上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

XXXXXX 进展调研课程论文

学院 XXXXXX 研究院

学号 XXXXXXX

姓名 XXXX

2025 年 1 月 15 日

目录

1 主题说明

结合自身研究领域，就 topic-代谢组学技术原理及前沿进展展开讨论，篇幅不限，中英文不限制。

2 正文

尽管从历史上的发展角度来看，科学的研究是哲学的边界，但是从现在的科学研究的角度来看，哲学的研究是科学的边界。

在生物学界，有很多的重大科研突破以及发现不乏是即兴的产物，而非是经过严谨的科学研究过程。然而，我一直相信，一篇好的学术论文，基本上属于下述三类成果之一：

- （1）具有重大理论意义
- （2）挑战了学界对某一重要生物学概念或方法的共通认识
- （3）为实践领域的某一中长期问题提出来可行的生物学解决方案

后两者，都带有浓郁的实践应用色彩，对于我专长的领域-生物信息学而言，或者需要长期且优质的数据分析，或者干脆要求作者是曾深度参与实际项目的研发人员，因此对个人而言暂时都是不大可行的。那么，要做出有趣的“好东西”，就必然得求诸理论。

我有一个高度理论化的脑袋。这几年来，我粗略地接触了很多生物科学领域、研讨了很多问题、学了很多东西，然而，真正参与思维模式构建的只有三种：计算生物学、统计学和系统生物学。在我看来，以基因组和蛋白质组学等为核心的分子生物学，不过是计算生物学的一个微小分支罢了。从最抽象的意义讲，三者的思维结构是共通的——以基于模型假设的理论框架和数据驱动的直觉为动力，以先验观念为前提，以规范推理为过程，以内融外贯通的宏大体系为奋斗目标。当然，在这里的“系统生物学”，显然更多体现了整合多层次生物数据的复杂性，而非单纯的数据分析或实验验证。相应地，我也形成了此种思考偏好，对理论和模型的思考可能深一点，实验看的得不少，但终归还是对经验和证据的思考少一点。从很大程度上说，理论思考比单纯的数据处理对于科学研究而言更为重要。

我为什么谈论这些，是因为从现实意义上去讲，代谢组学是干实验和湿实验离得最近的领域，而其理论框架的构建，则是我认为最有可能在短时间内做出突破的领域。究其根本，是因为代谢组学不同于中心法则中更上游的基因组学和蛋白质组学，它更接近于生物体的功能性研究，因而更需要理论的支撑。而且，代谢组学的数据量和数据质量，也是最适合于理论研究的。

代谢组学 (Metabonomics/Metabolomics) 是对某一生物体组分或细胞在某一特定生理时期或条件下所有代谢产物同时进行定性和定量分析，以寻找出目标差异代谢物，可

用于疾病早期诊断、疾病机理研究等。从逻辑上讲，基因组学是静态的，它表征的是可能发生什么 (what can happen)；转录组学捕获的是某一个时期某个状态下细胞所产生的转录本，表征的是可能正在发生的 (what appears to be happening), 而蛋白质作为生命活动的主要执行者，表征的是很多下游生物学现象、生物学事件或表型等实际发生的原因 (what makes it happen)。而代谢组学，是动态的，它表征的是生物体内代谢产物的实际变化，是生物体内生物学事件的最终结果 (what has happened or what actually happens)。

代谢组学的功能，就有点类似于 GPS，代谢组学不做任何假设，针对样本和参照物分别进行代谢物全谱检测，帮你寻找出差异代谢物，可能会有好几十个甚至更多，从它们所在的被上调或下调的代谢通路上去找到关键的代谢酶，再找到它们上游的调控基因，再进行后续的深入研究，是不是可以避免走一些弯路？所以说代谢组的假说-free 的组学模式研究 (从干实验角度来说)，以及处于生物事件下游终点的特点 (从湿实验角度来说)，两者相结合，是一个相当有效率的工具，也是一个相当 insightful 的研究视角。

在技术分析角度上，代谢组学所能做的，就是从数据中寻找规律，从规律中寻找差异，从差异中寻找假说，从假说中寻找答案。这点其实从本质逻辑上讲是和其他组学没有区别的：基因组学是从基因型到表型的逻辑链条，转录组学是从转录本到表型的逻辑链条，蛋白质组学是从蛋白质到表型的逻辑链条，而代谢组学则是从代谢产物到表型的逻辑链条。这些逻辑链条，都是从数据到规律、从规律到差异、从差异到假说、从假说到答案的逻辑链条，只是数据的类型和规律的性质不同而已。

下面是对代谢组学技术原理的一些讨论总结，其实万变不离其宗，更多的是一些回顾复盘。

一，代谢组学研究流程概述

1，样品采集与前处理

代谢组学研究通常从样品采集开始，包括血液、尿液、组织等多种生物基质。为保证后续结果的准确性，需要严格控制采样条件，如同一时间段采样、统一温度保存等。随后，对生物样品进行提取和前处理，常用方法包括有机溶剂沉淀蛋白（如甲醇、乙腈等）或液液萃取等。该过程的重点在于去除蛋白、脂质等杂质，以尽可能保留代谢物的完整谱信息 [1]。

2，代谢检测与数据获取

完整预处理之后，常使用液相色谱-高分辨率质谱 (LC-HRMS) 或气相色谱-质谱 (GC-MS) 进行检测。LC-HRMS 适用于极性代谢物的检测，GC-MS 适用于非极性代谢物的检测。两者结合可覆盖更广泛的代谢物种类。检测完成后，得到的数据通常为二维矩阵，其中行代表样品，列表示代谢物，元素为代谢物的相对丰度。常规通过质谱对分离后的代谢物特征峰进行检测的过程，能够获取的原始谱图包括保留时间、质量荷比 (m/z) 以及峰面积 (intensity) 等关键信息。也有研究采用核磁共振 (NMR) 等技术进行分子结构推断，但因为灵敏度与适用性等原因，NMR 在代谢组学研究中的应用较为有限，在定量分析中 LC-MS 依然最为常用 [2]。

3，数据预处理与归一化

获取原始谱图后，首先需要进行噪声过滤 (noise filtering) 和基线校正，以去除仪器噪声和漂移等干扰。接下来便是峰识别 (peak detection) 与峰对齐 (peak alignment)，将不同样本的同一代谢物峰进行匹配，构建最终的特征矩阵。5。目的是校正不同批次、不同样本间可能存在的进样量、检测灵敏度偏差，从而减少实验误差的影响。

二，代谢组学数据分析方法

1，无监督学习 (unsupervised learning)

在获得符合分析要求的特征矩阵后，常首先进行主成分分析 (principal component analysis, PCA) 或多维尺度分析 (Multidimensional scaling, MDS) 等无监督学习方法，以揭示样本间的差异和代谢物之间的相关性。PCA 是一种降维技术，通过将高维数据映射到低维空间，保留最大方差的方式，实现数据的可视化和解释。MDS 则是一种距离度量技术，通过计算样本间的相似性，将样本映射到低维空间，以便于观察样本间的关系。非线性映射 (Nonlinear mapping) 方法如 t-SNE 等、层次聚类分析 (Hierarchical clustering analysis,HCA) 等，也常用于代谢组学数据的降维和可视化，同样是常见的无监督手段，可以观察样本间聚类的模式。无监督方法的优势在于不需要预先定义类别，仅仅依据数据内在结构来发现组间差异或潜在分群，可以在研究早期快速获得全局概貌 [3]。

2，有监督学习 (supervised learning)

如果我们有先验的分组信息 (如健康组与患病组)，可以利用偏最小二乘判别分析 (PLS-DA)、支持向量机 (SVM)、随机森林 (Random forest) 等有监督学习方法，进行特征选择和分类预测。PLS-DA 是一种线性回归技术，通过最大化类间差异和最小化类内差异，找到最佳的分类超平面。SVM 是一种二分类模型，通过构建最优超平面，将不同类别的样本分开。随机森林则是一种集成学习方法，通过多个决策树的投票，实现分类预测。利用这些有监督算法来构建判别模型，寻找能够最大化区分两组的生物标志物 (biomarker)，并进行后续的生物学解释和验证，是代谢组学研究的重要目标之一。需要注意的是，尽管有监督模型可以帮助研究者从大量变量中筛选出区分度最好的代谢物，但是有监督模型易受到过拟合影响，通常需要配合交叉验证或独立测试集来评估模型稳健性 [4]。

3，差异代谢物筛选与注释

通过上述方法识别出的潜在差异代谢物，需要与已有的代谢物数据库，通常也是其他组学分析中常用的富集分析知识库 (如 METLIN、HMDB、KEGG 等) 进行比对，以获取准确的分子信息 (如结构、通路位置等)，并进行生物学功能注释。以转录组学分析最常见的 KEGG 数据库为例：先筛选出显著差异峰，再利用数据库进行匹配和通路富集分析，从而理解这些代谢物在特定生物学过程中的作用 [5]。

4，统计分析

在代谢组学数据分析中，统计分析是不可或缺的一环。常用的统计方法包括 t 检验、方差分析 (ANOVA)、卡方检验等，用于比较不同组间的代谢物丰度差异。此外，还有许

多专门用于代谢组学数据分析的统计方法，如非参数检验、多重比较校正等，以避免假阳性结果的出现。例如，对于两组间的差异分析，我们可以针对每个代谢物的峰面积进行统计检验，再结合多重假设检验校正，以确认哪些差异是真正具有统计学意义的 [6]。在代谢组学分析中，统计学方法贯穿始终。

三，代谢组学的研究与应用

1，疾病诊断与分型：

在转化医学领域，代谢组学与临床信息学相结合，可以帮助早期发现某些疾病特异性的代谢指纹 (metabolic fingerprint)，从而实现疾病的早期诊断与分型。例如，糖尿病、肿瘤等疾病患者的代谢组学特征与正常个体存在显著差异，通过分析这些差异代谢物，可以辅助医生进行疾病诊断和分型，提高诊断准确性和治疗效果。研究者可以进一步阐明疾病的潜在病理机制，并探索新的治疗靶标。而这些，事实上都是癌症多组学分析中常见的应用场景。

2，药物代谢与毒理学

药物研发过程需要考虑药物在体内的吸收、分布、代谢、排泄等 (ADME) 过程，以及药物对机体的毒性作用。通过代谢组学，我们可以全景式地检测机体对药物以及药物对机体代谢通路的影响。此外，在毒理学研究中，代谢组学也可以用来评估有毒化合物对机体代谢网络的扰动，从而为安全评估以及药物的毒副作用检测提供帮助 [7]。总结来说就是，代谢组学可以帮助研究者了解药物在体内的代谢途径、代谢产物及其浓度变化，从而指导药物的合理使用和剂量调整。同时，代谢组学还可以帮助研究者评估药物的毒性作用，发现潜在的毒性代谢物，为药物安全性评价提供重要参考。

3，营养学与个体化医疗

营养学领域也开始运用代谢组学方法来评估膳食干预对人体代谢通路的影响，为”个性化营养” 提供科学依据。例如，不同人群对相同膳食的成分的反应会存在差异，通过代谢轮廓分析可区分出相应模式的差异，涉及更科学的个体化营养方案 [8]。

四，回顾与反思

我们见过了太多太多的模板化的 cancer 多组学分析文献，遗憾的是代谢组学独木难成林，走得也是这种以量取胜为主的路径。

我们不去深究生物学机制研究方面的文献细节，因为这些文献往往涉及具体的生物过程和实验设计，而我们的目标是理解代谢组学的基本概念、研究方法及其应用。或者更重要的是，在统计过大多数的代谢组学文献之后，我能得到的初步结论就是，大多数文献依然是以医院作为主要研究单位，做的都是些工程性、分析性质的样本流程化工作。事实上，在后基因组学的现在，只要是涉及到癌症 (cancer) 的组学研究，实际上很难有突破性机制研究方面的文献发表。而代谢组，私以为在组学研究神坛中只不过是更贴生物学地气点，更贴近生物体功能性研究，理应更应该在” 真实性” 上为生物学研究赋能。但是，在组学研究神坛上，却往往被边缘化，甚至被边缘化到生物信息学领域，这无疑是一种遗憾。

我们缺的不是工程化研究，缺的是底层机制性研究。

如果代谢组学发展的主流趋势不能胜任，那么，我们就需要从理论上去思考，去寻找新的突破口。这也是我为什么要谈论多组学联合，代谢组与基因组、转录组、蛋白质组等多组学数据的融合分析已成为趋势，通过构建系统生物学模型，可以从多个层面从真实性上去逼近、去揭示生命活动的本质。而对海量异质性数据进行整合，需要更加高效的数据存储、计算和可视化方案，这对于生物信息学方法提出了新的考验 [9]。

多组学只是一种研究策略，而不是研究目的，同样我还要谈论的另外一点是机器学习，后者同样也只是一种分析策略，因为本质上机器学习使用的原始数据以及原始假设，都是基于统计学的视角去理解生物学事件，而统计学的本质是对数据的描述和推断，而非对生物学机制的解释。

参考文献

[1] Augustin Scalbert, Lorraine Brennan, Oliver Fiehn, Thomas Hankemeier, Bruce S Kristal, Ben van Ommen, Estelle Pujos-Guillot, Elwin Verheij, David Wishart, and Suzan Wopereis. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*, 5:435–458, 2009.

[2] Abdul-Hamid Emwas, Raja Roy, Ryan T McKay, Leonardo Tenori, Edoardo Saccenti, GA Nagana Gowda, Daniel Raftery, Fatimah Alahmari, Lukasz Jaremko, Mariusz Jaremko, et al. Nmr spectroscopy for metabolomics research. *Metabolites*, 9(7):123, 2019.

[3] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[4] Max Bylesjö, Mattias Rantalainen, Olivier Cloarec, Jeremy K Nicholson, Elaine Holmes, and Johan Trygg. Opls discriminant analysis: combining the strengths of pls-da and simca classification. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 20(8-10):341–351, 2006.

[5] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The kegg resource for deciphering the genome. *Nucleic acids research*, 32(suppl_1):D277–D280, 2004.

[6] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[7] Donald G Robertson. Metabonomics in toxicology: a review. *Toxicological sciences*, 85(2):809–822, 2005.

[8] Aoife O'Gorman and Lorraine Brennan. Metabolomic applications in nutritional research: a perspective. *Journal of the Science of Food and Agriculture*, 95(13):2567–2570, 2015.

[9] Biswapriya B Misra, Carl Langefeld, Michael Olivier, and Laura A Cox. Integrated omics: tools, advances and future approaches. *Journal of molecular endocrinology*, 62(1):R21–R45, 2019.