Metabolomic Profiling of Wines using LC/QTOF MS and MassHunter Data Mining and Statistical Tools

Application Note

Food Testing and Agriculture

Authors

Lucas Vaclavik, Ondrej Lacina, and Jana Haslova
Institute of Chemical Technology
Prague, Czech Republic

Jerry Zweigenbaum
Agilent Technologies, Inc.
2850 Centerville Road
Wilmington, DE 19808
USA

Abstract

A non-targeted metabolomic analysis approach to the classification of wine varieties was developed, employing LC/QTOF MS and MassHunter Workstation software. Molecular feature extraction, data filtering, and statistical analysis utilizing ANOVA and PCA identified 26 marker compounds that were then used to build a PLSDA prediction model. An overall accuracy of 95.6% in discriminating between Cabernet Sauvignon, Merlot and Pinot Noir red wine varieties was obtained using the model. This approach may be widely applicable to the analysis and characterization of foods.

Introduction

Wine is a beverage widely consumed throughout the world, with consumption estimated at 2.65 billion 9-liter cases in 2008, and wine consumption in the United States registered its 15th consecutive annual gain [1]. Thus wine is an important global food commodity of high commercial value. In order to protect this valuable market, wine authenticity control, mainly in terms of varieties, geographical origin and age, is continuously required to detect any adulteration and to maintain wine quality [2]. Wine is a complex matrix of water, alcohol, sugars and a wide range of minor organic and inorganic constituents. Many factors influence the concentration levels of these compounds, including grape variety, climate, wine-growing area, and, last but not least, the winemaking process. Establishing wine authenticity can thus be a very challenging proposition.

Agilent Technologies

In the past few years, the emerging field of metabolomics has become an important strategy in many research areas such as disease diagnostics, drug discovery and food science. It has been used for informative, discriminative, and predictive purposes associated with food quality and safety. Metabolomic studies aim at the comprehensive analysis of numerous targeted or non-targeted metabolites (compounds with molecular weight typically below 1000 daltons) in a biological sample. As the metabolites significantly differ in both physicochemical properties and abundance, these analyses require sophisticated analytical technologies.

One of the techniques most widely used in metabolomics is liquid chromatography/mass spectrometry (LC/MS). For processing of large chromatographic and/or spectral data sets typically generated by metabolomic analyses, effective software tools capable of rapid data mining procedures and aligning algorithms have to be used. Advanced chemometric tools for reduction of data dimensionality are also employed to maximize utilization of the information present in the multivariate records obtained from data mining, including principal component analysis (PCA) and Analysis of Variance (ANOVA).

This application note describes a non-targeted metabolomic analysis approach to the classification of wine varieties that employs liquid chromatography coupled to quadrupole time-of-flight mass spectrometry (LC/QTOF MS). The Agilent Series 1200 SL Rapid Resolution LC system was coupled to an Agilent 6530 accurate-mass Q-TOF MS with electrospray ionization (ESI) enabled with Agilent Jet Stream technology. MassHunter Workstation software, including Qualitative Analysis and Mass Profiler Professional, was used for molecular feature extraction and data filtering, followed by multivariate analysis by PCA and one-way ANOVA, eventually leading to the construction of a classification model using Partial Least Square Discrimination Analysis (PLSDA). The end result was the use of 26 compounds that enabled a predictive model that was 95.6% accurate in discriminating between Cabernet Sauvignon, Merlot and Pinot Noir red wine varieties. The results of this study have been published [3].

Experimental

Reagents and standards

The reagents and standards used were as specified in the previous publication [3]. HPLC grade methanol was purchased from Honeywell Burdick and Jackson (Muskegon, MI, USA), and deionized water was produced by a Milli-Q purification system (Millipore, Bedford, MA, USA). Ammonium formate, ammonium acetate, acetic acid, and formic acid, used as mobile phase additives (each of purity ≥ 99%), were supplied GFS Chemicals (Powell, OH, USA) using doubly distilled formic acid, acidic acid and ammonium hydroxide at the appropriate concentrations.

Samples

In total, 45 red wine samples of Cabernet Sauvignon (CS, n = 15), Merlot (M, n = 16), and Pinot Noir (PN, n = 14) varieties were purchased from reliable sources at Czech Republic and U.S. retail markets. Samples represented wines of various geographic origin (Australia, Bulgaria, Czech Republic, France, Germany, Hungary, Chile, Italy, Macedonia, Slovakia, Spain, and USA), and vintage (2004–2008), comprising a highly variable sample set. Wines were uncorked and an aliquot of sample was transferred into a 2-mL autosampler amber vial (filled to its capacity). Samples were stored in the dark at 2 °C, until LC/MS measurements were taken. Samples were analyzed in random order.

Instruments

This study was performed on an Agilent Series 1200 SL Rapid Resolution LC system coupled to an Agilent 6530 accurate-mass Q-TOF MS with electrospray ionization (ESI) enabled with Agilent Jet Stream technology. The instrument conditions are listed in Table 1.

To assure the desired mass accuracy of recorded ions, continuous internal calibration was performed during analyses with the use of signals at *m/z* 121.0509 (protonated purine) and *m/z* 922.0098 (protonated hexakis,1H, 1H, 3H-tetrafluoropropoxy, phosphazine or HP-921) in positive ion mode; in negative ion mode, ions with *m/z* 119.0362 (proton abstracted purine) and *m/z* 980.0164 (acetate adduct of HP-921) were used.

Table 1. LC and Mass Spectrometer Conditions

LC Run Conditions	
Column	Agilent ZORBAX Eclipse Plus C-18, 2.1 × 100 mm, 1.8 µm (p/n 959931-902)
Column temperature	40 °C
Injection volume	2 µL
Autosampler temperature	4 °C
Needle wash	Flush port (50:25:25 H_2O, IPA:MeOH:H_2O, 5 sec)
Mobile phase	Positive ionization mode: A = 5 mM Ammonium formate B = Methanol Negative ionization mode: A = 5 mM Ammonium formate B = Methanol
Flow rate	0.3 mL/min during gradient run 0.5 mL/min during column equilibration
Gradient	B = 5% to 65%, 0 to 13 min B = 65% to 100%,13 to 16 min B = 100% from 16 to 20 min B = 5% from 20 to 24 min (column equilibration)
Analysis time	20 min
MS Conditions	
Ion mode	Positive and negative, ESI+APCI multimode ionization
Drying gas temperature	300 °C
Vaporizer temperature	170 °C
Drying gas flow	11 L/min
Nebulizer pressure	40 psi
Capillary voltage	4500 V positive ion mode 3000 V negative ion mode
Skimmer voltage	65 V
Octapole DC1	47 V
Octapole RF	750 V
Fragmentor voltage	125 V
Spectra acquisition rate	1.4 spectra/second
MS/MS Conditions	
Ion mode	Positive ion
Isolation window	4 amu
Spectra acquisition rate	1.4 spectra/second

Data Processing and Statistical Analysis

MassHunter Workstation software, including Qualitative Analysis (version 3.01) and Mass Profiler Professional (version B.02.00), was used for processing both raw MS and MS/MS data, including molecular feature extraction, background subtraction, data filtering, statistical analysis by ANOVA and PCA, followed by the construction of the predictive classification model, molecular formula estimation, and database searching.

To perform subtraction of molecular features (MFs) originating from the background, analysis of a blank sample (deionized water) was carried out under identical instrument settings and background MFs were removed. Using background subtracted data, files in compound exchange format (CEF files) were created for each sample and exported into the Mass Profiler Professional (MPP) software package for further processing.

Results and Discussion

Wine analysis

This metabolomic profiling study (Figure 1) began with the analysis of a total of 45 red wine samples from around the world comprising three different categories based on the grape varieties used to produce them: Cabernet Sauvignon (15), Merlot (16), and Pinot Noir (14). To avoid any possible discrimination of metabolites present, no sample pre-treatment procedures such as extraction or dilution were performed prior to LC/ MS analysis, and the wine samples were injected directly onto the separation column. As this was an untargeted study, generic settings were applied to both LC separation and MS analysis to obtain profiles containing as many compounds as possible. Both reversed-phase chromatography and hydrophilic interaction liquid chromatography (HILIC) were used to separate the wide variety of compounds present in wine. In addition, Agilent Jet Stream technology enabled electrospray and multimode ion sources were used to collect the MS data. The multimode ion source provided simultaneous APCI and electrospray ionization. The most useful data was collected with the reversed-phase Agilent ZORBAX Eclipse Plus C18, 2.1 × 100, 1.8 µm particle size column.

The wine sample analyses were quite complex, exhibiting a multitude of peaks, each containing multiple compounds (Figure 2). In addition, the solvent blank contained contaminant peaks that needed to be subtracted in order to generate valid data. Data mining was therefore required to extract meaningful data from the results.

EXPERIMENTAL SETUP

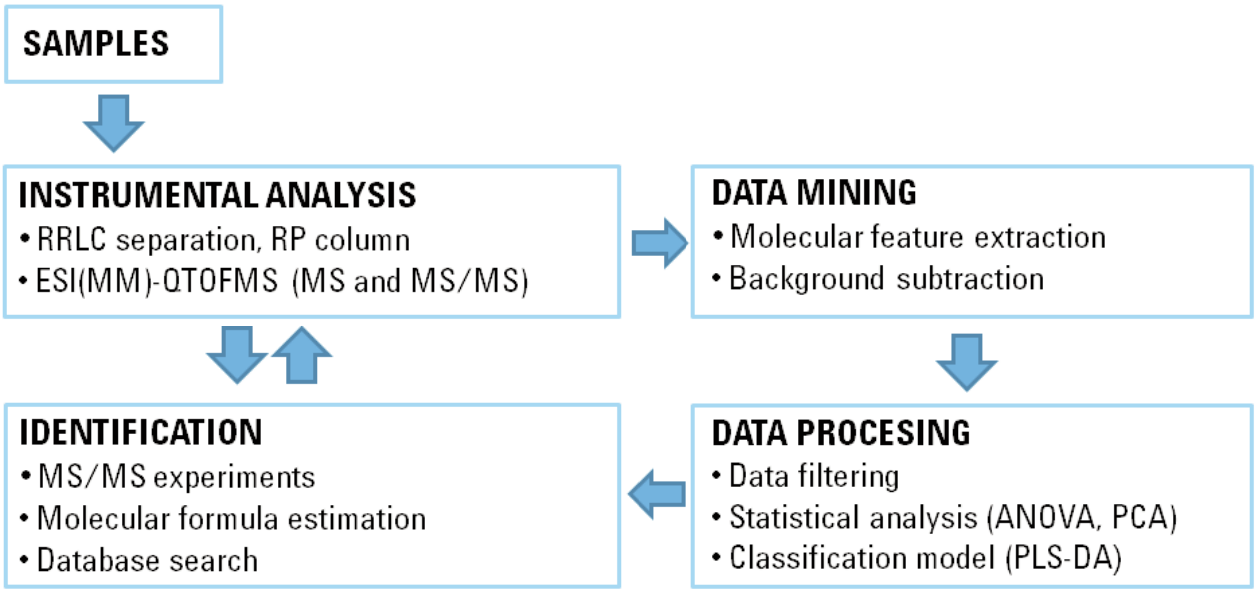

Figure 1. Workflow for a metabolomic study to generate a predictive model for wine classification and identify the marker compounds.

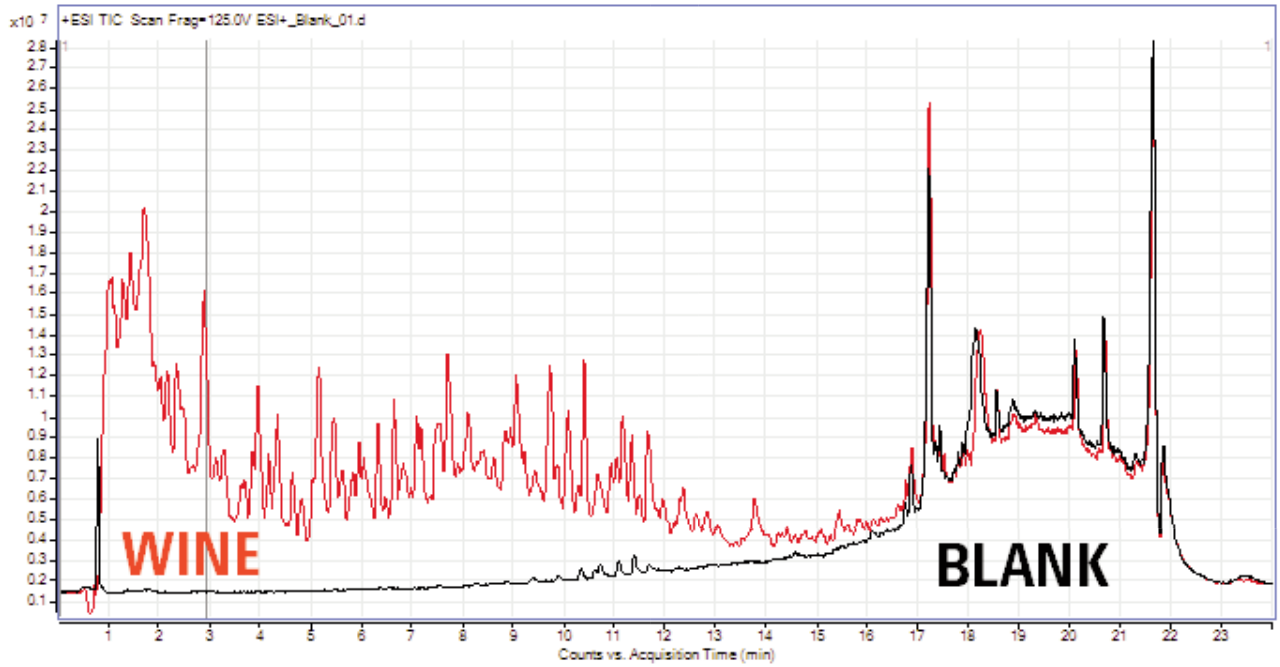

Figure2. Total Ion Current (TIC) chromatogram of a wine sample Ilustrating its complexity. Each peak contains several compounds. The blank must also be subtracted in order to properly interpret the data.

Data mining

The Molecular Feature Extractor (MFE) algorithm in the MassHunter workstation software was used to perform meaningful data mining. This algorithm removes data points that correspond to persistent or slowly-changing background, searches for features that have a common elution profile and groups ions into one or more *compounds* (features) containing *m/z* values that are related (correspond to peaks in the same isotope cluster, different adducts or charge states of the same entity). Molecular feature extraction also enables subtraction of chromatographic background caused by impurities in the mobile phase, extracting all compounds found in the solvent blank. The results are then used as a background subtraction dataset for all sample files, excluding all compounds that were in the blank from the evaluation of the samples.

The settings for the molecular feature extractor are important, because the algorithm has been *tuned* for different types of large and small molecules. Figure 3 shows the selection for the small molecules of interest in this study.

The masses found in the blank can be copied and pasted right into the MFE tab for automated background subtraction (Figure 4). Once all the settings are made for the data processing method, including the appropriate MFE algorithm and the threshold counts (Figure 3), it is saved and used to process all data files in a batch mode using the offline utility. The batch data processing includes finding compounds using MFE and then creating a compound exchange format (.CEF) file of the results. Those results are saved in a project folder that will then be used by Mass Profiler Professional for data filtering and statistical analysis.

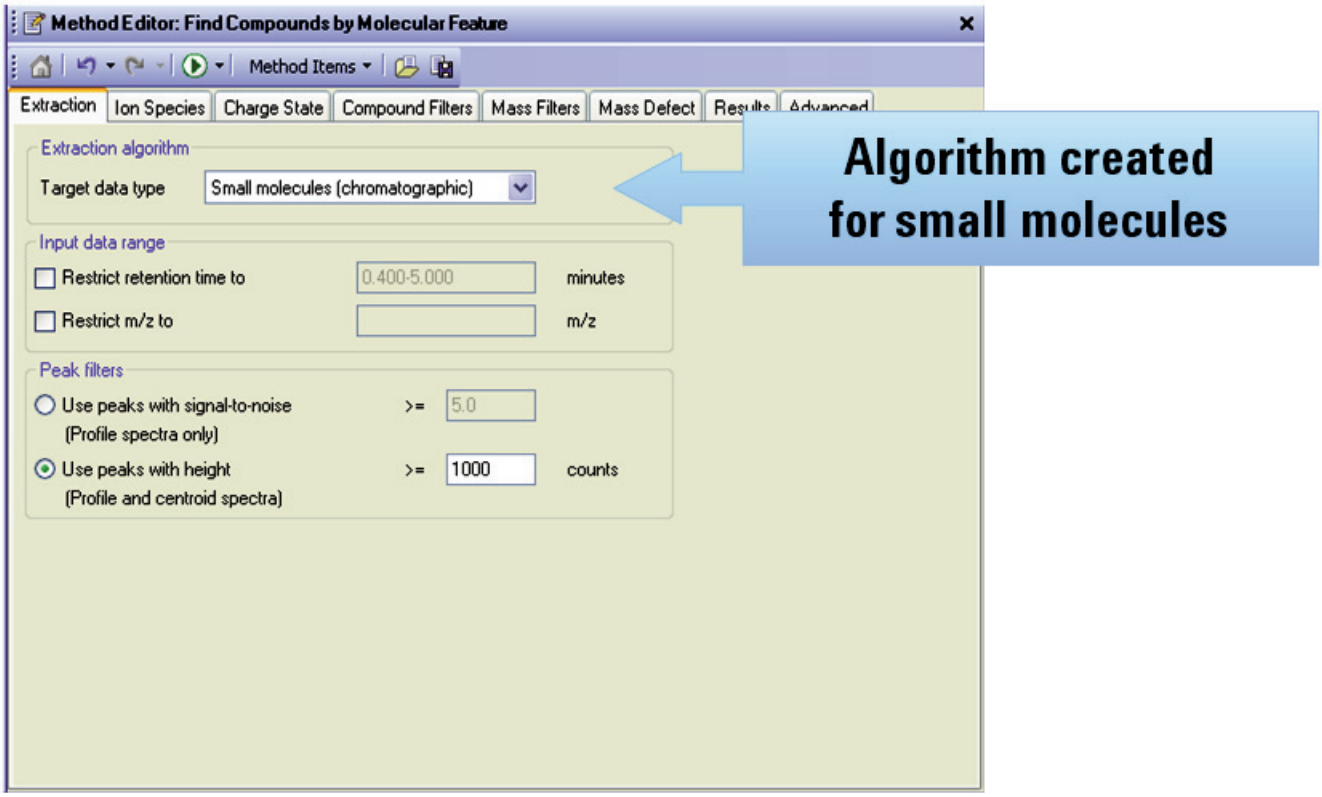

Figure 3. Settings used by the Molecular Feature Extractor (MFE) to extract entities (compounds) from the LC/QTOF MS data.

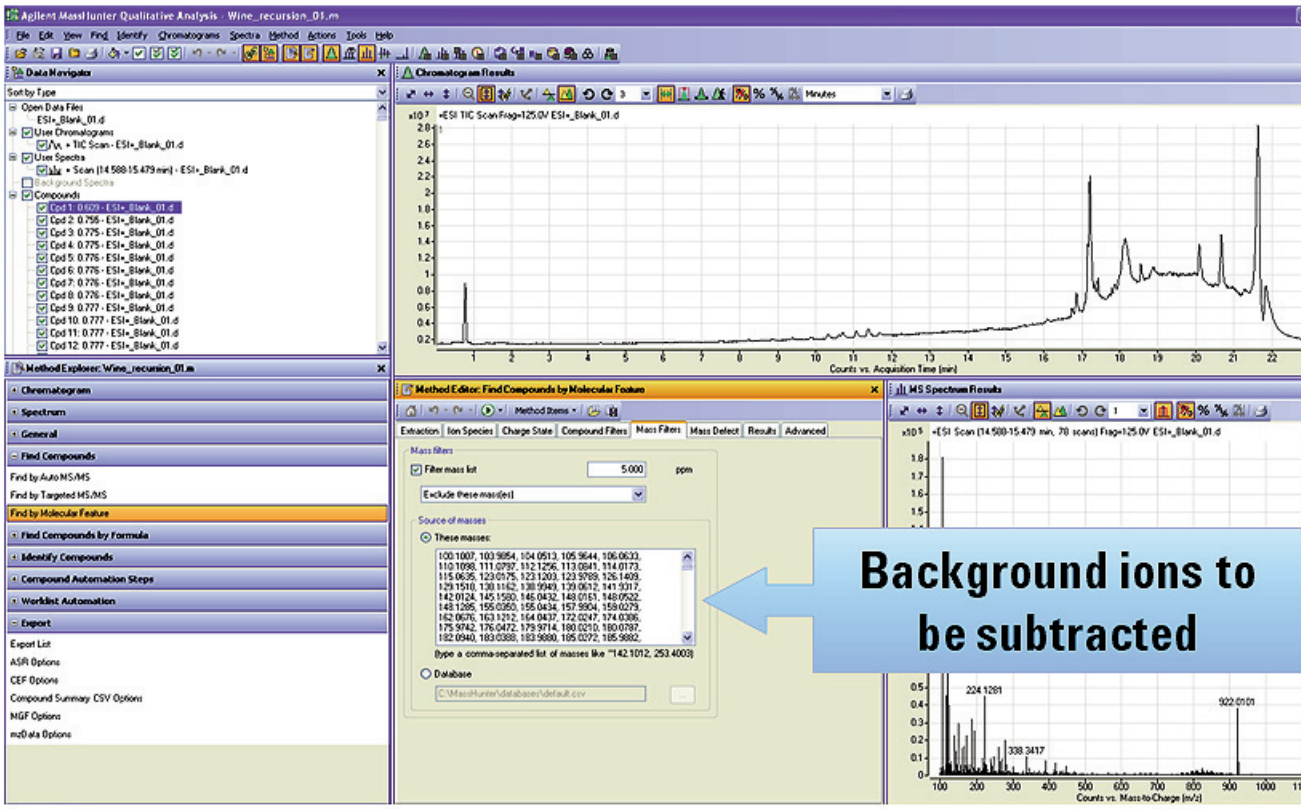

Figure 4. The resulting masses found in the blank can be copied and pasted right into the MFE tab for subtraction from all of the datasets.

Data processing

Data processing for metabolomic studies is often very tedious and time intensive when using complicated statistical software written to handle ASCI or text type results. Mass Profiler Professional (MPP) is ideal for the sophisticated data management, filtering, statistical analysis, interpretation, model creation, and prediction required to efficiently utilize metabolic data. It provides an easy-to-follow guided workflow that helps the user decide how best to evaluate the data. Expert users can go directly to the data processing they wish to use (see the Mass Profiler Professional brochure 5990-4164EN for further details).

MPP uses eight steps for data evaluation, starting with a summary report, revealing that the wine data set contains over 20,000 entities (or possible compounds) found throughout the samples. These are determined by the accurate mass of the cluster of ions and their chromatographic alignment. The next step is experiment grouping into classes, and in this case the three classes are Cabernet Sauvignon (CS), Merlot (M) and Pinot Noir (PN).

Data filtering

Entity filtering permits the creation of a higher quality data set so that subsequent multivariate analysis is more meaningful. The first filter determined which entities (compounds) were in at least one group 100% of the time (frequency analysis). That is, the compound must be in that group in all the samples. This frequency filter reduced the possible markers from 20506 entities to 663. However, by setting the frequency to 100%, some important markers could be filtered out. In the case of Pinot Noir, one compound was removed, even though it was present in 14 out of the 15 samples (Figure 5). Setting the frequency filter to 50% allows the inclusion of many more entities without excluding those like the compound in 14 of 15 Pinot Noir samples. Using the 50% filter reduced the number of entities from 20506 to 3600.

The next step was to filter the Analysis of Variance (ANOVA) results, which determine what level of variance is accepted as significant for a given entity. Using a probability p value of .05 (variance from one sample to another has a 95% probability that it is significant), the 3600 entities from the frequency filter were reduced to 40 significant compounds.

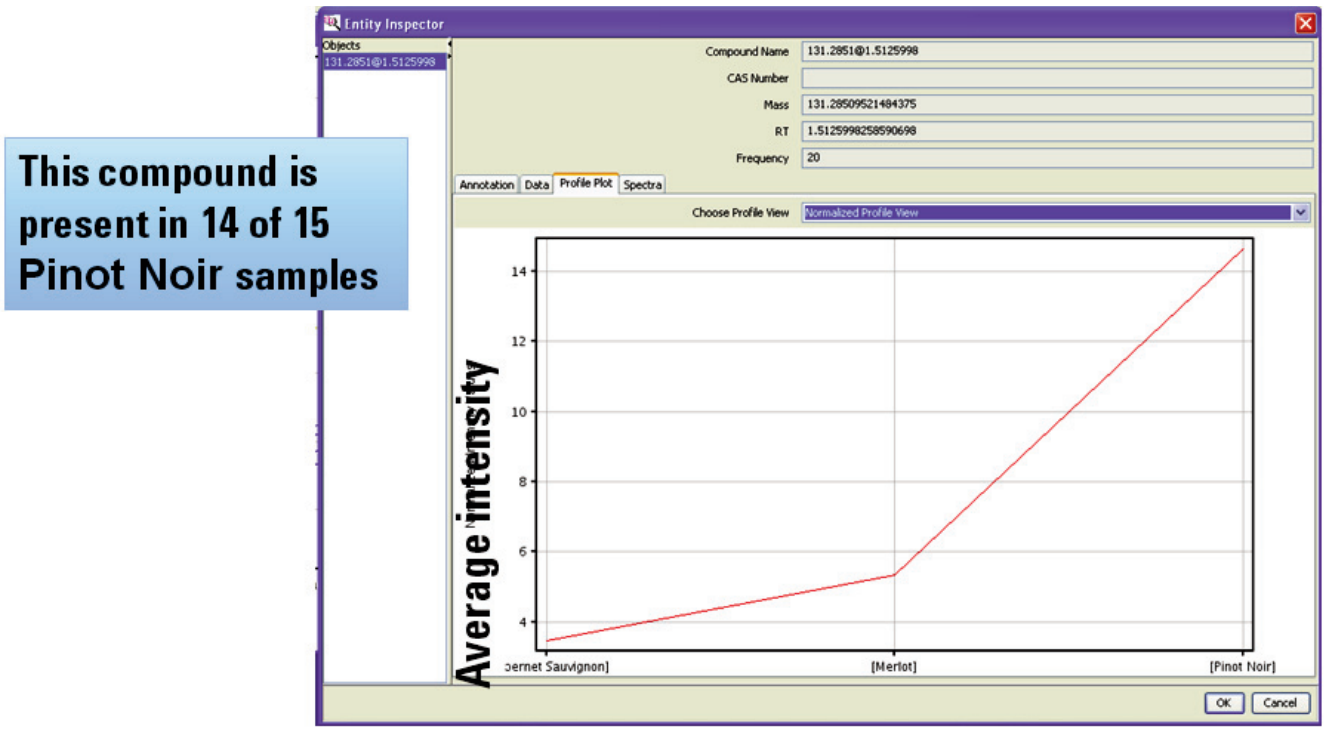

Figure 5. Using a 100% frequency filter can remove useful markers from the dataset, such as this compound that is present in Cabernet Souvignon and Merlot samples at very low frequency, but shows up in 14 of 15 Pinot Noir samples.

Fold Change was the final data processing filter applied, in order to identify entities with large abundance differences between the selected data classes, that is, those that differ in concentration by 2 fold, 3 fold, 4 fold, etc. between the three data classes (CS, M and PN). Examining the data at higher fold change than two, however, eliminated possible discriminating compounds of Cabernet and Merlot, and thus a 2-fold filter was applied (Figure 6). This reduced the data set from 40 to 26 possible compounds.

The next step in the processing of the data was recursion. Recursion allows the re-examination of the data to assure that each entity is a real peak and that those entities not found in a sample are not there. MassHunter Qualitative Analysis software automatically re-extracted the final group of 26 markers from the raw data to generate extracted ion chromatograms (EICs). A careful inspection of the resultant EICs was performed to eliminate false positives (not a real peak) and false negatives (a real peak is in a sample but was missed in the molecular feature extraction data mining step). Once the 26 entities were confirmed as real, statistical analysis was performed using Principle Component Analysis (PCA).

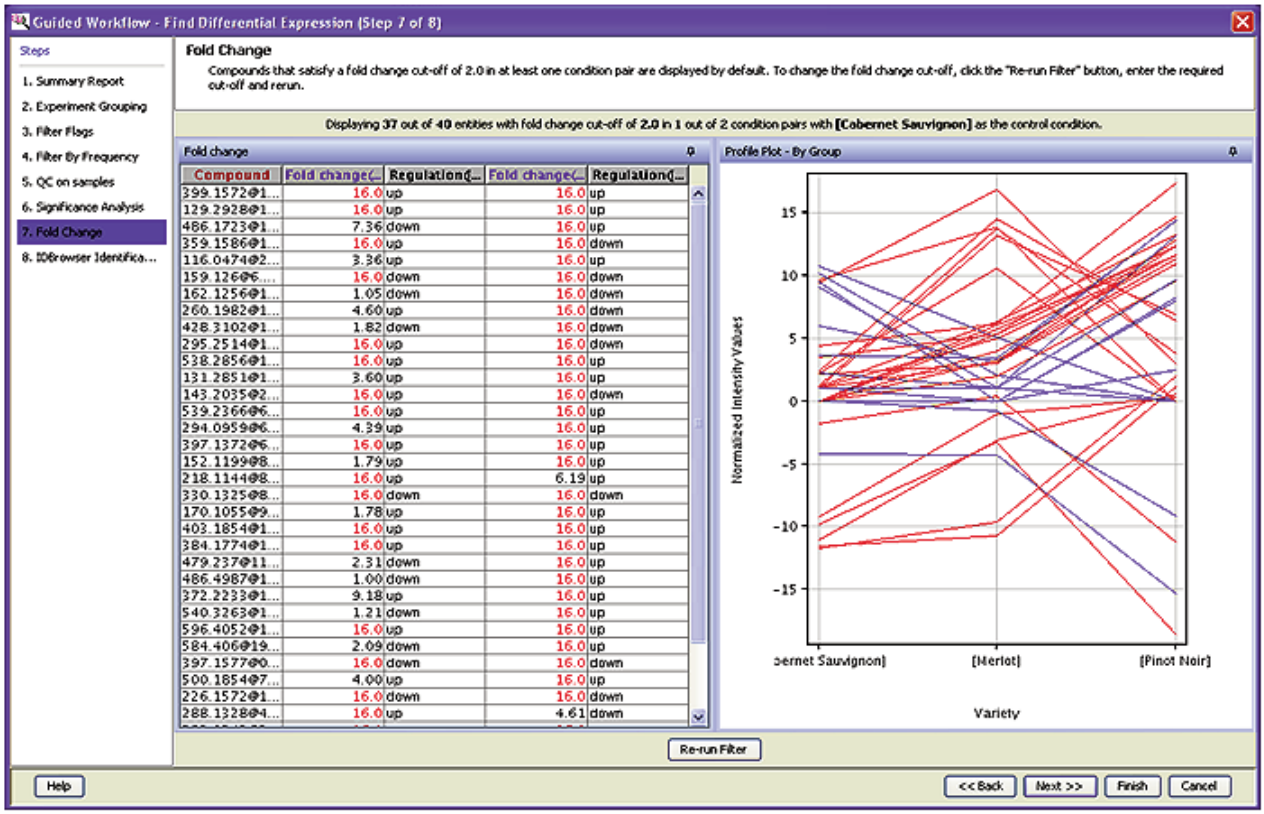

Figure 6. Fold Change Analysis is used to identify entities with significantly different abundance in selected classes. Examining the data at fold change higher than two eliminates possible Cabernet and Merlot markers, so a 2-fold filter was applied. This reduced the data set from 40 to 26 possible markers.

Statistical analysis

PCA is a frequently employed unsupervised multivariate analysis technique enabling data dimensionality reduction, while retaining the discriminating power in the data. It is performed using the transformation of measured variables into uncorrelated principal components, each being a linear combination of the original variables. The goal is to identify possible relationships within the classes of data. Performing PCA on the unfiltered data set and even the frequency-filtered data set did not reveal any relationships that would enable the data to be classified into varieties of wine. However, PCA of the 26 identified marker entities revealed distinctive grouping of the data into the three wine classes (Figure 7). Note that PCA does not make the statistical distinction between varieties; it only reveals that there are distinctions. The compounds that distinguish one variety of grapes from another were selected by the frequency filtering, ANOVA, and fold change filtering of the entities identified by the molecular feature extractor, and then qualified by the recursion analysis.

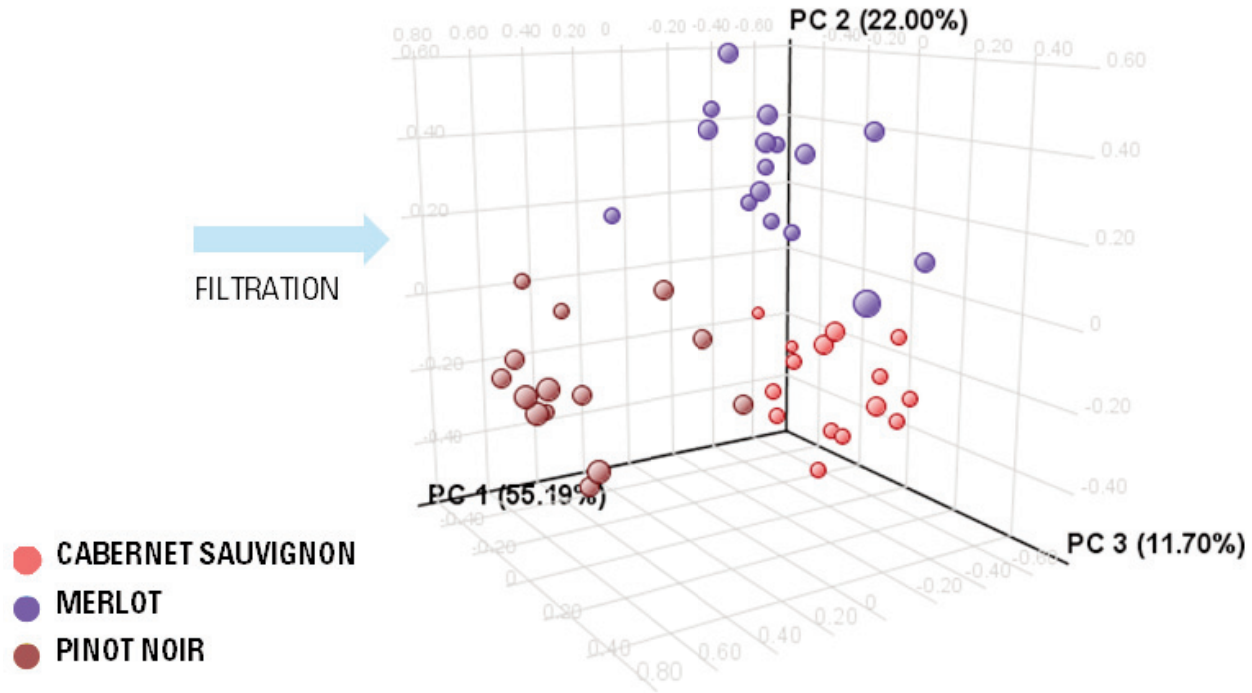

Figure 7. Principle Component Analysis (PCA) of the 26 markers remaining after the ANOVA and Fold Change filters were applied to the dataset. The markers for each of the three wine varieties group together well, indicating their utility for predicting the variety of an unknown wine.

Classification model

Having established three data classes with the filtered compounds that were selected through processing with Mass Profiler Professional, the next step was to create a model that can predict the variety of a wine. The Partial Least Square Discrimination Analysis (PLSDA) model in MPP best fit the mass spectral data. The first step in building the classification model was to train the model with the data.

The PLSDA algorithm produces a Confusion Matrix, which is a table with the true class in rows and the predicted class in columns. The diagonal elements represent correctly classified experiments, and cross diagonal elements represent misclassified experiments (Table 2). The table also shows the predictive accuracy of the model as the percentage of correctly classified experiments in a given class. The accuracy of this training set for each class, as well as overall, was 100%.

The next step was to test the model with the same data. Although redundant, this is a valid statistical procedure. The same class prediction model was used for the validation of the trained model. Note that Merlot was incorrectly identified in two cases, resulting in an accuracy of prediction for Merlot of 87.5%, and an overall predictive ability of 95.6% (Table 2). The use of more samples would likely improve the predictive ability of the model by improving the statistical power of the analysis.

Table 2. Confusion Matrix Illustrating the Classification Results Using the PLSD Model

	Cabernet Sauvignon (CS)	Merlot (M)	Pinot Noir (PN)	Accuracy (%)
Model training	**15**	**0**	**0**	
CS	0	16	0	**100.0**
M	0	0	14	**100.0**
PN				**100.0**
Recognition ability (%)				**100.0**
Model validation				
CS	15	0	0	**100.0**
M	1	14	1	**87.5**
PN	0	0	14	**100.0**
Predictive ability (%)				**95.6**

In order to demonstrate the predictive ability of the model, five wines were purchased that were not among the wines used to find the markers and develop the model. An additional wine was purchased whose identity was not revealed to the scientists conducting the experiment. Analysis of these wines and applying the classification model correctly predicted the variety of all five wines plus the unknown (Figure 8). These results demonstrate the feasibility of developing markers and using a model to accurately determine wine variety.

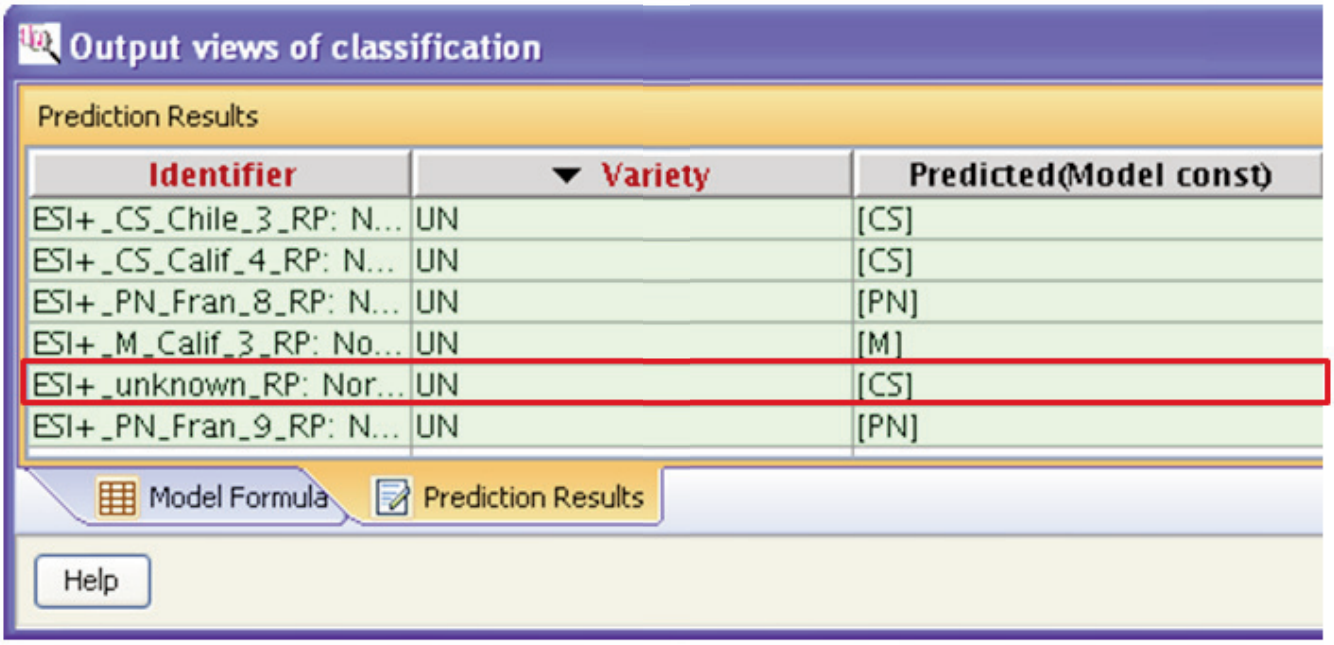

Figure 8. Table showing the results of the predictive model when applied to known wine samples, as well as one unknown. All samples were correctly classified.

Identification of wine markers

While it is not necessary to know the identity of the compounds used as wine markers, the availability of standards for the marker compounds could facilitate simpler tests for identification of wine variety. Using the single MS accurate mass data for the pseudo molecular ion and its isotopes, a molecular formula for each marker compound can be generated. Those formulae that provide a best fit of this data can be used to search private and public databases of possible compounds. Of the 26 markers, only one gave a result from a database. The molecular formula estimation and subsequent tentative identification of a selected marker compound for Pinot Noir (*m/z* 449.1078, RT 11.16 min), present in the final group of markers, were performed based on single MS. Using this data to search the PubChem database resulted in a suspect identification of this marker compound as cyanidin-3-O-glucoside, which is anthocyanidin pigment. Using the Q-TOF, accurate mass MS/MS of the *m/z* 449.1078 gave further indication that this is the identity of the marker compound. Without any idea of what a compound might be, even accurate MS and MS/MS data would be difficult to interpret to obtain a compound structure. Final confirmation would require a standard of the indicated compound.

Conclusions

Metabolomic studies are valuable tools for the profiling of complex food products such as wines. Using the highly accurate and reproducible data generated by the Agilent LC/Q-TOF MS system, a predictive model can be constructed to determine the variety of a wine. This may be broadly applicable to other foodstuffs. Sophisticated software tools such as MassHunter Qualitative Analysis and Molecular Profiler Professional can be used to conduct data mining, filtering and statistical analysis to identify markers specific to particular food varieties (such as wine varieties) and use those markers to build classification models that can determine the variety of a foodstuff. Such models can be highly accurate, with the overall accuracy of the wine variety classification model described here being 96.5%. While it is not necessary to know the identity of the marker compounds for the model to be accurate, the same instrument system can be used to generate MS/MS spectra that can enable identification of the marker compounds used in the model. If standards for the putative compounds are available, confirmation of the compound's identity can then be made.

References

1. Global Drinks Report: Wine Market Stagnates, Wine Spectator, Posted April 6 2009, http://www.winespectator.com/webfeature/show/id/Global-Drinks-Report-Wine-Market-Stagnates_4704.
2. P.R. Arhurst, M.J. Dennis, Food Authentication, Chapman-Hall, London, 1996.
3. L. Vaclavik, O. Lacina, J. Hajslova, J. Zweigenbaum. "The use of high performance liquid chromatography-quadrupole time-of-flight mass spectrometry coupled to advanced data mining and chemometric tools for discrimination and classification of red wines according to their variety.", Anal Chim Acta. **685**, 45-51 (2011).

For More Information

These data represent typical results. For more information on our products and services, visit our Web site at www.agilent.com/chem.

www.agilent.com/chem

Printed in the USA
June 22, 2011
5990-8451EN

Agilent Technologies