

Are the NCI/FDA Ovarian Proteomic Data Biased? A Reply to “Producers and Consumers”

Keith A. Baggerly*, Kevin R. Coombes, and Jeffrey S. Morris

Department of Biostatistics

U.T. M.D. Anderson Cancer Center

Houston, TX 77030-4009

Running Head: Bias or Biology?

* To whom correspondence should be addressed:

Keith Baggerly

Department of Biostatistics and Applied Mathematics

1515 Holcombe Blvd, Unit 447

Houston, TX 77030-4009

Phone: (713) 563-4290

Fax: (713) 563-4243

Email: kabagg@mdanderson.org

Abstract

Proteomic patterns derived from mass spectrometry have recently been put forth as potential biomarkers for the early diagnosis of cancer. This approach has generated much excitement, particularly as initial results reported by the NCI/FDA clinical proteomics program on SELDI profiling of serum suggested that near perfect sensitivity and specificity could be achieved in diagnosing ovarian cancer. However, more recent reports have suggested that much of the observed structure could be due to the presence of experimental bias. The NCI/FDA group has issued a rebuttal, “Producers and Consumers”, listing several objections to the findings of bias. In this paper, we address each of these objections in turn. While we are able to find structure separating cancers and controls, we continue to find evidence of substantial experimental bias. In the presence of bias, the mere presence of structure does not constitute proof that these spectra can be used for clinically meaningful tasks such as the diagnosis of cancer.

Background

Proteomic patterns derived from mass spectrometry have recently been put forth as potential biomarkers for the early diagnosis of cancer. Most of the attention has focused on the variant of mass spectrometry known as SELDI-TOF (surface-enhanced laser desorption and ionization time-of-flight) applied to samples derived from easily available biological fluids such as serum or urine. This approach has generated much excitement, particularly in light of results reported by the NCI/FDA clinical proteomics program. These results, initially reported in *The Lancet* (Petricoin et al. 2002), suggested that near perfect sensitivity and specificity could be achieved in diagnosing ovarian cancer. In addition to publishing these initial results, the NCI/FDA group have also made the raw spectra they used available on their web site, <http://ncifdaproteomics.com>. The data from the initial study were soon followed by data from two further SELDI serum studies on ovarian cancer, and most recently, with more high-resolution data derived from a different type of mass spectrometry (Qstar-TOF). In all cases, the posted results match or ex-

ceed those from the initial study. These latter datasets have now served as the basis for further papers showing various ways in which ever better separation between cancers and controls can be achieved (e.g., Alexe et al. 2004, Zhu et al. 2003, Conrads et al. 2003,2004).

Recently, however, two groups (Sorace and Zhan 2003, Baggerly et al. 2004a) have independently noted that much of the structure present may be due to experimental bias. If this interpretation is correct, then structure associated with bias could effectively obscure any meaningful biological information contained in the spectra. In the presence of confounding bias, the NCI/FDA ovarian spectra cannot be accepted as proof that proteomic profiling can reliably be used for cancer identification.

In response, the NCI/FDA group has issued a rebuttal (Petricoin et al. 2004) listing several objections to the findings of bias. The rebuttal notes that these findings “highlight the dangerous potential for error propagation that may arise if a disconnect is allowed to exist between the data producers and the data consumers”. The authors suggest that in order to “prevent the dissemination of inaccuracies and speculative conclusions, we believe that the producers of genomic and proteomic data should be intercalated more fully into the publication process, particularly when the focus of the publication is the analysis of data that the submitting authors have not generated”. This rebuttal has appeared in print as a commentary to the article of Sorace and Zhan (2003), and we refer to it in this article as “Producers and Consumers”.

Our goal is to address each point made in “Producers and Consumers” that relates to issues raised in Baggerly et al. (2004a). We do not dispute that one can mathematically analyze these spectra and find algorithms which differentiate cancer and control spectra. Rather, we contend that the differences between the cancer and control spectra may arise from factors that are not biologically relevant.

To clarify the notation, we note that there are three SELDI ovarian data sets under discussion:

- DS1: The initial data from the *Lancet* article,
- DS2: A second set of spectra derived from the same biological samples, but run on a

different chip type, and

- DS3: A third set of spectra derived from new biological samples but run on the same chip type as DS2.

All of the data are available from <http://www.ncifdaproteomics.com>.

The Objections and Our Responses

We will now try to address the specific objections identified in “Producers and Consumers”. The eight objections presented in “Producers and Consumers” that relate to points made by Baggerly et al. (2004a) are itemized below. After each, we respond with emphasis on our main contention: that the structure in these data are just as likely to reflect experimental bias as they are to reflect meaningful biological patterns of protein expression.

The first four objections relate to discussions about findings in the low m/z range of the proteomic spectra. Both Sorace and Zhan (2003) and Baggerly et al. (2004a) noted that it was possible to perfectly separate ovarian cancer patients from healthy women in DS3 using the intensities at just two m/z values: 2.79 and 245.2. As both of these values are in regions of the spectra widely thought to be unstable, and the first value appears too early to have any plausible biological explanation, experimental bias was suggested. The producers raise four objections to this finding.

1. Intensities at M/Z values below 1084 should not be trusted.

In “Producers and Consumers”, the authors state that their own analyses did not go down as far as m/z 2.79; values this low are not relevant. These numbers should not be trusted because they occur at m/z values lower than their lowest calibrant m/z value, which is 1084. M/Z values below this “should not be used for pattern classification”.

In response, we note that the m/z values identified in the initial *Lancet* paper and on the NCI/FDA website as separating patterns are as follows:

DS1: **534.82, 989.15**, 2111.71, 2251.18, 2465.02

DS2: **467.18, 500.85, 502.10, 664.92**, 12354.27

DS3: **435.46, 465.57**, 2760.67, 3497.55, 14051.98, 19643.41

Of these, 8/17 are below their lowest calibrant: 2/5 in DS1, 4/5 in DS2, and 2/7 in DS3. Further, Baggerly et al. (2004a) noted that the two lowest m/z values in DS3 are the most important for separation of cancers and controls.

In more recent papers, Alexe et al. (2004) exploit the low m/z region. While the m/z values used do not go below 200 because they were excluded by fiat, most of the structure of interest was found to exist below m/z 1000. Indeed, when m/z values below a threshold value t were excluded from consideration, “deterioration in the quality of the .. models becomes apparent for values of t above 500 or 1000”. Further, Zhu et al. (2003), cited in “Producers and Consumers” to refute some of the claims of Baggerly et al. (2004a), find most of their structure in the low m/z range (10 of the 18 m/z values they report are below 500).

We respectfully suggest that by the producers’ own standards, their own and many other analyses of DS1, DS2, and DS3 are questionable.

2. The randomization should preclude bias.

Despite the contention that values below the lowest calibration point are not reliable, “Producers and Consumers” maintains that the values at 2.79 cannot be due to bias, because the producers were blinded as to status and the samples were randomized with “cases and controls run intermingled on the same chips”.

First, we note that an identical comment about randomly commingling cases and controls was made in the *Lancet* paper with regard to DS1. One of the findings of Baggerly et al. (2004a) was that a subset of these samples had clearly not been randomized. This finding was not addressed in “Producers and Consumers”.

Second, a closer investigation of the initial file names of the DS3 spectra suggests that wholly disjoint sets of chips were used to produce the cancer and control spectra. Further, the type of

bias suggested by this result would easily account for the presence of structure in “noise” regions of the spectra. These results are described in more detail in Baggerly et al. (2004b).

3. Structure at 2.79 must be due to the low mass proteome.

Having discounted bias, “Producers and Consumers” concludes that the observed structure must be due to real biology associated with the low-mass proteome, which is currently not well understood.

While we concede that the low mass proteome has yet to be fully explored, we note that this explanation still seems odd with respect to the peak at 2.79. The signal is very weak, and there are no other peaks nearby in the spectrum. Even a metabolite should have a mass on the order of a single amino acid, and in this mass range there should be other artifacts present. However, we suggest a simpler explanation: as the assumption that the low m/z findings must be biologically relevant rests on the assumption that the data were randomized (addressed above), rejection of the prior assumption negates the latter.

4. We are confusing “noise” with “bias”.

In referring to our assertion that structure can be found in “noise”, “Producers and Consumers” comments that we are confusing “noise” with “bias”.

We see this as a semantic issue. We meant that we were able to find structure that separated ovarian cancer patients from healthy women in regions of the spectra that we believe to be unaffected by biology. These regions can be affected by electronic signals associated with machine operation, and we would call this electronic noise. By whatever name, the structure in question should not be present in the absence of experimental bias.

As the producers state, measurements in the low m/z range are not reliable; therefore, in our view, consistent structure in this range strongly supports the contention that these datasets are subject to significant experimental bias. Ultimately, in the presence of such bias, we cannot assume that any reported structure reflects a biologically relevant, reproducible pattern of

measured protein expression differences between cancer and control spectra.

The producers also raise four other objections that encompass DS1 and DS2, as well.

5. Extension of the presence of bias at 2.79 to the rest of the dataset or to other datasets is “judgmentally biased”.

In discussing the structure found in DS3, “Producers and Consumers” notes that Sorace and Zhan’s (2003) interpretation of bias at m/z 2.79 is extended to “the entire SELDI-TOF MS data set, including many other datasets that they did not in fact analyze”. These “broad conclusions are judgmentally biased and scientifically unfounded”.

We fail to see how extending the presumption of bias to the rest of the data set is judgmentally biased or even avoidable. Certainly with DS3, if structure at 2.79 shows that the samples were processed differently in some way, that difference should be expected to persist for all m/z values.

As to the latter part of the assertion regarding the other datasets, all three datasets are surveyed in Baggerly et al. (2004a), and the assertions of bias there are based on the analysis of all three. As our calculations are publicly available, we invite the scientific community to reproduce them in sufficient detail to be satisfied that they are not “scientifically unfounded”.

6. The SOP the producers follow with respect to calibration means that the data are correctly calibrated.

“Producers and Consumers” notes that while Baggerly et al. (2004a) “wondered .. about our calibration method, we adhere to strict SOPs whereby any TOF MS is calibrated at the beginning of every analysis”.

We do not dispute that a strict SOP was followed for calibration. However, we believe that the posted values are wrong. The posted spectra show m/z values corresponding to the default calibration that ships with the SELDI software. To us, this mistake suggests an error in file

export rather than a failure to attempt calibration, but an error, nonetheless.

As further evidence that a calibration problem exists, we note that in Conrads et al. (2003), where the NCI/FDA Qstar spectra were first described, Figure 4 of that paper shows Qstar and SELDI spectra derived from the same SELDI chip. The chips used for the Qstar experiment were of the same type as those used in DS2 and DS3. In the Conrads et al. (2003) figure, the maxima of the SELDI and Qstar spectra are roughly aligned, and we believe that the alignment shown there is correct. However, if we superimpose the location of the biggest SELDI peak from the Conrads et al. (2003) picture on a plot of the average cancer spectra from DS2 and DS3, we note that the posted maxima are hundreds of units away. This is shown in Figure 1a of this response. If we use the marked peaks in the Conrads et al. (2003) SELDI figure to supply an external calibration for DS2 and DS3, the peak locations are aligned even at m/z values not used in the calibration, as shown in Figure 1b.

The effect of using the default calibration is not slight. The m/z values for DS3 are off by about 2.5% in the vicinity of the biggest peak, and the m/z values for DS2 are off by about 3.9%. As the SELDI results are nominally accurate to within a few tenths of a percent, miscalibration this severe can actively mislead investigators performing database searches based on the reported m/z values.

7. One group can find transcendent structure, and another cannot.

“Producers and Consumers” notes that while Baggerly et al. (2004a) noted “the inability of features to transcend separate data sets”, a second article by Zhu et al. (2003) “concluded that transcendent features could be found”. The producers cite the latter publication as evidence that DS2 and DS3 contain reproducible biological structure.

Baggerly et al. (2004a) assumed that the errors in calibration described above should preclude the persistence of biological structure across datasets. We verified that the patterns supplied on the NCI/FDA web site did not represent reproducible structure across DS2 and DS3. But, given the offset, we did not conduct an exhaustive search. On the other hand, Zhu et al. (2003) noted

that when the 18 m/z values that were chosen to separate cancers from controls in DS2 were used in DS3, perfect separation was observed even though DS3 had been treated as a blinded test set.

This apparent contrast can in fact be easily resolved. The exact approach is detailed in Baggerly et al. (2004b), but the key point is simply that DS3 is so easy to correctly classify that near-perfect separation results were obtained using 18 m/z values *chosen completely at random*. Thus, biology is not required to explain the separation observed.

Further, when the patterns of protein expression at the 18 m/z values supplied are checked in both DS2 and DS3, the directionality of expression changes for 13 of the 18: if expression is higher in cancers in DS2, it is higher in controls in DS3. This suggests that a biological explanation is not only unnecessary to explain the findings in Zhu et al. (2003), it is actively precluded.

8. The focus of the objections has been the SELDI data, not the Qstar data.

The focus of the analysis in both Sorace and Zhan (2003) and Baggerly et al. (2004a) has been ovarian SELDI data, but the producers feel that the more recent high-resolution Qstar data is the current state of the art, and they suggest that more attention should be paid to the better data.

This observed focus is not due to a lack of interest in the Qstar data, but rather to the time lag associated with publication. However, as the Qstar ovarian spectra are derived from SELDI chips, biases that affect these chips can affect the Qstar data as well, so understanding how experimental design issues can affect the SELDI results is still relevant. We note that the file names of the DS3 SELDI spectra are identical to the file names of the Qstar spectra, which suggests to us that the DS3 chips were used in the Qstar experiment. If this is in fact the case, biases affecting the DS3 chips are even more directly relevant.

Further, while the Qstar data are of higher resolution, they also show signs of experimental bias. In Figure 2, we show a heat map of all of the Qstar spectra we have available, sorted by the file names supplied, in the vicinity of m/z 8602. This value is identified in Conrads et al. (2004) as being of use for distinguishing ovarian cancer patients from healthy controls, and a higher level of expression is observed for the cancer patient spectra. However, there is also a visible peak roughly 80 units lower in which expression is high for healthy women but just for half of the cancer patients. We believe that this is due to most of the control spectra being run before the cancer spectra, as the producers have noted that the ABI Qstar detector failed shortly after the study was complete. This interpretation is supported by the QA/QC results reported in Conrads et al. (2004). The initial Qstar run involved 248 spectra before the application of a QA/QC filter: 152 cancer and 96 control. After a QA/QC filter was applied to remove “lesser quality” spectra, 216 spectra remain: 121 cancer and 95 control. Of the 32 spectra identified as “lesser quality”, 31 are cancers. As noted in Conrads et al. (2004), “These mass spectra were all generated at the end of the experimental run”. If we assume that the cancers and controls have been randomly commingled so that exterior process degradation should not preferentially affect one group more than the other, the one-sided p-value associated with this result is $9.9810e-07$. This result is much easier to explain if we assume that cancers and controls were run on different sets of chips (as we believe to be the case with DS3) so that it is possible to run most of the controls before running the cancers.

These observations leave us concerned that some of the structure found in the Qstar data could also be confounded with nonbiological factors.

Concluding Remarks

The producers have claimed that the consumers are mistaken as to the presence of bias. We respectfully disagree. We are willing to revise our beliefs when features in the data that refute our claims are presented. Until that time, we must repeat our initial position: No one disputes that

structure can be found in all three datasets. However, the structure appears to be associated with strong evidence of experimental bias. As such, the demonstration of structure does not constitute proof that these spectra can be used for clinically meaningful tasks such as the diagnosis of cancer.

Finally, we address the producers' suggestion that all analyses performed by consumers should directly involve the producers so that all of the relevant information will be available. The producers are definitely to be commended for making their data available. However, while communication is a good thing, we believe that rather than relying on email and telephone exchanges, it would be far better to define standards for the publication of proteomic data so that all of the relevant information is available at the time of publication. The Microarray Gene Expression Data Society (MGED) has developed such a standard for microarray data: the Minimum Information About a Microarray Experiment (MIAME; Brazma et al. 2001, Spellman et al. 2002). Exactly what should be supplied in the proteomic equivalent is, we believe, a productive area for debate. Indeed, this was also the consensus of the participants at the Early Detection Research Network (EDRN) meeting on the analysis of SELDI/MALDI data (Seattle, 2004). In the interim, we note with respect to SELDI that current versions of the Ciphergen software support exporting the data in an XML format that could serve as a template for an eventual standard.

References

Alexe, Gabriela, Alexe, Sorin, Liotta, Lance A., Petricoin, Emmanuel, Reiss, Michael, and Hammer, Peter L. (2004). "Ovarian cancer detection by logical analysis of proteomic data". *Proteomics*, **4**:766-783.

Baggerly, Keith A., Morris, Jeffrey S., and Coombes, Kevin R. (2004a). "Reproducibility of SELDI-TOF Protein Patterns in Serum: Comparing Data Sets from Different Experiments". *Bioinformatics*, **20**:777-785.

Baggerly, Keith A., Morris, Jeffrey S., Edmonson, Sarah, and Coombes, Kevin R. (2004b).

“Signal in Noise: Can Experimental Bias Explain Some Results of Serum Proteomics Tests for Ovarian Cancer?”, Technical Report.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. (2001). “Minimum information about a microarray experiment (MIAME)-toward standards for microarray data”. *Nat Genet.*, **29**:365-71.

Conrads, Thomas P., Zhou, Ming, Petricoin III, Emmanuel F., Liotta, Lance, and Veenstra, Timothy D. (2003). “Cancer diagnosis using proteomic patterns”, *Expert Rev. Mol. Diagn.*, **3(4)**:411-420.

Conrads, Thomas P., Fusaro, Vincent A., Ross, Sally, Johann, Don, Rajapakse, Vinodh, Hitt, Ben A., Steinberg, Seth M., Kohn, Elise C., Fishman, David A., Whitely, Gordon, Barrett, J. Carl, Liotta, Lance A., Petricoin III, Emanuel F., and Veenstra, Timothy D. (2004). “High-resolution serum proteomic features for ovarian cancer detection”, *Endocrine-Related Cancer*, to appear June 2004.

Petricoin III, Emanuel F., Ardekani, Ali M., Hitt, Ben A., Levine, Peter J., Fusaro, Vincent A., Steinberg, Seth M., Mills, Gordon B., Simone, Charles, Fishman, David A., Kohn, Elise C., and Liotta, Lance A. (2002). “Use of Proteomic Patterns in Serum to Identify Ovarian Cancer”. *The Lancet*, **359**:572-577.

Petricoin III, Emanuel F., Fishman, David A., Conrads, Thomas P., Veenstra, Timothy D., and Liotta, Lance A. (2004). “Proteomic Pattern Diagnostics: Producers and Consumers in the Era of Correlative Science”. Comment on Sorace and Zhan. *BMC Bioinformatics*, <http://www.biomedcentral.com/1471-2105/4/24/comments>.

Sorace, James, and Zhan, Min (2003). “A data review and re-assessment of ovarian cancer serum proteomic profiling”. *BMC Bioinformatics*, 2003 Jun 09; 4:24. <http://www.biomedcentral.com/1471-2105/4/24>.

Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G,

Ball C, Lepage M, Swiatek M, Marks W, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert CJ Jr, Brazma A. (2002). "Design and implementation of microarray gene expression markup language (MAGE-ML)". *Genome Biol.*, **3**:RESEARCH0046.

Zhu, Wei, Wang, Xuena, Ma, Yeming, Rao, Manlong, Glimm, James, and Kovach, John S. (2003). "Detection of Cancer-Specific Markers amid Massive Mass Spectral Data". *PNAS*, **100**:14666-14671.

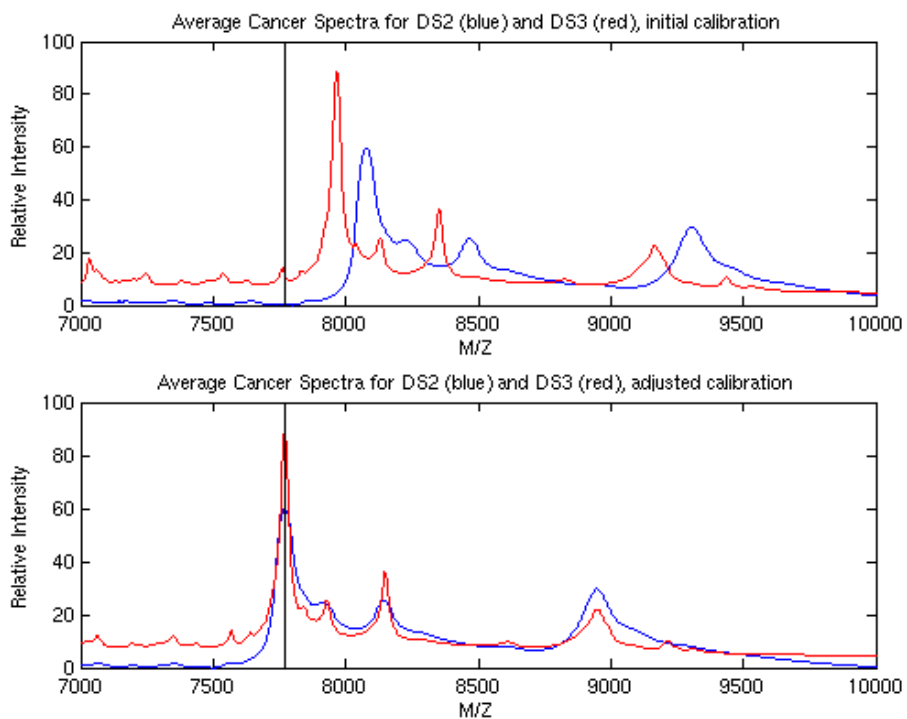


Figure 1: (a) The average cancer spectra from DS2 and DS3, with the location of the maximum peak from Conrads et al. (2003) shown. The posted spectra appear offset. (b) The corresponding average spectra after using the labeled peaks in the Conrads et al. (2003) figure to recalibrate the spectra. Agreement between DS2 and DS3 is now good throughout the region bracketed by calibrants.

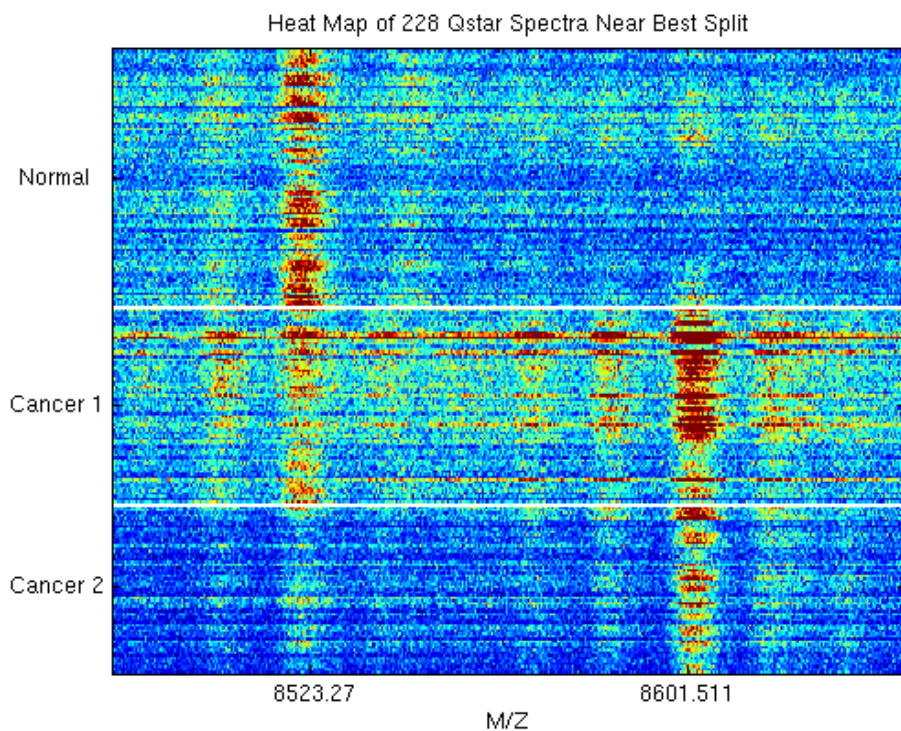


Figure 2: A heat map of the Qstar spectra we have available, sorted by file name, in the vicinity of m/z 8602. This m/z value is identified on the NCI/FDA website as useful for separating healthy women from ovarian cancer patients, and this separation is visible. However, roughly 80 Da below, there is a peak that serves to separate the healthy women and the first half of the ovarian cancer patients from the second half of the ovarian cancer patients.