Mass spectrometry data analysis: it's all in the preprocessing

Rubén Armañanzas[1], Yvan Saeys[2,3], Iñaki Inza[1], Miguel García-Torres[4], Yves Van de Peer[2,3], Concha Bielza[5], Pedro Larrañaga[5]

[1]Department of Computer Science and Artificial Intelligence
University of the Basque Country
{ruben,inza}@si.ehu.es

[2]Department of Plant Systems Biology, VIB
[3]Department of Molecular Genetics
Ghent University
{yvan.saeys,yves.vandepeer}@psb.ugent.be

[4]Laboratory for Space Astrophysics and Theoretical Physics
National Institute for Aerospace Technology
mgarciat@laeff.inta.es

[5]Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid
{mcbielza,pedro.larranaga}@fi.upm.es

1 Introduction

The identification of predictive biomarkers has gained much popularity in the area of the mass spectrometry (MS) data analysis. However, the path from the raw signal to the end outcome of the machine learning analysis is full of caveats and pitfalls.

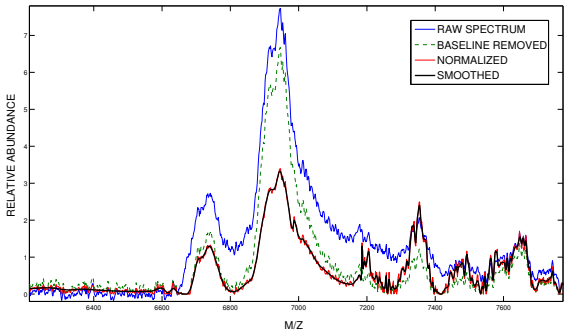
In this work, we present a workflow pipeline that ranges from parsing the raw MS data to the final generation of candidate features for proteomic profiling studies.

2 Preprocessing protocol

The most accepted formulation of a generic MS signal is

$$f(t) = B(t) + N \cdot S(t) + \varepsilon(t),$$

where $f(t)$ corresponds to the observed signal, $B(t)$ is an additive baseline component visually identifiable, and $S(t)$ corresponds to the expected true signal, which is modified by a normalization factor N. The last element, $\varepsilon(t)$, is an unknown noise component which groups the remaining variations.



Graphical example of the baseline removal, normalization and wave smoothing steps.

Baseline removal

The low range masses of a spectrum always presents an amplification of the intensity values due to chemical noise. The true signal must be estimated and then the difference between the observed and estimated signal should be removed. We propose the use of the *top-hat morphological operator* [Soille, 1999], since it needs little computation time and has proven its merits in the image analysis domain.

Normalization

MS spectra of similar samples are not always similarly quantified. Through normalization, all the spectra intensities are parsed into the same intensity ranges. We propose the use of local estimators over m/z windows [Meuleman *et al.*, 2008] with rescaling to the median value of the AUC (the TIC value).

Signal smoothing

Signal smoothing aims to alleviate the low resolution noise $\varepsilon(t)$. The technique most often used for this purpose is the wavelet denoising proposed by [Coombes *et al.*, 2007]. It makes use of the undecimated wavelet transformation to estimate the wavelet coefficients, which are then used to denoise the signal and obtain a smoothed signal.

Peak detection

Peak detection consists of distinguishing a m/z position corresponding to a true peak in the spectrum. At this stage, the aim is to screen a big number of peaks that later could be grouped into peakbins. To consider a point p as a peak, there must exist a point l (respectively r) on its left (respectively right), before the previous (next) peak that accomplishes two conditions: first, the value of the candidate point p must be higher than a sensitivity threshold and, second, the candidate point p must have a SNR $\geqslant 3$ within the intensity window framed by l and r. The algorithm estimates the SNR value as the ratio between the point's height and the median absolute deviation (MAD) in its $[l, r]$ window.

Peak assembly & quantification

The peak agglomeration tries to match the similar peaks detected over all the spectra. We propose the assembling of peak bins of different widths by means of the Pearson's linear correlation. The outcome signal values are quantified as the maximum value found on each bin.

3 Results

Four MS datasets were preprocessed (two SELDI and two MALDI-TOF) and the predictive accuracy of the outcome peakbins estimated on a 10-fold cross validation process. The classification paradigm was a continuous naïve Bayes classifier (normal distribution assumption). The process is repeated removing one of the preprocessing steps (where possible) and the results are compared by a t-test at three confidence levels.

	OVA	TOX	DGB	HCC
Samples	200	62	128	150
Phenotypes	(121/79)	(28/34)	(25/78/25)	(78/28)
m/z values	45200	45200	16075	36802
Full				
|bins|	1693±43.8	2686±136.1	176±8	234±10.1
Accuracy	94.50±7.62	71.71±16.72	81.24±7.32	88.65±4.53
BSRemoval				
|bins|	2500±69$^{\blacktriangle}$	N/A	193.6±3.71$^{\blacktriangle}$	650±28.55$^{\blacktriangle}$
Accuracy	93.50±3.35	–	75.28±16.91	65.38±13.23$^{\blacktriangle}$
Norm				
|bins|	668±12$^{\blacktriangle}$	2963±97.6$^{\blacktriangle}$	106±6.63$^{\blacktriangle}$	227±7.41$^{\diamond}$
Accuracy	82.44±8.74†	63.81±13.54	75.28±6.74$^{\diamond}$	76.03±17.81†
Smooth				
|bins|	N/A	N/A	181±4.53$^{\diamond}$	237±6.4
Accuracy	–	–	76.14±14.36	87.28±8.63
Assembly				
|bins|	6069±33.73$^{\blacktriangle}$	4051±89.8$^{\blacktriangle}$	452±11.21$^{\blacktriangle}$	1648±74.84$^{\blacktriangle}$
Accuracy	92.97±6.84	71.95±14.92	75.85±14.19	89.98±9.37

Significative differences at levels $^{\diamond}$ 0.90, † 0.95, and, $^{\blacktriangle}$ 0.99.

4 Conclusions

The above results illustrate the need for a complete preprocessing pipeline when dealing with MS datasets. Regarding the number of peakbins, some steps are especially mandatory, otherwise the final peakbin size exponentially grows and the problem becomes unaffordable. Also illustrative is the fact that the absence of baseline removal drastically reduces the predictive accuracy from 88% to 65% in one dataset.