Analytica Chimica Acta 829 (2014) 1–8

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data

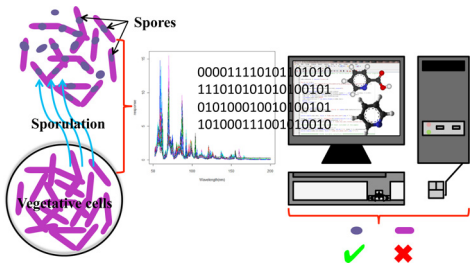
Piotr S. Gromski [a], Yun Xu [a], Elon Correa [a], David I. Ellis [a], Michael L. Turner [b], Royston Goodacre [a,*]

[a] School of Chemistry, Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK
[b] School of Chemistry, The University of Manchester, Brunswick Street, Manchester M13 9PL, UK

HIGHLIGHTS

- LDA, PLS-DA, SVM and RF analyses were applied to MS data.
- Double cross-validation using bootstrapping was employed to assess models.
- For all classifications, all bacteria were assessed with ~95% accuracy.
- Parsimonious modelling was used on a reduced set of mass ions and was more robust.
- The approaches developed are equally applicable to any multivariate data.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 23 January 2014
Received in revised form 24 March 2014
Accepted 27 March 2014
Available online 31 March 2014

Keywords:
Variable selection
Supervised learning
Bootstrapping
Double cross-validation
Pyrolysis mass spectrometry
Bacillus

ABSTRACT

Many analytical approaches such as mass spectrometry generate large amounts of data (input variables) per sample analysed, and not all of these variables are important or related to the target output of interest. The selection of a smaller number of variables prior to sample classification is a widespread task in many research studies, where attempts are made to seek the lowest possible set of variables that are still able to achieve a high level of prediction accuracy; in other words, there is a need to generate the most parsimonious solution when the number of input variables is huge but the number of samples/objects are smaller. Here, we compare several different variable selection approaches in order to ascertain which of these are ideally suited to achieve this goal. All variable selection approaches were applied to the analysis of a common set of metabolomics data generated by Curie-point pyrolysis mass spectrometry (Py-MS), where the goal of the study was to classify the Gram-positive bacteria *Bacillus*. These approaches include stepwise forward variable selection, used for linear discriminant analysis (LDA); variable importance for projection (VIP) coefficient, employed in partial least squares-discriminant analysis (PLS-DA); support vector machines-recursive feature elimination (SVM-RFE); as well as the mean decrease in accuracy and mean decrease in Gini, provided by random forests (RF). Finally, a double cross-validation procedure was applied to minimize the consequence of overfitting. The results revealed that RF with its variable selection techniques and SVM combined with SVM-RFE as a variable selection method, displayed the best results in comparison to other approaches.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Pyrolysis mass spectrometry (Py-MS) is a well-established analytical technology which has been used for the characterization

Abbreviations: LDA, linear discriminant analysis; PLS-DA, partial least squares-discriminant analysis; SVM, support vector machines; RF, random forests.

* Corresponding author. Tel.: +44 161 306 4480; fax: +44 2004636.
E-mail address: roy.goodacre@manchester.ac.uk (R. Goodacre).

http://dx.doi.org/10.1016/j.aca.2014.03.039
0003-2670/© 2014 Elsevier B.V. All rights reserved.

of microbial systems for several decades [1,2]. This is due to its high discriminatory ability [3], and it has been said to be a powerful 'fingerprinting' technique [4,5], which can be routinely applied to any organic material. As its names suggests, this technology involves the thermal degradation of complex organic molecules. This process takes place either in an inert atmosphere or within a vacuum, whereby complex organic substances are rapidly broken down into stable primary fragments (termed pyrolysate) which can then be measured by mass spectrometry. Depending on the type of pyrolysis used, rapid heating (from 300 to 1000 °C) of samples can occur on a metal filament or within a quartz sample tube/cuvette, and whilst this is a destructive method, it routinely allows for the reproducible analysis of samples of less than 1 mg [6].

Py-MS has been successfully used to discriminate complex tertiary mixtures of bacterial species (*i.e. Bacillus subtilis, Escherichia coli, Staphylococcus aureus*) [3], urinary tract infection bacteria [7], and many other microbial studies. More recently, these include changes in bacterial activated sludge populations [8], and the study of bacterial and fungal bioremediation systems [6]. Other more diverse applications include an assessment of olive oil adulteration [9,10]. Indeed, in combination with advanced statistical methods, and its ability to be applied to a diverse range of samples of biotechnological interest, it has in the past been termed an 'anything-sensor' [11], which has also been aptly demonstrated with numerous non-biological applications [12–14].

Due to the complexity of the output from Py-MS, where the data may be highly collinear and the majority of descriptors unrelated to the study, variable selection prior to pattern recognition is a crucial component of the analysis. This parsimonious approach to modelling [15] is needed in order to obtain robust and reproducible results. In the case of Py-MS for example, where we measure 150 mass-to-charge (m/z) intensities as input variables (range 51–200 m/z), it would indeed be desirable to construct a model with a reduced number of variables prior to classification [16–18].

Up until now, several studies have employed different approaches for feature selection and classification of Py-MS such as genetic algorithms, which have been successfully used as variable selection techniques in combination with multiple linear regression and partial least squares (PLS) regression [19]. Other statistical methods that have been applied to Py-MS data include variable selection in discriminant PLS analysis [20], PLS-discriminant analysis (PLS-DA) [21] which has been used for the discrimination of *B. subtilis* strains [22], Fisher's linear discriminant analysis (LDA) [23] which has been successfully employed to distinguish the difference between tobacco types [24], as well as support vector machines (SVM) [25–28] which was successfully implemented for the analysis of Py-GC–MS data [29].

The aim of this study is to compare various variable selection methods which are commonly used for the analysis of chemical data. To this end we employed LDA, PLS-DA, SVM and random forests (RF) [30], and until now the latter has not to our knowledge been used to analyse Py-MS data. The Py-MS data had been previously collected [17] from various bacteria belonging to different *Bacillus* species where the aim was to effect both species classification as well as being able to recognise the bacterial physiological state correctly: all bacteria were cultivated either as spores or as vegetative biomass.

2. Materials and methods

All statistical data analyses were performed using the R (2.15.0) [31] software environment. This language comprises a selection of packages suitable for different types of data and is available as free software in the public domain.

In this study, we used different approaches for feature reduction and classification of (1) the physiological states of the genus *Bacillus* (spores *versus* vegetative cells), as well as (2) differentiating seven *Bacillus* species (*B. amyloliquefaciens, B. cereus, B. licheniformis, B. megaterium*, *B. subtilis* (including *B. niger* and *B. globigii*), *B. sphaericus* and *Brevibacillus laterosporus*) [17]. This comparative analysis (Fig. 1) is based on the following combinations of variable selection methods and their respective classification algorithms: stepwise forward variable selection that has been combined with LDA [32]; PLS-DA and its variable importance for projection (VIP) coefficient [33]; SVM recursive feature elimination (SVM-RFE) to reduce the number of variables prior to SVM [34]; and finally, mean decrease in accuracy and Gini provided by RF [35]. These techniques have been used to establish variable importance, as well as to reduce input dimensionality and computational load/time.

2.1. Data

The original dataset used in this study for variable selection and classification was collected by Goodacre et al. [17] using Curie-point Py-MS from seven different types of *Bacillus*, where a detailed description of data collection and instrumentation can be found.

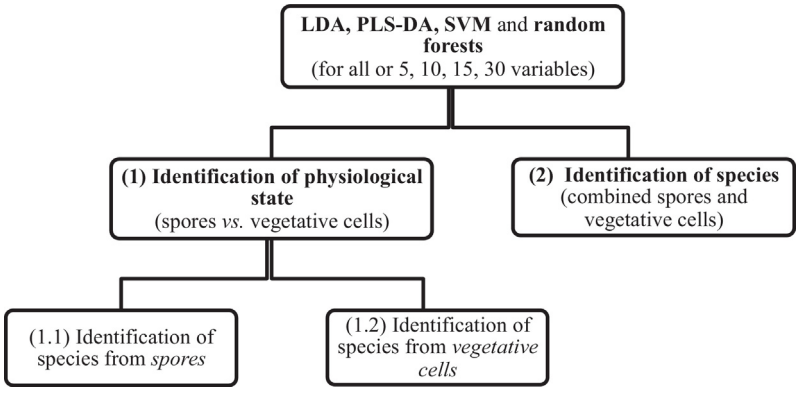


Fig. 1. The overall workflow of the studies for variable importance ranking for the analysis of four data sets. (1) Physiological state estimation where the prediction accuracy has been calculated for the separation of vegetative cells from spores. Following this separation two subsets were analysed for *Bacillus* speciation: (1.1) identification of *Bacillus* species from spores; (1.2) identification of *Bacillus* species from vegetative cells. (2) Illustrates the analysis of species when both spores and vegetative cells are analysed together in the same model.

The dataset consisted of 216 observations (108 spores and 108 vegetative cells) and 150 m/z input variables and was initially analysed using PLS-DA, neural networks, as well as rule induction approaches and genetic programming [17]. In addition, these data have also been analysed using a genetic algorithm–Bayesian network (GA–BN) in order to select important variables [36]. GA is based on the principles of Darwinian selection that relies on combination of chromosome selection, recombination and mutations. These algorithms attempt to find the relationship between input and output and turn the data into information that allows for the estimation of which variables are important [37–39]. Bayesian networks on the other hand described how each of the variables in the selection relates probabilistically to its parents by encoding distinctive combinations of possible distributions [40,41].

2.1.1. Data preparation

Predicting performance based on such a limited dataset is challenging, especially if one wants to build a good predictive model relying on already ranked variables. Hence, the dataset was randomly divided into three data sets based on principles of double cross-validation, as reported by Westerhuis et al. [42] using a bootstrapping resampling method (Fig. 2) [43,44]. Generally, bootstrapping is a statistical approach which is based on resampling of the original data to generate multiple 'new' partitions of the dataset (train and validation sets) and aims to provide a more accurate analysis [43,44]. With the intention of accurate validation of the results, only the training data sets were used for feature selection and not the validation nor test sets. Hence, the risk of overfitting is significantly reduced, as demonstrated by Brereton [45]. In order to evaluate the performance of the four models, a bootstrapping resampling procedure has been carried out for 100 repeated data splits [42,45]; further explanation of the bootstrapping procedure can be found in Supplementary Information (SI; Fig. S1). In addition, data scaling (auto-scaling) has been applied prior to the analysis to reduce the potentially dominating influence of a few input variables with much larger variations compared to the remaining ones. This auto-scaling method is a statistical procedure in which for each column (input variable) the mean value of that column is subtracted, followed by dividing the row entries for that column by the standard deviation within the same column. This results in scaling all the variables to a comparable scale [45,46]. Auto-scaling was performed on the training data only, and for the test data the means and standard deviations from the training data were used for this pre-treatment.

2.2. Variable selection

A selection of different feature selection approaches were employed and these were used to attempt to select the best variables needed for classification. This classification was based on the physiological status of the bacteria – spores *versus* vegetative cells – and these methods are briefly detailed below.

2.2.1. Stepwise forward variable selection for LDA model

Stepwise forward variable selection was used to perform feature selection prior to LDA. This approach begins with a preliminary model that includes the variables that best separates the groups within each dataset. In forward selection, Wilk's lambda criterion [32] was used to select which new variables would be included in the modelling. When a variable is added to the model and the p-value still displays statistical significance, then this feature is retained, else it is eliminated from the model. The process is completed when the addition of all new variables does not improve the model or when the model accuracy is 100% [32,47,48]. For variable selection, we used the *greedy.wilks* function within the package "klaR" [32].

2.2.2. Variable selection using VIP coefficient as provided by PLS-DA

PLS-DA is built on the rotation of X and Y components simultaneously in order to enable maximum separation between classes and to recognise which variables most influence class separation. The technique results in a set of scores and loadings vectors. These loadings correspond to PLS weighted sums of the absolute regression coefficients (B) computed separately (W) for each outcome. Therefore, PLS-DA can be used as a variable

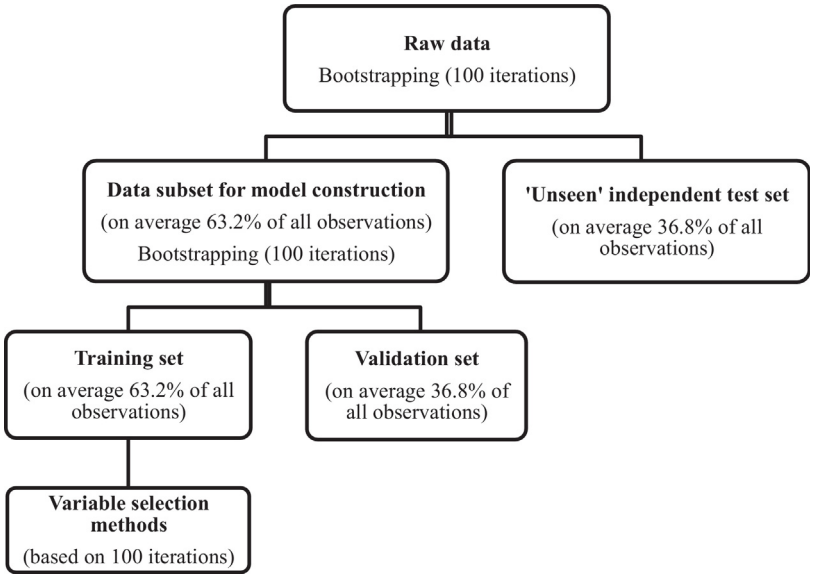


Fig. 2. Data preparation workflow that illustrates the data selection methods employed in this study, which incorporates bootstrapping with replacement and feature reduction approaches for variable selection. The resulting partition is of the Py-MS data collected from seven *Bacillus* species that were cultured either as vegetative cells or allowed to spore. Data comprise 216 observations (108 spores, 108 vegetative cells) and 150 m/z intensities (m/z 51–200) [17]. This process included double cross-validation in two resampling steps. The first was to generate 100 'unseen' test sets which one samples of independent data. The second was used to generate 100 calibration models incorporating variable selection (this was also performed 100 times).

selection method by ranking the most important loadings in decreasing order; that is to say the modulus of the weighting vector [21,49]. In this case study, we applied the generic function *varImp* within R that ranked variables in decreasing order of importance which relies on the VIP coefficient [33].

2.2.3. SVM-RFE

In contrast to forward variable selection, SVM-RFE excludes variables believed irrelevant using the SVM kernel weights after they are ranked in order of importance. This assessment is performed in the following steps as described in Guyon et al. [34] and Duan et al. [50]: (1) the model is trained using SVM on the previously generated training set; (2) the variables are ranked according to their weights in decreasing order; (3) the feature with the lowest weight is removed; (4) the process is repeated until classification accuracy decreases. SVM-RFE was performed in e1071 package "Misc Functions of the Department of Statistics (e1071), TU Wien" version 1.6 [51].

2.2.4. Variable selection using random forests

In this study we used the "randomForest" [35] package implemented in R. This package computes two measures of variable importance: mean decrease in accuracy over all classes (which was used to measure the importance of each variable to the classification) and mean decrease in Gini [52–55]. Further explanation of the approach that we have used is provided in SI where Fig. S2 is provided as typical example of ranking results and also in our recent studies on volatile analysis (paper under review).

2.3. Modelling process

As an integral part of the above variable selection process, different multivariate data analysis (MVA) algorithms were evaluated in terms of their ability to effect accurate discrimination. The following MVA methods were used.

2.3.1. LDA

LDA is a technique used for studying the relationship between a set of predictors *versus* a categorical response; hence, the approach is applied in statistical analysis as a dimensionality reduction technique during classification [23,56]. In general, LDA finds linear combinations of features that describes or separates two or more groups of samples by maximising inter-group variance whilst minimising intra-group variance. The key advantages of using LDA are its simplicity, good classification when the input features are linearly separable, and finally, it is simple and fast to implement. It is reported that LDA does not cope very well when large numbers of groups are to be classified, and this is due to difficulties in complex decision boundaries or data with high variance [23,56]. Note in this study our classifications included 2, 7 or 14 groups.

The *lda* function of the MASS package "Support Functions and Data sets for Venables and Ripley's MASS" was used to perform LDA [57]; this package is built in R. The *predict* function has been used to assess prediction accuracy for LDA, and indeed, all the classifiers we used. Straightforward graphics of LDA technique is presented in Fig. S3 (SI).

2.3.2. PLS-DA

PLS-DA is a supervised multivariate classification method used to improve the separation between different classes of observations. This method is based on PLS regression of a set of categorical variables (usually describing group, or class membership) or a set of continuous predictor variables. The method is closely related to (PLS-1) which finds the relationship between predictor variables and dependent variables by building models, one for each response. PLS-DA (also referred to as PLS-2) instead of building separate models for each response, is able to deal with multiple responses simultaneously *via* application on the unfolded class matrix and transforming the class vector into a matrix of zeros and ones. For example, if there are three classes then these would be encoded as [1 0 0], [0 1 0], and [0 0 1] [21,49]. PLS-DA was also computed within R using "caret" package written by Kuhn [33].

2.3.3. SVM

SVM works by representing samples as points in space, optimally separated by hyper-planes so that samples of distinct classes are divided by margins that are as wide as possible. The points positioning on the boundaries of these margins are so called "support vectors". For prediction, new samples are then projected into the same space and assigned to belong to a class based on which side of the margin they fall. The performance of SVMs relies upon the kernel selection. In this study, "linear" kernel was used, and this was chosen as it is the most commonly kernel function used to map data into a space, where the classes are linearly separable [26,28,58]. Fig. S4 in SI is provided as an additional support for visualisation of this process. SVM was performed in e1071 package "Misc Functions of the Department of Statistics (e1071), TU Wien" version 1.6 [51].

2.3.4. Random forests

Random forests (RF) are statistical learning techniques used for the classification or regression and estimation of variable importance based on multiple decision trees. In general, the approach is based on classification accuracy derived from growing an ensemble of trees, and using this collection of solutions by allowing them to vote for the most suitable class based on a characterised training dataset [30,54]. A more detailed description of RF approach can be found in SI. The "randomForest" [35] package was implemented in R.

3. Results and discussion

As described above, the main aim of this study was to assess different variable selection and multivariate data analysis algorithms for the classification of bacteria from their Py-MS metabolic profiles. In order to achieve this, a proper validation procedure was implemented; this is described above and shown as flow diagrams in Figs. 1 and 2.

Initially, four different MVA algorithms were assessed with variable selection methods that are commonly implemented and considered specific to each algorithm. The classification task here was to separate accurately sporulated cells from vegetative biomass. During this process, variables that were specific for this physiological classification were ranked and used for the further analysis of whether it was possible to use these analytes for the speciation of bacteria (see Fig. 1) from: (1.1) spores alone; (1.2) vegetative cells; and (2) the combination of both spores and vegetative biomass. This would allow us to establish whether differences in Py-MS spectra are a result of the presence of spores or vegetative cells predominating the biochemistry. If it did, then we would be unable to effect the identification of the seven different *Bacillus* species. During this process all 150 mass ions were also used, as these would contain additional information that was not physiologically specific. The reason this is important is that in a real world scenario – for example *Bacillus* could be used for bioterrorism or may be present as a food pathogen in yoghurt – these will not be collected as pure spores or pure vegetative cells but as a mixture of the two.

3.1. Physiological characterisations and variable selection

Initially we used all the Py-MS data and generated four different MVA models to identify the physiological status of the bacteria. As

Table 1
Prediction accuracy (%) of bacterial physiological state (spores *versus* vegetative cells) for the validation and test sets. Also shown are the accuracies for the different number of variables included in the analysis (5 or all 150 variables).

Numbers of variables selected	Dataset	LDA	PLS-DA	SVM	RF	
					MDA	MDG
5	Validation	98.97	97.89	99.64	98.98	98.98
	Test	98.69	97.47	99.40	97.44	97.44
All	Validation	–[a]	99.90	99.92	99.57	99.57
	Test	–	99.84	99.80	99.14	99.14

[a] Not possible to compute due to *collinearity*.

shown in Table 1, the prediction accuracy of all models was >99.9% which confirmed earlier findings that the Py-MS spectra contain pertinent physiological information. The next stage was to rank the input variables (m/z) in order to select important variables with respect to this physiological classification.

Remarkably, only five top rank variables were needed in order to achieve the accurate classification of spores from vegetative cells (Table 1); this accuracy was >97.5% for the 'unseen' data used to test the models. This accuracy dropped to *ca.* 88% when just two variables were used and these were m/z 105 and m/z 76. These have both been previously reported as being important [17,36], and m/z 105 is a pyridine ketonium ion generated during pyrolysis and electron impact of the spore biomarker dipicolinic acid (pyridine-2,6-dicarboxylic acid; DPA) [59]. These two variables were selected as being important by all five of the variable selection approaches.

During the variable selection process, many iterations were performed based on bootstrap partitioning of the Py-MS data (Fig. 2). Therefore, the training sets generated by bootstrapping varies in every iteration, and the variables selected as being important could differ depending on which sample subset they were selected from of [42]. However, an ideal variable selection method should not be very sensitive to such re-partitioning of the data if the important variables discovered by such a method were genuinely important. A total number of 100 iterations were performed, the frequencies for each variable appeared as top 30, 15, 10 and 5 most important variables were calculated and the top 20% of all variables are reported in Table S1. It is interesting to see that random forests had been the least data-dependent variable selections. For the top 30 important variables, 18 of those had 100% chance of occurrence on the list of top 30 most important variables by using MDG criterion and 17 by using MDA criterion; PLS-DA–VIP had also been very stable with 16 variables always appearing in the top 30. SVM-RFE was, however, seemingly rather sensitive to the random re-sampling, only 6 variables had always been ranked as top 30, and 15 had been selected in more than 90% of the iterations. Stepwise forward selection LDA had the most inconsistent top ranking variable lists, where no single variable was constantly chosen as the important variable and merely three variables had >50% frequency of being selected as the important variables. Similar observations can also be seen on the top 15, 10 and 5 rank lists. Such poor reproducibility of the LDA method may be due to the high number of variables and collinearity between them. If the aim is to discover the real underlying biological reason behind the separation between the classes (*e.g.*, different bacterial species), then rank consistency/reproducibility is probably the most important aspect to consider. If for a particular algorithm top variables change every time when a different dataset is used, these are not considered to be reproducible (that is to say the order is not consistent) and cannot be considered robust enough to infer any biological meaning.

Using the feature selection algorithms in this study on this large Py-MS dataset [17] which contains 150 variables, resulted in a reduction to 30 important variables, with only a slight cost in overall classification accuracy. Fig. 3 reports these selected variables and there is good agreement with the selected variables across all methods. We observed that 10 variables have been selected in all approaches (highlighted in red in Fig. 3). Additionally, we highlight the variables selected *via* two evolutionary methods as being important: GA–BN [36] indicated with bold font face and genetic programming (similar method to GAs but uses a tree structure) [17] highlighted by asterisks.

3.2. Classification of bacilli speciation

Following the above variable selection process based on bacterial physiology, the ranked variables were then used for classification of the seven different *Bacillus* species. In order to assess the information content of the selected variables with respect to *Bacillus* speciation of spores and vegetative cells, we generated five classification models: these were based on the first five variables, variables 1–10, 1–15, and 1–30, as well as using all 150 mass ions. The results of these classification models are displayed in Tables 2–4, and these include the outputs for the speciation classification analysis from spores or vegetative cells individually, and combinations of both sporulated and vegetative biomass. These tables report the average classification accuracies from 100 iterations (Fig. 2 shows the pictorial representation of this validation process).

3.2.1. Classification of Bacillus species from spores

In the next step, we compared the species prediction accuracy from *Bacillus* spores. As can be observed in Table 2, as we increase the number of input variables from 5 to 30, we observed a gradual increase in prediction accuracy for SVM and RF from ~85–88% to ~93.5% which only increased to 95% when all 150 inputs were used. Whilst this increasing accuracy is also seen for LDA and PLS-DA it starts from a much lower value of 77% or 66%, respectively when 5 variables are selected. This is perhaps a reflection of the different five variables that are selected (Fig. 3, Table S1) rather than the classifier *per se* as all models have excellent classification accuracies when $\geq$15 are used.

3.2.2. Classification of Bacillus species from vegetative cells

Table 3 shows a comparison of the species classification ability of the same four classification methods for the analysis of vegetative cells. In general, similar results as described above for the four different classifiers are observed. We noted that the classification accuracy for speciation from vegetative cells is generally better than that from spores for all methods and all input selections. This perhaps

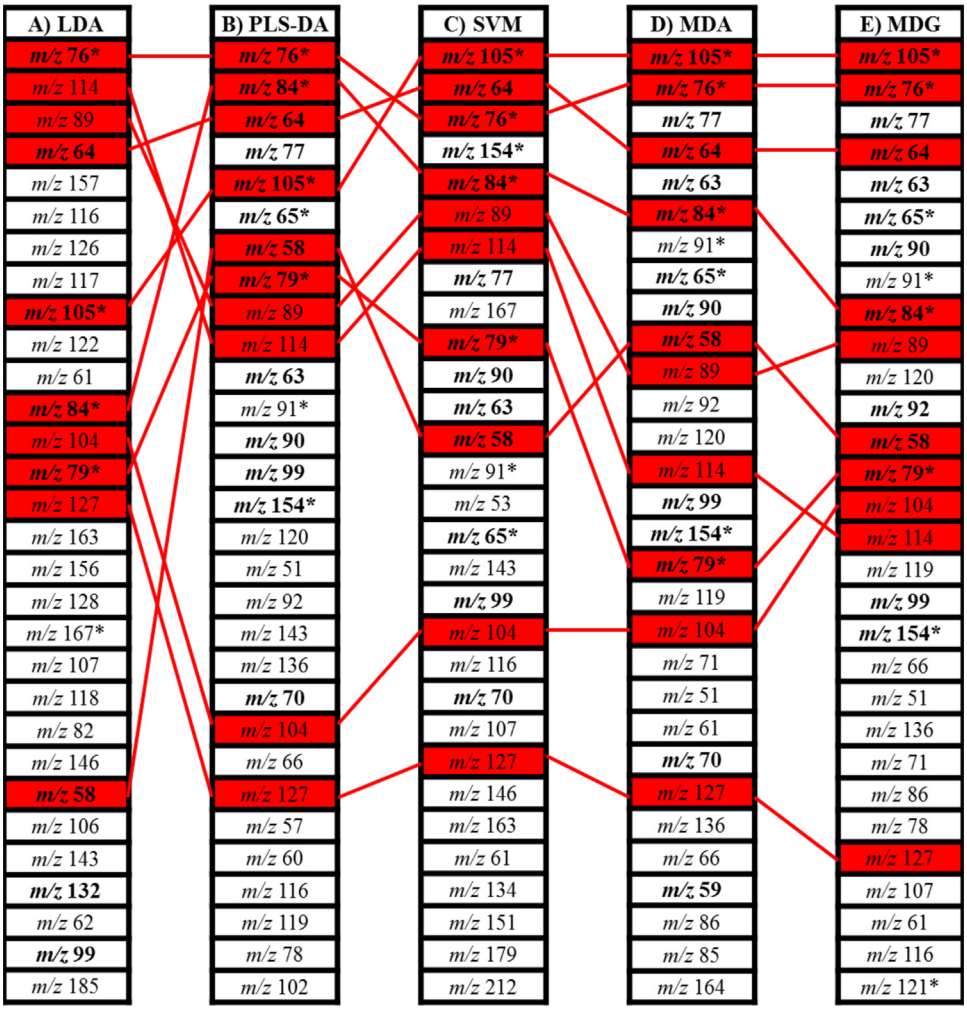


Fig. 3. The results of variable selection for the five feature selection approaches. These were calculated for objects produced by the training set for 100 iterations (the bottom LHS box in Fig. 2) after which the frequency of the variables selected was calculated and validated by: (A) stepwise forward variable selection for linear discriminant analysis; (B) PLS weighted sums of the absolute regression coefficients evaluated using *varImp* function for PLS-DA algorithm; (C) SVM recursive feature elimination; (D) MDA, mean decrease in accuracy provided by randomForest function; (E) MDG, mean decrease in Gini delivered by randomForest function. The lines joining the variables are those that are considered to be significant in all of the methods used. The bold features represent the variables selected by GA-BN in a previous study [36], and the variables with asterisks are those which were selected previously using genetic programming [17].

suggests that the Py-MS signals from vegetative cells have more species related information content than the signal from spores.

3.2.3. *Classification of* Bacillus *species from spores and vegetative cells combined*

In the above two classification systems for differentiation of the seven *Bacillus* species, we have used two separate models, one for spores (Fig. 1: model (1.1)) and the other for vegetative cells (Fig. 1: model (1.2)). This requires that we had a *priori* knowledge of whether we are performing Py-MS on a spore preparation or on vegetative cells. This is of course the case as our primary classifier (Fig. 1: model (1)) first differentiates the bacteria on the basis of their physiological status. Therefore, we finally combined physiological states with bacterial speciation

Table 2
The average prediction accuracy (%) of *Bacillus* species from spores for the validation and test sets. Displayed are the accuracies for the different number of variables included in the analysis (5, 10, 15, 30 and all 150).

Top variables selected	Dataset	LDA	PLS-DA	SVM	RF	
					MDA	MDG
5	Validation	77.39	65.88	85.04	88.27	
	Test	70.80	57.13	74.45	76.19	
10	Validation	84.57	84.98	91.55	90.69	90.46
	Test	84.14	77.92	82.85	81.34	80.92
15	Validation	89.87	86.93	92.14	92.02	91.59
	Test	83.46	80.04	85.25	83.78	81.63
30	Validation	92.44	91.94	93.48	93.36	93.69
	Test	86.14	85.93	86.65	85.97	85.50
All	Validation	–	95.53	95.00	94.58	
	Test	–	90.86	90.50	88.14	

Table 3
The average prediction accuracy (%) of *Bacillus* species from vegetative cells for validation and test sets. Displayed are the accuracies for the different number of variables included in the analysis (5, 10, 15, 30 and all 150).

Top variables selected	Dataset	LDA	PLS-DA	SVM	RF	
					MDA	MDG
5	Validation	85.37	68.82	83.34	86.53	
	Test	82.19	61.57	75.80	68.09	
10	Validation	88.69	84.00	92.50	92.97	91.47
	Test	82.94	77.13	86.32	86.83	83.04
15	Validation	89.33	90.83	93.70	93.93	94.23
	Test	81.70	84.11	88.76	87.23	87.59
30	Validation	94.56	93.01	95.04	95.18	94.80
	Test	88.55	87.83	90.57	89.40	89.54
All	Validation	–	96.44	95.62	94.67	
	Test	–	93.38	91.70	89.41	

Table 4
The average prediction accuracy (%) of *Bacillus* species from spores and vegetative cells combined for the validation and test sets. Displayed are the accuracies for the different number of variables included in the analysis (5, 10, 15, 30 and all 150).

Top variables selected	Dataset	LDA	PLS-DA	SVM	RF	
					MDA	MDG
5	Validation	78.56	49.72	77.96	87.08	
	Test	73.10	42.11	69.08	71.72	
10	Validation	86.50	69.89	90.27	91.35	90.45
	Test	81.98	63.17	82.50	82.72	80.96
15	Validation	88.00	80.91	92.11	92.79	92.56
	Test	81.47	73.66	86.05	85.14	83.88
30	Validation	94.63	90.26	93.83	93.79	93.90
	Test	90.53	83.04	87.35	87.13	86.79
All	Validation	–	95.11	95.23	94.36	
	Test	–	91.34	90.73	88.09	

which is reflected in 14 different classes (2 physiological states × 7 *Bacillus* species).

Despite the fact that a rather large number of classes are to be predicted, all classifiers work remarkably well, and when all variables are used prediction accuracies are above 95%. As described above, increasing the number of input variables from 5 to 30 also results in a concomitant increase in prediction accuracy. We note that the PLS-DA classifier when calibrated with only five inputs has a relatively low prediction accuracy of 50% compared with the other classifiers which are between 78 and 87%. Of course when 14 outputs are being predicted the 50% prediction accuracy from PLS-DA is encouraging, as a random classification is equivalent to a 1 in 14 chance of accurate bacterial species and physiological status prediction and is only 7%.

4. Concluding remarks

We have presented five variable selection approaches that were applied to pre-select the highly collinear metabolomics data generated by Py-MS prior to applying four different classification algorithms. Double cross-validation using bootstrapping with replacement was employed to split the MS dataset into training, validation and test subsets. This we believe is essential in order to assess objectively the predictive strength among the different MVA algorithms.

The classification scheme involved the assessment of both physiological status of the bacteria as well as then identifying which species of *Bacillus* they belong to. This was either performed using a hierarchical classifier: assessment of spores *versus* vegetative cells followed by bacterial speciation (Fig. 1: LHS branch); or a single classifier containing 14 groups for predicting these two characteristics simultaneous (Fig. 1: RHS branch).

Irrespective of the classification route all bacteria were assessed with ~95% accuracy and this was for all of the models performed. In practice, the classification model that uses a small number of variables should be preferred, because it is less computationally expensive to run and more parsimonious, and hence, more robust [15]. In addition, a smaller set of variables would be much easier to interpret and ease the task of (biochemical/physiological) knowledge discovery. For all the variable selection methods, we employed in this study, random forests appeared to be the most stable method while stepwise forward selection LDA had the worst reproducibility.

This paper has concentrated on chemometric methods applied to physiochemical data generate by Py-MS, and these chemometric methods could be equally applied to other bioanalytical methods used to classify bacteria – which include infrared and Raman spectroscopy as well as other forms of mass spectrometry (*i.e.* ESI-MS and MALDI-MS) [4,5]. Traditional microbiology methods used to detect spores include optical microscopy, while the classification of different bacteria is routinely achieved using morphological characteristics, both in terms of colony shape and cell morphology (rod-shaped, cocci, *etc.*), as well as a plethora of biochemical tests. Both of these require considerable expertise and the latter biochemical tests require incubation (typically 24–48 h) which slow down bacterial classification. Therefore, Py-MS (and indeed other bioanalytical methods) offers the advantage of speed for differentiation of spores *versus* vegetative cells and for bacterial classification of the correct genus and species.

In conclusion, whilst this paper has used pyrolysis-mass spectrometry as the test dataset, the variable selection methods and supervised learning approaches developed are equally applicable to any high dimensional multivariate or megavariate data. This would include MS- or NMR-based metabolomics,

spectroscopic investigations involving FT-IR and Raman spectroscopy, as well as data generated from chemical sensors. These will be an area of future investigations.

Acknowledgments

The authors would like to thank PhastID (grant agreement no: 258238) which is a European project supported within the Seventh Framework Programme for research and Technological Development and funding the studentship for PSG.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.aca.2014.03.039.

References

[1] D.B. Drucker, L.F. Gibson, Microbios 33 (1982) 93.
[2] W.J. Irwin, Journal of Analytical and Applied Pyrolysis 1 (1979) 3.
[3] R. Goodacre, M.J. Neal, D.B. Kell, Analytical Chemistry 66 (1994) 1070.
[4] D.I. Ellis, W.B. Dunn, J.L. Griffin, J.W. Allwood, R. Goodacre, Pharmacogenomics 8 (2007) 1243.
[5] D.I. Ellis, V.L. Brewster, W.B. Dunn, J.W. Allwood, A.P. Golovanov, R. Goodacre, Chemical Society Reviews 41 (2012) 5706.
[6] D. Melucci, S. Fedi, M. Locatelli, C. Locatelli, S. Montalbani, M. Cappelletti, Current Drug Targets 14 (2013) 1023.
[7] R. Goodacre, E.M. Timmins, R. Burton, N. Kaderbhai, A.M. Woodward, D.B. Kell, P.J. Rooney, Microbiology – UK 144 (1998) 1157.
[8] J.G. Green, J.M. Potter, Journal of Analytical and Applied Pyrolysis 91 (2011) 40.
[9] R. Goodacre, D.B. Kell, G. Bianchi, Nature 359 (1992) 594.
[10] R. Goodacre, D.B. Kell, G. Bianchi, Journal of the Science of Food and Agriculture 63 (1993) 297.
[11] R. Goodacre, D.B. Kell, Current Opinion in Biotechnology 7 (1996) 20.
[12] D. Cauzzi, G. Chiavari, S. Montalbani, D. Melucci, D. Cam, H. Ling, Journal of Cultural Heritage 14 (2013) 70.
[13] G. Chiavari, S. Montalbani, V. Otero, Rapid Communications in Mass Spectrometry 22 (2008) 3711.
[14] G. Chiavari, S. Montalbani, S. Prati, Y. Keheyan, S. Baroni, Journal of Analytical and Applied Pyrolysis 80 (2007) 400.
[15] M.B. Seasholtz, B. Kowalski, Analytica Chimica Acta 277 (1993) 165.
[16] S.J. Deluca, E.W. Sarver, K.J. Voorhees, Journal of Analytical and Applied Pyrolysis 23 (1992) 1.
[17] R. Goodacre, B. Shann, R.J. Gilbert, E.M. Timmins, A.C. McGovern, B.K. Alsberg, D.B. Kell, N.A. Logan, Analytical Chemistry 72 (2000) 119.
[18] A.P. Snyder, J.P. Dworzanski, A. Tripathi, W.M. Maswadeh, C.H. Wick, Analytical Chemistry 76 (2004) 6492.
[19] D. Broadhurst, R. Goodacre, A. Jones, J.J. Rowland, D.B. Kell, Analytica Chimica Acta 348 (1997) 71.
[20] B.K. Alsberg, D.B. Kell, R. Goodacre, Analytical Chemistry 70 (1998) 4126.
[21] M. Barker, W. Rayens, Journal of Chemometrics 17 (2003) 166.
[22] W. Cheung, Y. Xu, C.L.P. Thomas, R. Goodacre, Analyst 134 (2009) 557.
[23] R.A. Fisher, Annals of Eugenics 7 (1936) 179.
[24] T. Adam, T. Ferge, S. Mitschke, T. Streibel, R.R. Baker, R. Zimmermann, Analytical and Bioanalytical Chemistry 381 (2005) 487.
[25] F. Girosi, M. Jones, T. Poggio, Neural Computation 7 (1995) 219.
[26] V.N. Vapnik, IEEE Transactions on Neural Networks 10 (1999) 988.
[27] C.W. Hsu, C.J. Lin, IEEE Transactions on Neural Networks 13 (2002) 415.
[28] C.J.C. Burges, Data Mining and Knowledge Discovery 2 (1998) 121.
[29] S. Zomer, R.G. Brereton, J.F. Carter, C. Eckers, Analyst 129 (2004) 175.
[30] L. Breiman, Machine Learning 45 (2001) 5.
[31] R.D.C. Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, 2008.
[32] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press A Harcourt Science and Technology Company, London, 2000.
[33] M. Kuhn, Journal of Statistical Software 28 (2008) 1.
[34] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Machine Learning 46 (2002) 389.
[35] A. Liaw, M. Wiener, R News 2 (2002) 18.
[36] E. Correa, R. Goodacre, BMC Bioinformatics. 12 (2011) 33.
[37] L.B. Booker, D.E. Goldberg, J.H. Holland, Artificial Intelligence 40 (1989) 235.
[38] J.H. Holland, Scientific American 267 (1992) 66.
[39] J.R. Koza, Statistics and Computing 4 (1994) 87.
[40] E. Charniak, AI Magazine 12 (1991) 50.
[41] N. Friedman, D. Geiger, M. Goldszmidt, Machine Learning 29 (1997) 131.
[42] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J. P.M. van Duijnhoven, F.A. van Dorsten, Metabolomics 4 (2008) 81.
[43] B. Efron, Annals of Statistics 7 (1979) 1.
[44] B. Efron, G. Gong, The American Statistician 37 (1983) 36.
[45] R.G. Brereton, Trends in Analytical Chemistry 25 (2006) 1103.
[46] R.G. Brereton, Chemometrics: Data Analysis for the Laboratory and Chemical Plant, John Wiley & Sons, Chichester, 2003.
[47] A.J. Miller, Journal of the Royal Statistical Society: Series A (Statistics in Society) 147 (1984) 389.
[48] R.R. Hocking, Biometrics 32 (1976) 1.
[49] M. Haenlein, A.M. Kaplan, Understanding Statistics 3 (2004) 297.
[50] K.B. Duan, J.C. Rajapakse, H.Y. Wang, F. Azuaje, IEEE Transactions on NanoBioscience 4 (2005) 228.
[51] A. Karatzoglou, D. Meyer, K. Hornik, Journal of Statistical Software 15 (2006) 1–28.
[52] D.R. Cutler, T.C. Edwards, K.H. Beard, A. Cutler, K.T. Hess, Ecology 88 (2007) 2783.
[53] J.I. Gastwirt, Review of Economics and Statistics 54 (1972) 306.
[54] T.K. Ho, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 832.
[55] A. Liaw, M. Wiener, R News 2 (2002) 5.
[56] T. Hastie, A. Buja, R. Tibshirani, Annals of Statistics 23 (1995) 73.
[57] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, Springer, New York, 2002.
[58] S. Zomer, M.D.N. Sanchez, R.G. Brereton, J.L.P. Pavon, Journal of Chemometrics 18 (2004) 294.
[59] D.P. Cowcher, Y. Xu, R. Goodacre, Analytical Chemistry 85 (2013) 3297.