Wavelet-Based Preprocessing Methods for Mass Spectrometry Data

Jeffrey S. Morris

Department of Biostatistics and Applied Mathematics

UT M.D. Anderson Cancer Center

Overview

- Background and Motivation
- Preprocessing Steps
  - Denoising using Wavelets
  - Baseline Correction/Normalization
  - Peak Detection/Quantification
  - Working with Average Spectrum
- Virtual Mass Spectrometer
- Simulation Study
- Conclusions

Statistical Issues for Mass Spectrometry Experiments

- **Experimental Design**
  - Blocking/RANDOMIZATION – reduce possibility of systematic bias polluting the data.
- **Preprocessing**
  - Remove systematic artifacts/noise from data
  - Extract meaningful features (protein signal) : **nxp matrix**
- **Data Analysis/Discovery**
  - Analyze *n x p* matrix
    - Find which features are associated with exp. cond.
    - Build/validate classifier based on sets of features
    - Cluster samples/features
  - Lots of existing methods available for this

Statistical Model for Spectrum

$$Y_i(t_j) = B_i(t_j) + N_i S_i(t_j) + e_{ij}$$

Statistical Model for Spectrum

$$Y_i(t_j) = \overbrace{B_i(t_j)}^{\text{Baseline Artifact}} + N_i S_i(t_j) + e_{ij}$$

Statistical Model for Spectrum

$$Y_i(t_j) = \overbrace{B_i(t_j)}^{\text{Baseline Artifact}} + N_i \overbrace{S_i(t_j)}^{\text{Protein Signal}} + e_{ij}$$

Statistical Model for Spectrum

$$Y_i(t_j) = \overbrace{B_i(t_j)}^{\text{Baseline Artifact}} + \underbrace{N_i}_{\text{Normal-ization Factor}} \overbrace{S_i(t_j)}^{\text{Protein Signal}} + e_{ij}$$

Statistical Model for Spectrum

$$Y_i(t_j) = \overbrace{B_i(t_j)}^{\text{Baseline Artifact}} + \underbrace{N_i}_{\text{Normalization Factor}} \overbrace{S_i(t_j)}^{\text{Protein Signal}} + \underbrace{e_{ij}}_{\text{additive noise (detector)}}$$

$$e_{ij} \sim N\{0, \sigma^2(t_j)\}$$

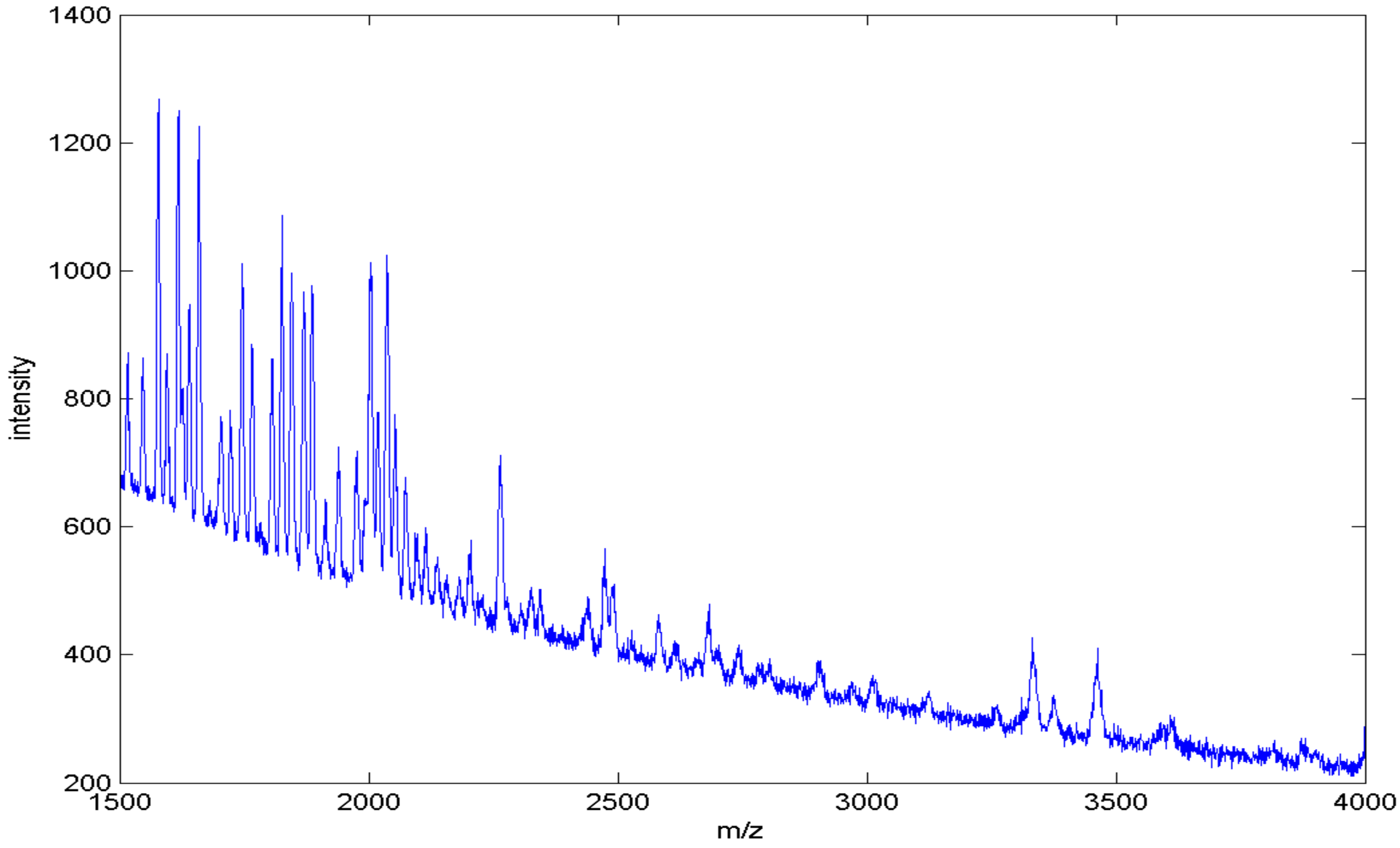
Preprocessing

- **Goal:** Isolate protein signal $S_i(t_j)$
  - Filter out baseline and noise, normalize
  - Extract individual features from signal
- **Problem:**
  - Baseline removal, denoising, normalization, and feature extraction are interrelated processes.
  - Where do we start?

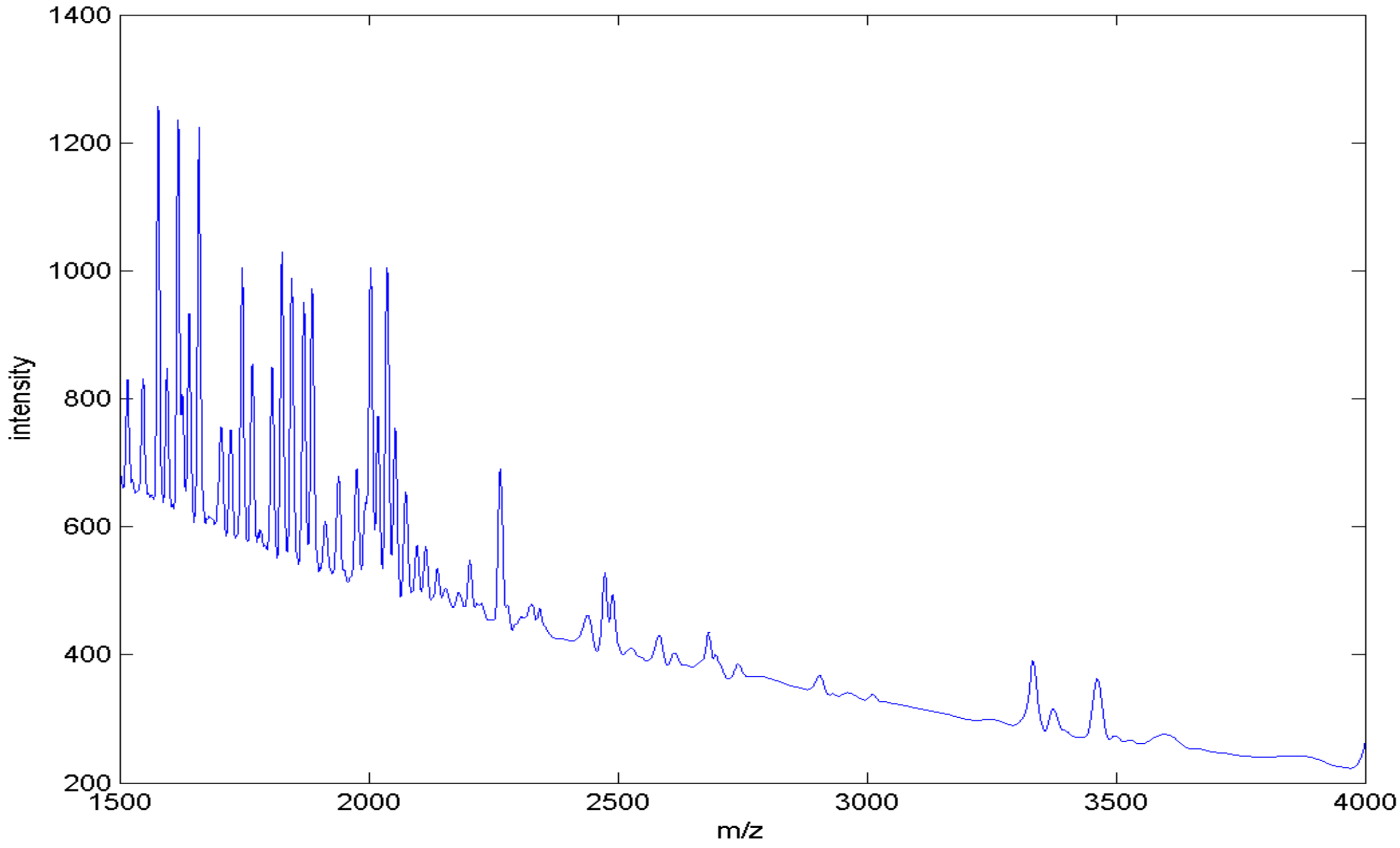
Denoising using Wavelets

- **First step:** Isolate noise using wavelets
  - Wavelets: basis functions that can parsimoniously represent spiky functions
  - Standard denoising tool in signal processing
- **Idea:** Transform from time to wavelet domain, threshold small coefficients, transform back.
  - **Result:** Denoised function and noise estimate
  - **Why does it work?** Signal concentrated on few wavelet coefficients, white noise equally distributed. Thresholding removes noise without affecting signal.
- Does *much* better than denoising tools based on kernels or splines, which tend to attenuate peaks in the signal when removing the noise.

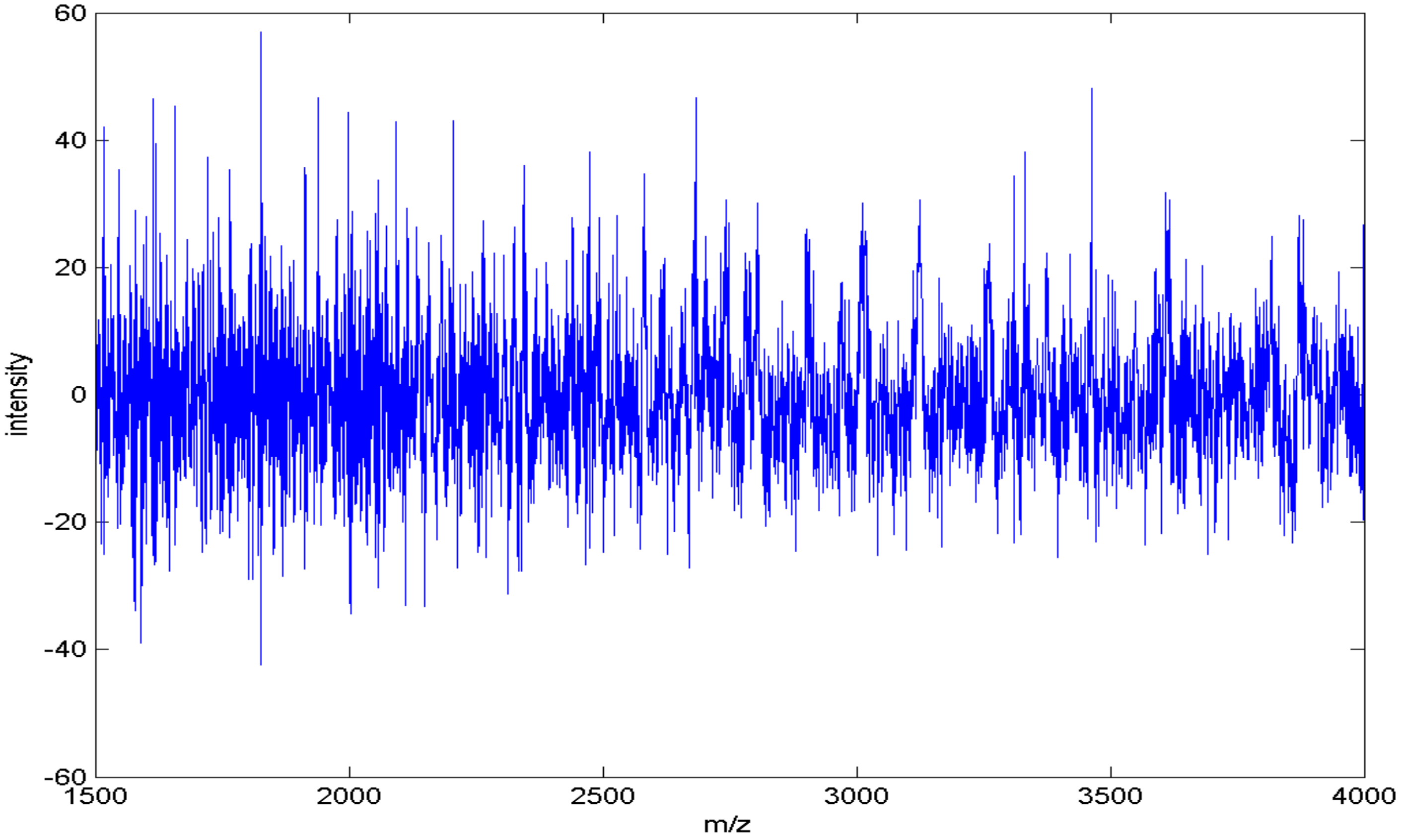
Raw Spectrum



Denoised Spectrum



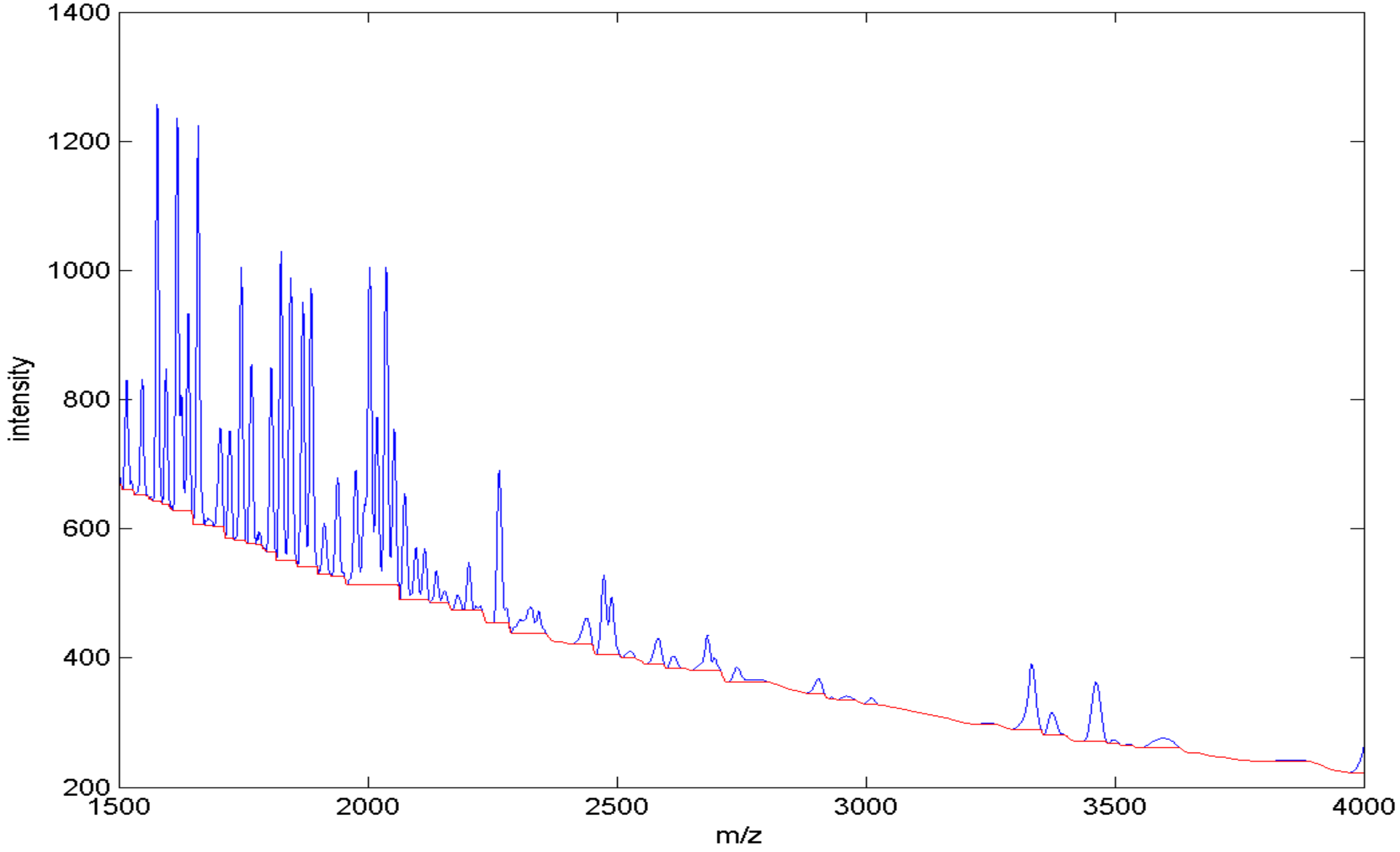
Noise



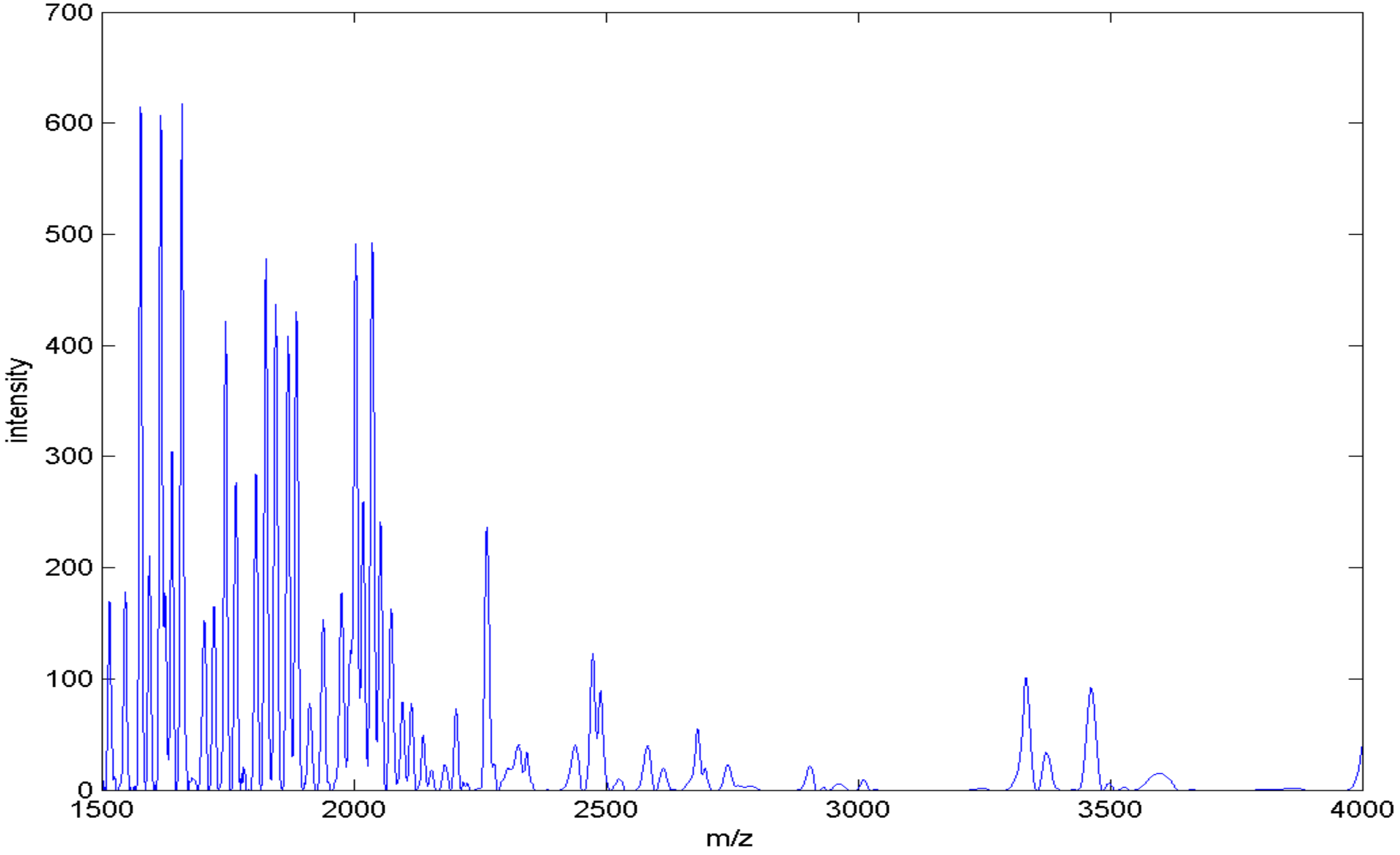
Baseline Correction & Normalization

- **Baseline:** smooth artifact, largely attributable to detector overload.
  - Estimated by monotone local minimum
  - More stably estimated after denoising
- **Normalization:** adjust for possibly different amounts of material desorbing from plates
  - Divide by total area under the denoised and baseline corrected spectrum.

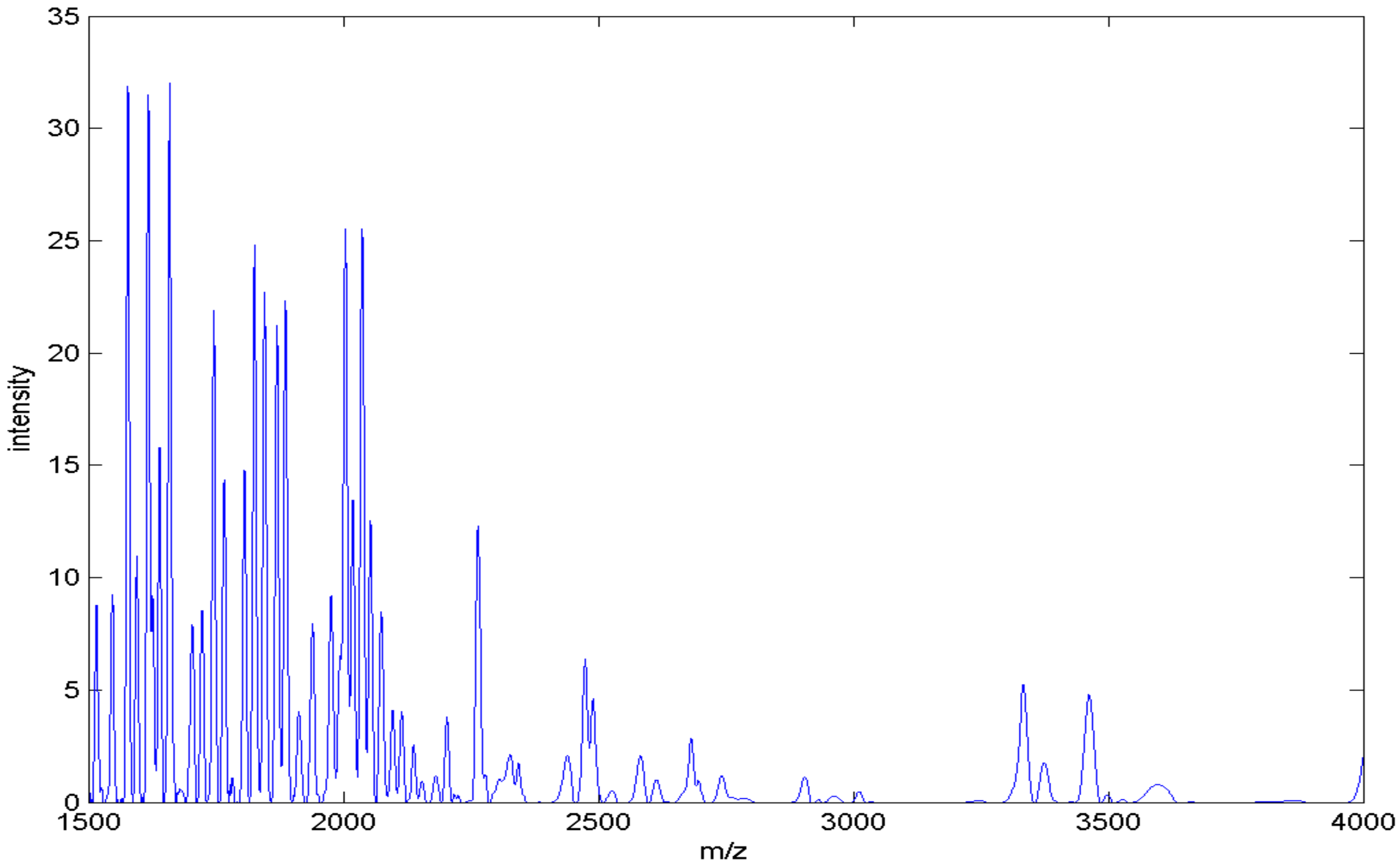
Baseline Estimate



Denoised, Baseline Corrected Spectrum



Denoised, Baseline Corrected, and Normalized Spectrum



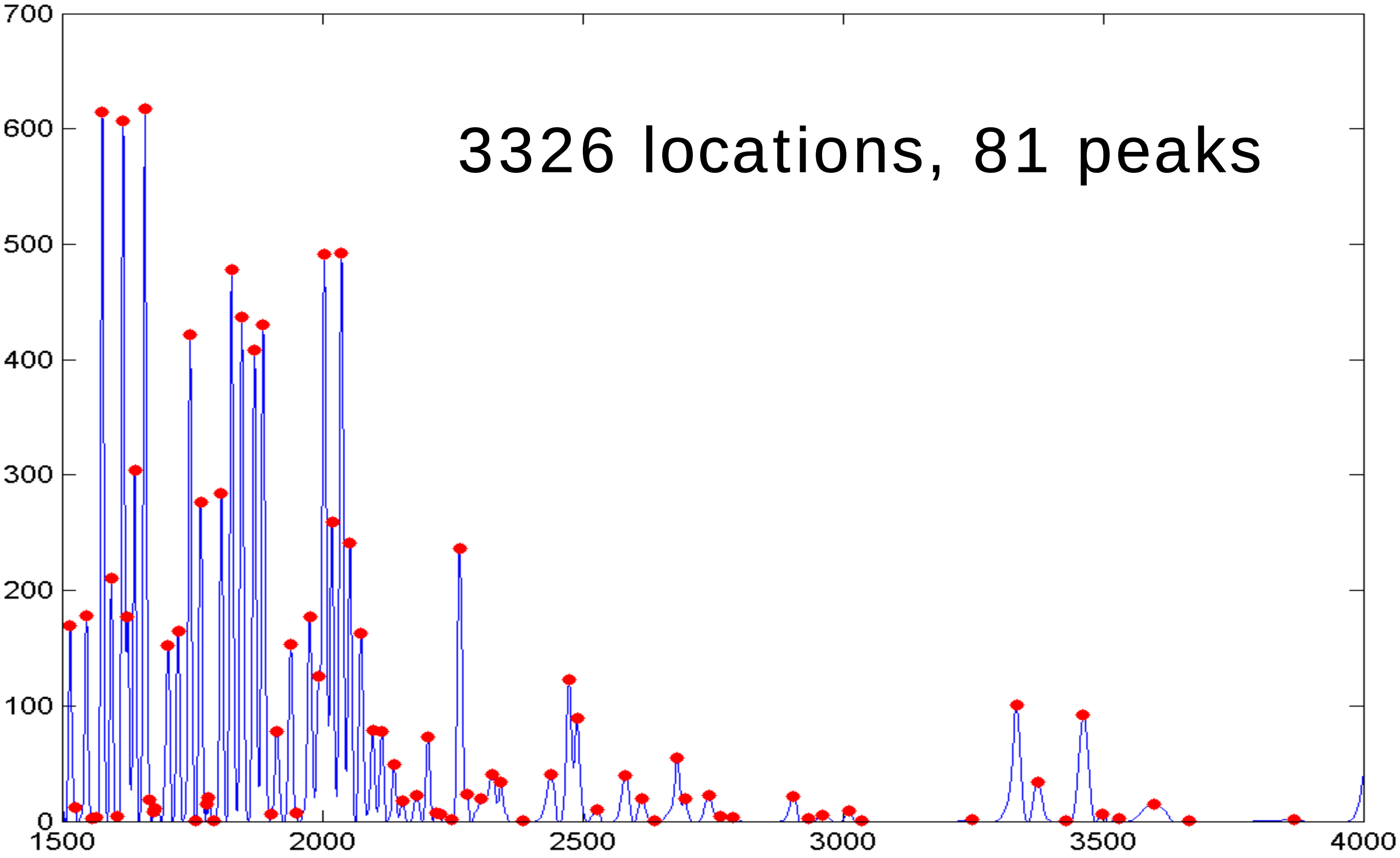
Protein Signal

- **Ideal Form of Protein Signal**: Convolution of peaks
  - Proteins, peptides, and their alterations
  - **Alterations**: isotopes; matrix/sodium adducts; neutral losses of water, ammonia, or carbon
- Limitations of instrument used means we may not be able to resolve all peaks.
- Advantages of peak detection:
  - Reduces multiplicity problem
  - Focuses on units that are theoretically the scientifically interesting features of the data.

Peak Detection

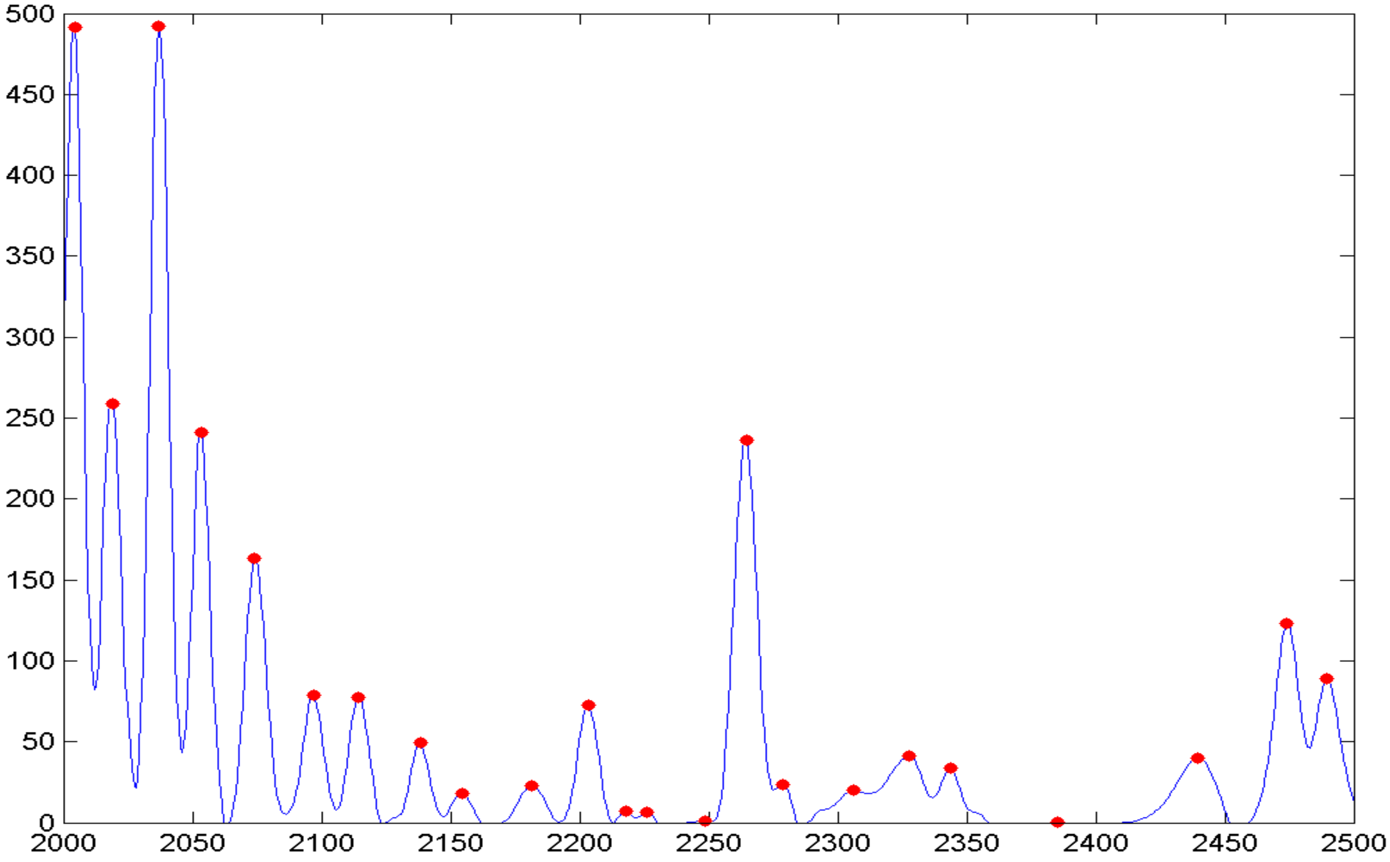
- Easy to do after other preprocessing
- Any local maximum after denoising, baseline correction, and normalization is assumed to correspond to a “peak”.
- May want to require S/N>δ to reduce number of spurious peaks.
  - We can estimate the noise process $\sigma(t)$ *by* applying a local median to the filtered noise from the wavelet transform.
  - Signal-to-noise estimate is ratio of preprocessed spectrum and noise.

Peak Detection

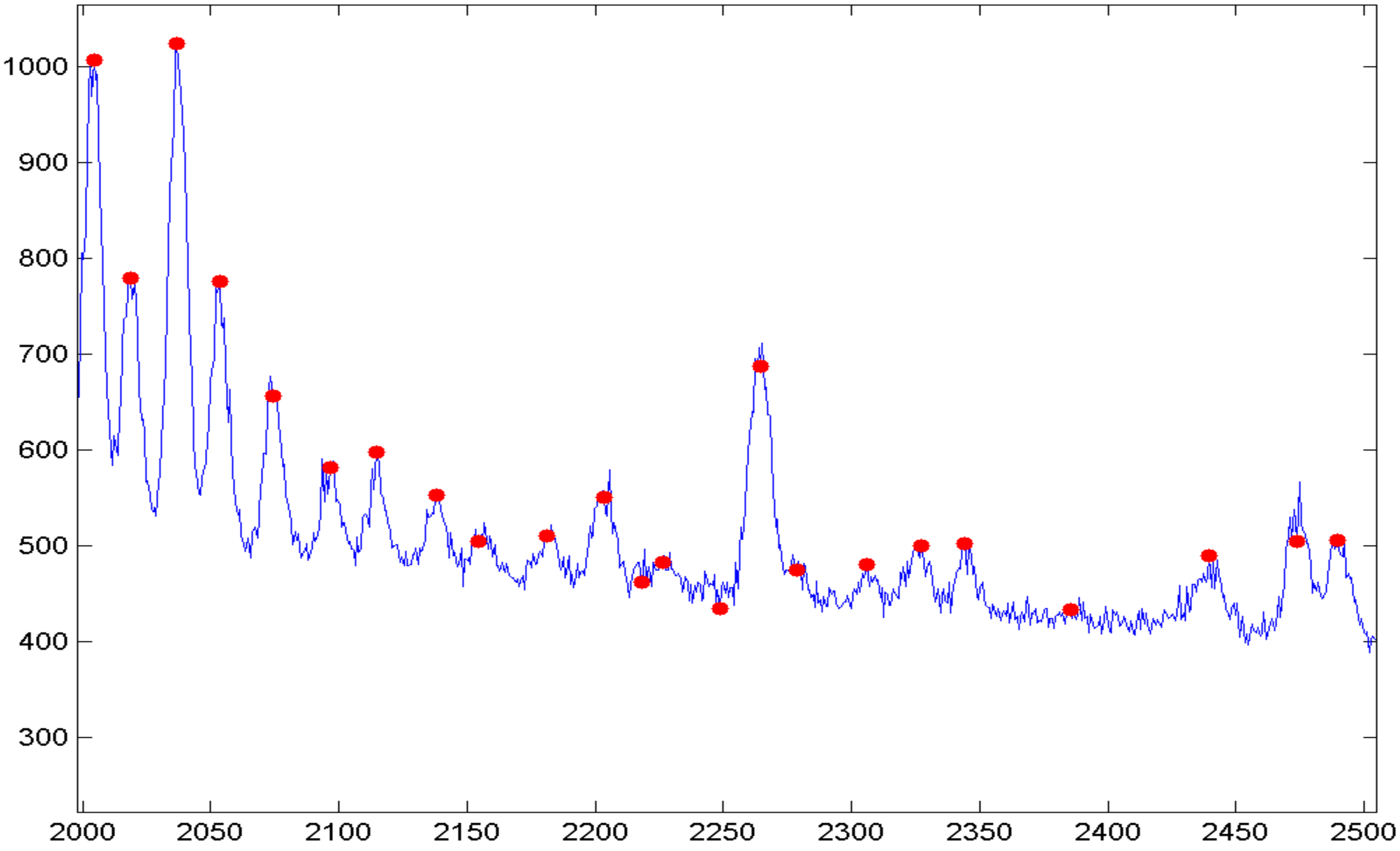


3326 locations, 81 peaks

Peak Detection (zoomed)



Raw Spectrum with peaks



Peak Quantification

- Two options:

1. Area under the peak: Find the left and right endpoints of the peak, compute the AUC in this interval.
2. Maximum intensity: Take intensity at the local maximum (may want to take log or cube root)

- Theoretically, AUP quantifies amount of given substance desorbed from the chip.
  - But it is very difficult to identify the endpoints of peaks

Peak Quantification

- The maximum intensity is a practical alternative
  - No need for endpoints, should be correlated with AUP
  - Physics of mass spectrometry shows that, for a given ion with m/z value *x*, there is a **linear relationship** between the **number of ions** of that type desorbed from plate and the **expected maximum peak intensity** at *x.*
- Problem with both methods:
  Overlapping peaks that are not deconvolvable
  - Local maximum at *t* contains weighted average of information from multiple ions whose corresponding peaks have mass at location *t*.
  - Major problem – short of formal deconvolution, have not seen simple solution to this problem.

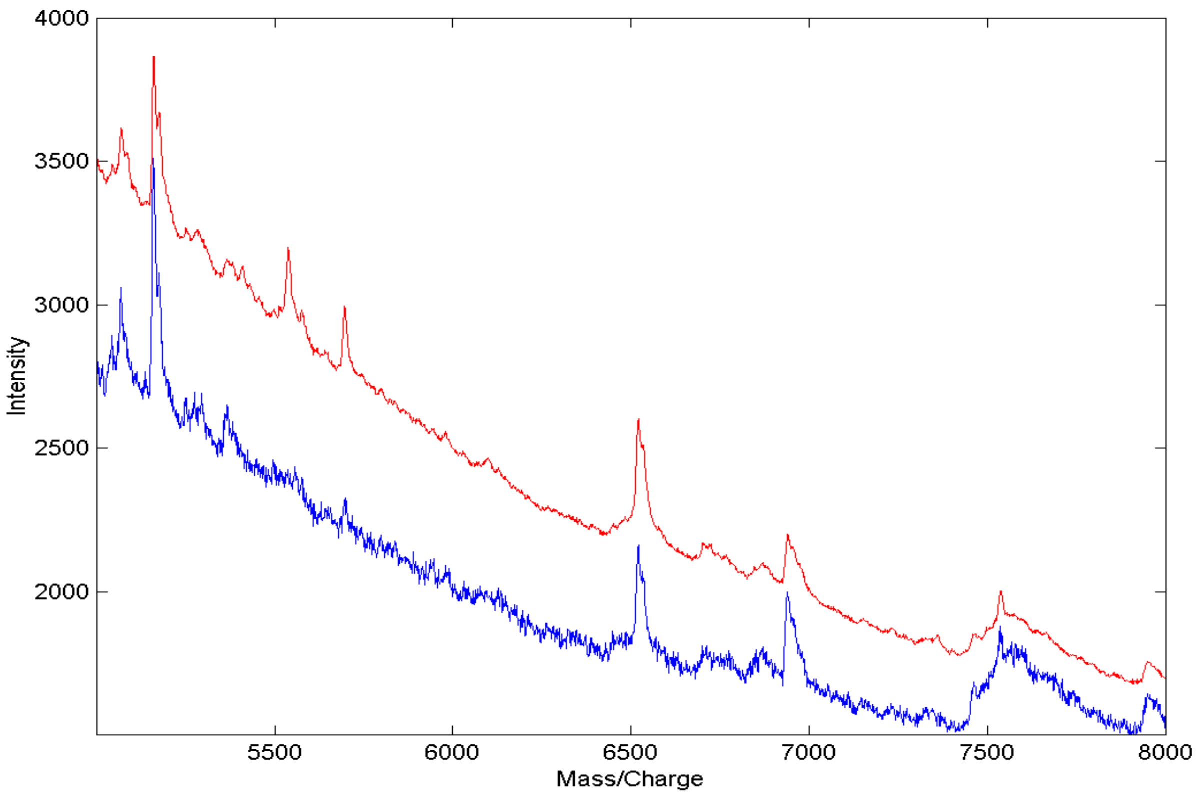
Peak Matching Problem

- If peak detection performed on individual spectra, peaks must be matched across samples to get n *x* p matrix.
  - Difficult and arbitrary process
  - What to do about “missing peaks?”
- **Our Solution:** Identify peaks on **mean spectrum** (at locations x_1, …, x_p), then quantify peaks on individual spectra by intensities at these locations.

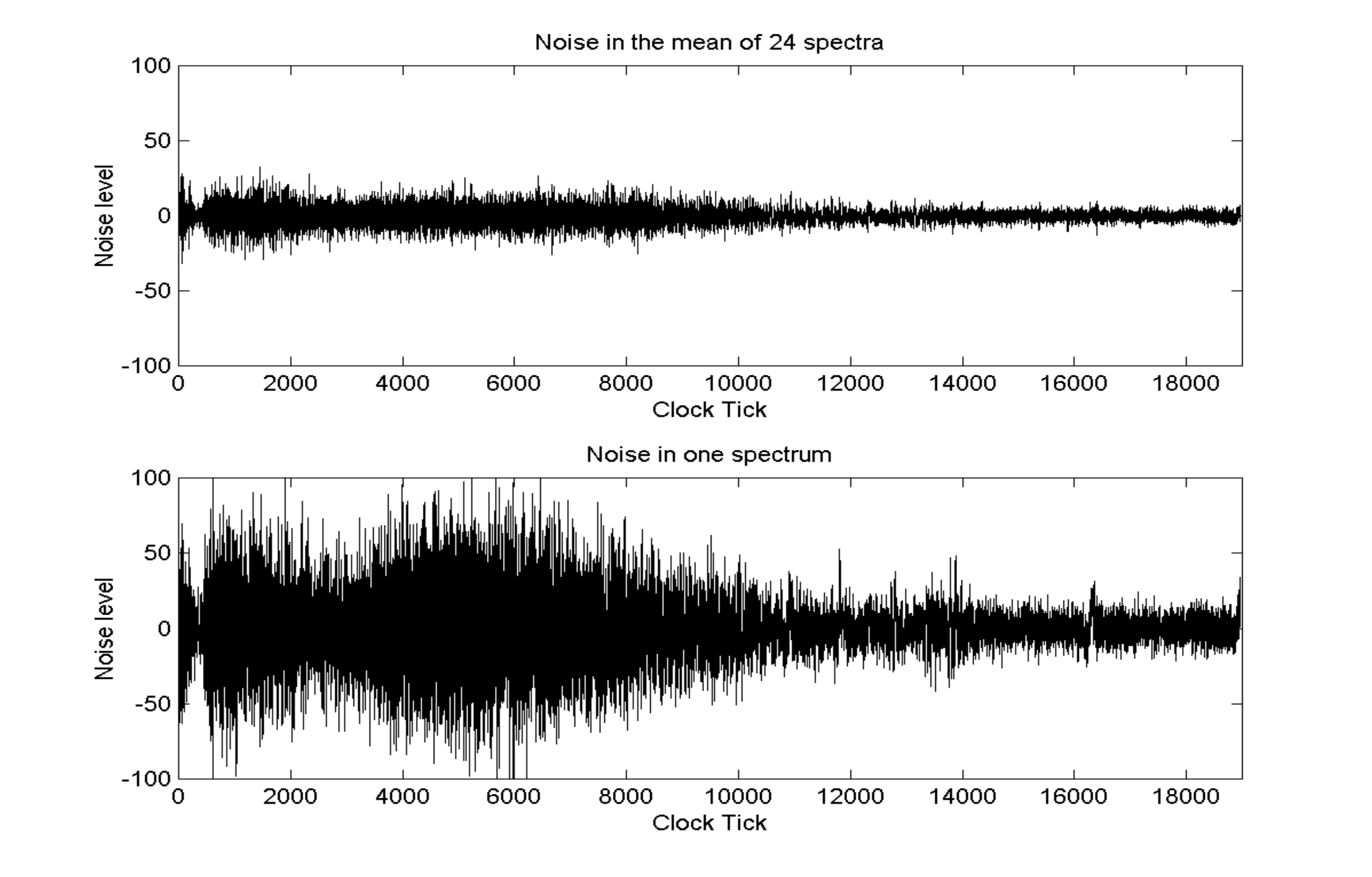
Advantages/Disadvantages

- Advantages
  - Avoids peak-matching problem
  - Generally more sensitive and specific
    - Noise level reduced by sqrt(n)
    - Borrows strength across spectra in determining whether there is a peak or not (signals reinforced over spectra)
  - Robust to minor calibration problems
- Disadvantage
  - Tends to be less sensitive when prevalence of peak < 1/sqrt(n).

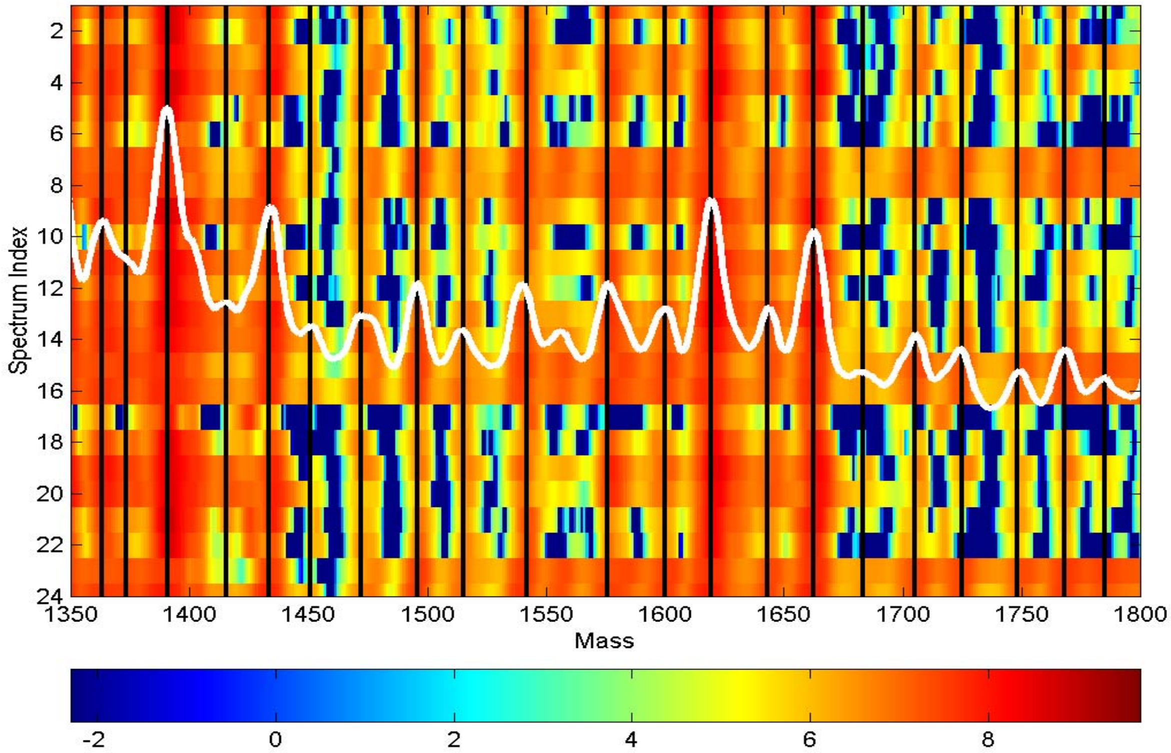
Noise reduced in mean spectrum



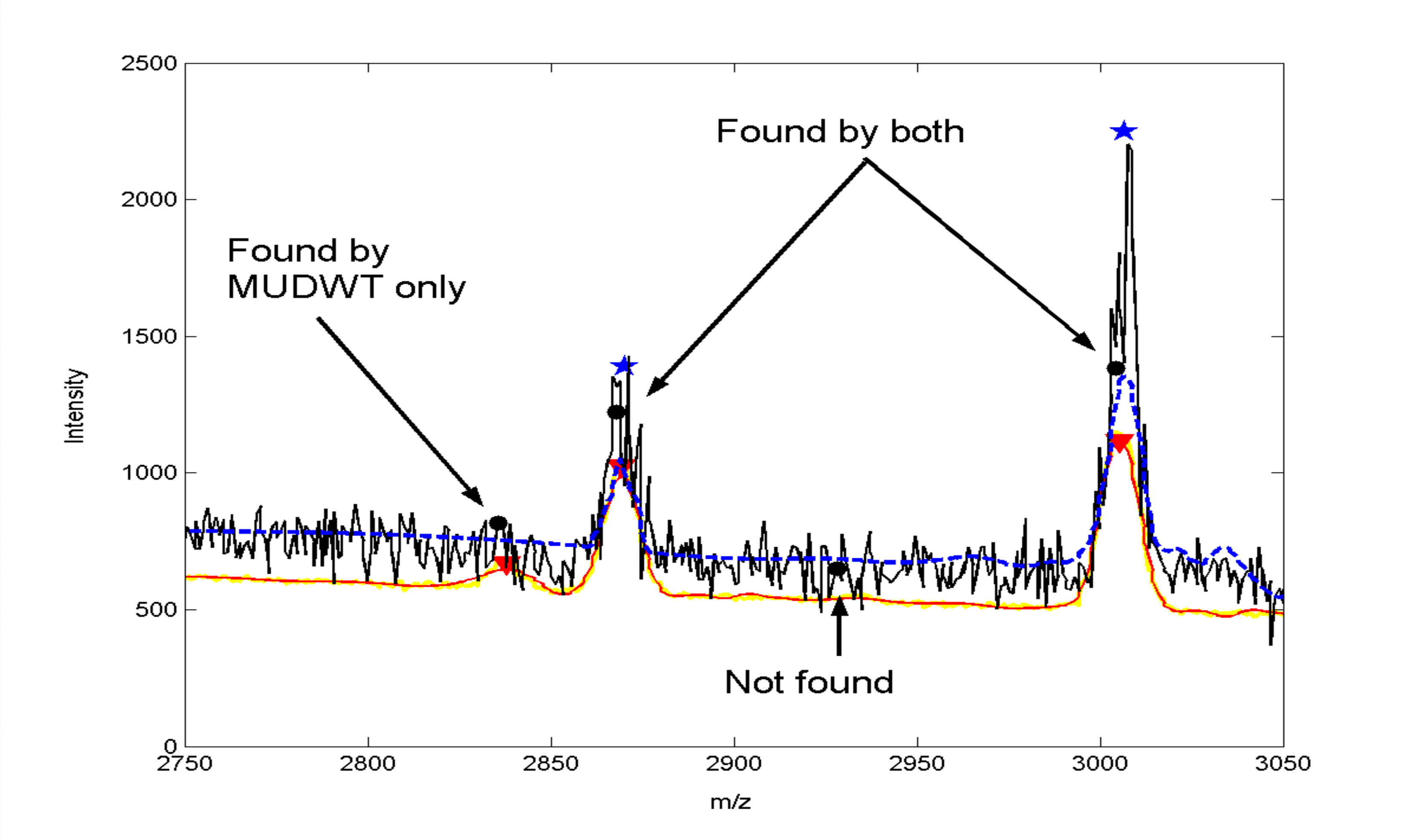
Noise reduced in mean spectrum



Peak detection with mean spectrum



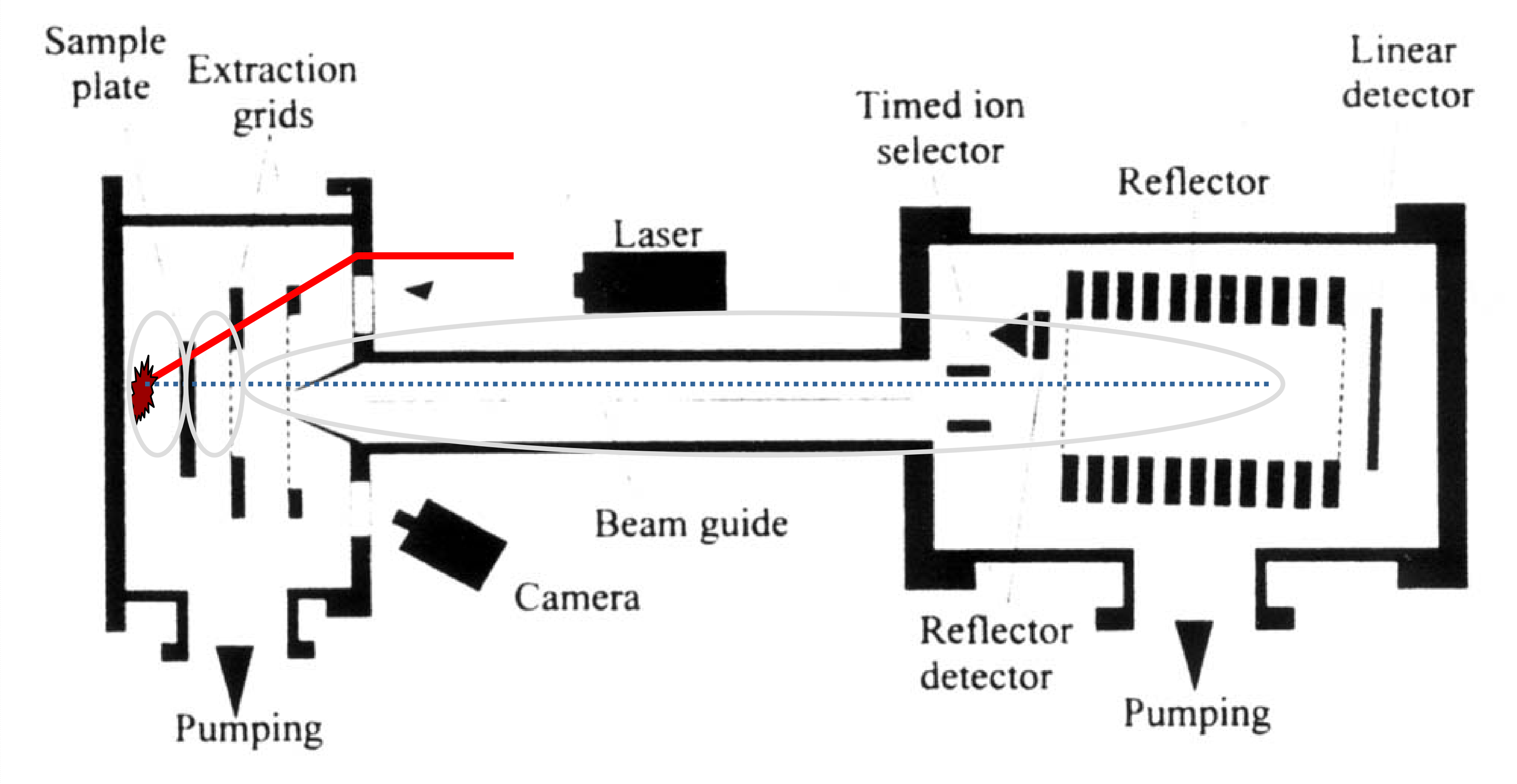
Sample Spectrum



Simulated spectra

- Difficult to evaluate processing methods on real data since we don’t know “truth”
- Have developed a simulation engine to produce realistic spectra
  - Based on the physics of a linear MALDI-TOF with ion focus delay
  - Flexible incorporation of different noise models and different baseline models
  - Includes isotope distributions
  - Can include matrix adducts, other modifications

MALDI-TOF schematic



Vestal and Juhasz. *J. Am. Soc. Mass Spectrom*. **1998**, *9*, 892.

Modeling the physics of MALDI-TOF

- Parameters

  D_1 = distance from sample plate to first grid (8 mm)

  V_1 = voltage for focusing (2000 V)

  D_2 = distance between grids (17 mm)

  V_2 = voltage for acceleration(20000 V)

  L = length of tube (1 m)

  v_0 = initial velocity ~ N(μ,σ)

  v_1 = velocity after focusing

  δ= delay time

- Equations

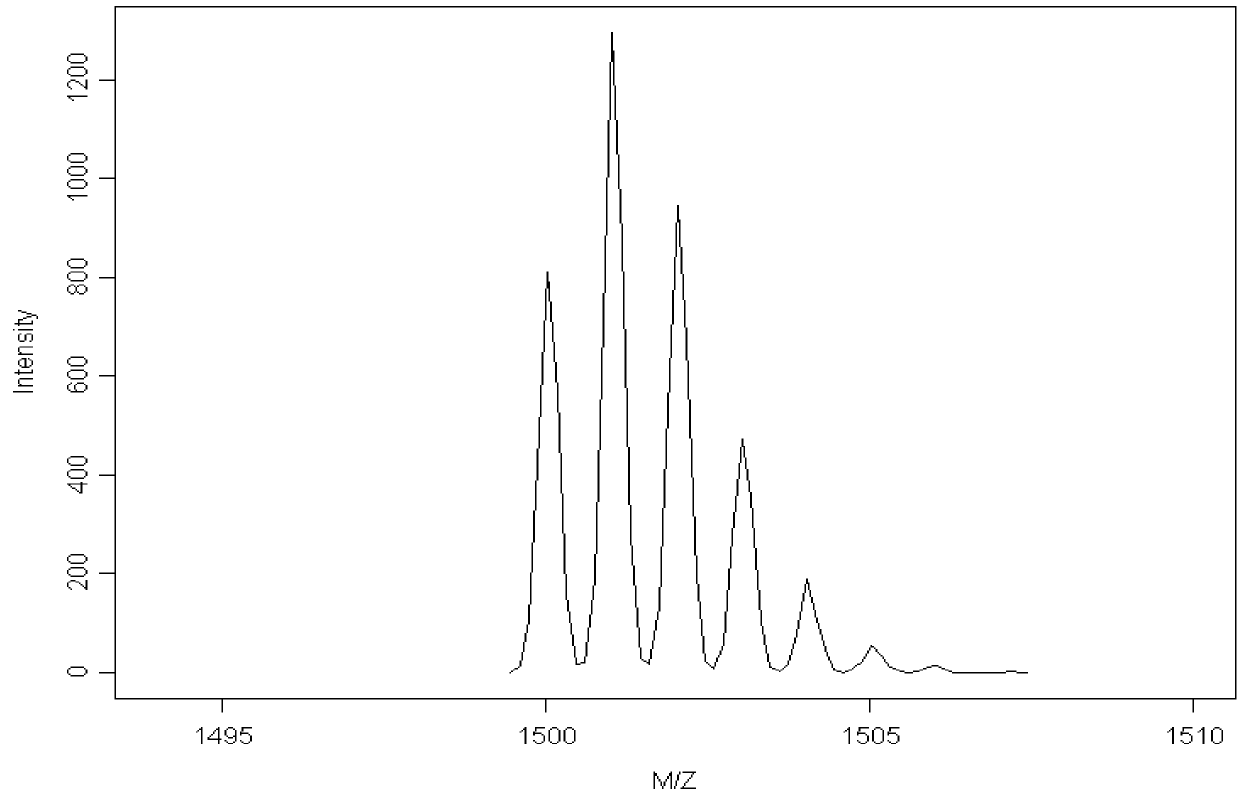
$$v_1^2 = v_0^2 + \frac{2qV_1}{mD_1}(D_1 - \delta v_0)$$

$$t_{DRIFT}^2 = L^2 / \left(\frac{2qV_2}{m} + v_1^2 \right)$$

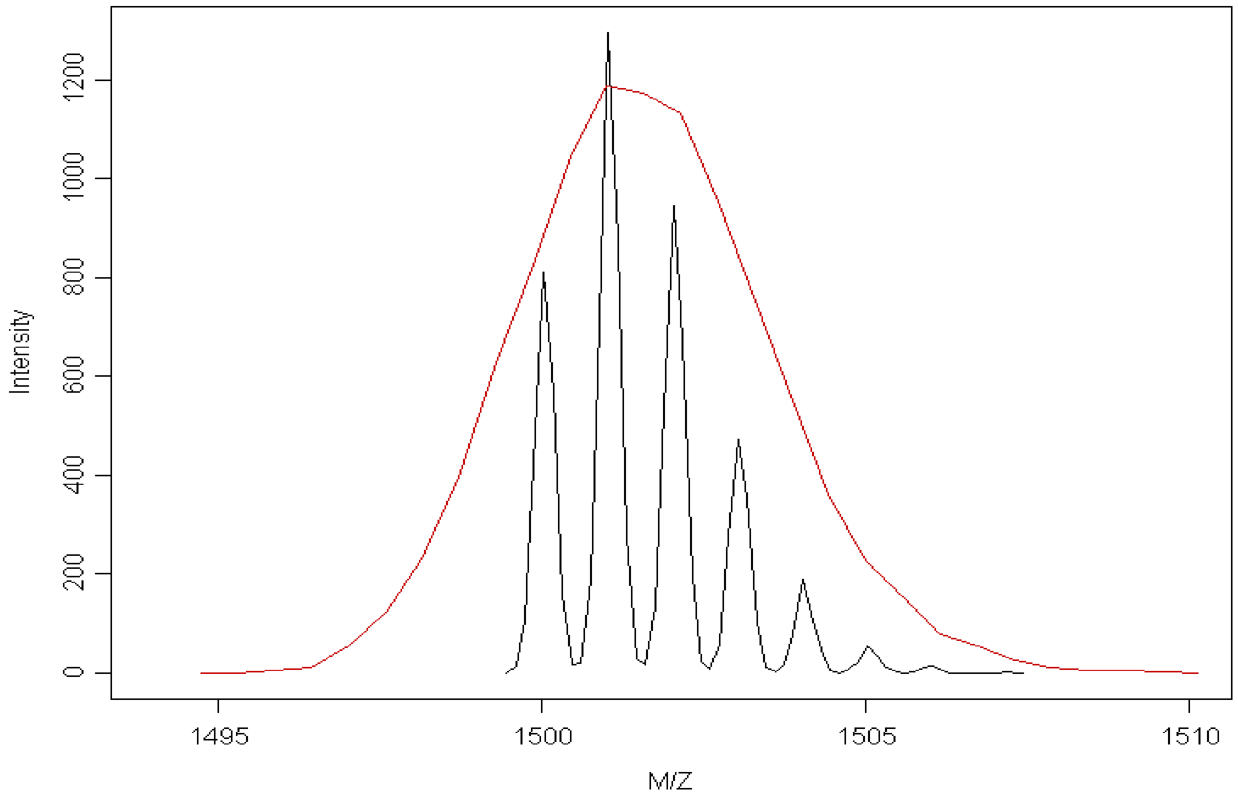
$$t_{ACCEL} = \frac{mD_2}{qV_2}\left(\frac{L}{t_{DRIFT}} - v_1 \right)$$

$$t_{FOCUS} = \frac{mD_1}{qV_1}(v_1 - v_0)$$

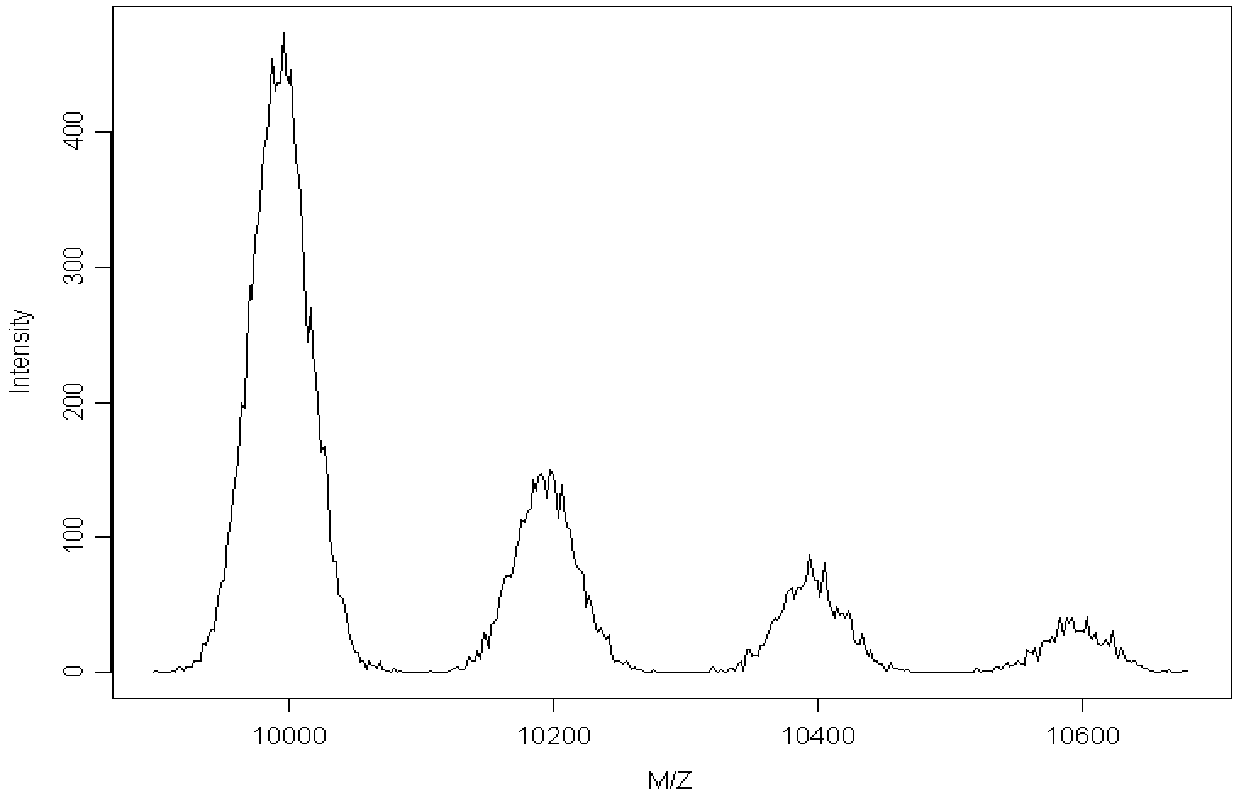
Simulation of one protein, with isotope distribution



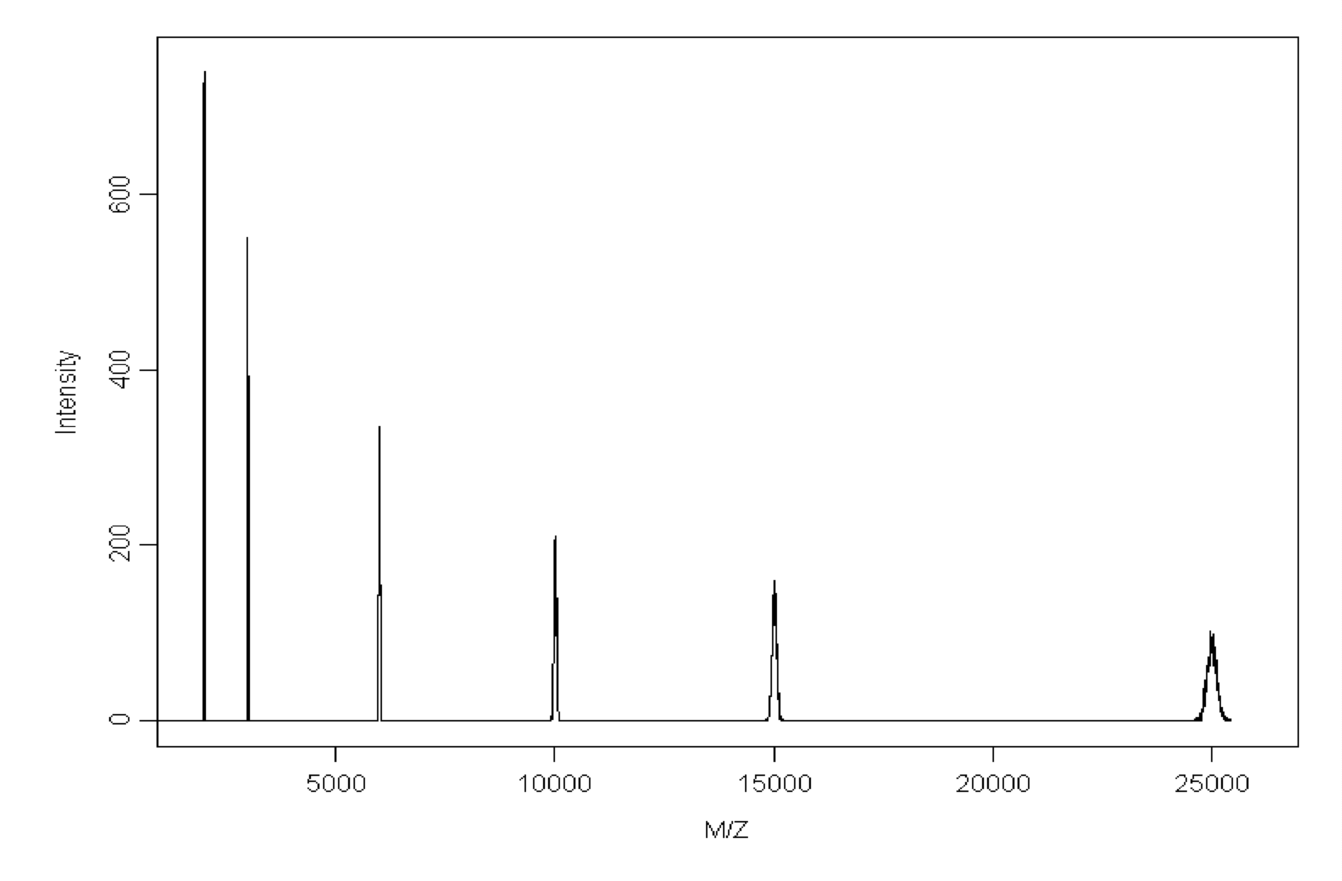
Same protein simulated on a low resolution instrument



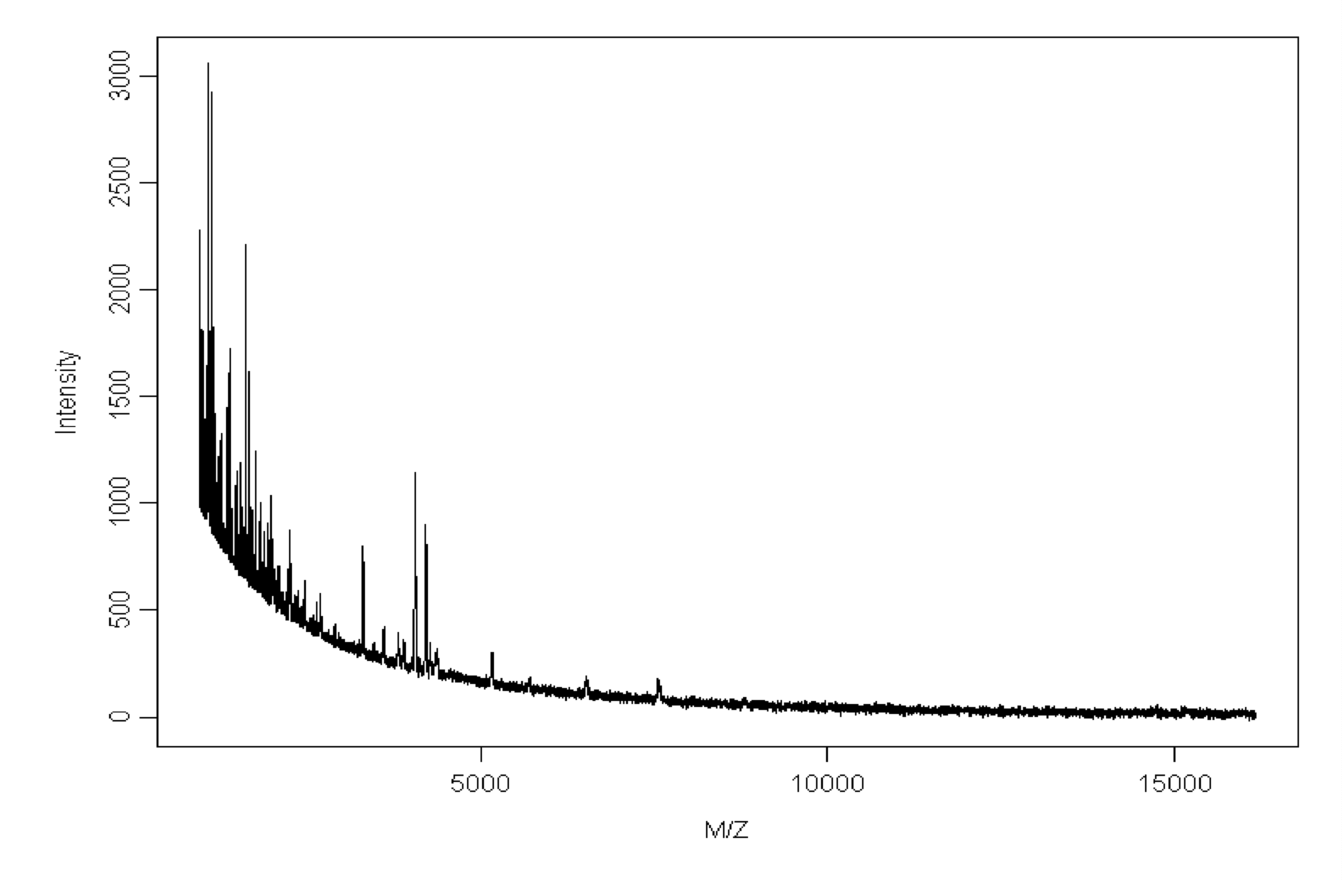
Simulation of one protein with matrix adducts



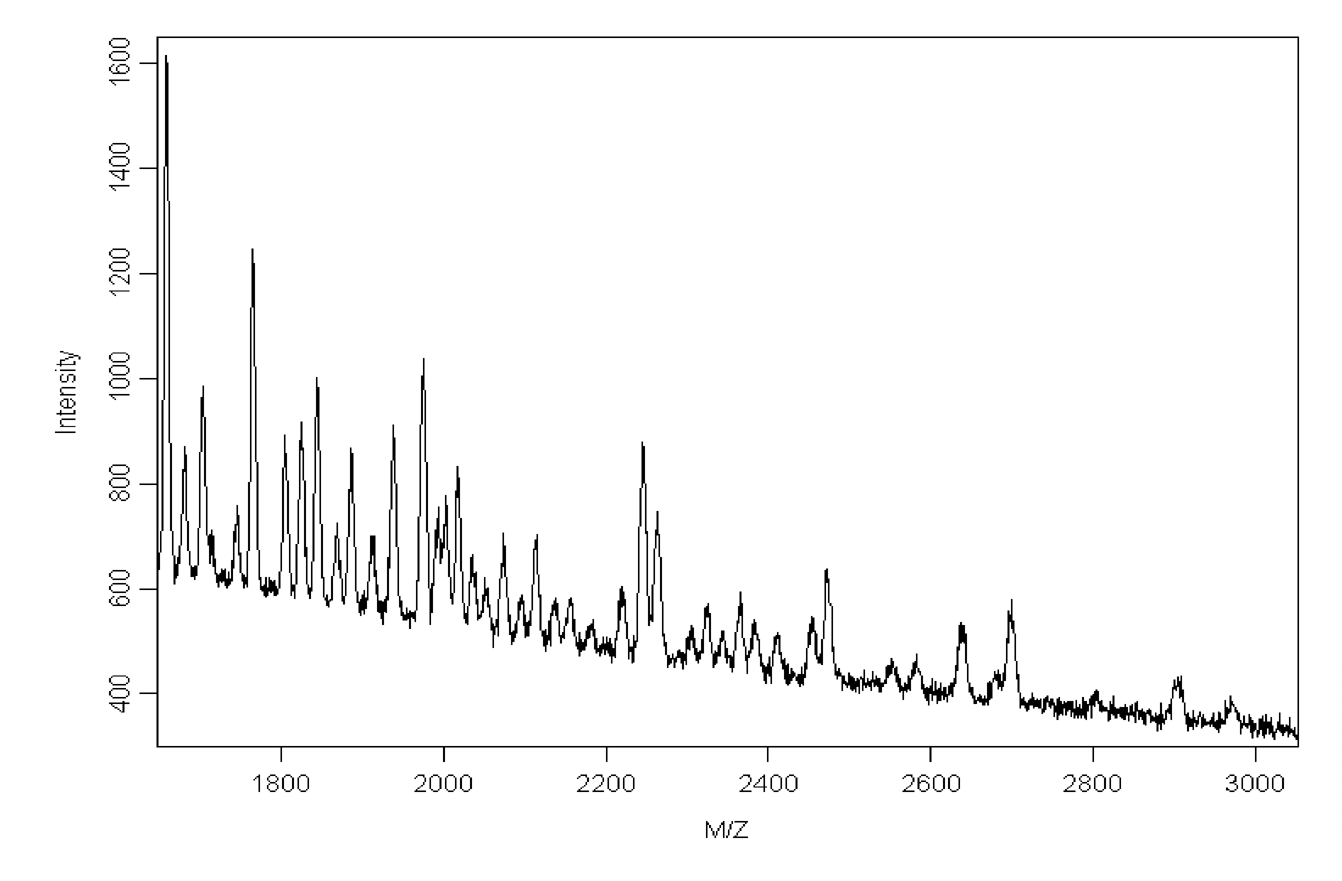
Simulated calibration spectrum with equal amounts of six proteins



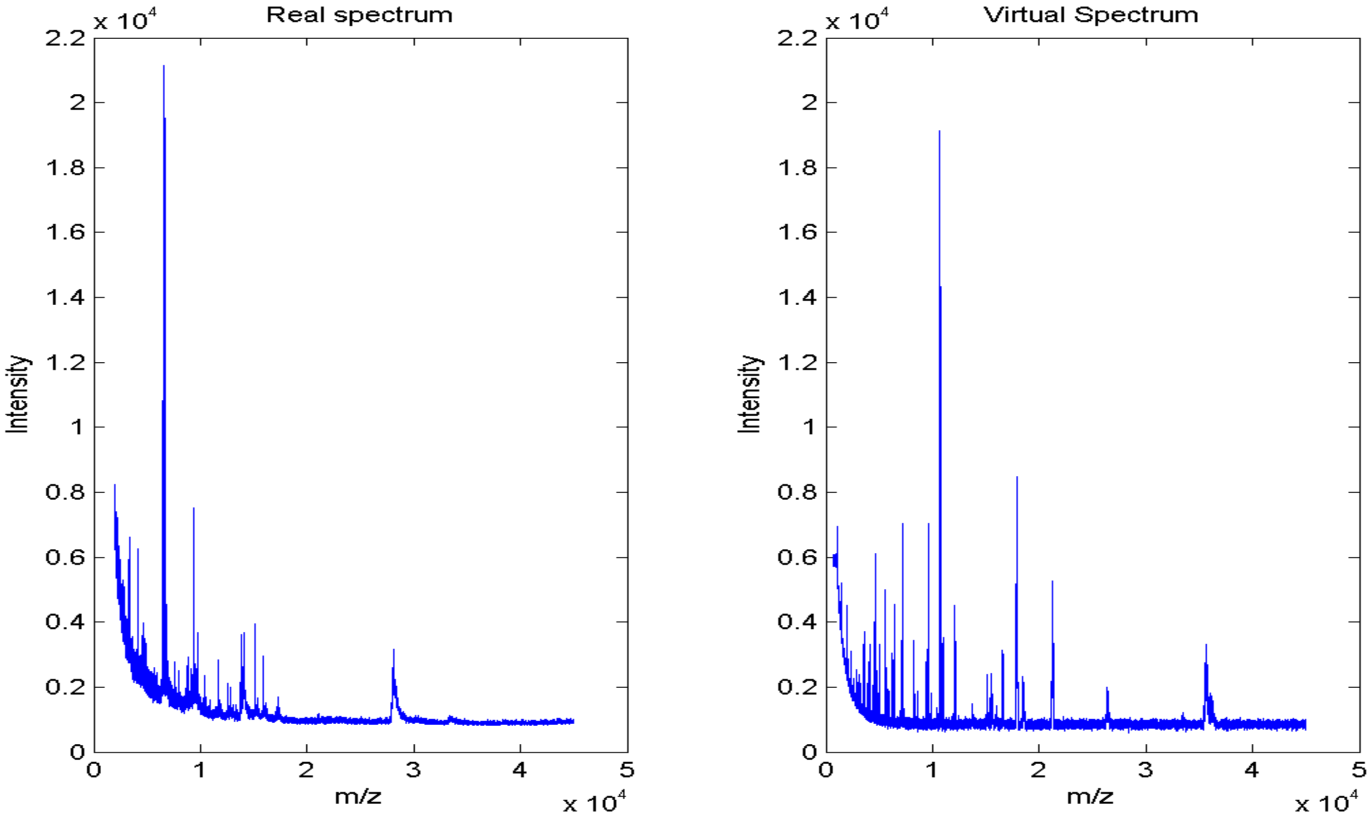
Simulated spectrum with a complex mixture of proteins



Closeup of simulated complex spectrum



Real and Virtual Spectra



Using Virtual Mass Spectrometer

- **Input: virtual sample**
  - proteins and peptides desorbed from sample
  - list of molecular masses w/ # of molecules
- **Output: virtual spectrum**
- **Simulation Studies: virtual population**
  - Defines distribution of proteins in proteome from which you are sampling
  - Assume *p* proteins; for each specify 4 quantities
    - major peak location (m/z of dominant ion)
    - prevalence (proportion of samples with protein)
    - abundance (mean # ions desorbed from samples w/ protein)
    - variance (var # of desorbed ions across samples w/ protein)

Simulation Study

1. Generated 100 random virtual populations based on MDACC MALDI study on pancreatic cancer.
2. For each virtual population, generated 100 virtual samples, obtained 100 virtual spectra.
3. Applied preprocessing and peak detection method based on individual and average spectra
4. Summarized performance based on sensitivity (proportion of proteins detected) and FDR (proportion of peaks corresponding to real proteins).
    - Tricky to do – see paper for details.

Simulation Results

Overall Results

	sensitivity	**FDR**	**pv***
SUDWT (indiv. spectra)	0.75	0.09	0.03
MUDWT (mean spectrum)	0.83	0.06	0.97

*pv=the proportion of simulations with higher sensitivity

Simulation Results

By Prevalence

π:	**<.05** (14%)	**.05-.20** (16%)	**.20-.80** (40%)	**>.80** (30%)
sensitivity **(SUDWT)**	0.43	0.74	0.81	0.82
sensitivity **(MUDWT)**	0.38	0.74	0.93	0.97
pv **(MUDWT)**	0.25	0.49	1.00	1.00

Simulation Results

By Abundance (mean log intensity)

log(μ):	**<9.0** (31%)	**9.0-9.5** (27%)	**9.5-10** (23%)	**>10** (19%)
sensitivity **(SUDWT)**	0.68	0.75	0.78	0.82
sensitivity **(MUDWT)**	0.78	0.84	0.85	0.88
pv **(MUDWT)**	0.97	0.89	0.84	0.78

Open problems: Preprocessing

- Better calibration?
  - Internal validation
- Better baseline correction?
- Alternative methods for normalization?
- Quality control/quality assurance?
- Best approach for quantification?

Open problems: Virtual Mass Spectrometry Instrument

- Include more alterations
  - Adducts and neutral molecule losses
  - Multiply-charged ions
- Develop more realistic model for baseline artifact
- Generalize to other instruments?

Acknowledgements

- Bioinformatics
  - Kevin Coombes
  - Keith Baggerly
  - Jing Wang
  - Lianchun Xiao
  - Spyros Tsavachidis
  - Thomas Liu
- Proteomics (MDACC)
  - Ryuji Kobayashi
  - David Hawke
  - John Koomen
- Ciphergen
  - Charlotte Clarke
- Biologists (MDACC)
  - Jim Abbruzzese
  - I.J. Fidler
  - Stan Hamilton
  - Nancy Shih
  - Ken Aldape
  - Henry Kuerer
  - Herb Fritsche
  - Gordon Mills
  - Lajos Pusztai
  - Jack Roth
  - Lin Ji