

Searching Model Space for SELDI-TOF Cancer Classifiers

Eric R. Siegel
UAMS Biostatistics

University of Arkansas for Medical Sciences, Little Rock AR 72205-7199

Abstract

To develop a classification model, we typically divide the data in half in order to train the model on one half and test it on the other. But when there are many potential predictor variables to choose from, as in cancer proteomics, and when the training phase employs automatic variable selection to choose them, the result is a model whose predictor variables (and associated parameter estimates) may depend on which samples end up in which half. To see how big a problem this is, I recently iterated the division, training, and testing process 10,000 times to get a distribution of models for classifying pancreatic cancer, using logistic regression with Forward Selection on intensities of 37 peaks from the SELDI-TOF mass spectra of 103 serum samples. Although classifiers contained from 3 to 7 peaks, every peak was selected at least once (minimum = 12) and no peak was selected 100% of the time (maximum = 9,279), implying that the peaks chosen by logistic regression with Forward Selection constitute a non-robust and unreliable panel for classification. From each sample's per-iteration probability of correct test-set prediction, I developed two distribution-wide measures of classification performance, and used them to identify a robust panel of SELDI peaks for cancer classification.

Keywords: SELDI, mass spectra, proteomics, classification.

Motivation

The principles of SELDI-TOF mass spectrometry, and its application to cancer proteomics and diagnostics, have been described in much more detail elsewhere (1, 2, 3, 4; see also Fig.1 for a brief explanation of how the technology works). My interest in the statistical analysis of SELDI-TOF data began when I joined with other researchers in a study, ultimately published as Bhattacharyya *et al.* (5), the objective of which was to use high-throughput protein profiling technology to identify biomarkers in the serum proteome for the early detection of resectable pancreatic cancer. In many respects, this paper is a companion to Bhattacharyya *et al.*, intended to describe in more detail some of the methods I developed for that study.

To understand how I came to develop my methods, some history must be disclosed. Initially, we were presented with a well-defined training set of serum samples from 35 patients with pancreatic cancer (cancer samples) and 35 patients with no evidence of this disease (normal samples), plus a blinded test set of 27 cancer samples and 28 normal samples. These

125 samples were used to generate 250 spectra that were subsequently accepted or rejected by an automated quality control procedure. Because we did not know better at the time, we accommodated space-and-time limitations by processing the training set on one day and the test set on a different day. The most obvious result of our confounding of Set with Day was that the quality control procedure rejected a noticeably greater percentage of spectra from the test set (41%) than from the training set (24%). A day-dependent imbalance in the rejection rate for cancer spectra vs. normal spectra was also obvious, as were day-dependent intensity differences in the spectra that were retained.

Then, as we assembled all our results for the manuscript and grant submission, we discovered that four of our training-set samples were from cancer patients whose consent forms did not cover multi-institutional sample sharing. We therefore had to exclude those samples from the analysis, and in particular, from the training set, which meant we had to start over from scratch. We chose to view this event as an opportunity for me to re-randomize the samples into training and test sets so as to homogenize the Day effects across both sets. But when I re-randomized, I did so twice, and obtained (via Forward Selection) classifiers with few predictors in common, both of which were markedly inferior to the original classifier we had trained on our subsequently excluded samples. This caused me to wonder: What would happen if I re-randomized many times and trained a new classifier each time? And would I be able to learn anything from the resulting distribution?

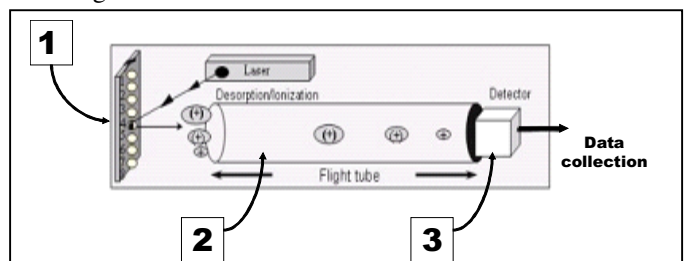
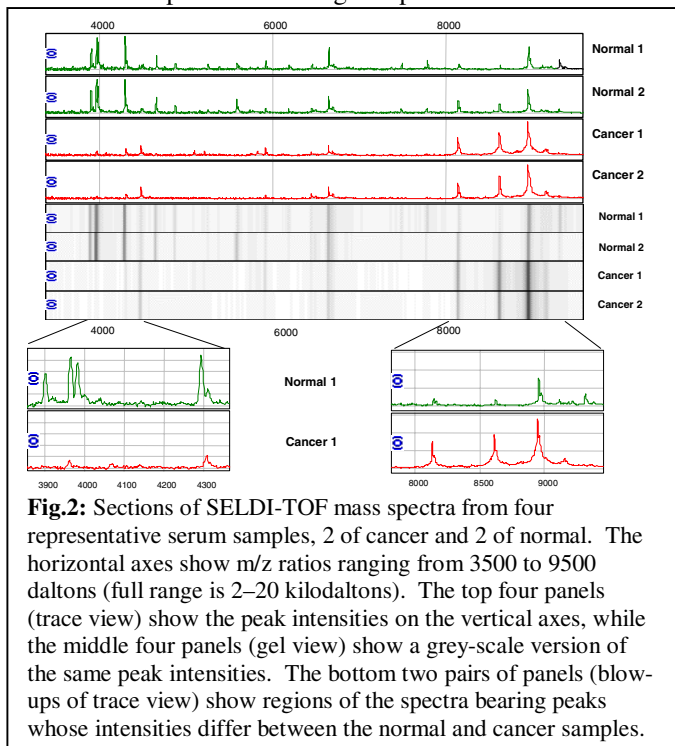


Fig.1: Principles of Surface-enhanced Laser Desorption-Ionization – Time-of-Flight (SELDI-TOF) mass spectrometry. In [1], a spot surface on the protein chip is irradiated by the laser, causing sample molecules to desorb and lose an electron, after which sample ions collect at one end of the flight tube. In [2], the collected sample ions enter the flight tube and are accelerated to the other end by an electric field. Because smaller ions accelerate more quickly, this separates the ions based on their mass-to-charge (m/z) ratio. When the ions reach the other end of the flight tube, [3], the detector measures how many there are and how long they took. The resulting data is ion current as a function of time, or equivalently as a function of the m/z ratio.

The work this paper is about to describe had two goals: (1) via resampling, to study how sensitive the SELDI peak selection process (Forward Selection) was to the random choice of samples to go into the training set, and (2) from the distribution of Forward Selection models obtained from resampling, to identify one or more combinations or "panels" of SELDI peaks that would tend to produce successful classifiers regardless of which samples went into the training set. Developing parameter estimates for a multivariate "final model" was not among my goals.

Methods

The SELDI-TOF mass spectrometer, ProteinChip® technology, and spectrum-processing software was provided by Ciphergen Biosystems, Inc., Fremont CA. Statistical analyses employed SAS version 8.2 (the SAS Institute, Cary NC). Patient characteristics, serum collection and processing, data acquisition, and univariate-analysis results have been described in more detail elsewhere (5). Briefly, 121 samples were applied in duplicate to Ciphergen IMAC-30 chip surfaces previously activated with 100 mM CuSO₄, then processed for laser desorption and collection of mass spectra using the Ciphergen PBS-II C mass analyzer. Samples were processed in two separate-day batches in order to accommodate space-and-time constraints. Ion currents between 2 and 20 kilodaltons were summed after baseline correction to yield a Total Ion Current (TIC) for each spectrum, and spectra were then normalized via multiplication to a common TIC. Spectra for which the TIC multiplication factor was greater than 2.0 or less than 0.5 were discarded so as to exclude spectra exhibiting low peak intensities or current



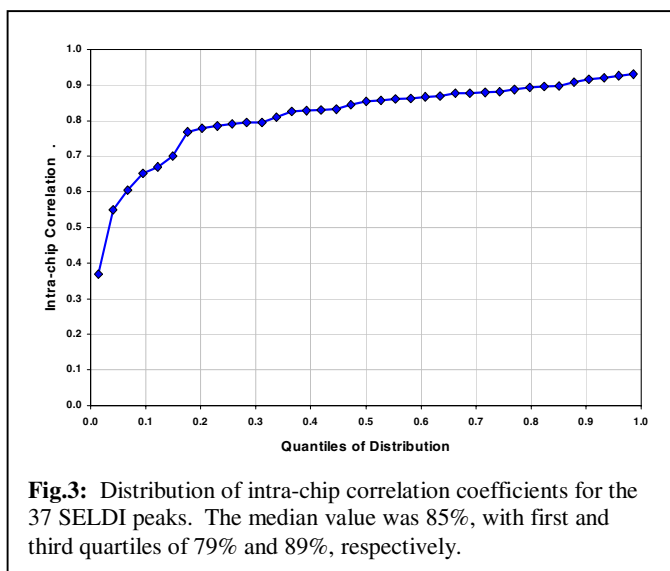
saturation, respectively. Of the 121 samples loaded in duplicate onto the chip surfaces, 103 samples (49 cancer, 54 normal) had at least one spectrum survive the quality-control criteria. Of these, 64 had both spectra survive, and 49 had one spectrum survive, yielding a total of 164 spectra for analysis. Representative spectra are shown in Fig.2.

Ciphergen's Biomarker Wizard software tool was used to identify as peaks those features within a mass window of 0.3% that were consistently present across a minimum of 20% of the spectra with a signal-to-noise ratio ≥ 2.5 . A total of 37 features between 2 and 20 kilodaltons were identified as SELDI mass peaks. For each peak, the spread in mass-to-charge (m/z) ratios was replaced with their median value, which was interpreted as the molecular weight of the peak.

The 64 samples having both spectra survive the quality-control criteria were subjected to variance-components analysis in order to determine each peak's Subject and Residual variance components. The intra-chip (i.e., intra-cluster) correlation coefficient (6) of each peak was estimated as the ratio of the Subject variance component to the sum of both components; Fig.3 shows their distribution. Because of the high intra-chip correlations found (median = 85%), peak intensities from the same sample were averaged to a single value per peak for subsequent statistical analysis.

Generation of Forward Selection models

The following procedure was repeated 10,000 times to generate the set of Forward Selection models. Resampling without replacement was used to partition cancer samples, and likewise normal samples, into equal-sized training and test sets. (In practice, the 2k+1 cancer samples were partitioned into k+1 training samples and k test samples. No effort was made to force equipartitioning of the separate-day batches into training and test sets.) After each resampling, samples falling into the test set had their classifications replaced with missing values, and the data set was subjected to multivariate



logistic regression with Forward Selection applied to all 37 SELDI peaks. Samples were weighted in the logistic regression according to whether average intensities derived from one or two spectra, using the following weight function:

$$weight = 1 + (n - 1)(1 - \rho_{median})$$

where ρ_{median} is the median intra-chip correlation determined from the variance-components analysis, and n is the number of values (1 or 2) from which the average intensity was calculated. (This weight function is structurally the same as the Variance Inflation Factor of Donner *et al.* (6), but with $1 - \rho$ replacing ρ .) The sum of training-set weights in each resampling was then adjusted to be equal to 52, the sample size of the training set. Leave-one-out cross-validation (LOOCV) was used to estimate the probability of correctly classifying each sample. Each probability is denoted as $\Pr(\text{correct}|i, j)$, with i indexing sample and j indexing iteration. Because samples had their classifications deleted when in a test set, $\Pr(\text{correct}|i, j)$ is a cross-validation probability for about half the combinations of i and j , and a test-set prediction probability for the remaining combinations of i and j . After each iteration, the train-test partition, the variables selected, and the $\Pr(\text{correct}|i, j)$ were cumulated and stored for further processing.

Performance measures

The number of times (out of 10,000) that each variable was chosen by Forward Selection was converted into a percentage, the variable's selection rate, and plotted via bar chart. The $\Pr(\text{correct}|i, j)$ from the Forward Selection models were used to define the following two estimates:

- the Per-patient Average Prediction Probability (PPAPP),

$$\bar{p}_{i\bullet} = \frac{1}{n_{j(i)}} \sum_{j \in \text{test set}} \Pr(\text{correct} | i, j)$$

with $n_{j(i)}$ the number of times sample i appeared in a test set; and

- the Expected Number Correctly Classified (ENCC).

$$N\tilde{p}_{\bullet j} = \sum_{i \in \text{test set}} \Pr(\text{correct} | i, j)$$

The PPAPPs of individual samples were used to describe, on average, how easy they were to classify correctly under Forward Selection. ENCCs were plotted against the quantiles of their distribution, and models having ENCCs near the maximum value of 51 were examined more closely.

Generation of additional model types

To generate the Forward Selection models, a fixed seed based on a historical date was used in the random number generator, so that the resampling process could be repeated using different model types on the same 10,000 train-test partitions. Inspection of variable selection rates and ENCCs suggested two additional model types to try, which are named and

described as follows:

- The 5-Best-Peak models, developed via multivariate logistic regression without variable selection on the five most frequently selected peaks from the Forward Selection models, and
- The 4-Peak-Pattern models, developed via multivariate logistic regression without variable selection on four peaks that were strongly over-represented in the subset of Forward Selection models having ENCCs within half a unit of the maximum.

Random realizations of both additional model types were generated by re-creating the 10,000 train-test partitions on which the Forward Selection models were generated, with the SAS Compare procedure used to confirm that re-creations were identical to originals. The $\Pr(\text{correct}|i, j)$ were cumulated, stored, and used to generate PPAPPs and ENCCs for both model types so that they could be compared to the Forward Selection models for evidence of differences, especially improvements, in classification performance. Inasmuch as the random realizations of each model type were developed using a different set of variables, the three model types can be interpreted as three different model spaces with a distribution of models in each space.

Results

Frequency of peak selection in Forward Selection models

The 10,000 iterations of resampling and logistic regression were first executed using Forward Selection on all 37 peaks. Anywhere from 3 to 7 peaks were selected into each model. The number of times each peak was selected was tallied and used to estimate a selection rate, with results shown in Fig.4. The five most frequently selected peaks had m/z ratios (selection rates) of 3966.8 (92.79%), 3983.1 (63.55%), 4309.4 (32.76%), 8951.7 (21.03%), and 5592.5 (17.13%), respectively. The noteworthy finding shown in this figure is that, although Forward Selection clearly chose some peaks far

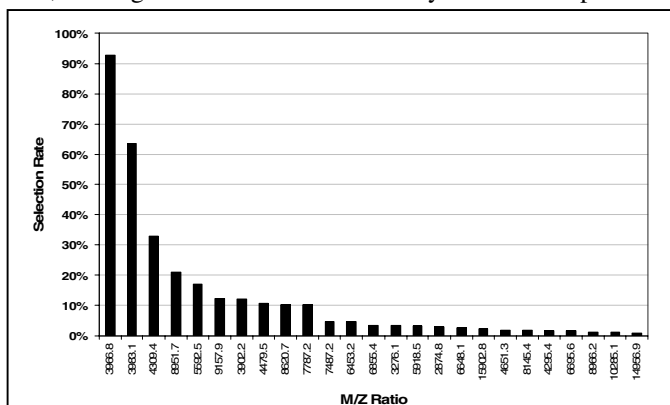


Fig.4: Selection rate of SELDI peaks into logistic-regression Forward Selection models. Peaks are sorted by descending selection rate. Only the 25 most frequently selected peaks are shown, but all 37 peaks had a non-zero selection rate, the minimum being 0.12%, and the maximum being 92.79%.

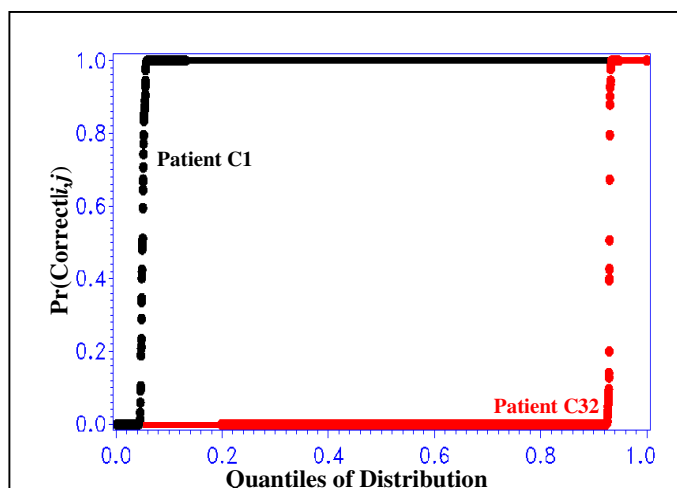


Fig.5: Distribution of test-set prediction probabilities for two samples. Probabilities were generated by logistic regression with Forward Selection on all 37 SELDI peaks. The vast majority of probabilities cluster very near 0 or 1. Probabilities of correct classification averaged to 95% for the sample from Patient C1, compared to 7% for the sample from Patient C32.

more often than others for classification, all peaks were selected at least once (the minimum was 12 times out of 10,000). Moreover, not even the most frequently selected peak was selected 100% of the time. If we equate a peak's selection rate with its importance to a classification model, then the resampling process evidently produced some training sets on which Forward Selection chose unimportant peaks, and other (possibly overlapping) training sets on which the procedure missed important peaks.

Test-set prediction probabilities in Forward Selection models

Test-set prediction probabilities, $\text{Pr}(\text{correct}|i,j)$, for individual samples were plotted against their quantiles for further examination. Two such plots are displayed in Fig.5. The vast majority of probabilities are within 10^{-6} of either zero or one, with only a few per sample assuming intermediate values. For the sample from Patient C1, the majority of probabilities were near 1, implying that this sample was easy to classify correctly in Forward Selection models. In contrast, the majority of probabilities for Patient C32's sample were near 0, implying that this sample was difficult to classify correctly in Forward Selection models. The average of probabilities for each sample is given by the per-patient average prediction probability (PPAPP; see under Methods), which was about 95% for Patient C1, compared to about 7% for Patient C32. Fig.6 shows how the PPAPP can be used to summarize how well or how poorly the Forward Selection models classify particular samples.

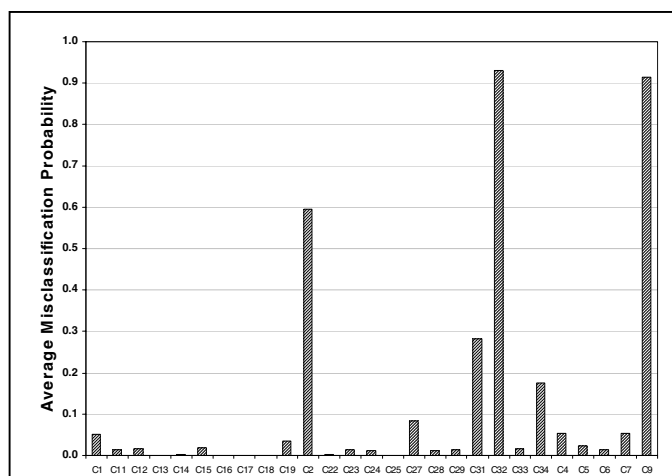


Fig.6: Misclassification rates (i.e., average misclassification probabilities) for samples under Forward Selection, showing that most samples have low misclassification rates, but a few have high misclassification rates. Shown are about half the Cancer samples; samples not shown have similar appearance. Misclassification rates are calculated as One Minus the PPAPP (Per-patient Average Prediction Probability).

ENCCs from Forward Selection models

Every test set contains data from 24 cancer patients and 27 normal patients, for a total of 51 samples to be classified. If, for each iteration j , we add up the test-set prediction probabilities, the result is a number between 0 and 51 that can be interpreted as the Expected Number Correctly Classified (ENCC; see under Methods). In Fig.7, the ENCC values from the Forward Selection models were plotted against the quantiles of their distribution. The "steps" at integer values of the ENCC in Fig.7 appear to result from the almost Bernoulli-like distributions of $\text{Pr}(\text{correct}|i,j)$ shown in Fig.5. The red circle in Fig.7 is drawn around 152 ENCC values that are above 50.5, or less than half a sample below the maximum possible value of 51.

It seemed natural to ask which peaks were selected into these unusually successful Forward Selection models, and whether particular peaks predominated. Accordingly, parameter estimates from the 152 Forward Selection models in question were exported to an Excel spreadsheet for further examination. Fig.8A, a schematic view of the Excel spreadsheet, shows that four peaks were selected overwhelmingly more often than the others into these unusually successful models. Fig.8B (a modified copy of Fig.4) shows that the four peaks in question had the first, second, fourth, and tenth highest selection rates in the Forward Selection model space. Those four peaks had m/z ratios (overall selection rates) of 3966.8 (92.79%), 3983.1 (63.55%), 8951.7 (21.03%), and 7787.2 (10.35%), respectively.

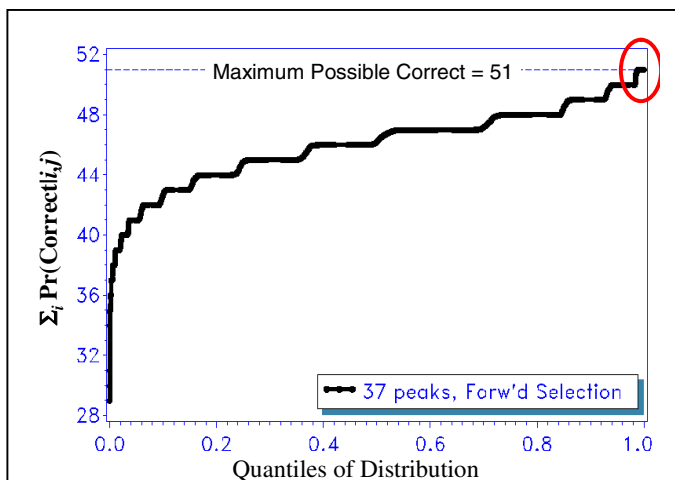


Fig.7: Distribution of Expected Number Correctly Classified (ENCC) by the Forward Selection models. For each iteration j , the ENCC is the sum of probabilities $\text{Pr}(\text{correct}|i,j)$ for patient samples (indexed by i) in the test set. The maximum possible correct (maximum possible ENCC) is equal to the number of test-set samples, which was always 51. The red circle calls attention to 152 Forward Selection models that approached to within 0.5 of the maximum possible correct.

Investigation of other model types suggested by Forward Selection results

Given the results of Fig.4, it was reasonable to ask whether we can improve on Forward Selection by restricting our attention to a reasonable number of the most frequently selected peaks. (To have a sample-to-feature ratio of at least 10, as suggested by Somorjai *et al.* (7), I chose five as a reasonable number based on a training set of 52.) And given the results of Fig.7 and Fig.8, it was equally reasonable to ask whether we can improve on Forward Selection by restricting our attention to the four peaks that predominated in those 152 unusually successful models. To answer these two questions, I regenerated the same 10,000 random partitions into training and test sets, but used for my logistic regressions only (a) the five best peaks of Fig.4, or (b) the four-peak pattern of Fig.8, both without variable selection. The ENCCs from these two model types were plotted against their quantiles along with those from the Forward Selection models, with results shown in Fig.9. Visual inspection of this figure discloses that, at almost all quantiles, the curve from 4-peak-pattern models is higher than that from 5-best-peak models, which in turn is higher than the curve from Forward Selection models. Also of interest was the number of models from each type that produced an ENCC within 0.5 of the maximum, as this denotes an expected value of less than 0.5 misclassification events. As noted earlier, 152 Forward Selection models had ENCCs within 0.5 of the maximum. The 5-best-peak models showed slight improvement, with 217 having ENCCs less than 0.5 below maximum. By contrast, the 4-peak-pattern models showed considerable improvement, with 3,280 having ENCCs within 0.5 of the maximum.

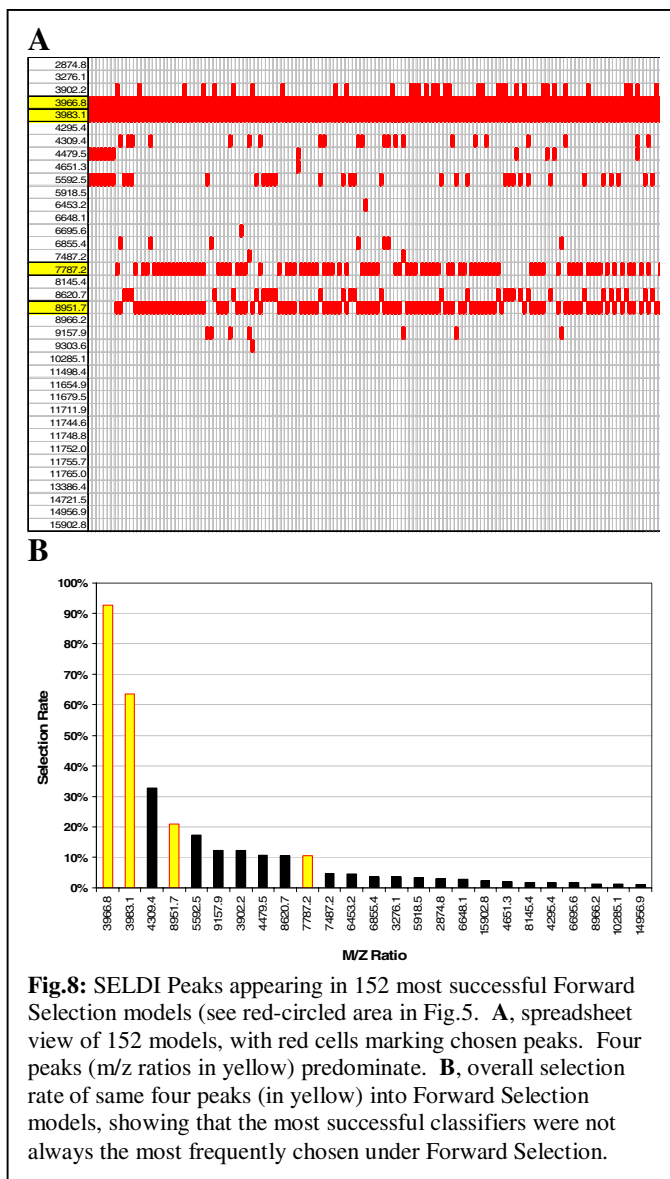


Fig.8: SELDI Peaks appearing in 152 most successful Forward Selection models (see red-circled area in Fig.5. **A**, spreadsheet view of 152 models, with red cells marking chosen peaks. Four peaks (m/z ratios in yellow) predominate. **B**, overall selection rate of same four peaks (in yellow) into Forward Selection models, showing that the most successful classifiers were not always the most frequently chosen under Forward Selection.

To further characterize the difference between the three model types, I computed the samples' PPAPPs from the Forward Selection models, the 5-best-peak models, and the 4-peak-pattern models, and converted the results into ROC curves. Fig.10 shows the region of the plot where the ROC curves for the three model types do not overlay one another. Areas under the ROC curves were 0.9845 for the Forward Selection models, 0.9962 for the 5-best-peak models, and 0.9996 for the 4-peak-pattern models, a rank order in conformance with Fig.9 results.

Discussion

Two major findings resulted from this work. The first finding was that the set of SELDI peaks chosen by Forward Selection was very sensitive to the random choice of samples that were allocated to the training set. Although the number of peaks chosen into each model varied from 3 to 7, all 37 peaks were

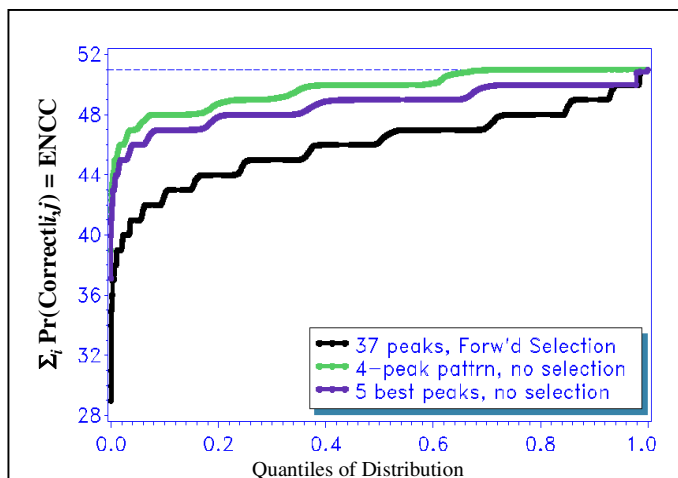


Fig.9: Distributions of ENCC from the three model types. Distribution medians (quartiles) are 46.19 (44.85–48.00) for Forward Selection models, 49.00 (48.00–49.99) for 5-best-peak models, and 50.00 (49.00–>50.99) for 4-peak-pattern models.

chosen at least once, and, although some peaks were chosen rather frequently, no peak was chosen 100% of the time. This result was surprising, but perhaps foreshadowed by the caveats and cautions of Somorjai *et al* (7). It certainly holds pessimistic implications for the reliability of a classifier developed via automated variable selection on a single randomly chosen training set. However, the second major finding was that, by repeatedly applying Forward Selection to a large number of randomly chosen training sets, we not only obtained a distribution of models, we were able to develop a way to search the distribution for models that, at least for this data set, had unusually successful classification rates. And by examining the variables selected into these superior classifiers, we were able to identify a panel of four SELDI peaks that proved, across 10,000 random train-test partitions sets, to possess much better discriminative power on the test sets than either Forward Selection or a second panel of the five SELDI peaks most frequently selected under Forward Selection. It is interesting that the four-peak panel consisted of the 1st, 2nd, 4th, and 10th most frequently selected peaks under Forward Selection: Had I only the results of Fig.4 to guide me, I would not have found this particular combination. Because the four-peak-pattern models tended to classify much better than the other two, we went on in Bhattacharyya *et al.* (5) to develop a multivariate logistic-regression classifier with parameter estimates for the four peaks in question (*m/z* ratios of 3966.8, 3983.1, 8951.7, and 7787.2). That "final model" was trained & tested on a 2:1 random split of the samples, and was done in order to meet anticipated reviewer demands for a classification rule. But while the four-peak panel was robust to the random choice of training sets, the parameter estimates they generated were not. For two of the peaks, in fact, the associated parameter estimates (data summaries not shown) were sometimes positive and sometimes negative, a fact that can make it difficult to decide whether a peak's elevated intensity predicts for or against an increased risk of cancer.

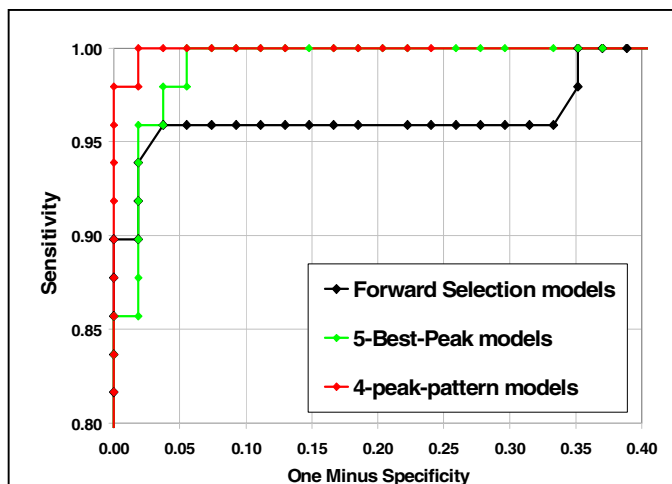


Fig.10: Empirical Receiver Operating Characteristics (ROC) curves for the three model types, developed using the Per-patient Average Prediction Probability (PPAPP) of samples under each model type. Only the region where the ROC curves do not overlay each other is shown. Areas under the ROC curves were 0.9845 for the Forward Selection models, 0.9962 for the 5-best-peak models, and 0.9996 for the 4-peak-pattern models.

The parameter estimates of the logistic-regression classifier in Bhattacharyya *et al.* (5) should therefore be taken with a grain of salt, based as they are on a single random train-test partition. In practice, however, the issue of parameter-estimate reliability is immaterial, because SELDI-TOF technology is too expensive to use directly for diagnostic screening. The real value of the work presented in this paper is that it allowed me and my collaborators (5) to discover a robustly discriminative panel of SELDI peaks that, with purification and sequencing, could lead to a cheap, simple, and reliable ELISA screen for pancreatic cancer.

I found the four-peak panel using the ENCC to search the model spaces for superior classifiers. I found the PPAPP to be far less useful a tool for this purpose. However, the PPAPP had an interesting property that is worth further comment. Fig.6 shows the misclassification rate (or average misclassification probability, equal to One Minus the PPAPP) for about half the cancer samples under Forward Selection. Although most samples had low misclassification rates, a few had high misclassification rates. The other half of the cancer samples, and all of the normal samples, were similar in that most of them also had low rates, but a few had high rates. Each sample's misclassification rate is the average of that sample's misclassification probabilities across the roughly 5,000 test sets in which that sample randomly appeared, which means that it estimates an expected value. That the misclassification rate varies so much from sample to sample, at least under Forward Selection, suggests that it may, in fact, estimate a parameter that characterizes a property of the sample. Statistical classification depends on all the samples having the correct class label, but sometime a sample that appears to be misclassified turns out to have the wrong class label instead. I suggest that the PPAPP and the

misclassification rate derived from it can be used as a tool to look for samples whose class labels need closer scrutiny. Concern can be raised with my use of logistic regression, a maximum-likelihood method, in conjunction with automated variable selection on so many variables. SAS users will know that, when one too many variables are added to the model, the Logistic Procedure will ordinarily terminate and issue a warning that "complete" or "quasicomplete" separation of points has been detected, and that the maximum likelihood may not exist. But complete separation of points denotes zero overlap between classes, which is exactly what we want our classifier to give us. And since I was interested in the variables selected, not their parameter estimates, I was not troubled by the absence of a likelihood maximum. Interestingly, when I decided to weight the average intensities, SAS stopped terminating & warning me of non-existent maxima. Instead, SAS always attained convergence, but always in a region of zero local curvature in at least one dimension. In other words, SAS always stopped on a ridge, not a peak, in the likelihood function. This may explain the wide variation in parameter estimates among the four-peak-pattern models that I noted earlier.

From both the four-peak-pattern models and the five-best-peak models, I have two sets of 10,000 parameter-estimate vectors, each set of which forms a sampling distribution. In one avenue of further research, I will index every vector in both sets with its ENCC, and then ask two questions. The first will be, are the parameter estimates of superior classifiers spread homogeneously throughout their distribution, or are they more concentrated in some regions of the distribution than in others? To set up the second question, we need to think of the two sets of ENCC-indexed parameter-estimate vectors as "parent distributions". For each parent distribution, we can form a "final model" from the centroid of parameter-estimate vectors of only the superior classifiers. If we apply those final models to the 103 samples, we get classification probabilities that are neither cross-validation probabilities nor test-set prediction probabilities.. The second question will be two-fold: (1), what exactly do the final-model classification probabilities represent, and (2) how do they compare to the classification probabilities from their parent distributions?

In another avenue of research, I will apply a different method, such as discriminant analysis, to my 10,000 train-test data partitions, and compare the results to those presented in this paper. Indications from principal-components plots of multivariate non-normality may make it necessary to train kernel-density-based discriminants in addition to linear discriminants once promising panels of SELDI peaks have been identified.

Finally, the methods presented in this paper were developed using one data set. It remains to be seen, of course, how successful my methods will be when applied to additional data sets. Finding out will be my third avenue of further research.

Acknowledgements

I wish to acknowledge the people who were co-authors with me on Bhattacharyya *et al.* (5), namely,

- Sudeepa Bhattacharyya and Larry Suva of the Center for Orthopaedic Research, Department of Orthopaedic Surgery, University of Arkansas for Medical Sciences, Little Rock, AR;
- Gloria M. Petersen and Suresh T. Chari of, respectively, the Divisions of (1) Epidemiology and (2) Gastroenterology & Hepatology, Mayo Clinic, Rochester, MN; and
- Randy S. Haun of the Department of Pathology, University of Arkansas for Medical Sciences, Little Rock, AR.

Without them, I would not have had the opportunity even to think about the problem, let alone develop methods to attack it. I also would like to thank Reid Landes of UAMS Biostatistics for reviewing this paper and suggesting improvements in its organization

References

1. Roboz J. Mass spectrometry in diagnostic oncoproteomics. *Cancer Invest.* 2005;23(5):465-478.
2. Seibert V, Ebert MP, Buschmann T. Advances in clinical cancer proteomics: SELDI-ToF-mass spectrometry and biomarker discovery. *Brief Funct Genomic Proteomic.* 2005 May;4(1):16-26.
3. Wulfkuhle JD, McLean KC, Paweletz CP, Sgroi DC, Trock BJ, Steeg PS, Petricoin EF 3rd. New approaches to proteomic analysis of breast cancer. *Proteomics.* 2001 Oct;1(10):1205-1215.
4. Ardekani AM, Liotta LA, Petricoin EF 3rd. Clinical potential of proteomics in the diagnosis of ovarian cancer. *Expert Rev Mol Diagn.* 2002 Jul;2(4):312-20.
5. Bhattacharyya S, Siegel ER, Peterson GM, Chari ST, Suva LJ, and Haun RS. Diagnosis of Pancreatic Cancer Using Serum Proteomic Profiling. *Neoplasia.* 2004 Sep-Oct;6(5):674-686.
6. Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. *American Journal of Epidemiology* 1981; 114:906-914.
7. Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics.* 2003 Aug 12;19(12):1484-1491.