Independent component analysis for the extraction of reliable protein signal profiles from MALDI-TOF mass spectra

Dante Mantini¹,², Francesca Petrucci³,⁴, Piero Del Boccio³,⁴,⁵, Damiana Pieragostino³,⁴,⁵, Marta Di Nicola³,⁴, Alessandra Lugaresi³,⁶, Giorgio Federici⁷,⁸, Paolo Sacchetta³,⁴, Carmine Di Ilio³,⁴ and Andrea Urbani³,⁴,⁵,*

¹Istituto Tecnologie Avanzate Biomediche (ITAB), Fondazione ‘G. d’Annunzio’, ²Dipartimento di Scienze Cliniche e Bioimmagini, Università ‘G. d’Annunzio’, ³Centro Studi sull’Invecchiamento (Ce.S.I.), Fondazione ‘G. d’Annunzio’, ⁴Dipartimento di Scienze Biomediche, Università ‘G. d’Annunzio’, Chieti-Pescara, ⁵Centro Europeo Ricerca sul Cervello (CERC), IRCCS-Fondazione S. Lucia, Roma, ⁶Dipartimento di Oncologia e Neuroscienze, Università ‘G. d’Annunzio’, Chieti-Pescara, ⁷Dipartimento di Medicina Interna, Università di Roma ‘Tor Vergata’ and ⁸Ospedale Pediatrico Bambino Gesù – IRCCS, Roma, Italy

Received on July 20, 2007; revised and accepted on October 16, 2007
Advance Access publication November 14, 2007
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Independent component analysis (ICA) is a signal processing technique that can be utilized to recover independent signals from a set of their linear mixtures. We propose ICA for the analysis of signals obtained from large proteomics investigations such as clinical multi-subject studies based on MALDI-TOF MS profiling. The method is validated on simulated and experimental data for demonstrating its capability of correctly extracting protein profiles from MALDI-TOF mass spectra.

Results: The comparison on peak detection with an open-source and two commercial methods shows its superior reliability in reducing the false discovery rate of protein peak masses. Moreover, the integration of ICA and statistical tests for detecting the differences in peak intensities between experimental groups allows to identify protein peaks that could be indicators of a diseased state. This data-driven approach demonstrates to be a promising tool for biomarker-discovery studies based on MALDI-TOF MS technology.

Availability: The MATLAB implementation of the method described in the article and both simulated and experimental data are freely available at http://www.unich.it/proteomica/bioinf/.

Contact: a.urbani@unich.it

1 INTRODUCTION

Independent component analysis (ICA) is a signal processing technique that can be used to separate distinct underlying signals from mixed recorded signals, on the basis of their statistical properties (Hyvärinen et al., 2001; Stone, 2004). The most common example for illustrating ICA refers to the so-called cocktail party problem. Consider a cocktail party where there are a number of people speaking in the same room. Assume further that there are several microphones in different positions, so that each records a mixture of the speech signals with slightly different weights. In this situation, a blind source separation is required, whether we aim at retrieving the different sound sources from the recordings without any a priori information. ICA is the technique able to do it, if there are at least as many microphones in the room as there are different simultaneous sound sources. In the last years, several efficient algorithms have been developed to solve the ICA problem (Bell and Sejnowski, 1995; Cardoso and Souloumiac, 1996; Hyvärinen, 1999; Ziehe et al., 2000). Furthermore, ICA has recently received attention because of its potential applications in several fields, as demonstrated in studies on specific problems dealing with speech recognition systems, financial time series analysis and biomedical signal processing (Back and Weigend, 1997; James and Hesse, 2005; Jung et al., 2001; Mantini et al., 2005; Smith et al., 2006). ICA techniques have been applied to a number of non-conventional situations, because the basic assumption of the ICA problem, i.e. the independence of the variables, is realistic in many circumstances, hence permitting a completely blind source separation, or feature retrieval. The successful applications of ICA have encouraged its use also for biological systems, which are in general more complex to analyze than non-biological ones (Frigyesi et al., 2006; Liebermeister, 2002; Scholz et al., 2004).

1.1 ICA theory

The ICA model can be mathematically described as:

$$X = AS$$  \hspace{1cm} (1)

where $X = [x_1, \ldots, x_n]^T$ is the matrix of $n$ observed signals, $S = [s_1, \ldots, s_m]^T$ is the matrix of $m$ underlying signals and $A$ denotes the $[n \times m]$ mixing matrix (Hyvärinen et al., 2001; Stone, 2004). It is a generative model, which means that it describes how the observed data are generated by a process of mixing the underlying signals $s_i$, whose estimates are named
independent components (ICs). The minimal required a priori information in the ICA model is the independence of the ICs. A solution for the ICA problem is possible if two additional conditions are met: the number of underlying signals is at most equal to the number of observed signals \((m \leq n)\), and the mixing matrix is full column-rank \((r(A) = m)\). In this case, the ICs can be retrieved by determining an \([m \times n]\) matrix \(W\), named unmixing matrix, such as

\[
S = WX
\]  

(2)

### 1.2 Pre-processing for ICA

In order to obtain \(W\) by estimating a minimum number of parameters, it is necessary to center and whiten the acquired data (Hyvärinen and Oja, 1997). Centering is performed by subtracting the \(j\)-th average value from each mixed signal \(x_j\), so that the underlying signals \(s_j\) become zero-mean; whitening is a linear transformation of matrix \(X\) into another matrix \(\tilde{X}\) whose observations are uncorrelated and with variances equal to unity. The most common method for whitening is the eigenvalue decomposition of the covariance matrix \(E X X^T = E D E^T\), where \(E\) is the orthogonal matrix of eigenvectors of \(E X X^T\) and \(D = \text{diag}(d_1, d_2, \ldots, d_n)\) is the diagonal matrix of its eigenvalues (Yang and Wang, 1999).

The whitened matrix \(\tilde{X}\) can be calculated as

\[
\tilde{X} = E D^{-1/2} E^T X
\]  

(3)

with \(D^{-1/2} = \text{diag}(d_1^{-1/2}, d_2^{-1/2}, \ldots, d_n^{-1/2})\).

Whitening transforms the unmixing matrix \(W\) into a new matrix \(\tilde{W}\), for which:

\[
\tilde{S} = WX = W E D^{1/2} E^T \tilde{X} = \tilde{W} \tilde{X}
\]  

(4)

\(\tilde{W}\) is orthogonal and allows minimizing the number of parameters to be estimated: instead of having to estimate all the coefficients of the original matrix \(W\), we only need to estimate the orthogonal matrix \(\tilde{W}\), containing a lower number of degrees of freedom.

After solving the ICA problem for (4), the unmixing matrix \(W\) can be computed as:

\[
W = \tilde{W} E D^{-1/2} E^T
\]  

(5)

### 1.3 Post-processing of ICA output

At the end of the ICA decomposition, the matrix \(W\) is calculated according to (5) and the matrix \(S\) according to (2). Subsequently, the matrix \(A\), being the pseudoinverse of any (not square) \(W\), is obtained with the formula

\[
A = (W^T W)^{-1} W^T
\]  

(6)

Typical post-processing of the ICA output consists of sorting the ICs with decreasing power level (James and Hesse, 2005). To this purpose, the power \(p_i\) of the \(i\)-th component across the matrix \(X\) is directly calculated from the \(i\)-th column of the matrix \(A\) as

\[
p_i = \sum_{j=1}^{n} a_{ij}^2
\]  

(7)

After sorting the components, the IC waveforms \(s_i\) and the related IC amplitudes \(A_i\) corresponding to the \(i\)-th column of the matrix \(A\), can be stored for further analysis.

### 1.4 Application to MALDI-TOF MS data

An intriguing application of ICA is the processing of proteomic signals, and in particular MALDI-TOF mass spectra. MALDI-TOF mass spectrometers are devices able to produce signals that correspond to the different time of flight of the analyzed proteins, ionized by means of a high-energy laser beam and accelerated with an electric field (Karas, 1996). The acquired spectra always present complex features, because they are composed by a number of overlapping peaks with different amplitudes, related to the abundance of the proteins, and are contaminated by artifacts of biological/physical origin (Gras et al., 1999), i.e. the baseline trend and the background noise. Due to the presence of these disturbances, very sensitive and accurate peak-detection methods, able to correctly separate protein signals from noise, are required. Several processing strategies have been proposed in the literature for analyzing MALDI-TOF data (Coombes et al., 2005; Gras et al., 1999; Mantini et al., 2007; Satten et al., 2004; Yasui et al., 2003); however, the problem of the potential detection of noise peaks as signals still has not been completely solved. This problem seriously limits the development of reliable proteomics tools for biomarker discovery and early disease diagnosis (Diamandis, 2004).

In this perspective, an ICA approach for extracting protein profiles from multi-subject MALDI-TOF MS data is proposed. According to the ICA theory, the observed signals will be the mass spectra and the protein profiles, assumed to be independent to each other, will correspond to the ICs. Each IC is expected to contain single peaks, or multiple peaks that are up- and down-regulated in the same manner across mass spectra. With regard to the basic assumptions for the solution of the ICA problem, it is worth noting that the mixing-matrix is always full-rank, since each mass spectrum cannot be obtained as a linear combination of the other spectra. In turn, a number of mass spectra at least equal to that of expected protein profiles are required for the ICA decomposition.

To the best of our knowledge, in this work ICA is used for the first time for high-dimensional proteomic data analysis, for the separation of the artifacts and for the direct resolution of protein signals. The ICA method has been validated on simulated data for assessing its capability of separating protein peaks, without noticeable signal distortion. Our findings show that the reliability of the proposed method is superior in reducing the false discovery rate of protein peak masses than those of classical methods in terms of peak detection. Moreover, it is demonstrated, using real serum and plasma samples obtained from a group of 30 patients, that the information extracted using ICA from MALDI-TOF data is valuable in the perspective of biomarker-discovery studies.

### 2 METHODS

#### 2.1 Simulated data

Synthetic data were prepared with the aim of reproducing signals with the same characteristics of real spectra obtained from MALDI-TOF...
MS devices. A set of 40 protein profiles was created; for each of them, 1–3 peaks with fixed relative abundance were combined in the same signal. The following equation (Foley, 1987) was adopted to simulate the MS peaks:

\[
x(z) = A_0 \exp \left( \frac{\sigma_p^2}{2\tau^2} \left( \frac{z - z_p}{\tau} \right) \right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{t^2}{2} \right) dt
\]

where \( z \) is the mass/charge (m/z) value, \( A_0 \) is the area of the peak, \( \tau \) is the time constant of the exponential decay, \( \sigma_p \) controls the tailing of the peak, \( z_p \) determines the position of the peak on the m/z axis, the ratio \( \tau/\sigma_p \) is a measure of its asymmetry, and \( h = \frac{1}{\sqrt{\pi}} \) for \( \sigma_p \). The parameters used in the simulation ranged between 180 and 700 for \( A_0 \), between 6000 and 18000 for \( z_p \), whereas it was set to 0.0172 for \( \tau \), and 0.0189 for \( \sigma_p \). Among the 40 protein profiles, 30 were generated using a single peak, whereas 6 and 4 were respectively obtained summing two and three peaks.

A synthetic MALDI-TOF MS dataset \( X_{\text{syn}} \) was prepared by means of a linear mixing of the 40 protein profiles, using realistic weights. A total number of 60 mass spectra with at most 40 peaks were produced; a specific signal (mass spectrum #33) contained an outlier peak, which was absent in the other mass spectra. In order to simulate realistic conditions, a decreasing baseline and a non-uniform background noise were added to each spectrum. Simulated MALDI-TOF mass spectra are shown in Figure 1.

\section*{2.2 Experimental data}

\subsection*{2.2.1 Sample preparation}

The experimental data, referring to both serum and plasma samples, were collected from 30 patients (age 28–40 years) affected by inflammatory auto-immune disease, who signed a written informed consent.

Equine myoglobin, dissolved in 0.1% trifluoroacetic acid (TFA) in deionized water, was used as calibrator. Sample preparation was performed by ZipTip resin (Millipore) C4 tips with sinapinic acid. The samples (20 \( \mu l \)) were first acidified by addition of 5 \( \mu l \) 1% TFA before loading and preparation with a sandwich layer method on MTP ground steel 384 (Bruker Daltonics). First, a sinapinic acid matrix seed layer was created by depositing a droplet (0.5 \( \mu l \)) of a saturated solution of sinapinic acid in 100% ethanol on the target. The C4 resin was first equilibrated by 0.1% TFA. Thereafter, the sample was trapped on the ZipTip resin and washed with TFA 0.1%; finally, the sample was eluted from the resin using 2 \( \mu l \) of a saturated solution of sinapinic acid in 30/70 ACN/0.1% TFA and spotted directly on MTP ground steel 384 (Bruker Daltonics) and subjected to MALDI-TOF MS acquisition (Birocchi et al., 2006).

\subsection*{2.2.2 Data acquisition}

Each matrix droplet was individually analyzed using a MALDI time-of-flight Bruker Reflex IV mass spectrometer, equipped with a nitrogen laser (337 nm), used in the linear mode under delayed extraction conditions (400 ns). The ion source and flight tube were evacuated by turbo pumps to a pressure lower than 6 \( \times 10^{-7} \) mbar. The laser spot was 57 \( \mu m \) high and 32 \( \mu m \) large. Data were collected using an Accelerating Voltage of 25 kV, Ion Source1 20 kV, Ion Source2 17 kV, Lens 9.60 kV. Spectra were collected for a mass range of 5–20 kDa at 1 GHz. The laser power was modulated between 20% and 40% in order to obtain less than 1000 ion counts for single acquisition run. Every single acquisition run was composed by 100 laser pulses at 5 Hz; multiple additions of single position acquisition run were employed to obtain a minimal spectrum intensity scale of 5 \( \times 10^3 \) ion counts. The resulting mass spectra showed a mass accuracy of 40 p.p.m. on average with peaks at 800/1200 FWHM. MALDI-TOF mass spectra from serum and plasma samples are illustrated in Figure 2.

\section*{2.3 ICA decomposition}

The generic data matrix \( X \), arranged with the spectrum IDs in rows and the intensities corresponding to the m/z values in columns, is the only necessary input for ICA. In turn, the results of the ICA decomposition are two matrices: the first is the matrix \( S \) of IC waveforms with the component IDs in rows and the intensities corresponding to the m/z values in columns; the second is the matrix \( A \) of IC amplitudes with the component IDs in columns and the spectrum weights in rows.

We used the FastICA algorithm (Hyvärinen, 1999) for the data decomposition into \( n \) ICs. We downloaded a free copy of the FastICA
gradient method, the fixed-point learning rule is used (Hyvärinen and Oja, 1997). During the estimate of the $n$-dimensional vector $w$, the learning rule will stop at a fixed point, for which $|n^T(hw - h - 1)|$ is sufficiently close to unity, and the linear combination $w^T X$ will be one of the required ICs.

After that one ICA basis vector is obtained, other ICA basis vectors are estimated by sequentially finding new basis vectors onto the subspace orthogonal to the one covered by the previous ones (Hyvärinen and Oja, 1997). The vectors obtained in this way are used to create the matrix $W$. This iterative procedure is stopped until convergence, or when the total number of the required ICs is achieved.

Since no exact assumption could be done regarding either the expected power or the morphology of signals and noise, the larger number suitable for the ICA model was chosen, i.e. the same number $n$ of the mass spectra.

### 2.4 Peak detection

The presence or the absence of peaks allowed to differentiate the signal components and the artifactual ones. As a consequence, a peak-picking algorithm was necessary for the identification of the protein signal profiles; to this purpose, the LIMPIC algorithm (Mantini et al., 2007) was run for each IC using a variable noise threshold set to 10σ. The LIMPIC method consists of signal smoothing, baseline subtraction, and peak-picking.

#### 2.4.1 Smoothing, baseline subtraction, measurement of noise

The smoothing was performed using a Kaiser filter, with a smoothing factor $p$ properly set in order to cover a range of 5 Da. The baseline drift $\epsilon$ was locally estimated from signal blocks having width of 150 Da; for each of them, the average intensity was calculated, so that a vector $w = [w_1, w_2, ... , w_N]$ of amplitude values was generated. Then, $w$ was associated to the vector $b = [b_1, b_2, ... , b_N]$ of the $m/z$ values corresponding to the central point of each interval. The values $w_k$, with $k = 1, ..., N$, characterized by rapid intensity variations were considered to be out of the baseline, hence they were disregarded. The baseline drift $\epsilon$ was estimated from the remaining points $(b_k, w_k)$, with $k = 1, ..., L$, by means of a linear interpolation, and then it was removed from the spectrum. The processed spectrum was used for the estimate of the residual noise level $\sigma$. This was calculated using the SD $g_0$ of the values included in same blocks considered for the baseline reconstruction; a polynomial interpolation of the points $(b_k, g_k)$, with $k = 1, ..., L$ were used to obtain $\sigma$.

#### 2.4.2 Peak-picking

A peak list was created after the peak-picking phase by finding the local maxima: if the point intensity was the highest among its nearest 3 points, a peak was detected in that position (Yasui et al., 2003); for our data, the parameter $f$ was chosen equal to 2, in order to cover a range of 0.5 Da. The peaks with intensity lower than 10σ were then eliminated from the peak list.

### 2.5 Biomarker identification

In order to obtain protein profiles from the experimental MALDI-TOF MS data acquired from serum and plasma, FastICA algorithm was run using $X_{ser}$ and $X_{plas}$, respectively. Conversely, the use of $X_{tot}$ allowed the simultaneous estimate of protein signal profiles in the two groups with similar or different abundance. After the separation of the ICs, the latter was statistically analyzed, in order to find those components that showed significant difference of protein concentration among serum and plasma. To this purpose, the Mann–Whitney U-test (Mann and Whitney, 1947) was performed for each IC, using the values contained in the columns of the matrix $A$. The Mann–Whitney U-test was chosen because it is not based on the assumption of that the two distributions are Gaussian, and it is concurrently able to assess the alternative hypothesis that they are statistically different. The Benjamini–Hochberg
method was used for multiple testing corrections (Benjamini and Hochberg, 1995). The test was considered significant at a chosen significance level ($\alpha = 0.05$). This approach was intended for biomarker discovery: when analyzing clinical samples from healthy and diseased populations, the protein signal profiles with imbalanced intensity across the two groups of mass spectra could be considered to be protein biomarkers.

### Results

Four IC waveforms separated with FastICA from the synthetic dataset are shown in Figure 3, along with a fraction of the mass spectrum that is included for comparison. Therefore, the correspondence of the peaks in the IC waveforms with the ones in the mass spectrum can be assessed; it is also evident that even the outlier peak at 18 kDa is separated as a component. The component amplitudes across the mass spectra can be observed from Figure 4; this information is provided in form of bar plot for five ICs; ICs #4 and #25 are single-peak components, and IC #13 is a double-peak component; they have variable and randomly distributed amplitudes across the mass spectra. Conversely, IC #28 can be directly identified from the associated bar plot the as the signal corresponding to the outlier peak, also shown in Figure 3. IC #43 is an example of noise component, which is separated by FastICA because it is statistically independent from the protein signal profiles. The noise components can be discriminated from the signal components by the LIMPIC algorithm for peak detection, since the noise components do not contain any peak. Therefore, the peak-picking method applied to the separated ICs allowed to perfectly detect the peaks that were present in the simulated dataset.

The ICA method was also tested on experimental data: FastICA was run with the serum mass spectra, with the plasma mass spectra, and then with all mass spectra. The outcomes in terms of detected peaks and hit-rate, defined as the ratio between the number of peaks using multi-subject data and the average number of peaks detected in the single spectra (Mantini et al., 2007), were compared with the ones of the in-house LIMPIC algorithm, and of the commercial algorithms APEX and CENTROID (Table 1). The hit-rate of the ICA method is always equal to unity, because no false positives are detected with this approach; on the other hand, the peak-picking performance is quite poor when using the separate serum and plasma groups, and is similar to those of LIMPIC only for the joint dataset. APEX and CENTROID present a larger number of peaks, but their hit-rate was lower than that of LIMPIC.

The ICs separated from the mixed dataset were qualitatively and quantitatively analyzed: six of them are illustrated in Figure 5. IC #1 is the component with the largest power, and has similar intensities in the serum and plasma groups; IC #4 has the characteristics of a signal with an outlier peak; IC #19 has all amplitudes close to unity, because it corresponds to the
myoglobin protein (H\(^+\) = 16 952.25 Da) that is used for the signal calibration. ICs #35 and #45 are double-peak components, for which the Mann–Whitney U-test is significant at the chosen significance level (P < 0.05, corrected). IC #53 seems to be a biological artifact: it has larger amplitudes for the serum group than for the plasma group (P < 0.001), but it is classified as a disturbance, because no peak above the noise threshold is detected in the associated waveform. The P-values about the protein ICs that were significantly different between the two sample groups are provided in Table 2. Specifically, the peaks that are associated in the literature with apolipoproteins CI, CII, CIII (Bondarenko et al., 1999) were observed to be differentially expressed in plasma and sera, in accordance with previous findings on human body fluids in presence of inflammatory auto-immune diseases (Hortin, 2006).

4 DISCUSSION
4.1 ICA decomposition
ICA has been extensively used for signal extraction tasks in the field of biomedical signal processing (James and Hesse, 2005). The separation of underlying signals that can be performed with this technique provides a tool for high-dimensional spectrometric data processing, where the signals of interest are generally contaminated by artifacts of both biological and physical origin: ICA has proved to extract reliable protein profiles using multi-subject MALDI-TOF mass spectra acquired in the same m/z range, which have been previously calibrated and normalized. It successfully separated underlying signals contained in the mass spectra; in particular, background noise and outlier peaks could be identified, and the real protein signals globally showed the same peaks contained into the mass spectra, with an increased signal-to-noise ratio (SNR). For this reason, it can be integrated with existing computational methods for peak detection, and can enhance their effectiveness.

ICA has also an intrinsic limitation: the optimal number of independent signals mixed in the mass spectra is unknown. Unless a dimension reduction procedure is previously performed (Yang and Wang, 1999), the typical ICA model assumes that the number of underlying signals is at most equal to the number of mass spectra (Hyvärinen et al., 2001; Stone, 2004): in this case, the ICA decomposition generally

---

**Table 1. Performance comparison of peak identification algorithms**

<table>
<thead>
<tr>
<th></th>
<th>ICA + LIMPIC</th>
<th>LIMPIC</th>
<th>APEX</th>
<th>CENTROID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serum Peaks</td>
<td>52</td>
<td>67</td>
<td>93</td>
<td>84</td>
</tr>
<tr>
<td>hit-rate</td>
<td>1</td>
<td>0.42</td>
<td>0.32</td>
<td>0.30</td>
</tr>
<tr>
<td>Plasma Peaks</td>
<td>49</td>
<td>84</td>
<td>121</td>
<td>113</td>
</tr>
<tr>
<td>hit-rate</td>
<td>1</td>
<td>0.40</td>
<td>0.30</td>
<td>0.25</td>
</tr>
<tr>
<td>Serum and plasma</td>
<td>89</td>
<td>88</td>
<td>143</td>
<td>128</td>
</tr>
<tr>
<td>hit-rate</td>
<td>1</td>
<td>0.47</td>
<td>0.41</td>
<td>0.35</td>
</tr>
</tbody>
</table>

ICA + LIMPIC, LIMPIC with variable noise threshold at 10\(r\), APEX and CENTROID with SNR = 10 were evaluated at for the serum dataset, for the plasma dataset and for the joint dataset. The number of detected peaks and the hit-rate were calculated for each peak-picking method.

**Table 2. List of the protein ICs which are characterized by significant statistical difference among serum and plasma groups**

<table>
<thead>
<tr>
<th>IC label</th>
<th>(m/z) (Da)</th>
<th>Serum intensity</th>
<th>Plasma intensity</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>IC #13</td>
<td>9139</td>
<td>0.503 ± 0.370</td>
<td>0.863 ± 0.713</td>
<td>0.048</td>
</tr>
<tr>
<td>IC #17</td>
<td>9715</td>
<td>0.986 ± 0.695</td>
<td>0.534 ± 0.658</td>
<td>0.042</td>
</tr>
<tr>
<td>IC #20</td>
<td>6434, 6633</td>
<td>0.566 ± 0.413</td>
<td>1.298 ± 1.038</td>
<td>0.047</td>
</tr>
<tr>
<td>IC #22</td>
<td>8917</td>
<td>0.634 ± 1.029</td>
<td>0.225 ± 0.338</td>
<td>0.008</td>
</tr>
<tr>
<td>IC #23</td>
<td>9127</td>
<td>0.331 ± 0.271</td>
<td>0.552 ± 0.416</td>
<td>0.024</td>
</tr>
<tr>
<td>IC #25</td>
<td>9629</td>
<td>0.367 ± 0.222</td>
<td>0.231 ± 0.202</td>
<td>0.044</td>
</tr>
<tr>
<td>IC #30</td>
<td>6439, 6636</td>
<td>0.246 ± 0.218</td>
<td>0.695 ± 0.621</td>
<td>0.007</td>
</tr>
<tr>
<td>IC #35</td>
<td>6430, 6629</td>
<td>0.281 ± 0.287</td>
<td>0.614 ± 0.494</td>
<td>0.029</td>
</tr>
<tr>
<td>IC #42</td>
<td>6451, 6648</td>
<td>0.018 ± 0.027</td>
<td>0.152 ± 0.247</td>
<td>0.038</td>
</tr>
<tr>
<td>IC #45</td>
<td>6881, 13762</td>
<td>0.158 ± 0.177</td>
<td>0.004 ± 0.024</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>IC #51</td>
<td>6941, 13882</td>
<td>0.101 ± 0.086</td>
<td>0.023 ± 0.047</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>IC #59</td>
<td>5601, 5757</td>
<td>0.078 ± 0.107</td>
<td>0.035 ± 0.046</td>
<td>0.049</td>
</tr>
<tr>
<td>IC #60</td>
<td>5069</td>
<td>0.092 ± 0.073</td>
<td>0.043 ± 0.031</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Average value and SD of the spectrum intensities, as well as the probability obtained with the Mann–Whitney U-test, are provided. For each IC, the \(m/z\) of the detected peaks is shown.

---

Fig. 5. Samples of 6 out of the 60 ICs separated from the serum and plasma MALDI-TOF mass spectra. For each IC, the plot of the waveform is shown; in addition, two bar plots containing the corresponding amplitudes for the serum and plasma mass spectra, respectively, are provided.
could produce a residual number of non-relevant ICs, depending on the correct number of protein profiles. As a consequence, the identification of the ICs becomes a difficult task.

4.2 Peak detection
The presence of peaks in the ICs can be considered as the main indicator for the identification of protein signal profiles. As a result, the integration of a peak-picking technique is required, not only for the characterization of the peaks contained in the signals produced by ICA, but also for the classification of the ICs into two different groups: the protein signals and the artifacts. By setting a conservative noise threshold equal to 10σ, we are able to automatically discriminate the disturbances. Conversely, 1–3 peaks are present in the ICs corresponding to real protein signals. With regard to the IC waveform, it is worth noting that the proteins associated to the peaks in a single component can be assumed to be dependent, because the increase or the decrease of their intensity is proportional across the analyzed mass spectra.

ICA provides signals that are present with different abundance in all spectra, and the peaks are detected in protein signals, which are largely uncontaminated by noise. As a result, we can assume that there are no false positives in the peak-picking procedure, and the hit-rate is always equal to unity. By contrast, the number of detected peaks is generally lower than that of other methods when using a limited number of mass spectra from serum or plasma (Table 1); the number of peaks is similar to that of LIMPIC only for the joint dataset corresponding to serum and plasma samples. We can infer that the main limitation of the proposed system is the requirement of a large number of mass spectra for achieving a sufficient number of detected peaks. This finding is consistent with the results obtained from the application of the ICA algorithms in other research fields (Hyvärinen et al., 2001; James and Hesse, 2005).

4.3 Biomarker discovery
The opportunity of isolating signal components, combined with the availability of amplitude values associated with single mass spectra, permits to use ICA for biomarker discovery. The non-parametric Mann–Whitney U-test has been used for assessing if the two distributions come from the same population. It does not only for the characterization of the peaks contained in the signals produced by ICA, but also for the classification of the ICs into two different groups: the protein signals and the artifacts. By setting a conservative noise threshold equal to 10σ, we are able to automatically discriminate the disturbances. Conversely, 1–3 peaks are present in the ICs corresponding to real protein signals. With regard to the IC waveform, it is worth noting that the proteins associated to the peaks in a single component can be assumed to be dependent, because the increase or the decrease of their intensity is proportional across the analyzed mass spectra.

ICA provides signals that are present with different abundance in all spectra, and the peaks are detected in protein signals, which are largely uncontaminated by noise. As a result, we can assume that there are no false positives in the peak-picking procedure, and the hit-rate is always equal to unity. By contrast, the number of detected peaks is generally lower than that of other methods when using a limited number of mass spectra from serum or plasma (Table 1); the number of peaks is similar to that of LIMPIC only for the joint dataset corresponding to serum and plasma samples. We can infer that the main limitation of the proposed system is the requirement of a large number of mass spectra for achieving a sufficient number of detected peaks. This finding is consistent with the results obtained from the application of the ICA algorithms in other research fields (Hyvärinen et al., 2001; James and Hesse, 2005).

5 CONCLUSION
It is the first time that ICA is used for the decomposition of protein signal profiles from MALDI-TOF mass spectra. In this case, data analysis is particularly difficult, because weak protein signals, represented by the true peaks in the acquired spectra, are generally contaminated by noise of biological and physical origin. A system for the reliable separation of protein signals has been developed, and used for improving classical peak detection methods. Although further verification on a larger clinical population is required, our findings suggest that the proposed system could be more effective than other open-source and commercial algorithms, when a sufficiently large number of mass spectra are available for analysis. In addition, the quantitative information on the peak intensity extracted with ICA could be used for the recognition of significant protein profiles by means of advanced statistical tests. From the perspective of a routine clinical employment of MALDI-TOF mass spectrometry, the proposed system might represent a further step toward the optimization of a standardized procedure for the automatic recognition of disease-state biomarkers.

ACKNOWLEDGEMENTS
We are particularly grateful to Enzo Ballone and Gian Luca Romani for continuous support and scientific discussion.

Conflict of Interest: none declared.

REFERENCES


