Published in final edited form as:

Anal Chem. 2016 September 06; 88(17): 8802–8811. doi:10.1021/acs.analchem.6b02222.

ADAP-GC 3.0: Improved Peak Detection and Deconvolution of Co-eluting Metabolites from GC/TOF-MS Data for Metabolomics Studies

Yan Ni[†], Mingming Su[†], Yunping Qiu[‡], Wei Jia[†], and Xiuxia Du[¶]

[†]University of Hawaii Cancer Center, Honolulu, HI 96813

[‡]Albert Einstein College of Medicine, Bronx, NY10461, USA

[¶]University of North Carolina at Charlotte, Charlotte, NC 28223, USA

Abstract

ADAP-GC is an automated computational pipeline for untargeted, GC-MS-based metabolomics studies. It takes raw mass spectrometry data as input and carries out a sequence of data processing steps including construction of extracted ion chromatograms, detection of chromatographic peak features, deconvolution of co-eluting compounds, and alignment of compounds across samples. Despite the increased accuracy from the original version to version 2.0 in terms of extracting metabolite information for identification and quantitation, ADAP-GC 2.0 requires appropriate specification of a number of parameters and has difficulty in extracting information of compounds that are in low concentration. To overcome these two limitations, ADAP-GC 3.0 was developed to improve both the robustness and sensitivity of compound detection. In this paper, we report how these goals were achieved and compare ADAP-GC 3.0 against three other software tools including ChromaTOF, AnalyzerPro, and AMDIS that are widely used in the metabolomics community.

Introduction

Gas chromatography-time-of-flight mass spectrometry (GC/TOF-MS) is one of the most widely used analytical platforms in untargeted metabolomics studies. The advantages of this platform include high chromatographic resolution, high mass measurement accuracy, and rapid spectrum acquisition. Like other analytical platforms such as nuclear magnetic resonance (NMR) spectroscopy and liquid chromatography mass spectrometry (LC-MS) that are commonly used in metabolomics studies as well, GC/TOF-MS produces large and complex datasets that require specialized computational tools for extracting qualitative and quantitative information of metabolites from the raw GC/TOF-MS data. The procedure of information extraction consists of multiple steps. Among these steps, detection of chromatographic peaks and deconvolution (i.e. computational separation) of compounds that co-elute from the chromatography are essential. This is because samples for untargeted metabolomics studies are very complex and usually contain hundreds of compounds that a

Correspondence to: Xiuxia Du.

This material is available free of charge via the Internet at <http://pubs.acs.org/>.

mass spectrometer is able to measure. These compounds produce chromatographic peaks of varying width and shapes due to their diverse physical and chemical properties, which makes it non-trivial for an algorithm to determine the apex and boundaries of all of the peaks in a data file in a robust fashion. In addition, the complexity of the samples causes co-elution to occur frequently despite advances in chromatography and makes it imperative to separate them computationally through deconvolution.

ADAP (Automated Data Analysis Pipeline) was developed as a complete and automated computational pipeline to deconvolute co-eluting compounds and extract metabolite information from GC/TOF-MS data. The first version of ADAP was equipped with the capabilities of peak detection, deconvolution, component-based alignment of samples, and library matching.¹ Deconvolution is achieved by clustering chromatographic peaks based on peak shape similarity. However, fragment ions that are produced by more than one co-eluting components are assigned to the most dominating component rather than assigned to all of the components that have produced these fragments. This issue was addressed in ADAP-GC 2.0 through a sequence of steps.² These steps include defining and detecting simple and composite peak features, selecting model peaks based on five metrics of peak quality, and utilizing constrained optimization to decompose composite peak features into a linear combination of simple ones. Compared to the first version of ADAP, ADAP-GC 2.0 substantially increased the accuracy of identification and quantification of co-eluting compounds.

Despite the increased accuracy in the extraction of metabolite information, the approaches that ADAP-GC 2.0 used for peak detection and deconvolution called for the appropriate specification of a number of analysis parameters. As a result, results from the deconvolution algorithm could vary due to differences in the parameters. In addition, ADAP-GC 2.0 has difficulty in extracting information of compounds that are in low concentration. These two limitations motivated us to develop ADAP-GC 3.0 with the goal to improve both the robustness and sensitivity of compound deconvolution. These two goals were achieved by improvements in the detection of chromatographic peaks and in the detection and selection of model peaks. In this paper, we report the details of these improvements and a comparative evaluation of ADAP-GC 3.0 against three other software tools that are used widely by the metabolomics community. These software tools include AnalyzerPro (SpectralWorks Ltd, UK), ChromaTOF (LECO Corporation, St. Joseph, MI, USA), and AMDIS (National Institute of Standards and Technology, Gaithersburg, MD, USA). ADAP-GC 3.0 is written in R and C and would be made open source and freely available to the metabolomics community.

Experimental procedures

Two different types of samples (named Sample I and II) were used for the development, evaluation, and validation of ADAP-GC 3.0. The samples were derivatized and analyzed on the GC/TOF-MS platform with the same protocol as that described in our previous publications. Briefly, after TMS derivatization, each 1 μ L aliquot of the derivatized solution was injected in splitless mode into an Agilent 6890N GC system (Santa Clara, CA, USA) that was coupled with a Pegasus HT TOF-MS (LECO Corporation, St. Joseph, MI, USA).

Separation was achieved on a DB-5 ms capillary column (30 m × 250 µm I.D., 0.25 µm film thickness; Agilent J&W Scientific, Folsom, CA, USA), with helium as the carrier gas at a constant flow rate of 1.0 mL/min. The temperature of injection, transfer interface, and ion source was set to 260 °C, 260 °C, and 210 °C, respectively. The GC temperature programming was set to 2 min isothermal heating at 80 °C, followed by 10 °C/min oven temperature ramps to 220 °C, 5 °C/min to 240 °C, and 25 °C/min to 290 °C, and a final 8 min maintenance at 290 °C. Electron impact ionization (70 eV) at full scan mode (m/z 40–600) was used, with an acquisition rate of 20 spectra per second in the TOF/MS setting.

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) ng/mL of each compound). With four pairs of co-eluting compounds in each sample, we were able to evaluate the overall performance of peak detection and deconvolution of ADAP-GC 3.0 and to compare it against ADAP-GC 2.0, ChromaTOF, AnalyzerPro, and AMDIS.
2. Urine samples with standard mixtures spiked in (Sample II): Sample II was prepared by spiking into a pooled urine sample the seven calibration curve samples of Sample I and an additional sample consisting of 0.1 µg/mL of each standard compound. Sample II was used for evaluating the performance of ADAP-GC 3.0 in terms of processing complex samples.

All of the raw mass spectrometry data files would be made available at <http://www.du-lab.org>.

Data analysis methods

Workflow

Figure 1 displays the workflow of ADAP-GC 3.0 which consists of five steps: (1) extraction and processing of extracted ion chromatograms (EICs), (2) detection of chromatographic peak features, (3) deconvolution, (4) alignment, and (5) QUAL/QUAN analysis. These five steps constitute the workflow of ADAP-GC 1.0 and ADAP-GC 2.0 as well, but with different sub-steps. As summarized in Figure 1, the major differences between ADAP-GC 3.0 and the previous two versions are in the detection of chromatographic peak features and deconvolution.

Detection of chromatographic peak features

This is the second step in the data processing workflow as depicted in Figure 1. In the first step, both the total ion chromatogram (TIC) and extracted ion chromatograms (EIC) have been extracted from the raw mass spectrometry data, denoised, and baseline corrected. From these chromatograms, peak features are to be detected in this second step. A chromatographic peak feature is defined as an observed, temporal, and bell-shaped signal intensity pattern in either the TIC or an EIC and could be numerically represented by the peak apex, the left boundary, the right boundary, and the signal intensity pattern between the two boundaries.

The peak detection method in ADAP-GC 2.0 relies on detecting the local maximum and local minimum within a window of pre-specified width. Since chromatographic peak features of different compounds in the same EIC could have very different widths, it is almost impossible to find one width parameter that will allow the detection of all of the real peak features. So the challenge in peak detection is to develop a method that is robust enough to detect peak features of varying peak width.

Toward this end, we have applied continuous wavelet transform in combination with the original local maximum and local minimum method to detect chromatographic peak features. Wavelet transform is a signal processing technique that can represent a one-dimensional temporal signal in a two-dimensional time-scale space. This redundant way of representing the 1D temporal signal in a 2D space facilitates the detection of not only the different frequencies that the signal contains, but the temporal location of the frequency components as well. As a result, wavelet transform has been applied widely in the analysis of non-stationary signals (i.e. the frequency content of the signal changes with respect to time). EICs and TIC that are extracted from GC/TOF-MS data are non-stationary signals. As such, results from the wavelet transform automatically provide information for locating the time interval where a chromatographic peak appears, regardless of the width of the chromatographic peak feature. This is the robustness we desire for the peak detection method. Continuous wavelet transform has been successfully applied in XCMS and MZmine 2 for processing LC/MS-based metabolomics data.³⁻⁶

In ADAP-GC 3.0, each peak feature that the wavelet transform-based method has detected is further examined for determining whether it is a simple peak feature or is part of a composite peak feature. A simple peak feature refers to a chromatographic peak feature (CPF) that results from a single component, whereas a composite peak feature results from summing signals of two or more components. A simple CPF has only one local maximum, and a composite CPF could have one, two, or more local maxima. Three ratios are used for the determination. The first two are the ratios of intensity values at the left and right boundary, respectively, to the intensity value at the peak apex. The third is the ratio of the difference in intensity values between the left and right boundary to the intensity value at the peak apex. If one or more of these three ratios for a peak feature is greater than the corresponding threshold, the peak feature will be considered as a part of a composite peak feature. Only those peak features that have all of the three ratios smaller than the corresponding threshold values are considered as simple peaks. These simple peaks will be candidates of model peaks for subsequent deconvolution (to be described). Figure 2A depicts two peaks (indicated by red circles) that are detected by the wavelet transform method and turn out to belong to a composite peak feature based on the ratio criteria.

Despite the advantage of the wavelet transform method in ensuring the robustness of peak detection, it could miss small peak features that are next to a dominant peak feature. Figure 2B depicts such an example. The three peak features indicated by blue circles are all missed by the wavelet transform due to the dominance of the peak indicated by the red circle. As a result, the peak feature indicated by the red circle in Figure 2B would be mistakenly considered as a simple peak feature and would then be a candidate for model peak for deconvolution.

To resolve this issue, ADAP-GC 3.0 further examines each simple peak feature that the wavelet method has detected by checking the surrounding area using the local maximum approach. Peak detection using the local maximum method is very sensitive when the width of the detection window is small. If there do exist small peaks that the wavelet method has missed, the peaks that are originally considered to be simple peaks are re-classified as composite peak features. As such, ADAP-GC 3.0 takes advantage of the robustness of the wavelet transform method and the sensitivity of the local maximum method to ensure that only truly simple peak features are selected as model peaks for final deconvolution.

Deconvolution

With peaks first detected from the TIC and all of the EICs and subsequently classified as either simple or composite, deconvolution can now proceed. Like deconvolution in ADAP-GC 2.0, deconvolution is carried out for each deconvolution window and proceeds through the entire elution duration. Within each window, the process of deconvolution consists of multiple steps that include: (1) determination of the number of components; (2) selection of model peaks; (3) decomposition of each composite peak feature into a linear combination of model peak features; (4) construction of pure spectrum for each component, and (5) correction of splitting issues.

Among these steps, selection of model peaks is the most critical. A model peak feature represents the elution profile of the corresponding compound. In the previous two versions of ADAP-GC, model peaks are required to meet a number of criteria including signal to noise ratio, sharpness, peak intensity, mass value, and similarity to a symmetric bell curve. However, after ADAP-GC 2.0 was completed, we have observed that asymmetric bell curves are not uncommon due to fronting peaks that could be caused by column overload, improper column installation, injection technique inconsistencies, or reverse solvent effect and tailing peaks that could be caused by inlet contamination, column blockage, improper column installation, solvent polarity mismatch, reverse solvent effect, or solvent effect violation. In addition, the combination of the aforementioned five criteria is so stringent that very good model peaks could be filtered out, which caused compounds went undetected. To resolve this issue, we have revised the strategy for selecting model peaks and determining the total number of components within each deconvolution window. These two tasks correspond to the aforementioned steps (1) and (2). Next, we describe these two steps in detail. Steps (3)–(5) are essentially the same as those in ADAP-GC 2.0 and we will only briefly summarize. Key steps in deconvolution are illustrated in Figure 3 using an example.

Determine the total number of components—This determination is accomplished by carrying out two steps of hierarchical clustering. During the first clustering, the retention time corresponding to the apex of all of the simple and composite peak features within a deconvolution window is clustered and clusters are determined using a relatively large dissimilarity threshold (i.e. elution time difference). This clustering serves an important purpose. That is to determine the minimal number of components within this deconvolution window and to ensure that hard-to-detect compounds, in particular low-concentration compounds, would be identified. If, after the subsequent second clustering (to be described next), it is found that, as a result of filtering, the total number of components is fewer than

that determined in the first round of clustering, ADAP-GC 3.0 would directly construct the mass spectrum based on the clusters obtained in the first clustering. An example will be shown in the Results section.

It is worth repeating that the first clustering is applied to the retention time corresponding to the apex of all of the simple and composite peak features within a deconvolution window. Subsequently, ADAP-GC 3.0 applies the second clustering to the elution profile (intensity values between peak boundaries) of peak features within each cluster obtained in the first clustering. Peak features that participate in the second clustering must meet two criteria. The first criterion is that the peak features are simple and, as a result, have a high likelihood of being unique. The second criterion is that the signal-to-noise ratio values are greater than a threshold and therefore the peak features are pure enough (i.e. having suffered negligible interference from noise) to serve as candidates for model peaks. In the second clustering, dot product between elution profiles of peak features is used as the similarity measure. As a result of the second hierarchical clustering, one or more sub-clusters could be produced for each cluster that is obtained in the first clustering. The total number of sub-clusters within the corresponding deconvolution window is the total number of components unless it is smaller than the total number of clusters obtained in the first clustering.

Select the model peak for each component—The next step is to determine the best model peak for each component in the sense of purity. All of the three versions of ADAP-GC take into account peak sharpness to evaluate the degree of purity of the candidate peaks for model peaks. However, version 1.0 and 2.0 do not take peak width into consideration. As a result, wide peaks could have high sharpness values due to cumulative summation of point-to-point change and the resulting sharpness values cannot truly reflect the sharpness characteristics of peaks.

To address this issue, version 3.0 modifies how the sharpness value is calculated. The modified sharpness bears a similarity to the sharpness value that AMDIS calculates, but is simpler. ADAP-GC 3.0 first defines the sharpness values between the maximum abundance A_{max} and an abundance value located n scans from the maximum A_n as:

$$\frac{A_{max} - A_n}{n}$$

The median sharpness values on each side are found and averaged. ADAP-GC 3.0 defines this average sharpness value as the sharpness of the entire peak feature.

AMDIS defines its sharpness as

$$\frac{A_{max} - A_n}{n \cdot N_f \cdot \sqrt{A_{max}}}$$

Compared to AMDIS's sharpness, ADAP's definition removed N_f and $\sqrt{A_{max}}$ where N_f is the noise factor. Three reasons are behind this modification: (1) AMDIS defines N_f to reflect the overall noisiness of each data file. In contrast, ADAP-GC 3.0 calculates the signal-to-

noise ratio of each peak feature and removes noisy peak features from participating in model peak selection by filtering them out using a signal-to-noise ratio threshold. (2) In GC/TOF-MS data, peaks of higher abundance values tend to be sharper and smoother and are better candidates for model peaks. Therefore, ADAP-GC 3.0 did not normalize the sharpness value using the maximum abundance value of the corresponding peak feature.

For each cluster obtained in the second clustering, ADAP-GC 3.0 calculates the sharpness value of all of the peak features. The peak feature with the largest sharpness value is selected as the model peak for the corresponding component.

Construct pure spectra and correct splitting issues—As in ADAP-GC 2.0, constrained optimization is applied for finding the optimal linear combination of model peaks to approximate each simple and composite peak feature. After decomposing all of the peak features within a deconvolution window, all of the resulting weights that correspond to the same model peak form a mass spectrum. The m/z values of the peak features give rise to the m/z -axis and the magnitude of the weights gives rise to the intensity axis. When there are two or more model peaks for the same component, two or more spectra are constructed for the same component and splitting issue occurs. Splitting issue affects the accuracy of compound identification and quantitation and must be corrected. ADAP-GC 2.0 corrects it by calculating the similarity between every pair of spectra that have been constructed and are close in retention time. For highly similar spectra, the corresponding model peaks are compared, the best one in terms of sharpness is kept, the other ones are discarded, and a second run of decomposition is carried out.

Overall computational workflow of deconvolution—The overall workflow is depicted in Figure 3. Figure 3A depicts the deconvolution window that spans from 8.62 min to 8.91 min and contains both simple and composite peak features. Figure 3B depicts the first clustering of retention time at peak apex. The clustering produced three clusters, and the singleton cluster consisting of mass 89 only was removed. Figure 3C–D depicts the second clustering that was applied to the elution profile of simple peak features in each remaining cluster that satisfied the signal-to-noise ratio requirement. With the dissimilarity value set at 15, this clustering produced one sub-cluster for each cluster obtained in the first clustering. This result indicated that there are two components within this deconvolution window. Figure 3E–F depicts the simple peak features in each cluster. The peak features of mass 158 and mass 142 have the highest sharpness values in their individual clusters and were selected as model peaks. Figure 3G–H depicts the mass spectra that were generated after decomposing each of the simple and composite peak features into a linear summation of the two model peak features. The two mass spectra were identified as iso-leucine and proline with matching scores being 846 and 988, respectively.

Results

Improvements of QUAL/QUAN analysis

In order to test the performance of compound identification and quantitation in ADAP-GC 3.0, we used it to analyze the data files of 27 standard compounds mixture and compared the results with those produced by ADAP-GC 2.0. Table 1 lists all of these 27 standard

compounds identified from Sample I and II using ADAP-GC 3.0 and the corresponding matching scores and R^2 values. Supplemental files compare ADAP3 against ADAP2_identification.csv, compare ADAP3 against ADAP2_quantitation.xlsx, and compare ADAP3 against ADAP2_spectra.pdf lists detailed identification and quantitation results obtained using ADAP-GC 3.0 and 2.0 for Sample I. For a total of six times, ADAP-GC 3.0 was able to identify three compounds that were in low concentrations and that ADAP-GC 2.0 failed to identify from Sample I. These compounds were histidine in data files S0.2 and S0.4, isocitric acid in data files S0.2, S0.4, and S0.8, and alloisoleucine in data file S0.2. Here S0.2 refers to the data file that corresponds to the concentration of compounds being 0.2 $\mu\text{g/mL}$ in Sample I and file naming is similar for other data files. In addition to the improvement in compound identification, the quantitation performance improved slightly as well. The average R^2 value calculated based on the deconvolution results obtained by ADAP-GC 3.0 is 0.997 whereas the R^2 value by ADAP-GC 2.0 is 0.992.

Figure 4 depicts the reason why ADAP-GC 3.0 was able to identify isocitric acid in data files S0.2 and S0.4 and histidine in data file S0.4. Due to the low concentration of both compounds, the quality of the chromatographic peaks was low and neither ADAP-GC 2.0 nor 3.0 was able to find model peaks for these compounds. However, ADAP-GC 3.0 benefited from the first round of clustering of the retention times of chromatographic peaks whereby two components have been found within the deconvolution window. Spectra were constructed directly from the chromatograms and were eventually identified as isocitric acid and histidine. The matching scores for isocitric acid were 851, 865, and 908 in S0.2, S0.4, and S0.8, respectively. The matching scores for histidine were 847 and 863 in S0.2 and S0.4, respectively. Scores for spectral similarity are calculated using the method described in the supplementary file scoring.pdf.

It is worth noting that citric acid and isocitric acid are isomers and are always found together. The difference between these two compounds is merely that the hydroxy group ($-\text{OH}$) is bound to a different carbon atom of each molecule. Large-scale separation of the two isomers has not been possible. The Figure in supplementary file reference spectra_citric acid vs isocitric acid.pdf depicts the reference spectra of these two compounds and it is clear that they share many mass peaks. The fact that ADAP-GC 3.0 was able to identify both of them in low concentration with high matching scores demonstrates that the strategy for deconvolution used by ADAP-GC 3.0 is successful.

Comparison with other software tools

We compared ADAP-GC 3.0 with three software tools that are equipped with the capability to do spectral deconvolution. These tools are AMDIS (Version 2.71), AnalyzerPro (Trial version 3.0.0.0), and ChromaTOF (version 4.34). Raw data in NetCDF format were analyzed independently by each software tool. The parameters used by all of the four software tools were adjusted appropriately to make their performance comparable (see supplementary file analysis_parameters.pdf).

Table 2 and Table 3 summarize the identification and quantitation results of 27 standard compounds in Sample I and II. Overall, ADAP-GC 3.0 and ChromaTOF produced the best and comparable results in terms of the number of compounds identified and their matching scores. ADAP-GC 3.0, AMDIS, AnalyzerPro, and ChromaTOF identified 25, 21, 20 and 24 standards, respectively, from the seven datasets of Sample I, and 27, 15, 25 and 27, respectively, from the eight datasets of Sample II. Their average matching scores of the 27 standard compounds are 919, 899, 873, and 924 in Sample I and 903, 906, 875, and 913 in Sample II, respectively.

Among the four pairs of co-eluting compounds (alloisoleucine and proline, cysteine and creatinine, asparagine and xylose, citric acid and isocitric acid) in Sample II, cysteine and creatinine co-elute at about 13.57 min with their peak apex only one to two scans apart (Figure 5A). In addition, they share most of their peak features. As a result, it is difficult to completely resolve them, which would affect the identification and quantitation results. In the urine samples, the peak apex of the two compounds were 12 to 30 scans apart, which made it easier to separate them during deconvolution. All of the four tools were able to identify cysteine in most of the data files.

In addition to cysteine and creatinine, histidine is another example of challenge for deconvolution. ADAP-GC 3.0 was able to identify it in all of the seven data files in Sample I, whereas AMDIS and AnalyzerPro were able to identify it in only one of the data files. ChromaTOF identified histidine in six of the files and failed in the data file with the lowest concentration (0.2 $\mu\text{g/mL}$). In this data file, peak feature of m/z 154 was found to be the only significant peak that was unique to histidine. However, its low signal-to-noise ratio and very low abundance compared to the co-eluting lysine at 21.95 min made it nearly impossible to be detected automatically (Figure 5B). ADAP-GC 3.0 was able to identify it, again because of the first round of clustering of the apex retention time.

All four software tools produced good quantitation results in Sample I with average R^2 values greater than 0.99. However, quantitation of standards in Sample II (standards spiked into urine samples) is more complex because there exist hundreds of metabolites with diverse biochemical properties and a wide range of concentrations. As a result, a total of 17, 10, 17, and 17 compounds out of 27 have R^2 values greater than 0.99 in Sample II for ADAP-GC 3.0, AMDIS, AnalyzerPro, and ChromaTOF, respectively. The lower R^2 values of several compounds indicate different degrees of impurity or inaccuracy of resolved mass spectra due to noise or co-eluting compounds. Among them, three standard compounds (i.e., creatinine, citric acid, and mannitol) have poor quantitation performance because they themselves exist in the urine samples and their high concentrations were above the dynamic range of the TOF-MS analyzer.

Conclusion

ADAP-GC 3.0 was developed to improve both the robustness and sensitivity of detecting compounds in untargeted GC/MS-TOF metabolomics data, in comparison to version 2.0. ADAP-GC 3.0 combines continuous wavelet transform and local maxima for detecting chromatographic peak features and uses a simple yet effective approach to selecting model

peaks. As a result, the total number of parameters that have to be pre-specified is reduced and compounds in low concentration could be detected. ADAP-GC 3.0 has been tested on samples consisting of only a mixture of standard compounds as well as on urine samples that are more complex than the mixture of standards.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Jiang W, Qiu Y, Ni Y, Su M, Jia W, Du X. Journal of proteome research. 2010; 9:5974–81. [PubMed: 20825247]
2. Ni Y, Qiu Y, Jiang W, Suttlemyre K, Su M, Zhang W, Jia W, Du X. Analytical chemistry. 2012; 84:6619–29. [PubMed: 22747237]
3. Tautenhahn R, Böttcher C, Neumann S. BMC Bioinformatics. 2008; 9:504. [PubMed: 19040729]
4. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. Analytical chemistry. 2012; 84:5035–9. [PubMed: 22533540]
5. Pluskal T, Castillo S, Villar-Briones A, Orešić M. BMC Bioinformatics. 2010; 11:395. [PubMed: 20650010]
6. Katajamaa M, Miettinen J, Orešić M. Bioinformatics. 2006; 22:634–6. [PubMed: 16403790]

Supporting Information Available

List of supplemental files:

- `scoring.pdf`
- `compare ADAP3 against ADAP2_identification.csv`
- `compare ADAP3 against ADAP2_quantitation.xlsx`
- `compare ADAP3 against ADAP2_spectra.pdf`
- `reference spectra_citric acid vs isocitric acid.pdf`
- `analysis parameters.pdf`

Figure 1.
Comparison of data processing workflows in ADAP-GC 1.0, 2.0, and 3.0.

Figure 2.

(A) An example of a composite peak feature successfully detected by the wavelet-transform method. Red circles indicate peak apexes and green circles indicate peak boundaries. (B) An example of a peak feature that was initially identified as a simple peak feature by the wavelet-transform method. It was then corrected as being a composite peak feature after the local maximum method was applied to examine the surrounding area. The red circle indicates the peak apex that the wavelet transform method originally detected. The blue circles indicate peak apex that the local maximum method found. The presence of the peaks denoted by blue circles indicates that the peak denoted by the red circle is a composite peak feature, not a simple one.

Figure 3.

Key steps of deconvolution in ADAP-GC 3.0 illustrated using a pair of co-eluting compounds from the sample of standards mixture (Sample I).

Figure 4.

(A1–A3). Identification of citric acid and iso-citric acid by ADAP-GC 3.0 at the lowest concentration 0.2 $\mu\text{g/mL}$ among the seven data files in Sample I. Citric acid and iso-citric acid share many peak features (A1) and only the simple peak feature of m/z 245 was qualified as the model peak for iso-citric acid (A3). The mass spectrum for citric acid was directly extracted at 19.81 min since the first round of clustering indicated that there was a compound around 19.81 min (A2). (B1–B3) Identification of citric acid and iso-citric acid by ADAP-GC 3.0 at concentration 0.4 $\mu\text{g/mL}$. Many peaks features were shared (B1). The first round of clustering indicated that there were at least two compounds within the deconvolution window (B2). Model peaks were found for both citric acid and iso-citric acid with m/z 183 for citric acid and m/z 245 for iso-citric acid (B3). (C1–C3). Identification of histidine by ADAP-GC 3.0 at concentration 0.4 $\mu\text{g/mL}$. Peak features of histidine were at much lower intensity values than the co-eluting lysine (C1) and only the simple peak feature for m/z 174 was selected as the model peak for lysine (C3). The mass spectrum for histidine

was directly extracted at 21.91 min since the first round of clustering indicated that there was a compound at around 21.91 min.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 5.

(A) Highly similar peak features of cysteine and creatinine in data file S0.6 of Sample I with concentration 0.6 $\mu\text{g/mL}$. It was very difficult to deconvolute them. (B) Peak features of histidine and lysine in the data file with the lowest concentration 0.2 $\mu\text{g/mL}$. Features of histidine were at much lower intensity values than those of lysine. The peak of m/z 154 was the only significant peak unique to histidine. The inset showed that it was very noisy and could not serve as a model peak. ADAP-GC 3.0 did manage to identify it based on the first round of clustering.

Table 1
Identification and quantification results of the 27 standard compounds from analyzing samples I and II using ADAP-GC 3.0

No.	Compound Name	Sample I (7 samples)					Sample II (8 samples)				
		RT (min)	Mass ^a	R ²	Score ^b	Count ^c	RT (min)	Mass	R ²	Score	Count
1	Pyruvic acid	5.17	174	0.999	933	7	5.17	174	0.977	939	8
2	Propanoic acid	5.34	117	0.999	981	7	5.34	117	0.996	976	8
3	β-Amino isobutyric acid	7.47	102	0.999	938	7	7.47	102	0.885	897	8
4	L-leucine	8.4	158	0.999	915	7	8.4	158	0.998	852	8
5	isoleucine	8.73	158	0.998	856	7	8.74	158	0.998	847	8
6	Proline	8.78	142	0.999	982	7	8.78	142	0.998	938	8
7	Glycic acid	9.34	73	0.998	974	7	9.34	189	0.996	968	8
8	Threonine	10.31	117	0.998	954	7	10.31	117	0.996	975	8
9	5-oxoproline	12.8	156	1	924	7	12.81	157	0.994	916	8
10	L-Cysteine [#]	13.57	73	1	842	2	13.54	307	0.373	715	8*
11	Creatinine [#]	13.57	73	1	867	5	13.59	115	0.347	968	8
12	Citrulline	14.85	73	0.997	947	7	14.85	142	0.994	925	8
13	D-Xylose	15.93	73	0.999	939	7	15.94	103	0.993	842	8
14	Asparagine [#]	16.15	116	0.998	756	7	16.16	116	0.992	785	8*
13(2)	D-Xylose [#]	16.16	103	0.994	959	7	16.17	103	0.998	965	8
15	1,4-Butanediamine	17.59	174	0.996	958	7	17.6	174	0.999	955	8
16	Glycerolphosphate	18.51	73	0.998	890	7	18.52	299	0.853	847	8
17	Chlorophenylalanine	18.95	218	NA ^d	954	7	18.96	218	NA	932	8
18	Citric acid [#]	19.81	183	0.997	933	7	19.85	273	0.891	946	8
19	Isocitric acid [#]	19.87	245	0.997	901	7*	19.89	245	0.978	834	8
20	L-Histidine [#]	21.93	154	0.995	893	7*	21.95	154	0.958	899	8
21	L-Lysine [#]	21.96	174	0.994	950	7	21.97	174	0.992	908	8
22	Mannitol	22.61	73	0.997	945	7	22.63	103	0.859	942	8
23	Galic acid	22.87	73	0.999	970	7	22.88	281	0.961	912	8
24	N-Acetyl glucosamine methoxime	25.97	202	0.998	888	7	25.96	129	0.996	848	8*
25	L-tryptophan	27.94	73	0.997	965	7	27.94	202	0.995	964	8

No.	Compound Name	Sample I (7 samples)					Sample II (8 samples)				
		RT (min)	Mass ^a	R ²	Score ^b	Count ^c	RT (min)	Mass	R ²	Score	Count
26	Adenosine	31.38	73	0.996	894	7	31.38	230	0.995	927	8
27	Guanosine	32.31	73	0.993	770	7	32.31	324	0.991	865	8
Average value				0.998	917				0.926	903	

* Specific improvements achieved by ADAP-GC 3.0 in terms of compound identification as compared to what is achieved by ADAP-GC 2.0.

^a quantitation mass.

^b Average matching score of the same compound identified in up to seven data files. Identification was accomplished by matching the spectra against a library of mass spectra for standard compounds.

^c The number of samples in which a compound is identified.

^d The R^2 value was not calculated because chlorophenylalanine is the internal standard.

[#] Four pairs of co-eluting compounds.

Table 2

Results of compound identification and quantitation from seven datasets of Sample I

No.	Compound Name	ADAP-GC 3.0						AMDIS						AnalyzerPro						ChromatOF					
		RT	Mass	N	Score	R ²	Mass	N	Score	R ²	Mass	N	Score	R ²	Mass	N	Score	R ²	Mass	N	Score	R ²			
1	Pyruvic acid	5.17	174	7	933	0.999	73	7	891	0.995	73	7	896	0.996	174	7	932	0.996							
2	Propanoic acid	5.34	117	7	981	0.999	73	7	971	0.998	73	7	962	0.999	117	7	978	0.999							
3	β-Amino isobutyric acid	7.47	102	7	938	0.999	102	7	943	0.999	102	7	934	0.999	102	7	944	0.999							
4	L-leucine	8.4	158	7	915	0.999	158	7	906	0.996	158	7	917	0.996	158	7	915	0.996							
5	isoleucine	8.73	158	7	856	0.998	158	7	843	0.996	158	6	839	0.996	158	7	851	0.994							
6	Proline	8.78	142	7	982	0.999	142	7	979	0.996	142	7	961	0.996	142	7	987	0.996							
7	GlycERIC acid	9.34	73	7	974	0.998	73	7	966	0.994	73	7	962	0.995	189	7	975	0.994							
8	Threonine	10.31	117	7	954	0.998	73	7	954	0.993	73	7	947	0.995	73	7	975	0.995							
9	5-oxoproline	12.8	156	7	924	1	156	7	924	0.993	156	7	899	0.996	156	7	928	0.996							
10	L-Cysteine	13.57	73	2	842	1	73	5	827	0.985	73	3	783	0.997	115	2	822	/							
11	Creatinine,	13.57	73	5	867	1	73	4	875	0.923	73	4	828	0.956	115	5	880	0.999							
12	Citrulline	14.84	73	7	947	0.997	73	7	895	0.994	73	4	892	0.996	142	7	945	0.991							
13	d-Xylose	15.93	73	7	939	0.999	73	6	894	0.995	73	4	872	0.999	103	7	936	0.996							
14	Asparagine	16.15	116	7	756	0.998	116	5	783	0.998	116	5	743	0.996	116	7	795	0.993							
13(2)	d-Xylose	16.16	103	7	959	0.994	103	7	927	0.981	73	5	833	0.997	103	7	968	0.989							
15	1,4-Butanediamine	17.59	174	7	958	0.996	174	7	954	0.985	174	7	929	0.992	174	7	960	0.992							
16	Glycerolphosphate	18.51	73	7	890	0.998	73	6	831	0.99	73	5	805	0.997	73	7	894	0.994							
17	Chlorophenylalanine *	18.95	218	7	954	/	73	7	932	/	73	6	900	/	218	7	954	/							
18	Citric acid	19.81	183	7	933	0.997	73	7	929	0.995	73	6	855	0.996	183	7	957	0.992							
19	Isocitric acid	19.87	245	7	901	0.997	73	7	867	0.994	73	4	819	1	245	7	909	0.992							
20	L-Histidine	21.92	154	7	893	0.995	73	1	735	/	154	1	802	/	154	6	882	0.993							
21	L-Lysine	21.96	174	7	950	0.994	73	7	943	0.985	73	7	887	0.988	174	7	950	0.989							
22	Mannitol	22.61	73	7	945	0.997	73	7	928	0.99	73	7	888	0.992	73	7	943	0.992							
23	Galic acid	22.87	73	7	970	0.999	73	7	948	0.996	73	7	890	0.994	281	7	984	0.993							
24	N-Acetyl glucosamine methoxime	25.97	202	7	888	0.998	73	5	863	0.996	73	6	788	0.992	73	7	915	0.996							
25	L-Tryptophan	27.94	73	7	965	0.997	73	7	959	0.99	73	6	954	0.995	202	7	966	0.991							
26	Adenosine	31.38	73	7	894	0.996	73	7	887	0.992	73	7	858	0.991	73	7	913	0.991							

No.	Compound Name	RT	ADAP-GC 3.0				AMDIS				AnalyzerPro				ChromaTOF																			
			Mass	N	Score	R ²	Mass	N	Score	R ²	Mass	N	Score	R ²	Mass	N	Score	R ²																
27	Guanosine	32.31	73	7	770	0.993	73	7	822	0.991	73	6	789	0.991	73	7	826	0.987																
Average			917				0.998				899				0.990				873				0.994				924				0.994			

Table 3

Results of compound identification and quantitation from eight datasets of Sample II

No.	Compound Name	ADAP-GC 3.0				AMDIS				AnalyzerPro				ChromaTOF			
		RT	Mass	Score	R ²	Mass	N	Score	R ²	Mass	N	Score	R ²	Mass	Score	R ²	
1	Pyruvic acid	5.17	174	939	0.977	73	8	932	0.973	73	6	910	0.977	174	941	0.977	
2	Propanoic acid	5.34	117	976	0.996	73	8	974	0.987	73	8	966	0.994	117	978	0.996	
3	β-Amino isobutyric acid	7.47	102	897	0.885	102	8	847	0.881	102	8	850	0.885	102	875	0.886	
4		L-leucine	8.4	158	852	0.998	158	8	904	0.998	158	8	910	0.998	158	897	0.998
5	isoleucine	8.74	158	847	0.998	73	8	836	0.994	158	8	831	0.998	158	831	0.998	
6	Proline	8.78	142	938	0.998	142	8	933	0.997	142	8	933	0.997	142	979	0.998	
7	Glycric acid	9.34	189	968	0.996	73	8	965	0.998	73	8	958	0.998	189	968	0.996	
8	Threonine	10.31	117	975	0.996	73	8	962	0.991	73	8	956	0.994	219	972	0.996	
9	5-oxoproline	12.81	157	916	0.994	73	8	938	0.983	156	8	913	0.995	156	948	0.994	
10	L-Cysteine	13.54	307	715	0.373	220	8	783	0.405	220	6	729	0.9	218	764	0.847	
11	Creatinine,	13.59	115	968	0.347	116	8	972	0.124	115	8	953	0.342	115	980	0.343	
12	Citrulline	14.85	142	925	0.994	142	8	890	0.991	73	6	882	0.99	142	919	0.994	
13	d-Xylose	15.94	103	842	0.993	73	8	929	0.848	73	6	849	0.952	103	921	0.993	
14	Asparagine	16.16	116	785	0.992	75	7	837	0.954	116	8	759	0.993	132	788	0.997	
13(2)	d-Xylose	16.17	103	965	0.998	73	8	966	0	73	8	946	0.998	307	963	0.997	
15	1,4-Butanediamine	17.6	174	955	0.999	73	8	923	0.817	174	8	916	0.999	174	956	0.999	
16	Glycerolphosphate	18.52	299	847	0.853	73	8	895	0.853	299	6	785	0.998	299	869	0.983	
17	Chlorophenylalanine	18.96	218	932	/	73	8	912	/	73	6	855		218	931	/	
18	Citric acid	19.85	273	946	0.891	73	8	978	0.149	73	8	940	0.84	273	970	0.702	
19	Isocitric acid	19.89	245	834	0.978	245	8	766	0.831	245	7	741	0.976	245	775	0.976	
20	L-Histidine	21.95	154	899	0.958	73	8	884	0.826	154	7	806	0.956	154	872	0.838	
21	L-Lysine	21.97	174	908	0.992	73	8	922	0.978	156	8	864	0.994	174	934	0.993	
22	Mannitol	22.63	103	942	0.859	73	8	949	0.152	73	8	917	0.967	319	945	0.937	
23	Galic acid	22.88	281	912	0.961	281	7	919	0.957	281	6	856	0.975	281	926	0.966	
24	N-Acetyl glucosamine methoxime	25.96	129	848	0.996	73	8	810	0.997	73	5	812	0.997	202	867	0.995	
25		L-Tryptophan	27.94	202	964	0.995	202	8	961	0.992	202	8	950	0.994	202	973	0.994
26	Adenosine	31.38	230	927	0.995	73	8	921	0.997	73	7	890	0.997	236	933	0.995	

No.	Compound Name	ADAP-GC 3.0				AMDIS				AnalyzerPro				ChromaTOF			
		RT	Mass	Score	R ²	Mass	N	Score	R ²	Mass	N	Score	R ²	Mass	Score	R ²	
27	Guanosine	32.31	324	865	0.991	73	8	847	0.998	73	7	820	0.995	324	877	0.99	
Average				903	0.926			906	0.803			875	0.952		913	0.94	