Bioinformatics Advance Access published August 30, 2007

Intersession Reproducibility of Mass Spectrometry Profiles and its Effect on Accuracy of Multivariate Classification Models

Richard Pelikan [1−3], William L. Bigbee [4], David Malehorn [4], James Lyons-Weiler [3−5] and Milos Hauskrecht [1−4]

[1]Department of Computer Science,[2]Intelligent Systems Program,[3]Department of Biomedical Informatics,[4]University of Pittsburgh Cancer Institute,[5]Genomics and Proteomics Core Laboratories, University of Pittsburgh, Pittsburgh PA 15260

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Motivation: The "reproducibility" of mass-spectrometry proteomic profiling has become an intensely controversial topic. The mere mention of concern over the "reproducibility" of data generated from any particular platform can lead to the anxiety over the generalizability of its results and its role in the future of discovery proteomics. In this study, we examine the reproducibility of proteomic profiles generated by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF-MS) across multiple data-generation sessions. We analyze the problem in terms of the reproducibility of signals, reproducibility of discriminative features, and reproducibility of multivariate classification models on profiles for serum samples from early lung cancer and healthy control subjects.

Results: Proteomic profiles in individual data-generation sessions experience within-session variability. We show that combining data from multiple sessions introduces additional (inter-session) noise. While additional noise can affect the discriminative analysis, we show that its average effect on profiles in our study is relatively small. Moreover, for the purposes of prediction on future (previously unseen) data, classifiers trained on multi-session data are able to adapt to inter-session noise and improve their classification accuracy.

Contact: pelikan@cs.pitt.edu,milos@cs.pitt.edu

1 INTRODUCTION

Mass spectrometry (MS) proteomic profiling has shown potential to quickly and effectively screen patients for disease. This is done by producing protein expression profiles from patients' tissue, blood, urine, saliva or other biofluid. Statistical machine learning techniques are applied to the resulting complex protein expression profiles in a process called predictive modeling. In typical case/control comparative studies, example profiles generated from biospecimens of diseased patients and healthy subjects are shown to the model in the training phase. New profiles from the screened subjects are evaluated by the model in the testing phase.

Earlier MS proteomic profiling studies stimulated significant enthusiasm (Petricoin *et al.*, 2002), discussion (Diamandis, 2003), and controversy (Ransohoff, 2005) in the general scientific community and among proteomics researchers. Potential confounding and bias in study design and analysis in initial studies (Baggerly *et al.*, 2004), were recognized early on and have been addressed in subsequent research (See Grizzle *et al.* (2005) for an overview).

Predictive modeling relies on the detection of potential biomarkers which may explain disease through previously understudied combinations of reproducible molecular measurements. The reproducibility of these surrogate biomarker patterns often comes into question; a pattern is not guaranteed to be replicated exactly within the same or other data generation session, or at a different laboratory. This results from the intrinsic variation introduced into the data by factors including, but not limited to, the biological nature of the samples and limitations of the MS technology.

Typical proteomic profiling studies attempted to minimize the effect of this variation by generating data in a single session. These data sets were produced in the 'ideal' environment where only a single instrument in a single laboratory produces all of the available data at the same time. As a result, potential factors of inter-session and inter-site biases were ignored. Despite encouraging classification results on these data sets, skepticism arose as to whether spectra generated during multiple sessions separated by variable intervals of time, or by a different laboratory, will be useful for predictive modeling applications. Promising inter-site reproducibility results were reported by (Zhang *et al.*, 2004; Semmes *et al.*, 2005). Inter-session reproducibility, however, remains a relatively open area of research.

The aim of our paper is to study the inter-session reproducibility of proteomic profiles generated by the same instrument over the course of 18 months. Inter-session reproducibility is the key to generalizability of classification results to any future sample analysis. Our study relies on proteomic profiles generated by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF-MS) of serum samples from early-stage lung cancer patients and healthy control subjects. The samples for 46 patients were repeatedly reanalyzed in four different sessions over the course of 18 months. Four groups of spectra generated during these sessions were the basis of this analysis.

In a clinical setting, samples are obtained sequentially rather than collectively, and any models developed should be applicable to samples produced in the future. Thus, a realistic model builds upon profiles from multiple data-generation sessions and is applied to data generated in new sessions. The study of such multi-session models and their characteristics is thus at the heart of this investigation.

We show that the performance of classification models built on profiles from multiple sessions is lower on average than the performance of the models built from single-session profiles. This

inter-session noise and possible biases appear to influence the profiles both on the level of individual peak signals as well as multivariate biomarker panels. However, the average effect of the inter-session noise (in our study) appears to be relatively low. Moreover, models which are trained on data from multiple sessions can adapt to this noise and improve their performance. This supports the notion that samples need not be run all at once, but rather may be collected on an as–produced basis.

2 MATERIALS AND METHODS

The samples analyzed in this work consist of 21 lung cancer and 25 control sera belonging to a larger pool of samples collected for the the University of Pittsburgh Cancer Institute (UPCI) lung cancer study (see below). The samples were originally selected to support a concurrent inter-site validation study. The samples were analyzed by SELDI-TOF-MS instrumentation at *four time periods* (June 2003, February 2004, November 2004 and January 2005), which we refer to as *sessions*. The sample selection process was similar to a prior inter-site validation study for prostate cancer (Semmes *et al.*, 2005). The selection occurred in February 2004 and was restricted to samples that were available in June 2003 and that had a sufficient number of aliquots remaining for later analyses. Hence the sample is not representative of population proportions of the parent UPCI study described below.

2.1 UPCI lung cancer clinical population

The parent study consisted of 115 newly diagnosed resectable non-small cell lung cancer (NSCLC) cases from the UPCI Lung Cancer Specialized Program of Research Excellence (SPORE) project and from 106 healthy controls matched by age, gender, smoking status and pack-year history. The NSCLC cases were distributed among patients as follows: age (40-49 4%, 50-59 18%, 60-69 34%, 70-79 36%, 80-89 8%); gender (men 59%, women 41%); smoking status (active smokers 40%, ex-smokers 49%, never smokers 11%); pack-year history ($<$30 23%, 30-59 29%, $>$60 37%); histopathology (adenocarcinoma 54%, squamous cell carcinoma 32%, other/unknown 14%); and stage (IA 21%, IB 23%, IIA 4%, IIB 15%, IIIA 21%, IIIB 7%, IV 9%). All case and control sera were collected and processed per a standardized protocol developed by the UPCI Lung Cancer SPORE. The samples were processed, divided into equal aliquots and stored at -80°C within 1 hour of collection in glass Vacutainer® tubes (BD Medical, Franklin Lakes, NJ). Only a subset of these samples are used in the reproducibility analysis pursued in this work and thus do not reflect the above population proportions. We note that the exact clinical population characteristics are less relevant to the study of inter-session reproducibility.

2.2 Preparation of serum for SELDI analysis

A fresh set of aliquots was used for each data production session. The protocols for the preparation and loading of serum samples for SELDI-TOF-MS analysis are specific for the ProteinChip® Arrays (Ciphergen Biosystems, Inc., Fremont, CA). Fully automated BioMek2000® protocols for processing of IMAC3 ProteinChip® Arrays are presently being utilized in the UPCI Clinical Proteomics Facility directed by Dr. Bigbee. Protocols for automated processing of these ProteinChip® Arrays, as well as performing mass spectrometry and preprocessing of the spectral data for analysis, have been derived and optimized from protocols implemented and validated in accord with an NCI EDRN validation study assessing the reproducibility of the SELDI-TOF-MS platform (Semmes *et al.*, 2005). Serum samples were denatured prior to processing on ProteinChip® Arrays. Twenty μl aliquots of serum were added into one well of a 96-well polystyrene microtiter plate, with 30 μl of 8 M urea/1% CHAPS in PBS. The serum-urea mixture was vortexed for 30 minutes at 4°C. One hundred μl of 1M urea / 0.125% CHAPS was then added to the serum/urea mixture and briefly mixed, followed by a 1:5 dilution of the serum/urea mixture with PBS. One hundred μl of the final diluted serum/urea mixture was then applied to one spot of a ProteinChip® Array, prepared as described below. Each serum sample was processed in duplicate in a blinded layout of combined case/control samples, together with a standard pooled serum sample (one spot on each ProteinChip® Array for quality assurance/control purposes).

2.3 Preparation and loading of ProteinChip® arrays

For pre-activation of the IMAC3 ProteinChip®, arrays were assembled into the Ciphergen Bioprocessor®, holding up to 12 chips, which allows for applying larger volumes of liquid to each array spot. The IMAC3 ProteinChip® Arrays were initially loaded with 50 μl of 100 mM $CuSO_4$ on each spot of the array. The chips were shaken on a TOMY MT-360 Micro Tube Mixer (Tomy Seiko Co., Ltd.), set at speed Form 20, Amplitude 7 for 5 minutes. Each array spot was then rinsed with 200 μl HPLC grade water, and aspirated. Fifty μl of 100 mM sodium acetate pH 4.5 were added to each array spot, and the chips shaken 5 minutes. The chips were rinsed with HPLC grade water, and then equilibrated twice for 5 minutes with 200 μl of PBS in each well. Equilibration buffers were aspirated prior to application of 100 μl of the denatured serum/urea mixture into each well of the Bioprocessor®; great care was taken to ensure no bubbles remained at the bottoms of the wells, occluding contact with the ProteinChip® surface. Serum mixtures were incubated with the ProteinChip® Arrays for 30 minutes at room temperature with shaking. The serum/urea mixtures were then discarded and the PBS washing step repeated twice, followed by 2 final rinses with HPLC water. The chips were removed from the Bioprocessor®, and air-dried at least 10 minutes but as long as overnight. The chips were stored in the dark at room temperature until SELDI-TOF-MS analysis. Immediately prior to analysis, 1.0 μl of a half-saturated solution of the energy absorbing molecule (EAM) sinapinic acid (Ciphergen Biosystems, Inc.) in 50% (v/v) acetonitrile, 0.5% trifluoroacetic acid was applied onto each spot of the array twice, letting the surface air dry 5 minutes between each EAM application. All chips spotted were read, as much as possible, in an uninterrupted run using the Ciphergen ChipReader® AutoLoader® device.

2.4 SELDI-TOF mass spectrometry analysis

The reacted ProteinChip® Arrays were analyzed using the PBSIIc ChipReader® instrument (Ciphergen Biosystems, Inc., Fremont, CA). The SELDI-TOF-MS spectra were collected by the accumulation and averaging of 192 laser shots from 16 positions across the diameter of the ProteinChip® Array spot, with warming shots not included. A laser intensity of 175-180 was used in a positive ion mode, ensuring that transient shot intensities were below saturation of the detector, with a detector sensitivity setting of 6, a focus lag time of 900 ns, employing mass deflection at 1000 Daltons. The protein masses were calibrated externally using the 7-in-1 purified peptide molecular mass standard (Ciphergen Biosystems, Inc.).

During each session, each sample was processed in duplicate, and each pair of replicates was averaged prior to further data pre-processing to create a mean profile for that pair. This resulted in a dataset of 184 spectra, with each of the 46 samples' profiles being produced once per session. These four datasets are the basis of our analysis.

2.5 SELDI-TOF-MS data preprocessing

Profile preprocessing aims to remove systematic noise and biases in the data while preserving the useful information content carried by the profiles. Typical MS profile pre-processing steps include: quality control, baseline correction, variance stabilization, normalization, alignment and smoothing. Profile preprocessing was performed using the Proteomics Data Analysis Package (PDAP), a collection of data analysis and visualization routines supporting the multivariate analysis of proteomic spectra and related biomarker discovery. PDAP has been developed at the Department of Computer Science, University of Pittsburgh. Results and methods developed within PDAP have been described in detail in three recent publications (Pelikan *et al.*, 2004; Hauskrecht *et al.*, 2005, 2007).

As each session of data was produced, case and control profiles were pre-processed together, but separately from spectra produced during other

sessions. Preprocessing through PDAP consisted of cube-root variance stabilization, baseline correction, intensity correction based on total ion current (TIC) in the range of 1.5 to 20 kDa, smoothing with Gaussian kernels, and profile alignment based on the mean spectra. No spectra failed to meet our quality control requirement that the TIC be within 2 standard deviations of the mean TIC across all spectra in a session.

2.6 Reproducibility analysis

Our experimental process evaluated the variability and reproducibility experienced by producing proteomic data in multiple sample-analysis and data-generation sessions separated by a large amount of time. Our analysis is divided into three steps which address the problem of inter-session reproducibility at different levels.

- We first examined the differences in signals from the same sample across multiple sessions. We defined a signal difference score to measure the discrepancies between signals from the same sample. We asked if the signal difference score for profiles from the same sample is significantly better than profiles from other samples. This would indicate that identical samples processed in multiple sessions experience more similarity to themselves than to other samples in the session, supporting the usage of profiles from multiple session for analysis purposes.
- Second, we asked whether discriminative information is affected by inter-session noise. We analyze this issue on the peak signal and multivariate levels, using differential expression and classifier accuracy metrics, respectively. The effect of intersession noise on these statistics is determined by comparing them on single-session and randomized multi-session data sets.
- Finally, we studied the predictive performance of multivariate models on future sessions. We asked by how much the performance of classification models deteriorates on future sessions with respect to their 'ideal' single-session performance. We also asked if performance of a multivariate model on future sessions can be improved if the model is trained on mixed-session data. The idea here is that if intersession variability exists, it can be learned through multi-session data, potentially leading to to accuracy gains over models trained on single sessions.

In the following we outline specific methods to test these objectives.

2.6.1 Reproducibility of profile signals No two MS profiles are exactly the same. Profiles may differ due to instrument noise, differences in sample preparation procedures, etc. Differences in profiles for the same sample are visible even if two profile replicates are generated in the same session, and even if they are placed on the same chip. The intra-session profile variation is well known and existing methods are robust enough to cope with it. The differences in profiles for the same sample across multiple data generation session are much less understood. The differences in the sample preparation at different times or instrument settings may effect the resulting profiles and contribute to possible inter-session biases and variability.

Figure 1 displays four MS profiles from the same sample that were generated in four different sessons. Although the shape of the profile may look similar, differences in relative intensities of peaks are apparent. Are these differences significant? Are these variations too strong to overcome so that the profiles from the same sample are useless and easy to confuse with profiles generated for other samples? To answer these questions we need to define a similarity (or distance) metric that helps us assess the differences among profiles. We would like MS profiles from the same sample to differ less across sessions than profiles from other samples. To achieve this goal we measure the similarity among a set $\mathbf{S}$ of k spectra using the average Euclidean distance d_E between all pairs of spectra:

$$d_E(\mathbf{S}) = \frac{1}{\frac{k(k-1)}{2}} \sum_{\forall 1 \le p < q \le k} \sqrt{\sum_{i=1}^{d} (p_i - q_i)^2} \qquad (1)$$

where p and q represent a pair of spectra from the subset of k replicate spectra generated from the same sample source. Intuitively, the signal difference score measures the sum of areas between all possible superimposed pairs of k spectra; smaller values indicate better similarity.

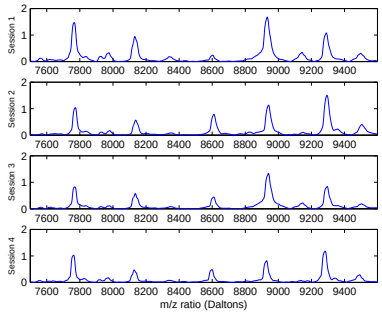

Fig. 1. MS profiles for a single sample across 4 different sessions. Changes are apparent in relative intensities of peaks.

We used the above signal difference metric first, to evaluate the similarity of spectral measurements from the same sample across multiple sessions and then, to determine that the differences from random collections of spectra from other patients are very different, and thus profiles that originate from the same sample are hard to confuse with other profiles.

A random permutation test (Good, 1994) was used to test the differences and their significance. We first estimated a distribution of signal difference scores for randomly grouped spectra. Random groupings were generated by shuffling the sample identities assigned to spectra in each session. The signal difference score was recalculated for each random profile grouping, and the process was repeated 1000 times to estimate the distribution of signal difference scores for randomly grouped spectra. Next, the signal difference score for profiles belonging to the correct samples was calculated. If the score is statistically significantly different with respect to the estimated distribution, we have greater confidence that signals from the same sample are similar to each other beyond random effects. This increases our confidence that profiles generated from multiple sessions are potentially useful for analysis.

2.6.2 Reproducibility of discriminatory signals Evaluating profile similarity across sessions helps assure us of the basic consistency (reproducibility) of spectra with respect to samples they represent. However, the differences in profiles across multiple sessions are apparent (see Figure 1). This leads to a concern that information potentially useful for disease detection purposes may be lost or at least significantly compromised if data from multiple sessions were used in the analysis. To assess the effect of the potential information loss we compare data mixed from multiple sessions to data generated from individual sessions and their discriminatory power.

The information that helps us discriminate between healthy (case) and diseased (control) profiles can be drawn from a single feature (peak) of the profile, or from a combination of multiple features. We measure the quality of discriminative information for a single feature (peak) by its *differential expression score*. The score quantifies the difference observed in a profile feature between case and control groups. Many criteria exist for measuring differential expression (Hauskrecht *et al.*, 2007). In this paper, we use the Fisher-like score, computed as $|\frac{\mu^{(+)} - \mu^{(-)}}{\sigma^{(+)} + \sigma^{(-)}}|$, where μ and σ represent the sample mean and variance of the feature, respectively. The signs $(+)$ and $(-)$ denote case and control samples, respectively.

Testing peaks' discriminatory information loss: To determine if the differential expression information is lost across multiple sessions we assumed that feature's differential expression follows a distribution across sessions. The distribution can be empirically estimated by randomly choosing each sample's spectrum from its replicate set. We generate 1000 randomized datasets and calculate a feature's differential expression score under each dataset to recover its empirical distribution.

If the profiles generated in a single session retain more discriminatory information, we expect their differential scores to be higher on average than the mixed-session distribution. We can test this by comparing the differences between the mean score for the mixed-session distribution and the score for the single session. We have four different sessions per sample and multiple spectra peaks. We use 100 peak regions, evaluate their single-session scores and compare their peak scores to the distributions generated for mixed session data. This process generates a distribution of score differences. If the single session spectra are 'better' we expect them to differ on average from 0. This difference and its significance can be assessed using the standard one-sided hypothesis testing framework.

Testing multivariate information loss: The reproducibility of differential information in individual features may be indicative of the reproducibility of discriminative information given by combinations of these features. However, this is not guaranteed. Are the feature combinations differently represented across sessions? If we mix data from different sessions, what is the effect on the discriminative pattern and the resulting predictive model? To answer these questions, we determine if the performance of a predictive model deteriorates on data mixed from several sessions, as opposed to data from the same data-generation session.

Performance of a predictive model is typically measured using accuracy (percentage of correct predictions), sensitivity and specificity, or area under the ROC curve statistics. In this work, we evaluate predictive models using their test set accuracy. Similarly, there are many classification models one may try to learn multivariate patterns. We use the linear Support Vector Machine model (Vapnik, 1995) to learn the relationship between diagnostic features and state of disease. This method has been used in previous studies (Pelikan *et al.*, 2004; Hauskrecht *et al.*, 2005, 2007) and is favored for its 'regularized' feature selection.

To assess the reproducibility of multivariate classification patterns across sessions, we generate 1000 random datasets such that each patient (sample) receives one of the profiles from its replicate set. Our goal is to analyze differences in the performance of classifiers on: (a) models trained and tested on profiles from multiple sessions, versus (b) models trained and tested on profiles from the same session. To measure test accuracies of models we first decide which patients (samples) will be used for training and testing purposes. Forty-six patients (samples) are split via random subsampling (Efron, 1987) so that 30% of the samples are in the test set. The spectra obtained for the remaining samples are used to train the predictive model. The split is always the same for both single-session and multi-session models. Test set accuracies of 1000 random models define a distribution of accuracy scores for multi-session data. This distribution can be compared to accuracy results for models trained and tested on four single sessions. However, four single sessions entries are not sufficient to make any strong conclusion. In addition, there is a chance a single train and test split may be biased. To eliminate these problems, we repeat the analysis for multiple (30) train–test splits. This lets us calculate 120 accuracy scores for single session models (30 per one session) and compare them to respective accuracy-score distributions defined by 1000 multi-session datasets. To assess the benefit or loss resulting from use of multi-session data, we compare the mean of their accuracy-score distribution to accuracies achieved by single-session models. To assess the global benefit or loss, we average the results over four different sessions.

2.6.3 Effect of multi-session data on generalization performance of predictive models In the 'ideal' analytical setup for proteomic profiling studies, a predictive model is trained and evaluated on data from the same session. It experiences only within-session noise and does not account for potential inter-session noise, should it be re-used for future prediction of profiles. However, in the practical setting of clinical screening, new samples may be processed on-the-fly, each at a different time and therefore experiencing unanticipated amounts of inter-session variability. Concerns about this inter-session reproducibility is related primarily to concerns over generalizability of predictive models that are extracted from past data sessions to profiles obtained in the future. We will analyze this aspect of the problem by learning predictive models that are tested on profiles from one target (test) session and trained on the profiles from the remaining three (training) sessions and by comparing them to the 'ideal' model trained and tested on the profiles from the same session.

We perform this analysis as follows. A target (test) session is chosen from the available four sessions. The remaining three sessions are used to train a (future) predictive model. Next, samples are divided via the random subsampling approach to training and testing samples, such that 30% of the samples are in the test sample set. The remaining samples are represented in the training sample set. Next, we generate 1000 multi-session training datasets by assigning each patient in the training sample set a profile from one of its training sessions and learn the models for each dataset. The models are tested on the test session samples and their accuracies define the distribution of (future) test accuracy scores for mixed-session data. The mean of the distribution is then compared to the accuracy achieved by the model trained on the same session as the test session. To provide additional assurance we repeat everything using 30 different train-test sample splits and average the results. This will let us compare the average future performance of mixed-session models to the 'ideal' model for one test session. The global performance can be assessed by averaging the results for four test sessions.

In our first comparison of (a) models trained on profiles from the three training sessions, versus (b) an 'ideal' model trained on profiles from the same session as the testing set, we expect the 'ideal' models to outperform the multi-session-trained models. Inter-session variability is not present in the ideal model and is therefore expected to cause a loss of performance. Our second aim is to compare models from group (a) versus (c) models trained on profiles from a single session other than the target session. The objective is to determine if predictive models trained on multi-session data can learn to adapt to inter-session noise and hence improve their performance when compared to models learned on single sessions.

We repeat the setup in the previous experiment to estimate the distribution of accuracy scores for the 1000 models trained on multi-session data. Accuracy scores are also obtained from models trained on one of the three single sessions. The difference between the mean accuracy of the multi-session models and single-session models are kept for a total of 3 differences. This process is repeated 40 times for each of the four target sessions. We repeat the hypothesis test to determine if the mean of these differences differs significantly from 0. In the case where multi-session models have the same generalization performance as single-session models, the mean of this distribution should not differ significantly from 0.

3 APPLICATION TO LUNG CANCER SERUM MASS SPECTRA

3.1 Signal reproducibility

We first examined whether proteomic spectra are reproducible across multiple sessions. We used the random regrouping test described in Section 2.6.1 to evaluate whether the signals from the same sample were more similar than signals from randomly chosen sample sets. Since we expect to find differences between case and control samples, this score was evaluated separately on respective subgroups of case and control spectra.

The histogram in figure 2 (left) indicates the average signal difference score (over all 59910 intensity measurements of the profiles) for the 21 cancer patients across all 4 sessions. A distribution of 1000 averages of 21 signal difference scores for randomly selected quadruplets of spectra is plotted as a reference. The score for the replicate spectra falls outside of the score distribution for randomly grouped spectra. The same behavior occurs with the control samples. Furthermore, we can assess the reproducibility of signal difference over a small region of the profile. The right panel of figure 2 displays the distribution of signal difference scores for the peak region at 8228 Da. The peak is less different among profiles from

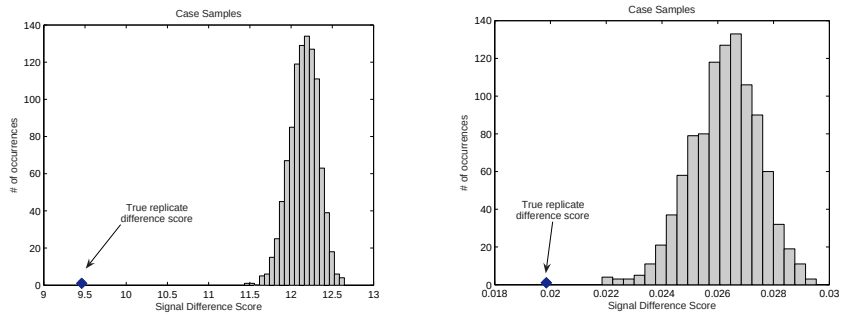

Fig. 2. Distributions of signal difference scores for random groupings of profiles for case samples. The left panel displays signal difference scores taken over the entire range of the signal (59910 features), while the right panel displays signal difference for a single feature at 8228 Da. The signal difference score for the true replicate spectra is plotted as a dot along the x-axis. The signal difference among the true replicates is much less than any observed signal difference among randomly grouped profiles. This shows that the observed greater similarity between a sample's true replicates is less likely to be due to random effects.

the same sample than from profiles randomly assigned to a sample. There is a statistically significant difference between the signal difference scores obtained from true and random replicates, at both the global and local (peak) signal level. This assures us that profiles from the same sample do not exhibit so much difference that they can be easily confused with profiles from a different sample. This encouraging result shows the reproducibility of proteomic profiles at the signal level.

3.2 Reproducibility of discriminative signals

We use the randomization framework from section 2.6.2 to determine whether differential expression scores obtained from mixed session data differ on average from the differential expression mined from single session datasets. These differences may assess the benefit or loss due to mixed-session analysis.

Figure 3 (left) displays the empirical distribution of differential expression scores for multi-session data of one prominent peak in the spectra. The distribution was obtained from 1000 random datasets such that each patient was randomly assigned a profile from one of the four sessions. The four marks indicate the differential expression scores obtained for profiles in four individual sessions.

We next determined the significance of these differences. To determine the amount of noise experienced over a range of features, we similarly analyzed the top 100 differentially expressed peak regions in the profiles. The mean was calculated for every feature's differential expression score distribution, as well as the score of the feature in the four single sessions. These four scores were subtracted from the mean and kept for each feature, resulting in a distribution of 400 differences. Figure 3 (middle) displays this distribution. If single session scores were biased (that is, better scores are produced by the single session analysis) we would expect to see the mean of this distribution to differ significantly from 0. In other words, we would expect to reject (at some significance level) the null hypothesis: the mean of differences is ≥ 0. Indeed, the mean of the distribution of differences was -0.0351, giving a p-value of 5.588×10^{-8} for the one-sided t-test, which leads to the rejection of the null hypothesis. Hence the amount of differential expression in single sessions appears to be better on average than in mixed-sessions. This shows that inter-session variability affects the measured differential information.

We expect this negative result to affect the performance (accuracy) of predictive models trained on multi-session data. The question is how big the effect really is. Earlier research studies considered it most ideal to learn from and evaluate their predictive models on data from a single session. We therefore compare the differences in accuracy between models trained on multi-session data versus models trained on single-session data.

Following the methods in section 2.6.2, we analyzed the accuracy of multi-session models versus single-session predictive models. Figure 3 (right) displays the distribution of differences between mean accuracies of multi-session and single-session predictive models. If better accuracies are achieved by predictive models for single-session data, we would expect to see the mean of this distribution to be below 0. Indeed, the mean of the distribution of differences was -0.0267 which once again indicates a loss that can be explained by additional inter-session variability. To confirm the difference we used a repeated resampling experiment proposed by Nadeau and Bengio (2003), estimating the 95% confidence interval around the mean of differences to be -0.0267 ± 0.0001. This experiment confirmed that this difference is indeed significant.

On average, there is about a 2.7% drop in accuracy when using multi-session data, demonstrating a relatively small (average) loss of reproducibility of multivariate discriminative patterns across multiple sessions. One should understand that this is an average assessment; the performance of an individual classifier may vary from session to session and also depends on how profiles from multiple sessions are mixed. On average, these mixed session models perform very well, achieving $84.27 \pm 2.15\%$ accuracy.

3.3 Generalization performance

Finally, we want to determine the effect the multi-session training has on predictive models which must generalize well to future, unseen profiles and sessions. The previous result demonstrates that intersession noise exists, but does not seem to greatly affect the performance of predictive models on average. However, the analysis used each session and did not try to assess the performance on future sessions. We use the methods in section 2.6.3 to analyze whether training predictive models on multi-session data generalizes well to profiles in future sessions and compare the performance of these models to 'ideal' predictive models trained and tested on single session data.

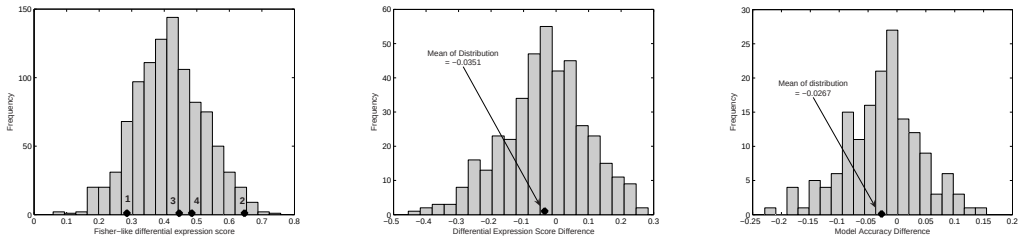

Fig. 3. Left panel: Distribution of differential expression scores under random regroupings of profiles for the peak region at 12.938 kDa. The differential expression score for the peak in each of the 4 individual sessions is plotted as a dot along the x-axis. Middle panel: distribution of differences between the mean of mixed-session Fisher score distributions and single session Fisher scores for 100 peak regions. The distribution has a mean of -0.0351 and p-value of 5.588×10^{-8} for the null hypothesis: the mean is equal to 0. Right panel: distribution of differences between the mean accuracies of models trained on multi-session data and accuracies of models trained on single-session data. The mean of this distribution falls below 0 (= -0.0267), indicating an on-average benefit of training from single-session data.

Figure 4 (left) displays a distribution of accuracy differences between the average of 1000 predictive models built from random multi-session training data and models trained on data that came from the session on which the model was tested. The mean of the distribution is -0.0231 which quantifies an overall average generalization accuracy loss one may expect to see by training the model on the mixed session data as opposed to the accuracy of the 'ideal' model. We analyzed the difference using an additional resampling test (Nadeau and Bengio, 2003) to compute the 95% confidence interval of the mean. The result of the mean falling within -0.0286 ± 0.0001 confirmed the difference is statistically significantly different. However, in terms of absolute numbers the accuracy loss with respect to the ideal model is not bad.

In a practical setting such as clinical screening, the training data will certainly not come from the same session as the testing session. This eliminates the possibility of having an 'ideal' predictive model. We repeated the previous experiment by examining the differences between the multi-session models and models trained on profiles from a single session other than the target session. This differs from the previous experiment since the single-session models lose the advantage of the 'ideal' environment. Inter-session noise must now be accommodated by both the multi-session and single-session-trained models.

Figure 4 (right) displays a distribution of accuracy differences between the average of 1000 predictive models trained on multi-session data and models trained on the remaining single sessions. The mean and 95% confidence interval of this distribution falls above 0 ($= 0.0289 \pm 0.0001$), indicating a benefit of training on multi-session data. The confidence interval is again computed using the repeated resampling test (Nadeau and Bengio, 2003), which confirmed the difference to be statistically significantly different. This result illustrates how training on multi-session data can allow the model to adapt to inter-session noise. The better a predictive model can adapt to inter-session noise, the more reproducible the performance will be on future data.

4 DISCUSSION

The objective of our paper was to investigate effects of inter-session variability of MS proteomic profiling data for serum samples obtained for early lung cancer patients and healthy control subjects over multiple (four) data-generation sessions. Such studies are critical for the acceptability of the MS profiling technology in clinical settings.

Through our experiments, we try to understand the global impact made by including data from multiple sessions on the accompanying analysis for the profiles. Since profile reproducibility is naturally imperfect, we expect some amount of inter-session variability to exist. Different sessions will yield different results, and the best choice of sessions to use is unclear. However, through averaging, we can quantify the effect of inter-session variability on our analysis and determine whether data from multiple sessions is useful.

First, our experiments show that the similarity among profiles for the samples reprocessed during different time periods is greater than that due to random chance. Second, we show that the discriminative information in profile peaks obtained from any single session is on average higher than the information mined from multi-session data. Third, we reinforce this result by performing multivariate classification analysis and by showing that classifiers tested on single-session data are better on average if they are trained on data generated from the same session as opposed to classifiers trained on multi-session data. However, our experiments showed no drastic loss of performance from training predictive models with multi-session data. Finally, we show that training classifiers on data from multiple sessions generalizes at minimal accuracy loss (about 2.86% loss with respect to the ideal classifier) to profiles generated in future sessions.

The samples selected for this study showed more robust differential features in their February 2004 spectra. This selection was made to achieve the goals of a concurrent inter-site validation study. We note that this selection may lead to an improved classification performance on the February 2004 spectra; data from other sessions are unaffected. Thus, average accuracy differences between single and multi session classifiers reported in the paper may become biased towards single session classifiers as compared to those expected under a fully random sample selection process. This does not affect our conclusions about effects of intersession variability (Sections 3.2 and 3.3), since promising differences reported in the paper would, under this bias, be obtained under less favorable settings.

The intersession reproducibility of profiles, both within and across sessions, can be influenced by many factors. Experimental conditions involving sample preparation and preprocessing should remain as consistent as possible. A careful study design free of confounding and standardized protocols for sample processing can help

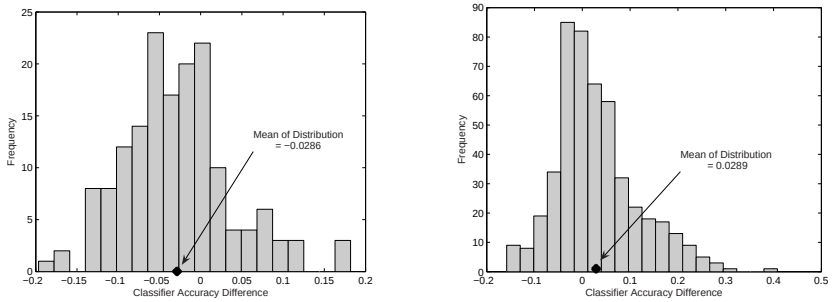

Fig. 4. Left panel: distribution of accuracy differences between predictive models trained on multi-session data and models ideally trained on data from the same single session as the target test session. The mean below 0 (= -0.0286) indicates an advantage of the ideally trained single-session models. Right panel: distribution of accuracy differences between the same predictive models trained on multi-session data and models trained on single-session data from sessions other than the target testing set. The mean above 0 (= 0.0289) indicates an advantage of training on multi-session data. This illustrates the ability of predictive models trained on multi-session data to adapt to inter-session noise.

to reduce sources of variability. All data analyzed in this study were obtained through standardized protocols implemented and validated in accord with an NCI EDRN validation study assessing the reproducibility of the SELDI-TOF-MS platform (Semmes *et al.*, 2005). The promising results reported here were achieved thanks to the strict adherence to these protocols.

Our study was based on serum sample aliquots stored over varying amounts of time. This raised concerns over possible sample degradation effects and the reproducibility of the information one can extract from them (Ranganathan *et al.*, 2006). However, a study by Grizzle *et al.* (2005) showed this effect to be relatively small. We confirmed these results indirectly by observing and measuring only a small amount of average inter-session variability. Additionally, if the degradation of samples was significant, one would expect too see its signs over time. In our case, data generated in February 2004 and January 2005 appear to exhibit stronger discriminative signals than data from sessions produced in June 2003 and November 2004. Thus, no immediate temporal relationship between time of processing and signal strength could be drawn, and observed inter-session differences are likely due to other causes.

This study aimed to assess the global (average) effects of inter-session variation on the reproducibility of profiles and their signals. We did not try to investigate and pinpoint profile regions that appear to be most vulnerable to inter-session variation. However, the fact that classifiers trained on multi-session data were able to adapt to inter-session biases suggest that such regions may exist. Future investigations of these relations may give additional insight on processes critical for the application of the MS profiling technology and may lead to further improvements in its reproducibility.

ACKNOWLEDGEMENT

This research was supported by Department of Defense grant USAMRAA W81XWH-05-2-0066, NCI grant P50 CA090440-06 and NLM grant 5 T15 LM007059-20. The data for this publication were made possible by Grant Number 1 UL1 RR024153 from the National Center for Research Resources (NCRR), a part of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NCRR or NIH.

REFERENCES

Baggerly, K. A., Morris, J. S., and Coombes, K. R. (2004). Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, **20**(5), 777–785.

Diamandis, E. P. (2003). Point: Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? *Clin Chem*, **49**(8), 1272–1275.

Efron, B. (1987). *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial & Applied Mathematics.

Good, P. (1994). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer.

Grizzle, W., Semmes, O., Bigbee, W., Zhu, L., Malik, G., DK, O., B, M., and U, M. (2005). The need for review and understanding of seldi/maldi mass spectroscopy data prior to analysis. *Cancer Informatics*, **1**.

Hauskrecht, M., Pelikan, R., Malehorn, D. E., Bigbee, W. L., Lotze, M. T., Zeh, H. J., Whitcomb, D. C., and Lyons-Weiler, J. (2005). Feature selection for classification of seldi–tof–ms proteomic profiles. *Appl Bioinformatics*, **4**(4), 227–246.

Hauskrecht, M., Pelikan, R., Valko, M., and Lyons-Weiler, J. (2007). Feature selection and dimensionality reduction in genomics and proteomics. In W. Dubitzky, M. Granzow, and D. Berrar, editors, *Fundamentals of Data Mining in Genomics and Proteomics*, pages 149–172. Springer.

Nadeau, C. and Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, pages 239–281.

Pelikan, R., Lotze, M., Lyons-Weiler, J., Malehorn, D., and Hauskrecht, M. (2004). *Serum Proteomic Profiling and Analysis*. Elsevier.

Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, **359**(9306), 572–577.

Ranganathan, S., Polshyna, A., Nicholl, G., Lyons-Weiler, J., and Bowser, R. (2006). Assessment of protein stability in cerebrospinal fluid using surface-enhanced laser desorption/ionization time-of-flight mass spectrometry protein profiling. *Clinical Proteomics*, **2**, 91–101.

Ransohoff, D. F. (2005). Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer*, **5**(2), 142–149.

Semmes, O. J., Feng, Z., Adam, B.-L., Banez, L. L., Bigbee, W. L., Campos, D., Cazares, L. H., Chan, D. W., Grizzle, W. E., Izbicka, E., Kagan, J., Malik, G., McLerran, D., Moul, J. W., Partin, A., Prasanna, P., Rosenzweig, J., Sokoll, L. J., Srivastava, S., Thompson, I., Welsh, M. J., White, N., Winget, M., Yasui, Y., Zhang, Z., and Zhu, L. (2005). Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. *Clin Chem*, **51**(1), 102–112.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Zhang, Z., Bast, R. C. J., Yu, Y., Li, J., Sokoll, L. J., Rai, A. J., Rosenzweig, J. M., Cameron, B., Wang, Y. Y., Meng, X.-Y., Berchuck, A., Van Haaften-Day, C., Hacker, N. F., de Bruijn, H. W. A., van der Zee, A. G. J., Jacobs, I. J., Fung, E. T., and Chan, D. W. (2004). Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. *Cancer Res*, **64**(16), 5882–5890.