Published in final edited form as:

Anal Chem. 2019 July 16; 91(14): 9069–9077. doi:10.1021/acs.analchem.9b01424.

ADAP-GC 4.0: Application of Clustering-Assisted Multivariate Curve Resolution to Spectral Deconvolution of Gas Chromatography–Mass Spectrometry Metabolomics Data

Aleksandr Smirnov[†], Yunping Qiu[‡], Wei Jia[§], Douglas I. Walker^{||}, Dean P. Jones[⊥], Xiuxia Du^{*,†}

[†]University of North Carolina at Charlotte, Charlotte, North Carolina 28223, United States

[‡]Albert Einstein College of Medicine, Bronx, New York 10461, United States

[§]University of Hawaii Cancer Center, Honolulu, Hawaii 96813, United States

^{||}Department of Environmental Medicine and Public Medicine, Icahn School of Medicine at Mount Sinai, New York, New York 10029, United States

[⊥]Emory University, Atlanta, Georgia 30322, United States

Abstract

We report a multivariate curve resolution (MCR)-based spectral deconvolution workflow for untargeted gas chromatography–mass spectrometry metabolomics. As an essential step in preprocessing such data, spectral deconvolution computationally separates ions that are in the same mass spectrum but belong to coeluting compounds that are not resolved completely by chromatography. As a result of this computational separation, spectral deconvolution produces pure fragmentation mass spectra. Traditionally, spectral deconvolution has been achieved by using a model peak approach. We describe the fundamental differences between the model peak-based and the MCR-based spectral deconvolution and report ADAP-GC 4.0 that employs the latter approach while overcoming the associated computational complexity. ADAP-GC 4.0 has been evaluated using GC–TOF data sets from a 27-standards mixture at different dilutions and urine with the mixture spiked in, and GC Orbitrap data sets from mixtures of different standards. It produced the average matching scores 960, 959, and 926 respectively. Moreover, its performance has been compared against MS-DIAL, eRah, and ADAP-GC 3.2, and ADAP-GC 4.0 demonstrated a higher number of matched compounds and up to 6% increase of the average matching score.

Graphical Abstract

*Corresponding Author Phone: (704) 687-7307; xiuxia.du@uncc.edu.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.9b01424.

Experimental information, software parameters, example of overestimating the number of components, details of data preprocessing for MCR, simulation of coeluting compounds, urine unit-mass-resolution data results, and standard-mixture high-mass-resolution data results (PDF)

The authors declare no competing financial interest.

Spectral deconvolution is an essential step in preprocessing untargeted, gas chromatography coupled mass spectrometry (GC–MS) metabolomics data. It computationally separates ions that are in the same mass spectrum but belong to coeluting compounds that are not resolved completely by chromatography. As a result of this computational separation, spectral deconvolution produces pure fragmentation mass spectra. Each mass spectrum consists of ions from a single compound, allowing matching of compounds through searching spectral libraries and relative quantitation of compounds by using intensity values of peaks in the pure mass spectra.

In the workflow of preprocessing GC–MS data, spectral deconvolution occurs after chromatographic peaks have been detected from extracted ion chromatograms (EICs). With the EIC peaks detected, traditional spectral deconvolution is usually carried out in two sequential steps.¹ The *component perception step* detects the presence of components and selects a model peak for each perceived component. At this stage, a component generally consists of multiple chromatographic peaks that have a very similar peak shape. From these peaks, a model peak is selected that can best represent the elution profile of the component. Subsequently, the *peak decomposition step* decomposes each detected EIC peak into a linear combination of the model peaks and constructs a pure fragmentation spectrum for each perceived component. This two-step approach is computationally efficient and provides model peaks that are similar to real EIC peaks in shape.

However, there is always the risk that a selected model peak has actually been produced by two or more coeluting components and therefore is inappropriate to serve as a model peak. This inappropriate model peak selection would cause incorrect pure fragmentation spectra for all the involved coeluting compounds and eventually errors in library matching and inaccuracy in relative quantitation. This phenomenon usually occurs when the signal from one of the coeluting compounds dominates and therefore it is hard to distinguish a composite EIC peak that is produced by two or more coeluting compounds from a unique EIC peak that is produced by a single compound. This is not uncommon in untargeted GC–MS metabolomics due to the high complexity of biological samples, especially when the mass resolution of the mass analyzer is low.

In recent years, there has emerged another class of spectral deconvolution methods that combine component perception and peak decomposition into one step. These methods use

the multivariate curve resolution (MCR) technique to construct model peaks of components and, therefore, can avoid the issue with selecting inappropriate model peaks in the traditional two-step spectral deconvolution. MCR is a well-established chemometric method to solve the ubiquitous problem of chemical mixture analysis,² and has been successfully used for spectral deconvolution in GC–MS,^{3–5} GC–FID,⁶ GC × GC–FID,⁷ GC × GC–qMS,⁸ GC × GC–HRTOFMS,⁹ and other gas chromatography analyses. Hantao et al. reviewed MCR methods and their applications to computational separation of chromatograms in complex samples.¹⁰ Some of the most frequently used MCR methods are MCR-ALS¹¹ and NMF,⁶ which iteratively construct approximations for the model peaks and spectra of components. These two methods have been used to analyze a wide variety of data sets from standards to human serum and urine samples.^{5,8,9,12–17}

However, while MCR methods have demonstrated their ability to computationally resolve chromatograms with coeluting components and avoid the issue of selecting inappropriate model peaks, the MCR methods have their drawbacks as well. One of such drawbacks is the fact that the solution of MCR-ALS, NMF, and some other MCR methods is not unique for the experimental data.¹¹ This inherent ambiguity of the MCR methods can be reduced by forcing the non-negativity, local rank, and unimodality constraints,¹¹ performing robust initialization,¹⁶ imposing sparse spectra,^{3,17} smoothing elution profiles,³ and other techniques. Another drawback of MCR is that developing an MCR-based spectral deconvolution that can be reliably applied to the entire chromatogram with various patterns of signal mixing is extremely challenging.⁹ The latter stems from both the enormous space of feasible MCR solutions and the high time complexity of iterative MCR methods.

We have addressed the above limitations of both the two-step and the MCR-based spectral deconvolution in the ADAP-GC 4.0 workflow to be described herein. The spectral deconvolution algorithm implemented in ADAP-GC 4.0 adapts the MCR approach and thus represents a significant shift in terms of spectral deconvolution principle from its predecessors^{18–20} that all use the traditional two-step approach. It combines several ideas from both approaches such as splitting the entire retention time range into deconvolution windows,^{18–20} adjusting the apex retention time of EIC peaks,¹ clustering EIC peaks to estimate the number of components,^{18–20} and applying the unimodality constraint to model peaks.³ Furthermore, ADAP-GC 4.0 addresses the problem of missing low-intensity peaks caused by imperfection of peak detection algorithms.²¹

To evaluate the performance of ADAP-GC 4.0, we have compared it against MS-DIAL,²² eRah,²³ and its immediate predecessor ADAP-GC 3.2²⁰ using both unit and high mass resolution GC–MS data. Overall, we demonstrate herein that ADAP-GC 4.0 produces results that are similar or better in terms of library matching scores and relative quantitation. It is relevant to note that all of the compared software tools are freely available with publications describing the algorithms, which allows us to compare different spectral deconvolution approaches. ADAP-GC 4.0 was not evaluated against commercial software tools such as ChromaTOF²⁴ or AnalyzerPro²⁵ because the principles of spectral deconvolution in these commercial software tools are mostly unpublished.

ADAP-GC 4.0 has been implemented in Java and is open-source. It has been incorporated into MZmine 2, a widely used graphical software tool for preprocessing GC- and LC-MS metabolomics data.^{21,26} ADAP-GC 4.0 source code and raw spectral data used in the software comparison can be found at <http://www.du-lab.org>.

EXPERIMENTAL INFORMATION

Unit Mass Resolution TOF GC-MS Data Sets.

Two sets of unit mass resolution data files were used in testing ADAP-GC 4.0: (1) Mixture of standard compounds (Sample I): Seven calibration curve samples with each containing 27 standard compounds were prepared at different concentrations ((0.2, 0.4, 0.6, 0.8, 1.0, 2.0, and 5.0) $\mu\text{g/mL}$ of each compound). With four pairs of coeluting compounds in each sample, we were able to evaluate the overall performance of peak detection and deconvolution of ADAP-GC 4.0. (2) Urine samples with standard mixtures spiked in (Sample II): Eight samples were prepared by spiking into a pooled urine sample with the seven calibration curve samples of Sample I and an additional sample consisting of 0.1 $\mu\text{g/mL}$ of each standard compound. Sample II was used for evaluating the performance of ADAP-GC 4.0 in terms of processing complex samples.

High Mass Resolution Orbitrap GC-MS Data Sets.

A series of 16 distinct mixtures containing a total of 260 standard compounds were analyzed for testing the detection and library matching of environmental pollutants using gas chromatography with high-resolution accurate mass detection. These standard compounds include brominated flame retardants, dioxins, furans, polychlorinated biphenyls, organonitrogen pesticides, pyrethroids, organophosphorous pesticides, and organochlorine pesticides.

See Supporting Information (SI) for details on preparing the unit mass resolution TOF GC-MS and high mass resolution Orbitrap GC-MS samples.

RESULTS AND DISCUSSION

Mathematical Principles of Spectral Deconvolution.

Let m be the total number of time points for each EIC and n the total number of EIC peaks. Let $m \times n$ matrix X represent all of the detected EIC peaks, wherein a column is the elution profile of an EIC peak corresponding to a particular m/z . The task of spectral deconvolution is to find $m \times I$ matrix C and $I \times n$ matrix S such that

$$X \approx C \times S \quad (1)$$

Each column of matrix C is the model peak of a component and each row of matrix S is the pure fragmentation spectrum of a component. Number I is the total number of components (also the total number of pure mass spectra) and is usually much smaller than the number of EIC peaks n . In each pure spectrum in S , the number of peaks with positive abundance

values is at most n . As a result of the spectral deconvolution in eq 1, each EIC peak $X[:, j]$ is decomposed into a linear combination of the components' model peaks $C[:, l]$ as follows:

$$X[:, j] \approx C[:, 1]S_{1j} + C[:, 2]S_{2j} + \dots + C[:, l]S_{lj}, j = 1, \dots, n \quad (2)$$

Hereafter, $X[:, j]$ denotes the j -th column of matrix X , similarly for matrices C and S .

The traditional two-step spectral deconvolution approach (used by AMDIS, MS-DIAL, and ADAP-GC 3.2) determines matrix C first. Each column of C is a model peak that is selected from real EIC peaks. Subsequently, each column of matrix S is determined by solving a separate optimization problem

$$S[:, j] = \arg \min_{\vec{s} \in \mathbb{R}_+^{l \times 1}} \|X[:, j] - C \times \vec{s}\|_2^2, j = 1, \dots, n \quad (3)$$

where $\|\cdot\|_2$ is the Euclidean vector norm and $\mathbb{R}_+^{l \times 1}$ denotes the space of all column vectors of length l with each element being non-negative. This traditional two-step approach chooses the model peak of a component from real EIC peaks and is computationally efficient.

In contrast, the MCR approach^{3,4,6,27} to spectral deconvolution (used by eRah and ADAP-GC 4.0) eliminates the step of selecting model peaks. Instead, it determines matrices C and S simultaneously by solving the following optimization problem

$$C, S = \arg \min_{\hat{C} \in \mathbb{R}_+^{m \times l}, \hat{S} \in \mathbb{R}_+^{l \times n}} \|X - \hat{C} \times \hat{S}\|_F^2 \quad (4)$$

where $\|\cdot\|_F$ is the Frobenius matrix norm and $\mathbb{R}_+^{m \times l}$ and $\mathbb{R}_+^{l \times n}$ represent the spaces of all non-negative matrices of dimensions $m \times l$ and $l \times n$ respectively. However, this approach is computationally more demanding and can produce model peaks that are of arbitrary shape when no shape constraints are applied. Next, we describe the inherent advantages of MCR and our approaches to address its disadvantages.

Advantages of MCR over Traditional Spectral Deconvolution Approach.

The MCR approach has two inherent advantages over the traditional approach. The first advantage is that MCR avoids the challenging task in the traditional approach that *selects* an EIC peak to be the model peak of a component. To qualify as a model peak, an EIC peak must have been produced by a single component. However, a model peak selected by the traditional approach could actually have been produced by two or more coeluting components, which would cause incorrect pure fragmentation spectra for all the involved coeluting compounds and eventually errors or low-confidence in library matching and inaccuracy in relative quantitation. This could occur because distinguishing these two types of EIC peaks is very challenging. The second advantage of the MCR approach lies in the

consequence of overestimating the number of components. The traditional approach needs this number of components to select model peaks and the MCR approach needs it to *construct* model peaks. When the number of components is overestimated, fragmentation spectra constructed by the MCR approach are affected much less and yield higher-confidence library matching results.

Figure 1 shows the resulting differences after the traditional and MCR approaches have been applied to deconvoluting two coeluting compounds L-histidine and L-lysine in file S1_1.cdf (see SI about the data files used herein). Histidine and L-lysine elute at 21.92 and 21.96 min, respectively. Because they elute in close proximity in time, EIC peaks for ions of the same m/z , such as peak P_{73} for $m/z = 73$, that are produced by both compounds are actually composite. If peak P_{154} and composite peak P_{73} are selected as model peaks for L-histidine and L-lysine (Figure 1A left), then the spectrum constructed for L-histidine will miss peak $m/z = 73$ together with similar peaks $m/z = 74$ and $m/z = 75$. Consequently, the resulting spectrum gave rise to a low library matching score (Figure 1A Middle).

This issue can be partially avoided by filtering out EIC peaks that are unqualified to serve as model peaks. Such a filtering is usually based on sharpness, signal-to-noise ratio, unimodality of EIC peaks, or a combination of those criteria.^{18,20} However, these criteria or their combinations may not be able to filter out all composite EIC peaks. For instance, the composite EIC peak P_{73} in Figure 1A left is unimodal, has a high sharpness and a high signal-to-noise ratio (as calculated by ADAP-GC 3.2) and therefore meets the requirements for sharpness, signal-to-noise ratio, and modality.^{18,20} However, it still causes construction of incorrect fragmentation spectra. Furthermore, the filtering step requires users to specify multiple parameters that general users usually do not know how to specify.

In contrast, the MCR-based spectral deconvolution completely eliminates the step of selecting model peaks. It determines model peaks as a solution to the optimization problem in eqn. 4. Figure 1B depicts the decomposition results for the coeluting L-histidine and L-lysine. Because MCR automatically produces model peaks that tend to resemble unique EIC peaks, the spectrum it constructs for L-histidine contains all peaks $m/z = 73, 74, 75$, and produces high matching scores with the correct library spectra for both compounds (Figure 1B middle and right).

Another advantage of MCR lies in the consequences of overestimating the number of components. This number is required by the traditional approach for selecting model peaks and by the MCR approach for solving eq 4. When the number of components is overestimated in the traditional two-step approach, the m/z of the model peak for the corresponding component would be completely missing from the pure fragmentation spectra of all other coeluting components, as a result of solving the optimization problem described by eq 3. This issue is referred to as *splitting*. ADAP-GC 3.0 addressed the splitting issue by performing a checking step after spectral deconvolution. If splitting does occur, it removes duplicate model peaks and repeats spectral deconvolution.¹⁹ However, this splitting issue checking adds several parameters that have to be specified by users. In contrast, spectra produced by the MCR approach usually do not miss important m/z peaks even when the number of components is overestimated (see SI), because this approach does not involve

selecting model peaks. As a result, the MCR approach tends to construct more accurate pure fragmentation spectra and is affected less by the overestimation of the number of components.

Despite these two inherent advantages, the MCR approach has its drawbacks as well. In particular, running MCR-based spectral deconvolution takes significantly more time than running the traditional two-step workflow. Furthermore, the MCR approach can produce model peaks that are very different from real EIC peaks in shape. To address these two drawbacks, we have implemented ADAP-GC 4.0 that could take full advantage of the strengths of MCR while overcoming its weaknesses.

Implementation of ADAP-GC 4.0.

The ADAP-GC 4.0 workflow (Figure 2) consists of three sequential steps after construction of EICs and detection of EIC peaks: (1) determination of deconvolution windows; (2) construction of model peaks in each deconvolution window by MCR, and (3) decomposition of each chromatogram that has been constructed in a deconvolution window into a linear combination of the model peaks, regardless of whether or not a peak has been detected in the EIC in the peak detection step prior to spectral deconvolution. Next, we describe each of these three steps in detail.

Deconvolution Windows.

ADAP-GC 4.0 splits the entire retention-time range into short disjoint intervals called deconvolution windows. Within each window, a separate MCR is carried out on the EIC peaks comprising that window. As a result, the deconvolution windows reduce the number of peaks and the number of time points participating in a single MCR, thus significantly improving the overall running time of the spectral deconvolution algorithm. Such deconvolution windows were used in several versions of ADAP-GC workflow.^{18,19,28} Here, we further improve the algorithm for choosing deconvolution windows by applying a sparse hierarchical clustering algorithm and selecting better window boundaries.

Previously, ADAP-GC 3.0 would use peaks detected in the total ion chromatogram (TIC) to determine deconvolution windows.¹⁹ Because that method required a separate set of peak-detection parameters specified by a user, it was replaced by the DBSCAN clustering algorithm applied to existing EIC peaks.²⁸ DBSCAN is a time- and memory-efficient clustering algorithm that requires fewer parameters than the TIC peak detection. However, its parameters are not very intuitive, so a user would have to try multiple values of DBSCAN parameters to produce short deconvolution windows especially when the number of detected EIC peaks is large. For that reason, in ADAP-GC 4.0 we use a clustering algorithm and a distance between EIC peaks that allow user to explicitly specify the maximum width of deconvolution windows.

In ADAP-GC 4.0, the dissimilarity between two EIC peaks with retention-time intervals $[a, b]$ and $[c, d]$ is defined by

$$D([a, b], [c, d]) = \max_{t_1 \in [a, b], t_2 \in [c, d]} |t_1 - t_2| \quad (5)$$

This definition combined with the complete-linkage hierarchical clustering of retention-time intervals permits construction of deconvolution windows no greater than a specified maximum width. Therefore, deconvolution windows are determined in four steps: (i) precluster the retention-time intervals of EIC peaks by finding groups of peaks that do not overlap in the retention time domain, (ii) perform the complete-linkage sparse hierarchical clustering²⁹ on each group, (iii) if some clusters overlap in the retention-time domain, then we adjust the boundary between those clusters by finding the minimum of the total ion chromatogram over the overlapping region; (iv) the adjusted boundaries of each cluster become the boundaries of deconvolution windows.

Number of Components.

Within each deconvolution window, ADAP-GC 4.0 performs an additional clustering of the EIC peaks in that window to determine the number of components. Determining the number of components has always been a challenging problem in both the spectral deconvolution and the general MCR. For instance, Stein used the sharpness values of EIC peaks by calculating their local maxima and the corresponding ranges of uncertainty,¹ Tsugawa used the ideal slope score and the second Gaussian derivative filter applied to sharpness values,³⁰ Domingo-Almenara used a multivariate matching filter method²³ to determine the number of components and their model peaks. Moreover, the cross-correlation and information-theoretic criteria³¹ can be used to assess the number of components. In the previous versions of ADAP-GC, EIC peak filtering and hierarchical clustering^{18,19,28} of the shape similarities between EIC peaks were used for that purpose. However, in ADAP-GC 4.0, we cluster the retention time of EIC peak apexes, because it does not require ad-hoc filtering, uses only one user-defined parameter and is more computationally efficient than the cross-correlation and information-theoretic criteria.

Before determining the number of components, we optionally adjust the retention times of EIC peak apexes. In the previous versions of ADAP-GC, the apex retention time of an EIC peak was calculated as retention time of the highest point in the elution profile of that EIC peak. That method is sufficient and preferred when the elution profile is comprised of many time points, as in the case of the GC–MS data acquired at the unit mass resolution. However, if the elution profile consists of only a few points, the retention times of EIC peaks should be calculated more accurately. For instance, the EIC peaks in Figure 3A are produced by two compounds eluting at the retention times 7.88 and 7.92 min. When the highest point in an elution profile is used for calculating the retention time of an EIC peak, our clustering algorithm has determined the presence of three components instead of two (Figure 3B). The latter happens because of the high variation of calculated retention times of the EIC peaks produced by the second compound. In order to improve the estimation of the number of components, the retention times of EIC peaks are adjusted by fitting a parabola into each EIC peak and using the top vertex of that parabola to calculate the retention time of an EIC peak (Figure 3D). Figure 3C shows that the adjusted retention times of EIC peaks have a

lower variation and the clustering algorithm correctly determines two components. Therefore, adjusting the retention times of EIC peaks leads to a more accurate estimation of the number of components in a deconvolution window. A similar approach was used by Stein,¹ where the top three points of EIC peaks are used to fit a parabola. Here, we use all data points above the half of the maximum intensity of an EIC peak. Using a variable number of the fitting points helps adjusting the apex retention time in both cases of a high and low number of time points that comprise an EIC peak.

Multivariate Curve Resolution.

After the number of components is determined, ADAP-GC 4.0 constructs the model peaks of components by performing the multivariate curve resolution (MCR) on the detected EIC peaks. The MCR algorithm iteratively searches for a solution of the minimization problem given by eq 4. Specifically, the new matrices C and S are calculated in each iteration, using the multiplicative update rules.¹² It was shown that by performing their updates multiple times, the derived matrices C and S would converge to a locally minimum solution of the optimization problem.³² However, that solution may not necessarily produce the best model peaks of components in terms of the library matching scores and relative quantitation. To achieve a better library matching and relative quantitation performance, we perform additional data preprocessing steps specific to the spectral deconvolution. For more details on these steps, see SI.

Decomposition of Chromatograms.

For each perceived component, the MCR algorithm constructs a model peak represented by a column of matrix C . At the same time, it decomposes each detected EIC peak into a linear combination of model peaks, thus constructing a fragmentation spectrum for each component represented by a row of matrix S . However, if matrix X in eq 4 does not contain a certain EIC peak, then that EIC peak will not be included into the constructed fragmentation spectra. This puts extremely stringent requirements on the peak detection step, which is executed prior to the spectral deconvolution: a peak detection algorithm must be able to detect all true EIC peaks and ignore false peaks produced by noise.^{21,23} Unfortunately, no peak detection algorithm is perfect.

To relax such stringent requirements on the peak detection step and be able to recover missed EIC peaks, ADAP-GC 4.0 decomposes all EICs including the EICs with no peaks detected, into a linear combination of model peaks and a horizontal baseline (Figure 2). As a result, a missed EIC peak can be recovered from a chromatogram and, therefore, the corresponding m/z would be present in the constructed pure spectra. This new workflow allows ADAP-GC 4.0 to extract EIC peaks of low signal-to-noise ratios using the information from EIC peaks of higher signal-to-noise ratios, that are produced by the same component.

Data Results.

We compared the new workflow to three freely available software tools: ADAP-GC 3.2,²⁸ MS-DIAL,³⁰ and eRah.²³ These tools have been developed in the last five years and represent the latest achievements in preprocessing GC-MS data. All three perform spectral

deconvolution but use different algorithms. ADAP-GC 3.2 performs the spectral deconvolution by filtering EIC peaks, clustering those peaks, selecting a model peak in each cluster, and then constructing the pure fragmentation mass spectra of components by decomposing EIC peaks into linear combinations of the model peaks.^{18–20} Spectral deconvolution in MS-DIAL is based on the algorithm first proposed by R. Dromey et al.³⁴ and used by many software tools including AMDIS³⁵ and MetaboliteDetector³⁶ that had been developed prior to MS-DIAL. The Dromey algorithm groups EIC peaks and chooses a model peak in each group, based on the sharpness values calculated for each EIC peaks. The most recent tool, eRah, performs MCR-based spectral deconvolution by the alternating-least-squares algorithm.⁵

Comparison of the library matching and relative quantitation results by the four software tools for the standard-mixture data produced at unit mass resolution, are presented in Table 1. To perform matching, we export constructed fragmentation spectra from the four software tools to MSP-files. Then, we use NIST MSPepSearch, a console version of NIST MS Search, to match the constructed fragmentation spectra against an in-house library produced by the same procedure that was used for generating the standard-mixture and urine data sets (see Experimental Section). The matching scores are calculated using the Simple Similarity Reverse search option and an intensity-based filter so that peaks of intensity below 5% of the highest peak in the spectrum do not participate in the library matching. We consider components matched if their matching scores exceed 800.

Regarding the library matching and relative quantitation results produced by ADAP 3.2 and 4.0, we observe their high similarity. However, ADAP-GC 4.0 was able to produce significantly higher matching scores for L-norlucine due to its ability to recover missing EIC peaks. Moreover, the main improvement of the new version of ADAP-GC lies in the number of user-specified parameters. Notice that ADAP-GC 3.2 requires its users to specify about 10 parameters and a list of excluded m/z values in order to produce the results listed in Table 1. In contrast, the spectral deconvolution algorithm in ADAP-GC 4.0 has only four parameters and does not require an exclusion of any m/z values.

The results produced by MS-DIAL slightly diverge from ADAP-GC 3.2 and 4.0 results. Specifically, the pairs of coeluting compounds asparagine/D-xylose and L-histidine/L-lysine were detected only in a small number of samples (Table 1). Indeed, Asparagine has not been matched with a score above 800 in any of the seven samples, while d-xylose was matched only in four samples out of seven with the average score 852. It seems that these library matching results are due to overestimating the number of components, which caused missing certain peaks in the constructed fragmentation spectra. Unfortunately, we could not come up with MS-DIAL parameters that would allow detecting the right number of components in the pairs asparagine/d-xylose and L-histidine/L-lysine while providing high library matching and relative quantitation results for other compounds. The results produced by eRah²³ demonstrate the highest inconsistency among the compared software tools. For instance the matching score for β -amino isobutyric acid in sample *S0.4.cdf* is 658, whereas the same compound in other samples is matched with the scores about 950. Asparagine was completely missed in sample *S0.8.cdf*, while the matching score for asparagine in the other

samples is about 810. The matching scores for L-histidine range from 609 to 932, and for guanosine from 449 to 914.

Table 2 contains the library matching results produced for standard-mixture data acquired at the high mass resolution. Similar to the unit-mass-resolution data, the pure fragmentation spectra are constructed by ADAP-GC 4.0, ADAP-GC 3.2, MS-DIAL, and eRah, exported to MSP-files, and matched by NIST MSPepSearch. However, the fragmentation spectra are matched against NIST MainLib and RepLib EI spectral libraries and the Simple Similarity matching score without the reverse option is used because of a large number of compounds in those libraries. As before, peaks of intensity below 5% of the highest peak in a spectrum do not participate in the library search, and a component is considered matched if its matching score exceeds 800. In Table 2, we do not provide R^2 values because only one replica of each sample was available for data preprocessing.

For the high-mass-resolution data, ADAP-GC 4.0 and ADAP-GC 3.2 again demonstrate almost identical results with a similar number of matched compounds and a slightly higher average matching score for ADAP-GC 4.0. The differences between ADAP-GC 4.0 and MS-DIAL are more significant. In fact, in several samples such as *PBDE_Tech_-Mixes.mzXML* and *Pest_mix_09.mzXML*, ADAP-GC 4.0 detected a larger number of compounds with a matching score above 800, while in other samples such as *Pest_-mix_08.mzXML* and *Pest_mix_10.mzXML*, MS-DIAL did a better job with detecting more compounds than ADAP-GC 4.0 did. Therefore, neither ADAP-GC 4.0 nor MS-DIAL has a clear advantage in terms of the number of detected compounds for high-mass-resolution data. However, we observe that ADAP-GC 4.0 typically produces higher matching scores for detected compounds. Regarding eRah, its library matching results for the high-mass-resolution data are consistently lower both in terms of the number of detected compounds and their matching scores. See SI for more details.

CONCLUSION

The ADAP-GC 4.0 workflow introduces a new MCR-based spectral deconvolution algorithm for preprocessing GC–MS metabolomics data. In this new algorithm, the entire retention time range is split into deconvolution windows to speed up computations, the peak apex retention times are adjusted to improve the inference of the number of components, and in each deconvolution window a separate MCR problem is solved with scaled EIC peaks and a baseline added to components. After model peaks are constructed, a decomposition of EICs is performed instead of decomposing EIC peaks, so that the spectral deconvolution can recover EIC peaks missed by the peak detection step.

Performance of the new spectral deconvolution algorithm and the whole workflow has been tested on GC-TOF data sets from a 27-standards mixture and urine with the mixture spiked in, and GC–Orbitrap data sets from mixtures of different standards. In all cases, ADAP-GC 4.0 demonstrates performance comparable to or better than other compared software tools (MS-DIAL, ADAP-GC 3.2, and eRah). Moreover, ADAP-GC 4.0 features significantly less user-defined parameters than ADAP-GC 3.2. Ref 33.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was financially supported by the United States National Science Foundation grant award 1262416 and National Institutes of Health grant award U01CA235507.

REFERENCES

- (1). Stein S J. Am. Soc. Mass Spectrom 1999, 10, 770–781.
- (2). Garrido M; Rius FX; Larrechi MS Anal. Bioanal. Chem 2008, 390, 2059–2066. [PubMed: 18320174]
- (3). Gao H-T; Li T-H; Chen K; Li W-G; Bi X Talanta 2005, 66, 65–73. [PubMed: 18969963]
- (4). Shao X; Liu Z; Cai W Analyst 2009, 134, 2095–2099. [PubMed: 19768219]
- (5). Domingo-Almenara X; Perera A; Ramírez N; Cañellas N; Correig X; Brezmes J J. Chromatogr A 2015, 1409, 226–233. [PubMed: 26210114]
- (6). Anbumalar S; Anandanatarajan R; Rameshbabu P Int. J. Comput. Appl 2013, 63, 1–10.
- (7). de Godoy LAF; Hantao LW; Pedroso MP; Poppi WJ; Augusto F Anal. Chim. Acta 2011, 699, 120–125. [PubMed: 21704766]
- (8). Omar J; Olivares M; Amigo JM; Etxebarría N Talanta 2014, 121, 273–280. [PubMed: 24607138]
- (9). Zushi Y; Hashimoto S; Tanabe K Anal. Chem 2015, 87, 1829–1838. [PubMed: 25572711]
- (10). Hantao LW; Aleme HG; Pedroso MP; Sabin GP; Poppi RJ; Augusto F Anal. Chim. Acta 2012, 731, 11–23. [PubMed: 22652260]
- (11). Tauler R J. Chemom 2001, 15, 627–646.
- (12). Vosough M; Salemi A Talanta 2007, 73, 30–36. [PubMed: 19071846]
- (13). Khayamian T; Tan G; Sirhan A; Siew Y; Sajjadi S Chemom. Intell. Lab. Syst 2009, 96, 149–158. Chimimetrie 2007, Lyon, France, 29–30 November 2007.
- (14). Jalali-Heravi M; Parastar H; Ebrahimi-Najafabadi H Anal. Chim. Acta 2010, 662, 143–154. [PubMed: 20171313]
- (15). Jalali-Heravi M; Moazeni RS; Sereshti H Journal of Chromatography A 2011, 1218, 2569–2576. [PubMed: 21429498]
- (16). Anbumalar S; Natarajan RA; Rameshbabu P Applied Mathematics and Computation 2014, 241, 242–258.
- (17). Yang R; Zhao N; Xiao X; Yu S; Liu J; Liu W J. Chemom 2015, 29, 442–447.
- (18). Ni Y; Qiu Y; Jiang W; Suttlemyre K; Su M; Zhang W; Jia W; Du X Anal. Chem 2012, 84, 6619–6629. [PubMed: 22747237]
- (19). Ni Y; Su M; Qiu Y; Jia W; Du X Anal. Chem 2016, 88, 8802–8811. [PubMed: 27461032]
- (20). Smirnov A; Jia W; Walker DI; Jones DP; Du X J. Proteome Res 2018, 17, 470–478. [PubMed: 29076734]
- (21). Myers OD; Sumner SJ; Li S; Barnes S; Du X Anal. Chem 2017, 89, 8689–8695. [PubMed: 28752757]
- (22). Tsugawa H; Cajka T; Kind T; Ma Y; Higgins B; Ikeda K; Kanazawa M; VanderGheynst J; Fiehn O; Arita M Nat. Methods 2015, 12, 523–6. [PubMed: 25938372]
- (23). Domingo-Almenara X; Brezmes J; Vinaixa M; Samino S; Ramirez N; RamonKrauel M; Lerin C; Diaz M; Ibanez L; Correig X; Perera-Lluna A; Yanes O Anal. Chem 2016, 88, 9821–9829. [PubMed: 27584001]
- (24). ChromaTOF <https://www.leco.com/products/separation-science/software-accessories/chromatof-software> (accessed January 8, 2018).
- (25). AnalyzerPro <https://www.spectralworks.com/analyzerpro.html> (accessed January 8, 2018).
- (26). Pluskal T; Castillo S; Villar-Briones A; Oresic M BMC Bioinf 2010, 11, 395.

- (27). Wang G; Ding Q; Hou Z TrAC, Trends Anal. Chem 2008, 27, 368–376.
- (28). Smirnov A; Jia W; Walker DI; Jones DP; Du X J. Proteome Res 2018, 17, 470–478. [PubMed: 29076734]
- (29). Nguyen T-D; Schmidt B; Kwoh C-K Pcedia Comput. Sci 2014, 29, 8–19.
- (30). Tsugawa H; Cajka T; Kind T; Ma Y; Higgins B; Ikeda K; Kanazawa M; VanderGheynst J; Fiehn O; Arita M Nat. Methods 2015, 12, 523–526. [PubMed: 25938372]
- (31). Hui M; Li J; Wen X; Yao L; Long Z PLoS One 2011, 6, 1–10.
- (32). Lee DD; Seung HS Algorithms for Non-negative Matrix Factorization In NIPS 2000; pp 556–562.
- (33). Myers OD; Sumner SJ; Li S; Barnes S; Du X Anal. Chem 2017, 89, 8696–8703. [PubMed: 28752754]
- (34). Dromey RG; Stefik MJ; Rindfleisch TC; Duffield AM Anal. Chem 1976, 48, 1368–1375.
- (35). AMDIS. <http://www.amdis.net/> (accessed January 22, 2019).
- (36). MetaboliteDetector <http://md.tu-bs.de/> (accessed January 5, 2018).

Figure 1.

Differences in spectral deconvolution results after the traditional and MCR approaches have been applied to deconvoluting two coeluting compounds L-histidine and L-lysine in file SI S1_1.cdf. (A) Results produced by the traditional approach. (B) Results produced by the MCR approach. (A Left) Peaks P_{154} and P_{73} are EIC peaks detected for $m/z = 154$ and 73 respectively, and are selected as model peaks for the two compounds; (A middle) The spectrum constructed for L-histidine completely misses peak $m/z = 73$ and produced a low matching score 680; (A right) The spectrum constructed for L-lysine produced a high matching score 923; (B left) The red and blue model peaks are constructed by MCR; (B middle and right) Both of the constructed spectra produced high matching scores 962 and 977. In (A) and (B), spectra in red and blue are produced by spectra deconvolution and spectra in black are from the spectral library produced on exactly the same GC-MS platform. Matching scores are calculated by *NIST MSSearch* using the reverse identity search.

Figure 2. ADAP-GC 4.0 data preprocessing workflow. The entire retention-time range is split into disjoint deconvolution windows and the MCR-based spectral deconvolution is carried out within each window.

Figure 3.

Clustering of apex retention times in file *Pest_mix_14.mzXML* before adjustment (B) and after adjustment (C). Apex retention time before adjustment is the retention time of the highest point of EIC peak. Apex retention time after adjustment is the retention time of a parabola fitted to the top points of EIC peak (D).

Table 1.
Comparison of Identification and Relative Quantitation Results by the Four Software Tools for the Standard-Mixture Data Acquired at Unit Mass Resolution

compound name	ADAP 4.0				ADAP 3.2				MS-DIAL				eKah			
	mass	score	qty	R ²	mass	score	Qty	R ²	mass	score	qty	R ²	mass	score	qty	score
pyruvic acid	5.17	73	941	7	0.999	174	935	7	0.999	174	923	6	0.920	73	942	7
propanoic acid	5.34	73	977	7	0.999	117	978	7	0.999	117	963	7	0.921	73	954	7
β -Amino isobutyric acid	7.47	102	972	7	1.000	102	971	7	0.999	102	972	7	0.919	102	915	6
L-norleucine	8.40	158	928	7	0.999	158	847	6	0.999	158	906	7	0.917	158	923	7
alloisoleucine	8.73	158	919	7	0.999	158	919	7	0.998	158	916	7	0.916	73	902	7
proline	8.78	142	989	7	0.999	142	987	7	0.999	142	991	7	0.917	142	974	7
glyceric acid	9.34	73	982	7	0.999	189	982	7	0.998	189	981	7	0.916	73	927	7
threonine	10.31	73	980	7	0.998	57	979	7	0.997	117	962	7	0.915	73	916	6
5-oxoproline	12.80	156	940	7	1.000	156	936	7	0.999	156	953	7	0.916	156	935	7
L-cysteine	13.57	73	959	1	NA	115	960	1	NA	115	950	6	0.867	73	932	5
creatinine	13.57	73	964	6	0.994	115	965	6	0.985	115	984	1	NA	73	951	2
citulline	14.84	73	997	7	0.997	70	994	7	0.997	70	933	7	0.916	73	946	7
d-Xylose	15.94	73	988	7	0.999	103	988	7	0.999	103	983	7	0.914	73	956	7
asparagine	16.14	73	899	5	1.000	116	880	6	0.997	75	667	0	NA	73	812	4
d-xylose	16.16	73	981	7	0.994	103	980	7	0.994	103	852	4	0.916	73	965	7
1,4-butanediamine	17.59	174	968	7	0.995	174	968	7	0.996	174	969	7	0.913	174	954	7
glycerolphosphate	18.50	73	966	7	0.999	73	957	7	0.998	75	931	7	0.895	73	947	7
chlorophenylalanine	18.95	73	985	7	NA	218	984	7	NA	218	962	7	NA	73	963	7
citric acid	19.80	73	993	7	1.000	273	982	7	0.998	75	973	7	0.995	73	956	7
isocitric acid	19.87	73	934	7	0.999	245	922	7	0.997	75	945	7	0.911	73	873	6
L-histidine	21.92	73	957	7	0.998	154	933	7	0.994	154	759	3	0.997	73	744	2
L-lysine	21.96	73	982	7	0.995	174	979	7	0.994	174	917	5	0.966	73	973	7
mannitol	22.61	73	974	7	0.997	103	974	7	0.996	103	976	7	0.913	73	925	6
galic acid	22.87	73	991	7	0.999	281	997	7	0.998	281	988	7	0.918	73	957	7
N-acetyl glucosamine methoxime	25.96	73	970	7	0.989	87	979	7	1.000	129	825	4	0.984	73	915	6
L-tryptophan	27.94	73	985	7	0.998	202	984	7	0.996	202	992	7	0.913	73	915	6

compound name	ADAP 4.0				ADAP 3.2				MS-DIAL				eRah			
	mass	score	qty	R ²	mass	score	Qty	R ²	mass	score	qty	R ²	mass	score	qty	
adenosine	31.38	73	923	7	0.997	230	922	7	0.994	236	919	7	0.908	73	915	7
guanosine	32.30	73	849	6	0.994	103	846	6	0.991	103	854	6	0.910	73	736	3
average			960		0.997		955		0.997		923		0.926		919	

Table 2.

Aggregated Library Matching Results for the Standard-Mixture Data Acquired at High Mass Resolution^a

sample name	ADAP 4.0		ADAP 3.2		MS-DIAL		eReh	
	qty	score	qty	score	qty	score	qty	score
dioxins	4	878	3	845	3	863	3	821
furans	4	904	3	875	3	890	3	878
PBB153	1	916	1	891	1	881	1	808
PBDE_Tech_Mixes	6	935	6	904	6	873	5	824
PCB_Cong_Calmix	14	950	14	936	14	918	14	905
PCB_Content_Eval_Mix1	6	960	6	947	6	933	6	880
PCB_Content_Eval_Mix2	3	958	3	931	3	916	3	901
Pest_Mix_08	16	897	14	906	14	890	15	849
Pest_Mix_09	40	927	38	920	38	907	37	835
Pest_Mix_10	25	897	22	905	23	906	24	834
Pest_Mix_11	28	927	26	931	25	907	26	914
Pest_Mix_12	35	918	32	916	33	905	31	895
Pest_Mix_13	27	954	27	948	27	916	27	924
Pest_Mix_14	9	952	9	966	9	946	9	938
Pest_Mix_15	24	921	23	929	23	888	22	866
Pest_Mix_16	8	916	8	926	8	876	6	835
average		926		917		901		869

^aFor individual results, see SI.