BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 9 2014, pages 1280–1286
doi:10.1093/bioinformatics/btu022

Data and text mining

Advance Access publication January 17, 2014

Automatic identification of mixed bacterial species fingerprints in a MALDI-TOF mass-spectrum

Pierre Mahé[1,*], Maud Arsac[1], Sonia Chatellier[2], Valérie Monnin[2], Nadine Perrot[2], Sandrine Mailler[2], Victoria Girard[2], Mahendrasingh Ramjeet[2], Jérémy Surre[2], Bruno Lacroix[1], Alex van Belkum[2] and Jean-Baptiste Veyrieras[1]

[1]BioMérieux SA, Unit Innovation and Systems, Marcy l'Etoile, 69280, France and [2]BioMérieux SA, Unit Microbiology, R&D Microbiology, La Balme Les Grottes, 38390, France

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry has been broadly adopted by routine clinical microbiology laboratories for bacterial species identification. An isolated colony of the targeted microorganism is the single prerequisite. Currently, MS-based microbial identification directly from clinical specimens can not be routinely performed, as it raises two main challenges: (i) the nature of the sample itself may increase the level of technical variability and bring heterogeneity with respect to the reference database and (ii) the possibility of encountering polymicrobial samples that will yield a 'mixed' MS fingerprint. In this article, we introduce a new method to infer the composition of polymicrobial samples on the basis of a single mass spectrum. Our approach relies on a penalized non-negative linear regression framework making use of species-specific prototypes, which can be derived directly from the routine reference database of pure spectra.

Results: A large spectral dataset obtained from *in vitro* mono- and bi-microbial samples allowed us to evaluate the performance of the method in a comprehensive way. Provided that the reference matrix-assisted laser desorption/ionization time-of-flight mass spectrometry fingerprints were sufficiently distinct for the individual species, the method automatically predicted which bacterial species were present in the sample. Only few samples (5.3%) were misidentified, and bi-microbial samples were correctly identified in up to 61.2% of the cases. This method could be used in routine clinical microbiology practice.

Availability and implementation: The complete dataset including both the reference database and the mock-up mixture spectra is available at http://archive.ics.uci.edu/ml/datasets/MicroMass.

Contact: pierre.mahe@biomerieux.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 15, 2013; revised on December 14, 2013; accepted on January 11, 2014

1 INTRODUCTION

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) (Anhalt and Fenselau, 1975) has been introduced as a genuine paradigm shifting technology in microbiology (Claydon *et al.*, 1996). Between 2005 and 2010, this has been facilitated by the release of commercial systems integrating instrument, reference bacterial spectral databases and user-friendly reporting software (Cherkaoui *et al.*, 2010; Martiny *et al.*, 2012). Although MALDI-TOF MS identification still requires to isolate bacterial colonies on solid culture media, it provides clear advantages: (i) it requires a small amount of reagents reducing cost and waste per sample and (ii) 93% of the isolates are identified within 24 h after inoculation versus 10% for classical techniques (Gaillot *et al.*, 2011; Tan *et al.*, 2012).

Identifying a species on the basis of a new MALDI-TOF MS spectrum is a complex multiclass prediction problem. Current commercial systems can identify several hundreds of bacterial species and a significant number of fungi (van Belkum *et al.*, 2012). This depends on efficient algorithms that ensure reliable identification results. To address this classification task, first approaches relied on similarity-based ranking algorithms (Jarman *et al.*, 2000), but the application of machine learning algorithms such as support vector machines and random forests has more recently been reported as well (De Bruyne *et al.*, 2011; Satten *et al.*, 2004; Villmann *et al.*, 2008). Every classification algorithm requires a reference (or training) database including multiple spectra from bacterial reference strains and species. The ongoing improvement of MALDI-TOF MS methods and databases should drive the introduction of innovative MS-based applications into clinical microbiology. This may include direct sample testing, microbial typing (Dieckmann and Malorny, 2011) and antimicrobial susceptibility testing (Burckhardt and Zimmermann, 2011).

MALDI-TOF MS for the detection of pathogens directly from clinical specimens is possible but only for a limited number of primary specimens and species. Improvement of such methods would constitute a major breakthrough in bacteriology [see Kok *et al.* (2011); Köhling *et al.* (2012) and references herein]. Currently, these approaches suffer from difficulties in sample preparation resulting in a low sensitivity and difficulties with polymicrobial samples. The available algorithms, tailored to identify bacteria from a pure colony, are not able to characterize a bacterial mixture (Fothergill *et al.*, 2013; Lagacé-Wiens *et al.*, 2012). Only a limited number of methods have been proposed to overcome this limitation, and their performance has not been studied in detail (Schleif *et al.*, 2011; Wahl *et al.*, 2002).

*To whom correspondence should be addressed.

We propose a new method for automatic bacterial mixture characterization from a single spectrum using a reference database developed for routine bacterial identification and evaluate it using an original *in vitro* spectral dataset.

2 METHODS

We consider a peak-list representation in which a mass spectrum is represented by a vector $x \in \mathbb{R}^p$: p is the number of channels or 'bins' involved in the peak-list representation, and each entry x_b is derived from the intensity of the peaks found in the b^{th} bin. Several schemes have been proposed to define such a representation (Coombes *et al.*, 2007). In this study, we rely on the approach embedded in the commercial VITEK-MS system (bioMérieux, France). We also assume to have at our disposal a reference database of mass spectra covering a panel of K bacterial species. Given this dataset, we address the problem of predicting which of the K reference bacterial species are actually present in a spectrum to analyze.

For that purpose, we model x as a positive linear combination of species-specific prototypes built from the reference dataset. This can formally be written as follows:

$$x = \sum_{i=1}^{K} \beta_i P_i + \epsilon \tag{1}$$

where $P_i \in \mathbb{R}^p$ is a prototype spectrum representing species i, the coefficient $\beta_i \geq 0$ accounts for the contribution of prototype i in explaining x and $\epsilon \in \mathbb{R}^p$ is a vector of independent and identically distributed (iid) random residuals assumed to be normally distributed. To define the species-specific prototypes P_i, we consider a minimum frequency threshold to introduce a peak at a given position in a prototype. Its intensity is then defined as the median intensity of the reference spectra (of the corresponding species) that exhibit this peak and is null otherwise.

As can be expected and as illustrated in Figure 1, the correlation between pairs of prototypes is in general not null and increases with their taxonomic proximity. Therefore, even if x involves a single bacterial fingerprint, the optimal decomposition of x according to (1), i.e. in terms of the residual sum of squares, may involve the contribution of several prototypes depending on their level of correlation with the prototype of the species actually present in the sample. To overcome this issue, we propose, in a spirit similar to Lindner and Renard (2012), to introduce $\gamma = [\gamma_1, \ldots, \gamma_K]$, the vector of unknown positive contributions of each of the K species to spectrum x and to redefine β_i as $\beta_i = \sum_j a_{ij}\gamma_j$, where a_{ij} quantifies the similarity between species i and j. The mixed spectrum model we consider is now expressed as follows:

$$x = \sum_{i=1}^{K}\sum_{j=1}^{K} a_{ij}\gamma_j P_i + \epsilon = \sum_{j=1}^{K} \gamma_j P_j^{(a)} + \epsilon \tag{2}$$

where $P_j^{(a)} = \sum_{i=1}^{K} a_{ij}P_i$. Therefore, it corresponds to the original one (1) after an appropriate redefinition of the prototypes, which we shall refer to as *adjustement*. If $a_{ii} = 1$ and $a_{ij} = 0$ for $i \neq j$, both models are equivalent.

For typical clinical specimens, we expect the number of distinct species found in a sample to be relatively small. To favor sparsity in the vector γ, we classically rely on the L1 penalty. More precisely, we consider the following optimization problem to estimate γ:

$$\hat{\gamma} = \underset{\gamma \in \mathbb{R}^{+K}}{\operatorname{argmin}} \; ||x - P^{(a)}\gamma||^2 + \lambda|\gamma|_1 \tag{3}$$

where $P^{(a)} = [P_1^{(a)}, \ldots, P_K^{(a)}] \in \mathbb{R}^{p \times K}$ is a matrix containing the adjusted prototypes and $|\gamma|_1 = \sum_{i=1}^{K} \gamma_i$ is the L1 norm of γ.

This problem is the standard Lasso problem with the addition of a positivity constraint. Solving problem (3) for increasing values of the parameter λ provides solutions with a decreasing number of non-null coefficients in γ, hence a collection of models achieving various trade-offs

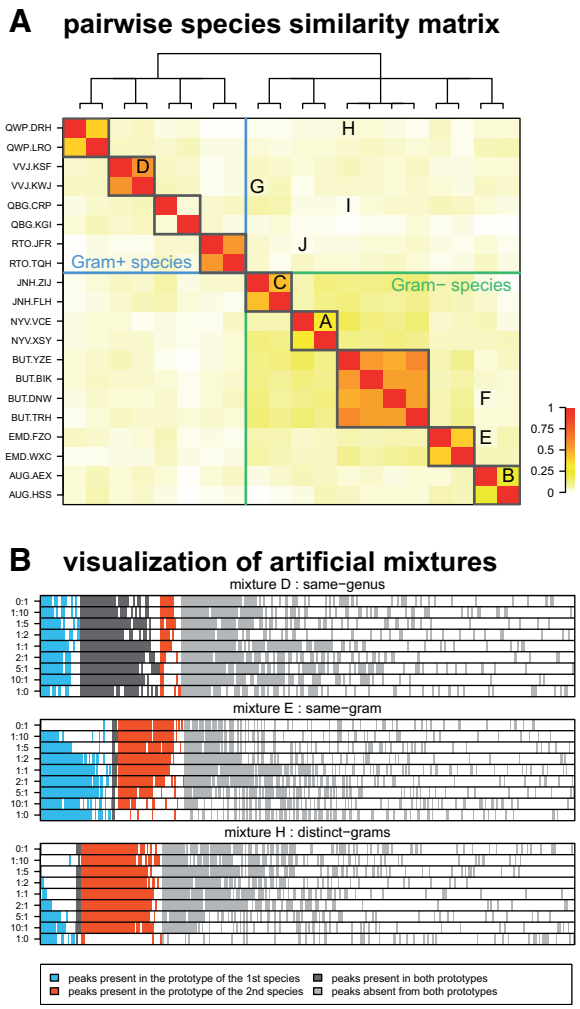

Fig. 1. The correlation between pairs of species prototypes increases with their taxonomic proximity, making combined fingerprints of species from the same genus harder to discriminate. Visualization of the two datasets used in this study. (**A**) *Reference database*. Pairwise species similarity matrix defined as the Jaccard coefficient between prototypes. Gray boxes identify genera and the eight (respectively 12) species shown in the upper-left (respectively lower-right) area correspond to Gram + (respectively Gram −) species, as illustrated by the reduced taxonomy shown above. The A–J letters indicate the position of the mock-up mixtures with respect to this reference dataset. Species are anonymized and represented by six-letter codes, shown on the y-axis, in which the three first letters encode the bacterial genera. (**B**) *Mixtures dataset*. Each panel illustrates a same-genus (top), same-Gram (middle) and distinct-Gram (bottom) bacterial mixture. Each figure shows nine spectra spanning the range of concentrations considered, ranging from the pure spectra of the first species (1:0 ratio) in the bottom, to the pure spectra of the second species (0:1 ratio) at the top, obtained for one of the two couples of strains considered. Peaks were ordered and colored depending whether they belonged to the corresponding species prototypes. We note in particular that for the same-genus mixture, the great majority of peaks that were found in the prototypes were found in both prototypes at the same time, reflecting the limited level of differentiation between species for this mixture. In contrast, the proportion of peaks present in one prototype or the other evolved consistently with the relative concentration of the species in the same-Gram mixture. This was less the case for this particular distinct-Gram mixture, as discussed in the experimental section

between reconstruction error (small λ values leading to many positive coefficients) and sparsity of the solution (high λ values leading to few positive coefficients). In practice, efficient algorithms make it possible to evaluate the entire regularization path of (3) and have access to the whole collection of solutions that can be reached when the parameter λ is varied. In this study, we used the LARS-EN algorithm (Zou and Hastie, 2005), which is implemented in the R `elasticNet` package and can be easily modified to accomodate an additional non-negativity constraint [see Efron *et al*. (2004: p.421), Equations 3.18 and 3.19].

A species is predicted to be part of a spectrum whenever its corresponding entry in the $\hat{\gamma}$ vector obtained from Equation (3) is positive. To choose where to stop on the regularization path and find the number of species prototypes achieving the appropriate trade-off between reconstruction error and sparsity of the solution, we base our model selection strategy on the Bayesian information criterion (BIC). The log-likelihood component of the BIC is derived from a standard linear regression model and is directly related to a least-squares residual, which can naturally be interpreted here as a spectrum reconstruction error. The BIC penalizes this error by the complexity of the reconstruction model, which corresponds here to the number of components involved in the reconstruction, and we select the model that minimizes the BIC along the regularization path. Please refer to the Supplementary Materials (Section S2) for further details about this model selection strategy.

Once the frequency threshold involved in prototype construction, the similarity criterion involved in their adjustment and the likelihood component of the BIC are defined, this method provides a fully automatic procedure to decompose a spectrum.

3 DATASET

We considered a reference panel of 20 Gram-positive and -negative bacterial species covering nine genera among which several species are hard to discriminate by mass spectrometry. Each species was represented by 11–60 mass spectra obtained from 7–20 bacterial strains, constituting altogether a dataset of 571 spectra obtained from 213 strains. The spectra were obtained according to the standard workflow used in clinical routine in which the microorganism was first grown on an agar plate for 24–48 h, before a portion of colony was picked, spotted on a MALDI slide and a mass spectrum was acquired. This dataset is described in more details in the Supplementary Materials (Section S1).

Based on this reference panel, a dedicated *in vitro* mock-up mixture dataset was developed. For that purpose, we considered 10 pairs of species of various taxonomic proximity: four mixtures (labeled A, B, C and D) involving species that belong to the same genus, two mixtures (labeled E and F) involving species that belong to distinct genera but to the same Gram type and four mixtures (labeled G, H, I and J) involving species that belong to distinct Gram types. Each mixture was represented by two pairs of strains, which were mixed according to the following nine concentration ratios: 1:0, 10:1, 5:1, 2:1, 1:1, 1:2, 1:5, 1:10 and 0:1. Two replicate spectra were acquired for each concentration ratio and each pair of strains, leading altogether to a dataset of 360 spectra, among which 80 were actually pure sample spectra. Further details about the sample preparation and spectra acquisition protocols are provided in the Supplementary Materials (Section S1). Figure 1 provides an illustration of the dataset, which is available online at the UCI machine learning repository.

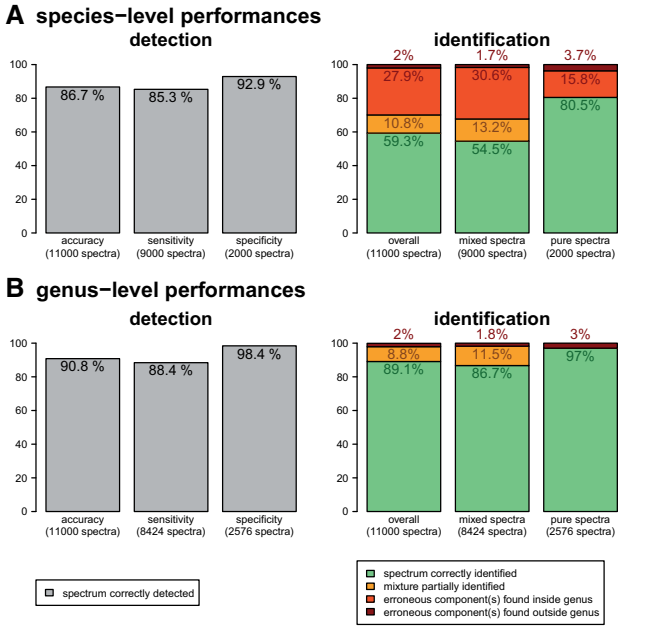

Fig. 2. Simulation study: mixture detection and identification performance. Panels **A** and **B**, respectively, show performances obtained at the species and genus levels. Within each panel, the left and right figures, respectively, show the mixture detection and identification performances defined in Section 4. In both cases, they are considered globally and within mixed and pure spectra. At the species-level, we distinguished between *within-genus* errors in which the erroneous components belonged to the same genera as the ones actually present in the sample and *outside-genus* errors otherwise

4 SIMULATION STUDY

In a first step, we relied on a simulation study to evaluate the behavior of the model and tune its free parameters. For that purpose, we generated a simulated dataset of 11 000 spectra using the additive linear model used for the decomposition algorithm and considered 30 configurations of the method. A detailed presentation of this simulation study and an analysis of the results obtained are provided in the Supplementary Materials (Section S3). These results lead us to select a configuration in which the peak frequency threshold was set to 0.4, the prototype adjustment was based on the Jaccard coefficient and an intercept was introduced in the linear model used to define the likelihood component of the BIC.

The main objective of our work was to evaluate the ability of the method to *detect* a bacterial mixture and to *identify* its components. We defined the sensitivity and specificity of mixture detection, respectively, as the proportion of mixed and pure spectra detected as such by the method. Detection of a mixture was considered to be successful if two or more components were predicted. A mixture was said to be correctly identified whenever both components, and only them, were detected, and partially identified when only one of the two components was detected. A misidentification occured whenever a component not part of the spectrum was detected.

Figure 2A shows the results obtained on the simulated dataset. In terms of detection (left panel), we noted that mixture detection was both sensitive and specific: >85% of the mixtures were

detected, whereas 7.1% of the pure spectra were mistaken for mixtures. Almost 64% of detected mixtures were correctly identified (54.5/85.3%), and this proportion increased to 86.6% for pure spectra (80.5/92.9%). We noted, moreover, that 13.2% of the mixtures were partially identified; hence that mixture identification was at least partially successful for almost 68% of the spectra. Around 31% (respectively 16%) of mixed (respectively pure) spectra were misidentified, but interestingly, we noted that the great majority of errors involved prediction of species that belonged to the same genera as the ones actually present in the samples. As expected, the detection and identification performance of the method was directly related to the taxonomic proximity of the mixed bacterial fingerprints: the closer the species, the more difficult to separate them from a mixed mass spectrum. Shifting the identification level to a higher taxonomic rank should therefore help to improve the performance of the method. Figure 2B shows that this was the case when considering identification at the genus level rather than at the species level. Although the approach could be applied using prototypes defined at the genus level, we derived this decomposition from that obtained at the species level by summing the species contributions among the various genera. The sensitivity and specificity of mixture detection, respectively, reached 88.4 and 98.4%, and >98% (89.1/90.8%) of the spectra correctly detected were correctly identified. Moreover, almost every mixed spectrum that was not detected as such was at least partially identified. Altogether, 89.1% of the spectra were correctly identified and only 2% of them were misidentified. We emphasize, however, that the proportion of 'pure' spectra involving a single component slightly increased when the dataset was analyzed at the genus level (23.4 versus 18.2% at the species level).

As described in the Supplementary Materials (Section S5), we noted, as expected, that the identification performance depended on the relative proportions of the species involved in mixed spectra. Although it remained steady when the minor species accounted for 30–50% of the mixture, a drop of 10% of correct identification was observed when it accounted for 10% of the mixture, at the benefit of the rate of partial identification. Interestingly, we observed, moreover, that the probability of partial identification was related to the proximity between species involved in the mixtures: the species tended to be closer for partially identified spectra (see Supplementary Materials, Section S5). The underlying phenomenon is intuitive: if two species fingerprints are close to each other, the number of peaks specific to each species is expected to be low (as for mixture D in Fig. 1B for instance), and our penalized approach will tend to favor a solution with a single prototype.

Finally, Figure 3A illustrates the relationship between the level of separability between species of a genus, and the proportion of within-genus misidentification obtained when the simulated dataset was analyzed at the species level. As expected, the closer the species, the higher the within-genus error rate. This suggested that the level of resolution could be made genus-specific: at the species-level for genera among which the species separability is sufficiently high and at the genus-level otherwise. In the following, we have adopted such a *hybrid* strategy with a minimum separability cutoff chosen empirically from the results shown in Figure 3A. The resulting identification peformances obtained on the simulated dataset are shown in Figure 3B. We noted that this hybrid-level strategy made it possible to obtain performances close to that obtained at the genus-level, hence markedly higher than the ones obtained at the species-level (84.6% of overall correct identification instead of 89.1 and 59.3% at the genus and species levels, respectively), with the benefit of maintaining a species-level resolution whenever possible. Although it also had the effect of increasing the proportion of 'pure' spectra involving a single hybrid-level component, 40.7% of the 22 000 components involved in the original 11 000 spectra remained analyzed at the species-level. The impact of this strategy on the 10 actual mixtures considered in the study is shown in Figure 3C, where we noted that only three mixtures remained analyzed at the species level, whereas four other ones ended up analyzed at the genus level, and the three remaining ones at a hybrid level, with one component defined at the species level and the other one at the genus level. Figure 3D puts the bacterial panel considered in this study in perspective with a larger one involving 742 bacterial and fungal species extracted from the database embedded in the VITEK-MS system. It revealed that few pairs of species are closer than the minimum distance cutoff considered in this study. Although it is hard to presume how many species would remain analyzed at the species-level if we applied this method to this larger database, it highlighted the fact that the reference panel considered in this study was a relatively challenging one. Interestingly, it showed, moreover, that the separability between species was not so different when the species belonged to the same genus or not. Therefore, this suggested that it may often be possible to distinguish between species of the same genus, which proved to be challenging with the panel of 20 species considered in this study.

5 RESULTS

Results obtained on the actual dataset analyzed at the hybrid level (as previously defined) are shown in Figure 4A. The most striking observation we made was a drop of 10% in the overall identification performance with respect to the simulation study (74.7 versus 84.6%). We noted, moreover, that this gap was even more serious for mixed spectra (61.2 versus 81.8%, hence a drop of >20%), whose proportion was lesser in the actual dataset (62.2 versus 77.7%). To better understand this issue, we resorted to a second simulation study meant to reproduce the actual mixture dataset. To do so, each mixture of the *in vitro* dataset was reproduced a hundred times. Two strains were considered each time, and the grid of relative concentrations considered to build the *in vitro* dataset was used, which therefore provided a simulated dataset of 18 000 spectra (see Supplementary Materials, Section S3). Results obtained on this dataset are shown in the panel B of Figure 4. We noted that the identification performance could be expected to be lower, as a drop of almost 9% in the identification performance of mixed spectra was observed with respect to the previous simulation study (73.1 versus 81.8%), although the grids of concentration ratios considered were slightly different. Nevertheless, a drop of 12% in terms of mixed spectra identification (61.2 versus 73.1%) remained between the actual and this second simulated dataset. Results obtained for the various mixtures considered are shown in the middle column of Figure 4. We noted that the performances obtained on the actual dataset were relatively consistent with

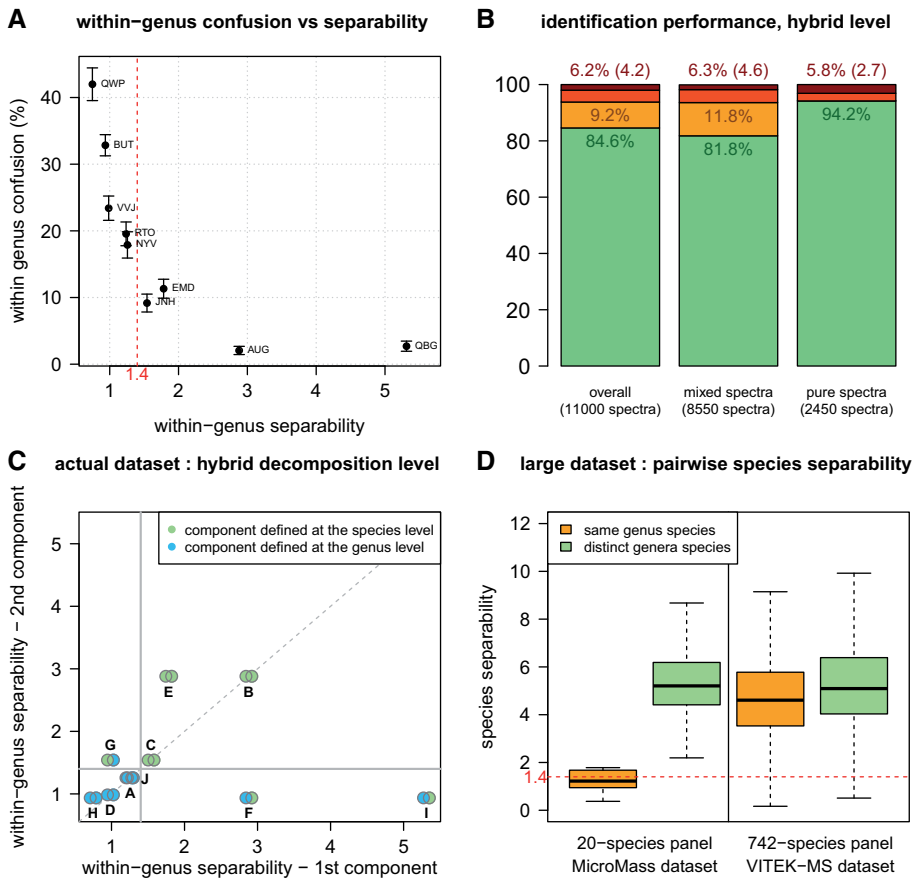

Fig. 3. Low discrimination between species fingerprints within a genus yields lower identification performance at the species level, but such cases are expected to be rare in routine settings. (**A**) *Impact of the level of species separability within a genus on the decomposition performance*. The *x*-axis shows the level of separability between the species of a given genus. This quantity reflects the extent to which species of a genus can be discriminated based on their MS fingerprints and was obtained by comparing each individual spectrum to the prototype of its reference species and to the prototypes of the species of the same genus. The *y*-axis shows the proportion of spectra involving a component of a given genus for which this component was misidentified at the species level but correctly identified at the genus level (i.e. the corresponding species was mistaken for another species of the same genus or predicted together with other species of the same genus). These two quantities were computed from the simulated dataset, as described in detail in the Supplementary Materials (Section S4). We noted that the proportion of erroneous decomposition greatly varied across genera and tended to decrease, whereas the separability between species of the genus increased. A minimum separability threshold of 1.4 seemed to be required to reasonably identify a mixture component at the species level. (**B**) *Hybrid identification performance*. This figure shows the performance obtained when the simulated dataset was analyzed at a hybrid level, where species for which the within-genus separability was <1.4, a threshold defined empirically from panel A, were analyzed at the genus level. (**C**) *Impact of the hybrid strategy on the actual mixture database*. Each mixture was represented according to the level of within-genus separability of its components (shown on the *x*-axis of panel A). The horizontal and vertical gray lines show the threshold of 1.4 used in this study to define a hybrid decomposition level, and the color code illustrates the level of resolution considered to analyze each mixture: mixtures A, D, H and J ended up being analyzed entirely at the genus level, mixtures B, C and E at the species level and mixtures F, G and I at a hybrid level, with one component defined at the species level and the other one at the genus level. Figure 1B illustrated the difficulty of analyzing mixture D at the species-level for instance. (**D**) *Pairwise species separability in a large bacterial panel used for routine identification*. The level of pairwise species separability was computed from the 20 species reference dataset involved in this study and from a larger dataset involving 742 bacterial species that was extracted from the database embedded in the VITEK-MS system (data not shown). The resulting distributions are shown as pairs boxplots gathering the values obtained for species belonging to the same genus (orange) or to distinct genera (green)

those obtained by simulation for six of the eight mixtures analyzed as such. For the two remaining mixtures, however, strong discrepancies were observed.

In the case of mixture G, we noted a much higher rate of erroneous identifications than expected. It turned out, however, that these errors were within-genus errors, involving prediction of the species `JNH.FLH` instead of, or together with the species `JNH.ZIJ` actually present in the sample. As described in the Supplementary Materials (Section S6), it seemed that the `JNH.ZIJ` strains involved in the mixtures were less specific to this species than were the majority of the strains of the reference dataset. Therefore, this suggested that the reference dataset did not capture the full extent of biological variability existing between strains of this species. We noted, moreover, from Figure 3C that the level of separability between these `JNH` species was only slightly higher above the threshold considered to define

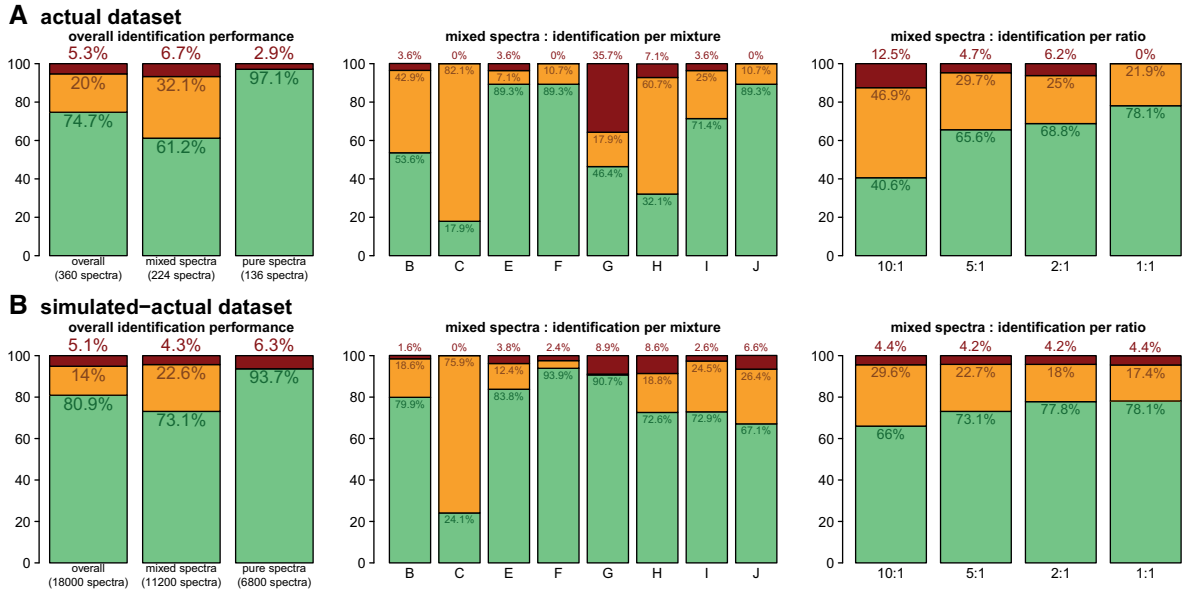

Fig. 4. Real dataset: mixture identification performance. Panels **A** and **B**, respectively, show the mixture identification performances obtained on the actual dataset and on the second simulated dataset, when they were analyzed at the hybrid level defined from Figure 3. The left-hand figures show the mixture identification performances obtained globally and on mixed and pure spectra. The center and right figures, respectively, detail the identification performance obtained for mixed spectra within each mixture and according to the relative concentration on the major species. At the hybrid level, mixtures A and D were considered to be 'pure' because they involved pairs of species of the same genera that were considered to be too similar

the hybrid level. Therefore, this suggested that these two species might have been analyzed at the genus-level if the reference dataset captured more biological variability.

Regarding mixture H, we noted a much higher rate of partial identification than expected. This mixture involved a component of each Gram type, and it turned out that the Gram-negative component was systematically detected, whereas the Gram-positive component was often absent (see Supplementary Materials, Section S9). To explain this behavior, we postulated that the mixture preparation, which relied on optical density measurements, might have suffered from species-specific biases, leading to either an underestimation or an overestimation of their actual concentrations. To validate this hypothesis, the pure bacterial suspensions used to generate the mixed samples were numerated by flux cytometry and plate counting, as described in the Supplementary Materials (Section S7). It turned out that for the same level of optical density (0.5 McFarland), five to seven times more cells were present in the Gram-negative suspension than in the Gram-positive one. This could be explained by the fact that the McFarland standard is known to be accurate to estimate concentrations of Gram-negative bacteria such as *Escherichia coli*, and that the cells of the Gram-positive species were larger than those of the Gram-negative one (see Supplementary Materials, Section S7). As a consequence of this overestimation of the Gram-positive concentration, a mixed sample thought to be balanced (1:1) was actually a five/seven-to-one mixture in favor of the Gram-negative component. Therefore, this meant that a higher proportion of mixed spectra could be expected to lie below the mixture limit of detection, which led to a higher rate of partial identification.

We noted also that while the performances obtained on the actual and simulated datasets were relatively consistent for same-genus mixtures (B and C), mixture B could be identified in more than half of the cases, whereas a much higher rate of partial identification was observed for mixture C. Interestingly, we noted from Figure 3C that the level of species separability within mixture B was much higher than within mixture C. Therefore, this highlighted once again that the level of species separability had a direct impact on the rate of partial identification, as discussed previously and in the Supplementary Materials (Section S5). Finally, Figure 4 (right) shows the identification performance obtained for mixed spectra depending on the relative concentrations of the species involved. We noted an overall discrepancy between the simulated and the actual performances, which could be explained in part by the poor results obtained for mixtures G and H, for the reasons described earlier in the text. We noted, however, that the gap increased when the mixtures were more unbalanced, which therefore suggested that the simulation model considered was optimistic for unbalanced mixtures. We emphasize, however, that the performances obtained on the real dataset were highly consistent with those obtained by simulation for balanced mixtures and that >78% of mixed spectra could be identified in this case.

6 CONCLUSION

We have developed a fully automatic procedure to characterize a polymicrobial sample on the basis of a single mass spectrum that uses the same reference database as the one used to identify pure cultures in clinical routine. Therefore, this method could be used in routine clinical microbiology practice. When evaluated on a *in vitro* mock-up dataset, the procedure exhibited an encouraging performance when the reference bacterial fingerprints were distinct enough, which is the case for most of the species of clinical

interest. Few spectra were misidentified and mixtures were always at least partially identified. More than 60% of the mixtures were detected and correctly identified.

The method was first evaluated by an extensive simulation study, which helped us in turn to optimally set its parameters. This simulation study revealed that the reference panel and the mock-up dataset considered in this study were relatively challenging. Interestingly, we noted also that the results obtained on the actual dataset were consistent with those obtained by simulation. Some discrepancies were observed for two mixtures, but they could be traced back: an insufficient biological coverage of the reference database in the first case and a bias in the sample preparation that lead to an overestimation of the concentration of a Gram-positive species. For these reasons, we believe that the performance truly attainable for this particular mixture panel lies between those obtained on the actual and simulated datasets and would be higher for a less challenging panel.

In terms of perspective, a natural question raised by this study is that of mixtures involving more than two species. Although we could not evaluate our method on real spectra in this setting, a simulation study described in Suppplementary Materials (Section S7) suggested that the approach could work reasonably well for balanced mixtures involving two to five species fingerprints. The proportion of mixtures correctly or partially identified was ~85% of more, and only a single component was missed in most cases of partial identification. We note also that an estimation of the relative concentrations c_i of the bacterial species $i = 1, \ldots, K$ predicted to be present in the sample can be empirically derived from the model parameters estimated by the Lasso, according for instance to $\widehat{c}_i = \hat{\gamma}_i / \sum_{j=1}^{K} \hat{\gamma}_j$. This possibility was not investigated in this work, but the results obtained suggested that when a mixture was correctly identified, this estimation could be reasonably accurate in some cases (see Supplementary Materials, Section S9).

Our future work will be to evaluate the relevance of this approach for blood culture or urine samples without preculture. This set up is likely to bring new challenges because the number of candidate bacterial species may be larger, and a stronger heterogeneity between the culture-based reference database and the spectra to analyze can be expected due to the matrix effects of blood or urine. Although these barriers have still to be overcome, the lessons learned in this study and the simulation framework proposed will be valuable assets toward this goal. Therefore, we think that our method provides an encouraging step toward future automated applications in clinical microbiology based on direct sample mass spectrometry.

Conflict of Interest: The authors are employees of bioMérieux, a company creating and developing infectious disease diagnostics, and in particular the VITEK-MS system involved in this study. No further potential conflicts of interest relevant to this article are reported.

REFERENCES

Anhalt,J. and Fenselau,C. (1975) Identification of bacteria using mass spectrometry. *Anal. Chem.*, **47**, 219–225.

Burckhardt,I. and Zimmermann,S. (2011) Using matrix-assisted laser desorption ionization-time of flight mass spectrometry to detect carbapenem resistance within 1 to 2.5 hours. *J. Clin. Microbiol.*, **49**, 3321–3324.

Cherkaoui,A. *et al.* (2010) Comparison of two matrix-assisted laser desorption ionization-time of flight mass spectrometry methods with conventional phenotypic identification for routine identification of bacteria to the species level. *J. Clin. Microbiol.*, **48**, 1169–1175.

Claydon,M. *et al.* (1996) The rapid identification of intact microorganisms using mass spectrometry. *Nat. Biotechnol.*, **14**, 1584–1586.

Coombes,K.R. *et al.* (2007) Pre-Processing mass spectrometry data. In: Dubitzky,W. *et al.* (eds) *Fundamentals of Data Mining in Genomics and Proteomics*. Springer, US, pp. 79–102.

De Bruyne,K. *et al.* (2011) Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. *Syst. Appl. Microbiol.*, **34**, 20–29.

Dieckmann,R. and Malorny,B. (2011) Rapid screening of epidemiologically important salmonella enterica subsp. enterica serovars by whole-cell matrix-assisted laser desorption ionization–time of flight mass spectrometry. *Appl. Environ. Microbiol.*, **77**, 4136–4146.

Efron,B. *et al.* (2004) Least angle regression. *Ann. Stat.*, **32**, 407–499.

Fothergill,A. *et al.* (2013) Rapid identification of bacteria and yeasts from positive-blood-culture bottles by using a lysis-filtration method and matrix-assisted laser desorption ionization-time of flight mass spectrum analysis with the saramis database. *J. Clin. Microbiol.*, **51**, 805–809.

Gaillot,O. *et al.* (2011) Cost-effectiveness of switch to matrix-assisted laser desorption ionization-time of flight mass spectrometry for routine bacterial identification. *J. Clin. Microbiol.*, **49**, 4412–4412.

Jarman,K. *et al.* (2000) An algorithm for automated bacterial identification using matrix-assisted laser desorption/ionization mass spectrometry. *Anal. Chem.*, **72**, 1217–1223.

Köhling,H. *et al.* (2012) Direct identification of bacteria in urine samples by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry and relevance of defensins as interfering factors. *J. Med. Microbiol.*, **61**, 339–344.

Kok,J. *et al.* (2011) Identification of bacteria in blood culture broths using matrix-assisted laser desorption-ionization sepsityper and time of flight mass spectrometry. *PLoS One*, **6**, e23285.

Lagacé-Wiens,P. *et al.* (2012) Identification of blood culture isolates directly from positive blood cultures by use of matrix-assisted laser desorption ionization-time of flight mass spectrometry and a commercial extraction system: analysis of performance, cost, and turnaround time. *J. Clin. Microbiol.*, **50**, 3324–3328.

Lindner,M.S. and Renard,B.Y. (2012) Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Res.*, **41**, e10.

Martiny,D. *et al.* (2012) Comparison of the Microflex LT and Vitek MS systems for routine identification of bacteria by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *J. Clin. Microbiol.*, **50**, 1313–1325.

Satten,G. *et al.* (2004) Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. *Bioinformatics*, **20**, 3128–3136.

Schleif,F.-M. *et al.* (2011) Hierarchical deconvolution of linear mixtures of high-dimensional mass spectra in microbiology. In: *Proceedings of the IASTED International Conference Artificial Intelligence and Applications.*

Tan,K. *et al.* (2012) Prospective evaluation of a matrix-assisted laser desorption ionization-time of flight mass spectrometry system in a hospital clinical microbiology laboratory for identification of bacteria and yeasts: a bench-by-bench study for assessing the impact on time to identification and cost-effectiveness. *J. Clin. Microbiol.*, **50**, 3301–3308.

van Belkum,A. *et al.* (2012) Biomedical mass spectrometry in today's and tomorrow's clinical microbiology laboratories. *J. Clin. Microbiol.*, **50**, 1513–1517.

Villmann,T. *et al.* (2008) Classification of mass-spectrometric data in clinical proteomics using learning vector quantization methods. *Brief. Bioinform.*, **9**, 129–143.

Wahl,K. *et al.* (2002) Analysis of microbial mixtures by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Anal. Chem*, **74**, 6191–6199.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B*, **67**, 301–320.