analytical chemistry

Article

pubs.acs.org/ac

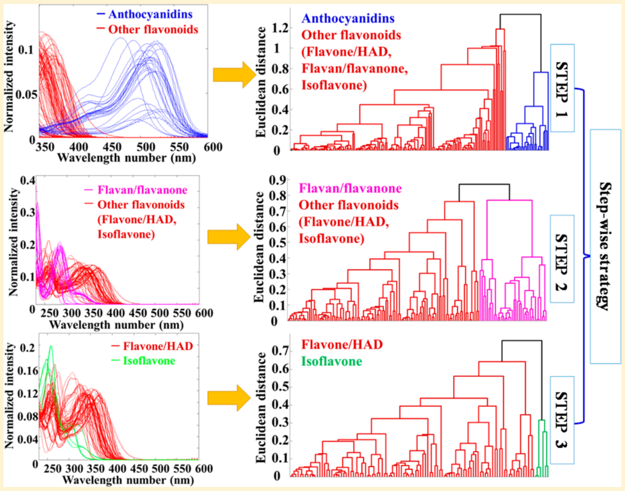
Development of a Comprehensive Flavonoid Analysis Computational Tool for Ultrahigh-Performance Liquid Chromatography-Diode Array Detection-High-Resolution Accurate Mass-Mass Spectrometry Data

Mengliang Zhang,[†] Jianghao Sun,[†] and Pei Chen[*]

Food Composition and Methods Development Lab, Beltsville Human Nutrition Research Center, Agricultural Research Service, United States Department of Agriculture, Beltsville, Maryland 20705-2350, United States

Ⓢ *Supporting Information*

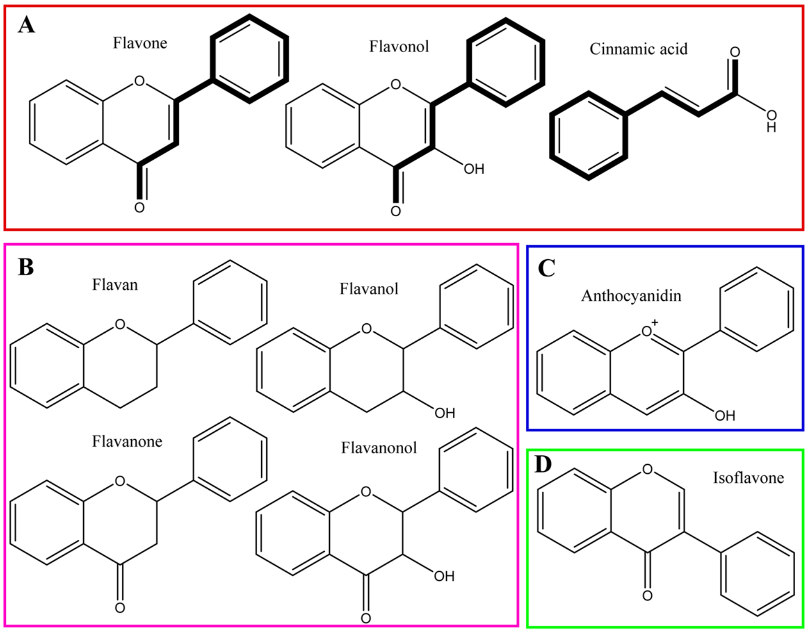
ABSTRACT: Liquid chromatography and mass spectrometry methods, especially ultrahigh-performance liquid chromatography coupled with diode array detection and high-resolution accurate-mass multistage mass spectrometry (UHPLC-DAD-HRAM/MS^n), have become the tool-of-the-trade for profiling flavonoids in foods. However, manually processing acquired UHPLC-DAD-HRAM/MS^n data for flavonoid analysis is very challenging and highly expertise-dependent due to the complexities of the chemical structures of the flavonoids and the food matrixes. A computational expert data analysis program, FlavonQ-2.0v, has been developed to facilitate this process. The program first uses UV−vis spectra for an initial stepwise classification of flavonoids into classes and then identifies individual flavonoids in each class based on their mass spectra. Step-wise identification of flavonoid classes is based on a UV−vis spectral library compiled from 146 flavonoid reference standards and a novel chemometric model that uses stepwise strategy and projected distance resolution (PDR) method. Further identification of the flavonoids in each class is based on an in-house database that contains 5686 flavonoids analyzed in-house or previously reported in the literature. Quantitation is based on the UV−vis spectra. The stepwise classification strategy to identify classes significantly improved the performance of the program and resulted in more accurate and reliable classification results. The program was validated by analyzing data from a variety of samples, including mixed flavonoid standards, blueberry, mizuna, purple mustard, red cabbage, and red mustard green. Accuracies of identification for all samples were above 88%. FlavonQ-2.0v greatly facilitates the identification and quantitation of flavonoids from UHPLC-HRAM-MS^n data. It saves time and resources and allows less experienced people to analyze the data.



Flavonoids are a group of phenolic compounds with various bioactivities and are widely distributed in plants. In various *in vitro* and *in vivo* models, they have exhibited diverse biological activities including anti-inflammatory, antiatherosclerotic, antitumor, antithrombogenic, antiosteoporotic, and antiviral effects.[1] Although dietary flavonoids may play an important role in human health, making recommendations on daily flavonoid intakes is very difficult. One of the important issues that limit progress in dietary flavonoid recommendations for consumers is the lack of appropriate analytical methods for the determination of flavonoids in foods and dietary intake levels.[2]

Profiling flavonoids in foods is challenging due to the fact that their structures are complex, their distribution and concentrations in plants vary greatly, and commercially available reference standards are limited.[3] Liquid chromatography/mass spectrometry (LC/MS) has become the most commonly employed method in flavonoid identification and quantification.[2,4] While technical advances such as ultrahigh-performance liquid chromatography-diode array detection-high-resolution accurate-mass multistage mass spectrometry (UHPLC-DAD-HRAM-MS^n) can provide much more detailed information for a sample, it also brings us a new challenge: the tremendous amounts of data to be analyzed. In recent years, the emergence of a few "omics" tools such as XCMS,[5] MZmine,[6,7] MetSign,[8] and MET-COFEA[9] have greatly facilitated data analysis using automated peak picking, peak alignment, peak integration, and database searching. However, they are designed for nontargeted metabolomics or metabolite profiling. They are inadequate for the analysis of a specific class of targeted plant secondary metabolites, such as flavonoids, due to the lack of specificity. Herein, FlavonQ-2.0v, a software program specifi-

Received: March 1, 2017
Accepted: June 25, 2017
Published: June 25, 2017



Group A: Flavone, flavonol, and hydroxycinnamic acid derivatives;

Group B: Flavan, flavanol, flavanone, and flavanonol;

Group C: Anthocyanidin;

Group D: Isoflavone.

Figure 1. Core structures of the main flavonoid classes and hydroxycinnamic acid derivatives.

cally designed for the analysis of flavonoids, has been developed.

FlavonQ-2.0v has made several important advances compared with its predecessor, FlavonQ. Like FlavonQ,[10] FlavonQ-2.0v features all the functions necessary to detect chromatographic peaks, integrate peak areas, interpret MS spectra, and produce qualitative and quantitative results. The important advance of FlavonQ-2.0v are (1) it is capable of analysis of all the major classes of flavonoids, including flavone/flavonol, flavan/flavanol, flavanone/flavanonol, isoflavone, anthocyanidins, and hydroxycinnamic acids (nonflavonoids) (Figure 1); (2) the program uses a chemometric pattern recognition method to classify the classes of the flavonoids by comparing the UV spectrum of a chromatographic peak to an UV−vis spectra library of 146 flavonoid and hydroxycinnamic acid standards; (3) the result obtained from the above-mentioned step is correlated with HRAM/MS^n spectra of that peak and searched against an in-house flavonoid database for tentative identification.

In this study, the stepwise approach of FlavonQ-v2.0 is explained and illustrated. The advantages of stepwise strategy with the projected difference resolution (PDR) method over conventional classification strategy is demonstrated. The program is validated with the analysis of samples spiked with flavonoids, mix standards, and plant extracts. The improved approach used in FlavonQ-2.0v is innovative, efficient, and highly effective.

■ MATERIALS AND METHODS

Chemicals and Plant Materials. Formic acid, HPLC grade methanol, and acetonitrile were purchased from Fisher Scientific. (Pittsburgh, PA). HPLC grade water was prepared from distilled water using a Milli-Q system (Millipore Laboratory, Bedford, MA). The reference standards for flavonoids and hydroxycinnamic acid derivatives were obtained from Sigma-Aldrich (St. Louis, MO), Chromadex, Inc. (Irvine, CA), Indofine Chemical Co. (Somerville, NJ), and Extrasynthese (Genay, Cedex, France). A list of 146 reference standards can be found in the Supporting Information.

Blueberry (*Vaccinium corymbosum* L.), mizuna (*Brassica juncea*), purple mustard (*Chorispora tenella*), red cabbage (*Brassica oleracea* L.), and red mustard green (*Brassica juncea*) were purchased from local grocery stores, and lyophilized immediately upon arrival and then ground and powdered.

UHPLC-DAD-MS Instrument. The UHPLC coupled with a diode array detector and LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific, San Jose, CA) was used. The chromatographic separation was achieved using a UHPLC column (200 mm × 2.1 mm i.d., 1.9 μm, Hypersil Gold AQ RP-C_{18}) (Thermo Fisher Scientific, Inc., Waltham, MA) with an HPLC/UHPLC precolumn filter (UltraShield Analytical Scientific Instruments, Richmond, CA) at a flow rate of 0.3 mL/min. UHPLC gradient and MS parameter settings were adapted from a previous study,[10] and the details can be found in the Supporting Information.

Sample Preparation. Each powdered sample (250 mg) was extracted with 5.00 mL of methanol/water (60:40, v/v) using sonication for 60 min at room temperature and the slurry mixture was centrifuged at 5000*g* for 15 min (IEC Clinical Centrifuge, Damon/IEC Division, Needham, MA). The supernatant was filtered through a 17 mm (0.45 μm) PVDF syringe filter (VWR Scientific, Seattle, WA), and 2 μL of the extract was used for each injection.

Data Format. MATLAB R2012b (MathWorks Inc., Natick, MA) was used to develop the program. All the calculations were performed on an Intel Core i7-4770 CPU at 3.4 GHz personal computer with 16 GB RAM running a Microsoft

Windows 7 Professional x64 operation system (Microsoft Corp., Redmond, WA).

The UHPLC-DAD HRAM MS data sets were acquired as RAW files. The DAD data were converted to text files from RAW files by Xcalibur plug-in tool, MSGet.[11] With an in-house algorithm, text files were read into MATLAB. For the MS data, the RAW files were first converted to mzXML by an open-source software package, ProteoWizard,[12] and then read into MATLAB by the built-in "mzxmlread" function in MATLAB bioinformatics toolbox.

■ RESULTS AND DISCUSSION

UV−vis Spectral Library of 146 Flavonoid Standards. First, 146 flavonoid and hydroxycinnamic acid derivative standards were analyzed using the UHPLC-DAD method and their UV−vis spectra were compiled into a UV−vis spectral library after they were normalized to unit vector length.[13] The 146 UV−vis spectra are shown in Figure 2A. As discussed in the previous paper,[10] flavonoid identification cannot be solely relied upon MS spectra and often requires the combination of multiple techniques such as chromatographic behavior, UV−vis spectrum, and HRAM-MS, and MS fragmentation information. Flavonoids have characteristic UV−vis absorbance profiles which come from different conjugated systems in the structures and can be used to distinguish isomers. For example, pelargonidin 3-*O*-glucoside (an anthocyanin), genistein 4′-*O*-glucoside (an isoflavone), and apigenin 7-*O*-glucoside (a flavonol glycoside) have exactly the same protonated or deprotonated ions in full scan MS spectra, and their fragmentation mass spectra are dominated by one or only a few fragments simply do not contain enough information to distinguish between them. The representative MS/MS spectra for the three flavonoids mentioned above are shown in Figure S1. However, they can be differentiated by their UV−vis spectra since the cinnamoyl structure in flavone, flavonol, and hydroxycinnamic acid derivatives have a strong UV absorbance band between 305 and 390 nm; however, anthocyanins are cations with a strong visible absorbance band at 450−550 nm (Figure 2A).[14,15]

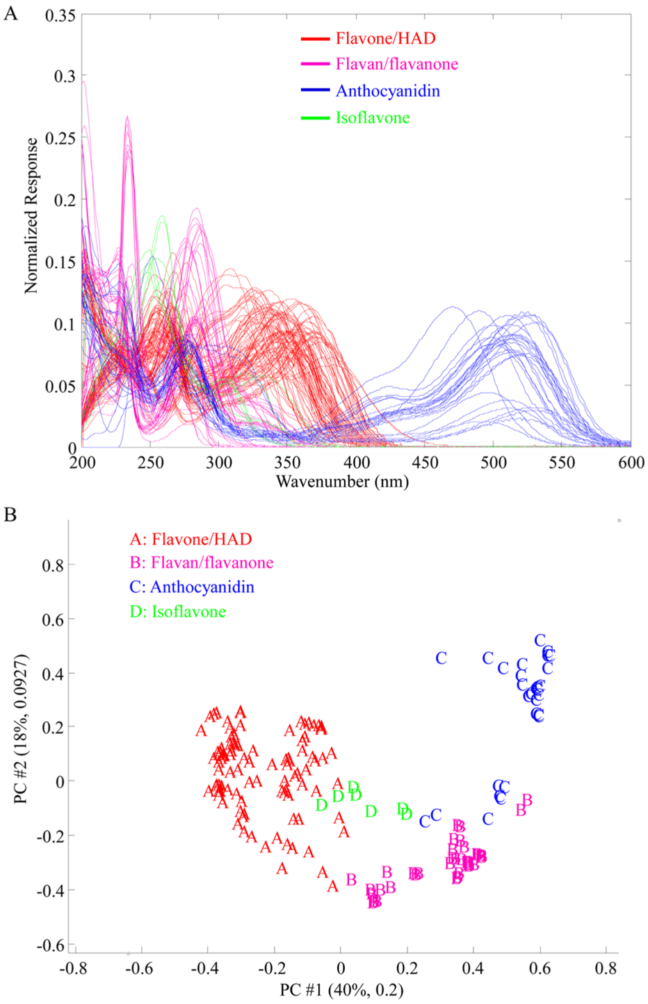


Figure 2. A total of 146 UV−vis spectra of flavonoids and HAD (A) and principal component analysis score plot for UV−vis spectra data of four classes (B).

The assignment of classes for flavonoids based on their UV−vis spectrum is a crucial step, especially for the identification of flavonoid isomers which belong to different flavonoid classes. In our previous study, UV−vis spectrum similarity analysis was used to assign the class of flavonoids for each chromatographic peak.[10] The UV−vis spectrum of a reference peak, either a spiked standard or an endogenous flavonoid peak, was selected and compared with that of all other chromatographic peaks. A threshold was set based on a trial-and-error procedure to filter out nondesired peaks. Particular care is needed to be taken for that method: (1) the reference peak had to be representative of the class of flavonoid as selection of a reference peak sometimes can be difficult especially for the classes of flavonoids which contain a great variety of substitution groups; (2) plant samples usually contain different classes of flavonoids; therefore, multiple reference peaks need to be selected to represent the different classes of flavonoids and multiple calculations are required since only one class of flavonoids could be classified by each calculation; and (3) the threshold for UV−vis spectral similarity analysis varies case by case. For example, the thresholds ranged from 50% to 90% for leek, curry leaf, chive, giant green onion, and red mustard green samples.[10] Thus, although the similarity analysis of FlavonQ worked well for the class of flavonols and their glycosides,[10] it is inconvenient to use the approach for identification of multiple classes.

Grouping 146 Reference Standards into Four Classes. A new strategy was developed in FlavonQ-2.0v to improve the similarity approach by using chemometric modeling and a UV−vis spectral library. The UV−vis spectral library was compiled from the UV−vis data of 134 flavonoids and 12 hydroxycinnamic acid derivatives (HADs) standards. Although HADs do not belong to flavonoid family, they were also included because their structures are similar to flavonoids and they are ubiquitous in plant with various bioactivities.[16] The standards were divided into four classes on the basis of the structural similarities of the aglycones (Figure 1): flavone/flavonol/HAD (class A), flavan/flavanol/flavanone/flavanonol (class B), anthocyanin (class C), and isoflavone (class D). Chemometric methods were employed to construct models for classification of different flavonoid classes based on the UV−vis spectral library, and the classifiers were used to predict the class of the flavonoid in unknown chromatographic peaks.

Options for Chemometric Models in FlavonQ-2.0v. Two methods, including soft independent modeling of class analogy (SIMCA)[17] and fuzzy optimal associative memory (FOAM),[18] were evaluated. Classification methods such as partial least-squares discriminant analysis (PLS-DA)[19] and the fuzzy rule-building expert system (FuRES)[20] were not used because they cannot be applied when only one class is known or present.[21] FlavonQ-2.0v was designed to be a versatile program which can classify not only single-class flavonoid but also multiclass flavonoids. Therefore, the classifiers like PLS-DA was not adopted in this study. In this program, it is the user's decision as to which group(s) of flavonoids will be used to build classification models. There are several advantages to making the flavonoid type selection adjustable. Flavonoids are usually synthesized through the phenylpropanoid metabolic pathway and several enzymes are involved in the biosynthesis. It is rare that a single plant sample contains all the enzymes for synthesis of all the classes of flavonoids. Limiting the flavonoid types in the sample can reduce the complexity of chemometric models and improve the model accuracy and reliability. For example, purple broccoli only contains flavonols and anthocyanins.[22] So if the chemometric model is built using only these two classes for the analysis of flavonoids in broccoli, it will simplify the data analysis and reduce the possibility of misclassifying them into other classes of flavonoids. Moreover, in some cases, only one class of flavonoids is the research focus (e.g., the isoflavones in soybean samples) and a chemometric model targeting the class of interest can be very efficient.

SIMCA and FOAM are commonly used as modeling methods, but they can be used in classification mode as well. Modeling methods exploit the similarities of the features within each independent class, therefore the test sample could belongs to none of the existing classes in the training sets. However, in classification mode, the test sample must be assigned to one of the classes in the training sets. When classifying an unknown UV–vis spectrum by the constructed SIMCA/FOAM models with more than one flavonoid class, three situations could be encountered: (a) it only belongs to one class; (b) it belongs to none of any classes; (c) it belongs to more than one class. The UV–vis spectra from real sample could be different from the UV–vis spectra in the library attributed to influence of environment (e.g., temperature, solvent) and possible coeluted compounds. In addition, the accuracy of classification also highly relied on the quality of the training set: the number and representativeness of flavonoids in the library (The in-house UV–vis library may not be able to represent all flavonoids in the tested plant materials). If modeling mode was used, some flavonoids that were not included in the library or their UV–vis spectra were distorted by other background influences could be misclassified as nonflavonoids which resulted in false negatives. Therefore, the winner-takes-all mode (classification mode) was used in both SIMCA and FOAM models to avoid situation (b). Statistic values (i.e., the combination of X-residuals and Hotelling's T^2 value for SIMCA and F-value for FOAM model) were calculated between the variance of an unknown UV–vis spectrum in chromatographic peak and each flavonoid class, and the unknown UV–vis spectrum was assigned to the best fit class of flavonoids (in another word, the most "similar" class of flavonoids with smaller X-residuals and Hotelling's T^2 value for SIMCA model or F-value for FOAM model). Although the winner-takes-all mode may result false positives, the result will be refined by using MS spectra. Similarly, for situation c, only the most "similar" class instead of multiple classes was assigned to an unknown UV–vis spectrum.

When only one class of flavonoids was selected to construct chemometric models, the statistic criteria (X-residuals and Hotelling's T^2 with 95% confidence intervals for SIMCA and $F^{0.05}$ for FOAM model) was used to define the limit of the class and reject nonflavonoids.

Projected Difference Resolution Method to Optimize Wavelength Range of UV–Visible Data. UV–vis spectra contain characteristic regions and noninformative regions. Chemometric models built directly using the UV–vis spectral data over the full scan range (200–600 nm) were not effective as shown in Figure 2B. Overlaps of the four classes were observed. It can be advantageous to identify and remove the noninformative regions because it improves the predictive ability and reduces complexity for chemometric models.[23] For example, dropping off the wavelength range between 200 and 220 nm in the UV region is a common practice to avoid the interferences caused by mobile phase and retain the most obvious features for flavonoids between 220 and 600 nm.[24]

Selection of the wavelength range used in chemometric models can affect the classification and is a challenging task because the spectra may have imperceptible distinctive features. Therefore, the wavelength range of UV–vis spectrum needs to be optimized in this study. One straightforward way to achieve this is to build chemometric models for different wavelength ranges, evaluate the models by cross-validation methods such as leave-one-sample-out method[25,26] and bootstrapped Latin partition method,[21] calculate the classification rates for the different wavelength ranges, and select the optimum range which gives the best classification rate. However, this calculation required hours to execute depending on which chemometric models and validation method were chosen and is not practical to use in the data processing program.

In this study, the projected difference resolution (PDR) method[27] was applied as an alternative to determine the optimum wavelength range. The PDR method measures the separation of two classes in multivariate data space and has been used successfully for selecting the optimal parameters for baseline correction, wavelet filters, and data transformation.[13,27,28] The larger the PDR values, the better the separation between two classes in the multivariate data space. For the assessment of multiple classes, the minimum PDR value of all the pairwise combinations was used to optimize the wavelength range.[13] The two most similar classes among multiple classes were considered as the most critical pairs for classification, so their PDR values were calculated under different wavelength ranges. For example, when we have four classes, for a specific wavelength range their PDR values in pairs (6 pairs) were measured, and the minimum PDR value of 6 pairs was used to indicate the separation of the two most similar classes among the four classes. Since the UV range is easily influenced by conjugated bonds and the higher range of UV–vis spectra (wavelength ≥250 nm) usually represents characteristic information for the structure of each flavonoid class, only the starting wavelengths (WLs) of UV–vis spectra was optimized in our study. Therefore, a series of test UV–vis spectral data sets were constructed with different starting WLs: Test-set-1 (200–600 nm), Test-Set-2 (201–600 nm), Test-Set-3 (202–600 nm)...Test-Set-301 (500–600 nm). For each wavelength range, the PDR values were calculated for the different classes of flavonoids. The wavelength range with the maximum PDR value represented the optimum wavelength range for the classification of the flavonoid classes. Compared to the optimization of wavelength range with chemometric models which required hours to execute, the PDR method only took seconds which saved considerable time.

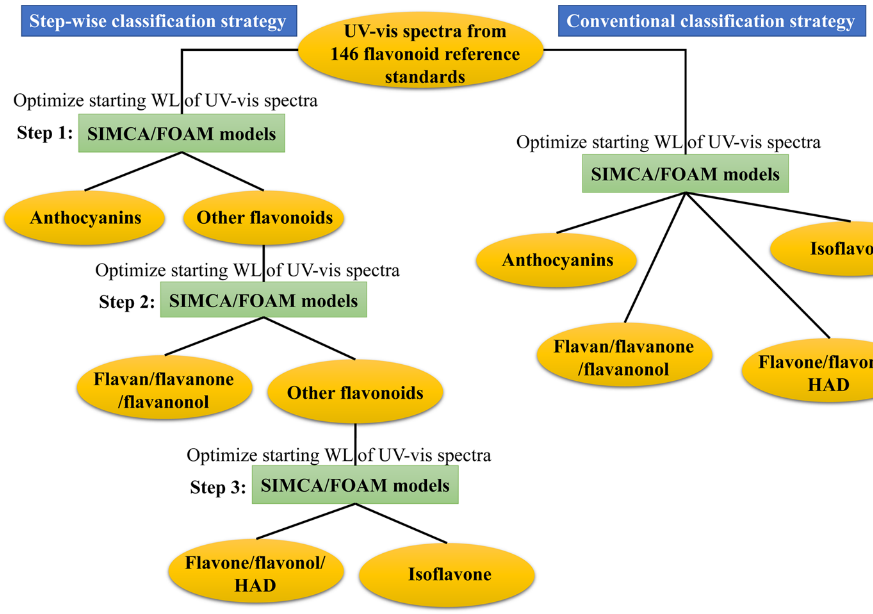

Step-wise classification strategy
UV-vis spectra from 146 flavonoid reference standards
Conventional classification strategy
Optimize starting WL of UV-vis spectra
Step 1: SIMCA/FOAM models
Anthocyanins
Other flavonoids
Optimize starting WL of UV-vis spectra
Step 2: SIMCA/FOAM models
Flavan/flavanone /flavanonol
Other flavonoids
Optimize starting WL of UV-vis spectra
Step 3: SIMCA/FOAM models
Flavone/flavonol/ HAD
Isoflavone
Optimize starting WL of UV-vis spectra
SIMCA/FOAM models
Anthocyanins
Isoflavone
Flavan/flavanone /flavanonol
Flavone/flavonol/ HAD

Figure 3. Flowchart for stepwise classification strategy and conventional classification strategy.

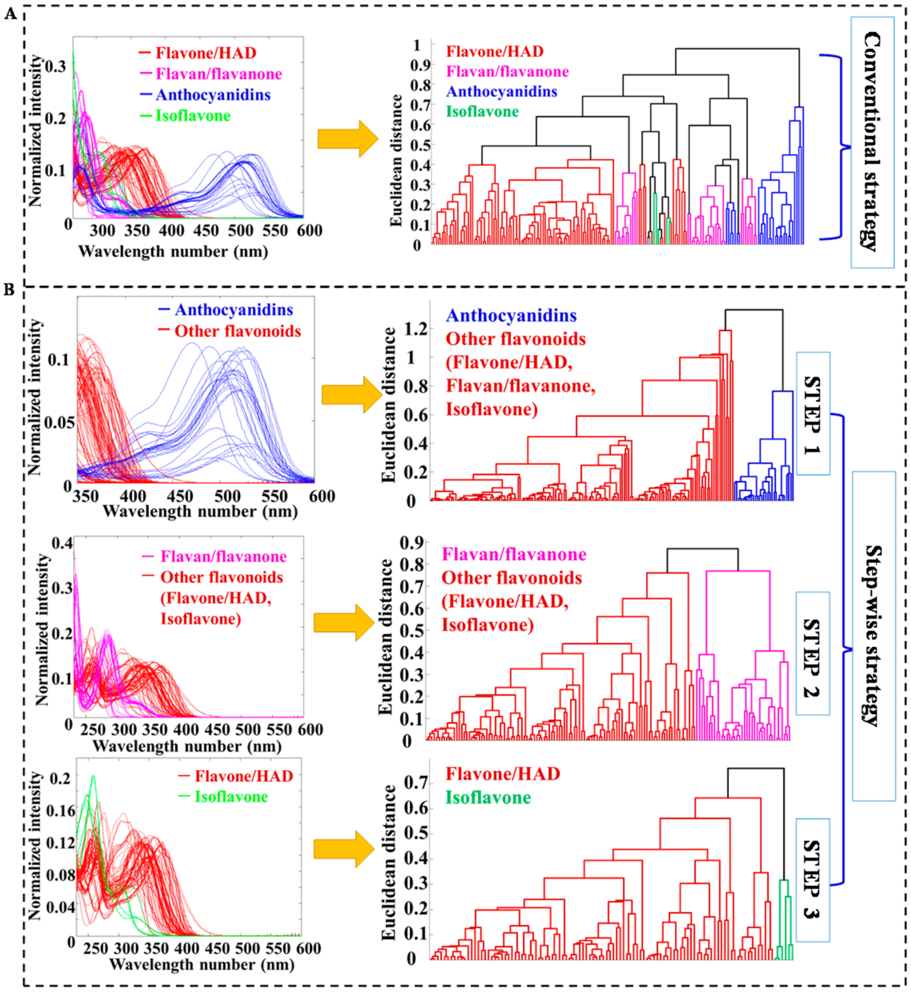


Figure 4. UV−vis spectra after wavelength range optimization (left) and dendrogramatic representations (right) of differentiation for four classes of flavonoids by conventional classification strategy (A) and stepwise classification strategy (B).

Classification of Flavonoids by Step-Wise Strategy. Step-wise classification was devised in this study to classify the UV−vis spectra of flavonoids in a novel way and the starting WLs of each step was optimized, respectively. A flowchart for

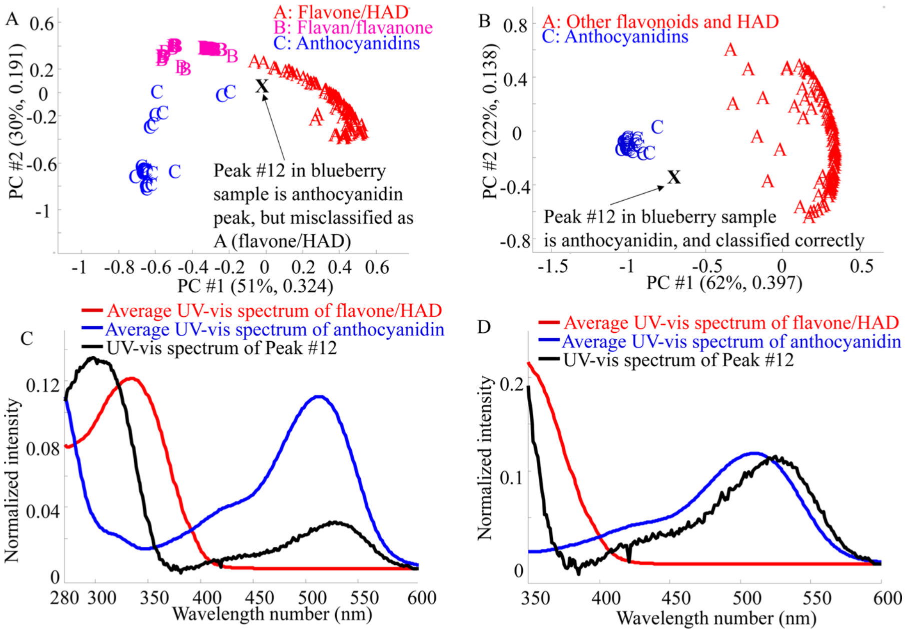


Figure 5. Principal component analysis score plot for UV−vis spectra data of three classes by conventional classification strategy (A) and by stepwise classification strategy-step 1 (B). Average UV−vis spectra of flavone/HAD and anthocyanidin and UV−vis spectrum of peak no. 12 in blueberry sample after starting WL optimization in conventional classification strategy (C) and in stepwise classification strategy-step 1 (D).

the two strategies of classification of four classes of flavonoids and the HADs is shown in Figure 3. Conventional classification strategy optimized universal parameters in data preprocessing and constructed one chemometric model by using all the data for the different classes. Step-wise classification strategy optimized data representation for each pair of classes and constructed multiple chemometric models. In each step, only two classes were defined and one group of flavonoids (class 1) was differentiated from other flavonoids (class 2). It is worth noting that either SIMCA or FOAM model can be selected for the classification, and the same model is used throughout the steps in stepwise classification.

Figure 4 shows the dendrograms based on Euclidean distances between spectra for two strategies outlined in Figure 3. Figure 4A shows that for a conventional classification strategy, even after the starting WL was optimized, the four classes were mixed with each other and none of classes was completely separated from others. However, the three dendrograms for a stepwise classification strategy (Figure 4B) demonstrated classification of each group of flavonoids into well-defined clusters. The benefit of stepwise classification strategy was proven by classification rates of SIMCA/FOAM models through leave-one sample out cross validation. With the conventional strategy, the best classification rates for the SIMCA and FOAM models were 99.3% and 95.6%, respectively. The classification rates were 100% for both the SIMCA and FOAM models using the stepwise classification strategy.

The order of flavonoid classes in stepwise classification process has great impact on the classification. For the four classes of flavonoids in Figure 2, 12 sequences were evaluated and their PDR values in each step were calculated based on the method in “Projected Difference Resolution Method to Optimize Wavelength Range of UV−vis Data” section. The results shown in Table S1 demonstrate that the flavonoid classification sequence used in Figures 3 and 4 is the optimal order in stepwise classification process: a relatively larger PDR value was achieved for the two most similar classes which indicates the better separation of the two classes in multivariate data space. Therefore, FlavonQ-2.0v separates anthocyanidins from the rest classes in the first step, then flavan/flavanone, and finally flavone/HAD and isoflavone in the stepwise classification process.

The application of the stepwise strategy eliminates some misclassifications of flavonoids in real samples. For example, peak no. 12 in blueberry sample (Figure S2) was manually identified as petunidin-3-*O*-arabinoside by the study of its mass spectra in both the positive and negative ionization modes.[29] When conventional classification was used, it was misclassified as flavone/HAD group (Figure 5A). The absorption band at 525 nm indicates that it is an anthocyanin instead of a flavone (Figure 5C). Peak no. 12 was successfully classified as anthocyanin when the stepwise classification strategy was applied (Figure 5B). Higher weight was given to the characteristic UV−vis band (525 nm) for anthocyanin by this strategy (Figure 5D). The stepwise strategy was more effective for classifying flavonoids based on their UV−vis spectra and, therefore, was adopted in this program. It is worth noting that isoflavones were not included in the chemometric model to study the flavonoids in this example because isoflavones are usually not found in blueberry.

Identification of Flavonoids Using In-House Database. After the chromatographic peaks were categorized into different classes of flavonoids, HRAM/MS^n data were used for putative identification of flavonoids and HAD. An in-house database was established in our lab which contained 5686 flavonoids and related compounds categorized into the four classes. The information for each compound such as chemical name, formula, accurate monoisotopic mass, protonated mass, deprotonated mass, and major product ions were included. Major product ions assigned to 4283 compounds were obtained by the observation of fragmentation mass spectra of flavonoids in our lab, METLIN mass spectrum database,[30] and the mass spectral library from Sumner’s group[31] or by predictions based on experience of experts and *in silico* fragmentation patterns using commercial software package HighChem Mass Frontier (Thermo Fisher Scientific Inc., San Jose, CA).

A selected number (user defined) of the most intense ions from the MS full scan spectrum of unknown chromatographic peaks were screened and matched with ions in the positive or negative mode from the in-house database. If the MS^n spectra for the ions in full scan spectrum were available in the data, the major product ions were searched through the MS^n spectra for matches. Multiple hits could be found after this searching process and all these candidate compounds were ranked in the result table based on the following priorities: candidate compounds with both precursor ions and product ions matched were ranked higher than others which were then ranked by mass errors in ascending order.

The program may provide multiple flavonoid candidates for a chromatographic peak and expertise in the field of flavonoid research is needed for affirmative identification (see examples in Tables S2 and S3). In the previous version of FlavonQ, the identification of flavonoids was based on a virtual mass spectrum database which was constructed by theoretically combining common aglycones and substitution groups.[10] For a single class of flavone/flavonol glycosides, it contained over 1.5 million possible combinations which, in most cases, have never occurred in the real world. In this study, all 5686 flavonoids and related compounds in the in-house database have been reported before. With this database, the computation speed of the program was faster and the identification results were more accurate. Flavonoids not included in this database could not be identified. However, potential flavonoid peaks (based on UV data) are flagged and can be manually identified and added to the database if needed.

Table 1. Comparison of Search Results for 25 Flavonoid and Hydroxycinnamic Acid Derivatives UHPLC-DAD-MS Data

	METLIN[a]		MetFrag (KEGG)[b]		FlavonQ-2.0v[c]	
	simple search	fragment search	parent ion search	fragment search	MS search	UV−vis and MS match
top 1 ranks[d]	2	3	5	6		18
top 5 ranks	10	12	15	18	8	25
no. of NA[e]	0	13	2	2	0	0
no. of candidate compounds[f]	663	54	221	213	162	113
avg no. of candidate compounds	26.5	2.2	8.8	8.5	6.5	4.5

[a] "Simple Search" and "Fragment Search" are two functions for METLIN database: "Simple Search" matches up precursor ions (20 ppm tolerance); "Fragment Search" matches up both precursor ions and selected fragment ions (up to 5 fragment ions) (http://metlin.scripps.edu). [b] KEGG database was selected for MetFrag search. "Parent Ion Search" and "Fragment Search" are functions for MetFrag Web tool: "Parent Ion Search" matches up precursor ions (20 ppm tolerance); "Fragment Search" matches up both precursor ions and MS/MS spectra (https://msbi.ipb-halle.de/MetFragBeta/). [c] "MS Search" matches up precursor ions and characteristic product ions (up to 1 for each precursor ion) in the MS spectra with in-house database; "UV−vis and MS Match" uses chemometric methods to determine the type of flavonoids before "MS Search". [d] Number of correct first ranked candidates. [e] Number of "none of correct candidates available". [f] Number of total candidate compound for all queries.

Comparison with METLIN and MegFrag Using Flavonoid UHPLC-DAD-MS Data Set. Two sets of mix reference standards (25 flavonoids and hydroxycinnamic acids) were analyzed by UHPLC-DAD-MS method as described previously. Their precursor ions and MS/MS spectra were manually input into METLIN database (http://metlin.scripps.edu) and MegFrag Web tool (https://msbi.ipb-halle.de/MetFragBeta). For METLIN, the precursor ions were searched under "Simple Search" function with 20 ppm tolerance; fragment search were performed under "Fragment Search" function with "Precursor M/Z" selected and up to 3 fragment ions were input for each search. KEGG database was selected in MegFrag Web tool and 20 ppm tolerance was set for "Parent Ion" search and "Fragmentation Processing". Besides of the proposed FlavonQ-2.0v data process pipeline, the data set were also processed in FlavonQ-2.0v by only searching precursor ions and characteristic product ion in MS spectra through in-house database without flavonoid classification based on UV−vis spectra. The results, given in Table 1, show that FlavonQ-2.0v generally performed better, indicated by higher number of correct first ranked candidates, a lower number of "none of correct candidates available", and a lower number of output candidates for each search. Take puerarin (an isoflavone) as an example, by searching $[M - H]^-$ (*m/z* 415.1029), METLIN outputs a list of 39 candidate compounds (puerarin ranked 28th) and MetFrag outputs 7 candidate (puerarin ranked fifth). For the 39 candidates from METLIN, 11 of them are nonflavonoids, with 22 flavone glucosides, 1 anthocyanidin, and 5 isoflavones. If "Fragment search" is applied, puerarin was not found because it has not been analyzed in METLIN. The 7 candidates from MetFrag includes 1 flavone, 2 isoflavones, and 4 nonflavonoids, and the MS/MS spectrum improved the ranking of puerarin from the fifth to the third out of 7. For FlavonQ-2.0v, 19 candidates were found without the use of UV−vis data (12 flavones, 5 isoflavones, and 2 anthocyanidins); only 5 candidates were found if UV−vis spectra were used for flavonoid classification and they were all isoflavones. The results were not unexpected due to the specificity of the FlavonQ program.[32] For example, METLIN includes 961 829 molecules among which about 14 000 metabolites have been individually analyzed and another 200 000 has *in silico* MS/MS data by May 2017. For the "Fragment Search" in METLIN, about half of the queries (13 of 25) in Table 1 returned "0 candidate" due to the lack of MS/MS data in the database. From Table 1, it has been observed that flavonoid classification based on UV−vis spectra can effectively narrow down the list of candidate compounds because there are some limitations for compound identification solely relied on MS/MS spectral comparison: for example, sometimes mass spectra are dominated by one or only a few fragments (e.g., a glycoside group loss, Figure S1) that can be explained by several candidates. Further examples and limitations of MS spectral library search are discussed extensively by Stephen Stein.[33]

Expansion of UV−Visible Library and In-House Database. As discussed in the previous sections, the accuracy of flavonoid identification based on UV−vis spectra and MS spectra can be improved by expanding the UV−vis library and in-house database. In our lab, the number of flavonoid UV−vis spectra continues to increase from several resources: acquisition of more flavonoid reference standards, isolated chromatographic peaks from plant materials, and reported spectra from peer-reviewed journals. The isolated chromatographic peaks should be pure and be identified and confirmed by mass spectrometric (HRMS, MS^n) and/or NMR methods,[34] the

DOI: 10.1021/acs.analchem.7b00771

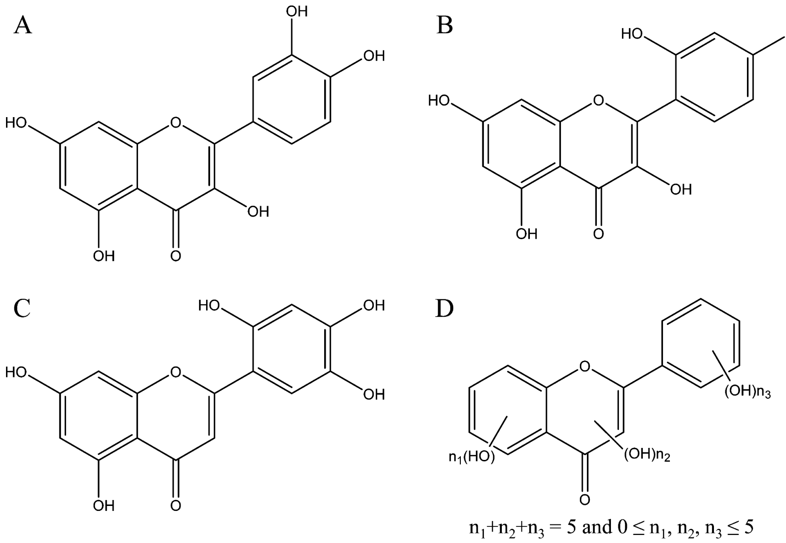


Figure 6. Chemical structures for quercetin (A), morin (B), hieracin (C), and pentahydroxyflavone (D).

reported UV−vis spectra should be validated by other independent laboratories. In this study, the UV−vis spectra in the library were exclusively collected from flavonoid reference standards, and more UV−vis data from other sources will be updated in the future release.

The product ions information in the in-house database can effectively target the correct hit of flavonoids especially when multiple isomers exist in the database for a particular precursor ion. Ideally the mass spectra library should contain MS^n spectra in both positive/negative modes and different collision energies. Such a library will be able to provide the most accurate identification of a compound such as Sumner's plant natural product MS library[31] and Compound Discoverer from Thermo Fisher Scientific. However, to construct such a library for over 5 000 flavonoids is not feasible for any single laboratory. Therefore, we are enhancing our in-house databases gradually by adding more characteristic product ions based on experiments and literatures. As the UV−vis library and in-house database expands, it will be effective automatically because FlavonQ-2.0v recalculates its chemometric models and searching results based on the updated library and database every time it executes.

Quantitation. The quantitation of flavonoids was performed using an external calibration curve with flavonoid reference standards and molar response factors as previously reported.[14,35] Ideally separate calibration curves should be used to quantify the flavonoids of each flavonoid class. For example, quercetin 3-*O*-rutinoside (rutin) for HADs and flavone/flavonol glycosides, catechin for flavan-3-ols and proanthocyanidins, hesperetin for flavanones, cyanidin 3-*O*-glucoside for anthocyanins, and genistein for isoflavones. The peak area integration method was demonstrated in the sample chromatogram (Figure S2) and different classes of flavonoids are represented by different colors. A brief identification, including major ion and formula, is provided for each flavonoid candidate chromatographic peak. The identification, peak areas, and tentative quantitation information were output automatically into spreadsheet which allowed the user to further analyze the results.

Performance of FlavonQ-2.0v. The performance of FlavonQ-2.0v was validated on samples spiked with flavonoid mixed standards and samples of plant extracts. FlavonQ-2.0v successfully identified all the flavonoid peaks in the flavonoid spiked mix standard samples. The results are shown in Tables S2 and S3. The results demonstrate the effectiveness of flavonoid identification by UV−vis and MS spectra. For example, apigenin was first classified by chemometric models in FlavonQ-2.0v as flavone/flavonol/HAD based on its UV−vis spectrum, so other isoflavone or anthocyanin isomers were excluded after this step. It was then identified as "Apigenin" in the flavonoid candidate list based on its precursor ion (*m*/*z* 269.0450 with error −1.59 ppm) and characteristic product ion (*m*/*z* 151, $^{1,3}A^-$) and it was distinguished from Baicalein which has characteristic product ion of *m*/*z* 169, $^{1,3}A^-$). In some cases, multiple candidate flavonoids were listed for a single peak since some flavonoid isomers with common product ions cannot be differentiated by the program. For example, quercetin, morin, and hieracin (Figure 6) are all flavonols and they have exact the same precursor ion (*m*/*z* 301.0348) and common characteristic product ion (*m*/*z* 151, $^{1,3}A^-$), so they were all reported in the result table.

FlavonQ-2.0v was also applied to the analysis of flavonoids in blueberry, mizuna, purple mustard, red cabbage, and red mustard green. The data were also analyzed manually. The FlavonQ-2.0v identification results were compared to those identified manually (Table S4). Among the 39 flavonoid candidate peaks, two anthocyanins, petunidin-3-*O*-arabinoside, and petunidin-3-*O*-glucoside, were misidentified by FlavonQ-2.0v as flavonol glucosides and flavanone glycoside, respectively. They were all small shoulder peaks (Peak no. 9 and no. 10 in Figure S2) and their UV−vis spectra were distorted by the close major peaks which led to the misclassification. This indicates that the chromatographic separation is critical for the correct identification of flavonoids. Another two peaks were labeled as "uncertain peaks" as the spectral data could not provide enough information for identification (peak no. 21 and no. 24 in Figure S2 and Table S4). The identification accuracy of flavonoids by FlavonQ-2.0v for plant materials is shown in Table 2. Overall, positive identifications were achieved for more than 88% of the flavonoid peaks using FlavonQ-2.0v.

The execution time of FlavonQ-2.0v was about 1 min for each sample after data format conversion. Construction of chemometric models using all 146 UV−vis spectra took about 30 s and the time was significantly reduced when fewer classes

DOI: 10.1021/acs.analchem.7b00771

Table 2. Flavonoid Identification Accuracy in Different Plants by FlavonQ-2.0v

plant name	no. of flavonoids identified[a]	no. of misidentification[b]	no. of uncertain peaks[c]	accuracy (%)
blueberry	39	2	2	89.7
mizuna	47	1	0	97.9
purple mustard	45	1	4	88.9
red cabbage	44	0	0	100.0
red mustard green	88	1	6	92.0

[a]Flavonoid peaks were identified by FlavonQ with s/n setting at 10. [b]Nonflavonoid peaks were identified as flavonoids. [c]Identity of peaks cannot be verified based on the data given.

of flavonoids were selected and fewer steps were conducted in the stepwise classification strategy. FlavonQ-2.0v was developed in MATLAB 2012b, but it is not necessary for the end user to install MATLAB to use FlavonQ-2.0v. MATLAB Compiler Runtime (MCR) is required to run FlavonQ-2.0v standalone application and is freely available at https://www.mathworks.com/products/compiler/mcr.html. The graphic user interface is shown in Figure S3. The UV−vis spectra of 146 flavonoid and HAD reference standards and in-house flavonoid database were compiled into FlavonQ-2.0v. This database will be continuously expanded and the chemometric models will become more reliable. Other common food constituents, such as simple phenolic compounds, phenyl alcohols, stilbenes, and lignans will also be included in the future. The in-house flavonoid database will be updated regularly as new flavonoids are found and reported.

CONCLUSIONS

A data processing tool for flavonoid analysis, FlavonQ-2.0v, was developed in this study. The program can classify the flavonoids using a chemometric model based on the UV−vis reference spectral library. The chemometric model used a novel stepwise classification strategy and data representation in each step was optimized by projected distance resolution (PDR) method. The stepwise classification strategy significantly improved the performance of the classifiers which resulted in more accurate and reliable classification of flavonoids. An in-house flavonoid database was implemented in the program for identification of flavonoids. FlavonQ-2.0v was validated by analyzing data from samples spiked with flavonoid mixed standards and blueberry, mizuna, purple mustard, red cabbage, and red mustard green extract samples. Accuracies of identification for all samples were above 88%. FlavonQ-2.0v greatly facilitates the identification and quantitation of flavonoids from UHPLC-HRAM-MS data. The automated computational tool is developed to assist, rather than replace, human expert. The result shows that it not only saves tremendous efforts for human experts but also allows less-experienced chemists to perform data analysis on flavonoids with reasonable results.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b00771.

Materials and reagents; data analysis results from FlavonQ-2.0v for spiked flavonoid standard mixtures; flavonoids identified by FlavonQ-2.0v from blueberry; fragmentation patterns in the MS spectra of three flavonoids; chromatogram for blueberry sample generated from FlavonQ-2.0v with classification and putative identification results labeled on the peaks; and graphic user interface for FlavonQ-2.0v (PDF)

AUTHOR INFORMATION

Corresponding Author

*Phone: +1 301 504 8144. Fax: +1 301 504 8314. E-mail: pei.chen@ars.usda.gov.

ORCID

Pei Chen: 0000-0002-1457-5177

Author Contributions

†M.Z. and J.S. contributed equally to this manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research is supported by the Agricultural Research Service of the U.S. Department of Agriculture, an Interagency Agreement Number AOD12026-001-01004 with the Office of Dietary Supplements at the National Institutes of Health. The John A. Milner Fellowship program by USDA Beltsville Human Nutrition Research Center and the NIH Office of Dietary Supplements is acknowledged for the support to Dr. Mengliang Zhang. We thank Dr. Peter de B. Harrington from Department of Chemistry and Biochemistry at Ohio University for providing Matlab routines for PDR, PCA, SIMCA, and FOAM functions. We thank Dr. Joseph M. Betz from NIH Office of Dietary Supplements and Dr. James M. Harnly from USDA for the careful revision of this article.

REFERENCES

(1) Nijveldt, R. J.; van Nood, E.; van Hoorn, D. E. C.; Boelens, P. G.; van Norren, K.; van Leeuwen, P. A. M. *Am. J. Clin. Nutr.* **2001**, *74*, 418−425.

(2) Balentine, D. A.; Dwyer, J. T.; Erdman, J. W., Jr.; Ferruzzi, M. G.; Gaine, P. C.; Harnly, J. M.; Kwik-Uribe, C. L. *Am. J. Clin. Nutr.* **2015**, *101*, 1113−1125.

(3) Satterfield, M.; Brodbelt, J. S. *Anal. Chem.* **2000**, *72*, 5898−5906.

(4) Johnson, A. R.; Carlson, E. E. *Anal. Chem.* **2015**, *87*, 10668−10678.

(5) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.

(6) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinf.* **2010**, *11*, 395.

(7) Katajamaa, M.; Miettinen, J.; Oresic, M. *Bioinformatics* **2006**, *22*, 634−636.

(8) Wei, X.; Sun, W.; Shi, X.; Koo, I.; Wang, B.; Zhang, J.; Yin, X.; Tang, Y.; Bogdanov, B.; Kim, S.; Zhou, Z.; McClain, C.; Zhang, X. *Anal. Chem.* **2011**, *83*, 7668−7675.

(9) Zhang, W. C.; Chang, J.; Lei, Z. T.; Huhman, D.; Sumner, L. W.; Zhao, P. X. *Anal. Chem.* **2014**, *86*, 6245−6253.

(10) Zhang, M.; Sun, J.; Chen, P. *Anal. Chem.* **2015**, *87*, 9974−9981.

(11) Kazusa DNA Research Insititute, *KOMICS*, 2008.

(12) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534−2536.

(13) Zhang, M.; Harrington, P. d. B. *Talanta* **2013**, *117*, 483−491.

(14) Lin, L. Z.; Harnly, J.; Zhang, R. W.; Fan, X. E.; Chen, H. J. *J. Agric. Food Chem.* **2012**, *60*, 544−553.

(15) Sun, J. H.; Lin, L. Z.; Chen, P. *Curr. Anal. Chem.* **2013**, *9*, 397−416.

(16) Chen, J. H.; Ho, C. T. *J. Agric. Food Chem.* **1997**, *45*, 2374−2378.

(17) Frank, I. E.; Lanteri, S. *Chemom. Intell. Lab. Syst.* **1989**, *5*, 247−256.

(18) Wabuyele, B. W.; Harrington, P. D. *Appl. Spectrosc.* **1996**, *50*, 35−42.

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

(20) Harrington, P. B. *J. Chemom.* **1991**, *5*, 467−486.

(21) Wang, Z.; Zhang, M.; Harrington Pde, B. *Anal. Chem.* **2014**, *86*, 9050−9057.

(22) Harnly, J. M.; Doherty, R. F.; Beecher, G. R.; Holden, J. M.; Haytowitz, D. B.; Bhagwat, S.; Gebhardt, S. *J. Agric. Food Chem.* **2006**, *54*, 9966−9977.

(23) Anderssen, E.; Dyrstad, K.; Westad, F.; Martens, H. *Chemom. Intell. Lab. Syst.* **2006**, *84*, 69−74.

(24) Bohm, B. A. In *Introduction to Flavonoids*; Harwood Academic Publishers: Amsterdam, The Netherlands, 1999; p 200.

(25) Zhang, M.; de B. Harrington, P.; Chen, P. *Curr. Chromatogr.* **2015**, *2*, 145−151.

(26) Zhang, M.; Zhao, Y.; Harrington, P. d. B.; Chen, P. *Anal. Lett.* **2016**, *49*, 711−722.

(27) Xu, Z. F.; Sun, X. B.; Harrington, P. D. *Anal. Chem.* **2011**, *83*, 7464−7471.

(28) Chen, P.; Lu, Y.; Harrington, P. B. *Anal. Chem.* **2008**, *80*, 7218−7225.

(29) Sun, J.; Lin, L. Z.; Chen, P. *Rapid Commun. Mass Spectrom.* **2012**, *26*, 1123−1133.

(30) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27*, 747−751.

(31) Lei, Z. T.; Jing, L.; Qiu, F.; Zhang, H.; Huhman, D.; Zhou, Z. Q.; Sumner, L. W. *Anal. Chem.* **2015**, *87*, 7373−7381.

(32) Nishioka, T.; Kasama, T.; Kinumi, T.; Makabe, H.; Matsuda, F.; Miura, D.; Miyashita, M.; Nakamura, T.; Tanaka, K.; Yamamoto, A. *Mass Spectrom.* **2014**, *3*, S0039−S0039.

(33) Stein, S. *Anal. Chem.* **2012**, *84*, 7274−7282.

(34) Qiu, F.; Fine, D. D.; Wherritt, D. J.; Lei, Z.; Sumner, L. W. *Anal. Chem.* **2016**, *88*, 11373−11383.

(35) Lin, L. Z.; Harnly, J. M. *J. Agric. Food Chem.* **2012**, *60*, 5832−5840.