A Call to Action: the Need for Standardization in Developing Open-Source Mass Spectrometry-Based Methods for Microbial Subspecies Discrimination

Chase M. Clark,[a] Brian T. Murphy,[a] Laura M. Sanchez[a]

aDepartment of Pharmaceutical Sciences, University of Illinois at Chicago, Chicago, Illinois, USA

KEYWORDS MALDI-TOF MS, bioinformatics, dereplication, microbial ecology

In the last decade, there has been a renewed push by academic researchers to create rapid and accurate techniques to differentiate, identify, and prioritize culturable microbial isolates. One such technique that continues to gain momentum among microbiologists is matrix-assisted laser desorption–ionization time of flight mass spectrometry (MALDI-TOF MS). It is an established, inexpensive technique commonly used to rapidly identify microbial taxa and differentiate culturable microbes. This technology has become commonplace in clinical and veterinary laboratories where rigorously validated methods are used in conjunction with commercially available reference databases to identify pathogenic microorganisms. However, the broader community, especially laboratories working with environmental microbes, typically cannot access the expensive software and databases. It is our opinion that this community, which relies on free and open-source software, currently lacks a coherent set of accepted experimental practices, including employment of internal standard strains, statistically driven determination of biological and technical replicates, and deposition of MS data into open-access repositories. Establishing guidelines would enable researchers to better compare microbial typing methods and advance our ability to group and delineate environmental isolates in an effective manner, particularly at the subspecies level.

Toward this end, we recommend that future studies should, at a minimum, employ the following guidelines. (i) When creating/validating methods, the reported accuracy/precision should be obtained using spectra from new biological replicates or closely related strains ("test data") that were not used to determine the parameters of the method ("training and validation data"). Though often a challenge, test spectra should be acquired using a different instrument/laboratory than was used to acquire training spectra. (ii) In addition to reporting culture duration, MALDI matrix type, and other similar variables, published studies should report the experimental design of collected data, including biological and technical replication, randomization, and blocking (designed to account for sources of variation and confounding, such as sample location on the MALDI target plate, day of data collection, chemical/medium ingredient batches, etc.) (1). (iii) Published studies should make data available for public use in both raw and standard open format (e.g. mzML) (2) in a repository such as the Mass Spectrometry Interactive Virtual Environment data repository (MassIVE; https://massive.ucsd.edu) (3).

We strongly recommend that guidelines ii and iii be considered for all microbially based MALDI-TOF MS publications, not just method development studies. Also, in addition to the aforementioned guidelines, the microbiology community would benefit from standardized and validated benchmark data sets, analogous to the metagenomics CAMI data sets and standards (4). For benchmarking, we suggest

Address correspondence to Brian T. Murphy, btmurphy@uic.edu, or Laura M. Sanchez, sanchelm@uic.edu.

The views expressed in this article do not necessarily reflect the views of the journal or of ASM.

following a pattern of logic similar to that of Vervier et al. (5) and hope to see further participation of corporations in creating open standards and data (6).

Lastly, we urge caution when benchmarking methods that were trained/validated on different data sets. Assertions such as "exceeding the taxonomic resolution of other methods" (7) must be made with caution, as there are many appropriate approaches to dereplicate/differentiate microorganisms using MALDI-TOF MS. While many methods continue to focus solely on protein *m/z* regions, we previously developed a freely available, open-source MALDI-TOF MS-based pipeline (IDBac) to group bacterial isolates by protein MS spectra (2 to 20 kDa) in addition to specialized metabolite MS spectra (<2 kDa), allowing us to achieve rapid and accurate subspecies dereplication of environmental microbial isolates (8–10). Therefore, when attempting to achieve subspecies resolution of isolates, we opine that it is advantageous to characterize microorganisms using as many orthogonal methods (metabolomics, genomics, proteomics) as possible. It is an exciting time for MALDI-TOF MS analytical and technical innovation, and in our opinion, we have only scratched the surface of its usefulness for both proteomics and metabolomics. Looking forward, establishing global reference data sets in addition to community standards for data analysis, data sharing, and method comparison will result in more accurate assessments of our ability to distinguish microbial strains at the subspecies level.

REFERENCES

1. Oberg AL, Mahoney DW. 2012. Statistical methods for quantitative mass spectrometry proteomic experiments with labeling. BMC Bioinformatics 13(Suppl 16):S7. https://doi.org/10.1186/1471-2105-13-S16-S7.
2. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Römpp A, Neumann S, Pizarro AD, Montecchi-Palazzi L, Tasman N, Coleman M, Reisinger F, Souda P, Hermjakob H, Binz P-A, Deutsch EW. 2011. mzML—a community standard for mass spectrometry data. Mol Cell Proteomics 10:R110.000133. https://doi.org/10.1074/mcp.R110.000133.
3. Wang M, Wang J, Carver J, Pullman BS, Cha SW, Bandeira N. 2018. Assembling the community-scale discoverable human proteome. Cell Syst 7:412–421.e5. https://doi.org/10.1016/j.cels.2018.08.004.
4. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E, Bremges A, Fritz A, Garrido-Oter R, Jørgensen TS, Shapiro N, Blood PD, Gurevich A, Bai Y, Turaev D, DeMaere MZ, Chikhi R, Nagarajan N, Quince C, Meyer F, Balvočiūtė M, Hansen LH, Sørensen SJ, Chia BKH, Denis B, Froula JL, Wang Z, Egan R, Don Kang D, Cook JJ, Deltel C, Beckstette M, Lemaitre C, Peterlongo P, Rizk G, Lavenier D, Wu Y-W, Singer SW, Jain C, Strous M, Klingenberg H, Meinicke P, Barton MD, Lingner T, Lin H-H, Liao Y-C, et al. 2017. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. Nat Methods 14:1063–1071. https://doi.org/10.1038/nmeth.4458.
5. Vervier K, Mahé P, Veyrieras J-B, Vert J-P. 2015. Benchmark of structured machine learning methods for microbial identification from mass-spectrometry data. arXiv 1506.07251.
6. Mahé P, Arsac M, Chatellier S, Monnin V, Perrot N, Mailler S, Girard V, Ramjeet M, Surre J, Lacroix B, van Belkum A, Veyrieras J-B. 2014. Automatic identification of mixed bacterial species fingerprints in a MALDI-TOF mass-spectrum. Bioinformatics 30:1280–1286. https://doi.org/10.1093/bioinformatics/btu022.
7. Dumolin C, Aerts M, Verheyde B, Schellaert S, Vandamme T, Van der Jeugt F, De Canck E, Cnockaert M, Wieme AD, Cleenwerck I, Peiren J, Dawyndt P, Vandamme P, Carlier A. 2019. Introducing SPeDE: high-throughput dereplication, and accurate determination of microbial diversity from matrix-assisted laser desorption–ionization time of flight mass spectrometry data. mSystems 4:e00437-19. https://doi.org/10.1128/mSystems.00437-19.
8. Clark CM, Costa MS, Sanchez LM, Murphy BT. 2018. Coupling MALDI-TOF mass spectrometry protein and specialized metabolite analyses to rapidly discriminate bacterial function. Proc Natl Acad Sci U S A 115:4981–4986. https://doi.org/10.1073/pnas.1801247115.
9. Clark CM, Costa MS, Conley E, Li E, Sanchez LM, Murphy BT. 2019. Using the open-source MALDI TOF-MS IDBac pipeline for analysis of microbial protein and specialized metabolite data. J Vis Exp 2019:e59219. https://doi.org/10.3791/59219.
10. Costa MS, Clark CM, Ómarsdóttir S, Sanchez LM, Murphy BT. 2019. Minimizing taxonomic and natural product redundancy in microbial libraries using MALDI-TOF MS and the bioinformatics pipeline IDBac. J Nat Prod 82:2167–2173. https://doi.org/10.1021/acs.jnatprod.9b00168.