The Analysis of Proteomic Spectra from Serum Samples

Keith Baggerly
Section of Bioinformatics
Department of Biostatistics
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

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
Example 1: Feb 16 ’02 Lancet

### 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://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.
Mass Accuracy is Poor?

A tale of 5 masses...

<table>
<thead>
<tr>
<th>Feb ’02</th>
<th>Apr ’02</th>
<th>Jun ’02</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>DS2</td>
<td>DS3</td>
</tr>
<tr>
<td>-7.86E-05</td>
<td>-7.86E-05</td>
<td>-7.86E-05</td>
</tr>
<tr>
<td>2.18E-07</td>
<td>2.18E-07</td>
<td>2.18E-07</td>
</tr>
<tr>
<td>9.60E-05</td>
<td>9.60E-05</td>
<td>9.60E-05</td>
</tr>
<tr>
<td>0.000366014</td>
<td>0.000366014</td>
<td>0.000366014</td>
</tr>
<tr>
<td>0.000810195</td>
<td>0.000810195</td>
<td>0.000810195</td>
</tr>
</tbody>
</table>
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

\[
\frac{m/z}{U} = a(t - t_0)^2 + b, \quad U = 20K, t = (0, 1, \ldots) \times 0.004
\]

Fitting it here

\[
a = 0.2721697 \times 10^{-3}, \quad b = 0, \quad t_0 = 0.0038
\]
What is the Calibration Equation?

The Ciphergen equation

\[ \frac{m/z}{U} = a(t - t_0)^2 + b, \quad U = 20K, \quad t = (0, 1, \ldots) \times 0.004 \]

Fitting it here

\[ a = 0.2721697 \times 10^{-3}, \quad b = 0, \quad t_0 = 0.0038 \]

These are the default settings that ship with the software!
Example 2: 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.
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

Disease Sample 19

Disease Sample 21
Unfractionating the Data

All samples; Simple base line correction, summed

15000
10000
5000
0
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2 x 10^5

Hemoglobin
Albumin
Antitrypsin
Transferrin
Immunoglobulin (Probably IgG)
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

<table>
<thead>
<tr>
<th>Peaks</th>
<th>MD</th>
<th>P-Value</th>
<th>Wrong</th>
<th>LOOCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>12886</td>
<td>2.547</td>
<td>$\leq 0.005$</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>8840, 12886</td>
<td>5.679</td>
<td>$\leq 0.01$</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>3077, 12886 74263</td>
<td>9.019</td>
<td>$\leq 0.01$</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5863, 8143 8840, 12886</td>
<td>12.585</td>
<td>$\leq 0.01$</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4125, 7000 9010, 12886 74263</td>
<td>23.108</td>
<td>$\leq 0.01$</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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 $41 \times 506$ 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
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.

- Dimension reduction is critical; both to avoid spurious structure and to focus our attention on peaks.

- There is structure in this data (some peaks have been confirmed, and their writeup is in progress) and it can be found!
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 we also won the competition...
Other Stuff

We were the only ones to notice the sinusoidal noise.

and the clock tick.

and we also won the competition...

and they were fixing the fractionation machine, and looking at the other stuff.
So What is the Peak?

They said that one was really clear, so we asked.
So What is the Peak?

They said that one was really clear, so we asked.

They wouldn’t tell us.
So What is the Peak?

They said that one was really clear, so we asked.

They wouldn’t tell us.

Like that’s going to stop us...
So What is the Peak?

They said that one was really clear, so we asked.

They wouldn’t tell us.

Like that’s going to stop us...

Our guess was that they were talking about the peak at 12886. Turns out we were wrong, but the following method should still work.
Go to us.expasy.org
Go to TagIdent and Query

Scan in Swiss-Prot can be restricted by the use of a keyword. This should be a keyword in the Swiss-Prot KW lines. If in doubt, consult the Swiss-Prot list of keywords.

For example, entering the following values:

- **pI:** 5.4  
- **Mw:** 35000  
- **pI range:** 0.25  
- **Mw range:** 20%  
- **OS or CC:** *planta*  
- **Keyword:** *chloroplast*

will return all Swiss-Prot entries for *plant chloroplast* proteins, with \(5.15 \leq pI \leq 5.65\) and \(28000 \leq Mw \leq 42000\).

When searching in Swiss-Prot, TagIdent removes signal sequences and/or pre-peptides (as documented in the Swiss-Prot feature table (FT lines)) before computing pI and Mw for each of the resulting chains.

The annotation in TrEMBL is done automatically; it is incomplete and not always correct. Thus information on TrEMBL FT lines is not used to process TrEMBL proteins into mature chains or peptides (i.e., pI and Mw are always computed for the whole sequence), and the use of a keyword is not allowed for searches in TrEMBL.
Check the Report

Results from TagIdent

The search in SWISS-PROT has been launched with the following values:
- pI = 7
- Mw = 4444
- delta-pI = 10.00
- delta-Mw = 44
- OS or OC = HUMAN
- KW keyword = ALL

Scan done on 28-Avr-2003.
Swiss-Prot Release 41.5 of 23-Apr-2003: 125236 entries

<table>
<thead>
<tr>
<th>Number of proteins found in the specified pI/Mw ranges</th>
<th>Swiss-Prot</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Scan in SWISS-PROT database (125236 entries)

3 proteins found in the specified pI/Mw ranges

- A2HS_HUMAN (P27665)
  - Alpha-2-HS-glycoprotein precursor (Fetuin-A) (Alpha-2-Z-globulin) (Fa-alpha-2-glycoprotein) (PRO2743)
  - pI: 10.83, MW: 4400.99 CONNECTING PEPTIDE

- CART_HUMAN (Q19598)
  - Cocaine- and amphetamine-regulated transcript protein precursor [Contains: CART(1-39); CART(32-39)]
  - pI: 4.49, MW: 4423.94 CART(1-39)

- UCN2_HUMAN (O96593)
  - Urocortin II precursor (Ucn II) (Stresscopin-related peptide) (Urocortin-related peptide)
  - pI: 10.15, MW: 4451.26 UROCORTIN II
Our Own Reports

On the *Lancet* data: Baggerly, Morris and Coombes (2003), accepted by *Bioinformatics* pending revisions.

On the Proteomics Data Mining Conference Data:


pdf preprints are available.
The Deluge

Bladder Cancer

Pancreatic Cancer

Leukemia

Colorectal Cancer

Brain 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, 3/25/03
Partners in Crime

Kevin Coombes
Jeff Morris
Jing Wang
David Gold
Lian-Chun Xiao

—
Ryuji Kobayashi
David Hawke
John Koomen