analytical chemistry

Article

Cite This: *Anal. Chem.* 2019, 91, 11482−11487

pubs.acs.org/ac

Model-Based Spectral Library Approach for Bacterial Identification via Membrane Glycolipids

So Young Ryu,*[,†] George A. Wendt,[†,‡] Courtney E. Chandler,[§] Robert K. Ernst,[§] and David R. Goodlett[§,∥,#]

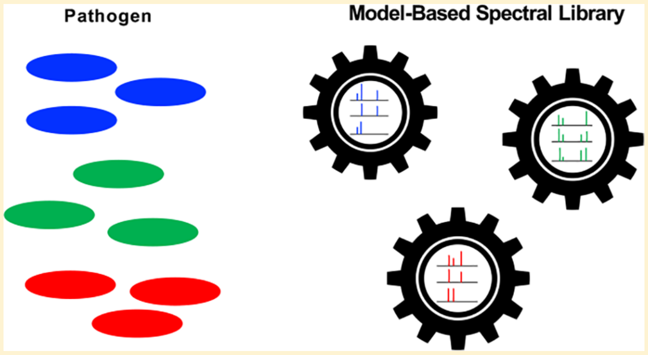
[†]School of Community Health Sciences, University of Nevada Reno, Reno, Nevada 89557, United States

[‡]Department of Epidemiology, School of Public Health, University of California Berkeley, Berkeley, California 94720, United States

[§]Department of Microbial Pathogenesis, School of Dentistry, University of Maryland, Baltimore, Maryland 21201, United States

[∥]International Centre for Cancer Vaccine Science, University of Gdansk, 80-308 Gdansk, Poland

Supporting Information



ABSTRACT: By circumventing the need for a pure colony, MALDI-TOF mass spectrometry of bacterial membrane glycolipids (lipid A) has the potential to identify microbes more rapidly than protein-based methods. However, currently available bioinformatics algorithms (e.g., dot products) do not work well with glycolipid mass spectra such as those produced by lipid A, the membrane anchor of lipopolysaccharide. To address this issue, we propose a spectral library approach coupled with a machine learning technique to more accurately identify microbes. Here, we demonstrate the performance of the model-based spectral library approach for microbial identification using approximately a thousand mass spectra collected from multi-drug-resistant bacteria. At false discovery rates < 1%, our approach identified many more bacterial species than the existing approaches such as the Bruker Biotyper and characterized over 97% of their phenotypes accurately. As the diversity in our glycolipid mass spectral library increases, we anticipate that it will provide valuable information to more rapidly treat infected patients.

Despite public health efforts to combat antimicrobial resistance, challenges remain in bacterial identification, in particular, related to organisms that are antimicrobial resistant.[1] To better address this problem and in turn more effectively control the spread of infectious diseases, it is essential to develop accurate, affordable, and timely diagnostic tools.[2] Profiling the Gram-negative glycolipid lipid A (and other bacterial membrane glycolipids from Gram-positive bacteria) by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) is a candidate for such a rapid and low-cost diagnostic tool.[3] Lipid A is the primary immunostimulatory component of lipopolysaccharide (LPS) and responsible for the toxicity of Gram-negative bacteria. Due to their diversity between species/phenotypes in the arrangement of fatty acyl side chains and sugar-associated functional groups, mass spectra generated from lipid A contain information to identify and characterize Gram-negative bacteria.[4] Circumventing the need for biological culture to produce a pure colony allows this glycolipid approach to be much faster and cheaper than currently used pathogen detection methods (e.g., morphological/biochemical method), as well as the protein-based MALDI-TOF MS approach.[5−10] The mass spectrometry approach of Leung et al.,[3] which is based on profiling bacterial glycolipids such as lipid A from Gram-negative microbes and related molecules from Gram-positive microbes, is also more cost-effective than the next generation sequencing approach that analyzes whole bacterial genomes.[11]

Bioinformatics tools exist that analyze MALDI-TOF MS protein-based mass spectra. For example, U.S. Food and Drug Administration (FDA) approved software such as Biotyper from Bruker Daltonics and Spectral Archive and Microbial Identification System (SARAMICS) from bioMérieux are currently used in hospital clinical laboratories.[7] Recognizing that these tools cannot differentiate closely related bacterial species (e.g., *Bacillus cereus* group),[12] new measures of spectral similarity and a statistical assessment of such identifications have been proposed. However, these tools are developed for information-rich protein-based MALDI-TOF MS data, not the glycolipid mass spectra which contain fewer peaks that are unique to species. To fully utilize glycolipid mass spectra for bacterial identification, it is essential to develop bioinformatic tools specific to glycolipid mass spectra such as those produced

Received: July 23, 2019
Accepted: August 1, 2019
Published: August 1, 2019

DOI: 10.1021/acs.analchem.9b03340

by lipid A and related Gram-positive molecules including lipoteichoic acid and cardiolipin.

Constructing meaningful theoretical lipid A mass spectra with reasonable complexity is very challenging. Wilson et al.[13] showed that a Cartesian product algorithm based on membrane glycolipid structure can, in theory, produce >2 billion molecular masses from the lipid A scaffold. However, we observe far fewer meaningful masses representing unique structures in real lipid A mass spectra. Here, we propose a spectral library approach that utilizes mass spectra generated by known lipid A structures and related glycolipids from Gram-positive bacteria. Since our algorithm is based on acquired data, we can develop an algorithm that reflects the stochastic nature of bacterial glycolipid ions. Such a spectral library concept has been previously used in proteomics. There, a peptide sequence is determined from a tandem mass spectrum by searching against previously assigned tandem mass spectral libraries of peptides.[14−16] The popular scoring approach used in this area involves several variations of dot product analysis. However, due to the lack of many meaningful masses representing unique structures of species in these glycolipid mass spectra, the traditional dot product approaches did not work well in analyzing membrane glycolipid-based mass spectra. Our previous work also shows that it is more suitable to use a machine learning technique for glycolipid mass spectra identification.[4]

In this work, we propose a model-based spectral library approach[17] for matching glycolipid mass spectra that we refer to as lipid A spectral library (LASL). Different from previously proposed spectral library approaches, LASL contains bacteria identification models instead of mass spectra or representative mass spectra. The machine learning model can select key ions in glycolipid mass spectra during its training runs. Thus, it can work better in identifying glycolipid mass spectra than algorithms designed for protein mass spectra. By using a model-based approach, LASL is complex enough to capture the apparent stochastic nature of glycolipid mass spectra better than using only one representative mass spectrum per bacteria.

Here, we first introduce LASL as a model-based spectral library approach and then discuss measures of uncertainty of bacterial identifications. Then, we demonstrate the performance of LASL using nearly a thousand glicolipid mass spectra. Finally, we discuss the limitations and potential of our approach.

MATERIALS AND METHODS

Data. For our analysis, we used the glycolipid mass spectral data set published by Leung et al.[3] that contained 906 mass spectra from various strains of six microbial species. We consider these 906 glycolipid mass spectra as a main data set. These mass spectra were generated by negative ion MALDI-TOF-MS analysis. In short, samples were grown in liquid culture and lipids isolated using the hot ammonium isobutyrate described by El Hamidi et al.[18] after which they were analyzed by MALDI-TOF MS in the negative ion mode. The data set included 404 mass spectra of *Acinetobacter baumannii* (AB, *A. baumannii*), 79 from *Enterobacter cloacae* (EC), 55 from *Enterococcus faecalis* (EF), 207 from *Klebsiella pneumoniae* (KP, *K. pneumoniae*), 78 from *Pseudomonas aeruginosa* (PA), and 83 from *Staphylococcus aureus* (SA). There were two phenotypes available in the data set: colistin-susceptible (cs) and colistin-resistant (cr). Colistin (also known as polymyxin E) is used as a major antibiotic for fighting Gram-negative infections. Since colistin is the last resort to treat patients infected by multi-drug-resistant bacteria (e.g., multi-drug-resistant *A. baumannii*[19]), the ability to accurately detect colistin-resistant bacteria and monitor their presence is essential. We denote colistin-susceptible *A. baumannii* and colistin-resistant *A. baumannii* as ABcs and ABcr, respectively. Similarly, we denote colistin-susceptible *K. pneumoniae* and colistin-resistant *K. pneumoniae* as KPcs and KPcr, respectively. Besides this main data set, we had a supplementary data set of four lipid A mass spectra generated from the following bacteria: *Clostridium difficile*, *Legionella bozemanii*, *Salmonella typhimurium*, and *Yersinia pseudotuberculosis*.

All mass spectra were converted to mzXML format using msconvert (v3.0.9393 ProteoWizard) and then processed using the MALDIquant (v1.16.2) and MALDIquantForeign (v0.10) R packages.[20] Specifically, the mass spectra were square-root-transformed and smoothed using a Savitzky−Golay filter.[21] Then, the baselines of mass spectra were corrected using the statistics-sensitive nonlinear iterative peak-clipping (SNIP) algorithm,[22] and peak intensities in mass spectra were normalized by their total ion current. The top K peaks were selected for the further analysis, where $K = 50$.[4] Then, we binned peaks by their mass-to-charge ratios with their bin sizes of 1 Da. The highest peak in each bin was selected. Their masses, (normalized) intensities, and ranks of intensities (across bins) were recorded.

Finally, we created decoy mass spectra, which did not belong to any species. Only a training set from the main data set was used for decoy spectra construction. For bacterial identification, two sets of decoy spectra were constructed. One set was used to train the model (N = 1,500), and another was used to test the model performance and measure false discovery rates (N = 10,000). Decoy mass spectra were created by extracting K (e.g., K = 50) random peaks from M mass spectra and randomly permuting their intensities. For example, if the K sampled peaks are expressed as (mz_1, $\mathrm{intensity}_1$), (mz_2, $\mathrm{intensity}_2$), ..., (mz_K, $\mathrm{intensity}_K$), then one example decoy spectrum can contain the following peaks: (mz_1, $\mathrm{intensity}_{30}$), (mz_2, $\mathrm{intensity}_{14}$), (mz_3, $\mathrm{intensity}_{11}$), ..., (mz_K, $\mathrm{intensity}_2$), where the original intensity values are mismatched with their m/z values. Our model performance was not too sensitive to the choice of M as long as M was not too small (e.g., M = 1). In this work, for each decoy spectrum, we randomly chose M to be an integer between 5 and 10. For the same purpose, we also constructed two sets of decoy mass spectra for AB phenotype identification and another two sets for KP phenotype identification.

Model-Based Spectral Library. The main data set was divided into test and train sets in a ratio of 2:1. For each set, we added decoy mass spectra, which did not belong to any species. Specifically, we added 1,500 mass spectra in the training set and 10,000 in the testing set. Adding decoy mass spectra in the training set improved the model performance, allowing identification of the correct species with higher confidence. Decoy mass spectra in the testing set did not overlap with ones in the training set, but were used to estimate p-values and false discovery rates.

Mass spectra in the training set were used to construct a model-based spectral library. We built bacteria/phenotype identification models using eXtreme gradient boosting (XGboost) with a logistic regression (binary classification) option.[23] One model was built for each microbial species. We treated mass spectra from bacteria of interest as positive cases

DOI: 10.1021/acs.analchem.9b03340
Anal. Chem. 2019, 91, 11482−11487

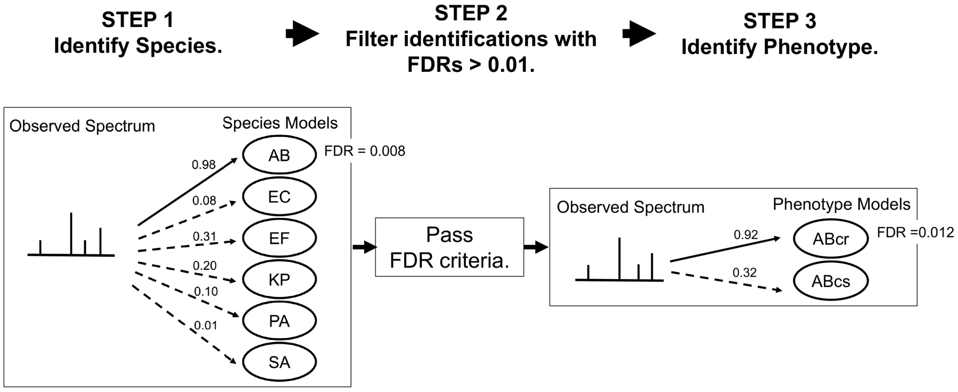


Figure 1. General workflow of LASL.

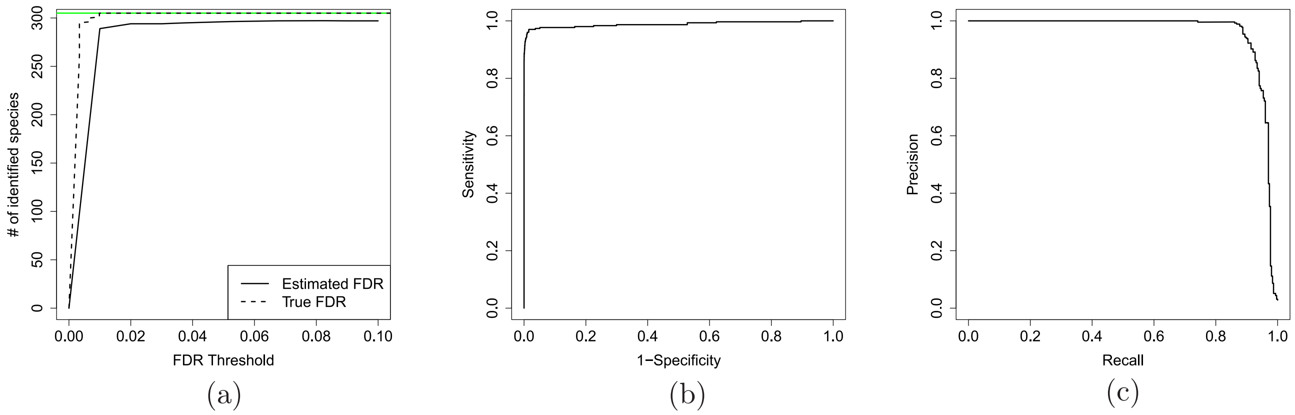


Figure 2. LASL performance in species identifications. (a) Estimated FDR threshold vs the number of identified species plot. The horizontal green line was the number of (nondecoy) mass spectra in the test set. The dotted line was based on the true FDR (tFDR) threshold. (b) Receiver operating characteristic (ROC) curve. (c) Precision-recall (PR) curve.

and mass spectra from other species and decoy mass spectra as negative cases. A total of six bacterial identification models were constructed. Similarly, two phenotype models also were built for AB and two other phenotype models were built for KP. The best tuning parameters for bacteria/phenotype models were selected using the 5-fold cross-validation and the grid search (see details about tuning parameters in Supporting Information).

Bacteria/Phenotype Identification. The general framework of bacteria/phenotype identification is displayed in Figure 1. Given a glycolipid mass spectrum, we first identified a bacterial species. If the bacterium was identified with high confidence (e.g., FDR < 0.01) and its phenotype models were available in the spectral library, we identified its phenotype. In detail, in step 1, we measured a predicted probability, p_b, that a given mass spectrum was from a microbial species b, where b represented species in the spectral library. In our setting, $\sum_{b \in \mathrm{SL}} p_b$ was not equal to one, where SL was a set of all the species in the spectral library because we chose not to use m-group classification models where $m > 2$. Noting that, in practice, a given mass spectrum may not be from microbial species in the spectral library, we intentionally added decoy mass spectra in the training set and used p_b as mere scores to choose the best species models. We called p_b as matching scores for the rest of the work. After a matching score of the given mass spectrum for each species model was estimated, the species with the highest matching score was assigned to the mass spectrum as a bacterial identification. We denote the top matching score as p_{b*}.

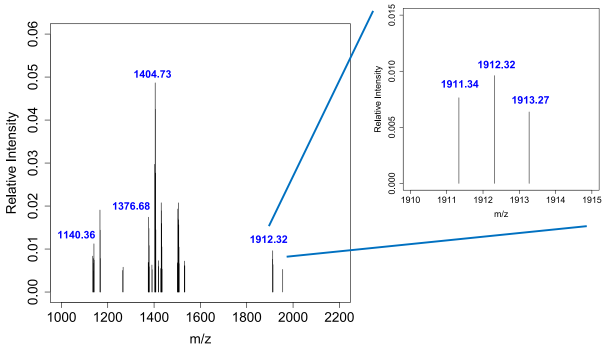
In step 2, we measured the uncertainty of bacterial species identifications. We note that the spectral library may not contain a microbial species of interest. Even when the library contains such a species, misidentifications can occur. Since it will be important to be certain about bacterial species identifications made from patients with infections, we calculated p-values and the corresponding false discovery rates (FDRs) for the bacterial species identifications and discarded identifications with FDRs > 0.01. The p-values were estimated using 10,000 decoy mass spectra in the test set:

$$p\text{-value} \approx \frac{\sum_{i=1}^{N_d} I(p_{d_i} > p_{b*})}{N_d} \quad (1)$$

where d represents the decoy mass spectrum, N_d is the number of decoy mass spectra, I is the indicator variable, p_{d_i} is the top matching score of the ith decoy mass spectrum, p_{b*} is the top matching score of an observed (nondecoy) mass spectrum. In other words, the p-value was calculated by dividing the number of decoy mass spectra with their top matching score greater than the top matching score of a given nondecoy spectrum by the total number of decoy mass spectra. The false discovery rates were estimated to correct multiple testing errors.[24]

For the glycolipid mass spectra identified as either AB or KP with high confidence (FDRs < 0.01), we identified their phenotypes in step 3 (Figure 1). We note that once

DOI: 10.1021/acs.analchem.9b03340
Anal. Chem. 2019, 91, 11482−11487



(a)

Feature	Gain	Coverage	Frequency
intensity @ 1911	0.328	0.141	0.041
intensity @ 1912	0.157	0.104	0.041
intensity @ 1405	0.084	0.077	0.053
m/z @ 1139	0.047	0.038	0.012
intensity @ 1729	0.041	0.064	0.041
rank intensity @ 1377	0.030	0.031	0.029
rank intensity @ 1912	0.028	0.036	0.018
intensity @ 1140	0.023	0.023	0.018
m/z @ 1367	0.017	0.044	0.029
intensity @ 1404	0.016	0.026	0.059

(b)

Figure 3. (a) Example mass spectrum for AB. The *m*/*z* values that were related to the top 10 important features were displayed. The zoomed spectrum was also shown. (b) Top 10 important features in the AB model. The intensity, the *m*/*z* value, and the rank of the intensity of the highest peak in each *m*/*z* bin were features considered to construct the species model.

phenotypes for other species become available, similar procedures can be incorporated. Similar to species identifications, the given mass spectrum was matched to the available phenotype identification models and their matching scores were calculated. A phenotype with the top matching score was assigned to a given mass spectrum. The corresponding *p*-value and FDR were estimated.

■ RESULTS AND DISCUSSION

LASL performed very well in identifying many species at low false discovery rates as shown in Figure 2a. Most LASL identifications had very low false discovery rates. At FDR < 1%, LASL identified about 95% of mass spectra. Out of 305 mass spectra, LASL identified 289 spectra at false discovery rates of 1% or less. Examples of correctly and incorrectly identified spectra are shown in Supporting Information (Figure S1). Since true identifications in the test set were known, we investigated how many mass spectra with FDRs < 0.01 were true identifications. Only one mass spectrum identified by LASL had a false identification at FDR < 1%. Furthermore, we also calculated a true false discovery rate (tFDR), which is a proportion of incorrect identifications of nondecoy spectra. The dotted line in Figure 2a is based on true false discovery rates. As shown here, our FDR estimations were close to true FDRs, and they were conservative estimations of tFDRs.

The proposed scores (p_{b*}) were also good at differentiating correct identifications from incorrect identifications (Figure 2b,c). We used decoy mass spectra in the test set to measure the discriminative power of the proposed scores. The matching scores (p_{b*}) in LASL were good at differentiating correct from incorrect identifications. The ROC (receiver operating characteristic) curve AUC (area under curve) was 98.84%. The precision-recall curve (PR) AUC was also very high with 96.01%.

LASL used multiple characteristics of the mass spectra to identify species. Among those, the top 10 important features were displayed in Figure 3 for the AB model. (The top 10 important features for other species/phenotypes can also be found in the Supporting Information.) LASL automatically chose the signature ions, which we denote as only those ions necessary and sufficient to correctly identify a microbe, and used them to identify the species. The characteristics of the signature ions were reproducible between both technical and biological replicates with some variations (see Figure S2 in the Supporting Information).

We note that there was no overlap in either decoy or nondecoy spectra between training and test sets. However, adding decoy spectra into the training set helped us identify more nondecoy spectra as shown in Table 1. Even without decoy spectra in the training set, LASL performed well. About 97% of top-ranked identifications among nondecoy spectra were correct (FDRs filtering was not applied at this stage). However, including the proposed decoy spectra in the train set, LASL performed even better. About 99% of top-ranked identifications among nondecoy spectra were correct identifications. In addition, including decoy spectra in the training process helped the model distinguish correct identifications from incorrect or decoy identifications. Without decoy spectra in the training process, LASL had poor precision-recall area under curve (PR AUC), while LASL with decoy-training had very good PR AUC. Specifically, LASL without decoy-training resulted in 93.80% ROC AUC and 56.51% PR AUC, while LASL with decoy-training resulted in 98.84% ROC AUC and 96.01% PR AUC.

Table 1. Performance Comparison in Bacterial Identifications with and without Decoy-Training[a]

	decoy-training	no decoy-training
correct top-ranked IDs, %	99.08	97.38
ROC AUC, %	98.84	93.80
PR AUC, %	96.01	56.51

[a] The proportion of correctly identified (nondecoy) spectra, area under curve (AUC) for receiver operating characteristic (ROC) curves and precision-recall (PR) curves, were used to compare the performance.

We further investigated the following alternative decoy spectra generation strategies: (1) A real number *D* was added to all of the *m*/*z* values of *K* peaks that were extracted from one spectrum, where *D* was a random number between −100 and 100; and (2) the intensity values of *K* peaks that were extracted from one spectrum were permuted where $K = 50$. At the estimated false discovery rate threshold of 1%, tFDRs were 0.35, 0.42, and 0.37% for the original decoy strategy, the alternative decoy strategy 1 (mass shift), and alternative decoy strategy 2 (intensity permutation), respectively. All of these

DOI: 10.1021/acs.analchem.9b03340

decoy strategies were conservative, but, among them, our original decoy strategy was the most conservative one. The proportions of correctly top-ranked identifications were 99.08, 98.69, and 97.38 for the original decoy strategy, the alternative decoy strategy 1 (mass shift), and alternative decoy strategy 2 (intensity permutation), respectively. Further investigation about the best way to construct decoy spectra is needed in the future.

Finally, we compared our proposed method to Biotyper[25] and bootstrap-based confidence scores.[12] We denote bootstrap-based confidence scores based on cosine and relative Euclidean distance as cosine and ieu, respectively. Details about the bootstrap-based confidence scores are shown in Supporting Information. LASL performed better than Biotyper, cosine, and ieu in various aspects. Since an FDR estimation strategy was developed for LASL, we compared the performance without making use of estimated false discovery rates. First, LASL was able to correctly identify more (nondecoy) mass spectra than the competing approaches. The proportion of correctly assigned bacteria for LASL was 99.08%, while the competing approaches produced results of 90.79, 90.49, and 84.27% for Biotyper, cosine, and ieu. Most importantly, LASL identified many more bacterial species than the competing approaches at true FDRs (tFDRs) < 0.01 (Table 2 and Figure 4). This comparison demonstrated the degree to which a glycolipid-specific bioinformatics tool could improve bacterial identifications from glycolipid mass spectra.

Table 2. Proportions of true bacterial identifications and the numbers of species identifications among LASL, Biotyper, cosine, and ieu. The proportions of correct top-ranked IDs were calculated before tFDR thresholds were applied

methods	correct top-ranked IDs, %	no. of IDs at tFDR < 1%
LASL	99.08	305
Biotyper	90.79	49
cosine	90.49	135
ieu	84.27	132

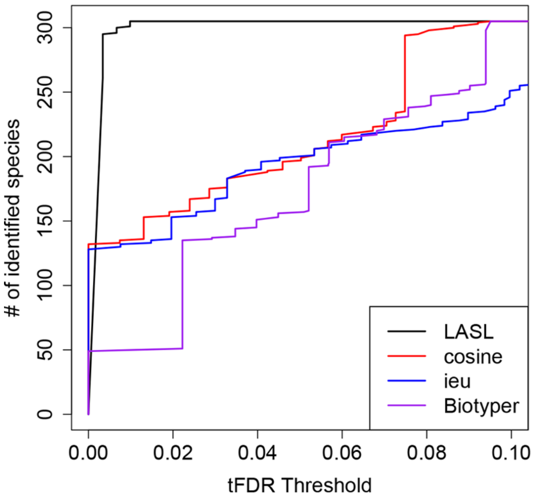


Figure 4. True FDR threshold vs the number of identified species plot comparing among LASL, Bioytper, cosine correlation (cosine), and intensity-weighted Euclidean distance (ieu).

LASL also performed well in identifying phenotypes when phenotypes of species were available in a spectral library (Table 3). At FDR < 1%, LASL identified phenotypes of 130 AB mass spectral entries, which were 97% of AB mass spectra in the test set. At the same threshold, 66 out of 67 KP mass spectra had their phenotype identifications at FDR < 1%. The area under curve calculations for ROC and PR were over 94% for both AB and KP phenotype identifications. We did not consider comparing our approach to Biotyper, cosine, and ieu in phenotype identifications since the number of confidently identified bacteria for Biotyper, cosine, and ieu were substantially smaller than LASL in the bacterial identification stage.

Table 3. Performance of LASL in Phenotype Identifications[a]

phenotype	ROC AUC, %	PR AUC, %	no. of IDs
ABcr vs ABcs	99.90	96.98	130 (134)[b]
KPcr vs KPcs	99.94	94.72	66 (67)[b]

[a]The number of identified phenotypes with FDR < 0.01, area under curve (AUC) for receiver operating characteristic (ROC) curves, and precision-recall (PR) curves were used to measure the performance. [b]The numbers in parentheses represent the total numbers of mass spectra identified as either AB or KP at FDR < 1% in the test set.

In this work, we proposed and tested a model-based spectral library approach for bacterial identifications using glycolipid mass spectra. LASL performed substantially better than the existing bioinformatics approaches in terms of accurately identifying and characterizing bacteria. However, LASL can identify only bacteria that are present in the spectral library. Thus, in the future, it is essential to build a spectral library that contains mass spectra from many different microbes. Noting that the mass spectrometry technology needed for this assay is relatively low-cost, widely distributed in hospital clinics, and easy to use, we anticipate that the diversity of bacteria in this library will increase rapidly in the future.

Another way to overcome the limitation of the existing small library with very few entries is to utilize false discovery rates. In practice, we may not know whether a bacterium of interest is present in a given spectral library, even when the library contains a wide variety of microbes. If a glycolipid mass spectrum of interest is not from bacteria in the spectral library, the best outcome would be that LASL assigns low matching scores (p_{b*}) and high false discovery rates to such spectra. Thus, the identification of those mass spectra would be discarded, not passing the FDR threshold (e.g., 1 or 5%). When we tested LASL with the supplementary data set, which contained no species from the spectral library, the matching scores for those identifications were very small ranging from 0.01 to 0.02. Their false discovery rates were larger than 5%. High false discovery rates or low matching scores of mass spectra do not necessarily imply that those spectra are not from bacteria in our spectral library. This is because glycolipid mass spectra of bacteria from the spectral library can have low matching scores due to the poor quality of mass spectra (e.g., low signal-to-noise). However, this demonstrated the potential use of our approach in practice in cases where our spectral library does not contain all bacteria. In the future, constructing decoy spectra from all of the currently available bacteria using theoretical lipid structures may enable us to more accurately measure false discovery rates for the identifications of bacteria that are not present in the spectral library. More investigation about decoy spectra generation strategies will help us use LASL in practice.

■ CONCLUSIONS

We developed and tested a model-based spectral library framework to analyze MALDI-TOF-MS data of bacterial membrane glycolipids such as lipid A from Gram-negative bacteria and related species from Gram-positive bacteria. The performance of LASL was demonstrated using human pathogens notorious as hospital-acquired infections (HAIs) and for the acquisition of resistance to antibiotics. With the proposed framework, the library can be extended easily, containing many more pathogens and organisms of general interest. As the microbial entries in the library increase, we believe that LASL will be able to provide valuable information for treatment decisions of infected patients ultimately helping to improve health care outcomes by decreasing morbidity/mortality rates as well as decreasing costs.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.9b03340.

Experimental details, supporting figures, and references (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: soyoungr@unr.edu. Tel.: +1 775-682-7116. Fax: +1 775-784-1340.

ORCID

So Young Ryu: 0000-0003-2347-7015

Notes

The authors declare the following competing financial interest(s): D.R.G. and R.K.E. have a significant financial interest in Pataigin LLC, the company developing diagnostic technology for rapid bacterial identification. All other authors declare no competing financial interests.

#To obtain data used in this work, e-mail: goodlett@maryland.edu.

■ ACKNOWLEDGMENTS

This publication was made possible by grants from the National Institute of General Medical Sciences (Grant GM103440 for S.Y.R. and G.A.W.; Grant 1R15GM126562-01 for G.A.W.; Grant 1R01GM111066-01 for D.R.G. and R.K.E.) from the National Institutes of Health. D.R.G. thanks the International Centre for Cancer Vaccine Science project of the International Research Agendas program of the Foundation for Polish Science cofinanced by the European Union under the European Regional Development Fund (MAB/2017/03) at the University of Gdansk. We also thank anonymous reviewers whose comments and suggestions helped improve and clarify this manuscript.

■ REFERENCES

(1) Willyard, C. *Nature* **2017**, *543*, 15.

(2) Khabbaz, R. F.; Moseley, R. R.; Steiner, R. J.; Levitt, A. M.; Bell, B. P. *Lancet* **2014**, *384*, 53−63.

(3) Leung, L. M.; Fondrie, W. E.; Doi, Y.; Johnson, J. K.; Strickland, D. K.; Ernst, R. K.; Goodlett, D. R. *Sci. Rep.* **2017**, *7*, 6403.

(4) Fondrie, W. E.; Liang, T.; Oyler, B. L.; Leung, L. M.; Ernst, R. K.; Strickland, D. K.; Goodlett, D. R. *Sci. Rep.* **2018**, *8*, 15857.

(5) Elssner, T.; Kostrzewa, M.; Maier, T.; Kruppa, G. Microorganism identification based on MALDI-TOF-MS fingerprints. In *Detection of Biological Agents for the Prevention of Bioterrorism*; Banoub, J., Ed.; NATO Science for Peace and Security Series A: Chemistry and Biology; Springer Science + Business Media: Berlin, Heidelberg, 2011; pp 99−113.

(6) van Belkum, A.; Durand, G.; Peyret, M.; Chatellier, S.; Zambardi, G.; Schrenzel, J.; Shortridge, D.; Engelhardt, A.; Dunne, W. M. *Ann. Lab. Med.* **2013**, *33*, 14−27.

(7) Mather, C. A.; Rivera, S. F.; Butler-Wu, S. M. *Journal of Clinical Microbiology* **2014**, *52*, 130−138.

(8) Pence, M. A.; McElvania TeKippe, E.; Wallace, M. A.; Burnham, C. A. *Eur. J. Clin. Microbiol. Infect. Dis.* **2014**, *33*, 1703−1712.

(9) Seng, P.; Drancourt, M.; Gouriet, F.; La Scola, B.; Fournier, P.; Rolain, J. M.; Raoult, D. *Clin. Infect. Dis.* **2009**, *49*, 543−551.

(10) Clark, A. E.; Kaleta, E. J.; Arora, A.; Wolk, D. M. *Clin. Microbiol. Rev.* **2013**, *26*, 547−603.

(11) Bertelli, C.; Greub, G. *Clin. Microbiol. Infect.* **2013**, *19*, 803−813.

(12) Yang, Y.; Lin, Y.; Chen, Z.; Gong, T.; Yang, P.; Girault, H.; Liu, B.; Qiao, L. *Anal. Chem.* **2017**, *89*, 12556−12561.

(13) Wilson, M. C.; Liang, T.; Yoon, S. H.; Leung, L.; Ernst, R. K.; Goodlett, D. R. A cartesian product approach to lipid A structure identification. 2015; http://goodlettlab.org/posters/2015_ASMS_Lisa.pdf.

(14) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. *Proteomics* **2007**, *7*, 655−667.

(15) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R. *Nat. Methods* **2008**, *5*, 873−875.

(16) Deutsch, E. W.; Perez-Riverol, Y.; Chalkley, R. J.; Wilhelm, M.; Tate, S.; Sachsenberg, T.; Walzer, M.; Käll, L.; Delanghe, B.; Böcke, S.; et al. *J. Proteome Res.* **2018**, *17*, 4051−4060.

(17) Ryu, S. Computer implemented methods and systems for identifying a species from mass spectra. U.S. Provisional Pat. 62/809285, 2019.

(18) El Hamidi, A.; Tirsoaga, A.; Novikov, A.; Hussein, A.; Caroff, M. *J. Lipid Res.* **2005**, *46*, 1773−1778.

(19) Cai, Y.; Chai, D.; Wang, R.; Liang, B.; Bai, N. *J. Antimicrob. Chemother.* **2012**, *67*, 1607−1615.

(20) Gibb, S.; Strimmer, K. *Bioinformatics* **2012**, *28*, 2270−2271.

(21) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627−1639.

(22) Ryan, C. G.; Clayton, E.; Griffin, W. L.; Sie, S. H.; Cousens, D. R. *Nucl. Instrum. Methods Phys. Res., Sect. B* **1988**, *34*, 396−402.

(23) Chen, T.; Guestrin, C. XGBoost: a scalable tree boosting system. *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, 2016; pp 785−794.

(24) Benjamini, Y.; Hochberg, Y. *Journal of the Royal Statistical Society. Series B (Methodological)* **1995**, *57*, 289−300.

(25) Mellmann, A.; Bimet, F.; Bizet, C.; Borovskaya, A.; Drake, R.; Eigner, U.; Fahr, A.; He, Y.; Ilina, E.; Kostrzewa, M.; et al. *Journal of clinical microbiology* **2009**, *47*, 3732−3734.