Supporting Information for 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*

E-mail: soyoungr@unr.edu

Contents

List of Figures

Tuning parameter selection for LASL

To select the best tuning parameters for bacteria/phenotype models, we used the 5-fold cross-validation and the grid search. The parameters considered were the number of maximum iterations, η which controlled the learning rate, γ which controlled the regularization, the maximum depth of the tree, and the minimum sum of instance weight needed in a child. Noting the imbalance between positive and negative cases in our dataset, we set the scale of positive weight as $\sum$(negative cases)/(positive cases). For example, for AB, the scale of positive weight was 6.81 (=(332+1500)/269) since there were 269 mass spectra from AB, 332 mass spectra from non-AB species, and 1,500 decoy mass spectra in the train set. In addition, we set the maximum delta step for each leaf output as one. For each tuning parameter, it took about 15 seconds to train the model using one core. In this paper, we explored >20,000 combinations of tuning parameters, and it took about 1.5 hours to train LASL using 64 cores.

Biotyper

The following is details about MALDI Biotyper Informatics package[1] (Bruker Daltonics, Bremen, Germany). MSPs (Main SPectra) are created by the standard MALDI Biotyper MSP Creation Method using mass spectra from the train set. In this study, each MSP represents one species. An MSP contains the average masses and the average intensities of the selected peaks (representing most reproducible and typical for a certain bacterial species) as well as the frequency of the peaks. Then, the MALDI Biotyper picks peaks using a Spectra Differentiation Filter Algorithm from which it generates mass lists for spectra in the test set. Then, these mass lists for spectra in the test set are compared by the MALDI Biotyper Pattern Matching Algorithm to the mass lists of all MSPs in the library. The Biotyper software generates a list of probable species identifications ranked by the scores generated by MALDI Biotyper Pattern Matching Algorithm. The score[2] (e.g. often called as log(score))

is based on the log_{10} of the product of three factors: the matches of the unknown spectrum (in the test set) against the MSPs in the library, the reverse matches of MSPs with the unknown spectrum, and the correlation of relative intensities of the unknown spectrum and the MSPs in the library. The product has a maximum value of 1,000, leading to a maximum (log-transformed) score of 3.

Bootstrap-based confidence scores

We compared our approach to recently published bacterial identification approaches for whole cell typing. This approach proposed confidence scores based on spectra similarity scores and a bootstrap approach. Specifically, one similarity score was a relative Euclidean distance weighted by peaks between spectrum a and b:

$$1 - \frac{\sum_{i \in W} eu_i(y_{ai} + y_{bi}) + \sum_{i=1}^{n_a} y_{ai} + \sum_{i=1}^{n_b} y_{bi}}{\sum_{i=1}^{n_a} y_{ai} + \sum_{i=1}^{n_b} y_{bi}} + \frac{\sum_{i \in W}(y_{ai} + y_{bi})}{\sum_{i=1}^{n_a} y_{ai} + \sum_{i=1}^{n_b} y_{bi}}, \quad (1)$$

where n_a and n_b were the number of peaks in a mass spectrum a and b, respectively, x_{ui} was an m/z value of ith peaks in a mass spectrum u, y_{ui} was an intensity value of ith peaks in a mass spectrum u, t represented a m/z tolerance (e.g., 1Da), W was a set of indices of common peaks between mass spectra a and b, and

$$eu_i = \frac{\sqrt{(x_{ai} - x_{bi})^2 + (y_{ai} - y_{bi})^2}}{\sqrt{max(x_{ai}, x_{bi})^2 t^2 + max(y_{ai}, y_{bi})^2}}. \quad (2)$$

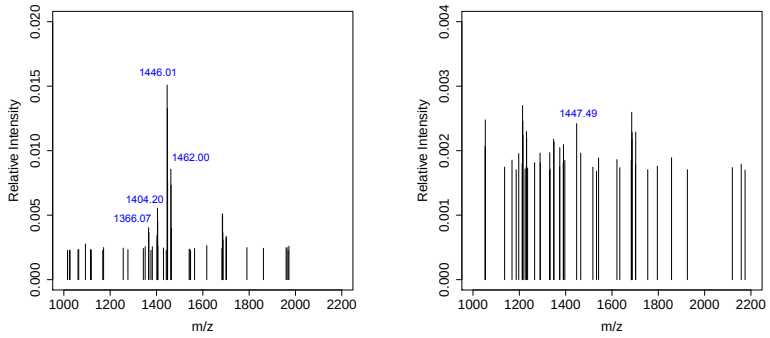
Another score was cosine correlation:

$$\frac{\sum_{i \in W} y_{ai} y_{bi}}{\sqrt{\sum_{i=1}^{n_a} y_{ai}^2} \sqrt{\sum_{i=1}^{n_b} y_{bi}^2}}. \quad (3)$$

The bootstrap-based confidence score[3] was calculated by dividing the number of the top

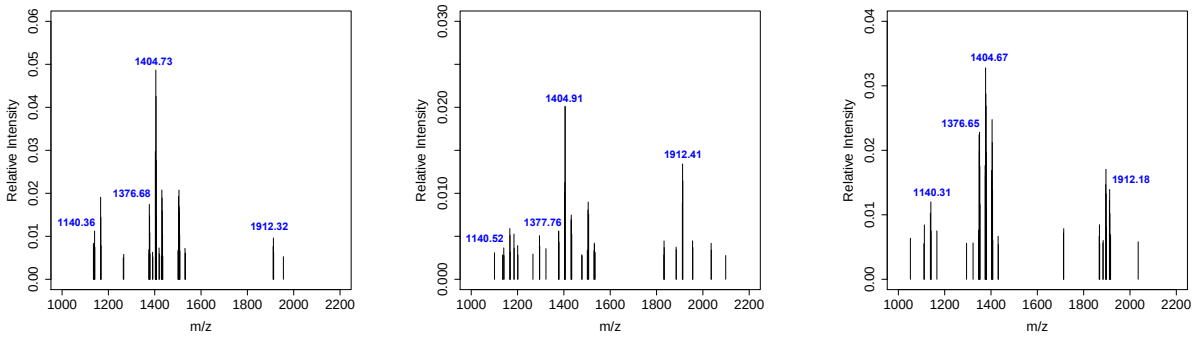
bootstrapped mass spectra that matched to the same species as a mass spectrum of interest by the total number of bootstrap spectra. The bootstrap mass spectra were constructed by sampling N peaks with a replacement from a mass spectrum of interest. Note that bootstrap mass spectra were different from our proposed decoy mass spectra since our decoy mass spectra were constructed from mass spectra from multiple species in the spectral library instead of from one mass spectrum of interest.

Supporting Figures



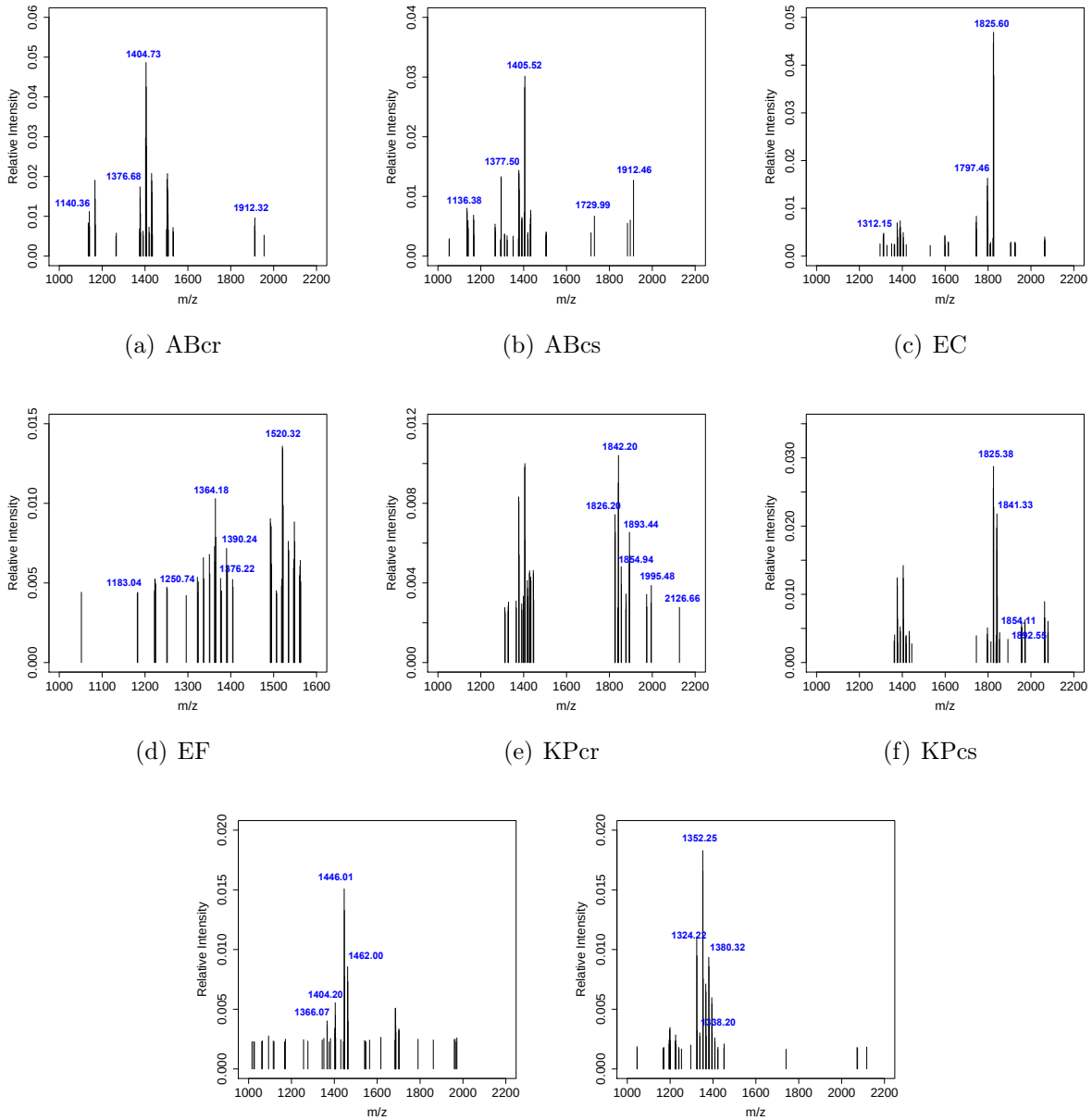
(a) Good Example (b) Bad Example

Figure S1: Examples of glycolipid profiling mass spectra. The masses of peaks selected by LASL were displayed in blue. (a) The given mass spectrum of PA was correctly identified as PA by LASL. (b) The given mass spectrum of PA was incorrectly identified as EF by LASL.



(a) AB (b) Tech Replicate (c) Biological Replicate

Figure S2: Examples of technical/biological replicates. The masses of peaks selected by LASL were displayed in blue. (a) An example spectrum from species AB. (b) A technical replicate of spectrum (a). (c) A biological replicate of spectrum (a).



(a) ABcr (b) ABcs (c) EC

(d) EF (e) KPcr (f) KPcs

(g) PA (h) SA

Figure S3: An example spectrum for each species/phenotype. The masses of peaks selected by LASL were displayed in blue. (a) An example spectrum for ABcr. (b) An example spectrum for ABcs. (c) An example spectrum for EC. (d) An example spectrum for EF. (e) An example spectrum for KPcr. (f) An example spectrum for KPcs. (g) An example spectrum for PA. (h) An example spectrum for SA.

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

(a) AB

Feature	Gain	Coverage	Frequency
intensity @ 1826	0.658	0.257	0.102
intensity @ 1798	0.061	0.062	0.041
rank intensity @ 1841	0.058	0.097	0.041
intensity @ 1825	0.044	0.045	0.061
rank intensity @ 1718	0.031	0.043	0.020
rank intensity @ 1308	0.021	0.028	0.020
rank intensity @ 1798	0.019	0.020	0.041
rank intensity @ 1366	0.014	0.074	0.041
m/z at @ 1312	0.012	0.049	0.020
m/z at @ 1972	0.010	0.025	0.020

(b) EC

Feature	Gain	Coverage	Frequency
intensity @ 1391	0.648	0.286	0.093
intensity @ 1183	0.055	0.058	0.023
intensity @ 1392	0.042	0.045	0.047
m/z @ 1204	0.036	0.129	0.070
intensity @ 1184	0.034	0.050	0.023
intensity @ 1364	0.024	0.038	0.093
m/z @ 1825	0.022	0.054	0.023
intensity @ 1521	0.020	0.024	0.047
rank intensity @ 1251	0.018	0.033	0.023
intensity @ 1377	0.016	0.033	0.047

(c) EF

Feature	Gain	Coverage	Frequency
intensity @ 1841	0.747	0.270	0.136
intensity @ 1842	0.137	0.155	0.106
rank intensity @ 2125	0.025	0.069	0.030
intensity @ 1993	0.011	0.038	0.015
intensity @ 1826	0.010	0.016	0.061
rank intensity @ 1854	0.010	0.004	0.045
m/z @ 1352	0.007	0.041	0.015
intensity @ 1892	0.005	0.011	0.030
rank intensity @ 1366	0.005	0.035	0.015
m/z @ 1367	0.005	0.041	0.015

(d) KP

Feature	Gain	Coverage	Frequency
intensity @ 1447	0.807	0.323	0.124
rank intensity @ 1367	0.037	0.108	0.053
intensity @ 1463	0.034	0.073	0.062
rank intensity @ 1337	0.029	0.155	0.097
m/z @ 1142	0.024	0.119	0.044
rank intensity @ 1446	0.014	0.014	0.044
intensity @ 1384	0.013	0.038	0.044
m/z @ 1009	0.010	0.005	0.027
rank intensity @ 1463	0.010	0.029	0.035
rank intensity @ 1404	0.003	0.003	0.044

(e) PA

Feature	Gain	Coverage	Frequency
intensity @ 1353	0.965	0.603	0.308
intensity @ 1325	0.022	0.238	0.231
m/z @ 1325	0.003	0.059	0.038
intensity @ 1351	0.003	0.042	0.077
intensity @ 1339	0.003	0.014	0.077
intensity @ 1381	0.001	0.007	0.077
intensity @ 1354	0.001	0.002	0.038
intensity @ 1352	0.001	0.003	0.038
intensity @ 1324	0.000	0.020	0.038
m/z @ 1272	0.000	0.011	0.038

(f) SA

Figure S4: Top 10 important features selected by LASL for the following species: (a) AB, (b) EC, (c) EF, (d) KP, (e) PA, and (f) SA.

Feature	Gain	Coverage	Frequency
intensity @ 1911	0.370	0.241	0.093
intensity @ 1404	0.092	0.076	0.028
intensity @ 2035	0.077	0.100	0.056
intensity @ 1895	0.075	0.024	0.037
intensity @ 1912	0.063	0.071	0.046
intensity @ 1403	0.059	0.054	0.037
intensity @ 1883	0.048	0.046	0.019
intensity @ 1405	0.045	0.042	0.019
intensity @ 1374	0.017	0.040	0.009
intensity @ 1505	0.017	0.011	0.009

(a) ABcr

Feature	Gain	Coverage	Frequency
intensity @ 1405	0.249	0.120	0.039
intensity @ 1911	0.202	0.066	0.036
intensity @ 1912	0.121	0.074	0.049
intensity @ 1404	0.096	0.124	0.049
rank intensity @ 1377	0.042	0.051	0.036
m/z @ 2035	0.028	0.037	0.018
rank intensity @ 1911	0.020	0.023	0.026
intensity @ 1729	0.018	0.014	0.023
intensity @ 1884	0.014	0.003	0.005
intensity @ 1376	0.013	0.024	0.018

(b) ABcs

Feature	Gain	Coverage	Frequency
intensity @ 1841	0.540	0.250	0.078
intensity @ 1842	0.089	0.132	0.044
intensity @ 1404	0.057	0.068	0.022
m/z @ 1391	0.024	0.036	0.007
rank intensity @ 1746	0.023	0.051	0.019
intensity @ 1993	0.021	0.034	0.007
rank intensity @ 2124	0.018	0.029	0.007
intensity @ 1973	0.016	0.025	0.033
intensity @ 1403	0.016	0.024	0.004
rank intensity @ 1404	0.015	0.031	0.044

(c) KPcr

Feature	Gain	Coverage	Frequency
intensity @ 1826	0.497	0.226	0.097
intensity @ 1842	0.251	0.152	0.086
rank intensity @ 1842	0.057	0.067	0.043
rank intensity @ 1331	0.044	0.098	0.043
rank intensity @ 1826	0.034	0.047	0.065
intensity @ 1786	0.023	0.093	0.043
intensity @ 1839	0.022	0.098	0.032
rank intensity @ 2307	0.022	0.051	0.032
rank intensity @ 1551	0.012	0.052	0.032
rank intensity @1825	0.009	0.017	0.054

(d) KPcs

Figure S5: Top 10 important features selected by LASL for the following phenotypes: (a) ABcr, (b) ABcs, (c) KPcr, and (d) KPcs.

References

(1) Mellmann, A.; Bimet, F.; Bizet, C.; Borovskaya, A.; Drake, R.; Eigner, U.; Fahr, A.; He, Y.; Ilina, E.; Kostrzewa, M. et al. High interlaboratory reproducibility of matrix-assisted laser desorption ionization-time of flight mass spectrometry-based species identification of nonfermenting bacteria. *Journal of clinical microbiology* **2009**, *47*, 3732–3734.

(2) Lartigue, M.-F.; Héry-Arnaud, G.; Haguenoer, E.; Domelier, A.-S.; Schmit, P.-O.; Van Der Mee-Marquet, N.; Lanotte, P.; Mereghetti, L.; Kostrzewa, M.; Quentin, R. Identification of Streptococcus agalactiae isolates from various phylogenetic lineages by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Journal of clinical microbiology* **2009**, *47*, 2284–2287.

(3) Yang, Y.; Lin, Y.; Chen, Z.; Gong, T.; Yang, P.; Girault, H.; Liu, B.; Qiao, L. Bacterial whole cell typing by mass spectra pattern matching with bootstrapping assessment. *Analytical Chemistry* **2017**, *89*, 12556–12561.