Probabilistic Disease Classification of Expression-Dependent Proteomic Data from Mass Spectrometry of Human Serum

Ryan H. Lilien[*,†], **Hany Farid**[*], and **Bruce R. Donald**[*,‡,§,¶]

Abstract

We have developed an algorithm called Q5 for probabilistic classification of healthy vs. disease whole serum samples using mass spectrometry. The algorithm employs Principal Components Analysis (PCA) followed by Linear Discriminant Analysis (LDA) on whole spectrum Surface-Enhanced Laser Desorption/Ionization Time of Flight (SELDI-TOF) Mass Spectrometry (MS) data, and is demonstrated on four real datasets from complete, complex SELDI spectra of human blood serum.

Q5 is a closed-form, exact solution to the problem of classification of complete mass spectra of a complex protein mixture. Q5 employs a probabilistic classification algorithm built upon a dimension-reduced linear discriminant analysis. Our solution is computationally efficient; it is non-iterative and computes the optimal linear discriminant using closed-form equations. The optimal discriminant is computed and verified for datasets of complete, complex SELDI spectra of human blood serum. Replicate experiments of different training/testing splits of each dataset are employed to verify robustness of the algorithm. The probabilistic classification method achieves excellent performance. We achieve sensitivity, specificity, and positive predictive values above 97% on three ovarian cancer datasets and one prostate cancer dataset. The Q5 method outperforms previous full-spectrum complex sample spectral classification techniques, and can provide clues as to the molecular identities of differentially-expressed proteins and peptides.

[*] Dartmouth Computer Science Department, Hanover, NH 03755, USA.

[†] Dartmouth Medical School, Hanover, NH 03755, USA.

[‡] Dartmouth Chemistry Department, Hanover, NH 03755, USA.

[§] Dartmouth Department of Biological Sciences, Hanover, NH 03755, USA.

[¶] Corresponding author: 6211 Sudikoff Laboratory, Dartmouth Computer Science Department, Hanover, NH 03755, USA. Phone: 603-646-3173. Fax: 603-646-1672. Email: `brd@cs.dartmouth.edu`

This work is supported by the following grants to B.R.D.: National Institutes of Health (GM 65982), National Science Foundation (IIS-9906790, EIA-0102710, EIA-0102712, EIA-9818299, and EIA-9802068), and the John Simon Guggenheim Foundation. H.F. is supported by an Alfred P. Sloan Fellowship, NSF CAREER award IIS-9983806, and a departmental NSF grant EIA-9802068.

1 Introduction[*]

Mass Spectrometry (MS) is a powerful tool for determining the masses of biomolecules and biomolecular fragments present in a complex sample mixture. The role of MS is similar to that played by 2D-gels in complex proteomic applications. Unlike gel electrophoresis, MS provides ultra-high resolution mass information. A mass spectrum consists of a set of m/z values and corresponding *relative intensities* that are a function of all ionized molecules present with that m/z ratio. The mass spectrum observed for a sample is thus a function of the molecules present. Experimental conditions that affect the molecular composition of a sample should therefore affect its mass spectrum. Mass spectrometry is therefore often used to test for the presence or absence of one or more molecules. The presence of such molecules may indicate a particular enzymatic activity, disease state, cell type, or condition. We refer to a solution from one of these 'states' containing one or more biomolecules as a *sample*.

Analysis of mass spectra by manual inspection has been feasible for samples containing a small number of molecules. These manual inspection techniques are impractical however, for samples containing a large number of protein fragments. Moreover, samples containing a large number of protein fragments tend to be the most interesting and have the potential to provide the most novel results. Recently, a number of algorithms have been developed to find spectral differences between mass spectra of samples taken from two separate conditions [7, 15, 34, 16, 14]. The discrimination of one condition from another by comparing their mass spectra is the goal of *Mass Spectrometry Classification Algorithms (MSCAs)*. Several MSCAs have been developed for human disease diagnosis as well as monitoring disease progression, regression, and recurrence [7, 25, 26, 5, 28, 3]. Given two states, we would like to know the answers to two questions: I) do molecular differences exist between the two states? and II) if molecular differences do exist, what molecules cause these differences? In mass spectrometry, question II) can be split into two parts: (a) what are the mass/charge ratios of the differently-expressed molecules? and (b) what are the molecular identities of the differently-expressed molecules? Given only the answers to I) and IIa), sample classification can, in principle, be performed. Many previous MSCAs [28, 3, 25, 7] answer I) but provide only a partial, incomplete answer to IIa). That is, many previous algorithms discover only a subset, rather than the full set, of discriminating m/z peaks. Moreover, the peaks in this subset are not guaranteed to be differentially expressed between the two states.

An MSCA can be tested empirically by comparing its accuracy to known MS data classifications, and by measuring its running time. Therefore, we compare results of our algorithm, Q5, to known assignments of MS data, and we also report that Q5 runs in minutes. This provides benchmarks, by calibration with ground truth. We motivate and define an objective error function by which linear classifiers of an MSCA can be evaluated. The classifier computed by Q5 is optimal under this error function with respect to the training set. Our characterization is twofold: (1) *Complexity* measures the running time of the algorithm. (2) *Correctness* measures how well Q5 minimizes the error function. We cast the MS classification problem into closed-form equations and then solve them using singular value decomposition, hence:

1. Q5 is a *combinatorially precise* algorithm: we can prove that the training runtime is $O(n^3 + n^2 r)$ and the testing runtime is $O(mrn)$, where n is the number of training spectra, m is the

[*] Abbreviations used: BPH, benign prostatic hypertrophy; CFES, closed-form exact solution; LDA, linear discriminant analysis; MALDI, matrix-assisted laser desorption/ionization; MS, mass spectrometry; MSCA, mass spectrometry classification algorithm; OC, ovarian cancer; PC, prostate cancer; PCA, principal components analysis; PPV, positive predictive value; PSA prostate specific antigen; SELDI-TOF, surface-enhanced laser desorption/ionization time of flight.

number of testing spectra, and r is the resolution of each mass spectrum.

2. Q5 always computes the optimal solution (with respect to the error function) using closed-form equations.

For exact algorithms such as ours, properties (1) and (2) can be proven mathematically. Hence, one can formally understand and analyze why a technique performs well, or poorly. We caution however, that *exact* does not necessarily imply perfect performance on biological data: it means the algorithm is guaranteed to optimize an objective error criterion that measures how well the (noisy) data is classified. In contrast, techniques such as genetic algorithms [28], neural networks [7, 14], and simulated annealing do not admit such guarantees: these methods have neither provable complexity nor correctness properties, and they are neither exact nor combinatorially precise.

We present a closed-form exact algorithm to answer questions I) and IIa) above. Moreover, Q5 computes the complete set of m/z peaks that are differentially expressed in one state vs. the other. This information is valuable because these peaks aid in the identification of proteins that are differentially present in each state. That is, a complete exact answer to IIa) is potentially helpful in determining the answer to IIb).

If each spectrum is sampled at the same m/z values then we can represent each spectrum as a point in an r-dimensional space, where r is the number of m/z values for which relative intensities are recorded per spectrum. We call this space *spectral-space*. Each spectrum is therefore represented in spectral-space by the point $(p_1, \ldots, p_r)$, where p_i is the relative intensity observed at the i^{th} m/z value. Similar spectra will inherently cluster in spectral-space. The assumption made by Q5 is that in spectral-space, healthy spectra form one cluster while disease spectra form a second, non-overlapping cluster (Figure 1). The hypothesis for classification is that any healthy spectrum lies closer to the healthy cluster than to the disease cluster (and vice-versa). Unclassified spectra can then be classified by assigning them to their nearest cluster. The confidence in each classification is thus a function of the distance of a sample to each cluster mean. In this representation it is reasonable to define an optimal linear discriminant using the hyperplane that maximizes the across-class variance while minimizing the within-class variance. Q5 computes, exactly, the hyperplane that satisfies this criterion.

Q5's ability to classify complex fragment mixtures was evaluated by testing its ability to discriminate the mass spectra of healthy vs. disease human serum samples. The two disease states examined in testing were ovarian and prostate cancer. Existing screening methods for both cancers carry a low positive predictive value (PPV) [18, 8]. When detected early, the 5-year survival rate increases for both cancers [30]. Improved screening techniques would be welcomed by the biomedical community.

1.1 Previous Work

An MSCA accepts as input a set of MS *training* spectra, together with their correct classifications. It outputs a *classifier (discriminant)* capable of classifying new mass spectra into one of the classes. These new spectra (called *test* spectra) have not been seen by the algorithm before and their classifications are unknown to the algorithm; the goal of the MSCA is to determine the correct classification based on the classifier constructed from the training set. *Classification verification* is the testing process by which the discriminant is evaluated for its ability to correctly classify test samples. MSCAs can be classified by the type of MS data processed, type of algorithm employed, and method of classification verification used. In the remainder of this section, we first describe

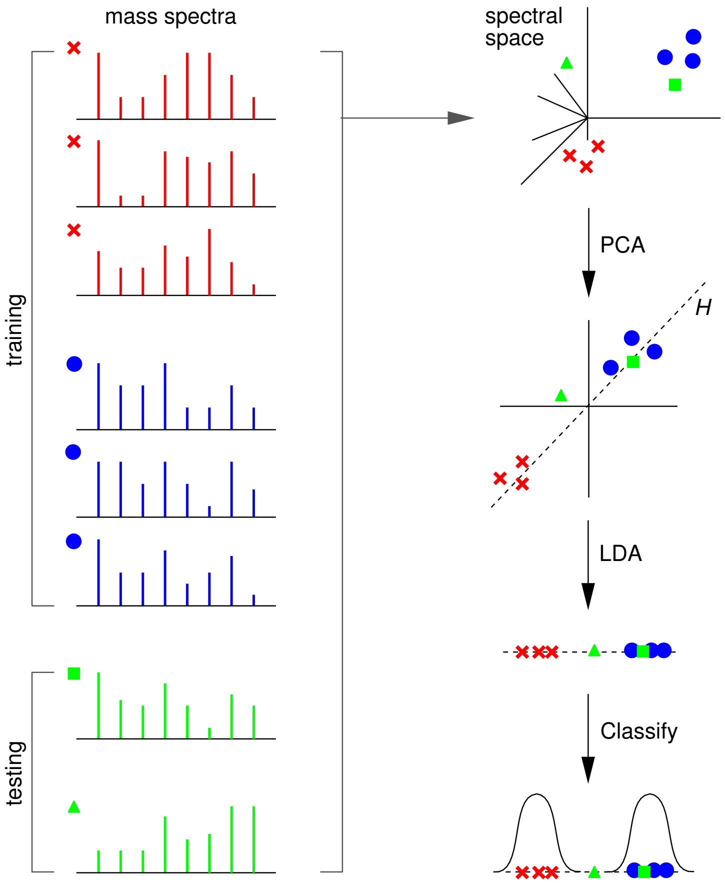

Figure 1: **Major Steps of the Q5 Algorithm.** The steps involved in building a two-class classifier are illustrated using simplified artificial spectra. On the left are training (×, ∘) and testing (□, △) spectra. Shown on the right, from top to bottom, are (1) the spectral space representation of each spectrum; (2) PCA: the result of dimensionality reduction (for simplicity we show the projection onto just the top two principle components); (3) LDA: the projection of each spectrum onto the discriminant surface H; and (4) the probabilistic classification. In this example the testing spectrum denoted by □ is classified as belonging to the class denoted by ∘, while the spectrum denoted by △ is unclassified.

a framework in which MSCAs can be compared. We then review previous work utilizing this framework. Finally we summarize the key differences between previous work and Q5.

Completeness of mass spectrum. Analysis may be performed on either *complete* or *partial* mass spectra. Complete mass spectra consist of the relative intensities of all m/z values acquired during MS data collection. This includes the relative intensities observed for all m/z values from 0 up to the upper limit of detection. An MSCA that processes complete mass spectra works with the entire recorded spectra: no values are “manually” excluded through preprocessing. When portions of a spectrum are excluded from consideration, we say a partial spectrum is generated.

Manual preprocessing. Frequently, spectra are manually preprocessed. In manual preprocessing, parts of the spectrum may be eliminated from consideration based on the magnitude of the relative intensity or prior (human) knowledge. This spectral manipulation produces a manually-processed partial spectrum. Modification of the peak intensities in a manner that imparts additional information represents another type of manual manipulation.

Sample source (the biological source of the MS sample)**.** Spectra may be obtained from either *simple* or *complex* fragment mixtures. Simple mixtures may contain one or only a small number of proteins and usually yield relatively "clean" spectra with fewer peaks. Complex fragment mixtures contain between tens to thousands of biological fragments and produce a commensurate number of m/z peaks. These peaks often present a challenge to MS analysis algorithms: a particular m/z peak may be the sum of many sub-peaks (contributions) from many different molecular fragments. Human serum (used in our application) is, for example, a complex fragment mixture.

Heuristic vs. Exact Classification Algorithms. *Heuristic* classification algorithms include approaches such as genetic algorithms, neural networks, and simulated annealing. These algorithms generally require multiple iterations to converge to a classifier; furthermore, the solution found by heuristic algorithms is not guaranteed to be optimal. In addition, many heuristic approaches are *non-deterministic*. Even when run on the same training set, the same non-deterministic algorithm often converges to a different discriminant. In contrast, MSCAs that utilize *closed-form exact solutions (CFES)* compute an exact solution using closed-form equations. CFES algorithms are computationally efficient; they are non-iterative and deterministic (i.e., always compute the same solution). Linear Discriminant Analysis (LDA), and thus Q5, are CFES algorithms.

Classification Verification. Classification algorithms must be verified, to confirm that the discriminant will properly classify samples that were *not* used in training. To test a classification algorithm, it is essential to perform multiple *leave-out* experiments, each with a different split between the training and testing (masked) sets. Often, there exists a split of samples into training and testing sets that performs significantly better than others. The performance statistics of the classifier against multiple different splits must therefore be reported. When only one or a small number of splits are tested, classification verification is said to be *partial.*

The use of a testing set also allows one to confirm that the discriminant has not been over-fit to the training data. If the discriminant has been overfit to the training spectra then one would expect excellent performance in the classification of the training spectra but poor performance in the classification of the testing spectra.

We now describe specific examples of existing MSCAs in previous research. Each example is discussed using the framework introduced above. We then describe Q5's advantages over previous techniques.

Heuristic Classification Techniques. Petricoin et al. give a heuristic MSCA on complete complex spectra with classification verification [28]. The method employs a genetic algorithm to select between 5 and 20 m/z peaks for use in classification. This MSCA was applied to SELDI[†]-TOF spectra of blood serum from 100 women with ovarian cancer, 100 women without cancer, and 17 women with benign gynecological disease. A genetic algorithm was trained on a set of spectra containing half of the cancer spectra and half of the normal spectra. The remaining 117 spectra were used in testing. A single training/testing split was performed; a sensitivity of 100%, specificity of 95%, and PPV of 94% was reported. Petricoin and co-workers have recently tested their MSCA against two additional sets of ovarian cancer and one set of prostate cancer SELDI mass spectra [2].

Adam et al. developed a decision tree based heuristic MSCA on partial complex spectra with partial classification verification for the diagnosis of prostate cancer [3]. A training set containing 85% of the total samples (n=326) was used to build the decision tree. The MSCA started with a

[†] *Surface-Enhanced Laser Desorption/Ionization (SELDI)* [17, 23] is a variant of the commonly used Matrix-Assisted Laser Desorption/Ionization (MALDI) [20] Mass Spectrometry. In SELDI MS, molecular samples are placed onto protein chips with selective affinity before ionization. Nonbinding molecules are washed off the chip and remaining molecules are ionized by laser bombardment. Ions are subsequently processed by a mass analyzer.

subset of 124 MS peaks and built a three-class decision tree using 9 of these. Partial testing using a single training/testing split resulted in a sensitivity of 83%, specificity of 97%, and PPV of 96%.

Another heuristic MSCA based on discriminant factorial analysis has been used to discriminate between betamethasone and dexamethasone [4]. Discriminant factorial analysis is an iterative technique that attempts to converge to the answer directly computed by LDA. Additionally, while not truly MSCAs, a number of papers have reported on heuristic techniques for the identification of differentially expressed m/z peaks. Artificial neural networks have been used to identify m/z peaks associated with astrocytoma [7] as well as bacteria involved in urinary tract infections [14].

Exact Classification Algorithms. Two recent works applied LDA to MS analysis. Miketova et al. performed LDA on a subset of peaks to differentiate Gram positive vs. Gram negative bacteria [24]. They present an exact algorithm on manually-processed partial complex spectra without classification verification. Their analysis used reduced dimensionality electron ionization mass spectra containing the relative intensities of 36 hand-picked m/z values. These 36 values surrounded 12 low-resolution mass peaks that had been shown in previous work to have discriminating power. The linear discriminant was computed on a training set of 36 sample spectra (18 Gram positive and 18 Gram negative). Although the computed discriminant was able to separate the training samples, its ability to classify novel samples was not evaluated.

Wagner et al. present an exact algorithm on manually-processed partial simple spectra with classification verification [32]. They performed TOF-SIMS (Time-of-Flight Secondary Ion MS) on a small number of proteins (12) each prepared as a single protein adsorbed film using one of two substrates. Replicate experiments were performed which generated spectra covering only the single amino-acid mass range of 0-200 m/z. Analysis was performed using two sets of peaks: the first set consisted of a preselected peak list while the second set contained all peaks with an intensity at least three times greater than the 0-200 m/z background region. They compared the discriminating power of principal components analysis (PCA), discriminant principal component analysis (DPCA), and LDA. Leave-one-out experiments were performed on multiple training/testing splits. The linear discriminant was used to predict the identity of an unknown single protein adsorbed film from its mass spectrum. Among their results, they showed that LDA and DPCA provide better discriminating power than PCA.

Goodacre et al. performed LDA on the Electrospray Ionization (ESI) mass spectra collected from 3 replicates of 6 different bacteria [13]. Partial spectra from 100-3050 m/z were used in analysis and no classification verification was performed. That is, no test spectra were classified using the computed discriminant.

In summary, of the existing MSCAs, those that use complete complex spectra [28, 3], do not use exact algorithms. Conversely, those MSCAs that are exact, do not operate on complete complex spectra [24, 32]. Moreover, only partial classification verification results have been reported for essentially all existing MSCAs. In these respects, Q5 differs from all previous MSCAs. Whereas Petricoin et al. and Adam et al. used heuristic methods, Q5 uses LDA, an exact method. In contrast to the work of Miketova et al. and Wagner et al., we do not remove from consideration parts of the recorded mass spectrum based on relative-intensity or *a priori* (human) knowledge. Our work utilizes affinity chip filtered human serum containing tens to thousands of proteins and protein fragments. Q5 uses complete mass spectra, sampled at 15154 (resp. 16382) m/z values over the range 0-20000 (resp. 0-22500), to compute a discriminant for ovarian (resp. prostate) cancer datasets. Whereas Wagner et al. classified unknown spectra by assigning them to the nearest class, we employ a novel probabilistic classification framework. For each unclassified testing spectrum, Q5 computes both the most likely class assignment as well as the probability that the unknown spectrum belongs to the specified class. Whereas only partial classification verification has been reported on existing MSCAs, Q5 is tested with several thousand training/testing splits. Q5 is, to our knowledge, the first closed-form exact solution to the problem of probabilistically classifying complete mass-spectra of a complex protein mixture. Finally, Q5 employs a discriminant back-projection algorithm to compute clues as to the molecular identities of differentially-expressed proteins and peptides.

2 Methods

We have designed and implemented Q5 to classify complex samples from mass spectrometry data. The major steps of Q5 are illustrated in Figure 1. In our algorithm, each spectrum is represented by a point in spectral-space, as described above. The set of all spectral points in spectral-space is dimensionality-reduced using *Principal Components Analysis (PCA)* [10]. In particular, PCA performs a transformation of spectral-space into a lower dimensional space with little or no information loss. A hyperplane, H, is then computed using *Linear Discriminant Analysis (LDA)* [11, 10]. The PCA dimensionality reduced sample points are projected onto H. The hyperplane H maximizes the across-class variance while minimizing the within-class variance of the projected sample points [11]. Thus, the LDA-computed hyperplane H satisfies our exactness criterion. As a result, classification is made easier in this projected space. Now suppose we wish to classify some new (test) spectra (that were not used in training). A test spectrum is first dimensionality-reduced by projecting onto the retained principal components. Next, it is projected onto the hyperplane H. Finally, if the classification confidence is above a threshold then the point is classified into the healthy or disease state. The confidence in classification is based on a symmetric Gaussian distribution centered at each class mean. The process of classifying two test spectra, represented by $\square$ and $\triangle$, is illustrated in Figure 1.

The three spaces used by Q5 are spectral-space, *PCA-space*, and *discriminant-space*. Spectral-space has been described above. PCA-space is the space spanned by the principal components retained from the PCA dimensionality reduction; discriminant-space H is the space spanned by the linear discriminant(s) computed from LDA. Discriminant-space has lower dimensionality than PCA-space which has lower dimensionality than spectral-space.

2.1 Principal Components Analysis

Principal component analysis (PCA) is often used in the analysis of points that are embedded in a high-dimensional space. PCA is a method for determining orthogonal axes of maximal variance from a dataset [10]. PCA is an *unsupervised* technique: the classification of each sample point is not considered in analysis. Sample points are zero-meaned and an eigendecomposition of the covariance matrix computed. The eigenvector associated with the i^{th} largest eigenvalue lies along the i^{th} principal component. Typically most sample point variance is captured by the first few principal components, (i.e., those with the largest eigenvalues). Projecting a dataset onto these largest principal components reduces sample dimensionality while maximally preserving variance. Two disjoint sets of points and the first PCA computed principal component (solid line) are shown in Figure 2A. PCA is used by Q5 for dimensionality reduction: it is not, and should not, be used to compute a linear separator directly. For example, the projection of the sample points from Figure 2A onto the first principal component are overlapping (Figure 2B) and are not classifiable. PCA is only used in Q5 to reduce the dimensionality of the sample points (with little or no information loss), as is required by LDA. See appendix A for details.

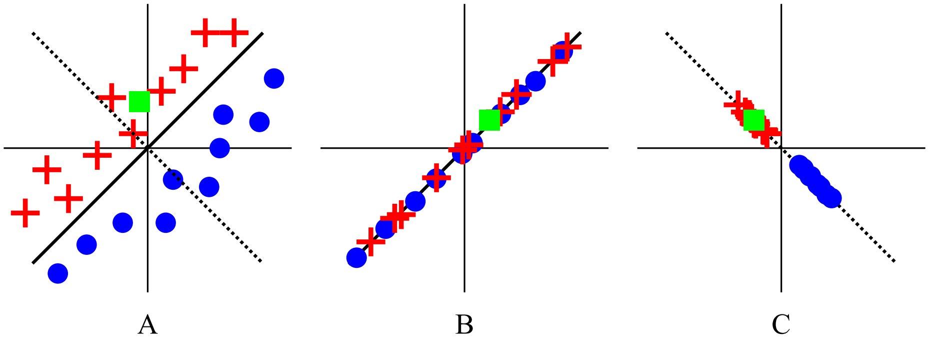

Figure 2: (A) Two disjoint sets of zero-meaned points in two-dimensional space. The first PCA component (solid line) and the LDA discriminant (dotted line) are shown. (B) Projection of both point sets onto the first principal component (solid line). This projection does not separate the two sets. (C) Projection of both point sets onto the LDA-computed discriminant (dotted line). The two sets of points are well separated. A test sample (green square) is easily classified by projecting onto the LDA-computed discriminant.

2.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [11, 10] of a system with k classes in d dimensions computes, in closed-form, $k-1$ orthogonal vectors, each of dimension d, which specify a hyperplane of dimension $k-1$. Projection of the sample points onto this hyperplane maximizes the between-class scatter and minimizes the within-class scatter (Figure 2C). For the purposes of sample classification, such a projection is clearly desirable, because it simultaneously reduces the dimensionality of the data and preserves the ability to discriminate one class from another. Whereas other discriminant-based approaches (i.e. Discriminant Factorial Analysis (DFA)) attempt to converge to the optimal separator through multiple iterations [4], LDA computes the optimal discriminant directly in closed-form.[‡] LDA is a *supervised* technique: the class membership of each sample is utilized in computing the discriminant. See appendix B for details.

2.3 Back-Projection

The LDA-computed linear discriminant can be back-projected from a PCA-space discriminant into a spectral-space discriminant. A spectral-space discriminant allows one to determine the m/z values of peaks used to differentiate between members of the two classes. This information is in principle useful in determining the molecular identities of differently-expressed biomolecules. The spectral-space linear discriminant, $\mathbf{e}^{\bullet}$, can be computed from the PCA-space linear discriminant, $\mathbf{e}$, by left-multiplying by the transpose of the principle component matrix, $\mathbf{V}$ (Eq. 8):

$$\mathbf{e}^{\bullet} = \mathbf{V}^T\mathbf{e}. \tag{1}$$

To determine which m/z values of the discriminant contribute most to classification, the spectral-space discriminant should be normalized by the average intensity of the zero-meaned spectra. Thus

[‡] A natural extension to the use of LDA is the use of Support Vector Machines (SVM) [31, 21, 33]. To this end, we replaced the LDA classifier with a non-linear SVM using a Gaussian kernel. Classification accuracy with the SVM was as good or slightly worse than the LDA. This suggests that, in our examples, the healthy/disease spectra are reasonably well-separated by a hyperplane, so that the benefits of a non-linear classifier will be, at best, minimal. Additionally, we note that back-projection to determine molecular identity (Sec. 2.3) is not straight-forward with SVMs.

an r-dimensional significance vector $\mathbf{s}$ can be computed, with components

$$s_i = \left|\mathbf{e}^{\bullet}_i \left(\bar{\mathbf{y}}_i - \boldsymbol{\mu}'_i\right)\right| \qquad (i = 1, \ldots, r) \tag{2}$$

where $\bar{\mathbf{y}}$ is the average cancer spectrum, $\boldsymbol{\mu}'$ is the all-class mean (Eq. 3), and r is the dimensionality of spectral-space. Each s_i thus represents the significance of the i^{th} m/z value of the mass spectrum for classification.

2.4 Probabilistic Classification

In the simplest case, a novel sample is classified into the class with the closest class mean. However, if the sample spectrum's projection into discriminant space is nearly equidistant to two or more class means then the confidence of classification should be reduced. Thus, a classifier should report not only the classification of a given sample but also the confidence in that classification. A probabilistic framework for reporting classification likelihoods was therefore implemented in Q5. The classification probability of each spectrum is computed from the distance in discriminant-space between the spectrum and the nearest class mean. See appendix C for details.

2.5 Q5 Testing

The testing of Q5 against each dataset consists of *D-experiments* and *D-runs*. In the first step of a D-experiment (Figure 3), the set of all n sample spectra is randomly partitioned into a training set T and a testing (masked) set M. Following this partition, Q5 performs PCA on the spectral points in set T. The result of PCA is that each spectrum is now a point in the $(n-3)$-dimensional PCA-space. The optimal separating hyperplane is next computed using LDA on the PCA dimensionality-reduced training spectra. The discriminant-space sample points from each class should be inherently clustered. The center of each cluster is then computed and used in probabilistic classification. A spectrum with a probability of classification less than a fixed threshold is not classified by Q5. A collection of s D-experiments is called a D-run. For each of the four datasets, four D-runs are performed with training sets consisting of 50%, 75%, 85%, and 95% of the total samples. For example, the 75% D-Run consists of s D-experiments; in each D-experiment a different random 75% of the total samples is assigned to the training set and the remaining 25% is assigned to the testing set. To illustrate the robustness of Q5, $s = 1000$ D-experiments were performed in each D-run. For each D-experiment the percent-classified, percent-correctly classified, positive predictive value, sensitivity, and specificity were computed using probability classification thresholds evenly sampled between 0.5 and 1.0. The mean and standard deviation of these values are computed for each D-run.

2.6 Algorithmic Complexity

Let n be the number of training spectra, m be the number of testing spectra, r be the resolution of each mass spectrum, and k be the number of classes. PCA requires computing the top n eigenvectors of the covariance matrix $\mathbf{C}$ built from the training spectra (Eq. 5). The eigenvectors of $\mathbf{C}$ can be computed efficiently using the $n \times n$ Gram matrix (Eq. 6). The Gram matrix $\mathbf{C}'$ can be constructed in $O(n^2 r)$ time and its eigenvectors computed in $O(n^3)$ time. Computing the eigenvectors of $\mathbf{C}$ from the eigenvectors of $\mathbf{C}'$ then requires $O(n^2 r)$ time. Projection of the training spectra onto the PCA basis requires $O(n^2 r)$ time. Computing the LDA discriminant entails computing the generalized eigenvectors of the $(n-3) \times (n-3)$ within- and between-class scatter matrices (Eqs. 10 and 11). This can be done in time $O(n^3)$. Projection onto the discriminant requires $O(kn)$ time. Finally,

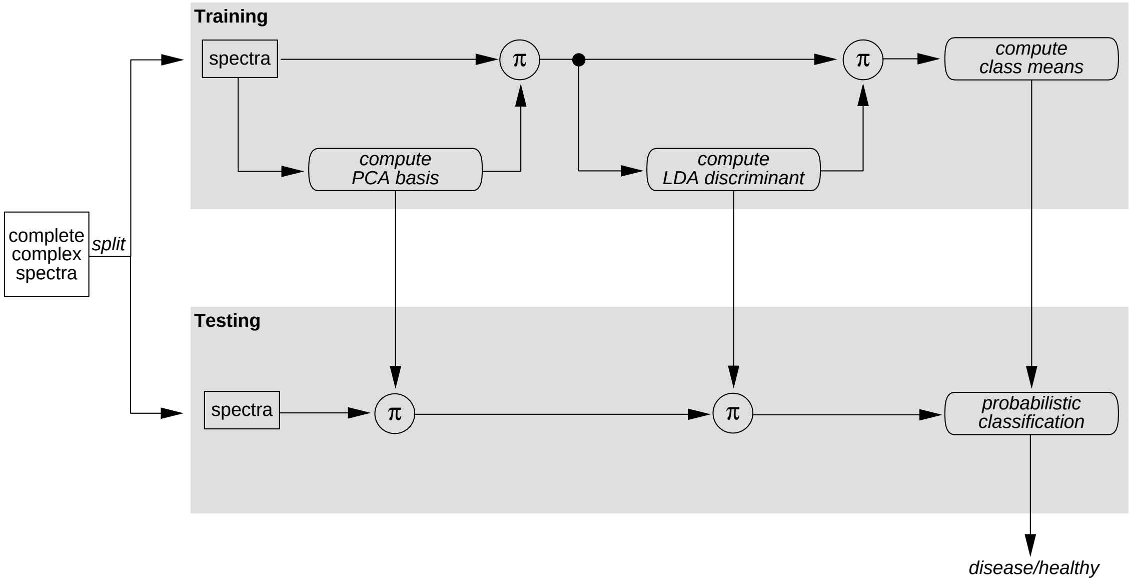

Figure 3: **One D-experiment.** A complete complex spectral dataset is first partitioned into training and testing sets. A PCA basis is then computed from the training spectra. These spectra are projected (denoted by π) onto the PCA basis creating the PCA-space representation. The spectra in this space are then projected onto the computed LDA discriminant. The class means and Gaussian probability distribution standard deviation, σ are computed from this representation. In the testing stage, the testing spectra are first projected onto the PCA basis, then onto the LDA discriminant and then classified.

since each spectrum can appear in at most one cluster, computing the class means requires $O(n)$ amortized time. Therefore training can be accomplished in time $O(n^3 + n^2r)$. The testing of m sample spectra can be performed in $O(mrn + mnk + mk^2)$ time: The all-class training mean can be subtracted in $O(r)$ time. Projection onto the PCA basis is $O(rn)$ time; projection onto the LDA discriminant is $O(nk)$ time. The nearest cluster mean is computed in time $O(k^2)$ and the classification probability computed in time $O(1)$. In our studies, $k \leq 3$ and $r \geq 15154$. Since we expect $k \ll r$, if we assume $k = O(1)$, the testing of one spectrum can be performed in time $O(rn)$ and m spectra can be classified in time $O(mrn)$. The PCA and LDA computations on the training set require 1.0 to 1.5 minutes of runtime on a Pentium 4 class workstation. Classification of a novel sample can be performed in under a second.

2.7 Implementation

Datasets were obtained from the NIH and FDA Clinical Proteomics Program Databank [2] and the Eastern Virginia Medical School [1]. Each spectrum in these datasets is contained in either an individual or a grouped file and is separated into either a healthy or a disease subdirectory. These datafiles are in either comma-delimited or Microsoft Excel format. Datafile reading, PCA, LDA, and probabilistic classification are implemented in MATLAB (Mathworks Inc, Natick, MA).

Each D-run is processed separately by Q5: the specified dataset is loaded and 1000 D-experiments are performed each with a random training/testing split. For each D-experiment, the training sample mean and discriminant-space projections of both the training and testing spectra are computed and saved. Subsequently, Q5 computes the PPV, sensitivity, specificity, percent correct, and percent

A

			Num Samples		
Data Set	SELDI Chip	Processing	Healthy	All Stages OC	Stage I OC
OC-H41	H4	Manual	100	100	24
OC-WCX2a	WCX2	Manual	100	100	-
OC-WCX2b	WCX2	Robotic	91	162	28

B

			Num Samples		
Data Set	SELDI Chip	Processing	Healthy	Benign Hypertrophy	Prostate Cancer
PC-IMAC-Cu	IMAC-3 (Cu)	Manual	81	78	168

Table 1: Details of (A) the three ovarian cancer datasets [28] and (B) the prostate cancer dataset [3] used in the testing of Q5.

classified for each D-experiment. These statistics are a function of the threshold used in probabilistic classification. Statistics were therefore computed for probability classification thresholds evenly spaced between 0.5 and 1.0 (see Figures 4 and 5).

3 Results and Discussion

Q5 was applied to classify three ovarian cancer and one prostate cancer dataset. In this section we report on the performance of the Q5 algorithm and compare these results, where possible, to previous MSCAs.

All datasets are complete complex spectra from SELDI-TOF MS experiments. Datasets were provided by Dr. Emanuel Petricoin III and Dr. George Wright Jr. The Petricoin group MS spectra were obtained from the NIH and FDA Clinical Proteomics Program Databank [2]. The Wright group MS spectra were obtained from the Eastern Virginia Medical School - Virginia Prostate Center [1]. We named each dataset by the cancer type screened (Ovarian Cancer (OC) or Prostate Cancer (PC)) and the SELDI affinity chip used in MS (H4 (Hydrophobic), WCX2 (Weak Cation Exchange - negative), or IMAC-Cu (Immobilized Metal Affinity Capture - coated with $CuSO_4$)). The healthy samples in the ovarian cancer datasets come from women at risk for ovarian cancer (the demographic most likely to use and benefit from serum screening) while the ovarian cancer positive samples come from women with tumors spanning all major epithelial subtypes and stages of disease. The samples from both OC-H4 and OC-WCX2a were manually prepared; the OC-WCX2b samples were prepared by a robotic instrument. Serum samples in the prostate cancer dataset [3] were collected and processed manually from men with normal prostates, benign prostatic hypertrophy, and all four stages of prostate cancer. The IMAC-3 affinity chip was coated with $CuSO_4$ and used in the SELDI MS experiment. All SELDI chips are produced by Ciphergen Biosystems (Freemont, CA, USA). The baseline was subtracted by the labs preparing the datasets OC-H4, OC-WCX2a, and PC-IMAC-Cu ; this results in some m/z peaks having negative relative intensities. The Petricoin group normalized the relative intensities of each sample in dataset OC-WCX2b to lie between 0 and 100. We performed no additional preprocessing on these datasets. The datasets are sampled at 15154 (resp. 16382) m/z values over the range 0-20000 (resp. 0-22500). All m/z points and relative intensities in the collected spectra are used in Q5 spectral analysis; none are discarded. Further details of the datasets are given in Table 1 and have been described previously [28, 2, 3].

The initial dimensionality of spectral-space (15154 for the ovarian cancer spectra [2], 16382 for the prostate cancer spectra [1]) is typically larger than the *intrinsic* dimensionality of the training set. Although the complete training spectra exist in 15154- or 16382-dimensional space, the intrinsic dimensionality of these points is bounded by the number of training samples. LDA

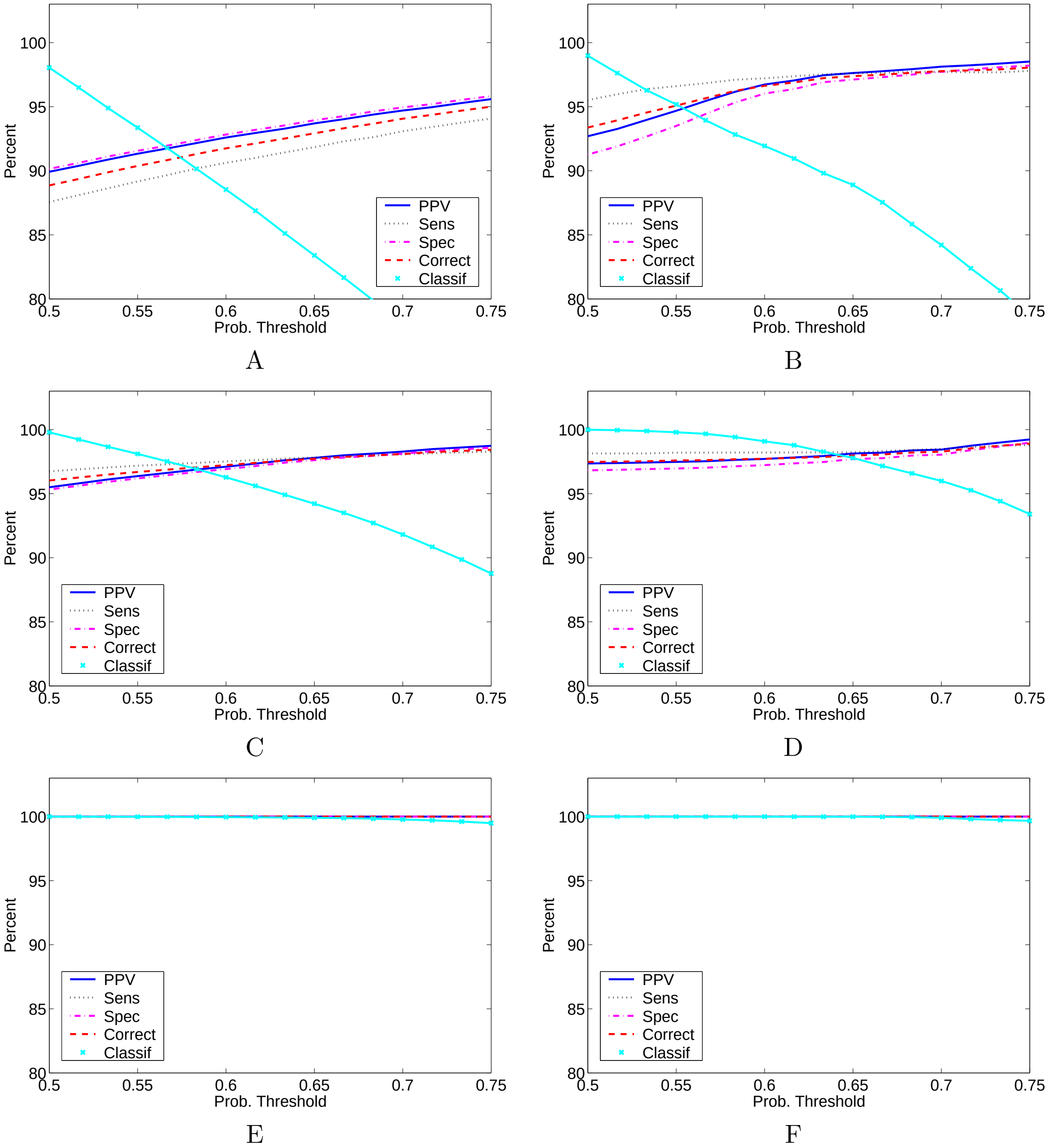

Figure 4: The probability classification threshold vs. percent-classified (Classif), percent-correctly classified (Correct), positive predictive value (PPV), sensitivity (Sens), and specificity (Spec) for six D-runs of Q5 on the ovarian cancer datasets. (A) OC-H4, 50% of samples used in training, (B) OC-H4, 95% of samples used in training. (C) OC-WCX2a, 50% of samples used in training, (D) OC-WCX2a, 95% of samples used in training. (E) OC-WCX2b, 50% of samples used in training, (F) OC-WCX2b, 95% of samples used in training. Increased probability classification thresholds increase Q5's percent-correctly classified, positive predictive value, sensitivity, and specificity while decreasing the percent-classified. Performance on the robotically prepared OC-WCX2b dataset is near perfect, see Table 2

Dataset	T%	PCT	% Corr	% Classif	PPV	Sens	Spec
OC-H4	50%	0.50	88.86 (3.01)	98.04 (1.45)	89.92 (3.71)	87.57 (5.12)	90.15 (4.04)
		0.63	92.46 (2.76)	85.46 (3.42)	93.25 (3.49)	91.36 (4.84)	93.49 (3.57)
		0.75	95.00 (2.51)	72.22 (4.29)	95.59 (3.33)	94.08 (4.42)	95.82 (3.28)
	75%	0.50	92.20 (3.62)	98.60 (1.55)	92.13 (4.71)	92.45 (5.51)	91.95 (5.15)
		0.63	95.53 (3.03)	87.83 (4.28)	95.44 (4.21)	95.70 (4.48)	95.35 (4.44)
		0.75	97.63 (2.34)	76.54 (5.41)	97.88 (3.22)	97.37 (3.62)	97.90 (3.20)
	85%	0.50	92.70 (4.43)	98.82 (1.89)	92.23 (5.86)	93.58 (6.27)	91.82 (6.64)
		0.63	96.15 (3.55)	88.61 (5.86)	96.16 (4.72)	96.33 (5.20)	95.95 (5.09)
		0.75	98.11 (2.64)	77.18 (7.65)	98.67 (3.19)	97.58 (4.43)	98.62 (3.33)
	95%	0.50	93.38 (7.34)	98.98 (3.09)	92.70 (10.00)	95.52 (8.98)	91.27 (12.43)
		0.63	97.18 (5.31)	89.96 (9.22)	97.40 (6.69)	97.51 (7.25)	96.81 (8.24)
		0.75	98.05 (4.72)	78.66 (12.79)	98.52 (5.45)	97.79 (7.57)	98.20 (6.73)
OC-WCX2a	50%	0.50	96.03 (1.87)	99.78 (0.48)	95.50 (2.97)	96.74 (2.36)	95.32 (3.25)
		0.63	97.54 (1.55)	95.05 (2.10)	97.55 (2.25)	97.71 (2.08)	97.36 (2.51)
		0.75	98.43 (1.36)	88.78 (3.06)	98.73 (1.74)	98.28 (1.96)	98.58 (2.01)
	75%	0.50	97.16 (2.19)	99.92 (0.40)	97.10 (3.02)	97.34 (3.12)	96.99 (3.22)
		0.63	98.07 (1.83)	97.09 (2.46)	98.30 (2.38)	97.96 (2.88)	98.19 (2.58)
		0.75	98.86 (1.58)	92.41 (3.61)	99.22 (1.74)	98.61 (2.59)	99.13 (1.97)
	85%	0.50	97.33 (2.85)	99.97 (0.30)	97.25 (3.83)	97.57 (4.01)	97.08 (4.16)
		0.63	97.98 (2.40)	97.78 (2.92)	98.13 (3.18)	97.98 (3.69)	97.99 (3.45)
		0.75	98.95 (1.94)	93.04 (4.74)	99.27 (2.12)	98.74 (3.19)	99.17 (2.39)
	95%	0.50	97.48 (4.76)	99.99 (0.32)	97.35 (6.37)	98.14 (5.88)	96.82 (7.74)
		0.63	97.85 (4.43)	98.34 (3.91)	97.92 (5.69)	98.23 (5.77)	97.46 (6.99)
		0.75	98.90 (3.27)	93.41 (7.70)	99.23 (3.54)	98.79 (4.93)	98.98 (4.79)
OC-WCX2b	50%	0.50	99.99 (0.08)	100.00 (0.07)	100.00 (0.05)	99.99 (0.12)	100.00 (0.10)
		0.63	100.00 (0.05)	99.93 (0.24)	100.00 (0.00)	100.00 (0.08)	100.00 (0.00)
		0.75	100.00 (0.03)	99.49 (0.61)	100.00 (0.00)	100.00 (0.04)	100.00 (0.00)
	75%	0.50	100.00 (0.00)	100.00 (0.05)	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)
		0.63	100.00 (0.00)	99.98 (0.19)	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)
		0.75	100.00 (0.00)	99.67 (0.67)	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)
	85%	0.50	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)
		0.63	100.00 (0.00)	99.98 (0.20)	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)
		0.75	100.00 (0.00)	99.71 (0.82)	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)
	95%	0.50	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)
		0.63	100.00 (0.00)	99.99 (0.23)	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)
		0.75	100.00 (0.00)	99.67 (1.50)	100.00 (0.00)	100.00 (0.00)	100.00 (0.00)

Table 2: Predictive results on all three ovarian cancer datasets. T%: Training Percent, PCT: Probability Classification Threshold, % Corr: Percent Correctly Classified, % Classif: Percent Classified, PPV: Positive Predictive Value, Sens: Sensitivity, Spec: Specificity. Values listed are means in percent, standard deviations are in parentheses.

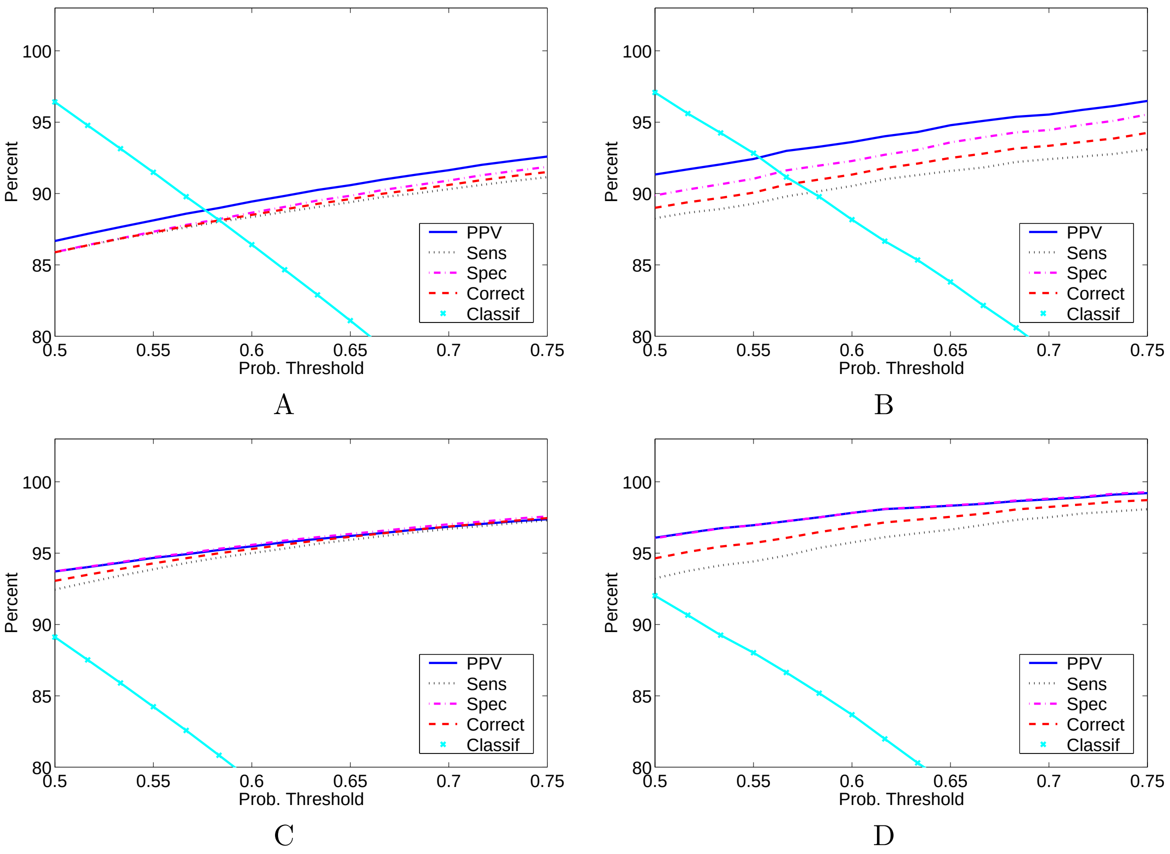

Figure 5: The probability classification threshold vs. percent-classified (Classif), percent-correctly classified (Correct), positive predictive value (PPV), sensitivity (Sens), and specificity (Spec) for two D-runs of Q5 on the PC-IMAC-Cu dataset. (A) 2-Class LDA, 50% of samples used in training, (B) 2-Class LDA, 95% of samples used in training. (C) 3-Class LDA, 50% of samples used in training, (D) 3-Class LDA, 95% of samples used in training. Increased probability classification threshold increases Q5's percent-correctly classified, positive predictive value, sensitivity, and specificity while decreasing the percent-classified.

can not be performed on a set of points in a space with dimensionality larger than the set's intrinsic dimensionality. That is, in order to guarantee a non-degenerate solution for LDA, the dimensionality of the data must be reduced to at most $n-k$ where n is the number of samples and k is the number of classes [11]. Therefore, since the intrinsic dimensionality of the training samples is no more than 95% of the total number of samples (327 in the largest dataset) we must project the spectra into a lower dimensional space. For this reason, PCA is performed on each training set. We use the $n-3$ largest principal components in dimensionality reduction since both two- and three- class LDA experiments are performed.

We now describe the results of running Q5 on the four datasets [28, 2, 3]. Q5 achieves performance results that compare favorably to previous work. For all datasets, a number of the training/testing splits result in 100% classification accuracy.

Ovarian Cancer. Q5 was applied to the three ovarian cancer datasets (OC-H4, OC-WCX2a, and OC-WCX2b) [2, 28]. The results of this analysis are given in Table 2 and Figure 4. For each dataset, a D-run was performed with training sets consisting of 50%, 75%, 85%, and 95% of the total number of sample spectra. Thus a total of 12,000 D-experiments were performed across these 12 D-runs. As one increases the probability classification threshold the percent-classified decreases.

At the same time, increasing the threshold increases the percent-correctly classified, sensitivity, specificity, and positive predictive value. Thus a higher threshold allows for increased classification accuracy at the cost of a decreased number of samples classified. A classification threshold exists that allows Q5 to classify 90.0% of the OC-H4 samples with a PPV of 97.4%, a sensitivity of 97.5%, and a specificity of 96.8%. Q5 achieves better performance statistics on the WCX datasets. Q5 can classify 93.4% of the OC-WCX2a samples with a PPV of 99.2%, a sensitivity of 98.8%, and a specificity of 98.9%. Q5's best performance is achieved on the OC-WCX2b dataset, classification is perfect in all 3000 D-experiments beyond the 50% training level. That is 100% of the samples are classified with a PPV of 100%, a sensitivity of 100%, and a specificity of 100%. It is worth noting that for each dataset tested there exists a probability classification threshold which achieves perfect classification in a majority of the D-experiments. The variance in classification performance illustrates the importance of reporting MSCA results on multiple different train/test splits.

Prostate Cancer. Q5 was then tested against the PC-IMAC-Cu prostate cancer dataset [3]. Q5 was used to compute both a two- and three- class discriminant. Each sample in the PC-IMAC-Cu dataset is classified as either Normal Healthy (NH), Benign Prostatic Hypertrophy (BPH), or Prostate Cancer (PC). For the two-class discriminant tests, both NH and BPH samples were considered 'healthy' while PC samples were considered 'disease'. The two-class discriminant tests consist of 4 D-runs, performed with training sets containing 50%, 75%, 85%, and 95% of the total number of sample spectra. As in the ovarian cancer tests, each D-run of the prostate cancer tests consisted of 1000 D-experiments. The percent-classified, percent-correctly classified, positive predictive value, sensitivity, and specificity are reported in Table 3 and Figures 5A and 5B. One set of three-class experiments were performed. In the three-class experiment each sample was classified as either NH, BD, or PC, Table 4. Similar to the other datasets, 4 D-runs were performed with training sets containing 50%, 75%, 85%, and 95% of the total number of sample spectra. The results of the three-class experiments are shown in Table 4 and Figures 5C and 5D. As was the case with the ovarian cancer classification, the prostate cancer classification showed a tradeoff between the accuracy and the percent of samples classified. In the 2-class experiments Q5 was able to classify 85.6% of the PC-IMAC-Cu samples with a PPV of 94.3%, a sensitivity of 91.3%, and a specificity of 93.0%. In the 3-class experiments Q5 classified 92.0% of the samples with a positive predictive value of 96.1%, a sensitivity of 93.2%, and a specificity of 96.1%. If we allow only 67.1% of samples to be classified Q5 achieves a PPV of 99.2%, a sensitivity of 98.1%, and a specificity of 99.3%. Table 5 shows three-way classification results.

For comparison, we forced Q5 to classify 100% of the training spectra. In this experiment each spectrum was classified into the class with the nearest class mean. The results of classifying 100% of the samples for each training percent are shown in the top row of each training percent in Tables 3 and 4. This 'complete' classification achieves an accuracy approximately equal to that obtained using a 0.5 probability classification threshold. We note that these accuracies are not as high as those achieved when a larger probability classification threshold is used. This illustrates that increased predictive accuracy can be achieved by not classifying 'ambiguous' spectra.

To verify that the discriminant computed by Q5 is not overly sensitive to outliers, we retrained the classifier for the ovarian cancer dataset OC-H4 using 75% of the data for training and misclassified 5% of this data in the training stage. With a probability classification threshold of 0.5, we are able to classify 97.80% of the testing data: 87.74% is correctly classified, with a PPV of 88.60%, a sensitivity of 86.82%, and a specificity of 88.65%. This compares reasonably well to the result with perfectly classified training data for the same dataset, where we are able to classify 98.60% of the testing data: 92.20% is correctly classified, with a PPV of 92.13%, a sensitivity of 92.45%, and a specificity of 91.95% (see Table 2). We conclude that a small number of misclassified training

Dataset	T%	PCT	% Corr	% Classif	PPV	Sens	Spec
PC-IMAC-Cu	50%		86.28 (2.71)	100.00 (n/a)	87.07 (3.65)	86.20 (3.96)	86.37 (4.46)
		0.50	85.88 (2.79)	96.42 (1.74)	86.67 (3.74)	85.88 (4.07)	85.89 (4.57)
		0.63	89.20 (2.62)	83.25 (3.09)	90.16 (3.70)	88.99 (3.83)	89.41 (4.38)
		0.75	91.51 (2.54)	69.24 (3.70)	92.59 (3.62)	91.15 (3.70)	91.87 (4.28)
	75%		88.72 (3.26)	100.00 (n/a)	89.60 (4.24)	88.46 (4.85)	88.98 (4.99)
		0.50	88.38 (3.35)	96.73 (2.09)	89.22 (4.38)	88.13 (4.98)	88.65 (5.14)
		0.63	91.61 (3.20)	84.51 (3.91)	92.80 (4.24)	90.99 (4.60)	92.25 (4.90)
		0.75	93.58 (3.12)	71.23 (4.74)	95.03 (4.02)	92.71 (4.54)	94.50 (4.71)
	85%		89.35 (4.22)	100.00 (n/a)	90.75 (4.97)	88.83 (6.21)	89.90 (5.86)
		0.50	89.04 (4.33)	96.98 (2.39)	90.44 (5.10)	88.52 (6.36)	89.62 (6.01)
		0.63	92.00 (3.96)	85.32 (4.92)	93.52 (4.73)	91.28 (5.86)	92.78 (5.53)
		0.75	94.11 (3.72)	71.87 (6.19)	95.85 (4.33)	93.10 (5.67)	95.24 (5.10)
	95%		89.31 (7.06)	100.00 (n/a)	91.64 (7.97)	88.60 (10.57)	90.11 (9.87)
		0.50	89.00 (7.23)	97.08 (3.92)	91.33 (8.28)	88.27 (10.81)	89.87 (10.15)
		0.63	92.05 (6.83)	85.57 (8.10)	94.25 (7.74)	91.27 (10.07)	92.98 (9.61)
		0.75	94.25 (6.35)	73.12 (10.35)	96.49 (6.70)	93.10 (9.89)	95.54 (8.52)

Table 3: 2-Class Q5 classification results on the prostate cancer dataset. T%: Training Percent, PCT: Probability Classification Threshold, % Corr: Percent Correctly Classified, % Classif: Percent Classified, PPV: Positive Predictive Value, Sens: Sensitivity, Spec: Specificity. The first row of each training percent does not utilize probabilistic classification and is shown for comparison; in this row we forced Q5 to classify 100% of the samples into the class with the nearest class mean. Values listed are means in percent, standard deviations are in parentheses.

Dataset	T%	PCT	% Corr	% Classif	PPV	Sens	Spec
PC-IMAC-Cu	50%		91.11 (2.35)	100.00 (n/a)	92.19 (2.97)	90.54 (3.47)	91.84 (3.37)
		0.50	93.06 (2.32)	89.12 (3.63)	93.71 (2.98)	92.45 (3.45)	93.73 (3.20)
		0.63	95.84 (1.92)	75.76 (4.24)	95.96 (2.69)	95.59 (2.83)	96.08 (2.74)
		0.75	97.45 (1.75)	60.17 (4.52)	97.38 (2.52)	97.30 (2.56)	97.57 (2.40)
	75%		92.47 (2.77)	100.00 (n/a)	93.55 (3.39)	91.67 (4.08)	93.41 (3.67)
		0.50	94.33 (2.58)	90.33 (3.66)	94.98 (3.33)	93.57 (3.84)	95.15 (3.34)
		0.63	96.96 (2.14)	78.36 (4.54)	97.29 (2.80)	96.44 (3.29)	97.46 (2.66)
		0.75	98.30 (1.86)	63.84 (5.31)	98.41 (2.49)	98.00 (2.93)	98.58 (2.23)
	85%		92.80 (3.51)	100.00 (n/a)	94.12 (4.09)	91.95 (5.32)	93.84 (4.53)
		0.50	94.58 (3.22)	90.81 (4.48)	95.50 (3.93)	93.66 (5.11)	95.57 (3.98)
		0.63	97.16 (2.60)	79.22 (5.44)	97.56 (3.30)	96.66 (4.21)	97.67 (3.17)
		0.75	98.56 (1.99)	65.27 (6.23)	98.73 (2.68)	98.26 (3.26)	98.84 (2.47)
	95%		93.14 (5.75)	100.00 (n/a)	94.78 (6.65)	91.90 (9.08)	94.46 (7.19)
		0.50	94.64 (5.59)	92.02 (6.56)	96.08 (6.33)	93.23 (9.09)	96.05 (6.45)
		0.63	97.29 (4.31)	80.63 (9.08)	98.17 (4.83)	96.29 (7.35)	98.19 (4.79)
		0.75	98.72 (3.26)	67.13 (10.30)	99.21 (3.64)	98.08 (5.82)	99.27 (3.40)

Table 4: 3-class Q5 classification results on the prostate cancer dataset. Positive predictive value, sensitivity, and specificity are measured with PC as the ‘positive’ result and either NH or BPH as the ‘negative’ result. The first row of each training percent does not utilize probabilistic classification and is shown for comparison; in this row we forced Q5 to classify 100% of the samples into the class with the nearest class mean. A sample is considered correctly classified if it is assigned to the proper class, (NH, BD, PC). T%: Training Percent, PCT: Probability Classification Threshold, % Corr: Percent Correctly Classified, % Classif: Percent Classified, PPV: Positive Predictive Value, Sens: Sensitivity, Spec: Specificity. Values listed are means in percent, standard deviations are in parentheses.

A

	Classification		
Spectra Type	NH	BPH	PC
NH	99.9 (0.5)	0.0 (0.0)	0.1 (0.5)
BPH	0.1 (0.4)	91.1 (6.2)	8.9 (6.2)
PC	0.4 (0.8)	4.0 (2.7)	95.6 (2.8)

B

	Classification		
Spectra Type	NH	BPH	PC
NH	100.0 (0.0)	0.0 (0.0)	0.0 (0.0)
BPH	0.0 (0.0)	95.2 (13.2)	4.6 (12.5)
PC	0.1 (1.4)	3.6 (7.2)	96.3 (7.4)

Table 5: 3-class classification results for the PC-IMAC-Cu dataset. (A) 50% Training with a 0.63 probability classification threshold. (B) 95% Training with a 0.63 probability classification threshold. Average performance is reported with the standard deviation in parentheses. NH: Normal Healthy, BPH: Benign Prostatic Hypertrophy, PC: Prostate Cancer.

samples has only a small effect on overall performance. One reason for this is that LDA relies on the variance of the projected data, and for a significantly large enough training set, a small number of outliers will not have a particularly adverse effect.

The consistency of the computed discriminants for each dataset was examined. Each discriminant is back-projected (Sec. 2.3) from PCA-space into spectral-space. The dot-product between all pairs of discriminants was computed. The normalized discriminants for each experiment fall within a small region of the 15154- or 16382-dimension unit-hypersphere (results not shown). This represents an advantage of Q5 over non-deterministic methods in that the Q5 computed discriminants are similar for all D-experiments.

Comparison of Results to Other MSCAs

Petricoin et al. [28] report classification statistics for only one training/testing split of the OC-H4 dataset; no performance statistics have been published for the OC-WCX2a and OC-WCX2b datasets. This makes a comprehensive comparison of Q5 to previous work difficult. The reported performance of the Petricoin group MSCA (on the OC-H4 dataset) lies within one standard deviation of the mean performance statistics of Q5. Q5 classification results on the OC-WCX2a and OC-WCX2b datasets were near perfect. There are, however, no published performance results of MSCAs on these datasets to which the Q5 results can be directly compared.

Q5's prostate cancer classification results can be compared to those of Adam et al. [3], who performed a single testing/training split and achieved a sensitivity of 83%, specificity of 97%, and PPV of 96%. Q5 achieved a higher sensitivity and a similar specificity and PPV (Tables 3 and 4). Additionally in a 3-way experiment, the decision tree of Adam et al. reports 100% of the normal healthy samples as normal, 93% of the BPH samples as BPH, and 83% of the PC samples as PC. Q5's 3-way classification results are better: Q5 detects a higher percentage of prostate cancer samples among those it is able to classify (Table 5). Hence Q5 outperformed the decision-tree based prostate cancer MSCA [3].

For the datasets tested, Q5 is able to determine that spectral differences do indeed exist between the healthy and disease states. These spectral differences are a function of molecular differences existing between the two sets of samples. Q5's LDA-computed discriminant provides the m/z values of peaks differentially present between the two states. Thus Q5 is able to provide full answers to questions I) and IIa) posed in the introduction (Sec. 1). In summary, we found that Q5 classification performed at or above the level of existing MSCAs. It is worth noting that all existing MSCAs reviewed here, including Q5, outperform the currently used clinical CA125 and PSA tests. The future for MSCAs in analyzing human blood serum appears promising.

Back-Projection

All MSCAs assume that some m/z peaks are differentially observed between the healthy and disease classes. Identification of m/z peaks with large class-specific relative intensity differences can, in principle, allow for the identification of biomolecules affected by the disease process. Most heuristic MSCAs base classification on a small number of m/z peaks. For example, Petricoin et al. [28] use 5-20 m/z peaks and Adam et al. [3] use 9 m/z peaks. Thus information on class-specific relative intensity differences for most m/z peaks is not available. An advantage of LDA is that the spectral-space discriminant can be used to compute a classification significance for *all* m/z values (Sec. 2.3, Eq. 2). Below, we show how to query a protein database using the discriminant peaks. The discriminant can also serve as supporting evidence for biomarkers discovered via other experimental techniques: the SELDI mass spectrum of a hypothesized serum biomarker can be checked for consistency with a discriminant.

To test the power of the back-projected discriminant for determining the identities of differently-expressed proteins and peptides, we took the largest m/z peaks from the normalized discriminant (the significance vector), interpreted them as masses, and looked up those masses in two protein databases (Table **??**). While the lookup is likely to yield some false positives due to mass-aliasing, the database lookup for the ovarian (resp. prostate) queries found 27 (resp. 39) proteins and peptide fragments that have been implicated in other human cancers, are growth factors, or are known serum or plasma proteins. While some of the SwissProt and TrEMBL entries are annotated with known or hypothesized function, many of the entries (particularly in the TrEMBL database) are of unknown function [6]. While these ‘lead’ proteins have masses consistent with the most significant discriminant peaks, we caution that the database lookup does not prove that these proteins are present in the serum. These ‘lead’ proteins can serve as the starting point for previously-described biomarker identification protocols. Perhaps the most interesting protein identified among those with known function for the OC-WCX2b query is TrEMBL entry[§] Q9BZK8, a 76AA protein of OCR1 (ovarian cancer-related protein 1). Other interesting results include: Q9NPJ2, a 36AA protein fragment of P53/TP52 (cellular tumor antigen); Q9NP09, a 36AA protein fragment of ERBB2 (polymorphism of the HER-2/neu oncogene); Q13262, a 44AA estrogen receptor fragment; Q9UH52/Q96B49, a 74AA protein fragment of OBTP (over-expressed in some breast tumors); and SwissProt entry PS2_HUMAN, an 84AA protein from TFF1/BCEI/PS2 (a breast cancer associated estrogen-inducible protein). The PC-IMAC-Cu search identified two known prostate cancer associated proteins: TrEMBL entry Q9GZR0, a 50AA protein fragment of SCN8A (voltage-gated sodium channel involved with metastatic human prostate cancer) and Q96P91, a 55AA protein fragment of PON1 (paraoxonase 1, associated with prostate cancer risk). Additional interesting results include: Q9NPJ2, a 36AA protein fragment of P53/TP52 (cellular tumor antigen) and Q12847, a 45AA protein fragment of TAP1 (tumor associated protein). Approximately 90 of the genes found in the ovarian cancer search and 70 of the genes found in the prostate cancer search have novel or unknown function. This raises the possibility that these genes may have a role or additional roles in ovarian or prostate cancer. The normalized discriminants are shown in Figure 6; the m/z peaks consistent with the masses of the described proteins are indicated.

Other methods for protein identification via mass spectrometry have been developed. Three main alternative techniques exist. In the first, the mass spectrum of a proteolytic digest determines a set of protein fragment masses that can be matched to a database [19]. In the second approach, a peptide's sequence is directly identified by tandem mass spectrometry (MS/MS) [9, 29]. In a third approach, tandem mass spectrometry may be applied to a proteolytic digest of the target proteins after these fragments have been separated via chromatography (e.g., liquid chromatography (LC)

[§] In this section, TrEMBL entries begin with 'Q' and have six characters.

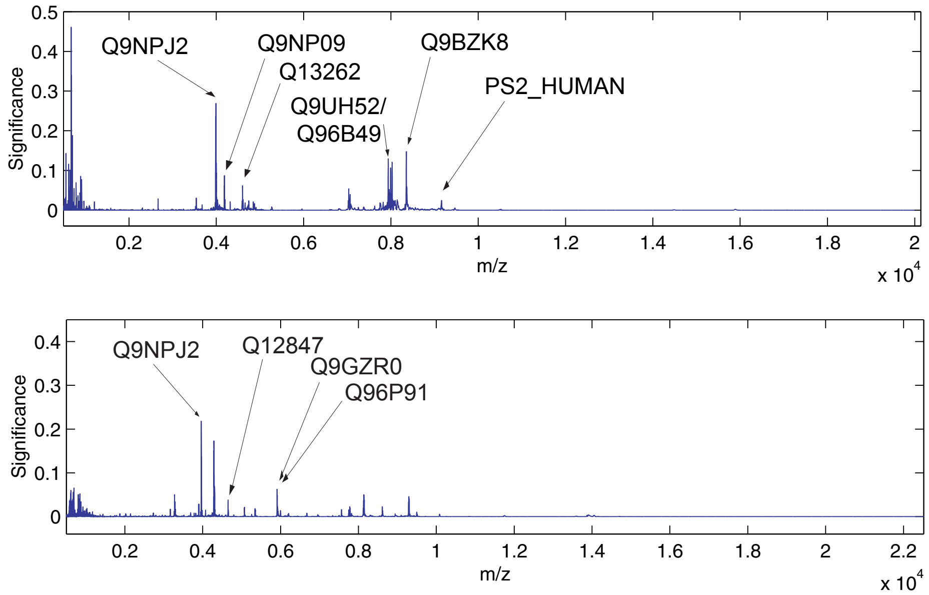

Figure 6: The normalized discriminant for ovarian dataset OC-WCX2b (top) and prostate dataset PC-IMAC-Cu (bottom) starting at an m/z value of 500. The location of the SwissProt and TrEMBL proteins noted in the text are indicated by their identification numbers for each discriminant (see text). These SwissProt and TrEMBL proteins are consistent with m/z peaks of the discriminant having significance for classification. Due to mass-aliasing, the database lookup does not prove these proteins present in the serum samples, but these proteins can serve as leads in the search for novel biomarkers.

MS) [22, 27]. In the last two approaches, one may use a database for sequence identification, or the peptide sequencing may be done *de novo*. Currently available whole-serum clinical cancer MS data is neither from a controlled proteolytic digest, nor from tandem mass spectrometry. Hence, we used a different approach to help identify the molecules most important in discrimination. It would be interesting, in future work, to extend Q5 to take advantage of the additional information experimentally available from controlled proteolytic digests or MS/MS.

Our work represents the first attempt to compute the molecular identities of the differentially-expressed proteins in datasets OC-WCX2b and PC-IMAC-Cu. Further investigation of our lead proteins and peptide fragments may enhance our understanding of the molecular basis of oncogenisis and could potentially lead to new therapeutic targets.

4 Conclusion

Mass spectrometry will soon play an important role in both the research lab and hospital clinic. For all but the simplest cases, manual analysis of complete complex spectra is impractical. This observation led to the development of a variety of MSCAs. Of the previous MSCAs, those that use complete complex spectra [28, 3], do not use exact algorithms. Conversely, those MSCAs that are exact [24, 32], do not operate on complete complex spectra. In contrast to previous work, Q5 uses PCA and LDA followed by probabilistic classification on complete complex SELDI-TOF mass spectra for the classification of healthy vs. disease serum samples. The use of a probabilistic classification framework increases the predictive accuracy of Q5. A tradeoff is shown between confidence in classification and the number of samples classified. Our solution is computationally efficient; it is non-iterative and computes the optimal linear discriminant using closed-form equations. Q5 thus represents a generally applicable technique. Although Q5 was tested against ovarian and prostate cancer it is reasonable to hypothesize that Q5 may be effective in the screening of other cancers and diseases. Q5 was tested against 2 cancer types and 4 datasets. Our results show that a classification threshold can be chosen for Q5 such that over 90% of the samples are classified with a sensitivity, specificity, and PPV near 100%. Q5 performed at or above the level of previous techniques while conferring all advantages of a closed-form exact solution. The consistently high level of performance on the testing spectra demonstrates that Q5 was not over-fit to the training spectra. We note that Q5's time complexity grows only linearly with the resolution of the mass spectra. Thus Q5 will scale well as higher-resolution spectra are collected.

Another advantage of Q5 is that the discriminant can be examined both to identify and to support the validity of novel biomarkers. Whereas previous complete complex spectra MSCAs discriminate using a small fraction of the total number of m/z peaks, Q5 computes all peaks that are differentially-expressed in one class vs. the other. We showed how Q5's discriminant back-projection technique can compute clues as to the molecular identities of differentially-expressed proteins and peptides.

Finally, we note that for MSCAs to be practical in a clinical setting, questions of reproducibility must be addressed. Ideally, a discriminant computed from one spectrometer should generalize to classify spectra collected on a different spectrometer, in a different laboratory. We have not addressed such reproducibility questions, which will be important for future work.

5 Supporting Material

The MATLAB code for Q5 is available at `http://www.cs.dartmouth.edu/~brd/Bio` and by contacting the authors. The software is distributed under the Gnu Public License [12].

6 Acknowledgments

We thank Dr. Emanuel F Petricoin III for providing access to his SELDI-TOF datasets. We thank Drs. A. Anderson, C. Bailey-Kellogg, and T. Lozano-Perez, Mr. C. Langmead, Mr. S. Lyu, Mr. A. Yan, Ms. E. Werner-Reiss, Dr. R. Mettu and all members of the Donald Lab for helpful discussions and comments on drafts.

References

[1] Eastern Virginia Medical School - The Virginia Prostate Center. http://www.evms.edu/vpc/seldi/index.html, 2002.

[2] NIH and FDA Clinical Proteomics Program Databank. http://clinicalproteomics.steem.com, 2002.

[3] BL. Adam, Y. Qu, J. Davis, M. Ward, M. Clements, L. Cazares, OJ. Semmes, P. Schellhammer, Y. Yasui, Z. Feng, and G. Wright Jr. 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.

[4] J.P. Antignac, B. Le Bizec, F. Monteau, and F. Andre. Differentiation of betamethasone and dexamethasone using liquid chromatography/positive elecrtospray tandem mass spectrometry and multivariate statistical analysis. *Journal of Mass Spectrometry*, 37:69–75, 2002.

[5] B. Austen, E. Frears, and H. Davies. The use of SELDI proteinchip arrays to monitor production of alzheimer's β-amyloid in transfected cells. *Journal of Peptide Science*, 6:459–469, 2000.

[6] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28:45–48, 2000.

[7] G. Ball, S. Mian, F. Holding, R. Allibone, J. Lowe, S. Ali, G. Li, S. McCardie, I. Ellis, C. Creaser, and R. Rees. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics*, 18:395–404, 2002.

[8] W. Catalona, J. Richie, F. Ahmann, M. Hudson, P. Scardino, R. Flanigan, J. deKernion, T. Ratliff, L. Kavoussi, B. Dalkin, W. Waters, M. MacFarlane, and P. Southwick. Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: Results of a multicenter clinical trial of 6,630 men. *Journal of Urology*, 151:1283–1290, 1994.

[9] V. Dancik, T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6:327–342, 1999.

[10] R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* Wiley, New York, NY, 1973.

[11] R. Fisher. The use of multiple measures in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[12] Gnu. The GNU general public license. http://www.gnu.org/licenses/licenses.html, 2002.

[13] R. Goodacre, J. Heald, and D. Kell. Characterisation of intact microorganisms using electrospray ionization mass spectrometry. *FEMS Microbiology Letters*, 176:17–24, 1999.

[14] R. Goodacre, É. Timmins, R. Burton, N. Kaderbhai, A. Woodward, D. Kell, and P. Rooney. Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. *Microbiology*, 144:1157–1170, 1998.

[15] T. Griffin, D. Goodlett, and R. Aebersold. Advances in proteome analysis by mass spectrometry. *Current Opinion in Biotechnology*, 12:607–612, 2001.

[16] T. Griffin, S. Gygi, T. Ideker, B. Rist, J. Eng, L. Hood, and R. Aebersold. Complementary profiling of gene expression at the transcriptome and proteome levels in saccharomyces cerevisiae. *Molecular & Cellular Proteomics*, 1:323–333, 2002.

[17] T. Hutchens and T. Yip. New desorption strategies for the mass spectrometric analysis of macromolecules. *Rapid Communications in Mass Spectrometry*, 7:576–580, 1993.

[18] I. Jacobs, S. Skates, N. MacDonald, U. Menon, A. Rosenthal, A. Davies, R. Woolas, A. Jeyarajah, K. Sibley, and D. Lowe and. Screening for ovarian cancer: A pilot randomised controlled trial. *The Lancet*, 353:1207–1210, 1999.

[19] P. James, M. Quadroni, E. Carafoli, and G. Gonnet. Protein identification by mass profile fingerprinting. *Biochemical and Biophysical Research Communications*, 195:58–64, 1993.

[20] M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10000 daltons. *Analytical Chemistry*, 60:2299–2301, 1988.

[21] J. Li, Z. Zhang, J. Rosenzweig, Y. Wang, and D. Chan. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry*, 48:1296–1304, 2002.

[22] A. Link, J. Eng, D. Schieltz, E. Carmack, G. Mize, D. Morris, B. Garvik, and J. Yates III. Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, 17:676–682, 1999.

[23] M. Merchant and S. Weinberger. Recent advancements in surface enhanced laser desorption/ionization time-of-flight mass spectrometry. *Electrophoresis*, 21:1164–1177, 2000.

[24] P. Miketova, C. Abbas-Hawka, K. Voorhees, and T. Hadfield. Microorganism Gram-type differentiation of whole cells based on pyrolysis high-resolution mass spectrometry data. *Journal of Analytical and Applied Pyrolysis*, (In press).

[25] C. Paweletz, J. Gillespie, D. Ornstein, N. Simone, M. Brown, K. Cole, Q.H. Wang, J. Huang, N. Hu, T.T. Yip, W. Rich, E. Kohn, W.M. Linehan, T. Weber, P. Taylor, M. Emmert-Buck, L. Liotta, and E. Petricoin III. Rapid protein display profiling of cancer progression directly from human tissue using a protein biochip. *Drug Development Research*, 49:34–42, 2000.

[26] C. Paweletz, B. Trock, M. Pennanen, T. Tsangaris, C. Magnant, L. Liotta, and E. Petricoin III. Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: Potential for new biomarkers to aid in the diagnosis of breast cancer. *Disease Markers*, 17:301–307, 2001.

[27] D. Perkins, D. Pappin, D. Creasy, and J. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551–3567, 1999.

[28] E. Petricoin III, A. Ardekani, B. Hitt, P. Levine, V. Fusaro, S. Steinberg, G. Mills, C. Simone, D. Fishman, E. Kohn, and L. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359:572–577, 2002.

[29] P. A. Pevzner, V. Dancik, and C. L. Tang. Mutation-tolerant protein identification by mass spectrometry. *Journal of Computational Biology*, 7:777–787, 2000.

[30] Seer cancer statistics review, 1973-1999. http://seer.cancer.gov/csr/1973_1999, 2002.

[31] V. Vapnik. *Statistical learning theory.* John Wiley and Sons, 1998.

[32] M. Wagner, B. Tyler, and D. Castner. Interpretation of static time-of-flight secondary ion mass spectra of adsorbed protein films by multivariate pattern recognition. *Analytical Chemistry*, 74:1824–1835, 2002.

[33] Z. Zhang, G. Page, and H. Zhang. Applying classification separability analysis to microarray data. In *Methods of Microarray Data Analysis: Papers from CAMDA '00*. Kluwer Academic Publishers, 2001.

[34] H. Zhou, J. Ranish, J. Watts, and R. Aebersold. Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. *Nature Biotechnology*, 20:512–515, 2002.

Appendix

A Principal Components Analysis

Each sample spectrum in the training set is represented as a column vector $\mathbf{x}$ $(\mathbf{x} \in X, |X| = n_x)$ (healthy) or $\mathbf{y}$ $(\mathbf{y} \in Y, |Y| = n_y)$ (disease). Here $|\cdot|$ is the number of elements in the specified set. Thus, n_x (resp. n_y) is the number of healthy (resp. disease) samples. Let $n = n_x + n_y$ be the total number of training samples; we assume all $\mathbf{x}$ and $\mathbf{y}$ vectors have dimensionality r (i.e., each mass spectrum is sampled at r points). The *all-class mean*,

$$\boldsymbol{\mu}' = \frac{1}{n}\left(\sum_{\mathbf{x}\in X}\mathbf{x} + \sum_{\mathbf{y}\in Y}\mathbf{y}\right), \tag{3}$$

is computed and subtracted from each sample, producing sets of zero-meaned samples X' and Y'. The columns of the $r \times n$ matrix $\mathbf{P}$ consist of all zero-meaned samples,

$$\mathbf{P} = \begin{bmatrix} X' & Y' \end{bmatrix}. \tag{4}$$

The $r \times r$ covariance matrix $\mathbf{C}$ can then be computed:

$$\mathbf{C} = \mathbf{P}\mathbf{P}^T. \tag{5}$$

The principal components are the eigenvectors, $\mathbf{v}_i$, of the covariance matrix $\mathbf{C}$. An eigendecomposition of $\mathbf{C}$ produces at most $w = \min(n, r)$ non-zero eigenvalues λ_i $(i = 1, ..., w)$ with corresponding normalized eigenvectors $\mathbf{v}_i$ such that $\mathbf{C}\mathbf{v}_i = \lambda_i\mathbf{v}_i$. Each eigenvalue λ_i is proportional to the variance of the original data in the direction of the i^{th} principal component. Frequently, the number (r) of points in each sample spectrum greatly exceeds the number (n) of samples. In this case, an alternate formulation of the covariance matrix may be preferable. When $r > n$, increased computational efficiency can be achieved by computing the eigenvectors of the $n \times n$ *Gram matrix* $\mathbf{C}'$,

$$\mathbf{C}' = \mathbf{P}^T\mathbf{P}. \tag{6}$$

The eigenvalues of $\mathbf{C}'$ equal the eigenvalues of $\mathbf{C}$, and the normalized eigenvectors of $\mathbf{C}'$ $(\mathbf{v}'_i$, $i = 1, ..., w)$ can be related to the normalized eigenvectors of $\mathbf{C}$ by

$$\mathbf{v}_i = \mathbf{P}\mathbf{v}'_i. \tag{7}$$

The largest principal components typically account for nearly all sample variance. Therefore, dimensionality reduction with PCA can be accomplished by sorting components by eigenvalue and then discarding the eigenvectors with the smallest corresponding eigenvalues. After discarding the eigenvectors, the w' eigenvectors that remain constitute the PCA basis. Once a set of principal components is selected as a basis, sample points can be projected onto these axes,

$$\begin{aligned} \mathbf{x}_p &= \mathbf{V}\mathbf{x}' \qquad \left(\mathbf{x}' \in X'\right), \\ \mathbf{y}_p &= \mathbf{V}\mathbf{y}' \qquad \left(\mathbf{y}' \in Y'\right), \end{aligned} \tag{8}$$

where the rows of matrix $\mathbf{V}$ are the retained eigenvectors $\mathbf{v}_i$ $(i = 1, ..., w')$, and $\mathbf{x}_p$ (resp. $\mathbf{y}_p$) are the PCA-space projections of each healthy (resp. disease) sample onto the w' principal components.

B Linear Discriminant Analysis

For simplicity we present a two-class LDA. Two disjoint sets of points and the LDA-computed discriminant (dotted line) are shown in Figure 2A. Projecting sample points onto the linear discriminant allows for point classification (Figure 2C). Higher-order LDAs can be employed to differentiate more than two classes: the generalization to k classes ($k > 2$) is straightforward [10].

After PCA-based dimensionality reduction, let column vectors $\mathbf{x}_p$ ($\mathbf{x}_p \in X_p, |X_p| = n_x$) (healthy) and $\mathbf{y}_p$ ($\mathbf{y}_p \in Y_p, |Y_p| = n_y$) (disease), of dimension w', be the training sample spectra from each of the two classes. The *within-class means* $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ are defined as:

$$\boldsymbol{\mu}_x = \frac{1}{n_x} \sum_{\mathbf{x}_p \in X_p} \mathbf{x}_p, \quad \text{and} \quad \boldsymbol{\mu}_y = \frac{1}{n_y} \sum_{\mathbf{y}_p \in Y_p} \mathbf{y}_p. \tag{9}$$

The *all-class mean* $\boldsymbol{\mu}$ is computed from X_p and Y_p similarly to Eq. (3). The *within-class scatter matrix* $\mathbf{S}_w$ is defined as:

$$\mathbf{S}_w \quad = \quad \mathbf{M}_x \mathbf{M}_x^T + \mathbf{M}_y \mathbf{M}_y^T, \tag{10}$$

where the columns of matrix $\mathbf{M}_x$ contain the zero-meaned PCA-space representation of healthy spectra $\mathbf{x}_p - \boldsymbol{\mu}_x$. Similarly, the columns of matrix $\mathbf{M}_y$ contain $\mathbf{y}_p - \boldsymbol{\mu}_y$. The *between-class scatter matrix* is defined as:

$$\mathbf{S}_b = n_x(\boldsymbol{\mu}_x - \boldsymbol{\mu})(\boldsymbol{\mu}_x - \boldsymbol{\mu})^T + n_y(\boldsymbol{\mu}_y - \boldsymbol{\mu})(\boldsymbol{\mu}_y - \boldsymbol{\mu})^T. \tag{11}$$

A *generalized eigenvector* $\mathbf{v}$, of $\mathbf{S}_b$ and $\mathbf{S}_w$ satisfies the equation $\mathbf{S}_b\mathbf{v} = \lambda \mathbf{S}_w \mathbf{v}$, (where λ is the eigenvalue). Let $\mathbf{e}$ be the maximal generalized eigenvalue-eigenvector of $\mathbf{S}_b$ and $\mathbf{S}_w$. The vector $\mathbf{e}$ is the optimal linear discriminant. The matrices $\mathbf{S}_w$ and $\mathbf{S}_b$ have size $w' \times w'$ and rank of at most $n - k$ (k is the number of classes). Therefore w' must be less than $n - k$ to avoid a guaranteed singularity in the eigendecomposition. This is the mathematical reason why sample points must be dimensionality-reduced to dimension less than $n - k$ before LDA is performed.

Once the linear discriminant is determined from the training set, the PCA-space vectors ($\mathbf{x}_p$, $\mathbf{y}_p$) are projected onto the linear discriminant to produce the discriminant-space representation of each spectrum (x_d, y_d). Note that for a two-class LDA, discriminant-space is one-dimensional, thus x_d is a scalar and is not typeset in boldface. Points in the discriminant-space of a k-class LDA ($k > 2$) are vectors and are thus typeset in boldface. Hence,

$$\begin{aligned} x_d &= (\mathbf{x}_p)^T \mathbf{e} \qquad (\mathbf{x}_p \in X_p) \\ y_d &= (\mathbf{y}_p)^T \mathbf{e} \qquad (\mathbf{y}_p \in Y_p) \end{aligned} \tag{12}$$

The points x_d should form one cluster while the points y_d should, ideally, form a separate non-overlapping cluster. Healthy and disease class means can be then be computed and used for classification. The spectral-space representation of a novel spectrum, $\mathbf{z}$, from the testing set can now be classified by computing z_d, the projection of $\mathbf{z}$ into the subspace spanned by the linear discriminant,

$$z_d = \left(\mathbf{V}\left(\mathbf{z} - \boldsymbol{\mu}'\right)\right)^T \mathbf{e}, \tag{13}$$

where $\mathbf{V}$ is the eigenvector matrix defined in Eq. (8). In Eq. (13), the PCA projection of the zero-meaned spectral-space representation $\mathbf{z}$ is projected onto the linear discriminant $\mathbf{e}$. The spectrum represented by $\mathbf{z}$ can then be classified based on proximity to the healthy and disease class means.

C Probabilistic Classification

Two-Class Probability. After projection onto the linear discriminant, let p_1 (resp. p_2) be the mean of class C_1 (resp. class C_2). Let z_d be the discriminant-space projection (Eq. 13) of a novel sample spectrum and let q be the midpoint between p_1 and p_2. Assume, without loss of generality, that z_d is closer to p_1 than p_2. That is, $d(z_d, p_1) < d(z_d, p_2)$ where $d(\cdot,\cdot)$ is the Euclidean distance. We define the probability that z_d belongs to C_1 as:

$$P(z_d \in C_1) = \exp\left[-\left(d(z_d, p_1)\right)^2 / \sigma^2\right], \tag{14}$$

where σ (the standard deviation of the Gaussian probability function) is chosen such that $P(q \in C_1) = 0.5$. Eq. (14) specifies a symmetric Gaussian probability density function centered at p_1 where the midpoint between p_1 and p_2 has a 50% probability of being classified into either C_1 or C_2.

The classification threshold $t \in [0.5, 1.0]$ is chosen such that a classification is made if and only if:

$$P(z_d \in C_i) > t \qquad (i = 1, 2). \tag{15}$$

If Eq. (15) is not satisfied then we consider z_d to be *ambiguous* and a classification is not made. A tradeoff is thus offered between the number of spectra classified and the accuracy of classification. Smaller values of t allow more samples to be classified at the cost of lower confidence in classification. Similarly, larger values of t classify fewer samples but with higher confidence.

k-Class Probability. In a k-class model ($k > 2$) we have classes C_i ($i = 1, ..., k$) and associated class means $\mathbf{p}_i$, where $\mathbf{p}_i$ is now a $(k-1)$-dimensional vector. The variance computed for the Gaussian probability density function of each class in the k-class model is not guaranteed to be the same for each class. That is, a different σ_i (the standard deviation) is defined for each class. Intuitively, the σ_i computed for class C_i in the k-class classifier is the smallest variance (σ) that would be computed if one were to compute a 2-class classifier (as described above) between class C_i and every other class C_j ($j \neq i$). To compute σ_i we first define a set of midpoints. Let $\mathbf{q}_{ij}$ be the midpoint between $\mathbf{p}_i$ and $\mathbf{p}_j$ and let $\mathbf{q}'_i$ be the midpoint closest to $\mathbf{p}_i$,

$$\mathbf{q}'_i = \operatorname*{argmin}_{\mathbf{q} \in Q_i} d(\mathbf{p}_i, \mathbf{q}), \tag{16}$$

where $Q_i = \{\mathbf{q}_{ij} | j = 1, ..., k; j \neq i\}$. Using $\mathbf{q}'_i$ we can now compute the σ_i such that the midpoint between two class means will have a probability of classification of 50%. σ_i satisfies the following equation:

$$\exp\left[-\left(d(\mathbf{p}_i, \mathbf{q}'_i)\right)^2 / \sigma_i^2\right] = 0.5. \tag{17}$$

The probability that a discriminant-space point $\mathbf{z}_d$ (where $\mathbf{z}_d$ has dimension $k-1$) belongs to class C_i is:

$$P(\mathbf{z}_d \in C_i) = \exp\left[-\left(d(\mathbf{z}_d, \mathbf{p}_i)\right)^2 / \sigma_i^2\right]. \tag{18}$$

As in the 2-class probabilistic framework, a classification threshold $t \in [0.5, 1.0]$ is specified such that a classification is made if and only if there exists an i such that $P(\mathbf{z}_d \in C_i) > t$. If this criterion is not satisfied we consider $\mathbf{z}_d$ to be ambiguous, and a classification is not made. By construction, it is not possible for a point to be classified into more than one class when the classification threshold t is chosen in the range $[0.5, 1.0]$. Note that the variances and classification probabilities computed for two classes using either the two-class model or the k-class model (with $k = 2$) are identical.