DEVELOPMENT AND EVALUATION OF A PROTOTYPE SYSTEM FOR AUTOMATED ANALYSIS OF CLINICAL MASS SPECTROMETRY DATA

By

Nafeh Fananapazir

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

August, 2007

Nashville, Tennessee

Approved:

Professor Constantin Aliferis
Professor Dean Billheimer
Professor Doug Hardin
Professor Shawn Levy
Professor Dan Liebler
Professor Ioannis Tsamardinos

TABLE OF CONTENTS

Page

Appendix

ACKNOWLEDGEMENTS

This work would not have been possible without the financial support of the National Library of Medicine or the Vanderbilt Medical Scientist Training Program. I am indebted to Dr. Cindy Gadd, Director of Educational Programs for the Department of Biomedical Informatics, who has been supportive in helping me achieve my career goals.

I am grateful to each of the members of my thesis committee for providing me with guidance in scientific research and in the structuring of this thesis. I would especially like to thank Dr. Constantin Aliferis, the chairman of my committee. He has taught me more as a teacher and mentor than I could possibly give adequate recognition for here. His support and encouragement have been invaluable.

Finally, I want to express appreciation and loving admiration for my wife, Katrina, and my son, Jameh, for their love and uplifting presence during these foundational years.

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

2DE	Two-dimensional Gel Electrophoresis
AUC	Area Under the Curve
FAST-AIMS	Fully Automated Software Tool for Artificial Intelligence in Mass Spectrometry
GEMS	Gene Expression Model Selector
KNN	K-Nearest Neighbor
m/z	Mass/Charge ratio
LSVM	Linear Support Vector Machine
MALDI	Matrix-Assisted Laser Desorption/Ionization
MB	Markov Blanket
MS	Mass Spectrometry
PSA	Prostate Specific Antigen
PSVM	Polynomial Support Vector Machine
ROC	Receiver Operating Characteristic
RFE	Recursive Feature Elimination
SVM	Support Vector Machine
TOF	Time of Flight
SELDI	Surface-Enhanced Laser Desorption/Ionization

CHAPTER I

INTRODUCTION

Mass spectrometry in context

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. In the past two decades, MS instrumentation and techniques have come to have a considerable impact on the way proteins and protein mixtures have been analyzed. Traditionally, once a protein of interest had been isolated through techniques such as two-dimensional gel electrophoresis (2DE), further protein analysis had been complicated by the inherent difficulty of ionizing proteins and peptides while minimizing fragmentation. Bolstered by the creation of ionization methods such as matrix-assisted laser desorption/ionization (MALDI), protein analysis benefits from sensitive, precise, direct measurement of large polypeptides effective, when linked with techniques such as time-of-flight mass analysis (TOF), over a large mass range.

MALDI-TOF analysis detects mass-to-charge (m/z) ratios of ionized proteins and MALDI spectra depict signal intensities associated with abundance and efficiency with which the proteins ionize. When an unknown protein is analyzed simultaneously with and calibrated against a standard of known mass, precise mass measurements of the unknown protein can thus be determined. MALDI involves ionization of a sample through exposure to a beam of laser-generated light onto a ultraviolet-absorbing matrix containing a small proportion of the sample to be analyzed. The composition of the matrix (small aromatic organic acids) varies depending on the application required and generally depends on the size of peptide fragments (ex. Sinapinic acid for the analysis of larger peptides and α-cyano-4-hydroxycinnamic acid (ACCA) for smaller peptides or peptides from tryptic digests). The ionized peptides separate according to mass and charge, their time of flight (TOF) increasing with m/z ratio. Surface enhanced laser desorption/ionization (SELDI) is a variation of MALDI. In SELDI-TOF MS, the surface of the sample chip acts as a separation step (e.g. via positive ion exchange, hydrophobicity, metal-binding) with the goal of allowing for increased sensitivity and reducing the need for expensive and time-intensive sample pre-processing and separation steps.

MS analysis and techniques have generated considerable excitement in the clinical domain as potentially powerful tools for the analysis of protein and peptide mixtures. Significant headway toward unambiguous identification of components of interest remains a significant challenge when analyzing very complex protein mixtures. The continued development and utilization of pre-processing separation techniques such as liquid chromatography (LC) [Covey 1986] and the use of tandem MS [Boyd 1994], which identifies peptide sequences both offer promise in helping to resolve such challenges.

MS studies in clinical bioinformatics

Within the last four years, several researchers 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 accidents . Published reports indicate remarkable potential for this technology to diagnose disease with minimally invasive testing procedures, low cost, and – in some cases – 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 [Anderson 2002].

Though the ultimate goal of many such studies may be to identify specific biomarkers for disease, most approach the problem as one of pattern recognition within mass spectra independent of identification of the peptides contributing to the discriminating patterns. In particular, blood serum appears to have attracted the attention of most clinical proteomics investigations, due to a combination of low cost and ease of sample collection and resulting promise as a screening tool. The success of such investigation depends primarily on the validity of the intuitive notion that – given blood necessarily traveling in close proximity to virtually every cell in the human body (tumor or otherwise) – a serum sample represents an informative passport stamped with useful molecular clues as to the state of the entire body. For example, in support of this notion, it is noted that prostate serum antigen (PSA) levels are regularly monitored (through non-MS, anti-PSA antibody techniques) in aid of prostate cancer detection.

Protocols for Clinical MS and Historical Considerations

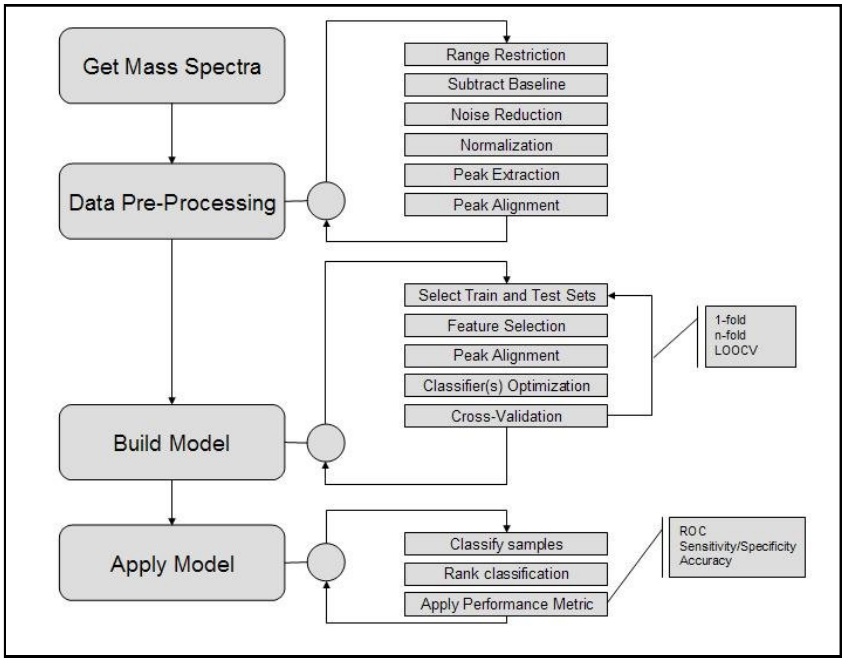


Figure 1: Abstracted protocol for MS data analysis

There is an exponentially increasing body of literature producing attempts at convincing methodologies for the analysis of MS data in the clinical domain. Major studies (2002-2004) are highlighted in Table 1; though no study incorporates all of the illustrated components, each falls within the data analysis paradigm summarized in Figure 1. Some milestones in the development of appropriate methodologies follows.

In February 2002, Petricoin et al. published the landmark paper “Use of proteomic patterns in serum to identify ovarian cancer.” [Petricoin 2002] In this study, mass spectra were generated and partitioned into a training and testing set. A genetic algorithm and self-organizing cluster analysis was applied to determine discriminatory signature patterns (m/z ratios that yielded good classification results in the training set). Corresponding signature patterns were then taken from samples in the test set and compared with the discriminatory pattern and a classification posited. The success of the genetic algorithm to classify was then evaluated in terms of sensitivity, specificity, and positive predictive value. Note that, in reference to Figure 1, pre-processing steps such as normalization, base-line correction, peak detection, feature selection, and peak alignment were not performed.

Table 1: Current methods used in MS data analysis

			Number of Samples				Study Design			
Reference	Pub. Date	Domain	Healthy	Diseased or Benign (non-Cancer)	Cancer	Total	Overall Study design[+]	Reported Pre-processing Steps[++]	Classifier[+++]	Metric[++++]
Petricoin	02/2002	ovarian cancer	100	16	100	216	1-fold		GA	sens/spec
Li	05/2002	breast cancer	41	25	103	169	100-fold	NR P	multivariate	sens/spec.
Adam	07/2002	prostate cancer	82	77	167	326	1-fold	NRBPA	DT	sens/spec
Qu	07/2002	prostate cancer	96	92	197	386	10-fold	NR P	ROC-analysis DT (boosted)	sens/spec.
Vlahou	02/2003	ovarian cancer	95	0	44	139	10-fold	NRPA	DT	acc
Yanagisawa	08/2003	lung cancer	14	0	79	93	LOOCV	RBPA	WFCCM	acc
Hilario	09/2003	lung cancer	17	0	24	41	10-fold	N PBA	DT KNN MLP Naïve Bayes	acc
Kozak	10/2003	ovarian cancer	56	19	109	184	1-fold	NR P	multivariate	sens/spec ROC acc
Won	12/2003	renal cancer	6	15	15	36	0-fold	RBPA	DT	sens/spec acc
Koopmann	02/2004	pancreatic cancer	60	60	60	180	30-fold	NRBP	multivariate	sens/spec ROC
Wadsworth	03/2004	head and neck cancer	102	0	99	201	1-fold	NR P	DT	sens/spec
Zhu	08/2004	liver cancer	25	25	20	70	1-fold	PA	DT	sens/spec PPV
Wong	08/2004	cervical cancer	27	0	35	62	1-fold	NRBPA	ROC-analysis	sens/spec positive/negative PV
Prados	08/2004	cerebrovascular accidents	0	21	21	42	10-fold	NRBPA	SVM KNN MLP	sens/spec
Vlahou	08/2004	bladder cancer	33	92	105	230	1-fold	N PA	DT	sens/spec
Yu	11/2004	colorectal cancer	92	35	55	182	10-fold	NR P	SVM NN	sens/spec

[+] Overall study design key: n-fold: n-fold cross-validation, LOOCV: leave-one-out cross validation

[++] Pre-processing key: N: normalization, R: range restriction, B: baseline subtraction, P: Peak detection and/or binning, A: Peak alignment

[+++] Classifier key: SVM: support vector machine, NN: neural network, GA: genetic algorithm, DT: decision tree, MLP: multi-layer perception, KNN: K-nearest neighbor, WFCCM: weighted flexible compound covariate method

[++++] sens: sensitivity; spec: specificity; acc: accuracy

Peak detection, alignment, and selection as applied to cancer classification were first reported in the literature by Adam et al. [Adam 2002]. A decision tree classifier was used. The peak detection procedure used, a potentially important pre-processing step, was dependent on the use of a proprietary algorithm embedded in Ciphergen SELDI software version 3.0. In July 2003, Coombes et al. [Coombes 2003] published the first publicly accessible peak detection algorithm (available for download in Matlab format). It appears that samples were obtained in replicate from pooled sources for healthy patients and cancer patients respectively. The number of patients pooled in each category does not appear to be disclosed, hindering evaluation of their methodology. Nevertheless, publication and relatively full disclosure of the algorithms involved serves as a healthy model for other researchers in the field [Coombes 2003]. Coombes peak detection algorithms have undergone further development, incorporating wavelet transformation, and exist as part of the publicly available Cromwell package [Coombes 2005].

In December 2003, Zhu et al. published an attempt to enhance the signal-to-noise ratio through smoothing of the spectra through use of a Gaussian filter, opening up interesting possibilities in terms of the application of signal analysis techniques to the MS domain [Zhu 2003].

Problems existing in current MS analysis

While recognizing the embryonic nature of such studies and the difficulty in navigating what can be seen as significant uncharted territory, it is important to point out that the studies listed in Table 1 tend to suffer from one or more (often avoidable) problems which may affect confidence in their specific results. These include: 1) Lack of disclosure of key methods components, preventing reproducibility [Adam 2002]; 2) Overfitting is a phenomenon in which a classification model has high predictive power on training data, but relatively low predictive power when applied to unseen future (testing) data. The ultimate case of overfitting occurred in a few cases in which there was no testing set at all, effectively evaluating the performance of the data analysis methodology on the basis of its ability to classify the data on which the methodology was applied [Rosty 2002] [Won 2003]. Perfect classification results may be obtained in this way, but the ability to generalize to unseen data is compromised. 3) One-time partitioning of the data, problematic in the sense that enthusiasm as to the strength of a particular result is tempered by not knowing whether the performance is affected by chance allocation of samples [Petricoin 2002], [Adam 2002]; 4) Lack of randomization when assigning to train and test categories, a key step in helping to ensure that the algorithms are learning to distinguish between the classes being studied rather than between mitigating biases associated with assignment to the training and testing categories. For example, a test set collected at a different time or location, or collected/analyzed with a different instrument, may create differences which are statistically significant, but of little value to analyzing the problem; 5) Some studies use accuracy, a performance metric that is sensitive to prior probability of a disease [Yanagisawa 2003]. If the proportion of disease samples in a dataset does not reflect the prevalence of disease within the population of study, a classifier can present itself in an artificially strong light when prevalence is not considered. In addition, the dependence of such metrics on prior probabilities complicates comparison of the strength of a certain data analysis methodology from dataset to dataset.

Sample-dependent challenges in biomarker detection

In the case of blood proteomics and the search for *a priori* unknown cancer biomarkers (or, more generally, the search for spectral motifs indicative of cancer), the proverbial search for the needle in the haystack is complicated considerably by the fact that the nature of the haystack (normal variations in blood composition) has yet to be fully distinguished from the nature of the needle (tumor biomarkers). In addition, even if discriminatory spectral motifs are discovered and appear to be robust in their ability to distinguish between normal and diseased samples, it remains to be determined whether such motifs are specific for the disease of interest or whether they represent a generalized response of the body to the presence of any of a range of pathological processes. For example, it has been noted that for several cancers (prostate, lung, breast) a strongly associated serum elevation in acute phase protein levels may be determined through MALDI analysis and, indeed, is sensitive and specific when identifying serum samples from cancer patients as compared with serum samples from healthy volunteers. Much of the value of this discriminatory power is lost in recognition of the fact that acute phase protein levels are elevated in response to a diverse range of diseases – including the generalized inflammatory response [Vejda 2002]. Just as with the earliest stages of cancer, some in this range may be subclinical (i.e. cannot be trivially filtered prior to screening); some such confounding conditions may have a prevalence that would result in an unwarranted and likely dangerous increase in the investigative and/or treatment measures that would ensue if such a “discriminatory” screening test were to be used.

Tissue and disease domains studied

At the time of research design and planning, publicly available MS datasets (for protein mixtures) largely represented MALDI or SELDI-MS analysis of blood serum. Blood is composed of two fractions: blood cells and blood plasma. Blood plasma consists mainly of water, mineral salts, and protein. The protein content in blood plasma falls into four categories, of which the first three listed make up the preponderance of the total protein content in blood plasma: 1) albumins; 2) globulins (includes immunoglobulins and lipoproteins); 3) fibrinogen (protein that polymerizes to form fibrin during blood clotting); 4) Low molecular weight (LMW) proteins. Blood serum refers to blood plasma from which the coagulation factors (e.g. fibrinogen) have been removed. It is the hope of proteomics research as applied to blood serum analysis and cancer diagnosis/classification, that information gleaned from a single locus (blood serum) can yield information about a diversity of loci (ovary, prostate, breast, lung, etc.)

Known proteins of higher molecular weight are almost invariably present from sample to sample and include those such as albumin which may be perceived as more likely to cloud effective analysis than to contribute. Therefore, it may seem advantageous to remove them (through centrifugation methods, for example) prior to MS analysis. However, in terms of sample preparation, simply removing protein of higher molecular weight may result in significant information loss; for example, albumin is known to bind and transport proteins of low molecular weight (cytokines, lipoproteins, and the proteolytic fragments of proteins with different histological origins). Therefore, such proteins with higher molecular weight can either be incorporated into the analysis (in which case their binding properties may interfere with detection of LMW proteins) or be subjected to solvent conditions that interfere with protein-protein interactions prior to removal. None of the datasets incorporated in this study seemed to take this into account before SELDI-TOF analysis.

Challenges in MS classification and biological discovery

Comparison of microarray with MS analysis

The analysis of high 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 sizes. As shown in figure 2, 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, and the convolution of intensity values for different peptides corresponding to the same mass-to-charge values. 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.

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 strong need for systems that will allow both high-quality first-pass analyses of MS data, as well as enhancing work of the data analyst.

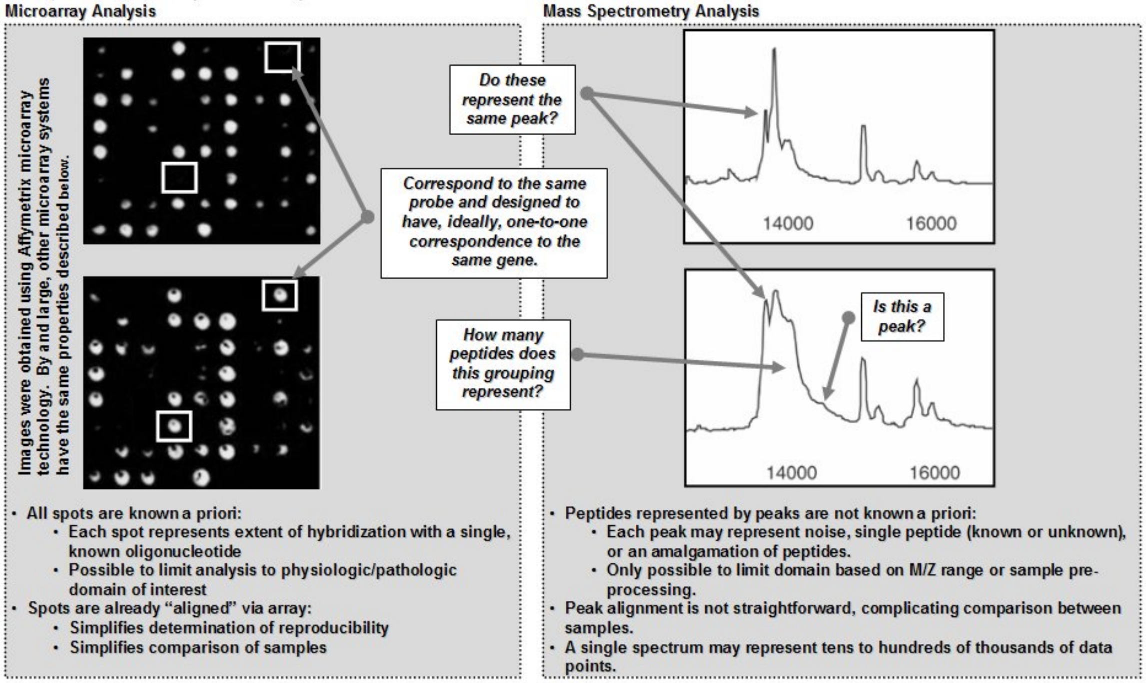
MS data analysis challenges

The typical MS spectrum, which upon visualization appears continuous, is in actuality represented by tens of thousands of discrete points each representing a unique m/z value and an intensity (representing relative mass abundance) at that m/z value. If considering each m/z point as a potential variable, the extraordinarily large number of variables (as compared with sample size) presents challenges the solutions to which are not trivial. Currently, there is no consensus as to how MS data should be treated in cancer diagnosis, and the development of appropriate and effective data analysis methods is an area of active research. Data analysis has extended to the domain of machine learning, the technology and study of algorithms through which machines can "learn" or automatically improve through experience. It remains the case that the theoretical foundation of the algorithms and statistical methods involved, their implementation, and their interdependency put the field of machine learning beyond the expertise of most researchers in the field of clinical cancer research.

Currently, much of this research is being done by expert biostatisticians, and a thorough *de novo* analysis of a single dataset may necessarily involve a considerable period of time on the order of days, weeks, or months. There is a need for clinician-scientists and other biomedical researchers without expertise in the field of machine learning to have access to intelligent software that permits at least a first pass analysis as to the diagnostic capabilities of data obtained from MS analysis.

Figure 2: Challenges in MS Classification/Biological Discovery

A Comparison of Microarray with MS Analysis



Microarray Analysis

Images were obtained using Affymetrix microarray technology. By and large, other microarray systems have the same properties described below.

Correspond to the same probe and designed to have, ideally, one-to-one correspondence to the same gene.

- All spots are known a priori:
  - Each spot represents extent of hybridization with a single, known oligonucleotide
  - Possible to limit analysis to physiologic/pathologic domain of interest
- Spots are already "aligned" via array:
  - Simplifies determination of reproducibility
  - Simplifies comparison of samples

Mass Spectrometry Analysis

Do these represent the same peak?

How many peptides does this grouping represent?

Is this a peak?

- Peptides represented by peaks are not known a priori:
  - Each peak may represent noise, single peptide (known or unknown), or an amalgamation of peptides.
  - Only possible to limit domain based on M/Z range or sample pre-processing.
- Peak alignment is not straightforward, complicating comparison between samples.
- A single spectrum may represent tens to hundreds of thousands of data points.

CHAPTER II

DEVELOPMENT/TECHNICAL DESCRIPTION OF FAST-AIMS

Existing automated systems for disease classification

The goal of this proposal is the creation of a decision support tool, FAST-AIMS: Fully Automated Software Tool for Artificial Intelligence in Mass Spectrometry, for use in the domain of cancer diagnosis. Although several systems in existence address focused aspects of the overall analysis, no publicly available (commercial or free) software system exists that accomplishes our goals - that is, a complete analysis of MS clinical data, beginning with raw spectra and ending with a diagnostic or prognostic model and an associated set of biomarkers. Machine learning techniques have been automated as applied to the domain of microarray analysis, and exist in varying stages of development. These include: 1) Gene Expression Model Selector (GEMS) a multicategory support vector machine tool developed by Alexander Statnikov, a graduate student in the Discovery Systems Laboratory at Vanderbilt University [Statnikov 2002]; 2) GeneCluster 2.0, a standalone application developed at MIT [Reich 2004]; 3) the web-based Gene Expression Data Analysis Tool (caGEDA) developed by the University of Pittsburg Medical Center (UPMC) [Patel 2004].

Development of a protocol schema for FAST-AIMS MS analysis

Peak detection and baseline subtraction

In effect, peak detection is a form of feature selection, in which the classification algorithms focus on the predictive value of few variables (also called features) relative to the entire range. For example, peak detection may indicate the existence of several hundred m/z points of interest, as compared with the original raw data including tens of thousands of possible m/z values.

Prior to development of FAST-AIMS, neither baseline subtraction nor peak detection were incorporated in our preliminary studies. However, while creating FAST-AIMS, the Coombes et al peak detection and baseline subtraction algorithms [Coombes 2003], together with Yasui et al's peak alignment algorithm were incorporated [Yasui 2003]. 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. The original Coombes algorithm for peak detection utilizes the following procedure:

1. Use first differences between successive time points to locate all local maxima and minima
2. Use the median absolute value of the first differences to define "noise"
3. Eliminate all local maxima whose distance to the nearest local minimum is less than the noise
4. Combine local maxima that are separated by fewer then T (default T=3) intervals or by less than M relative mass units (default M=0.05% of

smaller mass). Retain highest local maximum when combining nearby peaks.

5. Compute the slopes from the left hand local minimum up to the local maximum and from the local maximum down to the right-hand local minimum. Eliminate peaks where both slopes are less than half the value of the noise.

Feature selection

For the preliminary studies, Recursive Feature Elimination (RFE) was used in an attempt to reduce the number of variables (m/z values) for study. RFE ranks the features based on the weight learned by Support Vector Machine analysis [Guyon 2002]. The RFE algorithm iteratively removes the lowest half of these (features found least amenable to SVM separation on the basis of class) and proceeds to re-rank the variables based on SVM analysis of the remaining variables. RFE has been incorporated into FAST-AIMS.

Additionally, the HITON Markov Blanket induction algorithm [Aliferis 2003a] has been incorporated into FAST-AIMS. HITON determines the Markov Blanket (MB) of a given target variable (in this case, the presence or absence of illness); this involves identifying the smallest set of variables the values for which maximize the ability to predict that target. Effectively, this reduces noise (improving accuracy of results) as well as computation time through several orders of reduction in the size of a given set of features.

In independent experiments on several data domains (drug discovery, clinical diagnosis, text categorization, lung cancer microarray, MS prostate cancer) and feature set sizes, HITON and RFE algorithms were compared with a variety of peak selection procedures and were found to consistently select fewer peaks without loss of classification accuracy as compared with univariate, principal component-based, parameter-shrinkage, and wrapping methods [Aliferis 2003a].

Normalization

In preliminary experiments, normalization involved mapping of intensities to the range {min(y) → 0, max(y) → 1} by taking each intensity and mapping it to intensity-percentile/100. FAST-AIMS incorporates several other normalization techniques, the selection and sequence of which can be specified by the user [Fananapazir 2005].

Classification

During preliminary experiments, three classification algorithms were tested: K-Nearest Neighbor (KNN) [Fix 1951], Linear Support Vector Machine (LSVM), and Polynomial Support Vector Machine (PSVM) [Vapnik 1998]. While creating FAST-AIMS, multi-class SVMs were chosen 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 algorithms [Statnikov 2005a]. 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 [Guyon 2002].

Experimental design

A nested cross-validation design [Dudoit 2003] was used 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 [Statnikov 2005b].

Ten-fold cross validation was used in the preliminary studies. This experimental design requires that the data be split into ten mutually exclusive subsets. This was done in a stratified manner - class proportions approximated that of the original dataset. Each subset is taken in sequence to be the “test set” upon which the classification model derived from the remaining nine subsets (the “training set”) is evaluated. The parameters for the algorithms involved can be optimized within a given training set through iterative partitioning of the training set itself (nested cross-validation) and averaging the performance of parameter permutations. In this way, parameters for a given classifier can be optimized on the basis of their performance on the train-test set after having been trained on the train-train set. The selected optimized parameters, and the classification model they produce, can then be tested on the test set [Kohavi 1995].

For further discussion and comparison of stratified nested ten-fold cross-validation with other possible techniques, please refer to appendix B.

Performance metric

In preliminary experiments, the performance metric used to generate a performance score in each case was the Area Under the Curve based on the Receiver Operating Characteristic (ROC). The ROC curve is obtained by plotting sensitivity against 1 – specificity as the discrimination threshold is varied for a given classification model. The Area Under the Curve (AUC) can then be calculated.

One advantage to using ROC as a performance metric is that, given its dependence on sensitivity and specificity values, the ROC is insensitive to prior probabilities. In other words, the ROC metric is an evaluation of the classification model's ability to discriminate between classes, and (unlike metrics such as accuracy) represents a value that is independent of the relative distribution of the classes being studied within a given dataset.

In FAST-AIMS, the user can select either ROC or accuracy as a performance metric. Given that ROC analysis is limited to binary classification tasks, accuracy is used when there are three or more classes.

Technical description of FAST-AIMS

The graphical user interface for FAST-AIMS was developed using 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.

FAST-AIMS provides an intuitive wizard-like interface with defaults provided at every stage. In such manner, users need not be familiar with all steps of data analysis. The input for FAST-AIMS consists of an MS dataset. FAST-AIMS can automatically perform any 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. 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. Baseline subtraction, peak detection, and peak alignment are performed on the spectra within each training set. All steps that can be performed on each spectrum independently of the others (i.e., peak identification and normalization) are conducted for all of the data once. All steps that require consideration of multiple spectra (e.g., peak alignment and peak selection) are performed de novo for each sub-split of the data, so that these steps are not overfitted to the test spectra.
- A user-specified normalization sequence is applied within each training set.
- 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.

CHAPTER III

EVALUATION OF FAST-AIMS

Preliminary studies

To demonstrate proof of principle, most of the algorithms to be incorporated into the proposed FAST-AIMS system, including the data pre-processing steps outlined in Figure 1, were tested on the Matlab platform in a series of automated classification tasks. Three classifiers (KNN, LSVM, PSVM) and three methods of feature selection (all features, LSVM-RFE, PSVM-RFE) were applied to three publicly available datasets [Adam 2002] [Petricoin, Ardekani 2002] [Petricoin, Ornstein 2002].

The results of these preliminary classification studies are tabulated (Table 2). As might be expected, KNN performs relatively poorly given the likelihood of many low-relevance features. Nevertheless, the KNN classifier has been widely used as a benchmark in pattern recognition classification experiments because of its conceptual simplicity and its asymptotic behavior (its error is bounded by twice the Bayes error when the training set size approaches infinity [Cover 1967]). SVMs use a hyperplane to partition training data based on their class assignment [Vapnik 1998]. Features which are more amenable to partitioning are assigned higher weight – data which is non-separable can still be partitioned by minimizing the effects of a misclassification cost parameter. This quality allows SVMs to perform favorably relative to other classifiers on datasets with large dimensionality; the “curse of dimensionality” being somewhat circumvented.

Table 2: Results from Preliminary Classification Studies

A. Average AUC values†	Classifier	Feature Selection Method		
		All Features	LSVM - RFE	PSVM - RFE
Adam_Prostate_070102*	KNN	0.84425	0.96863	0.93739
	LSVM	0.99460	0.99595	0.99796
	PSVM	0.99659	0.99316	0.99747
Petricoin_Ovarian_021402**	KNN	0.88150	0.91382	0.79238
	LSVM	0.95455	0.91898	0.85350
	PSVM	0.94409	0.91564	0.84032
Petricoin_Prostate_070302***	KNN	0.85498	0.92219	0.83498
	LSVM	0.92981	0.93026	0.85102
	PSVM	0.93121	0.92788	0.83679
B. AUC Range†	Classifier	Feature Selection Method		
		All Features	LSVM - RFE	PSVM - RFE
Adam_Prostate_070102*	KNN	0.58847 - 0.97263	0.93157 - 1.00000	0.87928 - 0.99413
	LSVM	0.98534 - 1.00000	0.99022 - 1.00000	0.99413 - 1.00000
	PSVM	0.99120 - 0.99965	0.98240 - 1.00000	0.98925 - 1.00000
Petricoin_Ovarian_021402**	KNN	0.50588 - 0.99545	0.62273 - 1.00000	0.40455 - 0.96471
	LSVM	0.80000 - 1.00000	0.49091 - 1.00000	0.52727 - 0.99412
	PSVM	0.75455 - 1.00000	0.50909 - 1.00000	0.43636 - 0.99412
Petricoin_Prostate_070302***	KNN	0.59643 - 0.99667	0.66190 - 1.00000	0.28333 - 1.00000
	LSVM	0.57143 - 1.00000	0.57262 - 1.00000	0.36667 - 1.00000
	PSVM	0.59881 - 1.00000	0.59881 - 1.00000	0.31667 - 1.00000
C. Number of Features Selected		Number of Features Selected by Feature Selection Method		
		All Features	LSVM - RFE	PSVM - RFE
Adam_Prostate_070102*	Range (Avg.)	779 - 779 (779)	24 - 97 (65.2)	97 - 389 (194.1)
Petricoin_Ovarian_021402**	Range (Avg.)	15154-15154 (15154)	14 - 118 (49.9)	14 - 1894 (1421.9)
Petricoin_Prostate_070302***	Range (Avg.)	15154-15154 (15154)	14 - 236 (70.5)	3 - 59 (16.8)

* [Adam 2003], ** [Petricoin, Ardekani 2002], *** [Petricoin, Ornstein 2003]

The theory underlying SVMs has been well characterized, and implementations exist which allow for reasonable execution time.

In addition to demonstrating the feasibility of creating an automated software system, preliminary classification studies were helpful in finalizing the selection of algorithms to be incorporated into FAST-AIMS. SVMs, particularly polynomial support vector machines (PSVM), appear to perform particularly well as classifiers of MS data, influencing the decision to lend significant focus to SVM incorporation and parameter optimization in the development of FAST-AIMS. In terms of feature selection, RFE-LSVM greatly reduces the size of the feature set without sacrificing performance. In fact, in most cases, classification performance was improved through using RFE as a feature selection method.

Evaluation one: study of FAST-AIMS with multiple users

An evaluation study was designed such that the classification performance of FAST-AIMS would be compared with that of an expert biostatistician familiar with MS analysis. In designing the study, it was envisioned that FAST-AIMS users would represent a range of expertise; and study participants were recruited on this basis.

A dataset was selected [Banez 2003] and spectra were randomly assigned to a training set (n=108) and testing set (n=54). Class distribution (cancer, non-cancer) was maintained in each. The testing set (and associated class information) was strictly withheld from all users during the evaluation period. Each study participant was given access to the training data and given approximately one month to submit a classification model. Each was asked to estimate the performance of the submitted model (FAST-AIMS users were to rely on FAST-AIMS' built-in model-performance estimation functionality for this purpose). FAST-AIMS users were given a copy of a user manual [Fananapazir 2005] that was to be their only resource for instruction in the use of FAST-AIMS beyond a preliminary meeting (January 2005) describing the software (in general terms) and the evaluation study. Study participants were asked to work independently of one another.

Assuming successful generation of classification models with the tools at hand, each submitted model would then be applied to the withheld testing set and classification performance reported and compared using the ROC metric.

Description of Banez dataset

A strict requirement of the multiple-user study was that the dataset chosen for the purpose of evaluation was not to have been used in prior development and testing of FAST-AIMS. In addition, selection of the Banez prostate cancer dataset was based on it being of relatively large sample size and on relative lack of use in the public domain (reducing risk of chance prior exposure by those participating in the evaluation). The full dataset consists of a total of 162 SELDI-MS spectra (one spectrum per patient, 106 from patients with prostate cancer and 56 controls) taken from blood serum.

Description of evaluation study participants

Through the selection and participation of users of varying degrees of experience, and through comparison of FAST-AIMS users with non-FAST-AIMS users, the study design sought to provide a preliminary 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.

There were seven study participants: one biomedical informatics faculty member, one graduate student with experience in MS analysis, two biomedical informatics scientific programmers, one medical student with no experience in MS data analysis or machine learning techniques, one graduate student with experience in machine learning but no prior exposure to MS data, and an expert biostatistician.

Users were grouped as follows: 1) Users of FAST-AIMS: both those with no prior familiarity with FAST-AIMS or MS data (n=2) and those familiar with FAST-AIMS and MS data (n=4); 2) An expert biostatistician (n=1) with substantial prior exposure to analysis of MS data who was asked to produce a model independently of FAST-AIMS.

Table 3: Characteristics of classification models generated during multi-user evaluation

		Prior familiarity with FAST-AIMS and/or MS	**Computer time to generate model[++]**	**User time**	**Strategies Employed[+++]**	**Estimated performance (ROC)**	**Actual Performance (ROC)**
FAST-AIMS users	User 1	Y	8 hours	< 30 minutes	LOOCV BC, PD, PA AF, RFE, HITON SVM-gauss	0.810	0.802
	User 2	Y	9 hours		10-fold BC,PD, PA AF, HITON SVM-poly	0.773	0.779
	User 3	Y	19 hours		10-fold BC,PD, PA AF, HITON SVM-poly	0.760	0.773
	User 4	Y	3 hours		10-fold HITON SVM-poly	0.717	0.773
	User 5[†]	N	55 hours		10-fold BC,PD, PA AF, RFE, HITON SVM-gauss	0.786	0.777
	User 6[†]	N	22 hours		10-fold AF SVM-poly	0.789	0.773
Expert Biostatistician[+]				7 hours	UDWT, BC, WFCCM	0.808	0.811

[+] *Model developed independently of FAST-AIMS*

[++] *For FAST-AIMS users, time to generate model is computation time (not user time). User time was < 30 minutes in all cases.*

[+++] ***Strategies employed key:*** *n-fold: n-fold cross validation, LOOCV: leave-one-out cross-validation, BC: baseline correction (Coombes), PD: peak detection (Coombes), peak alignment (Coombes), AF: all features, RFE: recursive feature elimination, SVM-poly: support vector machine (polynomial kernel), SVM-gauss: support vector machine (gaussian kernel), UDWT: undecimated discrete wavelet transformation, WFCCM: weighted flexible compound covariate method*

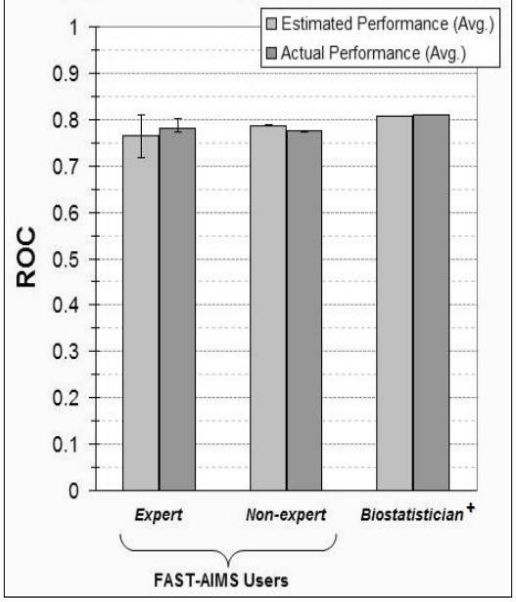
[†] *Users 5 and 6 had no exposure to FAST-AIMS or MS analysis prior to the evaluation study.*

Results for evaluation one

Results are detailed in Table 3 and summarized in Figure 3.

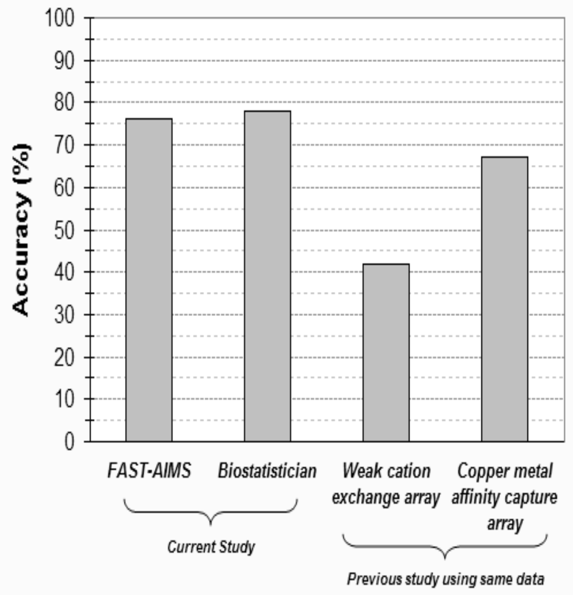
All FAST-AIMS users in the evaluation study were able to develop a classification model. In terms of classification performance, the expert biostatistician did have an edge over the users of FAST-AIMS. However, the value of FAST-AIMS is evident. First, the time that it takes for a user to enter the parameters for automatic model generation is significantly less than the time that it takes for an expert biostatistician to develop a working model. Second, the difference between the performance of the FAST-AIMS users and the expert biostatistician is not large, indicating that a non-expert, if using FAST-AIMS, may be able to approximate the performance of an expert.

Figure 3: Performance of Generated Models



+ Model developed independently of FAST-AIMS
++ Range bars represent ROC range (minimum to maximum)

Figure 4: Performance comparison between current study and previous study



We also note that FAST-AIMS users perform better (accuracy range: 76.1-80.4%) than the accuracy reported by either of two methods used to classify the same data in a previous paper (42% and 67%) [Banez 2003]. As such, even the non-expert user of FAST-AIMS may find his classification performance on or above par with work published by expert biostatisticians (see Figure 4).

Significantly, 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 greater than the ROC range of 0.663-0.699 reported in a recent landmark study determining the limitations of PSA screening. MS data seems to hold great potential as compared with PSA, no screening threshold for which yields high sensitivity and high specificity for prostate cancer screening in healthy men [Thompson 2005].

Results were published and presented for the 2005 AMIA Symposium [Fananapazir 2005].

Evaluation two: study of FAST-AIMS with multiple datasets

The remaining major deliverable in the thesis proposal consisted of carrying out a separate evaluation of FAST-AIMS performance on multiple datasets from the MS clinical domain. For comparison's sake, this evaluation was performed using the same three datasets incorporated into the preliminary studies.

In this evaluation, 10-fold cross-validation was used to generate the model with the best average performance (ROC) when averaged over all splits. The additional task of estimating performance when applied to new data was also specified when running the FAST-AIMS analysis. As compared with the preliminary studies, the FAST-AIMS evaluation included the following differences:

1. For the Petricoin ovarian dataset [Petricoin, Ardekani 2002] and the Petricoin prostate dataset [Petricoin, Ornstein 2002], Coombes baseline subtraction/peak-detection, and Yasui peak alignment were performed. The

Adam prostate dataset [Adam 2002] is a pre-processed dataset, baseline subtraction and peak detection/alignment already having been performed.

2. KNN was not used for classification.
3. HITON was included as a feature selection method.
4. Classifiers and feature selection methods were not considered separately when optimizing model parameters. Rather, the selection of the best classifier and feature selection combination was recognized as part of the optimization process and reported as components of a unified model.

Table 4: Results from multiple dataset evaluation of FAST-AIMS

	Estimated ROC Performance: Range (Average)	*Number of features: Range (Average)*	*Model Selected* †
Adam_Prostate_070102*	0.96825 – 1.0 (0.98110)	30 – 779 (212)	LSVM (Cost: 1000) LSVM-RFE (Cost: 100, # features: 90) Average ROC: 0.99883
Petricoin_Ovarian_021402**	0.88232 – 1.0 (0.95946)	20 – 15154 (3162)	LSVM (Cost: 1000) All Features Average ROC: 0.96128
Petricoin_Prostate_070302***	0.78112 – 0.99454 (0.92052)	100 – 15154 (8192)	PSVM (Cost: 1000, Deg: 2) All Features Average ROC: 0.98815

* [Adam 2003], ** [Petricoin, Ardekani 2002], *** [Petricoin, Ornstein 2003]

† SVM models have a cost parameter which permits some misclassifications. Increasing the value of C increases the cost of misclassifying points and forces the creation of a more accurate model that may not generalize well.

Results for evaluation two

Results for the multiple dataset evaluation of FAST-AIMS are summarized in Table 4 and represent two separate tasks.

In the first task, a classification model is generated for each dataset, representing the combination of methods and parameters that yielded the highest average ROC when averaged over each of the ten data splits.

A better estimate of performance when applied to new data is obtained through the nested cross-validation technique employed in the second task. Each training data split is used to generate a model in a manner similar to the task described in the paragraph above (using n-1 cross-validation that is blind to information contained in the withheld testing data split). Once the model for a given training split has been selected, it is applied to the associated testing split; the performance (ROC) is recorded. The average of these ten ROC values (for 10-fold cross-validation) is recognized as a better estimate of predictive power [table 4]. As such, the estimated ROC performance is the performance predicted in classifying new data if free to consider all permutations of the classification and feature selection methods (and associated parameters) selected by the user for consideration in model generation, and only indirectly represents the estimated performance for the specific combination of classifiers, feature selection methods, and parameters incorporated by the model generated by the previous task. This fact becomes clear when one considers that each of the ten splits used in cross validation may yield a different “best” classifier/feature-selection-method combination; none of which are guaranteed to be the same as the model generated by the first task.

CHAPTER IV

CONCLUSIONS AND DISCUSSION

This thesis describes work towards creation and evaluation of FAST-AIMS (Fully Automated Software Tool for Artificial Intelligence in Mass Spectrometry), a system for the automated development and evaluation of diagnostic models from clinical MS data. Two evaluations of this software are described; the first evaluation comparing the performance of models generated by several system users from a MS dataset to that of a model generated by an expert biostatistician; the second evaluation sought to use FAST-AIMS to develop classification models in three datasets and to help draw more general conclusions as to the value of such methods across multiple datasets in the clinical domain.

In the first evaluation, it is noted that when the classification models were applied to the withheld portion of the dataset, ROC analysis shows that FAST-AIMS allowed naïve and expert users alike to nearly match the performance of an expert biostatistician. In addition, results compared favorably with, and indeed proved superior to previous work on the same dataset. Initial attempts to evaluate the role of MS data in prostate cancer screening as compared with PSA seem to indicate that MS data has the potential to produce models that have superior sensitivity and specificity.

In the second evaluation, results were clearly demonstrative of the powerful classification power of machine learning techniques when applied to three MS datasets in the clinical domain. To what extent discriminatory power is attributable to relevant biological processes is an area of intense debate and continued research [Diamandis 2004].

Necessarily, the focus of this thesis has been fundamentally guided by dataset availability within the emergent area of clinical MS analysis of protein mixtures. The development and evaluation of software that employs powerful, robust machine learning techniques to classify cancer and non-cancer specimens has been described. In light of challenges discussed within the introduction, it is acknowledged that such classification, even if robust, can only be as meaningful as the datasets themselves allow. Certainly when applied to the datasets discussed, the machine learning methods employed by FAST-AIMS are able to distinguish between classes of disease in unbiased fashion. Powerful as they may be, such methods in and of themselves are limited by the quality of the information inherent in these datasets. For example, such classification methods are not able to ignore the classification "assistance" of differences attributable to non-biological factors such as might be created by biased sample collection or biased instrumentation conditions. Even when such bias is minimized, it is not possible to ignore class differences attributable to real but, in terms of clinical utility, trivial biological phenomena such as those manifested by general physiologic responses to disease. It is conceivable that a dataset could be produced that includes control samples associated with potentially confounding non-cancer disease processes, forcing FAST-AIMS to attempt identification of discriminatory patterns more specific for the cancer of interest. In general, the inclusion of appropriate controls is either incomplete, limited by lack of samples, perhaps not even considered, or thwarted by the current gap in knowledge as to what a range of appropriate controls would look like.

That being said, it is my strong belief that the true value of this study does not depend on any putative biological assertions or even, ultimately, on the quality of the current datasets. Until development of a modality with perceived greater screening potential, MS will likely continue gaining momentum in the clinical research domain, generating datasets with more samples, better consistency in sample collection, improved sample pre-processing, inclusion of well-thought out sample controls, etc. Yet – no matter how good the dataset – success in overcoming the enormity of the challenge posed by the classification task (and, certainly, success in biomarker identification) is dependent on the continued development of powerful, fast, non-overfitted techniques such as those incorporated by the software presented here.

Creation of user-friendly software will be important for the clinical use of mass spectrometry. Eventually, such software development will help create meaningful collaboration between those with expertise in MS analysis and other researchers whose expertise, though perhaps not greatly overlapping, would consequently find effective synergy and new avenues in the ultimate goal of reducing the significant societal burden imposed by cancer. The work presented in this thesis represents an initial step in recognition of the need for such a development.

APPENDIX A

MASTER'S THESIS DELIVERABLES

The following deliverables were presented to the Master's Thesis Committee and approved on April 2nd, 2004, and is reproduced from the written thesis proposal.

☑	**Task**
1. Allow importation of comma-separated data (text file)	
2. Allow selection of normalization procedure and parameters	
3. Allow selection of feature selection and parameters	
	a. RFE (recursive feature elimination)
4. Allow selection of peak detection, baseline subtraction and parameters	
	a. Coombes et al. peak detection (if time permits)
	b. Coombes et al. baseline subtraction (if time permits)
5. Allow selection of classifiers and parameters	
	a. KNN
	b. LSVM
	c. PSVM
	d. RBF-SVM (as time permits)
	e. DT (as time permits)
	f. NN (as time permits)
	g. Multi-category SVM (as time permits)
6. Employ "smart defaults"	
7. Ability to save and run classification parameters on "new" samples	
8. Output log of experiment details as a text file	
9. Move all executable files to disk and run independently of additional software	
10. Incorporate a "wizard-like" interface	

APPENDIX B

RESEARCH DESIGN OF FAST-AIMS' ANALYSIS

In the preliminary experimental design, as well as programming architecture of FAST-AIMS, three types of cross-validation (as opposed to bootstrapping) were considered: one-fold, n-fold, and leave-one-out cross-validation. Stratified vs. non-stratified and nested vs. non-nested methods were also considered. The decision to use nested, stratified ten-fold cross-validation was based on prior work related to bias-variance decomposition analysis.

According to bias-variance decomposition analysis, the generalization error for a classification model can broken down into three components: noise, bias, and variance [Aliferis 2006]. Noise refers to uncertainty inherent in a given dataset and represents the error of the optimal classification model from the class of all possible classification models. Bias refers to the difference in error between the optimal classification model and the best classification model from the class of all possible classification models given by a specific classifier (if a classifier is able to determine the optimal classification model, bias is zero). Variance refers to the difference in error between the best classification model that a specific classifier is able to produce and that of the actual model produced.

Prior work [Kohavi 1995] has shown that bootstrapping (in which instances are selected for training with replacement) demonstrates lower variance as compared with cross-validation, but demonstrates larger bias on specific datasets and higher computational cost [Efron 1997] [Braga-Neto 2004]. When implementing cross-validation, stratification was determined to reduce both bias and variance [Kohavi 1995]. Ten-fold cross-validation was determined to have very low bias across datasets tested, with higher-fold cross-validation (including LOOCV) offering little improvement. Lower-fold cross-validation (e.g. one-fold, 2-fold, 5-fold), though having lower computational cost, offer biases that are pessimistic. Ten-fold cross-validation was recommended as having the best trade-off between bias and computational cost [Kohavi 1995].

Nested n-fold cross-validation [Dudoit 2003], as compared with n-fold cross-validation, has been shown to be powerful in detecting overfitting and estimating generalization error conservatively for microarray and other high-dimensionality data [Statnikov 2005b] [Aliferis 2003b].

REFERENCES

Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z, Wright Jr. GL. *Serum Protein Fingerprinting Coupled with a Pattern-matching algorithm Distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men.* Cancer Research, 62, 3609-3614, 2002.

Alexe G, Alexe S, Liotta LA, Petricoin E, Reiss M, Hammer PL. *Ovarian cancer detection by logical analysis of proteomic data.* Proteomics 4, 766-783, 2004.

Aliferis CF, Tsamardinos I, Statnikov A. *HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection.* AMIA Annu Symp Proc. 21-5, 2003a.

Aliferis CF, Tsamardinos I, Massion P, Statnikov A, Hardin D, *Why Classification Models Using Array Gene Expression Data Perform So Well: A Preliminary Investigation Of Explanatory Factors*, Proceedings of the 2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS), June 23-26, 2003, Las Vegas, Nevada, USA, CSREA Press, 2003b.

Aliferis CF, Statnikov A, Tsamardinos I, *Challenges in the Analysis of Mass-Throughput Data: A Technical Commentary from the Statistical Machine Learning Perspective*, Cancer Informatics 2 133–162, 2006.

Anderson NG, Anderson NL. *The Human Plasma Proteome*. Molecular & Cellular Proteomics 1:845-867, 2002.

Banez LL, et al. *Diagnostic potential of serum proteomic patterns in prostate cancer*. J Urol. Aug;170(2 Pt 1):442-6, 2003.

Boyd R.K. *Linked-scan techniques for MS/MS using tandem-in-space instruments*. Mass Spectrometry Reviews 13 (5-6): 359-410, 1994.

Coombes, KR, Fritsche Jr. HA, Clarke C, Chen JN, Baggerly KA, Morris JS, Xiao LC, Hung MC, Kuerer HM. *Quality Control and Peak Finding for Proteomics Data Collected from Nipple Aspirate Fluid by Surface-Enhanced Laser Desorption and Ionization.* Clinical Chemistry, 49:10, 1615-1623, 2003.

Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, Hung MC, Kuerer HM. *Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform*. Proteomics, Vol. 5, No. 16, pp. 4107-4117, November 2005.

Cover TM, Hart PE. *Nearest neighbor pattern classification*. IEEE Trans. Inform. Theory, IT-13(1):21–27, 1967.

Covey TR, Lee ED, Henion JD. *High-speed liquid chromatography/tandem mass spectrometry for the determination of drugs in biological samples*. Anal Chem 58:2453-2460, 1986.

Diamandis E. *Commentary: Analysis of Serum Proteomic Patterns for Early Cancer Diagnosis: Drawing Attention to Potential Problems.* Journal of the National Cancer Institute, Vol. 96, No. 5, 353-356, March 3, 2004.

Dudoit S, van der Laan MJ, *Asymptotics of cross-validated risk estimation in model selection and performance assessment*, U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 126, February 5, 2003.

Fananapazir N. *FAST-AIMS User's Manual Version 1.00.02.* 2005.

Fananapazir N, Li M, Spentzos D, Aliferis CF. *Formative Evaluation of a Prototype System for Automated Analysis of Mass Spectrometry Data*. AMIA Symposium, 2005.

Fix E, Hodges JL. *Discriminatory analysis, nonparametric discrimination: Consistency properties*. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

Guyon IM, et al. *Gene selection for cancer classification using support vector machines*. Machine Learning 46:389-442, 2002.

Hilario M, Kalousis A, Muller M, Pellegrini C. *Machine learning approaches to lung cancer prediction from mass spectra.* Proteomics 2003, 3, 1716-1719.

Kohavi R. *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12): 1137–1143, 1995.

Patel S, Lyons-Weiler J. *caGEDA: a web application for the integrated analysis of global gene expression patterns in cancer.* Appl Bioinformatics. 3(1):49-62, 2004.

Petricoin III. EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishmman DA, Kohn EC, Liotta LA. *Use of proteomic patterns in serum to identify ovarian cancer*. The Lancet 2002, 359, 572-577.

Petricoin III. EF, Ornstein DK, Paweltz CP, Ardekani A, Hackett PS, Hitt BA, Velassco A, Trucco C, Wiegand L, Wood K, Simone CB, Levine PJ, Linehan WM, Emmert-Buck MR, Steinberg SM, Kohn EC, Liotta LA. *Serum Proteomic Patterns for Detection of Prostate Cancer.* Journal of the National Cancer Institute, 94:20, 1576-1578, 2002.

Quackenbush J. *Computational Analysis of Microarray Data*. Nature Reviews Genetics 2, 418-427, 2001.

Rappsilber J., et al. *Experiences and perspectives of MALDI MS and MS/MS in proteomic research*. International Journal of Mass Spectrometry 226:223-237, 2003.

Reich M, Ohm K, Angelo M, Tamayo P, Mesirov JP. *GeneCluster 2.0: an advanced toolset for bioarray analysis*. Bioinformatics. Jul 22;20(11):1797-8, 2004.

Rosty C, Christa L, Kuzdal S, Baldwin WM, Zahurak ML, Carnot F, Chan DW, Canto M, Lillemoe KD, Cameron JL, Yeo CJ, Hruban RH, Goggins M. *Identification of Hepatocarcinoma-Intestine-Pancreas/Pancreatitis-associated Protein I as a Biomarker for Pancreatic Ductal Adenocarcinoma by Protein Biochip Technology*. Cancer Research 2002, 62, 1868-1875.

Sorace JM, Zhan M. *A data review and re-assessment of ovarian cancer serum proteomic profiling*. BMC Bioinformatics 2003, 4:24.

Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. *A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis*. Bioinformatics. Mar 1;21(5):631-43, 2005a.

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. 2005b.

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.

Vapnik V. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Vejda S, Posovszky C, Zelzer S, Peter B, Bayer E, Gelbmann D, Schulte-Hermann R, Gerner C. *Plasma from cancer patients featuring a characteristic protein composition mediates protection against apoptosis*. Mol Cell Proteomics. 2002 May;1(5):387-93.

Wadsworth JT, Somers KD, Stack BC, Cazares L, Malik G, Adam BL, Wright GL, Semmes J. *Identification of Patients With Head and Neck Cancer Using Serum Protein Profiles.* Arch Otolaryngol Head Neck Surg 2004, 130:98-104.

Won Y, Song HJ, Kang TW, Kim JJ, Han BD, Lee SW. *Pattern analysis of serum proteome distinguishes renal cell carcinoma from other urologic diseases and healthy persons.* Proteomics 2003, 2:2310-2316.

Yanagisawa K, Shyr Y Xu BJ, Massion PP, Larsen PH, White BC, Roberts JR, Edgerton M, Gonzalez A, Nadaf S, Moore JH, Caprioli RM, Carbone DP. *Proteomic patterns of tumour subsets in non-small-cell lung cancer.* The Lancet 2003, 362:433-439.

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.

Yates. *Mass spectrometry and the age of the proteome*. Journal of Mass Spectrometry 1998, 33:1, 1-19.

Zhu W, Wang X, Ma Y, Rao M, Glimm J, Kovach JS. *Detection of cancer-specific markers amid massive spectral data.* PNAS 2003, 100:25, 14666-14671.