Cautions about Reproducibility in Mass Spectrometry Patterns: Joint Analysis of Several Proteomic Data Sets

Keith A. Baggerly, **Jeffrey S. Morris**, and **Kevin R. Coombes**
Department of Biostatistics
U.T. M.D. Anderson Cancer Center
1515 Holcombe Blvd, Box 447
Houston, TX 77030-4009
kabagg@mdanderson.org

February 26, 2003

Abstract

In a recent paper, Petricoin and colleagues [7] asserted that they could differentiate ovarian cancer serum samples from normal serum samples on the basis of mass-spectrometry proteomic profiles. Because the biological problem is important, they have now repeated the general experiment three times, with variations that include new types of ProteinChip arrays and new experimental samples. In reexamining the data from these experiments, we have encountered a series of problems that call into question the reproducibility of these results. Specifically, in one experiment there is evidence of a major shift in protocol mid-experiment. In another, structure in the noise regions of the spectra allows us to distinguish normal from cancer. Sets of features found to discriminate well in one experiment do not generalize to other experiments. Taken together, these and other concerns suggest that much of the structure being uncovered could be due to artifacts of sample processing, not to the underlying biology of cancer. We provide some guidelines for design and analysis in experiments like these to better ensure reproducible, biologically meaningfully results.

1 Introduction

Ovarian cancer is a frequently deadly disease, and the degree of morbidity is strongly linked to our inability to detect the tumors at an early stage. Neither X-rays nor MRIs are able to differentiate between cancers and benign cysts, surgical verification of cancer status is dangerous, and gene product assays (such as CA125) are not sensitive or specific enough. A simple, easily applied diagnostic test with high sensitivity and specificity would be of great utility.

In a recent paper in *The Lancet*, Petricoin and colleagues [7] reported finding patterns in mass-spectrometry (SELDI-TOF, Ciphergen) proteomic data that can distinguish between serum samples from normal women and serum samples from women with ovarian cancer, even when the cancers are at early stages. In their initial study, they started with 100 cancer spectra, 100 normal spectra, and 16 "benign disease" spectra. The cancer and normal sets were randomly split, and 50 cancer and 50 normal spectra were used to train a classification algorithm. The resulting model was used to classify the remaining spectra, and it correctly classified 50/50 of the cancers and 47/50 of the normals. It called the 16/16 benign disease "other" than normal or cancer[1]. These results are impressive, and have received a good deal of attention.

[1]Some numbers in the initial paper indicate 46/49 normals, and 16/17 benign disease; one benign disease sample was later determined to be normal.

The initial experiment and two related experiments by the same group of investigators have some very positive features. They collected enough samples to find real structure in the data. Splitting the data into training and validation sets allows for internal validation of the structure found, protecting somewhat against the tracking of random noise. Finally, all the data has been made publicly available (on the website http://clinicalproteomics.steem.com). Other studies using this technology are beginning to emerge but the papers do not mention easy access to the raw data [6, 8].

Our own analyses of these data have confirmed (in one case) their findings of the existence of structural features that strongly separate the normal from cancer samples, with a degree of separation that is well beyond what would be expected by random noise. Having found structure, however, a simple question arises: Is this structure is due to inherent biological differences associated with cancer, or due to artifacts associated with the technology? Changes that could introduce such artifacts include differential handling and/or processing of the samples, changes in the type of ProteinChip array, mechanical adjustments to the mass spectrometer itself, or a shift to a different machine or lab, among others. Either biology or artifact could account for strong systematic differences between the groups of spectra. The answer to the question of which is the driving force is crucial, since separation due to artifacts cannot be expected to generalize to future groups of patients.

We have conducted our own analysis of the three ovarian cancer data sets posted on their web site. Our findings suggest that while differences are being found within individual experiments, these differences do not generalize across experiments. This lack of generalizability indicates the need for careful experimental design and varying of several experimental conditions when conducting such studies.

Below, we introduce the data sets in more detail, and describe our reanalysis of the ovarian data. We then briefly summarize our findings.

2 Background

In order to discuss the analysis, we need to briefly review the nature of the data available on the website, the processing applied to the data, the function used for assessing the goodness of a feature set, and the method used for choosing feature sets.

The Data Sets. There are 3 data sets of ovarian mass spectra currently available on the web site. The first ovarian cancer data set, which was described in the initial paper, consists of 216 spectra, divided into 5 files: Training Cancer, Training Normal, Test Cancer, Test Normal, and Benign Disease. These spectra were obtained using the Ciphergen H4 ProteinChip array (since discontinued). These spectra have been baseline subtracted. The second ovarian cancer data set uses the same samples as above, run on the Ciphergen WCX2 ProteinChip array. Again, the spectra have been baseline subtracted. The third ovarian cancer data set contains 91 normal samples and 162 cancer samples. These samples were prepared robotically. These spectra have not been baseline subtracted. Each spectrum consists of a list of 15,154 m/z values and associated intensities. The m/z values are common across all spectra.

The Data Processing. In all cases, we believe their analysis was performed on the data sets before baseline subtraction (more details are given in the reanalysis and the discussion). Before comparison, all of the spectra in an experiment were normalized to have the same $[0, 1]$ intensity range as follows:

$$NV = \frac{V - \min(V)}{\max(V) - \min(V)}.$$

The Fitness Function. The "fitness" of a particular feature set containing N features is assessed using the associated scaled intensities to define locations in the N-dimensional unit cube as follows. Start with sample 1. If the Euclidean distance between sample 2 and sample 1 is less than $0.1 * \sqrt{N}$, put the samples into a common cluster and use the mean of the samples as the center center. If sample 2 is farther away, it starts a new cluster. Repeat the allocation of samples

as above until all samples are allocated to clusters. After all samples have been clustered, each cluster is labeled "cancer" or "normal" by majority vote, and the fitness is defined in terms of the number of samples correctly classified.

The Selection of Feature Sets. Feature sets are chosen for analysis using a genetic algorithm [4, 5]. Each run of the genetic algorithm starts with 1500 logical chromosomes (feature sets) of a set size ranging from 5 to 20 index values. The fitness of each feature set is assessed as above. New populations are then produced by preferentially combining pieces of the "most fit" members of the current generation. The process then evolves for 250 generations, with a mutation rate of 0.02% and random crossover locations. All 15,154 distinct features in a spectrum were available for inclusion in a feature set. There is no initial peak finding step.

3 Reanalysis

Baseline Correction Prevents Reproduction of Results. We began by looking at the Euclidean distance matrix from the first data set, using the intensities at the 5 reported m/z values and processing the spectra according to the normalization method given above. (Fig. 1). Two problems were immediately apparent. First, the distances between cancer samples and normal samples were not different from the distances between two cancer samples or between two normal samples. Ideally, we would like to see a "plaid" pattern, with small distances between samples of the same type and large distances between samples of different types. Such a pattern is visible in Figure 5 a, described in more detail below. Second, there are only 4 pairwise distances greater than $\sqrt{5}/10$, which is the cutoff distance for declaring a new cluster with 5 peaks, and these are all distances from one cancer to another cancer. Thus, the clustering approach described in the original paper will not work as desired as new clusters will effectively never be formed.

Include Figure 1 about here.

The problem lies in the fact that the posted data have been baseline subtracted (Fig. 2). The web page comments on this issue, noting that "this process creates negative intensities", but the situation is more serious. Baseline subtraction does produce negative intensity values (primarily in the low m/z region), but the problem is that this correction is an *irreversible* nonlinear operation. Given only the baseline subtracted values, it is impossible to reconstruct the raw values. This problem prevented us from reproducing their results on the first and second ovarian data sets and on the prostate cancer data set.

Include Figure 2 about here.

Based on our experience, the baselines of different spectra can be highly variable. They change from machine to machine and from day to day on the same machine. In general, the baseline signal is caused mostly by chemical noise from matrix molecules, with some contribution from electronic noise [1, 3]. The matrix noise contribution to the baseline signal is largest in the low m/z region.

Baseline correction also interacts with their chosen method of normalization. Normalizing to the range of the baseline corrected spectra is driven by the noise level in the matrix noise region as opposed to the natural zero intensity level of the machine, and introduces visible offsets that persist for the length of the spectra.

Our inability to reproduce their analysis using the posted data is disturbing. First, it shows that the reported results are not robust enough to withstand baseline subtraction. Second, it suggests that matrix noise in the low m/z region may be driving some of the structure.

Sample Processing Differences Cause Blatant Changes. Just because we cannot reproduce their results in data sets 1 and 2 does not mean that there is no structure to be discerned. The algorithm used in the original study was able to identify the 16 benign disease samples in data set 1 as "other" than normal or cancer. In looking at a "heat map" of all 216 spectra from data set 1, shown in Figure 3 (top), the benign disease spectra at the bottom are clearly distinct. Indeed, the cancer spectra and normal spectra show far greater similarity to each other than to the benign disease. Conversely, if we look at all 216 spectra from data set 2, shown in Figure 3 (bottom), we do not see this obvious separation. Because these are the

Include Figure 3 about here.

same biological samples, run on a different kind of ProteinChip array, this lack of separation is disturbing. If we look at both image maps together, however, we see that the benign disease spectra from data set 1 have profiles that are extremely similar to those of data set 2. This observation suggests that there was a change in protocol before the first set was complete.

One example of a protocol change that could produce results like this is a shift between chip types. Different chip surfaces, by design, bind different sets of proteins. However, the WCX2 chip used in the second experiment is nominally quite similar, in terms of the class of proteins it should bind, to the H4 chip that it replaced. Alternatively, maintenance or replacement of critical portions of the Ciphergen machine itself could cause similar changes that would be reflected in the need to recalibrate the formulas that transform the measured time-of-flight into estimates of the mass-to-charge ratio. Such technological differences can give rise to real differences in the spectra, but these differences are not biologically interesting.

We considered the possibility that an error had been made when the data sets were prepared for posting to the web, and that the benign disease spectra posted as part of the first data set were actually the same spectra posted with the second data set. To test this possibility, we compared the numerical values in the spectra. We found that none of the benign disease spectra in the first data set were numerically identical to any of the benign disease spectra in the second data set.

Data Set 3 is Offset Relative to Data Set 2. In order to see if we could generalize results across experiments, we tried to view data sets 2 and 3 (which used the same chip type) simultaneously. Even though data set 2 was baseline corrected, we hoped to use qualitative features of the spectra to assess similarity. Unfortunately, in attempting to match the indices of the major high m/z peaks for comparison, we found that the spectra from data set 3 were offset by roughly 50 to 60 clock ticks from the spectra in data set 2. (Figure 4). Converting to the m/z scale, an offset of this magnitude corresponds to an imprecision of more than 1%. However, the stated mass accuracy of the SELDI procedure is 0.1%. The observed offset between the data sets calls into question the stability of the procedure. A shift of this magnitude could cause the same protein to be identified differently in the two different experiments, obscuring the biology.

Include Figure 4 about here.

Separating Feature Sets Are Not Reproducible Across Experiments. In order for the results to be generalizable, feature sets found to be useful in one experiment should also be useful in another experiment. Because the chip surface was changed in going from data set 1 to data sets 2 and 3, the results from the first experiment cannot be compared to the other experiments. Because data sets 2 and 3 share a common chip surface, we assumed they should be comparable. The feature set reported for data set 2 contains 5 features. If we compute the Euclidean distance matrix for data set 3 using the intensities at these 5 features, the distance matrix clearly shows that the cancer samples and the normal samples have not been split apart (Fig. 5). The problem is not remedied by including an offset term; the distance plot produced is qualitatively similar to that shown here (data not shown). Testing the validity of the features found by analyzing data set 3 by applying them to data set 2 is more difficult, because of the baseline correction applied to data set 2. Thus, we checked the results one feature at a time. There were 7 features reported for data set 3. We found the single feature at m/z 435.46 to be the most useful in terms of splitting the cancer samples from the normal samples in data set 3. Checking the shapes of the spectra in this local region for both data sets, we found that there was clearly a visible separation between the sample types associated with the slope of a peak in data set 3. However, not only is there no clear separation in data set 2, but the shape in the region is no longer that of a peak but rather that of a valley. Moreover, there is clear evidence that the spectra were locally saturated before baseline subtraction (Fig. 6, flat regions of high intensity). Similar lack of agreement was found for the other features (data not shown). Again, the situation was not fixed by the inclusion of an offset term (offsets of 50, 55 and 60 clock ticks were tried; data not shown).

Include Figure 5 about here.

Include Figure 6 about here.

The Stated Method of Normalization Has Negligible Effects. With data set 3, which had not been baseline corrected, a quick look at the Euclidean distance matrix shows that we are able to reproduce their results; there is clear separation between the cancers and the normals

(Fig. 5 a). However, if we look closely at how the normalization method is affecting the data, we see very little change. Of the 263 spectra in this data set, all but one of the spectra has a maximum recorded intensity of 100, indicating saturation of the signal. The remaining sample has a maximum intensity of 99.7486. For most spectra, the range of saturation extends to about clock tick 2140, corresponding to an m/z value of about 398. We view (at least) all intensities at m/z values below this with suspicion. The minimum spectra intensities are almost all between 3.8 and 3.9, with no values falling outside the interval $[3.75, 3.96]$. Thus, normalizing to the range has very little effect if the data have not been baseline corrected. In light of this, we elected to work with the raw spectra (with no correction at all) in our first analysis.

Simple Tests Reveal Better Lower-Dimensional Separators, and Rank the Features They Supply. We feel that there are some problems with the fitness function and clustering methods used. Specifically: Classification accuracy is the only measure of fitness. No additional weight is given for larger separation, and no penalty is assessed for larger numbers of clusters. Euclidean distance does not adjust well to scale differences at different intensities. Thus, a consistent difference that is smaller in magnitude will be missed. The distance cutoff at $0.1 * \sqrt{N}$ is ad hoc. In jumping immediately to dimension 5 and higher, we miss the chance to find simpler explanations if such exist.

We applied a two-sample t-test to the difference between cancer samples and normal samples for every m/z value in data set 3. The most extreme t-values are huge in magnitude, with the largest in their list occurring at m/z 435.46, where the t-value is 22.3463. Using the intensities at this single feature, we can correctly classify 238 of the 253 samples. We note that, although the separation using this single m/z value is fantastic, this is the same value that failed to separate the spectra in data set 2. Moreover, the m/z value is located in what would normally be treated as the matrix noise range [3]. Looking at the t-values for the 7 chosen features individually suggests that some of them are far more important than others. The (m/z, t-value) pairs are shown below.

m/z	435.46	465.57	2760.67	3497.55	6631.70	14051.98	19643.41
t-value	22.346	-12.534	1.498	5.954	-3.501	6.081	-0.476

This ordering suggests that the first two peaks in their list are the most informative (this is addressed further below).

If we use the t-values to suggest particularly interesting features, we are further led to some values not in their list. The most extreme t-value, -27.0256, occurs at m/z 245.2, and the best single classifier is at m/z 244.9524 (with a t-value of -26.0531), where we misclassify only 5 of the samples. The m/z range in which these most extreme values fall, however, gives us substantial pause; recall that saturation spikes were seen for most of the spectra until about m/z 398.

We Can Achieve Perfect Classification with Noise. If the only measure of fitness is classification accuracy, then the search algorithm will not converge if there exist multiple feature sets that classify the data perfectly. Looking at the individual features where large t-statistics were observed, we considered the separations possible using only pairs of features. We quickly found two distinct pairs where perfect separation was possible using a straight line in Euclidean space. The first pair of m/z values is $(435.46, 465.57)$, which are the first two values in their list of 7. (Fig. 7 a). Note that both masses are less than 500. The second pair for which we found perfect separation was at m/z values of (2.79, 245.2), with t-values of $(-13.89, -27.0256)$ respectively (Fig. 7 b). These two values are clearly in the noise region; the first is even in the range before the machine may be recording stably.

Include Figure 7 about here.

The fact that we can find many perfect classifiers suggests that there may not be a unique best survivor from the genetic algorithm runs; multiple runs should reveal multiple optima. Given this, when reporting the results from a series of such runs it is useful to provide a synthesis of the overall results. A histogram of the frequency with which any given feature occurs in the entire set of runs, for example, would provide such a summary.

The fact that we can achieve perfect classification based on readings entirely in the noise region, however, is evidence of a more severe problem. There is no biological reason for this

difference: This finding is a machine artifact. The presence of this artifact suggests that there was a systemic difference in the way the groups of samples were processed. There should be no pattern in the noise region, and the chance of perfect separation of this many samples is too small for random chance to be a believable option. The existence of structure in a region of the raw spectra where there should be no structure, by itself, is mildly disturbing. When combined with other findings that the systematic differences caused by the technology can be large compared to the biological differences, the presence of this structure "raises the bar" that features elsewhere in the spectra must pass in order to be considered biologically meaningful.

Another Analysis. We performed an independent analysis of data set 3, to see if the features we identified would coincide with theirs. Because peak finding and baseline correction are intertwined, we dealt with them in an iterative fashion. We first found all local maxima in each spectrum, and recorded left and right "downslope extents" for each local maximum. We only included local maxima that were reached by climbing in both directions at least four steps larger than the noise. For these purposes, "noise" was taken to equal one-half of the median absolute value of the first differences of the entire spectrum. The "peaks" were removed from the spectrum and replaced with straight lines. We then fit a local minimum (using a window of 256 clock ticks) and subtracted these values as an initial estimate of the baseline. We then reran the peak finding and trimming steps, subtracted off the baseline, and reintroduced the peaks. After baseline correction, the individual spectra were normalized to have the same total ion current (summed intensities) in a region that excluded the matrix noise. We normalized using intensities above M/Z 1000. (Equivalent results were obtained normalizing to the region above M/Z 2000.) In order to restrict our analysis to definite peaks, we computed a local signal-to-noise ratio for each peak. For these purposes, the local noise was taken to equal the median absolute deviation from the median in a window of 256 clock ticks, and we retained peaks whose signal-to-noise ratio exceeded three. This procedure identified between 224 and 301 peaks per spectrum, with both a mean and a median of 266 peaks per spectrum.

We next aligned peaks across spectra, using the rule that peaks that were separated by less than 0.05% of the mass or by fewer than 3 clock ticks could not be distinguished. This procedure resulted in a list of 1093 peaks with signal-to-noise ratio greater than three in at least one spectrum. Of these peaks, 506 were seen at least 10 times; 240 were seen at least 50 times, and 78 were seen at least 100 times among the 263 spectra in the experiment. We only retained the 506 peaks that had been found at least 10 times for further analysis.

We performed two-sample t-tests to determine how well each individual peak distinguished the cancer samples from the normal samples. We visually inspected all peaks with absolute t-value greater than 10 and identified 15 significant peak regions, with m/z values equal to: 245, 434, 616, 886, 1531, 3201, 3928, 4743, 4782, 4898, 7380, 7760, 8033, 15892, and 16050. Only one of these peaks (m/z 434) also appears on their list.

The largest t-value is associated with the peak at m/z 245. The first three peaks on our list, including this one, are located well below the end of the matrix noise region in a part of the spectrum where saturation commonly occurs. The existence of significant peaks in this region of the spectrum even after baseline correction and normalization is troublesome, since it suggests the possibility that there are important differences in the way the samples were handled or processed. Significantly, none of the peaks that we found generalized to data set 2.

4 Discussion

In an effort to better understand the biological structure behind these results, we reanalyzed the data on the website from both the initial experiment and from two subsequent experiments on ovarian cancer. Unfortunately, instead of clarifying the issue, our analysis uncovered a series of problems suggesting a lack of generalizability.

In light of these findings, it is important to note that, had we been consulted, we would have enthusiastically supported the publication of the first data set. (We hope that we would have

detected the anomalous status of the benign disease samples before publication, but we are not certain of it.) The randomized 4x50 design with validation (50 of each group for training; 50 of each for validation) should have been adequate to detect real features—features that generalize from one data set to another—capable of distinguishing between the two groups of samples. It is only because the original authors have posted multiple data sets repeating the same basic experiment that the difficulty of obtaining reproducible results from this technology can be investigated.

The use of proteomic patterns in spectra to distinguish cancer samples from normal samples is a "black box" approach to the problem. Serum samples enter at one end of the black box; they pass through a complex process of protein extraction, sample preparation, mass spectrometry, and bioinformatic analysis; finally, a diagnosis emerges from the other end of the black box. Reproducibility of the proteomic patterns is critical to the success of this approach. The black box must yield the same results today and tomorrow; in a laboratory in Washington and a laboratory in Houston; on samples from the Mayo Clinic and on samples from the M.D. Anderson Cancer Center. The black box approach must rely on reproducibility because it does not provide an explanation or a mechanism to bolster its diagnosis. Consequently, the findings cannot be verified using independent technologies. By contrast, consider the current state of the art in applying microarrays to the problem of discovering biomarkers. Potential biomarkers are typically validated using other technologies such as Northern blots or RT-PCR.

In order to achieve the level of reproducibility required for a successful black box approach to the diagnosis of cancer, careful attention must be paid to measuring and controlling sources of variation in the procedure. A (very) incomplete list of such sources includes time (since results from a single machine can drift), temperature, humidity, the machine used, and the laboratory in which the experiment is conducted. A more complete list must be established, and experiments must be performed to estimate the magnitude of each effect. A good start in this direction is provided by Cordingley et al. [2]. Whenever possible, standard protocols should be drawn up to minimize the effect of irrelevant sources of variation. Sources that cannot be controlled must be repeatedly measured to account for them. Samples where these conditions have been altered should be included in the training set so that these changes do not drive the classification. The goal, of course, is to prevent major technological differences from overwhelming the biology associated with the outcome of interest.

Careful experimental design can help. By randomizing the samples, we can ensure that uninteresting factors — changes in the machine calibration, differences in chip quality, variations in the reagents — affect both kinds of samples equally and thus are not accidentally detected as biological differences. Keeping the operators blinded to the nature of the samples can also help ensure that systematic differences in processing do not occur inadvertently. (Ideally, "operators" here includes everyone who handles the sample from the nurse drawing blood to the technician performing the mass spectrometry.)

Results must also be carefully calibrated and revalidated after every shift in protocol. The same samples must be processed using both versions of the protocol, and the classification results confirmed. The results should remain robust with respect to major changes that are likely to occur but which noticeably affect the spectra.

On the analytical side, even within the black box paradigm, we believe that there are better ways to approach the analysis of proteomic spectra. For example, we believe that processing steps such as baseline correction are necessary with the current technology, since matrix distortions are often severe. We also believe that normalization is necessary after baseline correction has excluded the matrix noise region. We believe that dimension reduction by peak finding is useful, especially as we have found that many of the best "separators" tend to occur on the slopes of peaks rather than at the peaks themselves. (The gains in separability are slight and more than offset by the lack of interpretability.) Finally, while more involved statistical techniques can indeed find evidence of complex structure, we believe that low dimensional approaches using simple tests should be tried first. Given adequate sample sizes and randomization, it is reasonable to expect that if the data are properly processed then many different methods will

find the structure of interest.

The above discussion is predicated on the assumption that the black box approach should be preferred. We admit that we also find this assumption questionable. We suspect that pursuit and identification of some of the proteins involved in differentiating the samples might yield diagnostic tests that can be verified using other technologies and that are more generalizable to new data sets. This alternative approach also holds out the promise of providing explanations of the biological mechanism underlying the disease process.

We do believe there is biological information of interest to be had in proteomic spectra, but the level of external noise we see at present suggests the need for a great deal of caution making broad claims for the results.

Acknowledgements

The authors gratefully acknowledge support from the Michael and Betty Kadoorie Foundation and the State of Texas Tobacco Settlement Fund.

References

[1] Baggerly, Keith A., Morris, Jeffrey S., Wang, Jing, et al. (2002). "A Comprehensive Approach to the Analysis of MALDI-TOF Proteomics Spectra from Serum Samples", *Proteomics*, to appear.

[2] Cordingley, H. C., Roberts, S. L. L., Tooke, P., et al. (2003). "Multifactorial Screening Design and Analysis of SELDI-TOF ProteinChip Array Optimization Experiments", *BioTechniques*, **34**:364-373.

[3] Fung, Eric, and Enderwick, Cynthia (2002). "ProteinChip Clinical Proteomics: Computational Challenges and Solutions". *Biotechniques Computational Proteomics Supplement*, **34**:S34-S41.

[4] Goldberg, D.E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA.

[5] Holland, J.H. (1994). *Adaptation in Natural and Artificial Systems, 3e*. MIT Press, Cambridge, MA.

[6] Li, Jinong, Zhang, Zhen, Rosenzweig, Jason, Wang, et al. (2002). "Proteomics and Bioinformatics Approaches for Indentification of Serum Biomarkers to Detect Breast Cancer", *Clinical Chemistry*, **48(8)**:1296-1304.

[7] Petricoin III, Emmanuel F., Ardekani, Ali M., Hitt, Ben A., et al. (2002). "Use of Proteomic Patterns in Serum to Identify Ovarian Cancer". *The Lancet*, **359**:572-577.

[8] Rai, Alex J., Zhang, Zhen, Rosenzweig, Jason, et al. (2002). "Proteomic Approaches to Tumor Marker Discovery: Identification of Biomarkers for Ovarian Cancer", *Archives of Pathology and Laboratory Medicine*, **126**:1518-1526.

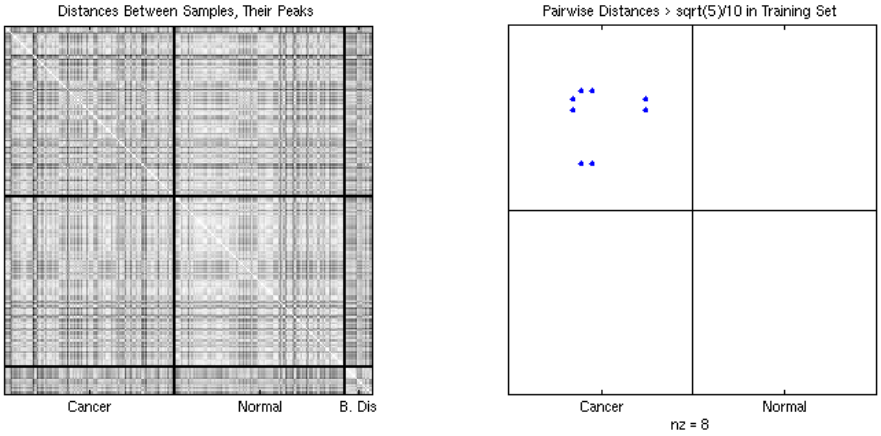

Figure 1: Euclidean distances between the 5-vectors of each sample evaluated at the peaks supplied, and a binary matrix indicating distances greater than $\sqrt{5}/10$. This feature set does not distinguish the two groups after baseline correction.

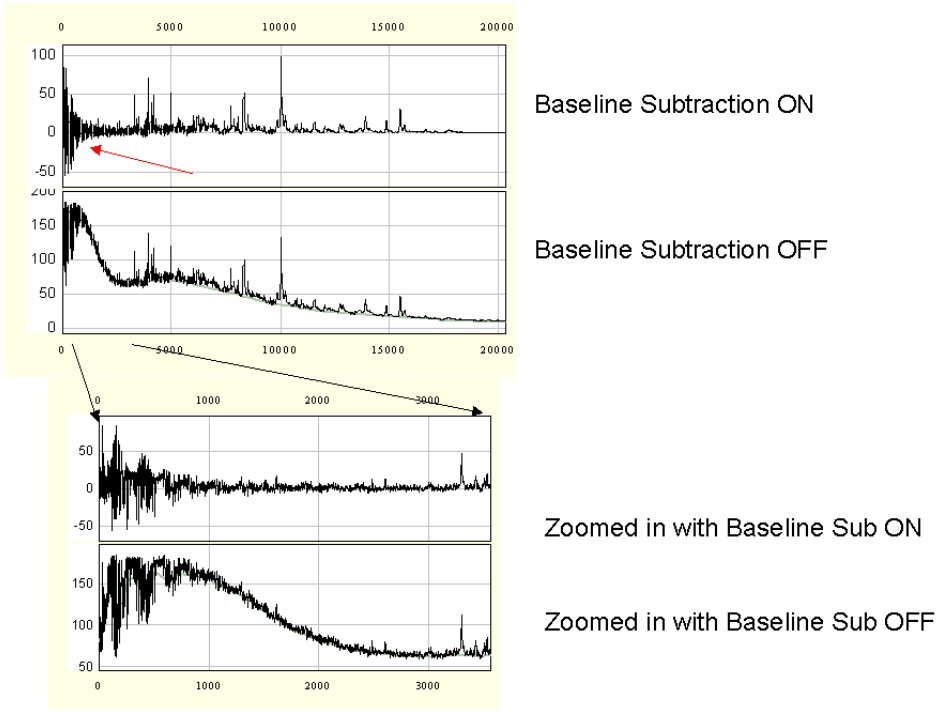

Figure 2: The effect of baseline subtraction on proteomic spectra. The operation is nonlinear and irreversible, and produces negative intensities at low m/z values. Normally, low m/z values are excluded from consideration due to known contamination with matrix noise.

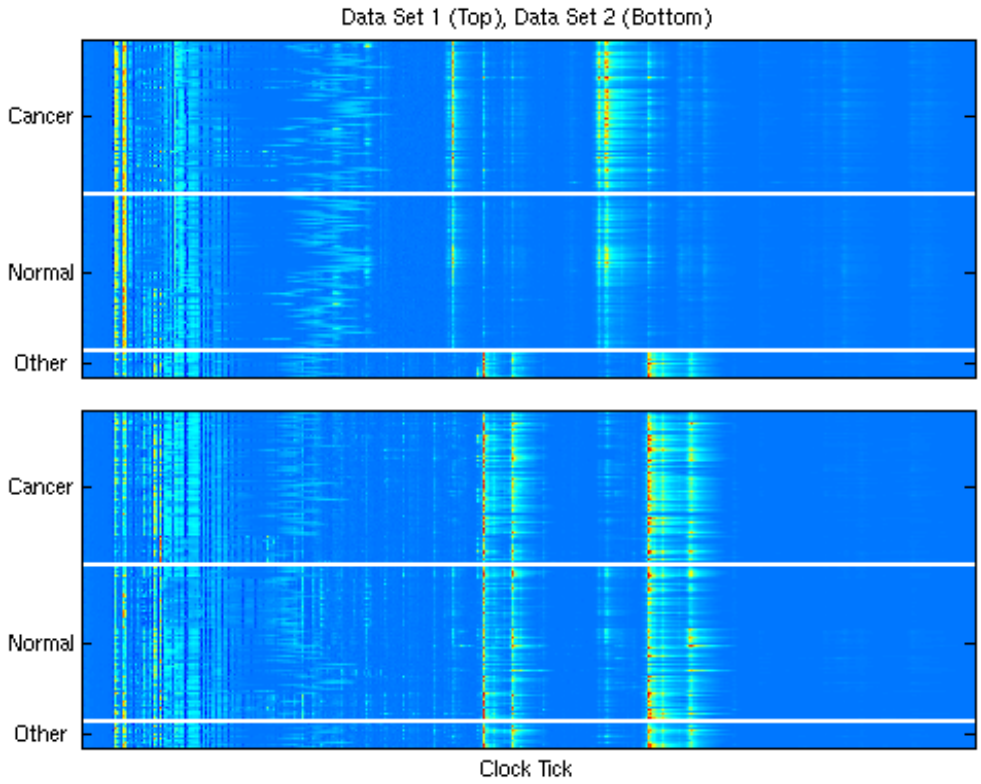

Figure 3: Heat map of all 216 samples from data set 1 (top), which were nominally run on the H4 chip, and of all 216 samples from data set 2 (bottom), which are nominally the same biological samples as data set 1, just run on the WCX2 chip. The gross break at the "benign disease" juncture in data set 1, and the similarity of the profiles to those in data set 2, suggests that a change in protocol occurred in the middle of the first experiment.

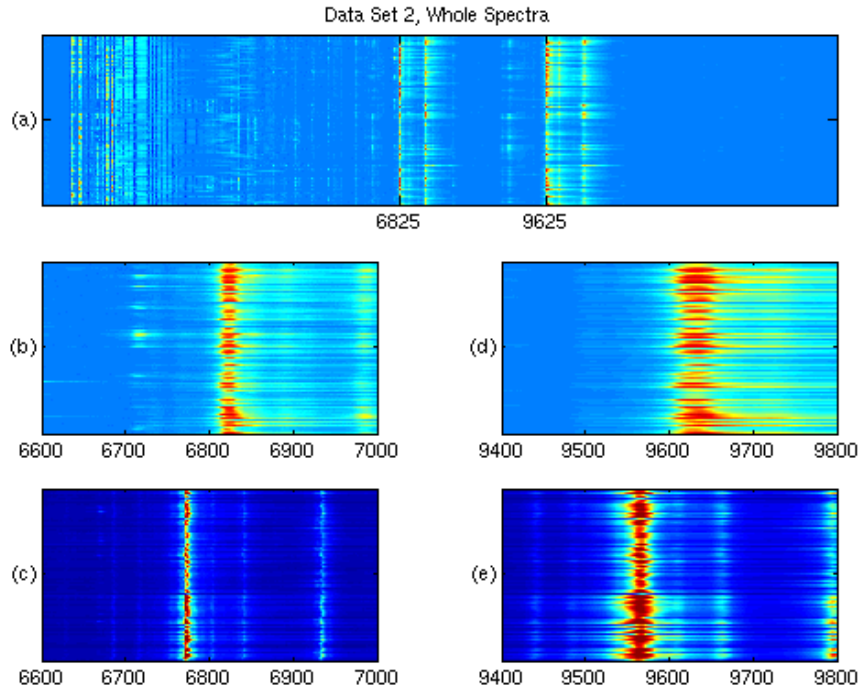

Figure 4: An attempt to align the spectra from data sets 2 and 3. (a) The whole spectra from data set 2, with two of the high mass peaks identified. (b) Zoom on the left peak region for data set 2, and (c) zoom on the same peak region for data set 3. (d) and (e) show the corresponding zooms for the right peak region. There is clearly an offset between data sets 2 and 3. The x-axis in all of the plots indicates the clock tick (of 15154); corresponding m/z offsets are more than 1%.

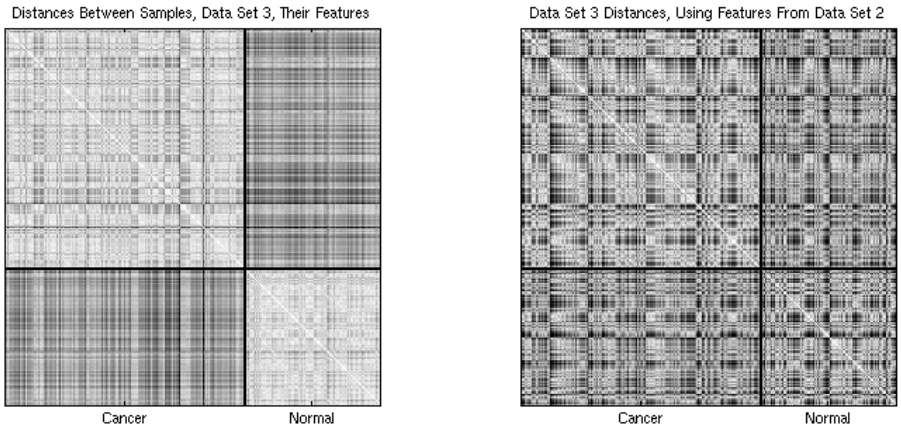

Figure 5: (a) Euclidean distance matrix using the 7 peaks identified for data set 3. The separation between cancer spectra and normal spectra is obvious. (b) Euclidean distance matrix for data set 3 using the 5 peaks identified for data set 2. The structure is effectively random, and there is no clear separation between cancers and normals. The data set 2 peaks do not separate data set 3 well.

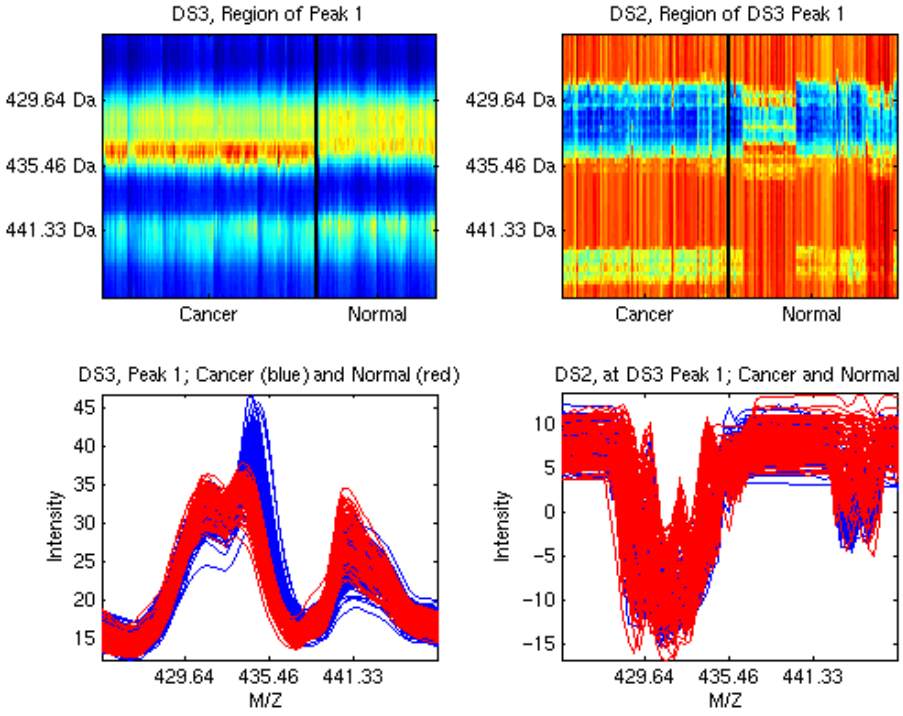

Figure 6: The region of the best separating peak for data set 3, shown for both data set 3 and data set 2. While this value does a good job of separating cancers from normals in data set 3, producing a visible peak, the corresponding region in data set 2 shows evidence of local saturation (flat tops) and reverse behavior with respect to what is high and what is low.

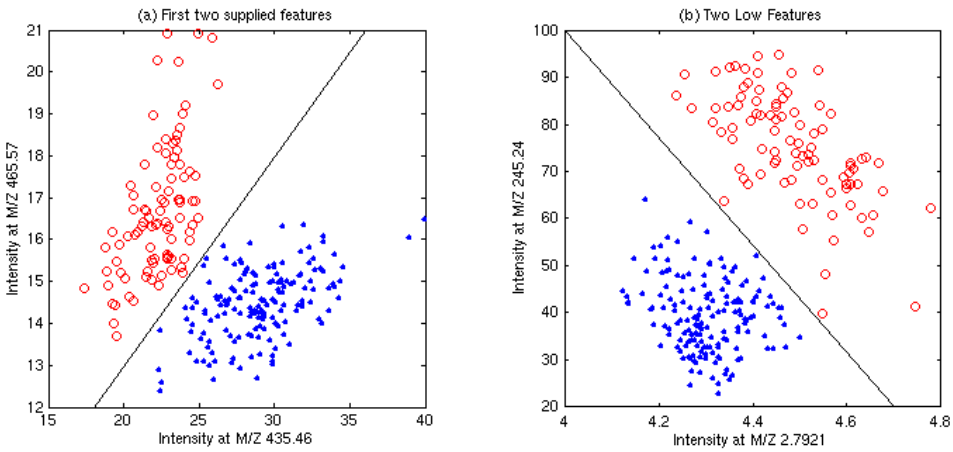

Figure 7: Separation of Cancers (dots) from Normals (circles) for data set three. (a) Perfect classification of tumors and normals using intensities at m/z values of (435.46, 465.57). These are the first two values in their list of 7, and are sufficient in and of themselves. Note that both masses are below 600, and thus questionable with respect to matrix contamination. (b) Perfect classification of tumors and normals using intensities at m/z values of (2.79, 245.2). Both of these values are clearly in the noise region, and the fact that we can perfectly separate the tumors from the normals here suggests the presence of some type of systemic bias.