Norges miljø- og
biovitenskapelige
universitet

Master’s Thesis 2016   60 ECTS
Department of Mathematical Sciences and Technology

Multivariate Classification Methods for Spectroscopic Data with Multiple Class Structure

Denis Tafintsev
Mathematical, Physical and Computational Sciences

Multivariate Classification Methods for Spectroscopic Data with Multiple Class Structure

Denis Tafintsev

December 2016

Abstract

The classification of microorganisms is an important task in many fields such as food production, medicine, biotechnology. Fourier transform infrared (FTIR) spectroscopy can provide comprehensive biochemical information about microorganisms via spectra. To extract the information, an appropriate chemometrics technique is needed to treat the data and get reliable classification results. From the very beginning it was known that utilizing hierarchical structure of the data is an advantage but might be a tedious and time consuming procedure. In this study we evaluate the best way for setting up a classifications scheme to identify microorganisms by FTIR spectroscopy. In this context our task was to classify ten different genera of food spoilage yeasts, which were cultivated in five different media and subsequently analyzed by (FTIR) spectroscopy. The methods, which were used in this study, are: principal component analysis (PCA), partial least squares discriminant analysis (PLSDA), Fisher liner discriminant analysis (FLDA), PLSDA and FLDA combined with HCA, PLSDA and FLDA combined with a one-versus-all (OVA) approach, PLSDA and FLDA combined with a one-versus-one (OVO) approach, and random forest (RF). The last method showed the best performance among the all methods we used. A validation success rate (SR) achieved by RF is equal to 97.5% for one of the media. The other successful methods are PLSDA combined with HCA and PLSDA applied directly to ten groups with SRs equal to 96.3% and 94.4%, respectively. Our results suggest that RF can be used for rapid identification of microorganisms even without utilizing a hierarchical structure in the data and can perform very accurately. Moreover, when using information from other blocks of data representing different cultivation media, the performance of RF was improved.

Table of Contents

List of Figures

List of Tables

1 Introduction

Due to a rapid development of analytical techniques during the last decades, enormous amounts of data are produced in the fields of life sciences. One of the rapidly growing classes of methods for analyzing different types of biological materials is spectroscopic techniques. For example, in the field of biochemistry methods such as infrared, Raman, MALDI-TOF, and fluorescence spectroscopy have been found to be very useful for analyzing biomolecules, including large molecules such as carbohydrates, lipids, proteins, and nucleic acids. In the field of microbiology, vibrational spectroscopy is widely used for identification, differentiation and in general characterization of microbial cells (Naumann, Helm, & Labischinski, 1991). The major advantage of these techniques is that they are non-destructive, fast, and that they provide rich information about chemical composition and chemical structure of the samples via spectra.

A very popular spectroscopic method for classification of microorganisms such as bacteria, yeasts, and fungi is a Fourier-transform infrared (FTIR) spectroscopy (Helm, Labischinski, Schallehn, & Naumann, 1991). FTIR spectroscopy is a vibrational spectroscopy technique, which is suitable for analyzing solid, liquid, and gas samples. FTIR spectra obtained from applying FTIR spectroscopy to microorganisms provide biochemical information about the chemical structure of the samples to be investigated. This information is a fingerprint-like signature, which allows identifying samples on genus, species, and strain level of taxonomy. In addition, FTIR spectra provide an own phenotypic structure and can potentially serve on its own or combined with other genetic and phenotypic techniques for developing a taxonomy (Duygu, Baykal, Acikgoz, & Yildiz, 2009).

Numerous statistical and chemometrics classification techniques for multivariate discrimination and classification can be found in the literature. Which method is to be applied in a certain situation may simply depend on which method is popular in a scientific field or community. Methods used for classification and discrimination of microorganisms by FTIR spectroscopy can be divided into unsupervised and supervised approaches. Unsupervised methods do not used any information about classes of microorganisms as opposed to supervised methods.

One of the most commonly used unsupervised methods in the field of multivariate analysis is Principal Component Analysis (PCA) (Hastie, Tibshirani, & Friedman, 2001). It is usually used either as a method to explore the data by means of loadings and scores plots or as a dimensionality reduction technique to reduce the dimensionality of classification problems**.** Hierarchical Cluster Analysis (HCA) is another widely spread unsupervised technique for analyzing FTIR spectra obtained from microorganisms, when an overview of the data complexity and structure is needed. Dendrograms obtained by HCA are used as hierarchical structures obtained by a data-driven approach. Such unsupervised methods are often used to explore data, to find grouping patterns present in a dataset, and to detect outliers before a classification scheme is established by supervised methods (Goodacre, Timmins, Rooney, Rowland, & Kell, 1996; Oust, Møretrø, Kirschner, Narvhus, & Kohler, 2004).

One of the first supervised approaches that was introduced for classification of microorganisms by spectroscopic data is ANN (Udelhoven, Naumann, & Schmitt, 2000; Wenning, Buchl, & Scherer, 2010). An ANN feed-forward network is a mapping of the inputs variables to the output classes. Typically, a three-layer network is a good model in most cases. In a three-layer network the input layer consists of variables (wavenumbers), the hidden layer represents a number of neurons and the output layer contains of the class identifiers. The connection functions are called activation functions and are often sigmoidal functions (Bishop, 1995). The use of ANN is very popular in the field of microbiology due to availability of a software package based on ANN and developed by Udelhoven, Novozhilov, and Schmitt (2003).

PLSDA is another example for a supervised method, which became very popular for classification problems in FTIR spectroscopy of microorganisms (Oust et al., 2004; Preisner et al., 2008; Coutinho, Sá-Correia, & Lopes, 2009). PLSDA finds the covariation between a data matrix X (FTIR data) and a reference data matrix Y (class identifiers) in order to find best class separation rules (Martens & Næs, 1989). In addition, PLSDA provides visualization opportunities by plotting scores and analyzing regression coefficients, which can be used to understand and interpret the classification results (Zimmermann, Tafintseva, Bağcıoğlu, Berdahl, & Kohler, 2016).

Another technique which is used for discrimination of the microorganisms in the field of biospectroscopy is linear discriminant analysis (LDA). LDA was developed by Fisher (1936) and became very popular, because the basic idea is intuitive, and the method is mathematically robust. According to Fisher’s criterion, a good separation can be found, when the ratio of a between-class scatter matrix to a within-class scatter matrix is maximized. LDA combined with HCA was used by Maquelin et al. (2003) to identify bacteria and yeast using data obtained from Raman and FTIR spectroscopy.

The Random Forest method (Breiman, 2001) is an emerging technique for classification problems and mostly used for omics data (Touw et al., 2013). In order to grow each tree in an ensemble, randomly selected samples are taken with replacement by a procedure called bootstrap (Breiman, 1996). To split each node in a tree, a small group of randomly selected variables are tested and the best variable is chosen. This random selection makes the chance of correlations among trees very low and prevents overfitting of the model. For classification, each sample to be classified is run throw the forest and a final decision is made by majority voting. Using the majority voting scheme allows reducing the overall classification error compared to a single decision tree (Maguire et al., 2012). RF can be an appropriate choice as a method that can solve a broad range of classification problems (Fernández-Delgado, Cernadas, Barro, & Amorim, 2014). In application to FTIR data on biofuels it showed a higher performance than LDA (Ollesch et al., 2013).

Two other methods used for classification which are based on PCA are SIMCA (Martens & Næs, 1989) and KNN (Adams, 1995). In SIMCA a PCA model is established on the data of each class and a new sample is projected into these PCA spaces in order to check distances to all the models. The unknown sample is assigned to the closest class among all. KNN method is a classification method based on PCA scores of the entire training data. PCA is used here to reduce the dimensionality and extract the most important information from the data. An unknown sample is projected into the PCA space and K neighboring points are found in the space, where K is a predefined parameter. The sample is assigned to a class by majority voting among the neighbor points. These methods were previously used for classification of microorganisms (Kansiz et al., 1999; Preisner et al., 2008).

Another very simple and straightforward method is based on calculation of Pearson correlation coefficients between spectra (Helm et al., 1991). It is a simple look-up method when a spectrum of an unknown sample is compared to all spectra in a reference library and the sample is identified to the group of closest samples. This approach is simple and does not require establishing any calibration models, thus facilitating the process of extending reference libraries of microorganisms. The method is used for classification of microorganisms by FTIR spectroscopy (Oberreuter, Seiler, & Scherer, 2002)

The identification of microorganisms by any phenotypic method, including FTIR, is complicated since many groups of microorganisms are presented, which are often very closely related. In addition, FTIR spectra represent high dimensional data characterizing each sample by thousands of variables. Thus, the data structure is very complex. One approach to avoid a high complexity of the classification problem is to utilize a hierarchical structure. Such a structure can either be available for example when a phylogenetic tree is available for the microorganisms given by microbial taxonomy. If such a structure is not available, it may be still advantageous to establish a data-driven hierarchical tree to reduce the multiclass classification problem to two-class (or few-classes) classification problems in each node of a classification tree. An example when a phylogenetic tree is used for classification analysis is presented by Liland, Kohler, and Shapaval (2014) where the authors classify moulds by FTIR data and use PLSDA classifier in each node of the tree. Udelhoven et al. (2000) used a phylogenetic structure of a data to set up a hierarchical classification system to identify bacteria and yeasts by ANN.

An example of a phylogenetic tree of moulds samples with four levels such as division, class, genus, and subgenus is presented in Fig. 1. In a classification situation, we can either use this hierarchical structure information given by taxonomy and build a classification model in each node or use classifiers on the entire level considering the classes on that level without taking into account the hierarchical structure. For example in Fig. 1 without using phylogenetic information to classify samples on the genus level, one classifier is needed to assign each sample into one out of nine genera. We can also build our own hierarchy using the data-driven approach and reveal the phenotypic structure that is available in the data. A problem of establishing classification models utilizing hierarchical structures becomes even more difficult when several phenotypic methods are used to identify microorganisms: if more than one phenotyping method is used, then multiple classifiers could be used at each node of a tree. How to combine such data is another very important question in this case. Thus, it is very tedious and time consuming to set up a classification scheme based on hierarchical structure (Liland et al., 2014)

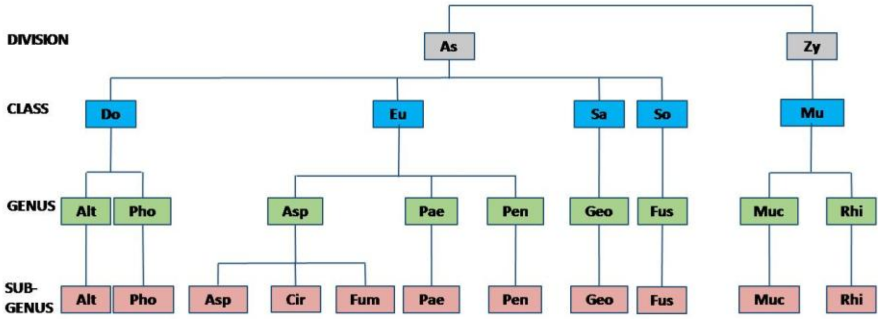


Figure 1: Example of a phylogenetic tree of moulds with division, class, genus and sub-genus as taxonomic levels.

The establishment of a classification tree is a time-consuming process. If a phylogenetic tree is readily available, it needs to be evaluated carefully if a hierarchy can be set up following this logic, or if it is more advantageous to set up new nodes in order to take into account the tree structure inherent in the data. The main goal of this study is to find the best way for setting up a classifications scheme for identifying microorganisms by FTIR spectroscopy. In order to investigate different ways to set up classification trees, we will consider different approaches for classification of microorganisms into a high number of groups based on FTIR spectroscopy. We will investigate the use of both single classifiers performing a classification into all classes at once and classifications where a classification tree is established. As an example for a one-classifier model, we will apply LDA and PLSDA. In order to investigate the possibility for establishing a classification tree we consider HCA method.

In this study we work with FTIR spectra obtained from food spoilage yeasts, which were analyzed and published by Shapaval et al. (2013). The yeasts were cultivated in five different media and thus we have five different blocks of FTIR data, each of them referring to a different cultivation medium. Different media contain different nutrients resulting in different phenotypic characteristics of the yeast cells (Shapaval et al., 2013). The data set consists of ten different genera. The phylogenetic tree structure is not available for this dataset.

The thesis is organized in the following way. The information about yeasts used for the analysis and FTIR data acquisition are given in section “Materials and equipment”. Classification and other methods used in this study are presented in the section “Methods”. The results of classification performances of all the methods are presented, compared, and discussed in section “Results and Discussion”. The conclusions are summarized in section “Conclusion”. Finally, section “Appendix” contains figures and graphs summarizing the results of classification performances of all the aforementioned methods applied to other four cultivation media, which were available in the study.

2 Materials and Equipment

2.1 Food Spoilage Yeast

The original dataset for this study included 12 different genera and 91 food spoilage yeast strains (Shapaval et al., 2013). In Table 1 the information about genera and number of species presented in each genus is shown.

Table 1: The names of the yeast genera and the number of species within each genus are presented.

Class	Genus name	Genus abbreviation	Number of species
1	*Candida*	*Can*	15
2	*Clavisporum*	*Cla*	1
3	*Debaryomyces*	*Deb*	1
4	*Hanseniaspora*	*Han*	3
5	*Issatchenkia*	*Iss*	1
6	*Lodderomyces*	*Lod*	1
7	*Metschnikowia*	*Met*	2
8	*Pichia*	*Pic*	6
9	*Rhodotorula*	*Rho*	4
10	*Saccharomyces*	*Sac*	3
11	*Torulaspora*	*Tor*	1
12	*Zygosaccharomyces*	*Zyg*	3

For growing food spoilage yeasts, five cultivation media were used coded as SAB, YPD, YMB, SD, and YEPD. Therefore five data subsets were obtained, one for each medium.

2.2 Equipment and Experimental Framework

To perform FTIR measurements, a High Throughput Screening eXTention (HTS-XT) unit combined with a Tensor 27 spectrometer (Bruker Optik GmbH, Germany) were employed. Each spectrum was recorded in a wavenumber range between 4000 and 500 cm^{-1} and 64 scans were averaged. More detailed information about growing conditions and sample preparation can be found in the paper of Shapaval et al. (2013).

Six Bioscreen runs which correspond to independent growth experiments were performed for each yeast strain grown on each medium. Two microcultivations, which correspond to biological replicates, were carried out in each Bioscreen run. Thereafter, from each biological replicate two samples were obtained for FTIR measurements to which we refer as technical replicates. The Opus software was used to identify bad quality spectra, which were subsequently removed from the dataset (Shapaval et al., 2013). Finally, the average number of spectra in each medium was equal to 2250.

For the data analysis Matlab R2013a (The MathWorks, Inc.) was used installed on Windows PC with 2.4 GHz double core processer and 4 GB RAM.

3 Methods

3.1 Data Preparation

At first, two groups with low sample size were removed from the dataset (throughout the thesis the words genus and group will be used interchangeably). Thereafter, the dataset was divided into calibration and validation subsets. The first Bioscreen run was used as an independent test set, which was kept aside during models' training. The other runs were used to calibrate the models. Each Bioscreen run is considered an independent experiment since the different Bioscreen runs were performed on different days.

3.2 Preprocessing

3.2.1 Averaging technical replicates

As was mentioned above two technical replicates were obtained for each biological replicate. It is done for several reasons: 1) to increase chances of obtaining good spectra which will go through the quality control; 2) to obtain more representative spectrum of a sample after averaging the two technical replicates; 3) to be able to use technical replicates in a majority voting scheme to improve prediction (Liland, Kohler, & Shapaval, 2014). In this study the technical replicates were averaged resulting in approximately 960 spectra in each medium.

3.2.2 Savitzky-Golay (SG) numerical algorithm

Applying the SG numerical algorithm to a spectroscopic data is a commonly used preprocessing step to inhibit spectral nose and increase signal properties relevant for further analysis. Moreover it can be used as bandpass filter allowing adjusting parameters such as window size, polynomial degree, and an order of a derivative to emphasize chemical features in a spectrum (Zimmermann & Kohler, 2013). The idea of SG procedure is that it approximates a spectrum within a moving window with a predetermined size by a polynomial using least squares criterion (Savitzky & Golay, 1964). When a value of a central point in the window is obtained a procedure repeats by moving the window one point further. Instead of calculating all coefficients of a fitted model, Savitzky and Golay (1964) suggested, to find a central point in the window, to use a set of integers in a weighting function and evaluate the central point by a convolution procedure. Furthermore, if observations are equally spaced, a set of integers can be found for calculation of derivatives of a least squares function as well. These integers do not depend on values of data points in the moving window and can be computed once for a particularly order of a polynomial and a specified window size. Thus, a computing time is drastically decreased using the SG numerical algorithm compared to conventional methods of calculating least squares fitted values and subsequently derivatives.

Utilizing the SG procedure, we computed the first order derivative of the spectra by using a window size of nine points and a polynomial of a third degree. The same parameters were used by Shapaval et al. (2013).

3.2.3 Selection of a spectral range

The selection of a spectral range is an important step in the preprocessing that allows finding relevant variables for data analysis and model training. Using spectral regions which do not contribute to discrimination or classification will unnecessary increase a computation time and lead to overfitting. The typical bands presented in a spectrum obtained from biological samples are associated with carbohydrates (1200 – 900 cm^{-1}), proteins (1700 – 1500 cm^{-1}), lipids (1760 – 1730 cm^{-1}) & (2950 – 2800 cm^{-1}), fatty acids (3000 – 2800 cm^{-1}), and water (3600 – 3000 cm^{-1}) (Zimmermann & Kohler, 2013). The region 1500 – 500 cm^{-1} is usually called a fingerprint region and the region 4000 – 1500 cm^{-1} is called a functional group region. We were interested to use information obtained from the spectra in both fingerprint and functional group regions. At the same time the water region was out of our because it does not provide information which is useful for discrimination and classification purposes. Furthermore the region 2800 – 1800 cm^{-1} is quite "flat", i.e. no bands present. Thus, the total spectral range used for the analysis was chosen to be 3100 – 2800 cm^{-1} & 1800 – 700 cm^{-1} as the one which contains relevant information for identification of microorganisms.

3.2.4 Extended multiplicative signal correction (EMSC)

A frequently used model-based preprocessing method in the field of vibrational spectroscopy is EMSC (Martens & Stark, 1991). This method allows separating physical and chemical variations presented in a spectrum by statistical estimation of parameters involved in a mathematical model representing the spectrum. Subsequently it is possible to analyze different variations independently.

The main parameter in a multiplicative signal correction (MSC) model (Ilari, Martens, & Isaksson, 1988) is a reference spectrum, which is obtained as mean of all spectra in the dataset or could probably be chosen according to some other logic. For each spectrum the MSC model looks like:

$$A_i(\tilde{v}) = \ a_i + b_i \cdot m(\tilde{v}) + E_i(\tilde{v}), \tag{1}$$

where a_i corresponds to a constant baseline effect, b_i corresponds to a multiplicative factor, $m(\tilde{v})$ is a reference spectrum, and $E_i(\tilde{v})$ is a residual term which corresponds to variations in a spectrum i that were not modeled.

When the parameters a_i and b_i are estimated by least squares regression, the corrected spectra are calculated by:

$$A_{i,corr}\,(\tilde{v}) = (A_i(\tilde{v}) - a_i)/b_i. \tag{2}$$

Eq. 1 takes into account a constant baseline effect. When non-constant baseline effects are present as for example in Raman spectroscopy (Kohler, Kirschner, Oust, & Martens, 2005), the MSC model can be further extended by adding linear and quadratic terms in the model to remove nonlinear baseline effects. The extended version is then called Extended Multiplicative Signal correction (EMSC) (Martens & Stark, 1991). The model can be given by:

$$A_i(\tilde{v}) = \ a_i + b_i \cdot m(\tilde{v}) + d_{1,i}\tilde{v} + d_{2,i}\tilde{v}^2 + E_i(\tilde{v}), \tag{3}$$

where $d_{1,i}\tilde{v}$ corresponds to linear baseline effects and $d_{2,i}\tilde{v}^2$ corresponds to quadratic baseline effects. This kind of model is called basic EMSC model.

After estimation of the parameters the corrected spectra is given by:

$$A_{i,corr}\,(\tilde{v}) = (A_i(\tilde{v}) - a_i - d_{1,i}\tilde{v} - d_{2,i}\tilde{v}^2)/b_i. \tag{4}$$

In this study we used the basic EMSC model for the correction of baseline and multiplicative effects. Once the EMSC model is established for a calibration dataset, the same model is used to correct the validation dataset. We applied first the SG numerical algorithm and then the EMSC normalization to the dataset, which was suggested as the most effective order by Zimmermann and Kohler (2013).

3.3 Principal Component Analysis (PCA)

PCA is one of the oldest and the most commonly used technique in a field of multivariate data analysis and multivariate statistics (Jackson, 1991). PCA allows extracting the most important information from a data by means of maximizing the explained variance. Furthermore, PCA helps to uncover hidden grouping patterns in a sample and a variable space by exploring score and loading plots, respectively. In case when a dimensionality reduction is necessary, PCA can do it by projecting the data set onto principal component directions.

Applying PCA, a data matrix can be represented as follows:

$$X = TP^T + E, \tag{5}$$

where X is a mean-centered data matrix with m rows (samples) and n columns (variables), T is an $m \times a$ matrix of scores, P is an $n \times a$ matrix of loadings, E is an $m \times n$ matrix of residuals, and a is a number of PCs used to approximate the data matrix. The columns of the matrix T are orthogonal while the columns of the matrix P are orthonormal.

The matrices T and P can be calculated applying singular value decomposition (SVD) of the data matrix X (Golub & Reinsch, 1970). If X is a real $m \times n$ matrix, then it can be decomposed as:

$$X = USV^T, \tag{6}$$

where U is a $m \times m$ orthogonal matrix, which consists of eigenvectors of XX^T, S is an $m \times n$ diagonal matrix, which consists of singular values of X, and V is an $n \times n$ orthogonal matrix, which consists of eigenvectors of X^TX.

If the rank of the matrix X is equal to r, then $a \leq r$. By using Eq.6, it can be shown that $T = U_aD$ and $P = V_a$, where U_a is $m \times a$ matrix, D is an $a \times a$ diagonal matrix with nonzero entries, and V_a is an $n \times a$ matrix.

In our study we used score plots to reveal and analyze grouping patterns in the data by usually plotting the first three principal components (PCs) in different combinations, which often explain most of the variance represented by the data matrix X. Higher components were also examined, but not shown because they did not show useful grouping patterns.

3. 4 Partial Least Squares Discriminant Analysis (PLSDA)

One of the classifiers, which were employed during this study, was PLSDA. Compared to unsupervised PCA, PLSDA uses label information about groups presented in the data. It has become a valuable technique in many situations in chemometrics (Höskuldsson, 1988). This method is a combination of PLS regression (PLSR) with discrimination rules designed for classification (Ballabio & Consonni, 2013). The idea of this approach is to find PLS components, which uncover the main covariation pattern within and between data matrices X and Y. Matrix Y consists of n rows (number of samples) and g columns (number of groups). Each row in Y contains 0s and 1s depending on a membership of a g-th group. Once PLS components are found by the non-linear iterative partial least squares (NIPALS) algorithm they can be used for estimation of regression coefficients by the following equation:

$$B = W(PW)^{-1}Q, \tag{7}$$

where W is a matrix of X loading weights, P is a matrix of X loadings, and Q is a matrix of Y loadings.

After the regression coefficients are found they can be used for prediction in the PLSR approach or for classification in PLSDA. A regression model is given by:

$$Y = B_0 + XB + E_y, \tag{8}$$

where E_y is a matrix of residuals and B_0 is given by:

$$B_0 = \bar{y} - \bar{x}B, \tag{9}$$

where $\bar{y}$ and $\bar{x}$ representing the mean of Y and X matrices, respectively.

In PLSDA a sample is assigned to a particular group, if an estimated value of Y is closer to 1 for that group than for other groups.

A more detailed description of PLS, particularly the NIPALS algorithm, PLS components interpretation, and usages of scores and loadings plots can be found in Höskuldsson (1988) and Wold, Sjöström, and Eriksson (2001).

3.5 Fisher Linear Discriminant Analysis (FLDA)

Another prominent classification technique that was used during this study is the Fisher Linear Discriminant Analysis (FLDA). It finds a classification line for two-class problem or a hyper-plane for multiclass problem the projection on which optimally separates the classes using variances of these classes. Fisher's criterion to be maximized in order to find an optimal separation in a multiclass problem is given by:

$$J(V) = \det(V^T S_B V) \, / \det(V^T S_W V), \tag{10}$$

where S_B is a between-class scatter matrix, S_W is a within-class scatter matrix, and V is an optimal projection matrix, which consists of eigenvectors of a matrix ${S_W}^{-1} S_B$.

The within-class scatter matrix is given by:

$$S_W = \sum_{i=1}^{g} \sum_{x_k \in class\ i} (x_k - \mu_i)(\, x_k - \mu_i)^T, \tag{11}$$

where x_k is a sample number k, μ_i is a mean of a group i, and g is a number of groups in a dataset.

The between-class scatter matrix is given by:

$$S_B = \sum_{i=1}^{g} n_i (\mu_i - \mu)\,(\mu_i - \mu)^T, \tag{12}$$

where n_i is a number of samples in a group i and μ is a mean of all samples.

In other words to find an optimal separation in the multiclass problem we need to seek a transformation matrix V, which maximizes a between-class scatter matrix and minimizes a within-class scatter matrix (Sugiyama, 2007). Once the projection matrix is found and a transformation of the data is completed, samples are classified by computing Euclidean distances from each sample to group means. A sample is assigned to a specific group, if the sample has the shortest distance to the group mean of the respective group.

3.6 Hierarchical Cluster Analysis (HCA)

HCA is a widely used technique in many fields such as biology, medicine, business and others (Tan, Steinbach, & Kumar, 2006). In FTIR spectroscopy it is used to find similarities or dissimilarities between samples of microorganisms (Wenning & Scherer, 2013).

We used an agglomerative hierarchical cluster procedure to establish a hierarchical tree structure. The latter is used for classification purposes followed by classifiers such as PLSDA or FLDA in each node. The agglomerative approach is also called as a “bottom-up” approach. It starts from a single observation considered as a cluster on a lowest level. The second step is to merge the two closest (with respect to chosen metric) clusters together and denote it as a new cluster. Thereafter the algorithm is looking for a second pair of closest clusters; it can be either two new samples or a sample and the cluster, which was obtained on the previous step. The procedure repeats until all the clusters are merged on a top level.

Another parameter of choice is a linkage criterion, which defines proximity between clusters. We used a Ward’s method that suggests merging two clusters for which the total within-cluster variance has a minimum increase. To establish a dendrogram we used group means and thereafter computed pairwise Euclidian distances between these group means.

Once the hierarchical tree was established, its structure was used in classification analysis. At each node from top to bottom of the tree we used PLSDA first. For comparison purposes later an FLDA classifier was used for the same tree structure.

3.7 One-Versus-All Approach (OVA)

It was reported that a One-Versus-All (OVA) scheme combined with a properly chosen classifier can deal very well with multiclass problems (Rifkin & Klautau, 2004). In this method a single group is set against other remaining groups and subsequently a classifier is established. Thereafter the procedure is repeated for a second, third, etc., group until g classifiers are established, where g is the number of groups. When a new sample needs to be classified it is run through all g classifiers and a label of a classifier which gives the largest value determines a label of the sample. That was the original idea of OVA approach. We used the idea in order to build a binary tree. Each of g classifiers is trained to separate one particular group against the rest. Once g classifiers are established we compare them and the classifier, which gave the best result (according to a predetermined criterion), is used at the first node of the binary tree. Afterwards the process is repeated for $g-1$ groups for the second node, $g-2$ for the third, etc., until $g-1$ classifiers are established for an entire OVA binary tree.

To choose an optimal classifier at each node we needed to compare 10 classifiers at node one, 9 classifiers at node two, etc. Since we had different amounts of samples in each group it was not meaningful to compare classifiers by considering total success rate (SR) obtained for a whole model. This is due to the fact that a single group has a much smaller sample size compared to the group of the remaining samples. In this case the SR of the bigger group predominates the total SR. Instead of this, we considered the minimum SR for each group in such classifiers. We considered the best classifier as the one, for which this minimum is highest among other minima in other classifiers. In the Matlab code instead of SR we used misclassification rate (MCR) and we were looking for a smallest maximum among other maxima. An explanation of this criterion is described schematically below.

Step 1. The algorithm starts from considering a matrix that consists of MCRs of all classifiers to be compared:

# of a classifier	1	2	3	4	5	6	7	8	9	10
MCR for a group # 1, %	10.34	26.32	16.67	3.41	11.54	10.98	14.29	2.57	1.15	20.47
MCR for a group # 2, %	1.61	0.13	19.36	4.26	0	12.83	0	3.89	0.56	0.54

⇓

Step 2. It finds the maximum MCR for each classifier out of two groups:

# of a classifier	1	2	3	4	5	6	7	8	9	10
MCR out of two groups, %	10.34	26.32	19.36	4.26	11.54	12.83	14.29	3.89	1.15	20.47

⇓

Step 3. It takes the classifier # 9 with a minimum MCR which is equal to 1.15%.

3.8 One-Versus-One Approach (OVO)

The One-Versus-One Approach (OVO) is another method for reducing multiclass classification problems to binary classification tasks, which has reported to be more accurate than the OVA approach (Furnkranz, 2002). In this method we need to train $g(g-1)/2$ classifiers, where g is a number of classes. Each classifier is build using data of two single groups. The final model is a collection of these classifiers. Thereafter, when a new sample is to be classified, each classifier gives one vote to its preferred class. Subsequently a majority voting scheme is used to make a final decision for the class the sample is to be assigned to. The disadvantage of this technique is that the number of classifiers to be trained increases faster than a number of classes, which requires long computational time. In our situation with 10 groups 45 classifiers were trained. If we double the number of groups, 190 classifiers need to be trained.

3.9 Random Forest (RF)

RF is a classifier which is based on an ensemble of decision trees (Breiman, 2001). A first step in the RF algorithm is to create a data sample set from the original dataset applying bootstrapping or in other words by random sampling with replacement m times, where m is a number of observations in the original dataset. It means that some samples are chosen not only ones and other samples may not be chosen at all. Once a new sample set is created about one-third of the samples are left out randomly and are not used in a decision tree construction. This subset is called “out-of-bag” (OOB) subset which is used for error estimation. For the construction an individual decision tree about two-third of the original samples will be used.

To construct each node of a decision tree, the RF method uses randomly selected variables. A number of variables which is selected by default is equal to $\sqrt{n}$, where n is a number of variables in a dataset. At each node of a decision tree a particular variable is chosen based on an information gain (Maguire et el., 2012). An information gain is an entropy reduction caused by splitting the data samples using particular variable (Mitchell, 1997), while entropy is a measure of an impurity of the samples to be separated at each node. RF algorithm uses so called Gini impurity which is given by:

$$i(N) = \sum_{i \neq j} \hat{P}\,(x \in \omega_i|N)\hat{P}(x \in \omega_j\,|N), \tag{13}$$

where $i(N)$ is the impurity of a node N, $\hat{P}(x \in \omega_i|N)$ is the fraction of training samples x at node N that are in a group ω_i.

A desired number of decision trees is built resulting in a “forest” of decision trees. When classifying, each sample will be passed through all the decision trees in the random forest and a majority voting scheme will be used to assign the sample to a most popular class.

3.10 Analysis of multiblock (MB) data

In this study, we have five different blocks of information according to five different cultivation media used for growing the microorganisms. All strains were grown in six independent cultivation runs. Thus, MB situation is obtained when each spectrum for a given medium has a corresponding spectrum of a strain grown on the other media in the same run. In other words the same sample is grown six times on five different cultivation media. It is known that using different media, which means using different nutrients to feed yeasts, provoke a different phenotype of yeast cells. The idea is to see if one medium can better separate some groups while another medium can probably better separate some other groups. In such case we expect that the combination of two media will improve an overall classification. We wanted to apply two-block setup to RF approach and check if the use of additional information will improve the identification success.

To use different matrices in a two-block setting, a row-to-row correspondence is necessary. In other words the same samples have to be presented and ordered in both data matrices in order to concatenate them. To have different combinations of two blocks, all five blocks have to be ordered. The samples in five data matrices were ordered alphabetically and samples which were not presented in at least one out of five media were removed from the analysis. A number of variables were doubled after merging two matrices together. In Fig.2 an explanation of the two-block set-up is shown.

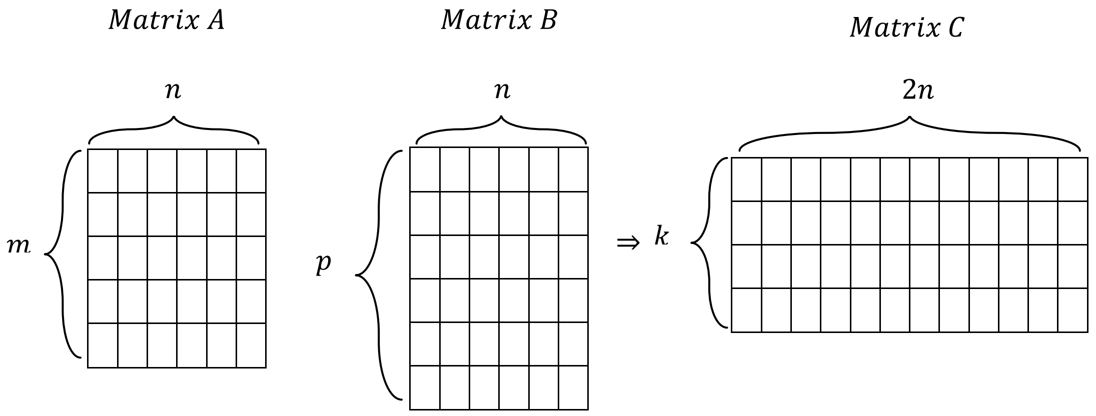


Figure 2: Merging two data matrices in order to perform two-block analysis. Where m and p correspond to number of rows in the matrices A and B, respectively, that need to be concatenated. k corresponds to a number of rows in concatenated matrix C. n corresponds to a number of variables.

4 Results and Discussion

4.1 Data Selected for Analysis

The original dataset, which we worked with consisted of five different blocks of data according to cultivation media (codes of the media are: SAB, YPD, YMB, SD, and YEPD), on which the food spoilage yeasts were grown. Each subset consisted of different amounts of spectra since 1) some of the yeasts did not grow well on each of the medium; 2) some of the spectra did not go through spectral quality control tests implemented by Opus software (Shapaval et al., 2013). The main results in this section will be presented for SAB medium, which we chose as an example and which we will refer to as the SAB dataset. The results for the other 4 media are presented in section Appendix.

The original dataset consisted of 12 different genera of food spoilage yeasts. The dataset was divided into calibration and validation subsets. The validation subset consisted of one out of six independent Bioscreen runs that represent independent growth experiments. First run was used for external model validation, while a calibration model was established on the other five runs. After preprocessing we realized that the genus *Iss* was presented only by two samples and the genus *Lod* only by one in a validation subset for media YMB and SD. Thus, due to the low sample size in the groups *Iss* and *Lod*, we decided to remove them being left with 10 genera.

4.2 Preprocessed spectra

After removing two groups from the dataset and before preprocessing, the SAB dataset consisted of 2061 FTIR spectra, which are shown in Fig. 3. As we can see from the graph, baseline variations and scaling variations are present in the spectra.

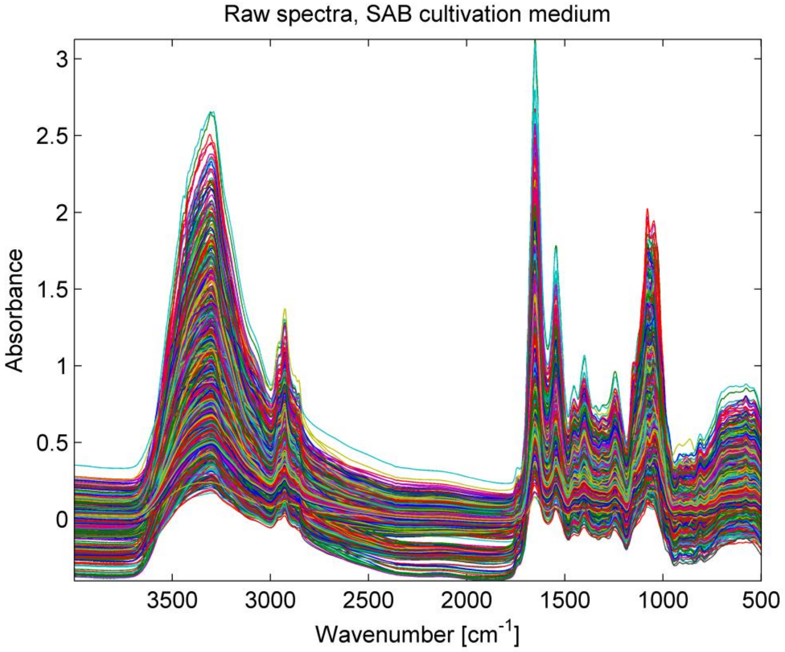


Figure 3: Raw spectra recorded in a range 4000 cm^{-1} and 500 cm^{-1} of yeasts grown on SAB medium.

Baseline variations can arise due to variations of the intensity in the light source, while scaling variations relevant to FTIR spectroscopic data is the result of differences in the sample thickness or effective optical path length (Kohler, Afseth, & Martens, 2010). These variations can influence the subsequent data analysis. Together with aforesaid effects other unwanted interferences can be present in the spectra which are not visible by naked eye but can cause difficulties in the interpretation of the spectra and instability of the established models. These interferences can be due to: 1) additional substances presented in the experimental environment, e.g. CO_2 and H_2O; 2) imperfective instruments; 3) different accompanying chemical components in the measured sample, e.g. contaminants (Zimmermann & Kohler, 2013).

To reduce such variations and strengthen analyte signals we used a combination of Savitzky-Golay (SG) algorithm and extended multiplicative signal correction (EMSC). The sequence, which was used for preprocessing is as follows: 1) averaging technical replicates; 2) first derivative by SG numerical algorithm using nine-points window size and third order polynomial (Savitzky & Golay, 1964); 3) spectral range selection 3100-2800 cm^{-1} and 1800-700 cm^{-1}; 4) EMSC with linear and quadratic terms (Afseth & Kohler, 2012).

The preprocessing was made separately for the calibration subset, where the EMSC model was established and applied later to the validation subset. After preprocessing, the calibration and validation subsets consisted of 799 and 162 samples, respectively, with 1454 variables in each. In Fig. 4 the preprocessed spectra for the calibration set are shown. The baseline effects are removed, which is mainly but not completely due to the derivative (Zimmermann & Kohler, 2013), and scaling variations have been reduced.

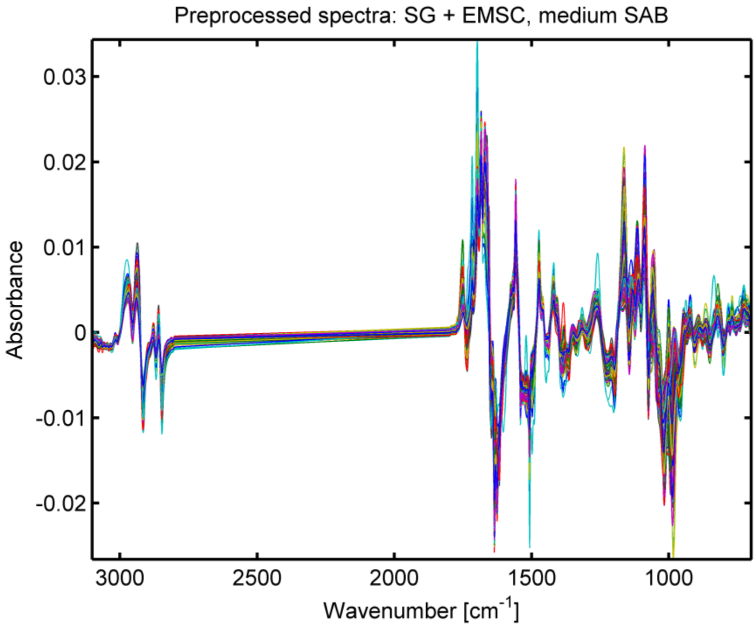


Figure 4: The SAB calibration dataset spectra preprocessed by SG and EMSC algorithms. The technical replicates are averaged. The 3100-2800 cm^{-1} & 1800-700 cm^{-1} spectral regions are selected for further analysis.

4.3 Principal Component Analysis (PCA)

To learn, explore and find clusters in the data, an unsupervised multivariate algorithm, PCA, was applied to the preprocessed dataset. Fig. 5 shows the score plot for the SAB dataset where we can see that the dataset is rather complex and groups are not easily separated except for the genera *Tor*, *Rho* and *Deb*. Furthermore in Fig. 6 where the score plot is presented for YEPD medium only the group *Tor* is well separated from the other groups suggesting that the structure can be even more complicated depending on the phenotypic variation in the data.

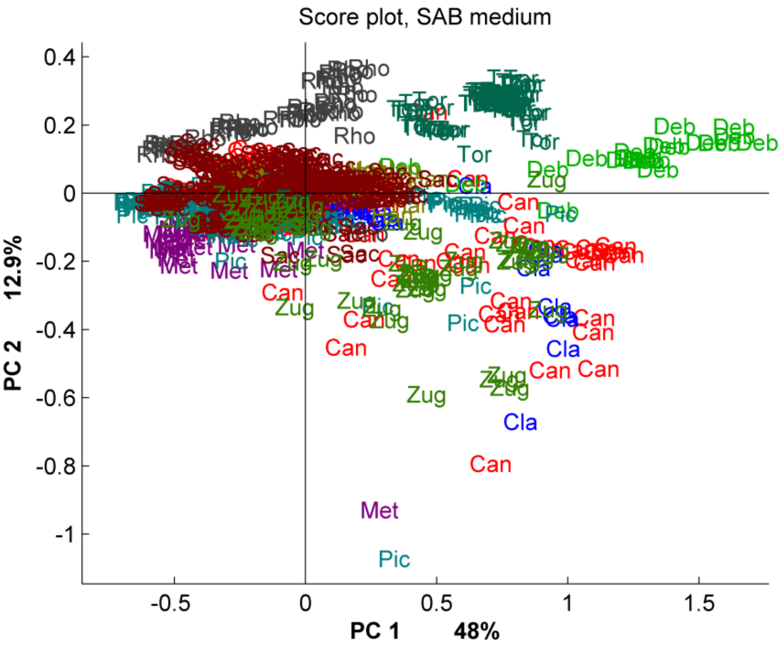


Figure 5: Score plot of PC1 vs. PC2 in PCA analysis of the SAB dataset.

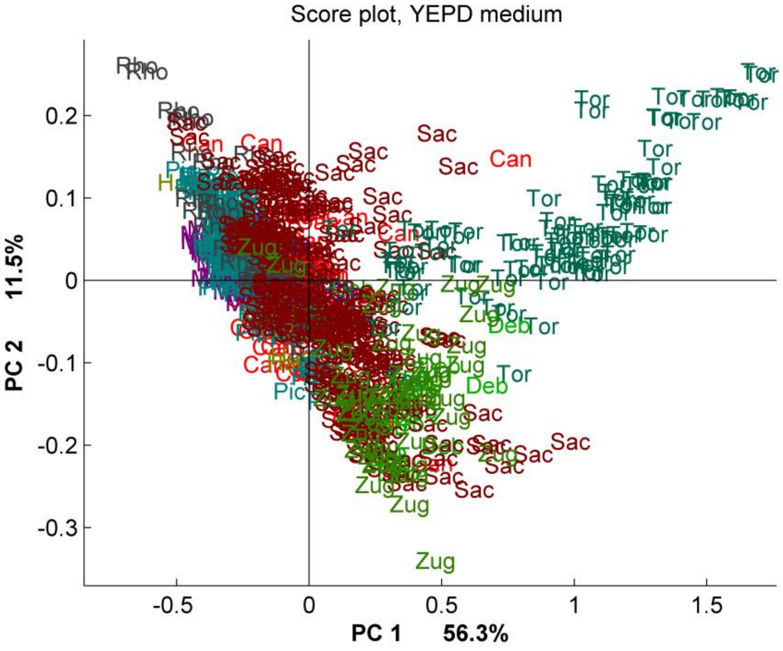


Figure 6: Score plot of PC1 vs. PC2 in PCA analysis of the YEPD dataset.

Looking closer to the species, which constitute the existing genera we could see that different genera contain different amounts of species. Some genera contain many species, some only few. Moreover, species of some genera constitute more homogeneous groups compared to others. Thus, *Tor*, *Rho* and *Deb* are well separated from the others because *Tor* and *Deb* consist only of one species each; while *Rho,* which contains four different species, has a very homogeneous structure. The biggest genus group *Can* contains 15 different species. It is scattered the most and that is probably why it is not an easy group to classify and identify (see results below).

As can be seen from Fig. 7 twelve PCs explain 95% of the variability in the SAB dataset. Furthermore, 30 PCs are needed to explain 99% of the variance. This proves the complexity of the dataset.

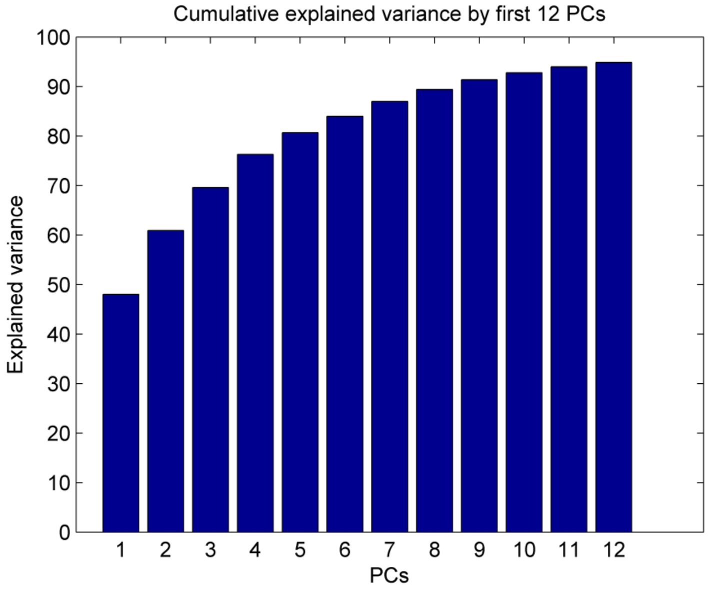
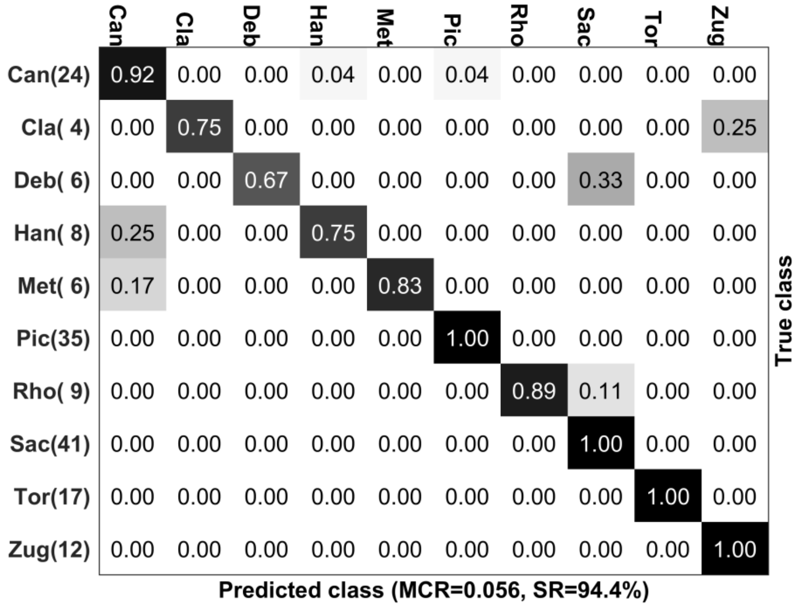


Figure 7: Cumulative explained variance by the first 12 PCs in a PCA analysis of the SAB dataset.

4.4 Partial Least Squares Discriminant Analysis (PLSDA)

A first method which was used for classification ten different genera, since it is frequently used technique in spectroscopy, is a partial least squares discriminant analysis (PLSDA) (Martens, & Næs, 1989) A model calibration was implemented by using live-one-run-out cross-validation approach in order to choose an optimal number of PLS factors. An optimal number of PLS factors is the one, which does not give significantly worse misclassification rate (MCR) than a minimum MCR. To find the significant differences, a build-in Matlab binomial cumulative distribution function `binocdf` was used. For model validation a first Bioscreen run was used. The validation results for SAB dataset are shown in Fig. 8. The overall success rate (SR) is equal to 94.4% which is relatively good result for such amount of groups. The lowest SR for a single group is equal to 67 % for a group *Deb.* Computation time for the model calibration is 30 seconds.

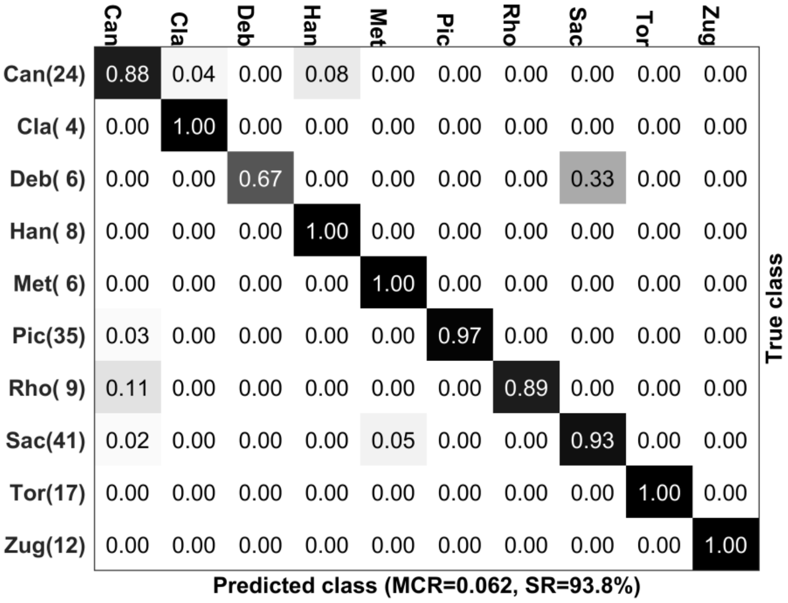


	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(24)	0.92	0.00	0.00	0.04	0.00	0.04	0.00	0.00	0.00	0.00
Cla(4)	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.25	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.17	0.00	0.00	0.00	0.83	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(9)	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.11	0.00	0.00
Sac(41)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Tor(17)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 8: Confusion matrix for the validation where the first run was used as an independent test set in the SAB dataset. The method used for classification is PLSDA. The numbers in parenthesis next to the group name correspond to a number of samples in each genus. MCR and SR correspond to the misclassification and success rate, respectively.

4.5 Fisher Linear Discriminant Analysis (FLDA)

Trying to improve classification results, we employed other classifiers such as PLS followed by Fisher linear discriminant analysis (FLDA) (Fisher, 1936). It was suggested by Barker and Rayens, (2003) that the PLS technique can be used for dimensionality reduction when a number of variables is much higher than a number of samples and a formal FLDA cannot be performed. This is the case in this study and for FTIR spectroscopic data in general. Hence we used PLS scores as predictors for FLDA. To calibrate the model a similar procedure was used as for PLSDA. The results from a model validation are shown in Fig. 9. The results are similar to results from PLSDA approach. The same tendency has the group *Deb* with two out of six samples misclassified, resulting in a SR equal to 67%. Smaller groups are better classified by FLDA. For example, *Cla* and *Han* have SR equal to 100% each whereas PLSDA performs better for big groups such as *Pic* and *Sac*.



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(24)	0.88	0.04	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00
Cla(4)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.03	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00
Rho(9)	0.11	0.00	0.00	0.00	0.00	0.00	0.89	0.00	0.00	0.00
Sac(41)	0.02	0.00	0.00	0.00	0.05	0.00	0.00	0.93	0.00	0.00
Tor(17)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 9: Confusion matrix for the validation where the first run was used as an independent test set in the SAB dataset. The method used for classification is FLDA. The numbers in parenthesis next to the group name correspond to a number of samples in each genus. MCR and SR correspond to the misclassification and success rate, respectively.

4.6 Cluster Analysis for Establishing a Hierarchical Tree

Since the data structure is complex, we employed a hierarchical tree approach in order to reduce the multiclass problem to a binary tree classification task. To establish a hierarchical tree we used Hierarchical Cluster Analysis (HCA). Instead of using original samples, group means were fed into the algorithm. Euclidian distances between group means were measured pairwise for all groups. Thereafter, a build-in Matlab function `linkage` was utilized with Ward's minimum variance criterion. A hierarchical tree, which was established from the SAB dataset where the first run is left aside, is shown in Fig. 10.

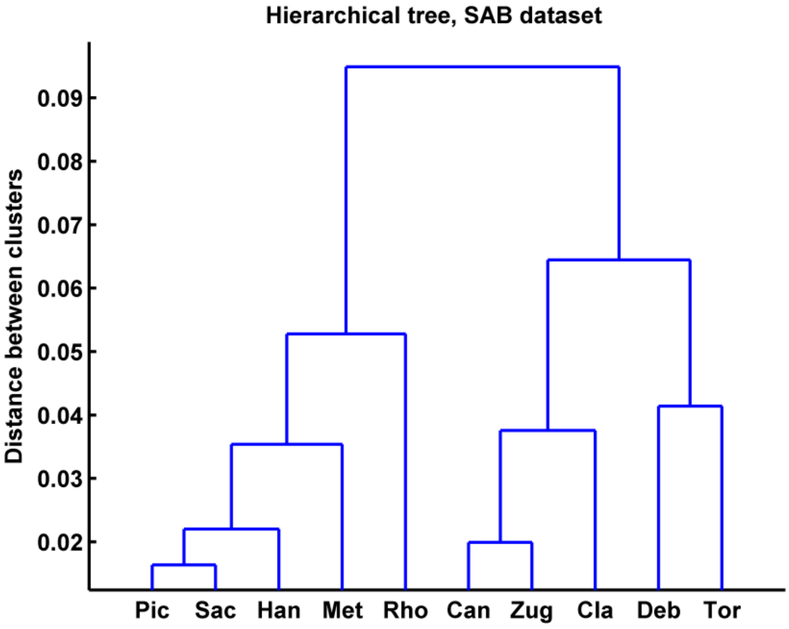
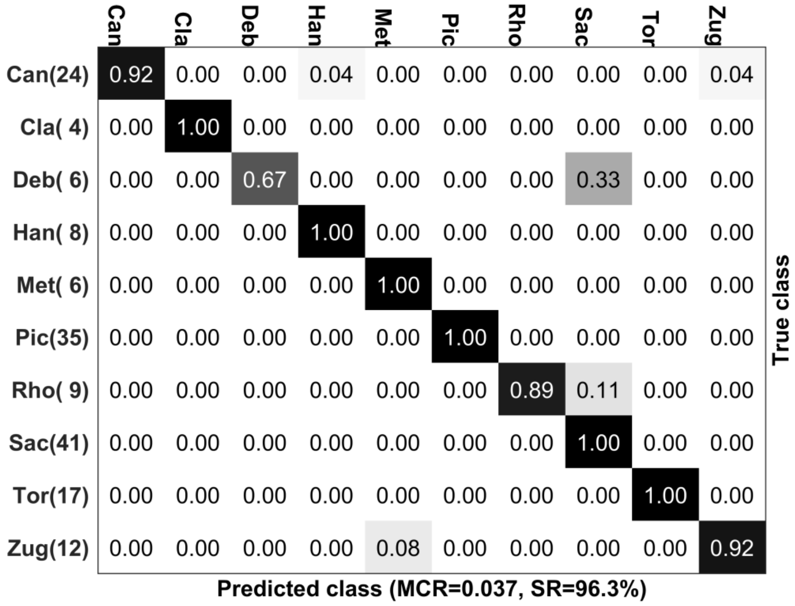


Figure 10: Hierarchical tree based on the group means for the SAB dataset where the first run was left aside. The established hierarchical structure was used for classification analysis using PLSDA and FLDA classifiers at each node.

PLSDA was applied to the hierarchical tree structure. In each node a PLS model was established. To calibrate each model, the same approach was used as for PLSDA applied directly to SAB dataset. Subsequently a model validation with independent run, which was kept aside, was performed to test an optimized model. The validation results are shown in Fig. 11 where first run was used as an independent test set.

An overall good classification is obtained with SR equal to 96.3%. The same two samples out of sixe in the validation set of the group *Deb* were misclassified as *Sac*. As a result the SR for genus *Deb* is equal to 67%. Looking closer to the score plot of the SAB data set we saw that two samples from the group *Deb* are right inside the *Sac* cloud (Fig. 12). As we will see below, this misclassification pattern will be repeated for other classification methods, which were used in this study suggesting that these two samples are probably wrongly assigned to *Deb* by biochemical analysis. This happens in microbiology that biochemical analysis based on the growth in different selective media has difficulties identifying phylogenetic unit of the microbial isolate, while FTIR spectroscopy can easily find it (Oust, Møretrø, Kirschner, Narvhus, & Kohler, 2004).



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(24)	0.92	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.04
Cla(4)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(9)	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.11	0.00	0.00
Sac(41)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Tor(17)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(12)	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.92

Figure 11: Confusion matrix for the validation where the first run was used as an independent test set in SAB dataset. The method for hierarchical tree establishment is HCA with PLSDA as a classifier in each node. The numbers in parenthesis next to the group name correspond to a number of samples in each genus. MCR and SR correspond to the misclassification and success rate, respectively.

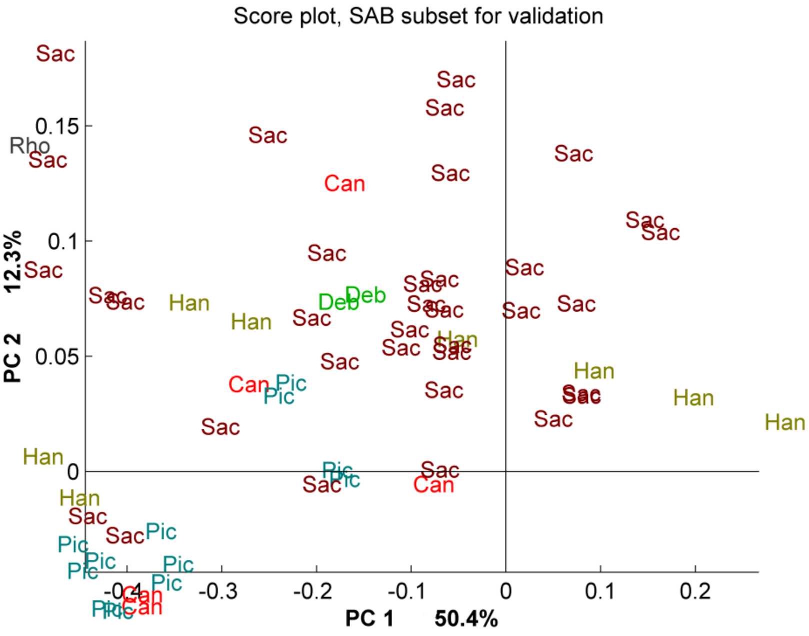
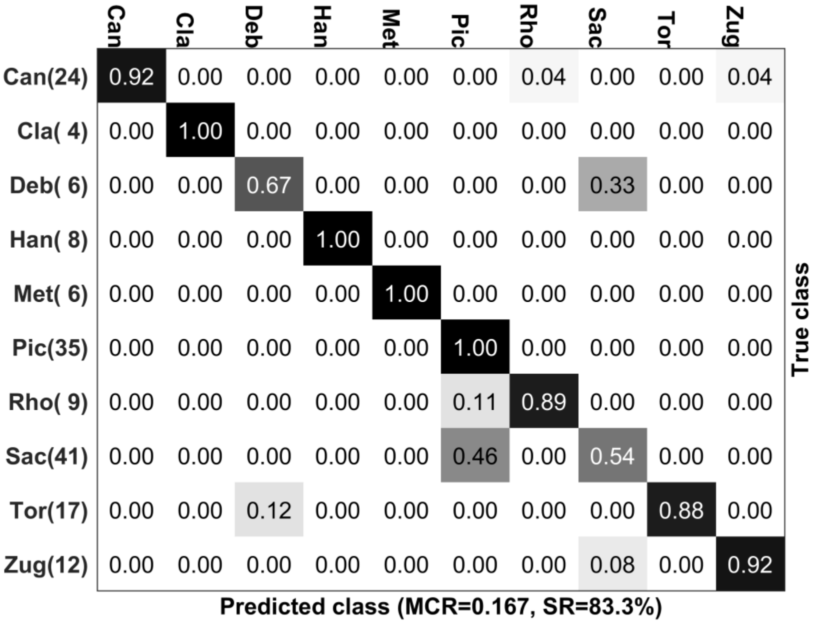


Figure 12: Zoomed score plot of PC1 vs. PC2 in PCA analysis of the SAB validation subset.

The same hierarchical tree structure was employed, which is shown in Fig. 10 together with FLDA performed on PLS scores. For model calibration and validation the same approaches were used as for PLSDA mentioned above.

A confusion matrix for run one, which was used as independent test set, is presented in Fig. 13. As we can see, the SR is equal to 83.3%, which is worse than when applying PLSDA. The worst classification results are shown to be for the samples of group *Sac* with SR equal to 54%. Thus, almost half of *Sac* samples were classified as *Pic*, which is an interesting result, because this is the last node in the hierarchical tree and PLSDA was able to achieve 100% SR for this problem. Similar for the group *Tor:* all samples were classified by PLSDA whereas FLDA misclassified two samples from *Tor* as *Deb*. It means that PLSDA in this case outperformed FLDA applied to PLS scores. The same two samples from *Deb* are misclassified as *Sac* by FLDA.



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(24)	0.92	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.04
Cla(4)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(9)	0.00	0.00	0.00	0.00	0.00	0.11	0.89	0.00	0.00	0.00
Sac(41)	0.00	0.00	0.00	0.00	0.00	0.46	0.00	0.54	0.00	0.00
Tor(17)	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.88	0.00
Zug(12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.92

Figure 13: Confusion matrix for the validation where the first run was used as an independent test set in SAB dataset. The method for hierarchical tree establishment is HCA with FLDA as a classifier in each node. The numbers in parenthesis next to the group name correspond to a number of samples in each genus. MCR and SR correspond to the misclassification and success rate, respectively.

4.7 PLSDA and FLDA in One-Versus-All (OVA) Approach

Another approach for building a classification tree which was tested is One-Versus-All (OVA). A binary split at each node was implemented with a one-versus-all classification rule, which is simple to implement and which was reported to perform as accurately as other more sophisticated approaches (Rifkin & Klautau, 2004). Each group is tested against all other groups; this is where the method has its name from. Two classifiers, PLSDA and FLDA performed on PLS scores, established at each node of a tree, were compared.

To optimize the number of PLS components and find the best group to be chosen for a node, i.e. which is the most easy to separate from the rest of the groups, a live-one-run-out cross-validation approach was used. Calibration results for SAB dataset are shown in Fig. 14, where percents of correctly classified samples for each node are presented. Validation was made on an independent test set, which was kept aside. For validation we allowed each sample to travel through the binary tree until it was classified to be a member of one of 10 different groups. A confusion matrix for validation is presented in Fig. 15.

Using FLDA performed on PLS scores as a classifier in OVA approach we obtained different structure of the tree compare to PLSDA in the same approach. The results from calibration model are shown in Fig. 16 and the confusion matrix for validation is shown in Fig. 17.

Applying PLSDA in OVA we obtained SR equal to 91.4% which is slightly better compared to FLDA with SR equal to 90.7%. The results of classification the groups are similar for both classifiers except for group *Han*, which was classified with SR equal to 100% by PLSDA and 75% by FLDA. The same samples of *Deb* were still misclassified as *Sac* by both classifiers.

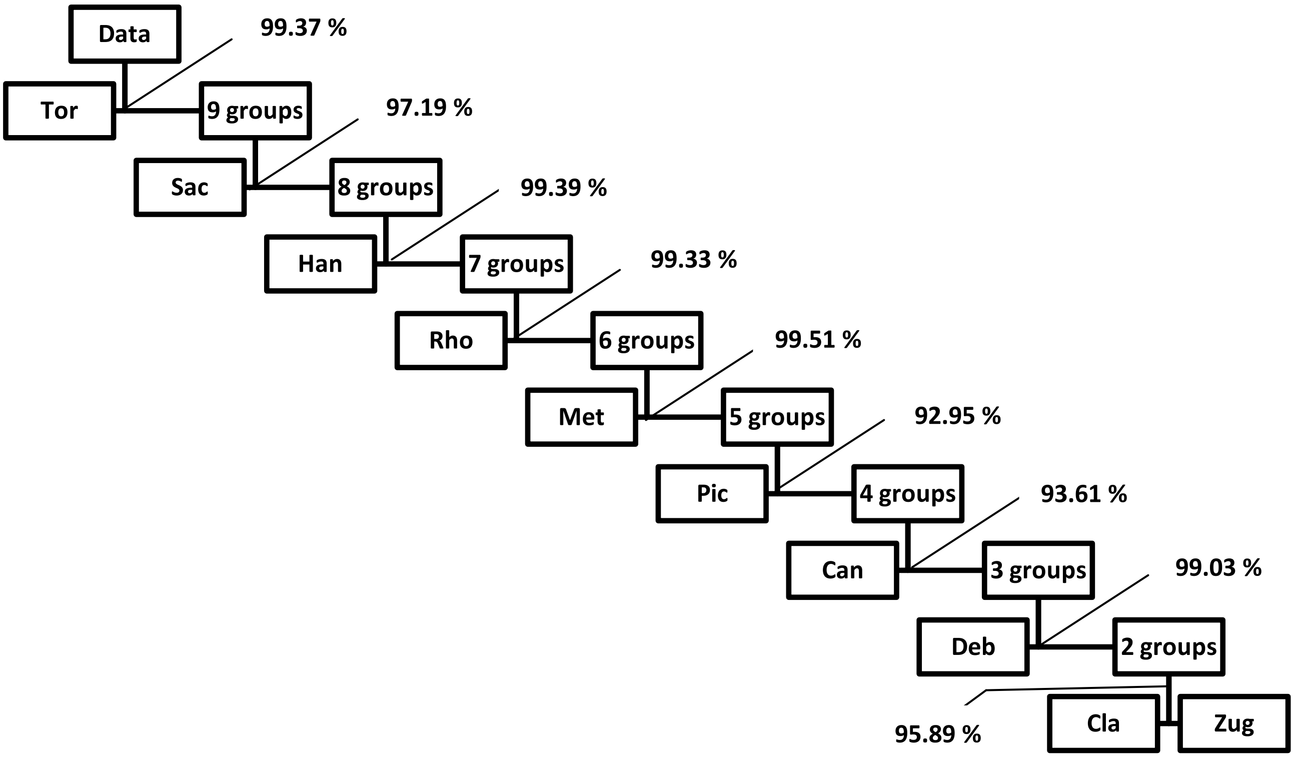
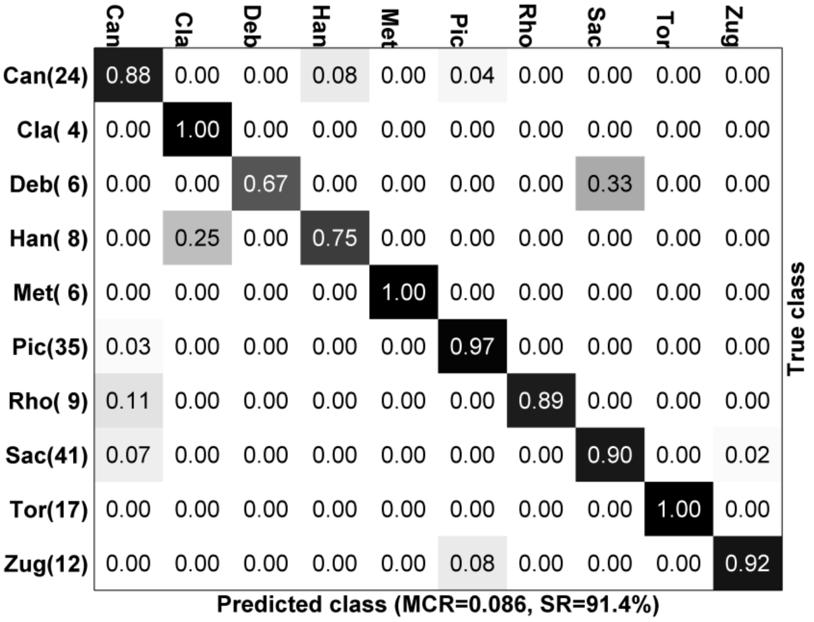


Figure 14: Calibration results for the binary tree built by one-versus-all approach with PLSDA classifier used at each node. The first run was used as independent test set in SAB dataset. The success rate is given for each node.



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(24)	0.88	0.00	0.00	0.08	0.00	0.04	0.00	0.00	0.00	0.00
Cla(4)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.25	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.03	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00
Rho(9)	0.11	0.00	0.00	0.00	0.00	0.00	0.89	0.00	0.00	0.00
Sac(41)	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.00	0.02
Tor(17)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(12)	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.92

Figure 15: Confusion matrix for the validation, where the first run was used as an independent test set in the SAB dataset. The one-versus-all approach was used for the tree building with PLSDA used as a classifier. The numbers in parenthesis correspond to a number of samples in each genus. MCR and SR correspond to the misclassification and success rate, respectively.

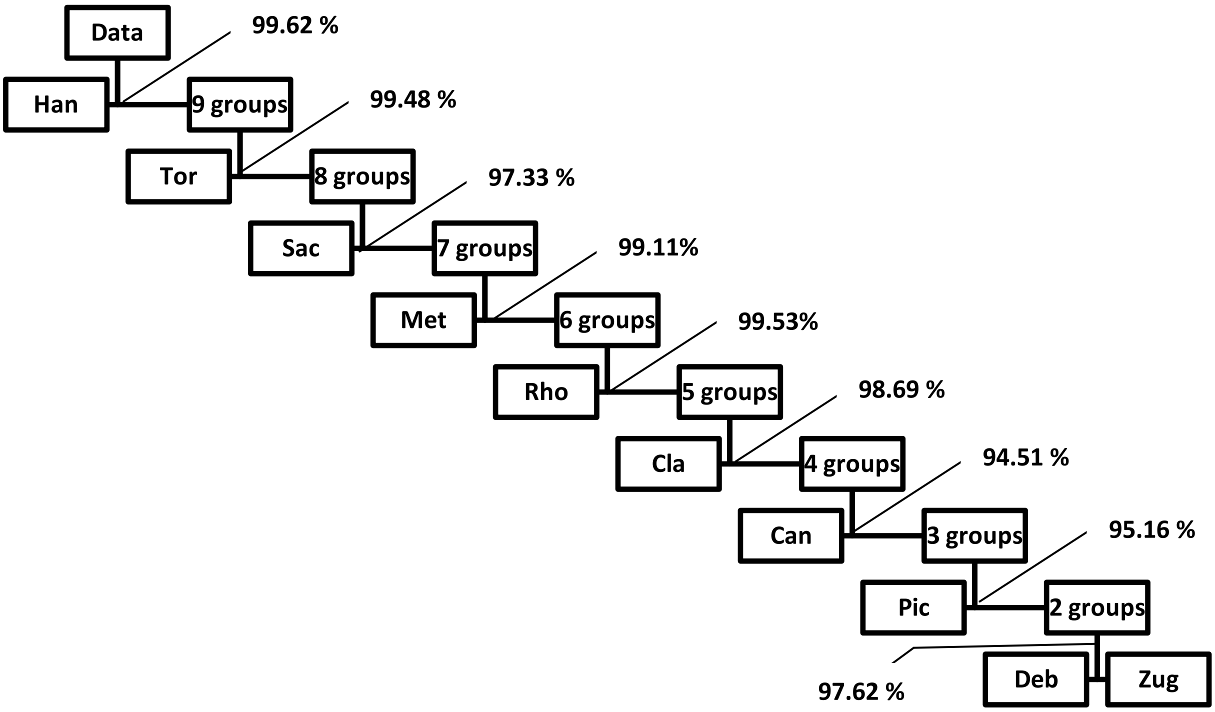
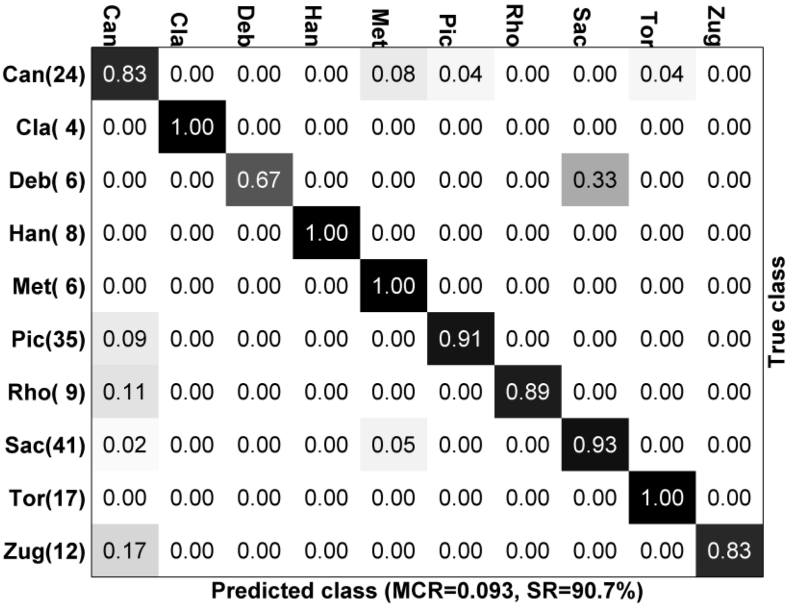


Figure 16: Calibration results for the binary tree built by one-versus-all approach with FLDA performed on PLS scores as a classifier used at each node. SR is given at each node. The first run was used as independent test set in the SAB dataset.



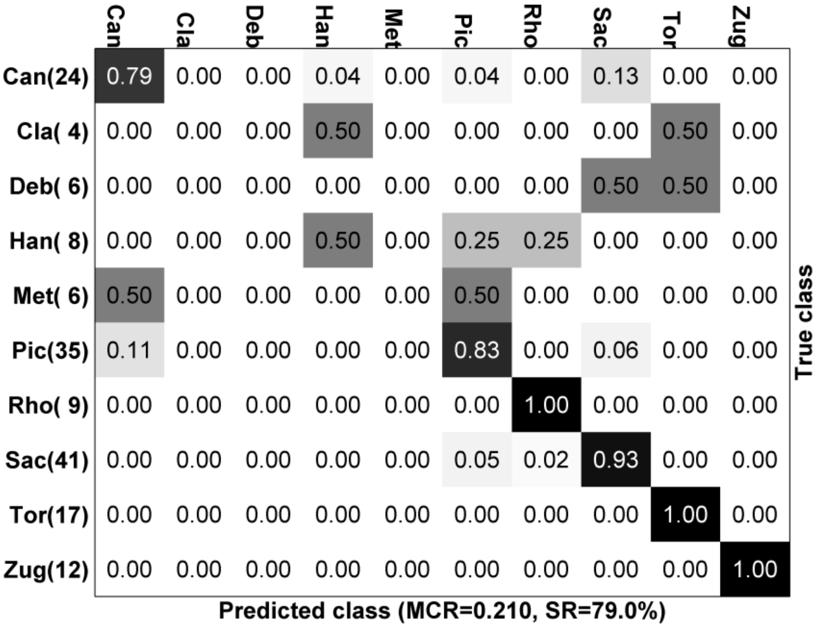
	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(24)	0.83	0.00	0.00	0.00	0.08	0.04	0.00	0.00	0.04	0.00
Cla(4)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.09	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00
Rho(9)	0.11	0.00	0.00	0.00	0.00	0.00	0.89	0.00	0.00	0.00
Sac(41)	0.02	0.00	0.00	0.00	0.05	0.00	0.00	0.93	0.00	0.00
Tor(17)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(12)	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.83

Figure 17: Confusion matrix for the validation where the first run was used as an independent test set in the SAB dataset. The one-versus-all approach was used for tree building with FLDA performed on PLS scores as a classifier. The numbers in parenthesis correspond to a number of samples in each genus. MCR and SR correspond to the misclassification and success rate, respectively.

4.8 PLSDA and FLDA in One-Versus-One (OVO) Approach

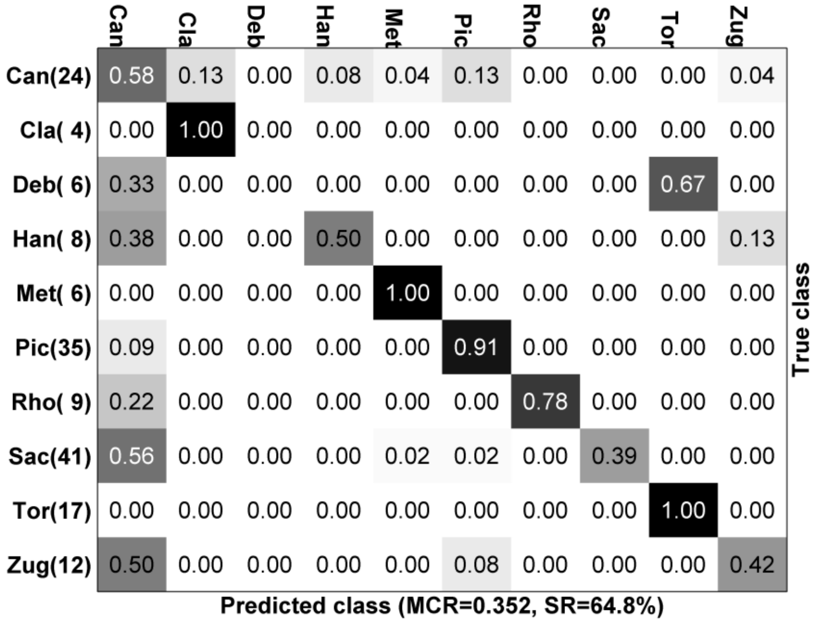
Another method that can solve a multiclass classification problem by reducing it to a combination of binary classifiers is an OVO approach (Mittal, Rani, & Ritambhara, 2016). This method requires training of $g(g-1)/2$ classifiers (g is a number of groups), each of which gives a vote in the classification. When a new sample is to be classified it is assigned by each classifier to one certain class giving thus one vote. To make a final decision for the sample a majority voting scheme is used.

For model calibration a live-one-run-out cross-validation approach was used to choose an optimal number of PLS components for each classifier. Both classifiers are compared: PLSDA and FLDA performed on PLS score. In Fig. 18 and 19 confusion matrices are shown for PLSDA and FLDA, respectively, for the validation dataset when the first run is kept out.



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(24)	0.79	0.00	0.00	0.04	0.00	0.04	0.00	0.13	0.00	0.00
Cla(4)	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.50	0.00
Deb(6)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.00
Han(8)	0.00	0.00	0.00	0.50	0.00	0.25	0.25	0.00	0.00	0.00
Met(6)	0.50	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00
Pic(35)	0.11	0.00	0.00	0.00	0.00	0.83	0.00	0.06	0.00	0.00
Rho(9)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(41)	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.93	0.00	0.00
Tor(17)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 18: Confusion matrix for the validation where the first run was used as an independent test set in the SAB dataset. The one-versus-one approach was used for the tree building with PLSDA used as a classifier. The numbers in parenthesis correspond to a number of samples in each genus. MCR and SR correspond to the misclassification and success rate, respectively.



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(24)	0.58	0.13	0.00	0.08	0.04	0.13	0.00	0.00	0.00	0.04
Cla(4)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.00
Han(8)	0.38	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.13
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.09	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00
Rho(9)	0.22	0.00	0.00	0.00	0.00	0.00	0.78	0.00	0.00	0.00
Sac(41)	0.56	0.00	0.00	0.00	0.02	0.02	0.00	0.39	0.00	0.00
Tor(17)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(12)	0.50	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.42

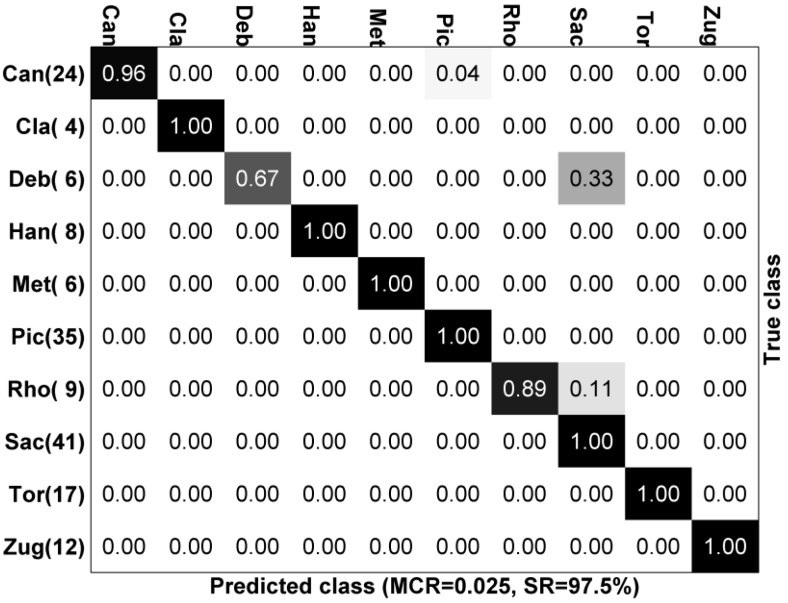
Figure 19: Confusion matrix for the validation where the first run was used as an independent test set in the SAB dataset. The one-versus-one approach was used for tree building with FLDA performed on PLS scores as a classifier. The numbers in parenthesis correspond to a number of samples in each genus. MCR and SR correspond to the misclassification and success rate, respectively.

A higher SR is obtained (SR=79.0%) by PLSDA compared to FLDA (SR=64.8%) in the OVO approach. At the same time three groups *Cla*, *Deb,* and *Met* were completely misclassified when applying PLSDA, whereas using FLDA only one group *Deb* was completely misclassified. Nevertheless the overall SR is better for PLSDA because big groups such as *Sac, Can,* and *Zug* classified better in this case and influence the total SR more than smaller groups such as *Cla, Deb,* and *Met* with SR equal to 0% in this case.

4.9 Random Forest (RF)

Random forest is a method of classification which can give better classification results compared to single tree classifiers such as CART (classification and regression trees) (Biau, Devroye, & Lugosi, 2008). To generate RF we used a Matlab function `TreeBagger` with default parameters and growing 1000 trees. In Fig. 20 a confusion matrix for the validation where the first run is kept out for the RF approach is shown.

An overall SR for the model is equal to 97.5% which is the best result according to approaches we used so far. Compared to second best method, which is PLSDA applied to hierarchical tree, we can see that for the group *Rho* the SR is the same as for RF. Group *Deb* still keeps SR equal to 67%. For other groups RF obtained higher SR than PLSDA applied to hierarchical tree.



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(24)	0.96	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00
Cla(4)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(9)	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.11	0.00	0.00
Sac(41)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Tor(17)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 20: Confusion matrix for the validation of RF where the first run was used as an independent test set in the SAB dataset. The numbers in parenthesis correspond to a number of samples in each genus. MCR and SR correspond to the misclassification and success rate, respectively.

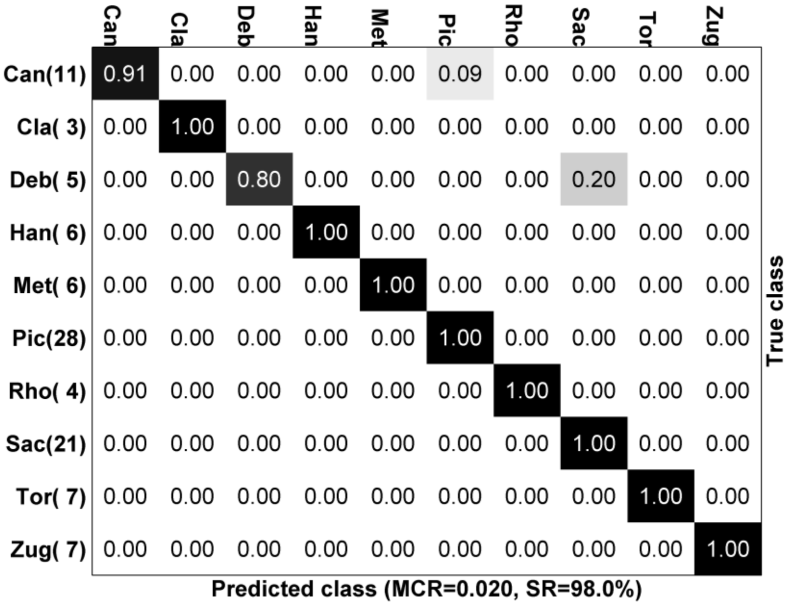
4.10 RF applied to multiblock (MB) data

In order to improve the performance of the RF approach a combination of two media was used for classification model establishment. In Table 2 the validation results are compared when both each medium separately and combinations of selected media were used for training. Each medium was reordered in a way that the amount of samples in each dataset is equal. For example if one particular sample was grown only in the YEPD medium and was not presented in the other four media, it was removed from the dataset. This procedure was applied for each sample. SR for medium SAB shown in Table 2 is lower than SR for medium SAB shown in Fig. 19, this is due to reordering, since the amount of samples in both calibration and validation subsets was decreased.

Table 2: Validation results obtained from RF approach applied to different media separately and to combinations of two media.

Medium	Success rate, %	Media	Success rate	Media	Success rate, %
SAB	95.9	SAB+YPD	96.9	YPD+SD	95.9
YPD	96.9	SAB+YMB	96.9	YPD+YEPD	95.9
YMB	92.9	SAB+SD	96.9	YMB+SD	95.9
SD	89.8	SAB+YEPD	95.9	YMB+YEPD	94.9
YEPD	93.9	YPD+YMB	98	SD+YEPD	92.9

As we can see from the Table 2 the best result gave the YPD medium with SR equal to 96.9%. A combination of YPD and YMB media gave a higher SR than each medium separately and it is equal to 98% (Fig. 21). By combining media YMB and SD with the lowest SRs equal to 92.9% and 89.8%, respectively, we obtained improvement in SR which is equal to 95.9%. An overall conclusion about using a MB data in the RF algorithm is that a combination of two data blocks used for model establishment performs better than at least one of the blocks in this combination used separately. In other words two blocks together can give higher SR than each block separately.



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(11)	0.91	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00
Cla(3)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(5)	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.20	0.00	0.00
Han(6)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(28)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(4)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(21)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Tor(7)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(7)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 21: Confusion matrix for the validation of random forest where the first run was used as independent test set in the combination of two datasets YPD and YMB media. The calibration and validation sets were smaller due the reordering necessary for the MB analysis. The numbers in parenthesis correspond to a number of samples in each genus. MCR and SR correspond to the misclassification and success rate, respectively.

4.11 Discussion

The main goal of the thesis was to find the best way for setting up a classifications scheme in order to identify microorganisms by FTIR spectroscopy. In the following the different methods, which were evaluated in the thesis, are discussed and compared. For comparison, we use several criteria such as simplicity of use for the purposes defined in this thesis, performance, and opportunity to extend a method to several blocks of data.

4.11.1 PCA

PCA was used to reveal the structure of the data set and get a first impression about existing clusters in the data. Score plots are a very useful tool for visualization of the data and detecting outliers. For example, in this study two samples from the group *Deb* were always misclassified as the group *Sac* by different methods used for classification. Looking closer to the score plot of SAB dataset we realised that these two samples are right in the *Sac* cluster.

So, we could conclude that two samples from the group *Deb* most likely belong to the group *Sac*. This can be due to a wrong classification by biochemical analysis (Oust, Møretrø, Kirschner, Narvhus, & Kohler, 2004). PCA combined with other methods can be used as basis for classification for example by projecting data onto principal component direction and thereafter using some criteria to assigning samples to the appropriate class. Such criteria are the Euclidian or the Mahalanobis distances to centroids. However,in the example considered in this thesis, the data structure is very complex and the groups are very heterogeneous. Groups are very scattered, many groups are on top of each other. Thus, more powerful supervised classification methods were needed, that take into account the variation in each group rather than a total variation as PCA as an unsupervised method seeks for. After analyzing explained variances by PCs and the score plots it became clear that applying PCA and calculating for example Euclidian distances to centroids would result in difficulties in classification.

4.11.2 PLSDA

The first supervised method chosen for the classification analysis is PLSDA. PLSDA searches for maximum covariance directions taking into account class information by using an indicator matrix. It can also handle a high dimensional data since it reduces dimensionality within the analysis. PLSDA has been extensively used in a field of FTIR spectroscopy for identification of samples of different types. Four closely related species of Lactobacilli have been analyzed by Oust et al. (2004) using PLSDA for FTIR data resulting in up to 100% correct classification depending on a choice of calibration and validation datasets. According to Preisner, Lopes, and Menezes (2008) model based on PLSDA was able to identify 100% of 32 tested samples of three different bacteria species based on FTIR spectroscopy. In our studywe applied PLSDA directly to the dataset with ten groups in order to classify them. The validation success rate for classification into ten genera was 94.4%. Compared to the previously published results, this classification result is very good, since the number of groups was much higher in our case. A PLSDA calibration model is fast to build (less than a minute) and the overall SR of the external model validation is comparable with the best models using other methods in this study such as RF. We observed that PLSDA has a tendency to classify bigger groups more accurately than smaller groups. This might be due to the model's optimization algorithm. When we calibrate and optimize the model in order to find an optimal number of PLS components, the procedure compares overall SRs of the model and does not take into account SRs of each particular group. Since we have groups of different sizes, bigger groups influence the overall SR of the model more that smaller groups. Consequently, the number of PLS factors used in the final model and the model's SR is predominated by the bigger groups. It should be investigated to what extend PLSDA could be modified to optimize PLSDA models with respect to the smaller groups to reduce the problem of unequal group sizes. This may be achieved by up-weighting the classification results of the smaller groups in the error calculation of the cross-validation procedure. This will force the algorithm to focus on the smaller groups. Another approach to avoid a dominant influence of big groups' classification results on the overall SR of a model is to consider a SR of each group in the model. A way to do it is described in section 3.7 "One-

Versus-All (OVA) approach" and consists of selecting the model which has better SR for its worst classified group as the best model.

An advantage of PLSDA is that it can be easily extended to a multiblock (MB) situation (Westerhuis, Kourti, & MacGregor, 1998). The use of MB data is an advantage for classification based on FTIR. Microorganisms can be grown on different, but well-defined media and subsequently analysed by FTIR spectroscopy (Shapaval et al., 2013).The rationale behind the idea of using different cultivation media for growth of microorganisms and FTIR spectroscopy lies in the fact that different nutrients in the media trigger different metabolic pathways. Thus, microorganisms grown on different media produce different macromolecules in the cells such as proteins, lipids, and carbohydrates. These molecules can be detected by FTIR. Thus, some groups that are close or overlap on one medium, may be potentially separated when they are grown on a different medium since they reveal a different phenotypic expression on different media. This approach is commonly used for the investigation of different strains in functional genomics, where growth parameters are used to separate strains grown on a high number of media (Warringer, & Blomberg, 2003). Since the phenotypic expression obtained by FTIR spectrocopy is high-dimensional compared to the growth phenotype, it is expected that a low number of media is enough to achieve high classification success based on FTIR spectroscopy. The use of MBPLSDA, when many different cultivation media are employed, is beyond the scope of this thesis and needs to be addressed in future.

4.11.3 FLDA

A method that is very often used in multivariate statistics for discrimination purposes is LDA. Many authors use LDA and FLDA notations interchangeably even though FLDA does not require samples in each class to be normally distributed or class co-variances to be equal as LDA method does. We stick to the FLDA notation since we in general do not require normally distributed data with equal class co-variances. In FLDA, eigenvalue decomposition is applied to scatter matrices in order to solve Fisher's optimization problem, which are required to be invertible. This requirement is not fulfilled for the highly collinear data obtained from FTIR spectroscopy. Therefore, FLDA alone cannot handle high dimensional data. It is necessary to apply regularization methods or reducing the dimensionality prior to FLDA. Maquelin et al. (2002) used FLDA performed on PCA scores to classify *Candida* species. PLS scores have also been used as predictors in FLDA for classification of filamentous fungi (Liland, Kohler, & Shapaval, 2014). In our work, PLS scores were used in order to pursuit the following two goals: (1) to reduce dimensionality and make data invertible and (2) to search for relevant group information in the data. The performance of FLDA is somewhat similar to PLSDA with 93.8% and 94.4% of the SRs for the external validation, respectively. An important result is that, FLDA based on PLS scores classifies smaller groups better than PLSDA itself, but bigger groups have smaller SRs in FLDA. The optimal number of PLS components required for the FLDA method is the same as for PLSDA. Since we use PLS scores to implement FLDA, the FLDA method can be extended to the use of several blocks of data in the same manner as PLSDA can be extended to MBPLSDA (Westerhuis, Kourti, & MacGregor, 1998). The same property is relevant for all the methods used in the study, except for random forest (RF) which is extended by different blocks differently. This is because all other classification methods considered in this study employ PLS modelling and thus can be extended to several blocks.

4.11.4 PLSDA and FLDA coupled with HCA

In the classification approach based on PLSDA and FLDA classifiers discussed in the previous section, all 10 groups of genera were classified at once, i.e. a potential hierarchical structure in the data was not taken into account. Therefore, we wanted to investigate if a hierarchical classification structure could further improve the results. Since the beginning of the use of FTIR spectroscopy for the identification of microorganism, it has been pointed out that utilizing a hierarchical structure may improve the classification results considerably. This has been discussed extensively in the context of ANN and their use in FTIR spectroscopy for classification of microorganisms (Udelhoven, Novozhilov, & Schmitt, 2003). Rebuffo, Schmitt, Wenning, von Stetten, and Scherer (2006) employed HCA for establishing a hierarchical structure of the data and subsequently used this structure in combination with ANN to identify *Listeria* species. The authors showed that using this approach, they could achieve 96% SR in external validation of the differentiation of five species. Another study by Rebuffo-Scheer, Schmitt, and Scherer, (2007) presented the results of classification *Listeria monocytogenes* serovars using HCA as preliminary technique for establishing hierarchical tree used by ANN. The SRs of identification O and H antigens were 98.8% and 91.6%, respectively.

We used the HCA approach for establishing hierarchical trees for our data (Fig. 10 in section 4.6 “Cluster Analysis for Establishing a Hierarchical Tree”). Two models were trained: one with PLSDA as a classifier in each node of the tree and one with FLDA for the same tree. In total nine classifiers, one in each node, were trained in each model. While it is tedious and time consuming to establish a tree, valuable insight into the data structure is obtained, which may result in a deeper understanding of the taxonomic relationships between the groups. The classification result obtained by PLSDA classifiers in a hierarchical tree was 96.3%, which is slightly higher than the result achieved with PLSDA for the classification into all groups by one model (94.4%).The improvement of the SR is not very big, but the problem of misclassification of small groups is eliminated. Thus, using PLSDA together with a hierarchical tree structure adds an important improvement, since it solves the problem that some groups may be represented by a small number of samples in a training set. However, these groups may be equally important for the identification of new unknown samples. The computational time for setting up the classification models is slightly increased to two minutes. For any practical purposes this increase of computation time is no obstacle. In contrast to PLSDA, the performance of FLDA in combination with hierarchical trees is worse compared to FLDA in a one-classifier model, where all 10 groups were classified at once. This lower SR is due to the fact that almost half of the samples of the biggest group *Sac* were misclassified as the group *Pic* on a lowest level of the hierarchical tree. The SRs for other groups using FLDA with the hierarchical tree are the same or slightly worse compared to FLDA without a hierarchical tree. Thereby, using hierarchical tree combined with FLDA is not preferred for a classification of the data set we used because it takes much more time to set up a hierarchical classification scheme compared to the one-classifier classification.

4.11.5 One-Versus-All (OVA)

The OVA approach is a very common approach in the machine learning community. It is usually used in a combination with support vector machines (SVM) as a classifier when a multiclass problem needs to be reduced to a binary task, which SVM is more suitable for (Vapnik, 1995). SVM by itself is used for example for classification of images, recognizing hand-written characters, and in biology for analyzing protein and DNA sequences (Hua & Sun, 2001). The OVA approach has not been used for the classification of microorganisms by FTIR spectroscopy to our knowledge. We wanted to use the OVA concept of creating a binary tree but using PLSDA and FLDA as the classifiers. The idea was to reduce the multiclass problem to several binary problems and to check if it can improve the classification. Our results show that the OVA approach of establishing a hierarchical structure is time consuming since very many models are to be established and the best is to be selected for each node. To establish the entire classification model either with PLSDA or FLDA used as classifiers we need around 20 minutes because in total 55 single models are trained for the nine nodes in total (see Methods for details). Both classifiers, PLSDA and FLDA, perform similarly well in the OVA approach, but the SR is slightly higher for PLSDA and equal to 91.4%, whereas the SR for FLDA is equal to 90.7%. As we can see, the performance of the methods is lower than when we used PLSDA or FLDA directly to classify ten groups at once. Based on the results we can conclude that it is not reasonable to employ the OVA approach combined with PLSDA and FLDA classifiers for the data set we used. However, this is an interesting approach and further analysis and development may be needed to adapt it to FTIR data.

4.11.6 One-Versus-One (OVO)

The OVO approach is a well known method for overcoming multiclass problems using SVM as a classifier. Some authors suggest that OVO scheme can perform better than OVA (Hsu & Lin, 2002) but some authors say the opposite (Rifkin & Klautau, 2004). Our results of applying the OVO approach cannot be directly compared to the results in the literature because the method has not been used for classification of microorganisms based on FTIR spectroscopy to our knowledge. The concept of OVO approach is to train classifiers for all possible combinations of pairs of groups presented in the dataset. Once $g(g-1)/2$ classifiers are trained, where g is a number of classes in a data set, each classifier gives one vote and the majority voting scheme is used to assign a sample to the winner class. We used PLSDA and FLDA methods here as classifiers. The number of single models to be trained in our case is equal to 45 and a computational time is around 8 minutes. The validation results showed a very low performance, when OVO was used in a combination with PLSDA and FLDA. The SRs were equal to 79.0% and 64.8%, respectively. Several groups were completely misclassified. According to Furnkranz (2002) the simple majority voting scheme in principal can lead to incorrect final classification. Assume the best-case scenario, when the classification model works well: from the 45 classifiers trained in our model, 9 classifiers can give a correct vote to a group number x. In this case the sample can at most get 8 votes for any other class y.In this situation the sampleis assigned to the group number x. However, in the unlucky situation, if each classifier performs poorly, applying majority voting will mislead. This may be what happened in our case: the 45 models established by PLSDA between only two groups might be too simple to classify samples correctly from a validation data set. Perhaps another classifier combined with OVO approach could perform better or a model complexity of the PLSDA classifier could be increased. To increase the complexity of the models an easy way is to set up a minimum number of PLS components to a certain value. In our study the results suggest that the OVO approach combined with either PLSDA or FLDA is not an optimal approach for identification of microorganisms based on FTIR spectroscopy.

4.11.7 Random forest (RF)

The last method, which we used for the classification of microorganisms, is RF. RF is a method that has been developed by Breiman, (2001). Only recently the method has become popular in biomedical FTIR microspectroscopy for the detection of cancerous tissues (Kallenbach-Thieltges, 2012). It was also used for biomarker identification in biofluids (Ollesch, 2013). RF is known to handle high-dimensional data very well. We wanted to evaluate RF applied to hierarchically structured data of microorganisms. Compared to other methods presented in this study, RF has the highest SR, which is equal to 97.5% for the SAB dataset. The performance of the RF is slightly better than PLSDA combined with HCA, which had the SR equal to 96.3%, however requires much less computational work and analysis to be done. It works relatively fast compared for example to PLSDA or FLDA in OVA or OVO approach. The time needed to generate 1000 trees, which is enough for establishing a stable and reliable classification model, is around two minutes. It has very high accuracy in classification of both small and big groups. After applying RF to two-block datasets from two cultivation media, the classification results could be further improved compared to one-block data. In Table 3 all classifiers are presented in diminishing order of SR.

Based on above mentioned properties of the methods and Table 3 we can say that RF is a simple in implementation and reliable method for identification of microorganisms. It fulfils all the criteria, which were set and requires minimum affords for the model optimization and analysis, whereas other tested methods require much more experience and effort. Fernández-Delgado, Cernadas,Barro, and Amorim (2014) tested 179 different classifiers and applied them to 121 different datasets from UCI Machine Learning Repository. However in their study authors did not analyze spectroscopic data but they showed that RF is the method that can solve different classification problems and can handle data from different origins.

Table 3: The comparison of classification approaches used in the study: the computational time and success rates (SR).

Classification approach	Computational time, min.	SR, %
RF	2	97.5
PLSDA+HCA	2	96.3
PLSDA	0.5	94.4
FLDA	0.5	93.8
PLSDA+OVA	20	91.4
FLDA+OVA	20	90.7
FLDA+HCA	2	83.3
PLSDA+OVO	8	79.0
FLDA+OVO	8	64.8

5. Conclusion

High-dimensional phenotyping data of microorganisms based on FTIR spectroscopy and MALDI-TOF are complex data and require powerful tools for the analysis. High-dimensional phenotyping data can be used to reveal taxonomic relations between microorganisms and to establish classification models for the identification of unknown samples. In this study, we considered a broad spectrum of methods, which can be applied for the analysis of hierarchically structured data such FTIR spectroscopic data of microorganisms. The aim was to investigate, which methods perform best with respect to identification success rate and simplicity of a model establishment. In total, we considered nine classification methods, namely PLSDA, FLDA, HCA PLSDA, HCA FLDA, OVA PLSDA, OVA FLDA, OVO PLSDA, OVO FLDA, and RF. It has been previously pointed out in the literature that the organization of data in a hierarchical structure with classifiers in every note, improves the classification results (Liland et al., 2014; Rebuffo et al., 2006; Zimmermann et al., 2016). This observation could be partially confirmed by our work. PLSDA applied to a tree established by HCA could improve the results of ordinary PLSDA, whereas FLDA combined with HCA could not show better performance than FLDA applied directly to the data. Notwithstanding, our results showed that the best classification method according to the predetermined criteria is RF, while RF is a method that does not require a predefined hierarchical structure. RF is relatively simple in implementation and can be used straightforward without optimizing too many parameters. This is of great advantage for users that are not experts in data analysis. The RF method is not prone to overfitting due to randomness of the selected variables. It does not require model calibration in a form that the other methods do and that has to be selected carefully. In RF, error estimation is done using a so-called "out-of-bag" set that is a part of a training set and thus eliminates the necessity to use an extra set for validation. The internal validation and the error estimation of other methods in this study are done by using leave-one-run-out cross-validation. The validation strategy chosen is the most conservative one for our data, since each run represents an independent repetition of the same experiment. RF is not the fastest method among the nine methods tested, but it works relatively fast: it takes around two minutes to create a forest of 1000 trees, which was sufficient for the data we investigated. The accuracy of RF was the highest among the methods we considered: 100% SR was achieved for seven groups out of ten and the total SR was equal to 97.5%. Another great advantage of RF is that it can be easily used for the analysis of multiblock (MB) data or heterogeneous data. The data used in our study contained several blocks of FTIR data obtained from the same samples. The microorganisms were grown on different, but well-defined media resulting in different phenotypic expressions of the microorganisms. FTIR spectra from the different phenotypic expressions were acquired. Adding these additional blocks of data for the classification analysis is a simple concatenation of the blocks for the RF method. The performance of the RF model was improved, when RF was extended by these additional blocks of information.

Finally, we conclude that using a hierarchical structure improves classification results in general, while RF can achieve the same or even better results without prior knowledge about the hierarchical structure of the data. It outperforms all the other methods we tested. A disadvantage of RF is that it does not have as rich visualization possibilities as methods based on latent variables. For example, PLS based methods allow exploring variable spaces via loading plots and correlation loading plots and samples spaces via score plots. In contrast, RF has the opportunity to select variables based on their importance (Touw et al., 2013). Therefore, a suggestion would be to use both methods that they can reinforce each other in order to establish a reliable model and to gain a comprehensive analysis of a data.

In future, it would be interesting to try other optimization algorithms for model calibration in PLSDA method that will take into account different group sizes as was mentioned in "Discussion" section. This was the biggest concern in the PLSDA method performance in this study: mostly the small groups were misclassified. Such approaches as OVA and OVO can be combined with other classifiers to check their performances. For example, using a support vector machines (SVM) combined with OVA approach analyzing attenuated total reflection FTIR (ATR-FTIR) spectra for diagnosis of gliomas resulted on average 93.75% and 96.53% of sensitivities and specificities, respectively (Hands et al., 2014). So, there is a high chance that these approaches can perform better for our data if other classifiers are considered. Different blocks of information, presented by different cultivation media, could be used for MB analysis with the methods we used in this study, such as PSDA and FLDA. Since we saw that RF could improve the classification by adding other blocks, the same should be valid for other classification approaches. The RF model can be optimized by using different number of variables used for growing every decision tree and by employing outlier detection property in order to obtain more accurate and robust models.

References

Adams, M. J. (1995). *Chemometrics in analytical spectroscopy.* Cambridge: The Royal Society of Chemists.

Afseth, N. K. & Kohler, A. (2012). Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems. 59*(6), 707-714. doi:10.1016/j.chemolab.2012.03.04

Ballabio, D. & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*. *16*(5), 3790-3798. doi:10.1039/c3ay40582f

Barker, M. & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*. *17*(3), 166-173. doi:10.1002/cem.785

Biau, G., Devroye, L., & Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research. 9*, 2015-2033. Retrieved from http://www.jmlr.org

Bishop, C.M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon press.

Breiman, L. (1996). Baggin predictors. *Machine Learning. 24*(2), 123-140. doi:10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Coutinho, C.P., Sá-Correia, I., & Lopes, J.L. (2009) Use of Fourier transform infrared spectroscopy and chemometrics to discriminate clinical isolates of bacteria of the *Burkholderia cepacia* complex from different species and ribopatterns. *Analytical and Bioanalytical Chemistry 394(*8), 2161-2171. doi:10.1007/s0021609-2908-4

Duygu, D.Y., Baykal, T., Acikgoz, I., & Yildiz, K. (2009). Review. Fourier transform infrared (FT-IR) spectroscopy for biological studies. *Gazi University Journal of Science 22*(3), 117-121. Retrieved from http://www.researchgate.net

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research 15,* 3133-3181. Retrieved from http://www.jmlr.org/papers/v15/delgado14a.html

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. *7*(2), 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x

Furnkranz, J. (2002). Round robin classification. *Journal of Machine Learning Research*. *2,* 721-747. doi:10.1162/153244302320884605

Golub, G.H. & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische mathematik. 14*(5), 403-420. doi:10.1007/bf02163027

Goodacre, R., Timmins, É.M., Rooney, P.J., Rowland, J.J., & Kell, D.B. (1996). Rapid identification of *Streptococcus* and *Enterococcus* species using diffuse reflectance-absorbance Fourier transform infrared spectroscopy and artificial neural networks. FEMS Microbiology Letters. 140, 233-239. doi:10.1111/j.1574-6968.1996.tb08342.x233-239

Hands, J.R., Dorling, K.M., Abel, P., Ashton, K.M., Brodbelt, A., Davis, C., ... & Baker, M.J. (2014). Attenuated total reflection Fourier transform infrared (ATR-FTIR) spectral discrimination of brain tumour severity from serum samples. *Journal of Biophotonics*. *7*(3-4), 189-199. doi:10.1002/jbio.201300149

Hastie, T., Tibshirani, R., & Friedman, J.H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer

Helm, D., Labischinski, H., Schallehn, G., & Naumann, D. (1991). Classification and identification of bacteria by Fourier-transform infrared spectroscopy. *Journal of General Microbiology*. *137*(1), 69– 79. doi:10.1099/00221287-137-1-69

Hsu, C.-W. & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks. 13*(2), 415-425. doi:10.1109/72.991427

Hua, S. & Sun, Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *Journal of Molecular Biology. 308(*2), 397-407. doi:10.1006/jmbi.2001.4580

Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*. *2*(3), 211-228. doi:10.1002/cem.1180020306

Ilari, J.L., Martens, H., & Isaksson, T. (1988). Determination of particle size in powders by scatter correction in diffuse near-infrared reflectance. *Applied Spectroscopy*. *42*(5), 722-728. doi:10.1366/0003702884429058

Jackson, J.E. (1991). *A users' guide to principal components*. New York: Wiley.

Kallenbach-Thieltges, A., Großerüschkamp, F., Mosig, A., Diem, M., Tannapfel, A. & Gerwert, K. (2013), Immunohistochemistry, histopathology and infrared spectral histopathology of colon cancer tissue sections. *Journal of Biophotonics. 6*(1), 88–100. doi:10.1002/jbio.201200132

Kansiz, M., Heraud, P., Wood, B., Burden, F., Beardall, J., & McNaughton, D. (1999). Fourier Transform Infrared microspectroscopy and chemometrics as a tool for the discrimination of cyanobacterial strains. *Phytochemistry*. *52*(3), 407-417. doi:10.1016/S0031-9422(99)00212-5

Kohler, A., Afseth, N.K., & Martens, H. (2010). Chemometrics in biospectroscopy. In E. Li-Chan, P.R. Griffiths, & J.M. Chalmers (Eds.), Applications of vibrational spectroscopy in food science (pp. 89-106). Chichester: John Wiley & Sons, Ltd.

Kohler, A., Kirschner, C., Oust, A., & Martens, H. (2005). Extended multiplicative signal correction as a tool for separation and characterization of physical and chemical information in Fourier transform infrared microscopy images of cryo-sections of beef loin. *Applied Spectroscopy*. *59*(6), 707-716. doi:10.1366/0003702054280649

Liland, K.H., Kohler, A., & Shapaval, V. (2014). Hot PLS – a framework for hierarchically ordered taxonomic classification by partial least squares. *Chemometrics and Intelligent Laboratory Systems*. *138*, 41-47. doi:10.1016/j.chemolab.2014.07.010

Maguire, A., Vega-Carrascal, I., Bryant, J., White, L., Howe, O., Lyng, F.M., & Meade, A.D. (2012). Competitive evaluation of data mining algorithms for use in classification of leukocyte subtypes with Raman microspectroscopy. *Analyst*. *140*, 2473-2481 doi:10.1039/c4an01887g

Maquelin, K., Choo-Smith, L.-P., Endtz, H.P., Bruining, H.A., & Puppels, G.J. (2002). Rapid identification of Candida species by confocal raman microspectroscopy. *Journal of Clinical Microbiology*. *40*(2), 594-600. doi:10.1128/JCM.40.2.594-600.2002

Maquelin, K., Kirschner, C., Choo-Smith, L.-P., Ngo-Thi, N. A., van Vreeswijk, T., Stämmler, M., Puppels, G. J. (2003). Prospective study of the performance of vibrational spectroscopies for rapid identification of bacterial and fungal pathogens recovered from blood cultures. *Journal of Clinical Microbiology*, *41*(1), 324–329. doi:10.1128/JCM.41.1.324-329.2003

Martens, H. & Næs, T., (1989). *Multivariate calibration*. Chichester: John Wiley & Sons, Ltd.

Martens, H. & Stark, E. (1991). Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*. *9*(8), 625-635. doi:10.1016/0731-7085(91)80188-f

Mittal, D., Rani, A. & Ritambhara (2016). Detection and classification of focal liver lesions using support vector machine classifiers. *Journal of Biomedical Engineering and Medical Imaging*. *3*(1). doi:10.14738/jbemi.31.1821

Naumann, D., Helm, D., & Labischinski, H. (1991). Microbiological characterizations by FT-IR spectroscopy. *Nature*. *351*(6321), 81-82. doi:10.1038/351081a0

Oberreuter, H., Seiler, H., & Scherer, S. (2002) Identification of coryneform bacteria and related taxa by Fourier-transform infrared (FT-IR) spectroscopy. *International Journal of Systematic and Evolutionary Microbiology* *52*(1), 91–100. doi:10.1099/00207713-52-1-91

Ollesch, J., Drees, S.L., Heise, H.M., Behrens, T., Bruning, T., & Gerwert, K. (2013). FTIR spectroscopy of biofluids revisited: an automated approach to spectral biomarker identification. *Analyst*. *138*. 4092-4102. doi:10.1039/c3an00337j

Oust, A., Møretrø, T., Kirschner, C., Narvhus, J.A., & Kohler, A. (2004). FT-IR spectroscopy for identification of closely related lactobacilli. *Journal of Microbiological Methods. 59*(2), 149-162. doi:10.1016/j.mimet.2004.06.011

Preisner, O., Lopes, J.A., & Menezes, J.C. (2008). Uncertainty assessment in FT-IR spectroscopy based bacteria classification models. *Chemometrics and Intelligent Laboratory Systems. 94*(1), 33-42. doi:10.1016/j.chemolab.2008.06.005

Rebuffo-Scheer, C.A., Schmitt, J., & Scherer, S. (2007). Differentiation of *Listeria monocytogenes* serovars by using artificial neural network analysis of Fourier-transformed infrared spectra. *Applied and Environmental Microbiology. 73*(3), 1036–1040. doi:10.1128/AEM.02004-06

Rebuffo, C.A., Schmitt, J., Wenning, M., von Stetten, F., & Scherer, S. (2006). Reliable and rapid identification of Listeria monocytogenes and Listeria species by artificial neural network-based Fourier transform infrared spectroscopy. *Applied and Environmental Microbiology. 72*(2), 994-1000. doi:10.1128/AEM.72.2.994-1000.2006

Rifkin, R. M. & Klautau, A. (2004). In defense of one-versus-all classification. *Journal of Machine Learning Research. 5*, 101-141. Retrieved from http://www.researchgate.net

Savitzky, A. & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry. 36*(8), 1629-1639.

Shapaval, V., Walczak, B., Gognies, S., Møretrø, T., Suso, H. P., Wold Åsli, A.,… Kohler, A. (2013). FTIR spectroscopic characterization of differently cultivated food related yeasts. *Analyst. 138*, 4129-4138. doi:10.1039/c3an00304c

Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research. 8*, 1027-1061. Retrieved from http:// http://www.csmining.org/cdmc2016/

Tan, P.N., Steinbach, M., & Kumar, V. (2006). Cluster analysis: Basic concepts and algorithms. In P. N. Tan, M. Steinbach, & V. Kumar (Eds.), *Introduction to data mining* (pp. 487-567). Boston, MA: Pearson Addison Wesley

Touw, W.G., Bayjanov, J.R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & van Hijum, S.A.F.T. (2013). Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics*, *14*(3), 315–326. doi:10.1093/bib/bbs034

Udelhoven, T., Naumann, D., & Schmitt, J. (2000). Development of a hierarchical classification system with artificial neural networks and FT-IR spectra for the identification of bacteria, *Applied Spectroscopy. 54*(10), 1471-1479. doi:10.1366/0003702001948619

Udelhoven, T., Novozhilov, M., & Schmitt, J. (2003). The NeuroDeveloper: a tool for modular neural classification of spectroscopic data. *Chemometrics and Intelligent Laboratory Systems. 66*(2), 219–226. doi:10.1016/S0169-7439(02)00161-2

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Warringer, J., & Blomberg, A. (2003). Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in *Saccharomyces cerevisiae.Yeast. 20*(1), 53-67. doi:10.1002/yea.931

Wenning, M., Buchl, N.R., & Scherer, S. (2010). Species and strain identification of lactic acid bacteria using FTIR spectroscopy and artificial neural networks. *Journal of Biophotonics. 3*(8-9), 493-505. doi:10.1002/jbio.201000015

Wenning, M. & Scherer, S. (2013). Identification of microorganisms by FTIR spectroscopy: perspectives and limitations of the method. *Applied Microbiology and Biotechnology. 97*(16), 7111-7120. doi:10.1007/s00253-013-5087-3.

Westerhuis, J. A., Kourti, T. & MacGregor, J. F. (1998), Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, *12*(5), 301–321. doi:10.1002/(SICI)1099-128X(199809/10)12:5<301::AID-CEM515>3.0.CO;2-S

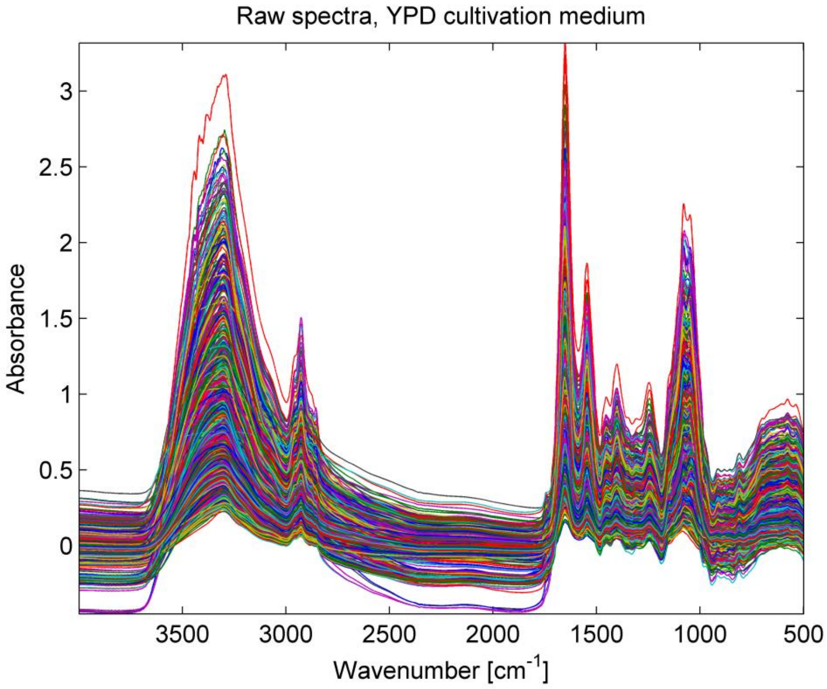
Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems. 58*, 109-103. doi:10.1016/s0169-7439(01)00155-1

Zimmermann, B. & Kohler, A. (2013). Optimizing Savitzky-Golay parameters for improving spectral resolution and quantification in infrared spectroscopy. *Applied Spectroscopy. 67*(8), 892-902. doi:10.1366/12-06723

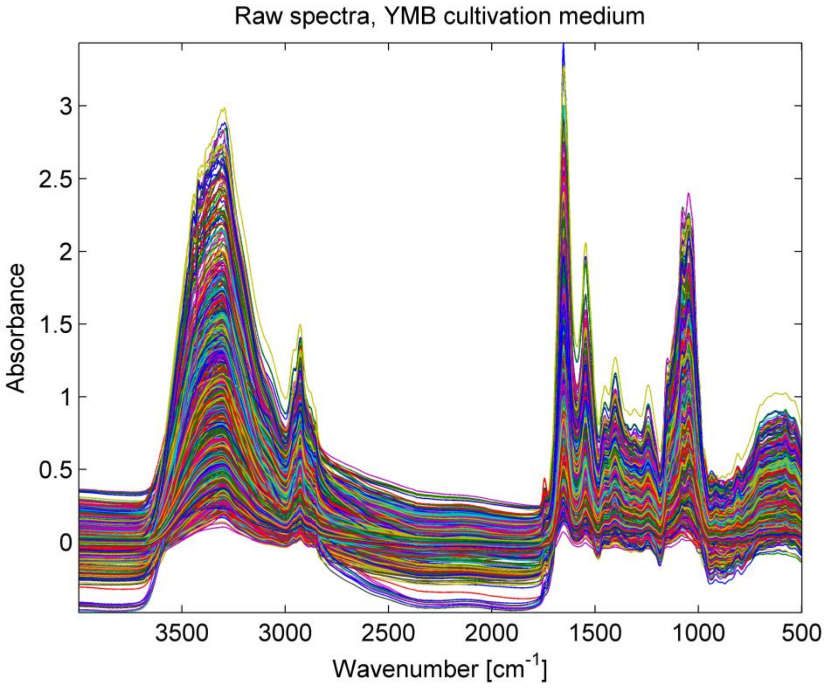
Zimmermann, B., Tafintseva, V., Bağcıoğlu, M., Berdahl, M.H., & Kohler, A. (2016) Analysis of allergenic pollen by FTIR microspectroscopy, *Analytical Chemistry 88*(1), 803–811. doi:10.1021/acs.analchem.5b03208

Appendix

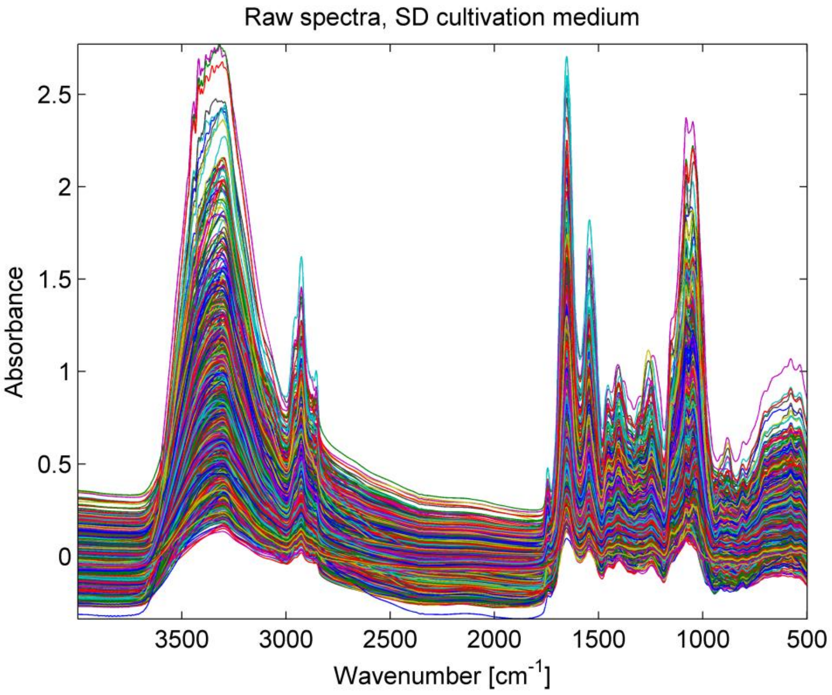
Results



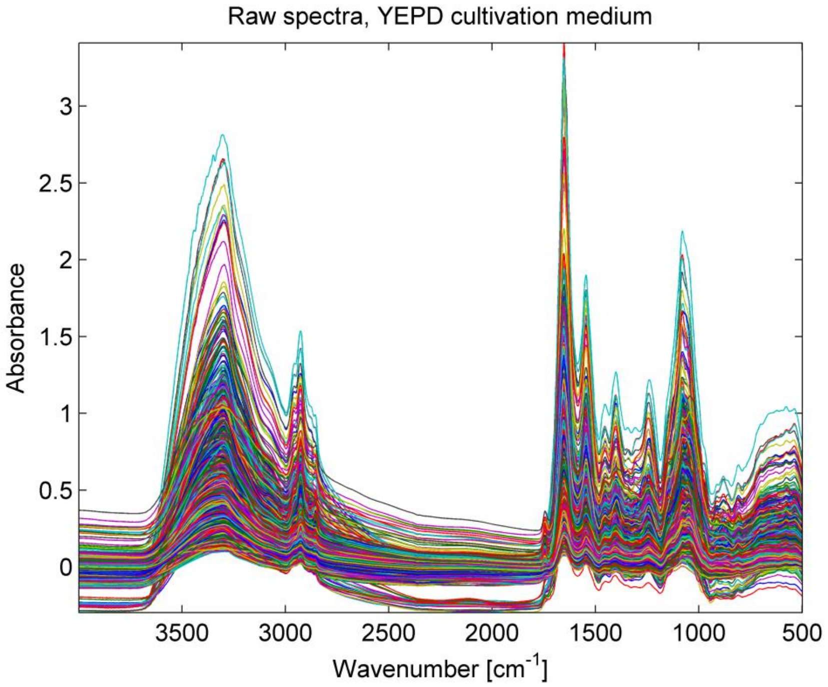
Raw spectra recorded in a range 4000 cm^{-1} and 500 cm^{-1} of yeasts grown on YPD medium



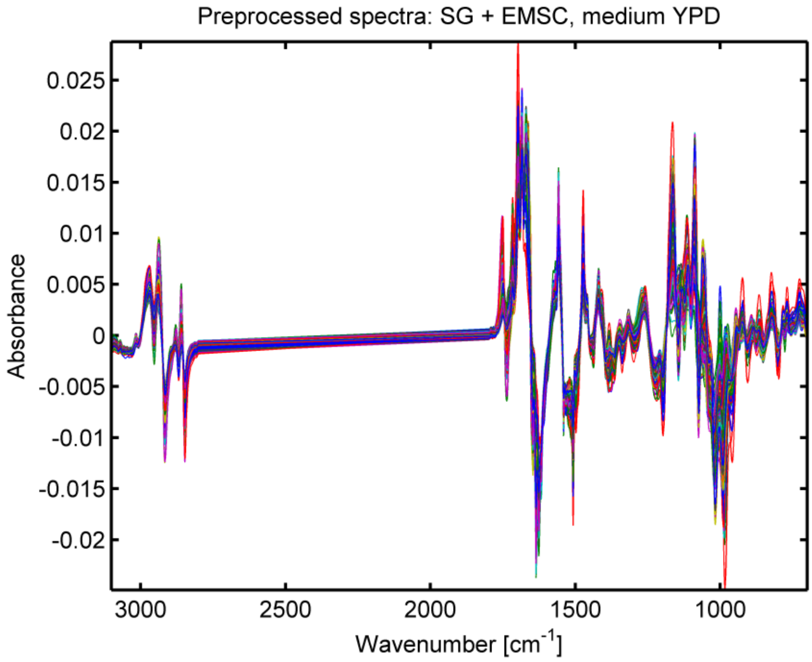
Raw spectra recorded in a range 4000 cm^{-1} and 500 cm^{-1} of yeasts grown on YMB medium



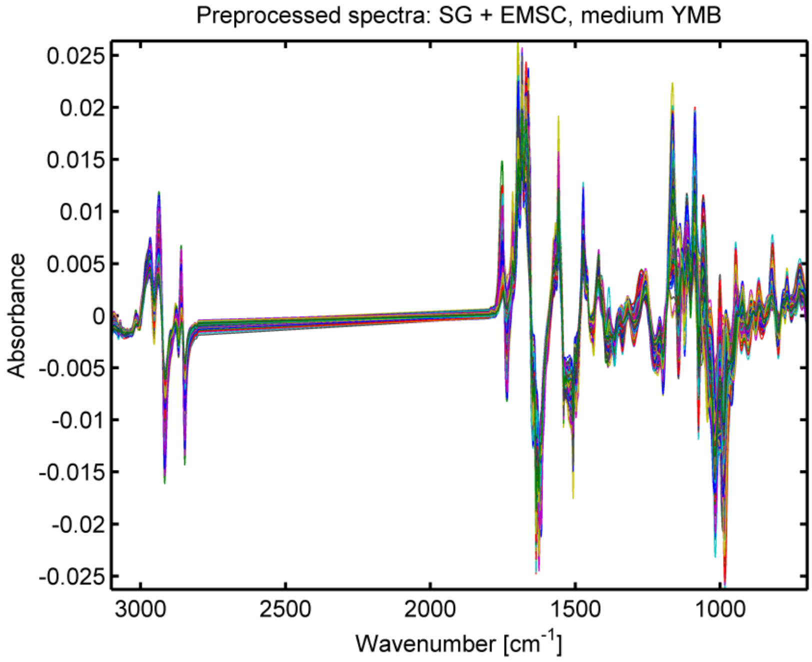
Raw spectra recorded in a range 4000 cm^{-1} and 500 cm^{-1} of yeasts grown on SD medium



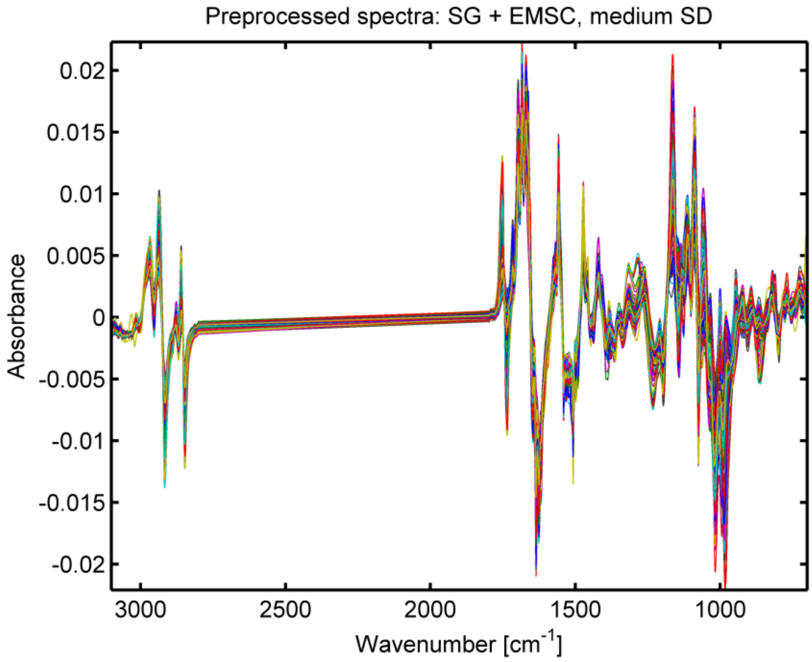
Raw spectra recorded in a range 4000 cm^{-1} and 500 cm^{-1} of yeasts grown on YEPD medium



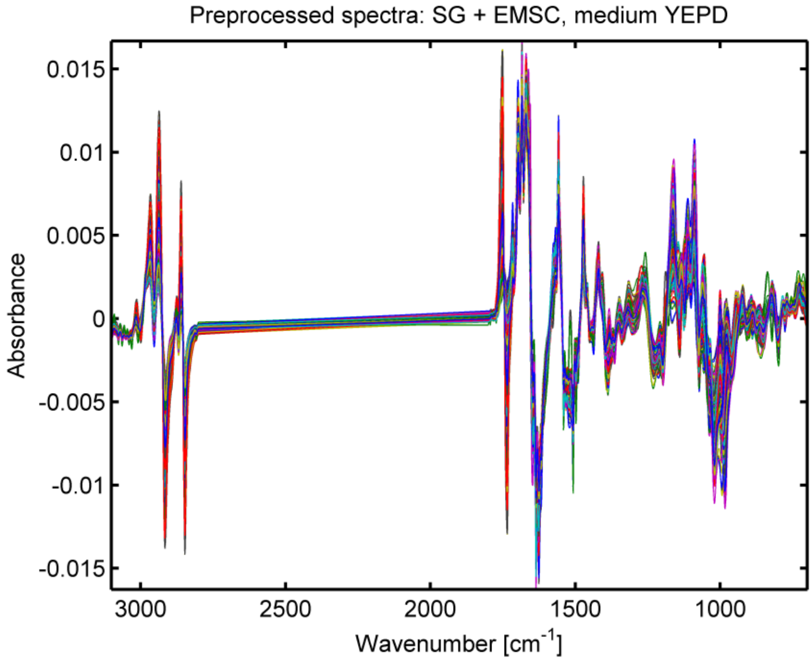
The YPD dataset spectra preprocessed by SG and EMSC algorithms. The technical replicates are averaged. The 3100-2800 cm^{-1} & 1800-700 cm^{-1} spectral regions are selected.



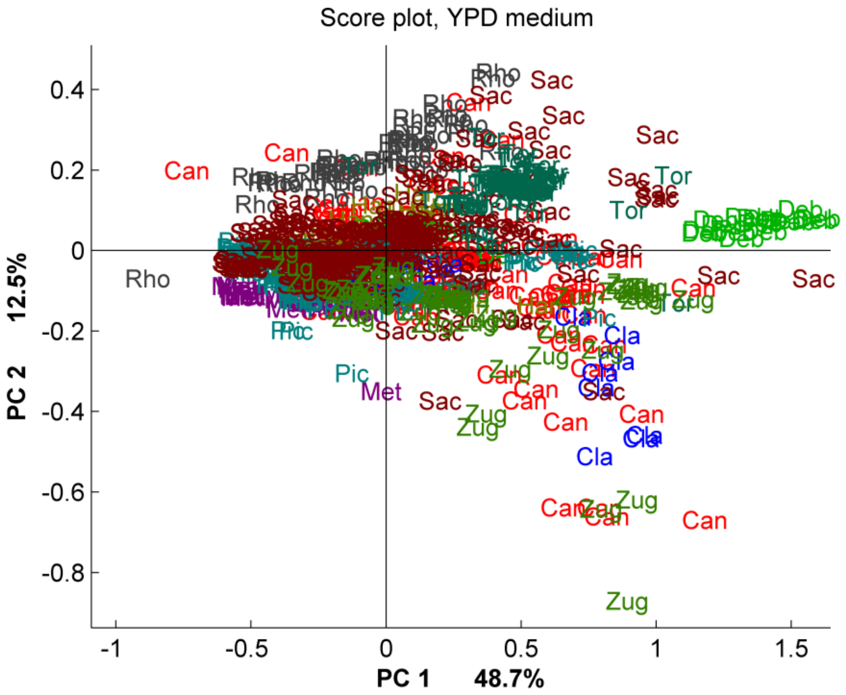
The YMB dataset spectra preprocessed by SG and EMSC algorithms. The technical replicates are averaged. The 3100-2800 cm^{-1} & 1800-700 cm^{-1} spectral regions are selected.



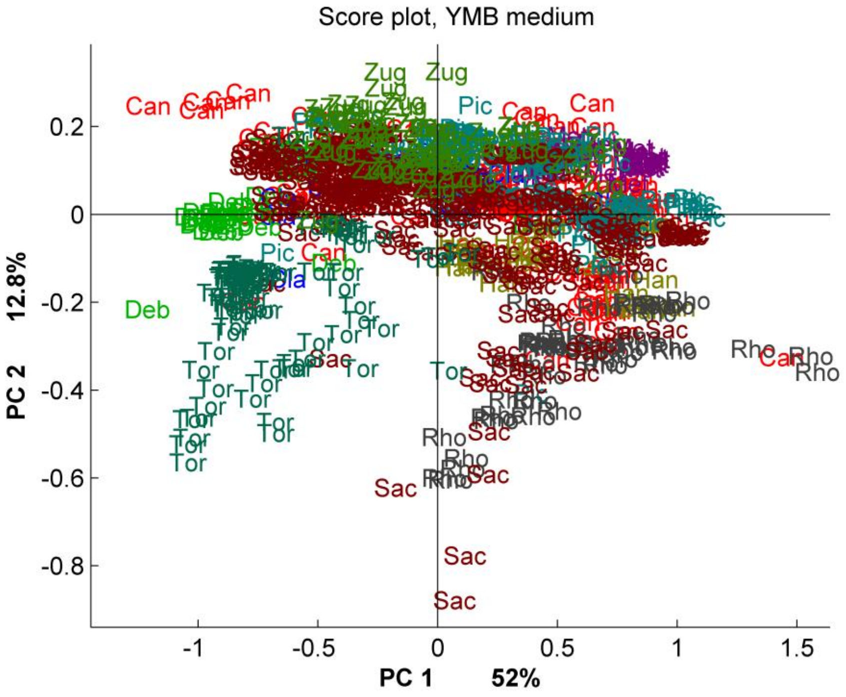
The SD dataset spectra preprocessed by SG and EMSC algorithms. The technical replicates are averaged. The 3100-2800 cm^{-1} & 1800-700 cm^{-1} spectral regions are selected.



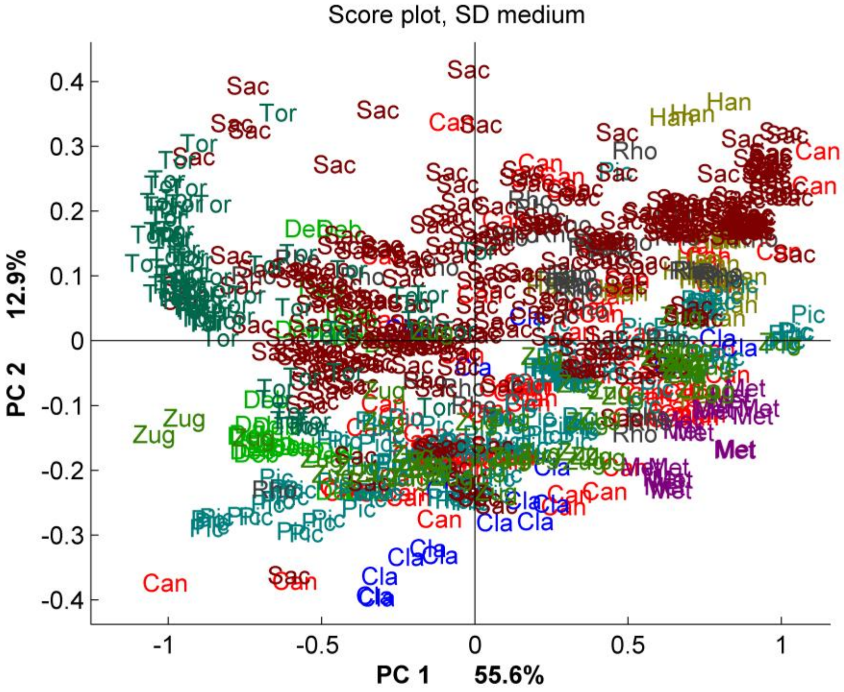
The YEPD dataset spectra preprocessed by SG and EMSC algorithms. The technical replicates are averaged. The 3100-2800 cm^{-1} & 1800-700 cm^{-1} spectral regions are selected.



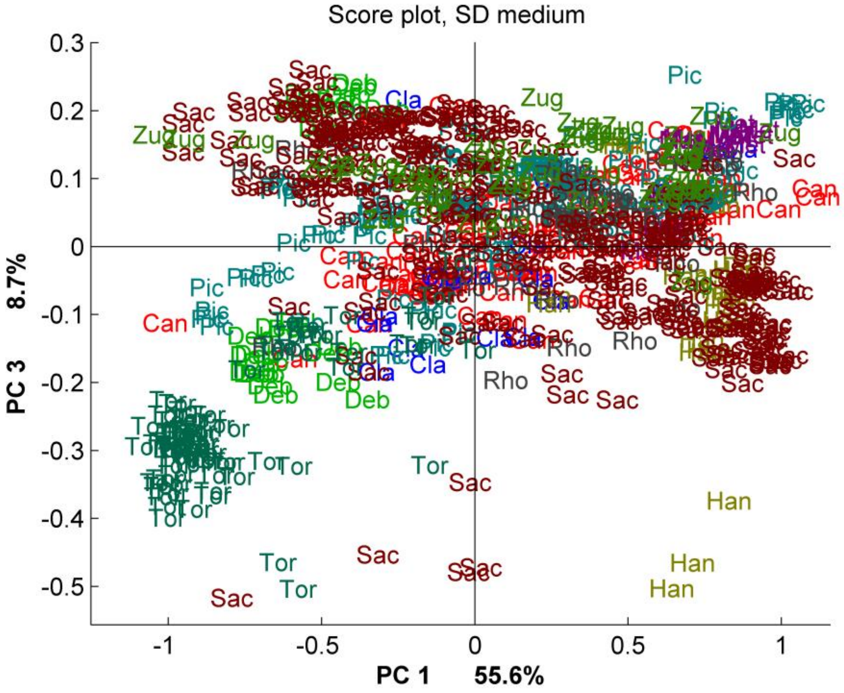
Score plot of PC1 vs. PC2 in PCA analysis of the YPD dataset preprocessed by SG and EMSC



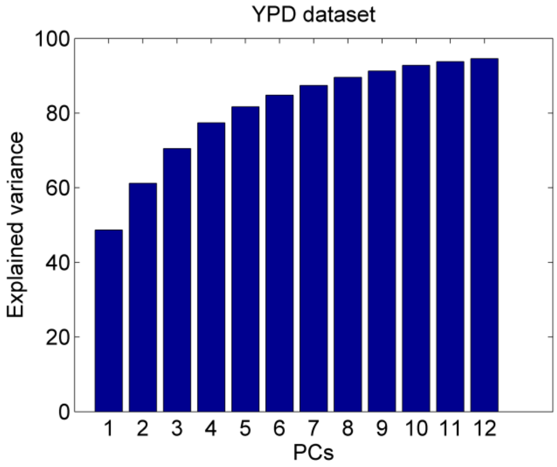
Score plot of PC1 vs. PC2 in PCA analysis of the YMB dataset preprocessed by SG and EMSC



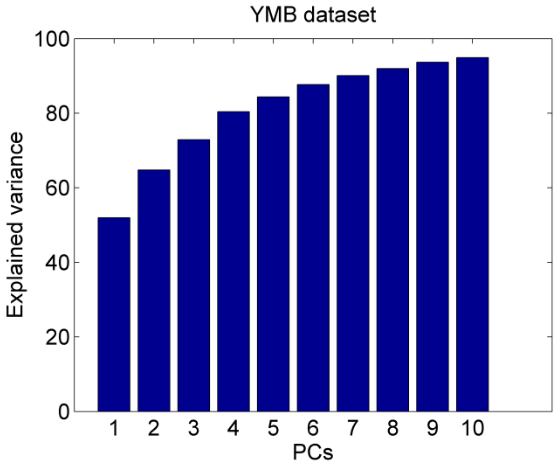
Score plot of PC1 vs. PC2 in PCA analysis of the SD dataset preprocessed by SG and EMSC



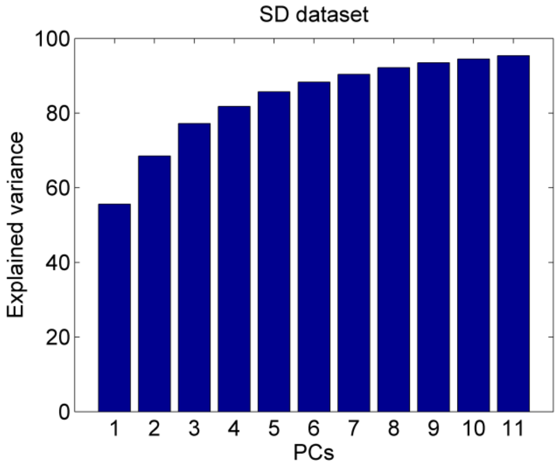
Score plot of PC1 vs. PC3 in PCA analysis of the SD dataset preprocessed by SG and EMSC



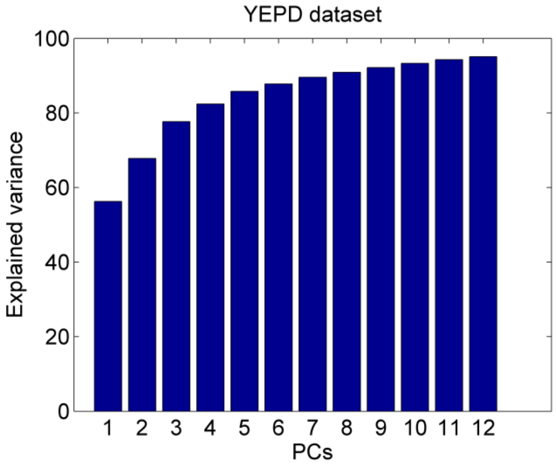
95% cumulative explained variance by the first 12 PCs in a PCA analysis of the YPD



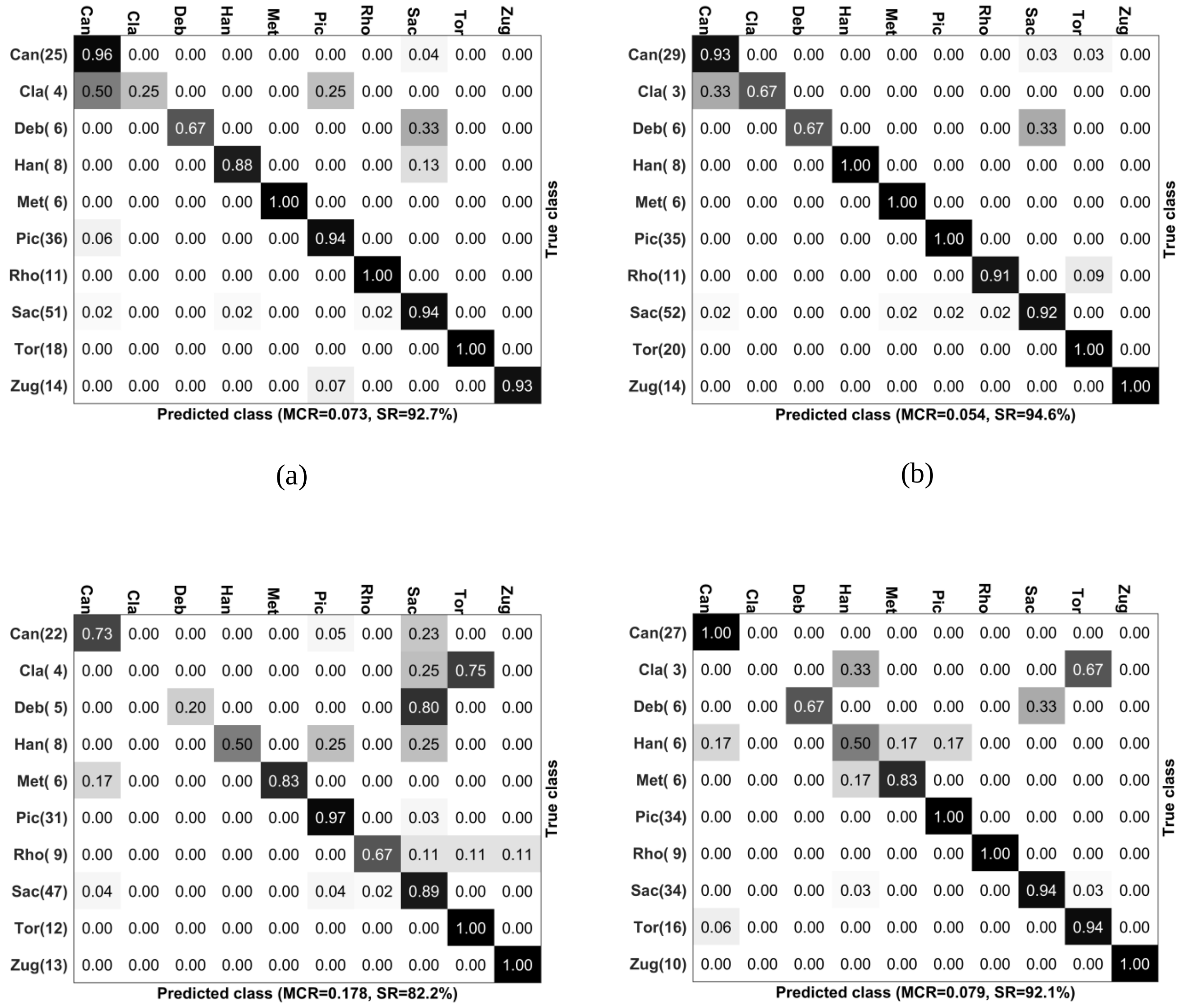
95% cumulative explained variance by the first 10 PCs in a PCA analysis of the YMB subset



95% cumulative explained variance by the first 11 PCs in a PCA analysis of the SD



95% cumulative explained variance by the first 12 PCs in a PCA analysis of the YEPD subset



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(25)	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00
Cla(4)	0.50	0.25	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	0.88	0.00	0.00	0.00	0.13	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(36)	0.06	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(51)	0.02	0.00	0.00	0.02	0.00	0.00	0.02	0.94	0.00	0.00
Tor(18)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(14)	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.93

(a)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(29)	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00
Cla(3)	0.33	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.09	0.00
Sac(52)	0.02	0.00	0.00	0.00	0.02	0.02	0.02	0.92	0.00	0.00
Tor(20)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(14)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(b)

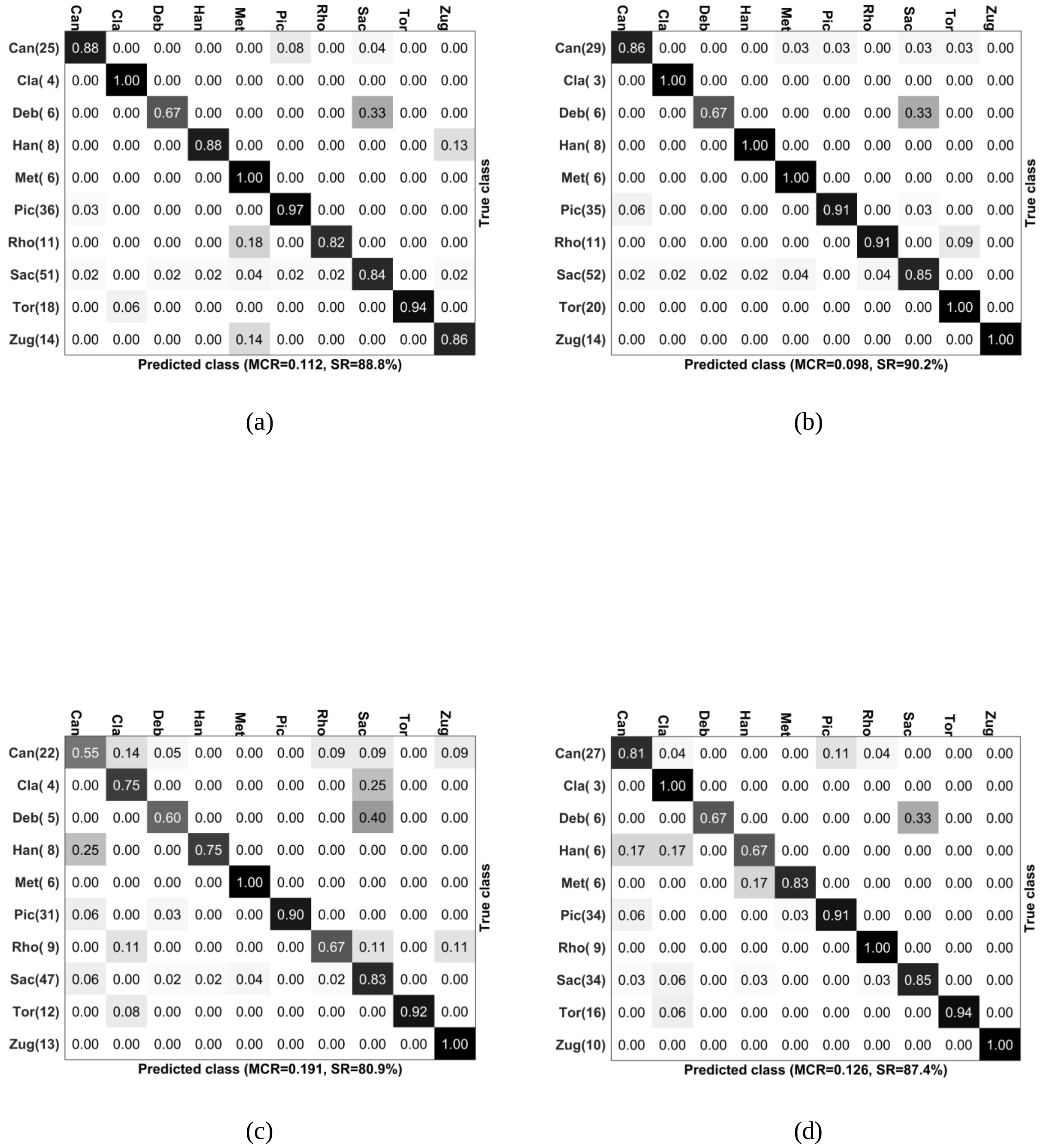
	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(22)	0.73	0.00	0.00	0.00	0.00	0.05	0.00	0.23	0.00	0.00
Cla(4)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.75	0.00
Deb(5)	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.80	0.00	0.00
Han(8)	0.00	0.00	0.00	0.50	0.00	0.25	0.00	0.25	0.00	0.00
Met(6)	0.17	0.00	0.00	0.00	0.83	0.00	0.00	0.00	0.00	0.00
Pic(31)	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.03	0.00	0.00
Rho(9)	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.11	0.11	0.11
Sac(47)	0.04	0.00	0.00	0.00	0.00	0.04	0.02	0.89	0.00	0.00
Tor(12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(13)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(c)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(27)	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cla(3)	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.67	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(6)	0.17	0.00	0.00	0.50	0.17	0.17	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.17	0.83	0.00	0.00	0.00	0.00	0.00
Pic(34)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(9)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(34)	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.94	0.03	0.00
Tor(16)	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00
Zug(10)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(d)

Confusion matrices for the validation where run one was kept out. Classification was done on different cultivation media (a) YPD, (b) YMB, (c) SD, (d) YEPD datasets. PLSDA classifier is used for the classification analysis.



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(25)	0.88	0.00	0.00	0.00	0.00	0.08	0.00	0.04	0.00	0.00
Cla(4)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	0.88	0.00	0.00	0.00	0.00	0.00	0.13
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(36)	0.03	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.18	0.00	0.82	0.00	0.00	0.00
Sac(51)	0.02	0.00	0.02	0.02	0.04	0.02	0.02	0.84	0.00	0.02
Tor(18)	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00
Zug(14)	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.86

Predicted class (MCR=0.112, SR=88.8%)

(a)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(29)	0.86	0.00	0.00	0.00	0.03	0.03	0.00	0.03	0.03	0.00
Cla(3)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.06	0.00	0.00	0.00	0.00	0.91	0.00	0.03	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.09	0.00
Sac(52)	0.02	0.02	0.02	0.02	0.04	0.00	0.04	0.85	0.00	0.00
Tor(20)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(14)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Predicted class (MCR=0.098, SR=90.2%)

(b)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(22)	0.55	0.14	0.05	0.00	0.00	0.00	0.09	0.09	0.00	0.09
Cla(4)	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00
Deb(5)	0.00	0.00	0.60	0.00	0.00	0.00	0.00	0.40	0.00	0.00
Han(8)	0.25	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(31)	0.06	0.00	0.03	0.00	0.00	0.90	0.00	0.00	0.00	0.00
Rho(9)	0.00	0.11	0.00	0.00	0.00	0.00	0.67	0.11	0.00	0.11
Sac(47)	0.06	0.00	0.02	0.02	0.04	0.00	0.02	0.83	0.00	0.00
Tor(12)	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.00
Zug(13)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Predicted class (MCR=0.191, SR=80.9%)

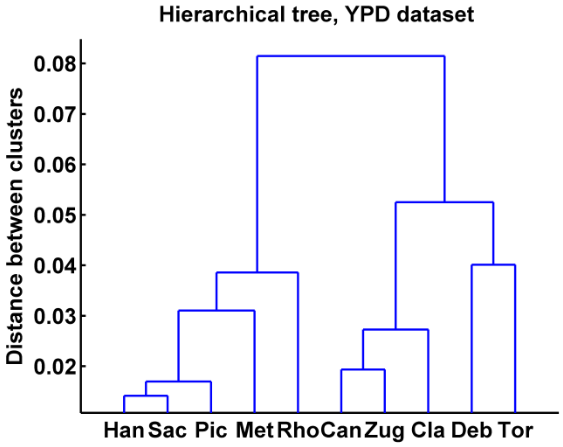
(c)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(27)	0.81	0.04	0.00	0.00	0.00	0.11	0.04	0.00	0.00	0.00
Cla(3)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(6)	0.17	0.17	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.17	0.83	0.00	0.00	0.00	0.00	0.00
Pic(34)	0.06	0.00	0.00	0.00	0.03	0.91	0.00	0.00	0.00	0.00
Rho(9)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(34)	0.03	0.06	0.00	0.03	0.00	0.00	0.03	0.85	0.00	0.00
Tor(16)	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00
Zug(10)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

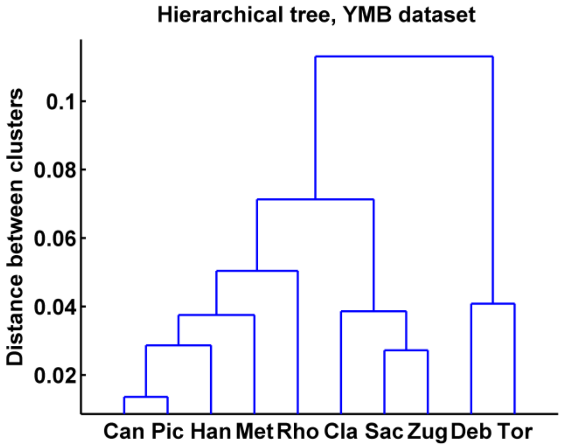
Predicted class (MCR=0.126, SR=87.4%)

(d)

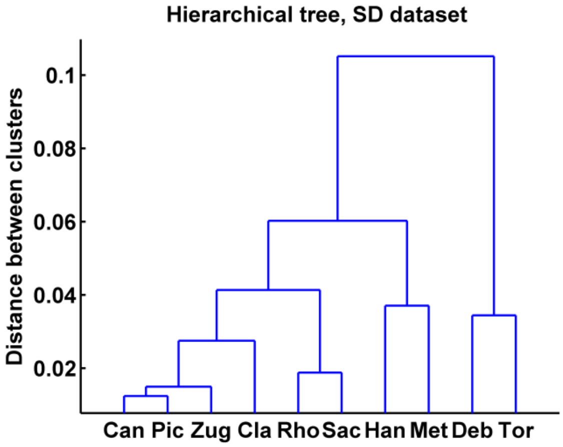
Confusion matrices for the validation where run one was kept out. Classification was done on different cultivation media (a) YPD, (b) YMB, (c) SD, (d) YEPD datasets. FLDA classifier is used for the classification analysis.



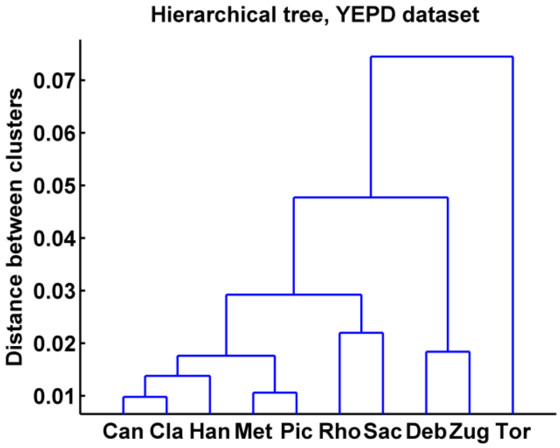
Hierarchical tree based on the group means for the YPD dataset



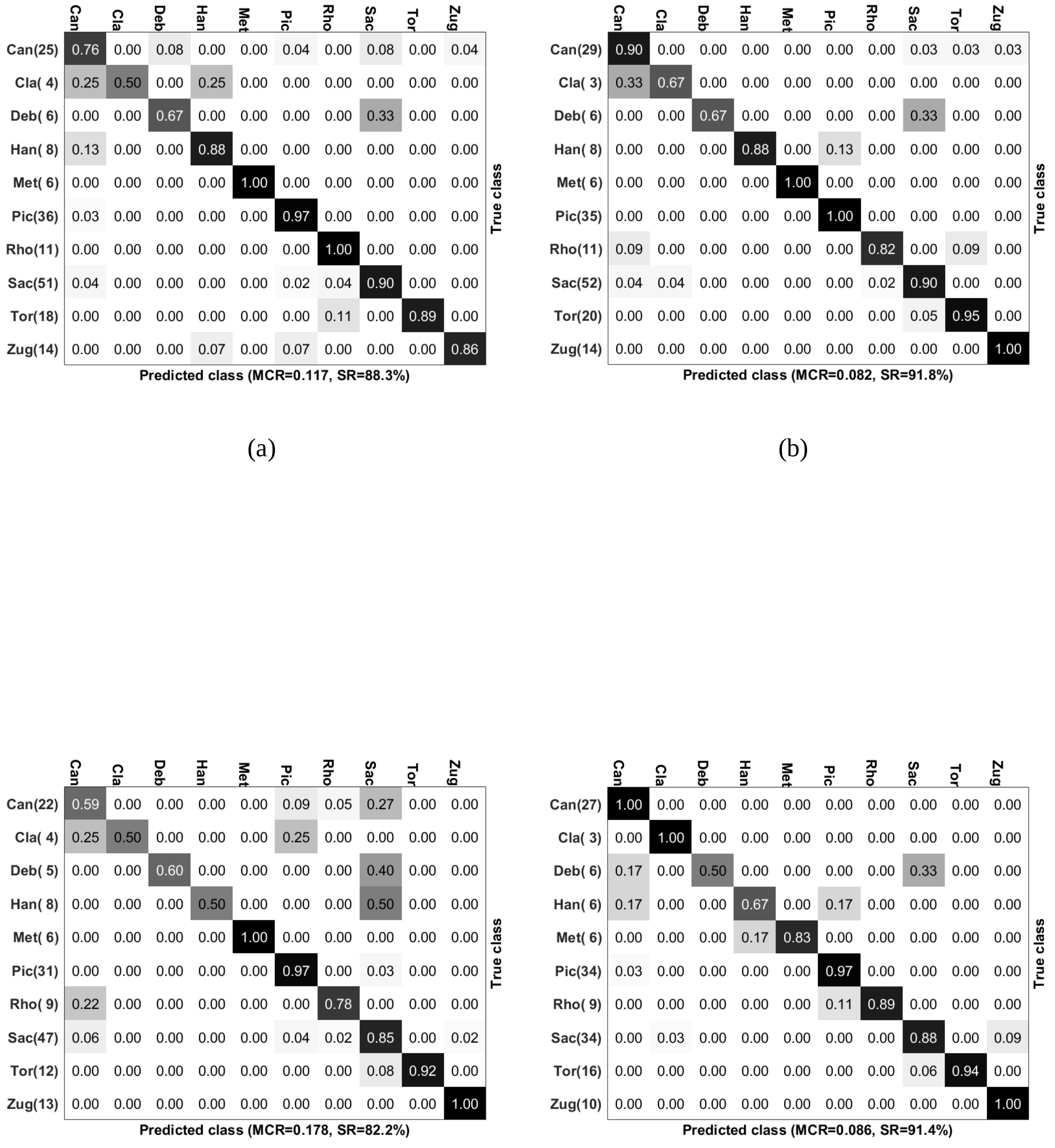
Hierarchical tree based on the group means for the YMB dataset



Hierarchical tree based on the group means for the SD dataset



Hierarchical tree based on the group means for the YEPD dataset



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(25)	0.76	0.00	0.08	0.00	0.00	0.04	0.00	0.08	0.00	0.04
Cla(4)	0.25	0.50	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.13	0.00	0.00	0.88	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(36)	0.03	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(51)	0.04	0.00	0.00	0.00	0.00	0.02	0.04	0.90	0.00	0.00
Tor(18)	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.89	0.00
Zug(14)	0.00	0.00	0.00	0.07	0.00	0.07	0.00	0.00	0.00	0.86

(a)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(29)	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.03
Cla(3)	0.33	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	0.88	0.00	0.13	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(11)	0.09	0.00	0.00	0.00	0.00	0.00	0.82	0.00	0.09	0.00
Sac(52)	0.04	0.04	0.00	0.00	0.00	0.00	0.02	0.90	0.00	0.00
Tor(20)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.95	0.00
Zug(14)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(b)

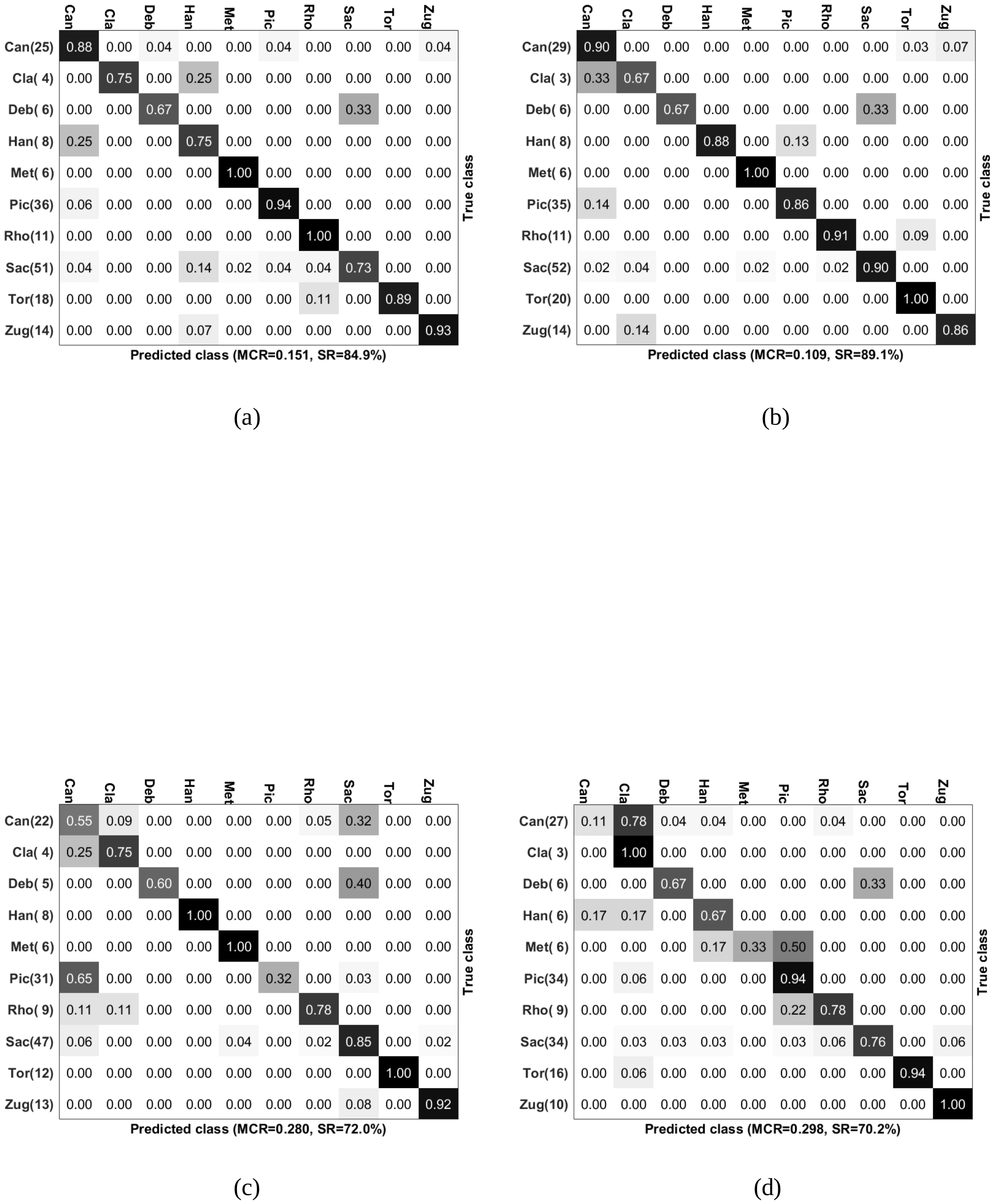
	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(22)	0.59	0.00	0.00	0.00	0.00	0.09	0.05	0.27	0.00	0.00
Cla(4)	0.25	0.50	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00
Deb(5)	0.00	0.00	0.60	0.00	0.00	0.00	0.00	0.40	0.00	0.00
Han(8)	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.50	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(31)	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.03	0.00	0.00
Rho(9)	0.22	0.00	0.00	0.00	0.00	0.00	0.78	0.00	0.00	0.00
Sac(47)	0.06	0.00	0.00	0.00	0.00	0.04	0.02	0.85	0.00	0.02
Tor(12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.92	0.00
Zug(13)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(c)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(27)	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cla(3)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.17	0.00	0.50	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(6)	0.17	0.00	0.00	0.67	0.00	0.17	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.17	0.83	0.00	0.00	0.00	0.00	0.00
Pic(34)	0.03	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00
Rho(9)	0.00	0.00	0.00	0.00	0.00	0.11	0.89	0.00	0.00	0.00
Sac(34)	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.88	0.00	0.09
Tor(16)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.94	0.00
Zug(10)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(d)

Confusion matrices for the validation where run one was kept out. Classification was done on different cultivation media (a) YPD, (b) YMB, (c) SD, (d) YEPD datasets. PLSDA classifier combined with HCA is used for the classification analysis.



True class	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(25)	0.88	0.00	0.04	0.00	0.00	0.04	0.00	0.00	0.00	0.04
Cla(4)	0.00	0.75	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.25	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(36)	0.06	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(51)	0.04	0.00	0.00	0.14	0.02	0.04	0.04	0.73	0.00	0.00
Tor(18)	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.89	0.00
Zug(14)	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.93

Predicted class (MCR=0.151, SR=84.9%)

(a)

True class	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(29)	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.07
Cla(3)	0.33	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	0.88	0.00	0.13	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.14	0.00	0.00	0.00	0.00	0.86	0.00	0.00	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.09	0.00
Sac(52)	0.02	0.04	0.00	0.00	0.02	0.00	0.02	0.90	0.00	0.00
Tor(20)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(14)	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.86

Predicted class (MCR=0.109, SR=89.1%)

(b)

True class	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(22)	0.55	0.09	0.00	0.00	0.00	0.00	0.05	0.32	0.00	0.00
Cla(4)	0.25	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(5)	0.00	0.00	0.60	0.00	0.00	0.00	0.00	0.40	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(31)	0.65	0.00	0.00	0.00	0.00	0.32	0.00	0.03	0.00	0.00
Rho(9)	0.11	0.11	0.00	0.00	0.00	0.00	0.78	0.00	0.00	0.00
Sac(47)	0.06	0.00	0.00	0.00	0.04	0.00	0.02	0.85	0.00	0.02
Tor(12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(13)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.92

Predicted class (MCR=0.280, SR=72.0%)

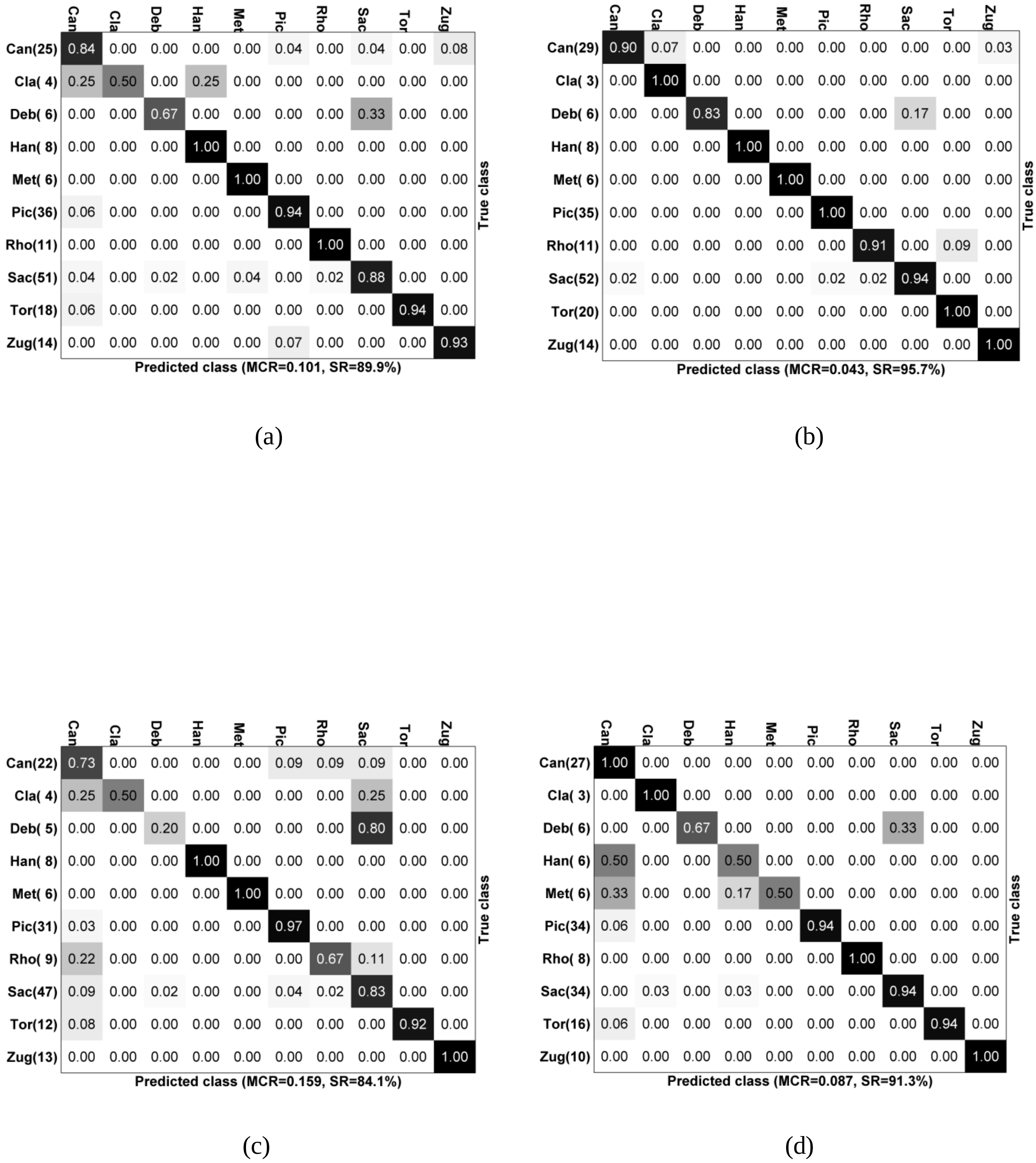
(c)

True class	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(27)	0.11	0.78	0.04	0.04	0.00	0.00	0.04	0.00	0.00	0.00
Cla(3)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(6)	0.17	0.17	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.17	0.33	0.50	0.00	0.00	0.00	0.00
Pic(34)	0.00	0.06	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00
Rho(9)	0.00	0.00	0.00	0.00	0.00	0.22	0.78	0.00	0.00	0.00
Sac(34)	0.00	0.03	0.03	0.03	0.00	0.03	0.06	0.76	0.00	0.06
Tor(16)	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00
Zug(10)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Predicted class (MCR=0.298, SR=70.2%)

(d)

Confusion matrices for the validation where run one was kept out. Classification was done on different cultivation media (a) YPD, (b) YMB, (c) SD, (d) YEPD datasets. FLDA classifier combined with HCA is used for the classification analysis.



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(25)	0.84	0.00	0.00	0.00	0.00	0.04	0.00	0.04	0.00	0.08
Cla(4)	0.25	0.50	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(36)	0.06	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(51)	0.04	0.00	0.02	0.00	0.04	0.00	0.02	0.88	0.00	0.00
Tor(18)	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00
Zug(14)	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.93

(a)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(29)	0.90	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
Cla(3)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.83	0.00	0.00	0.00	0.00	0.17	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.09	0.00
Sac(52)	0.02	0.00	0.00	0.00	0.00	0.02	0.02	0.94	0.00	0.00
Tor(20)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(14)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(b)

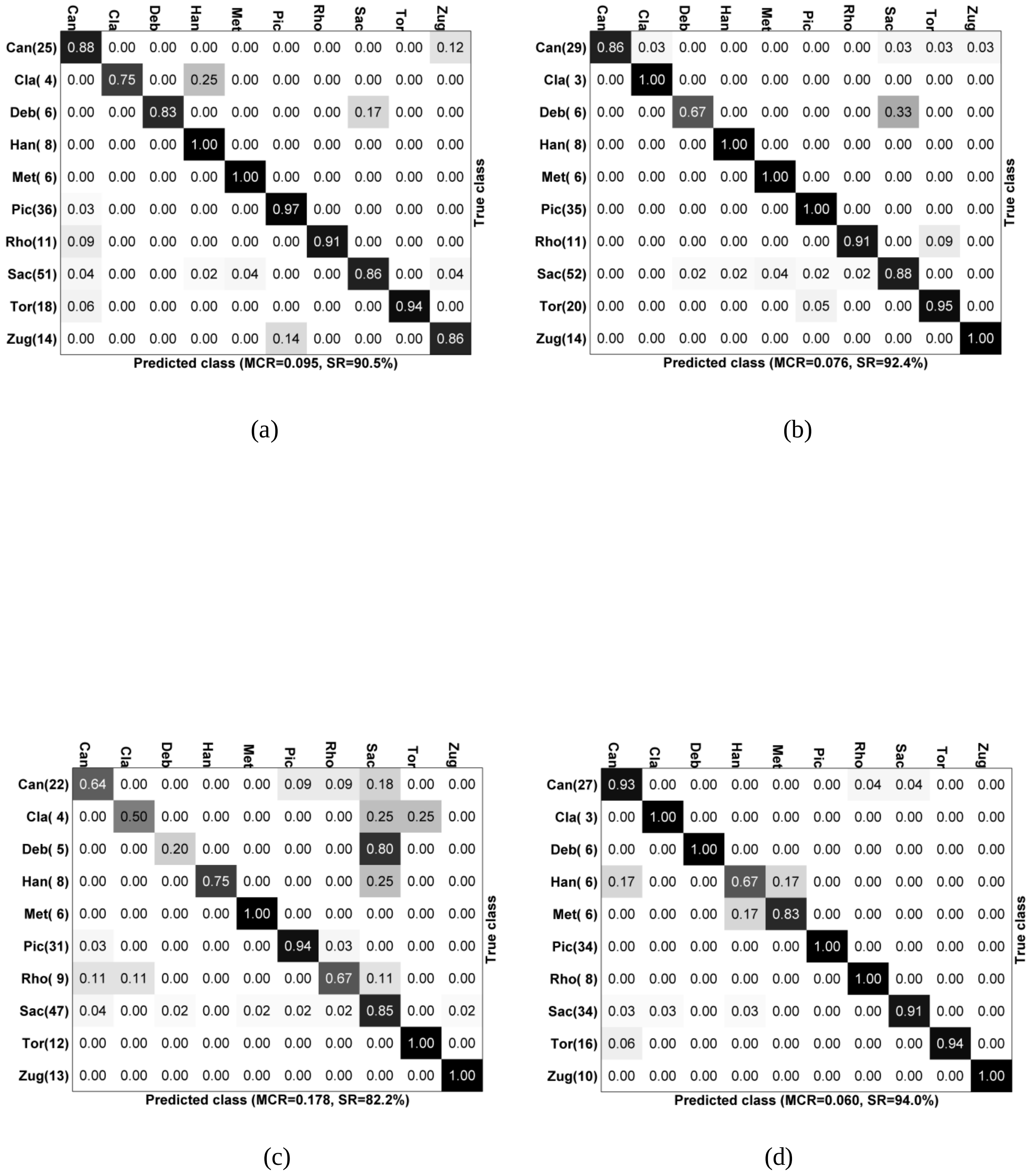
	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(22)	0.73	0.00	0.00	0.00	0.00	0.09	0.09	0.09	0.00	0.00
Cla(4)	0.25	0.50	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00
Deb(5)	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.80	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(31)	0.03	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00
Rho(9)	0.22	0.00	0.00	0.00	0.00	0.00	0.67	0.11	0.00	0.00
Sac(47)	0.09	0.00	0.02	0.00	0.00	0.04	0.02	0.83	0.00	0.00
Tor(12)	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.00
Zug(13)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(c)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(27)	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cla(3)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(6)	0.50	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.33	0.00	0.00	0.17	0.50	0.00	0.00	0.00	0.00	0.00
Pic(34)	0.06	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00
Rho(8)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(34)	0.00	0.03	0.00	0.03	0.00	0.00	0.00	0.94	0.00	0.00
Tor(16)	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00
Zug(10)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(d)

Confusion matrices for the validation where run one was kept out. Classification was done on different cultivation media (a) YPD, (b) YMB, (c) SD, (d) YEPD datasets. PLSDA classifier combined with OVA approach is used for the classification analysis.



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(25)	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12
Cla(4)	0.00	0.75	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.83	0.00	0.00	0.00	0.00	0.17	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(36)	0.03	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00
Rho(11)	0.09	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.00
Sac(51)	0.04	0.00	0.00	0.02	0.04	0.00	0.00	0.86	0.00	0.04
Tor(18)	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00
Zug(14)	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.86

Predicted class (MCR=0.095, SR=90.5%)

True class

(a)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(29)	0.86	0.03	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.03
Cla(3)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.09	0.00
Sac(52)	0.00	0.00	0.02	0.02	0.04	0.02	0.02	0.88	0.00	0.00
Tor(20)	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.95	0.00
Zug(14)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Predicted class (MCR=0.076, SR=92.4%)

True class

(b)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(22)	0.64	0.00	0.00	0.00	0.00	0.09	0.09	0.18	0.00	0.00
Cla(4)	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.25	0.25	0.00
Deb(5)	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.80	0.00	0.00
Han(8)	0.00	0.00	0.00	0.75	0.00	0.00	0.00	0.25	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(31)	0.03	0.00	0.00	0.00	0.00	0.94	0.03	0.00	0.00	0.00
Rho(9)	0.11	0.11	0.00	0.00	0.00	0.00	0.67	0.11	0.00	0.00
Sac(47)	0.04	0.00	0.02	0.00	0.02	0.02	0.02	0.85	0.00	0.02
Tor(12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(13)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Predicted class (MCR=0.178, SR=82.2%)

True class

(c)

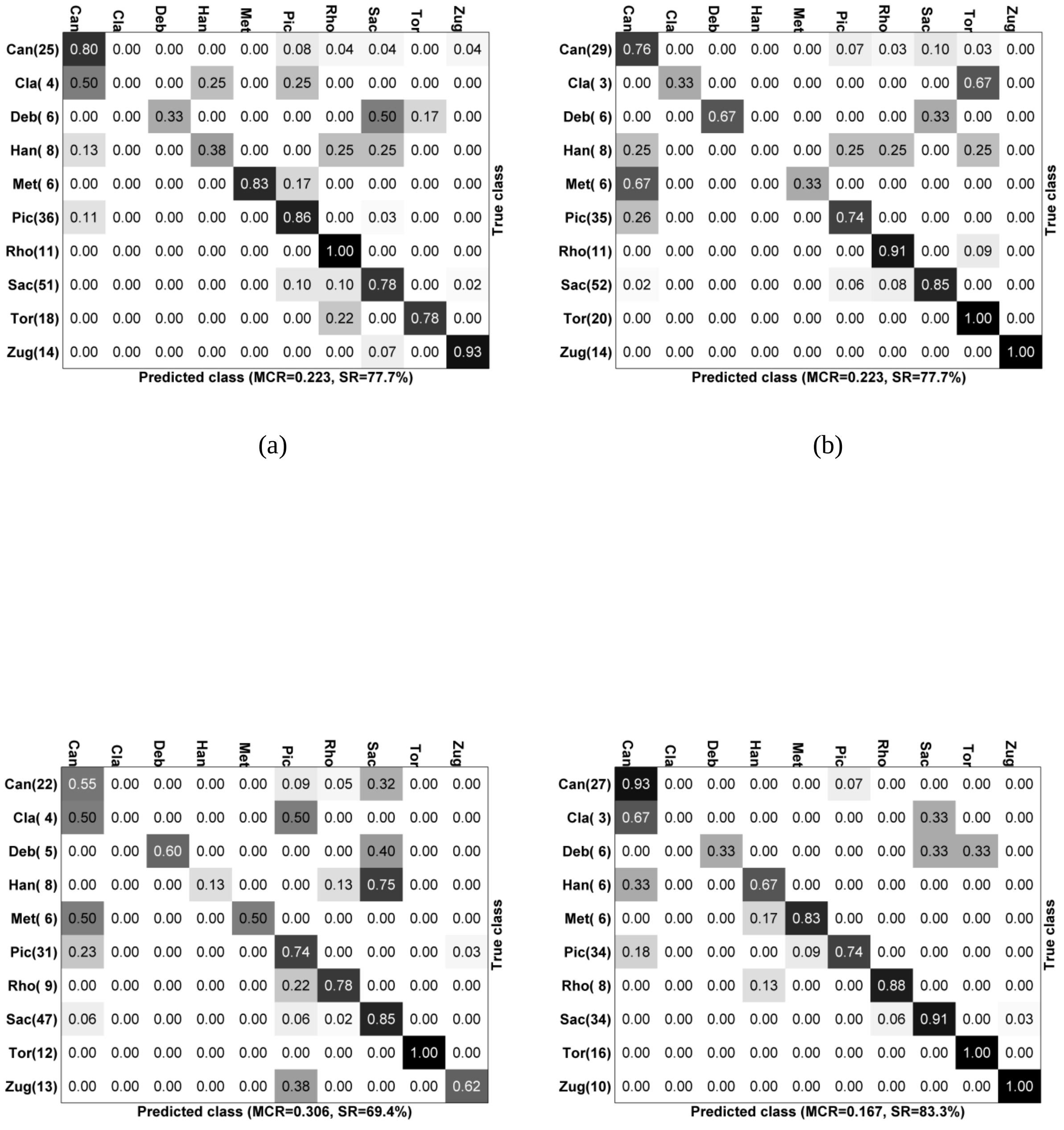
	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(27)	0.93	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.00	0.00
Cla(3)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Han(6)	0.17	0.00	0.00	0.67	0.17	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.17	0.83	0.00	0.00	0.00	0.00	0.00
Pic(34)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(8)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(34)	0.03	0.03	0.00	0.03	0.00	0.00	0.00	0.91	0.00	0.00
Tor(16)	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00
Zug(10)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Predicted class (MCR=0.060, SR=94.0%)

True class

(d)

Confusion matrices for the validation where run one was kept out. Classification was done on different cultivation media (a) YPD, (b) YMB, (c) SD, (d) YEPD datasets. FLDA classifier combined with OVA approach is used for the classification analysis.



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(25)	0.80	0.00	0.00	0.00	0.00	0.08	0.04	0.04	0.00	0.04
Cla(4)	0.50	0.00	0.00	0.25	0.00	0.25	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.50	0.17	0.00
Han(8)	0.13	0.00	0.00	0.38	0.00	0.00	0.25	0.25	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	0.83	0.17	0.00	0.00	0.00	0.00
Pic(36)	0.11	0.00	0.00	0.00	0.00	0.86	0.00	0.03	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(51)	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.78	0.00	0.02
Tor(18)	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.78	0.00
Zug(14)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.93

Predicted class (MCR=0.223, SR=77.7%)

(a)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(29)	0.76	0.00	0.00	0.00	0.00	0.07	0.03	0.10	0.03	0.00
Cla(3)	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.25	0.00	0.00	0.00	0.00	0.25	0.25	0.00	0.25	0.00
Met(6)	0.67	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.26	0.00	0.00	0.00	0.00	0.74	0.00	0.00	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.09	0.00
Sac(52)	0.02	0.00	0.00	0.00	0.00	0.06	0.08	0.85	0.00	0.00
Tor(20)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(14)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Predicted class (MCR=0.223, SR=77.7%)

(b)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(22)	0.55	0.00	0.00	0.00	0.00	0.09	0.05	0.32	0.00	0.00
Cla(4)	0.50	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00
Deb(5)	0.00	0.00	0.60	0.00	0.00	0.00	0.00	0.40	0.00	0.00
Han(8)	0.00	0.00	0.00	0.13	0.00	0.00	0.13	0.75	0.00	0.00
Met(6)	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00
Pic(31)	0.23	0.00	0.00	0.00	0.00	0.74	0.00	0.00	0.00	0.03
Rho(9)	0.00	0.00	0.00	0.00	0.00	0.22	0.78	0.00	0.00	0.00
Sac(47)	0.06	0.00	0.00	0.00	0.00	0.06	0.02	0.85	0.00	0.00
Tor(12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(13)	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.62

Predicted class (MCR=0.306, SR=69.4%)

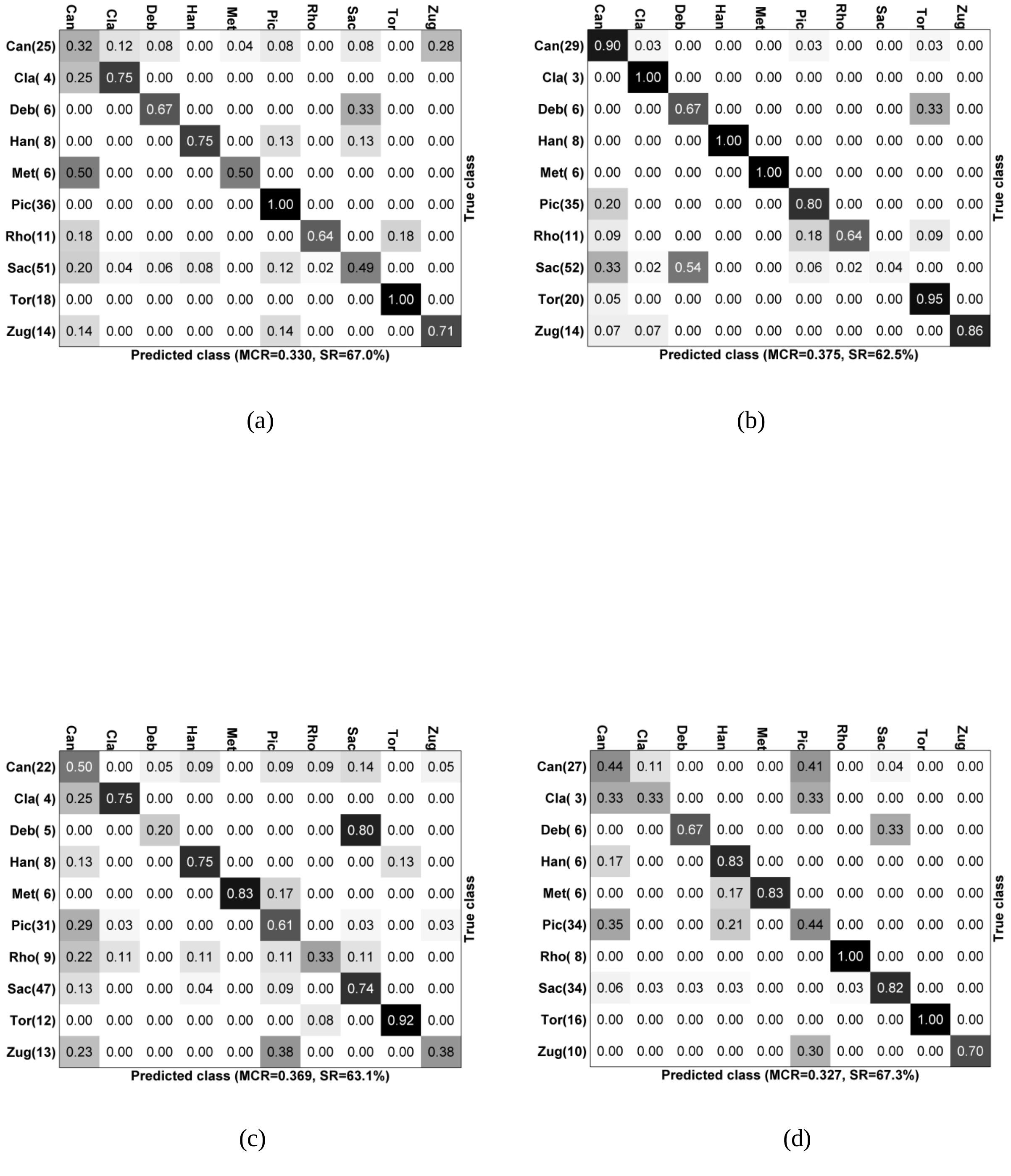
(c)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(27)	0.93	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00
Cla(3)	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Deb(6)	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.33	0.33	0.00
Han(6)	0.33	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.17	0.83	0.00	0.00	0.00	0.00	0.00
Pic(34)	0.18	0.00	0.00	0.00	0.09	0.74	0.00	0.00	0.00	0.00
Rho(8)	0.00	0.00	0.00	0.13	0.00	0.00	0.88	0.00	0.00	0.00
Sac(34)	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.91	0.00	0.03
Tor(16)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(10)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Predicted class (MCR=0.167, SR=83.3%)

(d)

Confusion matrices for the validation where run one was kept out. Classification was done on different cultivation media (a) YPD, (b) YMB, (c) SD, (d) YEPD datasets. PLSDA classifier combined with OVO approach is used for the classification analysis.



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(25)	0.32	0.12	0.08	0.00	0.04	0.08	0.00	0.08	0.00	0.28
Cla(4)	0.25	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	0.75	0.00	0.13	0.00	0.13	0.00	0.00
Met(6)	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00
Pic(36)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(11)	0.18	0.00	0.00	0.00	0.00	0.00	0.64	0.00	0.18	0.00
Sac(51)	0.20	0.04	0.06	0.08	0.00	0.12	0.02	0.49	0.00	0.00
Tor(18)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(14)	0.14	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.71

Predicted class (MCR=0.330, SR=67.0%)

True class

(a)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(29)	0.90	0.03	0.00	0.00	0.00	0.03	0.00	0.00	0.03	0.00
Cla(3)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.33	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.20	0.00	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.00
Rho(11)	0.09	0.00	0.00	0.00	0.00	0.18	0.64	0.00	0.09	0.00
Sac(52)	0.33	0.02	0.54	0.00	0.00	0.06	0.02	0.04	0.00	0.00
Tor(20)	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00
Zug(14)	0.07	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.86

Predicted class (MCR=0.375, SR=62.5%)

True class

(b)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(22)	0.50	0.00	0.05	0.09	0.00	0.09	0.09	0.14	0.00	0.05
Cla(4)	0.25	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(5)	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.80	0.00	0.00
Han(8)	0.13	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.13	0.00
Met(6)	0.00	0.00	0.00	0.00	0.83	0.17	0.00	0.00	0.00	0.00
Pic(31)	0.29	0.03	0.00	0.00	0.00	0.61	0.00	0.03	0.00	0.03
Rho(9)	0.22	0.11	0.00	0.11	0.00	0.11	0.33	0.11	0.00	0.00
Sac(47)	0.13	0.00	0.00	0.04	0.00	0.09	0.00	0.74	0.00	0.00
Tor(12)	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.92	0.00
Zug(13)	0.23	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.38

Predicted class (MCR=0.369, SR=63.1%)

True class

(c)

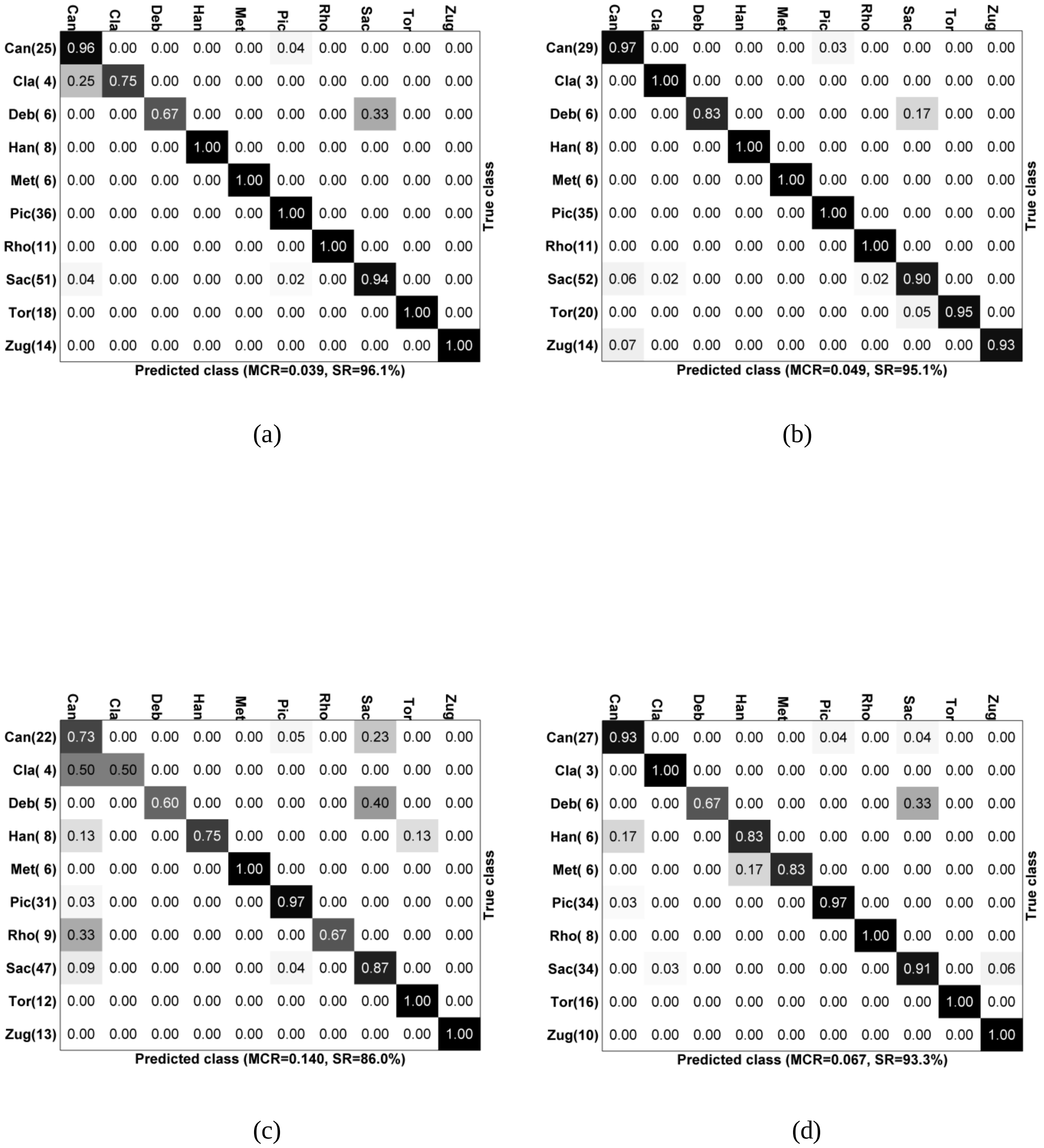
	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(27)	0.44	0.11	0.00	0.00	0.00	0.41	0.00	0.04	0.00	0.00
Cla(3)	0.33	0.33	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(6)	0.17	0.00	0.00	0.83	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.17	0.83	0.00	0.00	0.00	0.00	0.00
Pic(34)	0.35	0.00	0.00	0.21	0.00	0.44	0.00	0.00	0.00	0.00
Rho(8)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(34)	0.06	0.03	0.03	0.03	0.00	0.00	0.03	0.82	0.00	0.00
Tor(16)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(10)	0.00	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.70

Predicted class (MCR=0.327, SR=67.3%)

True class

(d)

Confusion matrices for the validation where run one was kept out. Classification was done on different cultivation media (a) YPD, (b) YMB, (c) SD, (d) YEPD datasets. FLDA classifier combined with OVO approach is used for the classification analysis.



	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(25)	0.96	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00
Cla(4)	0.25	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(36)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(51)	0.04	0.00	0.00	0.00	0.00	0.02	0.00	0.94	0.00	0.00
Tor(18)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(14)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Predicted class (MCR=0.039, SR=96.1%)

(a)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(29)	0.97	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00
Cla(3)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.83	0.00	0.00	0.00	0.00	0.17	0.00	0.00
Han(8)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(35)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Rho(11)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(52)	0.06	0.02	0.00	0.00	0.00	0.00	0.02	0.90	0.00	0.00
Tor(20)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.95	0.00
Zug(14)	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93

Predicted class (MCR=0.049, SR=95.1%)

(b)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(22)	0.73	0.00	0.00	0.00	0.00	0.05	0.00	0.23	0.00	0.00
Cla(4)	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(5)	0.00	0.00	0.60	0.00	0.00	0.00	0.00	0.40	0.00	0.00
Han(8)	0.13	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.13	0.00
Met(6)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Pic(31)	0.03	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00
Rho(9)	0.33	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.00	0.00
Sac(47)	0.09	0.00	0.00	0.00	0.00	0.04	0.00	0.87	0.00	0.00
Tor(12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(13)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Predicted class (MCR=0.140, SR=86.0%)

(c)

	Can	Cla	Deb	Han	Met	Pic	Rho	Sac	Tor	Zug
Can(27)	0.93	0.00	0.00	0.00	0.00	0.04	0.00	0.04	0.00	0.00
Cla(3)	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deb(6)	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Han(6)	0.17	0.00	0.00	0.83	0.00	0.00	0.00	0.00	0.00	0.00
Met(6)	0.00	0.00	0.00	0.17	0.83	0.00	0.00	0.00	0.00	0.00
Pic(34)	0.03	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00
Rho(8)	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sac(34)	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.06
Tor(16)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Zug(10)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Predicted class (MCR=0.067, SR=93.3%)

(d)

Confusion matrices for the validation where run one was kept out. Classification was done on different cultivation media (a) YPD, (b) YMB, (c) SD, (d) YEPD datasets. RF classifier is used for the classification analysis.

Norges miljø- og biovitenskapelig universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway