

Formative Evaluation of a Prototype System for Automated Analysis of Mass Spectrometry Data

N. Fananapazir¹, M. Li Ph.D.², D. Spentzos, M.D.³, C.F. Aliferis M.D., Ph.D.¹

¹Department of Biomedical Informatics, Vanderbilt University, Nashville, TN

²Biostatistics Core, Vanderbilt Ingram Cancer Center, Nashville, TN

³Beth Israel Diacones Medical Center, Harvard University, Boston, MA

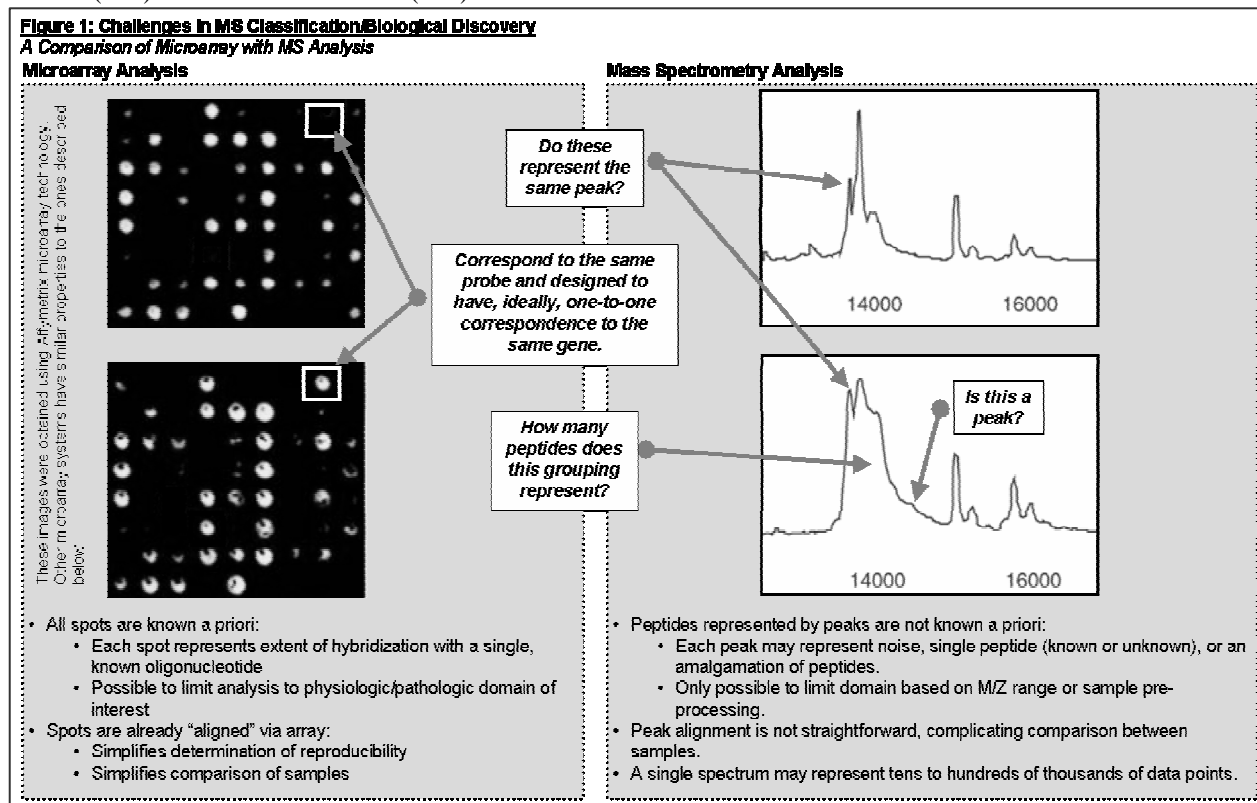
Abstract

Mass Spectrometry (MS) is emerging as a breakthrough mass-throughput technology capable of producing powerful clinical diagnostic and prognostic models and of identifying important disease biomarkers. Few individuals possess the necessary skills to carry out MS analyses competently, and access to such individuals is limited in most settings, hindering progress in this field. We seek to ease this burden by creating a fully automated system (FAST-AIMS) capable of analyzing mass spectra to produce high-quality diagnostic and outcome prediction models and identify related biomarkers. In the present report we introduce the system and conduct a formative evaluation in which 6 users apply it to a challenging dataset. FAST-AIMS' performance is compared to that of an expert statistician as well as to a previously published analysis by an independent group. In our experiments FAST-AIMS when used by both MS-sophisticated users (n=4) and naïve users (n=2) achieves

performance (a) comparable to our human expert, and (b) superior to the previously published manual analysis; in addition (c) the system's estimates future performance accurately, thus avoiding overfitting.

Background and Motivation:

Mass Spectrometry (MS) is a widely used technology capable of discriminating proteins and their post-digestion peptide products on the basis of mass and charge. Within the last three years, several research groups have explored the use of MS for clinical applications in the broad areas of early cancer detection, clinical diagnosis and clinical outcome prediction. Domains involve a variety of tissue types – blood serum, tissue biopsy, nipple aspirate fluid, pancreatic juice – in the analysis of a variety of cancers – ovarian, prostate, renal, breast, head and neck, lung, laryngeal, hepatic, cervical, pancreatic, colorectal, bladder – and non-cancers – hepatitis, and cerebrovascular accidents¹. Published reports indicate remarkable potential for this technology to



diagnose disease with minimally invasive testing procedures, low cost, and – in some cases – with unprecedented accuracy. It is expected by many that MS together with gene expression microarrays and related mass-throughput technologies will revolutionize medicine in the near future.²

The analysis of mass throughput data such as microarray data has involved severe modeling and statistical challenges, the most notable of which include multiple sources of data noise, very large numbers of predictor variables, lack of consistency in data-generating platforms, and small sample sizes³. The analysis of MS data introduces further analytic challenges, requiring additional pre-processing steps, the most notable of which are baseline correction of spectra, the detection of peaks corresponding to proteins and their peptide products, the alignment of peaks across spectra, the convolution of intensity values (denoting protein abundance) corresponding to the same mass-to-charge values (denoting protein mass). Additionally, in typical MS analysis, proteins are unknown a priori and identified by mass-to-charge ratios (as opposed to the use of known gene probes in microarray analysis); hence data-modeling tends to be heavily, if not exclusively, guided by the data and not by biological knowledge.⁴ Figure 1 compares MS with gene expression microarray analysis.

Contrary to microarray data analysis, where a multitude of systems exist for assisting both seasoned and inexperienced analysts, no such system currently exists that will automatically enable a statistically naïve user to create, from start to finish, diagnostic/early-detection models and selection of protein markers from MS data. Therefore, there is a pressing need for systems that will allow both high-quality first-pass analyses of MS data, as well as enhancing work of the data analyst.

In the present report we present 1) an early prototype system (FAST-AIMS: Fully Automated Software Tool for Artificial Intelligence in Mass Spectrometry) which aims at providing this functionality and 2) an initial evaluation of FAST-AIMS that demonstrates the system's ability to construct high-quality models in a fully-automated fashion when used by both experienced and inexperienced users.

Hypotheses:

Our specific hypothesis is that a fully automated system can be produced to create high quality models that generalize well when applied to unseen

MS data. Further, that this system is usable by researchers with varying degrees of expertise in data analysis and machine learning.

Methods

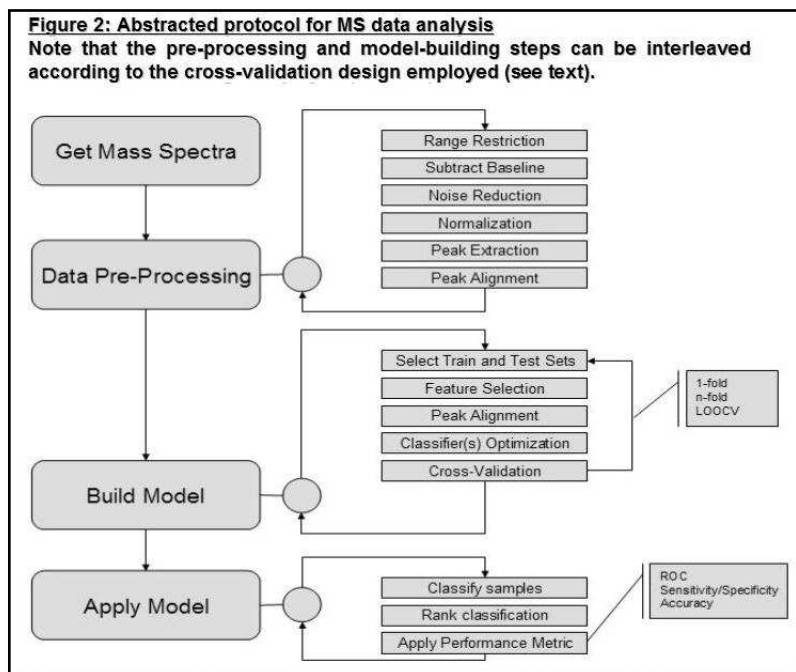
Review of the literature

As a first step toward constructing FAST-AIMS we broadly surveyed the techniques that have been used in the analysis of clinical applications of MS. A detailed side-by-side comparison of data analysis protocols, their medical domain of application, and literature references can be found in the on-line supplement of the present paper¹. A high-level protocol of MS analysis abstracted from this literature (and consistent with our experience in analyzing such data) is presented in Figure 2. We note that although several systems that address focused aspects of the overall analysis (albeit in a disjointed and often inconsistent manner) do exist, no publicly available (commercial or free) software system exists that accomplishes our goals - that is, a complete analysis of the data, starting with raw spectra and ending with a diagnostic or prognostic model and an associated set of biomarkers.

System Functionality & Component Algorithms

The high-level functionality of the system should provide for creation of diagnostic and prognostic models, estimation of their true (i.e., future) error, and ability to apply the models to new patients. Identification of biomarkers was also deemed important both for improving classification performance as well as for discovery purposes.

The generalized protocol of Figure 2 indicates the overall process for analysis. The specific algorithms chosen for each of the protocol steps are



as follows:

(1) Data-preprocessing: We chose the Coombes *et al* peak detection and baseline subtraction algorithm⁵, together with Yasui *et al*'s peak alignment⁶, since at the time of system design, these were among the very few publicly available, peer-reviewed preprocessing algorithms, and we have had good results using them in prior experiments.

(2) Feature selection: We chose two families of multivariate feature selection methods: the first is the HITON Markov Blanket induction algorithm, and the second is the Support Vector Machine (SVM)-based RFE algorithm^{7,8}. In independent experiments we have compared these two algorithms to a variety of peak selection procedures and found that they consistently selected fewer peaks without loss of classification accuracy compared to univariate, principal component-based, parameter-shrinkage, and wrapping methods.

(3) Classifiers: We chose multi-class SVMs because of their robust, high performance in published analyses of MS data and other mass-throughput data, most notably gene expression arrays in which SVMs outperformed all major pattern recognition algorithms⁹. SVMs have several additional attractive features including being able to handle arbitrarily complex functions, relative insensitivity to the curse of dimensionality, principled variable reduction, and an abundance of optimization methods – some empirical, and some theoretically-motivated.

(4) Model selection and future performance estimation: We apply a balanced nested n-fold cross-validation procedure in which the inner cross-validation is used to optimize parameters for the classifiers and conduct data pre-processing and peak selection, while the outer loop estimates the error of the resulting classifier. This design closely follows the powerful GEMS system for automated analysis of array gene expression data, in which it was shown via comparison to published analyses and cross-dataset evaluations that overfitting is avoided^{i,10}

The graphical user interface for FAST-AIMS was developed with Delphi 7.0 with all algorithms programmed in Matlab 6.5. Other than a downloadable executable, no additional software is required to run FAST-AIMS.

System Description

FAST-AIMS provides an intuitive wizard-like

ⁱ We note that this strict procedure is practically unattainable by hand, and as a result none of the published MS reports we have seen in the literature employ such strict measures to avoid overfitting.

interface with defaults provided so that users need not be familiar with all steps of data analysis. Given a MS dataset as input, FAST-AIMS can automatically perform one of the following tasks: a) generate a classification model by optimizing the parameters of classification and peak detection algorithms; b) estimate future classification performance of the optimized model; c) generate a model and estimate classification performance in tandem; d) apply an existing model to a new set of patients. In the process, the system also offers the option of identifying biomarkers that capture the classification tasks of interest and can be used to explore the underlying biological mechanisms.

In general, the analysis of the data is performed following the principle that all steps that can be performed on each spectrum independently of the others (i.e., baseline subtraction, peak detection and normalization) are conducted for all of the data once. All steps that require consideration of multiple spectra (e.g., peak alignment, peak selection, classifier building) are performed *de novo* for each sub-split of the data, so that these steps are not overfitted to the test spectra. (Overfitting leads to non-reproducible models and unrealistic performance estimates.) Below, we outline the main steps in the analysis as performed by FAST-AIMS:

- First, the system is given a series of spectra.
- The data is then split into multiple training and test sets.
- Pre-processing steps (i.e., baseline correction, peak detection and alignment, as well as normalization) are performed.
- One or more user-selected feature selection algorithms are then applied to each training set.
- The system then uses user-selected classifiers and selected range(s) of associated parameters from which to optimize the model.
- The system is now ready to optimize, select, and save a classification model based on the preceding steps. The user can specify the metric to be used for evaluation of the model (ROC or accuracy) after which the model is built and applied to the test set(s) in a task-dependent manner.
- All steps are logged and reported as the user navigates through the system.

Design of System Evaluation

To evaluate FAST-AIMS, a clinical MS dataset was chosen and 6 dataset-users of varying degrees of experience were recruited. The user subjects included one faculty of biomedical informatics and one graduate student with experience in MS analysis, one scientific programmer with experience in MS analysis and FAST-AIMS, one scientific programmer without prior exposure to MS analysis but involved

in FAST-AIMS' development, one medical student with no experience in data analysis, and one graduate student with experience in machine learning but no prior exposure to MS data. The first 4 users are designated as "expert" FAST-AIMS users, while the latter 2 are the "non-expert" FAST-AIMS users. The non-expert users were given a 1.5 hour hands-on session on how to use the system.

The dataset was not used during system development, and the identity of the dataset was masked from users during the evaluation. Selection of the dataset was based on it being of sufficient sample size to carry out a hold-out evaluation design of the system against the expert and lack of use in the public domain (thus reducing risk of chance prior exposure by those participating in the evaluation). The dataset¹¹ consisted of a total of 162 SELDI MS spectra (106 from patients with prostate cancer and 56 controls) taken from blood serum. The data was randomly assigned to a training set (n=108) and a testing set (n=54). The testing set (and associated class information) was withheld from all users during the evaluation period. Users were instructed to produce the best model possible based on the area under the ROC performance metric, and to estimate its performance.

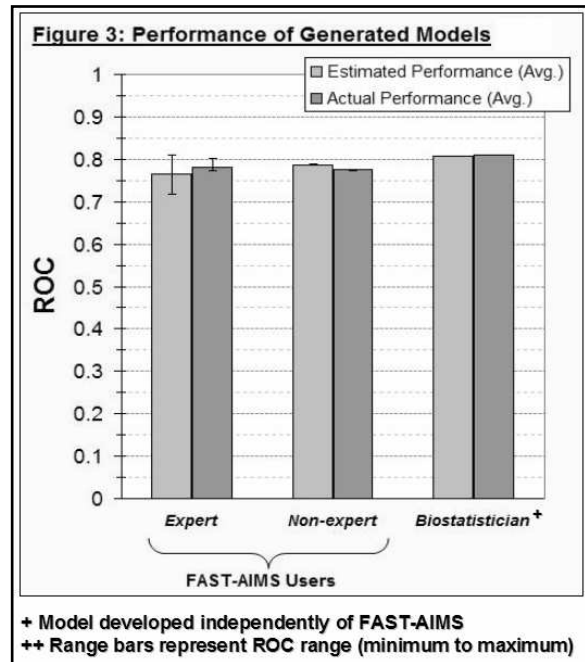
An expert faculty-level biostatistician with substantial prior exposure to analysis of MS data was also recruited to the study and was asked to produce a model independently of FAST-AIMS.

Each user and the biostatistician produced a model. Each model was then applied to the withheld testing data, and classification performance was calculated using the area under the ROC metric.

Through the selection and participation of users of different MS experience, and through comparison of FAST-AIMS users with non-FAST-AIMS users, the study design sought to provide an initial evaluation of the potential of such software to measure up to the performance of a highly qualified biostatistician, as well as to the performance of the human expert group associated with the original published analysis of the data.

Results

1. FAST-AIMS user models. All FAST-AIMS users in the evaluation study were able to use the system to develop a model. The time it took them to setup the system for analysis was less than 30 minutes. FAST-AIMS run for between 8 and 55 hours (median: 9 hours) depending on the complexity of the analysis and the available hardware. Table 2 of the online supplement gives details about the specific system settings employed by each user. In comparison the expert statistician logged 7 hours of data analysis work (in addition to about 3 hours of consultation



with the study authors for clarifying data characteristics).

2. Comparative performances. The average performance of each model grouping is presented in Figure 3. As can be seen, the differences between the performance of the FAST-AIMS users and the expert biostatistician was not large (mean .78 for FAST-AIMS vs. .81 for the expert).

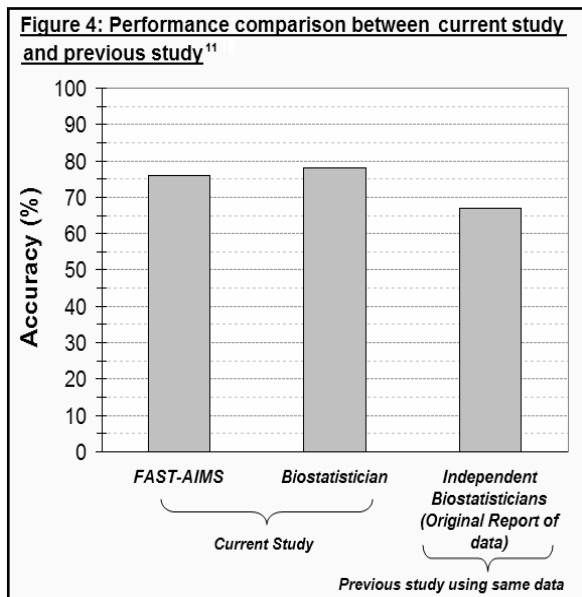
3. Overfitting. As presented in Figure 3, neither the expert nor the FAST-AIMS users overfitted their results, and the estimated performances for both groups were very accurate (within 1% of the true value).

4. Comparison to a previous human analysis with this dataset and to PSA. Both FAST-AIMS users and the present study's expert outperformed (accuracy range: 76.1-80.4%) the initial analysis of the original producers of the data¹¹ (best accuracy: 67%). (Figure 4).

Finally it is significant to notice that the area under the ROC values reported by all FAST-AIMS users (ROC range: 0.773 - 0.802) as well as by the biostatistician (0.811) are all significantly larger than the ROC range of 0.66-0.70 reported in a recent landmark study¹² of PSA screening. This suggests that MS may provide a much more promising screening test than the current state-of-the-art one, with significant public health implications.

Limitations and Future Research

FAST-AIMS, even in this early stage of development, clearly shows the feasibility and potential of fully automating the process of MS data analysis and modeling. We are in the process of expanding the evaluation by adding more datasets



and at least one more biostatistician. While many of the chosen algorithms in the system (e.g., peak selection algorithms) were included after extensive comparative evaluations of other possible alternatives, the same is not true for the classifiers and the data pre-processing methods employed. Our experiment showed that the statistical expert used different methods for pre-processing the data, for peak selection and for classification than the system. In future work, we will examine the effects of the different algorithms employed and their interactions in this and additional datasets. We will also conduct an extensive comparative evaluation of most major families of applicable algorithms in the clinical MS domain following the process established in the development of the GEMS system for microarray data analysis⁹. We believe that this systematic process and evaluation of robustness is also warranted for FAST-AIMS, and such an analysis will be feasible as diverse and improved algorithms for all steps in the generalized analysis protocol forming the backbone of FAST-AIMS are becoming available.

Conclusions

We introduced FAST-AIMS, a system for the fully automated development and evaluation of diagnostic models and biomarker discovery from clinical mass spectrometry data. We described a formative evaluation of FAST-AIMS, in which it was found that FAST-AIMS allowed naïve and expert users alike to quickly and with minimal effort match the performance of an expert biostatistician and exceed previously published results on the same data, while avoid overfitting.

Acknowledgments

Support from grants T15 LM07450-01 (NF) and LMRO1-7948-01 (CFA) is gratefully acknowledged. The authors wish to thank Yin Aphinyanaphongs, Kevin Maas, Yerbolat Dosbayev, and especially Alexander Statnikov for their valuable contributions to the work presented here.

References

- ¹ On-line supplement to: "N. Fananapazir, M. Li, D. Spentzos, C.F. Aliferis; *Formative Evaluation of a Prototype System for Automated Analysis of Mass Spectrometry Data*"; Available from <http://discover.mc.vanderbilt.edu/discover/public/supplements/fastaims>
- ² Aebersold R., Mann M. *Mass spectrometry-based proteomics*. Nature 422, 198-207 (2003).
- ³ Quackenbush J. *Computational Analysis of Microarray Data*. Nature Reviews Genetics 2, 418-427 (2001).
- ⁴ Rappsilber J., et al. *Experiences and perspectives of MALDI MS and MS/MS in proteomic research*. International Journal of Mass Spectrometry 226 (2003) 223-237.
- ⁵ Coombes KR, Fritsche Jr. HA, Clarke C., et al. *Quality Control and Peak Finding for Proteomics Data Collected from Nipple Aspirate Fluid by Surface-Enhanced Laser Desorption and Ionization*. Clinical Chemistry 2003, 49:10, 1615-1623.
- ⁶ Yasui Y, et al. *An Automated Peak Identification/Calibration Procedure for High-Dimensional Protein Measures From Mass Spectrometers*. J Biomed Biotechnol. 2003(4):242-248.
- ⁷ Aliferis CF, Tsamardinos I, Statnikov A. *HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection*. Proc. AMIA 2003 p. 21-5.
- ⁸ Guyon IM, et al. *Gene selection for cancer classification using support vector machines*. Machine Learning 2002 46:389-442.
- ⁹ Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. *A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis*. Bioinformatics. 2005 Mar 1;21(5):631-43.
- ¹⁰ Statnikov A, Tsamardinos I, Dosbayev Y, Aliferis CF. *GEMS: A System for Automated Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data*. Int J Med Inform. 2005.
- ¹¹ Banez LL, et al. *Diagnostic potential of serum proteomic patterns in prostate cancer*. J Urol. 2003 Aug;170(2 Pt 1):442-6.
- ¹² Thompson IM, et al. *Operating Characteristics of Prostate-Specific Antigen in Men With an Initial PSA Level of 3.0 ng/mL or Lower*. JAMA. 2005; 294:66-70.