Gene expression

MSclassifR: an R package for supervised classification of mass spectra with machine learning methods

Alexandre Godmer[1,2*], Yahia Benzerara[1], Nicolas Veziris[1,2], Mariette Matondo[3], Alexandra Aubry[2,4], Quentin Giai Gianetto[3,5]

[1] AP-HP, Sorbonne université, Hôpital Saint-Antoine, Département de Bactériologie, Paris, France ; [2] Sorbonne Université, INSERM, U1135, Centre d'Immunologie et des Maladies Infectieuses, Cimi-Paris, Paris, France; [3] Institut Pasteur, Université de Paris, Proteomics Platform, Mass Spectrometry for Biology Unit, UAR CNRS 2024 ; [4] AP-HP, AP-HP. Sorbonne-Université, Hôpital Pitié-Salpêtrière, Centre National de Référence des Mycobactéries et de la Résistance des Mycobactéries aux Antituberculeux, Paris, France ; [5] Institut Pasteur, Université de Paris, Bioinformatics and Biostatistics HUB, Computational Biology Department

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Classification of mass spectra is essential for identifying microorganisms from matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry. However, spectrally close organisms remain difficult to identify. In this context, we developed the MSclassifR R package to improve the classification of mass spectra. Its open code strengthens the reproducibility of analyzes in the community. **Results:** We applied the functions of our package to raw mass spectra from three different laboratories. The best workflow available in MSclassifR package achieves near 100% accuracy in all three datasets. Thus, MSclassifR constitutes an interesting alternative for reliable MALDI-TOF based diagnosis.

Availability: MSclassifR is freely available online from CRAN repository https://cran.r-project.org/web/packages/MSclassifR/index.html. Two vignettes illustrating how to use the functions of this package from real data sets are also available online to help users.

Contact: alexandre.godmer@aphp.fr **Supplementary information:** Supplementary material is available at *Bioinformatics* online.

1 Introduction

In clinical microbiology, the identification of pathogens is an essential step for the proper management of patients. This is particularly important for combating antibiotic resistance and providing adequate treatment to infected patients. To identify pathogens from biological cultures, Matrix Assisted Laser Desorption Ionization - Time of Flight (MALDI-TOF) mass spectrometry has become widespread worldwide because it (i) allows for faster pathogen identification (ii) is cheap and (iii) is a user friendly method (Hou, Chiang-Ni and Teng, 2019).

This technique consists in comparing a mass spectrum, acquired after the ionization of intact proteins of a bacterial culture coming from a biological sample, with referenced mass spectra profiles. For this purpose, spectrometer manufacturers provide easy-to-use commercial software. These solutions match the measured mass spectrum to previously measured mass spectra for known pathogens. Although this technique works well for distinguishing many species, the performance of commercial systems remains disappointing for distinguishing spectrally close organisms. This can occur between genetically close subspecies such as *Mycobacterium abscessus* complex (Panagea *et al.*, 2015; Suzuki *et al.*, 2015) or *Enterobacter cloacae* complex (Godmer, Benzerara, *et al.*, 2021).

To get the most out of the mass spectra provided by this technique, several studies have proposed using advanced machine-learning methods (Weis, Jutzeler and Borgwardt, 2020). They generally focus on a single machine learning method and attempt to predict the pathogen associated with a mass spectrum as precisely as possible from mass spectra libraries. Some provide the associated computer code (Asakura *et al.*, 2018) (Weis *et al.*, 2020) (Desaire and Hua, 2020) while many others do not, calling into question the reproducibility of the analyzes carried out (Weis, Jutzeler and Borgwardt, 2020). Several R packages already provide many well-documented methods for preprocessing MALDI-TOF mass spectra (Gibb, 2015) (Palarea-Albaladejo *et al.*, 2017), but additional steps must be applied to classify pre-processed mass spectra, the main ones being the selection of discriminating peaks, and the selection of an optimal mass spectra classification model. Each of these steps requires advanced knowledge and optimal parameter settings. In this context, we developed MSclassifR

to provide a software for accurately identifying some pathogens and strengthening the reproducibility of analyzes in the community.

2 Methods

MSclassifR offers R functions that concerns three main topics: (i) preprocessing of mass-spectra (ii) selection of discriminative mass-over-charge values (m/z) (Figure 1 A) (iii) estimation of a classification model to predict the category of a mass spectrum from shortlisted mass-over-charge values (Figure 1 B). We used MSclassifR on two data sets freely available online to determine an optimal data analysis pipeline leading to the prediction of the belonging of a mass spectrum to a category that is as precise as possible. The first dataset concerns *Klebsiella sp.* (Lasch, Stämmler and Schneider, 2016), the second is related to *Ecrobia sp.* (Wilke *et al.*, 2020). Additionally, we compared the performance of this pipeline with a commercial software from an in-house dataset concerning the *Mycobacterium abscessus* complex (Godmer, Aubry, *et al.*, 2021).

3 Results

We tested several variable selection methods combined with machine learning-based classification models on the two online datasets to define an optimal data analysis pipeline giving the best predictions. For each data analysis pipeline tested, a training set was used to select peaks and to estimate a classification model from these peaks with cross-validation. Next, the accuracy of the model is evaluated on a test set. This was performed with several training and test datasets to assess the robustness of the estimated accuracies. From our numerical experiments, it appears that peak selection before estimating a prediction model saves both computational time and prediction accuracy (Supplementary File 1). Furthermore, the use of a recursive feature elimination method based on random forests (RFE-RF) to select discriminant peaks and the subsequent use of a random forest-based classification model (RF) allows reaching near 100% accuracies with few variations, and with a low computation time on both datasets (Figure 1 B – Supplementary File 1). This could therefore be an interesting data analysis pipeline to apply to other datasets while keeping in mind that other methods give almost similar results and could give interesting results on other datasets (Figure 1 B).

Using the optimal data analysis pipeline previously defined, we applied it on the in-house dataset previously described (Godmer, Aubry, *et al.*, 2021), and compared it to a commercial software provided by a manufacturer: the Bruker® database (Library Mycobacteria v5.0) coupled to the MALDI Biotyper®. The identification the commercial system allowed us to identify a subspecies for only 60% of the mass spectra tested in our internal dataset. Additionally, the accuracy is only of 61% for the predicted subspecies (Supplementary File 2). For comparison, the best pipeline we defined previously (RFE-RF + RF) gives subspecies for all mass spectra with accuracies around 100% (mean=98.49%, s.d.= 2.5%) on test datasets using the framework we described previously (Figure 1 B – Supplementary File 1). It seems thus much more accurate.

Finally, these results show that MSclassifR is an interesting solution for routine research in the biology laboratory to identify pathogens from MALDI-TOF mass spectra. Two vignettes illustrating how to use the functions of this package from the datasets we used are available online to help users (https://agodmer.github.io/MSclassifR_examples/Vignettes/Vignettemsclassifr_Ecrobia.html; https://agodmer.github.io/MSclassifR_examples/Vignettes/Vignettemsclassifr_Klebsiella.html). Its code is open so that it contributes to the reproducibility of analyzes. It will be regularly updated to share new pipelines with interesting results for the community.

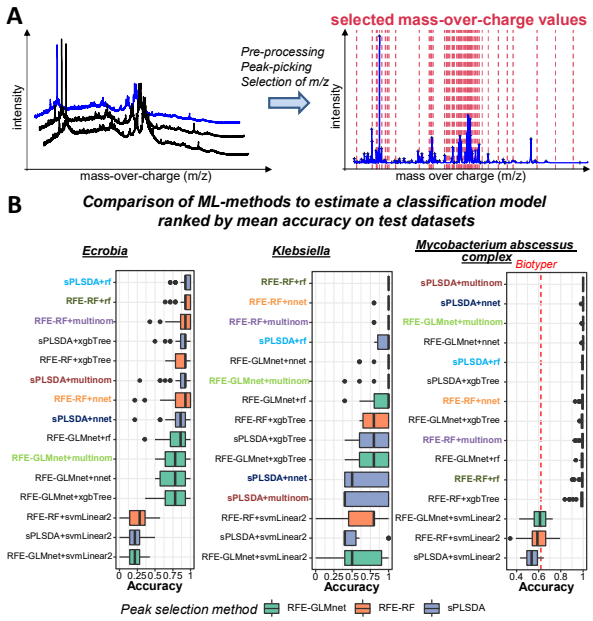

Fig 1: **A**: *MSclassifR proposes pipelines to preprocess mass spectra, and find discriminant mass-over-charge values.* **B**: *Accuracies obtained on test sets for the 3 datasets (Wilke et al., 2020), (Lasch, Stämmler and Schneider, 2016),* (Godmer, Aubry, *et al.*, 2021). *Top 3 methods are highlighted with different colors. Red dashed line named "Biotyper" is the accuracy level (61%) using Bruker® database (Library Mycobacteria v5.0) coupled to the MALDI Biotyper®.*

Conflict of Interest: none declared.

References

Asakura, K. *et al.* (2018) 'Rapid and easy detection of low-level resistance to vancomycin in methicillin-resistant Staphylococcus aureus by matrix-assisted laser desorption ionization time-of-flight mass spectrometry', *PLOS ONE*, 13(3), p. e0194212. doi: 10.1371/JOURNAL.PONE.0194212.

Desaire, H. and Hua, D. (2020) 'Adaption of the Aristotle Classifier for Accurately Identifying Highly Similar Bacteria Analyzed by MALDI-TOF MS', *Analytical Chemistry*, 92(1), pp. 1050–1057. doi: 10.1021/acs.analchem.9b04049.

Fangous, M.-S. *et al.* (2014) 'Classification algorithm for subspecies identification within the Mycobacterium abscessus species, based on matrix-assisted laser desorption ionization-time of flight mass spectrometry', *Journal of Clinical Microbiology*, 52(9), pp. 3362–3369. doi: 10.1128/JCM.00788-14.

Gibb, S. (2015) 'MALDIquant : Quantitative Analysis of Mass Spectrometry Data', pp. 1–16. Available at: http://strimmerlab.org/software/maldiquant/. (Accessed: 22 August 2018).

Godmer, A., Benzerara, Y., *et al.* (2021) 'Revisiting Species Identification within the Enterobacter cloacae Complex by Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry', Microbiology Spectrum, 9(1). doi: 10.1128/spectrum.00661-21.

Godmer, A., Aubry, A., et al. (2021) 'Spectral database of the subspecies of the Mycobacterium abscessus complex (MALDI-TOF Mass Spectrometry)'. doi: 10.5281/ZENODO.5793313.

Hou, T. Y., Chiang-Ni, C. and Teng, S. H. (2019) 'Current status of MALDI-TOF mass spectrometry in clinical microbiology', *Journal of Food and Drug Analysis*. No longer published by Elsevier, pp. 404–414. doi: 10.1016/j.jfda.2019.01.001.

Lasch, P., Stämmler, M. and Schneider, A. (2016) 'A MALDI-TOF Mass Spectrometry Database for Identification and Classification of Highly Pathogenic Microorganisms from the Robert Koch-Institute (RKI)', (October). doi: 10.5281/ZENODO.163517.

Palarea-Albaladejo, J. *et al.* (2017) 'MALDIrppa: quality control and robust analysis for mass spectrometry data', *Bioinformatics*. Edited by J. Hancock, 34(3), pp. 522–523. doi: 10.1093/bioinformatics/btx628.

Panagea, T. *et al.* (2015) 'Mycobacterium abscessus complex identification with matrix-assisted laser desorption ionization - Time of flight mass spectrometry', *Journal of Clinical Microbiology*, 53(7), pp. 2355–2358. doi: 10.1128/JCM.00494-15.

Suzuki, H. *et al.* (2015) 'A novel cluster of Mycobacterium abscessus complex revealed by matrix-assisted laser desorption ionization-time-of-flight mass spectrometry (MALDI-TOF MS)', *Diagnostic Microbiology and Infectious Disease*, 83(4), pp. 365–370. doi: 10.1016/j.diagmicrobio.2015.08.011.

Weis, C. *et al.* (2020) 'Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra', *Bioinformatics*, 36(Supplement_1), pp. I30–I38. doi: 10.1093/BIOINFORMATICS/BTAA429.

Weis, C. V., Jutzeler, C. R. and Borgwardt, K. (2020) 'Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review', *Clinical Microbiology and Infection*. Elsevier B.V. doi: 10.1016/j.cmi.2020.03.014.

Wilke, T. *et al.* (2020) 'Proteomic Fingerprinting Discriminates Cryptic Gastropod Species', *Malacologia*, 63(1), pp. 131–137. doi: 10.4002/040.063.0113.