analytical chemistry

Technical Note

pubs.acs.org/ac

RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data

C. D. Broeckling,*[,†] F. A. Afsar,[‡,⊥] S. Neumann,[∇] A. Ben-Hur,[‡] and J. E. Prenni*[,†,§]

[†]Proteomics and Metabolomics Facility, Colorado State University, Fort Collins, Colorado 80523, United States

[‡]Department of Computer Science, Colorado State University, Fort Collins, Colorado 80523, United States

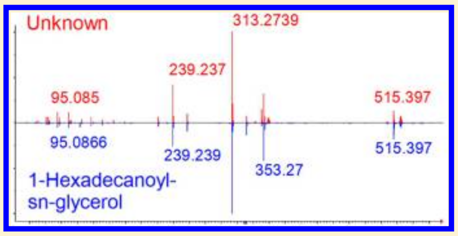
[§]Department of Biochemistry, Colorado State University, Fort Collins, Colorado 80523, United States

[∇]Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, 06108 Halle, Germany

[⊥]Department of Computer & Information Sciences, Pakistan Institute of Engineering & Applied Sciences, Islamabad, Pakistan

S *Supporting Information*

ABSTRACT: Metabolomic data are frequently acquired using chromatographically coupled mass spectrometry (MS) platforms. For such datasets, the first step in data analysis relies on feature detection, where a feature is defined by a mass and retention time. While a feature typically is derived from a single compound, a spectrum of mass signals is more a more-accurate representation of the mass spectrometric signal for a given metabolite. Here, we report a novel feature grouping method that operates in an unsupervised manner to group signals from MS data into spectra without relying on predictability of the in-source phenomenon. We additionally address a fundamental bottleneck in metabolomics, annotation of MS level signals, by incorporating indiscriminant MS/MS (idMS/MS) data implicitly: feature detection is performed on both MS and idMS/MS data, and feature−feature relationships are determined simultaneously from the MS and idMS/MS data. This approach facilitates identification of metabolites using in-source MS and/or idMS/MS spectra from a single experiment, reduces quantitative analytical variation compared to single-feature measures, and decreases false positive annotations of unpredictable phenomenon as novel compounds. This tool is released as a freely available R package, called RAMClustR, and is sufficiently versatile to group features from any chromatographic-spectrometric platform or feature-finding software.



Mass spectrometry (MS) has long been utilized for detecting and quantifying small molecules, particularly when coupled to separation tools such as gas chromatography (GC), liquid chromatography (LC), or capillary electrophoresis (CE). The strengths of these chromatographically coupled mass spectrometry platforms have been leveraged toward global metabolite profiling approaches, or metabolomics. The development of electrospray ionization (ESI)[1] was an important technological milestone, which allowed for the coupling of liquid separation methods to mass spectrometers. This development obviated the volatility requirement imposed by gas chromatography and supported development and expansion of both metabolomics and proteomics. Electrospray is considered a "soft" ionization technique, by which the molecular ion of the compound is generally more dominant than that achieved using "hard" ionization methods such as electron impact ionization (EI). However, the ESI process is imperfectly "soft" and does produce some degree of in-source fragmentation. Furthermore, secondary adducts, multimers, and fragmentation products of these can form during the ionization process, resulting in multiple observed ions representative of a single compound. These redundant signals are effectively utilized for EI spectra to allow for spectral-matching-based annotation metabolite signals.

Data analysis workflows that seek to detect mass signals in a nontargeted manner utilize both mass and retention time-based specificity—the resulting signal is commonly referred to as a "feature". In the absence of co-elution, one feature originates from a single compound. However, the reciprocal is largely untrue: a single compound can give rise to multiple features, as described above. Therefore, many metabolomics data processing tools, including both commercial and open-source tools, attempt to group features into spectra. Some grouping strategies are based on chemically meaningful and predictable patterns reflecting known phenomenon. However, this approach can be compromised by (i) interfering signals from co-eluting metabolites in complex samples that happen to look like fragments, adducts, or isotopes and (ii) unpredictable mass spectral fragments, adducts, or isotopes. As such, an unsupervised approach to grouping features is an attractive alternative. Previous tools including CAMERA,[2] AMDIS,[3] and MSClust[4] have attempted to address this issue, but none of these make full use of the nontargeted data. For example, CAMERA is biased toward the most abundant features and

Received: April 25, 2014
Accepted: June 13, 2014
Published: June 13, 2014

utilizes discrete binning by retention time. MSClust also looks for co-eluting and co-varying features and ultimately selects a representative "centrotype" feature for downstream statistical analysis—the majority of features are discarded. AMDIS works on a single data file, is generally not used for quantitation and does not utilize high-mass-accuracy data. Furthermore, all of these tools are designed for single-channel MS datasets.

Here, we report the development of a novel metabolomics workflow constructed around indiscriminant MS/MS (idMS/MS) data acquisition, which employs high-collision-energy fragmentation without precursor ion selection[5] and is acquired concurrently with low-collision-energy MS data. Our method is based on the premise that two features resulting from the same compound exhibit similarity in their retention times and a high correlation in their abundance profiles across different samples within a dataset. Based on this observation, we have developed a simple similarity function between features that allows us to use hierarchical clustering to generate the spectra of chemical compounds by grouping features from a single compound in a single cluster. Feature finding is conducted in both low- and high-collision-energy data, and a custom feature similarity score drives clustering of features into spectra suitable for informed manual interpretation, as well as automated database searching. This approach results in both in-source MS and idMS/MS spectra for all detected features and enables spectral matching to public, commercial, and custom spectral databases without additional experimentation.

EXPERIMENTAL SECTION

Sample Acquisition and Preparation. Equine cerebrospinal fluid (CSF) samples were obtained as previously described.[6] CSF was thawed at 4 °C, and 100 μL of CSF was precipitated with 400 μL of cold methanol. This solution was mixed thoroughly, incubated at −20 °C for 1 h, and spun at 12 000*g* for 15 min to remove proteins. The supernatant was transferred to autosampler vials for UPLC-MS analysis. The validation dataset consists of 50 urine samples, collected from Swedish males. Samples were prepared by thawing the urine at 4 °C, diluting with equal parts water, and centrifuging to remove particulates.

UPLC-MS Data Acquisition. Metabolome analysis of CSF and urine samples were accomplished using a Waters Acquity UPLC system coupled to a time-of-flight mass spectrometer (Xevo G2 Q-TOF MS). Five microliters (5 μL) of either protein-depleted CSF or diluted urine was injected onto an HSS T3 column (Waters, 1 mm × 100 mm, 1.7 μM), and eluted using a gradient of water to acetonitrile, each containing 0.1% formic acid. The gradient was held at 0.1% B for 1 min, ramped to 95% B over 12 min, and held for 3 min, before returning to 0.1% B and equilibrating for 3.9 min (20 min run time). The flow rate was held constant at 200 μL/min. Eluent was ionized via positive-mode electrospray ionization, with capillary voltage set to 2.2 kV, cone to 30 V, extraction cone to 2, with a source temperature of 150 °C and the desolvation nitrogen gas set to 350 °C at a flow rate of 800 L/h. Before acquisition, the instrument was calibrated via an infusion of sodium formate to within an error of 1 ppm. Mass accuracy was ensured via infusion of leucine enkaphalin lockmass, collected as a 0.5 s scan at a collision energy of 10 V every 20 s. Sample data were acquired in MS^E mode, with alternating scans (0.2 s/scan, m/z 50−1200) collected at collision energy of 6 V (MS) or using a CE ramp from 15 V to 30 V (idMS/MS). Each sample was injected in duplicate, with each set of injections being completely randomized for acquisition order. In addition, the samples were analyzed using data-dependent acquisition mode for traditional MS/MS experiments, with one DDA MS/MS spectrum acquired per MS scan, with a minimum precursor intensity threshold of 200 counts per second. All data were acquired in centroid mode.

Raw Data Conversion and Processing. Waters raw files were converted to cdf format using Databridge, which separates low-collision-energy MS and high-collision-energy idMS/MS data into two separate cdf files. The lockmass function data was discarded for this application. Feature detection (utilizing the centWave algorithm), an initial grouping step using a wide bandwidth (3), retention time correction, regrouping using a narrow bandwidth (1.5), and peak filling was performed using XCMS[7] (v. 1.32.0) in R[8] (v. 2.15). CAMERA[2] (v. 1.16.0) was used as a benchmark comparison, utilizing default values.

RAMClust Approach. The RAMClust approach was developed in Matlab and is currently fully implemented in R in a package called RAMClustR, and it is currently available via github (https://github.com/cbroeckl/RAMClustR). Implementation in R allowed an XCMS object to be used directly as input. The data within the XCMS object were extracted using the XCMS groupval function and was normalized to the total XCMS extracted ion signal (the quantile[9] method is an available option in RAMclustR). When a second collision energy level is used (as is possible with Waters MS^{E5} datasets utilized in this study), the user directs delineation of MS and idMSMS datasets using a tag located within the filename or filepath of the xcms object. RAMclustR is also capable of accepting properly formatted data matrices from other peak detection tools, with the only requirements being:

(1) no more than one sample (or file) name column and one feature name row;
(2) feature names that contain the mass and retention times, separated by a constant delimiter; and
(3) features in columns and samples in rows.

If both MS and idMS/MS data are to be imported, the feature names must be identical between the two datasets.

We developed a custom similarity matrix, which is the product of two Gaussian terms: one that considers the differences in retention times between two features and a second that considers the correlation between two features across all samples in the dataset. These two terms have widths defined by σ_t and σ_r, respectively, allowing flexibility in tuning correlational and retention time similarity decay rates independently based on the dataset and the acquisition instrumentation. The correlational relationship between two features can be described by either MS-MS, MS-idMS/MS, or idMS/MS-idMS/MS values, and we use Pearson's correlation to calculate similarity:

$$S_{ij} = \max\left\{\exp\left[-\frac{(1-c_{ij}^{\mathrm{MS1/MS1}})^2}{2{\sigma_r}^2}\right], \exp\left[-\frac{(1-c_{ij}^{\mathrm{MS2/MS2}})^2}{2{\sigma_r}^2}\right], \exp\left[-\frac{(1-c_{ij}^{\mathrm{MS1/MS2}})^2}{2{\sigma_r}^2}\right]\right\} \exp\left[-\frac{(t_i-t_j)^2}{2\sigma_t^2}\right]$$

where $c_{ij}^{\mathrm{MS1/MS2}}$ is the correlation coefficient between the ith and jth features of MS-idMS/MS ($c_{ij}^{\mathrm{MS1/MS1}}$ and $c_{ij}^{\mathrm{MS2/MS2}}$ are defined analogously) and σ_t and σ_r represent the widths of the corresponding Gaussian function. This captures our intuition that two features are similar if they are close in retention time and are correlated: both are required for two features to be grouped. Similarities were then converted to dissimilarities ($D_{ij} = 1 - S_{ij}$) for clustering. The output similarity matrix was then clustered using average linkage hierarchical clustering via that fastcluster package.[11] The dendrogram was then cut using the cutreeDynamicTree function in the package dynamicTreeCut.[12] For this application, the minimum cluster size is set to 2, dictating that only clusters with two or more features are returned, as singletons are impossible to interpret intelligently.

RAMclust similarity was calculated for the full feature matrix (within a user-specified maximum-allowed retention time window). Metabomolics datasets can generate thousands to tens of thousands of features, which can tax the memory of many desktop computers. To manage memory, we utilize the ff package[10] which allows for rapid temporary storage of large R objects using physical disk space rather than in memory, and process large data matrices in square blocks (2000 features at a time by default).

Cluster membership, in conjunction with the abundance values from individual features in the input data, were used to create spectra. Mass was derived from the feature mass, and the abundance for each mass in the spectrum was derived from the weighted mean of the intensity values for that feature. These spectra were then exported as an msp formatted document, which can be directly imported by NIST MSsearch, or used as input for MassBank[13] or NIST msPepSearch (http://peptide.nist.gov/software/ms_pep_search_gui/MSPepSearch.html) batch searching. Finally, the cluster membership was then used to create a third dataset, SpecData, which represented the MS level data after condensing feature intensities into spectral intensities using a weighted mean function, where the more-abundant signals contribute more to the spectral intensity.

■ RESULTS AND DISCUSSION

We developed and tested our approach using a UPLC-MS dataset of 38 samples of equine cerebrospinal fluid, and subsequently validated the approach in an independent urine dataset (see Figure S1 in the Supporting Information). XCMS[7] was used for feature finding, retention time correction, and alignment, and the resulting dataset was subsequently normalized to total XCMS signal intensity for each sample. The output data was then divided into low-collision-energy (MS) and high-collision-energy (idMS/MS) datasets, each with dimensions of row number equal to the number of injections and column number equal to the number of features (21060, for the CSF dataset). Each cell of these datasets represents the signal intensity at either low (MS) or high (idMS/MS) collision energy.

Any feature clustering tool must demonstrate accuracy to be useful in reducing redundancy without reducing biological coverage. One option to accomplish this is to compare the results of the clustering to a small panel of known compounds that are spiked into a sample. While this is a valid approach, it relies on the assumption that the chosen panel of compounds is representative of all the metabolites in a complex biological matrix. Thus, to increase the breadth of our validation experiments, we instead assessed the accuracy of the clustering by comparison against MS/MS spectra acquired using a traditional dependent acquisition (DDA) approach from the same CSF samples. All precursor ions that (i) could be mapped to a feature in the output dataset and (ii) contained more than 10 product ions were used as "valid" spectra for comparison. These spectra represented known precursor-product ion relationships from many of the major signals in the dataset, even if the identity of the compounds was unknown. The spectra created by RAMClust were then compared to the DDA spectra and the dot product spectral similarity score was calculated as a measure of accuracy, as described previously.[14] While the complexity of in-source and indiscriminant MS/MS signals is expected to be higher than DDA MS/MS spectra for the same compound, more-accurate clustering will still be revealed as relatively higher dot-product similarity scores between the RAMclustR reconstructed spectra and the mapped DDA MS/MS spectrum.

The RAMClust algorithm has several parameters that can be tuned by the user to improve clustering accuracy. Parameters σ_t and σ_r represent Gaussian tuning parameters of retention time similarity and correlational score, respectively, between feature pairs. The influence of these two parameters on the similarity is depicted in Figure 1. These tuning parameters will allow the algorithm to be used with MS data from any chromatographic platform. When idMS/MS data are available, correlational similarity can be calculated between two features, at the level of either MS vs MS, MS vs idMS/MS, or idMS/MS vs idMS/MS. While the MS-idMSMS correlation theoretically represents the CID event most directly, this relationship is subject to potential interfering signals in both data channels (MS and idMS/MS). In practice, a strong correlational relationship at any of the three levels represents strong evidence of precursor−product relationships; thus, the algorithm utilized the maximum correlational r-value of the three relationships.

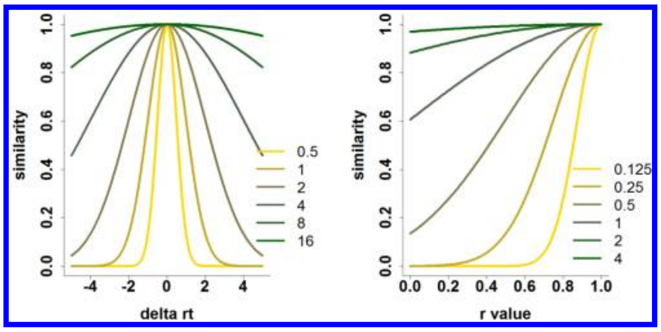


Figure 1. RAMClust is based on a custom feature similarity score, which is the product of two terms that capture similarity in retention time and correlation across samples. Each of the two terms has a tuning parameter associated with it that controls the width of the corresponding Gaussian: σ_t for retention time (left) and σ_r for the degree of (right). Increased values for the two σ terms decrease the rate of decay in the similarity score, as a function of either retention time difference or correlation r between pair of features.

The influence of σ_t and σ_r on the average spectral similarity between RAMClust and DDA spectra was rigorously evaluated at 441 combinations of parameter levels of σ_t and σ_r (Figure 2a). These results revealed a plateau of high spectral similarity at values of σ_t =2 and σ_r = 0.5 (Figure 2a). This σ_t value was approximately half the median peak width of the XCMS detected peak (max-min time for each individual peak in the xcms object), indicating that we can directly use XCMS input to set this parameter without user intervention: this holds true

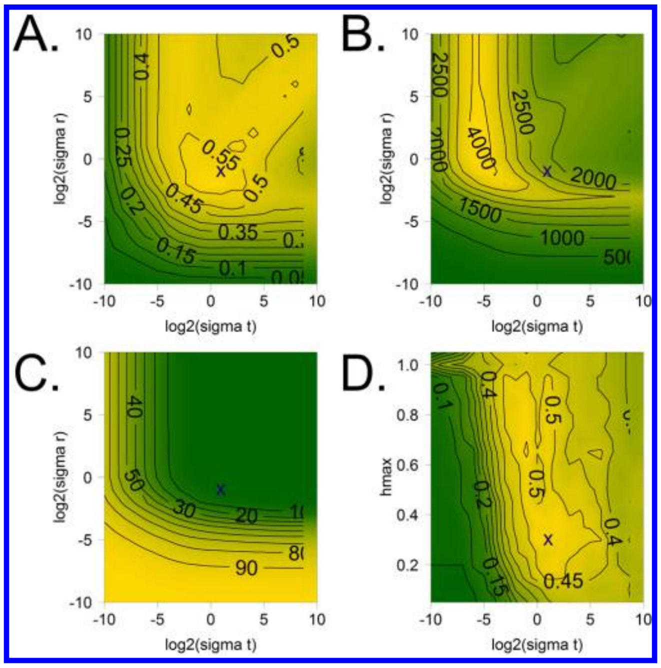


Figure 2. Influence on RAMClustR parameters, σ_t for σ_t for and hmax were systematically varied to examine the influence of these parameters on feature grouping accuracy, the number of clusters, and the number of ungrouped features (singletons). (A) RAMClust spectra generated using σ_t and σ_r values of 2 and 0.5 produce the strongest dot product similarity to DDA spectra, which represent validated precursor product relationships. This σ_t value is roughly half the median XCMS peak width, indicating that the σ_t value can be set automatically when XCMS data are used as the input. (B) Influence of σ_t and σ_r on the number of clusters with at least two features. The optimal values σ_t and σ_r (denoted with an "x"), as determined by the maximal dot product similarity, results in ~2500 clusters. (C) σ_t and σ_r values that are too selective results in fewer clusters, because of high singleton (features which cluster with no other features). (D) The dot product similarity scoring benefits from some precutting of the tree, as provided by the dynamicTreeCut algorithm, allowing us to set a default maximal cluster height of 0.3.

for an independent dataset of urine samples (see the supplementary material). Correlation is a scale-free statistic, and it should be platform-neutral; thus, we used our observed optimal value of 0.5 and can expect reasonable performance on any platform. Implementation of RAMclustR using parameters that maximized MS/MS similarity between reconstructed spectra and DDA spectra generated ~2500 clusters with at least two features (Figure 2b), and relatively few singletons (Figure 2c). This algorithm generated a large stable region, indicating that it is robust to small changes in parameter values. This stability generated a strong MS/MS similarity, even at "unreasonable" σ_t values (>200 s), as long as σ_r is proportionally high (Figure 2a). We interpret this as a scaling phenomenon, as the dynamicTreeCut algorithm is responsive to tree "shape" rather than an absolute height.[12] The dynamicTreeCut maximum height parameter was also examined in conjunction with σ_t, and it revealed that the tree pruning step benefited from some precutting (Figure 2d); thus, we employ a default value of 0.3 for this parameter. These parametrization rules make the algorithm extremely easy to use: when an XCMS object is used as input, the user needs to set none of these parameters, and when a dataset is imported from other software, only σ_t needs to be manually set. The output MS/MS similarity using default RAMclust similarity scores was used to compare results against the only other feature grouping tool in R: CAMERA. The results of this comparison indicated that RAMclust grouping of features resulted in spectra that are more similar to DDA spectra than the results generated from CAMERA's groupFWHM, groupCorr, and groupDen functions (see Table 1). This observation was validated on a second LC-MS dataset of urine samples: RAMClust grouping resulted in clustering output that better represents valid feature relationships and, consequentially, biological small molecule signals.

Table 1. Comparison between RAMclustR and CAMERA[a]

method	MSMS similarity[b]	nClus (>1)[c]	perSing[d]
CSF Dataset			
xsb ← groupFWHM(xset)	0.202	535	0.43
xsc ← groupCorr(xsb)	0.177	784	29.56
xsd ←groupDen(xsa)	0.043	39	0.00
RAMclustR(xset)	0.382	3248	15.47
Urine Dataset			
xsb ← groupFWHM(xset)	0.106	290	4.88
xsc ← groupCorr(xsb)	0.059	332	61.77
xsd ←groupDen(xsa)	0.020	35	0.00
RAMclustR(xset)	0.228	827	32.75

[a] The comparisons were performed using default values for both the CSF and Urine datasets. The first three rows in both the CSF and Urine datasets reflect CAMERA functions, while the final row reflects RAMClustR-based grouping. [b] MSMSsimilarity refers to the spectral similarity between mapped feature for which data-dependent MS/MS data were available and the reconstructed spectra from the output dataset defined in the "method" column. [c] nClus (>1) refers to the number of clusters with two or more features defined by the grouping method. [d] perSing is the percentage of all features in the data set that remain ungrouped (singletons).

The spectra produced via RAMclust grouping can be written to NIST MSP format for viewing and searching, and they can be submitted directly to the MassBank Database[13] batch search tool, submitted for batch searching to NIST msPepSearch, and/or viewed and searched via the NIST MSSearch program. All these tools offer the ability to generate and search against custom libraries of spectra, and our laboratory is creating libraries of in-source spectra toward this end. However, idMS/MS spectra re-created from the RAMClust algorithm and workflow were highly similar to authentic NIST MS/MS database spectra (see Figures 3a–c), demonstrating that this workflow can take full advantage of existing resources.

Since RAMClust-generated spectra accurately reflect spectra of authentic chemical standards, the intensity of the spectra themselves can be used as the quantitative unit for downstream statistical analysis. The intensity of the spectra were calculated using a weighted mean function of all the component features, such that each value in the resulting dataset represents the quantitative signal intensity value for each spectrum for each sample in the dataset. The use of spectra dramatically reduced analytical variation through an averaging of measurement noise, as compared to either the mean or median feature-based variation for each cluster (see Figure 3D).

■ CONCLUSIONS

Annotation of mass signals in nontargeted metabolomics experiments remains a significant bottleneck and is arguably one of the most important challenges to the field as confident metabolite identification is required for biological interpreta-

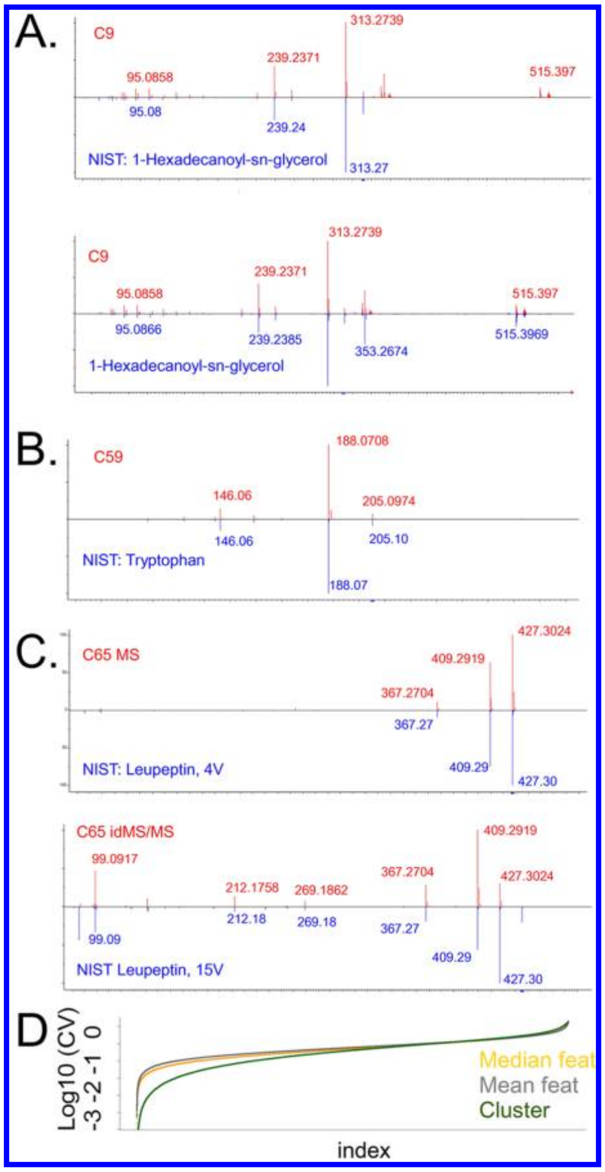


Figure 3. (A) Cluster membership and peak area data are used to generate spectra, which can be searched against spectra databases. The in-source low-collision-energy spectrum representing C9 was identified as hexadecanoyl-*sn*-glycerol (16:0 MAG) in the CSF samples, and shows a strong match to the NIST library spectrum representing this compound. However, the match is even stronger if all the in-source phenomenon are considered (bottom panel, standard run by the authors under identical analytical conditions). (B) Tryptophan in-source low-collision-energy spectrum can be identified with a high degree of confidence from either NIST MS/MS spectra (top) or a custom library spectrum (bottom). (C) Both low-collision-energy spectra (top) and high-collision-energy spectra (bottom) can be used for the same compound to increase the confidence of identification in the event that the MS spectrum is sparse, as demonstrated by leupeptin, a protease inhibitor added to the CSF samples before processing. (D) Clustering of features results in reduced analytical variation. The coefficient of variation (CV) of all individual compound measurements was calculated for all clusters, and compared to the median or mean feature CV for the features comprising those clusters. These ∼120 000 measures of variation indicate that the analytical variation for the majority of compound measurements is greatly reduced through aggregation into compound clusters or spectra.

tion. In this report, we demonstrate a novel workflow utilizing indiscriminant MS/MS data acquisition, expanded feature finding and a novel clustering algorithm to group features based on both low- and high-collision-energy data to generate spectra that are compatible with publically available spectral search tools. The workflow allows for more-efficient use of instrumentation, reduced feature redundancy and false discovery rate correction burden for downstream univariate statistical tests, improved analytical reproducibility, a more-automated annotation workflow, and greatly increased confidence in the annotations, compared to accurate mass-based searching alone. RAMClustR is available for download at https://github.com/cbroeckl/RAMClustR.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional information as indicated in text. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*Tel.: 970-491-2273. E-mail: Corey.Broeckling@colostate.edu (C. D. Broeckling).

*Tel.: 970-491-0961. E-mail: Jessica.Prenni@colostate.edu (Jessica E. Prenni).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

F.A.A. was funded by the Fulbright Scholarship Program of the U.S. Department of State and the Higher Education Commission of the Government of Pakistan.

■ REFERENCES

(1) Whitehouse, C. M.; Dreyer, R. N.; Yamashita, M.; Fenn, J. B. Electrospray Interface for Liquid Chromatographs and Mass Spectrometers. *Anal. Chem.* **1985**, *57*, 675−679 (DOI: 10.1021/ac00280a023).

(2) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **2012**, *84*, 283−289 (DOI: 10.1021/ac202450g).

(3) Halket, J. M.; Przyborowska, A.; Stein, S. E.; Mallard, W. G.; Down, S.; Chalmers, R. A. Deconvolution gas chromatography mass spectrometry of urinary organic acids—Potential for pattern recognition and automated identification of metabolic disorders. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 279−284 (DOI: 10.1002/(SICI)1097-0231(19990228)13:4<279::AID-RCM478>3.0.CO;2-I).

(4) Tikunov, Y. M.; Laptenok, S.; Hall, R. D.; Bovy, A.; de Vos, R. C. MSClust: A tool for unsupervised mass spectra extraction of chromatography−mass spectrometry ion-wise aligned data. *Metabolomics* **2012**, *8*, 714−718 (DOI: 10.1007/s11306-011-0368-2).

(5) Plumb, R. S.; Johnson, K. A.; Rainville, P.; Smith, B. W.; Wilson, I. D.; Castro-Perez, J. M.; Nicholson, J. K. UPLC/MS^E: A new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 1989−1994 (DOI: 10.1002/rcm.2550).

(6) Broccardo, C. J.; Hussey, G. S.; Goehring, L.; Lunn, P.; Prenni, J. E. Proteomic Characterization of Equine Cerebrospinal Fluid. *J. Equine Vet. Sci.* **2013**, http://dx.doi.org/10.1016/j.jevs.2013.07.013.

(7) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, And Identification. *Anal. Chem.* **2006**, *78*, 779−787 (DOI: 10.1021/ac051437y).

(8) R Core Team *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2014.

(9) (a) Bolstad, B. M. *preprocessCore: A collection of pre-processing functions, R package*, Version 1.22.0, 2014. (b) Brodsky, L.; Moussaieff, A.; Shahaf, N.; Aharoni, A.; Rogachev, I. Evaluation of Peak Picking Quality in LC-MS Metabolomics Data. *Anal. Chem.* **2010**, *82*, 9177–9187 (DOI: 10.1021/ac101216e).

(10) Adler, D.; Gläser, C.; Nenadic, O.; Oehlschlägel, J.; Zucchini, W. ff: memory-efficient storage of large data on disk and fast access functions. *R package version 2.2-13*, 2014, http://CRAN.R-project.org/package=ff.

(11) Mullner, D. *J. Stat. Software* **2013**, *53*, 1–18.

(12) Langfelder, P.; Zhang, B.; Horvath, S. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **2008**, *24*, 719–720 (DOI: 10.1093/bioinformatics/btm563).

(13) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, *45*, 703–714 (DOI: 10.1002/Jms.1777).

(14) Broeckling, C. D.; Heuberger, A. L.; Prince, J. A.; Ingelsson, E.; Prenni, J. E. Assigning precursor–product ion relationships in indiscriminant MS/MS data from non-targeted metabolite profiling studies. *Metabolomics* **2013**, *9*, 33–43 (DOI: 10.1007/s11306-012-0426-4).