Published in final edited form as:

Anal Chem. 2020 January 07; 92(1): 1050–1057. doi:10.1021/acs.analchem.9b04049.

The Aristotle Classifier Provides Accurate Identifications of Highly Similar Bacteria Analyzed by MALDI-TOF MS.

Heather Desaire^{1,*}, David Hua¹

¹Department of Chemistry, University of Kansas, Lawrence, Kansas 66045, United States

Abstract

MALDI-TOF MS has shown great utility for rapidly identifying microbial species. It can be used to successfully type bacteria and fungi from a variety of sources more rapidly and cost-effectively than traditional methods. One area where improvements are necessary is in the typing of highly similar samples, such as those samples from the same genus but different species or samples from within a single species but from different strains. One promising way to address this current limitation is by using advanced machine learning techniques. In this work, we adapt a newly developed machine learning tool, the Aristotle Classifier, to bacterial classification of MALDI-TOF MS data. This tool was originally developed for classifying glycomics and glycoproteomics data, so we modified it to be well-suited for assigning mass spectral data from bacterial proteins. The classifier exceeds existing benchmarks in classifying bacteria, and it shows particularly strong performance when the samples to be identified are highly similar. The combination of mass spectrometry data and tools like the Aristotle Classifier could ameliorate the ambiguities associated with challenging bacterial classification problems.

Graphical Abstract

Introduction

Machine learning methods are increasingly capable of interpreting massive data sets and correctly assessing complex patterns. They have been used to process images,¹ interpret speech,² and predict lifespan.³ While its virtually unconfined potential is almost confounding, machine learning is still a relatively untapped resource for interpreting mass

*Corresponding Author: Heather Desaire, phone: 785-864-3015, hdesaire@ku.edu.

Supporting Information

Supporting Information is available free of charge on the ACS Publications website: "The Aristotle Classifier for Bacteria" is the R code for the Aristotle Classifier used in this project.

spectrometry data. Emerging examples of applying machine learning to MS include: prediction of peptide sequences,^{4–6} (including the prediction of modified peptides),⁷ assigning cancer margin to image data,⁸ identifying metabolites,⁹ and improving data quality in proteomics experiments.¹⁰ Despite these notable examples, most researchers acquiring mass spectrometry data to identify disease states still do not employ machine learning methods for classifying the origin or disease state of their samples.

Researchers in the glycomics field, for example, still manually identify individual glycoforms that differentiate disease states from healthy ones,^{11–14} or they apply 100-year-old mathematical tools, such as principal components analysis (PCA)^{12,15,16} to evaluate the quality of their biomarkers. To advance data science methods in this field, we developed a machine learning method specifically for classifying MS data from glycomics experiments. The Aristotle Classifier accepts any type of MS-based glycomic or glycoproteomic data and classifies the spectra; for example, it can assign samples to a disease state.¹⁷ This tool has better predictive power than the commonly implemented data science methods in the field.¹⁷ After successfully applying the Aristotle Classifier to several different types of glycomics data sets, we next asked the question: Could this tool benefit other mass spectrometry researchers, outside the glycomics community?

One additional domain where machine learning could improve the utility of mass spectrometric analyses and impact human health is the diagnosis of microbial infections using mass spectrometry data.^{18,19} Commercial MALDI-TOF MS instruments to type bacteria are now on the market, including the Bruker Biotyper and the Biomereix Vitek, and these systems compare user-generated mass spectra of bacterial samples to prototypical spectra stored in spectral databases.^{20,21} While this approach is useful for discriminating a variety of sample types from each other, samples that are especially similar to each other are often impossible to assign using the on-board spectral databases. Particularly challenging samples include those from two different species of bacteria that originate from the same genus^{22–24} or those from the same species but with different phenotypes, such as samples that are either antibacterial resistant or not.¹⁸

In many cases, machine learning methods have provided researchers better ability to discriminate bacterial samples than what would be gained by simply comparing user-generated spectra to a set of prototypes.^{22–26} In one case, the use of the XGBoost classifier was pivotal in enabling the identification of polymicrobial species based on MS data of their membrane glycolipids.²⁵ In other examples, researchers have used a variety of machine learning strategies to discriminate samples of bacterial mixtures.^{24,26} Finally, support vector machines, random forest classifiers, and novel oversampling techniques have been implemented to improve the characterization of spectra from highly similar bacteria.²⁷

We sought to determine whether the Aristotle Classifier, which was developed for glycomics studies, could provide additional benefits to the field of bacterial identification. To test this hypothesis, we used nine different data sets containing MALDI-TOF MS data of bacterial proteins, where all the members of a given data set originate from the same genus. After classifying the data with the Aristotle Classifier, we compared the results to the current state

of the art. The Aristotle Classifier substantially outperformed existing methods for discriminating highly similar spectra from closely related bacteria.

Experimental

The data used in this study are available at the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/index.php>), and they are found by searching for the data set “MicroMass.” A group of 571 reference mass spectra found in the “pure spectra” datafile were used. These spectra are from 20 different bacterial species, 9 different genera, and 213 different strains.²⁴ The samples from each genus are considered an independent data set for these studies, where the overall goal is to determine sub-genus classification specificity. Each species is represented by 11 to 60 mass spectra from a collection of 7 to 20 bacterial strains.²⁴ The identity of each sample in the data set is provided in a separate file in the repository associated within the MicroMass project, named “pure.spectra.metadata.csv”. The labels in the metadata file were used to partition and assign the samples in the “pure spectra” data set.

The version of the Aristotle Classifier used to conduct all experiments herein is provided in the Supplemental Data. The classifier was built to run in R, and all testing was done in RStudio version 3.5.0. The original master data file, containing 571 samples with 1300 features, was transformed so that each sample was in a column and each feature was in a row, to be consistent with the overall framework of the Aristotle Classifier. The mass spectral signals were also normalized to adjust for run-to-run signal variation inherent in mass spectrometry experiments: All the ion counts for each feature (each peak) within a given sample were summed, generating a “total ion counts” value for each sample. Next, the ion counts for each feature (each peak) were divided by the total ion counts value for that sample. The resulting decimal numbers, which were all zero or greater, and all less than one, were used as the feature set.

In addition to the data matrix described above, several other input values are provided to the classifier, including the number of rows and columns in the data matrix, a vector called “Low”, a vector called “Hi,” and the number of aggregating replicates, “k”. During training experiments for each pairwise species comparison, the “Low” vector contained a list of all the columns in the matrix that corresponded to MS data from one particular species, and the “Hi” vector contained a list of columns that corresponded to its comparator species of the same genotype. The number of aggregated replicates (k) was set to 1000. To assure that results were consistent from day to day, each experiment was repeated 5 times, and the accuracy reported is the result that occurred most often during the five experiments. Most commonly, the accuracy would be identical for all five experiments.

In the five-fold cross validation experiments, samples within a single group (for example, the samples from the CPH species within the WQP genus) would be split into five sub-groups of equivalent or nearly equivalent size. Four of the five sub-groups would be used as labeled “training” samples, and the fifth group would be used as unlabeled “validation” samples. Only labeled samples would be used for feature selection: They would be used in the “Low”

and “Hi” vectors. Validation samples were scored based on the features selected by the training samples, and their overall score was reported in the Results vector.

For experiments where a non-zero value was used to distinguish Group 1 from Group 2, this threshold value was determined in the following manner: The Aristotle Classifier was implemented in its normal operational mode, and a numerical result was obtained for each sample. Next, an in-house script determined the threshold where the largest number of training samples would be scored correctly. The threshold was always located at the midpoint between two correctly scored training samples (one sample from each group). In cases where more than one threshold was available that satisfied these criteria, the threshold closest to zero was chosen.

Results and Discussion

Table 1 describes the nine data sets used for testing the utility of the Aristotle Classifier for bacterial typing. These data had been acquired previously²⁴ and are available at the UCI Machine Learning Repository. For each of the nine data sets tested, MS data for each sample originates from a single bacterial type. In eight of the nine test sets, two different species need to be discriminated; in one case, set 7, four different closely-related species are present and must be discriminated from each other. In some cases, the number of each species is fairly balanced – for example in Set 2, the majority and minority classes are very comparable, with 29 and 31 spectra available respectively. In some cases, class imbalance is more apparent: In Test Case 7, the minority class only contains 13% of the total examples in the data set. In each case, 1300 features are present. These features correspond to the abundances of different MS peaks that were in the MALDI-MS data of the acquired samples.

Description of the Aristotle Classifier and its Adaptation to Bacterial Data sets

The Aristotle Classifier, as originally described,¹⁷ contains two components. The first component accepts MS features for any given sample and multiplexes those features so that all possible ratios of two different peaks and three different peaks could be considered as features. This component was found to be useful for glycomics studies because a ratio of MS peaks is often more useful for discriminating glycans than any individual peak.¹⁷ The second component of the Aristotle classifier uses the features generated from part 1 to classify the samples, as described in more detail in a subsequent section herein. Mass spectra of glycans and of bacteria are very different, so before applying the Aristotle Classifier to bacteria, we first reassessed whether the feature generation step would be useful for classifying the bacterial samples. When typing samples based on their glycosylation, 10–30 different glycans are typically quantifiable; by contrast, the data sets studied herein have 1300 features per spectrum. This large difference in the number of features means that the component of the original classifier that multiplexed all the features prior to scoring would not be practical for use in bacterial typing. Considering all possible ratios of 1300 mass spectral features in the bacterial data sets would produce a feature set of >1,000,000 components per sample, which is an unwieldy size. Therefore, in assessing whether the Aristotle Classifier would be useful for typing bacterial species, we determined that the

classifier would have to score each spectrum based solely on the original 1300 features only and not consider the ratios of the peaks.

The scoring component of the Aristotle Classifier is pictorially described in Figure 1; the process contains five steps: 1) data set partitioning; 2) feature selection; 3) sample scoring; 4) aggregation; and 5) result reporting. The first step is for the classifier to select a total of eight samples with known sample type out of the collection of labeled samples in the data set; four are selected from each of two groups. In “training” experiments, all the samples are treated as labeled, and every sample in a given group has an equal likelihood of being selected in this step. In validation experiments, the validation samples are treated as unlabeled samples and are not ever selected for feature selection (Step 2.) After the subset of eight labeled samples are selected, they are used to determine which of the 1300 features are useful for discriminating the samples in Group 1 from those in Group 2. Each feature is assessed independently according to the following criteria: To be considered “useful,” one of two cases must be satisfied. Either all of the abundances for the four selected Group 1 samples have to be greater than all of the abundances for the four selected Group 2 samples, or all four of the selected Group 1 samples have to have lower abundances --for the given feature -- compared to all four of the selected Group 2 samples. The useful features are saved; the others are ignored. The useful features are used to score the *unselected* samples in Step 3, as described in Figure 1. Every useful feature contributes either a +1 or a -1 to the overall score for each scored sample. The feature contributes a +1 if the sample being scored has an abundance that is similar to the Group 1 samples; otherwise, the feature contributes a -1 to the sample score. Every saved feature is treated the same, either contributing a +1 or -1 to every scored sample, and the sum of the scores is computed. Finally, in Step 4, the entire process is repeated so that different combinations of labeled data can be used to select and score the remaining samples. A typical number of replicates would be 1000, although this is an adjustable parameter in the classifier.¹⁷ After combining the scores from all the aggregated experiments, the outcome (Step 5) is reported: Each sample receives a numerical score. If the score is >0, the sample is assigned to Group 1; if the score is <0, the sample is assigned to Group 2.

Figure 2 shows three examples of the output from the Aristotle Classifier in discriminating the MS data from different species. For each data set, the results are plotted with the sample number on the X axis and the score on the Y axis. The point along the X axis where the color changes indicates the point in the data set separating Group 1 samples from Group 2 samples. In each of the three cases shown in Figure 2, the vast majority of the samples in the training sets could be correctly assigned to their appropriate group (their species). For example, in Figure 2A, 28 of the 29 samples from the KSF group (Group 1) had scores of greater than zero, indicating their belonging with the KSF group. The scores for these samples appear in the left side of the figure; those in the upper left (green) quadrant are correctly assigned. On the right side of the figure, the scores for the remaining samples are present. Each of these samples is from the KWJ species, Group 2. In this case, all 31 of the samples are correctly assigned: all 31 samples received a negative score (<0), placing them in the lower-right (green) quadrant of the graph. In all, 59 of the 60 samples were correctly assigned to their species, resulting in a 98% training accuracy, as reported in Table 2; data set 2; Training Accuracy; “cut at zero”. The other two examples in Figure 2, which show

results for the 4th and 5th data sets, also demonstrate that the method is effective at classifying the training data for these samples. In Figure 2B, just five samples out of 54 were mis-assigned, resulting in a training accuracy of 91%, as shown in Table 2, data set 5, training accuracy, “cut at zero.” Figure 2C contains some of the least accurate results in the study: Here, 84% of the samples in data set 4 were correctly assigned, and four samples from each species were mis-assigned. These mis-assigned samples appear in the lower-left and upper right (red) quadrants on the figure. The results for the remaining training sets containing two species (set 1, 3, 6, 8, and 9) are reported on Table 2; training data; “cut at zero.” The Aristotle Classifier did an admirable job in assigning the samples: The overall accuracy for all the samples in these data sets was 93%.

Modification to the original classifier scoring protocol

Upon closer inspection of the training data in Figure 2 and the results for the other data sets studied herein, we realized that a modification to the original Aristotle Classifier scoring algorithm might improve the overall classification accuracy of this study. In the original description of the work, the threshold between the two groups being scored was always demarcated at “zero”. In other words, if the classifier output was a positive score for a given sample, that sample was always assigned as a member of Group 1, while a negative score relegated the sample to Group 2. In evaluating the data in Figure 2, it becomes apparent that some training sets of bacteria would have fewer missed assignments if a non-zero threshold were used to distinguish the two groups. For example, in Figure 2A, if the threshold that differentiates Group 1 from Group 2 were set at a value of -3500 , indicated by the dashed line on the figure, all the samples would be correctly assigned. Likewise, if the data in Figure 2C had been assigned using a threshold of positive 3500 , as indicated by a dashed line, then three of the mis-assigned samples could be correctly classified. The improvement in training accuracy is shown in Table 2: The data for Figure 2A and 2C respectively can be found in Table 2; data set 2 and 4 respectively; “Optimized Cut-Point.” Aside from data sets 2 and 4, other data sets’ training accuracy could also be improved if a non-zero threshold were chosen. Overall, five of the eight data sets shown in Table 2 have improved training accuracy when a non-zero value is used as the border that distinguishes the two species types. Furthermore, the overall weighted average of the training accuracy creeps up, from 93%, when zero was used as the threshold value to distinguish the two groups, to 95%, when the threshold value is optimally chosen to improve the training accuracy of each data set.

Clearly, choosing an optimal threshold for discriminating the two groups results in moderately higher accuracy in these training data. However, a question remains: Does this approach, of optimizing the threshold that separates Group 1 and Group 2, simply result in overfitting the data, or does it provide a true benefit, even for unlabeled data?

To determine the utility of introducing a custom threshold for each group, we ran a series of five-fold cross validations. In a five-fold cross-validation, 80% of the data is used for training (and determining the threshold between the two groups), and 20% of the data are treated as “unknowns” and not used in any way for model training. Only the scores of the unknowns, the remaining 20% of the data, are used to determine the method’s accuracy. During the five-fold cross-validation experiments, none of the “unknown” samples are used

in the feature-selection component of the algorithm, since they are all treated as unlabeled samples. (See Figure 1.) Therefore, in the five-fold cross-validation experiment, the labeled data sets were 20% smaller, since the validation samples were no longer part of the labeled training sets. The cross-validation experiment is conducted five times so that every sample is used as an unknown in one of the five tests. In each cross-validation experiment, samples from two different species, which both originate from the same genus, are compared, replicating the earlier training experiments.

Two different cross validations were conducted: In one case, the threshold between the two groups remained at zero; in the other case, the threshold was determined by selecting the value that optimally assigned the training data to their respective groups, and that threshold value was applied to score the validation samples. The results of the cross-validation experiments are shown in Table 2, under “Cross Validation Accuracy.” The two different columns in this section of the table are results for the two different experiments, where either “zero” was used as the threshold between the two species (“Cut at Zero” column), or an optimized threshold was determined, based on the training data (“Optimized Cut Point” column.)

Does the cross-validation experiment with the optimized cut point outperform the cross-validation experiment when zero was used as the threshold (“Cut at Zero”)? Some data sets are slightly more accurately assigned, and some fare worse. For example, data sets 2 and 3 are both assigned perfectly when an optimized cut point is used, while one sample is mis-assigned in each of these data sets when the assignments are made based on a cut-point of zero. For data sets 1 and 6, the inverse trend is observed: fewer samples are mis-assigned in the “cut at zero” experiment, where the samples’ identities are assigned based on whether or not the classifier’s final score was greater or less than zero. Overall using training data to determine a threshold score that discriminates Group 1 from Group 2 does not universally improve the accuracy of the method, in cross validation, over simply assigning the two groups based on whether their final scores are positive or negative. In fact, the cross-validation accuracy for the group of eight data sets dipped slightly, from 93% to 92%. Since no compelling benefit is demonstrated for these data, we focus henceforth on continuing to classify samples using the value of zero as the borderline between the two groups being scored, as described in Figure 1. When scoring other types of data sets, it may be possible that using an optimized threshold, as described herein, provides higher accuracy in cross-validation experiments than using a zero threshold. We would advise users of the classifier to check both possibilities, as we have done here, so that the classifier could be used in its optimal form.

While the cross-validation experiments did not confirm our original hypothesis, that using a non-zero cut point could improve the accuracy for these data sets, the experiments were useful for a different reason: They confirmed that the Aristotle Classifier can score rather small sets of data without training bias. Note that the overall classification accuracy of the training data, using the originally proposed cut-point at zero is 93%, and this value is equal to the overall *validation* accuracy when the cut-point is also set at zero, indicating no training bias or overfitting of the data. This lack of training bias is a signature feature of the Aristotle Classifier: Even though some of the data sets had very few examples in them – sets

1 and 3 had 37 and 30 samples respectively – the Aristotle Classifier correctly predicted how well the samples would be assignable in cross validation experiments.

Discriminating sets of bacteria with more than two species present.

While the Aristotle Classifier could be applied directly to the sample sets containing only two different species, we needed to develop a way to discriminate samples that had more than two components present. Group 7 (in Table 1) has four different bacterial species that are all from the same genus; the species are TRH, DNW, YZE, and BIK. Since the Aristotle Classifier is only designed to differentiate two groups at a time, its implementation to a classification problem with four different sample types requires some creativity. After testing several possible strategies for discriminating these groups, the optimal approach turns out to be the one shown in Figure 3. First, TRH is distinguished from the other three species present (BIK, DNW, YZE.) Then BIK is distinguished from the remaining two (DNW, YZE). Finally, the last two species (DNW, YZE) are distinguished from each other. While other solutions are also possible for discriminating multiple groups with a two-state classifier, this one performs well on these data. The overall accuracy for discriminating these four very similar bacterial types was determined using the same five-fold cross-validation strategy described above. In this case the threshold between the two groups being scored was always set to zero. The results showing how accurately each species was assigned in the cross-validation experiments are in Figure 3. Overall, 84.4% of the samples in this data set were correctly assigned to their species of origin.

In summary, the Aristotle Classifier, which was developed to distinguish glycosylation profiles, was modified and applied to a new problem: typing bacteria. In order to adapt the method to this new problem, two changes were made from the original description of the classifier: 1) Only the features for each sample were compared, not their ratios. This change was made because the bacterial data set has two orders of magnitude more features than glycan data sets. 2) A novel strategy for dealing with four different bacterial species was developed, as shown in Figure 3, and the strategy was effective at accurately assigning 84.4% of the samples in Group 7 in a five-fold cross validation. Additionally, we assessed the value of using a non-zero threshold to discriminate the two groups, but this approach did not offer a compelling benefit for assigning these particular data during a five-fold cross validation. Clearly, the Aristotle Classifier *can be applied* to bacterial typing. The more relevant question is *should it be applied?*

Comparison to current best practices

To address whether this approach offers advantages over existing methods, we compared it to the existing state-of-the-art. Figure 4 plots the percent of misclassified samples obtained when typing these 9 sets of bacteria using either the Aristotle Classifier or the state-of-the art benchmark protocol, where a prototypical spectrum is first generated from the training data then used for assigning the samples; the results for the benchmark method were reported in reference 24, Figure 3a. Based on the comparison shown in Figure 4, the Aristotle Classifier unquestionably stands out as a superior tool for bacterial typing at the species level. In six of the nine data sets tested, the Aristotle Classifier demonstrated a very substantial classification advantage. For example, the error rate for data set two plunged from over 20%

in the benchmark method to 2% for the classifier. Similarly, the Aristotle Classifier had less than half the error rate, compared to the existing state of the art, when scoring the very challenging data set 7, which has four different species types present. Furthermore, not a single example existed where the benchmark method outperformed the Aristotle Classifier; in the worst-case scenario, the two classifiers performed equally, on data sets 3 and 5.

Why did the Aristotle Classifier perform so well, compared to the benchmark method? The classifier is particularly well-suited to discriminating highly similar spectra. The level of spectral similarity was assessed earlier²⁴ by comparing each mass spectrum from the bacteria of a particular genus to the correct prototypical species spectrum and the prototype spectrum from the same genus but different species. The species of bacteria that are easier to score correctly have higher “separability.”²⁴ The three data sets that had the least separability (in other words, the data sets with samples that were most similar between the species to be classified) are data sets 1, 2, and 7.²⁴ In these three cases, the benchmark method had error rates of ~20–40%. By contrast, the Aristotle Classifier had two to 10-times fewer errors for these sets. In the three *most different* data sets, numbers 3, 5, and 9,²⁴ the Aristotle Classifier only outperformed the benchmark method in one case, set 9.

The fact that the Aristotle Classifier is particularly well-suited to discriminating highly similar data sets is not surprising to the developers of the method: The classifier was developed specifically to achieve this outcome. As the workflow in Figure 1 shows, the spectral features are not all treated equally: Those that discriminate the selected samples are used, while those that do not discriminate the selected samples are ignored. Furthermore, because the Aristotle Classifier selects four samples from each group to determine which peaks differentiate the samples, both the peak’s size difference between the two groups and the peaks’ variability within-each group are accounted for. By contrast, methods that use a single prototype spectrum cannot account for a peak’s variability within a given group.

What could be done to further improve upon these findings? Neither the sample sizes nor balance of the groups (the proportion in the minority class) appeared to negatively impact the classifier, so adding more samples to the groups will likely not lead to significant improvements in the outcome of the results. To improve upon these results, the next step may be to sub-divide the species into strains, since the hardest sets to assign correctly were those with high within-species variability. If enough samples of enough different strains were available, it may be possible to identify the bacteria at the strain level, instead of the species level; in this case, the problems arising from within-species variability would be removed, and identification might actually be easier in some cases. To test this hypothesis, we performed a preliminary proof-of-concept study on different strains from the same species.

Strain-specific analyses

We did a preliminary investigation into the potential of the classifier to be useful in assigning within-species strain types. In the MicroMass data set used in these studies, only one species had at least six example mass spectral files from two different strains: Both strains 126 and 188 are from the XSY species, which is in data set 6; and six examples are available for each strain. To gauge the utility of the classifier for strain-specific analyses, we attempted to

discriminate the 12 examples originating from these two strains using the same criteria as described for the other tests conducted herein. All twelve samples were correctly partitioned to their respective group, giving a 100% accuracy for this strain-specific comparison. (The classifier's output is shown in the TOC, the Table of Contents entry.) This proof-of-concept finding demonstrates the possibility that the classifier could be used to solve sub-species identification problems that are not yet tractable with other methods.

One important application of sub-species classification is the ability to determine whether or not a particular bacterial strain is antibiotic resistant; since the proof-of-concept study on sub-species classification was successful, an obvious next step would be to test the classifier's ability in determining whether or not a strain is resistant to antibiotics. Likewise, the classifier could be useful for enhancing the classification accuracy of complementary, MS-based bacterial screening methods that are now emerging, such as MRM profiling of lipids, which can be used to distinguish antibiotic resistance within given bacterial strains²⁸ or MALDI-MS analysis of glycolipid extracts, which is an alternative way to assess antibiotic resistance.²⁵ Both the MRM data sets and the glycolipid data sets typically have far fewer than 1300 features, but this should not be a limiting factor. Past work has shown that the classifier is effective even while using data sets with only 13 features;¹⁷ therefore, the classifier could be directly applied to both MRM data of bacterial lipids and MALDI data of glycolipids. In cases where <50 spectral features are present, the performance accuracy may be further boosted by considering ratios of the features, as we had demonstrated previously.¹⁷

In addition to applying this approach to other sample sets, we are determining the extent to which a large number of bacterial types can be distinguished from each other. Since past researchers have shown that distinguishing different genera is easier than distinguishing different species, we anticipate that all 20 species described herein (originating from nine different genera) could be discriminated effectively from each other using the MicroMass data set and the Aristotle Classifier. The optimal approach for applying the classifier to a group of 20 different species (or more) is currently under investigation. Even if it turns out that the Aristotle Classifier is not optimally effective at differentiating bacteria at the genus level, it would still be possible to characterize samples using a tiered approach, where a traditional classifier is used to assign the bacteria to the genus level and the Aristotle Classifier is used to complete the assignment, at the species level or perhaps even the phenotypic level (i.e. antibiotic-resistant or not.)

Conclusion

The Aristotle Classifier was adapted for classifying mass spectrometry data from highly similar bacterial species. Since the classifier was originally developed for glycomics data, several changes were necessary to optimally apply it to bacterial typing: First, MS peak abundances were used directly, instead of considering all possible peak ratios; second, a strategy was developed to apply the two-state classifier to a set of bacteria with four different species. Overall the method was demonstrated to be highly accurate and unbiased, with no drop in accuracy between training and cross validation studies. It outperformed the benchmark approach for assigning these data, and the most significant performance

enhancements were observed for groups of bacteria that had the least separability (highest similarity). Because of the dramatic performance enhancement of the classifier for assigning these data, and its demonstrated potential for correctly classifying strain-specific samples, the Aristotle Classifier is quite possibly an optimal tool for solving other outstanding pathogen classification problems where the fundamental challenge is to distinguish highly similar samples.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was funded by NIH grant R35GM136354 to HD. The authors are grateful to NIH for their support. We also acknowledge and appreciate Dr. Eden Go's contribution; she assisted with revisions.

References.

1. Cheng JZ; Ni D; Chou YH; Qin J; Tiu CM; Chang YC; Huang CS; Shen D; Chen CM Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Scientific Repts.* 2016, 6, 24454.
2. Nassif AB; Shahin I; Attili I; Azzeh M; Shaalan K Speech recognition using deep neural networks: A systemic review. *IEEE/Access*, 2019, 7, 19143–19165.
3. Liang M; Li Z; Chen T; Zeng J Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM Trans. Computational Biol. Bioinformatics*, 2005, 12, 928–937.
4. Trana NH; Zhanga X; Xinc L; Shanc B; Li M De novo peptide sequencing by deep learning. *Proc. Nat. Acad. Sci. USA* 2017, 114, 8247–8252. [PubMed: 28720701]
5. Kim M; Eetemadi A; Tagkopoulos I DeepPep: Deep proteome inference from peptide profiles. *PLoS Computational Biology*. 2017, 13, e1005661. [PubMed: 28873403]
6. Zhou XX; Zeng WF; Chi H; Luo CJ; Liu C; Zhan JF; He SM; Zhang ZF pDeep: Predicting MS/MS spectra of peptides with deep learning. *Anal. Chem* 2017, 89, 12690–12697. [PubMed: 29125736]
7. Zeng WF; Zhou XX; Zhou WJ; Chi H.; Zhan, J.F.; He, S. M. MS/MS spectrum prediction for modified peptides using pDeep2 trained by transfer learning. *Anal. Chem* 2019, 91, 9724–9731. [PubMed: 31283184]
8. Inglese P; McKenzie JS; Mroz A; Kinross J; Veselkov K; Holmes E; Takats Z; Nicholson JK; Glen RC Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer. *Chem. Sci* 2017, 8, 3500–3511. [PubMed: 28507724]
9. Shen H; Duhrkop K; Bocker S; Rousu J Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics*, 2014, 30(12), i157–i64. [PubMed: 24931979]
10. Berg P; McConnell EW; Hicks LM; Popescu SC; Popescu GV Evaluation of linear models and missing value imputation for the analysis of peptide-centric proteomics. *BMC Bioinformatics*, 2019, 20 (Suppl 2), 102.
11. Liu S; Cheng L; Fu Y; Liu BF; Liu X Characterization of IgG N-glycome profile in colorectal cancer progression by MALDI-TOF-MS. *Journal of Proteomics*, 2018, 181, 225–237. [PubMed: 29698801]
12. Russell AC; Simurina M; Garcia MT; Novokmet M; Wang Y; Rudan I; Campbell H; Lauc G; Thomas MG; Wang W The N-glycosylation of immunoglobulin G as a novel biomarker of Parkinson's disease. *Glycobiology*, 2017, 27(5), 501–510. [PubMed: 28334832]
13. Colhoun HO; Treacy EP; MacMahon M; Rudd PM; Fitzgibbon M; O'Flaherty R; Stepein KM Validation of an automated ultraperformance liquid chromatography IgG N-glycan analytical

method applicable to classical galactosaemia. *Ann. Clin. Biochem* 2018, 55(5), 593–603. [PubMed: 29444593]

14. Tan Z; Yin H; Nie S; Lin Z; Zhu J; Ruffin MT; Anderson MA; Simeone DM; Lubman DM Large-scale identification of core-fucosylated glycopeptide sites in pancreatic cancer serum using mass spectrometry. *J. Proteome Res.* 2015, 14(4), 1968–1978. [PubMed: 25732060]
15. Lee J; Hua S; Lee SH; Oh MJ; Yun J; Kim JY; Kim JH; Kim JH; An HJ Designation of fingerprint glycopeptides for targeted glycoproteomic analysis of serum haptoglobin: insights into gastric cancer biomarker discovery. *Anal. Bioanal. Chem* 2018, 410, 1617–1629. [PubMed: 29285644]
16. Lundstrom SL; Hensvold AH; Rutishauser D; Klareskog L; Ytterberg AJ; Zubarev RA; Catrina AI IgG Fc galactosylation predicts response to methotrexate in early rheumatoid arthritis. *Arthritis Research & Therapy*, 2017, 19(1), 182. [PubMed: 28793911]
17. Hua D; Patabandige MW; Go EP; Desaire H The Aristotle Classifier: Using the whole glycomic profile to indicate a disease state. *Anal. Chem.* 2019, 91(17), 11070–11077. [PubMed: 31407893]
18. Florio W; Tavanti A; Barnini S; Ghelardi E; Lupetti A Challenges in the diagnosis of microbial infections by MALDI-TOF mass spectrometry. *Front. Microbiol* 2018, 9, 1097.
19. Clark AE; Kaleta EJ; Arora A; Wolk DM Matrix-assisted laser desorption ionization-time of flight mass spectrometry: a fundamental shift in the routine practice of clinical microbiology. *Clin. Microbiol. Rev* 2013, 26 (3), 547–603. [PubMed: 23824373]
20. Mather CA; Rivera SF; Butler-Wu SM Comparison of the Bruker Biotyper and Vitek MS matrix-assisted laser desorption ionization-time of flight mass spectrometry systems for identification of mycobacteria using simplified protein extraction protocols. *J. Clin. Microbiol* 2014, 52, 130–138. [PubMed: 24172150]
21. Pence MA; McElvania TeKippe E; Wallace MA; Burnham CAD Comparison and optimization of two MALDI-TOF MS platforms for the identification of medically relevant yeast species. *Eur. J. Clin. Microbiol. Infect. Dis* 2014, 33, 1703–1712. [PubMed: 24800928]
22. Marin M; Cercenado E; Sanchez-Carrillo C; Ruiz A; Gonzalez AG; Rodriguez-Sanchez B; Bouza E Accurate differentiation of *Streptococcus pneumoniae* from other species within the streptococcus mitis group by peak analysis using MALDI-TOF MS. *Front. Microbiol* 2017, 8, 698. [PubMed: 28487677]
23. Carrasco G; Caballero JD; Garrido N; Valdezate S; Canton R; Saez-Nieto JA Shortcomings of the commercial MALDI-TOF MS database and use of MLSA as an arbiter in the identification of *Nocardia* species. *Front. Microbiol* 2016, 7, 542. [PubMed: 27148228]
24. Mahe P; Arsac M; Chatellier S; Monnin V; Perrot N; Mailler S; Girard V; Ramjeet M; Surre J; Lacroix B; van Belkum A; Veyrieras JB Automatic identification of mixed bacterial species fingerprints in a MALDI-TOF mass-spectrum. *Bioinformatics*, 2014, 30 (9), 1280–1286. [PubMed: 24443381]
25. Fondrie WE; Liang T; Oyler BL; Leung LM; Ernst RK; Strickland DK; Goodlett DR Pathogen identification direct from polymicrobial specimens using membrane glycolipids. *Scientific Repts.* 2018, 8, 15857.
26. Yang Y; Lin Y; Qiao L Direct MALDI-TOF MS identification of bacterial mixtures. *Anal. Chem.* 2018, 90, 10400–10408. [PubMed: 30091898]
27. Dentamaro V; Impedovo D; Pirlo G LICIC: Less Important Components for Imbalanced Multiclass Classification. *Information*, 2018, 9, 317.
28. Xie Z; Gonzalez LE; Ferreira CR; Vorsilak A; Frabutt D; Sobreira TJP; Pugia M; Cooks RG Multiple Reaction Monitoring Profiling (MRM-Profilig) of lipids to distinguish strain-level differences in microbial resistance in *Escherichia coli*. *Anal. Chem* 2019, 91(17), 11349–11354. [PubMed: 31398004]

Figure 1.
Description of the scoring algorithm of the Aristotle Classifier.

Figure 2.

Three examples of classification results. (A) Output from data set 2. (B) Output from data set 5. (C) Output from data set 4. The dots in green fields indicate samples that were correctly scored by the classifier when the threshold is set to zero. The dots in red fields indicate misclassifications. The dotted lines in A and C indicate a non-zero threshold that would optimally separate the training data.

Figure 3.

Protocol to score data set 7, with four different species present, using the Aristotle Classifier.

The data are assigned based on the outcome of three separate tests, as shown in the figure.

The overall accuracy was 84.4%.

Figure 4.

Comparison of misclassified samples in 9 data sets. The ability to correctly assign the samples within a genus but from different species is shown for 9 different data sets. The results of the benchmark analysis (orange bars) are originally provided in reference 24, Figure 3a. The Aristotle Classifier's performance is shown with blue bars. The Aristotle Classifier outperforms the benchmark method, especially in cases where the samples are difficult to assign.

Table 1:

Data sets Used to Discriminate Bacterial Species

Data set #	Genus	Species	# each	% smallest group
1	QWP	DRH	26	29.7
		LRO	11	
2	VVJ	KSF	29	48.3
		KWJ	31	
3	QBG	CRP	14	46.7
		KGI	16	
4	RTO	JFR	26	48.0
		TQH	24	
5	JNH	FLH	28	48.1
		ZIJ	26	
6	NYV	VCE	52	35.8
		XSY	29	
7	BUT	BIK	18	12.8
		DNW	32	
		TRH	31	
		YZE	60	
8	EMD	FZO	27	42.6
		WXC	20	
9	AUG	AEX	50	29.6
		HSS	21	

Table 2:

Percent Accuracy for Typing Bacteria

Data set	Training Accuracy:		Cross-Validation Accuracy	
	Cut at Zero	Optimize Cut Point	Cut at Zero	Optimize Cut Point
1	78	78	76	68
2	98	100	98	100
3	97	100	97	100
4	84	90	84	88
5	91	91	91	91
6	94	94	94	91
8	98	100	98	98
9	99	100	100	97
Total Overall	93	95	93	92