

Statistical methods in proteomics

Proteomics

- Diagnostic proteomics
 - Determines the presence and amount of proteins in tissue
- Structural proteomics
 - Studies the structure of specific proteins

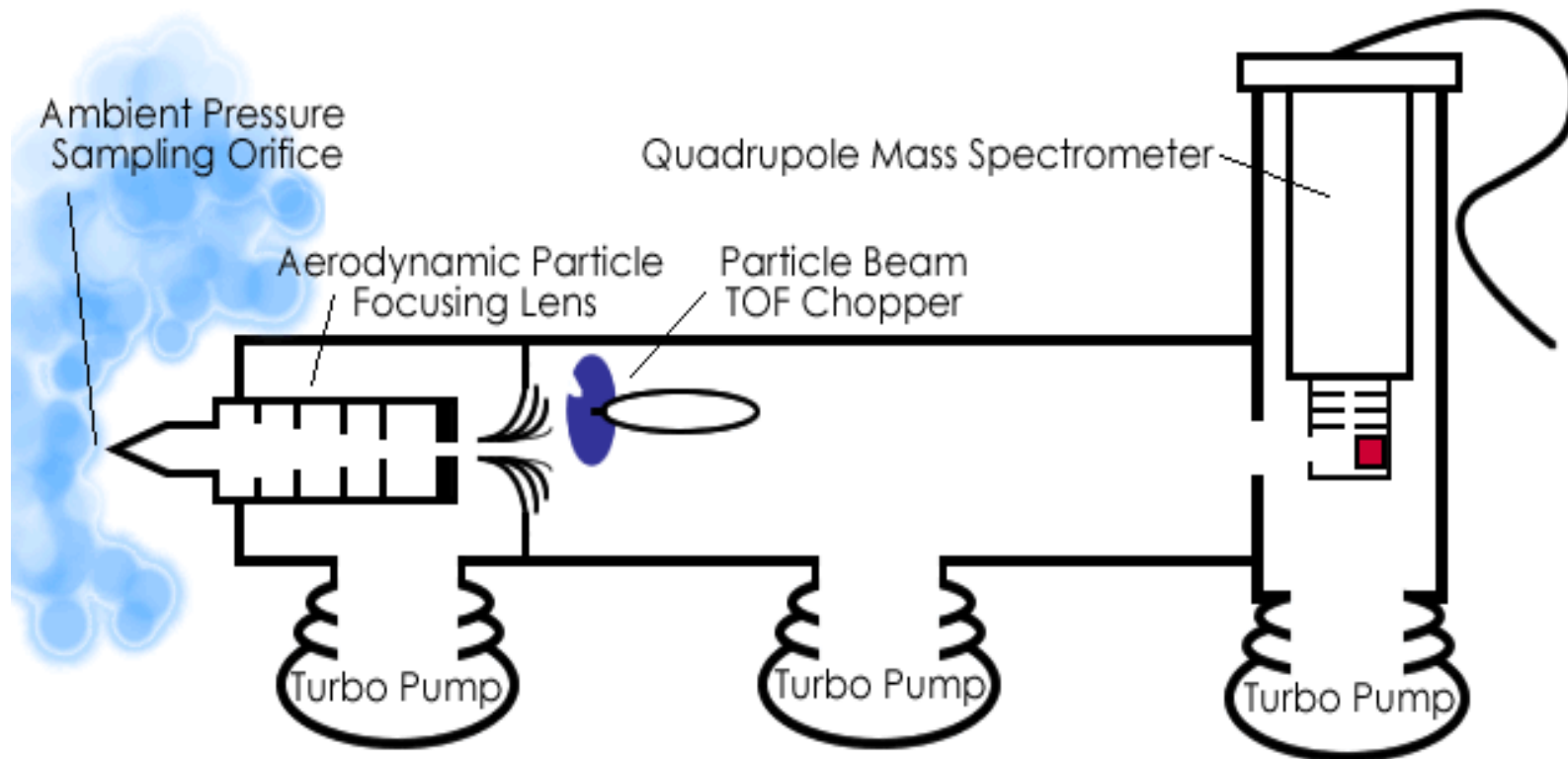
We focus on diagnostic proteomics.

Example

- 41 serum samples, 24 of which were from patients with disease
- 17 from patients without evidence of disease.
- Task: (1) find the most statistically significant ion peak differences between the two groups of spectra;
- or (2) separate the 41 unlabeled specimens into two groups based on their ion peak patterns.

Matrix-assisted laser desorption/ionization-time of flight

MALDI-TOF



Credit: Matt Thyson (Lexington, Massachusetts)
Source <http://cires.colorado.edu/~jjose/ams.html>

MALDI-TOF

The kinetic energy given to an ion is equal to the product of its charge z and the applied voltage E :

$$mv^2 \propto zE$$

Time of fly (TOF) for distance D is

$$t = D/v$$

$$t \propto (m/z)^{1/2}$$

E, D : same for all particles

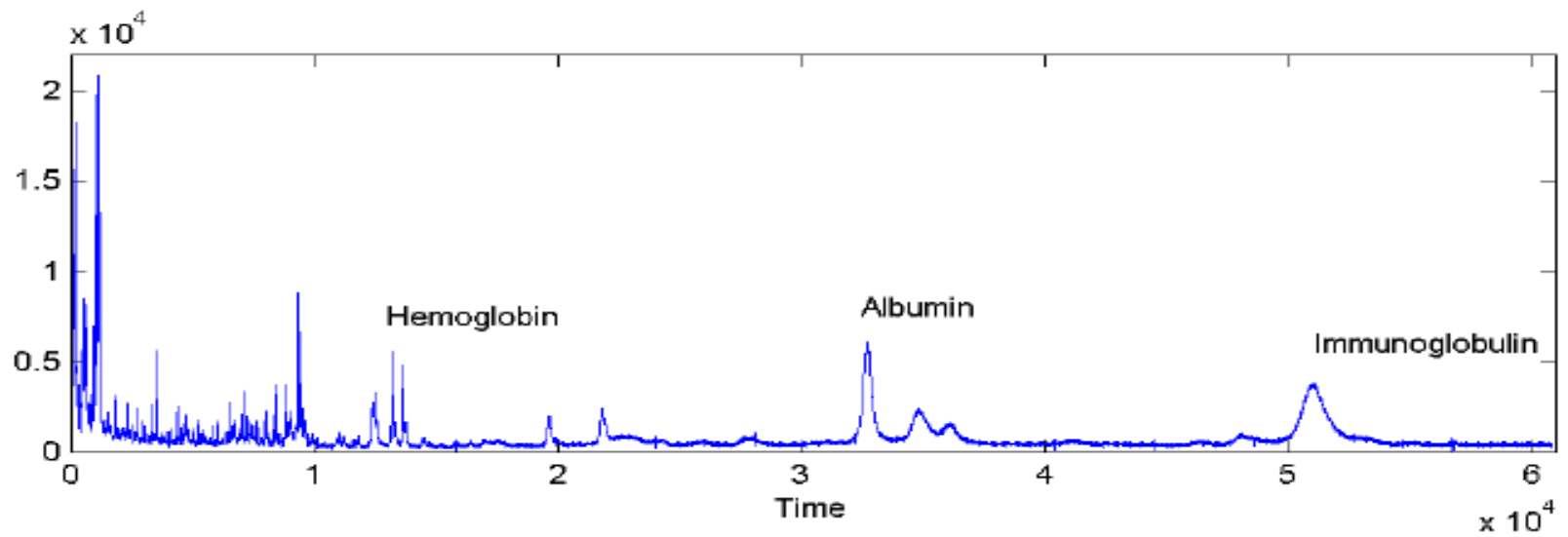
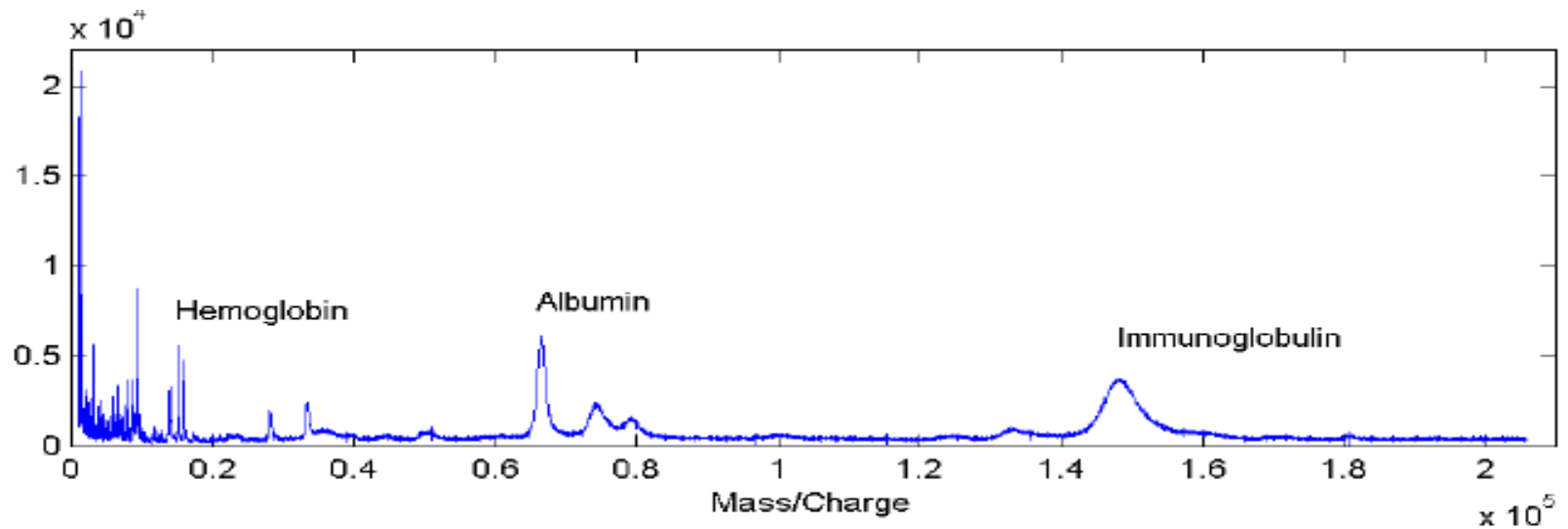
t : measurement

m/z : property of peptide (identity)

Height/area of the peak: quantity (y axis of spectrum)

} X axis of spectrum

- Location of peak (time or m/z) indicates type of protein
 - MS does not identify proteins directly. Several molecules may correspond to each point on the spectrum
- Intensity (height of peak) indicates amount



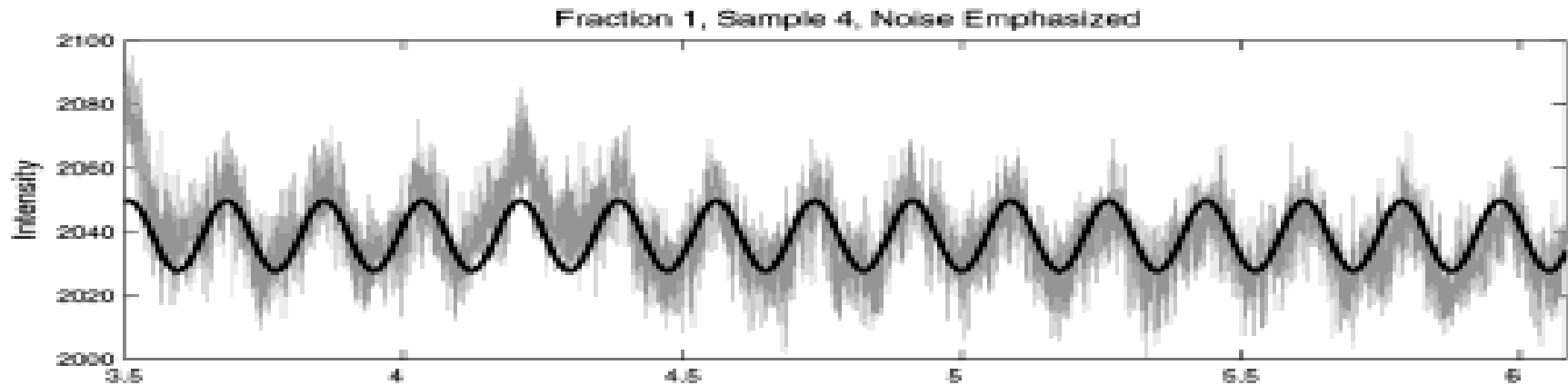
- Notice the scales in time units and in m/z are not linear

Preprocessing

- Sinusoidal noise removal
- Baseline subtraction
- Normalization
- Recombination

Preprocessing (I)

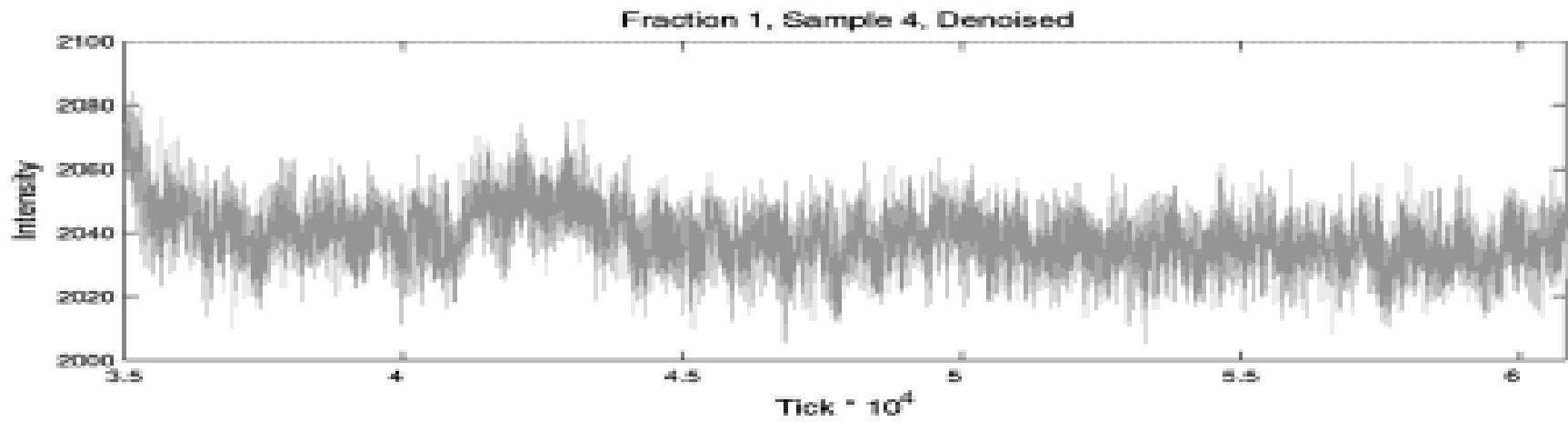
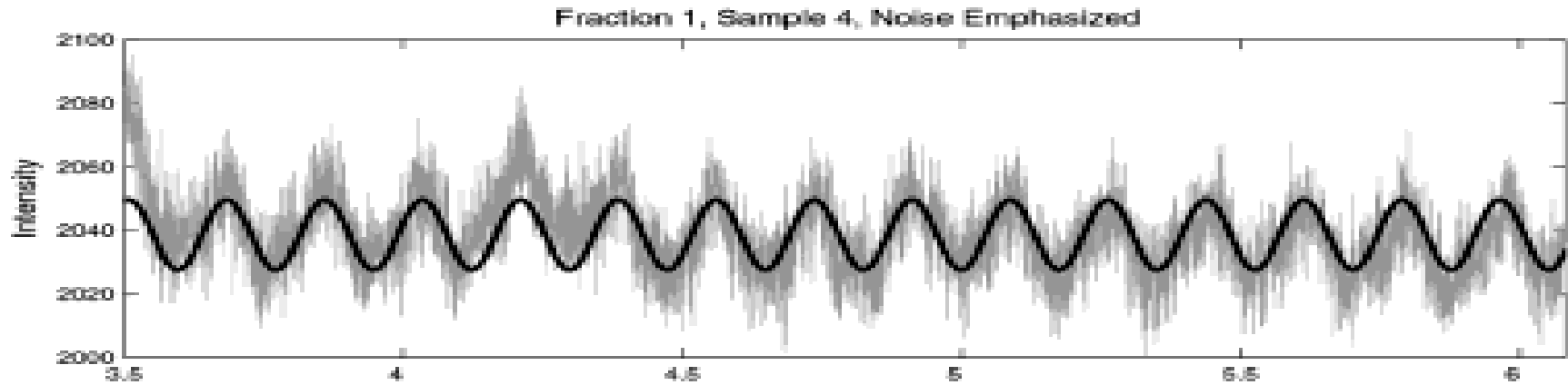
- Periodical patterns



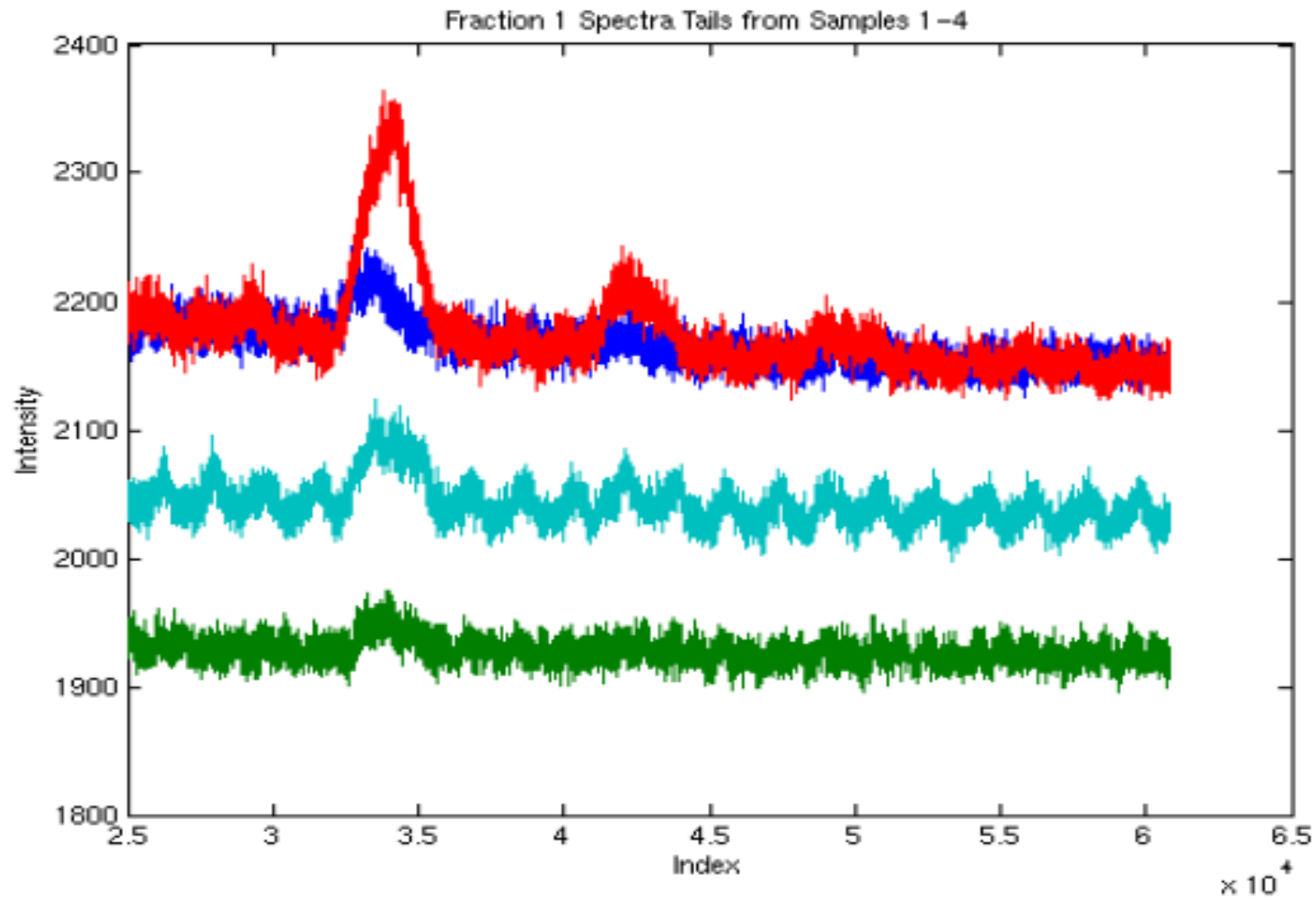
- Fourier transform at regions where larger peaks were absent
- Sinusoids of appropriate frequency were fit to the tails, extended to the full spectrum and subtracted

Preprocessing (I)

- Periodical patterns



Preprocessing (II)



Preprocessing (II)

- Different samples have different baseline
- Baseline is not constant within a sample
 - Local median
 - Local minimum
 - Window size

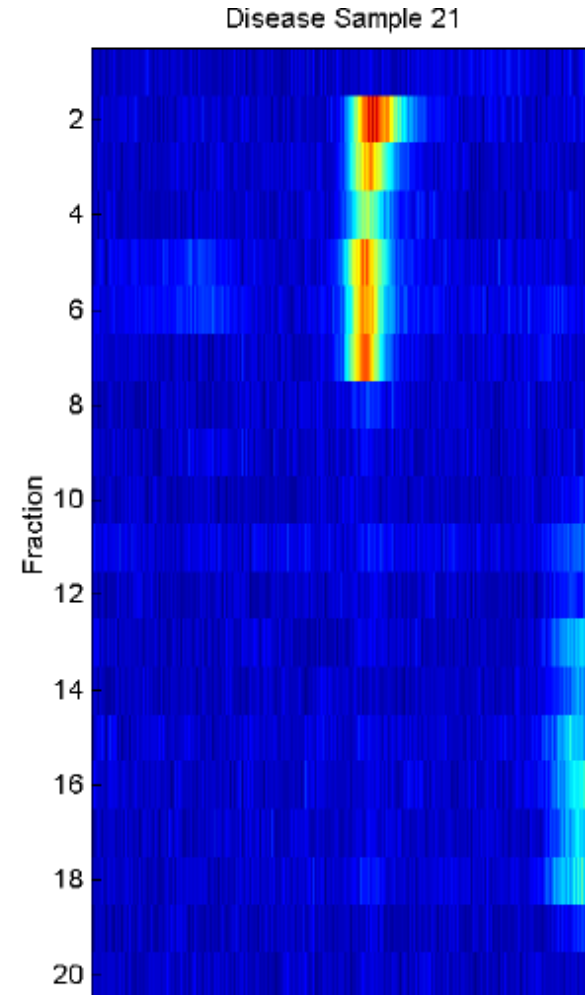
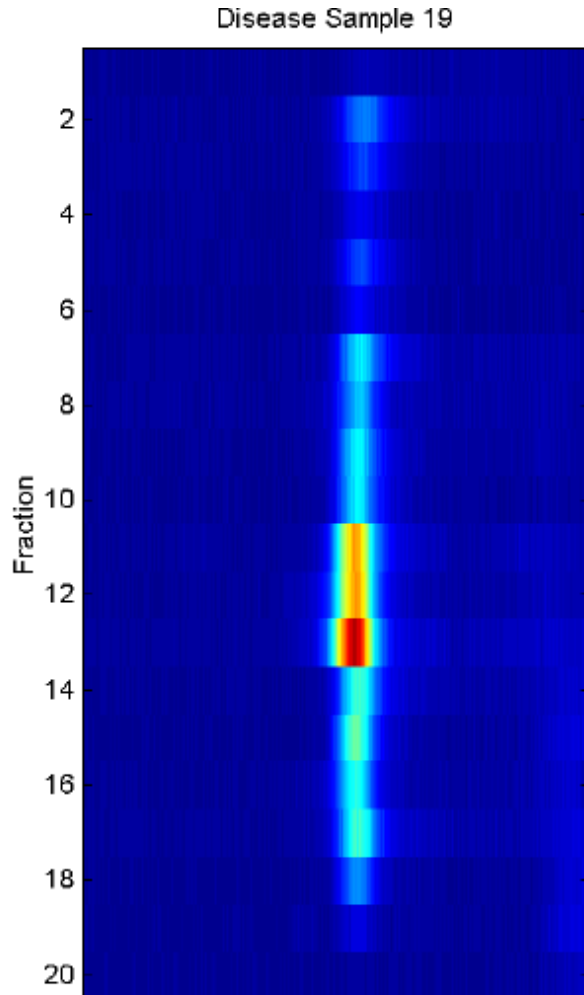
Preprocessing (III)

Total ion current normalization

- Divide each spectrum by total ion current (sum of all intensities)
- Similar to median normalization in microarrays

Preprocessing (IV)

- Recombining fractions

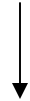


- introduction to mass spec

<http://masspec.scripps.edu/information/intro/>

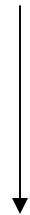
Data Reduction

- 60831 (time ticks) x 20 (fractions) x 41 (samples)



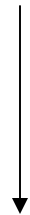
recombining fractions

60831 x 41



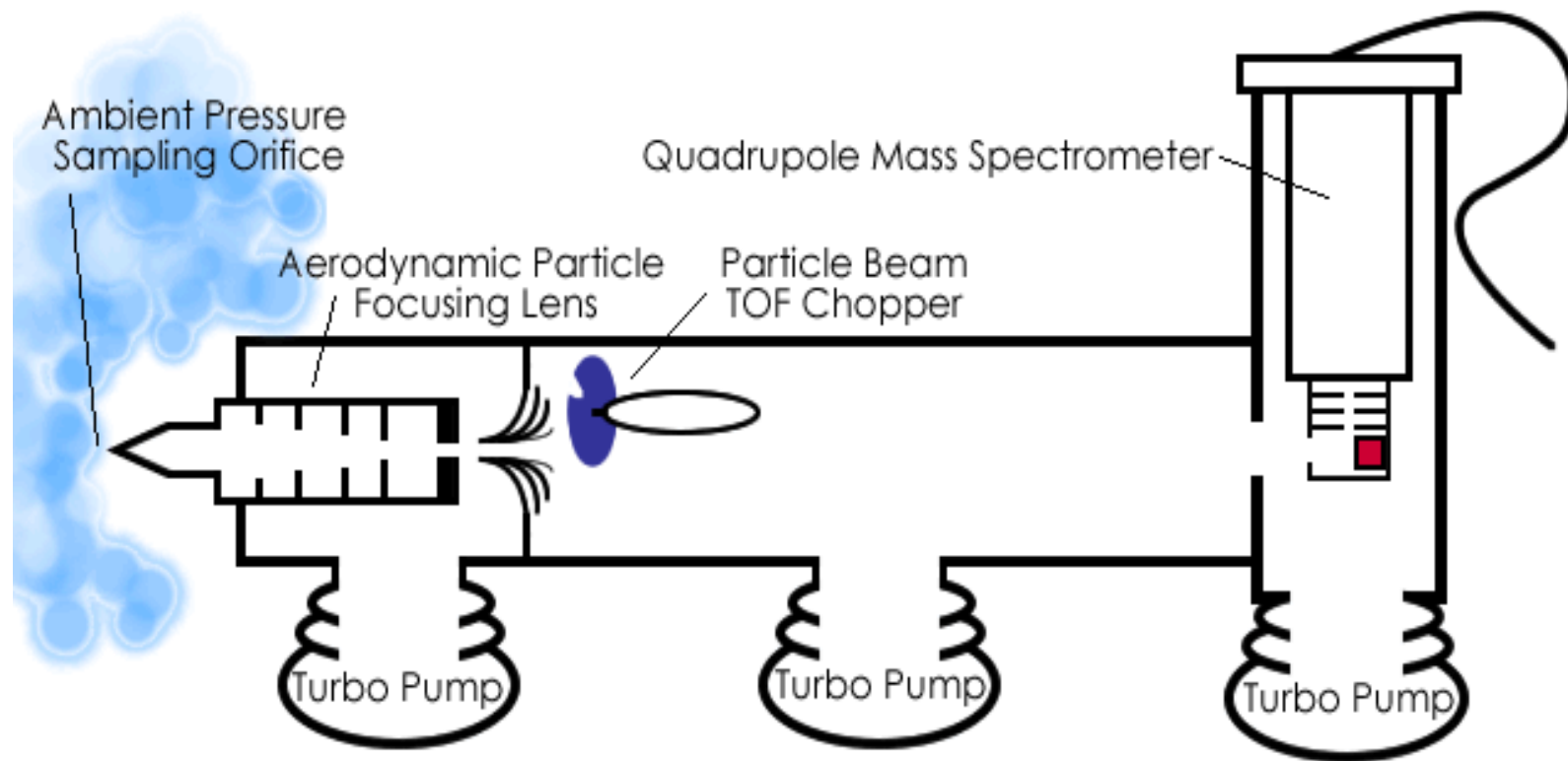
windowing and choosing
max. intensity for each spectrum

~2000 x 41



discard windows that do not
contain peak

506 x 41



Credit: Matt Thyson (Lexington, Massachusetts)
Source <http://cires.colorado.edu/~jjose/ams.html>