PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2018, 45(1): 23~35
www.pibb.ac.cn

宏蛋白质组学信息分析的基本策略及其挑战 *

徐洪凯[1,2)**] 闫克强[1,2)**] 何燕斌[1)] 闻 博[1)] 杨焕明[1,3)] 刘斯奇[1,2,3)***]

([1)] 深圳华大生命科学研究院，深圳 518083；[2)] 中国科学院北京基因组研究所，基因组科学与信息重点实验室，北京 100101；
[3)] 沃森基因组研究院，杭州 310008)

摘要 宏蛋白质组学是一门新型科学，它运用质谱技术规模化地采集自然界微生物种群的蛋白质信息，并结合多种组学数据，开展微生物种群的遗传特征及其生物功能的研究．宏蛋白质组学的信息分析与传统蛋白质组学方法有较大的不同，亟需拓展新的分析思路．由于宏蛋白质组的研究对象是复杂度极高的微生物样品，因此，需要构建尽可能囊括样本中所含微生物的基因组信息的物种数据库．面对庞大的数据库，必须考虑到分析过程中所消耗的计算资源和鉴定结果的质控标准，因此，需要高度优化库容量、搜库、假阳性控制等参数．鉴于宏蛋白质组数据中广泛存在复杂的同源蛋白质序列，因此，需要充分利用 NCBI 数据库中的分类信息进行匹配，并运用 LCA 算法过滤处理才能将蛋白质有效地归组到物种．本文立足于宏蛋白质组学信息分析，从宏蛋白质组的数据库建立、蛋白质归并、生物学意义发掘等几个方面着手，对该领域的发展现状、面临挑战以及未来研究方向进行了评述．

关键词 宏蛋白质组学，数据库，数据分析，蛋白归并，物种分析
学科分类号 Q51，Q71
DOI: 10.16476/j.pibb.2017.0187

微生物在整个地球生物圈中扮演着重要作用．微生物有很强的环境适应能力，具备足够的自身修复能力；微生物往往能够充分利用其寄生环境，发展复杂的微生物群落，形成微生物与环境互动的网络效应．一般认为，采用实验室细菌培养的方法，无法开展自然界微生物种群结构和遗传多样性的研究．这是因为一方面，仅有不到 1%的自然界微生物可能以实验室方法获得培养或纯化；而另一方面，通过人工手段培养的细菌可能丧失了自身特征，尤其是微生物菌落之间的相互关系．科学研究者很早就意识到这个问题，但是限于实验手段，自然界微生物群落结构及分布的研究一直进展缓慢．基因组学的兴起和 DNA 测序技术的发展，为解决这个难题奠定了理论和技术的基础．值得指出的是，二代测序技术(NGS)的出现，不仅提高了基因组测序的速度，极大地降低了测序成本；而且为基因组学开拓了新的应用领域．几乎与 NGS 同步，宏基因组(metagenomics)应运而生，它以环境样品中的微生物群体基因组为研究对象，以测序分析为研究手段，从大数据的角度解析微生物多样性、种群结构、进化关系、功能活性、种群互作、生存环境之间的复杂关系．NGS 能够对复杂微生物群体的基因组进行 DNA 测序，通过搜库或者 *de novo* 方法，获得环境微生物的基因信息总和，并实行功能注释，全面地揭示微生物种群的结构、功能、进化等特征[1]．

蛋白质是基因功能的直接体现者．在微生物宏基因组数据的基础上，进一步开展其相应的蛋白质组成及功能的研究，将能够深入发掘微生物群体中功能分子的丰度分布和功能取向．2004 年，Rodriguez-Valera[2]第一次提出了“宏蛋白质组”的概念．一般而言，宏蛋白质组学依据宏基因组数据，通过质谱技术采集特定环境条件下微生物种群中全体蛋白质信息，探索微生物组成以及代谢方式，揭示微生物与环境之间的相互关系．例如，在 2004 年，Wilmes 等[3]第一次利用双向电泳技术分

* 国家重点基础研究发展计划(973)(2014CBA02002, 2014CBA02005)资助项目．
** 并列第一作者．
*** 通讯联系人．
Tel: 0755-36307403, E-mail: siqiliu@genomics.cn
收稿日期：2017-09-05，接受日期：2017-10-20

离具有生物除磷作用的活性污泥中的微生物蛋白质，并利用质谱技术鉴定了这些分离的蛋白质，2005 年，Kan 等[4]采用液相 - 质谱联用技术分析了切萨皮克海湾的微生物群落．近年来宏蛋白质组学有关该领域的研究论文增长迅速(图 1)，它已经广泛地应用于众多领域，比如人类和动物的肠道微生物种群[5-7]、海洋环境微生物种群[8-9]、土壤微生物种群[10]等．

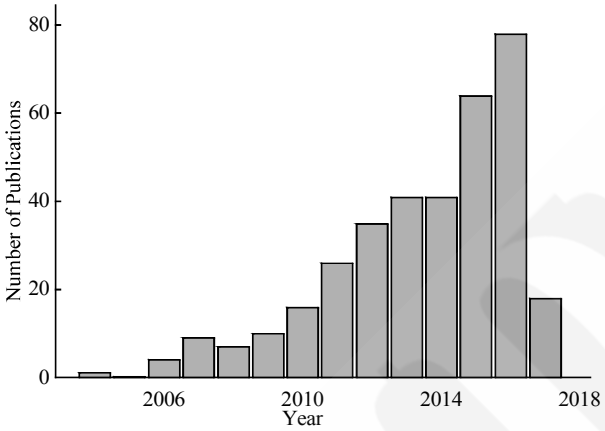

Fig. 1　Publications of metaproteomics

图 1　宏蛋白质组学年度发表论文数量(2004～2017)

(摘自 Web of Science，统计结果以“metaproteomics”和“metaproteome”为关键字进行搜索，截至 2017 年 04 月 30 日).

作为一门新兴学科，尽管适用于宏蛋白质组数据的信息分析工具和方法不断发展，但是这方面所遇到的挑战还是前所未有的．为了应对这些挑战，有必要深入了解一般蛋白质组与宏蛋白质组在生物信息学分析上的差异性．首先，一般蛋白质组的信息学分析所依据的基因组或转录组数据大都来自单一的生物体，MS/MS 与理论肽段之间的对应关系比较简单．宏蛋白质组的分析对象则相当复杂，含有大量不同类型的微生物，因此，在质谱数据搜库过程中就必须选择多个微生物基因组数据库来进行蛋白质鉴定．数据库选择是宏蛋白质组信息学分析关键因素之一：数据库过大，会影响鉴定的质量，而数据库太小，又会丢失许多肽段序列信息．其次，除了鉴定突变肽段需要较大的肿瘤基因组数据库之外，大多数条件下一般蛋白质组的信息分析所采用的搜索数据库都比较小．宏蛋白质组的分析却要面对庞大数据库．如果仍然采用传统搜库方法进行肽段鉴定，那将是耗时费力效率低下的过程，因此，宏蛋白质组数据库的搜库策略必须有所革新．再次，虽然理论上搜索引擎的结果不依赖于数据库的大小，但是实际上如果数据库悬殊太大，其假阳性的判据还是会受到影响．宏蛋白质组的肽段鉴定需要优化其假阳性的质控标准．又次，在一般蛋白质组信息分析中，基于肽段的蛋白质归并处理较为简单，而在宏蛋白质组数据中，由于微生物群体里菌群谱系相近，所测定的肽段既可能匹配到同一物种的不同蛋白上，又可能匹配到不同物种的蛋白上，增大了匹配的不确定性，导致鉴定结果可信度降低．最后，宏蛋白质组需要一套归并方法的评价系统．这是因为不当的归并策略不仅直接影响蛋白质的鉴定，也会造成蛋白质丰度分布、物种组成、功能分析的偏向性．

宏蛋白质组生物信息分析的基本工作流程如图 2 所示．从总体上看，这个流程与一般蛋白质组分析过程并无二致，然而，二者的操作细节却存在很大差异．本文以这个流程为线索，结合宏蛋白质组数据分析方法和技术难点，介绍目前该领域的发展状态和研究观点．

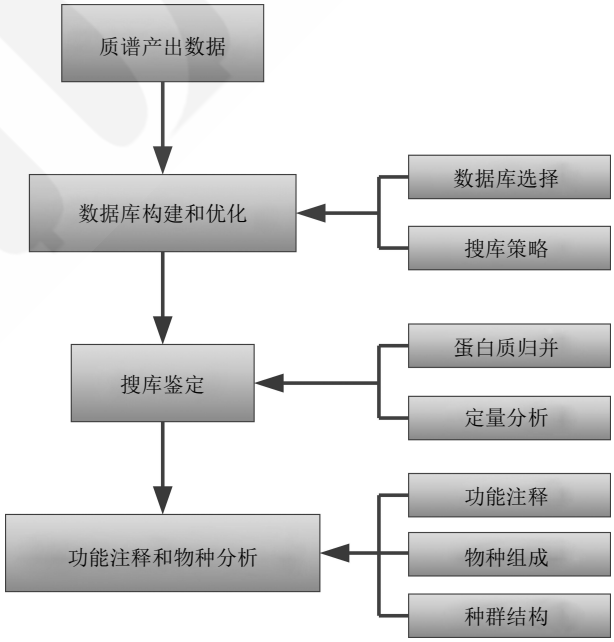

Fig. 2　The strategies of metaproteomics data analysis

图 2　宏蛋白质组数据分析基本策略

1　宏蛋白质组数据库的建立和搜索

1.1　数据库的选择

在蛋白质组学研究中，基于质谱信号的蛋白质鉴定方法之基本原则是，蛋白质数据库中肽段的理

论质量与质谱测定的实验数据之间应有完全的匹配，即所谓 Peptide Spectrum Match (PSM)原则．具体而言，是将肽段相对应的 MS/MS 谱图与数据库内的蛋白质的理论酶切肽段谱图相比较，通过算法对二者相似程度进行统计分析并予以评分，分值最高的理论肽段即作为鉴定结果．显然，这种策略在很大程度上有赖于合适的肽段数据库．对于一个生物物种而言，只要它的基因组已被测定而且相对应的基因注释也完成，那么它的蛋白质组的理论肽段数据库是比较容易建立的．但是对于一个成分复杂，而且各成分之间含量差异极大的混合微生物的宏蛋白质组分析而言，肽段数据库的建立就没有那么容易．一方面，理想的数据库应该包括样本所含所有微生物蛋白质序列信息，又不含其他多余物种的蛋白质信息，这样就需要足够大的数据库．事实上，由于环境样本的种群物种组成未知，不同分析样本之间的蛋白质差异巨大，构建理想数据库是非常困难的．另一方面，运用传统的质控方法对大数据库的搜索进行过滤可能导致大量的假阳性结果，而采用严苛的条件过滤假阳性则会产生许多假阴性结果．因此，如何选择合适的数据库是宏蛋白质组数据分析所面临的一个重大挑战．目前，三种选择策略是比较普及的．

1.1.1 选用合适的公共数据库进行蛋白质鉴定．常用的公共数据库主要包括 NCBI[11]、Uniprot[12]、Ensembl[13]等(表 1)．它们含有相对完整的物种信息，有助于鉴定样本中一些未知物种．但是，它们收集了大量其他物种的序列，容易造成搜索灵敏度下降，尤其是导致对某些菌种的信息偏好性，比如人体肠道致病菌种等[14]．

Table 1 Public reference databases in metaproteomics

表 1 宏蛋白质组中常用的公共数据库

数据库	描述	网址	文献
NCBI RefSeq	包含 62 739 个物种，16 172 490 条转录序列，70 427 238 条蛋白质序列	https://www.ncbi.nlm.nih.gov/refseq/	[11]
UniprotKB	包含细菌、古菌、病毒以及真核生物等在内的 57 860 个蛋白质组数据，共 68 493 254 条蛋白质序列	http://www.uniprot.org/uniprot/	[12]
UniRef	150 495 654 条分类信息	http://www.uniprot.org/uniref/	[15]
Ensembl	包括细菌(细菌 41 198，古菌 412)、真菌(388 个物种的 634 个基因组)、原生生物(114 个物种的 178 个基因组)等在内的基因组信息	http://asia.ensembl.org/index.html	[13]
GOS	海洋微生物，包括细菌、古细菌，病毒、真核生物. 全库 43GB，共包含 41 146 566 条蛋白质序列	http://data.imicrobe.us/project/view/26	[16−17]
HumanGut	13 个健康人的肠道微生物宏基因组	http://data.imicrobe.us/project/view/33	[18]
TARA	包含全球 68 个不同海域 243 个样本信息，样本来自每个海域的三个不同深度. 全库 13GB，共包含 4 千万条序列	http://ocean-microbiome.embl.de/companion.html	[19]
HPMC	人类胃肠道微生物宏基因组数据库，来自 1 800 个人类样本，包含超过 2 万个基因组信息	http://www.hpmcd.org/	[20]
HOMD	人类口腔微生物数据库，包含 1 322 个基因组信息，共 1GB	http://www.homd.org/	[21]

1.1.2 自定义配置获得数据库．如果研究者对样本物种组成有一定了解，则可能在公共数据库中筛选出样本可能存在的物种序列并组成新的数据库，这样能够有效缩减数据库大小，配置大小合适的数据库．例如，Russo 等[22]在研究淡水微生物种群功能与营养富集关系时，首先根据 16S rRNA 和 18S rDNA 测序方法得到样本分类信息，然后在公共数据库 Uniprot 中按物种信息筛选相关物种库文件，再自定义建库．有些研究也可根据采样地的环境信息从数据库中筛选相对应的样本信息，再进行自定义配置[8]．值得指出的是，人工选择物种一定带有某种偏好性，进而影响整体的数据鉴定质量，而且由于许多物种缺乏基因组数据，也就无法完成自定义配置数据库的建立．

1.1.3 以样本自身基因组或者转录组数据建库．由于 NGS 技术的广泛应用和测序费用大幅度下降，直接用被测微生物混合物的宏基因组或者宏转录组数据建库已成为可能．相比于前两种方法，这种方法得到的数据库与样本高度匹配，且不包含其他物种信息，有效地规避了前两种方法引起的误差．然而，该方法也有它的短处．首先，由于需要基因组或转录组测序，就必然需要大量的实验和计算工作，而 DNA 序列一般需要通过六框翻译形成蛋白质序列，使得序列数据量扩大了 6 倍；其二，原始测序数据在组装和拼接后，有些不能形成 contigs 或 scaffold 的短序列会被丢弃，进而影响鉴定正确

性．一般而言，相较于自定义基因组数据库，由于宏基因组的样本成分复杂，其测序结果质量低于自定义数据库数据质量．针对这一问题，Cantarel 等[23]采用测序样本未经拼接的序列直接作为数据库，获得了比拼接后数据库更多序列．考虑到未拼接序列过多，建库的计算量过大，May 等[24]在 Cantarel 方法的基础上发展了基于宏多肽组学的方法，用于建立相应的数据库．他们将测序得到的读长序列直接进行预测并翻译成多肽序列，对多肽序列进一步筛选和过滤，并建成相应样本的宏多肽组学数据库．据报道，相较于传统的基因组数据建库，该方法能鉴定到更多的肽段序列．

综上所述，在现行的宏蛋白质组信息分析中并没有一个广泛接受的数据库选择策略，在实践分析中，还是具体样本具体分析．一般而言，采用递进式选择方案，多点测试不同数据库的搜索结果，可能会获得较为理想的分析结果．

1.2 搜库策略

既然宏蛋白质组信息分析面临大数据库的搜索，那么随之而来的问题是传统搜库方法是否适应这种需求．答案是否定的．如果将传统单步搜库方法应用于宏蛋白质组数据，既需要较长的搜库时间，又存在较高的假阳性率．目前的宏蛋白质组研究广泛采用迭代搜库方法，它可以有效地缩减数据库大小，节省搜库时间，而且蛋白质鉴定的灵敏度和数量均有提升．例如，Jagtap 等[25]提出两步搜库法对数据库进行优化，其原理是在第一次搜库过程中不采用任何过滤参数设置进行搜库，并把第一次的搜库结果作为第二次搜索的库文件，然后根据库文件建立诱饵库进行常规搜库．通过如此处理使得库文件显著压缩，也降低了假阴性率和提高了鉴定数据数目和质量．在该基础上，Zhang 等[26]提出了三步迭代搜库法，通过增加搜库次数对库文件进行进一步的缩小和优化，提高搜索效率和鉴定结果的准确性．他们建立了用于肠道微生物种群研究的流程 MetaPro-IQ，在第一次搜库时不建立诱饵库，将搜库的结果文件作为新的库文件进行二次搜库，同时建立诱饵库和进行 FDR 过滤，二次搜库产生的结果文件再进行常规的搜库，将最终的搜库结果作为鉴定结果．此类迭代搜库方式极大地缩短了宏蛋白质组学在使用庞大数据库进行搜库鉴定所需要的时间．此外，Chatterjee 等[27]建立了一种基于分布式文件存储数据库 MongoDB 的建库搜索方法 ComPIL，将存储于蛋白质数据库 ProtDB 中的蛋白质序列进行模拟酶切后分别按照肽段质量和序列进行分组，并存储到 MassDB 和 SeqDB 中．这种方式能够将搜索任务平行分配到计算机集群中的多个节点，在计算节点资源充足时，数据库可以无限扩大，使得其用于大数据量、大数据库的宏蛋白质组搜库鉴定时有明显的优势．迭代搜库策略固然有效，然而究竟多少次迭代最为理想？其搜索结果的评价标准是什么？目前并未取得广为接受的质控判据．

1.3 质量控制

谱图质量高低不均、随机匹配的非正态分布、污染序列的存在等因素均可造成蛋白质鉴定结果的 PSM 假阳性匹配，所以，鉴定结果的质量控制是必要的．在传统蛋白质组搜库中，假阳性率控制是通过靶向库 / 诱饵库的质控方法得以实现的[28]．该方法的原理是构建一个虚拟的诱饵库，其具有与靶向库序列相似的氨基酸分布，长度分布，且与靶向库序列没有重叠的特点．构建诱饵库主要是通过随机序列库和反向序列库两种方法进行的．通过 PSM 匹配到诱饵库的假阳性匹配来计算整体 PSM 的置信度水平．由于靶向库和诱饵库的序列数目相等，匹配到靶向库和诱饵库的几率相等，通过计算假阳性率值(FDR)可以对鉴定结果做出可信评估，选定的 FDR 阈值来对鉴定数据进行判断，最终获得可信度高的匹配．FDR 评估是建立在合适数据库的基础之上的，可是 FDR 概念在宏蛋白质组学分析中遭遇了困难．问题的关键在于其数据库文件明显增大，而大的数据库会存在高度序列相似性，导致真假阳性结果的辩识更加困难，需要提高打分阈值来过滤掉假阳性结果；但是严格的过滤条件又会造成假阴性率上升．所以，传统的 FDR 衡量标准无法完全满足宏蛋白质组的搜库要求．从计算操作的角度来讲，构建大数据库的反库会消耗大量的计算资源，导致后期的搜库困难．为了解决这个问题，研究者一般从两个角度探索质控条件：a．选择大小合适且信息完整的数据库，并测试评估不同的 FDR 阈值；b．发展 FDR 非依赖性的质控方法，例如基于半监督机器学习算法的 Percolator 软件等正在逐步应用于鉴定结果的质控中[29]．

2 宏蛋白质组数据分析中的归并和定量

2.1 蛋白质鉴定

当前在宏蛋白质组研究领域，蛋白质的鉴定方法与传统方法基本是一致的，主要建立在数据库的

PSM 策略．其工作原理是将 MS/MS 谱图与数据库内的蛋白质经过理论酶切得到的虚拟谱图相比较，并通过一定的算法对二者相似程度进行打分，分值最高的图谱所对应的理论肽段即作为鉴定肽段．如果通过 *de novo* 方法获得短的蛋白质序列标签，再利用序列标签进行搜库鉴定，可以有效地缩减搜库时间．常用的 PSM 搜索软件有 MASCOT[30]、X!Tandem[31]、OMSSA[32]、SEQUEST[33] 和 Andromeda[34] 等．通过结合两种或多种软件进行数据库搜索，可以有效地提高蛋白质鉴定数目，减少假阳性鉴定结果．例如，OpenMS[35]、TPP[36]、Galaxy-P[37]、MPA[38] 等分析流程可以整合不同的搜库工具得到较为准确的鉴定结果．除上述搜库工具之外，一些专门针对宏蛋白质组数据鉴定的软件也发展起来．例如，搜库软件 ComPIL 解决了宏蛋白质组庞大数据量的搜库问题[27]．一种新的搜索算法 Sipros (http://siprosensemble.omicsbio.org/)使用大量的 CPU 计算资源进行混合并行运算，在开展大量数据搜索工作时具有很好的可扩展性，从而在宏蛋白质组数据集鉴定中显现优势[39]．除了数据库鉴定工具外，还有一些针对宏蛋白质组分析的综合应用软件也在逐步开发中．Galaxy-P 流程是由 Jagtap 等[37]开发的，专门用于宏蛋白质组学数据处理的流程，该流程集合了不同的模块，包括数据库预处理、谱图匹配、数据筛选过滤等．针对不同的生物学问题，用户可以灵活地选择适合的鉴定和后续分析方法．MPA 整合了多种数据库搜索工具，包含从谱图鉴定到物种、功能分析的整套流程模块．它还提供了用户可视化界面，方便用户调节分析流程并通过不同的可视化方法对结果进行分析[38]．针对宏蛋白质组学常见的宏蛋白质组的蛋白质搜索和鉴定软件列于表 2.

Table 2 Protein identification tools in metaproteomics

表 2 宏蛋白质组学中常用的蛋白质鉴定工具

	分析工具	授权	简介	相关宏蛋白质组学研究应用	文献
搜库工具	Mascot	商业	数据库搜索工具	不同季节海洋浮游细菌种群变化研究[9]、沼气池微生物种群研究[40]、淡水微生物种群研究[22]、土壤微生物种群研究[10]	[30]
	X!Tandem	开源	数据库搜索工具	人体肠道微生物种群研究[5]、海洋病毒种群功能研究[41]	[31]
	OMSSA	开源	数据库搜索工具	肥胖和正常体重人群肠道微生物种群比较[6]	[32]
	SEQUEST	商业	数据库搜索工具	近海上升流中藻类环境蛋白质组研究[8]、口腔唾液微生物种群研究[42]、河口水生微生物种群研究[43]、不同季节海洋浮游细菌种群变化研究[9]、海洋低氧区微生物种群研究[44]、不同营养环境海洋微生物种群研究[45]、人体肠道微生物种群研究[5, 7]、老鼠粪便微生物种群研究[46]	[33]
	Andromeda(MaxQuant)	开源	MaxQuant 软件的数据库搜索工具	口腔唾液微生物种群研究[47]、人体肠道微生物种群研究[5]	[34]
软件包	OpenMS	开源	包含 185 种数据分析工具，涵盖蛋白质组学分析各个层面	沼气池中互养的醋酸盐氧化细菌的研究[48]	[35]
	Maxquant	免费	操作相对简单，用于蛋白质鉴定及定量工作，是非标定量的常用软件	口腔唾液微生物种群研究[42]	[49]
	Trans-proteomic pipeline	开源	包括格式转换、谱图鉴定、定量、蛋白质归并等一系列工具	海洋病毒种群功能研究[41]	[36]
	Thermo protome discoverer	商业	包括多种搜库工具，可用于蛋白质鉴定和定量分析，PTM 鉴定，GO 功能注释分析等	河口水生微生物种群研究[43]、不同季节海洋 浮游细菌种群变化研究[9]、土壤微生物种群研究[10]	
	MetaProteomeAnalyzer (MPA)	免费	整合四种搜库工具结果，包括从蛋白质鉴定到功能注释等一系列分析工具	沼气池微生物种群研究[40]	[38]
	Sipros/ProRata*	开源	最初设计用于蛋白质组稳定同位素以及氨基酸突变的搜索，后被设计用于宏蛋白质组学数据搜索	人肠道微生物代谢功能和代谢活动研究[50]	[39]
	Galaxy-P*	开源	结合多种流程分析模块，主要用于宏蛋白质组数据分析	人口腔微生物种群研究[42]	[37]
	ComPIL*	开源	优化了数据库搜索方式，实现对大数据库的快速鉴定	人体肠道微生物种群研究[5, 7]	[27]

* 表示针对宏蛋白质组数据开发的软件和流程.

2.2 蛋白质归并

利用搜索引擎对质谱谱图进行搜库，会产生若干与二级质谱信号相匹配的肽段信息．由于一个肽段可能匹配到多个不同的蛋白质，因此根据这些肽段信息来进行蛋白质鉴定时，就需要确定这些肽段归属于哪一个蛋白质，即所谓蛋白质归并过程．在宏蛋白质组中，由于样品本身的复杂性，使检测到的肽段既有可能匹配到同一物种的不同蛋白质上，又有可能匹配到不同物种的蛋白质上，这显然增加了蛋白质归并的难度．因此，在宏蛋白质组研究中，普遍采用了“归组”的策略，即将序列相似度或者共享肽段数满足一定要求的蛋白序列归并到一个组中，以组为单位进行后续的物种分析和功能分析．目前针对宏蛋白质组学的蛋白质归并算法尚未成型，宏蛋白质组学数据分析中的归并处理仍旧沿用传统蛋白质组学的方法，比如 PAnalyzer[51]、ProteinProphet[52]、Protein-inference[53-54]、Scaffold[55]、IDPicker[56]以及一些特殊的算法流程等．一些较为常见的简并软件见表 3．

Table 3 Protein inference related tools in metaproteomics

表 3 宏蛋白质组学中常用的蛋白质归并相关的工具

分析工具	授权	运行平台	搜索工具	简介	文献
PAnalyzer	免费	Win/Linux/Unix/Mac OS	OMASSA X!Tandem	将肽段标记为 3 类，用来将鉴定到的蛋白质区分为 4 类，并可以对分类进行过滤，提高结论性蛋白质的可信度 https://code.google.com/archive/p/ehu-bio/wikis/PAnalyzer.wiki	[51]
ProteinProphet	开源	Linux/Unix	Mascot OMASSA X!Tandem	支持所有鉴定软件的输出格式，利用期望最大算法计算肽段检测置信度，并用来支持蛋白质可信度 http://proteinprophet.sourceforge.net/	[52]
Scaffold	商业	Win/Linux/Unix/Mac OS	兼容多个工具输出格式	利用不同搜索软件之间共有的结果来计算蛋白质可信度 http://www.proteomesoftware.com/products/scaffold/	[55]
IDPicker	开源	Win	兼容多个工具输出格式	根据酶切位点和电荷数，将匹配到的肽段分类，根据权重计算得分后，将每个类别中通过打分阈值的肽段用于蛋白质归并	[56]
Protein-inference	开源	Linux	OMSSA X!Tandem	基于 C++ 的分析方法，利用改进的线性规划模式缩减肽段丰度后，计算丰度推断值为 0 的蛋白质，即为样品中不存在的蛋白质 https://code.google.com/archive/p/protein-inference/	[54]
MassSieve	开源	Win32/Linux/Unix/Mac OS	兼容多个工具输出格式	基于 Java 语言编写的分析流程，利用输入文件中的肽段期望值或者相关系数进行过滤 https://www.ncbi.nlm.nih.gov/staff/slottad/MassSieve/	[57]
DBParser	开源	Win/Linux	Mascot	基于 Perl 语言编写的分析流程 https://omictools.com/dbparser-tool http://www.tycho.iel.unicamp.br/~tycho/apps/dbparser-files/	[58]
Fido (Percolator)	开源	Win32/Linux/Mac OS	兼容多个工具输出格式	独立软件 http://percolator.ms/ 整合软件包 http://crux.ms/	[59]

PAnalyzer 首先将质谱鉴定到的肽段依据蛋白质匹配情况进行分类标记：特异性肽段、差异性肽段、非差异性肽段等；然后依据此分类信息再进一步分类标记：结论性蛋白质(有特异性肽段支持的蛋白质)、模糊蛋白质组(共享差异性肽段的蛋白质)、不可区分蛋白质组(蛋白质组之间共享肽段的蛋白质)以及非结论性蛋白质(只有非差异性肽段支持的蛋白质)等；最后，在分类基础上逐一简并蛋白质[51].

ProteinProphet 可以用来处理几乎所有质谱鉴定软件输出的鉴定结果；它将所鉴定的肽段得分换算为肽段检测置信度，再转换为蛋白质鉴定置信度，然后依据蛋白质鉴定置信度进行蛋白质归并[52].

Protein-inference 是在线性规划模式蛋白质定量的基础上开发的一款归并算法[53-54]．何增有等[54]将蛋白质归并当作特殊的蛋白质定量问题，用蛋白

质定量的方法来解决蛋白质归并的难题．他们引入“蛋白质对肽段丰度值的贡献值”作为变量，建立起肽段丰度与蛋白质丰度推断值之间的换算关系；将肽段丰度值进行缩减后，可以通过计算得到丰度推断值为 0 的蛋白质，并将肽段归并到丰度推断值不为 0 的蛋白质中．

2.3 蛋白质定量

开展微生物种群功能的探究，仅仅依赖蛋白质鉴定层面信息是远远不够的，还需要结合定量层面信息．在宏蛋白质组的定量分析中常采用非标记和代谢标记两种．非标定量分为谱图计数法和信号强度法．谱图计数法基于丰度越高的蛋白质则在质谱鉴定时得到的一级谱图数量越多的假设．为了避免分子质量大的蛋白质可产生较多肽段谱图的计算偏差，Zybailov 等[60]提出了 NSAF(normalized spectral abundance factor)衡量指标，将蛋白质序列长度作为参考值对谱图计数结果进行校正．基于该方法的分析软件有 Scaffold、ProteoIQ、APEX、CENSUS[61]等．信号强度法则利用肽段信号强度表征肽段丰度，根据重构离子峰(XIC)的峰面积对肽段进行定量．基于该方法的分析软件有 MaxQuant[49]、OpenMS[35]等．代谢标记定量则是对微生物群落进行肽段标记，进而计算蛋白质的丰度．例如，蛋白质稳定同位素探测(protein-SIP)是通过标记微生物底物中的 C、N、S 等元素来研究重标同位素参入微生物蛋白质的相对同位素丰度(RIA)．由于稳定同位素的参入程度不同可导致肽段质量产生偏移，一般的搜库工具不能对标记肽段进行准确鉴定，Sachsenberg 等[62]开发了软件 MetaProSIP 用以对 Protein-SIP 标记肽段进行鉴定，同时准确计算 RIA 等．Zhang 等[63]建立了基于 ^{15}N 标记的肠道微生物种群定量方法，用以研究肠道微生物种群功能．代谢标记定量的常用分析软件有 TPP[64]、Patternlab[65]、Census[61−62]．

3 发掘宏蛋白质组数据所蕴含的生物学意义

3.1 宏蛋白质组的功能注释

与传统蛋白质组数据的功能分析相类似，宏蛋白质组学也利用基因功能数据库，对所鉴定蛋白质进行功能注释．较为常用的基因功能数据库列于表 4．进行功能注释时需要一些工具软件，较为常用的有 Blast2GO[66]、DAVID[67]等．Blast2GO 采用 GO 或 COG 数据库进行功能注释，而 DAVID 则从 GO 数据库扩展到 KEGG 功能通路以及蛋白质互作等多个方面进行注释．

Table 4 Databases for functional annotation

表 4 功能注释常用数据库

数据库	简介	URL	文献
Gene Ontology	注释信息包括细胞组分、分子功能和生物学过程三个部分.	http://geneontology.org/	[68]
KEGG	除蛋白质功能以外还包含已知代谢通路的信息.	http://www.genome.jp/kegg/ http://www.kegg.jp/	[69]
COG	包含原核生物(COG)和真核生物(KOG)的数据库.	http://www.ncbi.nlm.nih.gov/COG/ ftp://ftp.ncbi.nih.gov/pub/COG/	[70]
Swiss-Prot	UniprotKB 中经过评估的部分.	http://www.uniprot.org/uniprot/?query=reviewed%3Ayes/ http://www.uniprot.org/downloads/	[71]
SMART	自动鉴定与注释结构域，包括常规模式和基因组模式，分别包含不同的数据库.	http://smart.embl-heidelberg.de/	[72−73]
InterPro	用于蛋白质以及基因组的分类与自动注释的综合性数据库.	http://www.ebi.ac.uk/interpro/	[74]

3.2 依据宏蛋白质组数据开展物种分析

微生物的多样性、种群结构、进化关系、功能活性、功能互作以及与环境之间的相互关系等议题是宏基因组和宏蛋白质组学研究的重点．在这些议题中，首先应当明确样本中的微生物构成．宏基因组学已经开发了多种进行物种分析的方法，比如依赖 Blast 序列比对的 MEGAN[75]、MetaPhyler[76]，基于概率的 PhymmBL[77]，高精度高灵敏度的朴素贝叶斯分类工具 NBC[78]，利用 *k-mers* 的快速分类工具 Kraken、CLARK[79−80]等．宏蛋白质组学中的物种分析的基本思路来自于宏基因组学．目前宏蛋白质组学物种分析软件包括 Unipept[81]、MEGAN[75, 82]、

MPA[38]等．Unipept 和 MEGAN 引入了宏基因组学数据处理中所采用的最近公共祖先算法(LCA，lowest common ancestor algorithm)[83]，以此排除分类时会出现的分类错排、鉴定错误以及误差等，实现较为准确的物种分析．MPA 采用的非 LCA 分析方法是直接向蛋白鉴定结果中添加物种信息来实现分类．

Unipept(http://unipept.ugent.be)是一个网页版物种分析工具[81]．它利用 NCBI 分类数据库[84]和 UniProtKB[85]数据库，依据 NCBI 分类的层级结构和不同层级的排序信息对肽段序列进行搜索，运用 LCA 算法进行物种鉴定，选择保留特异性高的分类记录，避免分类信息缺失带来的鉴定误差．Unipept 输出的结果中不仅包含了肽段匹配到的蛋白质的基本信息，还包括鉴定到的蛋白质可能来源的物种分类信息．Unipept 可以生成的动态矩形树状图，直观地显示每个分类节点匹配到的肽段数目以及宏蛋白质组样品中存在的物种等信息．

MEGAN(https://github.com/danielhuson/megan-ce)是一个基于宏基因组软件所拓展的宏蛋白质组学数据分析工具[75, 82]．将宏蛋白质组数据鉴定到的蛋白质 BLAST[86]序列比对结果导入 MEGAN，再利用 LCA 算法将每条序列依据其匹配信息进行分类，可以产生在 NCBI 不同分类水平上匹配到该分类单元的序列数目．同时，MEGAN 还可以依据 InterPro2GO、SEED、eggNOG、KEGG 等功能注释进行分类．MEGAN 能够整合多种绘图工具，可以将所分类的结果以图表形式直观展示出来，同时也可以导出文本格式，便于对分类结果作进一步分析．

MPA(https://github.com/compomics/meta-proteome-analyzer)是一个可以在 Win、Linux 平台上运行的宏蛋白质组分析软件．它直接向蛋白质鉴定结果中添加分类信息等进行物种分析．通过多个不同的数据库搜索后，MPA 将相关的谱图信息、鉴定结果以及注释信息等关联的数据库进行整合存储；同时添加来自 Uniprot 的相关信息，NCBI 分类数据库中的分类信息，以及 KEGG 中代谢通路信息等．最后，MPA 综合这些信息开展物种鉴定．此外，MPA 含有开源的图形数据库 Neo4j(www.neo4j.com)，通过这个数据库可以生成代表性节点变量和相互关系的图表，便于有目的地过滤蛋白质[38]．

3.3 宏蛋白质组数据的整合分析

在微生物群落中，物种间广泛存在共生和互养现象．因此，单独进行某一物种深入研究并不能深刻揭示自然界微生物的发生和发展．宏基因组和宏蛋白质组的研究之所以得到研究者的青睐，正是因为它们所关注的对象是自然界存在的微生物混合群落，而不是实验室分离或培养的产物．宏基因组学已经被广泛应用到自然微生物环境研究之中．例如，从土壤环境微生物研究中发现新型活性分子，从人肠道微生物菌群的研究中发现肠道菌群与肥胖和Ⅱ型糖尿病等疾病的相关性等[87-90]．宏蛋白质组学的数据发掘更加注重对样本整体的生物特征的研究，注重物种组成与样本代谢功能的关系，注重样本与环境之间的互动．一般而言，获取宏蛋白质组数据之后，应根据已知的蛋白质鉴定和功能信息，对应不同的微生物物种，再结合蛋白质丰度和环境参数，寻找微生物群落可能的代谢方式特征，进一步扩展微生物群落与环境间的相互作用，揭示微生物物种的高度环境性的适应性．例如，Wilmes 等[3]首次利用双向电泳分离具有生物除磷作用的活性污泥中的微生物蛋白质，并利用质谱技术对分离到的蛋白质进行了鉴定．在这个基础上，Wilmes 等[91]的进一步实验发现，具有强生化除磷功能的污泥中参与聚羟基脂肪酸酯合成、脂肪酸氧化以及乙醛酸 / TCA 循环等代谢过程的蛋白质高表达，为强化生物除磷功能中的化学变化与代谢过程的关联提供了直接的证据，揭示了活性污泥中生物除磷的代谢模型，为改进污水生物处理起到了很好的指导作用．Colatriano 等[43]对不同深度的河口微生物进行宏蛋白质组研究．他们发现，不同深度的水域中含有不同的微生物种群，具有不同的蛋白质分布：表层微生物高丰度蛋白质以有机物转运蛋白质为主，而深层微生物的高丰度蛋白质多参与碳氮固定和循环．他们结合不同层面的环境因素，进一步阐明表层富营养水域的微生物主要功能在于吸收大小分子有机物，而深层寡营养区的微生物则主要通过合成通路将无机物转化为有机物．宏蛋白质组研究显著地提升了人们对河口微生物分布及其功能的了解，对于河口的环境治理提供了有价值的参考．宏蛋白质组学在消化道微生物中的研究应用也非常广泛．人消化道菌群功能的研究首先利用宏蛋白质组学方法建立复杂度较低的无菌小鼠肠道菌群模型，并依此来建立婴儿新生消化道菌群模型，在这个基础上最终完成成年人消化道菌群模型的建立[50]．Kolmeder 等[6]的肥胖人群结肠宏蛋白质组学的研究发现糖代谢相关蛋白质在肥胖患者和正常人群之间存在明显

差异，正常人群结肠宏蛋白质组中参与五碳六碳糖代谢的酶丰度更高，而在肥胖患者的结肠宏蛋白质组中丰度较高的是淀粉和果胶代谢相关的蛋白质．除蛋白质功能的差异之外，在物种分布上与正常人群相比在肥胖人群的结肠微生物菌群中的拟杆菌门细菌的含量较少却更活跃．另外研究还发现肠道菌群的变化还会对胰岛素的敏感度产生影响．人体的多种疾病，比如肥胖、Ⅱ型糖尿病等，与肠道微生物菌群有着密切的联系[90].

4 宏蛋白质组信息学发展前景

宏蛋白质组学在微生物种群功能研究中所发挥的作用正在受到科学界越来越多的关注．可以预计，宏蛋白质组学将是在未来的微生物种群研究中不可取代的重要工具之一．当然，与之相对应的宏蛋白质组信息学的发展需求也是迫切的．由于研究对象的特殊性，宏蛋白质组学数据分析必须建立一套与传统蛋白质组学不同的统计和数据模型，才能更好地解析数据，探索大数据所蕴含的生物学涵义．首先，由于测序成本的显著下降，大量的微生物种群的序列信息必然不断被揭示，微生物种群的基因组数据库将大大完善，这将促使宏蛋白质组的蛋白质鉴定速度和准确性得到革命性改变．越来越多的宏蛋白质组信息分析将会更加密切地结合宏基因组和宏转录组数据．如何有效地整合多组学数据，探索这些数据对于解析生物问题的贡献将是一个不可回避的信息科学的挑战．其次，随着质谱精度和通量的不断提高，谱图鉴定的结果更加可靠，同时高质量的谱图使得通过 *de novo* 方法进行蛋白质鉴定的准确性提升．但是，*de novo* 的肽段拼接算法仍存在很大的改进空间．如何建立一个非基因组数据依赖性的蛋白质鉴定分析技术仍然长路漫漫．其三，微生物群体的蛋白质组变化必然引发它的代谢网络改变，产生完全不同的代谢产物．目前，代谢组学的分析方法有了长足的发展，不过宏蛋白质组数据与代谢组数据的整合分析仍需要更多的关注，尤其是发展信息分析的工具．其四，目前常用的宏蛋白质组信息学的处理方法大多采用宏基因组或是传统蛋白质组的固有模式．例如，在蛋白质的定量数据分析上，基本的分析手段与一般定量方法无异，这样的处理可能忽略了微生物混合种群里不同基因组背景下微生物基因表达的特征性．又如，在微生物物种分析中，目前宏蛋白质组的分析方法基本沿用了宏基因组的算法，这些算法究竟是否符合蛋白质数据的分布特点并没有得到系统的评估．这样看来，一整套适应于宏蛋白质组数据处理的信息学系统还是受人期待，必有用武之地的．

参 考 文 献

[1] Gill S R, Pop M, Deboy R T, *et al*. Metagenomic analysis of the human distal gut microbiome. Science, 2006, **312** (5778): 1355-1359

[2] Rodriguez-Valera F. Environmental genomics, the big picture? FEMS Microbiology Letters, 2004, **231**(2): 153-158

[3] Wilmes P, Bond P L. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. Environ Microbiol, 2004, **6**(9): 911-920

[4] Kan J, Hanson T E, Ginter J M, *et al*. Metaproteomic analysis of Chesapeake Bay microbial communities. Saline Systems, 2005, **1**: 7

[5] Tanca A, Palomba A, Fraumene C, *et al*. The impact of sequence database choice on metaproteomic results in gut microbiota studies. Microbiome, 2016, **4**(1): 51

[6] Kolmeder C A, Ritari J, Verdam F J, *et al*. Colonic metaproteomic signatures of active bacteria and the host in obesity. Proteomics, 2015, **15**(20): 3544-3552

[7] Verberkmoes N C, Russell A L, Shah M, *et al*. Shotgun metaproteomics of the human distal gut microbiota. ISME J, 2009, **3**(2): 179-189

[8] Sowell S M, Abraham P E, Shah M, *et al*. Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. ISME J, 2011, **5**(5): 856-865

[9] Georges A A, El-Swais H, Craig S E, *et al*. Metaproteomic analysis of a winter to spring succession in coastal northwest Atlantic Ocean microbial plankton. ISME J, 2014, **8**(6): 1301-1313

[10] Zampieri E, Chiapello M, Daghino S, *et al*. Soil metaproteomics reveals an inter-kingdom stress response to the presence of black truffles. Sci Rep, 2016, **6**: 25773

[11] Pruitt K D, Tatusova T, Maglott D R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res, 2005, **33** (Database issue): D501-504

[12] Apweiler R, Bairoch A, Wu C H, *et al*. UniProt: the universal Protein knowledgebase. Nucleic Acids Res, 2004, **32** (Database issue): D115-119

[13] Yates A, Akanni W, Amode M R, *et al*. Ensembl 2016. Nucleic Acids Res, 2016, **44**(D1): D710-716

[14] Noble W S. Mass spectrometrists should search only for peptides they care about. Nat Methods, 2015, **12**(7): 605-608

[15] Suzek B E, Huang H, Mcgarvey P, *et al*. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics, 2007, **23**(10): 1282-1288

[16] Rusch D B, Halpern A L, Sutton G, *et al*. The Sorcerer Ⅱ Global

Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol, 2007, **5**(3): e77

[17] Parthasarathy H, Hill E, Maccallum C. Global ocean sampling collection. PLoS Biol, 2007, **5**(3): e83

[18] Kurokawa K, Itoh T, Kuwahara T, *et al*. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res, 2007, **14**(4): 169-181

[19] Pesant S, Not F, Picheral M, *et al*. Open science resources for the discovery and analysis of Tara Oceans data. Sci Data, 2015, **2**: 150023

[20] Forster S C, Browne H P, Kumar N, *et al*. HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. Nucleic Acids Res, 2016, **44**(D1): D604-609

[21] Chen T, Yu W H, Izard J, *et al*. The human oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database (Oxford), 2010, **2010**: baq013

[22] Russo D A, Couto N, Beckerman A P, *et al*. A Metaproteomic analysis of the response of a freshwater microbial community under nutrient enrichment. Front Microbiol, 2016, **7**: 1172

[23] Cantarel B L, Erickson A R, Verberkmoes N C, *et al*. Strategies for metagenomic-guided whole-community proteomics of complex microbial environments. PLoS One, 2011, **6**(11): e27173

[24] May D H, Timmins-Schiffman E, Mikan M P, *et al*. An alignment-free "metapeptide" strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. J Proteome Res, 2016, **15**(8): 2697-2705

[25] Jagtap P, Goslinga J, Kooren J A, *et al*. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics, 2013, **13**(8): 1352-1357

[26] Zhang X, Ning Z, Mayne J, *et al*. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. Microbiome, 2016, **4**(1): 31

[27] Chatterjee S, Stupp G S, Park S K, *et al*. A comprehensive and scalable database search system for metaproteomics. Bmc Genomics, 2016, **17**(1): 642

[28] Elias J E, Gygi S P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods, 2007, **4**(3): 207-214

[29] Gonnelli G, Stock M, Verwaeren J, *et al*. A decoy-free approach to the identification of peptides. J Proteome Res, 2015, **14**(4): 1792-1798

[30] Perkins D N, Pappin D J C, Creasy D M, *et al*. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 1999, **20**(18): 3551-3567

[31] Craig R, Beavis R C. TANDEM: matching proteins with tandem mass spectra. Bioinformatics, 2004, **20**(9): 1466-1467

[32] Geer L Y, Markey S P, Kowalak J A, *et al*. Open mass spectrometry search algorithm. J Proteome Res, 2004, **3**(5): 958-964

[33] Yates Ⅲ J R, Eng J K, Mccormack A L, *et al*. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem, 1995, **67**(8): 1426-1436

[34] Cox J, Neuhauser N, Michalski A, *et al*. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res, 2011, **10**(4): 1794-1805

[35] Rost H L, Sachsenberg T, Aiche S, *et al*. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods, 2016, **13**(9): 741-748

[36] Deutsch E W, Mendoza L, Shteynberg D, *et al*. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. Proteomics Clin Appl, 2015, **9**(7-8): 745-754

[37] Jagtap P D, Blakely A, Murray K, *et al*. Metaproteomic analysis using the Galaxy framework. Proteomics, 2015, **15**(20): 3553-3565

[38] Muth T, Behne A, Heyer R, *et al*. The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. J Proteome Res, 2015, **14**(3): 1557-1565

[39] Wang Y, Ahn T H, Li Z, *et al*. Sipros/ProRata: a versatile informatics system for quantitative community proteomics. Bioinformatics, 2013, **29**(16): 2064-2065

[40] Heyer R, Benndorf D, Kohrs F, *et al*. Proteotyping of biogas plant microbiomes separates biogas plants according to process temperature and reactor type. Biotechnol Biofuels, 2016, **9**: 155

[41] Brum J R, Ignacio-Espinoza J C, Kim E H, *et al*. Illuminating structural proteins in viral "dark matter" with metaproteomics. Proc Natl Acad Sci USA, 2016, **113**(9): 2436-2441

[42] Xie H, Onsongo G, Popko J, *et al*. Proteomics analysis of cells in whole saliva from oral cancer patients *via* value-added three-dimensional peptide fractionation and tandem mass spectrometry. Mol Cell Proteomics, 2008, **7**(3): 486-498

[43] Colatriano D, Ramachandran A, Yergeau E, *et al*. Metaproteomics of aquatic microbial communities in a deep and stratified estuary. Proteomics, 2015, **15**(20): 3566-3579

[44] Hawley A K, Brewer H M, Norbeck A D, *et al*. Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. Proc Natl Acad Sci USA, 2014, **111**(31): 11395-11400

[45] Morris R M, Nunn B L, Frazar C, *et al*. Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. ISME J, 2010, **4**(5): 673-685

[46] Wu J, Zhu J, Yin H, *et al*. Development of an integrated pipeline for profiling microbial proteins from mouse fecal samples by LC-MS/MS. J Proteome Res, 2016, **15**(10): 3635-3642

[47] Belstrom D, Jersie-Christensen R R, Lyon D, *et al*. Metaproteomics of saliva identifies human protein markers specific for individuals with periodontitis and dental caries compared to orally healthy controls. PeerJ, 2016, **4**: e2433

[48] Mosbaek F, Kjeldal H, Mulat D G, *et al*. Identification of syntrophic acetate-oxidizing bacteria in anaerobic digesters by combined

protein-based stable isotope probing and metagenomics. ISME J, 2016, **10**(10): 2405-2418

[49] Sano T, Huang W, Hall J A, *et al*. An IL-23R/IL-22 circuit regulates epithelial serum amyloid A to promote local effector Th17 responses. Cell, 2015, **163**(2): 381-393

[50] Xiong W, Abraham P E, Li Z, *et al*. Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. Proteomics, 2015, **15**(20): 3424-3438

[51] Prieto G, Aloria K, Osinalde N, *et al*. PAnalyzer: a software tool for protein inference in shotgun proteomics. BMC Bioinformatics, 2012, **13**: 288

[52] Nesvizhskii A I, Keller A, Kolker E, *et al*. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem, 2003, **75**(17): 4646-4658

[53] Huang T, Wang J, Yu W, *et al*. Protein inference: a review. Brief Bioinform, 2012, **13**(5): 586-614

[54] He Z, Huang T, Liu X, *et al*. Protein inference: A protein quantification perspective. Comput Biol Chem, 2016, **63**: 21-29

[55] Searle B C. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. Proteomics, 2010, **10**(6): 1265-1269

[56] Ma Z Q, Dasari S, Chambers M C, *et al*. IDPicker 20: Improved protein assembly with high discrimination peptide identification filtering. J Proteome Res, 2009, **8**(8): 3872-3881

[57] Slotta D J, Mcfarland M A, Markey S P. MassSieve: panning MS/MS peptide data for proteins. Proteomics, 2010, **10**(16): 3035-3039

[58] Yang X, Dondeti V, Dezube R, *et al*. DBParser: web-based software for shotgun proteomic data analyses. J Proteome Res, 2004, **3**(5): 1002-1008

[59] Serang O, Maccoss M J, Noble W S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. J Proteome Res, 2010, **9**(10): 5346-5357

[60] Zybailov B, Mosley A L, Sardiu M E, *et al*. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. J Proteome Res, 2006, **5**(9): 2339-2347

[61] Park S K, Venable J D, Xu T, *et al*. A quantitative analysis software tool for mass spectrometry-based proteomics. Nat Methods, 2008, **5**(4): 319-322

[62] Sachsenberg T, Herbst F-A, Taubert M, *et al*. MetaProSIP: automated inference of stable isotope incorporation rates in proteins for functional metaproteomics. Journal of Proteome Research, 2015, **14**(2): 619-627

[63] Zhang X, Ning Z, Mayne J, *et al*. *In vitro* metabolic labeling of intestinal microbiota for quantitative metaproteomics. Anal Chem, 2016, **88**(12): 6120-6125

[64] Palmblad M, Bindschedler L V, Cramer R. Quantitative proteomics using uniform (15)N-labeling, MASCOT, and the trans-proteomic pipeline. Proteomics, 2007, **7**(19): 3462-3469

[65] Carvalho P C, Yates Iii J R, Barbosa V C. Analyzing shotgun proteomic data with PatternLab for proteomics. Curr Protoc Bioinformatics, 2010, Chapter **13**: 1-15

[66] Conesa A, Gotz S, Garcia-Gomez J M, *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 2005, **21**(18): 3674-3676

[67] Huang Da W, Sherman B T, Lempicki R A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc, 2009, **4**(1): 44-57

[68] Ashburner M, Ball C A, Blake J A, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, **25**(1): 25-29

[69] Kanehisa M, Goto S, Sato Y, *et al*. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res, 2012, **40**(Database issue): D109-114

[70] Tatusov R L. A genomic perspective on protein families. Science, 1997, **278**(5338): 631-637

[71] Boutet E, Lieberherr D, Tognolli M, *et al*. UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. Plant bioinformatics: methods and protocols, 2007, 89-112

[72] Schultz J, Milpetz F, Bork P, *et al*. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci USA, 1998, **95**(11): 5857-5864

[73] Letunic I, Doerks T, Bork P. SMART: recent updates, new developments and status in 2015. Nucleic Acids Res, 2015, **43**(Database issue): D257-260

[74] Apweiler R, Attwood T K, Bairoch A, *et al*. InterPro--an integrated documentation resource for protein families, domains and functional sites. Bioinformatics, 2000, **16**(12): 1145-1150

[75] Huson D H, Auch A F, Qi J, *et al*. MEGAN analysis of metagenomic data. Genome Res, 2007, **17**(3): 377-386

[76] Liu B, Gibbons T, Ghodsi M, *et al*. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. Genome Biology, 2011, **12**(1): P11

[77] Brady A, Salzberg S. PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat Methods, 2011, **8**(5): 367

[78] Rosen G L, Reichenberger E R, Rosenfeld A M. NBC: the naive bayes classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics, 2011, **27**(1): 127-129

[79] Wood D E, Salzberg S L. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology, 2014, **15**(3): R46

[80] Ounit R, Wanamaker S, Close T J, *et al*. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. Bmc Genomics, 2015, **16**: 236

[81] Mesuere B, Devreese B, Debyser G, *et al*. Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. J Proteome Res, 2012, **11**(12): 5773-5780

[82] Huson D H, Mitra S, Ruscheweyh H J, *et al*. Integrative analysis of environmental sequences using MEGAN4. Genome Res, 2011, **21**(9): 1552-1560

[83] Schieber B, Vishkin U. On finding lowest common ancestors: Simplification and parallelization. SIAM Journal on Computing, 1988, **17**(6): 1253-1262

[84] Coordinators N R. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 2013, **41**(Database issue): D8-D20

[85] Wu C H, Apweiler R, Bairoch A, *et al*. The universal protein resource (UniProt): an expanding universe of protein information. Nucleic Acids Res, 2006, **34**(Database issue): D187-191

[86] Altschul S F, Gish W, Miller W, *et al*. Basic local alignment search tool. J Mol Biol, 1990, **215**(3): 403-410

[87] Rondon M R, August P R, Bettermann A D, *et al*. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol, 2000, **66**(6): 2541-2547

[88] Daniel R. The soil metagenome——a rich resource for the discovery of novel natural products. Curr Opin Biotechnol, 2004, **15**(3): 199-204

[89] Hu Y, Yang X, Qin J, *et al*. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. Nat Commun, 2013, **4**: 2151

[90] Qin J, Li Y, Cai Z, *et al*. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 2012, **490**(7418): 55-60

[91] Wilmes P, Wexler M, Bond P L. Metaproteomics provides functional insight into activated sludge wastewater treatment. PLoS One, 2008, **3**(3): e1778

The Strategies and Challenges in Metaproteomics Bioinformatics*

XU Hong-Kai[1, 2)**], YAN Ke-Qiang[1, 2)**], HE Yan-Bin[1)], WEN Bo[1)], YANG Huan-Ming[1, 3)], LIU Si-Qi[1, 2, 3)***]

([1)] *BGI-Shenzhen, Shenzhen* 518083, China;

[2)] *Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing* 100101, China;

[3)] *James D. Watson Institute of Genome Sciences, Hangzhou* 310008, China)

Abstract Metaproteomics is a new frontier of microbiological science that collects the proteomic data from microbes in nature using mass spectrometry and explores the corresponding genetic and biochemical mechanisms with systematical bioinformatics. In contrast to the traditional approach, metaproteomic informatics adopts new strategies, including algorithms, databases and searches. As the metaproteomic samples generally contain very complicated protein components, a large dataset with all the potential microbe genomes is basically required for searching peptides based on the signals of mass spectrometry, while such searching process is real time-consuming. Several considerable factors such as dataset capacity, searching strategy and false positive control, therefore, have to be carefully evaluated to achieve the better results of protein identification with an acceptable accuracy and efficiency. Meanwhile, except a common sequence merger in proteomic informatics, metaproteomics has to deal with the issues of vast sequence homologous and species grouping. Solving these problems relies on effective utilization to the public information gained from NCBI for species classification, and filtration treatment from sequence to species using LCA algorithm. Herein, we briefly introduce this field, including which is the basic informatics strategy of metaproteomics, what are the tough challenges in metaproteomic informatics, and how the technique difficulties are being solved in future.

Key words metaproteomics, database, data analysis, protein inference, species

DOI: 10.16476/j.pibb.2017.0187

---

* This work was supported by a grant from National Basic Research Program of China (2014CBA02002, 2014CBA02005).

**These authors contributed equally to this work.

***Corresponding author.

Tel: 86-755-36307403, E-mail: siqiliu@genomics.cn

Received: September 5, 2017 Accepted: October 20, 2017