Molecular BioSystems

Cite this: *Mol. BioSyst.*, 2012, **8**, 2845–2849

www.rsc.org/molecularbiosystems

METHOD

TOFwave: reproducibility in biomarker discovery from time-of-flight mass spectrometry data†

Marco Chierici,‡*[a] Davide Albanese,‡[a] Pietro Franceschi[b] and Cesare Furlanello[a]

Received 6th June 2012, Accepted 23rd July 2012
DOI: 10.1039/c2mb25223f

Many are the sources of variability that can affect reproducibility of disease biomarkers from time-of-flight (TOF) Mass Spectrometry (MS) data. Here we present TOFwave, a complete software pipeline for TOF-MS biomarker identification, that limits the impact of parameter tuning along the whole chain of preprocessing and model selection modules. Peak profiles are obtained by a preprocessing based on Continuous Wavelet Transform (CWT), coupled with a machine learning protocol aimed at avoiding selection bias effects. Only two parameters (minimum peak width and a signal to noise cutoff) have to be explicitly set. The TOFwave pipeline is built on top of the mlpy Python package. Examples on Matrix-Assisted Laser Desorption and Ionization (MALDI) TOF datasets are presented. Software prototype, datasets and details to replicate results in this paper can be found at http://mlpy.sf.net/tofwave/.

Introduction

Mass spectrometry (MS) based profiling allows the construction of molecular snapshots of a biological system and constitutes an excellent ground for systemic approaches with implications on the diagnosis of human diseases. Body fluids like blood serum can be routinely used to generate MS profiles that can be analyzed to identify potential disease biomarkers, such as metabolites, peptides, individual proteins, or sets of interacting proteins.[1,2] Due to the growth of technologies for MS profiling, instrumental platforms are capable of acquiring high throughput data with both high mass resolution and stability. The full exploitation of this potential requires the development of dedicated pipelines with high confidence and reproducibility.

Reproducibility is still a major concern for the discovery of predictive biomarkers in MS clinical studies.[3–5] Excluding batch effects, a fair fraction of the known sources of variability is due to non-explicit visual tuning in the preprocessing phase, which may explain why it is so hard to compare different algorithms for peak detection and quantification.[5] For reproducibility, the number of parameters should be reduced to a minimum and possibly depend more on technical specifications of the MS platform and less on manual fine tuning. Moreover, for predictive modeling from MS data, it is urgent to adopt Data Analysis Plans (DAPs) that ensure a valid estimate of accuracy and avoid overfitting by careful use of cross-validation protocols, as shown for large scale microarray studies.[6]

Here we introduce the TOFwave pipeline to address both reproducibility and predictive modeling issues with one software environment for TOF-MS data. For preprocessing, we propose a Continuous Wavelet Transform (CWT) module.[7] The multi-scale nature of the CWT solves the issue of tuning TOF-MS peak matching, since it automatically deals with peak-width variation in different mass spectrum regions. For data analysis, the TOFwave pipeline can use generic machine learning workflows for multivariate classification or regression and for the identification of a ranked set of predictive biomarkers (*e.g.* a MS peak profile). In particular, we propose here a DAP built according to model development practices defined in regulatory initiatives, such as the U.S. Food and Drug Administration's projects for reliable biomarker identification.[6]

The system is provided as a modular software pipeline, whose elements are all available as methods from the mlpy Python package, an Open Source environment for data analysis and machine learning for high throughput data.[8]

Materials and methods

The workflow (Fig. 1) is derived from an analysis protocol previously proposed for proteomic profiling.[9] The preprocessing, feature extraction and machine learning steps are built on top of the modular mlpy library. Although the workflow is presented here for a typical classification task such as a case/control study, it can be adapted to multi-class and regression problems by exploiting algorithms implemented in mlpy, such as multi-class (kernel) Support Vector Machines (SVMs) and Support Vector Regression (SVR).[8] The key components of TOFwave are described in the following.

[a] *Fondazione Bruno Kessler, Trento, Italy. E-mail: chierici@fbk.eu*
[b] *Biostatistics and Data Management - IASMA Research and Innovation Centre, Fondazione E. Mach, S. Michele all'Adige, TN, Italy*
† Electronic supplementary information (ESI) available. See DOI: 10.1039/c2mb25223f
‡ These authors contributed equally to this work.

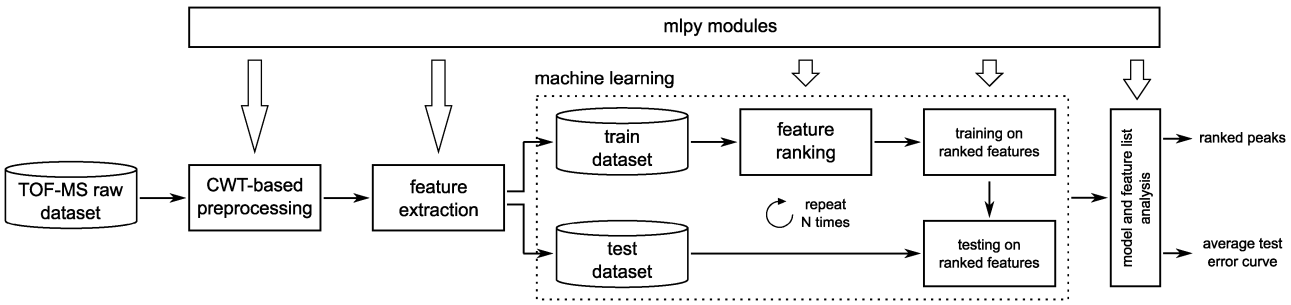

Fig. 1 The modular structure of the TOFwave workflow.

CWT-based preprocessing

Each raw spectrum first undergoes a de-noising and baseline removal phase, based on the CWT. In order to preserve the basic shape characteristics of TOF peaks, we chose the Mexican Hat mother wavelet.[10] De-noising and baseline removal steps are driven by Full Widths at Half Maximum Peak Height (FWHM) W_l and W_h, which represent the desired detail level by taking into account only spectral structures wid'er than W_l Da in regions with low mass-to-charge values, and structures wider than W_h Da in regions with high mass-to-charge values, respectively. A linear interpolation is performed in the range between the two parameters, obtaining $W(m/z)$.

Overall, the procedure consists of the following steps:

(1) each spectrum is transformed by CWT;

(2) de-noising is achieved by removing CWT coefficients that correspond to signal structures narrower than $W(m/z)$;

(3) baseline is corrected by removing CWT coefficients related to signal structures wider than $5W(m/z)$, a width that adequately represents the low frequency variation of the baseline in MALDI-TOF spectra;

(4) a de-noised and baseline-corrected version of the signal, $S(m/z)$, is reconstructed by inverse CWT (ICWT) on the updated coefficient matrix;

(5) the noise level N_l of the spectrum is estimated by:

(a) performing ICWT on the two lowest scales of the CWT coefficient matrix, obtaining $N(m/z)$;

(b) computing the 95-percentile on $N(m/z)$.

Each spectrum is then normalized by the estimated noise level, to obtain intensities in terms of signal-to-noise ratio $SNR = S(m/z)/N_l$. Fig. 2 shows the effects of different choices of detail levels for a sample spectrum of dataset B (see Application), demonstrating the flexibility of the proposed preprocessing pipeline.

Feature extraction

For each preprocessed spectrum, a list of peak locations is obtained by detecting intensity maxima that are above the SNR threshold (SNR_t). The peak locations of all samples are first merged, then clustered according to $W(m/z)$ by using a memory-saving implementation of centroid linkage hierarchical clustering. Common peaks are defined as the cluster centroids, computed by averaging peak locations belonging to the same cluster. Finally, peak heights are quantified as the maximum intensity of the normalized spectrum over each cluster.

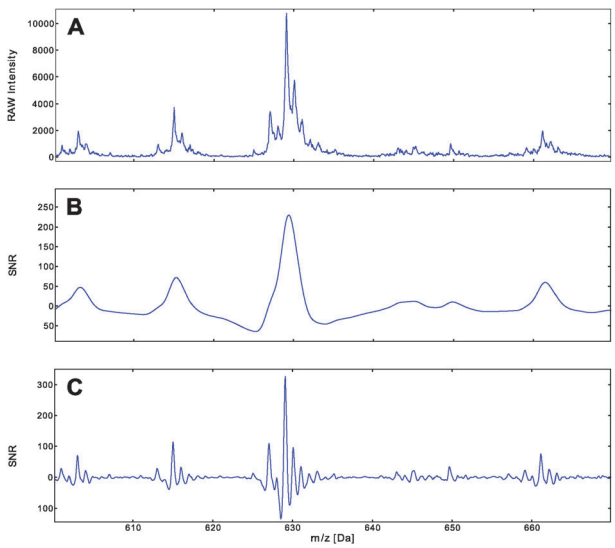

Fig. 2 Effects of different detail levels for a sample spectrum of dataset B (see Application). (A) Raw spectrum. (B) De-noised, baseline-corrected and normalized signal with $W(m/z) = 3$ (evaluated at 629 Da). (C) De-noised, baseline-corrected and normalized signal with $W(m/z) = 0.4$ (evaluated at 629 Da).

Machine learning

The machine learning module is run in a 100× random subsampling cross-validation (CV) schema (75–25% training-test proportion), while maintaining class labels proportion. For each training set, peaks are directly ranked according to feature weights computed by single-layer perceptron classifier after feature normalization; each test set is normalized according to the parameters of the training set. For the i-th CV subsample, a series of perceptron models built upon an increasing number of features (from the i-th ranked peak list)[11] are tested, obtaining a test error (TE) curve. An average test error (ATE) curve is finally computed, together with a unified peak list ranked by average positions.

Peak detection reproducibility

The performance of peak detection was evaluated on a validated and openly available MALDI-TOF dataset ("Aurum"), containing known purified and trypsin-digested proteins.[12] The performance test follows the approach proposed by Yang and colleagues.[13] We considered 200 spectra equally arranged in eight groups,

setting the detail level parameters to $W_l = 0.08$ (evaluated on the peak measured at $m/z = 1024.64$ Da) and $W_h = 0.28$ (evaluated at $m/z = 2886.31$ Da). We ran TOFwave peak detection on spectra in the same group with seven signal-to-noise thresholds ($SNR_t = 2, 5, 10, 20, 50, 100, 200$). We then assessed the performance in terms of sensitivity with respect to four false discovery rate (FDR) ranges ([0,0.1), [0.2,0.3), [0.4,0.5), [0.6,0.7)), with sensitivity and FDR computed using lists of ground-truth peaks.[13] For each group, we averaged the sensitivity values with associated FDR falling in the same range, thus obtaining eight average sensitivities for each FDR range. Finally, we compared the results with the sensitivities previously obtained[13] in the same setup by five public peak detection algorithms, namely "Cromwell",[14] "CWT",[15] "LMS",[16] "LIMPIC"[17] and "PROcess".[18] Results are displayed in Fig. 3, for increasing FDR. Sensitivity of TOFwave is found comparable to or higher than that of the alternative algorithms, with inferior variability in terms of interquartile ranges.

Additionally, to simulate typical tuning differences among users or laboratories, we repeated the previous analysis while decreasing and increasing the pair $W = (W_l, W_h)$ values by 10%, considering the additional setups $W-$ and $W+$, respectively. The comparison of sensitivity between $W-$, W and $W+$ is displayed in Fig. 4, for increasing FDR, showing that TOFwave peak detection is robust against small variations of (W_l, W_h).

Application

TOFwave was run on three MALDI-TOF datasets (A, B, C; details in ESI†). Datasets A and B were acquired in-house, to test the pipeline performance in the mass range typical of metabolic profiling. Sanguiin H6 ($C_{82}H_{54}O_{52}$, monoisotopic mass 1870.158 Da)[19] and methionine ($C_5H_{11}NO_2S$, monoisotopic mass 149.051 Da) have been spiked in half of the samples for datasets A and B, respectively. Dataset C is a proteomics pattern dataset consisting of spectra from 77 controls and 93 ovarian cancer patients.[20]

Results are presented in Table 1 in terms of Average Test Error (ATE) curves with 97.5% bootstrap (1000× resampling)

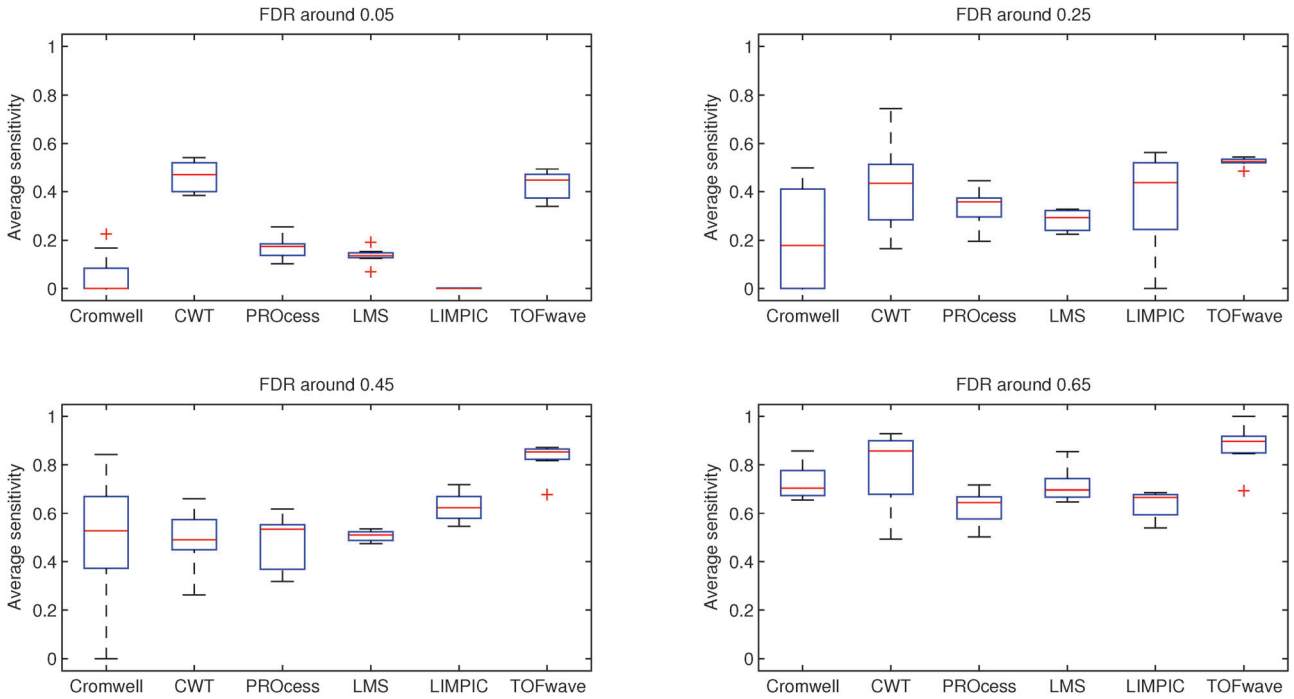

Fig. 3 Performance of TOFwave peak detection in comparison with different publicly available algorithms. Average sensitivity for different FDR ranges: [0,0.1) (FDR around 0.05, top left), [0.2,0.3) (FDR around 0.25, top right), [0.4,0.5) (FDR around 0.45, bottom left), [0.6,0.7) (FDR around 0.65, bottom right). The detail level parameters were set to $W = (W_l, W_h) = (0.08,0.28)$.

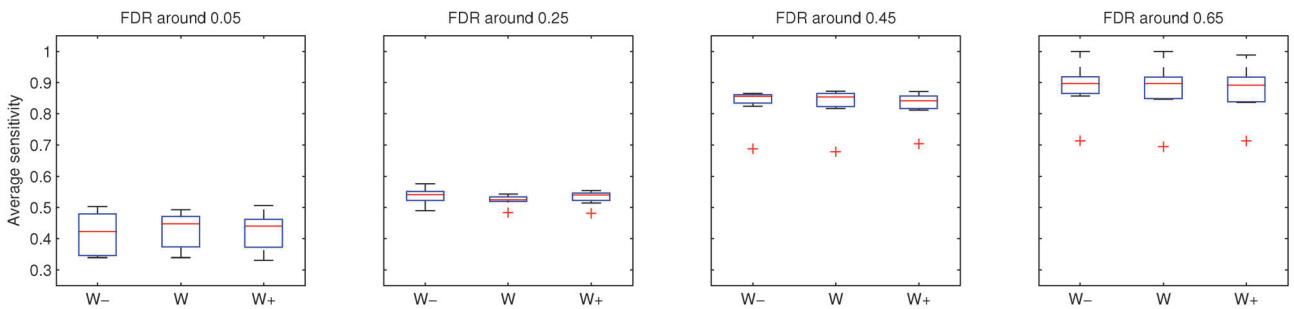

Fig. 4 Stability of TOFwave peak detection for 10% perturbation of the W setup (see Fig. 3), with $W- = 0.9W$ and $W+ = 1.1W$.

Table 1 Dataset A, $SNR_t = 2$: Average Test Error (ATE) with 97.5% bootstrap confidence interval (ATE_{min}, ATE_{max}) for the top-10 ranked peaks. *n*: number of peaks used in the model

n	ATE	ATE_{min}	ATE_{max}
1	0.006	0.000	0.014
2	0.003	0.000	0.009
3	0.002	0.000	0.008
4	0.013	0.007	0.021
5	0.015	0.007	0.024
6	0.033	0.020	0.048
7	0.034	0.022	0.049
8	0.046	0.033	0.061
9	0.045	0.032	0.060
10	0.044	0.031	0.059

Table 2 Dataset A, $SNR_t = 2$: top-10 ranked peaks (*m*/*z*, cluster centroids) with their associated cluster bounds (C_{min}, C_{max}) and average positions. *n*: peak rank

n	*m*/*z* [Da]	C_{min} [Da]	C_{max} [Da]	Avg. pos.
1	1910	1909.6	1911.7	1.03
2	1917	1915.7	1918.5	2.23
3	1894	1893.3	1896.1	2.75
4	1932	1930.7	1933.6	4.94
5	277.2	276.86	277.46	6.42
6	347.7	347.46	348.03	8.22
7	87.9	87.70	88.14	13.17
8	304.4	304.08	304.81	15.07
9	97.8	97.63	97.90	19.25
10	409.6	409.10	410.06	21.26

confidence intervals (ATE_{min}, ATE_{max}), and in Table 2 in terms of unified peak lists, ranked by average positions. To ensure that the procedure is not affected by systematic bias, each experiment was repeated by running the modeling step 10 times after having randomly permuted the sample labels, using the original class proportion ("random labels" in the following; details in ESI†).

As for dataset A, detail level parameters were set to $W_l = 0.4$ (evaluated at 629.2 Da) and $W_h = 4$ at 3148 Da (see ESI†). An average prediction error on test data ATE = 0.2% was obtained with three top-ranked peaks, with $SNR_t = 2$ (Table 1 and Table S2 (ESI†), Fig. 5). The analysis was repeated with $SNR_t = 4$ obtaining an Average Test Error of 0% on the three top-ranked peaks (Tables S3 and S4, Fig. S2, ESI†). In positive ion mode MALDI of large polyphenols is expected to yield mainly sodium and potassium adducts. As far as Sanguiin is concerned, *m*/*z* signals for these species are expected at *m*/*z* 1893.147 ($[M + Na]^+$), 1909.121 ($[M + K]^+$) and 1916.137 ($[M + 2Na]^+$). In Table 2 and Table S3 (ESI†) are presented the top-10 ranked features selected by the pipeline. Mass values represent the position of the cluster centroids after preprocessing with their associated cluster bounds. They can be related to the envelopes of the expected ionic species as can be seen in Fig. S1, ESI.† It is worth mentioning that cluster bounds, and thus the uncertainty about cluster centroids, may be reduced by selecting smaller W_l and W_h.

As for dataset B, detail level parameters were set to $W_l = 0.1$ at 171.55 Da and $W_h = 0.2$ at 644.6 Da (see ESI†). ATE = 0.3% was obtained with only one feature, and ATE = 1.1% with three features (Fig. 6, Tables S5 and S6 (ESI†)). In positive ion mode, methionine is expected as the $[M + H]^+$ ion at *m*/*z* 150.058 Da. Features corresponding to this ion and its ^{13}C contribution are indeed among the top 3 discriminating features (Table S5, ESI†).

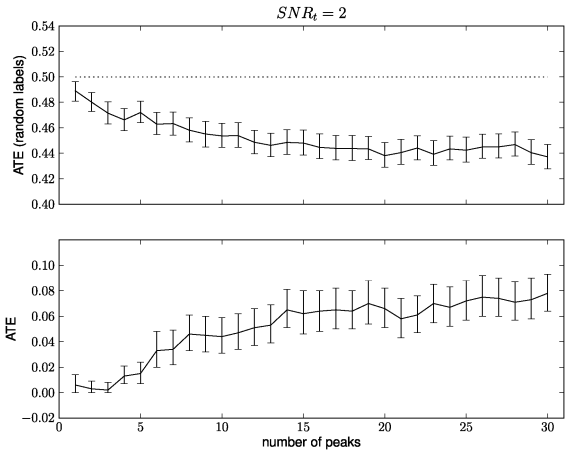

Fig. 5 Dataset A: Average Test Error (ATE) curves with 97.5% bootstrap confidence intervals (c.i.) for a complete preprocessing and classification experiment, with $SNR_t = 2$. Horizontal dotted line indicates the "no-information error rate", here defined as the ratio between the smallest class and the whole dataset size, corresponding to the error reached by classifying all samples as belonging to the most populous class.

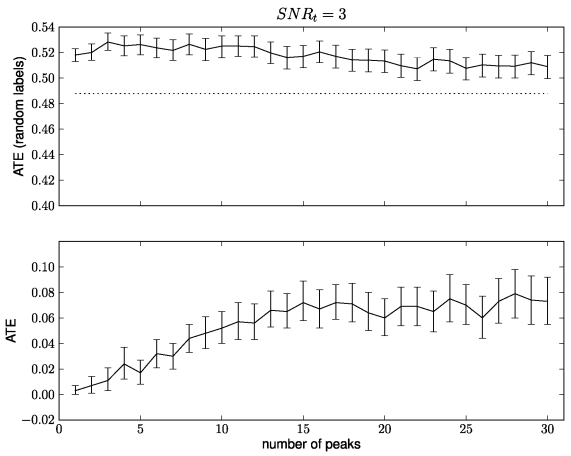

Fig. 6 Dataset B: Average Test Error (ATE) curves with 97.5% bootstrap c.i. for a complete preprocessing and classification experiment, with $SNR_t = 3$. Horizontal dotted line indicates the "no-information error rate".

An average prediction error on test data ATE = 36.6% for 10 top features was found on dataset C (details in ESI†).

Conclusions

In this work we presented TOFwave, a new pipeline for the identification of potential biomarkers in TOF-MS profiling experiments. TOFwave is provided as a modular software pipeline, freely available from http://mlpy.sf.net/tofwave. The pipeline couples TOF-MS data preprocessing with a predictive modeling workflow optimized to allow the control of selection bias and to avoid overfitting effects, thus improving reproducibility. Preprocessing parameters have been reduced to a minimum

and they are directly correlated to the physical characteristics of the spectra, exploiting the multiscale nature of the Continuous Wavelet Transform. The user thus gains full control on peak detection steps, avoiding any non-explicit visual tuning that may jeopardize reproducibility. Performance tests on the validated MALDI-TOF "Aurum" dataset proved the robustness of the preprocessing phase with respect to different choices of the parameters. Moreover, the performance of peak detection on the same dataset was found comparable to or higher than alternative publicly available algorithms in terms of sensitivity and false discovery rate. In our investigation we showed that TOFwave can be used on a wide range of MALDI-TOF spectra, with biomarkers ranging from small molecules (*e.g.* methionine, dataset B) to large metabolites (*e.g.* Sanguiin, dataset A), and even proteins or peptides (*e.g.* dataset C). This flexibility makes TOFwave suitable for biomarker identification at different *omics* levels, from metabolomics to proteomics.

Acknowledgements

The authors wish to thank the anonymous reviewers for their valuable comments and indications, which greatly helped to improve the quality of the manuscript. This work was supported by the EU FP7 Project HiPerDART and the PAT funded Project ENVIROCHANGE. Pietro Franceschi acknowledges Provincia Autonoma di Trento for the financial support under the Program Post Doc (Project STASMA) within the Programma Pluriennale per la Ricerca.

References

1 J. Sorace and M. Zhan, *BMC Bioinf.*, 2003, **4**, 24.
2 M. Wagner, D. Naik and A. Pothen, *Proteomics*, 2003, **3**, 1692–1698.
3 D. Ransohoff, *Nature*, 2005, **5**, 142–149.
4 K. Baggerly, K. Coombes and J. Morris, *Cancer Inf.*, 2005, **1**, 9–14.
5 J. Zou, G. Hong, X. Guo, L. Zhang, C. Yao, J. Wang and Z. Guo, *PLoS One*, 2011, **6**, e26294.
6 The MAQC Consortium, *Nat. Biotechnol.*, 2010, **28**, 827–838.
7 I. Daubechies, *Ten lectures on wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, 1992.
8 D. Albanese, R. Visintainer, S. Merler, S. Riccadonna, G. Jurman and C. Furlanello, arXiv:1202.6548.
9 A. Barla, G. Jurman, S. Riccadonna, S. Merler, M. Chierici and C. Furlanello, *Briefings Bioinf.*, 2008, **2**, 119–128.
10 C. Torrence and G. Compo, *Bull. Am. Meteorol. Soc.*, 1998, **79**, 61–78.
11 C. Furlanello, M. Serafini, S. Merler and G. Jurman, *BMC Bioinf.*, 2003, **4**, 54.
12 J. A. Falkner, M. Kachman, D. M. Veine, A. Walker, J. R. Strahler and P. C. Andrews, *J. Am. Soc. Mass Spectrom.*, 2007, **18**, 850–855.
13 C. Yang, Z. He and W. Yu, *BMC Bioinf.*, 2009, **10**, 4.
14 K. Coombes, S. Tsavachidis, J. Morris, K. Baggerly, M. Hung and H. Kuerer, *Proteomics*, 2005, **5**, 4107–4117.
15 P. Du, W. Kibbe and S. Lin, *Bioinformatics*, 2006, **22**, 2059–2065.
16 Y. Yasui, M. Pepe, M. Thompson, B. Adam, G. Wright, Jr., Y. Qu, J. Potter, M. Winget, M. Thornquist and Z. Feng, *Biostatistics*, 2003, **4**, 449–463.
17 D. Mantini, F. Petrucci, D. Pieragostino, P. Del Boccio, M. Di Nicola, C. Di Ilio, G. Federici, P. Sacchetta, S. Comani and A. Urbani, *BMC Bioinf.*, 2007, **8**, 101.
18 X. Li, R. Gentleman, X. Lu, Q. Shi, J. Iglehart, L. Harris and A. Miron, *Bioinformatics and Computational Biology solutions using R and Bioconductor*, 2005, pp. 91–109.
19 M. Gasperotti, D. Masuero, U. Vrhovsek, G. Guella and F. Mattivi, *J. Agric. Food Chem.*, 2010, **58**, 4602–4616.
20 B. Wu, T. Abbot, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams and H. Zhao, *Cancer Inf.*, 2006, **2**, 123–132.