Automatic Selection of Preprocessing Methods for Improving Predictions on Mass Spectrometry Protein Profiles

Richard C. Pelikan, MS, Milos Hauskrecht, PhD[2]
[1, 2]Departments of Biomedical Informatics, [2]Computer Science and [1,2]Intelligent Systems, University of Pittsburgh, Pittsburgh, PA

Abstract

Mass spectrometry proteomic profiling has potential to be a useful clinical screening tool. One obstacle is providing a standardized method for preprocessing the noisy raw data. We have developed a system for automatically determining a set of preprocessing methods among several candidates. Our system's automated nature relieves the analyst of the need to be knowledgeable about which methods to use on any given dataset. Each stage of preprocessing is approached with many competing methods. We introduce metrics which are used to balance each method's attempts to correct noise versus preserving valuable discriminative information. We demonstrate the benefit of our preprocessing system on several SELDI and MALDI mass spectrometry datasets. Downstream classification is improved when using our system to preprocess the data.

Introduction

Mass spectrometry (MS) protein profiling[1] is an analytical technique used to produce protein expression profiles from a biofluid such as serum, urine or saliva. The resulting protein expression profiles contain thousands of measurements of protein molecules and their relative abundances within the biofluid sample. Statistical machine learning techniques are applied to this data to create predictive models which have been used successfully to find diagnostic patterns. The diagnostic potential of this technology has drawn interest from the clinical community for screening of diseases such as cancers[2,3,4], diabetes[5] and Alzheimer's[6]. Accordingly, techniques which prepare the data, standardize and simplify the analytical process are also in demand, and are necessary to facilitate the introduction of such novel technologies into clinical workflows.

The MS equipment is subject to both artificial and biological sources of variation. Many of the artificial sources have been documented and studied[7,8]. With the growing interest in protein profiling technology, many labs created their own preprocessing routines to deal with the noise artifacts as manifested in their own data. While several good preprocessing routines were developed, it became less clear which methods to use, and when. The uncertainty about which kind of noise to expect in future data makes it difficult to anticipate which methods will work best, impeding development of a standard preprocessing system.

Research in comparing preprocessing methods for MS data is limited[9,10,11]. Some studies compare preprocessing based on how consistently "peak" features are retained in the preprocessed data. Others evaluate preprocessing in conjunction with the performance of predictive models trained with variously preprocessed data. While this facilitates the comparison of preprocessing methods, performance on a testing set should not be the only metric for selecting among methods. This introduces bias into the learning process and promotes methods which may overfit the data to the classifier. Instead, we can measure how each stage of a preprocessing system addresses its targeted noise sources, and optimize these metrics together with other quality metrics, such as data discriminability and reproducibility.

This work introduces a system which automatically selects appropriate methods for preprocessing mass spectrometry data. To address the problem of how to select the methods, we propose a number of preprocessing metrics which quantify how well the objectives of each preprocessing stage are met by individual methods. We define two types of metrics: *stage-wise* metrics reflect goals of individual stages, e.g. noise reduction, while *global* metrics help us to define overall objectives for the preprocessing. These metrics are balanced against each other, and the methods making acceptable tradeoffs between them are selected. These preprocessing metrics are a key advantage which enables the method selection process to be data-driven and facilitates the overall analysis. As in previous work[9,10], we evaluate our preprocessing system through classification performance by a predictive model. In our work, the predictive model makes a binary decision between presence and absence of a condition, e.g. lung cancer.

The next section describes each stage of preprocessing, their stagewise metrics, and the mechanisms used to select a method for each stage. The following section demonstrates the effectiveness of the system on several datasets comprised of data for various diseases. The conclusion summarizes the

advantages of this system, as well as additional areas for application and future directions for research.

Methods

We use the following notation and terminology to discuss protein profiles. A single proteomic mass spectrum, or *profile*, consists of *d features*. Each feature f_j is a datapoint (x_j,y_j) where y_j is the relative abundance of a molecule with molecular weight x_j in the sample. A *dataset* is a collection of profiles from different samples with n_- control (healthy) and n_+ case (diseased) profiles. The average and standard deviation of feature f_j over the dataset are given by μ_j and σ_j respectively. These statistics can be restricted to case or control profiles: μ_{j+}, μ_{j-}, σ_{j+}, σ_{j-}.

Preprocessing metrics

Our objective is to optimize the preprocessing of MS profiles. We are helped by multiple metrics (scores) that aim to quantify the objectives of preprocessing and compare differences in outcomes of various preprocessing methods. We first propose a *global* metric, the *differential ratio score*, which aims to retain discriminative information in profiles as much as possible. Next, we focus on proposing special *stage-wise* metrics defining the objectives of individual preprocessing stages.

We make the assumptions that discriminative information exists in proteomic profiles, and that the best preprocessing methods disturb this information as little as possible. Ideally, a preprocessing method should attempt to minimize the differences between profiles from the same class, while maximizing the differences between profiles from different classes.

The *Differential Ratio* (DR) score is a metric for estimating the strength of the dataset's discriminative signal: $DR = \text{diff}_{\text{inter}} / \text{diff}_{\text{intra}}$, where

$$\text{diff}_{\text{inter}} = \sum_{j=1}^{d} \left(\mu_{j+} - \mu_{j-}\right)^2$$

$$\text{diff}_{\text{intra}} = \sum_{i=1}^{n_+} \sum_{j=1}^{d} \left(\mu_{j+} - f_j\right)^2 + \sum_{i=1}^{n_-} \sum_{j=1}^{d} \left(\mu_{j-} - f_j\right)^2$$

The DR score is a global metric which is common across all stages of preprocessing. Below, we describe 5 commonly performed stages of preprocessing and the metrics which individually judge how well each stage is being performed.

Variance Stabilization is the initial stage in our preprocessing routine. Its purpose is to decouple the dependency of a feature's variance on its mean. This property is commonly referred to as *heteroscedacity*. Reducing heteroscedacity in the data is important since many statistical techniques assume covariates come from distributions with constant variance. Popular techniques for variance stabilization include taking the logarithm or n^{th} root of the data[12].

It is well-known that least-squares fitting does not work well in the presence of heteroscedacity. This can be seen in the large estimates of residuals if linear regression is applied to the feature variances and means. Thus, if a variance stabilization method performs better, the overall sum of residuals should be smaller. Moreover, the slope of the regression line should also be more horizontal, indicating less of a correlation between the mean and variance of covariates. The Heteroscedacity Retention (HR) score is computed as follows:

$$\boldsymbol{HR} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{d}} \left(\boldsymbol{R_j} + \boldsymbol{R_j} * \boldsymbol{m_{reg}}\right), \text{where}$$

$$\boldsymbol{R_j} = \left\| \boldsymbol{m_{reg}} * \boldsymbol{x_j} + \boldsymbol{b_{reg}} - \boldsymbol{y_j} \right\|$$

The terms m_{reg} and b_{reg} are the slope and bias resulting from linear regression of x on y. Examples of methods which are available to our system include the log and several n^{th}-root transformations, the generalized log transformation[12] and the AVAS transformation[13].

Baseline Correction refers to the correction of a constant background noise which raises the intensity of nonexistent measurements above zero. The observed effect is a vertical shift in the profile. These shifts can artificially increase the apparent differential expression of features, which can cause a predictive model to falter on data with a more correct baseline.

Signal-to-Noise Ratio (SNR) is a statistic used to estimate the proportion of signal governed by noise. The baseline should be a relatively noiseless, consistent signal, thus it should have high SNR. It should also not detract from the SNR of the original signal, as it should still contain delicate biological information. The baseline signal-to-noise (bSNR) score estimates the average SNR over the baseline-corrected data:

$$bSNR = \frac{1}{d} \sum_{j=1}^{d} \frac{\mu_{cj}}{\sigma_{cj}} * \frac{\mu_{bj}}{2\sigma_{bj}}$$

The c and b subscripts refer to terms calculated using the baseline-corrected profile, and the subtracted baseline, respectively. The bSNR score penalizes baseline methods which estimate widely varying baselines. A greater score hints that the proportion of signal to noise in the corrected profile is still greater than that removed in the baseline.

Normalization refers to adjusting profiles so that they appear to come from the same scale. Even profiles from the same class may appear as though they were generated from a different range of values. Common normalization techniques include rescaling feature values to the [0 1] range, or by scaling them by the total sum of all intensities in a profile.

Normalization aims to improve the homogeneity of profiles, and thus the DR score is a natural metric for it. Since all stages need a stagewise metric, we can balance the DR score with nearly any sensible metric. In this work, we reuse the HR score to grade normalization methods.

Smoothing refers to the correction of high-frequency noise which peppers every feature of the signal. Smoothing techniques characterize this noise as a separate additive component which either coincides with or deviates from a functional form.

The smoothing signal-to-noise (sSNR) score estimates to what degree the SNR of the profile improves after smoothing. It is calculated simply as the average SNR of all features:

$$sSNR = \frac{1}{d}\sum_{j=1}^{d}\frac{\mu_{cj}}{\sigma_{cj}}$$

The score increases with improved smoothing, which should remove only noise.

Alignment refers to the process of shifting profiles along the x-axis so that "peak" features in each profile appear at common points on the x-axis. Most alignment techniques work by identifying salient features in each profile that should coincide with the same point on the x-axis. Individual profiles are stretched and shrunk to optimize the alignment of these features, with a penalty function weighting the severity of each necessary adjustment.

The coefficient of variation (CoV) is the inverse of the SNR. It measures the dispersion of a variable and should be smaller when the measurements are more equal. The average CoV can be used as a metric to grade alignment quality. At the time of this work, the system supports alignment, but this feature was disabled in favor of allowing more time to the other stages of preprocessing, as the numbers of methods they evaluate are greater. See (**Table 1**) for an abridged list of methods available to our system for each stage.

Selection of Methods

We use a device called the Stagewise Performance (SP) curve to select a method by balancing stagewise metrics versus global metrics.

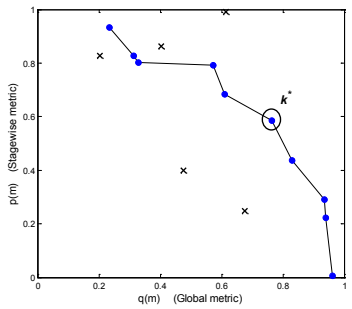

Figure 1. Example of SP curve with $\omega = 0.5$. Points on the "curve" and crosses represent evaluated methods for a stage of preprocessing. Crosses indicate methods which do not retain at least 80% of the peak features, and are therefore not considered. The selected method (circled) optimizes the weighted combination of stagewise (p) and global metrics (q).

Stage	Methods / References
Variance Stabilization	n^{th}-root, log[12], generalized log[12], ACE, AVAS[13]
Baseline Correction	Gamma[14], Monotone[8], Kernelized, Local minimum[15]
Normalization	Total Ion Current, quantile methods[16], 0-1 rescaling, MATLAB routines[17]
Smoothing	Fourier, Savitzky-Golay[18], Moving avg/median/geometric mean, wavelet decompositions[19,20]
Alignment	PAFFT/RAAFT[21], Dynamic time-warping, MATLAB routines [17]

Table 1. Abbreviated list of methods evaluated by our automatic preprocessing system. References are indicated for most methods.

For a single pre-processing stage with k competing methods, let p be the stagewise metric for that stage and q be the global metric over the entire preprocessing. Let m_k be the data resulting from preprocessing method k. The SP curve chooses the method k^* maximizing the following function:

$$k^* = \max_k(\omega\, q(m_k), (1-\omega)p(m_k))$$

where $0 \le \omega \le 1$ is a parameter introduced to weight the influence of the stagewise versus global metric. In this work, we set ω = ½ for every stage. (**Figure 1**) demonstrates the technique under these circumstances. To prevent trivial inflation of the DR score, methods which fail to reproduce at least 80%

of the peaks in the profile are not considered part of the "curve". The selected method is then chosen from the remaining points. The effect of ω on the preprocessing system has not yet been studied.

In addition to the competing methods, each stage is provided a "do-nothing" method, which acts as if no preprocessing is done for that stage. This way, the system allows itself to skip stages if all of the methods should score extremely poorly.

The preprocessing system proceeds through the 5 stages in the order above, using the SP curve to pick a single method per stage. The output of the system is the preprocessed data and the list of methods used to generate it. This allows additional data from the same generation to be preprocessed in the same manner.

Experiments and Results

We evaluated our system on four cancer datasets produced by Surface-Enhanced Laser Desorption Ionization (SELDI) Time-of-Flight (TOF) MS and two datasets produced by Matrix-Assisted (MALDI) TOF-MS technology. The MALDI data source exhibits different noise characteristics and can test the adaptability of our preprocessing system. In each dataset, case and control samples were matched based on clinical characteristics such as age, gender and smoking history. Our system requires that profile datasets are presented in a single text-file format, with each row containing delimited values for the features in that profile. Our experiments were performed on a Windows desktop with 2GB RAM.

We compared our automatic preprocessing system with three other preprocessing systems: without preprocessing, a baseline preprocessing procedure which has previously been published[15], and the PrepMS[22] preprocessing system. The baseline procedure was developed at that time to perform well based on empirical results.

Each dataset was split randomly into training and testing sets at a 70%/30% proportion. Our system is applied to the training data only, and retains only the five selected methods used for preprocessing. These five methods are then applied to the testing set. The other systems preprocess the entire dataset at once.

Each preprocessing system is evaluated by measuring the performance of a predictive model trained on the preprocessed training set and evaluated on the preprocessed testing data. Performance is measured in terms of area under the Receiver Operating Characteristic Curve (AUC)[23] for the classifier. Since our intention is not to evaluate the predictive model, we choose only a very basic classifier: a linear Support Vector machine. No feature selection is performed. We average the results from these experiments over 40 different train/test splits.

Data set \ method	None	Base	PrepMS	Auto
Breast Ca.	0.577	0.566	0.556	**0.622**
COPD	0.614	0.640	0.628	**0.670**
Panc. Ca.[15]	0.835	**0.900**	0.884	0.862
Lung Ca.	0.837	0.830	0.800	**0.880**
Diabetes	0.632	0.610	0.659	**0.757**
Liver Ca.[25]	0.965	0.963	0.967	**0.969**

Table 2. AUC of predictive models trained on variously preprocessed datasets from SELDI (top four rows) and MALDI (bottom 2 rows) platforms. In most cases, our automatic method selection system outperforms the competing systems. An AUC of 1.0 would indicate perfect classifier performance.

The average AUC achieved by predictive models on our SELDI-TOF and MALDI-TOF datasets is shown in (**Table 2**). In most cases, our automatic preprocessing method results in the best performance of predictive models. The pancreatic cancer dataset is an exception – since the baseline system was developed for empirical performance on this dataset, it may explain why it outperforms the automatic selection method. This suggests that each MS platform (and therefore each dataset) may have an unique optimal preprocessing method available. Likewise, many inappropriate methods may exist for a dataset. Static preprocessing methods such as our "baseline" method and PrepMS demonstrate that their effects on the breast and lung cancer datasets can be detrimental to discriminative signal.

Our preprocessing system is able to intelligently select those methods that will enhance discriminative signal, and this is seen in the improved AUC of predictive models learning from the automatically preprocessed data. In spite of the variation in noise sources and platform characteristics, our system is able to improve upon both SELDI and MALDI types of data without any adjustment from the user. Additional data platforms can easily be accommodated by adding preprocessing routines particular to that platform, and allowing the system to judge whether they are truly appropriate for enhancing discriminative signal while carefully removing noise.

Conclusion

We have presented a system for automatically selecting preprocessing methods to be used for

proteomic mass spectra. When no optimized preprocessing is available for a dataset, our system typically outperforms other "standard" methods for preprocessing, in terms of the performance of predictive models trained on the preprocessed data.

This system's primary advantage is the metric-based system for selection of methods. The stagewise versus global metric tradeoff contrasts other studies which only use a single metric, e.g. classification performance, to assess all stages of preprocessing. Our system can adjust to unexpected changes in the nature of noise, which becomes increasingly important as new data production technologies and protocols emerge. The system can also be applied to data requiring similar preprocessing, such as microarray and MRI images. New preprocessing techniques can always become part of the system, further enhancing it. Determining the appropriate weights per stage in the SP curve function is an intended future direction, and should also improve this system.

References

1. Callesen AK, Madsen JS, *et al.* Serum protein profiling by solid phase extraction and mass spectrometry: A future diagnostics tool? Proteomics. 2009;9:1428-41.
2. Semmes OJ, Feng Z, *et al.* Evaluation of serum protein profiling by seldi-tof-ms for the detection of prostate cancer. Clin Chem. 2005;51:102-12.
3. Watkins B, Szaro R, *et al.* Detection of early-stage cancer by serum protein analysis. American Laboratory. 2001;33:32-36.
4. Zhang Z, Bast R, , Jr, *et al.* Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. Cancer Res. 2004;64:5882-90.
5. Bergsten P. Islet protein profiling. Diabetes, Obesity and Metabolism. 2009;11:97-117.
6. Lopez MF, Mikulskis A, *et al.* High-resolution serum proteomic profiling of alzheimer disease samples reveals disease-specific, carrier-protein-bound mass signatures. Clin Chem. 2005;51:1946-54.
7. Malyarenko DI, Cooke WE, *et al.* Enhancement of sensitivity and resolution of seldi-tof mass spectrometry records for serum peptides using time-series analysis techniques. Clin Chem. 2005;51:65-74.
8. Coombes KR, Koomen JM, *et al.* Understanding the characteristics of mass spectrometry data through the use of simulation. Cancer Inform. 2005;1:41-52.
9. Wegdam W, Moerland PD, *et al.* Classification-based comparison of pre-processing methods for interpretation of mass spectrometry generated clinical datasets. Proteome Sci. 2009;7:19.
10. Floros X, Spyrou G, *et al.* Study on preprocessing and classifying mass spectral raw data concerning human normal and disease cases. 2006;390-401.
11. Cruz-Marcelo A, Guerra R, *et al.* Comparison of algorithms for pre-processing of seldi-tof mass spectrometry data. Bioinformatics. 2008;24:2129-36.
12. Durbin BP, Hardin JS, *et al.* A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics. 2002;18 Suppl 1:S105-10.
13. Tibshirani R. Estimating optimal transformations for regression via additivity and variance stabilization. Journal of the American Statistical Association. 1998;83:394-405.
14. Cytospec™ version 1.0. Lasch, P., 2004
15. Hauskrecht M, Pelikan R, *et al.* Feature selection for classification of seldi-tof-ms proteomic profiles. Appl Bioinformatics. 2005;4:227-46.
16. Meuleman W, Engwegen JY, *et al.* Comparison of normalisation methods for surface-enhanced laser desorption and ionisation (seldi) time-of-flight (tof) mass spectrometry data. BMC Bioinformatics. 2008;9:88.
17. Matlab version 2009b. Natick, MA: The Mathworks Inc., 2009
18. Savitzky A and Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry. 1964;36:1627-39.
19. P D, Lin SM, *et al.* Application of wavelet transform to the ms-based proteomics data preprocessing. BIBE 2007. 2007;680-86.
20. Coombes KR, Tsavachidis S, *et al.* Improved peak detection and quantification of mass spectrometry data acquired from seldi by denoising spectra with the undecimated discrete wavelet transform. Proteomics. 2005;5:4107-17.
21. Wong JW, Cagney G, *et al.* Specalign--processing and alignment of mass spectra datasets. Bioinformatics. 2005;21:2088-90.
22. Karpievitch YV, Hill EG, *et al.* Prepms: Tof ms data graphical preprocessing tool. Bioinformatics. 2007;23:264-5.
23. Hanley JA and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology. 1982;143:29-36.
24. Schölkopf B and Smola AJ. Learning with kernels : Support vector machines, regularization, optimization, and beyond. Adaptive computation and machine learning. 2002.
25. Goldman R, Ressom HW, *et al.* Candidate markers for the detection of hepatocellular carcinoma in low-molecular weight fraction of serum. Carcinogenesis. 2007;28:2149-53.