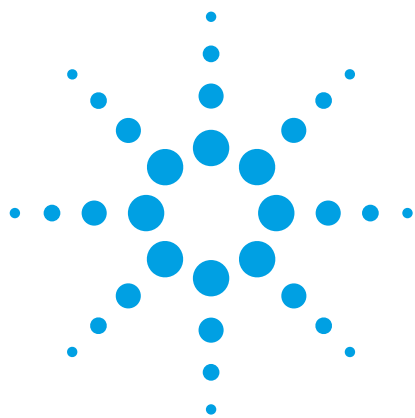




采用LC/QTOF MS 和 MassHunter 数据挖掘和统计工具对葡萄酒进行代谢组学分析

应用报告

食品检测和农业

作者

Lucas Vaclavik, Ondrej Lacina, Jana Haslova

Institute of Chemical Technology
Prague, Czech Republic

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

摘要

采用LC/QTOF MS和MassHunter工作站软件，建立了一种用于葡萄酒品种分类的非靶向代谢组学分析方法。利用分子特征提取、数据筛选和统计方法方差分析（ANOVA）和主成分分析（PCA），确定了26个标记化合物，并用其建立了一个PLSDA预测模型。然后使用该模型对赤霞珠、梅洛和黑比诺红葡萄酒品种进行鉴别，获得了95.6%的准确度。该方法可以广泛用于食品的分析 and 表征。

引言

葡萄酒是世界各地广泛消费的一种饮料。以9升一个包装计，2008年的消费量估计为26.5亿件。在美国，葡萄酒的消费量已连续15年保持增长[1]。因此，葡萄酒是一种商业价值高的重要的全球性食品。为了保护这一宝贵市场，需要对葡萄酒的真实性进行控制，主要包括对品种、原产地和酒龄等方面进行鉴定，以打击假冒伪劣和保证葡萄酒的品质[2]。葡萄酒是由水、酒精、糖，以及多种微量的有机和无机成分组成的复杂基体。许多因素影响这些化合物的浓度水平，包括葡萄的品种、气候、葡萄种植区域，最后，葡萄酒的酿造过程也相当重要。因此，建立葡萄酒的真伪分析方法是一项极具挑战性的任务。



Agilent Technologies

在过去几年中，代谢组学这一新兴技术已成为疾病诊断、药物发现和食品科学等许多研究领域的重要手段。它已被用于与食品质量和安全相关的信息获取、鉴别和预测。代谢组学研究的目的是对生物样品中众多靶向或非靶向的代谢物（通常分子量低于1000道尔顿的化合物）进行综合分析。由于这些代谢物的理化性质和丰度存在明显差异，因此需要复杂的分析技术进行分析。

在代谢组学研究中，液相色谱/质谱联用（LC / MS）是应用最为广泛的技术之一。为了处理代谢组学分析所获得的大量色谱和/或光谱数据，必须使用能够进行快速有效的数据挖掘和匹配运算的软件工具。能够对数据进行降维处理的先进的化学计量学工具也常被用于最大程度利用数据挖掘得到的多元记录信息，包括主成分分析（PCA）和方差分析（ANOVA）。

本应用报告描述了一个采用液相色谱/四极杆飞行时间质谱（LC/QTOF MS）技术进行葡萄酒品种分类的非靶向代谢组学分析方法。将Agilent 1200 SL系列快速高分离度LC系统与配备安捷伦喷射流技术电喷雾电离源（ESI）的Agilent 6530精确质量Q-TOF MS联用。使用MassHunter工作站软件，包括Qualitative Analysis和Mass Profiler Professional模块，进行分子特征提取、数据筛选，然后进行多元PCA分析和单因素ANOVA分析，最后使用偏最小二乘判别分析（PLSDA）构建一个分类模型。结果表明，使用由26种化合物建立的预测模型对赤霞珠、梅洛和黑比诺红葡萄酒进行品种鉴别，准确度为95.6%。这项研究结果已经发表[3]。

实验部分

试剂与标准品

试剂和标准品与之前发表的工作[3]相同。色谱纯甲醇购自Honeywell Burdick and Jackson公司（Muskegon, MI, USA），去离子水由Milli-Q纯化系统（Millipore, Bedford, MA, USA）制备。甲酸铵、醋酸铵、醋酸和甲酸作为流动相添加剂使用（纯度均≥99%），系GFS Chemicals公司产品（Powell, OH, USA），使用二次蒸馏的甲酸、醋酸和氢氧化铵配制成适当的浓度。

样品

从捷克共和国和美国的零售市场通过可靠渠道共购得45个红葡萄酒样品，包括赤霞珠（CS, N=15），梅洛（M, N= 16）和黑比诺（PN, N=14）。这些葡萄酒样品具有不同的原产地（澳大利亚、保加利亚、捷克共和国、法国、德国、匈牙利、智利、意大利、马其顿、斯洛伐克、西班牙和美国），酿造年份（2004-2008年），是一个有代表性的样本集。拔掉酒瓶塞后，取等量的葡萄酒置于2 mL自动进样器琥珀色样品瓶中（充至满容量）。样品在LC/MS分析前于2°C暗处存放。随机取样本进行分析。

仪器

本研究采用Agilent 1200 SL系列快速高分离度LC系统与配备安捷伦喷射流技术电喷雾电离源（ESI）的Agilent 6530精确质量Q-TOF MS系统。仪器条件见表1。

为了保证所记录离子预期的质量准确度，分析中需要不断进行内部校准，在正离子模式下采用 m/z 121.0509（质子化嘌呤）和 m/z 922.0098（质子化六（1H, 1H, 3H -四氟丙氧基）磷嗪或HP-921）信号；在负离子模式下，采用 m/z 119.0362（去质子嘌呤）和 m/z 980.0164（HP-921醋酸加合物）。

表 1. LC和MS 条件

LC 运行条件	
色谱柱	Agilent ZORBAX Eclipse Plus C-18, 2.1 × 100 mm, 1.8 µm (部件号 959931-902)
柱温	40 °C
进样量	2 µL
自动进样器温度	4 °C
进样针清洗	冲洗口 (50:25:25 H ₂ O, IPA:MeOH:H ₂ O, 5 sec)
流动相	正离子模式: A = 5 mM 甲酸铵 B = 甲醇 负离子模式: A = 5 mM 甲酸铵 B = 甲醇
流速	梯度运行时0.3 mL/min 柱平衡时0.5 mL/min
梯度程序	B = 5%-65%, 0-13 min B = 65%-100%, 13-16 min B = 100%, 16-20 min B = 5%, 20-24 min (柱平衡)
分析时间	20 min
MS 条件	
离子模式	正离子和负离子, ESI+APCI 多模式电离
干燥气温度	300 °C
气化温度	170 °C
干燥气流速	11 L/min
雾化器压力	40 psi
毛细管电压	正离子模式4500 V 负离子模式3000 V
锥孔电压	65 V
八极杆 DC1	47 V
八极杆 RF	750 V
碎裂电压	125 V
谱图采集速率	每秒1.4张谱图
MS/MS 条件	
离子模式	正离子
分离窗	4 amu
谱图采集速率	每秒1.4张谱图

数据处理和统计分析

MassHunter工作站软件, 包括Qualitative Analysis (version 3.01) 和Mass Profiler Professional (version B.02.00) 模块, 可以用来处理原始的MS和MS/MS数据, 包括分子特征提取、背景扣除、数据筛选和采用ANOVA和PCA进行统计分析, 然后进行分类预测模型构建、分子式推断和数据库检索。

为了对来源于背景噪音的分子特征 (MF) 进行扣除, 在相同的仪器设置下分析了一个空白样品 (去离子水)。将每份样品已扣除背景的数据创建为化合物交换格式 (CEF文件) 导入Mass Profiler Professional (MPP) 软件包作进一步处理。

结果与讨论

葡萄酒分析

该代谢组学研究 (图1) 首先分析了来自世界各地的总共45种红葡萄酒样品, 基于生产时采用的葡萄品种不同, 这些酒可以分为三类: 赤霞珠 (15)、梅洛 (16) 和黑比诺 (14)。为了避免对样品中存在的代谢产物产生任何可能的影响, 在样品进行LC/MS分析前, 我们没有对样品进行任何前处理, 如提取或稀释, 而是直接将葡萄酒样品注射到分离柱上。由于这是一个非靶向研究, 液相色谱分离和质谱分析采用一般性设置以使分析能够得到尽可能多的化合物。使用反相色谱和亲水作用液相色谱 (HILIC) 来分离葡萄酒中的多种化合物。此外, 安捷伦喷射流技术电喷雾和多模式离子源用于收集MS数据。多模式离子源可同时实现APCI和电喷雾电离。采用Agilent ZORBAX Eclipse Plus C18, 2.1 × 100, 1.8 µm粒径柱可得到最有价值的结果。

葡萄酒样品分析相当复杂, 会出现众多的色谱峰, 而每一个峰又含有多种化合物 (图2)。此外, 还需要扣除溶剂空白中的污染物峰, 以得到纯净的样品数据。然后再通过数据挖掘, 从这些有效结果中提取有意义的数据。

实验流程图

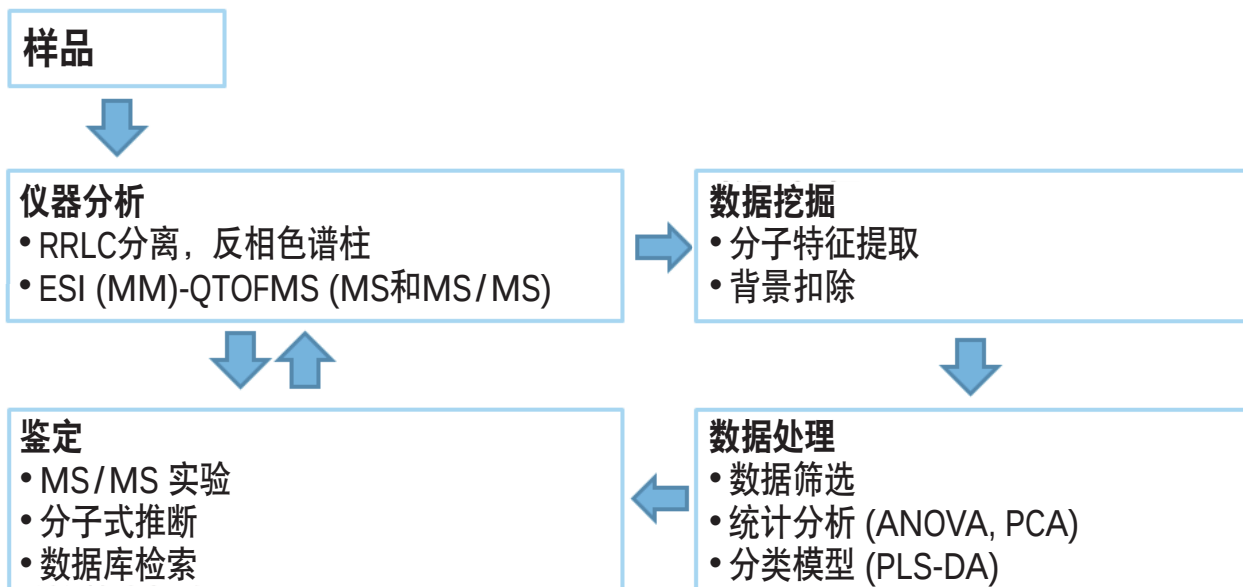


图1. 该代谢组学研究的工作流程, 目的是建立葡萄酒分类预测模型和鉴别标记化合物

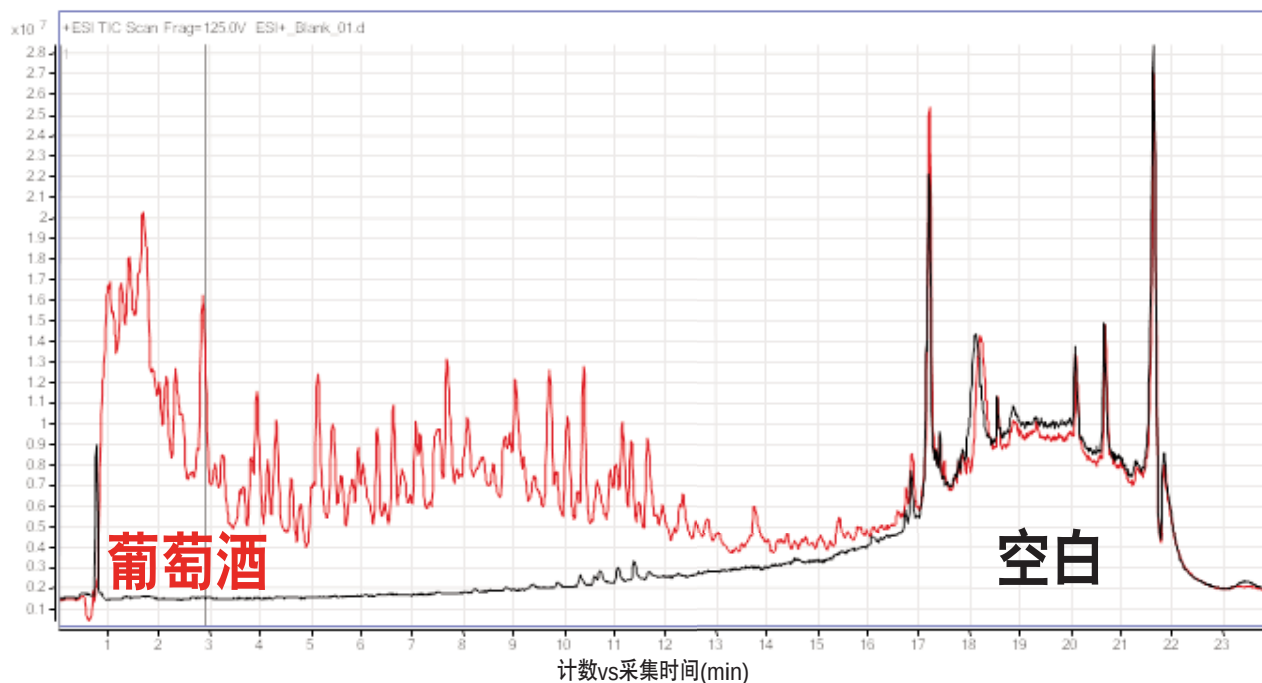


图2. 一个葡萄酒样品的总离子流 (TIC) 色谱图展现了它所含成分的复杂性。每个色谱峰又包含了多个组分, 必须扣除空白干扰以正确理解数据

数据挖掘

采用MassHunter工作站软件中的分子特征提取（MFE）算法来挖掘有意义的数据。该算法可以移除那些与持久的或缓慢变化的背景相关联的数据点，搜寻那些具有共同流出曲线的分子特征和属于一个或多个化合物（分子特征）的 m/z 值相关联的离子组群（对应同一个同位素簇中的多个峰，同一物质的不同加合物或电荷态）。分子特征提取也能扣除由流动相杂质造成的色谱背景干扰，提取溶剂空白中发现的所有化合物。提取的结果可作为所有样本数据文件的背景扣除数据集，从而在评价样品时扣除所用空白中所有化合物的影响。

分子特征提取的参数设置很重要，因为该算法针对不同类型的大、小分子提供不同的最优化方案。图3显示了该研究针对目标小分子所选择的参数设置。

将空白中发现的组分复制和粘贴到MFE标签栏，用于自动背景扣除（图4）。当所有数据处理方法的参数，包括适当的MFE算法和阈值计数（图3）设置完毕，即可将其保存，通过离线工具批处理所有的数据文件。数据批处理包括采用MFE寻找化合物，然后创建一个结果的化合物交换格式（.CEF）文件。这些结果将被保存在一个项目文件夹中，然后将被Mass Profiler Professional用来进行数据过滤和统计分析。

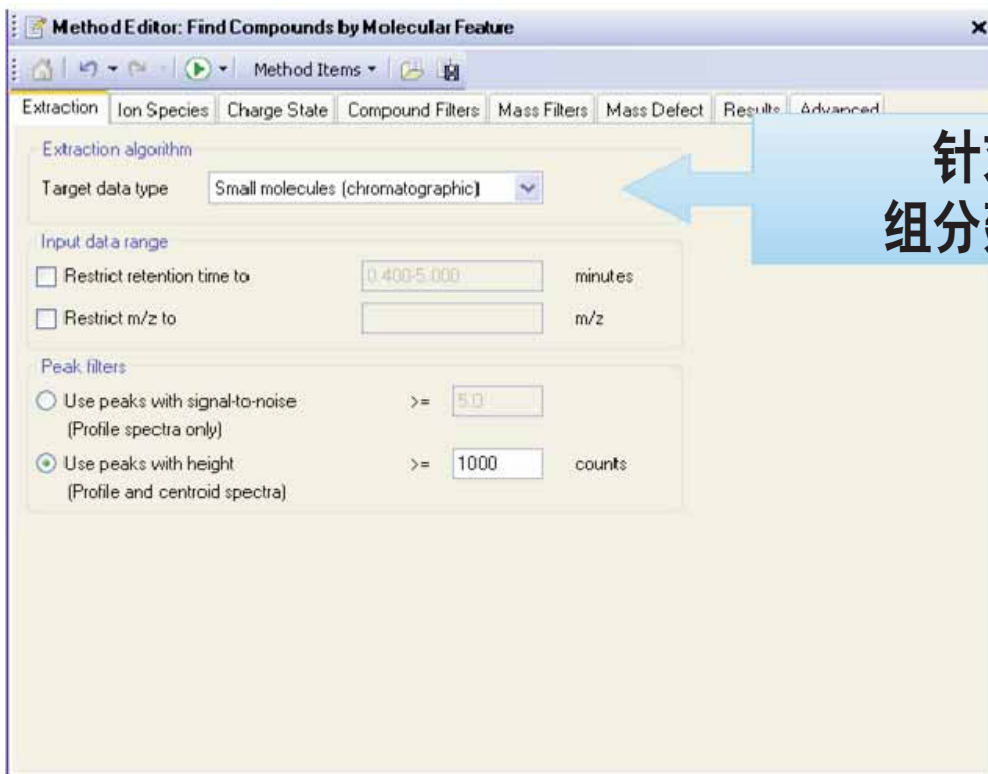


图3. 分子特征提取软件（MFE）所使用的设置，用于从LC/QTOF MS数据中提取实体（化合物）信息

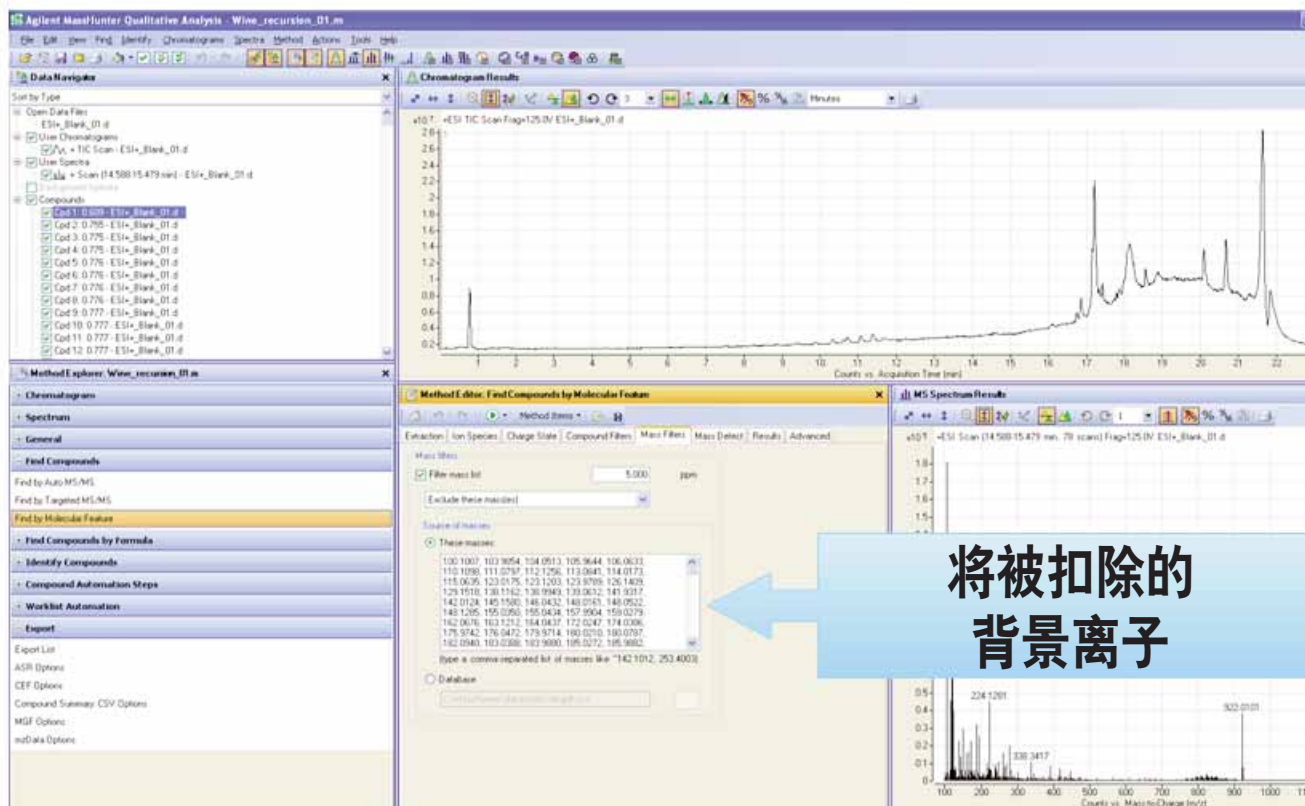


图4 将空白中发现的化合物复制并粘贴到MFE标签栏, 用于所有数据集的背景扣除

数据处理

当使用复杂的统计软件处理ASCII或文本类型表示的结果时, 代谢组学数据的数据处理往往会非常繁琐和费时。要有效利用代谢组学数据, 就需要能进行复杂的数据管理、筛选、统计分析、解释、创建模型和预测的工具, Mass Profiler Professional (MPP) 软件是一个理想的选择。它提供了一个易于遵循的指导工作流程, 帮助用户决定如何最好地去评估数据。专家级用户可以直接使用需要的数据处理方案 (详情见 Mass Profiler Professional 手册5990-4164CHCN)。

MPP使用八个步骤评估数据, 从一份总结报告开始, 揭示了该葡萄酒样品数据集中包含了超过20000个实体 (或可能的化合物) 信息。这些是通过离子簇的精确质量及其色谱匹配结果得到的。然后进行分类实验, 三类样品分别是赤霞珠 (CS)、梅洛 (M) 和黑比诺 (PN)。

数据筛选

组分筛选可以创建更高质量的数据集, 以使后续的多元分析结果更有意义。第一次筛选确定至少完全在一个组别中100%的时间 (频率分析) 都存在的实体 (化合物)。也就是说, 该化合物必须在该组所有的样本中。这种频率筛选把可能的标记物从20506个减至663个。不过, 设定频率为100%时, 一些重要的标记物也有可能被过滤掉。如15个黑比诺样品中14个均含有的一个化合物也被排除 (图5)。设置过滤频率为50%, 将允许保留更多的组分, 上述15个黑比诺样品中14个均含有的化合物将不再会被排除。使用50%频率过滤, 组分的数量从20506个减少到3600个。

下一步是进行方差分析 (ANOVA) 结果的筛选, 它决定了对于一个给定的化合物, 方差在什么水平被接受才有意义。使用概率值 $P=0.05$ (从一个样品到另一个样品的方差的显著性概率为95%), 频率筛选得到的3600个组分减少到40个有显著意义的重要化合物。

15个黑比诺样品中14个均含有该化合物。

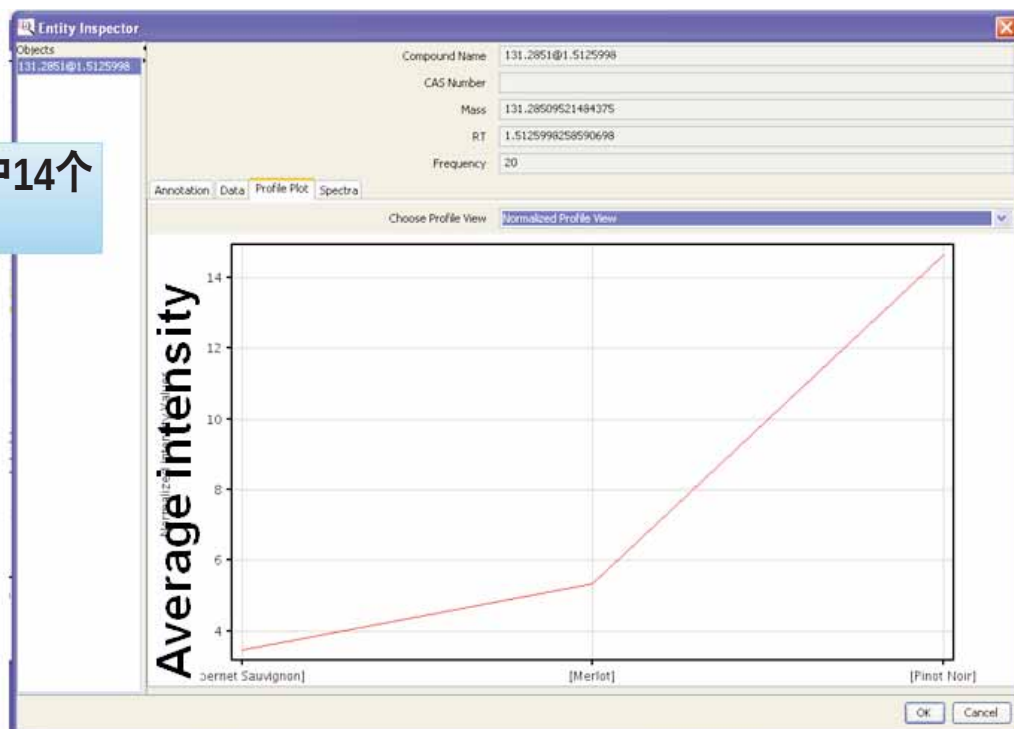


图 5. 使用100%的过滤频率可能会从数据集中排除有用的标记物，比如这种化合物虽在赤霞珠和梅洛样品中出现的频率很低，但在15个黑比诺样品的14个中存在

倍率分析是采用的最后一个数据筛选方法，以识别选定的数组之间丰度差异大的组分，即三组样品（CS，M和PN）中那些浓度相差2倍、3倍、4倍或更多的组分。如果采用高于两倍的倍率进行数据筛选，我们发现会漏掉赤霞珠和梅洛样品中可能的差异化化合物。因此，我们采用了2倍的倍率进行筛选（图6）。这使数据集中可能的化合物从40个减少到了26个。

数据处理的下一步是递归。递归允许重新检查数据，以确保每一个实体都是一个真实的峰，而不存在这些实体的样品就没有这些峰。MassHunter Qualitative Analysis软件能够自动重新从原始数据提取最终组的26个标记物生成提取离子色谱图（EIC）。对生成的EIC进行仔细检查，以消除假阳性（不是真实的峰）和假阴性（样品中存在真实的峰，但在分子特征提取和数据挖掘步骤中丢失了）。一旦26个实体被证实为真实的，即可采用主成分分析（PCA）对数据进行统计分析。

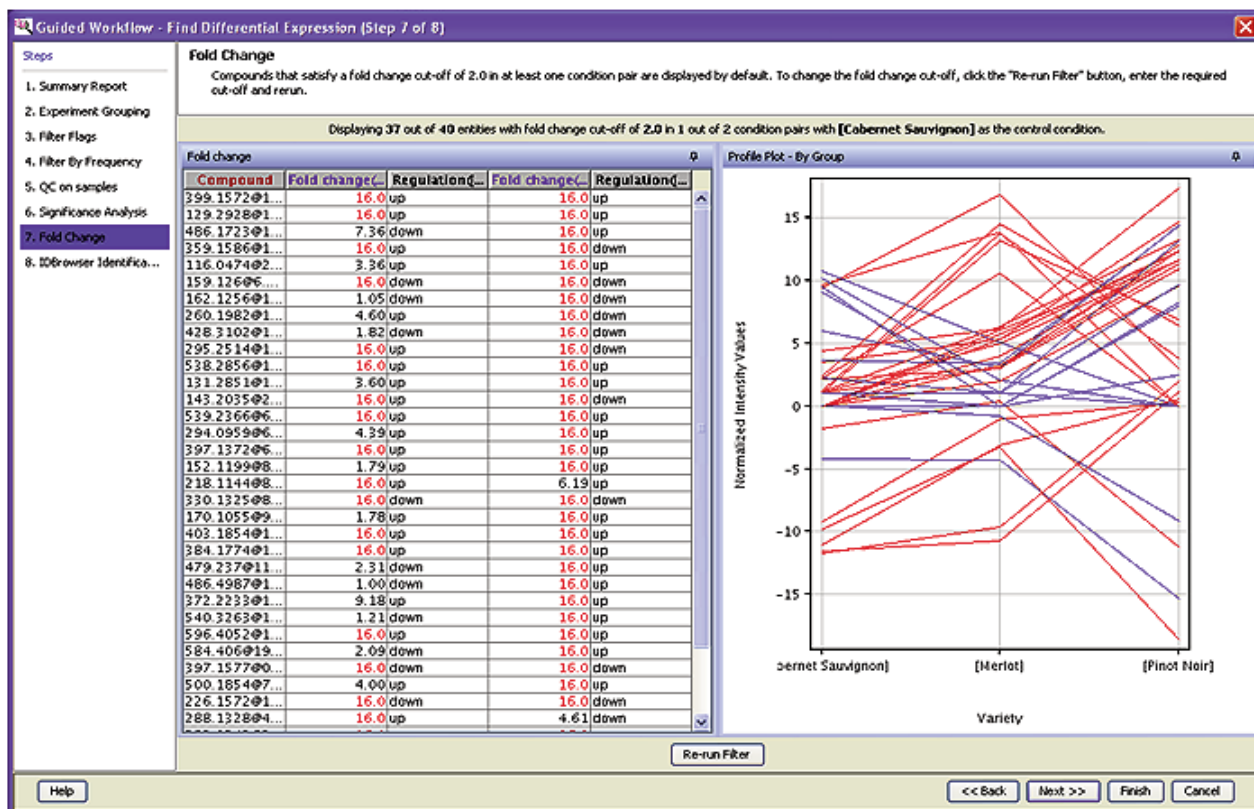


图6. 采用倍率分析识别选定组之间丰度有显著差异的实体。如果采用高于两倍的倍率进行数据筛选，我们会发现会漏掉赤霞珠和梅洛样品中可能的差异化合物。因此，我们采用了2倍的倍率进行筛选，这使数据集中可能的化合物从40个减少到了26个

统计分析

PCA是一种经常采用的自动多元分析技术，它可对数据进行降维处理，同时保留数据中的差异化信息。它是通过把测量的变量转化成正交的主成分，每个主成分又都是一个原始多变量的线性组合。我们的目标是识别不同类数据内可能存在的联系。对未经筛选的数据集甚至是经过频率过滤的数据集进行PCA分析，没有发现数据间的任何关联，因此无法用于葡萄酒的品种分类。但是，

对26个鉴定的标记物进行PCA分析，发现数据可以按照葡萄酒的类别明显聚为三组（图7）。注意，PCA分析并未故意制造葡萄酒种类间的统计差异，它只是揭示了这种客观存在的差异。通过对分子特征提取得到的实体进行频率过滤分析、方差分析，以及倍率分析，然后采用递归分析进行确认，我们选择出了能够用于葡萄品种鉴定的标记化合物。

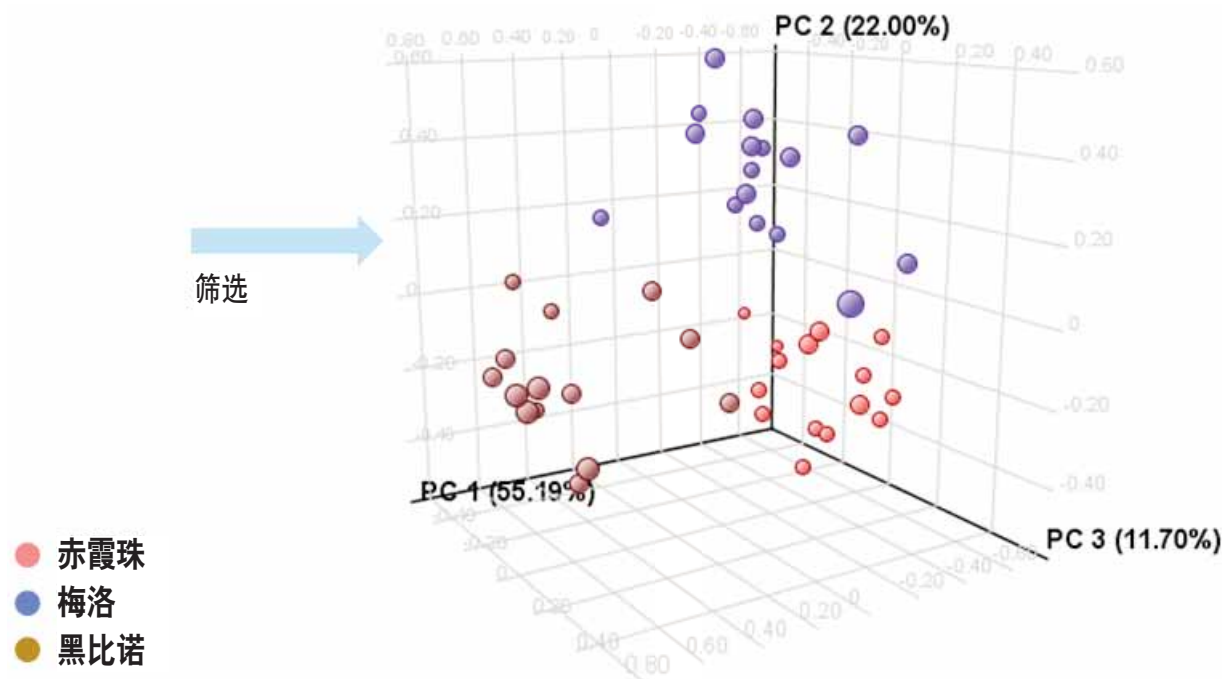


图 7. 对数据集通过方差分析和倍率分析筛选得到的26个标记物进行的主成分分析（PCA）。三类葡萄酒各自的标记物聚类良好，可用于预测未知葡萄酒的品种

分类模型

将Mass Profiler Professional软件筛选出的化合物数据分为三组，接下来就是要创建一个模型来预测葡萄酒的类别。MPP中的偏最小二乘判别分析（PLSDA）模型最适合分析这些质谱数据。建立分类模型的第一步就是要用数据对模型进行训练。

PLSDA算法产生一个混淆矩阵，表中的行表示真实的类别，列表示预测的类别。对角线上的元素表示正确分类的实验结果，交叉对角线上的元素表示错误分类的实验结果（见表2）。该表还列出了实验中针对某类葡萄酒以百分数表示的应用模型预测分类的准确度。可见，采用该训练集进行葡萄酒分类预测，准确度均为100%。

然后使用相同的数据对模型进行检验。虽然有些多余，但这是一个有效的统计过程。采用同样的分类预测模型验证其训练模型。注意，有两例样品错误地鉴定为梅洛酒，导致梅洛酒的预测准确度仅为87.5%，整体预测准确度为95.6%（见表2）。使用更多的样本进行模型训练，可通过提高模型的分析统计能力进而提高其预测准确度。

表 2. 使用PLSD模型预测分类结果的混淆矩阵

	赤霞珠 (CS)	梅洛 (M)	黑比诺 (PN)	准确度 (%)
模型训练				
CS	15	0	0	100.0
M	0	16	0	100.0
PN	0	0	14	100.0
识别率 (%)				100.0
模型验证				
CS	15	0	0	100.0
M	1	14	1	87.5
PN	0	0	14	100.0
预测率 (%)				95.6

为了证明该模型的预测能力，除了用于寻找标记物和建立预测模型的葡萄酒样品以外又购买了五种葡萄酒。还另外购买了一种葡萄酒，其类别对参与试验的科学家保密。通过分析这些酒并运用分类模型，五种葡萄酒样品以及未知酒样品的分类预测均正确无误（图8）。这些结果表明通过寻找标记物和使用预测模型准确鉴定葡萄酒类别的可行性。

Output views of classification		
Prediction Results		
Identifier	▼ Variety	Predicted(Model const)
ESI+_CS_Chile_3_RP: N...	UN	[CS]
ESI+_CS_Calif_4_RP: N...	UN	[CS]
ESI+_PN_Fran_8_RP: N...	UN	[PN]
ESI+_M_Calif_3_RP: No...	UN	[M]
ESI+_unknown_RP: Nor...	UN	[CS]
ESI+_PN_Fran_9_RP: N...	UN	[PN]

图 8. 使用预测模型对5种已知和1种未知种类葡萄酒的预测结果列表。所有样品均分类准确

葡萄酒标记物的鉴别

虽然没有必要知道用作葡萄酒标记物的化合物到底是什么，但是标记物标准品的使用会简化葡萄酒的品种判别测试。利用MS准分子离子及其同位素峰的精确质量数据，可获得每个标记化合物的分子式。最吻合此数据的那些分子式可用于检索相关私人 and 公共数据库。在这26种标记物中，只有一个从数据库中找到了结果。选定黑比诺最终筛选出的一个标记物 (m/z 449.1078, RT 11.16min)，通过其MS数据对其分子式进行估测和初步鉴定。然后使用这些结果来搜索PubChem数据库，推测该标记物为矢车菊素-3-O-葡萄糖苷，即花青素色素。使用Q-TOF分析，得到其MS/MS精确质量 m/z 449.1078，进一步表明该标记物即为花青素。如果化合物的其它信息完全未知，那么甚至精确的MS和MS/MS数据也将难以帮助我们获得该化合物的结构。标记物的最终确认需要该化合物的标准品。

结论

代谢组学研究在分析如葡萄酒等复杂食品时是一个有价值的工具。使用安捷伦LC/Q-TOF MS系统所产生的高准确度和重复性的数据，可以构建预测模型来确定葡萄酒的品种。这还可以推广应用到其它食品分析中。先进的软件工具如 MassHunter Qualitative Analysis和Molecular Profiler Professional，可以用来进行数据挖掘、筛选和统计分析，以确定特定食品品种（如葡萄酒）中的标记物，并使用这些标记物建立分类模型对食品进行分类。这些模型的预测可以非常准确，本文葡萄酒分类模型应用结果表明其总体准确度达到96.5%。虽然并不需要准确知道模型中的标记化合物到底是什么，但采用相同的仪器系统可生成MS/MS谱图数据，帮助我们对模型中的标记化合物进行鉴定。如果有推定化合物的标准品，该化合物即可得到确认。

参考文献

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).

更多信息

这些数据代表了典型的结果。如需了解更多有关我们产品和服务的信息，请访问我们的网站 www.agilent.com/chem/cn。

www.agilent.com/chem/cn

安捷伦对本资料可能存在的错误或由于提供、展示或使用本资料所造成的间接损失不承担任何责任。

本文中的信息、描述及技术规格等如有变更，恕不另行通知。

© 安捷伦科技（中国）有限公司，2011
中国印刷
2011年6月22日
5990-8451CHCN



Agilent Technologies