Electronic Supplementary Material (ESI) for Molecular BioSystems

TOFwave: Reproducibility in Biomarker Discovery from time-of-flight Mass Spectrometry Data

Electronic Supplementary Information (ESI)

Marco Chierici[1], Davide Albanese[1], Pietro Franceschi[2], and Cesare Furlanello[1]

[1]Fondazione Bruno Kessler, Trento, Italy
[2]Biostatistics and Data Management - IASMA Research and Innovation Centre, S. Michele all'Adige (TN), Italy

1 Introduction to TOFwave

This section is intended to be a short description of TOFwave usage. Detailed documentation, information and installation instructions can be found on `http://mlpy.sourceforge.net/tofwave`. TOFwave is free software and it can be downloaded on the project homepage.

TOFwave for Linux and Mac OS X is composed by three Python scripts, while for Windows (XP, Vista, 7) it is composed by three shell executables:

tofwave-viewer(.exe) is an utility to visualize spectra;

tofwave-cwtpre(.exe) performs the CWT-based preprocessing and outputs preprocessed spectra;

tofwave-analysis(.exe) performs feature extraction and modeling given the preprocessed spectra.

The `tofwave-cwtpre` parameters are bounds $l = (l_{left}, l_{right})$ [Da] and $h = (h_{left}, h_{right})$ [Da], defining widths W_l, W_h in the low and high mass regions as described in the main paper. Due to instrumental reasons, small variations in the choice of bounds l and h have negligible or no impact on the preprocessing and modeling phases.
Once the parameters l and h are chosen (for example, by visualizing spectra through `tofwave-viewer`), the procedure consists in running `tofwave-cwtpre`, followed by `tofwave-analysis` given a signal-to-noise threshold SNR_t. For example, assuming $l = (100, 110)$, $h = (2000, 2050)$ (and thus, $W_l = 10$ and $W_h = 50$), and $SNR_t = 3$:

```
$ tofwave-cwtpre inputfile.txt outputfile_cwtpre.txt 100,110 2000,2050
$ tofwave-analysis outputfile_cwtpre.txt results.txt -t 3 -p 30
```

where `inputfile.txt` contains raw TOF-MS spectra, `outputfile_cwtpre.txt` stores the preprocessed spectra, `results.txt` reports average test errors with confidence intervals and ranked peak list with average positions, and `-t 3` defines the SNR threshold $SNR_t = 3$. The option `-p 30` forces `tofwave-analysis` to perform modeling only on the first top-30 ranked peaks; this is convenient when one is interested in exploring only the first most discriminating peaks.

2 Dataset description

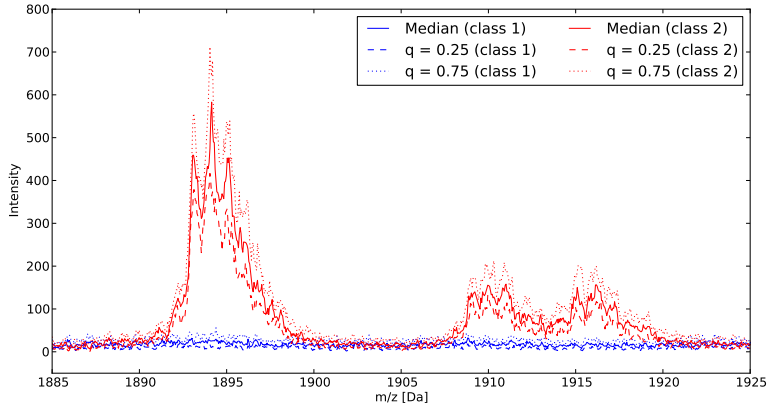
The software was evaluated on two metabolomic and one proteomic datasets, detailed in the following. An overview is reported in Table S1.

Metabolomic datasets. Two datasets were designed to test the performance of the proposed pipeline in the mass range typical of metabolic profiling (datasets A & B). All the spectra were acquired with a Bruker Ultraflex MALDI TOF TOF. All the standard mixtures were mixed with matrix solution, manually spotted and air dried.
In Dataset A, a methanol solution of the metabolites Quercetin 3,4'-diglucoside (monoisotopic mass: 626.148 Da) and t-resveratrol (monoisotopic mass: 228.243 Da) was spiked in the standard peptide mixture used for TOF calibration (Bruker). Sanguiin H6 (monoisotopic mass 1870.158 Da) (Gasperotti *et al.*, 2010) was added to the mixture in half of the samples. Spectra were acquired in linear positive ion mode by using DHB as matrix (Reed *et al.*, 2005). The solution was spotted 40 times (20 with Sanguiin and 20 without). A spectrum was obtained by averaging 500 laser shots per spot.
Dataset B probes the low end of the mass interval, where matrix peaks are extensively present. Methionine (monoisotopic mass 149.051 Da) was spiked in a methanol solution of 10 primary metabolites. Solution was spotted 41 times (21 with Methionine and 20 without) by using HCCA as matrix. Each spectrum was acquired in positive reflectron mode by summing 300 laser shots.
Datasets A and B were exported to comma separated format for data analysis and are available for download on the software homepage. Spectra of both datasets in the regions containing discriminating peaks are shown in Fig. S1.

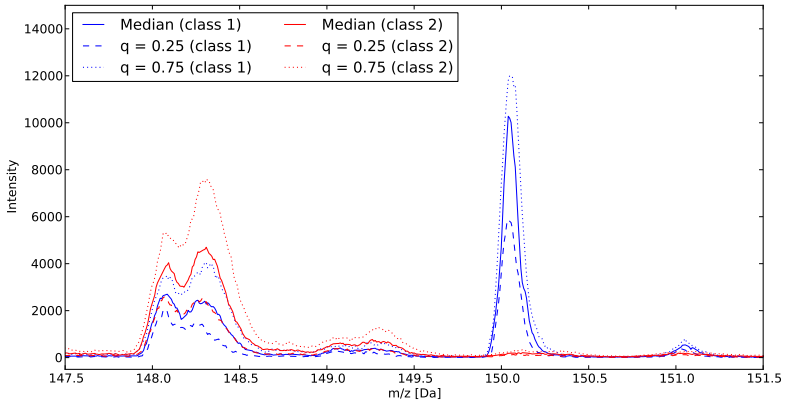
Proteomic dataset. A proteomic pattern dataset of spectra from 93 ovarian cancer and 77 control patients was considered (dataset C); details are given in the study by Wu and colleagues (Wu *et al.*, 2006). The dataset was produced at the Keck Laboratory with a Micromass M@LDI-L/R instrument, which can be used in either linear or reflectron modes of operation. We considered only the linear mode data, with masses in the 3450–28000 Da range. While no m/z locations that perfectly distinguish between cases and controls have been found in the literature (Wu *et al.*, 2006), we adopted this dataset because of its complementary range with respect to datasets A and B.

Dataset	Range [Da]		# Samples	
	min	max	Class 1	Class 2
A (metabolomic)	60	4392	20	20
B (metabolomic)	20	1000	21	20
C (proteomic)	3450	28000	93	77

Table S1: Overview of the three datasets. "Class 1" identifies either the class containing Sanguiin or Methionine in the mixture (for datasets A and B), or cancer patients (for dataset C).



(a)



(b)

Figure S1: Exploration of spectra in regions where Sanguiin (S1a) or Methionine (S1b) peaks are expected, by samples class. Continuous line: median of the intensities; dashed line: lower quartile of the intensities ($q = 0.25$); dotted line: upper quartile of the intensities ($q = 0.75$).

3 Results

The following scripts list the TOFwave commands used in the three experiments. We assume that the input files for datasets A, B, and C are named `sanguiin.txt`, `methionine.txt`, and `ovarian.txt`, respectively. The $ symbol identifies the command prompt.

Dataset A $W_l = 0.4$ (evaluated at 629.2 Da), $W_h = 4$ at 3148 Da, $\mathrm{SNR}_t = 2, 4$.

```
$ tofwave-cwtpre sanguiin.txt sanguiin_cwtpre.txt 629,629.4 3146,3150
$ tofwave-analysis sanguiin_cwtpre.txt results_sanguiin_2.txt -t 2 -p 30
$ tofwave-analysis sanguiin_cwtpre.txt results_sanguiin_4.txt -t 4 -p 30
```

Dataset B $W_l = 0.1$ at 171.55 Da, $W_h = 0.2$ at 644.6 Da, $\mathrm{SNR}_t = 3$.

```
$ tofwave-cwtpre methionine.txt methionine_cwtpre.txt 171.5,171.6 \
  644.5,644.7
$ tofwave-analysis methionine_cwtpre.txt results_methionine_3.txt -t 3 -p 30
```

Dataset C $W_l = 40$ at 5000 Da, $W_h = 150$ at 8565 Da, $\mathrm{SNR}_t = 1.5$.

```
$ tofwave-cwtpre ovarian.txt ovarian_cwtpre.txt 4980,5020 8490,8640
$ tofwave-analysis ovarian_cwtpre.txt results_ovarian_1.5.txt -t 1.5 -p 30
```

The "random labels" experiments are run by adding option `-r` to the `tofwave-analysis` commands above:

```
$ tofwave-analysis -r sanguiin_cwtpre.txt results_sanguiin_2_random.txt -t 2 -p 30
$ tofwave-analysis -r sanguiin_cwtpre.txt results_sanguiin_4_random.txt -t 4 -p 30
$ tofwave-analysis -r methionine_cwtpre.txt results_methionine_3_random.txt -t 3 -p 30
$ tofwave-analysis -r ovarian_cwtpre.txt results_ovarian_1.5_random.txt -t 1.5 -p 30
```

3.1 Dataset A

Table S2: Dataset A, $SNR_t = 2$, "random labels": average test error (ATE) with 97.5% bootstrap confidence interval (ATE_{min}, ATE_{max}). n: number of peaks used in the model.

n	ATE	ATE_{min}	ATE_{max}
1	0.489	0.481	0.496
2	0.480	0.473	0.488
3	0.472	0.463	0.480
4	0.466	0.458	0.475
5	0.472	0.464	0.481
6	0.463	0.455	0.472
7	0.463	0.454	0.472
8	0.458	0.449	0.468
9	0.455	0.445	0.465
10	0.454	0.445	0.463

Table S3: Dataset A, $SNR_t = 4$: average test error (ATE) with 97.5% bootstrap c.i. (ATE_{min}, ATE_{max}) on the top-10 ranked peaks (m/z, cluster centroids) with their associated cluster bounds (C_{min}, C_{max}) and average positions. n: number of peaks used in the model (for ATE) or peak rank (for m/z, cluster bounds and average position).

n	ATE	ATE_{min}	ATE_{max}	m/z [Da]	C_{min} [Da]	C_{max} [Da]	Avg. pos.
1	0.020	0.012	0.028	1910	1909.3	1911.6	1.00
2	0.004	0.000	0.012	1916	1915.2	1917.1	2.29
3	0.000	0.000	0.000	1894	1893.0	1896.1	2.71
4	0.016	0.008	0.025	277.3	277.00	277.69	6.75
5	0.034	0.021	0.048	304.4	304.08	304.73	7.42
6	0.042	0.029	0.058	87.7	87.29	87.86	9.51
7	0.038	0.027	0.051	1932	1931.0	1932.6	9.95
8	0.047	0.033	0.064	768.5	768.14	768.96	11.66
9	0.040	0.028	0.055	685.4	684.96	685.82	13.30
10	0.043	0.029	0.057	703.3	702.73	703.72	15.14

Table S4: Dataset A, $SNR_t = 4$, "random labels": average test error (ATE) with 97.5% bootstrap c.i. (ATE_{min}, ATE_{max}). n: number of peaks used in the model.

n	ATE	ATE_{min}	ATE_{max}
1	0.480	0.472	0.488
2	0.472	0.464	0.480
3	0.469	0.461	0.477
4	0.471	0.462	0.481
5	0.464	0.454	0.473
6	0.461	0.452	0.469
7	0.466	0.457	0.475
8	0.461	0.452	0.472
9	0.462	0.452	0.473
10	0.459	0.450	0.469

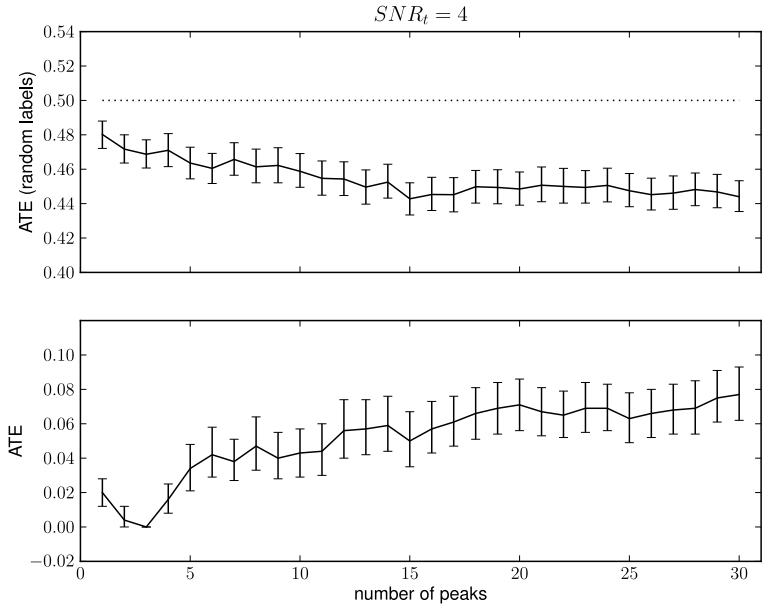


Figure S2: Dataset A. Average Test Error (ATE) curves with 97.5% bootstrap confidence intervals for a complete preprocessing and classification experiment, with $SNR_t = 4$. Horizontal dotted line indicates the "no-information error rate", here defined as the ratio between the smallest class and the whole dataset size, corresponding to the error reached by classifying all samples as belonging to the most populous class.

3.2 Dataset B

Table S5: Dataset B, $SNR_t = 3$: average test error (ATE) with 97.5% bootstrap c.i. (ATE_{min}, ATE_{max}) on the top-10 ranked peaks (m/z, cluster centroids) with their associated cluster bounds (C_{min}, C_{max}) and average positions. n: number of peaks used in the model (for ATE) or peak rank (for m/z, cluster bounds and average position).

n	ATE	ATE_{min}	ATE_{max}	m/z [Da]	C_{min} [Da]	C_{max} [Da]	Avg. pos.
1	0.003	0.000	0.007	151.1	151.02	151.18	1.00
2	0.007	0.001	0.014	150.1	149.99	150.18	2.00
3	0.011	0.003	0.021	152.1	152.01	152.16	3.33
4	0.024	0.012	0.037	85.96	85.925	86.016	4.34
5	0.017	0.008	0.027	89.5	89.44	89.57	6.12
6	0.032	0.021	0.043	321.2	321.12	321.32	6.20
7	0.030	0.020	0.040	135.42	135.405	135.435	7.34
8	0.044	0.033	0.055	328.2	328.19	328.32	9.32
9	0.048	0.036	0.061	275.2	275.16	275.29	11.62
10	0.052	0.040	0.065	354.4	354.27	354.45	11.68

Table S6: Dataset B, $SNR_t = 3$, "random labels": average test error (ATE) with 97.5% bootstrap c.i. (ATE_{min}, ATE_{max}). n: number of peaks used in the model.

n	ATE	ATE_{min}	ATE_{max}
1	0.518	0.513	0.523
2	0.520	0.514	0.527
3	0.528	0.522	0.535
4	0.525	0.518	0.533
5	0.526	0.518	0.534
6	0.524	0.516	0.531
7	0.522	0.514	0.530
8	0.526	0.518	0.534
9	0.522	0.514	0.531
10	0.525	0.516	0.533

3.3 Dataset C

ATE = 36.6% is obtained on dataset C with the first 10 ranked peak locations. The second top-ranked feature at $m/z = 3942$ Da is consistent with (Wu *et al.*, 2006); on the contrary, the peak at approximately 7935 Da shown in Fig. 6 in (Wu *et al.*, 2006) was ranked as the 12^{th} top feature, with average position 18.1 (Table S7).

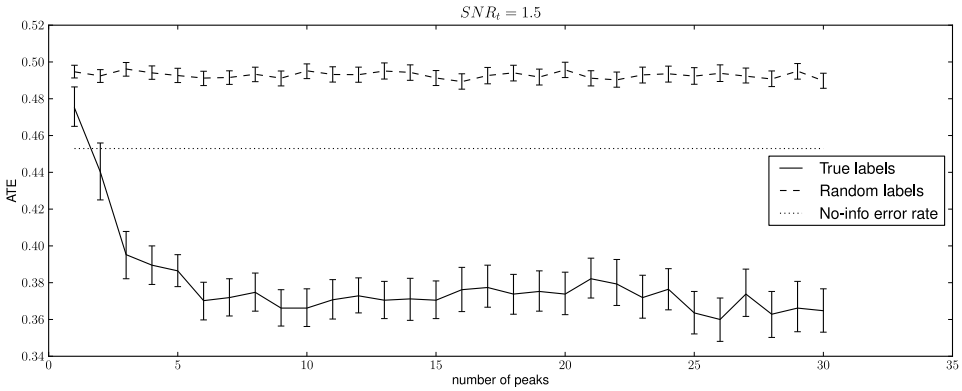


Figure S3: Dataset C. Average Test Error (ATE) curves with 97.5% bootstrap c.i. for a complete preprocessing and classification experiment, with $SNR_t = 1.5$.

Table S7: Dataset C, $SNR_t = 1.5$: average test error (ATE) with 97.5% bootstrap c.i. (ATE_{min}, ATE_{max}) on the top-15 ranked peaks (m/z, cluster centroids) with their associated cluster bounds (C_{min}, C_{max}) and average positions. n: number of peaks used in the model (for ATE) or peak rank (for m/z, cluster bounds and average position).

n	ATE	ATE_{min}	ATE_{max}	m/z [Da]	C_{min} [Da]	C_{max} [Da]	Avg. pos.
1	0.475	0.465	0.486	4060	4032.3	4108.2	2.29
2	0.440	0.425	0.456	3942	3897.6	3971.1	2.76
3	0.395	0.382	0.408	4205	4182.5	4212.5	5.45
4	0.390	0.379	0.400	8119	8085.1	8205.7	7.02
5	0.386	0.378	0.395	5897	5856.2	5940.3	9.17
6	0.370	0.360	0.380	6063	6060.1	6066.5	9.77
7	0.372	0.362	0.382	6430	6399.1	6478.5	15.08
8	0.375	0.365	0.385	9281	9176.0	9311.2	15.41
9	0.366	0.356	0.376	4776	4719.8	4805.1	15.87
10	0.366	0.356	0.377	4628	4593.0	4665.2	17.68
11	0.371	0.360	0.382	5341	5317.2	5354.4	17.76
12	0.373	0.364	0.383	7924	7910.1	7938.2	18.1
13	0.370	0.360	0.381	6214	6200.0	6218.1	18.11
14	0.371	0.360	0.382	6631	6559.4	6709.3	18.74
15	0.370	0.360	0.381	4459	4427.7	4480.9	20.03

Table S8: Dataset C, $SNR_t = 1.5$, "random labels": average test error (ATE) with 97.5% bootstrap c.i. (ATE_{min}, ATE_{max}). n: number of peaks used in the model.

n	ATE	ATE_{min}	ATE_{max}
1	0.495	0.491	0.498
2	0.492	0.489	0.496
3	0.496	0.492	0.500
4	0.494	0.491	0.498
5	0.493	0.489	0.497
6	0.491	0.487	0.495
7	0.492	0.488	0.495
8	0.493	0.489	0.497
9	0.491	0.487	0.495
10	0.495	0.491	0.499
11	0.493	0.489	0.497
12	0.493	0.489	0.497
13	0.495	0.491	0.499
14	0.494	0.490	0.498
15	0.491	0.487	0.495

References

Gasperotti, M., Masuero, D., Vrhovsek, U., Guella, G., and Mattivi, F. (2010). Profiling and accurate quantification of Rubus Ellagitannins and Ellagic acid conjugates using direct UPLC-Q-TOF HDMS and HPLS-DAD analysis. *Journal of agricultural and food chemistry*, **58**(8), 4602–4616.

Reed, J., Krueger, C., and Vestling, M. (2005). MALDI-TOF mass spectrometry of oligomeric food polyphenols. *Phytochemistry*, **66**(18), 2248–2263.

Wu, B., Abbot, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., and Zhao, H. (2006). Ovarian cancer classification based on mass spectrometry analysis of sera. *Cancer Informatics*, **2**, 123–132.