Biomarker Signature Discovery from Mass Spectrometry Data

Ao Kong, Chinmaya Gupta, Mauro Ferrari, Marco Agostini, Chiara Bedini, Ali Bouamrani, Ennio Tasciotti, and Robert Azencott

Abstract—Mass spectrometry based high throughput proteomics are used for protein analysis and clinical diagnosis. Many machine learning methods have been used to construct classifiers based on mass spectrometry data, for discrimination between cancer stages. However, the classifiers generated by machine learning such as SVM techniques typically lack biological interpretability. We present an innovative technique for automated discovery of signatures optimized to characterize various cancer stages. We validate our signature discovery algorithm on one new colorectal cancer MALDI-TOF data set, and two well-known ovarian cancer SELDI-TOF data sets. In all of these cases, our signature based classifiers performed either better or at least as well as four benchmark machine learning algorithms including SVM and KNN. Moreover, our optimized signatures automatically select smaller sets of key biomarkers than the black-boxes generated by machine learning, and are much easier to interpret.

Index Terms—MALDI/SELDI data, ovarian cancer, colorectal cancer, biomarker selection, automatic signature discovery

1 INTRODUCTION

Early detection and risk assessment of cancer are crucial for successful intervention strategies. Identification of validated biomarkers, as a non-invasive screening method, is recognized as a major breakthrough in cancer detection [1]. Human body fluids such as serum or plasma are easily accessible sources of biomarker discovery, due to their wide range of molecular components linked to cellular metabolism by-products. In proteome analysis, proteins with low molecular weight (LMW) are of great interest, because cellular and extracellular enzymatic events generate small fragments, playing an important role at the level of cancer-tissue micro-environment [2]. Mass spectrometry techniques are considered to be a promising approach for LMW biomarker discovery. Laser desorption ionization, either surface-enhanced (SELDI) or matrix-assisted (MALDI), are the two commonly used techniques to generate mass spectrometry data set. Each such mass spectrum provides thousands of mass-to-charge (m/z) ratios paired with corresponding peptide intensities. Experimental design for biomarker discovery requires adequate statistical sampling, appropriate handling and storage of serum or plasma samples until analysis, mass spectrometry processing of the samples, and detailed data analysis of the generated mass spectra. Our paper focuses only on this last step.

To handle the high dimensionality and inherent variability of biomedical mass spectra, “machine learning” algorithms have been applied to automatic discrimination between mass spectra. Unsupervised learning, such as clustering [3] or self-organizing maps [4], has been used to partition mass spectrometry data sets into homogeneous subgroups. Supervised learning, which analyzes pre-classified mass spectrometry data sets to generate automated classifiers, has been applied for cancer discrimination: artificial neural networks [5], K-nearest neighbors (KNN) [6], decision trees (DT) [7], [8], [9], [10], random forests (RF) [11], linear and quadratic discriminant analysis [12], support vector machines (SVM) [13], [14], [15]. Machine learning techniques have often yielded good classification accuracy, but they typically generate “black-box” classifiers, which are not easily interpreted biologically. To develop more pragmatic mass spectrometry classifiers, a key step is to automatically discover “signatures”, i.e., combinations of a small number of protein biomarkers strongly discriminating between cancer stages [16], [15], [17], [18].

We have developed new algorithms for automatic discovery of biologically interpretable signatures discriminating between various cancer states, by automated analysis of mass spectrometry data sets acquired from multi-stages cancer patients groups. We applied simulated annealing optimization techniques [19], [20], [21] to maximize discriminating power among all possible signatures.

We have tested our signature discovery algorithm on a new MALDI-TOF data set for colorectal cancer and two well-known ovarian cancer SELDI-TOF data sets. We have generated explicit signatures with high discriminating power between the various cancer patients groups involved in these data sets. We have compared performances between our optimized signature based classifiers and several benchmark machine learning techniques.

2 DATA SETS

2.1 Colorectal MALDI-TOF Data Set

One hundred and four colorectal cancer samples and 15 control samples were provided by first Surgical Clinic, Department of Surgical, Oncological and Gastroenterological Sciences, University of Padova, Italy. Between 2002 and 2005, the 104 cancer patients underwent surgeries and histopathological diagnosis. Among them, 27 were diagnosed with colorectal pre-cancer lesion (Ade-noma), 40 with Early Colorectal cancer (stage I or II), and 37 with Late Colorectal cancer (stage III or IV). The 15 healthy patients all received colonoscopy and were diagnosed to be unaffected.

A 10 ml blood sample was collected from each patient into a DB Vacutainer during the surgery or colonoscopy and transferred to the laboratory within 4 hours of collection, to be centrifuged at 3,000 rpm for 10 min. Plasma samples were then collected from the supernatant and stored in aliquots at −80°C in the Tumor Tissue Biobank of Surgical Clinic I as well as during transportation, until analysis.

For efficient removal of high molecular weight proteins and for specific isolation and enrichment of LMW species present in 15μl of plasma, we used a novel three steps size-exclusion strategy based on Mesoporous silica chips, fabricated by Dept of Nanomedicine (Methodist Hospital Research Institute, Houston, Texas) [22]. Mass spectra were acquired in linear positive-ion mode (range 800-10,000 “m/z” ratio) on a Voyager-DE-STR MALDI TOF Mass Spectrometer (Applied Biosystems, Framingham, MA) at Research Center of Protein Chemistry Core Laboratory (University of Texas Health, Houston, Texas). The manufacturer provided spectrometer accuracy was 0.3 percent. Only one blood plasma sample was extracted from each subject, but two “replicate” mass spectra were acquired from each blood plasma sample.

In total, 238 mass spectrometry replicates were acquired from four patients groups, with two mass spectrometry replicates per patient: the Control group CTR of 15 patients, the Adenoma group ADE of 27 patients with precancer lesions, the group ECR of 40
METHODS

2.2 Ovarian SELDI-TOF Data Sets

We have also tested our approach on two well-known mass spectrometry data sets of ovarian cancer, which can be freely downloaded from NCI-FDA clinical proteomics databank (http://home.ccr.cancer.gov/ncifdaproteomics/patterns.asp).

The data set “4-3-02” consists of 116 control (normal or benign patients) mass spectra and 100 ovarian cancer mass spectra, acquired manually, using a WCX2 protein chip and a Ciphergen PBS1 SELDI-TOF mass spectrometer with 0.1 percent accuracy. Each mass spectrum listed 15,154 m/z values in the 0-20,000 range. Baselines were removed prior to public access. This data set was studied by [4], [6], [23] who respectively reported discrimination accuracies of 97.5, 100 and 86.66 percent.

The data set “8-7-02” gathers 91 control and 162 ovarian cancer mass spectra, acquired by robotic hardware using a WCX2 chip but with an upgraded PBSII SELDI-TOF mass spectrometer with 0.1 percent accuracy. Each spectrum gave 15,154 m/z values in the 0-20,000 range. This data set was studied by [6], [24], [25] and [23], with reported discrimination accuracies of 100 percent.

3 METHODS

3.1 Pre-Processing

To lower data dimensionality and reduce acquisition “noise”, pre-processing of mass spectra is a standard first step, often implemented via commercial interactive softwares. Pre-processing principles are well known, but implementation details vary considerably, and are often not accessible in commercial softwares. For better context control, we have developed our own sequence of pipelined pre-processing steps for each raw mass spectrum.

1) Intensities are rescaled (normalized) so that total peak intensity equals 1.
2) Smoothing at current abscissa x is then implemented by moving average on a sliding window x ± u, where the fixed ratio u is user-defined.
3) Noise values are extracted by subtracting smoothed spectrum from raw spectrum. Local noise level at x is evaluated on a sliding window x ± v where the ratio v is user-defined.
4) Baseline is computed by moving medians within sliding windows x ± w where ratio w is user-defined. Baseline is then removed from smoothed spectrum. Using medians instead of means avoids discarding small peaks in the vicinity of large peaks.
5) Peaks above baseline are then extracted from the smoothed spectrum to retain only the “strong peaks”, for which “peak strength” (ratio of peak height to local noise level) is higher than a threshold thr fixed by the user.

3.2 Reference Biomarkers and Activation Frequencies

For a MALDI or SELDI spectrometer with manufacturer accuracy ρ, any peptide with “true” m/z ratio x will yield randomly shifted acquisition values within the “uncertainty window” x ± ρx. In view of these acquisition uncertainties, we fix a sequence of “reference biomarkers” Bjk at abscissas Bj = a(1 + jθ), with 0 ≤ j ≤ n where B0 = a and Bn = b are the smallest and largest observed m/z values among all spectra in our training data set.

We say that a reference biomarker B is “activated” by a mass spectrum M if at least one strong peak detected in M is positioned within the uncertainty window of B. Otherwise, we say that “B is not activated by M”.

Given two distinct groups G+ and G− of observed mass spectra, we characterize each reference biomarker B by its “activation frequencies” f+B and f−B, defined as percentages of mass spectra which respectively activate B within G+ and within G−.

3.3 Selection of Biomarkers with High Discriminating Power

One can prove that the most powerful test to discriminate between G+ and G− on the basis of the presence or absence of a single biomarker B is achieved by adequate thresholding for the ratio of activation frequencies f+B/f−B. This suggests quantifying the G+ detecting power of a biomarker B by D+B = f+B/f−B and its G− detecting power by the inverse ratio D−B = f−B/f+B. We say that B is a G+ biomarker if D+B > 1 and a G− biomarker if D−B > 1.

To seek optimal signatures combining small numbers of biomarkers, we first select a tentative target pool TP(2k) of 2k highly discriminating biomarkers, namely the k biomarkers with highest D+B and the k biomarkers with highest D−B.

Good generalization capacity requires k to be small while high discriminating power forces k to be large. So we successively implement our signature discovery algorithm for increasing values of k, and later select an optimal k.

3.4 Automatic Signature Discovery

Given a training data set involving two pre-classified groups of mass spectra G+ and G−, we seek to generate various automated classifiers of arbitrary mass spectra M and compare their performances. We systematically evaluate the performance level of any classifier by PERF(c, Sig) = (p+c + p−c)/2 where p+c and p−c are the frequencies of correct classification for the spectra belonging respectively to G+ and G−.

A signature Sig will be any fixed list of r biomarkers picked within the target pool TP(2k) just defined. Given any spectrum M, we count, within the signature Sig, the number a(M) of G+ biomarkers activated by M and the number v(M) of G− biomarkers which are NOT activated by M. The “G+ score” M for the signature Sig is then defined by s+(M, Sig) = a(M) + v(M))/r. Clearly, M will achieve a high G+ score if most of the G+ biomarkers belonging to signature Sig are actually present in M, and if simultaneously most of the G− biomarkers belonging to Sig are absent in M. The signature Sig will have high G+ detecting power if s+(M, Sig) is high for M in G+ and low for M in G−.

Exchanging the roles of G+ and G− symmetrically defines the “G− score” s−(M, Sig) of M for the signature Sig. Then Sig will have high G− detecting power if s−(M, Sig) is high for M in G− and low for M in G+. When one fixes a scoring threshold 0 < c < 1, the signature Sig determines a G+ classifier by assigning any observed spectrum M to G+ if s+(M, Sig) ≥ c and to G− otherwise. The performance level PERF(c, Sig) of this classifier is defined as above by its average frequency of correct classification (p+c + p−c)/2. For each Sig, there is an easily computable optimal threshold c = c* maximizing PERF(c, Sig) over all potential thresholds c. This maximized performance J*(Sig) = PERF(c*, Sig) will define the “G+ detecting power” of signature Sig.

Exchanging G+ and G−, as well as the scores s+(M, Sig) and s−(M, Sig) we similarly define the G− detecting power J*(Sig) of signature Sig.

Among all potential signatures Sig within our biomarkers target pool TP(2k), we will now seek two optimized signatures, Sig+ maximizing the G+ detecting power J+(Sig) and Sig− maximizing
the $G^-$ detecting powers $J^{-}(\text{Sig})$. But $J^{+}(\text{Sig})$ and $J^{-}(\text{Sig})$ have many local maxima, and the set of all potential signatures within $\text{TP}(2k)$ has very large cardinal. To solve this combinatorial challenge, we implement the separate maximizations of $J^{+}(\text{Sig})$ and of $J^{-}(\text{Sig})$ by Simulated Annealing as described next.

3.5 Optimized Signature Discovery by Simulated Annealing

Simulated annealing is a powerful stochastic descent technique to search for the global maximum of an “objective function” defined for configurations belonging to a very large discrete space [19], [20], [21]. Here, we implement a simulated annealing search to find an optimal signature $\text{Sig}^*$ maximizing the $G^+$ detecting power $J^{+}(\text{Sig})$ over the set of all potential signatures $\text{Sig}$ included in our target pool $\text{TP}(2k)$ of 2k biomarkers.

Each potential signature $\text{Sig}$ can be naturally coded as a binary sequence of length $2k$, where the coordinates equal to 1 correspond exactly to the biomarkers which belong to the list $\text{Sig}$. Starting from any initial signature, we successively visit each one of its $2k$ binary coordinates, and randomly modify this coordinate according to a “simulated annealing rule”. This procedure is repeated after each sequence of 2k steps.

At step $n$, the currently visited signature coordinate is randomly replaced by 0 or 1. This tentatively replaces the current signature $\text{Sig}_n$, by a new signature $\text{Sig}_{n+1}$, which is accepted or rejected with a precise probability explicitly computed in terms of $J^{+}(\text{Sig}_{n+1}) - J^{+}(\text{Sig}_n)$ and of a virtual “temperature” $T_n = (0.95)^n$ [19]. This iterative search essentially stops when $T_n \equiv 0$.

Our numerical implementation involves $200 \times 2k$ updates of signature coordinates per simulated annealing search. We implement multiple simulated annealing searches and retain the signature $\text{Sig}^*$ achieving the highest value for the $G^+$ detecting power $J^{+}(\text{Sig})$.

Similar simulated annealing searches are implemented to discover a signature $\text{Sig}^*$ maximizing the $G^-$ detecting power $J^{-}(\text{Sig})$.

3.6 Signature Based Patient Classification

After computing two optimal signatures $\text{Sig}^+$ and $\text{Sig}^-$, each mass spectrum $M$ can be characterized by its $G^+$ score $s^+(M) = s^+(M_1, \text{Sig}^+)$ and its $G^-$ score $s^-(M) = s^-(M, \text{Sig}^-)$. We then classify any new spectrum $M$ into $G^+$ if $s^+(M) \geq \alpha s^-(M) + \beta$, and into $G^-$ otherwise. The best coefficients $\alpha$ and $\beta$ are computed by maximizing the performance $(p^+ + p^-)/2$ of this signature based classifier over the training data set. Since $\text{Sig}^+$ and $\text{Sig}^-$ were restricted to be within the biomarker target pool $\text{TP}(2k)$, we denote by $\text{PERF}(2k)$ the performance level of this optimized signature based classifier.

When we have two replicate spectra $M_1, M_2$ per patient, we replace $s^+(M_1$) and $s^-(M_2)$ by their average $s^+$ and do the same for $s^-$, before constructing as above the best separator based on the sign of $s^* = (\alpha s^+ + \beta)$.

We progressively increase $k$ starting with a small value, and repeat the whole signature discovery procedure until $\text{PERF}(2k)$ reaches a satisfactory plateau.

Representing each mass spectra $M$ by the planar point $[s^+(M), s^-(M)]$ provides an efficient graphic display of $G^+$ and $G^-$ (see example in Section 4.3).

3.7 Cross Validation of Classifier Performance

To evaluate as defined above the performance of any classifier we need to estimate the frequencies $p^+$ and $p^-$ of correct decisions when the patient is resp. in $G^+$ or in $G^-$. Since our data sets have moderate size, we estimate $p^+$ and $p^-$ by a classical 10-fold cross validation analysis. So we randomly split the whole data set into 10 subsets and at each cross validation round, nine of these subsets are used for training a classifier and the 10th subset becomes the testing set. The percentages $p^+$, $p^-$ of correct decisions are computed within the left out 10th set. These training-testing steps are repeated 10 times, and we evaluate $p^+$, $p^-$ by averaging the 10 partial estimates just obtained. To capture the variability of these $p^+$ and $p^-$ estimates, this 10-fold cross validation is repeated for 100 such random partitions of the whole data set. This provides two samples of 100 estimates, one for $p^+$ and one for $p^-$. The mean and standard deviations of these two samples provide our final estimates of $p^+$ and $p^-$, as well as the associated 95 percent confidence intervals.

4 RESULTS FOR COLORECTAL CANCER MALDI DATA SET

Our colorectal cancer data set (see Section 2.1) involved 119 patients and 238 raw MALDI mass spectra (two replicates per patient): the control group CTR and the Colorectal Cancer group CRC, which was pre-classified according to cancer stage into three subgroups (Adenoma ADE, early cancer ECR, late cancer LCR). We have implemented our signature discovery algorithms for four discrimination tasks: CTR versus CRC, ADE versus ECR, ADE versus LCR, ECR versus LCR.

4.1 Pre-Processing Results

In our colorectal cancer data set, each raw spectrum listed roughly 37,000 distinct “m/z” ratios ranging from 800 to 10,000. Pre-processing was implemented as in Section 3.1, with pre-processing parameters set at $u = 0.0003$, $v = 0.017$, $w = 0.025$, $th = 2$. We thus detected an average of 330 strong peaks per spectrum. Fig. 1 displays the detected strong peaks for one typical raw mass spectrum within the m/z range [1,800, 2,000].

4.2 Biomarkers Target Pools

Our 842 reference biomarkers $B_i$ were positioned at successive abscissas $800 \times 1,003'$ with $j = 0, 1, 2, \ldots$. For each benchmark discrimination task CTR vs CRC, ADE vs ECR, ADE vs LCR, ECR vs LCR, we then successively extracted target pools $\text{TP}(2k)$ of 2k highly discriminating biomarkers, with $2k = 2, 4, \ldots, 40$.

For each benchmark task and each $k$, we computed two optimized signatures $\text{Sig}^+$ and $\text{Sig}^-$ within $\text{TP}(2k)$ and the performance level $\text{PERF}(2k)$ of the signature based classifier we have associated above to the pair of signatures $\text{Sig}^+$ and $\text{Sig}^-$. Signature based discrimination between cancerous and control group (CRC versus CTR) reached perfect performance level 100 percent for $2k = 8$. For discrimination between cancer stages ADE versus ECR, ADE versus LCR, ECR versus LCR, our signature based classifiers reached their respective performance plateaus for $2k = 30, 40, 32$.

4.3 Detailed Results for Discrimination between Adenoma and Early Colorectal Cancer

For signature based discrimination between $G^+ = \text{ADE}$ and $G^- = \text{ECR}$, we display in Fig. 2 the iterative maximization of $J^{+}(\text{Sig})$ by a simulated annealing search with a total of 6,000 temperature cooling steps. As above $J^{+}(\text{Sig})$ is the $G^+$ detecting power, with $G^+ = \text{ADE}$. As Fig. 2 indicates, $J^{+}(\text{Sig})$ reaches a maximum after roughly 4,100 simulated annealing steps and then stabilizes during the last 1,900 annealing steps. The optimal signature $\text{Sig}^* = \text{Sig}^{\text{ADE}}$ achieves correct classification of single spectrum replicates with frequency $p^+ = 91\%$ within $\text{ADE}$ and frequency $p^- = 95\%$ within $\text{ECR}$, yielding an $\text{ADE}$ detecting power of 93 percent. The second optimized signature $\text{Sig}^* = \text{Sig}^{\text{ECR}}$ achieves an $\text{ECR}$ detecting power of 95 percent.

Via the optimal signatures $\text{Sig}^{\text{ADE}}$ and $\text{Sig}^{\text{ECR}}$ we compute for each patient an $\text{ADE}$ score $s^{\text{ADE}} = s^+$ and an $\text{ECR}$ score $s^{\text{ECR}} = s^-$. The 67 patients belonging to either ADE or ECR can then be viewed as planar points with coordinates $(x = s^{\text{ADE}}, y = s^{\text{ECR}})$, as
4.4 Optimized Signatures for Colorectal Cancer Stage Discrimination

For each one of the colorectal cancer stage discrimination tasks ADE versus ECR, ADE versus LCR, ECR versus LCR, as well as for the much easier discrimination between cancerous and healthy patients CRC versus CTR, Table 1 exhibits separately the signature selected biomarkers discriminating in favor of each patient group. More precisely, for each discrimination task $G^+$ versus $G^-$, we pool together all the biomarkers belonging to at least one of the two optimized signatures $\text{Sig}^+$ and $\text{Sig}^-$; we then split this highly selective pool of biomarkers into two sets: the $G^+$ biomarkers and the $G^-$ biomarkers, which we display separately. The simultaneous presence of several such $G^+$ biomarkers then strongly points to $G^+$ patients, with a similar interpretation for $G^-$. With two replicates per patient on our colorectal cancer data set, the signature based classifiers just constructed for single replicates, are readily extended to pairs of replicates as outlined in Section 3.6. The performances $p^+$ and $p^-$ of these “patient level” classifiers are summarized in the “patient level” panel of Table 2, where the “single spectrum level” panel reports the performance based on single replicate classification. These basic “single spectrum” performance levels for our signature based algorithms will be compared below to the single spectrum performances of machine learning algorithms in Section 4.5.

Discriminating the control group CTR from the union CRC of all three colorectal cancerous groups ADE, ECR, LCR, was the easiest of the four tasks studied here, and was achieved with 100 percent accuracy, by very short signatures. The three discrimination tasks between the three cancer stages ADE, ECR, LCR required longer signatures, to achieve fairly good classification levels.

4.5 Performance Comparisons with Machine Learning Techniques

Colorectal cancer detection by machine learning applied to MALDI/SELDI data has been studied by several publications. For discrimination between colorectal cancer and controls, [26] used support vector machine learning to achieve correct classification frequencies of $p^+ = 83 \pm 4\%$ and $p^- = 89 \pm 3\%$, [27] used decision trees to reach $p^+ = 65\%$ and $p^- = 90\%$, and [28] applied K-nearest...
neighbors to reach \( p^+ = 60\% \) and \( p^- = 46\% \). With 100 percent accuracy for colorectal cancer versus control, our signature discovery algorithm clearly outperforms these three previously published machine learning results.

We are not aware of previous publications on discrimination between colorectal cancer stages based on MALDI/SELDI data. Since our colorectal cancer MALDI data is not published, we have thus implemented four benchmark machine learning algorithms on each one of our colorectal cancer stages discrimination tasks: support vector machines with Gaussian kernel, K-nearest neighbor, decision tree and random forest.

As in [12], these machine learning methods classically characterize each biomarker through the t-statistic \( t = \frac{m^+ - m^-}{\sqrt{s^+ + s^-}} \) where \( m^+, \sigma^+ \) and \( m^-, \sigma^- \) are the means and standard deviations of the biomarker intensities observed in \( G^+ \) and \( G^- \). The \( N \) biomarkers with the highest t-statistic are then selected as reference biomarkers. For \( N = 2, 4, 6, 8, \ldots \), we have implemented these four benchmark machine learning techniques on all our colorectal cancer discrimination tasks. In each case, we then selected the \( N \) for which machine learning performance reached a first plateau. The implementation used the SVM, KNN, DT and RF toolboxes provided by MATLAB. Note that for SVM, we have carefully selected the best values for two key learning parameters (cost “tradeoff” and gaussian kernel “scale”).

Performance comparisons between machine learning and our signature based discrimination are implemented for classification of single MALDI spectra, namely, at the “single spectrum level” of Table 2. Performances were evaluated by repeated 10-fold cross-validation as outlined in Section 3.7.

Among machine learning techniques, SVM is clearly superior to KNN, DT and RF. The number of biomarkers used by SVM is around 40 for the three delicate discriminations between cancer stages and 10 for the much easier cancer versus control task. Our signature based classifiers exhibit comparable performance levels with SVM for all the tasks, but offer two clear advantages over SVM: for comparable performances, our optimized signatures systematically involve less biomarkers than SVM, and the associated classifiers are fully and explicitly interpretable, which is not at all the case for the black box classifiers generated by SVM.

## 5 Results on Ovarian Cancer SELDI Data Sets

For all raw mass spectra in the published ovarian cancer SELDI data sets 4-3-02 and 8-7-02, pre-processing was implemented as

---

**TABLE 1**

<table>
<thead>
<tr>
<th>ADE vs ECR</th>
<th>ADE vs LCR</th>
<th>ECR vs LCR</th>
<th>CRC vs CTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1106 839</td>
<td>1146 819</td>
<td>921 819</td>
<td>1802 1202</td>
</tr>
<tr>
<td>1146 873</td>
<td>1192 839</td>
<td>1112 1029</td>
<td>3183 1439</td>
</tr>
<tr>
<td>1188 1261</td>
<td>1343 1246</td>
<td>1570 1032</td>
<td>3708 1652</td>
</tr>
<tr>
<td>1192 1288</td>
<td>1542 1261</td>
<td>1953 1210</td>
<td>4155 1657</td>
</tr>
<tr>
<td>1397 3610</td>
<td>1570 2651</td>
<td>2558 1213</td>
<td></td>
</tr>
<tr>
<td>1579 3985</td>
<td>1845 2773</td>
<td>2589 2143</td>
<td></td>
</tr>
<tr>
<td>1845 6283</td>
<td>2007 3844</td>
<td>4180 2331</td>
<td></td>
</tr>
<tr>
<td>1851 7984</td>
<td>4180 7254</td>
<td>5033 2628</td>
<td></td>
</tr>
<tr>
<td>2475 8630</td>
<td>4193 7587</td>
<td>6320 2807</td>
<td></td>
</tr>
<tr>
<td>3675 8813</td>
<td>5186 7984</td>
<td>6339 3856</td>
<td></td>
</tr>
<tr>
<td>5064 8032</td>
<td>8630 8813</td>
<td>6894 4944</td>
<td></td>
</tr>
<tr>
<td>6377 8813</td>
<td>6915 5829</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6416</td>
<td></td>
<td>7542 6226</td>
<td>8032 7889</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8376</td>
<td></td>
</tr>
</tbody>
</table>

For each discrimination task \( G^+ \) vs \( G^- \), the biomarkers belonging to either \( \text{Sig}^+ \) or \( \text{Sig}^- \) are first pooled together, before splitting this pool into \( G^+ \) and \( G^- \) biomarkers, displayed in two distinct columns. In the signature based classification of a new mass spectrum \( M \), the column under \( G^+ \) gathers the peak abscissas whose presence in \( M \) are strong indicators that \( M \) is of \( G^+ \) type. An analogous statement holds for \( G^- \). For each one of these four discrimination tasks, the number of key biomarkers selected by each optimal pair of signatures remains quite moderate, namely 23, 23, 29, and 8.

---

**TABLE 2**

<table>
<thead>
<tr>
<th>Single Spectrum level</th>
<th>ADE vs ECR</th>
<th>ADE vs LCR</th>
<th>ECR vs LCR</th>
<th>CRC vs CTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p^+ ) and ( p^- )</td>
<td>88±5% 80±5%</td>
<td>85±5% 88±4%</td>
<td>83±5% 83±4%</td>
<td>100±0% 100±0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient level</th>
<th>ADE vs ECR</th>
<th>ADE vs LCR</th>
<th>ECR vs LCR</th>
<th>CRC vs CTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p^+ ) and ( p^- )</td>
<td>87±5% 87±5%</td>
<td>89±6% 89±7%</td>
<td>86±4% 86±4%</td>
<td>100±0% 100±0%</td>
</tr>
</tbody>
</table>

Discrimination between \( G^+ \) and \( G^- \) based on our optimized signatures is evaluated by the frequencies \( p^+ \) and \( p^- \) of correct classifications within \( G^+ \) and \( G^- \). The \( p^+ \) and \( p^- \) estimates, and their 95 percent confidence intervals are obtained by repeating 100 times a 10-fold cross validation procedure. Performances on colorectal cancer data are given for single mass spectrum classification, as well as for two replicates classification (patient level).
above, with windows size parameters set at \( u = 0.0003, \ v = 0.017, \ w = 0.025 \) and peak strength threshold \( th = 2 \). On data set 4-3-02, the baseline had been removed prior to publication. In both data sets, all publicly accessible mass spectra had been pre-aligned to a fixed list of 15,154 peak positions, which we also adopted as our list of reference biomarkers.

Our signature based classifiers were then implemented separately on the SELDI data sets 4-3-02 and 8-7-02 to discriminate between Ovarian Cancer patients and Control patients. Classification performance reached a plateau at biomarker target pool size \( 2k = 22 \) for data set 4-3-02 and at target pool size \( 2k = 8 \) for data set 8-7-02. Final performances were evaluated as above by repeated 10-fold cross validations.

For the SELDI data set 4-3-02, signature based classifiers achieved correct decision frequencies of 95.4% for the ovarian cancer group and 94.4% for the control group. These performances compare quite well with published results obtained on the same data set by other algorithms, namely 96.5±3.5% and 93 ± 6% in \[4\], 96.5±3.5% and 97.5±2.5% in \[6\], 90 percent and 83.3 percent in \[23\]. For the SELDI data set 8-7-02, signature based classification reached 100 percent accuracy, which agrees with the results of previous studies \[6\], \[24\], \[25\], \[23\].

### 6 Conclusion

A natural goal for computer aided discrimination between various cancer stages or cancer types on the basis of proteomic m/z spectra acquired by MALDI or SELDI mass spectrometers is to select highly discriminating “signatures” gathering explicit small sets of biomarkers. This is a delicate task due to the well known repeatability and variability problems linked to proteomic mass spectra. Potential applications include software tools for early clinical diagnosis, as well as disease progression monitoring and evaluation of response to treatment.

A large number of machine learning studies have reportedly achieved good automated classification performances for specific data sets of proteomic mass spectra. Nevertheless, few machine learning algorithms have been incorporated into routinely and clinically usable software tools. This is in part due to the fact that machine learning generates “black box” classifiers with low biological interpretability.

In this paper we present and successfully test innovative algorithms to perform automated discovery of optimized short biomarkers signatures, to efficiently discriminate between given data sets of MALDI or SELDI mass spectra associated to various cancer stages. Our optimized signatures have the advantage of fairly direct biological interpretability.

Automated computer search for highly discriminating biomarkers signatures is an algorithmic problem with quite high computational complexity. This motivates our innovative use of simulated annealing optimization techniques to search for signatures with high discriminating power. We have combined this approach with efficient selection of target pools of potential biomarkers, and thus implemented powerful software tools for automated signature discovery.

We have first successfully tested our signature based mass spectrometry classifiers on a new experimental set of 238 MALDI-TOF mass spectra acquired from patients at various stages of colorectal cancer. Correct classification frequencies were good and compared quite favorably with the performance levels achieved on the same data set by four benchmark machine learning techniques: support vector machines, K-nearest neighbor, decision trees, random forest. Our signature discovery approach handles easy discrimination tasks (colorectal cancer versus control) certainly as well as all other machine learning techniques. For the more delicate discrimination between three colorectal cancer stages, the performances of our signature based classifiers were statistically indistinguishable from those of SVM, and these performance levels were clearly better than those for KNN, DT and RF. However, our optimized signatures had two key advantages over SVM: our signatures always required less biomarkers than SVM to achieve similar performances, and generated explicit and fully interpretable classifiers, instead of the black-box classifiers generated by SVM.

We have also tested our signature discovery techniques on two published data sets of SELDI-TOF mass spectra acquired from ovarian cancer patients. On these two ovarian cancer data sets, our signature based classifiers performed very well, and either matched or improved the performances achieved by other published techniques.

A key feature of our optimized signatures is that they do provide short lists of discriminating biomarkers identified by their m/z ratios. This is an important step to narrow down small sets of high priority biomarkers, to be targeted in further experimental studies or therapeutics research.

In future work, we intend to calibrate and test our signature discovery technique on MALDI and/or SELDI data sets acquired on breast cancer patients.

### Acknowledgments

This study was supported in part by grants from the CARIPARO and AIRC Foundation. Biological samples were collected and stored by the Tumor Tissue Biobank of Surgical Clinic I. Data acquisition was partially funded by University of Texas Health Sciences.

### References


Queries to the Author

Q1. Please check whether the affiliations for all the authors are ok as set.
Q2. Figure 1 is mismatch with provided pdf. We have followed source file. Please check.
Q3. Table 3 is not cited in the text. Please cite table 3 in the text at appropriate place/places.
Q4. Please provide names of all the author’s in the references.
Q5. Please provide page range for reference [24].