A WORKFLOW FOR PREPROCESSING AND PROTEOMIC BIOMARKER IDENTIFICATION ON MASS-SPECTROMETRY DATA

Michael Handler\textsuperscript{1+1}, Johannes D. Pallua\textsuperscript{2+2}, Georg Schäfer\textsuperscript{1+*}, Michael Netzer\textsuperscript{1}, Melanie Osl\textsuperscript{1}, Michael Seger\textsuperscript{1}, Bernhard Pfeifer\textsuperscript{1}, Michael Becker\textsuperscript{4}, Stephan Meding\textsuperscript{2}, Sandra Rauser\textsuperscript{2}, Axel Walsh\textsuperscript{5}, Helmut Klocker\textsuperscript{1}, Georg Bartsch\textsuperscript{3}, Christian W. Huck\textsuperscript{2}, Christian Baumgartner\textsuperscript{1}, and Günther K. Bonn\textsuperscript{2}

\textsuperscript{1} Institute of Electrical, Electronic and Bioengineering, University for Health Sciences, Medical Informatics and Technology, Austria
\textsuperscript{2} Institute of Analytical Chemistry and Radiochemistry, Leopold-Franzens University, Innsbruck, Austria
\textsuperscript{3} Department of Urology, Medical University of Innsbruck, Innsbruck, Austria
\textsuperscript{4} Bruker Daltonik GmbH, Bremen, Germany
\textsuperscript{5} Institute of Pathology, Helmholtz Center Munich, Neuherberg, Germany

\begin{abstract}
A core technology in proteomics is mass spectroscopy (MS) that permits the measurement of thousands of proteins/peptides simultaneously. Sophisticated data mining methods are necessary to identify highly predictive proteomic biomarker candidates in generated MS spectra that are specific to a certain disease. However, before analysis can be started the preprocessing of raw mass spectra is an essential task, mainly due to the presence of background signals in the spectra such as electrical and chemical noise. In this work we present a new data mining workflow for the identification of proteomic biomarker candidates using mass spectrometry data. The workflow includes two major steps: 1) the preprocessing of raw spectra, and 2) the identification of highly discriminating candidate masses using a 3-step feature selection approach by combining the advantages of efficient filter and effective wrapper techniques. With the proposed workflow we were able to identify putative candidate biomarkers in a life-threatening human disease using matrix-assisted laser desorption/ionization imaging MS (MALDI-IMS).

\end{abstract}

\begin{keywords}
Proteomics, Biomarker Discovery, Mass Spectrometry, Data Preprocessing, Data Mining, Human Disease

\end{keywords}

\section{Introduction}

Proteomics deals with the large-scale determination of protein expression patterns, protein interactions and protein pathways in blood serum or tissue cells. Proteomic biomarkers can show predictive abilities in the identification of diseases that were previously difficult to detect [1]. The discovery process of proteomic biomarkers generally includes the following steps: 1) experimental design, 2) study execution and sample collection, preparation and separation 3) mass spectrometry (MS) analysis, 4) biomarker identification and biological interpretation and 5) independent validation.

A variety of experimental designs are used in biomarker discovery studies. The most common design is the case-control study [2], where samples from two or more predefined groups, e.g. cases versus controls, are collected. Once sample collection, preparation, separation and MS analysis (liquid chromatography (LC) - MS or matrix-assisted laser desorption/ionization (MALDI) - MS have become the methods of choice for protein identification), have been carried out, a comprehensive technical review of generated data is necessary. This is mainly due to the presence of background signals in the spectra, i.e. internal fragments, electronic noise, chemical noise and ions originating from unknown fragmentation pathways [3].

Sophisticated data mining methods are now needed to preprocess collected raw mass spectra for generating denoised, baseline corrected and normalized spectra that allow for quality-assured processing of the subsequent targeted search for biomarkers in the data. For this discovery step, spectra are carefully separated into groups according to the study design (e.g. cases vs. controls). Supervised search and feature selection methods are usually applied to the identification of highly discriminating masses in the spectra when comparing samples from well phenotyped study cohorts. Subsequent discovery phases now include the verification, biological interpretation and validation of found biomarker candidates which is generally the most time consuming part of the whole biomarker discovery process.

Numerous workflows are described in the literature for the identification of biomarkers based on MS raw spectra using different modalities in preprocessing, classification and validation (e.g. [4, 5, 6]).

In this work we present a new data mining workflow...
for the identification of proteomic biomarker candidates in human tissue using MALDI - imaging mass spectrometry (IMS). The proposed workflow includes two major steps: 1) the preprocessing of raw mass spectra, and 2) the identification of highly discriminating biomarker candidates using a 3-step feature selection approach by combining the advantages of efficient filter and effective wrapper techniques.

2 Material and Methods

For the analysis of provided raw MALDI-IMS spectra a two-step strategy was developed consisting of a data preprocessing modality including binning and adjustment of m/z values, baseline correction, normalization, peak detection and quality assessment of spectra, coupled with a paradigm for the targeted identification of highly discriminating m/z values in the treated MS spectra (putative candidate biomarkers).

2.1 MALDI-IMS Experiments

Tissue specimens were cut on a cryostat at 12 µm thickness and mounted onto conductive indium-tin oxide (ITO) coated glass slides (Bruker Daltonik, Bremen, Germany). Matrix application has been done by the ImagePrep station (Bruker Daltonik, Bremen, Germany) using the standard methods provided with the instrument. The sections have been coated with sinapinic acid (SA). MALDI measurements were performed on an UltrafleXtreme instrument (Bruker Daltonik, Bremen, Germany) in positive linear mode. Histological staining after MALDI measurements and removal of the matrix was done according to standard protocols. For histopathological reevaluation and comparison of the imaging results the stained slides were scanned with the MIRAX DESK system (Carl Zeiss MicroImaging GmbH, Göttingen, Germany) and co-registered with the MALDI-IMS results.

2.2 Statistical Analyses and identification of discriminatory m/z peaks

Tissue type-associated spectra were selected using the FlexImaging 3.0 software (Bruker Daltonik). For this purpose, an experienced uropathologist evaluated the stained section and defined the regions of interest (ROI) for each tissue type. Extracted mass spectra of ROIs underwent recalibration on common “background” peaks (also known as spectral alignment) and normalization based on their total ion count in the observation mass range utilizing ClinProTools 2.2 software (Bruker Daltonik). An average spectrum across the spectra of each ROI was used to obtain the average peak intensities of the integration ranges. Significant differences in peak intensities between the histological groups were identified by visual inspection and the best discriminatory m/z peaks were selected. In a second bioinformatics statistical approach a feature selection algorithm was used to identify discriminatory biomarker candidates.

2.3 Preprocessing of MS spectra

Before computational analysis of MS spectra can be started, it is necessary to standardize all given source spectra to a common format to make spectra comparable. Therefore it is essential to channel the raw spectra through a preprocessing pipeline (see Figure 1). The basic concept of preprocessing mass spectra used in this work was introduced by Li et al. [4]. Some of the algorithms use parts of the PROcess library of the Bioconductor package of R [7]; the steps within the dashed box are a repetition of previous steps - see 2.3.8)

Figure 1. Preprocessing and evaluation steps (the white background indicates a Java implementation; the gray background indicates algorithms of the PROcess library [4] of the Bioconductor package of R [7]; the steps within the dashed box are a repetition of previous steps - see 2.3.8)
2.3.2 Adjustment of m/z values

To allow for consistent comparisons of mass spectra it is necessary to adjust all spectra to a defined m/z range. This interval is defined from the largest common minimal m/z value \( x_{\text{min}} \) to the smallest common maximal m/z value \( x_{\text{max}} \) of all treated spectra. In a next step m/z values less than \( x_{\text{min}} \) or greater than \( x_{\text{max}} \) are removed from the spectra. After this procedure the entire mass range of all spectra is equal and consistent (see Figure 2(a)).

2.3.3 Baseline correction

Due to chemical noise and ion overload, source spectra can contain elevated baselines which must be removed to make the spectra comparable [4]. To retrieve the baseline of a given spectrum, the estimated bottom of a given range must be determined. This bottom is estimated by local regression (local regression to the points below a certain quantile or by using the local minima of a given range around the m/z value) (see Figure 2(b)). This part of the workflow is executed using the bslhoff operation of the \textit{R} package \textit{PROcess} [4].

2.3.4 Normalization

To remove effects of experimental noise, which cause variations in the amplitude of spectra, the \textit{total ion normalization} procedure was used [4]. After selecting a cutoff value (minimal m/z value to start the normalization) the AUC (Area Under the Curve) is evaluated for each spectrum. If the intensities are given on equally spaced m/z values the AUC of a spectrum can be seen as the sum of all its intensities. In a second iteration step the intensities of all spectra \( U_i \) are equalized to the normalized spectrum

\[
N_i(x) = U_i(x) \cdot \frac{AUC_M}{AUC(U_i)}
\]

where

\[
AUC(U_i) = \sum_{j=1}^{n} U_i(j)
\]

denotes the AUC of spectrum \( U_i \), \( n \) denotes the number of m/z values in spectrum \( U_i \) and \( AUC_M \) is the median AUC of all spectra (see Figure 2(c)).

2.3.5 Peak detection

Next, peak detection (peaks represent specific, abundant polypeptides in the sample [4]) is carried out. These peaks are subsequently used for quality assessment (see 2.3.6) and the alignment to given internal standards (see 2.3.7). The function \text{isPeak} of the \textit{PROcess} package was used [4]. The peaks are detected based on three different measures: a) a signal-to-noise ratio related threshold b) a minimal intensity threshold, and c) a minimal shape ratio of a putative peak.

In contrast to the work of Li et al. [4] we developed a procedure allowing parallelization of this step.

2.3.6 Quality assessment of spectra

To filter out spectra of poor quality another preprocessing step is needed. For this modality a routine similar to the quality function of the \textit{PROcess} package [4] was used. This routine was modified in such a way that each spectrum is evaluated for its quality separately by allowing multithreaded execution.

For the evaluation of spectral quality three measures are determined:

- \textit{Quality}: Ratio of the AUC of the spectrum compared to its AUC after subtraction of an estimated noise envelope.
- \textit{Retain}: Number of points with intensity greater than 5 times the noise envelope over the total number of points in the spectrum.
- \textit{Peak}: Ratio of the number of detected peaks in each spectrum to the average number of detected peaks over all compared spectra.

2.3.7 Alignment to internal standards

Because of measurement variations, peaks in different spectra that correspond to the same protein may be located at different m/z values [4]. To align the spectra to each other it is necessary to define reference m/z values, which either occur naturally in all given spectra or by peaks representing internal standards. An internal standard is a compound added to a sample in known concentration to facilitate the qualitative identification and/or quantitative determination of the sample components [8]. In the described approach two internal standards were used. Before the alignment can be done the two peaks of the natural/internal standard need to be detected. Using linear interpolation the alignment is calculated as follows:

\[
\hat{x} = (x - IS_l) \cdot \frac{IS_u - \hat{IS}_l}{IS_u - IS_l} + \hat{IS}_l
\]

where \( x \) is the original m/z value, \( \hat{x} \) is the aligned m/z value, \( IS_l \) and \( IS_u \) denote the m/z values of the corresponding lower and upper internal standard peaks of the processed spectra, and \( \hat{IS}_l \) and \( \hat{IS}_u \) hold the actual m/z values of the lower and upper internal standard (see Figure 2(d)).

2.3.8 Recalibration of preprocessed spectra

Due to the removal of poor quality assessed spectra and alignment to internal standards, parts of the initial preprocessing modality, i.e. binning, adjustment and normalization, have to be repeated.
2.3.9 Generation of mean spectra

As a final step, a mean spectrum is computed if multiple spectra per sample are available. This step reduces noise and further improves the signal-to-noise ratio in the average spectrum.

2.4 Identification of biomarker candidates

After preprocessing, the spectra have comparable m/z and intensity values. Consequently, the spectra can be formally defined as a set of tuples, \( T = \{(c_j, m) | c_j \in C, m \in M\} \) with \( C = \{case, control\} \), where \( C \) is the set of class labels and \( M \) is the set of features (m/z values in the spectrum). We further developed an algorithm published by our group [9], identifying those masses in \( M \) showing a highest discriminatory ability according to \( C \). These masses are identified by employing a three-step feature selection procedure.

In the first step, a filter approach selects relevant features (m/z values). However, the result set of features from step 1 contains regions of adjacent features that are highly correlated, as most of them are redundantly representing the same information of the spectra. Therefore, in step 2 we identify a representative for every region in the spectra. In step 3, a wrapper approach further reduces the dimensionality of the result set of step 2 by optimizing the discriminatory ability.

2.4.1 Step 1: Selecting relevant features

In our experimental setup we use the Student t-test as filter approach and the resulting P-value as measure for the discriminatory ability. Additionally, we calculate a second parameter \( \Delta \) representing the ratio of the mean intensities in each class \( \bar{x}_{c_i} \) relative to the maximum intensity \( I_{\text{max}} \) in all spectra. The parameter \( \Delta \) is defined by

\[
\Delta = \frac{|\bar{x}_{c_1} - \bar{x}_{c_2}|}{I_{\text{max}}}
\]

The parameter \( \Delta \) is important to ensure that differences in the intensity can also be technically detected. A feature is defined as relevant, if \( P\text{-value} < \alpha \) and \( \Delta > \Gamma \). The parameters \( \alpha \) and \( \Gamma \) have to be set by the user. In our experimental setup we set \( \alpha = 0.1 \) and \( \Gamma = 0.03 \).

2.4.2 Step 2: Selecting region representatives

We use a forward selection method as depicted in Listing 1 to identify a representative for every region in the spectra. The representative feature is the feature with the highest quality (i.e. discriminatory ability according to used filter method from step 1) within the region. The size \( s \) of the region is a relative measure depending on the index of the feature representing the m/z value. This is due to technical reason as many different fragments of peptides with low molecular weight were causing many narrow peaks in the region of low m/z values. In our experimental setting we set \( s = 3 \) as proposed in [9].

**Listing 1**

```plaintext
procedure representatives (dataSet DS): featureSet rep 
for all descendingly ranked features \( f_i \) in DS do 
    rs: = 0.5 * rs(\( f_i \)); 
    if no feature \( f_j \) exists in rep with index(\( f_j \)) - rs < index(\( f_i \)) < index(\( f_i \)) + rs rep.add(\( f_i \)); 
return rep;
```

2.4.3 Step 3: Selecting the best features

To reduce the number of features we use a wrapper-based approach using a classifier and a search strategy to find a smaller feature subset while keeping the discriminatory ability at least constant. In our experimental setting we applied logistic regression [10] as classifier and modified binary search (MBS) [9] as search strategy and use the area under the ROC curve as measure for the discriminatory ability. The resulting features represent highly discriminating m/z values in the spectra. Finding the local maximum near the selected m/z values ensures that the feature is representing a peak in the spectra.

3 Results

3.1 Experimental data

For experimental evaluation of our developed workflow a data set of a life-threatening human disease was provided. This data set comprises 19,260 MALDI-IMS spectra of 14 diseased-form samples and 15 harmless-form samples.

3.2 Preprocessing

Key preprocessing steps are illustrated in Figure 2, showing adjustment, baseline correction, normalization, and alignment of spectra. For more details see figure legend.

3.3 Identification of biomarker candidates

By applying our developed data mining workflow, which allows for the preprocessing of raw mass spectra and the subsequent search for highly discriminating masses in treated spectra, four candidate masses could be identified when comparing diseased-form vs. harmless-form samples.

Table 1 shows the identified masses and the estimated mean AUC values using logistic regression analysis. We
Figure 2. Visualization of preprocessing steps

(a) **Adjustment**: Dashed lines represent the common minimal and common maximal m/z values identified in all spectra. The mass range between both borders represents the spectra after adjustment.

(b) **Baseline substraction**: Gray indicates the spectrum before baseline substraction, black after baseline substraction.

(c) **Normalization**: Gray shows the spectrum before normalization, black after normalization.

(d) **Alignment**: Dashed lines represent the internal standards. The upper plot shows the unaligned spectrum, the lower the aligned spectrum. Additionally, the zoomed areas around the internal standards are displayed.
Table 1. Best discriminating mass peaks: harmless-form (reference class) versus diseased-form.

<table>
<thead>
<tr>
<th>Mass</th>
<th>Mean AUC</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>0.95</td>
<td>0.05</td>
</tr>
<tr>
<td>M2</td>
<td>0.95</td>
<td>0.05</td>
</tr>
<tr>
<td>M3</td>
<td>0.85</td>
<td>0.076</td>
</tr>
<tr>
<td>M4</td>
<td>0.85</td>
<td>0.076</td>
</tr>
</tbody>
</table>

used 10-fold cross validation to statistically validate our results. Note that the absolute m/z values are replaced by mass symbols due to confidentiality issues (a patent application is in preparation). M1 and M2 in particular demonstrated the highest predictive values in terms of sensitivity and specificity, thus exhibiting interesting candidature in the context of the investigated disease.

4 Discussion & Conclusion

In this work we have presented a two-step computational workflow to identify proteomic biomarker candidates using mass spectrometry data. The first step includes the preprocessing of raw mass spectra to remove background and noise signals in the spectra. Inadequate or incorrect preprocessing may result in low quality spectra that do not allow for a statistically significant and biologically relevant analysis of data [11].

In the second step, biomarker candidates are identified using a 3-step feature selection approach by combining the advantages of filter and wrapper techniques. In our experimental setting we applied a Student t-test as filter approach. The t-test has some advantages compared to the originally proposed information gain, as the resulting p-value represents the effect size between two data distributions which is of biological interest. As classifier we choose logistic regression analysis, a common method in biomedical data analysis, which does not require advanced parameterization [12]. Because our method is a generic framework, the filter and classification algorithms can be replaced accordingly, and thus adapted to specific problems. Our approach delivers (i) a list of potential candidate masses as well as (ii) a subset of this list showing the highest discriminatory ability in terms of the AUC. The AUC is a commonly used global index of diagnostic accuracy which is also able to deal with imbalanced data. Currently, our proposed setting for the search for significant masses is defined for dichotomous class problems, however it can easily be extended to multiple classes.

With our proposed approach we were able to find four biomarker candidates discriminating best between tissue samples of the harmless- and diseased form using MALDI-IMS. Note that the workflow can be generally applied to data obtained from other MS instrumentations used in biomarker discovery studies.

Technically, our parallelizable approach allows for an efficient computation of multiple spectra in shorter time, an advantage when preprocessing large amounts of spectra using a computational cluster environment.

5 Acknowledgement

This work was funded by the IMGuS project from the AWS, the Österreichische Nationalstiftung für Forschung, Technologie und Entwicklung, and partly by the GENAU Project: Bioinformatics Integration Network III (BIN III). The authors would like to thank doctors Wolfgang Horninger, Jasmin Bektic and Viktor Skradsky for sample collection for the Biobank, and Konstantin Halikias, Sören-Oliver Deininger and Arnd Ingendoh from the company Bruker Daltonik GmbH, Bremen, Germany for their aid. We also gratefully acknowledge Mag. C. Seifarth for sectioning and for the assistance with sample preparation.

References


