生物医学大数据现状分析

林　丽
吉林大学白求恩医学院　吉林　长春　130021
DOI：

【摘要】 生物医学是一门新兴的前沿交叉学科，它综合了医学、生命科学和生物学的理论和方法而发展起来.近年来随着先进仪器装备与信息技术等越来越广泛和深入的整合到生物技术中来，生物医学研究中越来越频繁的涉及到大数据存储和分析等信息技术。大数据时代的来临对生物医学研究产生了重大影响.其中，一个重要发展趋势就是由假设驱动向数据驱动的转变。数十年来分子生物学水平上的实验目的是获得结论或者是提出一种新的假设而现在基于海量生物医学大数据，可以对海量数据的研究来探索其中的规律，直接提出假设或得出可靠的结论.随着先进的生物分析技术的不断推出和更新，生物医学数据迅速积累。

【关键词】 生物医学；大数据微生；物群落

0　引　言

生物医学是应用生物医学信息、医学影像技术、基因芯片、纳米技术、新材料等技术的学术研究和创新交叉领域.随着以“社会一心理一生物”为代表的大医学模式的提出和系统生物学的发展，形成了现代系统生物医学面向生物医学的系统生物学研究是与21世纪生物技术技术和大数据技术密切相关的领域，是关系到提高医疗诊断水平和人类健康的重要研究领域.

1　生物医学大数据研究的特点

1.1　大数据的重要性

高通量的研究思路和相关数据生产方式的飞跃是大数据产生的主要因素.大数据经历着从概念到小范围技术实践，最终到广泛接受并成为一个新兴研究方向的历程.2008年9月，Nature杂志率先出版了由HOWc等人[4]所发论文组成的“大数据专刊”表明大数据的影响已触及自然科学、社会科学和工程学的各个领域.200年10月，The Fourth Paradigm：

Data Intensive Scient沪。Discovery}s}一书的出版，显示出与大数据关系密切的数据密集型科学发现范式已被确立和广泛认可.2011年2月，Science杂志推出Overpeck等人[f}l所发表文章组成的“数据处理专刊”.2012年5月，联合国发布大数据政务白皮BigDataforDevelopment：Challenges & Opportunities，体现了大数据领域的研究计划在国家战略层面的重要性.2014年Science杂志推出“Big biological impacts frombig data”等一系列评论，也明确无误地表明了生物学相关研究已进人大数据时代.

在大数据时代，庞大繁杂的数据以及对数据的研究对社会、科技、经济的发展将发挥支撑促进作用.大数据本身是一种潜在的战略性资源，具有小规模数据无法匹及的趋势预测潜力，大数据的分析和应用才能将这些资源的效益真正释放出来.美国、欧盟等已在国家和地区层面开展了大数据研究和发展计划，将大数据研究提升到国家和国际重大战略层面.2013年5月9日美国总统奥巴马签署了一项行政命令，要求政府帮助公众和企业更容易取得政府持有的数据，从而促进美国的创新和经济增长.2013年7月习近平总书记在中国科学院考察时指出“大数据是工社会的‘石油’资源”.谁掌握了大数据以及大数据的研究技术，谁就掌握了主动权.尤其是在生物医学等事关人类健康和命运的研究领域，对相关大数据的研究就是对健康领域未来的掌握.

1.2　生物医学大数据的特点

以高通量测序仪器、单细胞检测装备和实时动态图像系统为代表的新一代生物分析平台已经和正在为生物医学研究提供海量数据，而要充分利用蕴藏于海量数据中的深刻规律，大数据驱动的研究策略必不可少[9].大数据至少包含3层含义(}3 V}(图1}：数据量大(volume of data)，处理数据的速度快

(velocityof processing the data)，数据源多变(variability of datasources)．这是那些依赖大数据进行分析和预测过程的重要特征[} o}．具体到生物医学大数据研究而言，大数据研究的 3V 特点体现如下：第一，生物医学数据量大．通常对于一个样本的人体基因组和转录组(多组织多时间点)测序数据量会分别超过 100 和 30GB 基于 3 GB 人类基因组和 10－30 倍测序深度)．考虑到一次试验中通常会涉及到数百个甚至上万个人体样本，相关的数据量产出十分巨大．第二，研究对于处理结果准确性和处理速度均有高要求．如个性化医疗["]，就具有较高的时效性要求，而单细胞测序及诊断等＋ z}，对突变位点和功能模块的鉴别准确性要求较高．第三，相关源数据来源多变且具有较大的异质性．同时生物医学数据的分析和解释通常会

利用到 NCBI 系列[13]等通用数据库以及 UniProt (www．uniprot．org)等专业数据库．源数据和数据库的异质性，会导致数据缺失、数据矛盾等问题的遍存在，成为相关大数据整合与分析的瓶颈．正是因为生物医学研究具有典型的 3V 特点，所以需要依靠大数据思维和数据分析策略对生物医学数据进行深人挖掘．

2　生物医学大数据的典型应用

典型的生物医学数据包括癌症、个性化医疗等数据，其呈现形式包括功能基因组、单细胞、宏基因组(又称元基因组)数据等．所有这些数据存储于 NCBI 或 EBI 等大型通用数据库中．同时随着高通量测序技术的发展和应用以及生物技术与信息技术的融合，NCBI 等大型通用数据库中生物医学数据类型和数据规模不断增大．

2.1　现有大型通用生物医学数据库

现有生物医学大型通用数据库包括美国 NCBI 的 GenBank、欧洲的 EBI、日本的 DDBJ 等．针对于某些特定数据或研究对象的数据库如 U ni－Prot (蛋白数据库)、MG－RAST(微生物数据库)也正在快速发展．这些都是从事生物信息数据的管理、汇聚、分析、发布等工作的大型数据库．近年来，随着高通量测序技术的发展等，这些大型数据库数据量不断激增，如表 1 所示．

2.2　个人基因组以及个性化医疗

2008 年 11 月 6 日，Nature 杂志刊登了“第一个亚洲人基因组图谱’，论文，封面名为“你的生命掌握在你手中’．“第一个亚洲人基因组图谱’，的完成是医学方面的重要成就，这意味着未来 5－10 年，一个人只需要花很少的费用就可以拥有自己的基因组图谱．可以预见未来，医生可以依据这个基因组图谱对病人进行更精确地诊断和治疗，更可能在发病前就进行必要的十预．甚至连药物都可以根据这个基因图谱为一个人单独设计．可以说这是“你的生命掌握在你自己手中”，“个人基因组时代已经来临”的先兆．基因组图谱结合对基因表达调控等与医学有关知识，可以对人类认识疾病的发病过程，对疾病的抵抗性研究将带来新思路．有了“基因组图谱’，不仅对疾病治疗有作用，更重要的是在发病前人们就可以十预、预防这些疾病了．这样，治病对人们来说将不再是千篇一律了。

3　结论

在基础研究领域，除高通量基因组和转录组测序产生的数据外，代谢组、蛋白质组等领域也正在极速增长，而细胞表型、代谢过程、致病基因等的分析都函需将不同类型的数据加以整合和解构，从中挖掘出深刻而又非显而易见的生物学规律．

【参考文献】

[1]张娟．医学图像配准中相似性测度的研究[D]．南方医科大学，2014．

[2]于海容，姜安丽．国外叙事医学教育发展及其对护理学的启示[J]．中华护理杂志，2014，49(01)：83－86．