Chemometrics and Intelligent Laboratory Systems 104 (2010) 20–27

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Standard machine learning algorithms applied to UPLC-TOF/MS metabolic fingerprinting for the discovery of wound biomarkers in *Arabidopsis thaliana*

Julien Boccard [a], Alexandros Kalousis [b], Melanie Hilario [b], Pierre Lantéri [c], Mohamed Hanafi [d], Gérard Mazerolles [e], Jean-Luc Wolfender [a], Pierre-Alain Carrupt [a], Serge Rudaz [a,*]

[a] School of Pharmaceutical Sciences, University of Geneva, University of Lausanne, Switzerland
[b] Artificial Intelligence Lab, Computer Science Department, University of Geneva, Switzerland
[c] Laboratory of Analytical Sciences, University of Lyon, University Claude Bernard Lyon I, CNRS, Villeurbanne, France
[d] Unité de Recherche en sensométrie et Chimiométrie, ONIRIS, Nantes, France
[e] INRA-UMR 1083 SPO, Montpellier, France

ARTICLE INFO

Article history:
Received 7 September 2009
Received in revised form 18 February 2010
Accepted 8 March 2010
Available online 15 March 2010

Keywords:
Data mining
Machine learning
Mass spectrometry
Metabolomics
Arabidopsis thaliana
UPLC-MS

ABSTRACT

Metabolomics experiments involve the simultaneous detection of a high number of metabolites leading to large multivariate datasets and computer-based applications are required to extract relevant biological information. A high-throughput metabolic fingerprinting approach based on ultra performance liquid chromatography (UPLC) and high resolution time-of-flight (TOF) mass spectrometry (MS) was developed for the detection of wound biomarkers in the model plant *Arabidopsis thaliana*. High-dimensional data were generated and analysed with chemometric methods.

Besides, machine learning classification algorithms constitute promising tools to decipher complex metabolic phenotypes but their application remains however scarcely reported in that research field. The present work proposes a comparative evaluation of a set of diverse machine learning schemes in the context of metabolomic data with respect to their ability to provide a deeper insight into the metabolite network involved in the wound response. Standalone classifiers, i.e. J48 (decision tree), kNN (instance-based learner), SMO (support vector machine), multilayer perceptron and RBF network (neural networks) and Naive Bayes (probabilistic method), or combinations of classification and feature selection algorithms, such as Information Gain, RELIEF-F, Correlation Feature-based Selection and SVM-based methods, are concurrently assessed and cross-validation resampling procedures are used to avoid overfitting.

This study demonstrates that machine learning methods represent valuable tools for the analysis of UPLC-TOF/MS metabolomic data. In addition, remarkable performance was achieved, while the models' stability showed the robustness and the interpretability potential. The results allowed drawing attention to both temporal and spatial metabolic patterns in the context of stress signalling and highlighting relevant biomarkers not evidenced with standard data treatment.

1. Introduction

Metabolomics is devoted to the examination of the entire system of metabolites present in a given cell, tissue type or organism. Rather than focusing on components separately, it aims to understand how biological systems behave at a functional level as the result of the integration and interaction between individual components that can be concurrently monitored [1]. Despite the emergence of robust technologies, the identification of individual metabolites still remains difficult. Exploiting this highly complex information has become a challenging issue in systems biology [2]. As a well-established model in plant biology, *Arabidopsis thaliana* constitutes an attractive basis for the development of non-targeted metabolite fingerprinting methods in plants and since metabolites are increasingly recognised as important signalling molecules, a metabolomic study was envisaged to assess its response to wound.

A non-targeted UPLC-MS metabolite fingerprinting approach was selected because of its sensitivity, resolution and throughput capacity [3]. With the development of high-throughput methods for the analysis of complex biological systems ("Omics" approaches), values for concentrations of thousands of compounds are available in a single experiment. Comparing samples has become a problem of high dimensionality [4] and previous studies showed successful applications of multivariate analysis to extract and display relevant information from HPLC-MS [5,6] and UPLC-MS data [3,7]. Multivariate statistical modelling (such as PCA and PLS) is classically used for a better understanding of the relationships between concentrations and samples. Its use becomes however problematic when a high number

* Corresponding author. School of Pharmaceutical Sciences, University of Geneva, 20 Bd d'Yvoy, 1211 Geneva 4, Switzerland. Tel.: +41 22 379 65 72; fax: +41 22 379 68 08.
URL: serge.rudaz@unige.ch (S. Rudaz).

0169-7439/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2010.03.003

of variables are measured, while datasets not only become larger in size, more complex or when non-linear relationship is expected. Within this context, machine learning methods, and more specifically classification algorithms, are promising alternatives to address these issues.

This work falls within the area of metabolomic fingerprinting and stress biomarker discovery since the main goal was to identify significant changes at the metabolic level. The analytical approach is non-targeted and the identification of induced metabolites relied on well-known classification algorithms and feature selection methods, covering various underlying learning assumptions. A comparative study was performed to assess their ability to cope with specific problems in metabolomics, such as noise, outliers and high dimensionality. Resampling procedures, namely cross-validation, were used to estimate the predictive performance of the algorithms. The stability of the resulting models, with respect to perturbations of the training set, was also investigated.

2. Experimental

2.1. Solvents

Methanol (MeOH) from VWR (Leuven, Belgium) and isopropanol (IPA) from SDS (Peypin, France) were used for the extraction. LC-MS grade acetonitrile (ACN) and water were obtained from Fisher Scientific (Loughborough, UK). Formic acid of LC-MS grade from Sigma-Aldrich (Steinheim, Germany) was used as the eluent additive.

2.2. Plant samples

A. thaliana (Ecotype Columbia-0) were grown for 7 weeks at 22 °C, and 70% relative humidity in a controlled greenhouse environment to ensure the consistency of the metabolic fingerprints. Half of the rosette leaves were crushed across the apical lamina with forceps (30–40% of the leaf area), incubated in the light for defined timings, harvested and immediately frozen in liquid nitrogen.

2.3. Extraction procedure

Leaf tissues of approximately 500 mg were ground and extracted with 5 mL isopropanol (IPA) using a ball mill (ball diameter 2 cm, frequency 30 Hz, time 2 min) (Retsch MM200, Schieritz & Hauenstein AG, Arlesheim, Switzerland). IPA extracts were purified by SPE (Waters Sep-Pak C18, Vac 1 cc, 100 mg). After cartridge conditioning (1 mL MeOH, 1 mL MeOH–H_2O (85:15% v/v)), 5 mg of extract was diluted in 500 µL of MeOH–H_2O (85:15% v/v), loaded and washed with 1 mL MeOH–H_2O (85:15% v/v) to remove chlorophyll and other lipophilic pigments. The residue was finally dissolved in 200 µL of MeOH–H_2O (85:15% v/v).

2.4. LC-MS analyses

LC-MS analyses were performed on a Micromass-LCT premier time-of-flight (TOF) mass spectrometer (Waters, Mass., USA) with an electrospray interface and coupled with an Acquity UPLC system (Waters, Mass., USA). ESI conditions: capillary voltage 2800 V, cone voltage 40 V, MCP detector voltage 2650 V, source temperature 120 °C, desolvation temperature 250 °C, cone gas flow 10 L/h, and desolvation gas flow of 550 L/h. Detection was performed in negative ion mode in the m/z range 100–1000 with a scan time of 0.25 s in centroid mode.

A solution of Leucine-Enkephalin (Sigma-Aldrich, Steinheim, Germany) at 5 µg/mL was infused through the Lock Spray™ probe at a flow rate of 20 µL/min with the help of a second LC pump (Shimadzu LC-10ADvp, Duisburg, Germany) for the Dynamic Range Enhancement (DRE) lockmass.

The separations were carried out on Waters Acquity UPLC columns at 35 °C (BEH C18: 1.0×50 mm and 2.1×150 mm, 1.7 µm) with the following solvent system: A = 0.1 vol.% formic acid–water and B = 0.1 vol.% formic acid–ACN. For the 1.0×50 mm column, the gradient elution was performed at a flow rate of 300 µL/min using: 5% B for 0.3 min, 5 to 98% B in 6.7 min and holding at 98% B for 3 min.

2.5. Datasets

UPLC-TOF/MS was performed on 72 plant extracts with experimental groups corresponding to unwounded controls (Ctrls) or wounded specimens (Wnd), the latter being harvested at different time points, namely after 90 min (Wnd90), 3H (Wnd3H), 6H (Wnd6H) and 24H (Wnd24H). Wounding time points were based on a previous study. Moreover each wounding time was equally separated into two subgroups of eight specimens, distinguishing local (L) from distal (D) leaf samples [8]. Thus, the dataset had a total number of 72 samples falling into nine different classes, namely Ctrl, Wnd90L, Wnd90D, Wnd3HD, Wnd3HL, Wnd6HD, Wnd6HL, Wnd24HD and Wnd24HL, where each class contained eight samples (see Fig. 1A).

Three groups of datasets were created in order to evidence both the temporal and spatial development of the wound defence signalling by detecting accurate induction patterns. First, the complete dataset including all nine classes and thus all 72 instances, was investigated. In the second group, Local and Distal specimens from a specific time point, were included (eight samples per class) and compared to the control specimens (eight samples). Thus, at each time point T, a dataset involving three classes, i.e. Ctrl, WndT-L, and WndT-D, and a total of 24 instances was built. Four datasets (#1–4) composed this group. In the third group, the Local and Distal specimens from two subsequent time points, T_i and T_j were compared. Each dataset had therefore four distinct classes, i.e. WndT_i-L, WndT_i-D, WndT_j-L, and WndT_i-D, and a total of 32 samples. The third group included six datasets (#5–10). Fig. 1B gives a brief description of the different datasets.

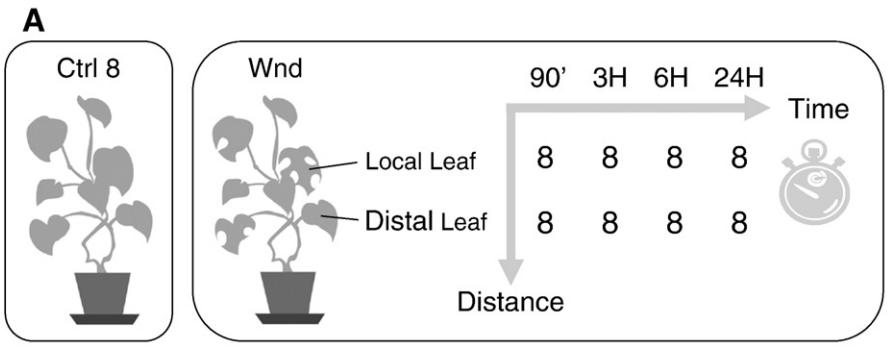
3. Machine learning schemes

A comparative benchmark of several classification algorithms and feature selection methods was proposed by covering a choice of particular learning principles to reveal their specific strengths and applicability. Several parameter settings were assessed for each machine learning algorithm. Standard implementations of the classification algorithms and feature selection methods were performed with the Waikato Environment for Knowledge Acquisition (WEKA, University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/ml/weka/) [9]. Learning algorithms were evaluated in terms of stability [10] by analysing the classification models they produced. Ideally, learning algorithms building models insensitive to perturbations of the training sets offer more confidence when examining the resulting models in order to identify relevant markers.

3.1. Learning algorithms and parameters

3.1.1. Decision trees

Decision trees are univariate logic-based classification algorithms able to sort the training examples using feature values based on a divide-and-conquer strategy. Each node involves testing a particular attribute of the dataset and the branches correspond to values that this variable can take. Decision trees are usually quick algorithms evaluating attributes independently. The J48 decision tree algorithm was used with four different M parameter values [11], defining the minimum number of examples in each node of the tree ($M = 2, 4, 6, 8$), high values of M corresponding to general and simple models. Random forest (RF) decision tree ensembles are collections of individual decision trees obtained by a bagging procedure (bootstrap aggregating). All trees are



B

Dataset #	Ctrls	Wnd90		Wnd 3H		Wnd 6H		Wnd 24H		Number of	Number of
		local	distal	local	distal	local	distal	local	distal	Classes	Instances
G	x	x	x	x	x	x	x	x	x	9	72
1	x	x	x							3	24
2	x			x	x					3	24
3	x					x	x			3	24
4	x							x	x	3	24
5		x	x	x	x					4	32
6		x	x			x	x			4	32
7		x	x					x	x	4	32
8				x	x	x	x			4	32
9				x	x			x	x	4	32
10						x	x	x	x	4	32

Fig. 1. (A) Plant sample class distribution associated with their spatial and temporal characteristics. (B) Classification dataset composition.

applied in parallel to build a consensus predicted class as the most frequent output produced by the independent trees [12]. RFs of $N = 10$, 20, 30, 40 and 50 trees were also tested with 8 features.

3.1.2. k-nearest neighbours

The *k*-nearest neighbours (kNN) algorithm is an instance-based learning scheme that relies on the regrouping of instances with similar properties in the high-dimensional space of the data. By examining the class labels of other instances positioned in the neighbourhood of an unknown sample, kNN intends to predict its class value by a majority vote. Indeed, the memorised training set is searched for the most strongly resembling observations and classifications schemes are created from the instances themselves. Rather than inferring classification rules, the learned concept is encapsulated in the training set without explicitly illustrating the data patterns. The kNN algorithm [13] was applied with a number of nearest neighbours *k* that took values in the set [3,5,7,10], low *k* values generating complex models.

3.1.3. Support vector machines

Support vector machines (SVMs) are increasingly popular classification methods with high generalisation abilities [14,15]. They rely on the selection of a small number of critical boundary instances called support vectors to build a hyperplane separating the classes. SVMs were originally developed by Vapnik [16] and rely on a loss function allowing the simultaneous minimisation of both the empirical prediction error and the model's complexity. The Sequential Minimal Optimization (SMO) algorithm [14–16], was applied with a linear kernel and five different values for the *C* parameter, $C = 0.01$, 0.1, 1, 10, 100.

3.1.4. Multilayer perceptron

Multilayer ANNs create non-linear classifiers by the use of a hierarchical structure involving three types of neurons, i.e. input, hidden and output units linked together by a network of connections of adjustable weights. The back propagation algorithm using a gradient descent process was used to minimise the error function. A multilayer perceptron (MLP) [17], with one hidden layer of $N = 3$, 5, 7 or 10 units was experimented.

3.1.5. Radial Basis Function network

The RBF networks represent another form of feed-forward neural network. They possess three layers where the hidden one is composed of neurons applying a radial activation function. The output is computed as a weighed sum of the hidden units [18]. Each hidden unit constitutes a specific point in the input space whose activation value decreases with a Gaussian distance function. A Gaussian Radial Basis Function network (RBF) [18] with 2, 10 and 20 centres was applied. This implementation of the RBF networks uses the *k*-means clustering method as basis function. Symmetric multivariate Gaussians are fit to the data from each cluster and logistic regression was then applied.

3.1.6. Naive Bayes

Naive Bayes is a probabilistic algorithm relying on an explicit probability model by allocating a probability to each class that corresponds to the product of the individual probabilities of every attribute value. The predicted class label then corresponds to the class with the greatest probability. Despite being very simple and assuming independence among the variables, NB performs often well in the presence of complex problems, making it useful in practice [19]. The Naïve Bayes classifier was assessed with a Gaussian distribution function [13,19].

3.2. Feature selection

Dimensionality reduction is often very useful when mining large datasets and the selection of a subset of representative features retaining the salient characteristics of the data is therefore a fundamental issue. The number of features is a key aspect that determines the size of the hypothesis space of the data [20]. In practice, most classifiers' performance can be improved thanks to prior variable selection, as accuracy is often negatively influenced by a

high number of irrelevant variables. This is particularly the case in highly multivariate biological data such as metabolomic data and circumventing the inclusion of irrelevant attributes having negative effects could provide a more compact and interpretable image of a phenomenon. By combining complementary filter selection and learning algorithms relying on distinct learning principles, performance can be improved thanks to the strengths of both methods.

Five well-known feature selection algorithms were examined as a pre-processing step prior to the application of the classification algorithms to analyse their ability of improving the classification accuracy. Some parameters were used at fixed values for each feature selection method. Classification performance was estimated using a ten fold cross-validation with an inner cross-validation loop for the classification algorithm's parameter automatic tuning. Additionally, most of the feature selection methods require the number of features to select to be given *a priori*; nevertheless there is no way to estimate that number in advance. A similar inner cross-validation procedure, estimating the classification performance of models built on feature sets of different cardinalities was employed in order to choose the appropriate number of features to select, i.e. the one giving rise to the best classification performance [21]. Subsequently, the feature selection algorithm was reapplied on the complete set of the outer fold where the number of features to select was the one that resulted in the best performance in the inner cross-validation procedure. The following feature set cardinalities were experimented: 16, 32, 64, 128, 256 and 512.

3.2.1. RELIEF-F

The instance-based feature ranking scheme RELIEF-F (Recursive Elimination of Features) [22] was evaluated with $k=10$ nearest neighbours. It provides the advantage to efficiently assess each variable in the context of other attributes. The algorithm randomly selects instances from the original data and detects their nearest neighbour of the same and the opposite class. These objects are used to compare the values of each variable and adjust its score. A positive or negative update depends on the ability of the variable to separate either objects from the same class, from different classes or both. RELIEF-F uses the average value of the k nearest neighbours of each class for the evaluation of a given variable and a class contribution is weighted by its prior probability.

3.2.2. SVM-based selection

SVM-based feature selection methods are wrappers based on SVM as target learning algorithm to estimate the predictive value of subsets of attributes. SVM coefficients are used as indicators of relevance and the discriminatory power of each feature is estimated by the norm of its weight in the model. A cross-validation procedure is usually applied to evaluate the classifier's accuracy with a specific subset. A linear kernel with $C=1$ was used to build two SVM-based feature selection models, based on a single application of the linear SVM (SVMOne) or an iterative execution of the algorithm where the lowest ranked feature is removed at each loop (RFSVM), respectively.

3.2.3. Information Gain

The Information Gain (IG) is a univariate method that selects features on the basis of the information contribution related to the class variable without considering feature interactions. It relies on the difference of entropy before and after splitting the data according to a given variable [23].

3.2.4. Correlation-based feature selection

Correlation-based feature selection (CFS) [24], assesses subsets of features on the basis of both the predictive merit of each feature and its degree of correlation with other features already included in the selected subset. The balance promotes a high correlation with the class information and a low level of inter-correlation within the subset. CFS was applied using forward selection (iteratively adding variables to existing subsets until convergence) to ensure the smallest sizes of subsets. It performs a parsimonious selection and can therefore drastically reduce the number of features in a highly inter-correlated dataset without loss of prediction performance.

3.3. Classification performance

Numerous measures of performance are available, such as sensitivity, specificity and Kappa coefficient [25], the most common and comprehensible being the classification accuracy, i.e. the percentage of correctly classified instances. Since no preference for sensitivity or specificity was shown, the accuracy measure was used here to evaluate the classification performance of the different classification algorithms on each dataset. Pairwise resampled t-tests [26] were performed to compare each algorithm's results against every other within each dataset, based on a significance threshold of $p<0.05$.

The parameters tested for each classifier or feature selection algorithm are summarised in Table 1. Automatic parameter tuning was performed by selecting the appropriate parameter setting for each algorithm based on classification accuracy using a nested cross-validation procedure [21].

4. Results and discussion

Plant specimens were analysed by UPLC-MS experiments with a fast chromatographic gradient to get a short analysis time of less than 7 min allowing high sample throughput while ensuring a satisfactory resolution and low ion suppression levels. The hyphenation of UPLC separation and MS detection produces data having an intrinsic 2D structure. A sample is characterised by a set of mass spectra recorded in sequence, that contain all the ions detected for a given time lapse. UPLC-MS raw data constitute therefore large but sparse data matrices. A pre-processing procedure was applied to convert these raw data into peak vectors, i.e. information-rich objects containing the m/z values and chromatographic times of detected peaks, therefore concentrating the whole information content (Fig. 2).

This pre-processing procedure was performed with the MakerLynx™ software. Centroid data were used and a baseline correction was applied to correct sample drift. Peak detection was performed by considering ion intensities measured in the same m/z and retention time interval (0.05 Da and 0.5 min) as signals describing the same ion among samples. A global data table characterising each of the 72 samples by a series of 1068 variables corresponding to intensities associated with a precise m/z signal measured at a specific retention time was obtained.

Table 1
Classification and feature selection algorithms and parameters (CA = classification algorithm, FS = feature selection).

Algorithm	Reference	Type	Learning scheme	Parameter	Values
J48	[11]	CA	Decision tree	*M*	2, 4, 6, 8
RF	[12]	CA	Decision tree ensemble	*N*	10, 20, 30, 40, 50
KNN	[13]	CA	Nearest neighbours	*K*	3, 5, 7, 10
SVM	[16]	CA	Support vector machine	*C*	0.01, 0.1, 1, 10, 100
MLP	[17]	CA	Artificial neural network	*N*	3, 5, 7, 10
RBF	[18]	CA	Artificial neural network	*B*	2, 10, 20
NB	[19]	CA	Probabilistic method	–	–
RELIEF	[22]	FS	Nearest neighbours	*K*	10
IG	[23]	FS	Entropy	–	–
SVMOne	[16]	FS	Support vector machine	*C*	1
RFSVM	[16]	FS	Support vector machine	*C*	1
CFS	[24]	FS	Correlation	–	–

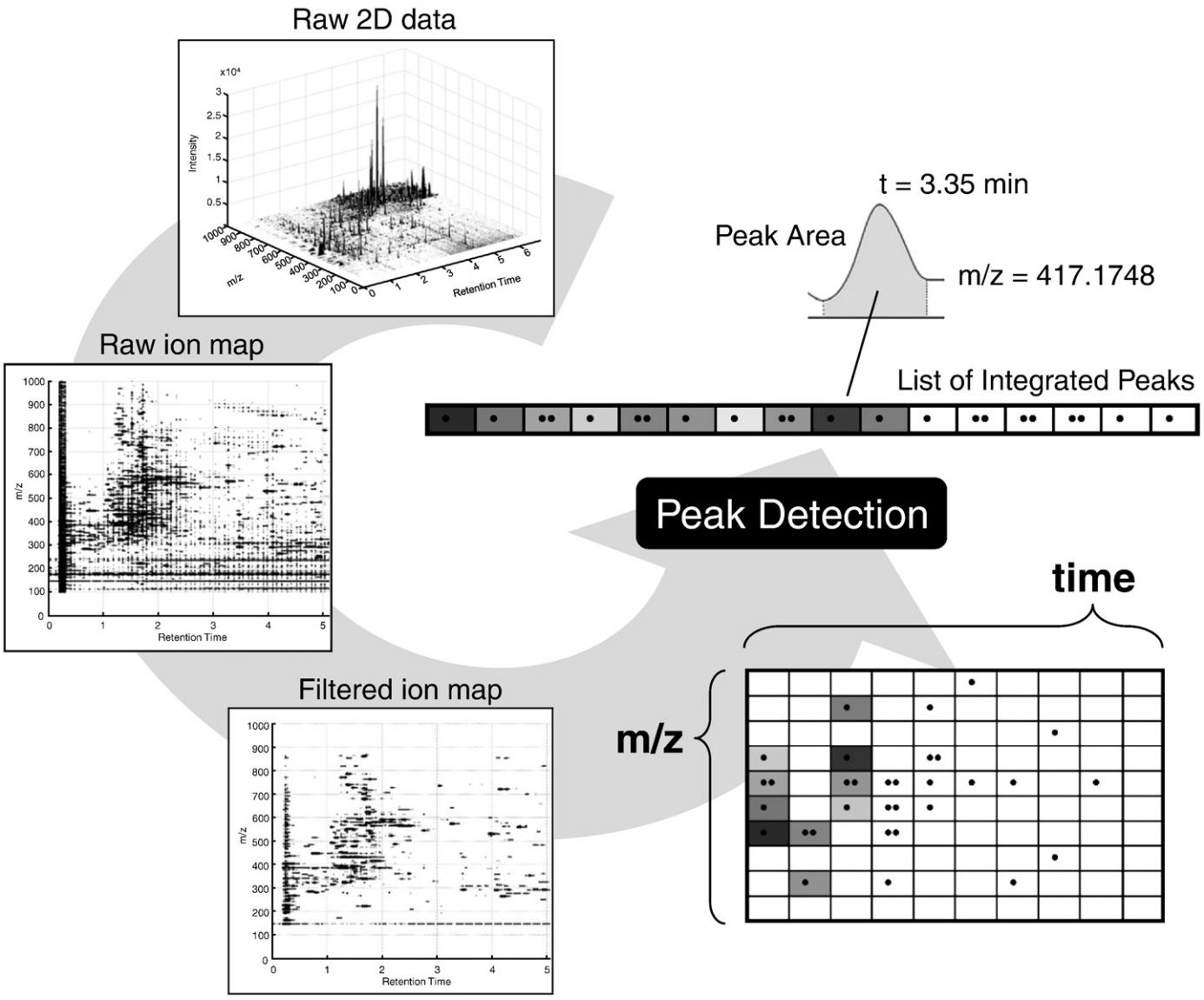


Fig. 2. Data pre-processing methodology concentrating the information content of the data.

Diverse classification datasets (see Fig. 1B) were built with respect to both the spatial and temporal development of the wound response and unit variance scaling was performed as pre-treatment.

Principal Components Analysis (PCA) was initially performed on the global dataset (G) as a preliminary step of data examination to assess the samples' distribution with respect to their class labels. Control leaf samples were clearly separated on the first principal plane (PC1 16% vs. PC2 12%) but no underlying structure could be evidenced within the wounded specimens (Fig. 3). Indeed, principal components correspond to the directions that maximise the variance and there is no guarantee that these dimensions are discriminant [27].

As a consequence of the complex character of the biological extracts and the subtleness of variations in such non-targeted analyses, the detection of minor but significant biomarkers among constitutive highly expressed compounds remains a challenging analytical and statistical problem. Several data subsets containing samples corresponding to specimens measured at different time points following the wound induction were derived from the original dataset. This procedure intends to construct classification models able to uncover factors that vary not only between wounded and control specimens, but also between different time points after wounding and between local and distal leaves.

4.1. Classification performance with the full sets of metabolites

The evaluation of several well-known machine learning algorithms applied on each classification dataset and the original full feature dimensionality constituted the first set of learning experiments. Despite the small number of instances within each class, the overall classification performance was noticeably good. The results gave rise to three groups of classification accuracy. The lower scores were achieved by J48, RBF and KNN; NB and RF achieved better performance, while MLP and SVM gave the highest percentages of correctly classified instances. The full results are given in Table 2.

Regarding learning principles, decision trees are fundamentally different algorithms compared to the others. When the latter examine all the features together in parallel, decision trees evaluate the features sequentially. The worst results over all datasets were obtained with J48, while RF performed reasonably well. NB produced interesting results, most notably for clear situations such as datasets comparing Control samples (datasets 1–4) with other classes. When important interactions between variables are expected, parallel methods are generally more robust and constitute therefore a better choice when building classification models of metabolomic data [4]. The necessity to examine more than one variable at a time was confirmed by

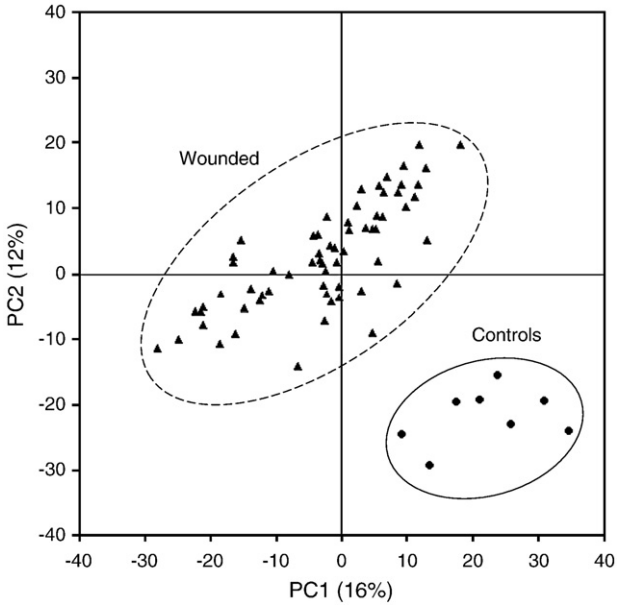


Fig. 3. PCA score plot (PC1 16% vs. PC2 12%). Control plants (Ctrls) symbolised by circles and wounded plants (Wnd) by triangles.

obtaining the best average performance with two parallel methods, namely MLP and SVM. Surprisingly, RBF networks presented poor results compared to other methods. This method was demonstrated to provide complex mappings but the high dimensionality and multiple correlations characteristic of metabolomic data did not allow obtaining satisfying results.

Performance was further compared by resampled *t*-tests [26] to rank the methods on a statistical basis. This confirmed the advantage of using MLP or SVM. Nonetheless, MLP computing time requirements constituted a noteworthy drawback compared to SVM.

Since the SVM algorithm achieved the best classification performance, an examination of the stability of the corresponding classification models, with respect to changes in the classification datasets, was envisaged [28]. The rank correlation coefficients showed that the learned models were particularly stable since Spearman's rank coefficients were above 90% for all datasets (data not shown).

4.2. Classification performance with feature selection

With data of high dimensionality such as metabolomic UPLC-MS data, feature selection algorithms provide a way to reduce dimensionality. This reduction can lead to simpler and more interpretable classification models, reduce the computational complexity of the subsequent classification model building phase and potentially improve the predictive performance. The final classification models were built on the feature sets of reduced dimensionality and the feature selection algorithms were coupled with each of the classification algorithms examined in Section 3.1.

Table 2
Classification performance accuracy on the full metabolite dimensionality (in percentage).

Dataset	MLP	RF	SVM	KNN	NB	J48	RBF	Average
1	97.8%	96.7%	97.0%	93.0%	91.5%	93.3%	61.3%	90.1%
2	93.0%	96.7%	95.5%	92.7%	95.8%	91.3%	65.3%	90.1%
3	97.2%	98.7%	99.2%	92.7%	86.8%	79.5%	70.7%	89.2%
4	99.7%	98.5%	100.0%	95.8%	94.8%	90.7%	86.3%	95.1%
5	91.7%	81.4%	87.7%	71.2%	81.0%	80.6%	63.3%	79.6%
6	91.8%	80.5%	90.1%	65.5%	82.4%	88.2%	62.8%	80.2%
7	95.6%	91.1%	94.3%	86.4%	89.4%	84.8%	75.3%	88.1%
8	66.0%	55.7%	67.6%	44.6%	61.8%	36.7%	40.3%	53.2%
9	97.4%	91.9%	96.1%	93.8%	90.8%	83.5%	59.0%	87.5%
10	90.3%	90.9%	91.8%	85.1%	84.9%	82.7%	65.4%	84.4%
G	75.4%	70.2%	77.5%	63.4%	66.8%	58.3%	57.4%	67.0%
Average	90.5%	86.6%	90.6%	80.4%	84.2%	79.1%	64.3%	82.2%

The classification results achieved were noteworthy but considering the good performance of the algorithms with the full metabolite dimensionality and the small size of the training sets, feature selection resulted in enhancements of the prediction accuracy, particularly for the worst classification results (e.g. RBF and kNN). Similarly to the classification results without feature selection, accuracy performance was compared by resampled *t*-tests. The results indicated the RELIEF and IG feature selection methods associated with SVM or MLP as the best combinations with similar performance. SVMOne and RFSVM gave rise to modest improvements for most of the classification algorithms.

The average cardinalities of the selected feature sets were between 230 and 300 for IG (235), RELIEF (281), SVMOne (286) and RFSVM (292). CFS had a remarkably different behaviour producing feature sets of much lower cardinality which nevertheless resulted to classification models with high predictive power. CFS focuses on the reduction of the correlation of the selected features which could be a possible explanation of the low cardinalities of the feature sets it produced, i.e. the original feature set contains highly redundant and correlated features. In some sense, the CFS subsets contain prototypical features of clusters of highly correlated variables. This was particularly the case for datasets 1 to 4, all of which involved control plant samples as one of the classes. For these datasets, an average of 35 features was sufficient to obtain almost perfect classification accuracies (with the exception of the models produced by J48 that had considerably lower performance). This was considered as a clear indication of high levels of redundancy within the feature subsets selected by IG, RELIEF, SVMOne, and RFSVM, since a much smaller number of features is sufficient to produce classification models of perfect accuracy.

The feature selection methods that resulted in the largest number of significant improvements over the different classification datasets and algorithms were IG and RELIEF. In fact, these two feature selection methods resulted in better performance even when compared to the more sophisticated SVM-based feature selection methods (RFSVM and SVMOne). However, careful consideration is needed when using the IG feature selection method due to its univariate nature that might miss features that are strongly informative only in the presence of other variables. The full feature selection results are provided as supplementary material.

4.3. Signalling of the wound response

4.3.1. Dataset assessment

The classification results indicate a family of easy classification problems, i.e. datasets including either samples from the Control (89–95% without feature selection, and >95% with) or from the Wnd24H classes (85–88% and 90–93%); both classes apparently have quite distinctive metabolic profiles. The other wounded plant groups, namely Wnd90, Wnd3H and Wnd6H, presented apparently more similar analytical metabolomic profiles as the corresponding classification problems resulted in lower predictive performance averages (accuracies of around 80% and 90%), with the most difficult cases to separate comparing Wnd3H vs. Wnd6H (dataset #8, accuracies of 53% and 61%).

The dataset *G* containing all classes and thus providing a global view of the biological experimentation and the combination of SVM with RELIEF resulted in a predictive accuracy as high as 83%.

4.3.2. Specific vs. global situations

Having established the pertinent information content of most of the datasets through the high predictive performance obtained with prior feature selection, two approaches were then explored concurrently for further investigation of the biological relevance. On one

hand, a closer analysis of the learned models involving specific situations, i.e. datasets #1–10 including three or four experimental classes, was achieved. This part intended to gain a comprehensive in-depth understanding of the discriminating features with respect to definite time spans of wound induction with their local and distal specificity. On the other hand, a global assessment of the impact of feature selection was performed on the whole dataset (G) to provide a restricted non-redundant list of metabolites over all samples. The first approach is illustrated by the results of the SVM models built on dataset #1 that are presented here in more detail. An external validation procedure relying on an independent test dataset acquired several weeks after the first set of experiments was applied. It corresponded partially to dataset #1 with six control and six locally wounded leaves harvested after 90 min (Wnd90L). The different classifiers and feature selection–classifier combinations, learned previously from dataset #1, were used to predict the labels of the new samples and obtained the same levels of excellent performance, see Table 3.

The combination of the RELIEF feature selection and the SVM classification method was chosen according to the previous classification results presented in Section 3.2 and SVM coefficients were used as indicators of predictive power. The higher the average coefficient of a given feature over the ten cross-validation folds, the more important is the feature. The absolute weight value of a given feature reflects then the relative importance, provided that all the features have been normalised and a linear kernel is used. A list ranking the variables by their predictive merit could thus be generated according to their SVM coefficients. Thanks to the high-mass accuracy of TOF/MS detection, some of the most predictive variables were identified as known biomarkers [29]. As an example, previously described stress metabolites from the oxylipin family could be highlighted, such as m/z 209.1167 (jasmonic acid, JA, rank 6), m/z 322.2011 (JA-Ile, rank 15), m/z 352.1754 (HOOC-JA-Ile, rank 3), m/z 225.1125 (HO-JA, rank 25), m/z 237.1493 (OPC-4, rank 27) or m/z 338.1959 (HO-JA-Ile, rank 13). Corresponding formiate (FA), sodium formiate (Na-FA) and sodium (Na) adducts, i.e. m/z 406.1834 (HO-JA-Ile, Na-FA, rank 12), m/z 417.1748 (JA-Glc, FA, rank 4), m/z 374.1579 (HOOC-JA-Ile, Na, rank 11) or m/z 277.1041 (JA, Na-FA, rank 16) were also identified. All these m/z were ranked on the top of the list (< rank 30) generated with the SVM models learned on the feature sets selected by the RELIEF algorithm. The most relevant features for the comparison between Control, Wnd90L and Wnd90D specimens are reported in Table 4. This table summarises the ranks of each feature with respect to a given pairwise classification problem. The comparison of these ranks can draw attention on either ubiquitous (local or distal) or specifically localised signals (local or distal). While the former stand on top of the first two ranking lists and not the third, the latter are expected to have a high rank on the third column and only one of the two others. From these considerations, it appears clearly that all identified known oxylipins induced after 90 min are strictly local regulators. On the other hand, distal signalling molecules remain only poorly described and features accounting for the Wnd90D class distinction were therefore hardly assignable to known metabolites. These compounds constitute however relevant candidates for further complementary targeted studies, as described by Glauser et al. [30].

Table 3
External test set prediction accuracy for dataset #1 ($n = 12$, 6 controls and 6 locally wounded leaves).

Test set	MLP	RF	SVM	KNN	NB	J48	RBF
No selection	100.0%	100.0%	100.0%	100.0%	91. 7%	83.3%	50.0%
InfoGain	100.0%	100.0%	100.0%	100.0%	100.0%	83.3%	100.0%
RELIEF	100.0%	100.0%	100.0%	100.0%	91.7%	91.7%	100.0%
SVMOne	100.0%	100.0%	100.0%	100.0%	91.7%	91.7%	100.0%
RFSVM	100.0%	100.0%	100.0%	100.0%	91.7%	100.0%	100.0%
CFS	100.0%	100.0%	100.0%	100.0%	91.7%	83.3%	100.0%

Table 4
SVM mean weight coefficients rankings for dataset #1.

Cvs90L	Rank Cvs90D	90Lvs90D	Feature name	Compound ID	Localisation
1	18	731	m/z 903.4354		Local/distal
2	820	4	m/z 353.157		Local
3	175	30	m/z 352.1754	HOOC-JA-Ile	Local
4	585	3	m/z 417.1748	JA-Glc (FA)	Local
5	113	144	m/z 994.466		Local/distal
6	583	7	m/z 209.1167	JA	Local
7	358	21	m/z 561.1325		Local
8	3	392	m/z 768.4452		Local/distal
9	87	194	m/z 307.0754		Local/distal
10	4	455	m/z 340.0573		Local/distal
11	223	24	m/z 374.1579	HOOC-JA-Ile (Na)	Local
12	937	1	m/z 406.1834	HO-JA-Ile (Na-FA)	Local
13	590	2	m/z 338.1959	HO-JA-Ile	Local
14	19	823	m/z 361.1619		Local/distal
15	951	6	m/z 322.2011	JA-Ile	Local
16	790	10	m/z 277.1041	JA (Na-FA)	Local
17	51	282	m/z 863.4111		Local/distal
18	981	5	m/z 445.2004		Local
19	56	163	m/z 823.4427		Local/distal
20	982	8	m/z 435.1723		Local
21	81	308	m/z 783.4091		Local/distal
22	15	263	m/z 881.4387		Local/distal
23	439	18	m/z 327.2143		Local
24	11	279	m/z 383.1129		Local/distal
25	272	46	m/z 225.1125	HO-JA	Local
26	40	617	m/z 851.4712		Local/distal
27	980	11	m/z 237.1493	OPC-4	Local
28	936	12	m/z 306.092		Local
29	35	908	m/z 447.048		Local/distal
30	48	973	m/z 813.4297		Local/distal
…	…	…	…	…	…
243	17	38	m/z 775.473		Distal
316	59	47	m/z 274.0726		Distal
326	53	35	m/z 334.0918		Distal
359	28	33	m/z 397.0893		Distal
443	39	9	m/z 387.1154		Distal
…	…	…	…	…	…

The second approach intended the global investigation of the G dataset. It was performed by applying both the RELIEF and the CFS algorithms to generate either a comprehensive picture of the induction phenomenon for the former or a limited list of non-redundant representative discriminating metabolite patterns for the latter. CFS automatically determines the number of features to select, i.e. 53 for the global dataset, while for RELIEF the 113 most important features were retained, as that number was simply the average feature subset size selected by the combination of RELIEF and SVM.

Performing PCA on the RELIEF selected features and retaining the first two principal components provided a very clear distribution of the different sample classes along two orthogonal temporal and spatial dimensions (Fig. 4). Additionally, the samples of the test set could be successfully associated with their respective classes. The class separation was remarkable but not surprising since the selected features on which PCA was performed were retained precisely in view of maximizing the class distinction. However, the comparison with the PCA score plot based on the full feature space (Fig. 3) illustrates the discriminative power of the selected features. Moreover, previous studies revealed wound regulators in a temporal perspective without bringing to light the spatial development of wound signalling [3,5].

On the other hand, the CFS list highlighted non-correlated prototypical metabolite patterns of high relevance acting either as temporal or spatial regulators of the wound response. This parsimonious selection allowed the collection of valuable information to drive correlation analysis of the whole metabolite pool and guide the targeted investigation of specific compounds in a subsequent metabolic profiling step as an alternative strategy to explore the data complexity [30].

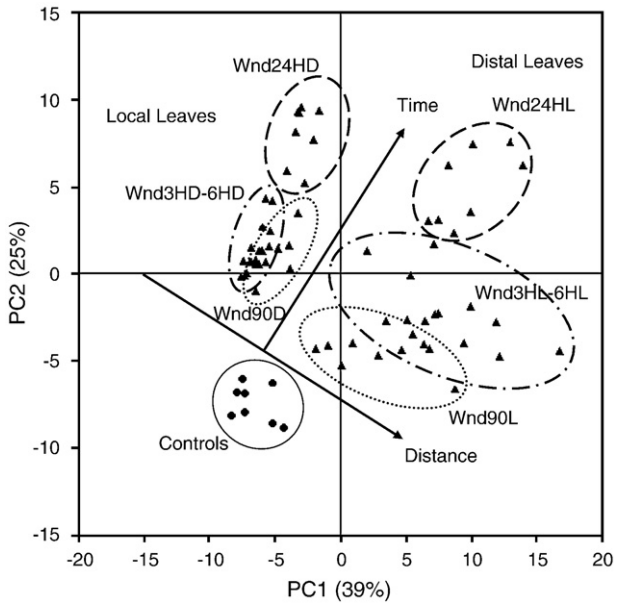


Fig. 4. PCA based on the most discriminating features as these were determined by RELIEF (113 features).

Taken together the results of the two data investigation approaches indicate that machine learning provided relevant tools to consider metabolic data and allowed the detection of signalling events occurring at different times and between distinct parts of a wounded plant.

5. Concluding remarks

Metabolic fingerprinting constitutes the very first exploratory phase of biomarker discovery. The potency of UPLC-TOF/MS for the sensitive detection of wound biomarkers provides a high-throughput analysis for non-targeted approaches in metabolomics and generates a large amount of complex data involving networks of inter-correlated metabolites. In the present study, machine learning algorithms including feature selection and classification algorithms demonstrate a great potential to increase the biological knowledge as they can help to automate tasks and provide relevant predictions through classification models. A careful analysis of the models allowed identifying the factors involved in the response to wounding in a spatial and temporal perspective. Feature selection prior to learning was beneficial to common learning algorithms by building simpler and more interpretable models. The classification models built with the RELIEF–SVM combination highlighted both known metabolite patterns and original compounds. Additionally, CFS provided a restricted list of highly relevant prototypical metabolites. Further investigations of these molecules by metabolic profiling and target analysis are envisaged to correlate these results with the biochemical events governing both local and systemic response of plants to wounding.

Acknowledgements

The Swiss National Science Foundation is thanked for supporting this work (grant 205320-107735 to JLW and SR). We thank Elia Grata and Gaétan Glauser for the UPLC-MS metabolomic data and their helpful comments. The work reported in this paper was partially funded by the European Commission through EU project e-LICO (FP7-231519).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chemolab.2010.03.003.

References

[1] L.W. Sumner, P. Mendes, R.A. Dixon, Plant metabolomics: large-scale phytochemistry in the functional genomics era, Phytochemistry 62 (2003) 817.
[2] R.D. Hall, Plant metabolomics: from holistic hope, to hype, to hot topic, New Phytol. 169 (2005) 453.
[3] E. Grata, J. Boccard, D. Guillarme, G. Glauser, P.A. Carrupt, E.E. Farmer, J.L. Wolfender, S. Rudaz, UPLC-TOF-MS for plant metabolomics: a sequential approach for wound marker analysis in *Arabidopsis thaliana*, J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 871 (2008) 261.
[4] W. Weckwerth, K. Morgenthal, Metabolomics: from pattern recognition to biological interpretation, Drug Discov. Today 10 (2005) 1551.
[5] J. Boccard, E. Grata, A. Thiocone, J.Y. Gauvrit, P. Lanteri, P.A. Carrupt, J.L. Wolfender, S. Rudaz, Multivariate data analysis of rapid LC-TOF/MS experiments from *Arabidopsis thaliana* stressed by wounding, Chemometric Intell. Lab. Syst. 86 (2007) 189.
[6] J. Trygg, J. Gullberg, A.I. Johansson, P. Jonsson, T. Moritz, Chemometrics in metabolomics — an introduction, Biotechnol. Agriculture and Forestry 57 (2006) 117.
[7] E. Grata, D. Guillarme, G. Glauser, J. Boccard, P.A. Carrupt, J.L. Veuthey, S. Rudaz, J.L. Wolfender, Metabolite profiling of plant extracts by ultra-high-pressure liquid chromatography at elevated temperature coupled to time-of-flight mass spectrometry, J. Chromatogr. A 1216 (2009) 5660.
[8] E. Grata, J. Boccard, G. Glauser, P.A. Carrupt, E.E. Farmer, J.L. Wolfender, S. Rudaz, Development of a two-step screening ESI-TOF-MS method for rapid determination of significant stress-induced metabolone modifications in plant leaf extracts: the wound response in *Arabidopsis thaliana* as a case study, J. Sep. Sci. 30 (2007) 2268.
[9] E. Frank, M. Hall, L. Trigg, G. Holmes, I.H. Witten, Data mining in bioinformatics using Weka, Bioinformatics 20 (2004) 2479.
[10] A. Kalousis, J. Prados, J.C. Sanchez, L. Allard, M. Hilario, Distilling classification models from cross validation runs, , 2004, p. 113.
[11] J.R. Quinlan, Improved use of continuous attributes in C4.5, J. Artif. Intell. Res. 4 (1996) 77.
[12] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5.
[13] D.W. Aha, D. Kibler, M.K. Albert, Instance-based learning algorithms, Mach. Learn. 6 (1991) 37.
[14] S.S. Keerthi, E.G. Gilbert, Convergence of a generalized SMO algorithm for SVM classifier design, Mach. Learn. 46 (2002) 351.
[15] J. Platt, How to implement SVMs, Ieee Intell. Syst. 13 (1998) 26.
[16] V.N. Vapnik, An overview of statistical learning theory, Ieee Trans. Neural Netw. 10 (1999) 988.
[17] S.I. Gallant, Perceptron-based learning algorithms, Ieee Trans. Neural Netw. 1 (1990) 179.
[18] T. Poggio, F. Girosi, Networks for approximation and learning, Proceedings of the Ieee, 78, 1990, p. 1481.
[19] D.J. Hand, K.M. Yu, Idiot's Bayes — Not so stupid after all? Int. Stat. Rev. 69 (2001) 385.
[20] T.M. Mitchell, Machine Learning, McGraw Hill Higher Education, New York, 1997.
[21] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273.
[22] M. Robnik-Sikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Mach. Learn. 53 (2003) 23.
[23] S. Kullback, An application of information theory to multivariate analysis, Ann. Math. Stat. 23 (1952) 88.
[24] M.A. Hall, Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning, Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc, 2000, p. 359.
[25] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. 76 (1971) 378.
[26] C. Nadeau, Y. Bengio, Inference for the generalization error, Mach. Learn. 52 (2003) 239.
[27] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417.
[28] A. Kalousis, J. Prados, M. Hilario, Stability of feature selection algorithms: a study on high dimensional spaces, Knowl. Inf. Syst. 12 (2007) 95.
[29] E.E. Farmer, E. Almeras, V. Krishnamurthy, Jasmonates and related oxylipins in plant responses to pathogenesis and herbivory, Curr. Opin. Plant Biol. 6 (2003) 372.
[30] G. Glauser, J. Boccard, S. Rudaz, J.-L. Wolfender, Mass spectrometry-based metabolomics oriented by correlation analysis for wound-induced molecule discovery: identification of a novel jasmonate glucoside, Phytochem. Anal. 21 (2010) 95.