

综 述

代谢组学分析及数据处理技术

卢红梅, 梁逸曾

(中南大学 化学化工学院, 湖南 长沙 410083)

摘 要: 用于代谢组学数据获取的分析技术主要包括 NMR、FT-IR、MS 及其与色谱等联用技术。这些分析手段正在使代谢组学产生海量数据。处理、分析和整合这些数据需要专门的数学、统计和信息学知识和工具。主要综述了代谢组学中所采用的分析手段、由其所获取的多维代谢组学原始数据的处理方法, 代谢组学所采用的数据标准及数据库技术、统计分析方法, 以及相关方面的发展需求。

关键词: 代谢组学; 分析技术; 数据处理

中图分类号: G353.11 **文献标识码:** A **文章编号:** 1004-4957(2008)03-0325-08

The Development of Analytical Technologies and Data Mining in Metabolomics

LU Hongmei, LIANG Yizeng

(Chemistry & Chemical Engineering College, Central South University, Changsha 410083, China)

Abstract: Metabolomics refers to the global analysis of cellular metabolites with various analytical technologies. The most widely applied technologies in metabolomics include NMR, FT-IR, MS and hyphenated chromatography. Like other 'omics' research, metabolomics generates a mass of data. The handling, processing and analyzing of these data are challenging, and specialized mathematical, statistical and informatics tools are required. The major analytical technologies, the processing methods for the raw data, the standards, database technologies and statistical methods used in metabolomics and the development are reviewed in this paper with 116 references.

Key words: metabolomics; analytical technology; data processing

代谢组学是对生物或细胞内所有低相对分子质量代谢产物同时进行定性和定量分析的一门新学科^[1-2]。它是以组群指标分析为基础, 以高通量检测 and 数据处理为手段, 以生物标记物 (biomarkers) 和生物标记模式 (biomarker pattern) 的发现、信息建模与系统整合为目标的系统生物学的一个分支。近 10 年来, 在疾病标记物研究、药物开发、毒性评价、营养、植物和微生物代谢网络和代谢工程研究、环保等应用方面已经取得了一定的成绩^[3-6]。其现有研究主要分为 3 个层次:

(1) 代谢指纹分析 (metabolic fingerprinting): 代谢指纹分析以整个图谱 (指纹图谱) 代表特定细胞或组织的特定代谢模式, 不分离鉴定具体单一组分, 是一种快速、高通量、全局分析手段^[7-8]。其图谱通常是直接通过光谱技术, 比如核磁共振 (NMR)^[9-11]、傅立叶变换红外光谱 (FT-IR)^[12-13]、傅立叶变换离子回旋共振质谱仪 (FTICR-MS) 或质谱 (MS)^[7] 分析获得, 因此样本制备相对简单。它通常结合模式识别和判别技术用来分类代谢指纹和识别不同模式指纹的某些特征, 是在不同状态判别筛选、疾病诊断及其特定代谢模式的发现中最有用的方法^[3, 14-20]。

(2) 代谢物谱图分析 (metabolite profiling): 代谢物谱图分析是对数种预设的靶标进行定量分析, 如与某些代谢路径有关的特定代谢物组^[4]。它在新药研发中应用广泛。最近研究者认为细胞内的代谢物谱图有帮助识别基因的功能, 尤其是应用于那些没有明显显型的突变^[19, 21-23], 大量突变代谢物谱图的研究积累成果可能导致潜在代谢网络的发现^[24]。

(3) 靶标分析 (targeted analysis): 靶标分析用于定性定量分析少数与特定代谢反应相关的代谢物,

收稿日期: 2007-04-02; 修回日期: 2007-07-18

基金项目: 国家自然科学基金资助项目 (20745005); 湖南省自然科学基金资助项目 (07JJ4004)

第一作者: 卢红梅 (1976-), 女, 湖南溆浦人, 讲师, 博士, Tel: 0731-8830824, E-mail: hongmeilu@mail.csu.edu.cn

其分析的前提是必须知道目标代谢物的结构和合适的分析方法,是一种真正的定量方法^[8, 25]。它需要严格的样本制备和分离。靶标分析最大的局限性在于需要事先已知分析目标,并且可获得该目标化合物纯标准品。

1 分析技术

大量分析技术用于代谢组学的分析研究^[8, 26-28],主要包括 NMR、MS及 MS联用技术和 FT-IR等。

1.1 质谱 (MS)及其联用技术

由于具有快速、高灵敏度、选择性定性定量、可识别代谢物和可测量多种代谢物等性质,MS已经在代谢组学研究中成为首选技术^[8, 29-31]。

1.1.1 气相色谱-质谱 (GC-MS) 采用 GC-MS可以同时测定几百个化学性质不同的化合物,包括有机酸、大多数氨基酸、糖、糖醇、芳胺和脂肪酸^[32],该分析技术被专家称为最宝贵的分析手段。尤其是最近发展起来的二维 GC (GC×GC)-MS技术,由于其具有分辨率高、峰容量大、灵敏度高及分析时间短等优势,而备受代谢组学研究者的青睐^[27]。另外,GC-MS最大的优势是有大量可检索的质谱库^[31]。

1.1.2 液相色谱-质谱 (LC-MS) 相对于 GC-MS,LC³/MS能分析更高极性和更高相对分子质量的化合物^[33]。LC-MS的一个很大的优势是,它大多数情况下不需要对非挥发性代谢物进行化学衍生。过去 10 年,LC³/MS技术中的软电离方式使得质谱仪更加完善和稳健,尤其是最近几年,整体式毛细管柱和超高效液相色谱 (UPLC)^[34-37]在分离科学中的应用为复杂的生物混合物提供了更好的分离能力;另一方面,现代离子阱多级质谱仪的发展使 LC-MS可提供未知化合物的结构解析信息^[38],因此 LC-MS在代谢组学研究中越来越受青睐。

1.1.3 毛细管电泳-质谱 (CE-MS) 相对其它分离技术,CE-MS具有几个重要的优势:高的分离效率、微量样本量 (平均注射体积 1~20 nL)以及很短的分析时间。CE被用于代谢物的目标和非目标分析^[39-41],包括分析无机离子、有机酸、氨基酸、核苷及核苷酸、维生素、硫醇、糖类和肽类等。CE-MS的最大优点是它可在单次分析实验中分离阴离子、阳离子和中性分子,因此 CE可以同时获得不同类代谢物的谱图^[42-45]。这使得它成为高通量非目标分析代谢组学研究中一个很有吸引力和发展前景的分析技术。

1.2 核磁共振 (NMR)

NMR非常适合于代谢组学研究,因为它是唯一既能定性,同时又能在微摩尔范围定量大量有机化合物的技术^[6, 46-50]。NMR是非破坏性的,样本可以再用于进一步的分析。NMR的样本制备简单且易自动化。但是复杂混合物的 NMR谱的解析非常困难^[6, 46, 51]。对全面的代谢谱图分析,NMR最大的缺陷是灵敏度相对较低,从而使得它不适合分析大量的低浓度代谢物。

1.3 傅立叶变换红外光谱 (FT-IR)

FT-IR因基本不需要样本的制备、分析时间短、非破坏性等特点而被广泛用于高通量筛选和生物样本的整体分析^[52]。大量研究表明,尽管 FT-IR不如质谱灵敏,但是作为代谢指纹分析手段在快速检测和疾病诊断等方面具有很好的潜力^[53]。当然,FT-IR也有一些缺陷,如水吸收峰很强,必须通过样本的脱水或数据处理等方法去除。

以上所列分析技术只是代谢组学分析中一些常用技术,另外还有薄层色谱^[54]、高效液相色谱/二极管阵列 (PDA)^[55]或者电化学检测器联用^[56-57]、表型微阵列^[58]和各种蛋白阵列。每种技术都有其优缺点,多项技术或多种检测器的在线组合或平行联用可显著增强单个生物样本中不同极性、不同相对分子质量代谢物的分析,以及降低定量下限。

2 数据处理

正如其它组学研究一样,代谢组学也正在产生海量数据^[26]。处理、分析和管理这些多维的大数据集需要专门的数学、统计和信息学工具^[59-60]。目前,市场上尚无满足代谢组学数据处理分析需求的软件,可供选择的基本还是一些手动算法或半自动的软件,并且大部分处于研究初期。代谢组学数据处

理主要包括以下几个方面:原始数据的处理、数据和信息的管理、统计分析、数据标准化、代谢物及路径识别、数据集成和代谢网络的数学模拟。

代谢组学数据分析处理的困境已经引起了国际上代谢组学研究专家们的一致重视,目前,世界各地的代谢组学专家对如何为代谢组(从植物、微生物到人体样本)学中的数据处理与分析(生物信息学问题)找寻合适的方法和工具进行了探讨。研究与讨论的焦点集中于以下3个方面:

2.1 原始数据处理

在原始数据处理方面,基于不同分析技术所产生的数据分别出现了一系列的相关软件和算法。

2.1.1 质谱及其联用数据处理 这类软件主要是针对 GC-MS、LC-MS数据提出来的。通常情况下,处理一套原始代谢组学质谱及其联用数据包括除噪、色谱解析、峰检测、色谱匹配、化合物定性定量。该步骤已经成为代谢组学质谱及其联用数据分析中最具挑战性和最耗时的步骤,急需自动处理方法。

就 GC-MS数据处理而言,近年来出现的一些相应软件主要有两类,一类是由仪器商提供的软件,如 Leco公司的 ChromaTOF(<http://www.leco.com>)和 Ion Signature Technology公司的 IST for GC/MS(<http://www.ionsigtech.com>)等。这类软件对数据格式要求很严格,通常不能处理其它仪器产生的数据。另一类是独立于仪器的软件,这些包括 AMDIS(<http://chemdata.nist.gov/mass-spc/amdis/>)和 AnalyzerPro(<http://www.spectralworks.com>)等。这类软件能读取通用的 NetCDF、CSV、Excel或一些通用仪器的数据。本实验室采用标准品混合物的 GC-MS数据对其中一些软件进行评价,结果发现所有的这些自动软件在代谢物的识别上两类错误(false positive和false negative)出现的机率很高,解析所得的质谱的准确度很低,解析重复性不够而且不能识别低浓度代谢物^[61]。另外,Styczynski等^[62]最近提出的 SpectConnect(<http://spectconnect.mit.edu/>)可以在没有质谱库的情况下,系统、自动地进行代谢物的检测及生物标记物的发现,但是该软件要求分析必须具备重复样本。

类似于 GC-MS的数据处理,LC-MS数据处理软件主要分两类,一类是由仪器商提供的软件,如 Waters公司开发的 MarkerLynx软件(<http://www.waters.com/>)具有原始数据处理功能,并带有统计分析软件及测试仪器性能标准物的代谢组学应用包;Agilent公司开发的 MassHunter(<http://www.chem.agilent.com/Scripts/PDSLocal.asp?lPage=38489&LL=Y&CountryCode=CN>)以及 Applied Biosystems/MDS Sciex公司开发的 MarkerView(<http://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catD=601522>)等。第2类是可独立于仪器商的软件,如最近发展起来的 MZmine(<http://mzmine.sourceforge.net/index.shtml>),是一个用于处理代谢组学和蛋白组学 LC-MS数据的平台独立软件^[63]。该软件可以添加新的算法,并且可以处理其它类型的 MS联用数据,如 GC-MS和 CE-MS等仪器所产生的代谢组学数据。类似的软件还有美国 Samuel Roberts Noble Foundation提出的软件 MET-DEA(<http://www.noble.org/plantbio/ms/MET-DEA/index.html>)^[64]和美国 Scripps Research Institute开发的 XCMS(<http://metlin.scripps.edu/download/>)^[65]等。

在不同代谢样本的峰匹配、大量代谢物定量等方面近来也出现了一些相关的软件。如 MetaQuant(<http://bioinformatics.org/metaquant/>),是基于 Java,可以用来同时定量几百个代谢物的软件^[66];由荷兰生物系统基因组学中心与其合作单位提出的 metAlign(www.metalign.nl),可以分析比较、匹配不同的质谱数据集;美国 California 生物化学与分子生物研究所提出的 COPSPARI(<http://www.biomechanic.org/comspari/>),主要用于不同色谱或质谱谱图比较的可视化,同时包含了一个将 NetCDF格式数据转化为 ASCII格式的程序^[67];英国牛津大学提出的 SpecAlign(<http://physchem.ox.ac.uk/~jwong/specalign/>),提供许多通用的原始数据预处理功能,包括标准化、扣背景、平滑等,更重要的是它在软件中添加了新的多个质谱完全匹配算法,该算法也可用于色谱或光谱的匹配^[68-69]。

另外也有一些相关方面的独立算法出现,如在联用数据解析方面,本实验室提出的一些多变量曲线分辨算法^[70-73];Jonsson等^[17,74-75]提出的一些类似的处理 GC-MS数据的算法;以及成分检测算法(CODA)^[76](www.acdlabs.com)或窗口质谱选择法(WMSM)^[77]等。在克服系统误差、峰的识别与匹配等方面,Sysi-Aho等^[78]提出了采用多个内标化合物计算代谢组学样本中每类分子最优标准化因子以克服系统误差的 NOMIS法;Kanani等^[79]在保证代谢组学分析的高通量特性的情况下,也提出了相应

的新算法以克服 GC-MS 代谢组学样本分析中原始样本衍生破坏原始代谢物浓度与衍生物峰面积的一一对应关系而带来的系统误差; Dixon 等^[80]提出了用于大型 GC-MS 代谢组学数据峰识别及匹配的自动算法; Kazmi 等^[81]也提出了用于多个样本质谱匹配的新算法等。

尽管目前已出现了一些商业化或公开化的数据处理软件和方法,但是没有一个真正解决了代谢组学原始数据处理问题。能独立于仪器商、并且可灵活地添加新算法的标准数据处理软件必将推进代谢组学的发展。

2.1.2 NMR 数据处理 NMR 技术用于代谢组学研究,通常与模式识别等方法一起使用,为了得到准确的模式识别结果,就必须进行数据前处理。当然,相对于质谱及其联用数据来说, NMR 的原始数据处理要简单一些,它主要包括数据转换、标准化、匹配等步骤。针对这些问题,研究者已经提出了一些软件和算法。英国 Birmingham 大学 Viant 等^[82-83]采用 matlab 编写的软件 ProMetab 可以处理一维 NMR 谱图和二维 NMR 谱图的投影图,它可按用户所定义的宽度在化学位移方向将数据分割成小的片段,用户可删除不想要的片段,并可将特定的片段重新连接为新的数据,从而降低 pH 值对化学位移的影响。软件还包含了各种数据标准化和转换方法,并提出了一种新的、可消除数据异方差性数据的转换方法,从而可增加弱峰在数据中的权重。Ammann 等^[84]提出的 StePSM 方法则可用于识别 ¹H NMR 代谢组学样本组间不同的代谢物。为了消除因温度、pH 值、离子强度或仪器的不稳定等实验条件波动而引起的化学位移和峰形变化, Davis 等^[85]提出了采用非抽样小波变换的可调整窗口法 (adaptive binning),改进了模式识别效果。Stoyanova 等^[86]则提出了一个自动的方法来消除这些实验条件波动而引起的化学位移和峰形变化。Forshed 等^[87]还对早先发展的匹配方法——线性插值和基于峰识别的匹配方法进行了比较研究,并提出了采用遗传优化算法来进行 NMR 峰匹配^[88]。Lee 等^[89]在 Forshed 等工作的基础上,又提出了采用集束搜索算法 (beam search algorithm) 进行峰匹配的方法。

2.2 数据标准及数据库技术

代谢组学数据的管理很多方面非常类似于微阵列数据 MAME 协议^[90],但是其要求更加繁多。每个代谢组学实验通常包含了两套数据,一套是实验数据本身;另一套是实验方法。后者即元数据 (描述数据的数据),包括实验设计、样本特性、分析前处理信息、分析技术相关信息和数据处理细节。它与实验数据本身一样重要,如何采用合适的数据库存储和管理它们对今后重复该实验条件、比较不同实验室的结果以及数据挖掘非常重要。另一方面,很多情况下,单个生物样本采用不同的分析技术得到几组数据,如何存储和管理这些代谢组学数据也是一个问题。

总的来说,目前发展起来的或国际上正在致力发展的代谢组学数据库,主要有以下几种:

(1) 实验室特定数据库:可存储实验方法和元数据,所含信息非常详细,可采用标准格式输出数据给其它数据库。如 Fiehn 实验室的 SetupX 和 BinBase 软件 (<http://fiehnlab.ucdavis.edu/>)。

(2) 物种数据库:这类数据库存储与某一物种相关的、已报道的、相对简单的代谢物图谱。可为其它实验 (如蛋白组学、转录组学等) 提供一个数据源。

(3) 通用代谢谱:这类数据库存储的是代谢标准谱。这些标准谱是通用的,因此它们应该可以用于不同数据库、不同代谢组学研究平台之间的比较。这类数据库很复杂,它们包含了很多物种在不同的生理状况下已报道的代谢谱。

(4) 特定物种的已知代谢物库:这类数据库基于特定物种,列出了所有在该生物中已观测到的代谢物,这些代谢物可能是在不同的生理状态下检测到的。如 HMDB (<http://www.metabolomics.ca/>); 英国 Manchester 大学的酵母系统生物网络研究小组的酵母代谢物谱 (www.yesbn.org) 等。

(5) 已知代谢物库:它不针对具体的物种,包含所有曾经以某种方式检测到的所有代谢物。数据的存储可能以单个生物水平进行,也可能按其它分类方法进行。

(6) 标准生物化学数据库:其提供的是已有的生物化学信息,大部分的信息来自文献。目前已经有很多这类数据库,其中最通用的就是 KEGG (<http://www.genome.jp/kegg/>),该数据库包括了大量通路信息、基因组信息和化学信息;另外如 ExPASy (<http://www.expasy.ch/cgi-bin/search-biochem-index>),提供从序列到结构,以及两维凝胶等蛋白质操作相关的全套服务;EcoCyc (<http://ecocyc.org/>) 描述大肠杆菌的生化关系,包括基因、代谢途径、信号释放途径和运输蛋白等;MetaCyc

(<http://metacyc.org/>), 阐述了以微生物为主的 150 多种生物体中的代谢途径, 包含了从大量的文献和网上资源中得到的代谢途径、反应、酶和底物的资料, 其与 EcoCyc 在大肠杆菌资料库上有重复; IRIS (<http://www.iris.irri.org/>)、AraCyc (<http://www.arabidopsis.org/biocyc/>) 分别为水稻和拟南芥的有关数据库; BRENDA (<http://www.brenda.uni-koeln.de/>) 是一个基于原始文献的酶和代谢信息的综合性的数据库, 该数据库含有来自于 9 800 多种有机体的 83 000 种不同酶的数据, 它们按约 4 200 种 EC 号码进行分类, 这些数据摘自约 46 000 篇参考文献, 参考文献可与 PubMed 相链接; Reactome (<http://www.reactome.org/>) 能够帮助研究者对基因组规模数据进行观察和分析, 这些观察和分析不是孤立的, 而是与生物学路径问题等内容联系的; PUMA2 (<http://compbio.mcs.anl.gov/puma2/>), 可以进行基因和代谢网路的高通量进化分析, 它包含上千个原核和真核基因, 可采用各种生物信息学工具对序列数据进行互动分析。类似的数据库还有最近发展起来的 METLIN (<http://metlin.scripps.edu/>)^[60] 以及 BMRB (<http://www.bmrwisc.edu/>) 和 MMCD (<http://mmcd.mrfam.wisc.edu/>) 等。

(7) 代谢组数据库与国际标准: 理想化的代谢组学数据库应该非常全面, 并且充分灵活, 随着分析技术的不断发展, 可以收录新的格式数据, 并且和平行组学 (包括转录组学和蛋白组学) 的对应实验数据兼容。目前, 已经有一些公开的数据库、数据管理系统、数据分析和可视化工具, 如 AMet (<http://www.amet.org/>), 一个植物代谢组学实验数据模拟软件, 涵盖大部分植物代谢组学研究工作的网站, 包含了这些工作开展的时间, 甚至还有详细的实验步骤, 并将代谢物信息标准化, 以便于研究者交流; SMRS (<http://www.smrsgroup.org/>), 提供一个开放的代谢组学实验报告方式和标准的数据文件转换格式, 讨论实验元数据的哪些信息需要记录与保存; MeMo (<http://dbkgroup.org/memo/>), 是一个代谢组学模拟软件包, 采用 SQL/XmL 数据库技术管理功能基因组学中的代谢组学数据; MSI (<http://msi-workgroups.sourceforge.net/>) 是一个国际组织, 致力于定义整个代谢组学研究所需的最小信息, 其分为 5 个亚机构: 元数据管理组、化学分析组、数据处理组、本体论 (Ontology) 组和格式转化组; 另外还有 DOME (<http://medicago.vbi.vt.edu/>) 包含有许多关于代谢物的原始数据和分析结果, 可进行代谢组学、蛋白组学和转录组学的数据管理, 该系统包括了一个子系统参照物数据库 B-Net, B-Net 有望获取文献中已有的特定种类的分子生物学背景信息, DOME 也并入了一系列数据统计分析和可视化的工具 (如 BROME, standing for BRowser for OMEs)^[59, 91]; 另外, 最近发展的 BioSpider (www.biospider.ca) 就如何挖掘基因、蛋白和代谢物不同数据库相互间的信息, 并整合特定分子的所有信息做出了一些贡献^[92]; 这类数据库还有 MapMan (<http://gabi.rzd.de/projects/MapMan/>)、KaPPA-View (<http://kp.v.kazusa.or.jp/kappa-view/>) 和 MeNet (<http://metnet.vrac.iastate.edu/>) 等。

2.3 统计分析

除了采用合适的分析技术和数据库管理技术获得和存储数据外, 如何采用统计分析方法将这些数据转化成信息以了解细胞代谢的变化是一个重要的问题。代谢组学数据是多变量的, 它是对大量个体 (对象) 中许多不同代谢物 (变量) 的观察结果。每个变量可以看作一个不同的维, 如果有 n 个变量 (代谢物), 那么每个对象就可以看作是 n 维空间中的一个点。代谢组学数据集与其它组学数据集一样, 它们所含的变量数要多于样本数。在一个典型的组学实验中, 几百到几万个变量被测量, 但是通常样本数目却很小。要对这些数据进行统计分析, 以获得数据中不相关的特征, 减少变量数是非常重要的。通用的方法可分为两大类: 无监督和有监督算法^[91]。

无监督方法主要用于采用原始谱图信息或预处理后的信息对样本进行归类。该方法将得到的分类信息和这些样本的原始信息 (如药物的作用位点或疾病的种类等) 进行比较, 建立代谢产物与这些原始信息的联系, 筛选与原始信息相关的标记物, 进而考察其中的代谢途径。无监督方法中应用最广泛的是主成分分析 (PCA) 和层级聚类^[93-94]。另外还有自组织网络 (SOM)、软独立建模分类法 (SMCA) 和 K 近邻法 (KNN)^[94-95] 等。

有监督的方法用于建立类别间的数学模型, 使各类样品间达到最大的分离, 并利用建立的多参数模型对未知的样本进行预测。有监督的方法包括 ANOVA、人工神经网络 (ANNs)、偏最小二乘 (PLS)、判别函数分析 (DFA)^[19, 96-97]、演化算法与 PLS 和 DFA 一起使用等^[18, 98-102]。

目前, 统计分析方面已经存在很多软件及算法, 如 MATLAB (The Mathworks, Natick, MA, USA),

GNU Octave(<http://www.octave.org/>)和 R(<http://www.r-project.org/>)等都提供了相应的多变量分析工具包或算法,用户可进行灵活运用,但是它们需要用户熟悉其程序语句。另外还出现了一些图形界面多变量分析软件工具,如瑞典的 Evince(<http://www.umbio.com/products/evinced/software/software.asp>)和 SMCA(<http://www.acc.umu.se/%7Etnkjtg/chemometrics/softlinks.html>),美国的 Unscrambler(<http://www.camo.com/rt/Products/Unscrambler/unscrambler.html>), Pirouette(<http://www.infometrix.com/software/pirouette.html>)和 S-Plus(<http://www.insightful.com/products/spplus/default.asp>)等。但是这些软件还不能满足代谢组学研究的需要。代谢组学需要的是灵活、可扩展、免费可公开的、图形界面友好、易为一般用户所接受的多变量分析工具。最近 PyChem^[103]在这方面迈出了尝试性的一步,该软件不但具有大量的多变量分析功能(包括主成分分析 PCA、判别函数分析 DFA、偏最小二乘回归 PLS1、聚类分析 CA 和大量的用于特征选择的遗传算法工具),而且还包括了一些多变量数据预处理算法。

3 展 望

代谢组学是一个相对新兴的学科,但是它已经成为基因组学和系统生物学研究中不可分割的一部分。将来的发展还需要分析科学和生物信息学的改进。新的分析技术的发展主要集中在分辨率的提高、样本全面性分析的改进、分析速度和通量的增加以及新仪器的发展,包括微流控器件^[104-105]、UPLC^[106-108]、多维色谱-各种检测器(如 GC×GC-TOF-MS^[109-110]、GC×GC-Quadrupole-MS^[111]或 LC×LC-MS或 LC-CE-MS^[112-114]和 LC-NMR-MS^[115-116])。数据处理主要集中在数据的预处理、分析、标准化和管理方面。总的来说,代谢组学的发展需要更有力的成分分析设备,更多种分离分析手段的整合,更强大的数据挖掘和分析工具,更完备的代谢组学数据库。

致谢:感谢英国 Manchester 大学多学科生物中心 Douglas Kell 教授、Goodacre Roy 教授及其博士后 Warwick Dunn 的帮助。感谢英国生物技术与生物科学研究委员会中国伙伴合作项目(China Partner Award from British Biotechnology and Biological Science Research Council)的资助。

参考文献:

- [1] FIEHN O. [J]. Comparative and Functional Genomics, 2001, 2(3): 155 - 168
- [2] GOODACRE R. [J]. Drug Discovery Today, 2004, 9(6): 260 - 261.
- [3] ELLIS D I, GOODACRE R. [J]. Analyst, 2006, 131(8): 875 - 885
- [4] FIEHN O, KOPKA J, TRETHEWEY R N, et al. [J]. Anal Chem, 2000, 72(15): 3573 - 3580
- [5] MASHEGO M R, RUMBOLD K, DE MEY M, et al. [J]. Biotechnology Letters, 2007, 29(1): 1 - 16
- [6] VANTM, ROSENBLUM E, TJEERDEMA R. [J]. Environmental Science & Technology, 2003, 37(21): 4982 - 4989.
- [7] ALLEN J, DAVEY H M, BROADHURST D, et al. [J]. Nature Biotechnology, 2003, 21(6): 692 - 696
- [8] DUNN W B, ELLIS D I. [J]. Trac-Trends in Analytical Chemistry, 2005, 24(4): 285 - 294
- [9] CHOI H K, CHOI Y H, VERBERNE M, et al. [J]. Phytochemistry, 2004, 65(7): 857 - 864
- [10] KM H, CHOI Y, ERKELENS C, et al. [J]. Chemical & Pharmaceutical Bulletin, 2005, 53(1): 105 - 109
- [11] LAMERS R J, WESSELS E C, VAN DE SANDT J J, et al. [J]. Journal of Nutrition, 2003, 133(10): 3080 - 3084
- [12] GDMAN E, GOODACRE R, EMMETT B, et al. [J]. Phytochemistry, 2003, 63(6): 705 - 710
- [13] JOHNSON H E, BROADHURST D, KELL D B, et al. [J]. Applied and Environmental Microbiology, 2004, 70(3): 1583 - 1592
- [14] FLORIAN C L, PREECE N E, BHAKOO K K, et al. [J]. Cancer Research, 1995, 55(2): 420 - 427.
- [15] HMMLREICH U, SOMORJAI R L, DOLENKO B, et al. [J]. Applied and Environmental Microbiology, 2003, 69(8): 4566-4574
- [16] JARVIS R M, GOODACRE R. [J]. Bioinformatics, 2005, 21(7): 860 - 868
- [17] JONSSON P, GULLBERG J, NORDSTROM A, et al. [J]. Anal Chem, 2004, 76(6): 1738 - 1745
- [18] KELL D. [J]. Molecular Biology Reports, 2002, 29(1): 237 - 241
- [19] RAAMSDONK L M, TEUSINK B, BROADHURST D, et al. [J]. Nature Biotechnology, 2001, 19(1): 45 - 50
- [20] SCHOLZ M, GATZEK S, STERLING A, et al. [J]. Bioinformatics, 2004, 20(15): 2447 - 2454
- [21] HALL R, BEALE M, FIEHN O, et al. [J]. The Plant Cell, 2002, 14(7): 1437 - 1440
- [22] OLIVER S G, WILSON M K, KELL D B, et al. [J]. Trends in Biotechnology, 1998, 16(9): 373 - 378
- [23] PHELPS T J, PALUMBO A V, BELAEV A S. [J]. Current Opinion in Biotechnology, 2002, 13(1): 20 - 24
- [24] VILLAS-BÔAS S G, MOXLEY J F, ÅKESSON M, et al. [J]. Biochemical Journal, 2005, 388(2): 669 - 677.

- [25] SHULAEV V. [J]. Briefings in Bioinformatics, 2006, 7(2): 128 - 139.
- [26] GOODACRE R. [J]. Journal of Nutrition, 2007, 137(1): 259S - 266S.
- [27] LENZ E M, W LSON ID. [J]. Journal of Proteome Research, 2007, 6(2): 443 - 458.
- [28] SUMNER L W, MENDES P, DIXON R A. [J]. Phytochemistry, 2003, 62(6): 817 - 836.
- [29] DETMER K, ARONOV P A, HAMMOCK B D. [J]. Mass Spectrometry Reviews, 2007, 26(1): 51 - 78.
- [30] GLNSKIM, WECKWERTH W. [J]. Mass Spectrometry Reviews, 2006, 25(2): 173 - 214.
- [31] HALKET J M, WATERMAN D, PRZYBOROWSKA A M, et al [J]. Journal of Experimental Botany, 2005, 56(410): 219 - 243.
- [32] ROESSNER U, WAGNER C, KOPKA J, et al [J]. The Plant Journal, 2000, 23(1): 131 - 142.
- [33] MAURER H H. [J]. J Chromatogr. B: Biomedical Sciences and Applications, 1998, 713(1): 3 - 25.
- [34] BENDAHL L, HANSEN S H, GAMMELGAARD B, et al [J]. Journal of Pharmaceutical and Biomedical Analysis, 2006, 40(3): 648 - 652.
- [35] CASTRO-PEREZ J, PLUMB R, GRANGER J H, et al [J]. Rapid Communications in Mass Spectrometry, 2005, 19(6): 843 - 848.
- [36] JOHNSON K A, PLUMB R. [J]. Journal of Pharmaceutical and Biomedical Analysis, 2005, 39(3/4): 805 - 810.
- [37] LEANDRO C C, HANCOCK P, FUSSELL R J, et al [J]. J Chromatogr. A, 2006, 1103(1): 94 - 101.
- [38] TOLSTIKOV V, FIEHN O. [J]. Analytical Biochemistry, 2002, 301(2): 298 - 307.
- [39] PERRETT D, ALFRZEMA L, HOWSM, et al [J]. Biochemical Society Transactions, 1997, 25(11): 273 - 278.
- [40] PERRETT D, ROSS G. [J]. Trac-Trends in Analytical Chemistry, 1992, 11(4): 156 - 163.
- [41] TERABE S, MARKUSZEWSKIM, NOUE N, et al [J]. Pure and Applied Chemistry, 2001, 73(10): 1563 - 1572.
- [42] BRITZMCKBB N P, TERABE S. [J]. Chemical Record, 2002, 2(6): 397 - 404.
- [43] SOGA T, MAIZUMI M. [J]. Electrophoresis, 2001, 22(16): 3418 - 3425.
- [44] SOGA T, OHASHI Y, UENO Y, et al [J]. Journal of Proteome Research, 2003, 2(5): 488 - 494.
- [45] SOGA T, UENO Y, NARAOKA H, et al [J]. Anal Chem, 2002, 74(10): 2233 - 2239.
- [46] KRISHNAN P, KRUGER N J, RATCLIFFE R G. [J]. Journal of Experimental Botany, 2005, 56(410): 255 - 265.
- [47] LE GALL G, COLQUHOUN I J, DAVIS A L, et al [J]. Journal of Agriculture and Food Chemistry, 2003, 51(9): 2447 - 2456.
- [48] PAN Z Z, RAFTERY D. [J]. Analytical and Bioanalytical Chemistry, 2007, 387(2): 525 - 527.
- [49] RATCLIFFE R G, SHACHAR-HILL Y. [J]. Annual Review of Plant Physiology and Plant Molecular Biology, 2001, 52(1): 499 - 526.
- [50] VERPOORTE R, CHOI Y H, KIM H K. [J]. Phytochemistry Reviews, 2007, 6(1): 3 - 14.
- [51] GRIFFIN J. [J]. Current Opinion in Chemical Biology, 2003, 7(5): 648 - 654.
- [52] KELL D B. [J]. Current Opinion in Microbiology, 2004, 7(3): 296 - 307.
- [53] DIEM M, BOYDSTON-WHITE S, CHIRBOGA L. [J]. Applied Spectroscopy, 1999, 53(4): 148A - 161A.
- [54] TWEEDEDALE H, NOTLEY-MCROBB L, FERENCIT. [J]. Redox Report, 1999, 4(5): 237 - 241.
- [55] FRASER P, PINTO M, HOLLOWAY D, et al [J]. Plant Journal, 2000, 24(4): 551 - 558.
- [56] GAMACHE P, MEYER D, GRANGER M, et al [J]. Journal of the American Society for Mass Spectrometry, 2004, 15(12): 1717 - 1726.
- [57] KRISTAL B, VIGNEAU-CALLAHAN K, MATSON W. [J]. Methods in Molecular Biology, 2002, 186: 185 - 194.
- [58] BOCHNER B, GADZINSKI P, PANOMIROS E. [J]. Genome Research, 2001, 11(7): 1246 - 1255.
- [59] SAITO K, DIXON R, WILLMEZ L. Plant Metabolomics[M]. 1st ed Berlin Heidelberg: Springer-Verlag, 2006: 141 - 154.
- [60] WENTE J, NORDSTROM A, MORITA H, et al [J]. Journal of Proteome Research, 2007, 6(2): 459 - 468.
- [61] LU Hongnei, DUNN W B, SHEN Hailin, et al [J]. Trac-Trends in Analytical Chemistry, 2008, 27, in press.
- [62] STYCZYNSKIM P, MOXLEY J F, TONG L V, et al [J]. Anal Chem, 2007, 79(3): 966 - 973.
- [63] KATAJAMAA M, MIETTINEN J, ORESIC M. [J]. Bioinformatics, 2006, 22(5): 634 - 636.
- [64] BROECKLING C D, REDDY I R, DURAN A L, et al [J]. Anal Chem, 2006, 78(13): 4334 - 4341.
- [65] SMITH C A, WENTE J, O'MALLE G, et al [J]. Anal Chem, 2006, 78(3): 779 - 787.
- [66] BUNK B, KUCKLICK M, JONAS R, et al [J]. Bioinformatics, 2006, 22(23): 2962 - 2965.
- [67] KATZ J E, DUMLAO D S, CLARKE S, et al [J]. Journal of the American Society for Mass Spectrometry, 2004, 15(4): 580 - 584.
- [68] WONG J W H, CAGNEY G, CARTWRIGHT H M. [J]. Bioinformatics, 2005, 21(9): 2088 - 2090.
- [69] WONG J W H, DURANTE C, CARTWRIGHT H M. [J]. Anal Chem, 2005, 77(17): 5655 - 5661.
- [70] KVALHEIM O M, LANG Yizeng. [J]. Anal Chem, 1992, 64(8): 936 - 946.
- [71] LANG Yizeng, KVALHEIM O M. [J]. Anal Chim Acta, 1994, 292(1/2): 5 - 15.
- [72] LANG Yizeng, KVALHEIM O M. [J]. Fresenius Journal of Analytical Chemistry, 2001, 370(6): 694 - 704.
- [73] LANG Yizeng, KVALHEIM O M, KELLER H R, et al [J]. Anal Chem, 1992, 64(8): 946 - 953.

- [74] JONSSON P, JOHANSSON A, GULLBERG J, et al [J]. *Anal Chem*, 2005, 77(17): 5635 - 5642
- [75] JONSSON P, STENLUND H, MORITZ T, et al [J]. *Metabolomics*, 2006, 2(3): 135 - 143
- [76] WINDIG W, PHALP J M, PAYNE A W. [J]. *Anal Chem*, 1996, 68(20): 3602 - 3606
- [77] FLEMING C M, KOWALSKI B R, APFEL A, et al [J]. *J Chromatogr. A*, 1999, 849(1): 71 - 85
- [78] SYSTÄHO M, KATAJAMAA M, YETUKURI L, et al [J]. *BMC Bioinformatics*, 2007, 8(1): 93 - 109
- [79] KANANI H H, KLAPAM I [J]. *Metabolic Engineering*, 2007, 9(1): 39 - 51
- [80] DIXON S J, BRERETON R G, SONI H A, et al [J]. *Journal of Chemometrics*, 2006, 20(8/10): 325 - 340
- [81] KAMISA, GHOSH S, SHIND G, et al [J]. *Metabolomics*, 2006, 2(2): 75 - 83
- [82] PUROHIT P V, ROCKE D M, VANTM R, et al [J]. *OMICS*, 2004, 8(2): 118 - 130
- [83] VANTM R [J]. *Biochemical and Biophysical Research Communications*, 2003, 310(3): 943 - 948
- [84] AMMANN L P, MERRITT M. [J]. *Metabolomics*, 2007, 3(1): 1 - 11
- [85] DAVIS R A, CHARLTON A J, GODWARD J, et al [J]. *Chemometrics and Intelligent Laboratory Systems*, 2007, 85(1): 144 - 154
- [86] STOYANOVA R, NICHOLLS A W, NICHOLSON J K, et al [J]. *Journal of Magnetic Resonance*, 2004, 170(2): 329 - 335
- [87] FORSHED J, TORGER P J O, ABERG K M, et al [J]. *Journal of Pharmaceutical and Biomedical Analysis*, 2005, 38(5): 824 - 832
- [88] FORSHED J, SCHUPPE-KOISTINEN I, JACOBSSON S P. [J]. *Anal Chim Acta*, 2003, 487(2): 189 - 199
- [89] LEE G-C, WOODRUFF D L. [J]. *Anal Chim Acta*, 2004, 513(2): 413 - 416
- [90] BRAZMA A, HINGAMP P, QUACKENBUSH J. [J]. *Nature Genetics*, 2001, 29(4): 365 - 371
- [91] MENDES P. [J]. *Brief Bioinformatics*, 2002, 3(2): 134 - 145
- [92] KNOX C, SHRIVASTAVA S, STOTHARD P, et al Biospider: A web server for automating metabolome annotations[C]. In ALTMAN R B, DUNKER A K, HUNTER L, et al (eds), *Pacific Symposium on Biocomputing*, 2007: 145 - 156
- [93] ALTER O, BROWN P O, BOTSTEN D. [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2000, 97(18): 10101 - 10106
- [94] EISEN M B, SPELLMAN P T, BROWN P O, et al [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 1998, 95(25): 14863-14868
- [95] TAMAYO P, SLONIM D, MESIROV J, et al [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 1999, 96(6): 2907 - 2912
- [96] CHURCHILL G [J]. *Biotechniques*, 2004, 37(2): 173 - 175, 177.
- [97] MUSUMARRA G, BARRESI V, CONDORELLI D, et al [J]. *Biological Chemistry*, 2003, 384(2): 321 - 327.
- [98] PENA-REYES C, SIPPER M. [J]. *Artificial Intelligence in Medicine*, 2000, 19(1): 1 - 23
- [99] GOODACRE R. [J]. *Journal of Experimental Botany*, 2005, 56(410): 245 - 254
- [100] HARRIGAN G G, GOODACRE R. *Metabolic profiling*[M]. 1st ed Boston: Kluwer Academic Publishers, 2003: 239 - 256
- [101] MENDES P, CAMACHO D, DE LA FUENTE A. [J]. *Biochemical Society Transactions*, 2005, 33(6): 1427 - 1429
- [102] SHIH, PAOLUCCI U, VIGNEAU-CALLAHAN K, et al [J]. *OMICS*, 2004, 8(3): 197 - 208
- [103] JARVIS R M, BROADHURST D, JOHNSON H, et al [J]. *Bioinformatics*, 2006, 22(20): 2565 - 2566
- [104] GARCIA C, HENRY C [J]. *Anal Chem*, 2003, 75(18): 4778 - 4783
- [105] WANG J, CHEN G [J]. *Talanta*, 2003, 60(6): 1239 - 1244
- [106] NORDSTROM A, OMALLE G, QIN C, et al [J]. *Anal Chem*, 2006, 78(10): 3289 - 3295
- [107] SHERMA J. [J]. *The Journal of AOAC International*, 2005, 88(3): 63A - 67A.
- [108] WREN S [J]. *Journal of Pharmaceutical and Biomedical Analysis*, 2005, 38(2): 337 - 343
- [109] ADAHCHOUR M, JOVER E, BEENS J, et al [J]. *J Chromatogr. A*, 2005, 1086(1/2): 128 - 134
- [110] DALLUGE J, BEENS J, BRINKMAN U. [J]. *J Chromatogr. A*, 2003, 1000(1/2): 69 - 108
- [111] ADAHCHOUR M, BRANDT M, BAIER H, et al [J]. *J Chromatogr. A*, 2005, 1067(1/2): 245 - 254
- [112] MONDELLO L, TRANCHIDA P, STANEK V, et al [J]. *J Chromatogr. A*, 2005, 1086(1/2): 91 - 98
- [113] SHELTON E [J]. *Journal of Pharmaceutical and Biomedical Analysis*, 2003, 31(6): 1153 - 1166
- [114] STRONK T, ORTIZ M, BULTA, et al [J]. *J Chromatogr. B*, 2005, 817(1): 49 - 66
- [115] CORCORAN O, SPRAUL M. [J]. *Drug Discovery Today*, 2003, 8(14): 624 - 631
- [116] YANG Zheng [J]. *Journal of Pharmaceutical and Biomedical Analysis*, 2006, 40(3): 516 - 527.