Systematic and Applied Microbiology 34 (2011) 20–29

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning

Katrien De Bruyne[a,b], Bram Slabbinck[a,c], Willem Waegeman[c], Paul Vauterin[b], Bernard De Baets[c], Peter Vandamme[a,*]

[a] Laboratory of Microbiology, Ghent University, K.L. Ledeganckstraat 35, 9000 Ghent, Belgium
[b] Applied Maths N.V., Keistraat 120, 9830 Sint-Martens-Latem, Belgium
[c] KERMIT, Dept. of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure Links 653, 9000 Ghent, Belgium

ARTICLE INFO

Article history:
Received 17 February 2010

Keywords:
MALDI-TOF MS
Bacteria
Species
Identification
Data processing
Machine learning techniques
Leuconostoc
Fructobacillus
Lactococcus

ABSTRACT

At present, there is much variability between MALDI-TOF MS methodology for the characterization of bacteria through differences in e.g., sample preparation methods, matrix solutions, organic solvents, acquisition methods and data analysis methods. After evaluation of the existing methods, a standard protocol was developed to generate MALDI-TOF mass spectra obtained from a collection of reference strains belonging to the genera *Leuconostoc*, *Fructobacillus* and *Lactococcus*. Bacterial cells were harvested after 24 h of growth at 28 °C on the media MRS or TSA. Mass spectra were generated, using the CHCA matrix combined with a 50:48:2 acetonitrile:water:trifluoroacetic acid matrix solution, and analyzed by the cell smear method and the cell extract method. After a data preprocessing step, the resulting high quality data set was used for PCA, distance calculation and multi-dimensional scaling. Using these analyses, species-specific information in the MALDI-TOF mass spectra could be demonstrated. As a next step, the spectra, as well as the binary character set derived from these spectra, were successfully used for species identification within the genera *Leuconostoc*, *Fructobacillus*, and *Lactococcus*. Using MALDI-TOF MS identification libraries for *Leuconostoc* and *Fructobacillus* strains, 84% of the MALDI-TOF mass spectra were correctly identified at the species level. Similarly, the same analysis strategy within the genus *Lactococcus* resulted in 94% correct identifications, taking species and subspecies levels into consideration. Finally, two machine learning techniques were evaluated as alternative species identification tools. The two techniques, support vector machines and random forests, resulted in accuracies between 94% and 98% for the identification of *Leuconostoc* and *Fructobacillus* species, respectively.

Introduction

A range of physiological, serological, biochemical, chemotaxonomic, and more recently, genomic methods such as 16S rRNA gene sequence analysis and multilocus sequence analysis are typically applied for the identification of bacteria [36]. Nevertheless, new technologies for accurate and rapid identification of bacteria are essential to different fields in microbiology. A rapid, high-throughput identification method, matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry (MALDI-TOF MS), has been introduced in bacterial taxonomy [17,28,40] and successfully applied to a number of taxa [3,16,21,30,41,42]. MALDI-TOF MS has the ability to measure peptides and other compounds in the presence of salts and to analyze complex peptide mixtures, making it an ideal method for measuring non-purified extracts and intact bacterial cells. Different experimental factors, including sample preparation, cell lysis method, matrix solutions and organic solvents, affect the quality and reproducibility of bacterial MALDI-TOF MS fingerprints [26,28,34,38,42,44]. Also, differences in instrumental performance, mass range, and mass resolution have profound effects on the obtained spectra [46]. Bacterial MALDI-TOF MS spectra display common characteristics, such as sharp signal peaks above the baseline noise level and heteroscedasticity of the noise level associated to the baseline value. All of these factors hamper a straightforward data analysis of the raw mass spectra and call for specific data analysis procedures. In this paper, we perform a spectrometric analysis and data preprocessing to tackle the

Abbreviations: ACN, acetonitrile; ACTH, adrenocorticotropic hormone; CHCA, α-cyano-4-hydroxycinnamic acid; *F.*, *Fructobacillus*; FN, false negative; FP, false positive; *k*NN, *k*-nearest neighbour; *L.*, *Leuconostoc*; *Lc.*, *Lactococcus*; MALDI-TOF, matrix-assisted laser desorption ionization-time-of-flight; MDS, multi-dimensional scaling; MS, mass spectrometry; PMCC, Pearson product moment correlation coefficient; RF, random forest; SA, sinapinic acid; SVM, support vector machine; TFA, trifluoroacetic acid; TP, true positive; UPGMA, unweighted pair group method with arithmetic mean.

* Corresponding author.
E-mail address: Peter.Vandamme@UGent.be (P. Vandamme).

0723-2020/$ – see front matter © 2011 Elsevier GmbH. All rights reserved.
doi:10.1016/j.syapm.2010.11.003

above-mentioned problems. From this point on, data analysis can be performed, ultimately leading towards identification. Several data analysis strategies can be applied. Simple data analysis techniques such as similarity calculation are typically performed or one can focus on more advanced techniques such as machine learning methods.

The use of machine learning techniques for microbial species identification purposes is still limited and is dominated by the application of artificial neural networks (ANNs), which have covered data resulting from different analysis techniques, such as fatty acid methyl ester analysis [18,37,47], Fourier-transformed infrared spectroscopy [4,19,20,31], and genetic and proteomic fingerprinting [24,33,39]. However, research on ANN analysis of MALDI-TOF mass spectra for bacterial identification is also limited [7]. In this paper, we evaluate two other popular machine learning techniques: random forests (RFs); and support vector machines (SVMs). Only a small number of RF studies handle MALDI-TOF MS data for the identification of bacteria, while SVMs have not been applied to MALDI-TOF MS data yet [22,30,35,45]. Until present, the majority of machine learning research papers have focused on the identification of particular species, mainly pathogens or species with a clinical or nutritional impact, and do not include species identification schemes within a specific and/or restricted taxon. This is, however, surprising as machine learning techniques could contribute significantly to the field of microbial taxonomy, given their advantages. By learning from the data patterns, machine learning techniques maximally exploit the information embedded in the data. Learning is typically fast and can be regarded as a very important aspect in keeping identification schemes up-to-date with current standings in microbial taxonomy. The methods easily handle multi-dimensional data and can offer solutions where standard analysis methods fail. Finally, different algorithms exist which are implemented in a multitude of open-source software programs and, thus, free to be applied by every scientist.

In this study, we present a data analysis approach for mass spectra obtained by different sample preparation methods, using the 4800 Plus MALDI TOF/TOF™ Analyzer (Applied Biosystems, CA, USA). An objective criterion for data quality scoring is developed and an optimal protocol for the analysis of mass spectral fingerprints is obtained by two different identification methods. First, identification based on similarity searches against identification libraries is performed, and second, two machine learning techniques are evaluated for the identification of *Leuconostoc* and *Fructobacillus* species. Subsequently, a data quality criterion and common data analysis methods are used to develop and evaluate a reference MALDI-TOF MS data set for the genus *Lactococcus*. Finally, the results, advantages and prospects of this spectrometric analysis for bacterial identification are discussed.

Materials and methods

Bacterial strains

For the present study, 59 strains representing all species within the genera *Leuconostoc* (see Table S1), *Fructobacillus* (see Table S1), and *Lactococcus* (see Table S2) were included. For each species, several strains from diverse isolation sources were selected, if available. Rehydration and cultivation of freeze-dried cultures were performed according to the instructions of the BCCM™/LMG Bacteria Collection (Ghent University, Belgium). Strains from the research collection (characterized by R-numbers) were stored in microbank cryogenic vials TM (Pro-Lab Diagnostics, ON, Canada) at −80 °C. After resuscitation, strains were grown aerobically on MRS at 28 °C for 24 h and checked for purity. Prior to MALDI-TOF MS analysis, the strains were subcultured at least twice, or until a homogeneous cell culture was observed.

MALDI-TOF mass spectrometry

Sample preparation

The protocol included the use of α-cyano-4-hydroxycinnamic acid (CHCA) (5 mg/ml) in a 50:48:2 acetonitrile (ACN):water:trifluoroacetic acid (TFA) matrix solution [44]. Chemicals were of high performance liquid chromatography (HPLC) grade quality. To obtain cellular extracts, 1 μL of bacterial cells (manipulated by Looplast® inoculation loops) was washed in HPLC grade water and ethanol. 70% formic acid and pure ACN were added in a 1:1 (v/v) ratio to the bacterial pellet, and the mixture was vortexed for 30 s. The supernatant, called the 'cell extract', obtained after microcentrifugation, was transferred into a new tube. For all strains examined, both cell extracts (1.5 μL) and homogeneous cell smears of the isolated colonies were transferred to the spot sites on a 384-well stainless steel target plate. Cell extracts and cell smears were air-dried for about 10 min. The matrix–organic solvent mixture (1 μL) was added to the spots and allowed to dry (referred to as the dried droplet method). Each sample was spotted, at least, in duplicate, to verify reproducibility. The samples were allowed to air-dry at room temperature, inserted into the mass spectrometer and subjected to MALDI-TOF MS analysis. In addition to the cell smear and cell extract methods, additional sample preparation methods, as described previously [38] were tested on a small number of strains. These included heat treatment (15 min at 95 °C) of the cell extracts and cell smears, sonication (30 s, 0.3 MHz) of intact cells and the so-called sandwich method [44].

MALDI-TOF MS sample analysis

Prior to analysis, the mass spectrometer was externally calibrated with a peptide mix of angiotensin I, glu-fibrinopeptide B, adrenocorticotropic hormone (ACTH) (18-39), insulin (bovine), ubiquitine, and cytochrome c. The matrix solution and external calibration peptide mix were mixed in a 1:1 (v/v) ratio. One microlitre of this mixture was spotted on the designated calibration spots on the 384-well target plate. Internal calibration was not performed because such calibration signals would interfere with the bacterial spectrum [14]. The 4800 Plus MALDI TOF/TOF™ Analyzer (Applied Biosystems, Framingham, MA, USA) was used in the linear mode. The mass spectrometer uses a 200-Hz frequency tripled Nd:YAG laser, operating at a wavelength of 355 nm. Ions generated by the MALDI process were accelerated at 20 kV through a grid at 19.3 kV into a short, linear, field-free drift region into the detector. The detector system, an electron multiplier, detected and counted the generated ions by converting their kinetic energy into an electric current, proportional to the number of ions present. For each spot, 40 sub-spectra for each of 50 randomized positions within the spot (2000 spectra/spot) were collected and presented as one main spectrum. MALDI-TOF mass spectra were generated in the mass range 2–20 kDa. Laser intensity was set between 3600 and 3800 V, obtaining signal intensities between 5×10^2 and 1×10^4. Data were collected in an automated fashion using random sampling over the sample spot to minimize the effects of operator bias when manually searching for sweet spots.

Data pre-processing

Raw data were extracted as t2d files from the 4800 Plus MALDI TOF/TOF™ Analyzer. The t2d files were imported in the Data Explorer 4.0 software (Applied Biosystems, CA, USA) and transformed to text files. These text files consisted of an array containing the signal intensity for each 0.5 m/z value. A Data Explorer script

(freely available on request) was written to export the peak list of each of the samples from one spot set, to separate text files, which were used as input files for the BioNumerics 6.0 software package (Applied-Maths, Sint-Martens-Latem, Belgium).

The description of data analysis procedures using the BioNumerics 6.0 software contains BioNumerics-specific terminologies. For detailed information please refer to the BioNumerics manual [2]. The Data Explorer 4.0 peak list consisted of a spectral intensity value for every 0.5 Da, which was loaded into the BioNumerics software. Using these data points, a densitometric curve was reconstructed. This workflow was integrated into a script to facilitate the import of the normalized peak lists as densitometric curves. Since all measurements were performed after calibration of the 4800 Plus MALDI TOF/TOF™ Analyzer, data could be considered normalized, and additional normalization of the experiment type was not performed.

Each spectrum, defined by a first custom MALDI-TOF BioNumerics experiment type, had a resolution of 40,000 points. Only spectra that showed peak intensities higher than 5×10^2 and without repetitive signals in the spectrum (electric noise due to non-ionization of the sample spot material) were kept for further analysis. This quality scoring was done by visual inspection. To reduce the noise and to obtain spectra that needed less computational power, each spectrum of this experiment type was converted by applying a fivefold reduction in resolution of the normalized tracks (from 40,000 pts to 8000 pts). These reduced spectra were defined by a second custom BioNumerics experiment type (called 'Maldi2') and still contained enough resolution to obtain an accurate sampling of each peak. Therefore, this down-sampling did not eliminate any valuable information. Additionally, a background correction was applied. The minimum value in a moving window of 10 values around each data point was determined, and the average of these minimum values in a moving window of 20 values around each data point was calculated as its related background intensity. Subtracting these background intensities from the respective spectra finally resulted in the Maldi2 experiment type spectra. Subsequently, noise calculation was performed, average spectra were calculated and peak classes were defined. Based on these peak classes, the set of spectra was converted to the final character data set, which was also scaled to a binary data set. This methodology is explained in detail in the following paragraphs.

A BioNumerics script was developed for calculating the noise of each spectrum. Noise is determined as the standard deviation calculated on the total spectrum. For this purpose, an average spectrum was constructed by combining the average values in a moving window of 5 values around each data point. The standard deviation between this average spectrum and the original Maldi2 spectrum was calculated. Dividing this standard deviation by the average spectrum intensity provides the noise (%), which is an objective measure of data quality. The effect of noisy signals on the identification accuracy was evaluated using a Jackknife test (data sampling) and the *k*-nearest neighbour method (*k*NN, data identification). An automated Jackknife test was performed to assess the standard error of each analysis. The principle of the Jackknife method is to take out one entry and to identify this entry against the different groups, in this case, defined as species. This identification was repeated for all entries, and the percentage of correct group identification is a measure of the internal stability of that group [2]. The *k*NN method was applied for identification with $k = 3$. Furthermore, profile matching was restricted to spectra obtained from strains different from the one analyzed. Identification was considered positive if two out of the three matches were concordant. Ultimately, the threshold to reject a spectrum was set to 5%.

Additionally, taking into account all spectra, an average spectrum was created for each strain, regardless of the sample preparation method or culture conditions used. Prior to averaging the spectra, the variability in spectral intensity was eliminated by normalization. Peaks on these spectra were automatically detected using the BioNumerics band search tool (band search filter: minimum profiling of 2.0%; grey zone of 0%; relative to max. value enabled; minimum area of 0%; shoulder sensitivity of 0) [2]. The creation of peak classes was performed on the average spectra, using the BioNumerics band matching tool (optimization of 0%; position tolerance of 0.05%; change towards end of fingerprint of 0%; minimum height of 0%; minimum surface of 0%) [2]. Using these peak classes, band matching of the total data set was performed, and bands within the defined band classes of the experiment type were considered to create a reduced character data set. The resulting data set was then scaled to a binary data format. These mathematical procedures were integrated in different BioNumerics scripts.

Data analysis

Different similarity measures were evaluated to perform numerical analyses on the Maldi2 spectra and the final character data set. Similarities were calculated using curve-based (Pearson Product Moment Correlation Coefficient, PMCC), as well as band-based (Jaccard) measures. The two measures resulted in comparable *k*NN identification results. The Euclidean distance was also considered for the character data set. Please note that for binary data the Euclidean distance corresponds to the Hamming distance. For all analyses, clustering was performed using the unweighted paired-group method with arithmetic mean (UPGMA) method.

Multi-dimensional scaling (MDS) was used for the visualization of the likeliness of data, for example, for exploring similarities or dissimilarities. The MDS algorithm starts with a matrix of data similarities, and then assigns a location to each data point in the *N*-dimensional space using a non-linear least squares fit, minimizing the distances between the data points. The resulting data positions can be displayed by a 3D visualization [11].

Principal component analysis (PCA) is another dimensionality reduction method where linear combinations are composed from the different peaks [15]. Initially, the linear combination representing the largest amount of variability in the data was chosen and defined as the first principal component (PC). Next, subsequent linear combinations were composed that were orthogonal to the previous PCs, repeatedly based on the combination representing the highest variance. A simple visualization of the PCA was achieved by plotting the variance and accumulated variance of the PCs as ranked by the amount of represented variance in a so-called scree plot. From such plot, the number of PCs needed to cover a certain percentage of variability in the data could be determined.

Machine learning

Two popular machine learning techniques were evaluated for identification purposes: support vector machines (SVMs); and random forests (RFs). In a supervised setting, both techniques calculate multiple mathematical functions describing the boundaries between the different species in the data set and were mainly chosen because they can easily handle high-dimensional and non-linearly separable data, such as MALDI-TOF data.

SVMs were constructed by the program R, using the e1071 package. The Gaussian RBF kernel was selected and the values of the two model parameters *C* and gamma were optimized by a grid search. For *C*, the range $[2^{-5},2^{15}]$ was evaluated in steps of 2^2; for gamma, the range $[2^{-15},2^5]$ was evaluated in steps of 2^2. SVM construction in this R package is based on the LibSVM software package [9].

RFs were constructed, using the code available at the website of Breiman [6]. RFs are based on two main parameters: the num-

ber of trees, and the number of split variables at each node of the tree. These parameters were optimized using a grid search, and this optimization was based on the out-of-bag error. The number of trees was optimized over a range of [1000,4000] in steps of 250. The number of split variables was optimized over all peaks, starting from one peak, and by considering a step size of five.

To achieve good generalization of the data and prevent any overfitting, tenfold nested-cross validation with stratification was performed [27,43]. An outer tenfold cross-validation splits the data set into ten stratified subsets wherein each subset was sequentially used for testing and the others for training. Optimization of the different model parameters for training was achieved by an inner tenfold cross-validation. The same principle as in outer cross-validation was used. As RFs have a very low tendency to overfit due to a convergence of the generalization error, the outer cross-validation was performed for parameter optimization and ultimate testing [5].

For performance evaluation, the results of the ten test sets, as generated by the outer-cross-validation, were joined together in a so-called pooled test set. As such, analysis was performed based on all MALDI-TOF profiles present in the data set. From the pooled test set, a multi-class contingency or confusion matrix was calculated. The presence of fifteen species allowed for easy visualization and interpretation of these matrices. A global identification accuracy was calculated by expressing the main diagonal as a percentage of the full data set size. This metric was biased due to the imbalanced nature of the data set, since each species was represented by a different number of spectra. Therefore, statistics were also calculated for each class by decomposing the multi-class confusion matrix in n two-class confusion matrices, in which each of the n classes or species were regarded as the positive class, and all other species as negative. For each two-class confusion matrix or, thus, for each species, an F-score was calculated as $F\text{-score} = (2 \times \text{sensitivity} \times \text{precision})/(\text{sensitivity} + \text{precision})$, where $\text{sensitivity} = TP/(TP + FN)$, and $\text{precision} = TP/(TP + FP)$, with TP the number of true positive identifications, FN the number of false negative identifications, and FP the number of false positive identifications. Finally, an average F-score was calculated over all species.

Results and discussion

Sample preparation

Different sample preparation methods were evaluated to obtain reproducible and informative bacterial spectra. When using the cell smear method in the present study, 1 μL of matrix solution was additionally added, resulting in enhanced signal intensity. Apart from the analysis of cell smears and cell extracts, several other methods have been reported to improve spectra from Gram-positive bacteria [38,44]. In the present study, some of these sample preparation methods were evaluated on a small scale (data not shown). Although other studies do report enhanced spectral qualities [44], no clear improvement in spectrum quality could be observed in this study. Washing of the cells to remove salts did not result in additional peaks or elevated intensities of the spectra signals, which conforms to the observations of Williams et al. [44] and Dieckmann et al. [13]. However, Williams et al. [44] reported that heat treatment of the cells resulted in more peaks, while this was not observed by our experiments.

The sample amount was proven to be critical for MALDI-TOF MS analysis. Too little or too much sample can drastically affect the mass spectrum [29], and makes the difference between success and failure. As tested on multiple strains (data not shown), a more controlled and reproducible method is obtained by using bacterial serial dilutions. Starting from $OD_{610\,nm}$ 1, dilutions up to 1:20 resulted in reproducible spectra. When too little sample was used, the instrument was unable to detect peaks with sufficient signal intensities, resulting in a noisy spectrum. When using too much sample, the detector was saturated and dominant peaks were created. Moreover, too much sample pollutes the MALDI-TOF MS system, interfering with its accuracy and shortening the time between maintenance. The optimal bacterial concentration for MALDI-TOF MS has earlier been quantified by Liu et al. [29] as 4 mg bacterial cells/30 μL matrix solvent solution.

Characteristics of MALDI-TOF mass spectra

MALDI-TOF MS ion intensities usually do not correlate to the relative amount of each component, unless the analyte is carefully prepared and calibrated. However, the relative ion intensities may correspond when studying analytes that are similar in mass and with the same functional groups [23]. The reproducibility of the ion intensities is illustrated in Fig. S1. Spectral profiles of cellular extracts of *F. pseudoficulneus* R-35156, obtained from the same agar plate, contained comparable relative ion intensities (Fig. S1A and B). After subsequently subculturing the strain for three times, the profile (Fig. S1C) showed some variation in ion intensities compared to the other profiles, although the main peak intensities were similar. After storage at −80 °C for one month, the same culture was grown and re-analyzed (Fig. S1D). The main peaks of the spectra were conserved, although differences in peak intensities were observed. Comparable results were obtained using other strains (data not shown). Variations in bacterial growth and corresponding ion intensities were also observed in other studies [10,25,28,46]. Additionally, analyzing inter-laboratory reproducibility, Wunschel et al. [46] concluded that, in the same laboratory, different ion intensities were detected, although the main peak list was generally stable, whereas for other laboratories, additional peaks were detected under the same strict experimental conditions, even when using the same bacterial suspension.

Data processing and clustering

Initial data evaluation

Initially, we performed a multi-dimensional scaling (MDS) analysis based on the final character data set obtained by different sample preparation methods (see Fig. S2). From this figure, no bias in the distribution of the spectra can be directly correlated to the sample preparation methods (e.g., cell smear and cell extract method), which suggests that MALDI-TOF MS data are not preferentially clustered according to the sample preparation method. For the total data set (Fig. 1, left), as well as for each of the species examined, an intermixed image of spectra obtained by the cell smear method (grey dots) and cell extract method (black dots) was observed in the MDS. Yet, for some species, like *L. mesenteroides*, a separation between spectra obtained by the cell extract method and the cell smear method was observed (Fig. 1, right).

Additionally, MDS was performed on spectra generated with the cell smear method and the cell extract method, obtained from cells grown on TSA and MRS, to check for medium-specific clustering. Within spectra of e.g., the *L. citreum* strains, a clear grouping between the sample preparation methods, as well as from the growth media was observed in the MDS (Fig. S3). For e.g., *L. lactis* on the other hand, clustering according to the sample preparation method was also observed, although a coherence tending to cluster according to growth medium was not that distinct as for *L. citreum* (Fig. S3). Clustering according to the growth medium, probably due to slight variations in the expression of proteins, was

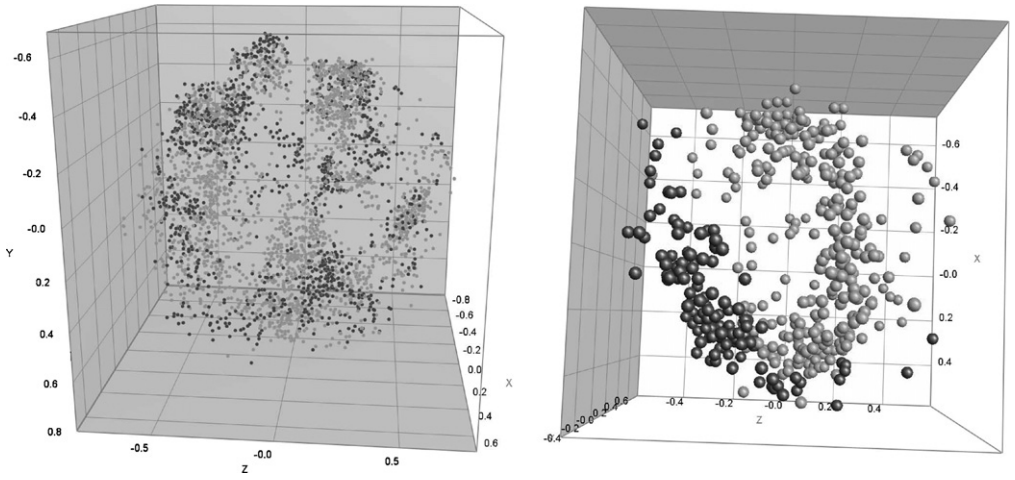

Fig. 1. Left: Multi-dimensional scaling (MDS) representing Maldi2 spectra obtained by the cell smear method (grey dots) and the cell extract method (black dots). Right: MDS of Maldi2 spectra obtained from *L. mesenteroides* strains, by both the cell smear (grey dots) and the cell extract (black dots) method.

observed in the present study for only a few species and has also been previously described [3,13,34].

MALDI-TOF MS analysis and its potential for species delineation within Leuconostoc and Fructobacillus

In a first part of the study, spectra of all 31 *Leuconostoc* and 8 *Fructobacillus* strains grown in standard conditions (24 h, MRS, 28 °C) were generated, using both the cell smear method and the cell extract method. Additionally spectra of the same strains, but grown for prolonged incubation times (48 h, 72 h), and on a different culture medium (TSA) were obtained. In total, 3585 MALDI-TOF mass spectra were generated. The quality of each spectral profile was scored by a noise value. For all entries, noise values between 0 and 35% were obtained.

Starting from the binary data set, different data analyses were performed to reveal a discriminatory effect in the data and to assess how well cluster analysis can be applied to these data. Firstly, a distance matrix was calculated using the Hamming distance, which shows the separability of the different species (see Fig. S4). From the heat map of this matrix, high similarities (grey shadings) within the spectra obtained from one species (as uniquely indicated by the colours on the axes) can be clearly distinguished. Moreover, when looking at the species block of *L. mesenteroides*, three different zones with relatively low distances can be seen to correspond to the three subspecies of *L. mesenteroides*. Furthermore, this figure also indicates low similarities between the spectra of the different species. Although peak intensities are not further considered by the conversion of continuous data to binary data, only taking the binary data into account seems very promising. In addition to the distance matrix calculation, a clustering of the peaks has been performed, and visualized in the heat map in Fig. S5. This clustering clearly shows that many peaks are involved in the delineation of the different species, as exhibited by the many white areas in the heat map, which correspond to a value of one. As a second analysis technique, principal component analysis (PCA) has been applied. A cumulative scree plot (see Fig. S6), denoting the variance and cumulative variance of the top-*n* PCs, shows that only about 60% of the variance are represented in the first twenty principal components. This underscores the observation that many peaks are not correlated and are needed to obtain good discrimination between the different species. From the heat map of the distance matrix and the peak clustering, one can conclude that an integrated analysis should not be restricted towards the most discriminating peaks or peak combinations, but preference should be given to the whole peak profile (2–20 kb).

Towards a MALDI-TOF MS identification library for Leuconostoc and Fructobacillus strains

As species-specific information was demonstrated by the previous analyses, the next step was the evaluation of band- and character-based analyses of MALDI-TOF mass spectra for the classification of bacterial strains within the genera under study. To score the identification potential of this approach, an automated Jackknife test and the *k*-nearest neighbour method were applied. Three identification classes were considered: a match; a mismatch; and an undetermined spectrum. The results provided insight into the accuracy of the MALDI-TOF MS data set for species identification within the groups studied. Identification results are given in Table 1. Species represented by only one strain (e.g., *L. gelidum*, *L. holzapfelii* and *F. ficulneus*) were omitted from the analysis. Ideally, if all spectra were correctly identified, each species should have a 100% identification score. Apart from the percentage matches, the table provides information on the misidentified spectra. For example, *L. pseudomesenteroides* profiles were identified as *L. pseudomesenteroides* in only 74.6% of the analyses, and were mainly misassigned to *L. mesenteroides* (17.0%), and to a lesser extent to *L. lactis* (2.3%), *L. durionis* (1.7%), *L. fallax* (1.7%) and *L. gasicomitatum* (0.5%). Based on these analyses, the total percentages of matches, mismatches, and undetermined spectra calculated from Maldi2 spectra and using the PMCC were 84.9%, 14.3%, and 1.8%, respectively. To conclude, this group separation statistic provided a rapid and reliable method to evaluate the stability of the defined groups, and resulted in an identification reliability of 84.6%.

We also evaluated different identification libraries: one based on spectra obtained by the cell smear method; one based on spectra obtained by the cell extract method; and one combined library including spectra of both methods. For each analysis, the percentage matches, mismatches, and undetermined spectra were determined (Table S3). Within each of the three identification libraries, analysis of both the Maldi2 experiment type and the derived character set yielded comparable identification scores. Additionally, identification libraries covering spectra obtained from bacteria grown from a single culture medium yielded no higher identification scores than identification libraries covering spectra obtained from bacteria grown on different culture media. The identification library covering spectra obtained by the cell extract method yielded slightly higher identification scores than the library covering spectra obtained by the cell smear method (Table S4). However, one should consider the extra work necessary to prepare extracts against the few percentages additional positive identifications. Although not performed in this study, one could

Table 1
Identification results of the Jackknife test and *k*-nearest neighbour analysis (*k* = 3, PMCC, Maldi2 spectra). Only species of the genera *Fructobacillus* and *Leuconostoc* are considered with more than 1 strain. The rows denote the correct species (T) and columns the predicted species (P). Row and column indices correspond to the species *F. durionis* (1), *F. pseudoficulneus* (2), *L. carnosum* (3), *L. citreum* (4), *L. fallax* (5), *L. gasicomitatum* (6), *L. kimchii* (7), *L. lactis* (8), *L. mesenteroides* (9) and *L. pseudomesenteroides* (10). The final column (*U*) gives the percentage of undetermined spectra. An overview of the total percentage of matches, mismatches, and undetermined spectra is given below.

*k*NN	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	U
T1	87.88	0.00	0.00	0.00	0.30	0.00	0.00	4.24	1.21	4.55	1.82
T2	0.00	99.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.47
T3	0.00	0.00	99.15	0.21	0.00	0.21	0.00	0.21	0.21	0.00	0.00
T4	0.00	0.00	0.00	97.66	0.18	0.54	0.00	0.00	1.44	0.00	0.18
T5	0.00	0.00	0.00	0.00	90.39	0.00	1.42	3.20	2.14	0.00	2.85
T6	0.00	0.00	0.00	1.61	0.00	88.76	0.00	0.23	1.83	6.88	0.69
T7	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
T8	0.16	0.00	0.00	0.47	0.95	0.00	0.00	93.06	3.79	0.79	0.79
T9	17.28	0.00	10.48	2.59	3.89	1.62	0.00	4.75	53.24	1.40	4.75
T10	1.69	0.00	0.00	0.00	1.69	0.56	0.00	2.26	16.95	74.58	2.26
Matches	83.94										
Mismatches	14.30										
Undetermined	1.76										

perform a duplicate analysis of the unknown sample and combine the identification results and their respective quality scores. This will provide an additional tool to the researcher to estimate the identification reliability. Moreover, when relying on computer-based identification systems, one could certainly use the combined dataset including spectra obtained from both sample preparation methods and from different culture media, because training of a computer model on this data set takes all observed variation in the data into account, without losing identification efficiency. When using computer-based identification tools, it is important to include the expected (or some degree of) variation that will be possibly observed in later identification studies.

It is also interesting to evaluate the identifications in the perspective of MDS. MDS of the spectra obtained from *Leuconostoc* and *Fructobacillus* strains by the cell smear method as well as the cell extract method is presented in Fig. S7. As stated earlier, 84.6% of the MALDI-TOF MS profiles were correctly identified at species level when using the combined identification library (Table 1). The high coherence of the different species groups is reflected in Fig. S7, presenting the species as differently coloured clusters. Within this 3D space, species clusters are clearly separated. The *L. mesenteroides* cloud is widespread over the total volume in one long cloud, indicating that subspecies do cluster relatively together.

Validation on the MALDI-TOF analysis technique on the genus Lactococcus

As MALDI-TOF mass spectra proved useful for the identification of strains belonging to the genera *Leuconostoc* and *Fructobacillus*, the same experimental setup was used to extend the MALDI-TOF MS library for identification of bacterial strains within the genus *Lactococcus*. Twenty strains (Table S2) were aerobically grown on MRS at 28 °C. Prior to analysis, the strains were subcultured at least twice, or until a homogeneous cell culture was observed. 686 MALDI-TOF mass spectra of these twenty *Lactococcus* strains were measured and further analyzed, as described in 'Materials and methods' section. PMCC-based analysis of the Maldi2 spectra resulted in 94.2% correct identifications taking species and subspecies levels into account, 5.6% mismatches and 0.2% undetermined spectra (Table 2). Again, one species, *Lc. piscium*, only represented by a single strain, was omitted from the analysis. For *Lc. garvieae*, *Lc. lactis* and *Lc. raffinolactis*, almost no misidentifications were observed. All *Lc. lactis* spectra were correctly identified at the species level.

The obtained cluster analysis and MDS of the genus *Lactococcus* (Fig. S8) clearly supports the separated species clusters. The subspecies of *Lc. lactis* are grouped together in the MDS but, as indicated previously, correct subspecies identification was obtained for 88% of the *Lc. lactis* spectra. Cluster analysis on the mass spectra of the Maldi2 experiment type, using the PMCC, resulted in species-specific clusters for the three genera studied. Other studies have reported on the reflection of the genus phylogeny, as obtained through 16S rRNA gene sequence analysis, in the obtained MALDI-TOF MS dendrograms [3,13,34]. This was not observed for our data.

Machine learning

The data analyses discussed in the previous section were performed to resolve a discriminatory effect in the data. These analyses also give an indication on how well machine learning techniques could discriminate between the MALDI-TOF mass spectra of the different species or classes. SVMs and RFs were applied to the binary peak data and the test results showed a very high species identification performance. The multi-class confusion matrices, as resulting from identification on the pooled test set, are reported in Tables 3 and 4. Each row represents the true species while the columns denote the identified species. From both tables, it can

Table 2
Identification results of the Jackknife test and *k*-nearest neighbour analysis (*k* = 3, PMCC, Maldi2 spectra). Only species of the genus *Lactococcus* are considered with more than 1 strain. The rows denote the correct species (T) and columns the predicted species (P). Row and column indices correspond to the species *Lc. garvieae* (1), *Lc. Lactis* subsp. *cremoris* (2), *Lc. lactis* subsp. *hordriae* (3), *Lc. lactis* subsp. *lactis* (4), *Lc. plantarum* (5) and *Lc. raffinolactis* (6). The final column (*U*) gives the percentage of undetermined spectra. An overview of the total percentage of matches, mismatches, and undetermined spectra is given below.

*k*NN	T1	T2	T3	T4	T5	T6	*U*
T1	98.99	1.01	0	0	0	0	0
T2	0	67.92	0	30.19	0	0	1.89
T3	0	0	98.25	1.75	0	0	0
T4	0	5.81	4.65	89.53	0	0	0
T5	0	0	2.04	0	95.92	2.04	0
T6	0	0	0	0	0	100	0
Matches	94.21						
Mismatches	5.64						
Undetermined	0.15						

Table 3
Identification results of a random forest (RF) model on the binary peak data. The rows denote the correct species (T) and columns the predicted species (P). Identification percentages are given. Species of the genera Fructobacillus and Leuconostoc are considered. T indices correspond to the species *F. durionis* (1), *F. ficulneus* (2), *F. fructosus* (3), *F. pseudoficulneus* (4), *L. carnosum* (5), *L. citreum* (6), *L. fallax* (7), *L. gasicomitatum* (8), *L. gelidum* (9), *L. holzapfelii* (10), *L. inhae* (11), *L. kimchii* (12), *L. lactis* (13), *L. mesenteroides* (14) and *L. pseudomesenteroides* (15).

RF	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
T1	96.97	0	0	0	0	0	0	0.61	0	0	0	0	0.61	1.82	0
T2	0	94.87	0	0	1.28	0	0	0	0	0	1.28	1.28	0	0	1.28
T3	0	0	95.38	0	3.85	0	0	0	0	0	0	0	0	0.77	0
T4	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
T5	0.21	0	0	0	97.01	1.28	0	0.21	0	0	0	0	0	1.28	0
T6	0	0	0	0	0.72	98.20	0	0.18	0	0	0.18	0	0	0.72	0
T7	0	0	0	0	5.69	2.49	87.90	1.07	0	0	0	0.36	0	2.14	0.36
T8	0	0.46	0	0	1.38	0	0.23	96.79	0	0	0	0	0	0.46	0.69
T9	0	0	0	0	4.62	0	0	13.85	15.38	0	0	0	0	0	66.15
T10	0	0	0	0	5.00	0	0	0	0	87.50	0	0	2.50	5.00	0
T11	0	0	0	0	1.60	0	0	0	0	0	98.40	0	0	0	0
T12	0	0	0	0	0	0	0	0	0	0	0	96.88	0	0	3.13
T13	0.32	0	0.16	0	2.05	0.63	1.10	0.16	0	0	0	0	93.85	1.10	0.63
T14	0	0	0	0	2.33	0.61	0.61	0	0	0	0.30	0	0	96.16	0
T15	1.13	0	0	0	3.39	1.69	2.26	0	1.13	0	0	1.13	0	9.60	79.66

clearly be seen that overall species discrimination is very high. One species, *L. gelidum* (column T9), is badly predicted, giving mainly false predictions as *L. pseudomesenteroides* (column T15). Spectra of *L. gelidum* LMG 18297^T measured after a certain date, all clustered within the *L. pseudomesenteroides* cluster, which points to a contamination. After growth of the cryo stock of LMG 18297^T, the authenticity of the strain was verified with *pheS* gene sequence analysis [12]. The strain was identified as *L. pseudomesenteroides*. This result confirmed that cross-contamination occurred during the study. Furthermore, in the RF experiments, some MALDI-TOF mass spectra of different species were predicted as *L. carnosum* (column T5), although, on the other hand, this trend was not seen in the SVM experiments. No obvious argument is found for explaining this observation. In developing a species identification scheme, one is mostly interested in evaluating this scheme by calculating a global performance measure. We have calculated an accuracy value, which corresponds to the sum of all correct predictions (main diagonal of matrix) to the data set size. In addition, an *F*-score is calculated by averaging the *F*-scores resulting from comparing each species to the others (see also 'Materials and methods' section). These values are reported in the bar diagram visualized in Fig. 2. Very high identifications are achieved. The accuracy resulting from the SVM experiments is 98.4%, while the RF experiments result in 94.1%. The average *F*-score obtained by the SVM and RF experiments is 96.8% and 89.7%, respectively. Note that the accuracy metric is somewhat biased due to the imbalanced nature of the data set. Therefore, the average *F*-score gives a more reliable view on the identification performance of both machine learning techniques. We can conclude that MALDI-TOF MS analysis, combined with a thorough preprocessing and intelligent prediction models can raise the identification performance of the species described within the genera *Fructobacillus* and *Leuconostoc*.

In the previous section, species identification was evaluated by *k*NN analysis using a Jackknife test. Note that in the described testing setup, identification was performed only against the spectra of different strains, which resulted in the removal of some single-strain species. Therefore, no objective comparison can consequently be made for the results of the machine learning experiments, which identify spectra using the learned mathematical functions representing the intra-species variation among the different strains of all species. Therefore, we chose to perform a second *k*NN identification analysis, using a Jackknife test and all spectra of all species. In other words, in this setting identification is made possible on spectra of the same strain. Note also the different approach between both identification methods: whereas *k*NN analysis identifies spectra by data matching, machine learning techniques identify spectra by learning mathematical functions. The resulting *k*NN identification is given in Table 5. The percentage of matches equals 98.0%. Compared to the results of the *k*NN experiment reported in Table 1, identification was greatly improved, as this experiment resulted in 83.9% matches. Although the presence of species-specific features was hereby confirmed, this also clearly

Table 4
Identification results of a support vector machine (SVM) model on the binary peak data. The rows denote the correct species (T) and columns the predicted species (P). Identification percentages are given. Species of the genera *Fructobacillus* and *Leuconostoc* are considered. Row and column indices correspond to the species *F. durionis* (1), *F. ficulneus* (2), *F. fructosus* (3), *F. pseudoficulneus* (4), *L. carnosum* (5), *L. citreum* (6), *L. fallax* (7), *L. gasicomitatum* (8), *L. gelidum* (9), *L. holzapfelii* (10), *L. inhae* (11), *L. kimchii* (12), *L. lactis* (13), *L. mesenteroides* (14) and *L. pseudomesenteroides* (15).

SVM	P1	PS2	P3	P4	P5	P6	P7	P8	P9	P10	P11	PS12	P13	P14	P15
T1	99.39	0	0	0	0	0	0.30	0	0	0	0	0	0.30	0	0
T2	0	94.87	0	0	0	0	2.56	0	0	0	0	0	0	2.56	0
T3	0	0	99.23	0	0	0	0.77	0	0	0	0	0	0	0	0
T4	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
T5	0	0	0	0	99.79	0.21	0	0	0	0	0	0	0	0	0
T6	0	0	0	0	0	99.28	0.18	0.18	0	0	0.36	0	0	0	0
T7	0	0	0	0	0	0	98.22	0	0	0	0	0	1.07	0.71	0
T8	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
T9	0	0	0	0	0	0	0	0	76.92	0	0	0	0	0	23.08
T10	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0
T11	0	0	0	0	0	0.80	0	0.80	0	0	98.40	0	0	0	0
T12	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
T13	0	0	0	0	0	0.16	0	0	0	0	0	0	98.58	0.79	0.47
T14	0	0	0	0	0	0	0	0	0	0	0	0	0.51	99.49	0
T15	1.13	0	0	0	0	0	0.56	0	10.73	0	0	0	0.56	1.69	85.31

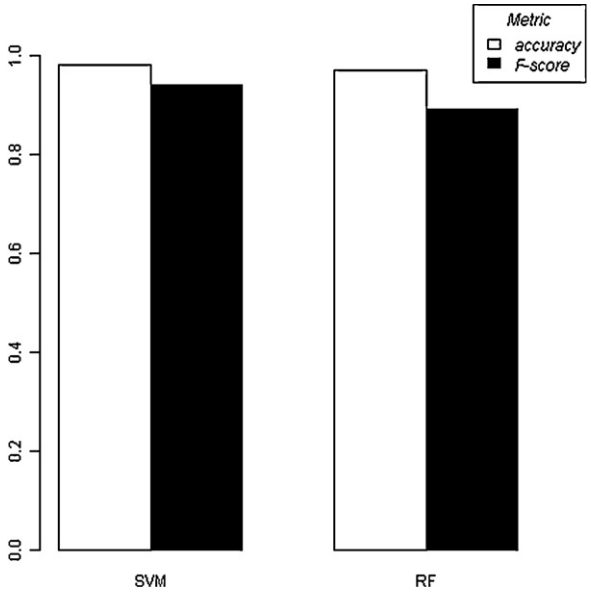

Fig. 2. Bar diagram with the accuracy (white bar) and *F*-score (black bar) measures as calculated from identification with the random forest (RF) and support vector machine (SVM) models.

showed the presence of a certain degree of variation between the spectra of different strains within a single species. Clear examples are species *L. mesenteroides* and *L. pseudomesenteroides*. Please note that the same cross-contamination between *L. gelidum* LMG 18297^T and a *L. pseudomesenteroides* strain is also reflected in these results. When comparing the measures with the results of the machine learning experiments, it is important to note again that the percentage of matches/mismatches/undetermined is also somewhat biased as each species is represented by a different number of spectra. Comparison should, therefore, only be done between the percentage of matches and the accuracy metric. Globally, SVMs perform better, while RFs perform worse than *k*NN.

While MALDI-TOF MS analysis exhibits very high resolution at the species level in the considered genera, machine learning techniques maximally exploit the peak patterns embedded in the MALDI-TOF mass spectra. This identification approach has also different advantages for extending the bacterial scope. It enables rapid identification (range of seconds), easy handling, and, importantly, also allows for very fast retraining of the constructed models. The latter aspect is important for two main reasons: a rapidly changing microbial taxonomy and a growing amount of data, which is easily delivered by the MALDI-TOF MS analysis technique. Storing this data in a MALDI-TOF MS database sets an ideal starting point for the development of identification schemes in multiple genera. Remark that this kind of databases is already commercially exploited e.g., BioTyper and Saramis [1,8]. However, identification by these systems is achieved by entry matching against the corresponding database and the calculation of similarity scores. Our approach, however, exploits all spectral information and variability in the database to build an identification model, and subsequently, to allow the identification of the unknown MALDI-TOF MS profiles.

Given the high species resolution of the MALDI-TOF MS analysis, the rapid and massive data generation, and the advantages of the intelligent machine learning techniques, it is clear that the combination of the spectrometric analysis and the learning aspect results in a very efficient and effective alternative tool for bacterial species identification.

Comparison of MALDI-TOF MS analysis and multi-locus sequence analysis

Concerning the taxonomic resolution, both multi-locus sequence analysis (MLSA) and MALDI-TOF MS allow species level identification [12,32]. Both techniques have the potential to identify at the subspecies level, while for MALDI-TOF MS also the identification at the strain level has already been described [12,22,42]. When handling a large number of bacteria to be identified, the potential of a technique to be automated can become an important point of consideration. Whereas MLSA requires genomic DNA, and therefore an extra cultivation step, MALDI-TOF MS can directly be applied on a single colony. Moreover, when comparing workload and speed, MALDI-TOF MS is clearly most suited for fast automated identification. Concerning data portability, MLSA has one main advantage over MALDI-TOF MS: the universal character of sequences. For MALDI-TOF MS analysis, successful inter-laboratory analyses have already been reported [46], and commercial database systems for identification are available which rely on the inter-laboratory comparison of spectra [1,8]. The main disadvantage of MLSA resides in the fact that prior

Table 5
Identification results of the Jackknife test and *k*-nearest neighbour analysis (*k* = 3) (PMCC, Maldi2 spectra). All strains of all species of the genera *Fructobacillus* and *Leuconostoc* are considered. The rows denote the correct species (T) and columns the predicted species (P). Row and column indices correspond to the species *F. durionis* (1), *F. ficulneus* (2), *F. fructosus* (3), *F. pseudoficulneus* (4), *L. carnosum* (5), *L. citreum* (6), *L. fallax* (7), *L. gasicomitatum* (8), *L. gelidum* (9), *L. holzapfelii* (10), *L. inhae* (11), *L. kimchii* (12), *L. lactis* (13), *L. mesenteroides* (14) and *L. pseudomesenteroides* (15). The final column (*U*) gives the percentage of undetermined spectra. An overview of the total percentage of matches, mismatches, and undetermined spectra is given below.

*k*NN	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	U
T1	98.79	0	0.61	0	0	0	0	0	0	0	0	0	0	0	0.30	0.30
T2	0	97.44	0	0	0	1.28	1.28	0	0	0	0	0	0	0	0	0
T3	0	0	98.46	0	0	0	0	0	0	0	0	0	0	0	0	1.54
T4	0	0	0	99.53	0	0	0	0	0	0	0	0	0	0	0	0.47
T5	0	0	0	0	99.36	0.21	0	0.21	0	0	0	0	0.21	0	0	0
T6	0	0	0	0	0	99.10	0	0.36	0	0	0.18	0	0	0	0	0.36
T7	0	0	0	0	0	0	99.29	0	0	0	0	0	0	0.36	0	0.36
T8	0	0	0	0	0	0	0	99.08	0	0	0	0	0	0	0	0.92
T9	0	0	0	0	0	0	0	0	78.46	0	0	0	0	0	21.54	0
T10	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
T11	0	0	0	0	0	0.80	0	0	0	0	98.40	0	0	0	0	0.80
T12	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0
T13	0.16	0	0	0	0	0	0.63	0	0	0	0	0	97.16	1.10	0.32	0.63
T14	0	0	0	0	0.21	0	0.11	0	0	0	0	0	0.64	98.73	0	0.32
T15	0	0	0	0	0	0	0.56	0	7.34	0	0	0	0	4.52	86.44	1.13
Matches	97.95															
Mismatches	1.59															
Undetermined	0.46															

sequence knowledge is required to design and optimize primer sequences. MALDI-TOF MS on the other hand does not require any knowledge prior to analysis. Taking into account only qualitatively good sequences, one can definitely state that the reproducibility of sequences is better than that of the obtained spectra, because ion intensities can fluctuate among spectra. Not the ion intensities but rather their ratios of intensities should, therefore, be taken into account. The prime investment for both systems (16 capillary sequencer vs. linear MALDI-TOF MS system) is in the same price range, whereas the operational costs vary considerable, with MALDI-TOF MS being an ideal high-throughput, low-cost analysis system.

Conclusion

The minimal sample preparation, sample acquisition, and the speed of the data acquisition combined with its potential for high throughput sample automation, make MALDI-TOF MS a valuable screening and rapid identification method. Whole-cell spectra produced by MALDI-TOF MS have taxonomically characteristic features that can be used to differentiate bacteria at genus, species and subspecies levels, even though only a small portion of the bacterial proteome is detected.

In the present study, the CHCA matrix combined with a 50:48:2 ACN:water:TFA matrix solution was selected to perform the MALDI-TOF MS analyses. When analyzing the spectra obtained by the cell smear method or the cell extract method, ion intensities were not perfectly reproducible for all peaks, although the relative intensities of the main peaks were comparable. The amount of sample was found to be critical for MALDI-TOF MS analysis, since too little sample resulted in noisy spectra, and too much sample gave rise to overly dominant peaks in the spectrum. Multi-dimensional scaling of the spectra illustrated the presence of species-specific information. Data analysis included calculation of the noise score. Based on this noise score, data was withdrawn or rejected for further analysis. Preprocessed profiles, as well as a derived binary character set were successfully used for species identification within the genera *Leuconostoc*, *Fructobacillus*, and *Lactococcus*. Identification scores obtained by analysis of spectra of the cell extract method only, were slightly higher. However, the extra workload to prepare an extract should be considered against the few additional positive identifications. Numerical analysis of (nearly all) spectra, obtained under controlled experimental conditions (24 h bacterial growth, 28 °C, MRS), resulted in species-specific clustering for each of the three genera. The use of the machine learning techniques random forests and support vector machines for the identification of MALDI-TOF mass spectra proved to be very successful.

Automated MALDI-TOF mass spectra acquisition coupled to preprocessing techniques has resulted in high quality data sets. Combined with common data analysis methods and automated machine learning techniques, we have developed a rapid and relatively simple method for the characterization and identification of lactic acid bacteria within the genera *Leuconostoc*, *Fructobacillus*, and *Lactococcus*.

Acknowledgements

The authors want to thank Griet Debyser and Pablo Moerman for technical support, and Freek Spitaels for generating spectra which were partially used in this study.

This work was supported by the Federal Research Policy (Action for the promotion of and Cooperation with the Belgian Coordinated Collections of Microorganisms (C3/00/17)) to K.D.B., the Belgian Science Policy (C3/00/12 and IAP VI-PAI VI/06) to B.S., and the Research Foundation – Flanders to W.W.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.syapm.2010.11.003.

References

[1] 2010 SARAMIS Product Versions, AnagnosTec GmbH.
[2] 2009 BioNumerics Manual, Version 5.1, Applied Maths NV, Sint-Martens-Latem.
[3] Barbuddhe, S.B., Maier, T., Schwarz, G., Kostrzewa, M., Hof, H., Domann, E., Chakraborty, T., Hain, T. (2008) Rapid identification and typing of listeria species by matrix-assisted laser desorption ionization-time of flight mass spectrometry. Appl. Environ. Microbiol. 74 (17), 5402–5407.
[4] Bosch, A., Miñán, A., Vescina, C., Degrossi, J., Gatti, B., Montanaro, P., Messina, M., Franco, M., Vay, C., Schmitt, J., Naumann, D., Yantorno, O. (2008) Fourier transform infrared spectroscopy for rapid identification of nonfermenting Gram-negative bacteria isolated from sputum samples from cystic fibrosis patients. J. Clin. Microbiol. 46 (8), 2535–2546.
[5] Breiman, L. (2001) Random forests. Mach. Learn. 45 (1), 5–32.
[6] Breiman, L. 2004 Random Forests , http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.
[7] Bright, J.J., Claydon, M.A., Soufian, M., Gordon, D.B. (2002) Rapid typing of bacteria using matrix-assisted laser desorption ionisation time-of-flight mass spectrometry and pattern recognition software. J. Microbiol. Methods 48, 127–138.
[8] Bruker Daltonik Gmbh (2010) Application Note #MT-80 Microorganism Identification and Classification based on MALDI-TOF MS Fingerprinting with MALDI Biotyper.
[9] Chang, C., Lin, C. (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.
[10] Claydon, M.A., Davey, S.N., Edwards-Jones, V., Gordon, D.B. (1996) The rapid identification of intact microorganisms using mass spectrometry. Nat. Biotechnol. 14 (11), 1584–1586.
[11] Cox, T.F., Cox, M.A.A. 2001 Multidimensional Scaling, Chapman and Hall, London.
[12] De Bruyne, K., Schillinger, U., Caroline, L., Boehringer, B., Cleenwerck, I., Vancanneyt, M., De Vuyst, L., Franz, C.M.A.P., Vandamme, P. (2007) *Leuconostoc holzapfelii* sp. nov., isolated from Ethiopian coffee fermentation and assessment of sequence analysis of housekeeping genes for delineation of *Leuconostoc* species. Int. J. Syst. Evol. Microbiol. 57 (12), 2952–2959.
[13] Dieckmann, R., Helmuth, R., Erhard, M., Malorny, B. (2008) Rapid classification and identification of salmonellae at the species and subspecies levels by whole-cell matrix-assisted laser desorption ionization-time of flight mass spectrometry. Appl. Environ. Microbiol. 74 (24), 7767–7778.
[14] Donohue, M.J., Smallwood, A.W., Pfaller, S., Rodgers, M., Shoemaker, J.A. (2006) The development of a matrix-assisted laser desorption/ionization mass spectrometry-based method for the protein fingerprinting and identification of Aeromonas species using whole cells. J. Microbiol. Methods 65 (3), 380–389.
[15] Duda, R.O., Hart, P.E., Stork, D.G. 2001 Pattern Classification, John Wiley and Sons, New York, p. 654.
[16] Fagerquist, C.K., Yee, E., Miller, W.G. (2007) Composite sequence proteomic analysis of protein biomarkers of *Campylobacter coli*, *C. lari* and *C. concisus* for bacterial identification. Analyst 132 (10), 1010–1023.
[17] Fenselau, C., Demirev, P.A. (2001) Characterization of intact microorganisms by MALDI mass spectrometry. Mass Spectrom. Rev. 20 (4), 157–171.
[18] Giacomini, M., Ruggiero, C., Calegari, F., Bertone, S. (2000) Artificial neural network based identification of environmental bacteria by gas-chromatographic and electrophoretic data. J. Microbiol. Methods 43, 45–54.
[19] Goodacre, R., Timmins, E.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 Microbiol. Lett. 140 (2–3), 233–239.
[20] Goodacre, R., Timmins, E.M., Burton, R., Kaderbhai, N., Woodward, A.M., Kell, D.B., Rooney, P.J. (1998) Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. Microbiology 144, 1157–1170.
[21] Grosse-Herrenthey, A., Maier, T., Gessler, F., Schaumann, R., Böhnel, H., Kostrzewa, M., Krüger, M. (2008) Challenging the problem of clostridial identification with matrix-assisted laser desorption and ionization-time-of-flight mass spectrometry (MALDI-TOF MS). Anaerobe 14 (4), 242–249.
[22] Hettick, J.M., Kashon, M.L., Slaven, J.E., Ma, Y., Simpson, J.P., Siegel, P.D., Mazurek, G.N., Weissman, D.N. (2006) Discrimination of intact mycobacteria at the strain level: a combined MALDI-TOF MS and biostatistical analysis. Proteomics 6 (24), 6416–6425.
[23] Hillenkamp, F., Peter-Katalinic, J. 2007 MALDI MS: A Practical Guide to Instrumentation, Methods and Applications, Wiley-Vch, Weinheim, p. 345.
[24] Iversen, C., Lancashire, L., Waddington, M., Forsythe, S., Ball, G. (2006) Identification of *Enterobacter sakazakii* from closely related species: the use of artificial neural networks in the analysis of biochemical and 16S rDNA data. BMC Microbiol. 6 (28), 1–8.
[25] Jackson, O.L., Jr. (2001) MALDI-TOF mass spectrometry of bacteria. Mass Spectrom. Rev. 20 (4), 172–194.

[26] Jaskolla, T.W., Karas, M., Roth, U., Steinert, K., Menzel, C., Reihs, K. (2009) Comparison between vacuum sublimed matrices and conventional dried droplet preparation in MALDI-TOF mass spectrometry. J. Am. Soc. Mass Spectrom. 20 (6), 1104–1114.
[27] Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selectionL. In: Fourteenth International Joint Conference on Artificial Intelligence, Montréal, Canada, Morgan Kaufmann.
[28] Lay, J.O., Jr., Liyanage, R. (2006) MALDI-TOF Mass Spectrometry of intact bacteria. In: Wilkins, C.L., Lay, J.O., Jr. (Eds.), Identification of Microorganisms by Mass Spectrometry, John Wiley & Sons, Hoboken, NJ, p. 352.
[29] Liu, H., Du, Z., Wang, J., Yang, R. (2007) Universal sample preparation method for characterization of bacteria by matrix-assisted laser desorption ionization-time of flight mass spectrometry. Appl. Environ. Microbiol. 73 (6), 1899–1907.
[30] Moura, H., Woolfitt, A.R., Carvalho, M.G., Pavlopoulos, A., Teixeira, L.M., Satten, G.A., Barr, J.R. (2008) MALDI-TOF mass spectrometry as a tool for differentiation of invasive and noninvasive Streptococcus pyogenes isolates. FEMS Immunol. Med. Microbiol. 53 (3), 333–342.
[31] Mouwen, D.J.M., Capita, R., Alonso-Calleja, C., Prieto-Gömez, J., Prieto, M. (2006) Artificial neural network based identification of *Campylobacter* species by Fourier transform infrared spectroscopy. J. Microbiol. Methods 67 (1), 131–140.
[32] Pan, C., Xu, S., Zhou, H., Fu, Y., Ye, M., Zou, H. (2007) Recent developments in methods and technology for analysis of biological samples by MALDI-TOF-MS. Anal. Bioanal. Chem. 387 (1), 193–204.
[33] Piraino, P., Ricciardi, A., Salzano, G., Zotta, T., Parente, E. (2006) Use of unsupervised and supervised artificial neural networks for the identification of lactic acid bacteria on the basis of SDS-PAGE patterns of whole cell proteins. J. Microbiol. Methods 66 (2), 336–346.
[34] Ruelle, V., El Moualij, B., Zorzi, W., Ledent, P., Pauw, E.D. (2004) Rapid identification of environmental bacterial strains by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid Commun. Mass Spectrom. 18 (18), 2013–2019.
[35] Satten, G.A., Datta, S., Moura, H., Woolfitt, A.R., Carvalho, M.G., Carlone, G.M., De, B.K., Pavlopoulos, A., Barr, J.R. (2004) Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. Bioinformatics 20 (17), 3128–3136.
[36] Sintchenko, V., Iredell, J.R., Gilbert, G.L. (2007) Pathogen profiling for disease management and surveillance. Nat. Rev. Microbiol. 5 (6), 464–470.
[37] Slabbinck, B., De Baets, B., Dawyndt, P., De Vos, P. (2009) Towards large-scale FAME-based bacterial species identification using machine learning techniques. Syst. Appl. Microbiol. 32 (3), 163–176.
[38] Smole, S.C., King, L.A., Leopold, P.E., Arbeit, R.D. (2002) Sample preparation of Gram-positive bacteria for identification by matrix assisted laser desorption/ionization time-of-flight. J. Microbiol. Methods 48 (2–3), 107–115.
[39] Tuang, F.N., Rademaker, J.L.W., Alocilja, E.C., Louws, F.J., De Bruijn, F.J. (1999) Identification of bacterial rep-PCR genomic fingerprints using a backpropagation neural network. FEMS Microbiol. Lett. 177 (2), 249–256.
[40] Van Baar, B.L. (2000) Characterisation of bacteria by matrix-assisted laser desorption/ionisation and electrospray mass spectrometry. FEMS Microbiol. Rev. 24, 193–219.
[41] Vanlaere, E., Sergeant, K., Dawyndt, P., Devreese, B., Vandamme, P. (2006) Identification of *Burkholderia cepacia* complex using MALDI-TOF mass spectrometry. J. Cyst. Fibr. 5 (Suppl. 1), S34–S134.
[42] Vargha, M., Takats, Z., Konopka, A., Nakatsu, C.H. (2006) Optimization of MALDI-TOF MS for strain level differentiation of *Arthrobacter* isolates. J. Microbiol. Methods 66, 399–409.
[43] Varma, S., Simon, R. (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7 (1), 91.
[44] Williams, T.L., Andrzejewski, D., Lay, J.O., Musser, S.M. (2003) Experimental factors affecting the quality and reproducibility of MALDI TOF mass spectra obtained from whole bacteria cells. J. Am. Soc. Mass Spectrom. 14 (4), 342–351.
[45] Williamson, Y.M., Moura, H., Woolfitt, A.R., Pirkle, J.L., Barr, J.R., Carvalho, M.D.G., Ades, E.P., Carlone, G.M., Sampson, J.S. (2008) Differentiation of *Streptococcus pneumoniae* conjunctivitis outbreak isolates by matrix-assisted laser desorption ionization-time of flight mass spectrometry. Appl. Environ. Microbiol. 74 (19), 5891–5897.
[46] Wunschel, S.C., Jarman, K.H., Petersen, C.E., Valentine, N.B., Wahl, K.L., Schauki, D., Jackman, J., Nelson, C.P., White, E.t. (2005) Bacterial analysis by MALDI-TOF mass spectrometry: an inter-laboratory comparison. J. Am. Soc. Mass Spectrom. 16 (4), 456–462.
[47] Xu, M., Voorhees, K.J., Hadfield, T.L. (2003) Repeatability and pattern recognition of bacterial fatty acid profiles generated by direct mass spectrometric analysis of in situ thermal hydrolysis/methylation of whole cells. Talanta 59, 577–589.