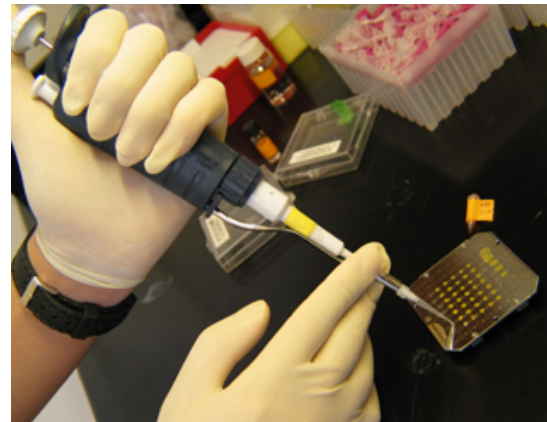
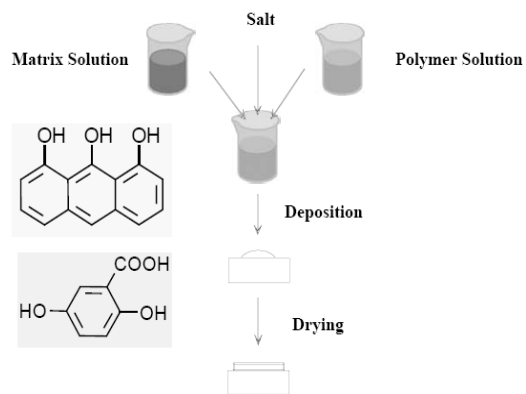


MALDI-TOF



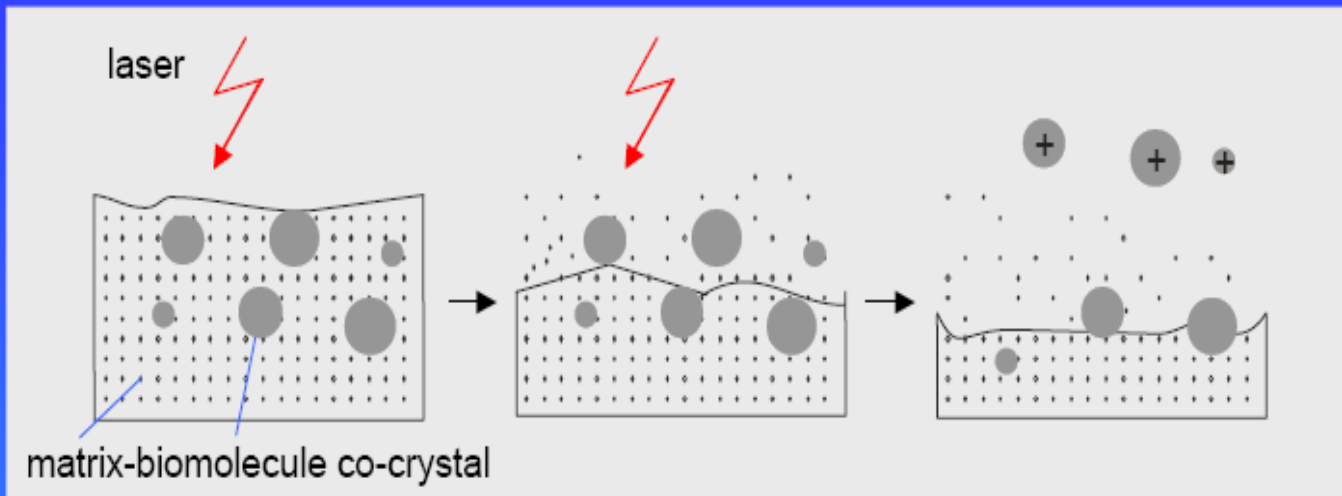
Step 1: Sample preparation

Sample Preparation for MALDI-TOF MS

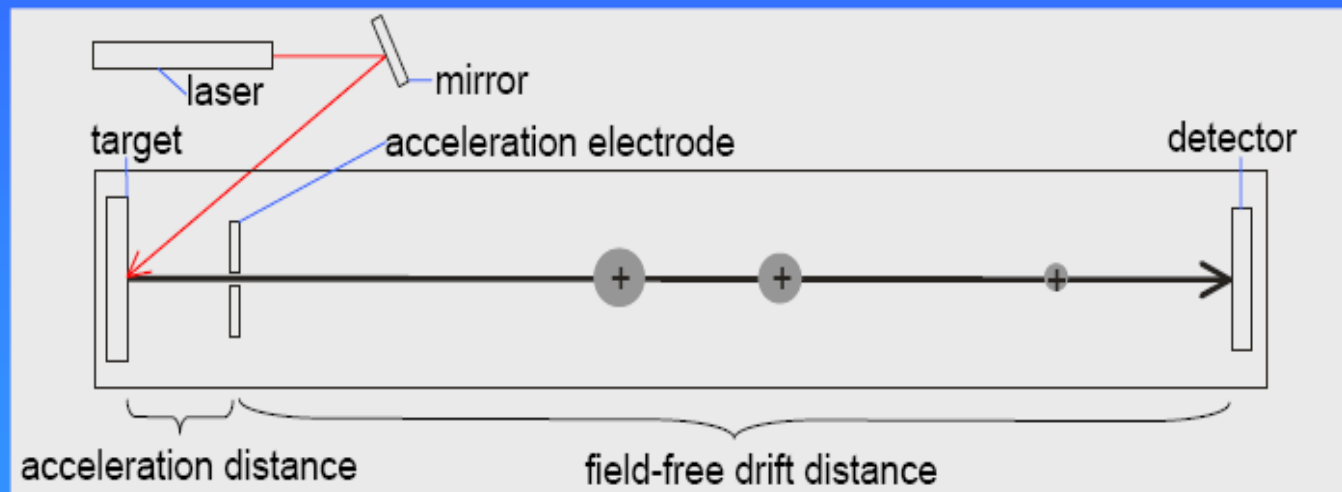


MALDI-TOF mass spectrometry

Matrix
Assisted
Laser
Desorption
Ionization



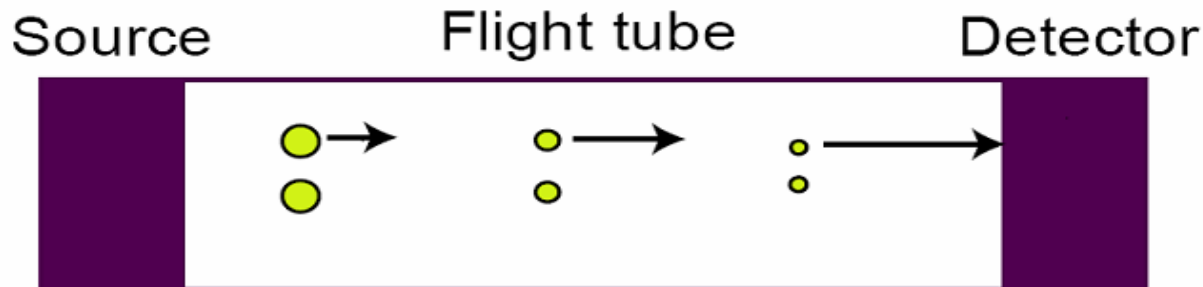
Time
Of
Flight



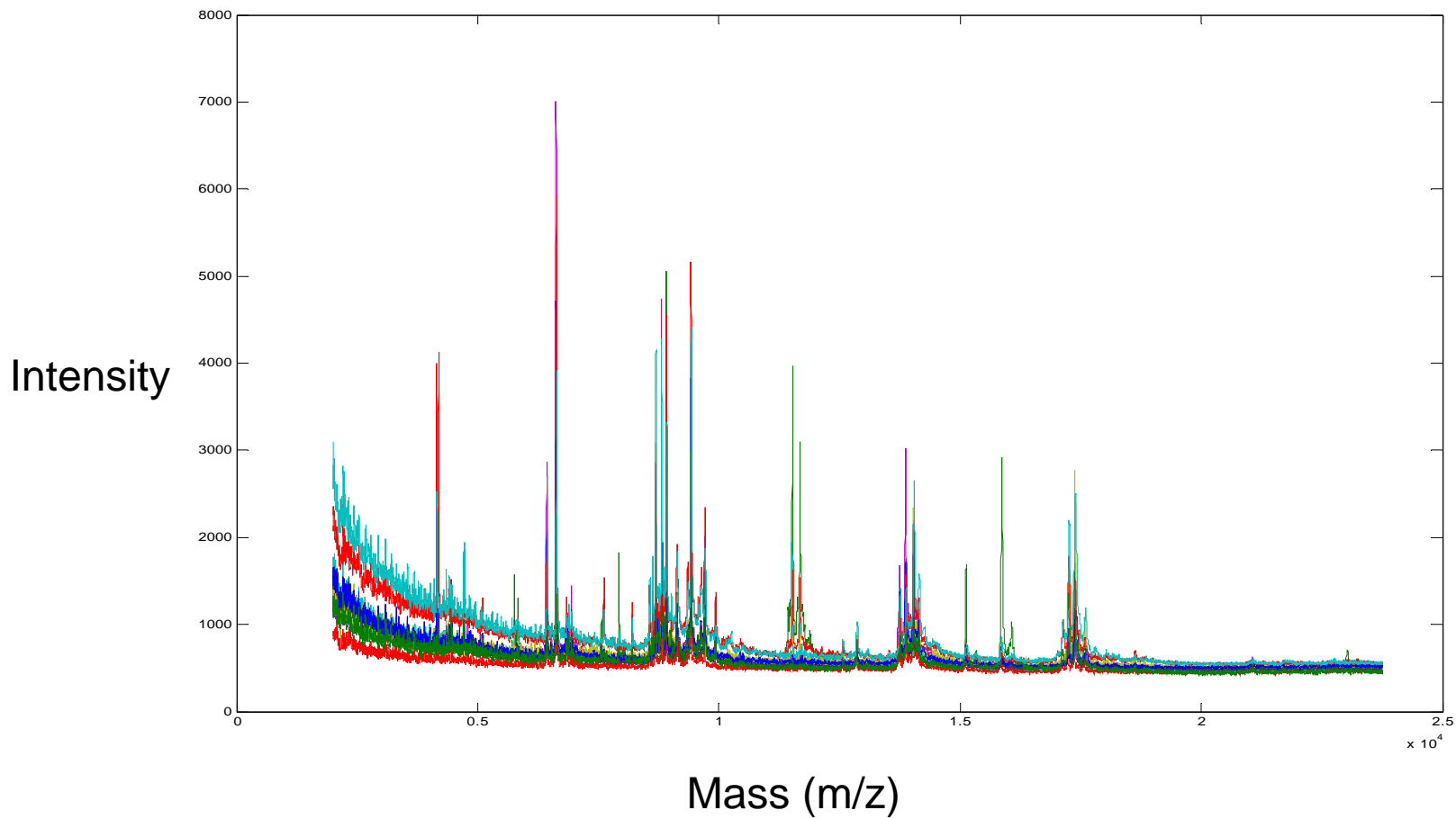
Principle Idea of MALDI-TOF MS

- Upon laser irradiation all molecules obtain similar energy
- Convert electric energy to kinetic energy
- **Time Of Flight (TOF)** separates ions based on size(mass/charge, m/z)

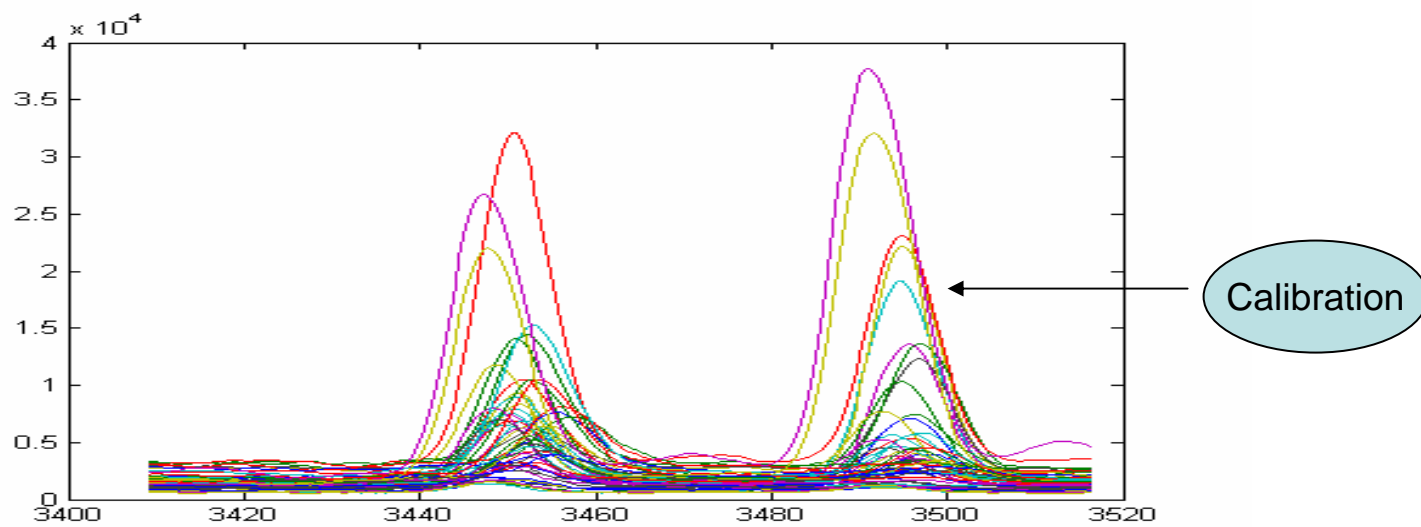
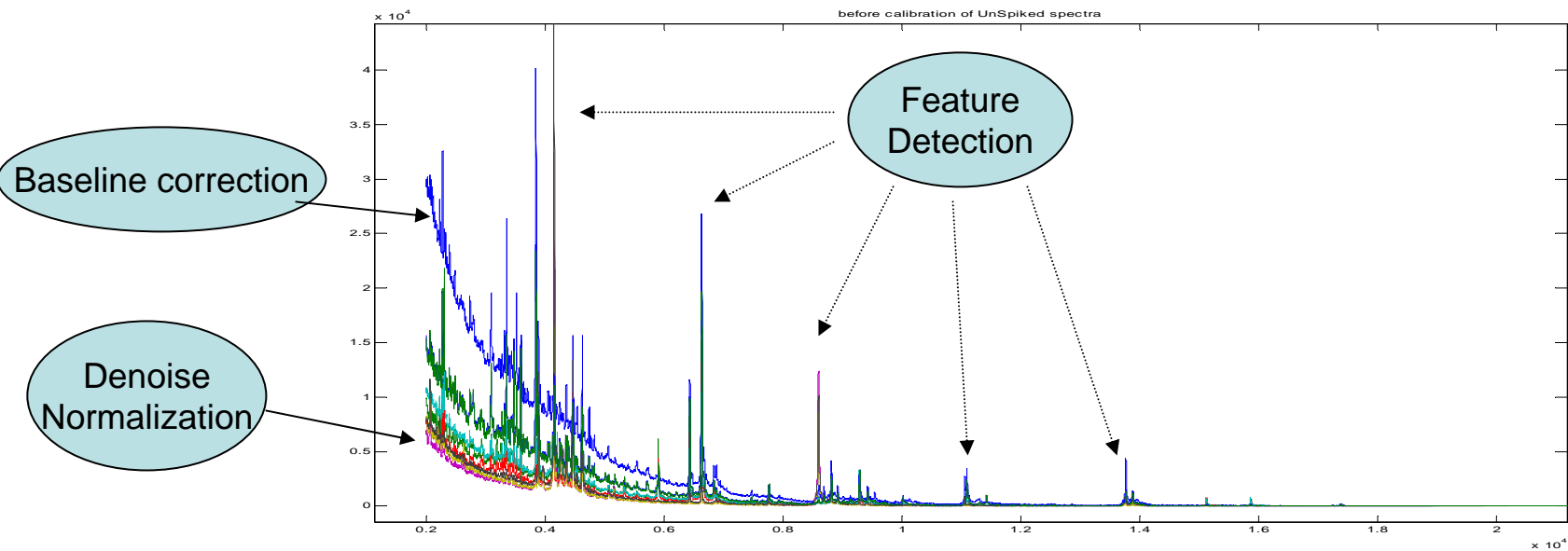
$$\text{TOF} : (m/z)^{1/2}$$



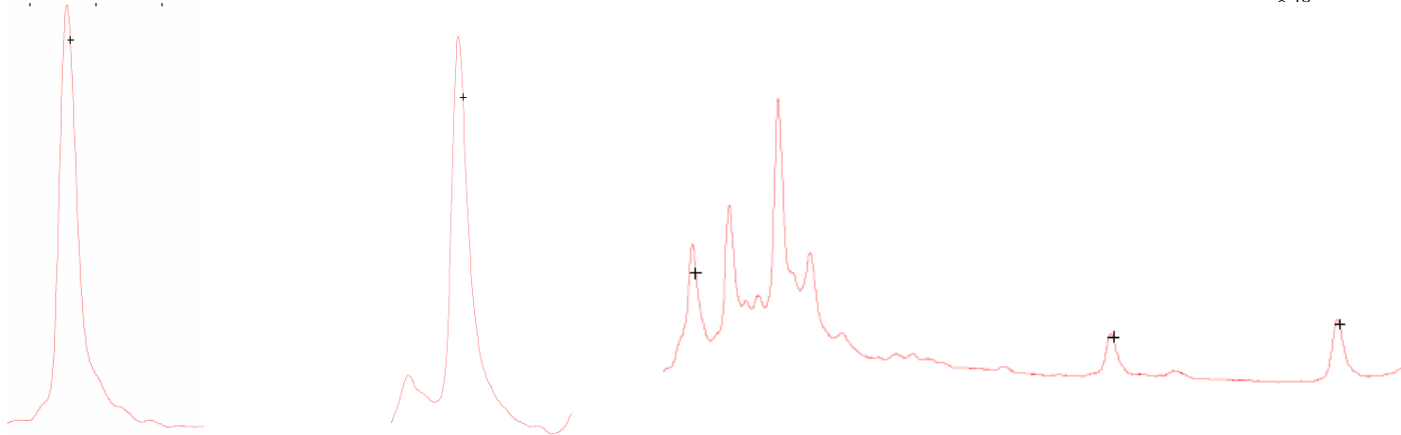
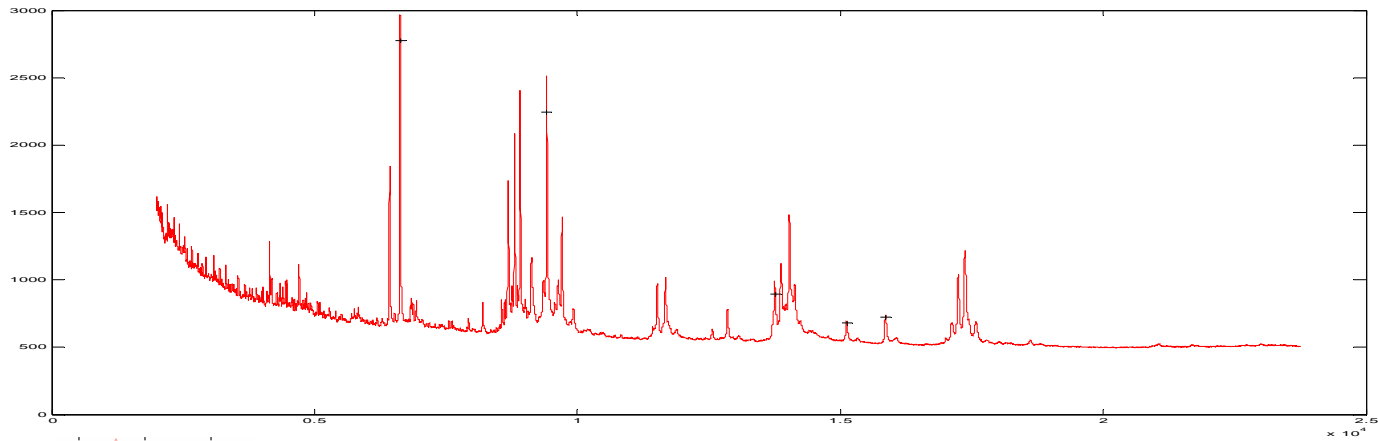
Raw Spectra



- MALDI-TOF
- **Data preprocessing for raw spectra**
- Build a prediction model from training set
- Model validation



Peak Calibration



- (1) m/z values around some known proteins
- (2) show clear bell-shape

Convolution Based Calibration

- Calibrate each spectrum with the known peaks. Max $h(t)$ happens when f and g overlap the most.
- The optimum shift is obtained by maximizing the sum of convolution values on the multiple peak locations.

Note: all process are on the time domain.

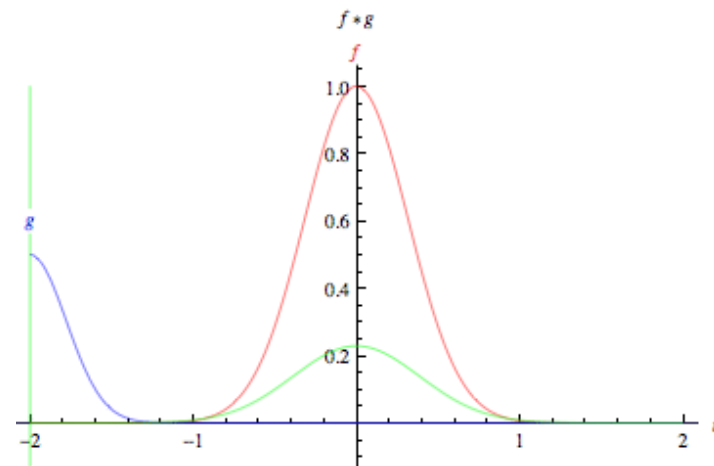
shift to right : $t_{\text{original}} < t_{\text{max}}$

shift to left : $t_{\text{original}} > t_{\text{max}}$

keep the same : $t_{\text{original}} = t_{\text{max}}$

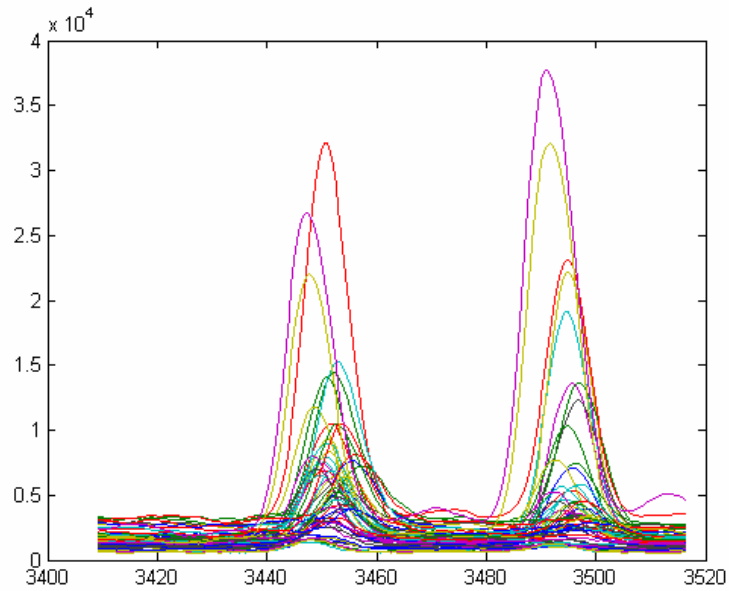
$f(t)$: observed peak

$g(t)$: ideal peak (normal distribution)

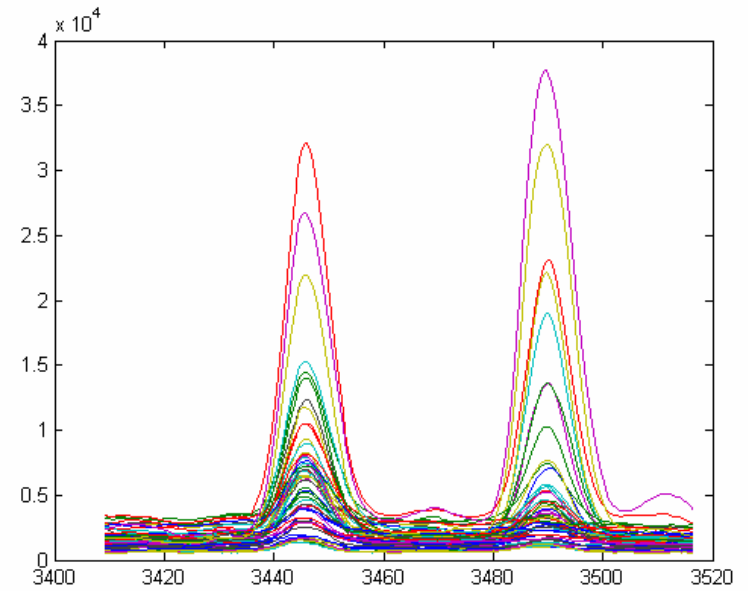


$$h(t) = (f * g)(t) \equiv \int_{t_1}^{t_2} f(\tau)g(t-\tau)d\tau = \int_{t_1}^{t_2} f(t-\tau)g(\tau)d\tau$$

Calibration

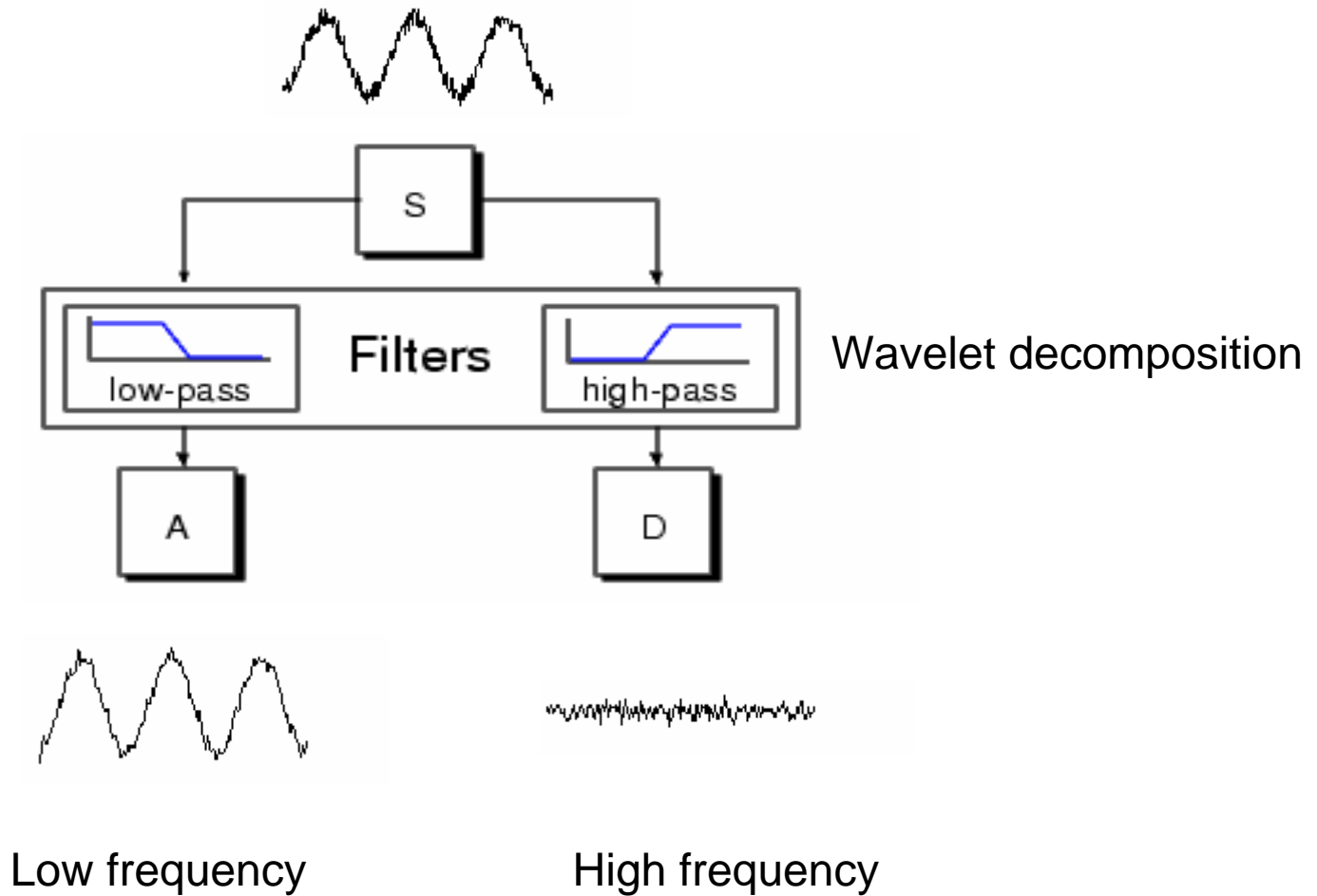


Before Calibration



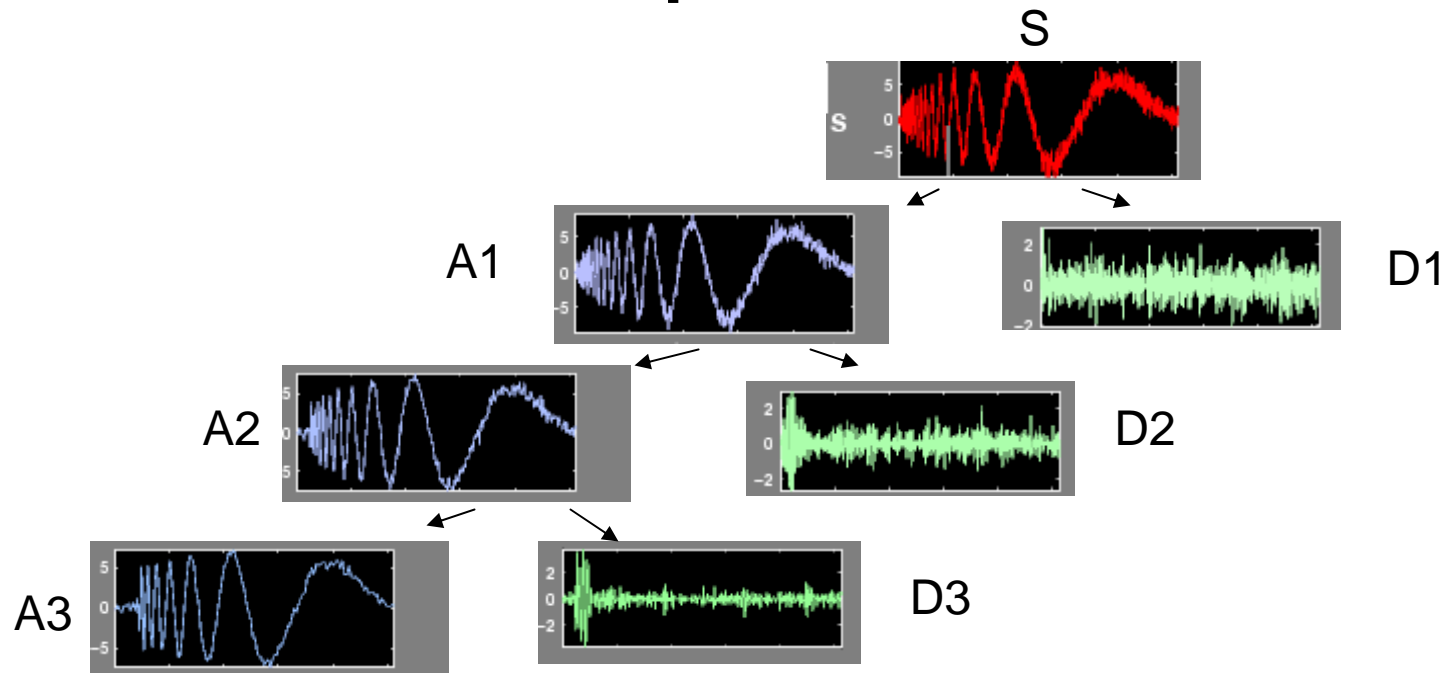
After Calibration

Wavelet Denoising



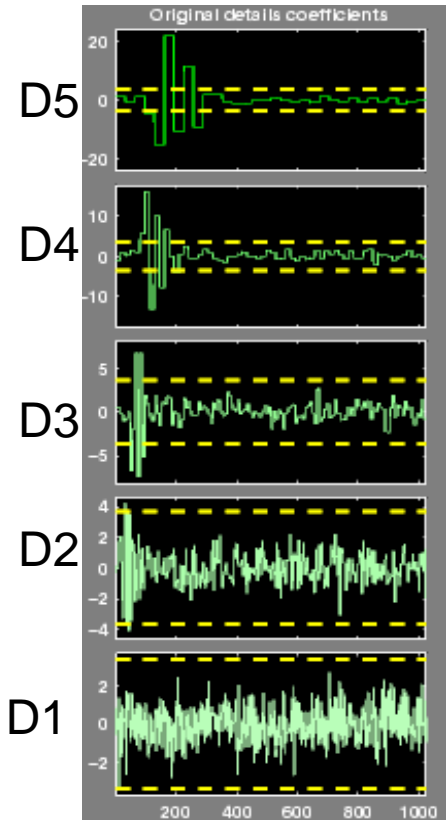
$$S = A + D$$

Wavelet Decomposition Tree

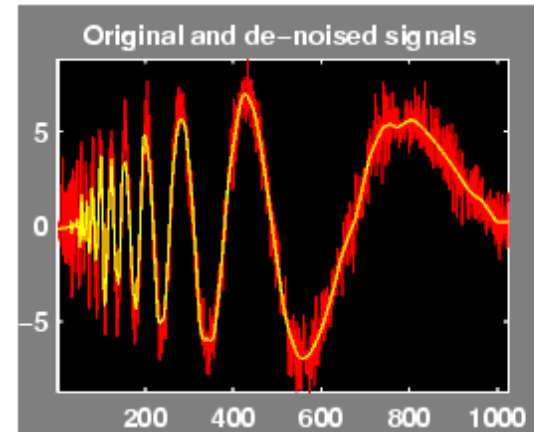


$$\begin{aligned} S &= A_1 + D_1 \\ &= A_2 + D_2 + D_1 \\ &= A_3 + D_3 + D_2 + D_1 \end{aligned}$$

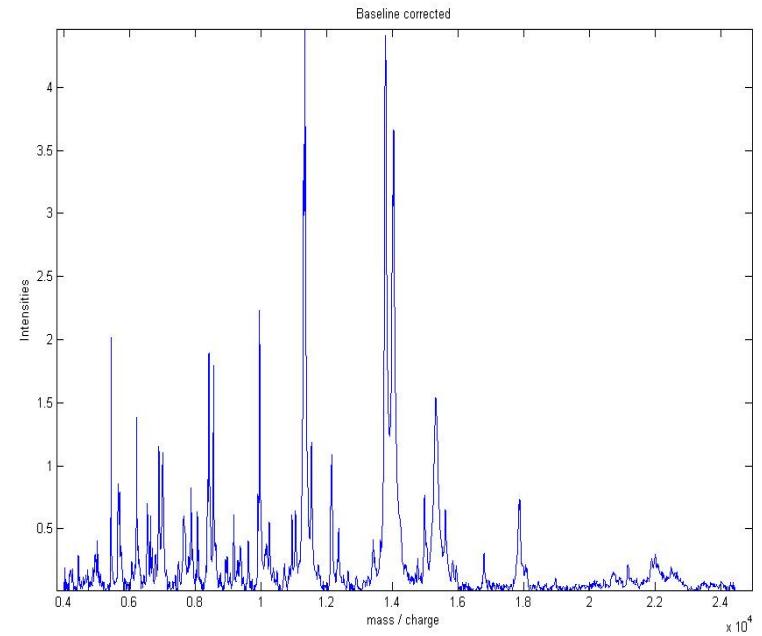
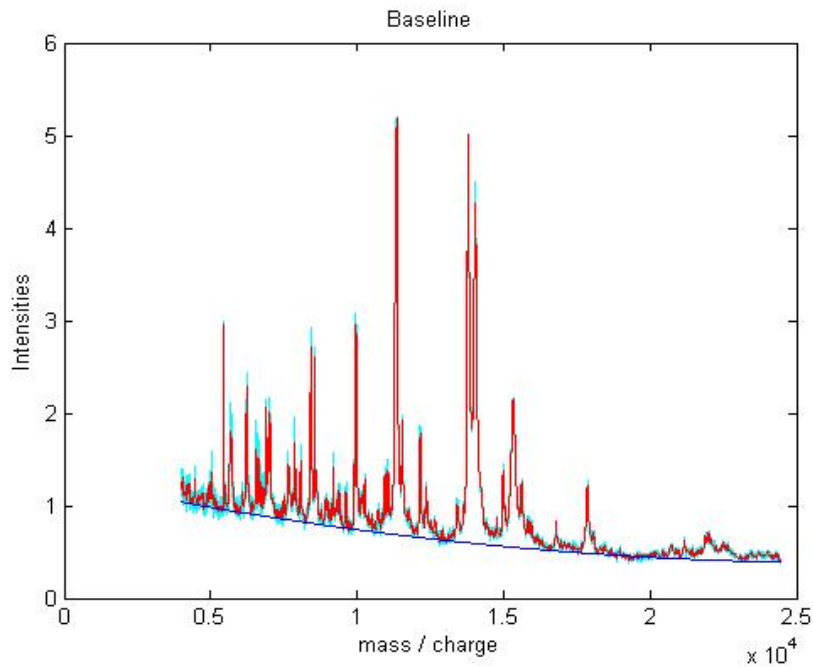
Wavelet Denoise



Remove noise by
thresholding



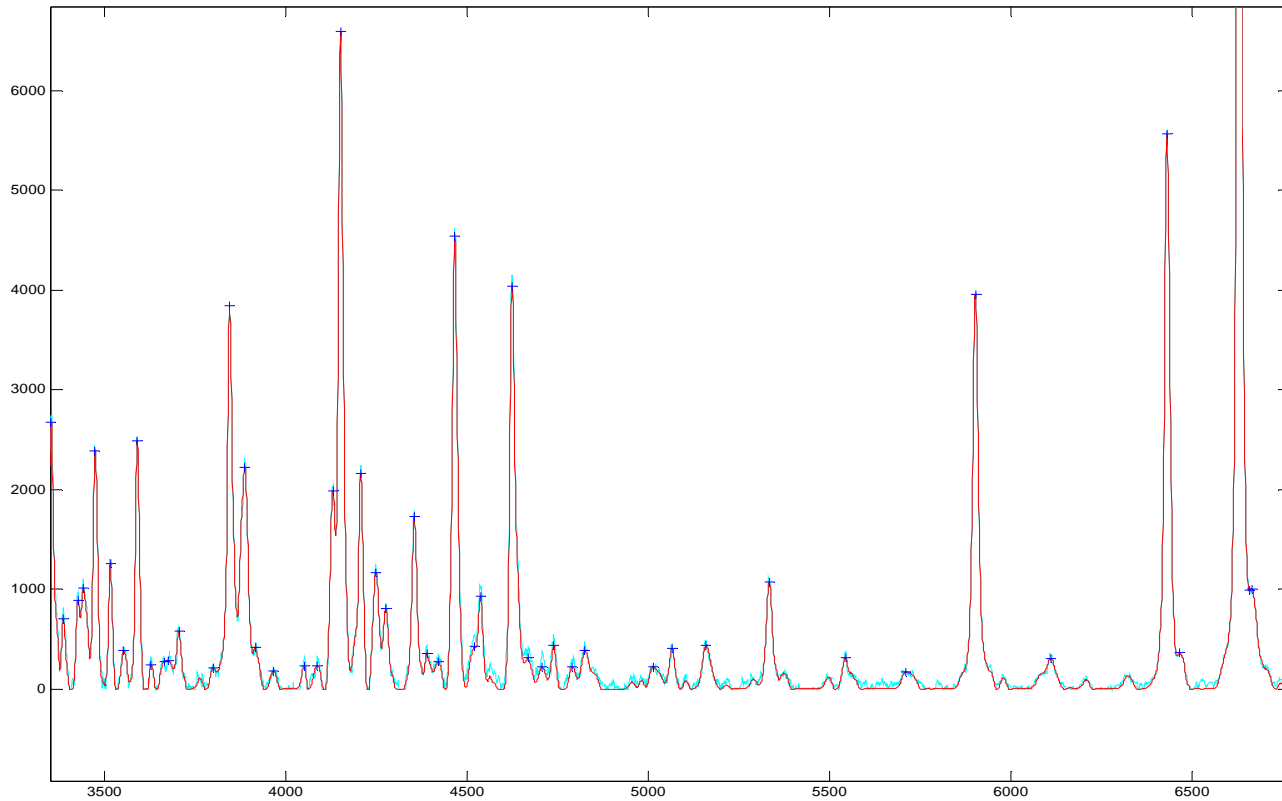
Baseline Correction

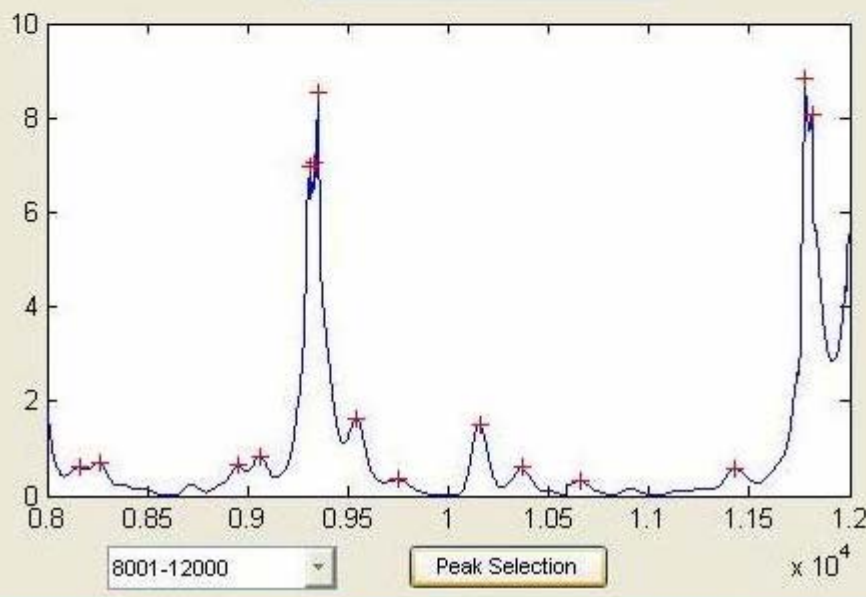
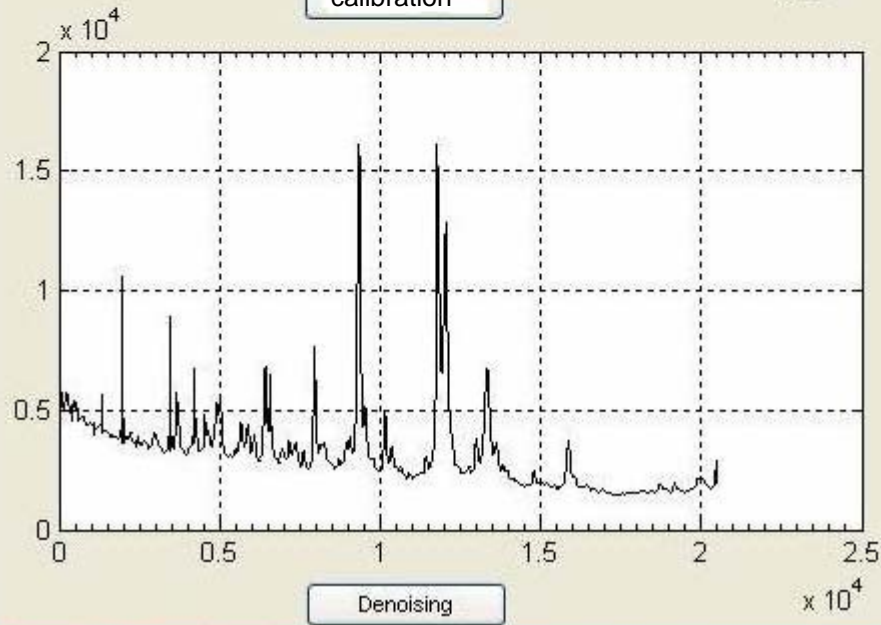
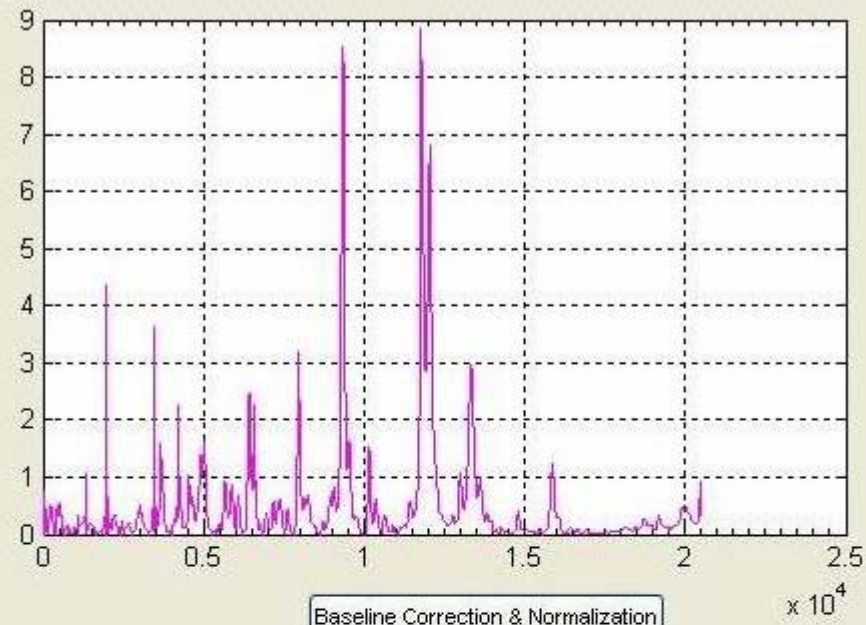
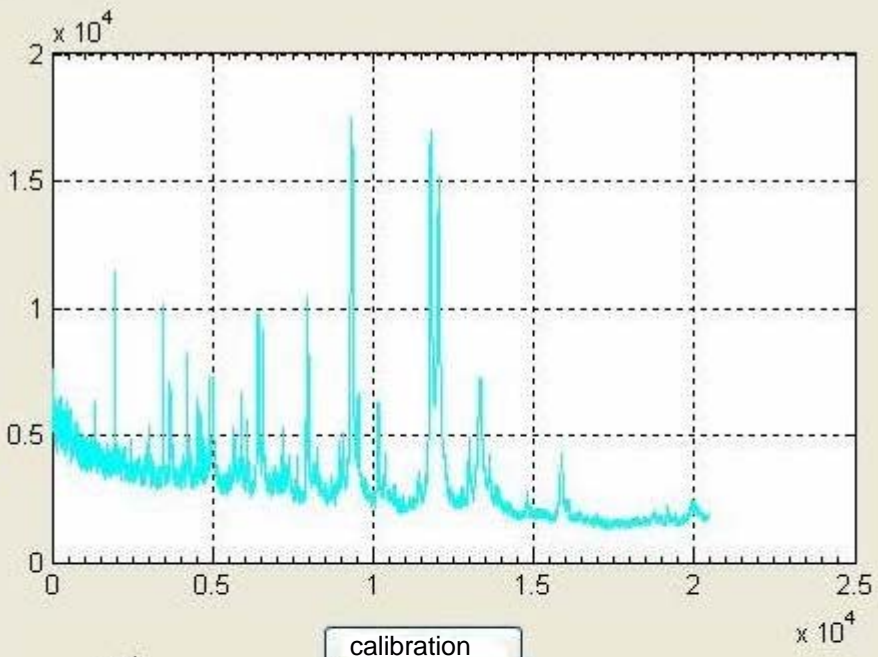


Spline curve to fit the local minima

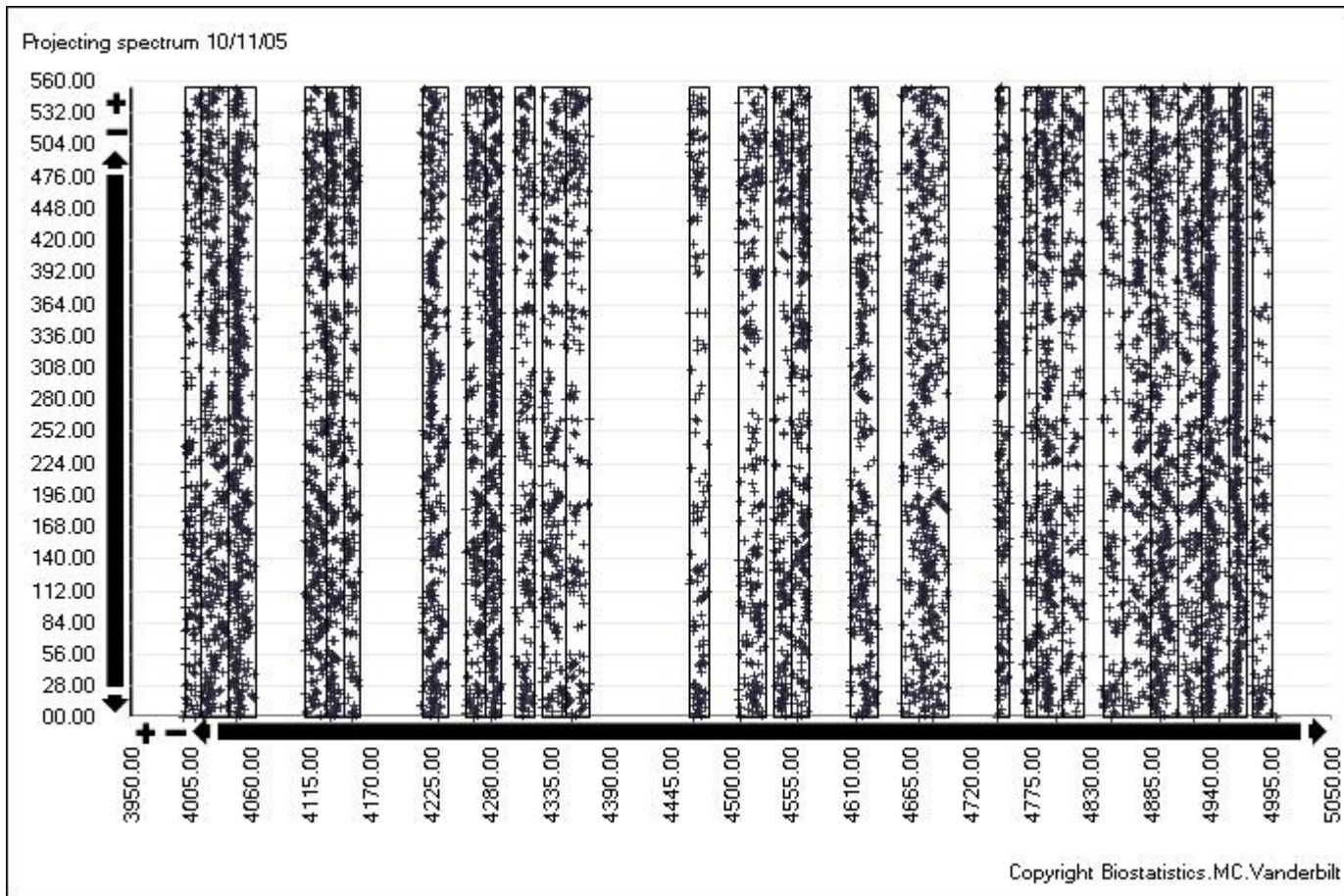
Peak Detection

- (1) local maxima
- (2) pass signal/noise cutoff to filter out small peaks



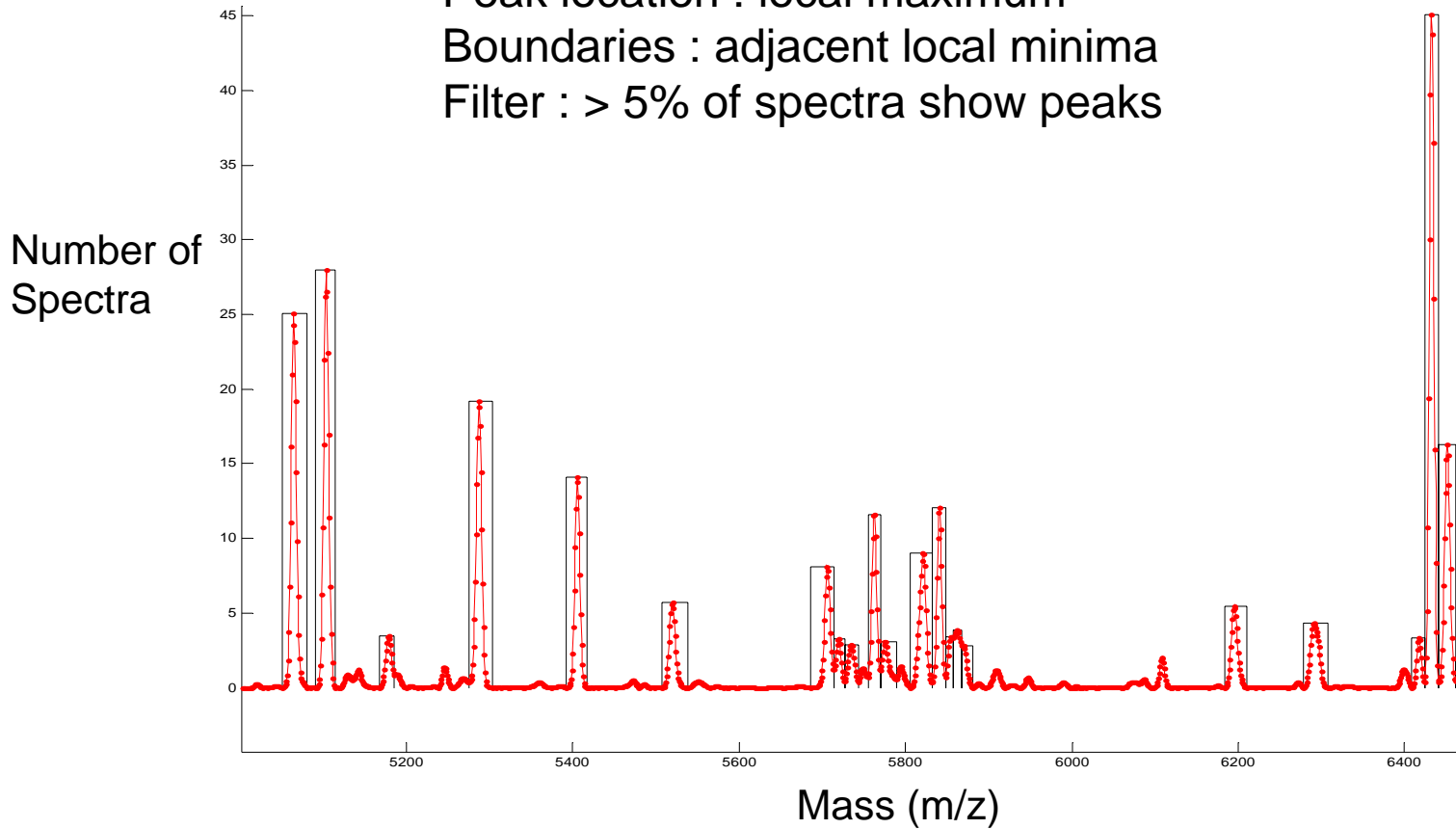


Peak Distribution



Common Peaks Finding

Peak location : local maximum
Boundaries : adjacent local minima
Filter : > 5% of spectra show peaks



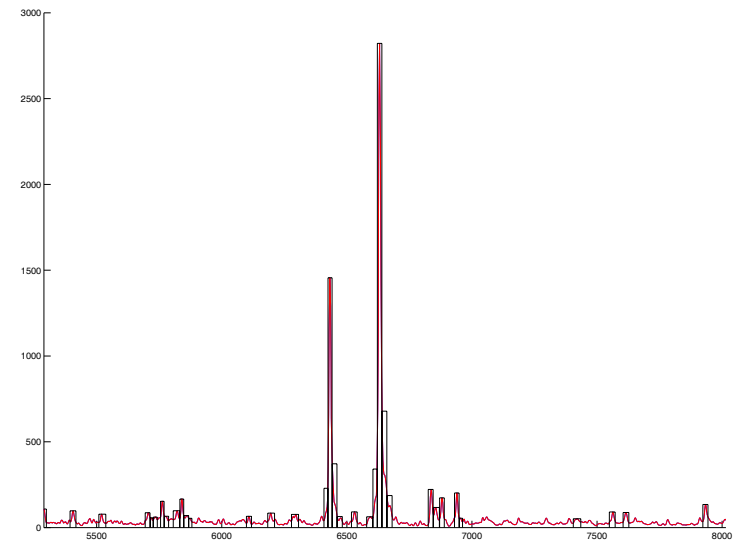
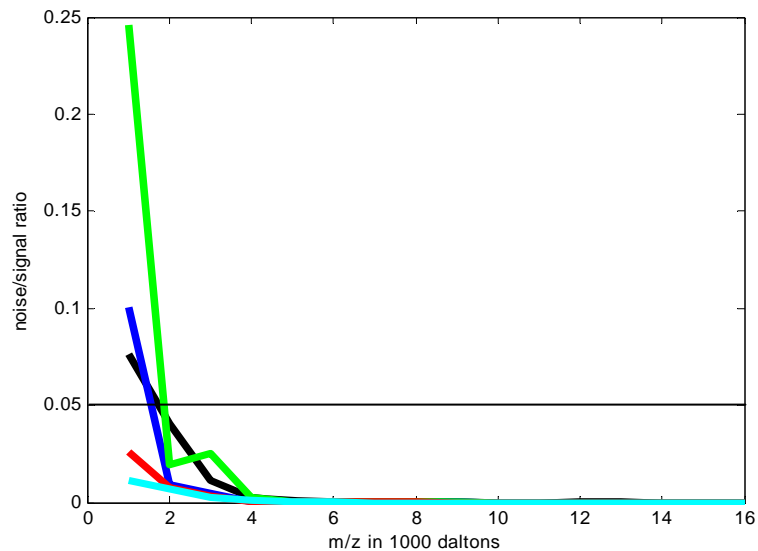
Kernel Density of Peaks Distribution

Feature Quantification

- Normalization : standardize the AUC for all spectra to the median AUC
- Peak intensity: AUC within peak boundaries

Threshold Selection

- Denoise: wavelet threshold
- Peak detection: signal to noise ratio
- Common peak finding: bandwidth of KDS



Training Set

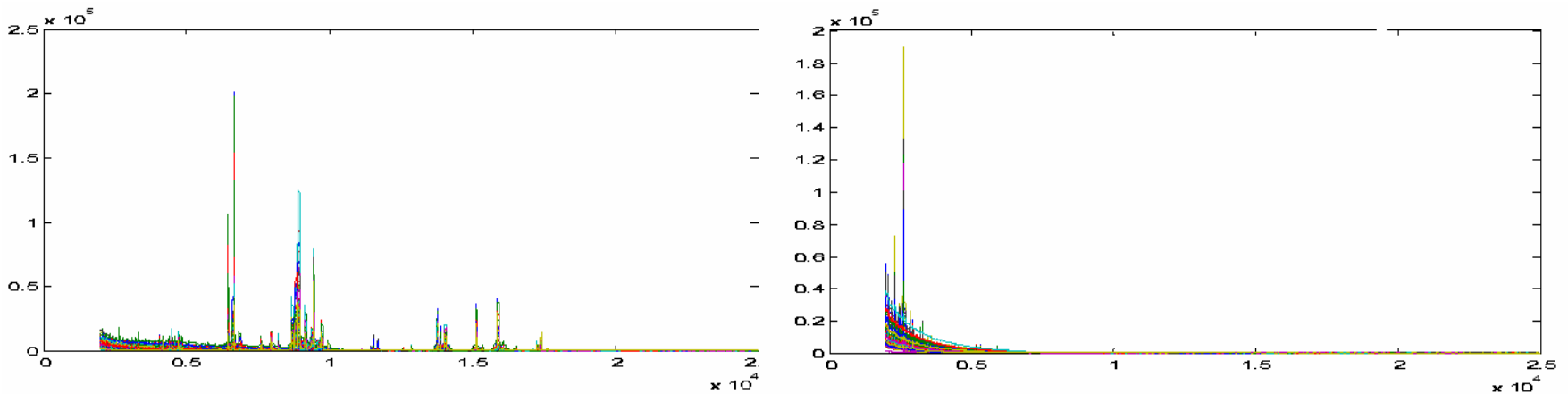
- 35 pretreatment (EGFR+VEGF) serum samples from stage IIIB/IV NSCLC; 14 male, 21 female; age range 36-72
- Experimental Design

	Day1	Day2	Day3
Blood sample	pt1-pt35	pt1-pt35	pt1-pt35
Replication	3 replications each pt	3 replications each pt	3 replications each pt

105 samples were **randomly** spotted in two 64-well plates each day.

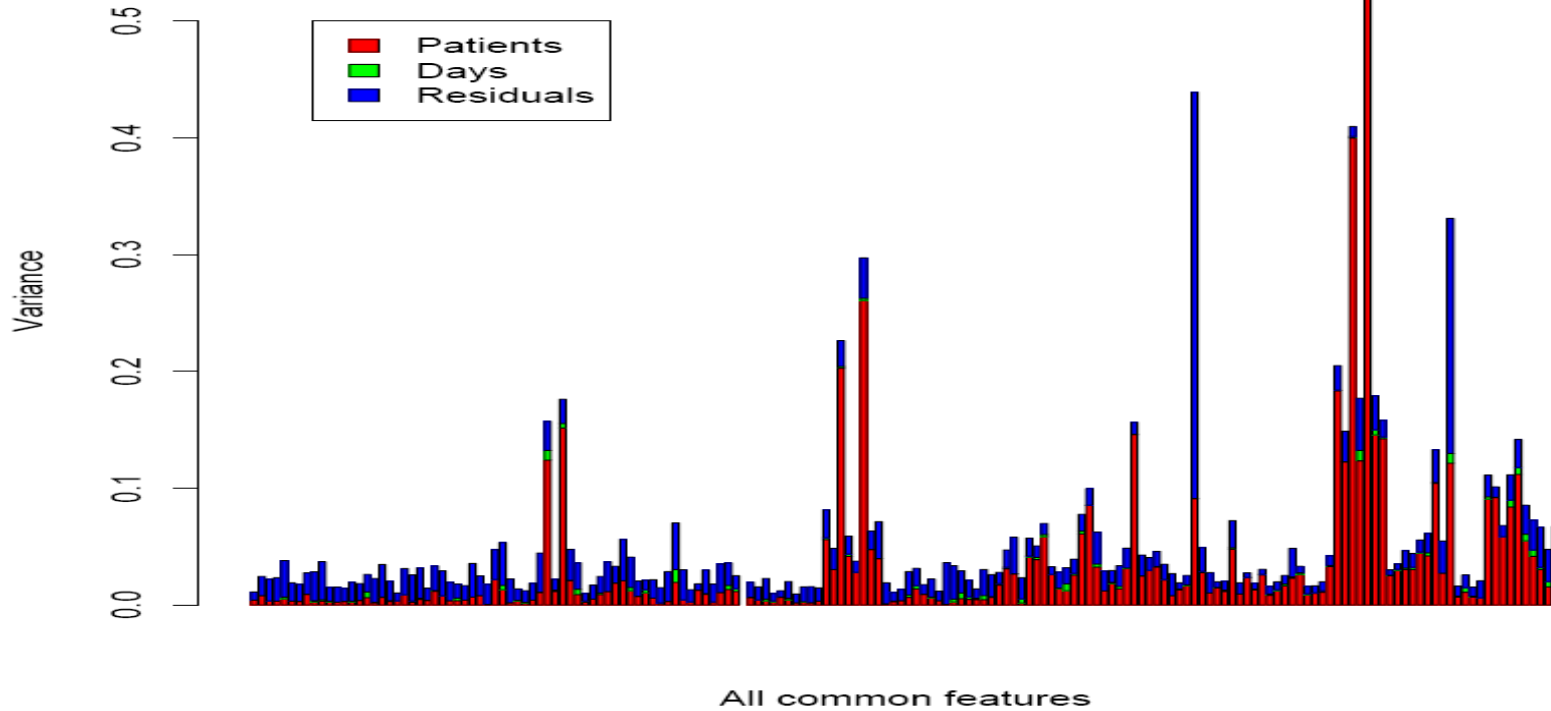
Training Set

290 good spectra; 25 bad spectra



174 features (3000-20,000 m/z) after data preprocessing

Variance Components



- MALDI-TOF
- Data preprocessing for raw spectra
- **Build a survival prediction model from training set**
- Model validation

Procedure of Constructing Survival Prediction Model from Training Set

(1) Feature selection

- CPH model

$$h_i(t | x_{j,i}) = h_{j,0}(t) \exp(\beta_j x_{j,i})$$

$x_{j,i}$: intensity for feature j of patient i

- FDR cutoff 0.05 : 11 features associated with survival

(2) Create a compound score as a prediction index

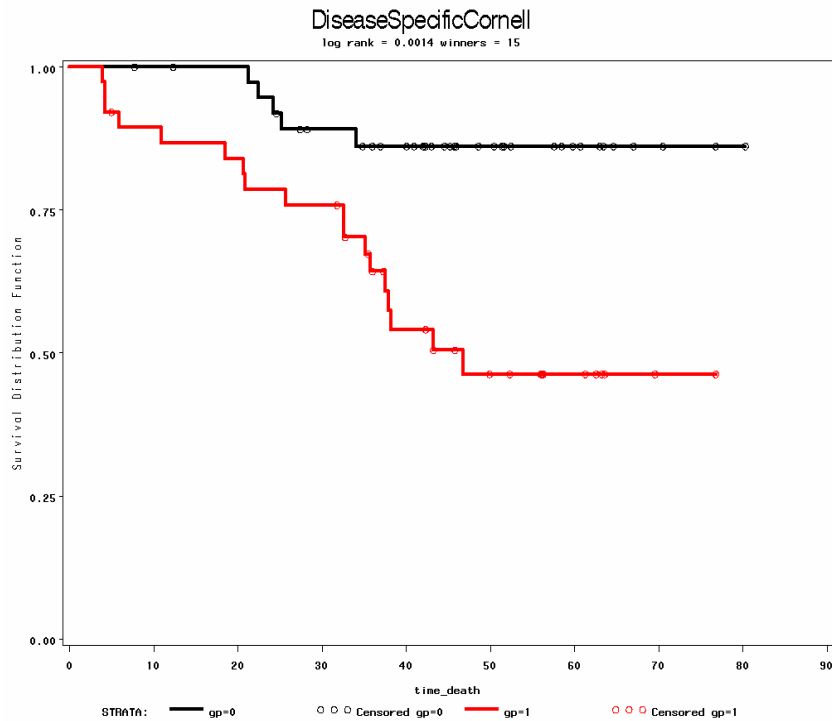
$$s_i = \sum_{j=1}^k (\text{sign of } \beta_j) w_j x_{j,i} \quad w_j : \text{Wald statistics}$$

(3) Prediction model : predicted hazard rate

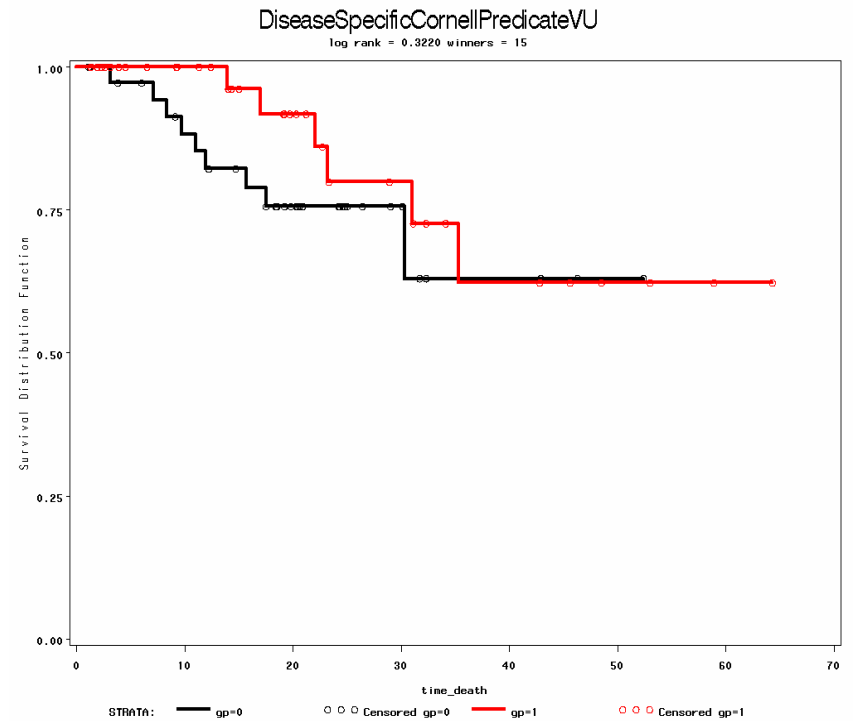
$$h_i(t | s_i) = h_0(t) \exp(0.0235 s_i)$$

- MALDI-TOF
- Data preprocessing for raw spectra
- Build a prediction model from training set
- **Model validation**

Overfitting



Training set



Independent test set

Model Validation

Goal : evaluate how the model performs in the future dataset

- External validation: independent test set
- Internal validation: training set

External Validation

Data Preprocessing for Independent Test Set

- Calibration
- Baseline correction
- Wavelet denoise
- Normalization : median AUC from training set
- Peak location and boundaries from training set

Prediction for Independent Test Set

Freeze w_1, w_2, \dots, w_k and φ from training set

Compound score as a predictor

$$s'_i = \sum_{j=1}^k (\text{sign of } \beta_j) w_j x'_{j,i} \longrightarrow \text{Association with survival outcome.}$$

CPH : predicted hazard rate

$$h'_i(t) = h_0(t) \exp(0.0235 s'_i)$$

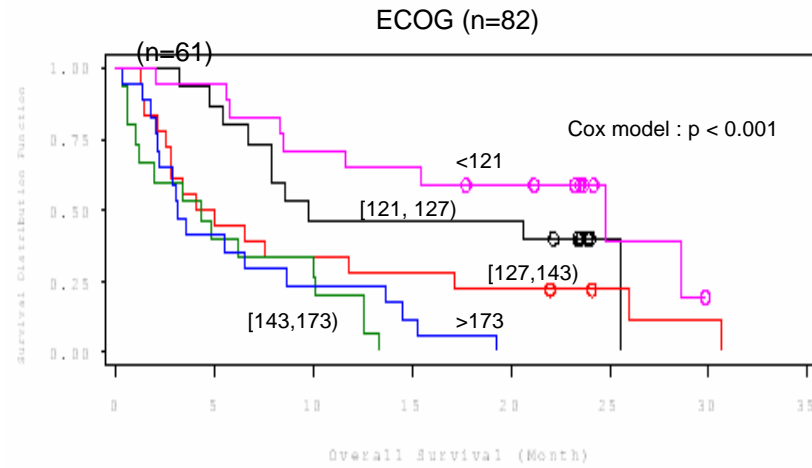
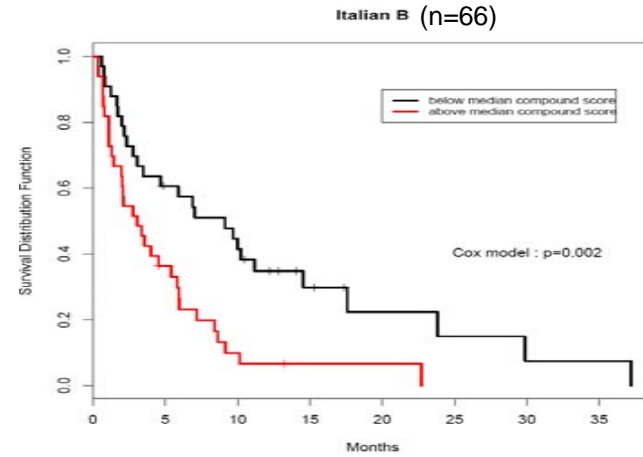
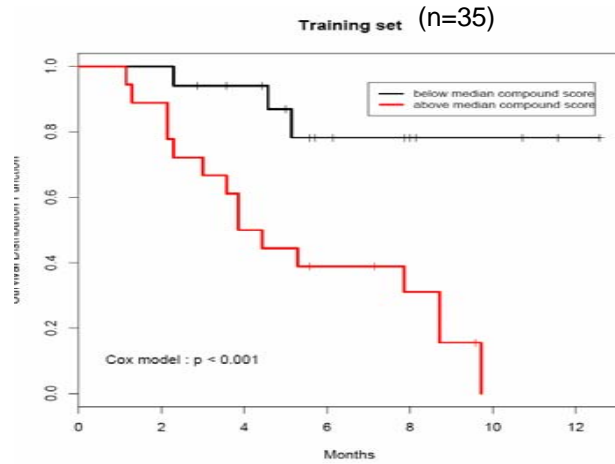
Validation of Predictive Model

- What to validate ?
 - Predictive ability
- C-index (Harrel et al. 1982) : measure the agreement between predicted and observed survival time for two subjects.
C-index ranges from 0 to 1.
1: perfect prediction; 0.5 :random prediction;
0: opposite prediction

Independent Test Sets

- Eastern Cooperative Oncology Group (ECOG) (n=82)
- Italian (n=66)

KM Survival



C-index

Training set	0.77
Test set	0.62 (Italian)

Training C-index : the C-index on the training set

Generalized C-index : the C-index on the independent test set

Internal Validation

Goal : estimate the generalized C-index through the internal validation process

- Data splitting
- K-fold cross validation
- Bootstrap

Focus on the procedure after data preprocessing

Data Splitting



- (1) Build a prediction model based on training set
- (2) Compound score for test set: winners and Wald statistics from training set
- (3) Generalized C-index: calculate C-index from test set

K-fold Cross Validation

1. Test	2. Train	3. Train	4. Train	5. Train
---------	----------	----------	----------	----------

1. Train	2. Test	3. Train	4. Train	5. Train
----------	---------	----------	----------	----------

1. Train	2. Train	3. Test	4. Train	5. Train
----------	----------	---------	----------	----------

1. Train	2. Train	3. Train	4. Test	5. Train
----------	----------	----------	---------	----------

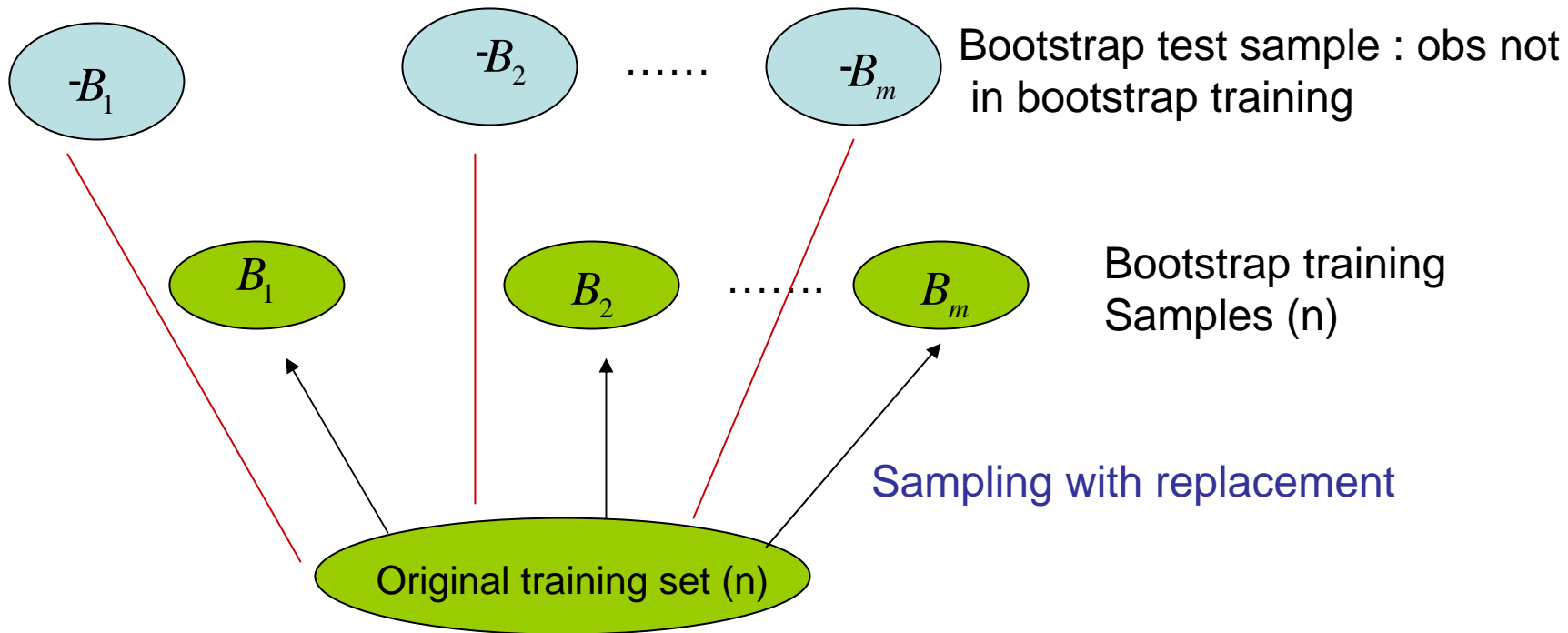
1. Train	2. Train	3. Train	4. Train	5. Test
----------	----------	----------	----------	---------

Sample size n
divided into $K = 5$ parts,

C-index for each test set

Combine C-index from all test sets to get the estimate the generalized C-index

Bootstrap



Bootstrap

- Generalized C-index

$$C^{(0.632+)} = (1-W) C_{training} + W \frac{1}{m} \sum_{i=1}^m C_{T_i(test)}$$

$$\gamma \text{ (non - informative C - index)} = 0.5$$

$$R \text{ (relative overfitting rate)} : \frac{\frac{1}{m} \sum_{i=1}^m C_{T_i(test)} - C_{training}}{\gamma - C_{training}} \quad R \in [0, 1]$$

$$W(\text{weight}) = \frac{0.632}{1-0.368R} \quad w \in [0.632, 1]$$

C-index(0.632+) ranges from C-index(0.632) if there is minimum overfitting (R=0) to $\frac{1}{m} \sum_{i=1}^m C_{T_i(test)}$ if there is maximum overfitting (R=1)

C-index

Method	NSCLC (n=35)	Overfit Example (n=77)
Training set	0.77	0.69
Bootstrap	0.71	0.53
Indep test set	0.62 (Italian)	0.28

Reproducibility of MALDI-TOF MS

Winners of Case Study (11) EGFR+VEGF	Winners of JNCI 2007 (8) EGFR
4121	5843
4596	11445
4720	11529
4821	11685
5720	11759
5841	11903
11441	12452
11528	12579
11684	
11731	
11902	

Take Home Message

- Good experimental design
- Precisely follow the protocol of MALDI-TOF
- MALDI-TOF can detect the true signal

Acknowledgements

Stuart Salmon

Shuo Chen

Roy Herbst

Anne Tsao

Hai Tran

Alan Sandler

David Carbone

Dean Billheimer

Yu Shyr

Ju-Whei Lee

Pierre Massion

Julie Brahmer

Joan Schiller

Thao P. Dang