Enhancing the analysis of mass-spectrometry proteomic profiles using prior knowledge and past data repositories

Milos Hauskrecht and Richard Pelikan
Computer Science Department and Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260
{milos, pelikan}@cs.pitt.edu

Abstract

Whole-sample mass-spectrometry proteomic profiling based on SELDI-TOF-MS technology has lead to many promising results in detection of various types of cancer and other diseases. However, the majority of SELDI-TOF-MS disease studies performed to this day do not attempt to identify protein species responsible for these promising results. The limitation of the protein identification is that it requires secondary lab-based analysis which increases the cost of the study, and that at the end, the identifications may not lead to any new biologically important result. To address the problem, our work focuses on computational approaches that provide early insights on the identity of putative biomarker signals found with a SELDI-TOF-MS instrument. We present two computational methods to achieve this goal: (1) labeling of mass-spectra peaks corresponding to high-abundance protein species, and (2) evaluation of disease-specific signals with the help of profiles from past case/control studies. The key benefit of our methods is that they can provide early characterization of discriminative signals while working directly with whole-sample profiles. As a result, they can be used to filter-out some of the MS signal candidates so that subsequent identification procedures are directed towards analysis of signal species that are likely to yield new information.

1 Introduction

Whole-sample mass-spectrometry (MS) proteomic profiling technology based on surface enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI-TOF-MS) provides an inexpensive assay with a great promise for screening for the presence of a disease, and/or for the assessment of benefits of different patient-management interventions. The potential efficacy of serum protein profiling for cancer classification has been demonstrated on multiple studies, including human breast (Li et al. 2002), colon (Watkins et al. 2001), head and neck (Wadsworth at al. 2004), lung (Zhukov et al. 2003, Xiao et al., 2003), ovarian (Petricoin et al. 2002a, Jones et al. 2003), bladder (Vlahou 2003,2004) and prostate cancer (Wright at al 1999, Adam et al. 2001, Petricoin et al. 2002b). Other promising sample sources include urine (O'Riordan et al. 2004), saliva (Hu et al. 2005) and spinal fluid (Sickmann et al. 2002).

The existence of high-accuracy classification models for whole-sample low-resolution SELDI-TOF-MS profiles, aside from their role in disease detection, is important also for biomarker discovery – the selection of one or a small set of MS profile peaks capable of discriminating between case and control samples. In general, a peak biomarker can be associated with any protein species present in the sample that discriminates well between case and control profiles in the study. However, to fulfill the ultimate goal of discovery proteomics it is critical to assess the identity of the biomarker peak so that follow-up studies and validation of the signal can be

conducted. The protein identification task is typically addressed through secondary lab-based analyses, most often 2D gel electrophoresis followed by the mass-spectrometry TOF-TOF sequencing and identification steps. Unfortunately, this process incurs additional costs and delays, and precious resources may be spent on identification of biomarker signals that are unlikely to bear any fruits. Our hypothesis is that key initial insights on peaks in whole-sample profiles and their potential usefulness as biomarkers can be obtained early and relatively inexpensively from the low-resolution proteomic data with the help of prior knowledge and/or data collected for other case/control studies.

In this work we present and describe two computational methods that let us obtain early insights on potential biomarker-bearing peaks: (1) labeling of peaks corresponding to species with high expected abundance in the sample specimen, and (2) evaluation of potential disease-specific signals with the help of profiles from past case/control studies. Our objective is not to devise procedures that yield complete information about every peak; instead we seek as much characterization as is possible given the limits of the SELDI-TOF-MS technology and available external knowledge. Such a computational analysis can provide a good initial understanding of discriminatory performance of MS case/control profile signals and can be used to steer the application of follow-up protein identification procedures so that these are applied only to signals that are likely to reveal some new information.

2 SELDI-TOF-MS profiling

Surface enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI-TOF-MS), developed by Ciphergen Biosystems, Inc. (Fremont, CA), is used for the mass analysis of compounds such as proteins, peptides and nucleic acids (with masses up to 200 thousand Daltons) within solutions such as serum, urine, or cell lysates. Profiles may be determined for whole sera using whole serum ('neat spotting'), or using fractionated samples, with profiles determined for each sample fraction. SELDI operates by capturing compound(s) of interest on a chip. The surface of the chip possesses affinity characteristics such as ion-exchange, hydrophobicity, or antigen/antibody. Compounds of interest then dock onto the surface through these affinity interactions. Contaminants are removed by washing. After addition to the chip of "energy absorbing molecules" (aka, matrix) the remaining bound substances are analyzed under high vacuum by laser desorption/ionization time of-flight mass spectrometry. The time of flight through the vacuum is converted to provide inferred molecular weight information. An example of a profile is shown in Figure 1.

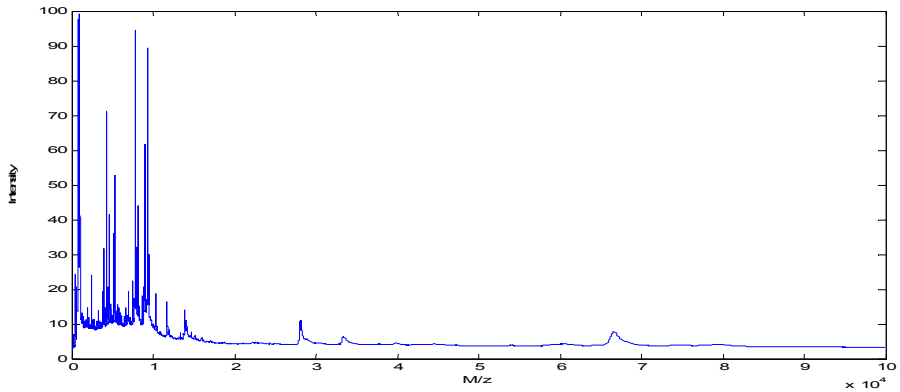

Figure 1: A typical SELDI-TOF-MS profile showing mass to charge [*m/z*] ratio versus relative ion intensity. Note the relative abundance of species below 20,000 Daltons.

Profile data obtained from the low-resolution SELDI-TO-MS instrument are often corrupted and subject to multiple systematic biases and sources of error. These problems are addressed during the preprocessing phase that precedes the interpretive data analysis. Profile preprocessing includes computational steps taken to remove a variety of noise signals in the data with the expectation that most of the useful information content carried by the profiles is preserved. MS profile pre-processing consists of: profile quality control, baseline correction, variance stabilization, normalization, alignment and profile smoothing. The details about these methods can be found in Coombes et al. (2005, 2006) or Hauskrecht et al (2005).

Differential and classification analysis

A typical disease study involves two patient groups: case (disease) group and control (normal) group, and their corresponding MS profiles. The aim of the interpretive data analysis is to identify signals in spectra that carry the differences in between the two groups. Univariate differential analysis (Baldi & Long 2004, Hauskrecht at al 2006) considers individual features (peaks) and their impact on the case/control discrimination, while the multivariate classification analysis (Ball et al 2002, Qu et al 2002, Yasui 2003, Hauskrecht et al 2005) examines the discrimination potential of combination of multiple peaks (peak panels).

Figure 2 shows a statgram built for the pancreatic case/control study (Hauskrecht et al 2005). The differential potential of a peak is measured by the –log of the p value of the Wilcoxon ranksum test. The arrows points to two peak regions with the highest Wilcoxon-based differential score. The differences in the mean case and the mean control profiles for the two groups in these two regions are apparent.

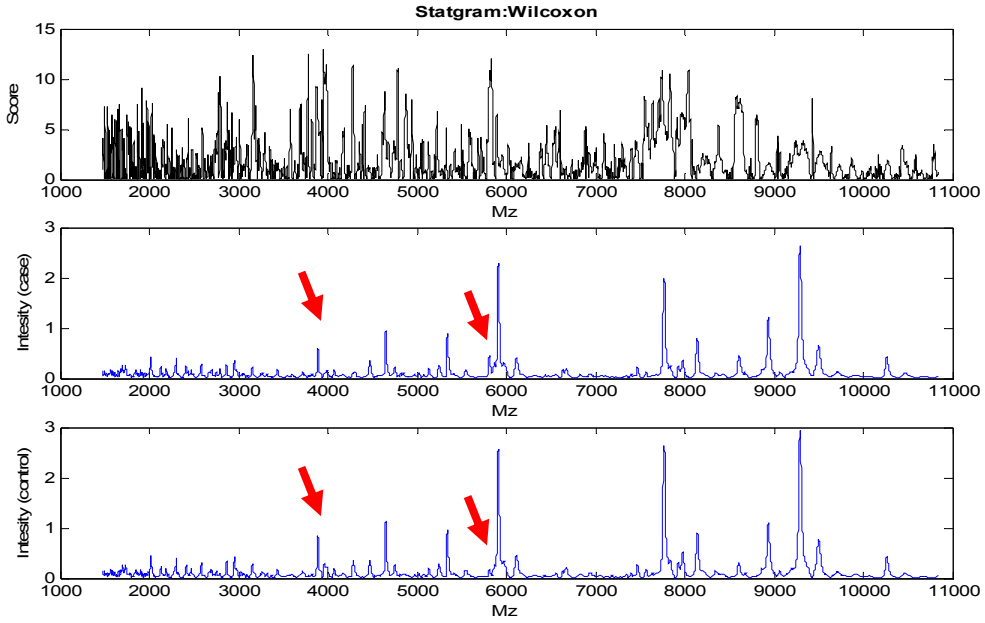

Figure 2. A statgram for the case/control pancreatic cancer study. The differential score is based on the p-value of the Wilcoxon-ranksum test. Peaks with high differential score are labeled by arrows.

A typical result of the differential analysis is a set of discriminative peaks; each peak accompanied by its univariate differential score. In the multivariate classification analysis, a set

of peaks together with predictive classification statistics (accuracy, sensitivity, specificity or area under the ROC curve) the peaks achieve on some classification model (e.g., a support vector machine, CART, etc.) is reported. In both cases, the analysis leads to a set of peaks representing signals with the strongest discriminative power and some quantitative assessment of their discriminative potential.

3 Computational methods for characterization of discriminative peaks in the low-resolution whole-sample spectra

The interpretive data analysis lets us assess the potential of individual peaks for differentiating case/control subjects on the level of MS profile signals. However, this analysis does not provide any insight about the species that stand behind these signals. To fulfill the ultimate goal of discovery proteomics, the discriminative signal species suggested by the study need to be identified. The objective of our work is to develop computational methods that can provide early identification and characterization of MS profile signals and this without the need for secondary protein identification analyses. Such a characterization can provide an early understanding of the discriminative potential achieved by profiles in a specific study, or can at least help us to narrow down the number of discriminative signals that should be subjected to a more detailed secondary protein identification interrogation and analysis.

We focus on two computational solutions to achieve our goal:

- labeling of peaks corresponding to highly abundant protein species (Section 3.1);
- evaluation of the specificity of discriminative peaks with the help of profiles collected for other disease studies (Section 3.2).

In the first case, we hope we can reliably identify some of the peaks in the MS profiles directly by relying on the knowledge of expected protein abundances in the sample. In the second case, our hope is that other disease studies may help to reveal more information about the nature of potential biomarker signals.

3.1. Peak-labeling for low-resolution whole-sample proteomic profiling

The goal of peak labeling is to annotate peaks in the spectra with corresponding proteins. The general peak labeling problem is complicated by the existence of multiple protein species with the same or close to the same mass, which are, given the MS instrument precision, very hard to distinguish. In addition, the ionization process may cause proteins to take on double or triple charge causing the signal of the same species to be recorded multiple times at different mass to charge positions (one half, one third of the original mass to charge value respectively) increasing possible signature overlaps. The overlap problem is further complicated by the fact that a pure (unmodified) protein may undergo possible post-translational modifications, each of which leads to a shift of the protein signature (peak) in the spectra. Due to the large number of possible signature overlaps, an attempt to consider all possible protein species and their modifications as peak candidates does not to lead to any feasible peak-labeling solution.

The low-resolution MS profiling technology is limited in terms of the number of species it can reliably detect and measure. The number of measurements and peaks defining a typical MS profile gives us very a rough upper limit on the number of species we can directly observe. The total number of measurements for the SELDI-TOF-MS technology is in the neighborhood of 65,000 measurements. Many of these measurements are aggregated into peaks that tend to refer to the same underlying signal. Considering only profile peaks, the number of species we can detect

counts at most few thousands. For example, the feasible region for detection in SELDI-TOF-MS profiles (1,500-40,000 Daltons) yields approximately 3000 peaks. This number is significantly smaller than the number of proteins in the Swiss-PROT or other proteomic databases, and we did not even count multiply charged ions and possible protein modifications. Given this, it is unrealistic to believe that every possible species with a specific molecular mass can be seen or detected by the MS technology. Some of the species may not bind well to the matrix surface and they are washed away when the sample is processed and rinsed. But most of all, many species occur in the sample in such a low concentration that their detection becomes very unlikely. At the end, it is the higher abundance species in the sample we are more likely to detect in the whole-sample profile.

The species and their abundance in the sample depend on the specimen analyzed in the study. Naturally, more is known about common sample media (serum, plasma) and their expected composition in terms of their high abundant species (Anderson & Anderson 2004). Such a prior knowledge can be used to select a set of protein candidates and their modifications we expect are more likely to be seen in the sample. We assume this set is provided as an input to our peak-labeling procedure.

Peak-labeling procedure

The goal of peak-labeling is to annotate peaks by assigning protein labels to peaks (Figure 3). We assume we have finite set of protein labels $L=\{l_1, l_2, ..., l_k\}$ that make sense in context of the specimen and instrument sensitivity. We assume a profile is defined by a set of peaks $Q=\{ q_1, q_2, ..., q_m\}$ where each peak i is defined by its mass and intensity components (x_i, y_i). The protein labels are used to annotate peaks in the profile. We assume the following restrictions on the peak labeling procedure: a protein label cannot be assigned to more than one peak, and peak labels must be assigned to peaks in order of their expected mass. Given these restrictions the peak labeling problem can be viewed as the problem of assigning peaks to different protein labels (Figure 3). We note that it is not necessary that all peak labels are assigned a peak, and of course, many peaks may remain unassigned.

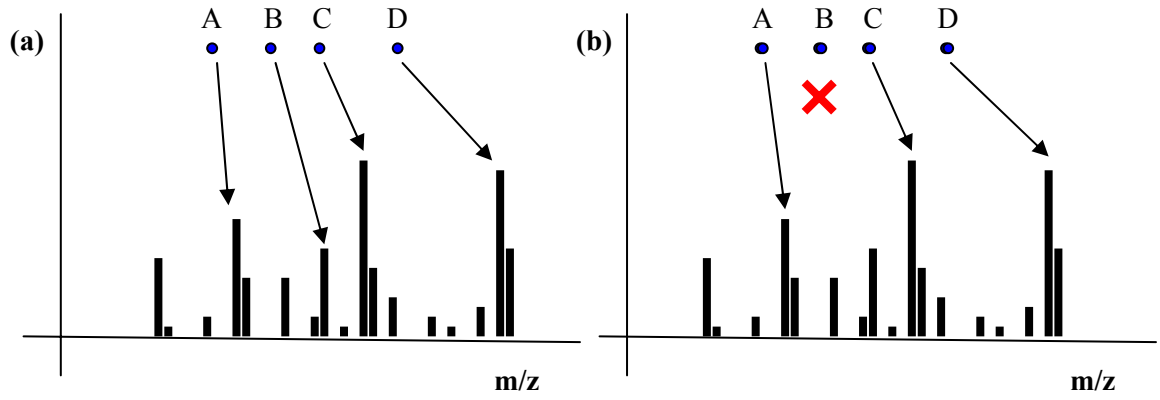

Figure 3. The basics of the peak-labeling procedure. (a) A protein label (A,B,C,or D in the figures) is assigned to at most one MS signal peak. The assignment must respect the expected mass of the protein species used. (b) Not all peak labels may receive a peak.

As in general, there are more ways of assigning peaks to labels, we need a model that lets us measure the quality of each individual labeling. The best labeling is then achieved by optimizing this measure. In (Pelikan & Hauskrecht 2007) we proposed a new probabilistic model that

measures the quality of the peak-labeling by considering both the location and the intensity components of the peak. The novelty of the approach stems from the incorporation of the peak intensity component into the model. The expected benefit of such a model is its improved peak labeling performance over more traditional peak-location-only approaches. Intuitively, if we consider an assignment of a label to a peak we need to consider not only its expected location, but also its expected intensity. This may prevent mislabeling of peaks due to possible imprecision in peak's m/z position and the fact that more than one peak may provide a match for a protein species.

The model

Our probabilistic model consists of three key components (Pelikan & Hauskrecht 2007):

- *Detection component* defining the probability a protein species is detectable as a peak in the profile;
- *Peak-location component* defining the probability a protein species is detected at a specific m/z location;
- *Peak-intensity component* defining the probability of relative intensity measurements for protein species considered by the peak labeling procedure.

These three components let us define the joint distribution:

$$p(l_1 = (x_1, y_1), l_2 = (x_2, y_2), \ldots l_k = (x_k, y_k))$$

of protein labels being assigned to peaks with specific location and intensity measurements. A special *null* value is used to denote an "empty" assignment representing the case in which the species is not detected in the profile.

Probability of detection. The detection component of the model is represented by the probability $p_{0,i}$ a protein species l_i is detectable as a peak in the profile Our model assumes the chance of k protein species showing up/or not showing as peaks in the profiles are independent of each other. For example, the probability that none of the k species is detected in the profile is

$$p(l_1 = null, l_2 = null, \ldots l_k = null) \sim \prod_{i=1}^{k} (1 - p_{0,i})$$

Peak-location component. A peak corresponding to a protein species may not be recorded exactly at its expected mass-to-charge location; instead it may be recorded in its close neighborhood. This may lead to a situation in which more than one peak may be associated with a protein. We assess the quality of the protein-location match using a Gaussian density model:

$$p(l_i^{(x)} = x_i) \sim \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(- \frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right)$$

where μ_i is the expected location of protein's peak and σ_i defines the variance of the location due to the instrument precision. The joint probability of multiple protein-location matches is then defined as:

$$p(l_1^{(x)} = x_1, l_2^{(x)} = x_2, \ldots l_k^{(x)} = x_k) \sim \prod_{i=1}^{k} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(- \frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right) \qquad (1)$$

Peak-intensity component. The fact that there can be more than one peak location match for a specific protein may lead to many incorrect peak labelings. To improve the accuracy of detection we enhance the model with the peak intensity information and the knowledge of expected protein composition and their relative abundance in the sample. Intuitively, if the species A is expected to be more abundant than B in the sample, we expect the intensity of the peak for A (integral peak intensity) to be is higher than the intensity for the peak B. We incorporate the knowledge of expected relative abundance through a Dirichlet distribution model. More specifically, the probability of a specific protein-to-peak-intensity assignment is:

$$p(l_1^{(y)} = y_1, l_2^{(y)} = y_2, \ldots l_k^{(y)} = y_k) \sim Dir(\tilde{\mathbf{y}} \mid \alpha_1, \alpha_2, \ldots \alpha_k) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \tilde{y}_i^{(\alpha_i - 1)} \qquad (2)$$

where $\alpha_1, \ldots \alpha_k$ are hyperparameters of the Dirichlet distribution and

$$\tilde{y}_i = \frac{y_i}{\sum_{j=1}^{k} y_i} \qquad (3)$$

represent relative abundances observed in the specific protein-peak assignment. Figure 4 illustrates the nature of the problem. We need to assess the quality of each protein-to-peak assignment in terms of its intensity measurements while relying on the knowledge of the expected abundances of protein species in the sample.

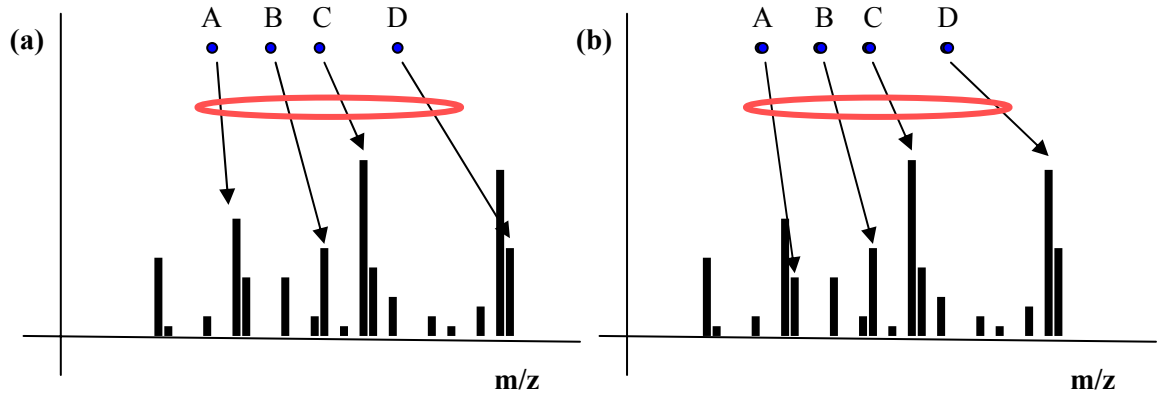

Figure 4. Relative abundance component of the peak-labeling model. All assignments of protein labels to peaks (two assignments are shown) are covered by the Dirichlet distribution model built from knowledge of expected relative abundances.

Given the three model components, the joint probability of all (in-order) protein label assignments is:

$$p(l_1 = (x_1, y_1), l_2 = (x_2, y_2), \ldots l_k = (x_k, y_k)) \sim \prod_{i=1}^{k} p_{0,i} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right) \tilde{y}_i^{(\alpha_i - 1)} \qquad (4)$$

while all out of order assignments are explicitly enforced to 0.

Since a proteomic profile consists of many peaks and only some of them are assigned to protein labels we assume the probability of all profiles consistent with the assignment is uniform and the probability of all inconsistent profiles is 0. This rule defines the conditional probability:

$$p(Q \mid l_1 = (x_1, y_1), l_2 = (x_2, y_2), \ldots l_k = (x_k, y_k))$$

of a profile Q.

Peak-labeling optimization procedure

Having defined a probabilistic model, our goal is to identify the protein-to-peak assignment L^* with the highest posterior for the observed profile Q:

$$\begin{aligned} L^* &= \arg\max_L p(l_1 = (x_1, y_1), , \ldots l_k = (x_k, y_k) \mid Q) \\ &= \arg\max_L p(l_1 = (x_1, y_1), \ldots l_k = (x_k, y_k) \mid Q)\, p(l_1 = (x_1, y_1), \ldots l_k = (x_k, y_k)) \end{aligned} \qquad (5)$$

Since the conditional probability $p(Q \mid l_1 = (x_1, y_1), \ldots l_k = (x_k, y_k))$ is uniform for all profiles respecting the labeling, it is sufficient to optimize $p(l_1 = (x_1, y_1), \ldots l_k = (x_k, y_k))$ that is consistent with the observed profile. The advantage of this optimization criterion is that it is *nearly-decomposable* along individual protein labels (Equation 4). Note that if the score was fully decomposable along protein species, we would be able to solve the optimization problem through the dynamic programming approach. The full decomposability of the probabilistic score in Equation 4 is, however, limited by global dependencies of the peak intensity component (Equation 3). To overcome this problem, we approximate the components used in the calculation of the normalization constant through an iterative heuristic optimization procedure. The procedure starts with an initial (heuristic) assignment of peak intensities to labels. These estimates are then fixed which makes the score decomposable and amenable to the dynamic programming solution. Once the assignment of proteins-to-peaks is found the initial intensity assignment is corrected and the new dynamic programming optimization is performed. This is repeated till no new intensity assignments are found in the previous step.

Experiments

To test our method we used data generated from the virtual MALDI-TOF mass spectrometer proposed by Morris et al. (2005). A set of 100 simulated spectra was generated with 16 controlled spiked-in peptides. The relative concentrations of these peptides were chosen arbitrarily and retained as information to be used by the identification procedure. Our task was to label peaks in the spectra correctly (true positive), while avoiding labeling peaks which may appear as a result of noise (false positive).

To assess the quality of our peak-labeling method we evaluated its precision-recall (PR) curve (Davis & Goodrich 2006) and related statistics. In our task, precision refers to the fraction of label-assigned peaks which are matched to the correct label, while recall refers to the fraction of the 16 labels which were correctly assigned to a peak. We have tested two versions of our peak-labeling method: a baseline version that relies only on the expected mass of the species and our improved version that combines the knowledge of the expected mass together with their abundance information. The PR curves for the two methods were obtained by varying the parameters of the model and are shown in Figure 5.

The PR curve for the peak-location method is completely dominated by the PR curve for the abundance-enhanced method. . The area under the method's PR curves (AUC) for the peak location method is 0.57877, while the abundance-enhanced method achieves an AUC of 0.73866. Similarly, the maximum *F*-measure, obtained by the method using only peak information was 0.52257, versus 0.66667 when including relative abundance information. These results show the contribution of relative abundance information greatly improves the accuracy of the peak-labeling method.

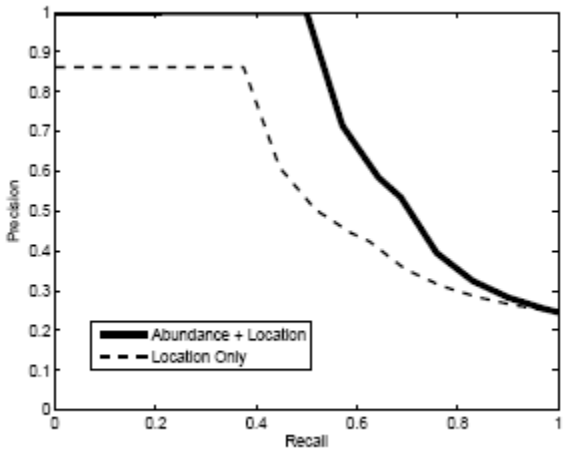

Figure 5. Precision recall (PR) curves for two different peak-labeling methods: the solution based on the location component only and the solution that takes into account both the location and the abundance information. The curves demonstrate the improved performance of the more informed method. The area under the PR curves (AUC) for the peak location method is 0.57877, while the abundance-enhanced method achieves an AUC of 0.73866.

3.2. Analysis of discriminative peaks using data from past case-control studies

The construction of bioinformatics data repositories has been advocated for a long time and there are ongoing efforts to store and share the bioinformatics data among researchers from variety of platforms and studies. Data repositories can be used to combine samples from multiple studies that target the same disease and use them to achieve a higher power of the analysis. Our computational approach is slightly different: we use the data generated from studies for other diseases to evaluate the specificity of putative MS biomarker signals for the study at hand. The approach does not attempt to identity discriminative signals; instead it seeks to answer the following questions: Is the discriminative peak present in other studies? Does it carry any discriminative information also for other (related) diseases? How specific is the signal?

To demonstrate the potential of past data repositories for understanding of existing discriminative signals consider two studies performed on two types of cancer: the lung and the pancreatic cancer. The data in these studies can be analyzed individually and a set of putative biomarker signals can be identified for both of them. However, when analyzed individually, it is not really clear how important these biomarker candidates are for the detection of each disease. The main concern is that a discriminative signal identified in a study data may reflect differences in between healthy subjects and patients suffering from a larger group of diseases, and hence may not define a biomarker suitable for the detection of the target disease only. For example, analyses of samples for many disease studies may reveal increased levels of Serum amyloid A (SAA) and its corresponding peak signature when compared to healthy controls. Such a peak, though promising from the viewpoint of case/control discriminability, may not lead to a good disease-specific biomarker.

Figure 6 shows the results and benefits of the combined analysis on two cancer studies. The analysis is performed using univariate differential score based on the p-value of the Wilcoxon-ranksum test. The smaller values reflect better discriminability of case and control profiles. The results show that two peaks in the vicinity of 9500 Daltons have a good discriminatory potential for both the lung and the pancreatic cancer. Hence these are likely to represent cancer biomarkers but they are unlikely to provide any strong evidence with respect to the lung or the pancreatic disease. On the other hand, a large peak in the neighborhood of 9300 Daltons appears to provide a very good discriminative potential for the pancreatic cancer, but not for the lung cancer. Hence, it is a good pancreatic cancer specific signal candidate. This simple example demonstrates the possibilities of computational analysis of the MS proteomic signals for multiple disease studies and insights the analysis may offer.

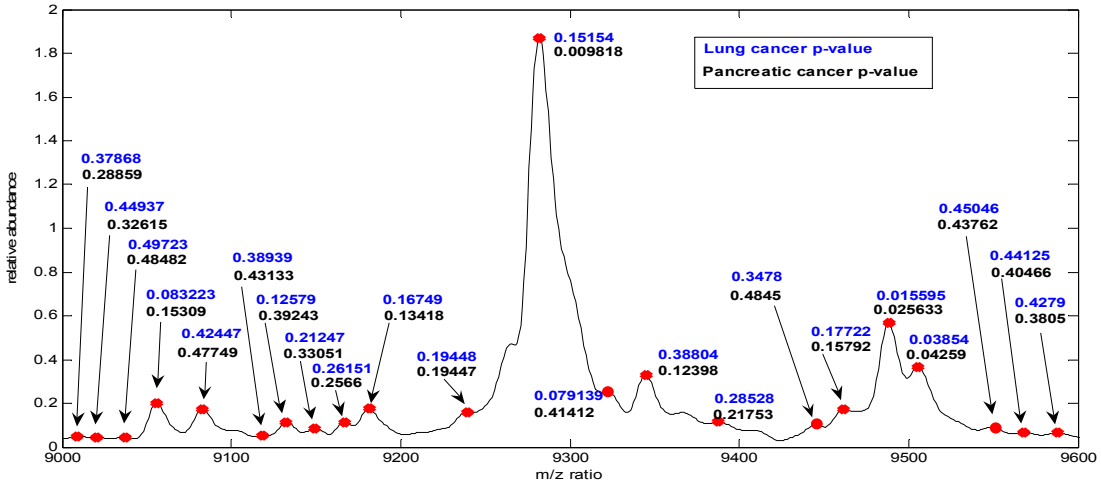

Figure 6. Comparison of differential p-values for peaks identified in the lung cancer dataset with the data from the pancreatic cancer study. Two peaks in the vicinity of 9500 Daltons exhibit a very good discrimination for both types of cancer, while the peak in the vicinity of 9300 Daltons is a very good discriminator for the pancreatic cancer data.

4 Conclusions

We have described two computational methods for post-interpretive analysis of whole-sample MS proteomic profiles for case control studies that is aimed to illuminate the nature of underlying discriminative signals with the help of external knowledge and data obtained for other disease studies.

Our peak-labeling procedure relies on information about the mass of highly abundant protein species which are expected to be found in a particular sample medium. The probabilistic model that supports the peak-labeling procedure comes with a number of parameters that need to be estimated from external knowledge sources including protein databases and literature. This process is not always straightforward and may require additional auxiliary calculations. For example, the mass-to-charge location of a protein can be estimated from the known protein sequences by summing up the average isotopic amino acid residue weights for the sequence. The double and triple charge peak locations can be identified from the parent protein and the signatures are expected to be seen at one half and one third of their mass on the mass-to-charge axis (Morris et al. 2005). Known protein modifications and their sequence information can be extracted from the protein databases and used to calculate a modified peak location using the same approach as applied for the parent protein.

The parameters of the detection and the relative abundance components of our model cannot be mined in the proteomic databases and must be estimated from the literature. The relative abundance component builds upon the knowledge of protein composition of the specific sample medium. For example, the paper by Anderson & Anderson (2004) gives a list of the top 70 high-abundant protein components occurring in the plasma together with their expected concentrations. Such knowledge can be directly translated into our relative abundance model. Efforts to characterize proteomes of other specimen include urine (Rassmusen et al. 1996) and saliva (Hu et al. 2005). Finally, not all high abundant proteins in the sample may be detectable by the instrument. For example, some of the species may not bind well to the surface of the chip. Our preliminary experiments on spiked-in proteomic spectra (Pelikan & Hauskrecht 2007) were performed assuming a fixed prior probability of occurrence of all species and their modifications that are most abundant in the plasma (based on the Anderson & Anderson 2004 study). However, we expect these probabilities may be further refined as we gain and incorporate more knowledge into the procedure. We expect the knowledge of affinity of chip surfaces to proteins, expected proportions of singly and multiply charged ions, or expected proportions of parent and modified proteins in the sample will greatly enhance our ability to identify the protein species.

Acknowledgements

This work was supported by Department of Defense grant USAMRAA W81XWH-05-2-0066, National Library of Medicine grant 5 T15 LM007059-20 to the Pittsburgh Biomedical Informatics Training Program and NCI grant P50 CA090440-06.

References

Adam, BL, Qu, Y. Serum Protein Fingerprinting Coupled with a Pattern-matching Algorithm Distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men. In *Cancer Research Journal*, volume 62, pages 3609-3614, July 1, 2002.

Adam BL, Vlahou A, Semmes OJ, Wright GL Jr. Proteomic approaches to biomarker discovery in prostate and bladder cancers. *Proteomics*. 1:1264-70, 2001.

Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects, *Molecular Cell Proteomics*, vol. 1, no. 11, pp. 845–867, Nov 2002.

Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF Protein Patterns in Serum: Comparing Data Sets from Different Experiments. *Bioinformatics*, 20(5):777-85, 2004.

Baldi, P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized *t* -test and statistical inferences of gene changes. *Bioinformatics,* Vol. 17 no. 6, pp. 509-519, 2001

Ball G, Mian S, Holding F, Allibone RO, Lowe J, Ali S, Li G, McCardle S, Ellis IO, Creaser C, Rees RC. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumors and rapid identification of potential biomarkers. *Bioinformatics*. 18(3):395-404, 2002.

Coombes KR, Baggerly KA, and Morris JS. Pre-Processing Mass Spectrometry Data. *Fundamentals of Data Mining in Genomics and Proteomics,* W Dubitzky, M Granzow, and D Berrar, eds. Springer, 79-99, 2006.

Davis J, Goodrich M, The relationship between Precision-Recall and ROC curves, *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

Hauskrecht, M., Pelikan, R., Malehorn, D. E., Bigbee, W. L., Lotze, M. T., Zeh, H. J., Whitcomb, D. C., and Lyons-Weiler, J. Feature selection for classification of SELDI-TOF-MS proteomic profiles. *Applied Bioinformatics*, 4(4):227-246, 2005.

Hauskrecht M, Pelikan R, Valko M, Lyons-Weiler J. Feature Selection and Dimensionality Reduction in Genomics and Proteomics. In *Fundamentals of Data Mining in Genomics and Proteomics*, eds. Berrar, Dubitzky, Granzow. Springer, Fall 2006.

Hu S, Xie Y, Ramachandran P, Ogorzalek Loo RR, Li Y, Loo JA, Wong DT. Large-scale identification of proteins in human salivary proteome by liquid chromatography/mass spectrometry and two-dimensional gel electrophoresis-mass spectrometry. *Proteomics.* Apr;5(6):1714-28, 2005.

Jones MB, Krutzsch H, Shu H, Zhao Y, Liotta LA, Kohn EC, Petricoin EF 3rd. Proteomic analysis and identification of new biomarkers and therapeutic targets for invasive ovarian cancer. Proteomics. 2:76-84, 2002.

Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem*. 48:1296-304, 2002.

Morris JS, Coombes KR, Koomen J, Baggerly KA, and Kobayashi R, Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum, *Bioinformatics*, vol. 21, no. 9, pp. 1764–1775, 2005.

O'Riordan E, Orlova TN, Mei J J, Butt K, Chander PM, Rahman S, Mya M, Hu R, Momin J, Eng EW, Hampel DJ, Hartman B, Kretzler M, Delaney V, Goligorsky MS. Bioinformatic analysis of the urine proteome of acute allograft rejection. *J American Soc Nephrology*, Dec;15(12):3240-8, 2004.

Pelikan R, Hauskrecht M: In-silico protein identification methods for whole-sample proteomics. submitted to *IEEE Transactions on Computational Biology and Bioinformatics,* 2007

Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA. Use of proteomic patterns in serum to identify ovarian cancer. Lancet. 2002 359:572-7, 2002a.

Petricoin E, Ornstein DK. Serum Proteomic Patterns for Detection of Prostate Cancer. *Journal of the National Cancer Institute,* Vol. 94, No. 20, 2002b.

Qu Y, Adam BL, Yasui Y, Ward MD, Cazares LH, Schellhammer PF, Feng Z, Semmes OJ, Wright GL Jr. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clinical Chem.* 48:1835-43, 2002.

Rasmussen HH, Orntoft T, Wolf H, Celis JE. Towards a comprehensive database of proteins from the urine of patients with bladder cancer," *J Urol*, vol. 155, no. 6, pp. 2113–2119, Jun 1996.

Sickmann A, Dormeyer W, Wortelkamp S , Woitalla D, Kuhn W, and Meyer HE. Towards a high resolution separation of human cerebrospinal fluid. *J Chromatogr B Analyt Technol Biomed LifeSci*, 771(1-2): 167-196, May 2002.

Yasui Y., Pepe M., Thompson M.L., Adam B.-L., Wright G.L., Qu, Y., Potter J.D., Winget M., Thornquist M., Feng Z. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, 4, 449-463, 2003.

Vlahou A, Giannopoulos A, Gregory BW, Manousakas T, Kondylis FI, Wilson LL, Schellhammer PF, Wright GL Jr, Semmes OJ. Protein profiling in urine for the diagnosis of bladder cancer. *Clinical Chemistry*, Vol 50, No. 8, pages 1438-41, 2004.

Wadsworth JT, Somers K, Stack B, Cazares L, Malik G, Adam B-L, Wright GL, Semmes OJ. Identification of Patients With Head and Neck Cancer Using Serum Protein Profiles. In *Archives of Otolaryngol Head and Neck Surgery*, Vol. 130, 2004.

Watkins B, Szaro R, Ball S, Knubovets T, Briggman J, Hlavaty JJ, Kusinitz F, Stieg A, Wu Y-J. Detection of Early Stage Cancer by Serum Protein Analysis. In *American Laboratory,* 2001.

Wright, GW Jr, Cazares LH, Leung SM, Nasim S, Adam BL, Yip TT, Schellhammer PF, Gong L, Vlahou A. Proteinchip(R) surface enhanced laser desorption/ionization (SELDI) mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. Prostate Cancer Prostatic Dis. Dec 2(5/6):264-276, 1999.

Xiao XY, Tang Y, Wei XP, He DC. A preliminary analysis of non-small cell lung cancer biomarkers in serum. *Biomed Environ Sci.* 16:140-8, 2003.

Zhukov TA, Johanson RA, Cantor AB, Clark RA, Tockman MS. Discovery of distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. *Lung Cancer.* 40:267-79, 2003.