DOI: 10.13376/j.cbls/2015138

文章编号：1004-0374(2015)08-0995-05

植物代谢组学数据分析和数据库

申国安，段礼新，漆小泉*

(中国科学院植物研究所植物分子生理学重点实验室，北京 100093)

摘　要：植物代谢组学是近年发展起来的进行植物生物学研究的新兴学科，目前的难点和热点是数据分析和信息挖掘。现针对植物代谢组学数据分析和信息挖掘过程中涉及的几个主要方面进行了概述，内容包括数据特点、分析流程、软件和数据库。系统地介绍了植物代谢组数据分析预处理、常用统计分析原理和适用条件、数据分析中的注意事项、常用软件和数据库的优缺点，最后探讨了植物代谢组数据分析的问题和新发展动态。

关键词：植物；代谢组学；数据分析；统计方法；软件；数据库

中图分类号：Q94-3；R969.1　　**文献标志码：**A

Plant metabolome data mining and databases

SHEN Guo-An, DUAN Li-Xin, QI Xiao-Quan*

(Key Laboratory of Plant Molecular Physiology, Institute of Botany,
Chinese Academy of Sciences, Beijing 100093, China)

Abstract: Plant metabolomics is a newly emerging ‘omics’, which has been widely used in plant science research. Data analysis and data mining are the hotspots and the most difficult aspects of plant metabolomics research. In this review, we summarized the major aspects of data analysis and data mining, including the characteristics of metabolomics data, procedure of data analysis, major software packages and public databases. Also data pre-processing, the statistic principle and methods used in metabolomics data analysis, and the characteristics of major software and public databases were introduced in detail. Finally, we discussed the current situation and prospect of plant metabolomics data analysis.

Key words: plant; metabolomics; data mining; statistic; software; database

植物代谢组学的研究对象是植物细胞中所有的小分子代谢成分[1]。代谢组学使用高通量检测和数据处理手段，通过信息建模与系统整合[2-4]，研究在不同基因型和不同环境条件下代谢物整体的变化和差异。代谢组学与基因组学、转录组学、蛋白质组学一起构成了系统生物学。植物代谢组学是研究基因功能，阐明代谢途径及其调控机制，探索各种植物生物学现象的重要工具[5-7]。植物代谢组学广泛地应用于植物生理变化、基因功能、代谢调控网络等研究领域[3,6,8-11]。

随着各种分析仪器的精度和质量提高，高通量地收集代谢组数据已经不再是植物代谢组研究中的主要问题。随之而来的是如何处理海量数据，消除实验误差，从大量的数据中挖掘规律，获得有价值的信息，进而研究植物的生长发育和环境适应，揭示农作物高产、优质、高抗的代谢基础，这是目前植物代谢组学研究的难点和热点。

1　植物代谢组学数据的特点

相比于人、动物、细菌、真菌等生物，植物中的代谢物种类更多[7]，超过 20 万种[11-12]，特别是富含植物特有的次生代谢物，包括萜类、黄酮类、生物碱等[10]。每个样品一般都会检测出几百个，甚至上千个峰或代谢物[9]。另外，随着数据采集手段

收稿日期：201-01-04；**修回日期：**2015-02-05

基金项目：国家重点基础研究发展计划(“973”项目)(2013CB127005)；国家高技术研究发展计划(“863”计划)(2012AA10A304-3)；国家自然科学基金项目(31200227)

***通信作者：**E-mail: xqi@ibcas.ac.cn；Tel: 010-62836949

变得越来越便捷，实验人员常常会分析比以往更多的样品，这使得代谢组学需要处理的数据量越来越大，普通的分析方法难以适用，需要一些特殊的分析方法[13]。

代谢组学数据的变异程度非常高[14]，植物体内代谢物含量在不同个体和不同环境下显示巨大差异。此外，在样品的制备过程中，通常要经过很多步骤，操作中难免引入误差。仪器的性能在连续的样品分析过程中，也有很大的波动，如质谱在连续长时间的样品分析过程中，离子发生器上污垢积累越来越多，效能会逐渐降低，表现为检测出的样品信号强度梯度下降。在代谢组学研究过程中，需要保证高灵敏度、高通量、无偏性的要求，并减少背景的干扰[15]，这样才能发现真正有意义的差异或变化规律。

2 植物代谢组学数据分析流程

植物代谢组学研究的主要技术手段包括核磁共振 (NMR) 波谱、液相色谱 - 质谱 (LC/MS)、气相色谱 - 质谱 (GC/MS) 等[15-16]。无论使用哪种技术手段，最后得到的都是一系列色谱、质谱、化学位移的定量数据，样品之间的差异就由这些定量数据决定[17]。植物代谢组数据的基本分析流程包括数据的预处理、数据统计分析和数据库搜索比对等主要步骤。

2.1 数据预处理

样品制备、实验操作、仪器运行的波动常常造成随机误差。数据预处理可以消除噪音，减少误差，提高后续数据分析的准确性。数据预处理主要包括降低噪声、校正基线、归一化、数据标准化等[14,18-19]。

仪器在运行的过程中会因为电压不稳定等因素而引入随机噪声，所以，首先要对数据进行去噪和平滑处理。常用的平滑处理方法包括匹配滤波和移动窗平均滤波[19]。实际测量的谱图基线强度常常大于零，需要校正。常用的基线校准方法是将同一谱图的所有数据值减去最小值，使基线为零[19]。

代谢组数据经常会存在缺失的现象，如果不加以处理，必然影响数据分析结果。一般使用组内和全部样本的平均值代替空值[19]。

数据归一化 (normalization) 处理是把有量纲的数据变为无量纲形式，目的是为了不同单位和量级的数据可以相互比较。方法有很多种，最常用的一种是 0-1 标准化 (0-1 normalization)：将原始数据减去最小值，然后除以最大值与最小值的差值，通过这样的线性变换，把原始数据变为 (0, 1) 之间的小数。另外一种是 Z-score 标准化 (zero-mean normalization)：将原始数据减去平均值，然后除以标准差，经过这样变换的数据符合标准正态分布，即均值为 0，标准差为 1。除了这两种方法，实践中还经常使用代谢物的相对浓度，即每个代谢物除以样品的总浓度或者内标浓度，以此来校正个体差异或其他因素对代谢物绝对浓度的影响[14,19]。

数据中心化可以消除数据的常数偏移量，可以对坐标原点做变换，一般是均值中心化，即每个变量减去该变量的平均值[14,19]。

数据分析常常要求数据符合正态分布，如果不符合，可以采用适当的非线性数据变换，使得偏态分布的数据符合正态分布，常用的方法有对数转换和平方根反正弦转换[14,19]。对数转换还可以缩小数据的差异，使得数据更适合后续分析和图形化展示。

不同的预处理方法会对统计分析结果产生不同的影响，应该根据具体的研究目的、数据类型以及统计分析方法，选择适当的预处理方式[14]。

2.2 单变量分析

数据的差异可能是因为随机误差造成的，也可能的确是实验处理条件不同引起的。代谢组学分析首先需要对数据的差异进行显著性检验，以判断数据的变异是否由随机波动引起。代谢组学数据分析中常用 T 检验和 U 检验比较两组数据差别是否显著[20]。如果样本数较多，并且符合正态分布，一般使用 U 检验。当样本数比较少，变量符合正态分布，使用 T 检验[19]。当对多个样本代谢组数据之间的差异进行统计显著性分析时，需要使用方差分析[20]。方差分析需要满足两个条件：样本符合正态总体分布和各组样本的总体方差相等。整体随机性可由 D 法、W 法或卡方检验法验证，方差齐性可由 Barlett 或 Levene 法检验[19]。

如果两个样本总体分布相同，但是不确定是正态分布时，可以使用曼 - 惠特尼 U 检验 (Mann-Whitney U, 也称为 Wilcoxon 秩和检验) 考察两个总体是否差异显著。当不能确定两个总体是否相同时，应该使用 Z 检验或 Wald-Wolfowits 检验[19]。

2.2 多变量分析

多变量分析可以分为有监督的方法 (supervised method) 和无监督的方法 (unsupervised method) 两种[3,21]。无监督的方法直接对原始数据进行归类，发现数据的整体特点，以及各变量之间的内在关系，主要包括主成分分析 (principal component analysis, PCA)[22-23]、聚类分析 (hierarchical cluster analysis, HCA)[24]、自组织映射神经网络算法 (self-organizing

maps, SOMs)[3]。有监督的方法包括偏最小二乘法 (partial least square, PLS)[25]、线性判别分析法 (linear discrimination analysis, LDA)、支持向量机法 (support vector machine, SVM)[26]、人工神经元网络等(Artificial neural networks, ANN)[27]。

PCA 是代谢组学数据分析中使用最为广泛的一种方法 [28]。主成分分析将原始的多个变量通过线性变换为少数几个综合指标 (即主成分)，从而实现大规模复杂数据的特征提取和降维。PCA 的分析结果用得分图展示，每个点代表一个独立样本，分布点越靠近，说明这些样品中所含有的代谢物组成和浓度越接近 [17]。如果存在异常离散的数据点，需要考虑数据的质量是否存在一定的问题。PCA 分析还可以得到载荷因子图，它显示样品之间或者组间存在的主要差异代谢物，它们可作为生物标记物 [29]。

聚类分析法 (hierarchical clustering analysis, HCA) 也是一种无监督模式识别方法，常被用于代谢组学数据分析，它根据相似性把样本划分为不同的类 [9]。聚类的基础是计算样品两两之间的距离或相似性，距离的计算方法有欧氏距离、明氏距离和绝对值距离，相似性通常用夹角余弦和相似性系数表示 [19]。通常使用最短距离法 (nearest neighbor)、最长距离法 (furthest neighbor)、类间平均链锁法 (between-groups linkage)、类内平均链锁法 (within-groups linkage)、中间距离法 (median clustering)、重心法 (centroid clustering) 等计算类与类之间的距离 [18]。最短距离法适用于长条状或 S 形的嵌套聚类，最长距离法、中心法和类平均法适用于椭球形的类 [19]。根据距离计算方式和采用的聚类原则，聚类分析又分为很多方法，这些方法总的目的是使类内样本之间距离最小，类之间的样本距离相对较大。聚类分析过程通常包括以下步骤：数据收集，并且收集相应的变量；产生一个相似矩阵；决定把目标总体细分为几类，及其对每一种类别相应的定义；实施聚类分析，产生结果 [3]。聚类分析与其他分析方法，比如小波变换，联合用于代谢组分析，可以进一步提高分类正确率 [3]。

自组织映射神经网络算法 (SOM) 的基本原理是将多维数据当做几何学节点，较近的节点组成相邻的类，从而使多维的数据模式聚成二维节点的自组织映射图 [18]。SOM 数据输入顺序影响聚类结果，改进的批次自组织映射图法 (BL-SOM) 克服了这个缺点。SOM 分析适合代谢组数据与其他组学数据的整合分析 [31]。

偏最小二乘判别分析 (PLS-discriminant analysis, PLS-DA) 是一种有监督的模式识别方法。在实际情况下，往往已经知道某些样品应该归为一类，将这些信息整合到模式识别算法中，可以明显改善分析结果。偏最小二乘判别分析计算与 PLS 相似，只是在分析时对样品强行分组，有利于发现组间的异同 [17]。偏最小二乘判别分析是代谢组学数据分析的一种常用方法 [14]，集中了多元线性回归分析、主成分分析、典型相关分析等优点。

3 植物代谢组学数据分析软件

目前代谢组分析的软件非常多，有免费开源的，也有商业收费的，这些软件各有优缺点。在实际工作中，很难使用单一的软件完成所有的分析，一般会同时使用几种功能互补的软件，以满足复杂分析的需要，最常用的软件有以下几种。

XCMS 是一个开源免费的 R 软件包，主要用于 LC-MS 数据的处理，也可以处理 GC-MS 数据，是目前使用最广、引用文献最多的一个软件。XCMS 接受 NetCDF、mzXML、mzData 等数据格式。XCMS 的特点是参数非常多，这使得分析工作相对灵活，但同时也给初学者带来了困难。XCMS 现在已经发布了基于云计算的在线版本 (http://xcmsonline. scripps. edu)，不需要下载安装，可以直接把数据提交到该网站，并在线查看分析结果。

AMIDS (automated mass spectral deconvolution and identification system，http://www.amdis.net) 是一个优秀的 GC-MS 数据分析软件。不足之处是分析的时候需要标准代谢物数据库，能批量解卷积数据，但是不能多样本对齐。

MET-IDEA (metabolomics ion-based data extraction algorithm) 软件是美国 Noble Foundation 为了进行植物代谢组研究而开发的一个免费数据分析软件，界面友好，操作容易。该软件能够快速对大批量的原始质谱数据进行提取，转换为半定量数据，克服了代谢组学研究中的一个难题 [32]。虽然软件的分析效果还不错，但是没有解卷积功能。与 XCMS 类似，只是使用单个离子作为代谢物的代表，没有利用整个质谱图数据信息。

除了以上这些优秀的免费软件，一些仪器公司开发的商业软件也使用非常广泛，如沃特世 (Waters) 公司的 MassLynx 软件、力可 (Leco) 公司的 Chroma TOF 软件、珀金埃尔默 (Perkin Elmer) 公司的 Turbomass 软件等。另外，常用植物代谢组数据统计分析软件还有 SAS、SPSS、Matlab、R 等。

4 植物代谢组学相关数据库

代谢组数据的分析离不开各种数据库，常用的代谢组学数据库包括参考谱图数据库，如 MassBank、METLIN、NIST 等；代谢途径数据库，如 KEGG、PlantCyc、MetaCye 等；化合物信息数据库，如 PubChem、ChemSpider 等；代谢组学实验信息管理型数据库，如 SetupX、Sesame LIMS 等 [19,33]。其中最著名的包括以下几个。

美国国家标准与技术研究院 (The National Institute of Standards and Technology, NIST) 化学数据库是一个使用非常广泛的数据库，免费提供各种化合物化学和物理性质的查询 (http://srdata.nist.gov/gateway/)，可通过名字、分子式、CAS 登录号、相对分子质量等查找。数据库包括 79.6 万张质谱图，包含 66.7 万个化合物，其中 74.6 万张谱图带有化学结构 [19]。自动批量查询 NIST 数据库和数据采集软件已经被成功开发出来，解决了 NIST 不能批量查询，及查询后有效数据的自动采集问题 [34]。

MassBank (http://www.massbank.jp/) 是一个小分子化合物质谱数据库。数据库包含 2 337 种标准代谢物，10 286 种挥发性的天然和人工合成化合物，以及 679 种合成药物的各种质谱图，对化合物鉴定非常有用 [35]。使用者可以查询全部数据库，也可以限定条件查询指定的数据库。MassBank 在进行质谱搜索的时候，相似性使用权重余弦相关系数计算，其中权重值根据质荷比和峰强度进行了优化。另外，MassBank 合并了相同化合物在不同分析条件下获得的质谱数据，这样可以显著改善识别的准确性 [35]。

METLIN (http://metlin.scripps.edu/index.php) 代谢组数据库收集了大量的代谢物信息以及串联质谱数据。代谢物信息包括质谱、化学式、结构等，而且与 KEGG 数据直接链接。METLIN 包含 1 万多种代谢物的串联质谱信息，5 万多张高精度 ESI-QTOF MS/MS 信息，超过 16 万种预测的碎片结构，是世界上最大的数据库，而且数据还在不断增加。研究人员可以通过精确相对分子质量、单个或多个碎片、中性丢失、全谱进行搜索。除了手动搜索，还可以通过 XCMS online 自动搜索，如果没有精确的匹配，系统会返回近似的结果。

KEGG (Kyoto Encyclopedia of Genes and Genomes, www.genome.jp/kegg/) 是一个综合的生物信息学研究平台。KEGG 的 PATHWAY 数据库集合了各种代谢通路，主要用于查询生物代谢物分子的相互作用和反应网络。KEGG 的化学信息类数据库包含了各种生物代谢物和反应的知识。

PlantCyc 9.5 (http://www.plantcyc.org) 数据库目前提供超过 350 种植物共有或特有的 1 117 条代谢通路信息，包含代谢通路、催化的酶和基因，以及各种植物代谢物，这些信息都来源于文献报道，并经过人工校正。PlantCyc 还整合了各种植物代谢通路数据库，包括 MetaCyc 数据库中所有的植物代谢通路。

5 展望

植物代谢组学仍然是一个快速发展的领域，新的研究方法和应用不断出现，如近年迅速发展的 mQTL[1,36-37] 和 mGWAS[38-40] 分析，就是把代谢组学分析与其他生物学分析手段相结合，拓展了植物代谢组学研究的领域和深度。植物代谢组学与其他分析手段的结合，将会把植物研究提升到一个前所未有的深度，为更深入研究植物代谢途径和关键调控位点，进一步开发利用植物重要功能代谢物和营养成分，提供一种全新的宏观研究方法 [6,41-44]。将代谢组学跟多种组学系统地结合起来，相互验证，从整体的角度全面理解植物代谢，将是植物代谢组学发展的方向 [3,6,33]。

目前的分析软件功能仍然需要完善，能够同时完成数据解卷积、定量定性、多样本对齐、统计分析、数据库搜索、图形化展示的软件是从事代谢组学研究人员的梦想。虽然分析软件很多，但是功能多偏重于一个方面，研究人员必须使用多个软件才能完成整个分析，能够同时满足多种分析需要的软件很少。还有就是分析软件常常不能跟数据库查询自动对接，这方面也需要加强。

植物代谢组学研究的一个瓶颈是代谢物成分的鉴定。在植物代谢组学研究中，对未知化合物的结构解析一直是一个非常重要但也非常棘手的问题 [11]。虽然有非常多的信息，但是因为缺乏化学标准品，大量的植物代谢物结构仍然没有被鉴定，已知结构的化合物只占很少的比例。加强代谢物实体库的分离，系统鉴定植物中的每一个代谢物，是彻底解决这个问题的重要途径。

[参 考 文 献]

[1] 段礼新, 解丽霞, 薛震, 等. 植物代谢组学分析平台建立及其应用[C]. 杨凌: 全国植物生物学大会, 2012

[2] Dettmer K, Hammock BD. Metabolomics-a new exciting field within the "omics" sciences. Environ Health Perspect, 2004, 112(7): 396-7

[3] 尹恒, 李曙光, 白雪芳, 等. 植物代谢组学的研究方法及其应用. 植物学通报, 2005, 22(5): 532-40
[4] 杨军, 宋硕林, Jose CP, 等. 代谢组学及其应用. 生物工程学报, 2005, 21(1): 1-5
[5] Fiehn O, Kopka J, Dörmann P, et al. Metabolite profiling for plant functional genomics. Nat Biotechnol, 2000, 18(11): 1157-61
[6] 淡墨, 高先富, 谢国祥, 等. 代谢组学在植物代谢研究中的应用. 中国中药杂志, 2007, 32(22): 2337-41
[7] Saito K, Matsuda F. Metabolomics for functional genomics, systems biology, and biotechnology. Annu Rev Plant Biol, 2010, 61: 463-89
[8] 邱德有, 黄璐琦. 代谢组学研究: 功能基因组学研究的重要组成部分. 分子植物育种, 2004, 2(2): 165-77
[9] 高敏, Saxena PK, 刘春朝. 植物代谢组学研究进展. 西北植物学报, 2005, 25(2): 405-12
[10] 王莉, 张艳霞, 史玲玲, 等. 功能基因组学和代谢组学技术在植物次生代谢物合成及调控研究中的应用. 北京林业大学学报, 2007, 29(5): 153-9
[11] 张凤霞, 王国栋. 植物代谢组学应用研究: 现状与展望. 中国农业科技导报, 2013, 15(2): 28-32
[12] Goodacre R, Vaidyanathan S, Dunn WB, et al. Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol, 2004, 22(5): 245-52
[13] 杨辉华, 任洪军, 李灵巧, 等. 基于Sector/Sphere 平台的GC-MS多样本并行对齐算法实现. 计算机应用, 2013, 33(1): 215-8
[14] 柯朝甫, 张涛, 武晓岩, 等. 代谢组学数据分析的统计学方法. 中国卫生统计, 2014, 31(2): 357-9
[15] 周秋香, 余晓斌, 涂国全, 等. 代谢组学研究进展及其应用. 生物技术通报, 2013, 1: 49-55
[16] Okazaki Y, Saito K. Recent advances of metabolomics in plant biotechnology. Plant Biotechnol Rep, 2012, 6(1): 1-15
[17] 阿基业. 代谢组学数据处理方法: 主成分分析. 中国临床药理学与治疗学, 2010, 15(5): 481-9
[18] 焦旭雯, 赵树进. 药用植物代谢组学的研究进展. 广东药学院学报, 2007, 23(2): 228-31
[19] 漆小泉, 王玉兰, 陈晓亚. 植物代谢组学: 方法与应用[M]. 北京：化学工业出版社, 2011: 120-81
[20] Fiehn O. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. Comp Funct Genomics, 2001, 2(3): 155-68
[21] 许国旺, 杨军. 代谢组学及其研究进展. 色谱, 2003, 21(4): 316-20
[22] Fiehn O. Metabolomics-the link between genotypes and phenotypes. Plant Mol Biol, 2002, 48: 155-71
[23] Wei J, Xie GX, Zhou ZT, et al. Salivary metabolite signatures of oral cancer and leukoplakia. Int J Cancer, 2011, 129(9): 2207-17
[24] Keun HC, Bbels TMD, Bollard ME, et al. Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. Chem Res Toxicol, 2004, 17(5): 579-87
[25] Wang C, Kong HW, Guan YF, et al. Plasma phospholipid metabolic profiling and biomarkers of type 2 diabetes mellitus based on high performance liquid chromatography/electrospray mass spectrometry and multivariate statistical analysis. Anal Chem, 2005, 77(13): 4108-16
[26] Trygg J, Holmes E, Lundstedt T. Chemometrics in metabonomics. J Pmteome Res, 2007, 6(2): 469-79
[27] Lindon JC, Holmes E, Nicholson JK. Pattern recognition methods and applications in biomedical magnetic resonance. Prog Nucl Magn Reson Spectrosc, 2001, 39: 1-40
[28] Ciosek P, Brzózka Z, Wróblewski W, et al. Direct and two-stage data analysis procedures based on PCA, PLS-DA and ANN for ISE-based electronic tongue-Effect of supervised feature extraction. Talanta, 2005, 67(3): 590-6
[29] 郭宾, 戴仁科. 代谢组学及其研究策略和分析方法进展. 中国卫生检验杂志, 2007, 17(3): 554-63
[30] 夏金梅, 吴晓建, 元英进. 代谢物组学信息挖掘WT-HCA方法. 化工学报, 2007, 58(7): 1783-91
[31] Hirai MY, Yano M, Goodenowe DB, et al. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. Proc Natl Acad Sci USA, 2004, 101: 10205-10
[32] Broeckling CD, Reddy IR, Duran AL, et al. MET-IDEA: data extraction tool for mass spectrometry-based metabolomics. Anal Chem, 2006, 78(13): 4334-41
[33] 周国艳, 胡望雄, 徐建红, 等. 整合多个组学(omics)分析植物代谢产物及其功能. 浙江大学学报: 农业与生命科学版, 2013, 39(3): 237-45
[34] 谢为博, 邓克俭. NIST Chemistry Webbooks数据库的批量查询和数据采集. 计算机与应用化学, 2004, 21(2): 314-16
[35] Horai H, Arita M, Kanaya S, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom, 2010, 45(7): 703-14
[36] Matsuda F, Okazaki Y, Oikawa A, et al. Dissection of genotype-phenotype associations in rice grains using metabolome quantitative trait loci analysis. Plant J, 2012, 70(4): 624-36
[37] Hill CB, Taylor JD, Edwards J, et al. Whole-genome mapping of agronomic and metabolic traits to identify novel quantitative trait loci in bread wheat grown in a water-limited environment. Plant Physiol, 2013, 162(3): 1266-81
[38] Riedelsheimer C, Lisec J, Czedik-Eysenberg A, et al. Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc Natl Acad Sci USA, 2012, 109(23): 8872-7
[39] Chen W, Gao YQ, Xie WB, et al. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet, 2014, 46(7): 714-21
[40] Li HH, Liu J, Liu HJ, et al. Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat Commun, 2014, 5: 3438
[41] Hall R, Beale M, Fiehn O, et al. Plant metabolomics: the missing link in functional genomics strategies. Plant Cell, 2002, 14(7): 1437-40
[42] Hirai MY, Klein M, Fujikawa Y, et al. Elucidation of gene-to-gene and metabolite-to-gene networks in *Arabidopsis* by integration of metabolomics and transcriptomics. J Biol Chem, 2005, 280(27): 25590-95
[43] Hirai MY, Sugiyama K, Sawada Y, et al. Omics-based identification of *Arabidopsis* myb transcription factor regulating aliphatic glucosinolate biosynthesis. Proc Natl Acad Sci USA, 2007, 104(15): 6478-83
[44] Saito K, Matsuda F. Metabolomics for functional genomics, systems biology and biotechnology. Annu Rev Plant Biol, 2010, 61: 463-89