Available online at www.sciencedirect.com
ScienceDirect
Systematic and Applied Microbiology 32 (2009) 163–176
SYSTEMATIC AND APPLIED MICROBIOLOGY
www.elsevier.de/syapm

Towards large-scale FAME-based bacterial species identification using machine learning techniques

Bram Slabbinck[a,*], Bernard De Baets[a], Peter Dawyndt[b], Paul De Vos[c]

[a]*Research Unit Knowledge-based Systems, Faculty of Bioscience Engineering, Ghent University, Coupure links 653, 9000 Ghent, Belgium*
[b]*Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, 9000 Ghent, Belgium*
[c]*Laboratory of Microbiology, BCCM^{TM}/LMG Bacteria Collection, Ghent University, K.L. Ledeganckstraat 35, 9000 Ghent, Belgium*

Received 13 August 2008

Abstract

In the last decade, bacterial taxonomy witnessed a huge expansion. The swift pace of bacterial species (re-)definitions has a serious impact on the accuracy and completeness of first-line identification methods. Consequently, back-end identification libraries need to be synchronized with the List of Prokaryotic names with Standing in Nomenclature. In this study, we focus on bacterial fatty acid methyl ester (FAME) profiling as a broadly used first-line identification method. From the BAME@LMG database, we have selected FAME profiles of individual strains belonging to the genera *Bacillus*, *Paenibacillus* and *Pseudomonas*. Only those profiles resulting from standard growth conditions have been retained. The corresponding data set covers 74, 44 and 95 validly published bacterial species, respectively, represented by 961, 378 and 1673 standard FAME profiles. Through the application of machine learning techniques in a supervised strategy, different computational models have been built for genus and species identification. Three techniques have been considered: artificial neural networks, random forests and support vector machines. Nearly perfect identification has been achieved at genus level. Notwithstanding the known limited discriminative power of FAME analysis for species identification, the computational models have resulted in good species identification results for the three genera. For *Bacillus*, *Paenibacillus* and *Pseudomonas*, random forests have resulted in sensitivity values, respectively, 0.847, 0.901 and 0.708. The random forests models outperform those of the other machine learning techniques. Moreover, our machine learning approach also outperformed the Sherlock MIS (MIDI Inc., Newark, DE, USA). These results show that machine learning proves very useful for FAME-based bacterial species identification. Besides good bacterial identification at species level, speed and ease of taxonomic synchronization are major advantages of this computational species identification strategy.

Keywords: *Bacillus*; Bacteria; Fatty acid methyl ester; Gas chromatography; Identification; Machine learning; *Paenibacillus*; *Pseudomonas*; Species; Taxonomy

Abbreviations: ANN, artificial neural networks; AUC, area under the ROC curve; BAME, bacterial FAME; FAME, fatty acid methyl ester; FN, false negative; FP, false positive; FPR, false positive rate; Pr, precision; RF, random forests; ROC, receiver operating characteristic; Se, sensitivity; SVM, support vector machines; TN, true negative; TP, true positive; TPR, true positive rate; TSBA, trypticase soy broth agar.

*Corresponding author.
E-mail addresses: Bram.Slabbinck@UGent.be (B. Slabbinck), Bernard.DeBaets@UGent.be (B. De Baets), Peter.Dawyndt@UGent.be (P. Dawyndt), Paul.DeVos@UGent.be (P. De Vos).

0723-2020/$ - see front matter
doi:10.1016/j.syapm.2009.01.003

Introduction

Our knowledge about the bacterial landscape is continuously evolving. This is clearly demonstrated by the growing list of validly published species in the International Journal of Systematic and Evolutionary Microbiology. From January 2005 to March 2008, not less than 1888 bacterial species have been (re)defined [13]. Given this rapid change in taxonomy, back-end identification libraries of first-line identification methods need constant updates. Because gas chromatographic whole-cell fatty acid methyl ester (FAME) analysis is cheap, easy and automated, many laboratories routinely use this chemotaxonomic technique for the identification of bacterial species. FAME analysis usually relies, however, on commercial identification systems such as the Sherlock Microbial Identification System (MIS, Microbial ID Inc. (MIDI), Newark, DE, USA) for which the back-end identification libraries are only updated every few years and only cover part of all known species. The accuracy of bacterial species identification is therefore highly compromised, making this update latency a major drawback of the Sherlock MIS system. In this paper, we explore the realization of an up-to-date FAME-based bacterial species identification system powered by machine learning techniques.

Qualitatively, the fatty acid composition of bacterial species is highly conserved, and significant changes only take place over considerable periods of time. Quantitatively, the measured fatty acid composition of a particular strain is stable, given highly standardized culture conditions. More than 300 fatty acids have already been found in bacteria. Differences in chain length, positions of double bonds and the binding of functional groups make them very useful taxonomic markers [10,27]. The first evidence suggesting that fatty acids could be used for the identification of bacteria was reported by Abel et al. [1] and Kaneda [25]. The use of whole-cell FAME analysis for bacterial identification has already been applied successfully for a wide range of bacterial taxa [8,12,21,23,24,26,36–38,41,48–50,53]. Furthermore, FAME-based identification of bacteria by machine learning techniques has already been investigated for a multitude of bacterial species and genera [4,17,18,45]. Nonetheless, from a taxonomic perspective, the development of a bacterial species identification tool requires the inclusion of a maximal number of species for each genus considered and a sufficient number of strains per species. A first attempt towards FAME-based bacterial species identification on a genus-wide scale has been undertaken on the genus *Bacillus* by artificial neural networks [46]. In the present study, we have extended this approach by considering the genera *Bacillus*, *Paenibacillus* and *Pseudomonas* and by evaluating some other machine learning techniques as well.

According to the List of Prokaryotic names with Standing in Nomenclature as published in March 2008, the genera *Bacillus*, *Paenibacillus* and *Pseudomonas* consist, respectively, of 145, 86 and 117 validly published species (excluding synonyms and species reassigned to other genera) [13]. The genus *Bacillus* comprises a heterogeneous group of aerobic, endospore-forming, Gram-positive, rod-shaped organisms. *Bacillus* species can be found in various environments and many of these species are of industrial, clinical and commercial interest due to the production of spores and a variety of interesting components [3,29,30]. It is well known that due to the limited discriminative power of FAME analysis, it is not possible to distinguish all species of the genus *Bacillus* [24]. In 1994, a reclassification within the genus *Bacillus sensu lato* based on 16S rRNA sequence analysis has led to the proposal of the genus *Paenibacillus*. The group of strains emerging from *Bacillus sensu lato* was originally referred to as *Bacillus* RNA group 3. The genus *Paenibacillus* comprises a heterogeneous group of facultative anaerobe or strictly aerobe, endospore-forming, Gram-positive, rod-shaped organisms, of which most species have peritrichous flagella. Some strains are known to be important pathogens of insects [2,42]. FAME analysis of 11 *Paenibacillus* species revealed different species groups [20]. The genus *Pseudomonas* comprises a heterogeneous group of aerobic, non-spore forming, Gram-negative, rod-shaped organisms with polar flagella. The genus *Pseudomonas* is well known from the opportunistic and clinically important *Pseudomonas aeruginosa* and various plant pathogenic pathovars such as *Pseudomonas syringae* [30,40]. The taxonomy of the so-called *P. syringae* pathovars that are of practical importance has been under discussion for several decades. Regarding this species and its various pathovars, a DNA–DNA hybridization study of the different pathovars proposed nine genomospecies of which some are linked to other validly published species [16]. Evaluation of the discriminative power of FAME analysis for the identification of pseudomonads also made clear that FAME data do not allow to distinguish all *Pseudomonas* species from each other [47,51].

Even though microbial taxonomy is changing rapidly, the development of up-to-date identification libraries can be realized by computer systems and back-end databases. Based on a laboratory information management system and a FAME database, we have analyzed the identification at genus and species level of members of three genera by three machine learning techniques: artificial neural networks (ANN), support vector machines (SVM) and random forests (RF). This new approach shows a moderate to high identification performance and outperforms the commercial Sherlock MIS system.

Materials and methods

Bacterial strains and FAME analysis

To assure reproducible and interpretable results, all strains of the genera *Bacillus*, *Paenibacillus* and *Pseudomonas* were grown under standard conditions unless are as stated in Table S1. The protocol designed by MIDI Inc. (Newark, DE, USA) for the construction of the Sherlock MIS TSBA identification library has been followed. For aerobes, such as *Bacillus*, *Paenibacillus* and *Pseudomonas*, this protocol recommends 24 h of growth on trypticase soy broth agar (TSBA) at a temperature of 28 °C. Following growth, on average 40 mg bacterial cells are harvested in the overlapping region of quadrants three and four of the streaked plates. Next, FAMEs are extracted by a four-step procedure of saponification, methylation, extraction and sample cleanup, and are finally analyzed by gas chromatography. Whole-cell bacterial FAME profiles resulting from gas chromatographic analysis following these standard growth conditions are further indicated as standard FAME profiles.

Collaborative FAME analysis research at the Laboratory of Microbiology (Ghent University, Belgium) as well as the BCCM^{TM}/LMG Bacteria Collection (Ghent University, Belgium) has resulted in the BAME@LMG database currently containing more than 67,000 bacterial FAME profiles. From this database, only those validly published *Bacillus*, *Paenibacillus* and *Pseudomonas* species represented by at least three standard FAME profiles were selected. With respect to the *Pseudomonas syringae* pathovars, the genomospecies taxonomy as suggested by Gardan et al. [16] was followed. In March 2008, the BAME@LMG database contained 961, 378 and 1673 standard FAME profiles of 74 *Bacillus* (51%), 44 *Paenibacillus* (51%) and 95 *Pseudomonas* species (81%), respectively. An overview of the selected strains and FAME profiles is given in Table S1. From this database, separate *Bacillus*, *Paenibacillus* and *Pseudomonas* data sets were created. These three data sets were also merged into a single complete data set. Two versions of the complete data set were created by annotating the FAME profiles by their genus and species name, or by their genus name only. This resulted in a data set with 213 classes and a data set with 3 classes, respectively. The separate *Bacillus*, *Paenibacillus* and *Pseudomonas* data sets contained, respectively, 71, 46 and 94 FAMEs; whereas the combined data set contained 105 FAMEs. Each FAME profile is represented as a vector of the different relative FAME peak area percentages as calculated by the Sherlock MIS using the TSBA50 peak naming method.

Sherlock MIS

Routine identification of bacterial species based on FAME analysis is traditionally performed by commercial systems such as Sherlock MIS (MIDI Inc., USA). According to the taxonomy as published in March 2008, the Sherlock MIS TSBA50 identification library contains, respectively, 30 (21%), 18 (21%) and 31 (26%) validly published *Bacillus*, *Paenibacillus* and *Pseudomonas* species. In order to make a reliable comparison of our new identification approach with Sherlock MIS, of all species present in the data set, the identification results of only those species present in both the Sherlock MIS TSBA50 library and the selected BAME@LMG data sets were analyzed. For each identification strategy, identification is evaluated by considering the genus and species name associated with the highest identification output value. For Sherlock MIS, this is the species with the highest similarity index (SI) value. Subsequently, the number of correct identifications is averaged for each species, possibly represented by a different number of FAME profiles. However, in our approach, only FAME profiles correctly identified at genus level are considered for subsequent species identification. Finally, for each genus, a global average is calculated over all species.

Experimental design

Two identification strategies have been evaluated. These strategies are schematically represented in Fig. 1. In the stratified identification strategy, genus and species identification is performed by separate identification models. Identification is performed first at genus level, followed by identification at species level. For genus identification, the FAME profiles are only annotated by genus name. At species level, a species identification model is generated based on a genus-specific data set. This data set comprises only FAME profiles of the species of a particular genus, which are annotated by genus and species name. The second approach is the straight identification strategy. One single species identification model is generated based on the complete data set, in which the FAME profiles are annotated both by genus and species names.

For comparison with Sherlock MIS, the stratified identification strategy has been considered. For each species identification model, the training and test set combination resulting in the highest AUC value (see also below) is considered. Ultimately, identification at genus level is achieved by merging the respective training and test sets and by training and testing a genus identification model based on the merged data sets. As such, it is possible to identify the same profiles for genus and species identification and to rule out those profiles incorrectly identified at genus level. Importantly, the same machine learning technique is considered for all species identification models and the genus identification model. Even though it is possible that different machine learning techniques result in better performance on

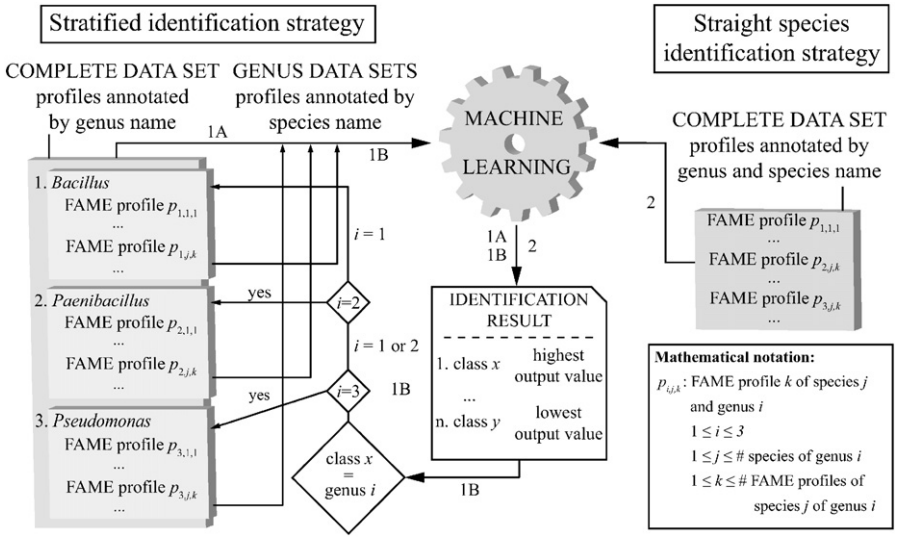

Fig. 1. Schematic presentation of the experimental design. 1A/B. *Stratified identification strategy*. 1A. Genus identification is performed by the genus identification model. This model relies on the complete FAME data set in which the profiles are annotated by genus name (dark grey box). 1B. For each genus, a species identification model is built based on the FAME profiles corresponding to that specific genus. The respective FAME profiles are annotated by species name (light grey boxes). In both cases, each profile is labelled with the genus or species name associated with the highest output value. However, species identification is only performed for the genus associated with the highest output value following genus identification. 2. *Straight species identification strategy*. The complete data set of FAME profiles is annotated by genus and species name (dark grey box). Identification is performed by a single identification model. Each profile is labelled with the genus and species name associated with the highest output value.

different data sets, we have chosen to focus on the same technique for ease of implementation.

Machine learning

Terminology

Classification and identification are terms used with different meanings in the fields of microbiology and machine learning. Whereas identification is similarly interpreted as assigning existing class labels to unknown bacterial organisms or data points, classification should be interpreted differently. In a taxonomic context, bacterial classification refers to the grouping of bacterial organisms based on genotypic and phenotypic similarities [30]. However, in machine learning, and more specifically in supervised classification, the goal of classification is to statistically describe the relationship between data points of various classes, given the class labels, by generalizing the observed trends on the data [31]. In the context of this paper, we refer to classification as the process of building computational models to distinguish between the FAME profiles of the different bacterial genera and species.

Artificial neural networks

ANNs are mathematical models inspired by biological neural systems, being constructed as complex networks of interconnected neurons. The architecture of an ANN is composed of information processing units, neurons, which are interconnected by weighted links. This architecture is typically visualized in layers. In this study, a feed-forward ANN is used with three neuron layers: one input layer, one hidden layer and one output neuron layer [31]. A feed-forward architecture implies a signal transfer from the input neurons towards the output neurons without feedback. Both the hidden neurons and the output neurons calculate a weighted sum of their input signals, known as the neuron activation. An activation function maps each activation value into a specific interval. The functions used are the sigmoid (s) and the bipolar sigmoid (b) functions, which map values into the intervals [0,1] and [−1,1], respectively. The best combination of functions used on the hidden and output neurons must be empirically determined [31,46]. Each combination of an activation function f_1 for the hidden neurons and an activation function for the output neurons f_2 is further denoted as 'ANN f_1/f_2'. The number of input and the number of output neurons correspond to the number of FAMEs and the number of bacterial species present in the data set. The input neuron values correspond to the relative FAME peak area percentages, while the output neuron values correspond to a score for each bacterial species or genus. The hidden neurons govern the power and the complexity of the ANN as these neurons determine the number of connections and weights in the network. The optimal number of hidden neurons is set to the number

resulting in the minimum mean squared validation error over a range of different numbers of hidden neurons [11].

As each FAME profile is labelled with a genus name or a genus and species name, an error function is minimized by calculating the difference or error between the ANN output values and the target values [5]. Minimization is achieved using the backpropagation method, which involves three stages: feed-forward of the training data, calculation and backpropagation of the error, and adjustment of the weights. Optimal classification requires different phases or epochs. Optimization of the backpropagation method is done by the resilient propagation algorithm. To ensure generalization of the data and to prevent under- and overfitting, the early stopping method halts ANN training at the epoch corresponding to the minimum validation error [5,11,14,31,43]. Identification of FAME profiles using a trained ANN is done by labelling the profile with the genus and/or species name corresponding to the output neuron with the highest output score [11,14].

The ANNs used in this study are implemented by the Neural Network Toolbox 4.0.3 of the MATLAB R14 software package.

Random forests

RF is an ensemble method based on bagging. Ensemble methods generate multiple classifiers and aggregate the results. Specifically, RF is an ensemble of classification trees. At each node of a classification tree, the best split is chosen among a subset of features randomly chosen at that node. In contrast, standard classification trees split each node based on all features available. RF performs very well compared to many other classifiers and is robust against overfitting [7]. To achieve optimal classification of the data, two parameters need to be optimized: the number of randomly chosen features at each node (*Nf*) and the number of trees to be grown in the forest (*Nt*). Optimization of *Nt* is done by varying the number of trees from 1000 to 4000 by steps of 250 and by setting *Nf* on its default value, which equals the root of the number of FAMEs. The number leading to the minimum test error becomes the final number of trees. Optimization of *Nf* is done by setting *Nf* equal to its default value, twice the default value and half the default value. The number resulting in the minimum test error is selected [7,28]. The first step in the RF algorithm is the generation of *Nt* bootstrap samples from the original data set. These samples contain about two-thirds of the FAME profiles of the original data set. The remaining profiles are used as test set. For each bootstrap sample an unpruned tree is grown by the method previously described. Next, for each sample, the corresponding test profiles are predicted by the corresponding tree. At the end of the run, and over all samples, take the class which got most of the votes every time a given profile was present in a test set. The proportion of misclassifications averaged over all test profiles is taken as the overall error rate. New data can be predicted by aggregating the predictions of the *Nt* trees and selecting the label with the highest number of votes [7].

The RFs used in this study are generated by the RandomForest software package [7].

Support vector machines

SVMs map the FAME profiles, referred to as feature vectors in this context, to a higher dimensional feature space in which the mapped data is maximally separated by a hyperplane. On each side of this hyperplane two maximally separated, parallel hyperplanes are constructed. A larger distance or margin between both hyperplanes leads to a better generalization and classification of the data. The main principle of SVM theory corresponds to finding the feature vectors on the maximum margin hyperplanes. These points are called the support vectors, as once the SVM is trained, a significant proportion of the vectors can be discarded by retaining only the support vectors for identification. Although SVMs try to linearly separate the data in feature space, class distributions mostly overlap. This problem is resolved by allowing misclassifications and minimizing the corresponding penalties while maximizing the margin [6]. A more general solution is obtained by introducing a kernel function in the training algorithm. By using this kernel function, it is possible to apply non-linear functions to separate the different classes. In this study, the radial basis function kernel ('SVM RBF') and the linear kernel ('SVM lin') are used.

For the identification of bacteria, we are confronted with multiple species or classes. This multi-class classification problem is solved by considering a one-versus-others setting implying the construction of n two-class SVMs, with n the number of species or genus classes considered. Consequently, each SVM separates one class from all the others. Optimization of the SVM parameters is achieved by grid search. FAME profiles are assigned the label corresponding to the class with the highest output value [6].

The SVMs in this study are implemented by the libSVM software package [9].

Validation

The initial data set is randomly split into a training set and a test set. The test set contains about one-third of the FAME profiles of each bacterial species and is used to test the classifiers on their ability to correctly identify the FAME profiles. This process is repeated ten times. In the ANN and SVM implementation, stratified cross-validation is performed to prevent overfitting and to ensure generalization during the training process.

Cross-validation is not performed with RF as this technique is robust against overfitting.

Statistical analysis

Statistical analysis is performed on the identification output scores of the respective machine learning technique. An output score corresponds to the probability of belonging to a certain class, which, in this study, represents either a species or a genus.

The choice of metrics for statistical analysis of the identification results is very important as these metrics should be insensitive to changes in the number of FAME profiles per species or per genus, and in the number of species per genus. In this study, three statistical metrics are used: receiver operating characteristic (ROC) curves, the mean sensitivity (Se) and the mean precision (Pr).

Receiver operating characteristic curves are highly appropriate due to their insensitivity to changes in class distribution [15,52]. ROC curves are built from a confusion matrix that results from a two-class classification. As we try to distinguish each species from all other species, the associated matrices are generated following the one-versus-others method. This implies that for each class or species, the corresponding profiles are labelled positive and the profiles of the other classes negative [44]. Next, for each one-versus-others classification and a certain threshold value, a confusion matrix is built by considering the output values of all test profiles. As such, a true positive (TP) corresponds to a positive profile with a value above the threshold. Accordingly, a value under the threshold leads to a false negative (FN). Dividing the total number of TP by the sum of the total number of TP and FN results in the true positive rate (TPR). A negative profile with an output value above the threshold is a false positive (FP). In the other case, the profiles are true negatives (TN). Analogous to the TPR, the corresponding false positive rate is calculated (FPR). Plotting the TPR of the classifier on the Y-axis and the FPR on the X-axis of a two-dimensional graph results in a point showing how well the model classified the data for a certain choice of threshold value. Instead of visualizing the performance of a classifier by a single point, a ROC curve is created by varying the threshold between the minimum and the maximum output values. In this study, this corresponds to the values 0 and -1, and 1. The variation step is set by ranking all FAME profiles in the test set based on their output values. Hereby, the output value of each FAME profile corresponds to a threshold value and, thus, to an additional point in the ROC graph. The diagonal line $y = x$ represents the strategy of randomly guessing classes. To compare different classifiers, in practice it is common practice to use the area under the ROC curve (AUC). Statistically, the AUC represents the probability that a classifier will assign a randomly chosen positive instance a higher score than a randomly chosen negative instance [15]. Classification of n classes in a one-versus-others setting implies the calculation of n AUC values. As an overall performance measure, we calculate the mean of the different AUC values, analogously to Hand and Till [19], who calculated an overall AUC value as the mean AUC for each pair of classes.

Similar to the calculation of the overall AUC, the overall Se and Pr are calculated. By calculating for each species the Se and Pr and averaging it over all species, a performance metric is obtained showing the Se and Pr variation over all species. Instead of using a threshold value to calculate the number of TP, the winner-take-all rule assigns the species or genus label with the highest output value to each FAME profile. Similarly as described above, the TP, FP, FN and TN values are calculated. Se corresponds to the percentage of positive data points that are predicted correctly, while Pr corresponds to the percentage of positive results that are true positives.

As each technique is performed ten-fold, this finally leads to a mean overall AUC and a mean overall Se and Pr. These metrics will be further denoted as AUC, Se and Pr.

Results

Stratified identification strategy

In the stratified identification setting, genus identification is performed preliminary to species identification. A detailed report of the accuracy values is given in Table 1.

Fig. 2 shows the accuracy of the *Bacillus*, *Paenibacillus* and *Pseudomonas* species and genus classification. As could be expected, the three machine learning techniques have resulted in a very high FAME-based genus classification performance.

Figs. 3 and 4 show the identification performance of each machine learning technique. At genus level, among all experiments, the highest Se value of 0.994 ± 0.005 and Pr value of 0.993 ± 0.008 is attained with the SVM RBF model. The multi-class confusion matrix of this experiment is shown in Table 2. At species level and for each genus considered, the identification accuracies are different for the three machine learning techniques. Among all *Bacillus* species identification experiments, RF has resulted in the highest Se value of 0.885 ± 0.216 and Pr value of 0.926 ± 0.143. However, these values are not resulting from the same experiment. For *Paenibacillus* species identification, among all experiments, the highest Se value and Pr value are achieved by RF: 0.974 ± 0.092 and 0.981 ± 0.081, respectively. Among all *Pseudomonas* species identification experiments, RF has led to the highest Se value of 0.689 ± 0.350 and Pr value

Table 1. Overview of the results for genus and species identification.

Metric	Identification strategy	ANN, *s*/*s*	ANN, *b*/*s*	ANN, *b*/*b*	ANN, *s*/*b*	RF	SVM, RBF	SVM, lin
AUC	Stratified ID							
	Genus	0.992	0.992	0.991	0.993	**0.998**	0.997	0.996
		(0.004)	(0.006)	(0.006)	(0.003)	**(0.001)**	(0.002)	(0.002)
	Bacillus	0.966	0.972	0.966	0.964	**0.988**	0.981	0.977
		(0.011)	(0.008)	(0.008)	(0.011)	**(0.007)**	(0.005)	(0.008)
	Paenibacillus	0.970	0.971	0.971	0.965	**0.990**	0.976	0.983
		(0.008)	(0.014)	(0.007)	(0.013)	**(0.015)**	(0.013)	(0.005)
	Pseudomonas	0.944	0.951	0.950	0.937	**0.987**	0.979	0.979
		(0.008)	(0.005)	(0.006)	(0.014)	**(0.003)**	(0.003)	(0.003)
	Straight ID	0.971	0.973	0.974	0.964	**0.991**	0.988	0.988
		(0.003)	(0.004)	(0.004)	(0.005)	**(0.002)**	(0.002)	(0.002)
Se	Stratified ID							
	Genus	0.979	0.978	0.979	0.978	0.977	**0.985**	0.979
		(0.006)	(0.007)	(0.009)	(0.005)	(0.006)	**(0.005)**	(0.006)
	Bacillus	0.753	0.740	0.731	0.748	**0.847**	0.544	0.457
		(0.028)	(0.024)	(0.024)	(0.036)	**(0.021)**	(0.053)	(0.036)
	Paenibacillus	0.753	0.734	0.724	0.749	**0.901**	0.610	0.551
		(0.047)	(0.037)	(0.045)	(0.046)	**(0.040)**	(0.068)	(0.044)
	Pseudomonas	0.551	0.537	0.501	0.523	**0.673**	0.281	0.272
		(0.039)	(0.037)	(0.035)	(0.047)	**(0.014)**	(0.028)	(0.026)
	Straight ID	0.669	0.669	0.633	0.634	**0.732**	0.239	0.232
		(0.021)	(0.016)	(0.032)	(0.021)	**(0.015)**	(0.011)	(0.01)
Pr	Stratified ID							
	Genus	0.9835	0.981	0.984	0.984	0.982	**0.989**	0.983
		(0.004)	(0.007)	(0.004)	(0.003)	(0.004)	**(0.003)**	(0.006)
	Bacillus	0.812	0.798	0.803	0.748	**0.908**	0.829	0.751
		(0.03)	(0.026)	(0.031)	(0.036)	**(0.013)**	(0.043)	(0.055)
	Paenibacillus	0.815	0.796	0.775	0.803	**0.947**	0.775	0.800
		(0.049)	(0.036)	(0.043)	(0.062)	**(0.018)**	(0.264)	(0.247)
	Pseudomonas	0.671	0.669	0.645	0.643	**0.851**	0.708	0.688
		(0.023)	(0.041)	(0.031)	(0.038)	**(0.023)**	(0.035)	(0.021)
	Straight ID	0.745	0.757	0.728	0.718	**0.882**	0.661	0.634
		(0.026)	(0.022)	(0.031)	(0.021)	**(0.009)**	(0.019)	(0.031)

Classification performance is indicated by the area under the ROC curve (AUC). Identification performance is indicated by sensitivity (Se) and precision (Pr). Results are reported for two identification strategies: stratified and straight identification (ID). The stratified identification strategy performs identification at genus level and at species level for the three genera: *Bacillus*, *Paenibacillus* and *Pseudomonas*. Three machine learning techniques are used: artificial neural networks (ANN) with a sigmoid (*s*) and/or bipolar sigmoid (*b*) activation function on the hidden and output neurons (f_1/f_2), random forests (RF) and support vector machines (SVM) with RBF and linear (lin) kernel. Performance values and standard deviations are reported. Highest performance values are indicated in bold face.

of 0.887 ± 0.201. These values are not resulting from the same experiment. Overall, the three machine learning techniques have resulted in a very high FAME-based genus identification performance. In the case of species identification, RF and ANN have outperformed SVM with a clear and distinct advantage for RF. From these statistics, it can be concluded that it is possible to achieve a moderate to good accuracy for FAME-based *Bacillus*, *Paenibacillus* and *Pseudomonas* species identification by machine learning.

Straight species identification strategy

Table 1 and Figs. 2–4 also show classification and identification accuracies of ANN, RF and SVM for the data set covering the species of all three genera. This strategy concerns a straight species identification in which genus identification is not considered. Generally, a good FAME-based species classification performance has been achieved by the three techniques. Among all species identification experiments, the highest Se value

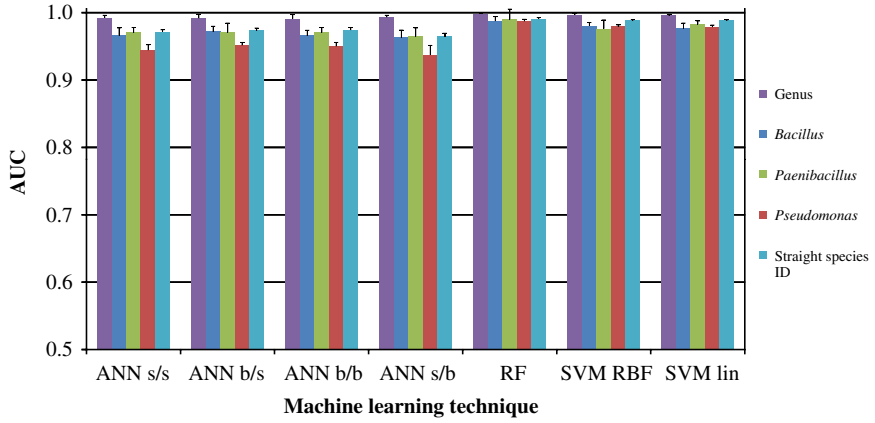

Fig. 2. Overview of the area under the ROC curve (AUC) values resulting from species classification of the genera *Bacillus*, *Paenibacillus* and *Pseudomonas*, genus classification and overall species classification. The different machine learning techniques used for classification are artificial neural networks (ANN) with a sigmoid (*s*) and/or bipolar sigmoid (*b*) activation function on the hidden and output neurons, random forests (RF) and support vector machines (SVM) with a RBF and linear kernel.

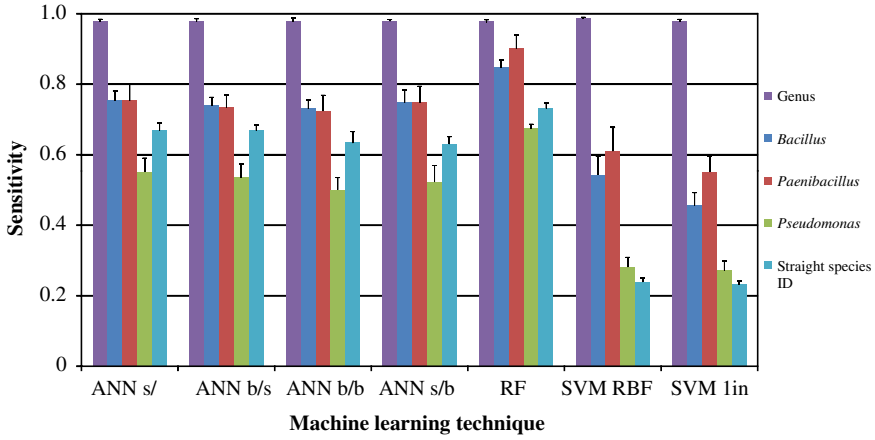

Fig. 3. Overview of the sensitivity values resulting from species identification of the genera *Bacillus*, *Paenibacillus* and *Pseudomonas*, genus identification and overall species identification. The different machine learning techniques used for identification are artificial neural networks (ANN) with a sigmoid (*s*) and/or bipolar sigmoid (*b*) activation function on the hidden and output neurons, random forests (RF) and support vector machines (SVM) with a RBF and linear kernel.

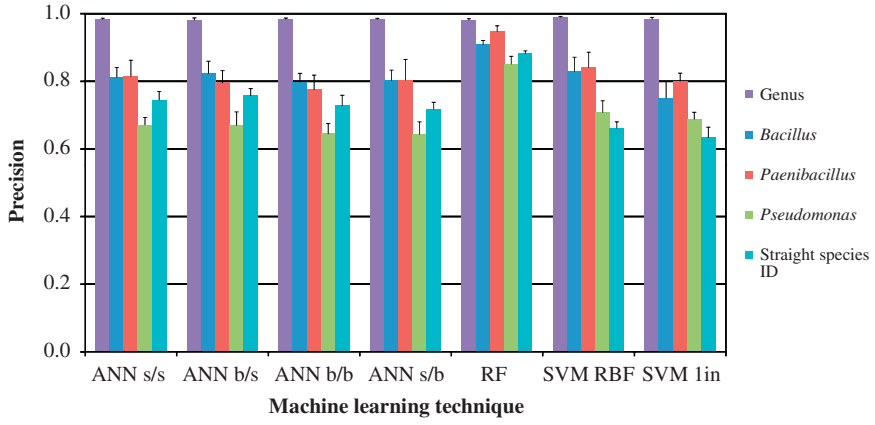

Fig. 4. Overview of the precision values resulting from species identification of the genera *Bacillus*, *Paenibacillus* and *Pseudomonas*, genus identification and overall species identification. The different machine learning techniques used for identification are artificial neural networks (ANN) with a sigmoid (*s*) and/or bipolar sigmoid (*b*) activation function on the hidden and output neurons, random forests (RF) and support vector machines (SVM) with a RBF and linear kernel.

and Pr value have been 0.764 ± 0.332 and 0.898 ± 0.186, respectively. Generally, RF and ANN have outperformed SVM. From these statistics, it can be concluded that FAME-based species identification RF results in a moderate accuracy.

Comparison with Sherlock

Fig. 5 shows a comparison between the *Bacillus*, *Paenibacillus* and *Pseudomonas* species identification accuracies obtained with the RF technique in the stratified strategy setup and with the TSBA50 identification library of Sherlock MIS. For each genus, species identification is done by selecting the RF model resulting in the highest AUC value. Identification is only based on those species present in both the BAME@LMG data set and the Sherlock MIS identification library. Moreover, species identification is only considered following correct identification at genus level. For example, a *Bacillus* profile is not further taken into account when it is identified as *Paenibacillus* in the genus identification model. Identification at genus level is performed by RF and performance has been nearly perfect. Three *Bacillus* profiles have been rejected due to identification as *Paenibacillus* and one *Paenibacillus* profile has been rejected due to identification as *Bacillus*. For the three genera, a distinct gap is observed between the RF identification performance and the Sherlock MIS performance. On average, 78.28%, 94.49% and 75.65% of the *Bacillus*, *Paenibacillus* and *Pseudomonas* species have been correctly identified by RF where Sherlock MIS by the TSBA50 identification library has only achieved 55.77%, 51.22% and 27.00% species identification accuracy. Generally, the machine learning approach has significantly outperformed the commercial identification system for species identification in the three bacterial genera.

Table 2. Multi-class confusion matrix resulting from genus identification by the best SVM experiment.

Bacillus	317	2	1
Paenibacillus	1	125	0
Pseudomonas	0	0	557

The number of correct predictions are presented on the main diagonal, the other cell values show the number of incorrect predictions. Row labels correspond to the true genus names, column labels correspond to the predicted genus names.

Discussion

Stratified identification strategy

The results obtained indicate that, when considering FAME data, the three machine learning techniques RF, ANN and SVM result in nearly perfect genus identification. A first approach towards FAME-based identification of bacterial genera by machine learning has been performed by Bertone et al. [4] and Giacomini et al. [17,18] who successfully identified a limited number of marine and environmental bacteria at genus level by ANN. The researchers have concluded that FAMEs are good biomarkers for bacterial genus identification and that it would be worthwhile to build a FAME-based bacteria identification system at species level. In a taxonomic context, a first in-depth study on FAME-based species identification by machine learning techniques has been performed for the genus *Bacillus* by Slabbinck et al. [46]. From this study, we concluded that species identification by FAME data and machine learning techniques is very promising, taking into

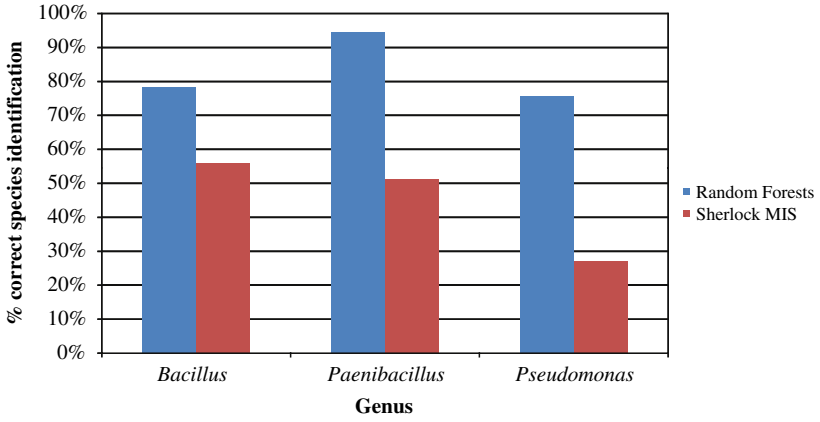

Fig. 5. Comparison of the identification performance by random forests and Sherlock MIS (MIDI Inc., Newark, DE, USA). For each genus, the random forests model resulting in the highest identification performance is used.

account the limited resolution of FAME analysis for species discrimination. The present study has extended the scope by evaluating species identification in the genera *Bacillus*, *Paenibacillus* and *Pseudomonas* by three machine learning techniques: RF, ANN and SVM. These genera have been selected because two genera (*Bacillus* and *Paenibacillus*) belong to the same phylum and are distantly related to the third genus (*Pseudomonas*) which belongs to a different phylum. From a genus-wide identification perspective, all genera are represented in the BAME@LMG database by a sufficient number of species to cover at least half of the validly published species. When considering genus classes only, the three genera *Bacillus*, *Paenibacillus* and *Pseudomonas* can easily be distinguished from each other based on FAME data. Furthermore, analysis of the multi-class confusion matrices show that misclassifications of the FAME profiles are mainly due to misclassifications of *Bacillus* profiles as *Paenibacillus*, and conversely (example given in Table 2). This result was expected as both genera are evolutionary more related to each other than to the genus *Pseudomonas*. Nonetheless, genus identification is surprisingly good when taking into account that the genera *Bacillus* and *Paenibacillus* were reported as hard to distinguish from each other based on numerical FAME analysis [20,24]. Even though, as different bacterial genera can possibly be hard to distinguish based on FAME data, the strategy of selecting the highest output value will fail when extending the taxonomic scope towards dozens of bacterial genera. Therefore, the development of an alternative scoring and weight mechanism will become indispensable for reliable genus and species identification.

Kämpfer [24] concluded that fatty acid analysis has a potential for species differentiation within the genus *Bacillus*. The application of machine learning techniques for FAME-based *Bacillus* species identification has supported this hypothesis. The identification results indicate that species of the genus *Bacillus* can be distinguished and that the application of RF has resulted in the best identification accuracy. From a taxonomic perspective, some species are closely related and are consequently assigned to a species group such as the *Bacillus subtilis* and *Bacillus cereus* groups. Integration of this prior knowledge into computational classification models has confirmed that wrong identifications are mostly due to identifications as species of the same group [46]. Moreover, species that are more distantly related through evolution might also show highly similar FAME patterns. Taking into account this information about species groups and species distinctness, and the presence of 74 *Bacillus* classes, the identification accuracy achieved by RF can be considered as remarkably high.

Similar to the FAME analysis of Kämpfer [24], Heyndrickx et al. [20] concluded that FAME analysis allows genus identification and identification of *Paenibacillus* species into several species groups. As about one-fourth of the species in the genus *Paenibacillus* has been validly published since January 2006, no in-depth study of species discrimination by FAME analysis has previously been performed. Our identification results show that species in the genus *Paenibacillus* can be distinguished from each other based on their FAME profiles and machine learning techniques.

Identification results show that *Pseudomonas* species are harder to distinguish compared to those of *Bacillus* and *Paenibacillus*. Fatty acid analysis of pseudomonads has been a matter of discussion for several decades [22,32–35,39,53]. Two broad studies on this issue were reported by Stead et al. [47] and Vancanneyt et al. [51] showing that analysis of whole-cell fatty acid fingerprints of pseudomonad strains revealed major groups and subgroups corresponding well to the groupings based on DNA–DNA and DNA–rRNA hybridization techniques. The strains of rRNA group I in the study of Vancanneyt et al. [51] represented 29 different and validly described *Pseudomonas* species which could be grouped into four major FAME subgroups. Subgrouping of various phytopathogenic species was also found. The authors demonstrated that whole-cell fatty acid data show some qualitative and quantitative differences among the various subgroups and concluded that some species can only be distinguished based on smaller quantitative differences [47,51]. Nonetheless, machine learning techniques clearly take advantage of these quantitative differences as the RF identification results show that on average 68.90% of the species are assigned a correct species label. The above-mentioned issues are, however, not the only reason for a lower identification percentage. As mentioned in the introduction, taxonomy of the genus *Pseudomonas* has been under revision for several decades. In particular, the taxonomic position of various pathovars is still under discussion. In the present study, we have chosen to follow the *Pseudomonas syringae* taxonomy as proposed by Gardan et al. [16]. Both the limitations of whole-cell FAME analysis and the uncertainties in the taxonomic position of various *Pseudomonas* species seem like the two main reasons of the lower identification percentage. Nonetheless, RF maximally exploits the FAME analysis resolution to distinguish *Pseudomonas* species on a genus-wide scale.

From all the machine learning techniques evaluated in this study, the best identification accuracy has been achieved by RF while SVM has resulted in the worst accuracy. The application of SVMs has resulted in a very poor identification performance, which can possibly be due to two main reasons. Besides the choice of an inappropriate kernel, it is possible that the many species classes with only few FAME profiles per class play a crucial role in SVM classification. Even though SVM

analysis has been performed using the one-versus-others approach, the LibSVM software actually performs a one-versus-one classification. The main disadvantage of this method is bad performance when handling two classes which contain only few data points per class. This is the case in our setup as the species classes contain only few FAME profiles. Consequently, FAME-based SVM identification is biased to a certain degree. Consequently, future identification will be performed by the random forests technique. Future work on the evaluation of the power of FAME analysis to discriminate between one and more species and its integration into identification models will ultimately improve FAME-based species identification and contribute to a better first-line identification.

Straight species identification strategy

Instead of considering a layered identification system, it is also possible to build a single model including all species classes. However, results indicate that the identification accuracy of this approach is confined by the genus *Pseudomonas* which has the largest number of species and profiles and, thus, has a larger weight in the calculation of the statistical metrics. Besides this, additional issues should be taken into account. The complete data set comprises 213 classes of which some have few data points per class. This leads to a much harder classification task compared to only 3 genus classes with many data points per class and, subsequently, 74, 44 and 95 species classes. Moreover, considering the rapidly evolving taxonomy, retraining of the complete model will become necessary in order to achieve up-to-date identification. More classes also result in a longer training time. In contrast, in a layered system only those species identification models which correspond to updated data sets need to be retrained. As such, dropping the genus identification model should not be an option as a layered system clearly results in better identification performance and better scalability.

Comparison with Sherlock MIS

Up until now, identification of FAME profiles has traditionally relied upon commercial identification systems such as Sherlock MIS (MIDI Inc., USA). Even though Sherlock MIS is considered one of the standard technologies for routine FAME-based bacterial identification, the commercially exploited identification system has one main disadvantage when aiming at genus-wide identification. As mentioned in the Materials and methods section, the TSBA50 library contained only 30 of the 142 validly published *Bacillus* species, 18 of the 86 validly published *Paenibacillus* species and 31 of the 112 validly published *Pseudomonas* species. By making use of the BAME@LMG database and machine learning techniques, we have been able to partially fill this gap and to respond to the dynamic character of taxonomy by rapidly creating new data sets and training new up-to-date identification models. Given the fact that Sherlock MIS is a pioneer in FAME analysis, this system is a good benchmark to compare the power of both identification systems. However, reliable benchmarking is only possible by taking into consideration only those *Bacillus*, *Paenibacillus* and *Pseudomonas* species that are present in both the BAME@LMG data set and the TSBA50 identification library. In the present study, a FAME profile has been identified at genus level and, subsequently, at species level by one of the species identification models following successful genus identification. Based on the highest output scores, significantly better identification results are achieved by the machine learning approach. The RF method correctly identifies on average 78.28% of the *Bacillus* species, 94.49% of the *Paenibacillus* species and 75.65% of the *Pseudomonas* species. This is in contrast to the Sherlock MIS which correctly identifies only 55.77%, 51.22% and 27% of the respective species. The main reason for this gap can be found in the different approach of identification. Sherlock MIS calculates correlation values between unknown FAME profiles and the TSBA50 identification library entries based on the Mahalanobis distance where, in contrast to MIDI, machine learning techniques take advantage of learning from the data. Based on the knowledge inside the different data classes, machine learning techniques learn to distinguish the different classes from one another. Next, probability values are given to an unknown FAME profile of belonging to each class. It is clear that machine learning really takes advantage of learning from the data in contrast to the naive Sherlock MIS approach of comparing each FAME profile with each library entry.

Besides this, it should be mentioned that Sherlock MIS includes significantly more genera in its identification libraries. This makes the identification potentially more prone to wrong identification results. Consequently, comparison with Sherlock MIS will become more reliable when more genera are implemented in our identification scheme.

Conclusion

With this study, the next step has been taken towards a computational genus-wide species identification system based on whole-cell FAME data. FAME-based genus and species identification is evaluated using the machine learning methods SVM, ANN and RF. The three machine learning techniques have shown a similar and nearly perfect identification performance at genus level. At species level, experiments have demonstrated

that RF is the best technique for species identification with each of the three genera. Besides this, RF has also several advantages as opposed to ANN and SVM such as robustness against overfitting and optimization of a small number of parameters. Consequently, further work on various other genera and species will be performed by the RF technique in a stratified identification strategy. Considering the limited discriminative power of FAME analysis for species identifications and ongoing discussions about *Bacillus*, *Paenibacillus* and *Pseudomonas* taxonomy, a moderate to high identification performance has been achieved. Compared to the commercial Sherlock MIS (MIDI Inc., USA), the identification performance of *Bacillus*, *Paenibacillus* and *Pseudomonas* species has been significantly improved.

As bacterial taxonomy is rapidly evolving, flexible solutions are required to achieve up-to-date first-line bacterial species identification. In this paper, we have presented a machine learning approach to tackle this problem. Up-to-date and accurate identification are two of the main advantages of this approach as opposed to the Sherlock MIS. Nonetheless, the current approach has some drawbacks. According to the List of Prokaryotic names with Standing in Nomenclature as published in March 2008, *Bacillus*, *Paenibacillus* and *Pseudomonas* comprised 145, 86 and 117 validly published species, respectively [13]. The data set extracted from the BAME@LMG database contained 961, 378 and 1673 standard FAME profiles of 74 *Bacillus* species, 44 *Paenibacillus* species and 95 *Pseudomonas* species, respectively. Particularly for the *Bacillus* and the *Paenibacillus* data set, only half of the validly published species have been included for training of the machine learning techniques. Based on the BAME@LMG database alone, we are thus still far away from a complete genus-wide bacterial species identification. No single computational FAME analysis has been performed on this scale yet. Moreover, knowledge about the heterogeneity of each species is limited by the restricted number of strains and FAME profiles present in the BAME@LMG database. These drawbacks are, however, inherent to research performed at a single institute which can be seen as 'data-restricted' as well as to the rapidly evolving taxonomy. Consequently, this problem may only be solved by future cooperation between different research institutes performing bacterial FAME analysis under the same standardized conditions. Even though cooperation is not straightforward, it should not be a huge obstacle as the proposed approach would benefit all cooperating parties and would improve bacterial species identification in many microbiology-related fields. Moreover, the ultimate solution for the problem lies in building a public FAME database. Where gene and genome sequence databases are hugely expanding in number and content, databases of phenotypic data are still far behind.

As the advantages of machine learning techniques are fast training and learning, and the ability to handle large data sets, future work in FAME-based bacterial species identification by machine learning techniques will involve the implementation of more genera and species. Increasing the number of genera and species will, however, make training of new identification models a harder but challenging computational task, and will lead to more error-prone results. The degree of reduced identification power will, however, depend on the number of genera and species described/included in the new system, but also on the intra- and inter-genus/species variation of the additional and new taxa. The one strain–one taxon descriptions do not provide this natural variation and microbiologists should be discouraged to create such new taxa because of their weak phenotypic discrimination. Hereby, it is also important to note that the expansion of the identification system will be limited as most bacteria do not grow under the same standardized growth conditions or are even unculturable. Next to this, future work will also need to integrate an alternative scoring and weight mechanism to obtain reliable species identification as, in a stratified identification system, species identification fully relies on the power of genus identification. A third important future task corresponds to the high-throughput sequencing methodologies which know a powerful growth since the last decades. At present, we are also investigating the correlation between FAME data and 16S rRNA data at genus and species levels. By the integration of this biological knowledge in the identification models, we are aiming to enforce the presented bacterial species identification system to resolve its inability of distinguishing between very closely related operational taxonomic units.

Acknowledgements

This research is funded by the Belgian Science Policy (BELSPO, Projects C3/00/12 and IAP VI-PAI VI/06). B.S. would like to thank Liesbeth Lebbe and other co-workers of the Laboratory of Microbiology (Ghent University, Belgium) for providing additional data.

Appendix A. Supplementary materials

The online version of this article contains additional supplementary data. Please visit doi:10.1016/j.syapm.2009.01.003.

References

[1] K. Abel, H. Deschmertzing, J.I. Peterson, Classification of microorganisms by analysis of chemical composition I.

Feasibility of utilizing gas chromatography, J. Bacteriol. 85 (5) (1963) 1039–1044.
[2] C. Ash, F.G. Priest, M.D. Collins, Molecular identification of ribosomal-RNA group 3 bacilli (Ash, Farrow, Wallbanks and Collins) using a PCR probe test – proposal for the creation of a new genus *Paenibacillus*, Anton. Leeuw. Int. J. G. 64 (3–4) (1993) 253–260.
[3] R. Berkeley, M. Heyndrickx, N. Logan, P. De Vos, Applications and Systematics of *Bacillus* and Relatives, Blackwell Publishing, Oxford, 2002.
[4] S. Bertone, M. Giacomini, C. Ruggiero, C. Piccarolo, L. Calegari, Automated systems for identification of heterotrophic marine bacteria on the basis of their fatty acid composition, Appl. Environ. Microbiol. 62 (6) (1996) 2122–2132.
[5] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, New York, 1995.
[6] C.M. Bishop, Pattern Recognition and Machine Learning, 1st ed., Springer, New York, 2006.
[7] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[8] L. Čechová, E. Durnová, S. Šikutová, J. Halouzka, M. Němec, Characterization of spirochetal isolates from arthropods collected in South Moravia, Czech Republic, using fatty acid methyl ester analysis, J. Chromatogr. B 808 (2004) 249–254.
[9] C. Chang, C. Lin, LIBSVM: a library for support vector machines, Software available at 〈http://www.csie.ntu.edu.tw/~cjlin/libsvm〉.
[10] P. Dawyndt, M. Vancanneyt, C. Snauwaert, B. De Baets, H. De Meyer, J. Swings, Mining fatty acid databases for detection of novel compounds in aerobic bacteria, J. Microbiol. Meth. 66 (3) (2006) 410–433.
[11] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley, New York, 2001.
[12] E. Eerola, O. Lechtonen, Optimal data processing procedure for automatic bacterial identification by gas–liquid chromatography of cellular fatty acids, J. Clin. Microbiol. 26 (9) (1988) 1745–1753.
[13] J.P. Euzéby, List of bacterial names with standing in nomenclature: a folder available on the Internet, Int. J. Syst. Bacteriol. 47 (1997) 590–592.
[14] L. Fausett, Fundamentals of Neural Networks: Architectures, Algorithms and Applications, Prentice-Hall, Englewood Cliffs, NJ, 1994.
[15] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (2006) 861–874.
[16] L. Gardan, H. Shafik, S. Belouin, R. Broch, F. Grimont, P.A.D. Grimont, DNA relatedness among the pathovars of *Pseudomonas syringae* and description of *Pseudomonas tremae* sp. nov. and *Pseudomonas cannabina* sp. nov. (*ex* Sutic and Downson 1959), Int. J. Syst. Bacteriol. 49 (1999) 469–478.
[17] M. Giacomini, C. Ruggiero, F. Calegari, S. Bertone, Artificial neural network based identification of environmental bacteria by gas-chromatographic and electrophoretic data, J. Microbiol. Meth. 43 (2000) 45–54.
[18] M. Giacomini, S. Bertone, F.C. Soumetz, C. Ruggiero, An advanced approach based on artificial neural networks to identify environmental bacteria, Int. J. Comput. Intell. 1 (2) (2004) 96–103.
[19] D.J. Hand, R.J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, Mach. Learn. 45 (2001) 171–186.
[20] M. Heyndrickx, K. Vandemeulebroecke, P. Scheldeman, K. Kersters, P. De Vos, N.A. Logan, A.M. Aziz, N. Ali, R.C.W. Berkeley, A polyphasic reassessment of the genus *Paenibacillus*, reclassification of *Bacillus lautus* (Nakamura 1984) as *Paenibacillus lautus* comb. nov. and of *Bacillus peoriae* (Montefusco et al. 1993) as *Paenibacillus peoriae* comb. nov., and emended descriptions of *P. lautus* and of *P. peoriae*, Int. J. Syst. Bacteriol. 46 (4) (1996) 988–1003.
[21] J. Heyrman, J. Mergaert, R. Denys, J. Swings, The use of fatty methyl ester analysis (FAME) for the identification of heterotrophic bacteria present on three mural paintings showing severe damage by microorganisms, FEMS Microbiol. Lett. 181 (1999) 55–62.
[22] S. Ikemoto, H. Kuraishi, K. Komagata, R. Azuma, T. Suto, H. Murooka, Cellular fatty acid composition in *Pseudomonas* species, J. Gen. Appl. Microbiol. 24 (1978) 199–213.
[23] T.J.J. Inglis, M. Aravena-Roman, S. Ching, K. Croft, V. Wuthiekamm, B.J. Mee, Cellular fatty acid profile distinguishes *Burkholderia pseudomallei* from avirulent *Burkholderia thailandensis*, J. Clin. Microbiol. 41 (10) (2003) 4812–4814.
[24] P. Kämpfer, Limits and possibilities of total fatty acid analysis for classification and identification of *Bacillus* species, Syst. Appl. Microbiol. 17 (1994) 86–98.
[25] T. Kaneda, Biosynthesis of branched chain fatty acids. I. Isolation and identification of fatty acids from *Bacillus subtilis* (ATCC 7059), J. Biol. Chem. 238 (4) (1963) 1222–1228.
[26] P. Kotilainen, P. Huovinen, E. Eerola, Application of gas–liquid chromatographic analysis of cellular fatty acids for species identification and typing of coagulase-negative Staphylococci, J. Clin. Microbiol. 29 (2) (1991) 315–322.
[27] C. Kunitsky, G. Osterhout, M. Sasser, Identification of microorganisms using fatty acid methyl ester (FAME) analysis and the MIDI Sherlock Microbial Identification System, in: M. Miller (Ed.), Encyclopedia of Rapid Microbiological Methods, PDA, Bethesda, 2006, pp. 1–18.
[28] A. Liaw, M. Wiener, Classification and regression by random forest, R News 2 (3) (2002) 18–22.
[29] N.A. Logan, P. De Vos, Genus I. *Bacillus*, In: P. De Vos, G.M. Garrity, D. Jones, N.R. Krieg, W. Ludwig, F.A. Rainey, K.-H. Schleifer, W.B. Whitman (Eds.), Bergey's Manual of Systematic Bacteriology, vol. 3: The Firmicutes, 2nd ed., Springer, New York, 2009.
[30] M.T. Madigan, J.M. Martinko, P.V. Dunlap, D.P. Clarck, Brock Biology of Microorganisms, 12th ed, Pearson Education Inc., San Francisco, 2009.
[31] T.M. Mitchell, Machine Learning, McGraw-Hill, Boston, 1997.

[32] C.W. Moss, S.B. Samuels, R.E. Weaver, Cellular fatty acid composition of selected *Pseudomonas* species, Appl. Microbiol. 24 (4) (1972) 596–598.
[33] C.W. Moss, S.B. Dees, Identification of microorganisms by gas chromatographic–mass spectrometric analysis of cellular fatty acids, J. Chromatogr. 112 (1975) 595–604.
[34] C.W. Moss, S.B. Dees, Cellular fatty acids and metabolic products of *Pseudomonas* species obtained from clinical specimens, J. Clin. Microbiol. 4 (6) (1976) 492–502.
[35] C.W. Moss, Gas–liquid chromatography as an analytical tool in microbiology, J. Chromatogr. 203 (1981) 337–347.
[36] G.M. Mukwaya, D.F. Welch, Subgrouping of *Pseudomonas cepacia* by cellular fatty acid composition, J. Clin. Microbiol. 27 (12) (1989) 2640–2646.
[37] A.G. O'Donnell, M.R. Nahaie, M. Goodfellow, D.E. Minnikin, V. Hájek, Numerical analysis of fatty acid profiles in the identification of Staphylococci, J. Gen. Microbiol. 131 (1985) 2023–2033.
[38] G.J. Osterhout, V.H. Shull, J.D. Dick, Identification of clinical isolates of Gram-negative nonfermentative bacteria by an automated cellular fatty acid identification system, J. Clin. Microbiol. 29 (9) (1991) 1822–1830.
[39] H. Oyaizu, K. Komagata, Grouping of *Pseudomonas* species on the basis of cellular fatty acid composition and the quinone system with special reference to the existence of 3-hydroxy fatty acids, J. Gen. Appl. Microbiol. 29 (1) (1983) 17–40.
[40] N.J. Palleroni, The road to the taxonomy of *Pseudomonas*, in: P. Cornelis (Ed.), *Pseudomonas*: Genomics and Molecular Biology, Caister Academic Press, Norfolk, 2008, pp. 1–18.
[41] M. Pineiro-Vidal, F. Pazos, Y. Santos, Fatty acid analysis as a chemotaxonomic tool for taxonomic and epidemiological characterization of four fish pathogenic Tenacibaculum species, Lett. Appl. Microbiol. 46 (5) (2008) 548–554.
[42] F.G. Priest, Genus I. *Paenibacillus*, In: P. De Vos, G.M. Garrity, D. Jones, N.R. Krieg, W. Ludwig, F.A. Rainey, K.-H. Schleifer, W.B. Whitman (Eds.), Bergey's Manual of Systematic Bacteriology, vol. 3: The Firmicutes, 2nd ed., Springer, New York, 2009.
[43] M. Riedmiller, H. Braun, A direct adaptive method for faster backpropagation learning: the RPROP algorithm, in: Proceedings of the IEEE International Conference on Neural Networks, San Francisco, USA, 1993, pp. 586–591.
[44] R. Rifkin, A. Klautau, In defense of one-vs-all classification, J. Mach. Learn. Res. 5 (2004) 101–141.
[45] C. Ruggiero, M. Giacomini, F. Calegari, R. Berti, S. Bertone, L. Casareto, Interpretation of gas chromatographic data via artificial neural networks for the classification of marine bacteria, Cytotechnology 11 (1993) S83–S85.
[46] B. Slabbinck, B. De Baets, P. Dawyndt, P. De Vos, Genus-wide *Bacillus* species identification through proper artificial neural network experiments on fatty acid profiles, Anton. Leeuw. Int. J. G. 94 (2) (2008) 187–198.
[47] D.E. Stead, Grouping of plant-pathogenic and some other *Pseudomonas* spp. by using cellular fatty acid profiles,, Int. J. Syst. Bacteriol. 24 (2) (1992) 281–295.
[48] D.E. Stead, J.E. Sellwood, J. Wilson, I. Viney, Evaluation of a commercial microbial identification system based on fatty acid profiles for rapid, accurate identification of plant pathogenic bacteria, J. Bacteriol. 72 (1992) 315–321.
[49] M. Steele, W.B. McNab, S. Read, C. Poppe, L. Harris, A.M. Lammerding, J.A. Odumeru, Analysis of whole-cell fatty acid profiles of verotoxigenic *Escherichia coli* and *Salmonella enteritidis* with the Microbial Identification System, Appl. Environ. Microbiol. 63 (2) (1997) 757–760.
[50] S. Van den Velde, K. Lagrou, K. Desmet, G. Wauters, J. Verhaegen, Species identification of corynebacteria by cellular fatty acid analysis, Diagn. Micr. Infec. Dis. 54 (2) (2006) 99–104.
[51] M. Vancanneyt, S. Witt, W. Abraham, K. Kersters, H.L. Frederickson, Fatty acid content in whole-cell hydrolysates and phospholipid fractions of Pseudomonads: a taxonomic evaluation, Syst. Appl. Microbiol. 19 (1996) 528–540.
[52] G.M. Weiss, F. Provost, Learning when training data are costly: the effect of class distribution on tree induction, J. Artif. Intell. Res. 19 (2003) 315–354.
[53] D.F. Welch, Applications of cellular fatty acid analysis, Clin. Microbiol. Rev. 4 (4) (1991) 422–438.