

[Short Communication] DiagnoTop: A Computational Pipeline for Discriminating Bacterial Pathogens without Database Search

Diogo Borges Lima, Mathieu Dupré, Marlon Dias Mariano Santos, Paulo Costa Carvalho, Julia Chamot-Rooke

▶ **To cite this version:**

Diogo Borges Lima, Mathieu Dupré, Marlon Dias Mariano Santos, Paulo Costa Carvalho, Julia Chamot-Rooke. [Short Communication] DiagnoTop: A Computational Pipeline for Discriminating Bacterial Pathogens without Database Search. Journal of The American Society for Mass Spectrometry, 2021, 32 (6), pp.1295-1299. 10.1021/jasms.1c00014 . pasteur-03441711

HAL Id: pasteur-03441711
https://pasteur.hal.science/pasteur-03441711

Submitted on 23 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

ACS Partner Journal

Journal of the American Society for Mass Spectrometry

pubs.acs.org/jasms

Short Communication

DiagnoTop: A Computational Pipeline for Discriminating Bacterial Pathogens without Database Search

Diogo Borges Lima,[¶] Mathieu Dupré,[¶] Marlon Dias Mariano Santos, Paulo Costa Carvalho,* and Julia Chamot-Rooke*

Cite This: *J. Am. Soc. Mass Spectrom.* 2021, 32, 1295−1299

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Pathogen identification is crucial to confirm bacterial infections and guide antimicrobial therapy. Although MALDI-TOF mass spectrometry (MS) serves as foundation for tools that enable rapid microbial identification, some bacteria remain challenging to identify. We recently showed that top-down proteomics (TDP) could be used to discriminate closely related enterobacterial pathogens (*Escherichia coli*, *Shigella*, and *Salmonella*) that are indistinguishable with tools rooted in the MALDI-TOF MS approach. Current TDP diagnostic relies on the identification of specific proteoforms for each species through a database search. However, microbial proteomes are often poorly annotated, which complicates the large-scale identification of proteoforms and leads to many unidentified high-quality mass spectra. Here, we describe a new computational pipeline called DiagnoTop that lists discriminative spectral clusters found in TDP data sets that can be used for microbial diagnostics without database search. Applied to our enterobacterial TDP data sets, DiagnoTop could easily shortlist high-quality discriminative spectral clusters, leading to increased diagnostic power. This pipeline opens new perspectives in clinical microbiology and biomarker discovery using TDP.



KEYWORDS: *top-down proteomics, enterobacterial pathogens, clinical microbiology, diagnostics*

Downloaded via INST PASTEUR DE PARIS on November 23, 2021 at 09:57:33 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

The current method used in hospitals for rapid bacterial identification rely on computational tools tailored toward matrix-assisted laser desorption ionization−mass spectrometry (MALDI-TOF MS).[1] The diagnostic is based on comparing an experimental spectrum obtained in the 2000−20 000 Da range generated from intact proteins from a bacterial colony and profiles of known bacteria stored in a database.[2] Although the approach can characterize bacterial pathogens in more than 80% of the cases, it fails when the database lacks appropriate profiles or when challenged with closely related pathogens that give similar protein patterns. Moreover, it does not bring any information on the presence of virulence of resistance determinants. There is, therefore, a need for new strategies that overcome these limitations. We recently demonstrated that top-down proteomics (TDP) could be successfully employed to discriminate closely related pathogenic bacterial strains such as the enterobacteria *Escherichia coli*, *Shigella*, and *Salmonella*.[3] Our approach relied on LC-MS/MS analysis of intact proteins obtained from bacterial cultures and the subsequent identification of specific discriminative proteoforms. Identification was achieved using Prosight, a widely adopted database search engine,[4] and a proteoform database generated from Uniprot. We noted that the use of Uniprot was biased because of the high variability in the quality of the available microbial proteomes. For reference species, such as *E. coli* K12, many reviewed protein sequences are accessible and allowed reliable identification of proteoforms. In contrast, the databases available for poorly studied microbes contained mostly nonreviewed, missing, or misannotated sequences. In addition, these protein databases usually contain only a little or no information on amino acid variations or modifications of proteins. These shortcomings justify why many high-quality MS/MS spectra remain unidentified and opens room for new tools to significantly increase the repertoire of discriminant proteoforms.

To circumvent this limitation, we present a new computational pipeline, called DiagnoTop, that uses spectral clustering and pattern recognition to statistically list discriminant spectral clusters based on MS/MS data, without database search. Our pipeline relies on Top-Down Garbage Collector (TDGC) to serve as a gatekeeper and guarantee a high-quality spectra feed to DiagnoProt,[5] a software previously introduced by us, that shortlists discriminative spectral clusters[6] and tailored here for the analysis of TDP data. Applied to our enterobacterial TDP data set, DiagnoTop achieves a higher number of discrim-

Special Issue: Focus: Top-Down Proteomics: Technology Advances and Biomedical Applications

Received: January 13, 2021
Revised: April 8, 2021
Accepted: April 8, 2021
Published: April 15, 2021



ACS Publications

© 2021 American Society for Mass Spectrometry. Published by American Chemical Society. All rights reserved.

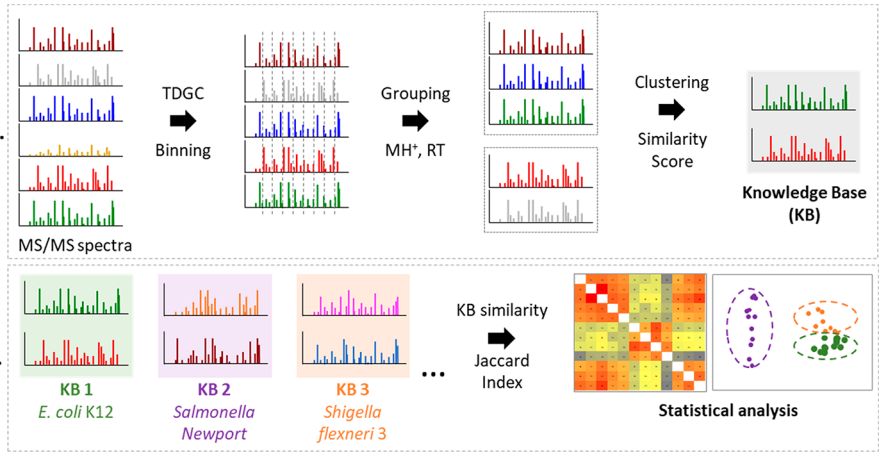


Figure 1. DiagnoTop workflow, including (A) spectral clustering to create the 12 knowledge bases (KB) and (B) comparison of all KBs.

inative features than the widely adopted search engine approach.

EXPERIMENTAL SECTION

Computer Setup. DiagnoTop is composed of TDGC and DiagnoProt; both components are programmed in C# and require .NET Framework 4.8 to be installed in a computer with at least 16 GB RAM and Windows 10. TDGC is available at http://patternlabforproteomics.org/tdgc/, and DiagnoProt at http://patternlabforproteomics.org/diagnotop/. Our software support MS2, MGF format files, and can read directly from Thermo RAW files. Our pipeline can handle various activation approaches, e.g., Higher-Energy C-trap Dissociation (HCD), Electron Transfer Dissociation (ETD), and ETD combined to HCD (EThcD).

TDP Data Set. We used DiagnoTop on a TDP data set previously published by us[3] and available in ProteomeXchange[7] (PXD019247). Briefly, 12 enterobacterial strains from the Collection of Institut Pasteur were investigated: four *Salmonella enterica enterica* (serotype Enteritidis, Typhimurium, Newport, and Muenchen), one *Shigella sonnei*, two *Shigella flexneri* (serotypes 2a and 3), and five *E. coli* (O157:H7, O157:H7 with Shiga-toxin 1 (stx1), and Shiga-toxin-2 genes, O26:H11 with stx1 and eae genes, O26:H11 with eae gene, and MG155 K12). All bacteria were cultured overnight and subcultured for ∼4 h at 37 °C in LB medium to be harvested at late exponential growth phase. Cell lysis was performed in PBS (1×) with protease inhibitors (PMSF, 1 mM; EDTA, 1 mM) by mechanical disruption.

LC analyses were conducted on a Dionex Ultimate 3000 RSLC nanosystem (Thermo-Scientific) using a 150 min LC gradient and in-house packed C4 columns (5 μm porous spherical particles of 300 Å pore size, Reprosil). MS analyses were performed with an Orbitrap Fusion Lumos mass spectrometer fitted with a nano-electrospray ionization source using the Intact Protein Mode (2 mTorr). The MS method includes full MS scans acquired at 60K resolving power (at m/z 400) with a scan range set to 500−1750 m/z, two μscans per MS scan, an AGC target value of 5×10^5, and maximum injection time (MIT) of 50 ms. Top 4 ions with an intensity threshold $>1 \times 10^5$ were isolated with 1.2 m/z width, fragmented with EThcD (10 ms, 10%) and then added to a dynamic exclusion window for 60 s. MS/MS scans were acquired at 60K resolving power, with 2 μscans, an AGC target value of 5×10^5, and MIT of 250 ms. All experiments were performed in biological triplicates.

The 36 corresponding raw files were processed with ProSight PC v4.1 (Thermo Scientific) and Proteome Discoverer v2.4 (Thermo Scientific) using the ProSight PD 3.0 node. Spectral data were first deconvoluted and deisotoped using the cRAWler algorithm. Spectra were then searched using a two-tier search strategy with searches against an Uniprot XML database created from the sequences identified in the individual TDP analysis of the 12 strains after removing duplicates (1516 protein entries and 10 425 proteoforms). Search 1 consists of a ProSight Absolute Mass search, and Search 2 is a ProSight Biomarker search (tolerance of 10 and 5 ppm for MS 1 and MS2, respectively). Identifications with *E*-values better than 1×10^{-10} (−log (*E*-value) = 10) and between 1×10^{-10} and 1×10^{-5} were respectively considered confident and medium hits. A 1% proteoform spectrum match (PrSM)-level FDR was employed.

DiagnoTop Workflow. Figure 1 summarizes the DiagnoTop workflow. The first step was to use the quality control filter (TDGC) on all MS/MS spectra to select the high-quality ones. TDGC uses an automatic scoring function that considers the distribution of isotopic envelopes and signal-to-noise ratios throughout the spectrum to classify whether it is of high-quality or not. We used TDGC with its default parameters. For the task at hand, TDGC also helped speed up the processing and increase clustering precision (data not shown).

In what followed, DiagnoTop was used to generate the Knowledge Base (KB). We define KB as a collection of spectral clusters derived from the various bacterial strains. We define a spectral cluster as a set of similar MS/MS spectra grouped according to a similarity function, in our case, the spectral angle.[8,9] To compute the spectral angle, the spectra were first vectorized using bins of 0.02 m/z width. Only spectra whose precursor ion retention time was within a 10 min tolerance and the precursor mass differ at most by 3.5 Da were considered for belonging to the same cluster. This aimed to group all similar MS/MS spectra arising from the fragmentation of the same proteoform.

The second part of DiagnoTop includes evaluating the separability between bacterial strains. For this, we used the Jaccard index with the information on number of shared clusters for each condition; these results can be used to generate a PCA. We also provide a clustergram that displays

https://doi.org/10.1021/jasms.1c00014
J. Am. Soc. Mass Spectrom. 2021, 32, 1295−1299

the number of similar spectral clusters between each pairwise comparison and a dendrogram for the bacterial strains.

RESULTS AND DISCUSSION

We processed the 36 raw files obtained for the TDP analysis of our 12 enterobacteria using the DiagnoTop pipeline described above. We determined the optimal spectral angle threshold for clustering by repeating the DiagnoTop routine by varying the spectral angle from 0.25 to 0.60 with a 0.05 step. We selected the spectral angle that maximized the Bhattacharyya distance (BD)[10] applied to the set of points resultants from a 3D PCA mapping for different bacteria. We recall that the BD is a measure of the dissimilarity (or similarity) between different probability distributions that measures the relative closeness of the two samples. Figure 2 shows the BD resulting from the various spectral angles verified and suggests 0.4 as optimal. We hypothesize 0.4 should be adequate for most TDP data set using a similar setup as ours.

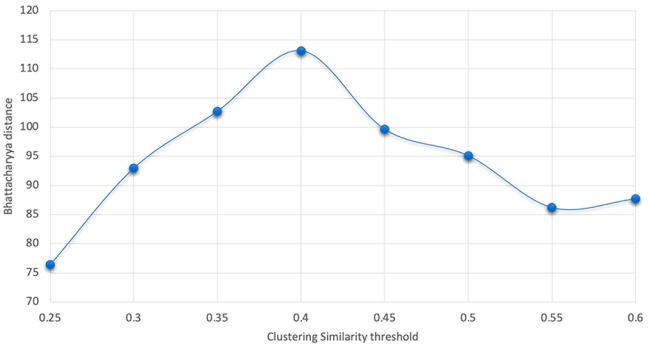


Figure 2. Bhattacharyya distance versus spectral clustering angle when comparing *E. coli* versus *Salmonella* TDP data.

TDGC shortlisted 82 878 high-quality spectra when considering all our mass spectrometry runs. Table 1 summarizes the number of discriminant spectral clusters obtained using DiagnoTop and the discriminant proteoforms identified by ProsightPD 3.0. For example, Table 1 indicates that *Salmonella* enteritidis contains 207 unique spectral clusters in at least two replicates, i.e., spectra found in two or more biological replicates of *Salmonella* enteritidis and in no other strain. We considered as discriminant only spectral clusters (DiagnoTop) or proteoform identifications (ProsightPD) found exclusively for a bacterial strain. As shown in Table 1, it is clear that DiagnoTop overperforms ProsightPD, whatever the conditions used, for the number of discriminant features. The DiagnoTop computing time was 6h30 while Prosight PD took approximately 10 days to create the proteoform databases and perform the searches.

Figure 3 shows a clustergram generated considering the discriminative spectral cluster provided by DiagnoTop as input to HeatMapper[11] using Kendall's Tau distance. Figure 3 shows that strains belonging to the same species cluster together, with *Salmonella* clustering further than *E. coli* and *Shigella*, as expected. Dupre et al.'s[3] dendrogram, generated only by considering Prosight, shows the same disposition as ours achieved with DiagnoTop. Figure 4 shows a PCA generated with DiagnoTop's data that corroborates with the clustergram.

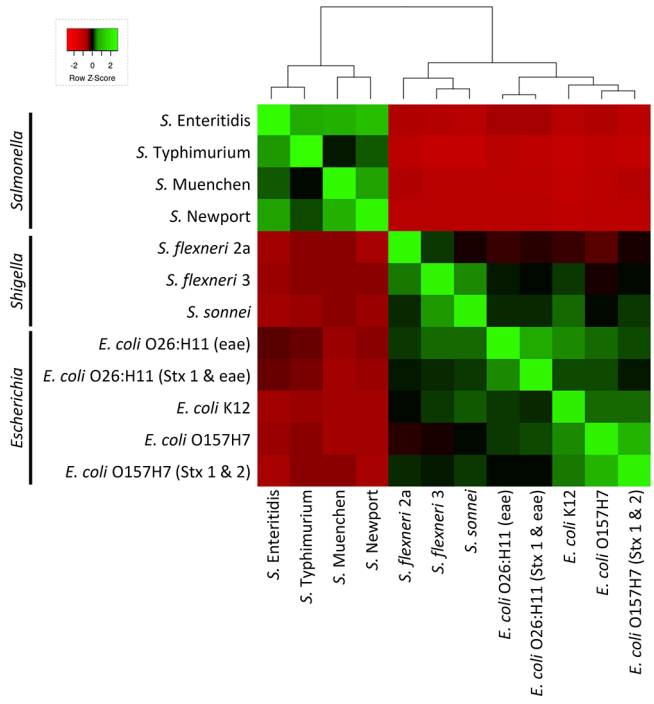


Figure 3. Clustergram generated using DiagnoTop results and HeatMapper.

Table 1. DiagnoTop and Prosight PD result summary

	DiagnoTop						prosight PD			
	1 replicate		2 replicates		3 replicates			Exclusive proteoforms		
strains	exclusive	total	exclusive	total	exclusive	total	total proteoforms	1 replicate	2 replicates	3 replicates
Salmonella enteritidis	1537	3151	207	1821	75	1437	1190	521	53	10
Salmonella typhimurium	2091	3319	544	1772	244	1255	863	375	76	15
Salmonella	1017	2280	146	1409	64	1165	578	139	31	12
Salmonella Muenchen Newport	1097	2197	145	1245	58	1027	473	147	18	6
Shigella sonnei	1007	2316	96	1405	44	1221	432	84	26	11
Shigella flexneri 2a	1168	2306	142	1280	73	1080	532	189	21	7
Shigella flexneri 3	1037	2301	78	1342	30	1149	430	77	17	4
E. coli O157H7 (Stx 1 and 2)	992	2318	101	1427	36	1256	454	58	16	6
E. coli O157H7	784	2178	78	1472	25	1281	604	105	17	8
E. coli O26:H11 (Stx 1 and eae)	983	2390	141	1548	65	1307	499	97	12	2
E. coli O26:H11 (eae)	1157	2666	187	1696	82	1425	644	175	40	10
E. coli MG1655 (K12)	689	1944	61	1316	24	1180	478	82	22	13

https://doi.org/10.1021/jasms.1c00014
J. Am. Soc. Mass Spectrom. 2021, 32, 1295−1299

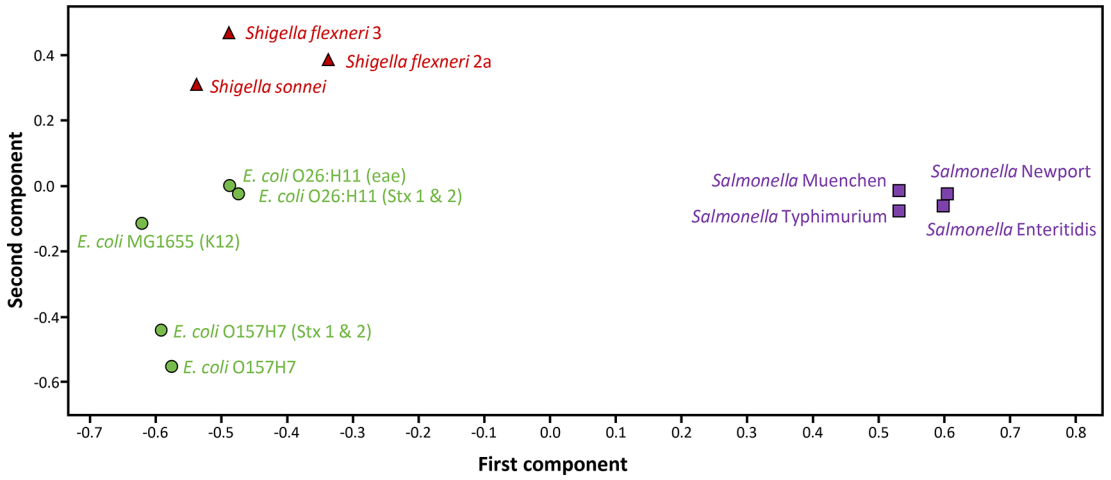


Figure 4. Principal component analysis generated with DiagnoTop.

CONCLUSION

We introduced a computational pipeline named DiagnoTop that compares large TDP data sets and identifies discriminative spectral clusters without a database search. Applied to the analysis of 12 enterobacterial strains that cannot be differentiated with MALDI-TOF MS, DiagnoTop converged, in all cases, to a high number of exclusive spectral clusters that can be used to discriminate all strains from each other. We show that, in significantly less time, we reported many more discriminative spectral clusters than Prosight PD, a widely adopted commercial solution to analyze TDP data and characterization of proteoforms. DiagnoTop paves the way to classifying unknown bacterial strains (many times lacking sequence databases) by comparing discriminative spectral clusters to those previously stored in a spectral database. Thus, we believe that DiagnoTop holds great promise in the future for clinical microbiology applications.

AUTHOR INFORMATION

Corresponding Authors

Julia Chamot-Rooke − *Mass Spectrometry for Biology Unit, CNRS USR2000, Institut Pasteur, Paris 75015, France;* orcid.org/0000-0002-9427-543X; Email: julia.chamot-rooke@pasteur.fr

Paulo Costa Carvalho − *Laboratory for Structural and Computational Proteomics, Carlos Chagas Institute Fiocruz, Paraná, Curitiba CIC 81350-010, Brazil;* orcid.org/0000-0001-6530-3350; Email: paulo@pcarvalho.com

Authors

Diogo Borges Lima − *Mass Spectrometry for Biology Unit, CNRS USR2000, Institut Pasteur, Paris 75015, France;* orcid.org/0000-0001-6056-0825

Mathieu Dupré − *Mass Spectrometry for Biology Unit, CNRS USR2000, Institut Pasteur, Paris 75015, France;* orcid.org/0000-0002-1845-0048

Marlon Dias Mariano Santos − *Laboratory for Structural and Computational Proteomics, Carlos Chagas Institute Fiocruz, Paraná, Curitiba CIC 81350-010, Brazil;* orcid.org/0000-0002-1178-1266

Complete contact information is available at: https://pubs.acs.org/10.1021/jasms.1c00014

Author Contributions

¶D.B.L. and M.D. contributed equally. D.B.L., P.C.C., and M.D.M.S. are developers of DiagnoTop. M.D. generated the mass spectra data and the ProsightPD analysis. All authors participated in the data analysis, wrote the manuscript, and approved the final version of the text.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was financially supported by the Institut Pasteur, the CNRS, the Agence Nationale de la Recherche (PathoTOP project ANR-15-CE18-0021), CNPq, Inova Fiocruz, and CAPES. It has also received funding from the European Union's Horizon 2020 Research and Innovation Programme through the European Joint Programme One Health EJP under Grant Agreement no. 773830 and the European EPIC-XS project no. 823839.

REFERENCES

(1) Demirev, P. A.; Fenselau, C. Mass spectrometry for rapid characterization of microorganisms. *Annu. Rev. Anal. Chem.* **2008**, *1*, 71−93.

(2) Welker, M. Proteomics for routine identification of microorganisms. *Proteomics* **2011**, *11* (15), 3143−3153.

(3) Dupre, M.; Duchateau, M.; Malosse, C.; Borges-Lima, D.; Calvaresi, V.; Podglajen, I.; Clermont, D.; Rey, M.; Chamot-Rooke, J. Optimization of a Top-Down Proteomics Platform for Closely Related Pathogenic Bacterial Discrimination. *J. Proteome Res.* **2021**, *20* (1), 202−211.

(4) Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y. B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.* **2007**, *35*, W701−706.

(5) Silva, A. R. F.; Lima, D. B.; Leyva, A.; Duran, R.; Batthyany, C.; Aquino, P. F.; Leal, J. C.; Rodriguez, J. E.; Domont, G. B.; Santos, M. D. M.; Chamot-Rooke, J.; Barbosa, V. C.; Carvalho, P. C. DiagnoProt: a tool for discovery of new molecules by mass spectrometry. *Bioinformatics* **2017**, *33* (12), 1883−1885.

(6) Lima, D. B.; Silva, A. R. F.; Dupre, M.; Santos, M. D. M.; Clasen, M. A.; Kurt, L. U.; Aquino, P. F.; Barbosa, V. C.; Carvalho, P. C.; Chamot-Rooke, J. Top-Down Garbage Collector: a tool for selecting high-quality top-down proteomics mass spectra. *Bioinformatics* **2019**, *35* (18), 3489−3490.

(7) Perez-Riverol, Y.; Alpi, E.; Wang, R.; Hermjakob, H.; Vizcaino, J. A. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics* **2015**, *15* (5−6), 930−949.
(8) Carvalho, P. C.; Xu, T.; Han, X.; Cociorva, D.; Barbosa, V. C.; Yates, J. R., III YADA: a tool for taking the most out of high-resolution spectra. *Bioinformatics* **2009**, *25* (20), 2734−2736.
(9) Bandeira, N. Spectral networks: a new approach to de novo discovery of protein sequences and posttranslational modifications. *BioTechniques* **2007**, *42* (6), 687−689.
(10) Bhattacharyya, A. On a Measure of Divergence between Two Statistical Populations Defined by Their Probability Distributions. *Bulletin of the Calcutta Mathematical Society* **1943**, *35*, 99−109.
(11) Babicki, S.; Arndt, D.; Marcu, A.; Liang, Y.; Grant, J. R.; Maciejewski, A.; Wishart, D. S. Heatmapper: web-enabled heat mapping for all. *Nucleic Acids Res.* **2016**, *44* (W1), W147−153.

https://doi.org/10.1021/jasms.1c00014
J. Am. Soc. Mass Spectrom. 2021, 32, 1295−1299