BIOINFORMATICS

Vol. 00 no. 00 2006
Pages 1–7

Difference Detection in LC-MS Data for Protein Biomarker Discovery

Jennifer Listgarten [a], Radford M. Neal [a,b], Sam T. Roweis [a], Peter Wong [c] and Andrew Emili [c,d]

[a]Department of Computer Science, [b]Department of Statistics, [c]Banting and Best Department of Medical Research, [d]Program in Proteomics and Bioinformatics, University of Toronto, Toronto, Ontario, M5S 3G4, Canada

ABSTRACT

Motivation: There is a pressing need for improved proteomic screening methods allowing for earlier diagnosis of disease, systematic monitoring of physiological responses, and the uncovering of fundamental mechanisms of drug action. The combined platform of LC-MS (Liquid-Chromatography-Mass-Spectrometry) has shown promise in moving toward a solution in these areas. In this paper we present a technique for discovering differences in protein signal between two classes of samples of LC-MS serum proteomic data without use of tandem mass spectrometry, gels, or labeling. This method works on data from a lower-precision MS instrument, the type routinely used by and available to the community at large today. We test our technique on a controlled (spike-in) but realistic (serum biomarker discovery) experiment which is therefore verifiable. We also develop a new method for helping to assess the difficulty of a given spike-in problem. Lastly, we show that the problem of class prediction, sometimes mistaken as a solution to biomarker discovery, is actually a much simpler problem.

Results: Using precision-recall curves with experimentally extracted ground truth, we show that i) our technique has good performance using 7 replicates from each class, ii) performance degrades with decreasing number of replicates, iii) the signal that we are teasing out is not trivially available (*i.e.*, the differences are not so large that the task is easy). Lastly, we easily obtain perfect classification results for data in which the problem of extracting differences does not produce absolutely perfect results. This emphasizes the different nature of the two problems and also their relative difficulties.

Availability: Our data is publicly available as a benchmark for further studies of this nature, along with supplementary information, at http://www.cs.toronto.edu/∼jenn/LCMS.

Contact: jenn@cs.toronto.edu

1 INTRODUCTION

Much proteomic work to date has concentrated on the use of tandem mass spectrometry (also written MS/MS) in which the sequences of some peptides in serum are found, and analysis is based on a list of peptides [3]. While such an approach has been successful for problems such as building protein catalogues, this approach has shortcomings when applied to biomarker discovery, in which the goal is to find what differences exist between two classes of samples (*e.g.*, cancer versus healthy). We refer to this as the problem of *difference detection*. Since the list of peptides derived from tandem mass spectrometry experiments is never complete for complex mixtures, there are likely to always be protein signals of interest that are missed [1, 9]. One can avoid this problem by looking at all of the raw data from the LC-MS experiment rather than just the list of sequenced peptides, and without reliance on concurrent MS/MS sequencing. Once regions (time, *m/z*) that are different have been identified, these can be characterized by MS/MS [13]. While the present paper seeks to provide a statistical/computational technique for biomarker discovery from LC-MS data, such a step is but one of many in the grand goal of true biomarker discovery, which must address issues ranging from sample collection to a feedback loop with basic biology for validation. For more discussion of these issues we refer to reader to [4, 7] and references therein.

Our goals in this paper are to show that in a realistic experiment using human blood serum, with a controlled spike-in of known peptides, we can elicit signals of interest between two classes, *relative to a known ground truth*, using data from a low-precision mass spectrometry instrument. We show that our ability to achieve this is affected by number of replicates available. The difficulty of our spike-in difference detection problem is assessed to ensure non-triviality of the problem. We also demonstrate that the problem of class prediction is easier than that of difference detection. Lastly, we provide a benchmark data set to the community for problems of this nature; such a data set is currently unavailable, to our knowledge.

Related Work

Recently, several approaches to the specific problem of difference detection between two classes of LC-MS samples, without use of MS/MS or chemical/isotopic labeling, have been published. These approaches typically involve a suite of algorithms, starting from data pre-processing such as filtering, background-subtraction and alignment along the LC time axis, and then move on to one of two approaches i) detect and quantify peaks to then do a differential peak analysis (*e.g.*, [1, 11]), or ii) do a peak-free, 'signal-based', differential analysis to find regions of interest that can then be further studied (*e.g.*, [13, 9]), as we do here. In the former approach, peak detection is carried out on one LC-MS run at a time, without the advantage of leveraging across samples, so features could be lost that might be captured by a signal-based approach, since the latter need not first capture discrete features within each LC-MS run [13, 5]. The relative merits of each approach are likely closely linked to the precision of the MS instrument being used. In [1, 11], high-precision instruments are used with a peak-based approach, while [13, 9], with lower-precision data, use a signal-based approach.

Prakash *et al.* [9] use a signal-based approach to the problem of looking for one spike-in peptide added to a base mixture of 4 peptides – a very simple mixture relative to a more realistic setting involving human blood serum. With just their simple mixture of five proteins, Prakash *et al.* note that even "these mixtures are surprisingly complex" and that "thousands of peptide-like features are observed". They align their data, and nicely characterize their scoring function which quantifies the number of shared peaks between mass spectra, but they stop short of using any statistical tests when doing difference detection, relying on differences in ion amplitude (shown in [13] to be sub-optimal), and they do not quantify their difference detection results, using instead qualitative plots. America *et al.* [1] use a peak-based approach with 2D LC-MS to find differences between tomatoes at various stages of ripening. While their suite of algorithms appears to be very solid, they do not show how well they do relative to any ground truth. A very comprehensive study was conducted in [13] where several different spike-in mixtures, consisting of tryptic digests of roughly 8 proteins, were added to a base mixture of 1000 known tryptic peptides, and differences detected. ROC curves were used to evaluate detected differences with respect to ground truth, but it is not clear how their analysis would generalize to human serum, which we tackle here. Silva *et al.* perform a detailed exploration of spike-in proteins to human blood serum using data from a very high-precision MS instrument [11]. In this study, we show the ability to detect known spike-in proteins in human serum with data from a much lower-precision instrument – the more common 'workhorse' type mass-spectrometer that is widely available and affordable to the proteomics community at present, making our work of immediate and widespread utility.

2 APPROACH TO DETECTION

Our approach to difference detection relies heavily on our previous work on time alignment and normalization [6], in which we introduce the CPM (Continuous Profile Model). Alignment in LC time is one of the major obstacles to reliably detecting differences [1, 6, 9] and thus this step is critical. The work presented here extends our earlier work by showing that the CPM can form the basis for reliable difference detection in LC-MS data.

Alignment in LC-Time

The CPM, based on an HMM (Hidden Markov Model), performs multiple alignment and normalization of time series data such as LC-MS data. One can think of the HMM in the CPM as containing a series of hidden states, each of which represents some underlying 'true' time, to which each observed time point in each LC-MS run is ultimately assigned. The alignment in time is dictated by which observed time gets probabilistically mapped to which hidden state. The states are called 'hidden' because before the algorithm is run on the data we do not know which observed time points map to which states. In addition to the hidden time states mentioned, hidden states are also augmented by 'scale' states, which allow scaling of the signal locally in time. This mapping to both time and scale states performs alignment and normalization concurrently. Training, during which the best parameters for the HMM are found, is performed by maximum likelihood using the Expectation-Maximization algorithm. Both training and later use of the model (*i.e.*, alignment extraction) are performed efficiently in HMMs by use of dynamic programming. The CPM has the advantages that no template is required, all experiments are aligned simultaneously (thereby leveraging information across all experiments), normalization is concurrent with alignment, and the model is probabilistically formulated, making it amenable to principled extensions (*e.g.*, multi-class alignment in which it is not assumed that all samples are from the same class during alignment.)

Whereas in [6] only the TIC (Total Ion Count) were used for alignment, we have since extended the model to use multiple *m/z* bins at each time point, where each bin is designed to contain roughly equal ion abundance over all samples being aligned. On previous data sets we found four *m/z* bins to provide better alignment than one bin, but without the computational burden of more bins, and without much loss in quality of alignment with respect to using more bins. We thus used four *m/z* bins in this work.

Another minor change we have made to the model in [6] was to remove the hidden scale states, replacing them with 'scaling' spline parameters. In this modified setting, the HMM states now correspond only to 'true time' states, and are no longer augmented by scale states, so that now a path through the HMM states tells us how to warp time in one LC-MS run, but not how to scale the ion amplitude (*i.e.*, normalize it). The scaling is instead performed by adding a new parameter vector to the model, which consist of a spline (and we use one spline per observed time series). A spline is a simple idea in its simplest form (and the one used here). Imagine you have a a set of experimental measurements (*e.g.*, temperature at various time points). Then you might try to fit a straight line to the points so as to characterize or visualize the trend, or if that fails, perhaps a quadratic line. A more general form of trend line would be a spline, in which you pick a few 'control points' at fixed times, between which you interpolate (*e.g.* linearly) the temperature. On this basis, you can then fit the temperature at each control point, using all of the data. The control points are fixed ahead of time, using for example a constant time interval. Thus, instead of scaling our LC-MS data at every time point, the values at various control points of our scaling spline are fitted, using all of the data, with the assumption that we will linearly interpolate between them. This modification has three implications: i) we are no longer estimating the distribution over scale states, since we are no longer treating them as hidden variables, rather, we are obtaining a point estimate of spline control point values, ii), because we use fewer spline control points than latent times, the scaling becomes more global in nature, and iii) the algorithm can run faster because we are obtaining a point estimate of the parameters rather than a full distribution (and also using relatively few of them). This modification achieves roughly the same effect (*i.e.*, puts the different LC-MS runs into similarly good ion amplitude correspondence with one another), but with less computational burden. Details of both changes are available on our web site (see 'Availability' section).

Lastly, although the original CPM was intended to align replicate data, we here use the model to align samples from two different, though similar, (as would be expected in a biomarker discovery type experiment) classes and find it to work well. To accommodate this change, we apply normalization concurrently with alignment during model training, but do not include the learned local normalizations in the final alignment, since if a peak exists in only one class, such normalization would coerce the two classes to look more similar than they are. Instead, after use of the CPM (with spline scaling) for alignment, we simply do a single global normalization. That is, we apply one scaling factor to each LC-MS run so that the total

ion abundances from each LC-MS run are identical after normalization. Figure 1 shows the data set used in the paper before and after alignment by the CPM with a global normalization applied to the aligned data after alignment.

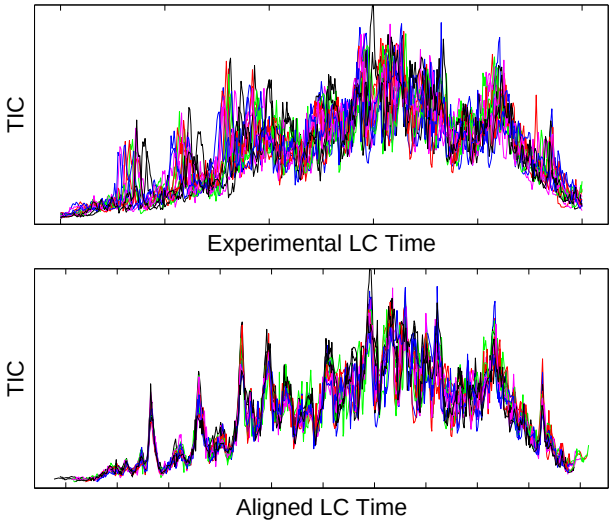

Fig. 1. Unaligned (top) and aligned (top) TIC of the 14 LC-MS runs used in this paper (note that alignment was not done on the TIC, we use TIC for display purposes only). 'Unaligned' actually refers to data that has been coarsely pre-processed with simple linear offsets (shifts) and global scaling.

Detecting Differences Between Classes

After applying the CPM to all samples from each of two classes (*e.g.*, cancer versus not cancer), we have in hand a set of *aligned data matrixes*. That is, we have, for each LC-MS run k, a matrix of data, $\mathcal{M}^k$ where $\mathcal{M}^k_{\tau,i}$ indexes the normalized ion abundance at the τ'th aligned time point for the i'th *m/z* bin in the k'th observed LC-MS run. Note that prior to use of the CPM, all *m/z* values are mapped to bins of width $\frac{1}{2}$ *Th* which are retained throughout all other processing (similarly done in [10, 13, 6]). In our data set, we had roughly 1000 rows corresponding to 1000 time points, and 2400 columns corresponding to *m/z* bins of width 0.5 *Th* spanning 400-1600 *Th* (which were then further reduced to four bins as mentioned earlier).

Even for very good alignments, one can usually observe regions in two different LC-MS runs that appear to correspond to each other, but that are not perfectly and completely overlapping. As such, we have found it useful to apply a small local smoothing in time and *m/z* to each data matrix. The smoothing in *m/z* helps to overcome any possible rounding issues in the *m/z* binning. This smoothing is implemented by convolving a 2D Hamming filter of length 25 in aligned time points (corresponds to roughly 12 observed LC time points because aligned time is roughly twice as dense as observed time – see [6]), and length 3 *m/z* bins (corresponds to 1.5 *Th*) to each of the data matrixes.

The size of the Hamming filter used (25 x 3) was chosen based on data from similar experiments using the following intuition: If we do not smooth at all, then some *m/z* bins across experiments will not be in good correspondence with one another, and likewise, some LC times will not be in good correspondence with one another. Thus, if we use one of these LC-MS runs and ask how well it 'predicts' the other replicates in this class, it will 'predict less well' because of the regions of poor correspondence. Going to the other extreme, if we heavily smooth this same LC-MS run, and then ask how well it 'predicts' the others, its signal will be completely smeared out, and it will not 'predict' them well. In between these two regimes lies an amount of smoothing for which this one LC-MS run optimally 'predicts' the others. More formally, we use the Total Variation Distance (TVD) between a smoothed LC-MS run and the other runs to see how well it 'predicts' the other runs. The TVD is simply a symmetric measure of how much two probability distributions diverge. To be able to use the TVD, we pretend that each data matrix, $\mathcal{M}^k$ is a distribution by forcing its components to sum to one. Formally, our method operates as follows: i) normalize each $\mathcal{M}^k$ to sum to 1, ii) apply Hamming smoothing to one LC-MS run k', iii) measure how 'close' this smoothed run, k', is to every other non-smoothed run by computing the TVD (Total Variation Distance – measures distance between distributions), iv) do this for each run in turn and sum together the TVD values, v) repeat for variously sized Hamming filters. Too little smoothing produces larger TVDs because of mismatched peaks, while too much smoothing washes out the signal so that the smoothed run is 'further' from the non-smoothed signals and the TVD is greater. In between these regimes lies the optimal amount of smoothing.

A Spatial Test Statistic for Difference Detection We devise a simple statistical test, based on a t-statistic (first applied to mass spectrometry data in [12]), which is applied to the data set. First we compute a (Welch) t-statistic at each *(time, m/z)* in the set of data matrixes, $\{\mathcal{M}^k\}$. Let $C^r_{\tau,i}$ be the class mean over all samples in class r at the τ'th time point and i'th *m/z* bin, *i.e.*, $C^r_{\tau,i} \equiv \frac{\sum_{\{k|k\in r\}} \mathcal{M}^k_{\tau,i}}{N_r}$, where N_r is the number of samples in each class. Similarly, let $V^r_{\tau,i}$ be the class standard deviation over all samples in class r at the τ'th time point and i'th *m/z* bin, *i.e.*, $V^r_{\tau,i} \equiv \sqrt{\frac{\sum_{\{k|k\in r\}} (C^k_{\tau,i} - \mathcal{M}^k_{\tau,i})^2}{N_r - 1}}$. Also, let $V'_{\tau,i} \equiv \sqrt{V^1_{\tau,i}/N_1 + V^2_{\tau,i}/N_2}$ be the pooled standard deviation. Then we calculate a signed t-statistic, $t_{\tau,i} \equiv \sqrt{\frac{C^1_{\tau,i} - C^2_{\tau,i}}{V'_{\tau,i}}}$. If $V^1_{\tau,i} = 0$ and $V^2_{\tau,i} = 0$ and $C^1_{\tau,i} = C^2_{\tau,i}$ then [1] we set $t_{\tau,i} = 0$. After this point-wise statistic is calculated, we modify it to account for the fact that we know, *a priori*, that signal of interest in this type of data will span more than one time point, because elution off the LC column is not instantaneous, and possibly more than one *m/z* bin, because of isotope shoulders. We therefore calculate what we call a *spatial* t-statistic, $t'_{\tau,i}$, by performing 2D smoothing on the matrix consisting of t-statistics. This is implemented by convolving the matrix of values $t_{\tau,i}$ with a 2D Hamming filter of length 10 in aligned time (*i.e.*, 5 in experimental time) and length 3 *m/z* bins, corresponding to our prior beliefs about peak sizes of interest. Note that regions of interest which are smaller than this filter can still be captured if their signal is strong enough. That is, this method does not immediately throw out smaller peaks, as is the case in [13] where a hard threshold

[1] This occurs when there is no ion abundance in any of the samples at this time and *m/z*.

of length in time is used, it only requires that such regions provide more evidence for themselves. Without this last step of moving to a spatial t-statistic, we observed that the ability to reliably detect differences across classes was diminished, as can be observed in Figure 2. While smoothing the t-statistic removes the possibility of using corresponding t-statistic p-values, these p-values would in any case provide only a ranking given the massive scale of multiple testing, so nothing is lost in this regard.

To detect differences between classes, we simply look for areas with the largest magnitude spatial t-statistics, $t'_{\tau,i}$. This provides a ranked set of hypotheses. To get a sense of whether these test statistics are providing real information, one can permute the class labels (spike-in/base-mixture) and calculate the resulting null distribution estimate. While it is possible that each (time, *m/z*) may require its own null distribution, we have found a single distribution to be sufficient (and hence can use the spatial t-statistics to rank hypotheses). Figure 3 shows how the distributions of spatial t-statistics in our problem compares to a null distribution estimated from 100 random permutations of the class labels. The real test statistic distribution is skewed toward negative values – here class 1 was base mixture only, while class 2 was base mixture plus spiked-in peptides, so we expect to see more negative signal.

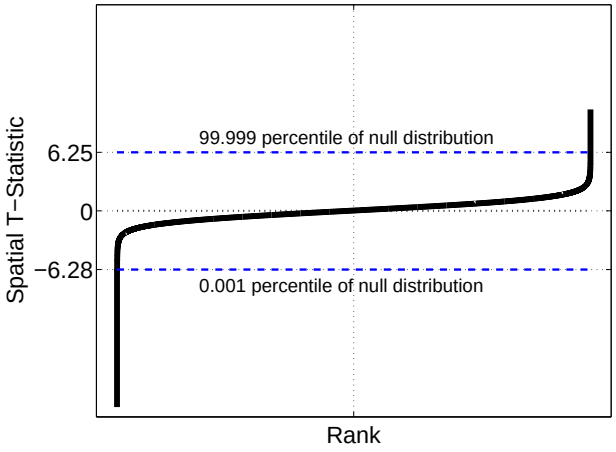

Fig. 3. The solid line shows all 2.4×10^6 sorted spatial t-statistics for our data set ($\sim$1000 time points $\times$ 2400 *m/z* values). The dashed lines show the 0.001 (top) and 99.999 (bottom) percentiles of the null distribution estimated from 100 random permutations of the class labels (producing a total of 2.4×10^8 test statistics). There are 86 spatial t-statistics above the top dashed line, and 468 below the bottom dashed line.

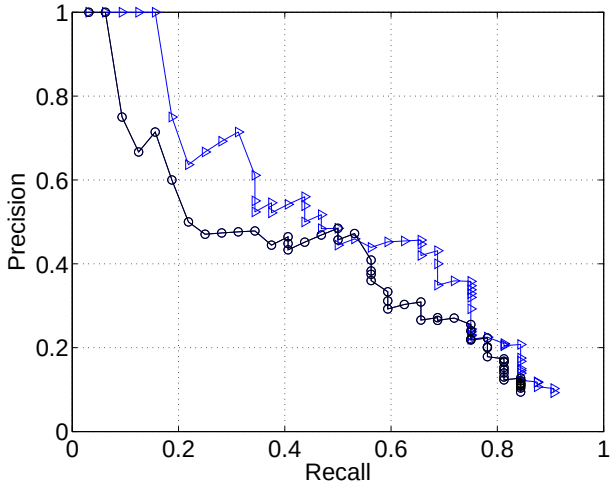

Fig. 2. Precision-recall curves when a spatial t-statistic is used (triangle marker) versus a regular t-statistic (circle marker) when using seven replicates from each class. The curve for the spatial t-statistic is the same as that in Figure 4 below, with seven replicates.

Comparison to Ground Truth

To characterize the performance of our algorithm, we ran the spiked-in reference peptides in buffer-only (no serum) through the LC-MS to obtain eight 'ground-truth' runs. Clearly, extracting a single ground truth from these runs is itself not a trivial problem. We might have used our CPM alignment/normalization to help extract the ground truth, but we sought a method which was completely independent from that which was used to do difference detection. Because the reference peptides are relatively few in number, peak detection was not overly difficult, though we note that it is still unlikely to be perfect. We extracted ground truth peaks in a fairly simple manner which involves four main steps, i) thresholding, ii) clustering, iii) *m/z* extraction, iv) voting across eight ground truth LC-MS runs.

Note that our ground truth comparisons use *m/z* , not time, since time is not consistent, producing a large dependency on alignment, which we wished to bypass for comparison with ground truth. However, the main signal of interest, and that with the highest precision and accuracy, lies in *m/z* , not time. We did however visually verify that detected differences did correspond to the spiked-in peptides in both time and *m/z*. Additionally, when our method is used in a setting that does not involve comparison to ground truth, it is simple to obtain LC-time estimates for all detected differences by mapping back from HMM hidden time states to real observed time and averaging over replicates to obtain a time window.

Formally, for each ground truth run, w, we i) created a data matrix, $\mathcal{M}^w$, by binning the *m/z* as described earlier, ii) smoothed the matrix using 2D Hamming filter of length 10 in aligned time (corresponds to 5 time points in experimental time) and 3 *m/z* bins, iii) normalized the total ion abundance in each matrix to be 1, iv) thresholded the abundance at each (time, *m/z*), so that if the abundance was less than T_0, then we set $\mathcal{M}^w_{i,\tau} = 0$, v) for all remaining non-zero entries in $\mathcal{M}^w$, call a peak any group of tuples (i, τ) that are connected (*i.e.*, are linked by a path of $\mathcal{M}^w_{i',\tau'} > 0$, where link is defined with respect to a neighbourhood consisting of the eight surrounding points in (time, *m/z*)). We calculated the *m/z* value of each peak to be the weighted-average *m/z* value of all tuples assigned to that peak, where the weights are proportional to each tuple's ion abundance. The threshold, T_0, was chosen on the basis of histograms of the data, and in such a manner as to attempt to cut off background noise while maintaining the greatest possible signal (we used $T_0 = 5 \times 10^{-5}$). We set this threshold low (retaining as much as possible) knowing that we would force replicate LC-MS ground

truth runs to later agree, reducing the possibility of spuriously detected peaks. Finally, we calculated the 'strength' of each peak as the sum of the ion abundance of all tuples assigned to it. If several peaks had the same *m/z*, their strengths were added to obtain the strength for that *m/z*. This strength attribute is later used to see if 'stronger' peaks are those identified more easily (Figure 5).

Given a list of *m/z*'s for each of the eight ground truth replicates, we then clustered the *m/z* values using complete-linkage, Euclidean distance hierarchical clustering, using a distance cut-off that produced clusters such that no two *m/z* values assigned to a cluster differ by more than 1 *Th*. Lastly, each unique cluster is voted on in a binary fashion from the 8 ground truth peak extractions (using a tolerance of 1 *Th*), and those clusters with at least 6 votes were considered to be in our final ground truth. In this manner, we extracted 32 ground truth *m/z* values. The strength of each of these was determined by the average strength of the peak corresponding to that *m/z* value, over the LC-MS replicates that voted for it.

We also had access to theoretical predictions for our tryptically digested peptides. While these theoretical predictions are not a bad guide, ultimately, it is unknown which charge-state ions will be observed, whether non-tryptic peptides will be found [9], and what the ionization efficiency of each tryptic peptides is (and hence whether it will be observed). We note however, that our experimentally extracted ground truth peaks contained 7 of the 8 theoretical predictions.

Lastly, we note that our naive method of extracting ground truth peaks cannot be applied to the actual difference detection problem itself, since in realistic settings, using for example serum, there are so many peptide peaks that it is impossible to segment them from background in the naive way described in this section, and more sophisticated methods would need to be applied.

Precision-Recall Curves

In order to compute precision-recall curves for the difference detection task, we found regions of interest and compared their *m/z* values to our extracted ground truth list. *Precision* reflects what proportion of our detected differences appear in the ground truth, while *recall* tells us what proportion of the ground truth values we actually managed to recover.

Specifically, we i) choose a threshold, R_0, for the spatial-t statistic, and refer to all *m/z* values with a t-statistic larger in magnitude than R_0 as 'on', ii) find all unique 'on' *m/z* values, iii) obtain the number of true positives by counting how many of our 'on' *m/z* appear in our extracted ground truth list within a tolerance of 1 *Th*, (and which also appear in the correct direction – negative in this experiment) not allowing more one than true positive count per ground truth peak, iv) for all 'on' *m/z* which did not match to a ground truth peak, cluster them (again, complete-linkage, Euclidean distance hierarchical) such that every cluster contains only *m/z* values no more than 1 *Th* apart, v) use the number of unique clusters as the number of false positives. We repeated these steps for decreasing thresholds, R_0 to trace out the precision-recall curve, stopping when the precision reached 10%.

Note that precision-recall curves differ significantly from ROC curves. While a diagonal line on an ROC curve shows that a classifier is guessing no better than random, a diagonal line on a precision/recall curves shows that one can, for example, extract 50% of known ground truth values while incurring a false positive rate of 50%, which may be a non-trivial achievement.

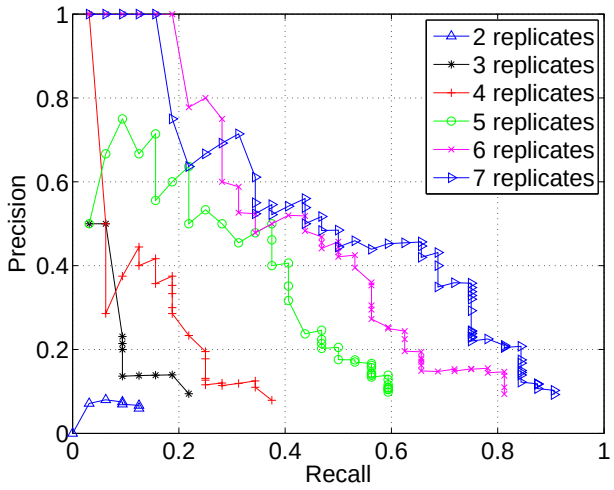

Fig. 4. Precision-Recall curves for the difference detection analysis. Each curve shows the results when a different number of replicates was used, from two through to seven.

3 EXPERIMENTS AND RESULTS

We tested our difference detection technique on a spike-in LC-MS data set where class 1 consisted of a base mixture of human serum, and class 2 consisted of this same base mixture of serum, along with three known peptides (see Laboratory Methods for details). We show later that the amount of spike-in was not trivially easy to find, as was intended by design of the experiment. For each data set, serum-only runs were alternated with the spike+serum runs (seven of each class) to control for potential time and order-dependent biases, and all samples were run on the same column.

Laboratory Methods

Frozen human serum was thawed on ice. A 2 μl aliquot ($\sim$160 μg protein) was denatured and reduced for 30 min on ice by adding 5 volumes (10 μl) of 8M urea (pH 8.5), 1 mM fresh DTT. This was followed by 5 volumes (10 μl) of acetonitrile and 50 μl of 100 mM Ammonium Bicarbonate, 1 mM $CaCl_2$. The samples were then incubated and digested for two days with 10 μl of covalent trypsin beads (Poroszyme; Applied Biosystems) at 30°C with rotation. Prior to analysis, the samples were further diluted with 70 μl of Buffer A (5% ACN, 0.05% HFBA). For the spiking runs, the reference peptides standard was added at 1.0 pmol (Peptide Calibration Standard #206195, Bruker Daltonics Inc.) to the sample immediately prior to LC-MS. 5 μl (0.2 nmol total protein) of the diluted serum samples were analyzed using standard capillary scale LC-MS profiling methods as described in [10]. Briefly, a quaternary HPLC pump was interfaced using the electrospray ionization method to an LCQ quadrupole ion trap tandem mass spectrometer (Thermo Finnigan; San Jose, CA). The samples were loaded onto a 150 μm i.d. fused silica capillary micro-column (Polymicro Technologies; Phoenix, AZ) bearing a fine nozzle created with a P-2000 laser puller (Sutter Instruments; Novato, CA) and packed with 8 cm of 5 μm Zorbax 300SB-C18 resin (Agilent Technologies, Mississauga, ON, Canada). The ion trap mass spectrometer was operated in dual cycling mode, cycling from precursor scanning to dynamic

data-dependent MS/MS scan mode, even though we did not use the MS/MS data in our experiments, and this mode in fact degrades the quality of the data.

Difference Detection Results

Figure 4 shows the precision-recall curves obtained when using 7 replicates from each class down to only 2 replicates from each class. Using 7 replicates from each class, we obtain very reasonable results, with for example a recall of 50% matching to a precision of 50%. Secondly, as expected, with a decreasing number of replicates, our ability to achieve good results systematically deteriorates. Note that alignment in LC-time was done using all 7 replicates from both classes, and only for the latter part of the analysis, (that which uses the spatial t-statistic) was the number of replicates varied. One would expect that if the entire analysis had been done on varying number of replicates that the systematic differences would be more extreme, as less information would be available with which to produce good alignments.

Assessing The Difficulty of the Problem

If the amount of reference peptides spiked in were very large, the difference detection problem might be trivially easy. How can one get at how difficult our problem is? Well, if the amount of spike-in were very large relative to the base mixture, then the spike-in would swamp the signal of the base mixture, and a simple *1D analysis* looking at the ion count at each *m/z*, irrespective of time (*i.e.*, summing out the ion count out over time) would be able to tease out all differences of interest. We apply such a technique here, and show that the 1D analysis is inferior to the full *2D analysis* described in Section 2.

Formally, the 1D analysis involves converting each data matrix, $\mathcal{M}^k$, to a vector representing the total ion count at each *m/z* bin. Thus we define the Total Time Ion Count (TTIC), for *m/z* bin i for the k'th LC-MS run, T_i^k, as $T_i^k \equiv \sum_\tau \mathcal{M}_{i,\tau}^k$. Then the 1D analysis is exactly the same as the 2D analysis, only it operates on $\{T_i^k\}$ rather than on the set $\{\mathcal{M}_{i,\tau}^k\}$. A direct comparison of the 1D versus the 2D approach is shown in Figure 5. The 2D approach allows us to more precisely find subtler regions of interest. To assess the stability of this relationship, we redid this analysis seven times, each using a different subset of the six replicates per class, and found the pattern to be somewhat unstable in the region of recall less than 0.3, indicating there are no statistically significant differences in this region. However, for recall larger than 0.3, the pattern was stable, suggesting a statistically significant difference for regions of larger recall. Thus we have provided evidence that our spike-in experiment was not trivially easy.

We want to emphasize that the point here is not to contrast two difference detection methods (as they are essentially the same), but to show that the signal of interest is indeed swamped out, to a certain extent, by baseline peaks, indicating that the amount of spike-in is not so large as to rise above this, and thus trivially easy to find.

Predictive Modeling

While we were not able to achieve perfect difference detection results, (where perfect would mean always having precision=1 and recall=1), we were nevertheless able to easily build a perfect supervised prediction model for the class of each sample. We used regularized LR (logistic regression [2]) as a prediction model with features consisting of every (time, *m/z*) in the aligned data matrixes, $\{\mathcal{M}^k\}$, described earlier[2]. We used L2 regularization, which is similar to putting a Gaussian prior with spherical covariance over the LR weights.

Over a range of 7 orders of magnitude for the regularization term, we always achieved perfect sensitivity/specificity as measured by leave-one-out cross validation (*i.e.*, a perfect ROC curve). This demonstrates that the problem of classification is far easier than that of difference detection, and that good classification performance provides little information about the ability to do comprehensive difference detection. Of course if one were unable to build a good classifier, this would suggest that the task of difference detection would likely also be difficult. Conversely, if one were able to successfully perform difference detection, then one might reasonably expect to be able to build a good classifier.

4 DISCUSSION AND CONCLUSION

We have presented and evaluated a technique for discovering differences in protein signal in serum, between two classes of samples of LC-MS data, without use of tandem mass spectrometry, gels, or labeling. Along the way, we demonstrated how to show that a particular spike-in difference detection problem is not trivial, and in particular that our difference detection problem was not trivial. We also showed that the problem of supervised prediction, sometimes mistaken as a solution to biomarker discovery, is actually a much simpler problem. Our difference detection technique works on data from lower precision mass-spectrometers, the kind that are widely available and affordable to the proteomics community at present, making our work of immediate and widespread utility. Lastly, we hope that by making our data set publicly available, that further papers of this nature can include comparisons to previously published results, as, at present, this is a difficult, if not impossible, task.

ACKNOWLEDGMENT

We thank the reviewers and Daniel Knapp for suggestions on how to improve the manuscript.

REFERENCES

[1]A.H. America, J.H. Cordewener, M.H. van Geffen, A. Lommen, J.P. Vissers, R.J. Bino, and R.D. Hall. Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional LC-MS. *Proteomics*, 2:641–53, 2006.

[2]T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[3]T. Kislinger and A. Emili. Going global: protein expression profiling using shotgun mass spectrometry. *Curr Opin Mol Ther.*, 5:285–293, 2003.

[4]Jennifer Listgarten and Andrew Emili. Practical proteomic biomarker discovery: taking a step back to leap forward. *Drug Discovery Today*, 10:1697–1702, 2005.

[5]Jennifer Listgarten and Andrew Emili. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular and Cellular Proteomics*, 4:419–434, 2005.

[6]Jennifer Listgarten, Radford M. Neal, Sam T. Roweis, and Andrew Emili. Multiple alignment of continuous time series. In L. K. Saul et al, editor, *Advances in Neural Information Processing Systems*, volume 17. The MIT Press, 2005.

[2] For computational efficiency, we actually do LR in an equivalent 14-dimensional feature space obtained from PCA (Principal Components Analysis) which gives identical results to doing LR in the original feature space [8].

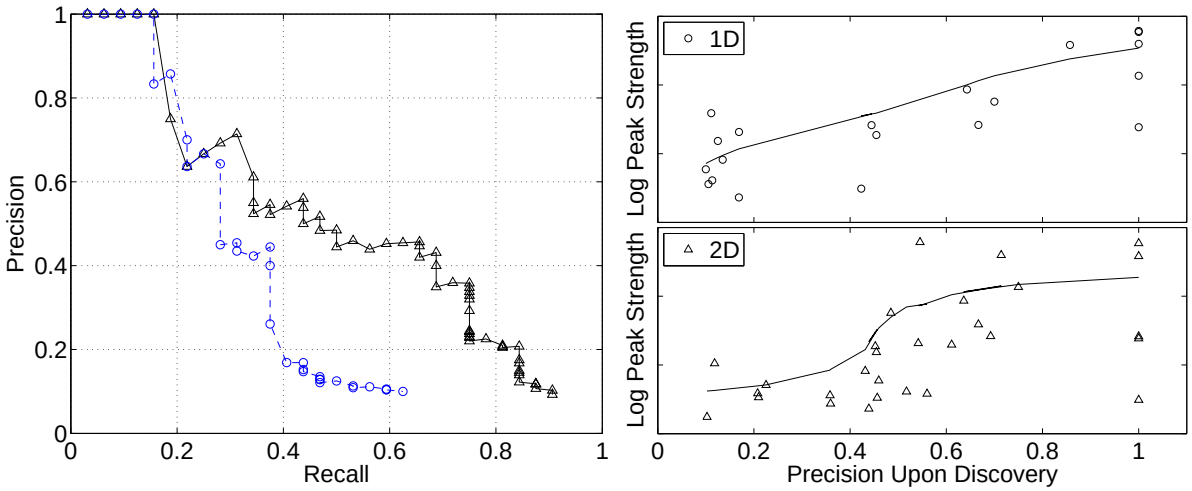

Fig. 5. Left: Precision-Recall curve for 1D analysis (dashed-circle) versus a full 2D analysis (solid-triangle). Right: The precision level when each peak is detected along with the strength of the peak. Overlaid is a lowess-smoothed (span=17) version of the plotted data points to more easily visualize the trends. As expected, stronger peaks are discovered at higher precision levels, though the trend is noisy. Both plots used seven replicates from each class. The 1D plot contains 20 ground truth peaks, while the 2D plot contains 29.

[7]James Lyons-Weiler. Standards of excellence and open questions in cancer biomarker research: An informatics perspective. *Cancer Informatics*, 1:1–7, 2005.

[8]R. M. Neal and J. Zhang. Classification with Bayesian neural networks and Dirichlet diffusion trees. In I. Guyon et al, editor, *Feature Extraction, Foundations and Applications*. Physica-Verlag, Springer, 2006.

[9]A. Prakash, P. Mallick, J. Whiteaker, H. Zhang, A. Paulovich, M. Flory, H. Lee, R Aebersold, and B. Schwikowski. Signal maps for mass spectrometry-based comparative proteomics. *Molecular and Cellular Proteomics*, doi:10.1074/mcp.M500133-MCP200, 2006.

[10]Dragan Radulovic, Salomeh Jelveh, Soyoung Ryu, T. Guy Hamilton, Eric Foss, Yongyi Mao, and Andrew Emili. Informatics platform for global proteomic profiling and biomarker discovery using liquid-chromatography-tandem mass spectrometry. *Mol Cell Proteomics*, 10:984–997, 2004.

[11]Jeffrey C. Silva, Richard Denny, Craig A. Dorschel, Marc Gorenstein, Ignatius J. Kass, Guo-Zhong Li, Therese McKenna, Michael J. Nold, Keith Richardson, Phillip Young, and Scott Geromanos. Quantitative proteomic analysis by accurate mass retention time pairs. *Analytical Chemistry*, 77:2187–2200, 2005.

[12]W. Wang, H. Zhou, H. Lin, S. Roy, T.A. Shaler, L.R. Hill, S. Norton, P. Kumar, M. Anderle, and C.H. Becker. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical Chemistry*, 75:4818–4826, 2003.

[13]Matthew C. Wiener, Jeffrey R. Sachs, Ekaterina G. Deyanova, and Nathan A. Yates. Differential mass spectrometry: A label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. *Analytical Chemistry*, 76:6085–6096, 2004.