analytical chemistry

Article

pubs.acs.org/ac

Automatic Spectroscopic Data Categorization by Clustering Analysis (ASCLAN): A Data-Driven Approach for Distinguishing Discriminatory Metabolites for Phenotypic Subclasses

Xin Zou,[†,‡] Elaine Holmes,[§,∥] Jeremy K Nicholson,[§,∥] and Ruey Leng Loo*[,‡,§]

[†]Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China

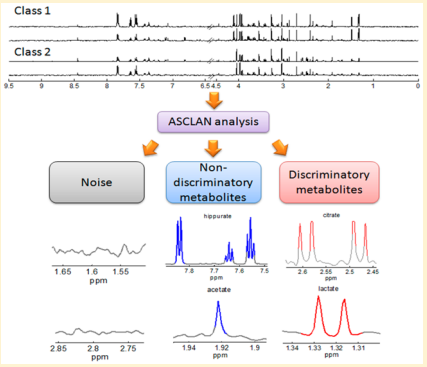
[‡]Medway Metabonomics Research Group, Medway School of Pharmacy, Universities of Kent and Greenwich, Chatham Maritime, Kent, ME4 4TB, U.K.

[§]Section of Biomolecular Medicine, Division of Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College London, London SW7 2AZ, U.K.

[∥]MRC-NIHR Phenome Centre, Imperial College London, London SW7 2AZ, U.K.

Supporting Information

ABSTRACT: We propose a novel data-driven approach aiming to reliably distinguish discriminatory metabolites from nondiscriminatory metabolites for a given spectroscopic data set containing two biological phenotypic subclasses. The automatic spectroscopic data categorization by clustering analysis (ASCLAN) algorithm aims to categorize spectral variables within a data set into three clusters corresponding to noise, nondiscriminatory and discriminatory metabolites regions. This is achieved by clustering each spectral variable based on the r^2 value representing the loading weight of each spectral variable as extracted from a orthogonal partial least-squares discriminant (OPLS-DA) model of the data set. The variables are ranked according to r^2 values and a series of principal component analysis (PCA) models are then built for subsets of these spectral data corresponding to ranges of r^2 values. The Q^2X value for each PCA model is extracted. K-means clustering is then applied to the Q^2X values to generate two clusters based on minimum Euclidean distance criterion. The cluster consisting of lower Q^2X values is deemed devoid of metabolic information (noise), while the cluster consists of higher Q^2X values is then further subclustered into two groups based on the r^2 values. We considered the cluster with high Q^2X but low r^2 values as nondiscriminatory, while the cluster with high Q^2X and r^2 values as discriminatory variables. The boundaries between these three clusters of spectral variables, on the basis of the r^2 values were considered as the cut off values for defining the noise, nondiscriminatory and discriminatory variables. We evaluated the ASCLAN algorithm using six simulated ^{1}H NMR spectroscopic data sets representing small, medium and large data sets (N = 50, 500, and 1000 samples per group, respectively), each with a reduced and full resolution set of variables (0.005 and 0.0005 ppm, respectively). ASCLAN correctly identified all discriminatory metabolites and showed zero false positive (100% specificity and positive predictive value) irrespective of the spectral resolution or the sample size in all six simulated data sets. This error rate was found to be superior to existing methods for ascertaining feature significance: univariate t test by Bonferroni correction (up to 10% false positive rate), Benjamini–Hochberg correction (up to 35% false positive rate) and metabolome wide significance level (MWSL, up to 0.4% false positive rate), as well as by various OPLS-DA parameters: variable importance to projection, (up to 15% false positive rate), loading coefficients (up to 35% false positive rate), and regression coefficients (up to 39% false positive rate). The application of ASCLAN was further exemplified using a widely investigated renal toxin, mercury II chloride ($HgCl_2$) in rat model. ASCLAN successfully identified many of the known metabolites related to renal toxicity such as increased excretion of urinary creatinine, and different amino acids. The ASCLAN algorithm provides a framework for reliably differentiating discriminatory metabolites from nondiscriminatory metabolites in a biological data set without the need to set an arbitrary cut off value as applied to some of the conventional methods. This offers significant advantages over existing methods and the possibility for automation of high-throughput screening in "omics" data.



Metabolic profiling of biological samples using proton nuclear magnetic resonance (^{1}H NMR) spectroscopy and mass spectrometry (MS) generates complex metabolic phenotypes that can be mined to uncover important biological information about the biological system.[1,2] One of the critical

Received: October 23, 2015
Accepted: May 5, 2016
Published: May 5, 2016

DOI: 10.1021/acs.analchem.5b04020
Anal. Chem. 2016, 88, 5670–5679

aspects of data mining is to accurately extract the important spectral features that contribute to metabolic distinctions between biological classes using multivariate data analysis techniques, such as orthogonal partial least squares-discriminatory analysis (OPLS-DA).[3] Typically, this involves the use of multivariate parameters that provide a measurement of the relative contribution of each spectral variable to the class separation within a data set, such as loading weights,[4,5] variable importance in the projection (VIP),[6−8] loading coefficients,[9] and regression coefficients.[1]

In recent years, metabolic profiling studies have been applied to complex data sets, such as those generated from large-scale epidemiological cohorts investigating risk of cardiovascular diseases[1,10] and cancers[11] or to investigate metabolic variation in response to drug treatment in terms of drug toxicity,[12] drug metabolism,[13] and drug effect[14] in both human[13,15] and animal[11] studies. Particularly in the case of human studies, the diversity in genetic and environmental background has been shown to make the discovery of genuine discriminatory metabolites for disease state or response to therapeutic intervention challenging in terms of selecting true biomarkers from biological noise. A series of statistical spectroscopic correlation techniques, such as statistical total correlation spectroscopy (STOCSY),[16] statisical heterospectroscopy (SHY),[17] and subset optimization by reference matching (STORM)[18] have proven useful for enhancing biomarkers recovery from data with inherent variation. However, although these algorithms can identify correlated structures from subsets of samples within a model, thereby aiding structural elucidation of biomarkers, compositional variability relating to substructure within sample classes still proves to be problematic. More recently, a variant of statistical spectroscopic techniques, statistical homogeneous cluster spectroscopy (SHOCSY)[19] in combination with OPLS-DA has been developed to ensure reliable biomarker recovery in data sets that show a high degree of variation in response or a dichotomized response.

An increasingly common strategy for assigning significance to metabolic features in spectral data sets is to combine univariate approaches with multivariate approaches. For example, top ranked spectral features extracted from discriminant analyses are subjected to a t test analysis corrected for multiple testing using methods, such as Bonferroni correction to control the family wise error rate (FWER),[1,24] or Benjamini−Hochberg correction to control the false discovery rate (FDR).[7] Neither of these approaches are optimal for spectroscopic data sets; Bonferroni correction is often deemed to be too stringent for metabolic profiling studies due to the colinearity of the metabolic signatures, which serves to increase the false negative discovery rate,[20] while Benjamini−Hochberg correction has been shown to increase the false positive discovery rate. Chadeau-Hyam et al.[21] proposed the metabolome wide significance level (MWSL) approach to control for the family wise error rate by the use of permutation testing. Unlike the Bonferroni and Benjamini−Hochberg methods, the MWSL approach calculates a cutoff p-value to differentiate discriminatory from nondiscriminatory spectral variables based on the given data set. In terms of the performance, the MWSL has been shown to generate comparable results to the Bonferroni correction approach. However, to date, there has been no general consensus in the data analysis strategy for defining rules for the selection of discriminatory metabolites. Given the increasing number of application of metabolic profiling studies aiming to define biological or clinical phenotypes and to identify potential discriminatory biomarkers reflecting those phenotypes, a reliable and objective method for biomarker selection is required.

Here, we propose automatic spectroscopic data categorization by clustering analysis (ASCLAN), a novel data-driven pipeline aiming to objectively extract discriminatory metabolic signatures based on the strength of association of each variable with the biological classes. The ASCLAN algorithm is based on first constructing an OPLS-DA model to achieve optimal differentiation of the biological classes and establish correlation coefficients between each spectral variable and the OPLS-DA model scores (r^2). The r^2 value thus gives the relative contribution of a given spectral variable to the class separation within the data set. The ASCLAN algorithm subsequently creates an additional feature (Q^2X), by building PCA models corresponding to different subsets of spectral variables with different r^2 values each at a 0.1 increment. These metrics of spectral features (r^2 and Q^2X) are then used to categorize the spectral variables, by K-means clustering analysis in a two-step fashion, into three groups corresponding to noise, nondiscriminatory metabolites and discriminatory metabolites/candidate biomarkers. The cluster of spectral variables with low Q^2X and r^2 are considered noise; while cluster with high Q^2X but low r^2 values are considered nondiscriminatory metabolites; and cluster with both high r^2 and Q^2X values are discriminatory metabolites/biomarkers.

The ASCLAN algorithm was evaluated using six simulated ^{1}H NMR spectra data sets with different sample sizes and spectral variable resolutions and was subsequently exemplified using a biological data set corresponding to a mercury II chloride renal toxicity study in a rat model.

MATERIALS AND DATA ANALYSIS

Data Sets. *Simulated ^{1}H NMR Spectral Data.* To evaluate the ASCLAN algorithm for its ability to accurately differentiate discriminatory variables from nondiscriminatory variables from spectral data sets, six simulated data sets were generated with different sample sizes and spectral variable resolutions designed to emulate paraquat-induced renal toxicity as an exemplar disease condition. These simulated spectral data sets were designed to represent a two class problem, corresponding to urine samples from a toxic versus control state. Within each simulated data set, the class representing paraquat toxicity contained higher signal intensities representing lactate (δ1.32, doublet (d), δ 4.10, quartet (q)), and L-alanine (δ 1.46, d, δ 3.76, q) and lower signal intensities for creatinine (δ 3.05, singlet (s), δ 4.05, s) and citrate (δ 2.53, dd, δ 2.65, dd) when compared to the control class. Sample sizes of 50, 500, and 1000 spectra in each class were chosen to represent typical studies of small, medium, and large metabolic profiling data sets, respectively. For each of these data sets, we generated two levels of spectral resolutions 0.005 and 0.0005 ppm to represent a lower and full resolution data sets, respectively. The lower resolution data sets consisting of 2000 spectral variables, of which 45 spectral variables constitute the four discriminatory metabolites (alanine, lactate, creatinine, and citrate). The reduced resolution data sets were generated by binning the full resolution data sets with 20 000 spectral variables (consist of 466 discriminatory variables). The means and variances of the signal intensities for these metabolites are shown in the Table S-1. The concentrations of the remaining metabolites were simulated using the software default parameters for the whole data set introducing nonsystematic variance across the

DOI: 10.1021/acs.analchem.5b04020
Anal. Chem. 2016, 88, 5670−5679

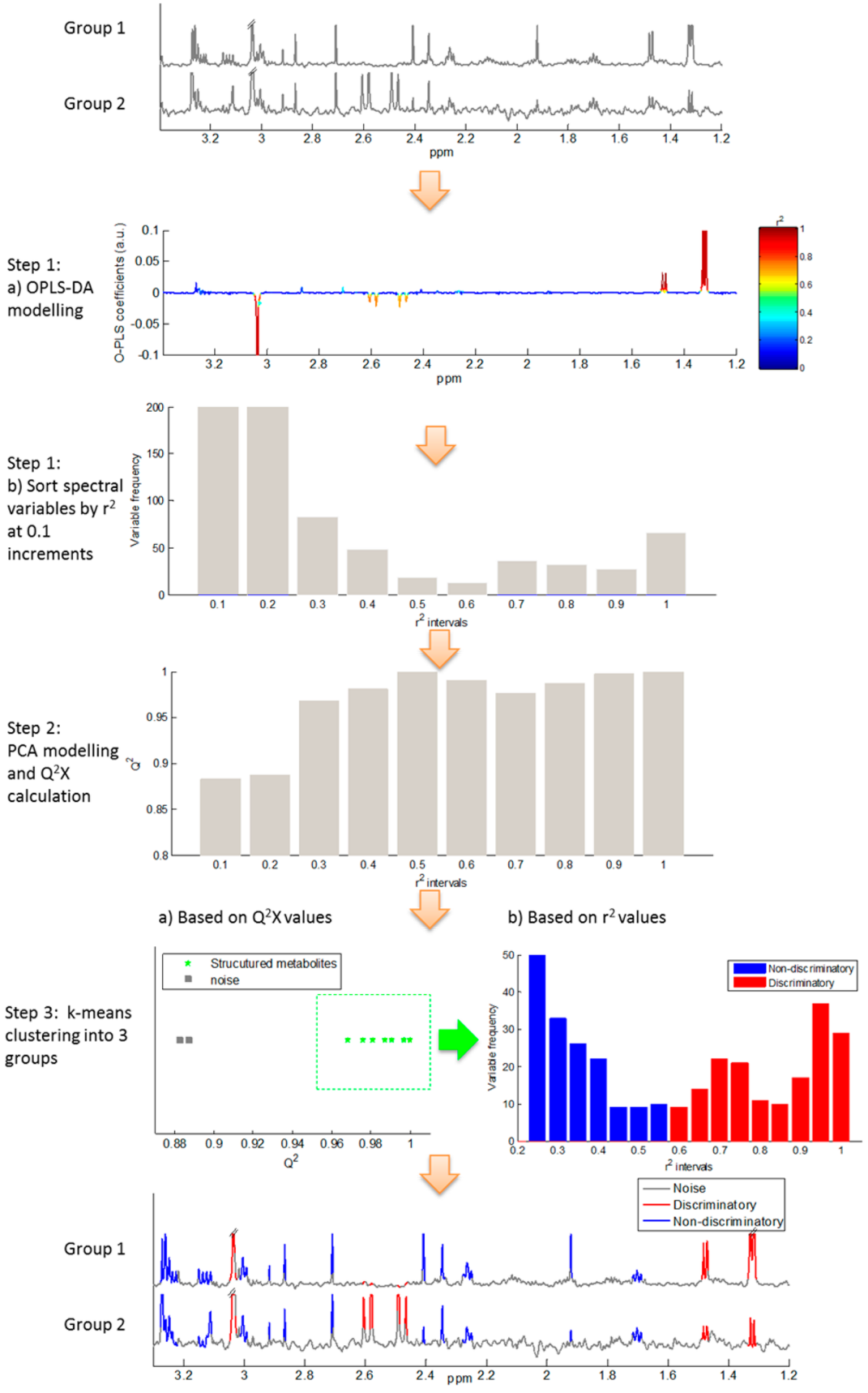


Figure 1. Schematic diagram of the ASCLAN algorithm for a data set consisting of two biological classes.

data set. The simulated spectra covered a chemical shift region from δ 0 to δ 10 and the software default parameters were employed to generate random peak shift between spectra (with a standard deviation set to 0.05 ppm at pH 7.4) to mimic true

DOI: 10.1021/acs.analchem.5b04020
Anal. Chem. 2016, 88, 5670−5679

natural variation in the data set. These data set were simulated by MetAssimulo software[22] using ^{1}H NMR metabolite signals information from the Human Metabolome DataBase (HMDB).[23]

Renal Toxicity Model. Mercury II chloride ($HgCl_2$) is widely used model of acute renal proximal tubular toxicity.[24] The accumulation of Hg^{2+} in the renal tubules can lead to extensive damage to the proximal tubular epithelium and renal failure associated with low molecular weight proteinuria, calciuria, phosphaturia, and general amino aciduria.[24−26] A data set consisting of ^{1}H NMR urine spectra acquired at 600 MHz from Sprague−Dawley rats (N = 10) 24 h after treatment with a dose of $HgCl_2$ at 0.75 mg/kg in 0.9% saline and the predose were used. Previous studies have shown that $HgCl_2$ toxicity is characterized by lower urinary excretion of citrate, hippurrate, succinate and 2-oxoglutarate and increased excretion of glucose, organic acids (3-D hydroxybutyrate, lactate) and amino acids (e.g., valine, alanine).[24,27−29] The animal husbandry and spectral acquisition parameters have been detailed in previous publications.[24] In this study, we compared the urine samples collected predose and 24 h post administration of $HgCl_2$.

NMR Spectroscopy. An aliquot of 400 μL from each urine sample was added to 200 μL of sodium phosphate buffer (0.2 M Na_2HPO_4 in H_2O and 0.2 M NaH_2PO_4 in 80:20 H_2O:D_2O, pH 7.4) containing 1 mM sodium 3-trimethylsilyl-[2,2,3,3-$^2H_4$]-propionate (TSP) and 3 mM sodium azide. Samples were centrifuged at ∼1800 g for 5 min to remove any solid debris. ^{1}H NMR spectra of urine were acquired using a Bruker AVANCE 600 MHz spectrometer and were measured at 300 K using a Bruker flow-injection system. A standard pulse sequence using the first increment of a NOE sequence to achieve irradiation of the water frequency during the mixing time and relaxation delay was employed. The total spectral acquisition time was ∼4 min per urine sample.

Pretreatment of NMR Spectra. ^{1}H NMR urine spectra were phased and baseline corrected and referenced to the chemical shift of TSP at δ 0.0. The region between water and urea resonances δ 4.5−δ 7.0 was excluded, leaving chemical shift regions between δ 0.7−4.5 and δ 7.0−9.0, giving a total of 9494 variables and a resolution of 0.0006 ppm. Spectra with poor water suppression and distorted baselines were identified visually and subsequently excluded from subsequent analysis. The resulting numbers of samples involved in the analysis were N = 7 for baseline (t_0) and N = 8 for 24 h postdose (t_{24}).

Automatic Spectroscopic Data Categorization by Clustering Analysis (ASCLAN). *Algorithm Developed Consisted of Three Steps. Step 1: OPLS-DA Modeling.* An OPLS-DA model was calculated for each data set and the model was considered to be valid when the 7-fold cross validation (CV) Q^2Y statistic, providing a measure of predictivity of the model, was significantly higher than the Q^2Y obtained by a permutation test based on 100 iterations (p < 0.05). For models returning a valid Q^2Y statistic, the loading weights, r^2, was calculated for each spectral variable. The r^2 was calculated as

$$r^2 = \left(\frac{t^{\mathrm{T}}X_i}{s_t s_{X_i}}\right)^2$$

where s_t and s_{X_i} are the standard deviations of score vector t and spectral variable X_i, respectively. T indicates the transpose of the vector. The r^2 values of all spectral variables were then normalized giving values of zero to one, where one represents maximal contribution of the variable relating to the class of the samples.

The spectral data were then sorted according to the r^2 values for each spectral variable and ranked from lowest to highest in 0.1 increments giving a total of 10 spectral variable subsets. Thus, groups of variables within the subsets corresponding to the higher r^2 values were more likely to be associated with class of toxicity.

Step 2: PCA Modeling. A separate PCA model was constructed for each of the spectral variable subsets previously sorted according to the r^2 values. For each PCA model (corresponding to r^2 increments of 0.1), a 7-fold cross validation statistic, Q^2X value was calculated. Typically, the PCA models consist of subset of spectral variables with low r^2 values, tend to generate a low Q^2X value as these spectral variables usually consist of random noise or variables that lack metabolic information. Conversely, the calculated Q^2X values for the PCA models tend to improved when the PCA models are generated using spectral variables that are rich in metabolic information relating to a systemic perturbation. Thus, after step 1 and 2 described above, each spectral variable is effectively characterized by an r^2 value, corresponding to the loading weight as obtained from the OPLS-DA model, and a Q^2X value representing the 7-fold cross validation statistics as extracted from the PCA model.

Step 3: Assignment of Discriminatory Metabolites by K-Means Clustering. K-means clustering[30] of the "proxy" spectral features (in terms of Q^2X and r^2 values) was performed by minimum Euclidean distance criterion in a two-step approach. Initially, the spectral variables were clustered into two groups based on the Q^2X values. The cluster that consists of spectral variables showing lower Q^2X (and typically low r^2) was deemed to be devoid of any metabolic information relating to the class of sample. These spectral variables were considered to correspond to noise regions within the spectra and were excluded from further analysis. The cluster that consists of spectral variables showing higher Q^2X values was then further subclustered into two groups based on the r^2 value of each variable. The application of this K-means clustering approach thus enabled the grouping of spectral variables based on the inherent metabolic features of the spectral variable, rather than relying on the user to arbitrary defining a cutoff value, into three clusters: (i) noise regions of the spectra showing low Q^2X and r^2 values; (ii) nondiscriminatory metabolites, consisting of spectral variables corresponding to structured signals that do not contribute to the differentiation between classes, these variables often show high Q^2X but low r^2 values; and (iii) discriminatory metabolites, these spectral variables consist of structured signals, which contribute to the differentiation between classes. These discriminatory metabolites tend to show high Q^2X and r^2 values. The boundaries of these three clusters, based on the r^2 values, were considered as the inherent cut off values for noise, nondiscriminatory and nondiscriminatory metabolites regions.

A schematic diagram describing the ASCLAN algorithm is shown in Figure 1. All calculations and the ASCLAN algorithm were written in MATLAB (R2012a, Mathworks, Natick, USA) environment. The ASCLAN encrypted matlab code is available in Supporting Information. The authors will also illustrate how to use the ASCALN code in more details in author's webpage.

Data Analysis. All spectral data from both the simulated data sets and the $HgCl_2$ data set were normalized by the

DOI: 10.1021/acs.analchem.5b04020
Anal. Chem. 2016, 88, 5670−5679

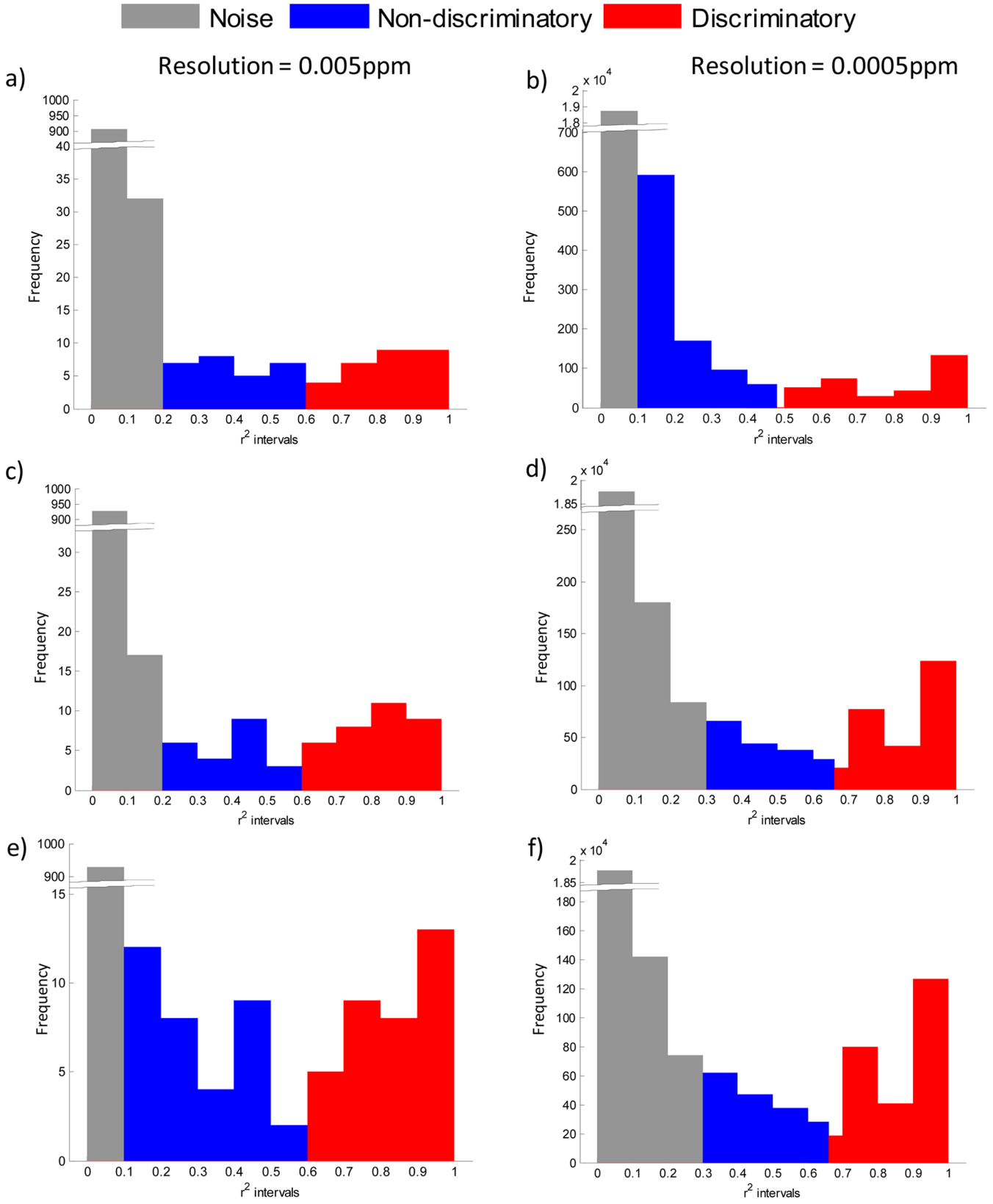


Figure 2. Distribution of loading coefficients, r^2, indicating noise, nondiscriminatory and discriminatory metabolites for simulated data sets with reduced spectral resolution (panel on the left) and full spectral resolution (panel on the right) and for data set sizes of 50 samples per group (a and b), 500 per group (c and d), and 1000 per group (e and f).

probabilistic quotient normalization (PQN)[31] method using the median spectrum as a reference. The spectroscopic data were mean centered and scaled to unit variance prior to any data analysis. The ASCLAN algorithm was applied to distinguish nondiscriminatory and discriminatory spectral variables in each simulated ^{1}H NMR data set. The ability for

DOI: 10.1021/acs.analchem.5b04020
Anal. Chem. 2016, 88, 5670−5679

Table 1. Performance Metrics for Sensitivity, Specificity, and Positive Predictive Value (PPV), Expressed in %, Comparing ASCLAN Method to Conventional Methods Based on OPLS-DA Parameters and Univariate *t*-Testing Adjusting for Multiple Correction Testing

		reduced resolution (0.005 ppm)			full resolution (0.0005 ppm)		
method	performance	N = 50 per group	N = 500 per group	N = 1000 per group	N = 50 per group	N = 500 per group	N = 1000 per group
ASCLAN	sensitivity	93.3	82.2	77.8	72.3	56.7	57.3
	specificity	100	100	100	100	100	100
	PPV	100	100	100	100	100	100
VIP	sensitivity	100	100	100	99.8	99.8	99.4
	specificity	87.2	95.4	95.7	83.7	93.6	94.3
	PPV	27.0	50.6	52.3	12.8	27.0	29.6
loading coefficient	sensitivity	100	100	100	99.8	1	99.8
	specificity	84.6	65.8	56.3	87.3	71.1	62.2
	PPV	23.4	12.1	9.7	15.8	7.4	6.0
regression coefficient	sensitivity	100	100	100	99.8	1	99.8
	specificity	83.9	64.9	55.6	86.1	70.0	61.1
	PPV	22.6	11.8	9.6	14.6	7.4	5.8
Benjamini–Hochberg	sensitivity	100	100	100	99.8	100	99.8
	specificity	87.9	73.2	65.5	87.6	72.7	64.6
	PPV	16.0	7.9	6.3	16.1	8.0	6.3
Bonferroni	sensitivity	100	100	100	94.9	99.8	99.6
	specificity	98.6	92.9	88.0	99.2	94.4	89.6
	PPV	61.6	24.5	16.7	74.2	29.8	18.6
MWSL	sensitivity	100	93.3	97.8	86.7	97.1	88.9
	specificity	99.5	99.7	99.7	100	100	99.9
	PPV	83.3	87.5	89.8	93.3	92.3	93.5

the algorithm to correctly identify discriminatory metabolites was compared to the existing methods using various parameters commonly used to show the importance of a variable within the OPLS-DA model. This includes the VIP, loading coefficient and regression coefficient. The VIP for each spectral variable is calculated by

$$\mathrm{VIP}_j = \sqrt{\frac{J}{\sum_{p=1}^{P} \mathbf{SS}_p(Y)} \sum_{p=1}^{P} r_{pj}^{\,2} \mathbf{SS}_p(Y)}$$

where J is the number of spectral variables, P is the number of correlated variables in spectral data, $r_{pj}^{\,2}$ is the loading weight of the pth latent variable of the jth spectral variable, $\mathbf{SS}_p$ is the percentage of dummy matrix Y explained by the pth latent variable. Here, we considered a spectral variable as discriminatory when the VIP value was greater than one.[32] For methods based on the OPLS-DA loading coefficients and regression coefficients, the significance of a spectral variable was estimated by jack-knifing resampling method as described in Wiklund et al.[9] A spectral variable was considered to be discriminatory if the confidence interval of the loading coefficient and regression coefficient did not include zero. In addition, we also compared the accuracy in differentiating discriminatory metabolites using univariate t test analysis adjusting for multiple comparisons by Bonferroni correction, Benjamini–Hochberg correction and metabolome wide significance level (MWSL) methods. The Bonferroni correction considered a spectral variable as a discriminatory metabolite when the p-value <0.05/total number of spectral variables. For Benjamini–Hochberg correction, the spectral variables were sorted according to their p-values in an ascending order. The p-value for each spectral variable was adjusted by (0.05× rank of the variable)/total number of spectral variables. The p-value for MWSL was obtained by permutation and was calculated by $(0.05 \times M)$th smallest value in Q, where M is the number of permutations, and $Q = (q_1, q_2, \ldots q_M)$, where q corresponds to the smallest p-value of all spectral variables obtained by a permutation.

The ability of the ASCLAN approach in reliably extracting and differentiating the four discriminatory metabolites (lactate, alanine, citrate, and creatinine) from noise or nondiscriminatory metabolites was initially assessed using six simulated data sets. The accuracy of the ASCLAN method was then compared to the six conventional approaches mentioned above. In addition, the overall performance was also evaluated by the percentage of sensitivity, specificity and the positive prediction value (PPV), and these were calculated using the equations as shown below

$$\text{sensitivity} = \frac{\mathrm{TP}}{P} \times 100\%$$

$$\text{specificity} = 1 - \frac{\mathrm{FP}}{N - P} \times 100\%$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\%$$

where true positive (TP) corresponds to the spectral variables correctly identified as discriminatory metabolites and false positive (FP) indicates spectral variables incorrectly identified as discriminatory metabolites. P is the number of spectral variables for the discriminatory metabolites and N the total number of spectral variables within the data set. In the simulated data sets with reduced resolution, where N = 2000 and P = 45 (consisting of 16 spectral variables for lactate, 9 for alanine, 10 each for citrate and creatinine) and, for full resolution, N = 20 000 and P = 466 (consisting of 156 spectral variables for lactate, 96 for alanine, 108 for citrate and 106 for creatinine). PPV indicates the probability of correctly

DOI: 10.1021/acs.analchem.5b04020
Anal. Chem. 2016, 88, 5670–5679

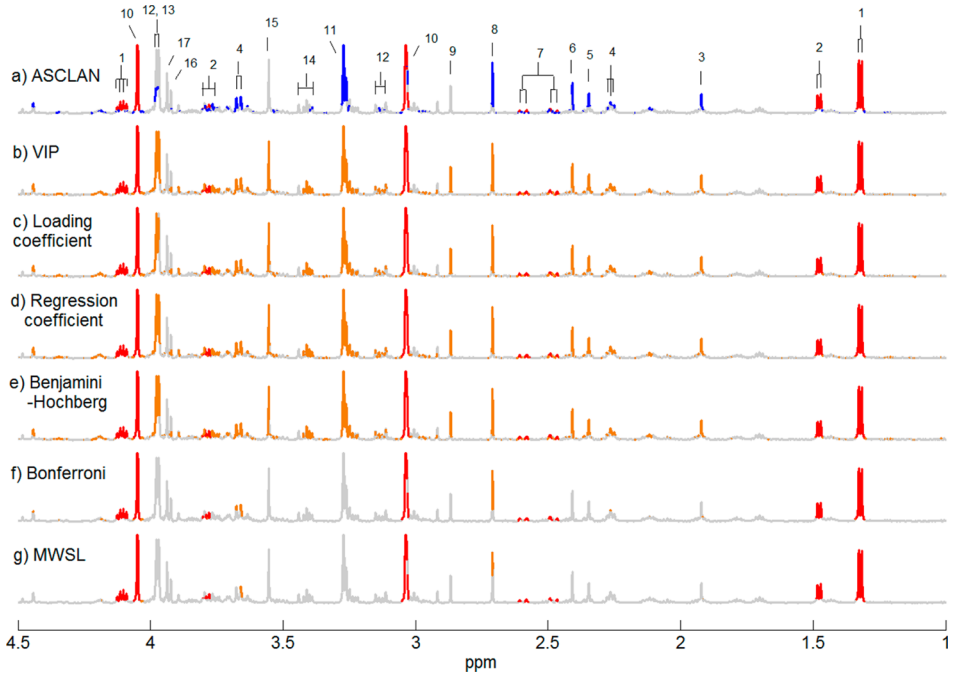


Figure 3. Spectra color coded by the ability of (a) ASCLAN, (b) VIP, (c) loading coefficient, (d) regression coefficient, (e) Benjamini–Hochberg, (f) Bonferroni, and (g) metabolome wide significance level, MWSL methods to correctly distinguish discriminatory metabolites from nondiscriminatory metabolites based on $N = 50$ spectral in each group. These full-resolution median spectra are color coded as follows: red for discriminatory metabolites, brown for false positive, and gray for nondiscriminatory variables including noise for all methods, except for ASCLAN where the noise regions (gray) are distinguished from signals representing nondiscriminatory metabolites (blue). In the simulated data sets, the discriminatory metabolites show increased signal intensities of lactate and L-alanine and a reduced creatinine and citrate in the paraquat toxicity group. Key: 1, lactate; 2, L-alanine; 3, acetate; 4, phenylactylglutamine; 5, *p*-cresol sulfate; 6, succinate; 7, citrate; 8, dimethylamine; 9, trimethylamine; 10, creatinine; 11, trimethylamine-*N*-oxide; 12, L-histidine, 13, hippurate; 14, taurine; 15, glycine; 16, creatine; 17, glycolic acid.

identifying the spectral variables that are discriminatory metabolites for the data sets.

RESULTS AND DISCUSSION

Performance of ASCLAN for Identifying Discriminatory Metabolites Compared to Standard Methods Using Simulated Data Sets. Six simulated data sets with different sample sizes and spectral resolutions were used to assess the ability of ASCLAN to correctly identify spectral variables as discriminatory metabolites. The OPLS-DA models constructed for all simulated data sets showed all the models were valid with Q^2Y statistics >0.85, and permutation test p-value $<10^{-5}$ (Table S-2). Using the boundaries between each K-means cluster for noise, nondiscriminatory metabolites and discriminatory metabolites, we found the r^2 cutoff values for discriminatory metabolites varied from data set to data set and were affected by the resolution of the data set, particularly for small data sets ($N = 50$), where the r^2 cut off values for discriminatory metabolites was 0.60 for reduced resolution and 0.49 for full resolution, Figure 2. As the sample size in each group increased, the r^2 cutoff values for discriminatory metabolites remained stable, with r^2 cutoff values of 0.60, for data sets with reduced resolution. However, for full resolution data sets the r^2 cutoff values slightly increased to 0.66 for the medium and large data sets. This demonstrates that the ASCLAN approach is versatile and capable of defining optimal cut off values based on inherent features within the data sets Based on the results from the six simulated data sets, we found the ASCLAN approach were able to differentiate discriminatory signals from nondiscriminatory signals, Table S-3. It can be seen that the ASCLAN method correctly considered the majority of the signal variables associated with the four discriminatory metabolites from the data sets with reduced resolution: for lactate ($N = 16$), ASCLAN correctly identified all 16 variables for the small data set; 11 variables for the medium size data set and 12 variables for the large data set; for alanine ($N = 9$), all 9 variables for the small data set; 8 variables for the medium size data set and 6 variables for the large data set; for citrate ($N = 10$), 9, 8, and 8 variables for the small, medium and large data sets respectively; and for creatinine ($N = 10$), 9 variables were correctly identified irrespective of the sample size. This gave an overall sensitivity of 93.3%, 82.2%, and 77.8% for the small, medium and large data set, respectively, Table 1. A similar trend was observed for full resolution data sets although the larger data sets show slightly lower sensitivity (between 56.7% to 72.3%) compared to the small data set (between 77.8% to 93.3%). The sensitivity of ASCLAN was generally lower than all of the other six methods, Table 1. The lower sensitivity of ASCLAN was attributed to the tails of the discriminatory metabolites being considered as nondiscriminatory because of their lower r^2 values in comparison to the peaks of the discriminatory metabolites with higher r^2 values. The summary results for the data set with full resolution and $N = 50$ in each group is shown in Figure 3. Despite the relatively lower sensitivity of ASCLAN compared to the six existing methods, the ASCLAN algorithm, which is based on a data driven approach, was reliable in correctly identifying all four discriminatory metabolites in all six simulated data sets. Although ASCLAN (Figure 3a) considered some signal variables (tail ends of "real" signals) as noise regions, we considered this to be noncritical as the determination of noise or nondiscriminatory metabolites can be easily verified visually. Moreover, the ASCLAN approach did

DOI: 10.1021/acs.analchem.5b04020
Anal. Chem. 2016, 88, 5670–5679

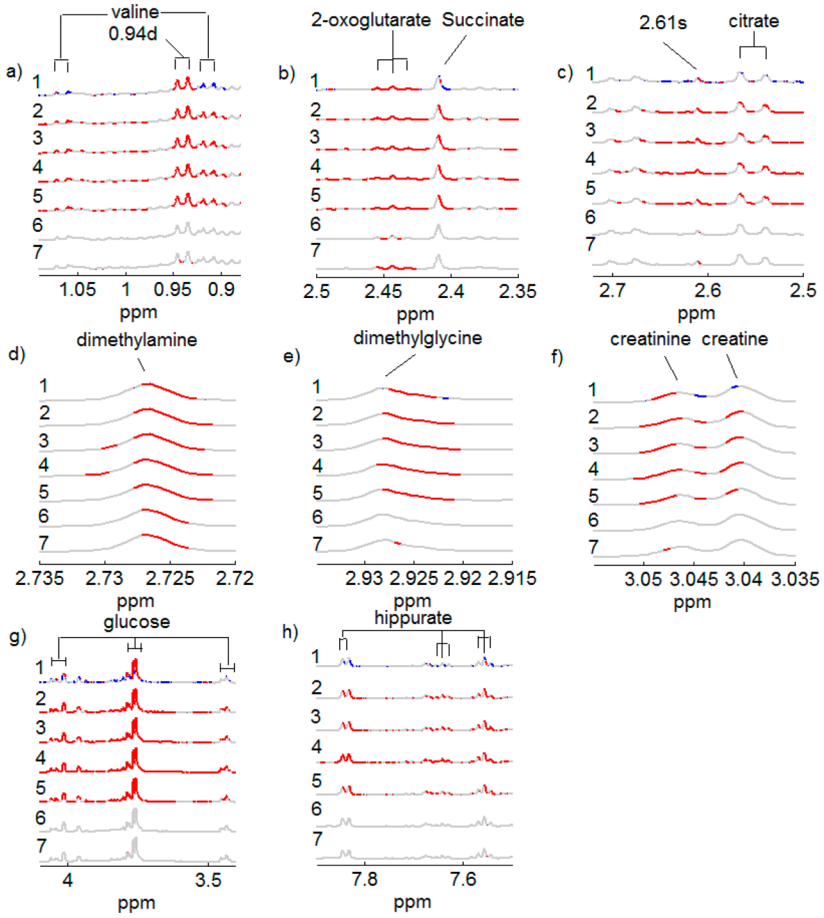


Figure 4. Median spectrum summarizing the discriminatory metabolites identified by the ASCLAN method for (a) valine and an unknown discriminatory metabolite at δ 0.94 d, (b) 2-oxoglutarate and succinate, (c) unknown metabolite at δ 2.61 s, (d) dimethylamine, (e) dimethylglycine, (f) creatinine, (g) glucose, and (h) hippurate. The median spectrum represents the results for (1) ASCLAN, (2) VIP, (3) loading coefficient, (4) regression coefficient, (5) Benjamini–Hochberg, (6) Bonferroni, and (7) metabolome wide significance level, MWSL methods. The median spectrum was color coded as follows: red for discriminatory metabolites and gray for nondiscriminatory metabolites for all other methods except for ASCLAN where gray is used for noise and blue is used for nondiscriminatory metabolites.

not consider any of the noise or nondiscriminatory variables as discriminatory, giving a zero false positive for all six simulated data sets, irrespective of the sample sizes or spectral resolutions, Table S-3, and therefore specificity (true negative rate) and PPV (accuracy) of 100%, Table 1. This zero false positive results outperformed all the methods based on OPLS-DA parameters: VIP (ranges from 2.1% to 15.4%), loading coefficient (from 7.5% to 35.1%), and regression coefficient (from 7.9% to 39.0%), as well as by multiple correction methods, Benjamini–Hochberg (from 7.1% to 35.4%), Bonferroni (from 0.8% to 10.4%), and MWSL corrections (between 0.1% and 0.4%). We considered this zero false positive result from ASCLAN is particularly important for metabolic profiling studies as the goal for these studies is typically to identify potential diagnostic biomarkers of disease or response to a therapeutic intervention, thus the focus is on clinical robustness and reliability rather than sensitivity, although high sensitivity is desirable. Furthermore, the zero false positive of ASCLAN was unaffected by the resolution of the spectral variables and the sample size, again reflecting the robustness of the method. This was not the case for the other methods except for MWSL where the level of false positive selection remained low for all data sets. Despite the lower false positive rates of MWSL, the dimethylamine at δ 2.73 was considered as discriminatory variables when in reality incorrect. Other methods, showed considerably larger number of spectral signals that have been incorrectly considered as discriminatory variables (Figure 3b–f), Tables 1 and S-3.

Application of ASCLAN to a Rat Renal Toxicity Study. Having applied the ASCLAN algorithm to simulated data sets, we then validated the algorithm using real data to identify the discriminatory metabolites characterizing $HgCl_2$ toxicity in a rat model comparing the predose urine samples versus 24 h postdose $HgCl_2$ administration when overt renal tubular toxicity was present as verified by histology data. The validity of the OPLS-DA model was indicated by high Q^2Y values (Q^2Y = 0.96) and permutation testing p = 0.026. The ASCLAN algorithm considered 1096 spectral variables with $r^2 \geq 0.68$ as discriminatory metabolites. The number of spectral variables identified by ASCLAN was considerably higher than Bonferroni (114 spectral variables) and MWSL (216 spectral variables); but much lower than Benjamini-Hochberg (4161 spectral variables), VIP (4368 spectral variables), loading coefficient (4705 spectral variables), and regression coefficient (5855 spectral variables). The ASCLAN algorithm successfully identified metabolites that are known to be associated with $HgCl_2$ toxicity. These include increased excretion of valine (δ 0.92 (d), δ 1.06(d)), glucose (δ 3.4–4.1), and creatinine (δ

DOI: 10.1021/acs.analchem.5b04020
Anal. Chem. 2016, 88, 5670–5679

3.05 (s), δ 4.05 (s)), with decreased excretion of hippurrate (δ 3.97 (d), δ 7.55, triplet (t), δ 7.64 (t), δ 7.73 (d)), succinate (δ 2.41 (s)), 2-oxoglutarate (δ 2.44 (t)), dimethylglycine (δ 2.93 (s)), and dimethylamine (δ 2.73 (s)).[28,29] ASCLAN also considered δ 2.61 (s) (an unknown metabolite which is highly correlated with δ 0.94 (d)) as a discriminatory metabolite. This unknown metabolite has been previously identified as being characteristic of $HgCl_2$ induced toxicity.[24] Chadeau-Hyam et al.[21] have previously reported the performance of MWSL to be similar to that of Bonferroni correction method. Here, we also reported comparable outcomes between Bonferroni corrections and MWSL. Both the Bonferroni and MWSL methods have considered the unknown metabolites at δ 2.61 s, 2-oxoglutarate and dimethylamine as discriminatory metabolites but were unable to consider the other known metabolites related to renal toxicity (Figure 4).[24] The Benjamini–Hochberg, VIP, loading coefficient and regression coefficient methods identified all discriminatory metabolites identified by ASCLAN. However, these methods also indicated a considerably large number of spectral variables that deem lacking any metabolic features as discriminatory metabolites. In view of this, we consider ASCLAN to be a good platform for distinguishing discriminatory metabolites from the data set, without the necessity of allocation of an arbitrary cutoff point. The whole procedure can be applied automatically and without the input of the investigator allowing for a less subjective and more transferrable means of identification of biomarkers. The ASCLAN approach is not as conservative as Bonferroni and MWSL approaches, which potentially increases false negative rate; but more stringent than Benjamini–Hochberg, VIP, loading coefficient and regression coefficient methods, which is often used in metabolic profiling studies and has been shown to generate large number of false positive signals.

The ASCLAN algorithm, which we have developed here, may be applied on its own for data analysis of highly homogeneous data sets, as demonstrated in the simulated data sets and the rat renal toxicity study. However, for highly heterogeneous data sets, as it has been increasingly reported,[33] we envisage that the ASCLAN may be applied in conjunction with some of the newer data analysis frameworks such as SHOCSY.[19] In this instance, SHOCSY is first applied as a "cleaning up" step to identify samples with similar features (homogeneous group) and indifferent features (heterogeneous group). Then, an OPLS-DA model is constructed using the homogeneous group before ASCLAN is applied to the "clean" OPLS-DA model to categorize the spectral variables into noise, nondiscriminatory or discriminatory variables. In doing so, this removes the need to rely on the application of an arbitrary cutoff value using loading regression and adjusting by Bonferroni corrections, as it was done in the SHOCSY algorithm. The advantage of applying SHOCSY and ASCLAN sequentially in this way would enable genuine discriminatory metabolites, particularly those with low intensity, to be more easily discovered as these are not obscured by the heterogeneous group. This concept is based on a similar concept to that demonstrated by Posam et al.[18] Nonetheless, to ensure feasibility of such a strategy, this would need to be further validated using appropriate simulated data sets.

CONCLUSION

The ability to robustly distinguish discriminatory metabolites in metabolic profiling studies is critical to ensure correct biological interpretation. ASCLAN offers an alternative and reliable data-driven approach for extracting discriminatory metabolites without requiring the investigators to define an arbitrary selection criterion but instead make a cut-point decision based on the inherent spectral features. This ability is not affected by the sample size or the spectral resolution of the data sets as demonstrated using the simulated data sets. Moreover, the application of ASCLAN delivered zero false positive results, which outperformed the existing methods including Benjamini–Hochberg, Bonferroni, MWSL, OPLS-DA VIP, OPLS-DA loading coefficient, and OPLS-DA regression coefficient methods. The ASCLAN approach was successfully applied to a renal toxicity study in a rat model and the metabolites known to be attributed to the renal toxicity were identified. We have demonstrated that this data-driven approach, ASCLAN, offers an attractive data analysis framework that can be applied to biological data sets for reliable extraction of discriminatory metabolites. We propose its further validation and use in high-throughput screening for discriminatory metabolites in metabolic profiling studies.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.5b04020.

Encrypted matlab code with brief instructions (ZIP)

Table S-1, means and variances in the signal intensities of metabolites; Table S-2, OPLS-DA modeling results for the simulated data sets; Table S-3, results comparing ASCLAN approach to conventional methods (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: r.loo@kent.ac.uk.

Notes

The encrypted ASCLAN matlab code can be freely downloaded for use in data analysis. However, the use and redistribution of the code, in whole or in part, for commercial purposes requires explicit permission of the authors and explicit acknowledgment of original publication. We ask that users who use the ASCLAN approach to cite the ASCLAN, STOCSY and SHY papers as well as the STOCSY patent (US20070043518, US7835872 and US7373256) in any resulting publications.

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

R.L.L. acknowledges support from the MRC New Investigator Grant Award (G1002151) and X.Z. is a postdoc working on the grant. The authors thank the colleagues who performed the experiments and collected the NMR data for COMET study.

REFERENCES

(1) Holmes, E.; Loo, R. L.; Stamler, J.; Bictash, M.; Yap, I. K.; Chan, Q.; Ebbels, T.; De Iorio, M.; Brown, I. J.; Veselkov, K. A.; Daviglus, M. L.; Kesteloot, H.; Ueshima, H.; Zhao, L.; Nicholson, J. K.; Elliott, P. *Nature* **2008**, *453*, 396–400.

(2) Chan, E. C.; Koh, P. K.; Mal, M.; Cheah, P. Y.; Eu, K. W.; Backshall, A.; Cavill, R.; Nicholson, J. K.; Keun, H. C. *J. Proteome Res.* **2009**, *8*, 352–61.

(3) Bylesjo, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. *J. Chemom.* **2006**, *20*, 341–351.

DOI: 10.1021/acs.analchem.5b04020
Anal. Chem. 2016, 88, 5670–5679

(4) Cloarec, O.; Dumas, M. E.; Trygg, J.; Craig, A.; Barton, R. H.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. *Anal. Chem.* **2005**, *77*, 517−26.
(5) Coen, M.; Hong, Y. S.; Clayton, T. A.; Rohde, C. M.; Pearce, J. T.; Reily, M. D.; Robertson, D. G.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *J. Proteome Res.* **2007**, *6*, 2711−9.
(6) Su, S. L.; Duan, J. A.; Wang, P. J.; Liu, P.; Guo, J. M.; Shang, E.; Qian, D. W.; Tang, Y. P.; Tang, Z. X. *J. Proteome Res.* **2013**, *12*, 852−865.
(7) Nevedomskaya, E.; Pacchiarotta, T.; Artemov, A.; Meissner, A.; van Nieuwkoop, C.; van Dissel, J. T.; Mayboroda, O. A.; Deelder, A. M. *Metabolomics* **2012**, *8*, 1227−1235.
(8) Mehmood, T.; Liland, K. H.; Snipen, L.; Saebo, S. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 62−69.
(9) Wiklund, S.; Johansson, E.; Sjostrom, L.; Mellerowicz, E. J.; Edlund, U.; Shockcor, J. P.; Gottfries, J.; Moritz, T.; Trygg, J. *Anal. Chem.* **2008**, *80*, 115−22.
(10) Wang, L.; Hou, E.; Wang, L.; Wang, Y.; Yang, L.; Zheng, X.; Xie, G.; Sun, Q.; Liang, M.; Tian, Z. *Anal. Chim. Acta* **2015**, *854*, 95−105.
(11) Patel, S.; Ahmed, S. *J. Pharm. Biomed. Anal.* **2015**, *107C*, 63−74.
(12) Coen, M.; Goldfain-Blanc, F.; Rolland-Valognes, G.; Walther, B.; Robertson, D. G.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *J. Proteome Res.* **2012**, *11*, 2427−40.
(13) Clayton, T. A.; Baker, D.; Lindon, J. C.; Everett, J. R.; Nicholson, J. K. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 14728−33.
(14) Winnike, J. H.; Li, Z.; Wright, F. A.; Macdonald, J. M.; O'Connell, T. M.; Watkins, P. B. *Clin. Pharmacol. Ther.* **2010**, *88*, 45−51.
(15) Smith, E. A.; Macfarlane, G. T. *Microb. Ecol.* **1997**, *33*, 180−8.
(16) Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. *Anal. Chem.* **2005**, *77*, 1282−9.
(17) Crockford, D. J.; Holmes, E.; Lindon, J. C.; Plumb, R. S.; Zirah, S.; Bruce, S. J.; Rainville, P.; Stumpf, C. L.; Nicholson, J. K. *Anal. Chem.* **2006**, *78*, 363−71.
(18) Posma, J. M.; Garcia-Perez, I.; De Iorio, M.; Lindon, J. C.; Elliott, P.; Holmes, E.; Ebbels, T. M.; Nicholson, J. K. *Anal. Chem.* **2012**, *84*, 10694−701.
(19) Zou, X.; Holmes, E.; Nicholson, J. K.; Loo, R. L. *Anal. Chem.* **2014**, *86*, 5308−15.
(20) McDonald, J. H. *Handbook of Biological Statistics*, 2nd ed.; Sparky House Publishing: Baltimore, 2009.
(21) Chadeau-Hyam, M.; Ebbels, T. M.; Brown, I. J.; Chan, Q.; Stamler, J.; Huang, C. C.; Daviglus, M. L.; Ueshima, H.; Zhao, L.; Holmes, E.; Nicholson, J. K.; Elliott, P.; De Iorio, M. *J. Proteome Res.* **2010**, *9*, 4620−7.
(22) Muncey, H. J.; Jones, R.; De Iorio, M.; Ebbels, T. M. *BMC Bioinf.* **2010**, *11*, 496.
(23) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. *Nucleic Acids Res.* **2013**, *41*, D801−7.
(24) Holmes, E.; Cloarec, O.; Nicholson, J. K. *J. Proteome Res.* **2006**, *5*, 1313−20.
(25) Nicholson, J. K.; Timbrell, J. A.; Sadler, P. J. *Mol. Pharmacol.* **1985**, *27*, 644−651.
(26) Nicholson, J. K.; Kendall, M. D.; Osborn, D. *Nature* **1983**, *304*, 633−5.
(27) Gartland, K. P.; Bonner, F. W.; Nicholson, J. K. *Mol. Pharmacol.* **1989**, *35*, 242−250.
(28) Holmes, E.; Bonner, F. W.; Nicholson, J. K. *Comp. Biochem. Physiol., Part C: Pharmacol., Toxicol. Endocrinol.* **1996**, *114*, 7−15.
(29) Holmes, E.; Bonner, F. W.; Sweatman, B. C.; Lindon, J. C.; Beddell, C. R.; Rahr, E.; Nicholson, J. K. *Mol. Pharmacol.* **1992**, *42*, 922−930.
(30) Dasgupta, S.; Freund, Y. *IEEE Trans. Inf. Theory* **2009**, *55*, 3229−3242.
(31) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. *Anal. Chem.* **2006**, *78*, 4281−90.
(32) Cho, H. W.; Kim, S. B.; Jeong, M. K.; Park, Y.; Miller, N. G.; Ziegler, T. R.; Jones, D. P. *Int. J. Data Min. Bioin.* **2008**, *2*, 176−92.
(33) Clayton, T. A.; Lindon, J. C.; Cloarec, O.; Antti, H.; Charuel, C.; Hanton, G.; Provost, J. P.; Le Net, J. L.; Baker, D.; Walley, R. J.; Everett, J. R.; Nicholson, J. K. *Nature* **2006**, *440*, 1073−1077.

DOI: 10.1021/acs.analchem.5b04020
Anal. Chem. 2016, 88, 5670−5679