BIOINFORMATICS

Vol. 20 no. 5 2004, pages 777–785
DOI: 10.1093/bioinformatics/btg484

Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments

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, USA

Received on July 14, 2003; revised on October 14, 2003; accepted on October 16, 2003
Advance Access publication January 29, 2004

ABSTRACT

Motivation: There has been much interest in using patterns derived from surface-enhanced laser desorption and ionization (SELDI) protein mass spectra from serum to differentiate samples from patients both with and without disease. Such patterns have been used without identification of the underlying proteins responsible. However, there are questions as to the stability of this procedure over multiple experiments.

Results: We compared SELDI proteomic spectra from serum from three experiments by the same group on separating ovarian cancer from normal tissue. These spectra are available on the web at http://clinicalproteomics.steem.com. In general, the results were not reproducible across experiments. Baseline correction prevents reproduction of the results for two of the experiments. In one experiment, there is evidence of a major shift in protocol mid-experiment which could bias the results. In another, structure in the noise regions of the spectra allows us to distinguish normal from cancer, suggesting that the normals and cancers were processed differently. Sets of features found to discriminate well in one experiment do not generalize to other experiments. Finally, the mass calibration in all three experiments appears suspect. Taken together, these and other concerns suggest that much of the structure uncovered in these experiments 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 ensure better reproducible, biologically meaningfully results.

Availability: The MATLAB and Perl code used in our analyses is available at http://bioinformatics.mdanderson.org

Contact: kabagg@mdanderson.org

INTRODUCTION

There has been much recent interest in using patterns in SELDI proteomic mass spectra derived from serum to differentiate samples from patients both with and without disease. However, there are questions as to the stability of this procedure over multiple experiments. An illustration of the potential power of the proteomic technique is apparently provided by ovarian cancer. Ovarian cancer is frequently a deadly disease, and its 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 invasive, and gene product assays (such as CA125) have never been shown to be effective in screening programs. A simple, easily applied diagnostic test with high sensitivity and specificity would be of great utility.

In a recent paper in *The Lancet*, Petricoin *et al*. (2002) reported finding patterns in SELDI-TOF proteomic spectra that can distinguish between serum samples from healthy women and those 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, with 50 cancer and 50 normal spectra used to train a classification algorithm. The resulting algorithm was used to classify the remaining spectra. It correctly classified 50/50 of the cancers and 47/50 of the normals in the validation subset. It called 16/16 of the benign disease 'other' than normal or cancer[1]. These results are impressive and have received a good deal of attention.

The initial experiment and two related experiments by the same group of investigators have some very positive features: (1) the investigators collected enough samples to find real structure in the data; (2) splitting the data into training and validation sets allows for internal validation of the structure found, protecting somewhat against the tracking of random noise; and (3) all the data has been made publicly available (on the website http://clinicalproteomics.steem.com). One or two other groups have been willing to make their data available (Adam *et al*., 2002), but this laudable practice is not yet universal.

*To whom correspondence should be addressed.

[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.

Through our own analyses of these data, we 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 well beyond what would be expected by random noise. The question is whether 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 instrument or lab, among others. The answer to the question of whether biology or artifact is the driving force is crucial: the former has scientific and medical implications and should be reproducible, while the latter means only that the statistical analysis of separating signal from noise will be more difficult.

Our findings suggest that while there are differences within individual experiments, these differences are not the same across experiments. This lack of agreement indicates the need for careful experimental design, for varying experimental conditions when conducting such studies, and for better methods of external calibration.

BACKGROUND

In this section we review the nature of the available data, 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. All the methods discussed in this section are those that are presented on the website http://clinicalproteomics.steem.com and used in the initial analysis by Petricoin *et al*. (2002). These are not our own methods.

SELDI mass spectrometry

In brief, surface-enhanced laser desorption and ionization time of flight (SELDI-TOF) mass spectrometry begins with applying a biological sample (such as serum) to a precoated stainless steel slide. This coating 'enhances' the surface to bind preferentially a particular class of proteins based on their physiochemical properties. Different coatings give different 'chip types' which bind to different classes of proteins. The sample is further mixed with an energy absorbing matrix (EAM) such as sinnapinic acid, which causes the entire mixture to crystallize as it dries. The sample is then put into a vacuum chamber and the crystal is hit with a laser, causing the proteins to desorb and ionize when the matrix absorbs the energy produced at the wavelength of the nitrogen laser. This produces ionized protein molecules in the gas phase. A brief electric field is then applied to accelerate the ions down a flight tube, and a detector at the end of the tube records the time of flight. Given the time of flight, the known length of the tube and the voltage applied, the mass-to-charge ratio (m/z value) of the protein can be derived. A typical spectrum consists of the sequentially recorded numbers of ions arriving at the detector (the intensity) coupled with the corresponding m/z value. Peaks in the intensity plot ideally correspond to individual proteins. We distinguish between a peak (a local maxima in the observed spectra) and a feature (the observed intensities at a particular m/z value). A set of spectra will have thousands of features, but only a small fraction of these would correspond to peaks. More details are available in Siuzdak (1996) or de Hoffman and Stroobant (2002).

The datasets

There are three datasets of SELDI ovarian mass spectra derived from serum currently available on the website, which we shall refer to as datasets 1–3, respectively.

(1) Dataset 1, which was described in the initial paper (Petricoin *et al*., 2002), consists of 216 spectra divided into five files: Training Cancer, Training Normal, Test Cancer, Test Normal and Benign Disease. These spectra were obtained using the Ciphergen H4 ProteinChip array. These spectra were baseline-subtracted.

(2) Dataset 2 uses the same 216 samples as above, but run on the Ciphergen WCX2 ProteinChip array. Again, the spectra were baseline-subtracted.

(3) Dataset 3 contains a new set of samples, 91 normal and 162 cancer. The split into training and test groups is not reported here. These samples were prepared robotically, whereas the samples in the previous datasets were prepared by hand. The samples were run on the WCX2 array which was also used for dataset 2. Finally, these spectra have not been baseline subtracted.

Each spectrum is reported in a text file with two columns. The first column consists of a list of 15 154 mass-to-charge ratios (m/z values) and the second column gives the associated intensities. The m/z values reported are common across all spectra and all datasets.

The data processing

Despite the fact that baseline-subtracted data are provided on the website for the first two datasets, we believe all their analyses were performed on the datasets before baseline subtraction. More details are given in the analysis and discussion. Before comparison, all the spectra in an experiment were normalized to have the same [0, 1] intensity range as follows. For a single spectrum, Let V_i denote the raw intensity at the i-th m/z value, $i \in \{1, \ldots, 15\,154\}$, and let $V_{\min}$ and $V_{\max}$ denote the smallest and largest observed intensities in the spectrum, respectively. Then the normalized intensity NV_i is given by

$$NV_i = \frac{V_i - V_{\min}}{V_{\max} - V_{\min}}.$$

The fitness function

The 'fitness' of a particular feature set containing N features is assessed using the associated scaled intensities to define

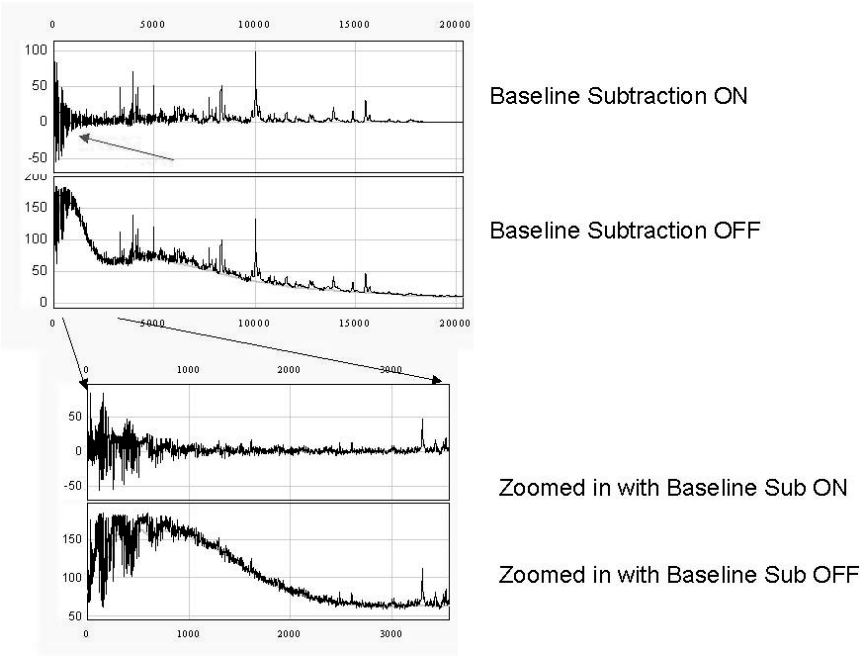

Fig. 1. The effect of baseline subtraction on proteomic spectra. The operation is nonlinear and irreversible, and produces negative intensities at low m/z values. Normally, low m/z values are excluded from consideration due to known contamination with matrix noise.

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. 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 (Goldberg, 1989; Holland, 1994). Each run of the genetic algorithm starts with 1500 logical chromosomes (feature sets) of a 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 are available for inclusion in a feature set. There is no initial peak finding step.

The best feature sets

For each of the experiments, a single feature set that allows the cancers to be separated from the normals is reported.

OUR ANALYSIS OF THE SELDI OVARIAN CANCER SERUM DATASETS

Our goals in looking at all of the ovarian datasets were to check the initial results and to find features separating cancers and normals that were stable across experiments. However, we encountered some major problems.

Baseline correction prevents reproduction of results

Using the spectra in dataset 1, we processed the data according to the methods described above. Then, using the processed intensities at the 5 m/z values in the feature set reported on the website, we computed the distances between all pairs of spectra and assembled these in a distance matrix. 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 look for 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 Fig. 4a, described below.) Second, there were only four pairwise distances greater than $\sqrt{5}/10$, which is the cutoff distance for declaring a new cluster with five 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 never be formed.

The problem is that the posted data have been baseline subtracted (Fig. 1). The web page comments on this issue,

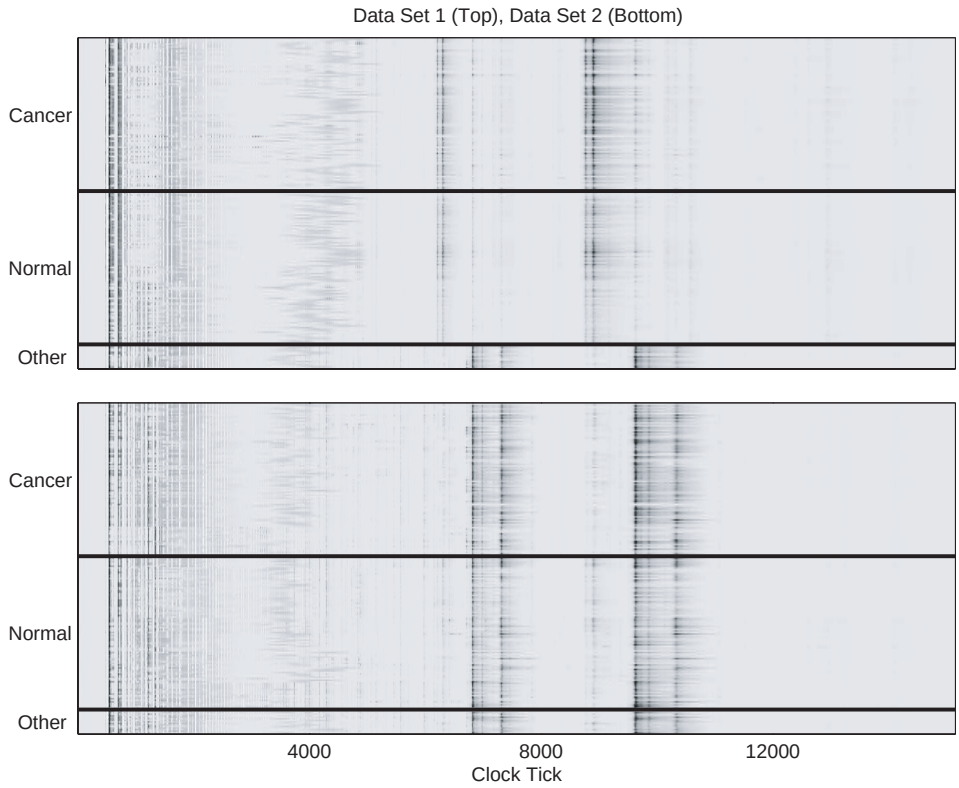

Fig. 2. Heat map of all 216 samples from dataset 1 (top), which were run on the H4 chip, and of all 216 samples from dataset 2 (bottom), which are the same biological samples as dataset 1, just run on the WCX2 chip. The gross break at the 'benign disease' juncture in dataset 1, and the similarity of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

noting that 'this process creates negative intensities'. But a more serious problem is that this correction is an irreversible nonlinear operation. Given only the baseline-subtracted values, it is impossible to reconstruct the raw values. It is possible to turn off the baseline correction within the Ciphergen software if the raw binary files (XPT files) are supplied, but this approach cannot be taken using just the text files at hand. This problem prevented reproduction of their results for the first and second ovarian datasets. We were able to reproduce their results on dataset 3, which was not baseline-corrected.

Sample processing differences cause blatant changes

The algorithm used in the original study was able to identify the 16 benign disease samples in dataset 1 as 'other' than normal or cancer. A 'heat map', in which all the spectra are plotted side-by-side and regions of higher intensity are shown by darker bands, is shown for the 216 spectra from dataset 1 in Figure 2 (top).

The benign disease spectra at the bottom are clearly distinct from the rest. Indeed, the cancer spectra and normal spectra show far greater similarity to each other than either does to benign disease. Conversely, the 216 spectra from dataset 2, shown in Figure 2 (bottom), do not show such a separation. Because these are the same biological samples, run on a different kind of ProteinChip array, this lack of separation is disturbing. Considering both image maps together, however, we see that the benign disease spectra from dataset 1 have profiles that are extremely similar to those of dataset 2. This observation suggests that there was a change in protocol before the first set was complete. This is disturbing, as the Lancet article states that 'positives and controls were run concurrently, intermingled on the same chip and on multiple chips; the operators were unaware of which was which'.

We considered the possibility that an error had been made when the datasets were prepared for posting to the web, and that the benign disease spectra posted as part of dataset 1 were actually the same spectra posted with dataset 2. To test this possibility, we compared the numerical values in the spectra. We found that none of the benign disease spectra in dataset 1 were numerically identical to any of the benign disease spectra in dataset 2.

Dataset 3 is offset relative to dataset 2

To see if we could generalize results across experiments, we tried to view datasets 2 and 3 (which used the same chip type) simultaneously. Even though dataset 2 was

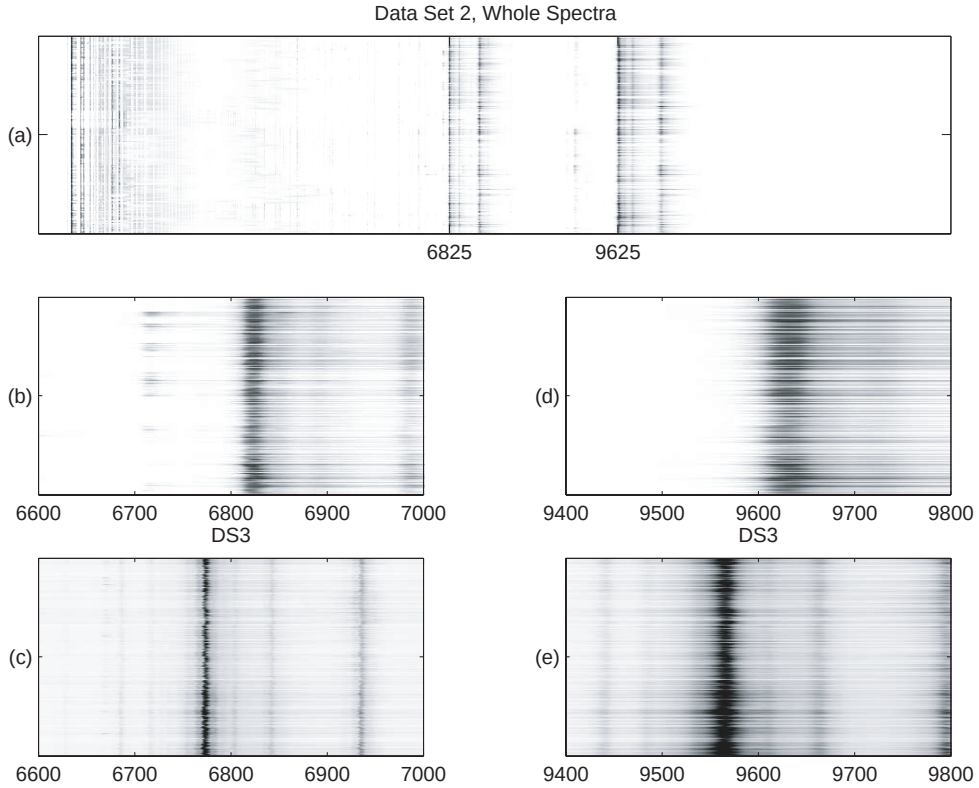

Fig. 3. An attempt to align the spectra from datasets 2 and 3. (**a**) The whole spectra from dataset 2, with two of the high mass peaks identified. (**b**) Zoom on the left peak region for dataset 2 and (**c**) zoom on the same peak region for dataset 3. (**d**) and (**e**) show the corresponding zooms for the right peak region. There is clearly an offset between datasets 2 and 3. The x-axis in all of the plots indicates the clock tick (of 15 154); corresponding m/z offsets are more than 1%.

baseline-corrected, we hoped to use qualitative features of the spectra to assess similarity. We attempted to match the indices of the major high intensity peaks for comparison. We found that the spectra from dataset 3 were offset by roughly 50–60 clock ticks from the spectra in dataset 2 (Fig. 3).

Converting into 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%.

In order to bring datasets 2 and 3 to better agreement, we fit a quadratic offset to the masses of dataset 3 by least squares using four peaks for calibration. The four peaks were chosen from the average spectra for datasets 2 and 3, at time indices (6773, 7269, 9566, 10 260) in dataset 3 and (6823, 7321, 9632, 10 337) in dataset 2, respectively. Matching at these four peaks introduces an offset of nearly 200 Da at the low end of the m/z region.

Separating feature sets are not reproducible across experiments

For 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 between dataset 1 and datasets 2 and 3, the results from the first experiment cannot be compared with those of the other two. Because datasets 2 and 3 share a common chip surface, we assumed they should be comparable. The feature set reported for dataset 2 contains five features. When we computed the distance matrix for dataset 3 using the intensities at these five features, the distance matrix clearly showed that the cancer samples and the normal samples had not been split apart (Fig. 4b).

The problem is not remedied by including either a linear or a quadratic offset term to correct for the calibration problem noted above; the distance plots produced are qualitatively similar to that shown in Figure 4b (offset data not shown). Testing the validity of the features found by analyzing dataset 3 by applying them to dataset 2 is more difficult, because of the baseline correction applied to dataset 2. Thus, we checked the results one feature at a time. There were seven features reported for dataset 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 dataset 3. Checking the shapes of the spectra in this local region for both datasets, we found that there was clearly a visible separation between the sample types associated with the slope of a peak in dataset 3. However, not only is there no clear separation in dataset 2, but the shape in the region is no longer that of a peak but rather that

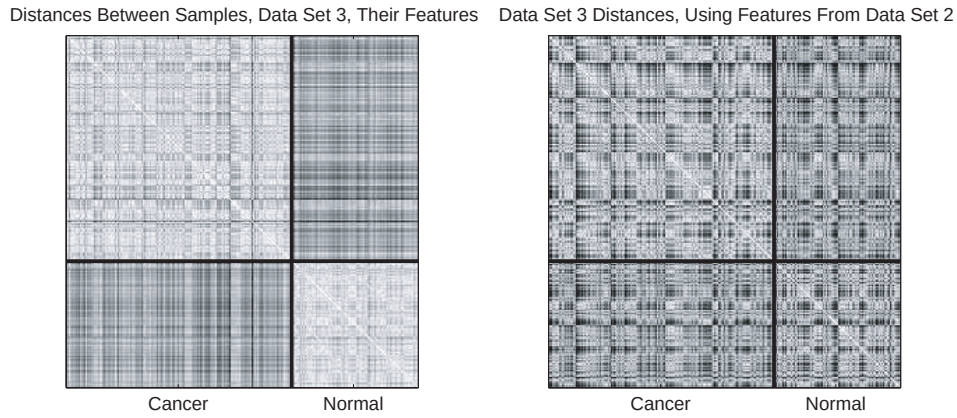

Fig. 4. (**a**) Distance matrix using the seven peaks identified for dataset 3. The separation between cancer spectra and normal spectra is obvious. (**b**) Euclidean distance matrix for dataset 3 using the five peaks identified for dataset 2. The structure is effectively random, and there is no clear separation between cancers and normals. The dataset 2 peaks do not separate dataset 3 well.

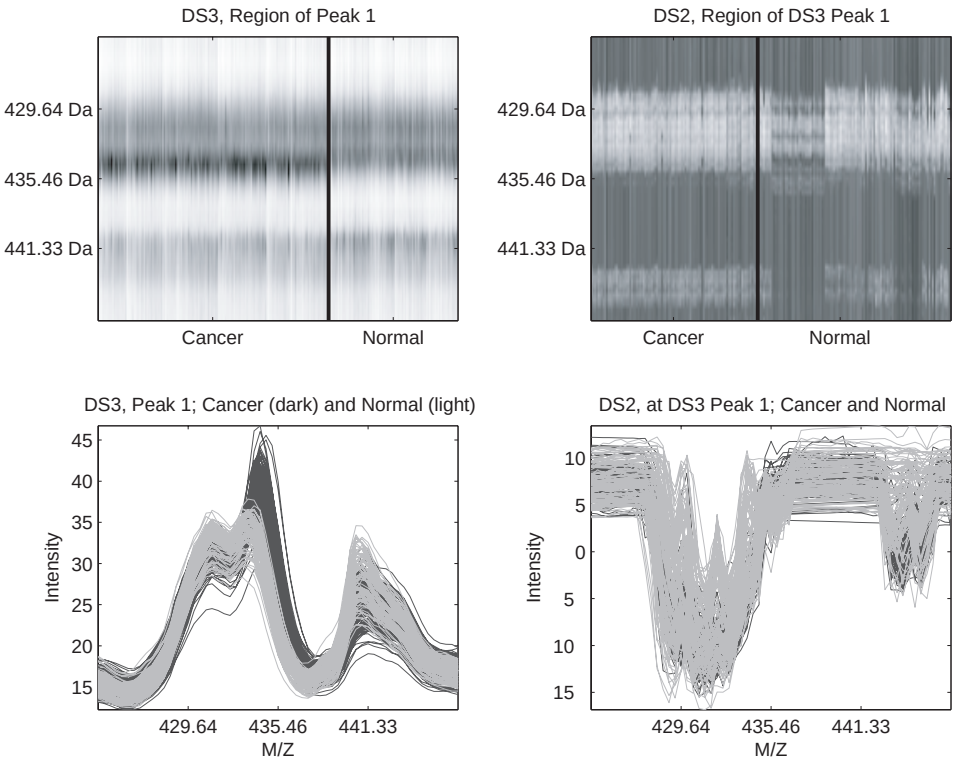

Fig. 5. The region of the best separating peak for dataset 3, shown for both dataset 3 and dataset 2. While this value does a good job of separating cancers from normals in dataset 3, producing a visible peak, the corresponding region in dataset 2 shows evidence of local saturation (at tops) and reverse behavior with respect to what is high and what is low.

of a valley. Moreover, there is clear evidence that the spectra were locally saturated before baseline subtraction (Fig. 5, flat regions of high intensity).

We found similar lack of agreement for the other features (data not shown). Again, we attempted to correct the situation by including an offset term. This did not result in qualitatively better agreement (we tried linear offsets of 50, 55 and 60 clock ticks and the quadratic fit described above; data not shown).

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. We elected

Table 1. Two-sample t-values for the normal/cancer separation in dataset 3 at the 7 m/z values supplied on the website

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

to search for the presence of such sets in dataset 3. We initially processed the data as described, but we noted very little change when we looked at how the normalization method affected the data. Of the 253 spectra in dataset 3, 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. The corresponding minimum intensities are almost all between 3.8 and 3.9, with no values falling outside the interval [3.75, 3.96]. In light of this, we elected to work with the raw spectra (with no correction at all).

We searched for the single features and pairs of features that best separated normals from cancers. In order to restrict our attention to the features most likely to achieve good separation, we applied a two-sample t-test to the difference between cancer samples and normal samples for every m/z value in dataset 3. We paid special attention to the 7 m/z values supplied on the website. The most extreme t-values are huge in magnitude, with the largest in the list of seven occurring at m/z 435.46, where the t-value is 22.3463. Using the intensities at this single feature, we correctly classified 238 of the 253 samples. We note that this is the same value that failed to separate the spectra in dataset 2. Looking at the t-values for the seven reported m/z values individually suggests that the first two are more important than the others (Table 1). Using the t-values to suggest particularly interesting features, we were led to several values not in the initial 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 misclassified only five of the samples.

Looking only at the individual features where large t-statistics ($|t| > 10$) 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 seven. Both masses are less than 500. The second pair of m/z values is (2.79, 245.2), with t-values of (-13.89, -27.0256), respectively. These two m/z values are clearly in the noise region; the first may be in the range before the instrument is recording stably.

The fact that there is structure in the noise has recently been discussed in depth by Sorace and Zhan (2003).

Another analysis

We performed an independent analysis of dataset 3, to see if the features we identified would coincide with theirs. Details of our processing (baseline correction, normalization and peak selection) are available from our website, http://bioinformatics.mdanderson.org. After processing, 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 12 and identified nine significant peak regions, with m/z values equal to: 64, 245, 434, 618, 886, 1531, 3010, 3200 and 8033. Only one of these peaks (m/z 434) also appears on their list. There are several good separators above the noise region of the spectra, and we believe that good separation could be achieved using only these, but we are still suspicious of these results here given the structure in the noise.

The first four peaks on our list are located well below the end of the matrix noise region in a part of the spectrum where saturation commonly occurs. None of the peaks that we found generalized to dataset 2.

The mass calibration for all three datasets is suspect

We noted earlier that the vectors of m/z values supplied for all three datasets were the same. As these values are derived from the calibration equation relating time-of-flight to m/z, this identity implies that the same mass calibration was in place throughout, over a period of several months. We then looked more closely at the precise m/z values. The first few such values are $-7.86\text{E}-05$, $2.18\text{E}-07$, $9.60\text{E}-05$, 0.000366014 and 0.000810195. We note that these are the precise values that the Ciphergen software coming with our PBS-II instrument returns when we use the factory default calibration settings, which apply before a calibration sample has been run.

DISCUSSION

In an effort to understand better 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 agreement across experiments.

We found that baseline correction prevented reproduction of their initial results, suggesting that the initial analysis was performed on the raw spectra. 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 instrument, and introduces visible offsets that persist for the length of the spectra.

We believe some form of baseline correction is useful. Baselines of different spectra can be highly variable. They

change from instrument to instrument and from day to day on the same instrument. In general, the baseline signal is caused mostly by chemical noise from matrix molecules, with some contribution from electronic noise (Fung and Enderwick, 2002; Baggerly *et al.*, 2003). The matrix noise contribution to the baseline signal is largest in the low m/z region. Given this, our inability to reproduce their analysis using the posted data is worrisome. 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.

Just because we cannot reproduce their results in datasets 1 and 2 does not mean that there is no structure to be discerned. Comparing datasets 1 and 2 showed that there was a change in protocol in the middle of dataset 1. One example of a protocol change that could produce the observed results is a shift between chip types. Different chip surfaces, by design, bind different sets of proteins. Alternatively, maintenance or replacement of critical portions of the Ciphergen instrument 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 observed an offset between datasets 2 and 3 that was substantially larger than the nominal precision of the procedure. The observed offset between the datasets 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.

We were unable to use features from dataset 2 to separate normals from cancers in dataset 3, which uses the same chip type. Part of this is due to the calibration problem, but this is exacerbated by the fact that most of the features reported are in the very low m/z range for all three datasets. The low mass range is prone to unstable calibration and to other artifacts such as signal saturation. We note that the features supplied for dataset 1 are also in the lower end of the mass range.

We were able to find multiple feature sets that perfectly classified the samples in dataset 3, and at least one such feature set that lies wholly in the noise region of the spectra. The fact that we can achieve perfect classification based on readings entirely in the noise region, however, is evidence of a problem. There can be no biological reason for the difference at m/z 2.79: It is pure instrument artifact. The presence of such an artifact suggests that there was a systematic difference in the way the groups of samples were processed. There should be no pattern in the noise region, and perfect separation of so many samples is essentially impossible by random chance alone. Structure in noise, by itself, is mildly disturbing. When combined with other findings that the systematic differences caused by the technology can be large compared with 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.

The reported m/z values correspond to those that would be supplied if no external calibration had ever been applied to the system. This is a problem, in that the mass values are likely to be inaccurate, and can differ by substantially greater amounts than the nominal accuracy of the instrument (which presumes regular calibration) would imply. Calibration appears to be something of a 'stealth issue'. It can introduce offsets between labs even if both labs have calibrated their spectra, if, e.g. different choices are made for the calibrants employed. This problem is worse if the m/z values need to be extrapolated beyond the observed mass range of the calibrants. This issue of extrapolation is another reason why the very low mass range of the spectra is suspect, as the mass predictions can be unstable.

It is important to note that there are many aspects of this study that we applaud, and had we been consulted, we would have likely supported the publication of the first data set based on the separation rates achieved. (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 4×50 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 dataset to another—capable of distinguishing between the two groups of samples. It is only because the original authors have posted multiple datasets 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.

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 instrument can drift), temperature, humidity, the instrument 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*. (2003). 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 instrument 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, processing steps such as baseline correction are necessary with the current technology, since matrix distortions are often severe. Normalization is necessary after baseline correction, and the matrix noise region should be excluded. Dimension reduction by peak finding is useful, especially since 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, we think that there are some problems with the fitness function and clustering methods described in Petricoin *et al*. (2002). Specifically: classification accuracy is the only measure of fitness used. 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*. Finally, in jumping immediately to dimension 5 and higher, we miss the chance to find simpler explanations if such exist. While more involved statistical techniques can indeed find evidence of complex structure, low-dimensional approaches using simple tests should be tried first. Given adequate sample sizes and randomization, if the data are properly processed then various methods will find the structure of interest if it is there to be found.

The above discussion is predicated on the assumption that the black box approach is preferred. This assumption is 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 datasets. This alternative approach also holds out the promise of providing explanations of the biological mechanism underlying the disease process.

Proteomic spectra are promising for scientific discovery. But sufficient external noise can lead to false discovery.

ACKNOWLEDGEMENTS

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

REFERENCES

Adam,B.-L., Qu,Y., Davis,J.W., Ward,M.D., Clements,M.A., Cazares,L.H., Semmes,O.J., Schellhammer,P.F., Yasui,Y., Feng,Z. and Wright,G.W.,Jr (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.*, **62**, 3609–3614.

Baggerly,K.A., Morris,J.S., Wang,J., Gold,D., Xiao,L.-C. and Coombes,K.R. (2003) A comprehensive approach to the analysis of MALDI-TOF proteomics spectra from serum samples. *Proteomics*, **3**, 1667–1682.

Cordingley,H.C., Roberts,S.L.L., Tooke,P., Armitage,J.R., Lane,P.W., Wu,W. and Wildsmith,S.E. (2003) Multifactorial screening design and analysis of SELDI-TOF ProteinChip array optimization experiments. *Biotechniques*, **34**, 364–373.

de Hoffman,E. and Stroobant,V. (2002) *Mass spectrometry: Principles and Applications*. John Wiley, New York, NY.

Fung,E. and Enderwick,C. (2002) ProteinChip clinical proteomics: computational challenges and solutions. *Biotechniques Comput. Proteomics Suppl.*, **34**, S34–S41.

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

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

Petricoin,E.F.,III, Ardekani,A.M., Hitt,B.A., Levine,P.J., Fusaro,V.A., Steinberg,S.M., Mills,G.B., Simone,C., Fishman,D.A., Kohn,E.C. and Liotta,L.A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, **359**, 572–577.

Sorace,J.M. and Zhan,M. (2003) A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics*, **4**, 24. http://www.biomedcentral.com/1471-2105/4/24

Siuzdak,G. (1996) *Mass Spectrometry for Biotechnology*. Academic Press, San Diego, CA.