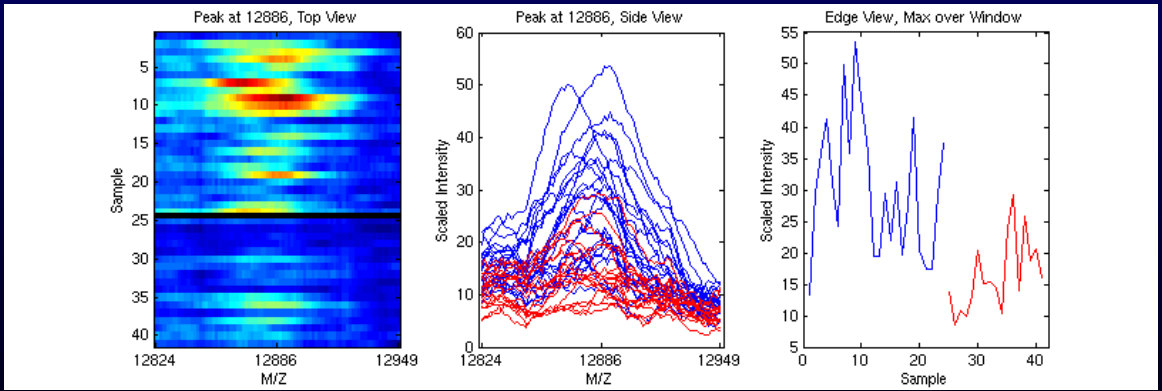
The Analysis of Proteomic Spectra from Serum Samples

Keith Baggerly

Biostatistics & Applied Mathematics
MD Anderson Cancer Center



What Are Proteomic Spectra?

DNA makes RNA makes Protein

Microarrays allow us to measure the mRNA complement of a set of cells

Mass spectrometry allows us to measure the protein complement (or subset thereof) of a set of cells

Proteomic spectra are mass spectrometry traces of biological specimens

Why Are We Excited?

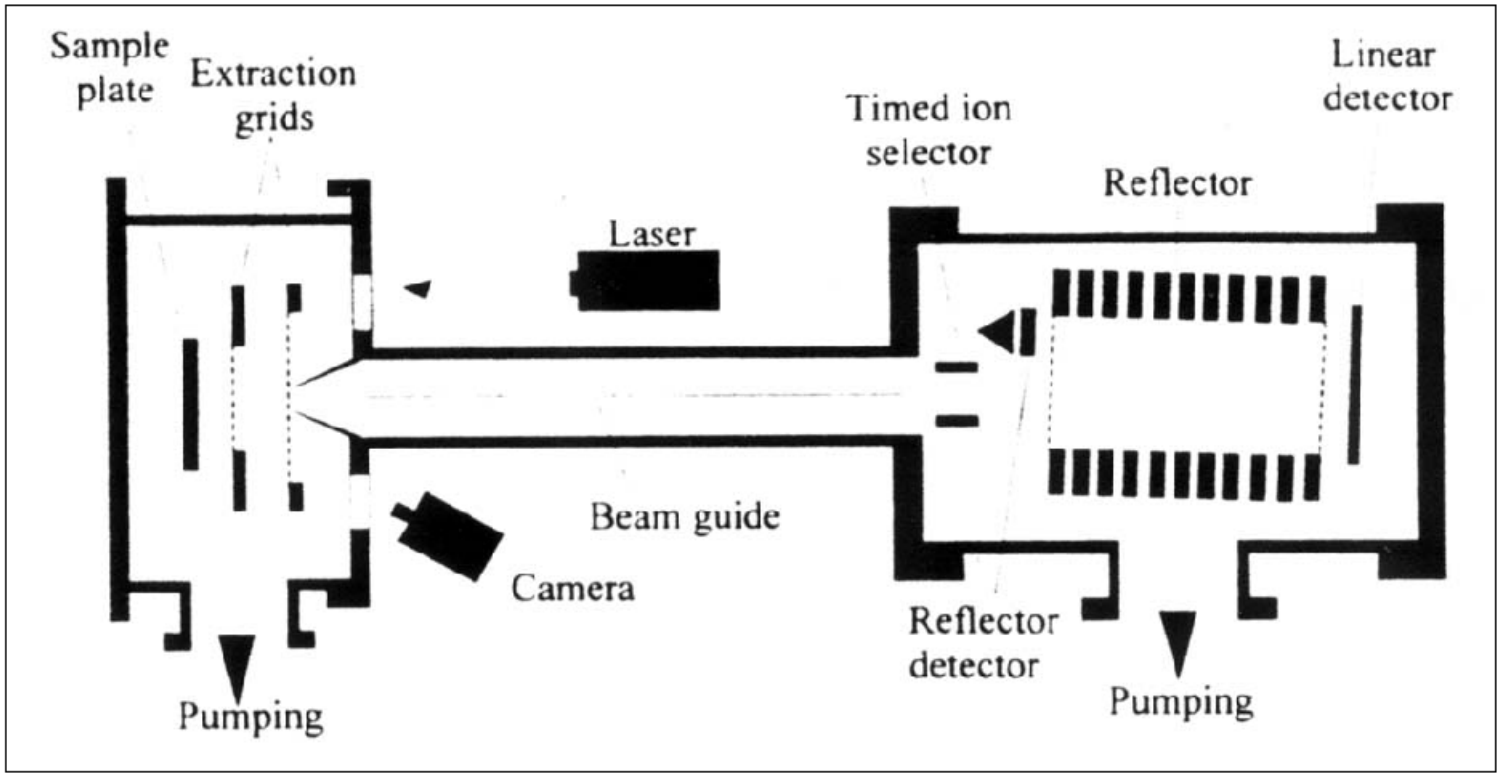
Profiles at this point are being assessed using serum and urine, not tissue biopsies

Spectra are cheaper to run on a per unit basis than microarrays

Can run samples on large numbers of patients

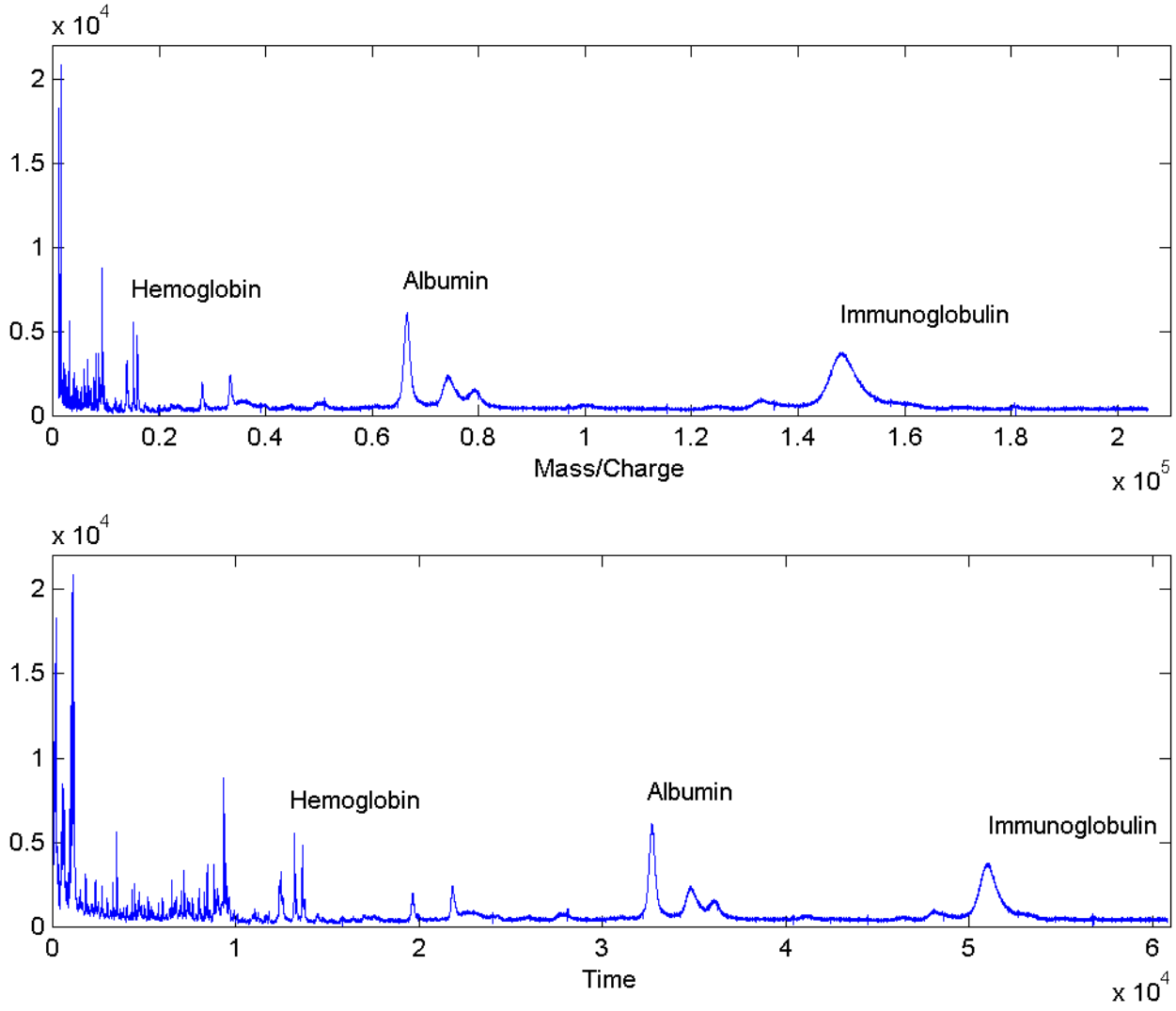
How Does Mass Spec Work?

Block Diagram of a MALDI-TOF

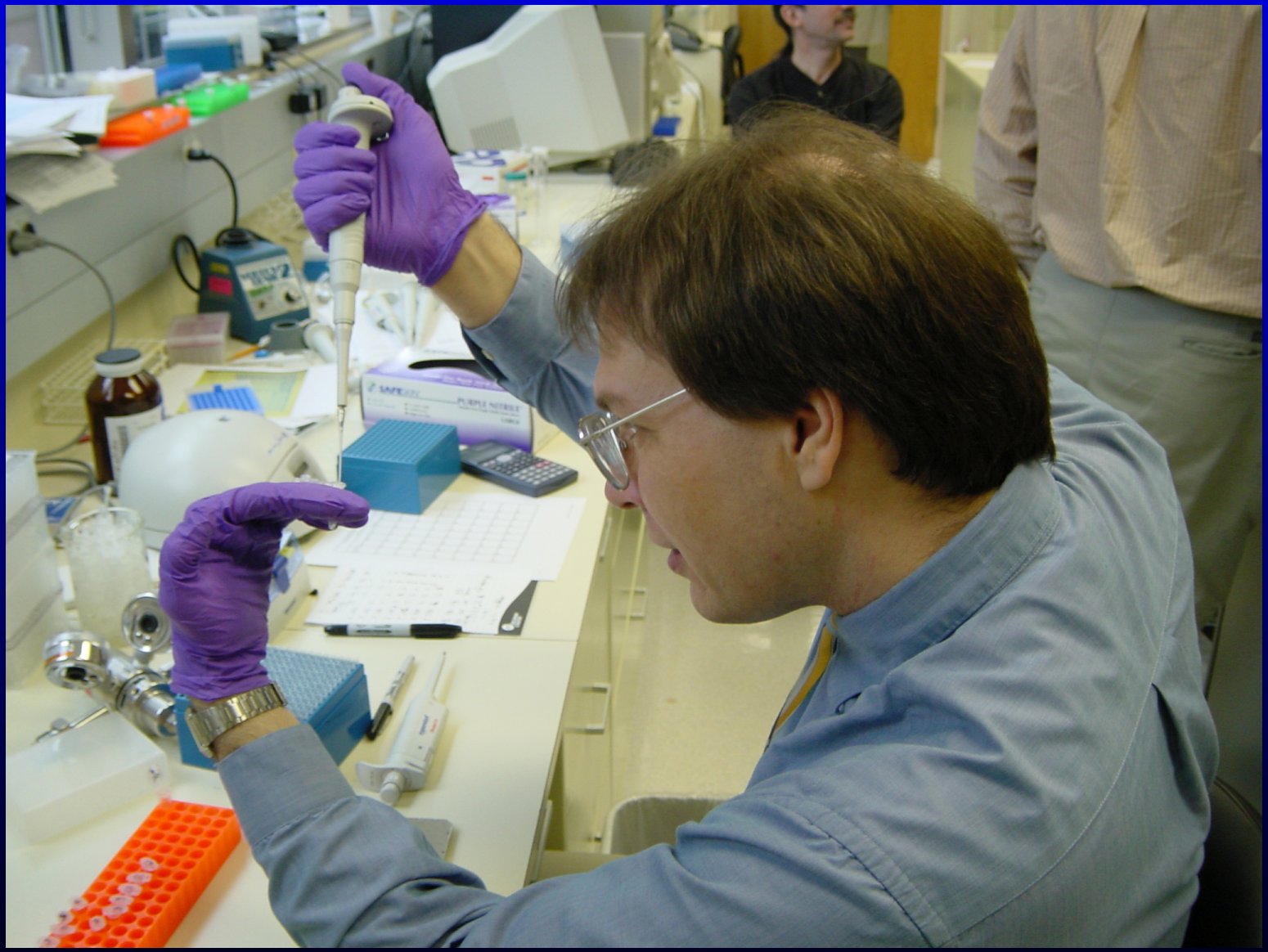


Vestal and Juhasz. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 892.

What Do the Data Look Like?

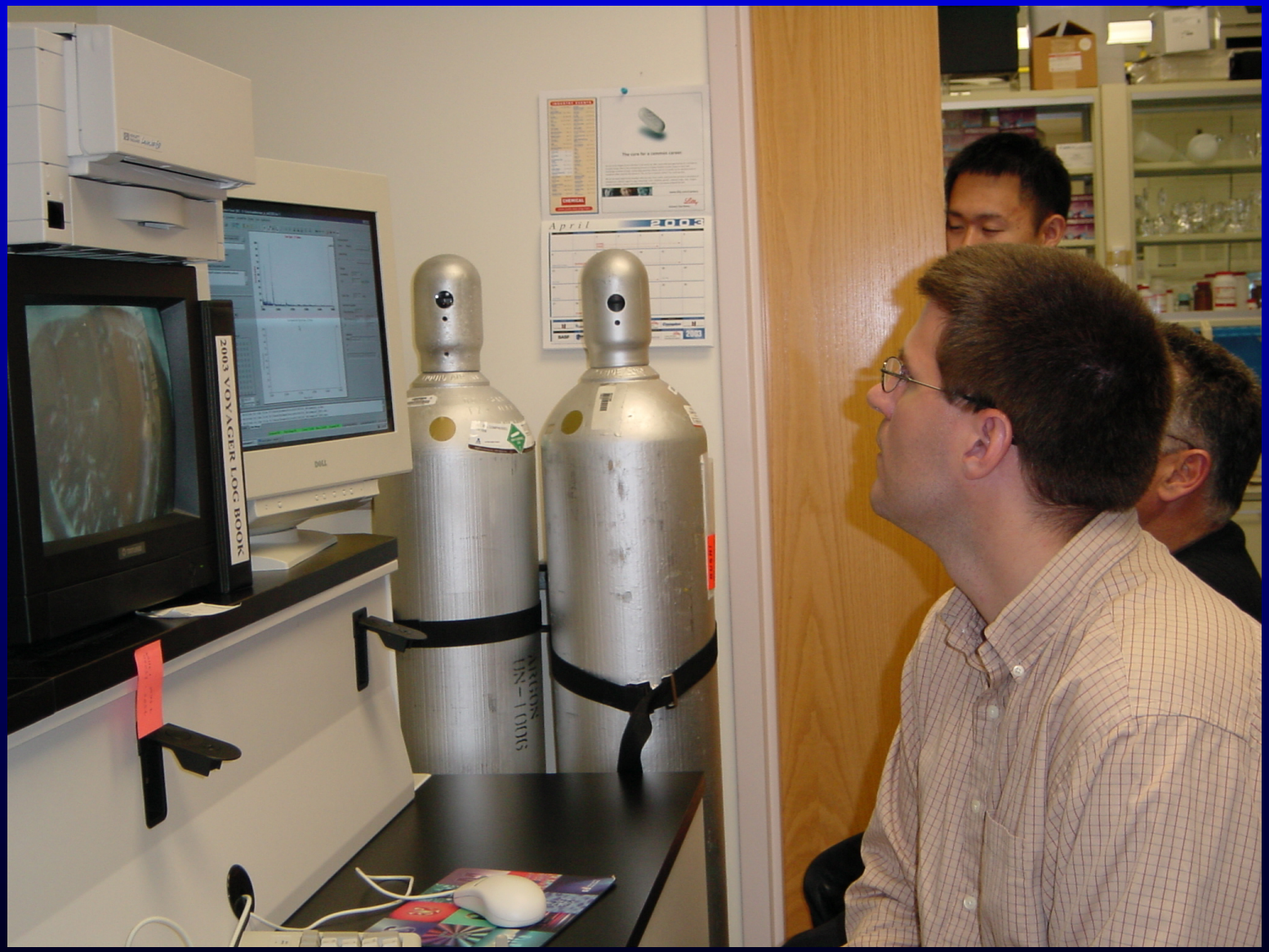


Learning: Spotting the Samples



What the Guts Look Like

Taking Data



Some Other Common Steps

Fractionating the Samples

Changing the Laser Intensity

Working with Different Matrix Substrates

SELDI: A Special Case

www.ciphergen.com

Precoated surface performs some preselection of the proteins for you.

Machines are nominally easier to use.

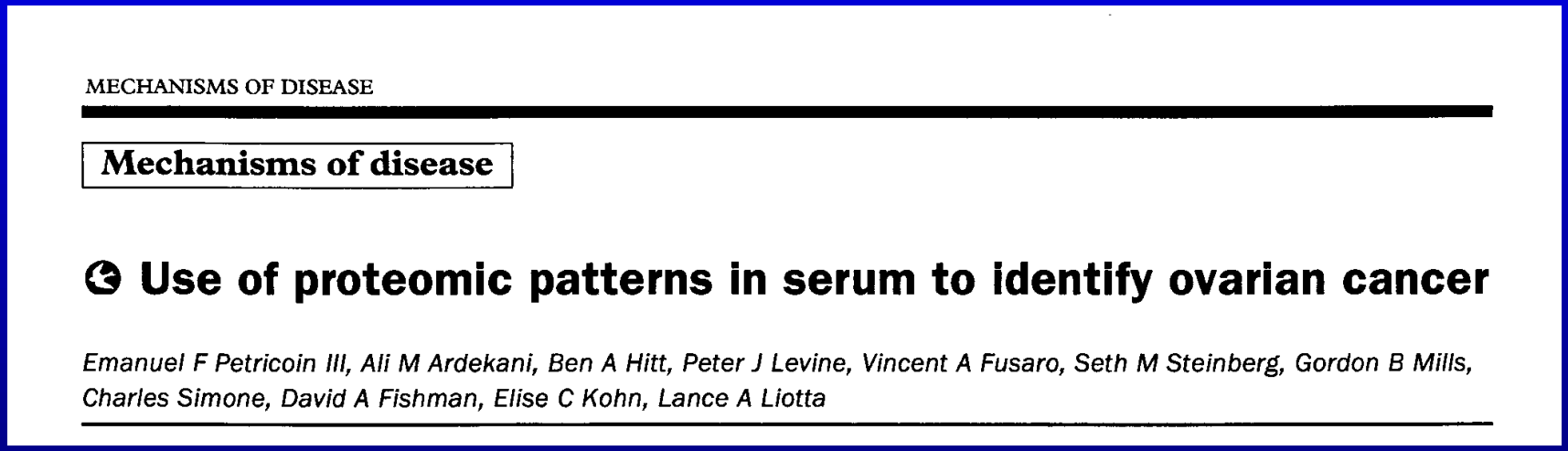
A Tale of Two Examples

Example 1 – Learning from the literature (SELDI)

Example 2 – Testing out our understanding (MALDI)

A story in pictures

A SELDI Example: Feb 16 '02 Lancet



MECHANISMS OF DISEASE

Mechanisms of disease

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

- 100 ovarian cancer patients
- 100 normal controls
- 16 patients with "benign disease"

Use 50 cancer and 50 normal spectra to train a classification method; test the algorithm on the remaining samples.

Their Results

- Correctly classified 50/50 of the ovarian cancer cases.
- Correctly classified 46/50 of the normal cases.
- Correctly classified 16/16 of the benign disease as “other”.

Data at http://www.ncifdaproteomics.com (used to be at http://clinicalproteomics.steem.com)

Large sample sizes, using serum

The Data Sets

3 data sets on ovarian cancer

Data Set 1 – The initial experiment. 216 samples, baseline subtracted, H4 chip

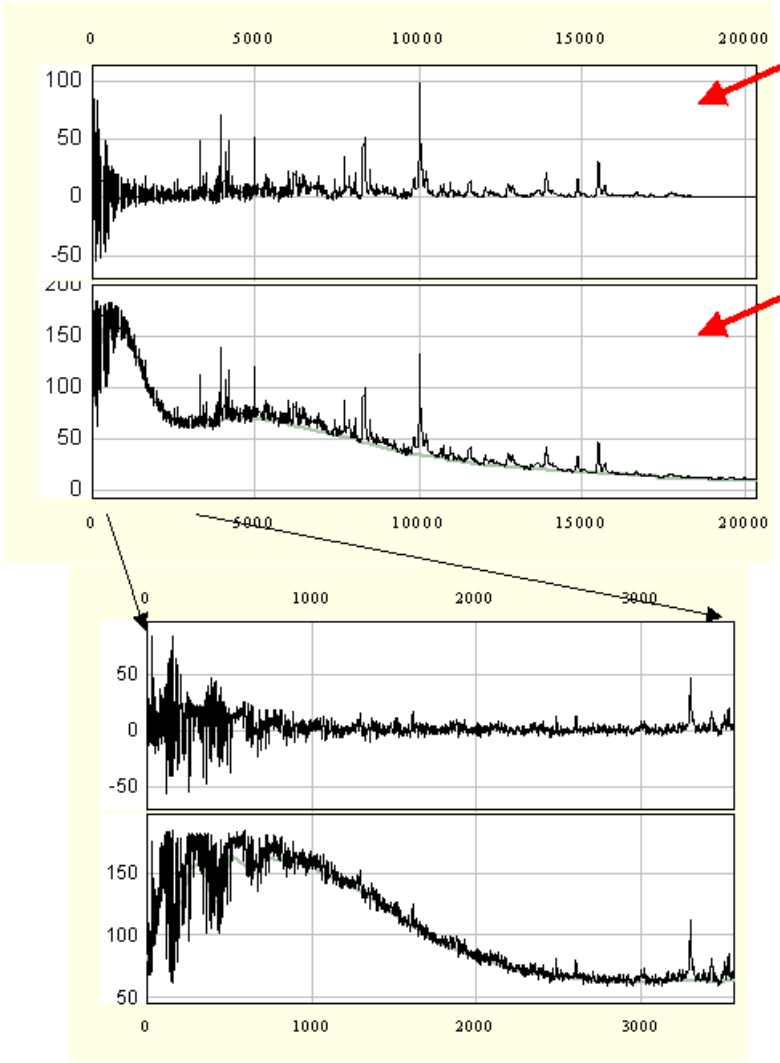
Data Set 2 – Followup: the same 216 samples, baseline subtracted, WCX2 chip

Data Set 3 – New experiment: 162 cancers, 91 normals, baseline NOT subtracted, WCX2 chip

A set of 5-7 separating peaks is supplied for each data set.

We tried to (a) replicate their results, and (b) check consistency of the proteins found

We Can't Replicate their Results (DS1 & DS2)



They posted this

Baseline Subtraction ON

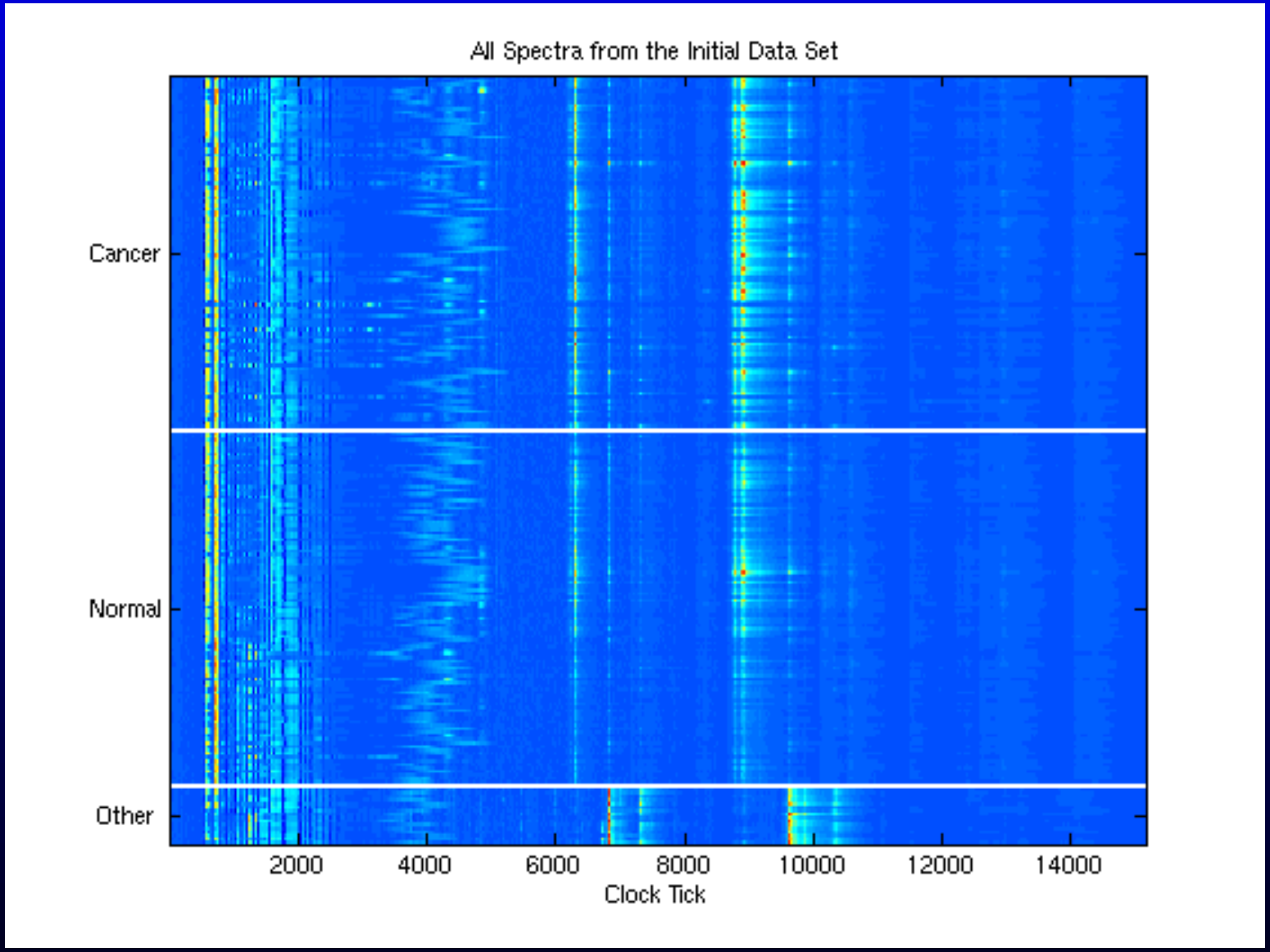
They analyzed this

Baseline Subtraction OFF

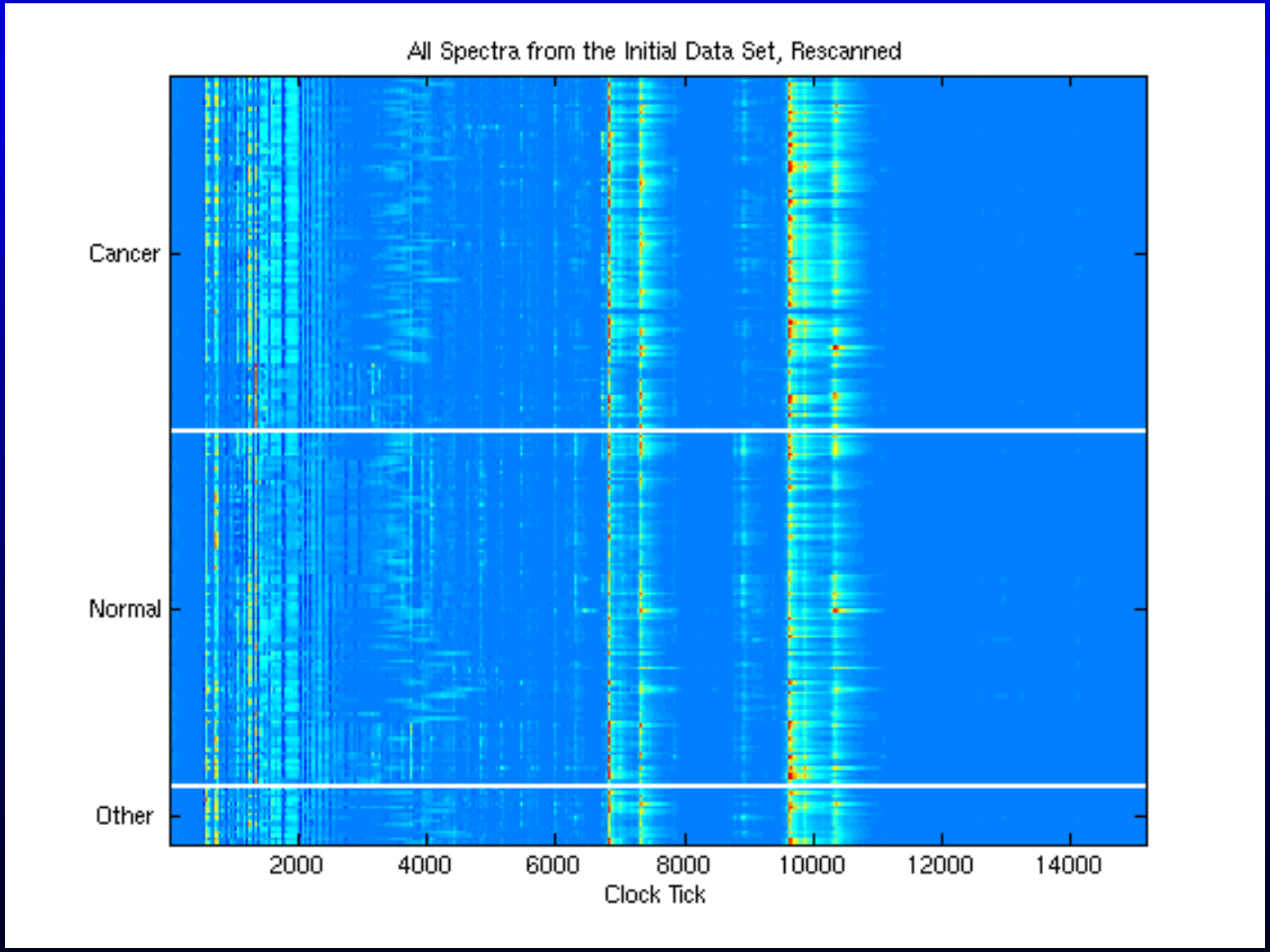
Zoomed in with Baseline Sub ON

Zoomed in with Baseline Sub OFF

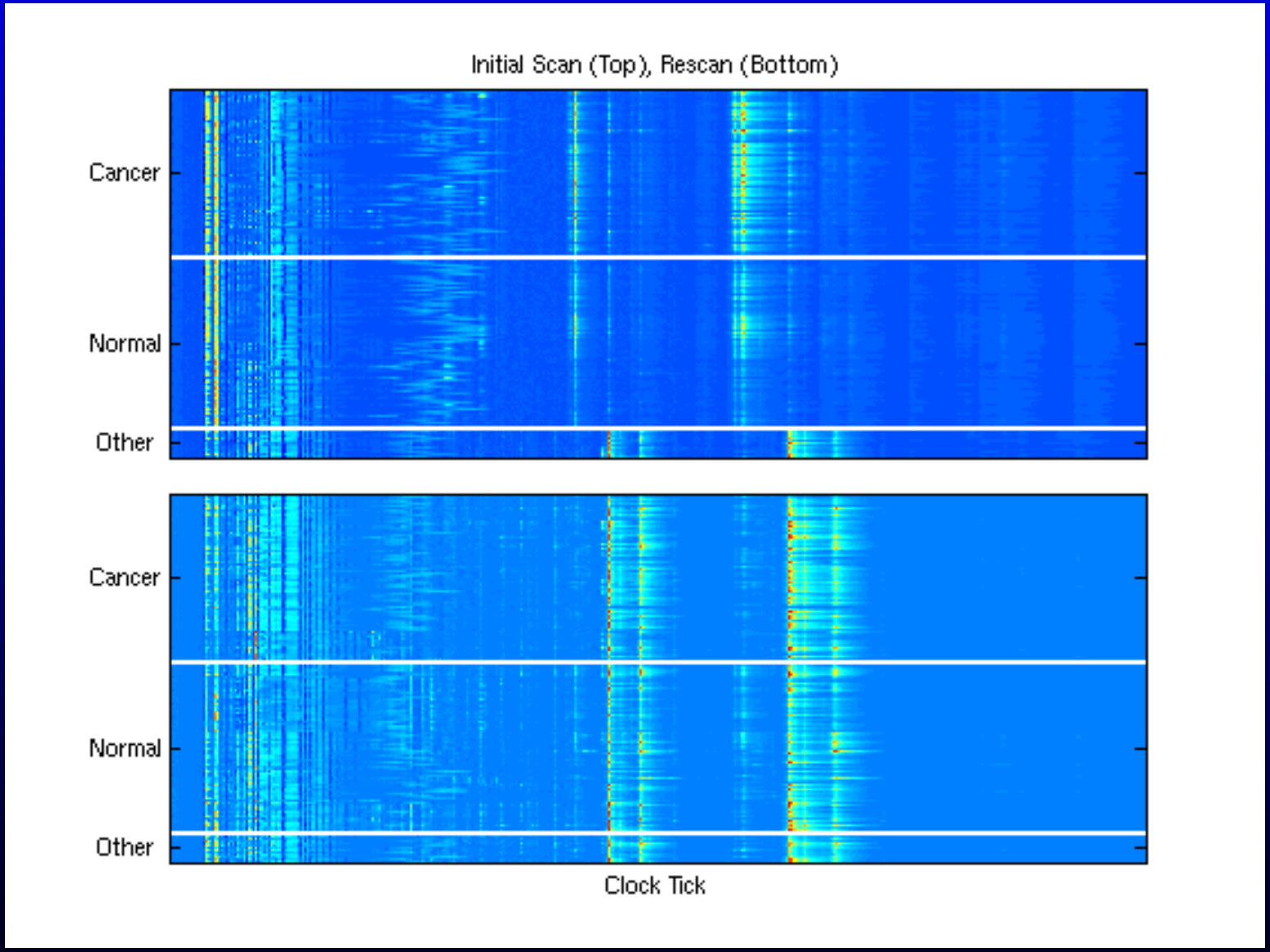
Some Structure is Visible in DS1



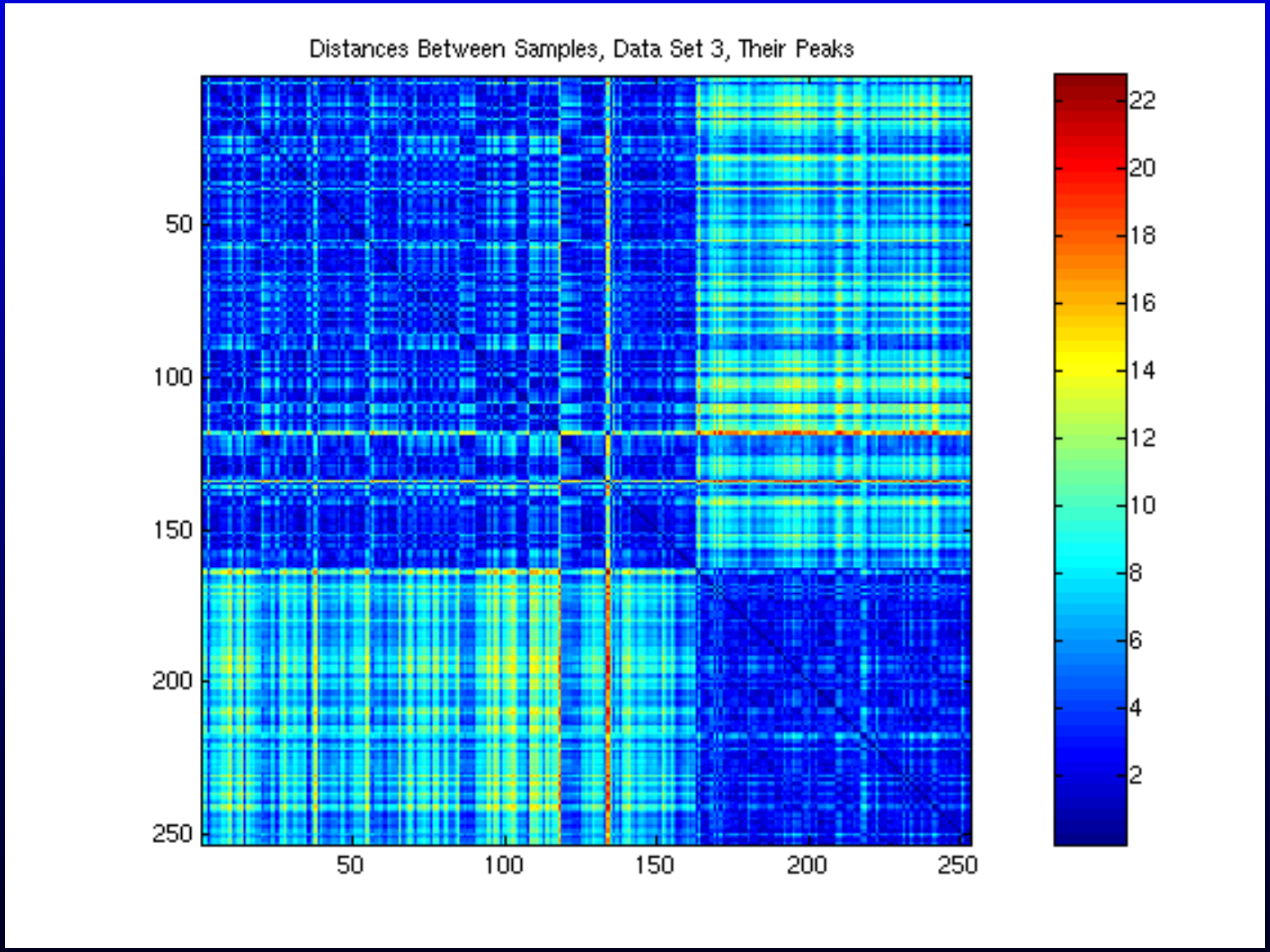
Or is it? Not in DS2



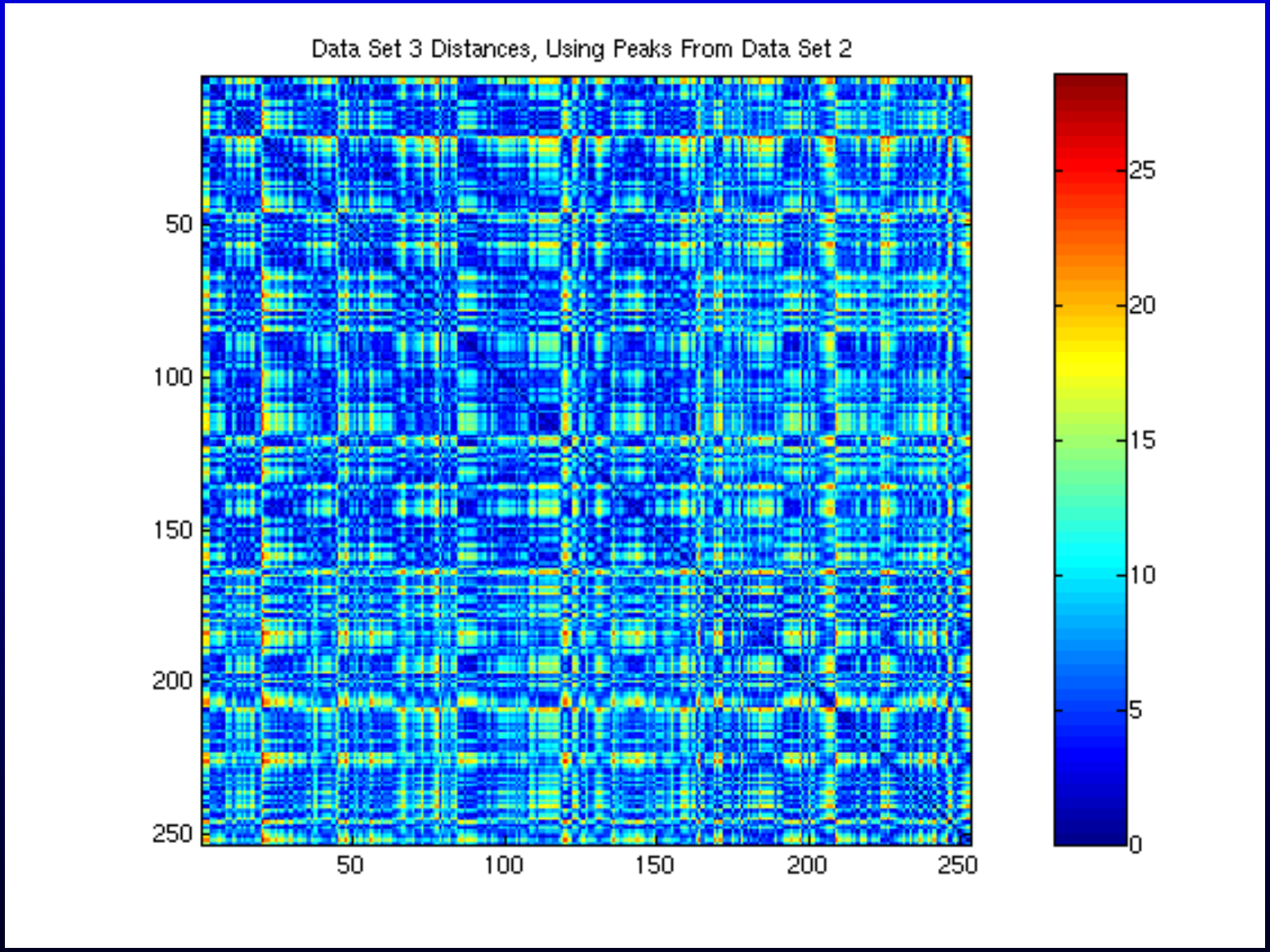
Processing Can Trump Biology (DS1 & DS2)



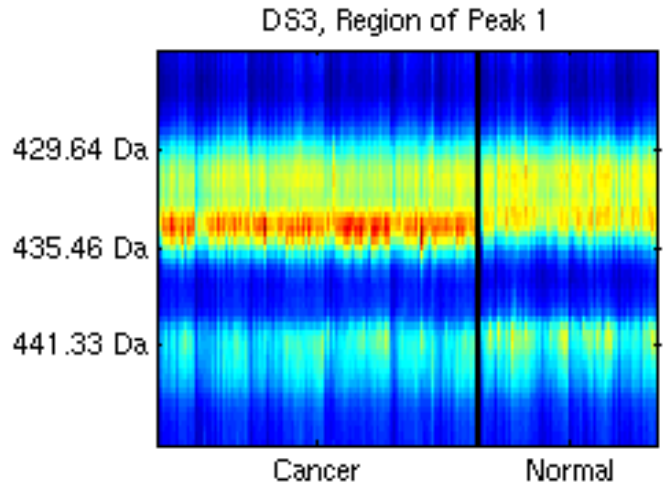
We Can Analyze Data Set 3!

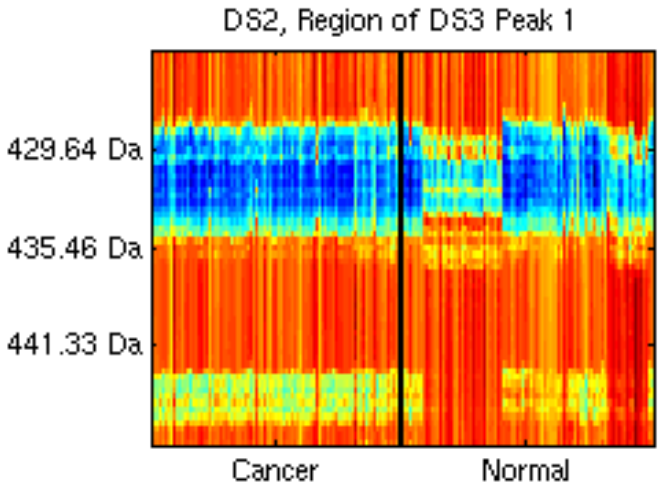


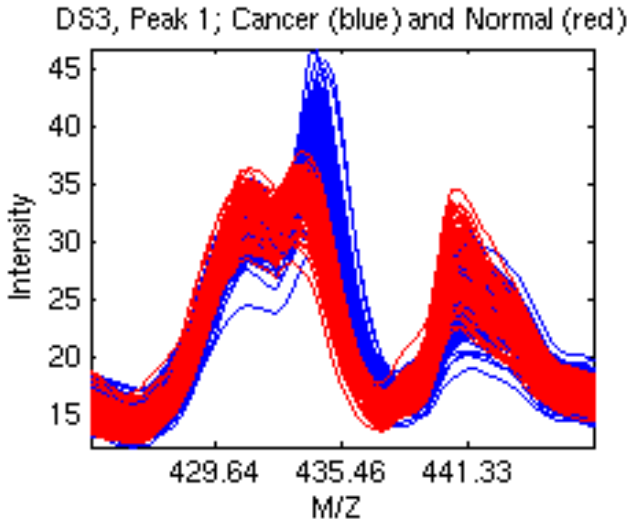
Do the DS2 Peaks Work for DS3?

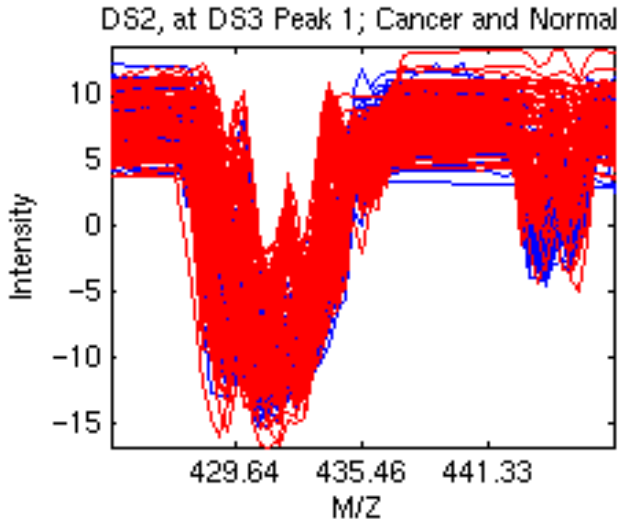


Do the DS3 Peaks Work for DS2?

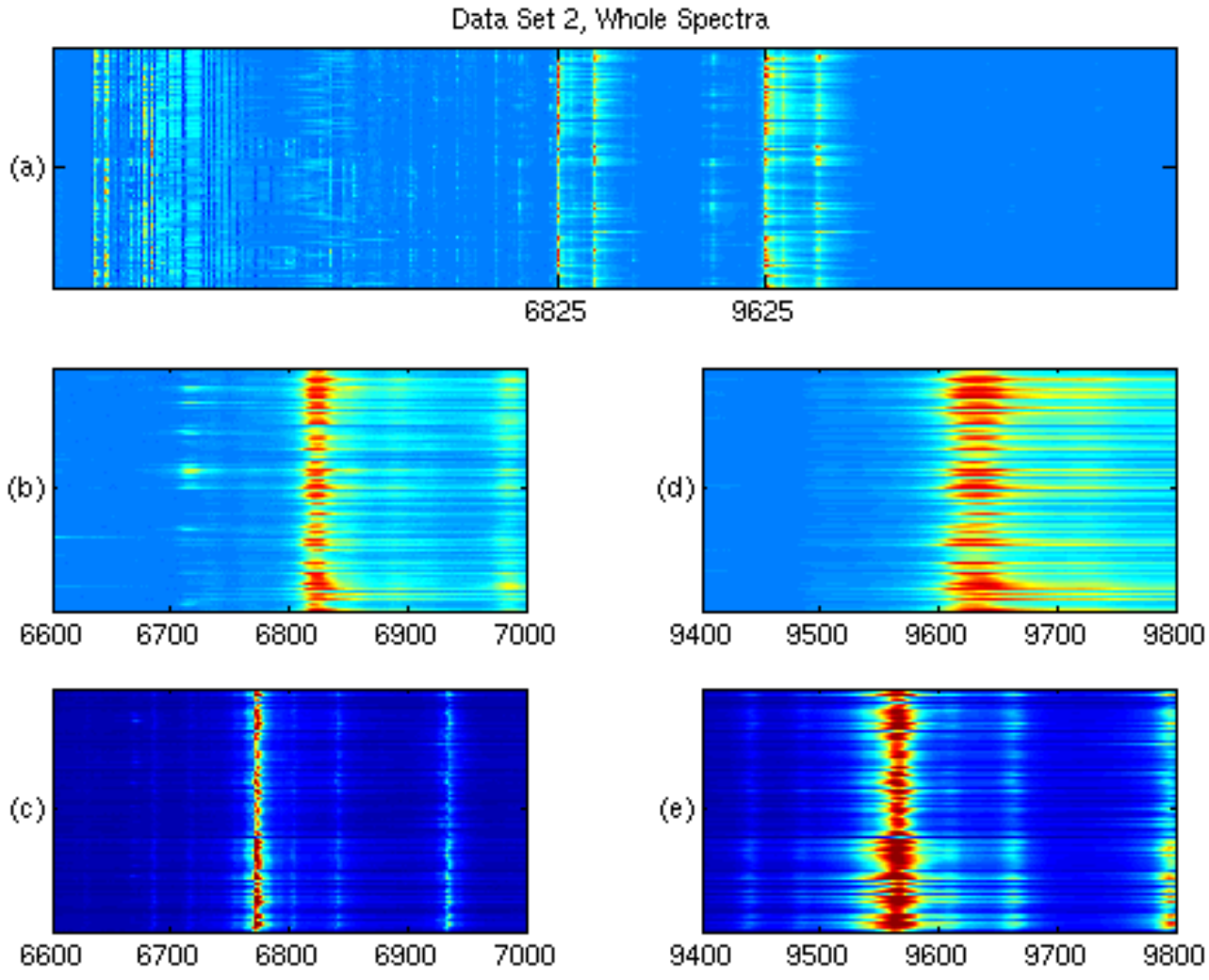




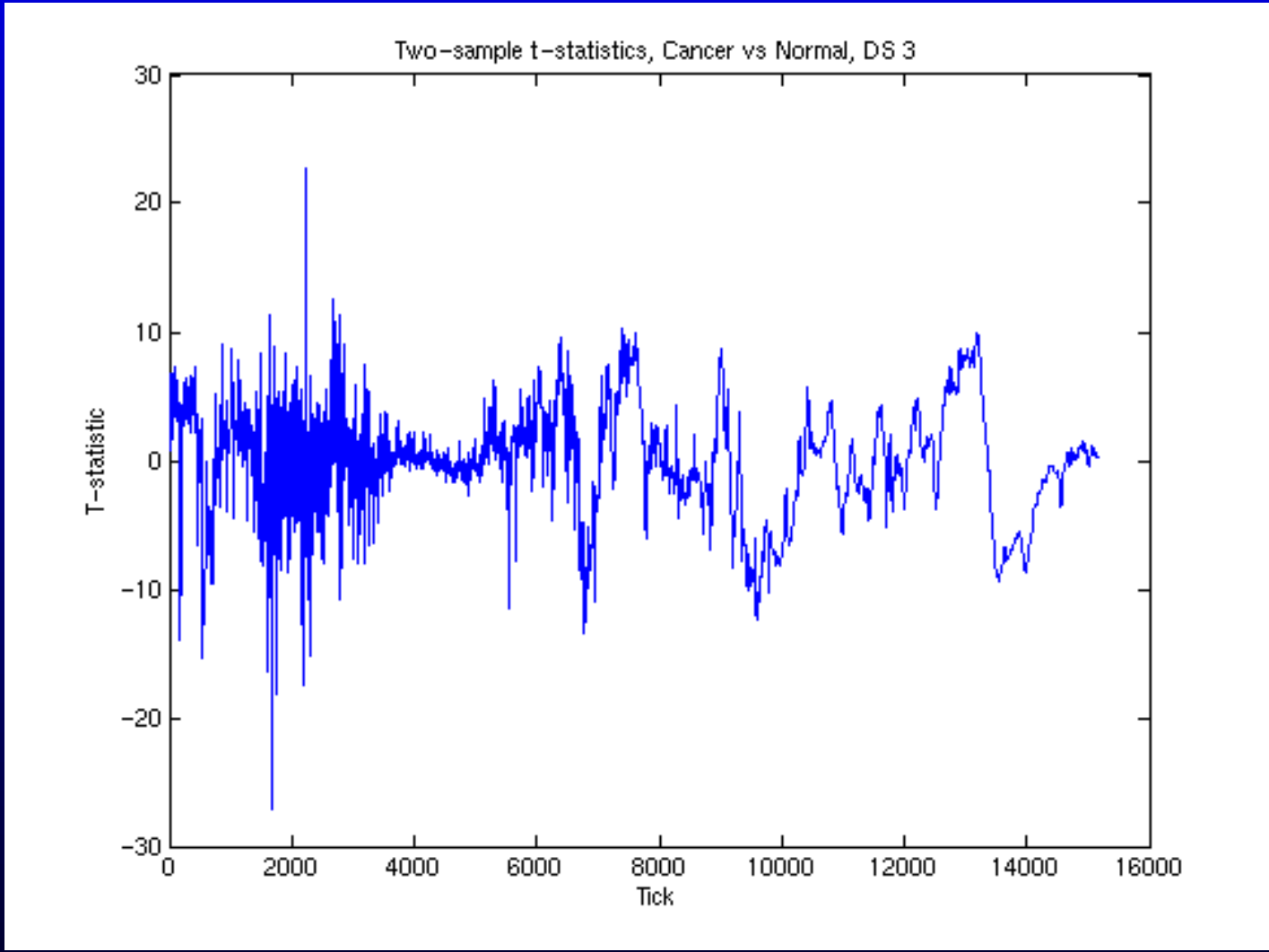




Peaks are Offset

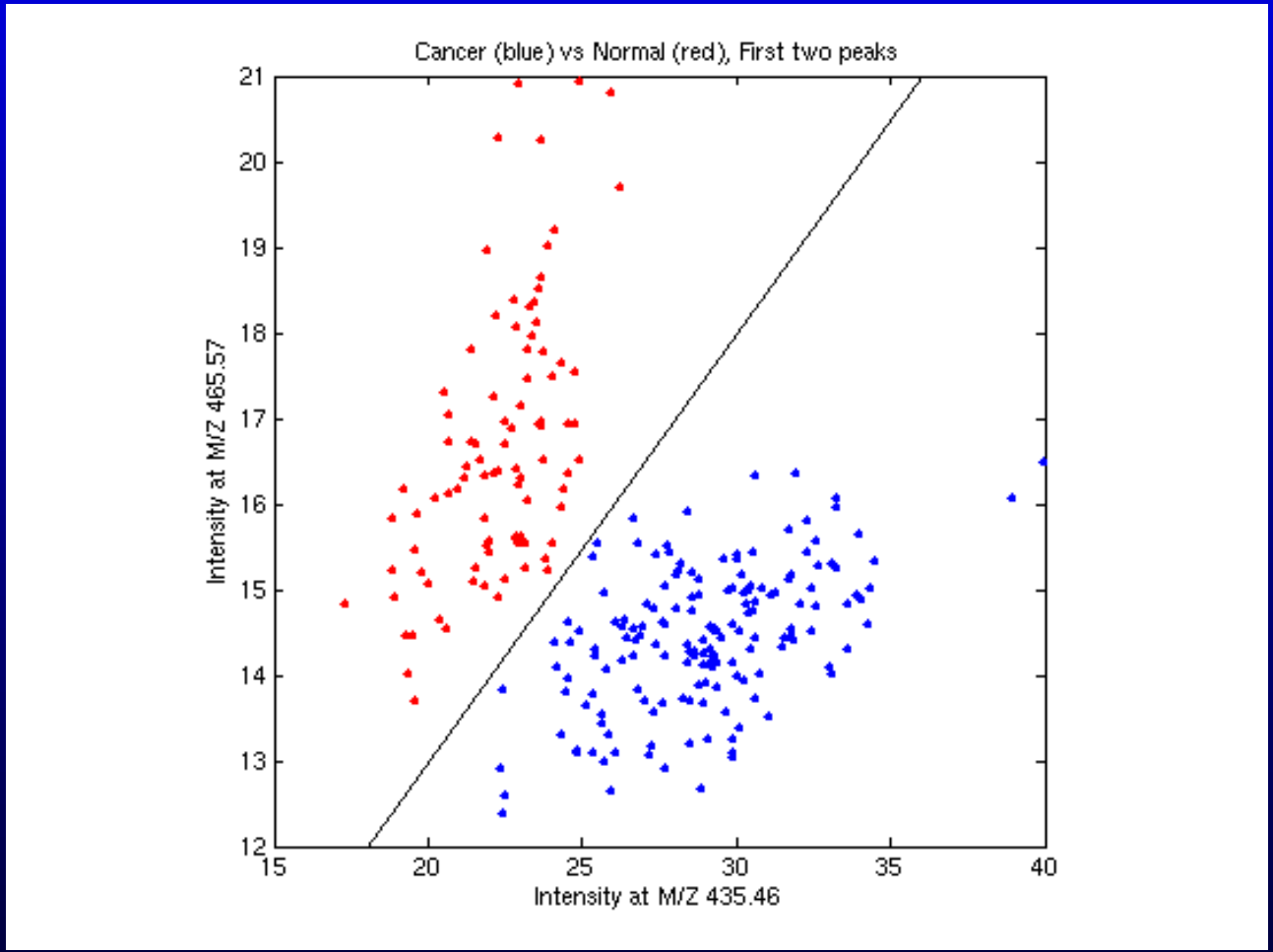


Which Peaks are Best? T-statistics



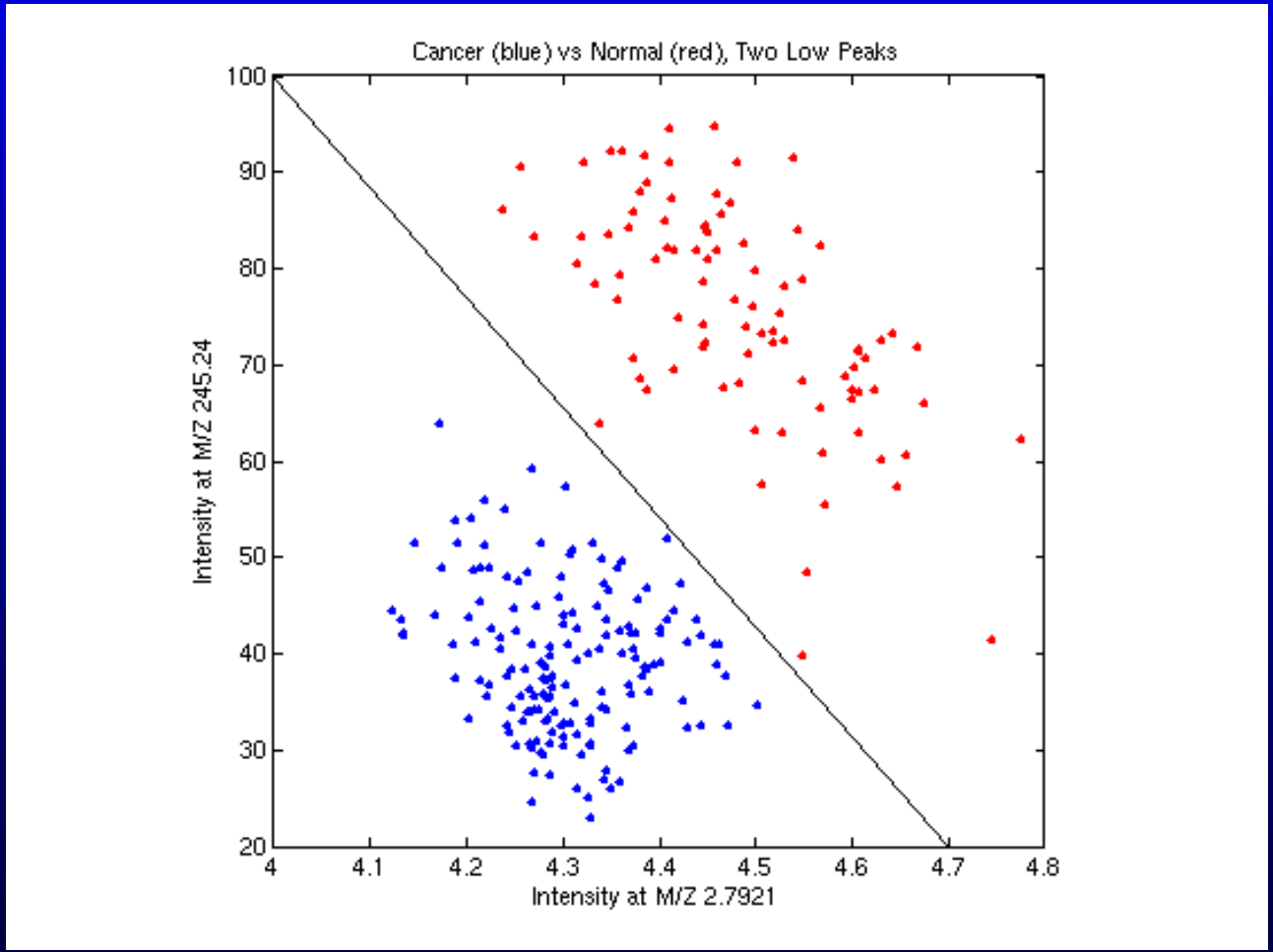
Note the magnitudes: t-values in excess of 20 (absolute value)!

One Bivariate Plot: M/Z = (435.46,465.57)



Perfect Separation. These are the first 2 peaks in their list, and ones we checked against DS2.

Another Bivariate Plot: M/Z = (2.79,245.2)



Perfect Separation, using a completely different pair. Further, look at the masses: this is the noise region.

Perfect Classification with Noise?

This is a problem, in that it suggests a qualitative difference in how the samples were processed, not just a difference in the biology.

This type of separation reminds us of what we saw with benign disease.

(Sorace and Zhan, BMC Bioinformatics, 2003)

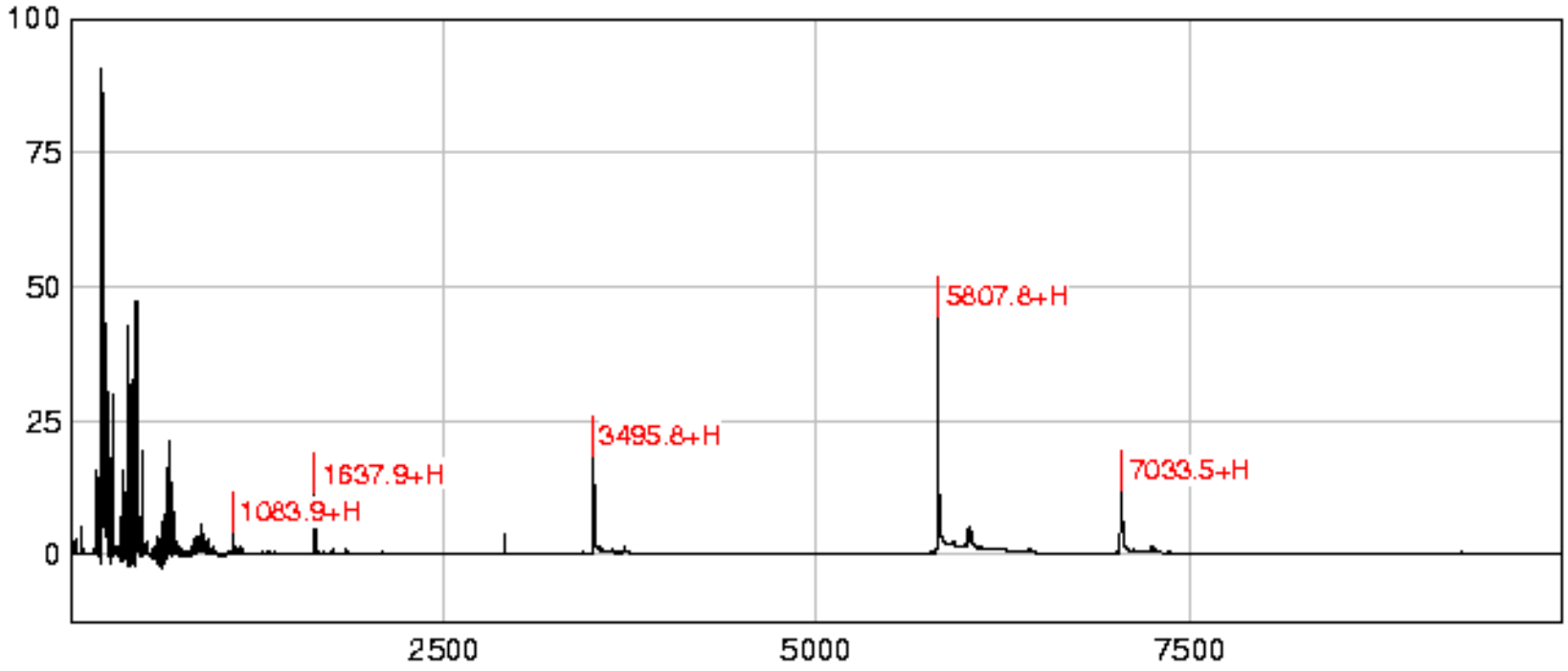
Mass Accuracy is Poor?

A tale of 5 masses...

Feb ’02 DS1	Apr ’02 DS2	Jun ’02 DS3
-7.86E-05	-7.86E-05	-7.86E-05
2.18E-07	2.18E-07	2.18E-07
9.60E-05	9.60E-05	9.60E-05
0.000366014	0.000366014	0.000366014
0.000810195	0.000810195	0.000810195

How are masses determined?

Calibrating known proteins



Calibration is the Same?

M/Z vectors the same for all three data sets.

Machine calibration the same for 4+ months?

What is the Calibration Equation?

The Ciphergen equation

$$\frac{m/z}{U} = a(t - t_0)^2 + b, \quad U = 20K, t = (0, 1, ...) * 0.004$$

Fitting it here

$$a = 0.2721697 * 10^{-3}, \quad b = 0, \quad t_0 = 0.0038$$

What is the Calibration Equation?

The Ciphergen equation

$$\frac{m/z}{U} = a(t - t_0)^2 + b, \quad U = 20K, t = (0, 1, ...) * 0.004$$

Fitting it here

$$a = 0.2721697 * 10^{-3}, \quad b = 0, \quad t_0 = 0.0038$$

These are the default settings that ship with the software!

Other issues

Prostate Cancer

Q-star data different

clinical trials?

A MALDI Example: Proteomics Data Mining

41 samples, 24 with lung cancer*, 17 controls.

20 fractions per sample.

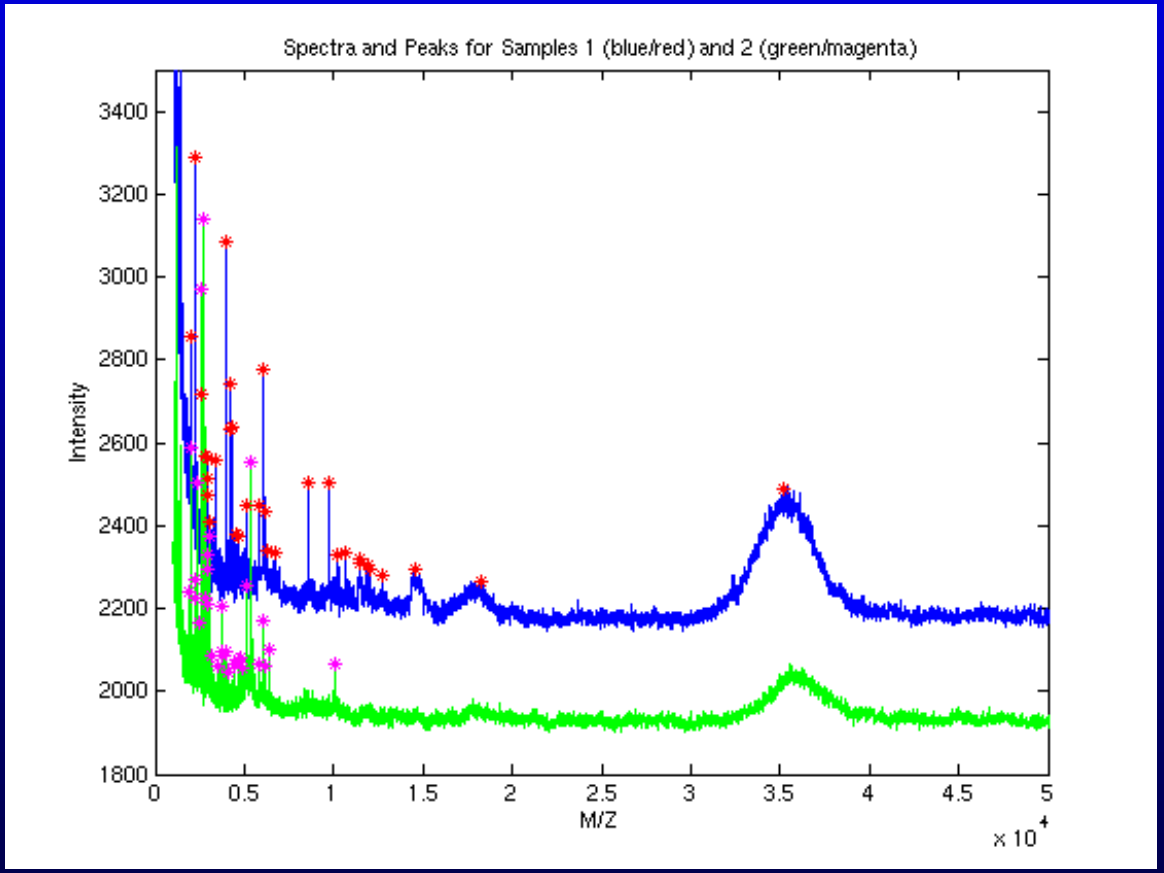
Goal: distinguish the two groups; we know this can be done due to the “zip effect”.

Data used to be at

http://www.radweb.mc.duke.edu/cme/proteomics/explain.htm

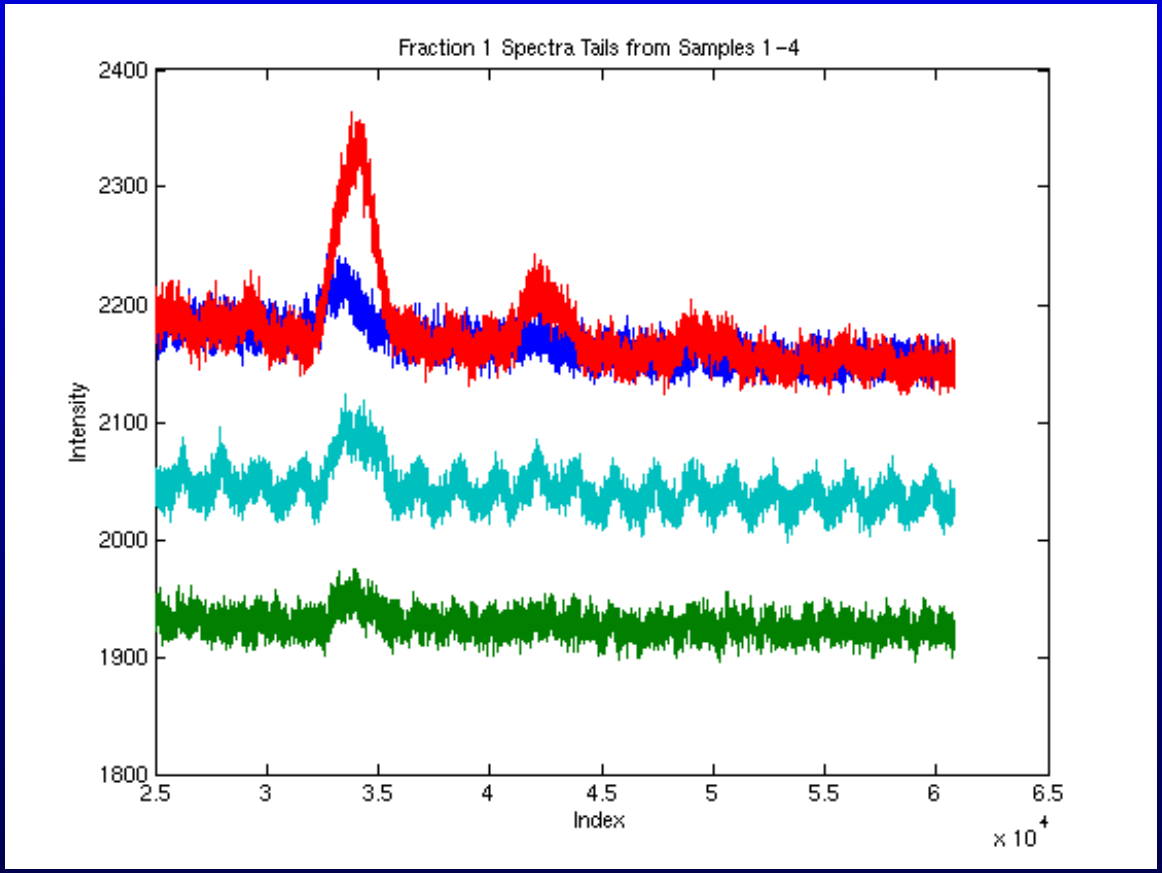
but the site has been retired. Send email to Ned Patz or Mike Campa at Duke if interested (Campa002@mc.duke.edu, patz0002@mc.duke.edu).

Raw Spectra Have Different Baselines



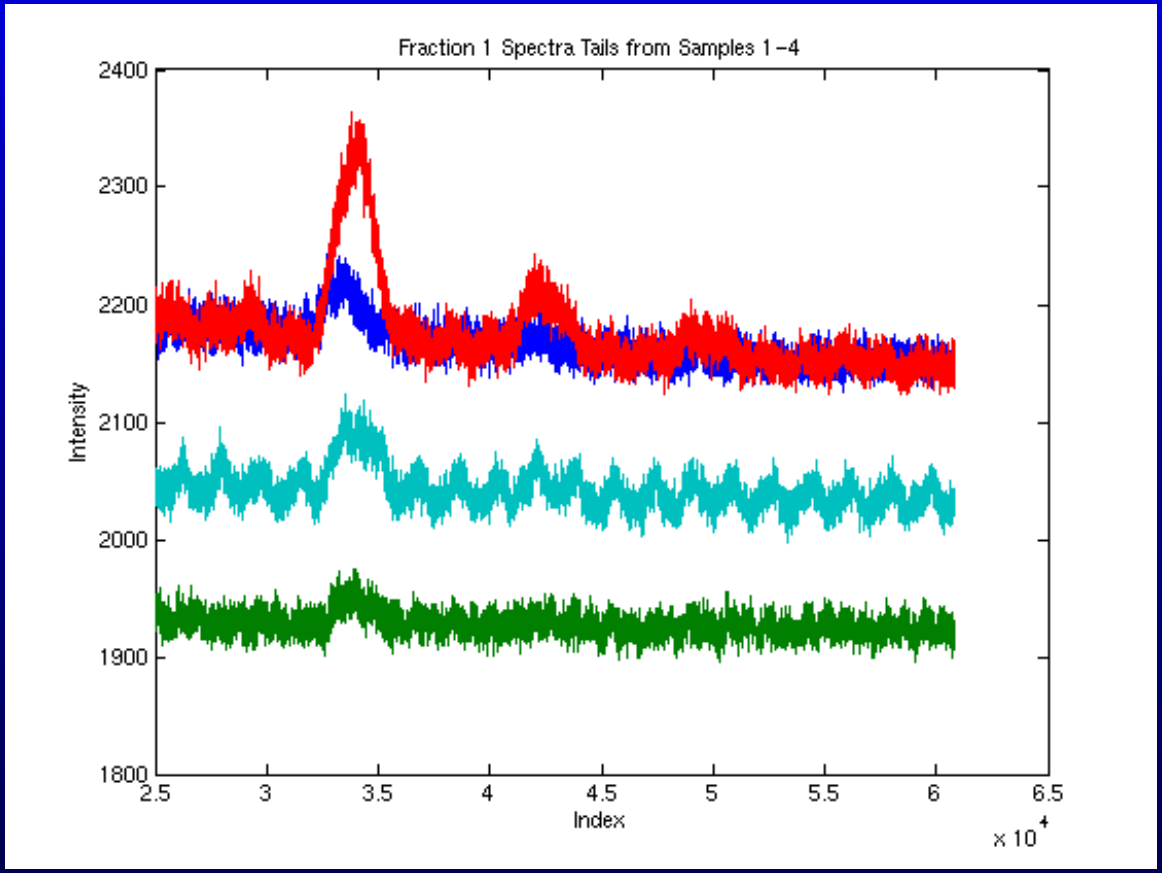
Note the need for baseline correction and normalization.

Oscillatory Behavior...



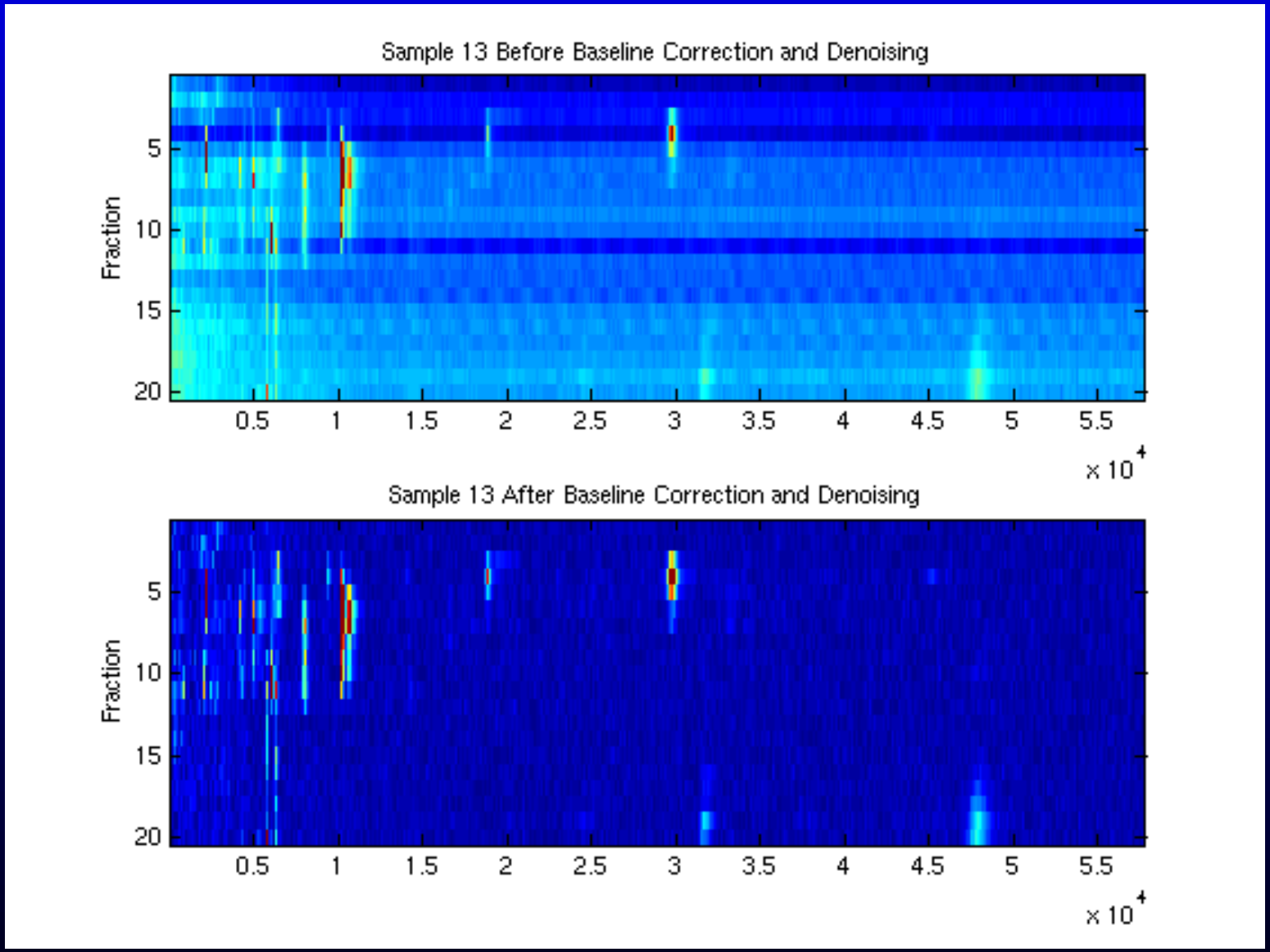
Roughly half the spectra have sinusoidal noise.

Oscillatory Behavior...

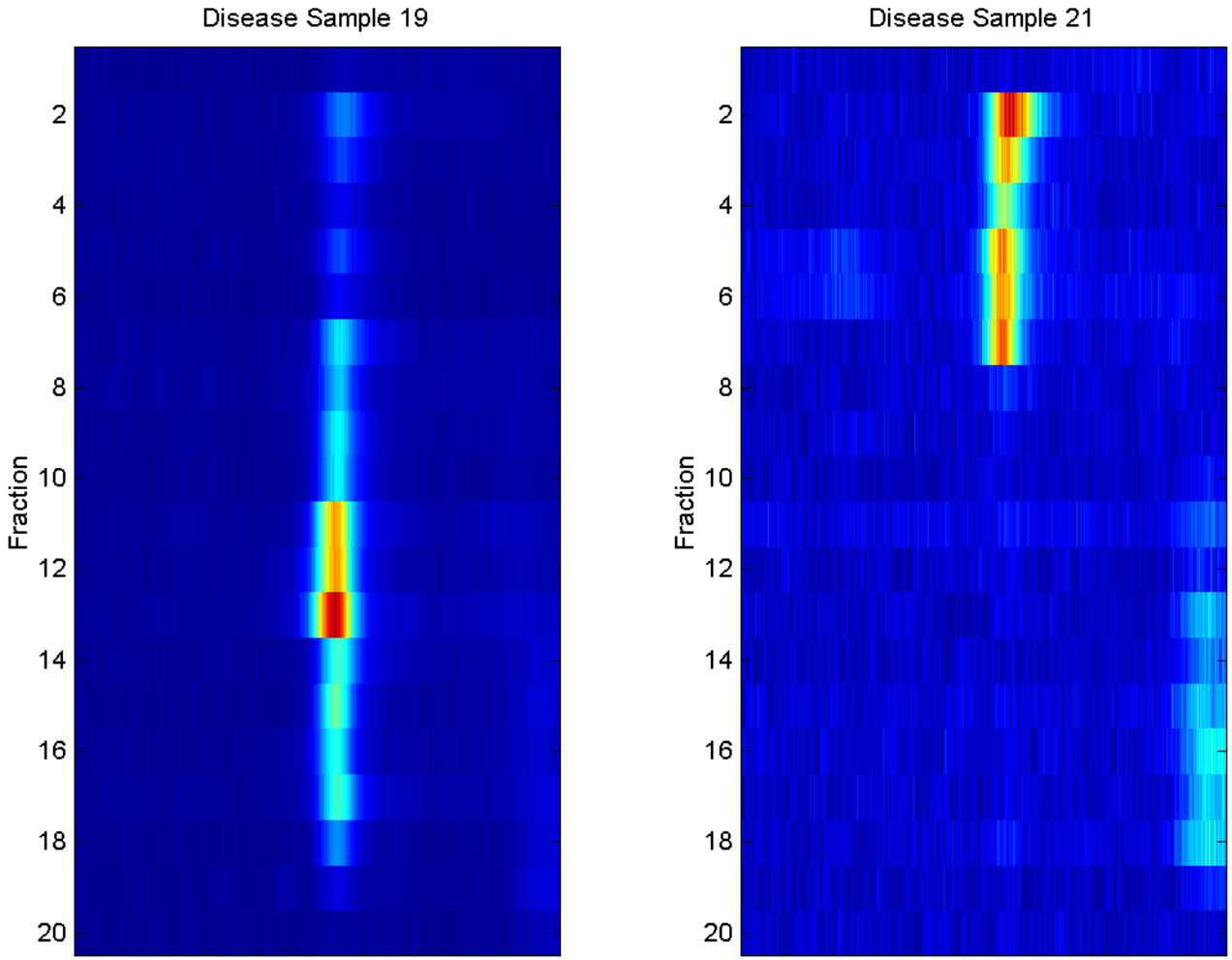


Roughly half the spectra have sinusoidal noise. We're seeing the A/C power cord.

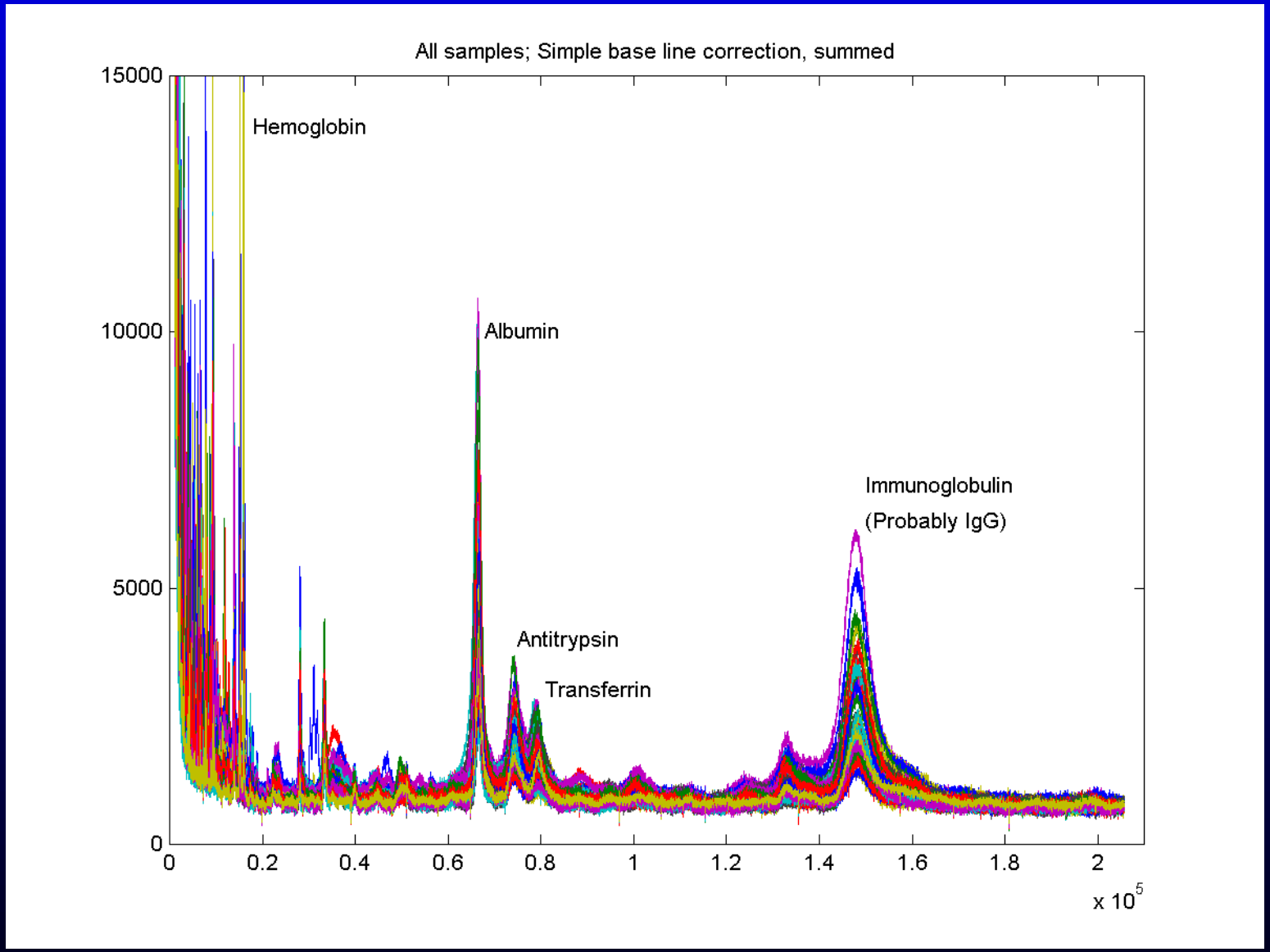
Baseline Adj: Fraction Agreement, Before & After



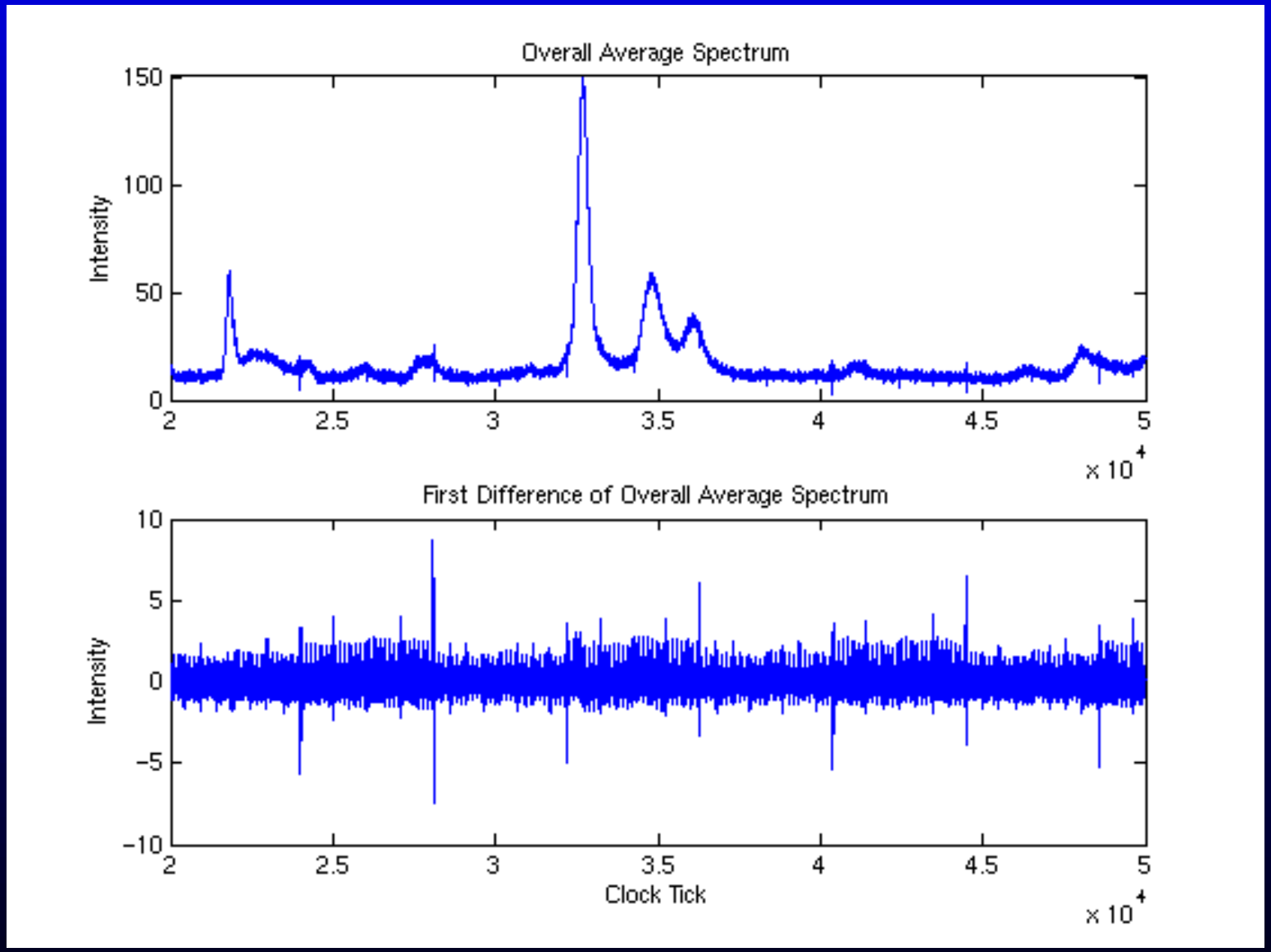
Fractionation is Unstable



Unfractionating the Data

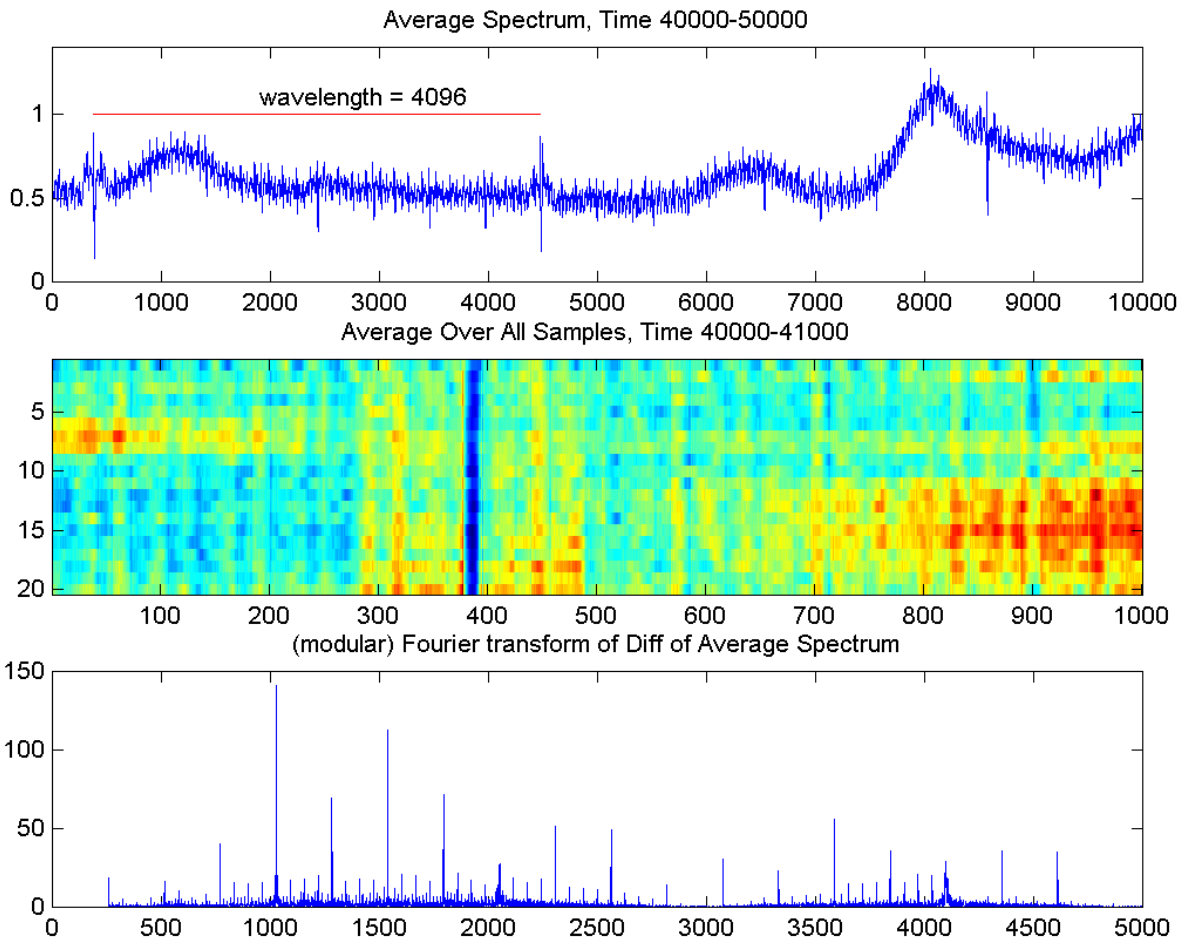


The Overall Average Shows Spikes. Difference It.



Computer Buffer?

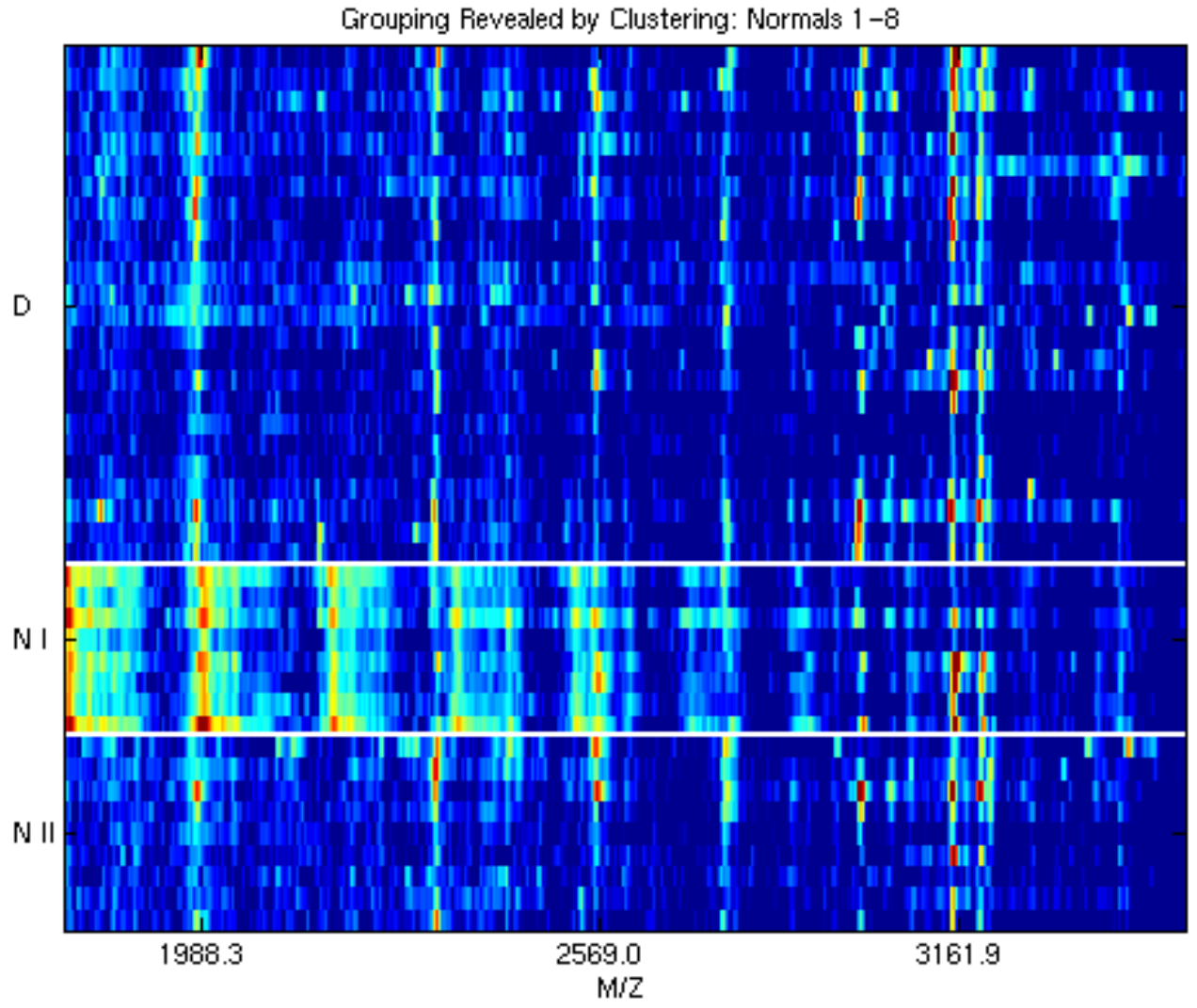
Spike spacing has a wavelength of $4096 = 2^{12}$.



Are We Done Cleaning Yet?

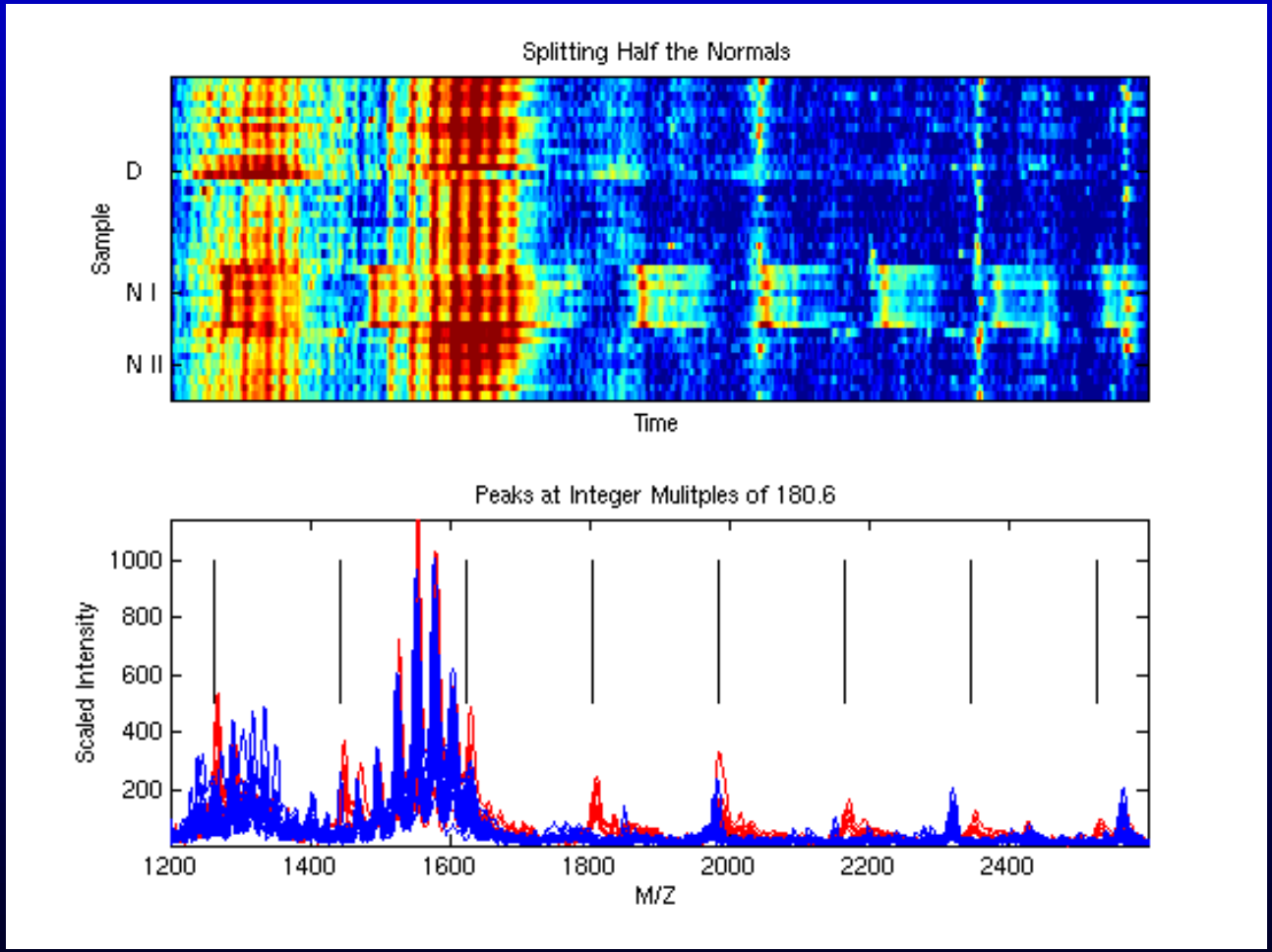
Give the problem a chance to be easy, try some simple clustering.

PCA Splits off Half the Normals



Peaks at Integer Multiples of M/Z 180.6!

This suggests a polymer. No Amino Acid dimers fit.



Cleaning Redux

- Baseline Correction and Normalization
- Inconsistent Fractionation
- Computer Buffers
- Polymers in some Normal Spectra
- Peak Finding (Use Theirs)

Data reduced to 1 spectrum/patient, with 506 peaks per spectrum.

Find the Best Separators

Peaks	MD	P-Value	Wrong	LOOCV
12886	2.547	≤ 0.005	11	11
8840, 12886	5.679	≤ 0.01	5	6
3077, 12886 74263	9.019	≤ 0.01	3	4
5863, 8143 8840, 12886	12.585	≤ 0.01	3	3
4125, 7000 9010, 12886 74263	23.108	≤ 0.01	1	1

There are 9 values that recur frequently, at masses of 3077, 4069, 5825, 6955, 8840, 12886, 17318, 61000, and 74263.

P-values are not from table lookups!

Testing Reality (Significance)

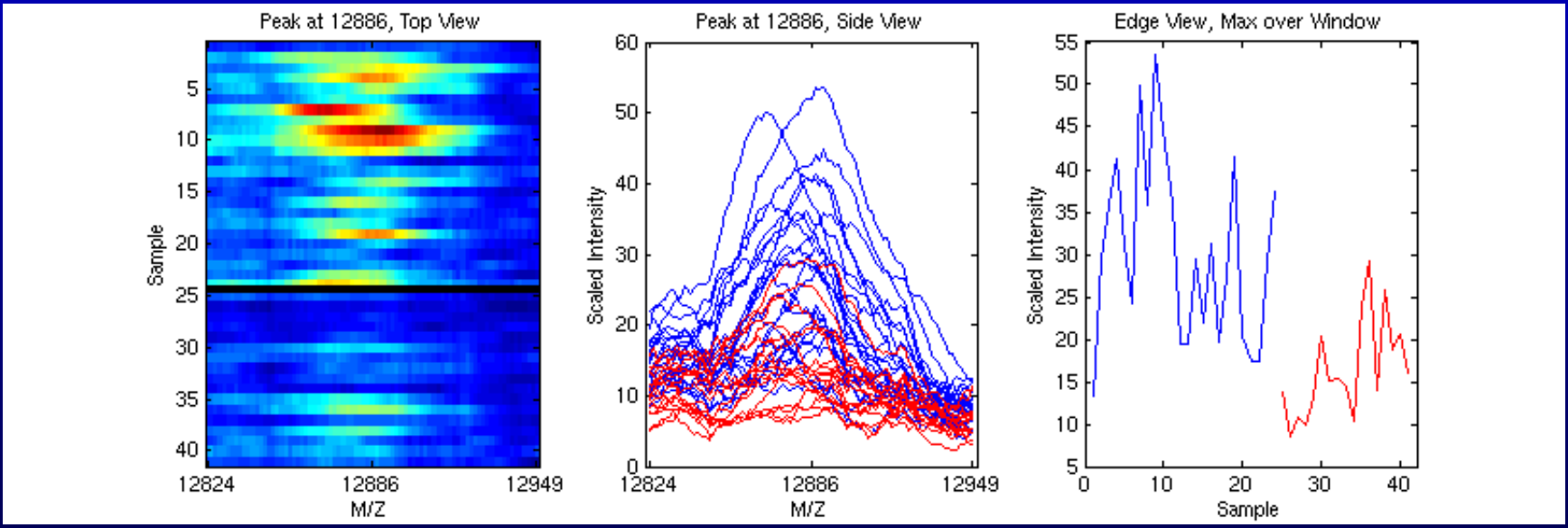
Generate a bunch of “random noise” data matrices, each 41×506 in size.

For each matrix, split the 41 noise “samples” into groups of 24 and 17.

Repeat our search procedure on the random data, and see how well we can separate things.

The Eyeball Test

We applied one last filtering step and actually *looked* at the regions identified. All 9 peaks listed above passed the eye test.



Blue lines = Cancers

Red lines = Controls

Other Stuff

We were the only ones to notice the sinusoidal noise.

Other Stuff

We were the only ones to notice the sinusoidal noise.

and the clock tick.

Other Stuff

We were the only ones to notice the sinusoidal noise.

and the clock tick.

and they were looking at power cables and other stuff.

Other Stuff

We were the only ones to notice the sinusoidal noise.

and the clock tick.

and they were looking at power cables and other stuff.

and they gave us a nice shiny plaque!

The Deluge at MDA

Brain Cancer

Bladder Cancer

Leukemia

Pancreatic Cancer

Breast Cancer

Several show real structure, several show processing effects.

“If you’re not working on a proteomics project, you will be soon!”
Kevin Coombes to Bioinf section at MDA

The Punchlines

- There is no magic bullet here. (Sorry.)
- Data preprocessing is extremely important with this type of data, and there is still much room for improvement.
- Use Simple Tests and Pictures
- Insist on Good Experimental Design
- There is structure in this data and it can be found!

Our Own Reports

On the *Lancet* data:

Baggerly, Morris and Coombes (2004), *Bioinformatics*, **20(5)**:777-785.

On the Proteomics Data Mining Conference Data:

Baggerly, Morris, Wang, Gold, Xiao and Coombes (2003), *Proteomics*, **3(9)**:1677-1682.

More methodology:

Coombes et al (2003), *Clinical Chemistry*, **49(10)**:1615-1623.

pdf preprints are available.

Partners in Crime

Kevin Coombes, Jeff Morris

———-

Jing Wang, David Gold, Lian-Chun Xiao

———-

Ryuji Kobayashi, David Hawke, John Koomen

Index