新时代背景下生物医学技术与的生物医学大数据的分析

陈平力
汕头大学医学院图书馆　广东　汕头　515041
DOI：

【摘要】 基于目前生物医学大数据的现状我们可以发现，生物医学大数据的研究正处于蓄势待发状态：适应于生物医学大数据的软硬件平台、大数据存储、大数据分析挖掘等方法等还不成熟，制约着生物大数据的研究.然而一旦相关研究获得突破并有所优化和应用，将会全方位地支撑生物医学大数据的深入解构；进而有助于对医学现象的趋势分析和预测，服务于相关的遗传疾病研究、公共卫生监控、医疗与医药开发等广泛生物医学应用.

【关键词】 单细胞；医学图像；数据挖掘

0 引　言

随着生物分析技术和计算技术的快速发展，生物医学产生了大量的数据.21 世纪以来，随着高通量 DNA 测序的技术发展和逐步应用，生命科学领域的数据量正在极速增长.1977 年实现了}－X174 噬菌体全基因组测序；2000 年，人类基因组草图被绘制完成. 21 世纪尤其是 2010 年以来，随着新一代测序技术的发展，更大数量级的基因组数据产出日渐增(从 GB，TB 级到 PB，EB 级)：Illumina 公司最新的推出的 HISEQ X TEN 测序仪 3 天内测序约 1.8 TB 的碱基数.大规模的基因组数据的分析和管理正在成为推动生命科学创新的重要源泉.

1 生成海量大数据的先端生物技术

生物医学大数据的研究依赖于高通量、高质量的数据生成线段生物技术和相关仪器.目前生物医学大数据的来源主要有 3 种：}1) DNA 测序仪器；(2)高

通量高精度质谱仪；(3)高通量高精度表观型分析仪器.

DNA 测序仪器：454，Illumina，PacBio 等新一代测序技术的问世，带来生物医学领域的革命.新一代测序技术能够较经济地对基因组进行高效准确的测序.随着实验技术的成熟和数据分析算法的开发，新一代测序技术不仅大量应用在生物医学数据研究[38]，而且在一些复杂的医学研究项目中也得到实际应用[39].相关大数据数量级远超过太字节 TB)级别.

高通量高精度质谱仪：蛋白质组学是继基因组测序计划后崛起的一门新兴学科，逐渐成为后基组时代的研究前沿和热点领域.而代谢组学是继基因组、蛋白组后发展起来的一门学科，主要研究的是作为各种代谢路径的底物和产物的小分子代谢物，在疾病诊断、新药研发、毒理方面都有非常大的应用.近年来随着研究人员对蛋白组学、代谢组学的不断重视以及质谱技术的高速发展，高通量高精度质谱仪产生了越来越多的生物医学数据.相关大数据累积的数量级也已超过太字节别 .

高通量高精度表观型分析仪器：首先，随着荧光蛋白标记等标记式检测方法、红外和拉曼等非标记式细胞检测方法、单细胞操纵等技术的发展，荧光流式细胞分选仪(}EAC}9}、活体单细胞拉曼分选仪(}BACg}等单细胞分析和操控平台日渐成熟.由于单物种群体或群落中的单细胞数量巨大，相关单细胞表观型数据量大且积累迅速.其次，单细胞操纵技术的成熟开启了单细胞测序序幕，相关的海量测序数据将会迅速被生成。最后，高分辨显微图像的海量生成，迅速地积累了更为海量的生物医学大数据。

2 生物医学大数据的分析

生物医学大数据可以分为大数据存储和大数据

分析两方面，其中大数据存储服务于大数据的深人分析.当今生物医学中的典型大数据包括各类基组数据、宏基因组数据和单细胞数据以及生物医学图像数据等.

2.1 基因组数据分析

在高通量数据生成和系统化数据分析方面，目前国际上对组学数据的高通量生成和系统化分析已经初步形成了若十通用流程.在高通量基因组和转录组数据生成方面，454，Solexa，PacBio 等新一代测序技术的引人和推广，配合高通量数据分析方法，使更加细致、深人的基因组和转录组数据分析成为可能.在标准化数据分析流程方面，包括华盛顿大学的 Tophat－Bowtie－Cufflink 系列，华大基因的 soAP 系列，以及商业化的 CL，CB10 系列(http://www.clcbio.com)等.这些系统化分析流程整合了基因组、转录组和部分表观基因组等数据的分析，极大地推动了生物系统的相关快速、标准化和深人的研究(在此不一一赘述).随着高通量测序数据的快速积累，更高水平上的基因组数据整合、挖掘与可视化等分析要求也在提高.必须通过适应于大数据分析的软硬件系统优化、分析流程的整合、交互式可视化分析平台的建设等方法来实现.

2.2 蛋白质组数据分析

对于蛋白组学，以高分辨多级串联质谱为代表的质谱分析技术日趋稳定；通过收集海量的高分辨率一维质谱(M 9}和二维质谱(MS/Mgt 数据，一些大规模的蛋白组定性和定量分析工作也已完成[o_}o.目前蛋白质组学研究向着研究对象更全面(如全面的一级质谱数据独立获取((DIA)数据研究等}4″)和研究规律更深人(如整合不同组学数据进行机制性研究等}20′＞的方向发展(在此不一一赘述).尤为重要的是，随着高通量数据生成和系统化数据分析方法的日臻成熟，组学研究发展中的必然要求(如基于细胞内全组分多样性与相互作用的“全局性”分析要求、翻译后修饰等重要调控过程解析要求、表观型调控解析要等)被提出，不同层面组学数据之间的融合分析变得益发重要与紧迫.

2.3 宏基因组数据分析

生物医学相关的微生物群落大数据分析和数据挖掘任务数量的指数型增长趋势.目前，NCBI(http://www.ncbi.nlm.nih.gov)，MG－RAST}48]以及 CAMERA}4}}中公开的宏基因组项目超过 10000 个，包含高达数百 TB 的数据.保守预计 2014－2015 年国际上每年会有上万个相关数据分析任务(每个任务中的样本数量从个位数到上千个)，因此其科研和应用市场需求十分旺盛.其次，每个微生物群落大数据分析项目的数据量也在增加：宏基因组数据分析项目的平均数据量达到了 10 GB－1 TB 量级.如此巨大的数据量对数据分析的效率和准确性也提出了较高要求.

3 结论

适应于生物医学大数据的软硬件平台，大数据存储，大数据分析挖掘等方法的提出、优化和应用，将会全方位地支撑生物医学大数据的深人解构和相关研究对象的趋势分析和预测，进一步服务于相关的遗传疾病、公共卫生、医疗与医药等广泛生物医学应用.

【参考文献】

[1]代涛，黄菊，马晓静.国际全科医生制度发展历程：影响因素分析及政策启示[J].中国卫生政策研究，2015，8(02)：1－7.

[2]薛晓芳. 知识可视化理论、方法和工具及军事医学应用研究[D].中国人民解放军军事医学科学院，2014.