大数据

生物学变革新契机

Big data
a new opportunity for biology

张 旭/主编

科学出版社

大数据：生物学变革新契机

张　旭　主编

科学出版社
北　京

内 容 简 介

随着信息技术的发展，世界正在由资本经济时代向数据经济时代过渡，大数据作为一种新的资源，在社会的各个方面发挥着重要的作用，有力推动着社会经济的发展。随着大数据研究与应用投入的不断加大，生物大数据带来了生物产业的一次变革，创造出巨大的经济价值和社会价值，并已成为全球生物产业发展的新助力，给生物产业的发展带来划时代的意义。本书正是呈现生物大数据的历史变革及产生重大影响。全书共分 6 部分，首先阐述了大数据时代已经来临的历史背景，主要国家对生物大数据发展进行的战略布局，生物大数据带来的革命性意义，生物大数据开发与利用的关键技术，生物大数据的未来市场，生物大数据时代的发展困境。

本书适用于不同的读者群。从事生物大数据研究的所有研究者、教育者和学生均能从中获益；政府各基金资助部门的管理者、政策制定者亦会从中受益；即使是普通读者，也能从中一窥生物大数据研究的重大变革，了解生物大数据变化趋势对人类的重大贡献。

图书在版编目（CIP）数据

大数据：生物学变革新契机 / 张旭主编.—北京：科学出版社，2016.1

ISBN 978-7-03-046189-6

Ⅰ.①大… Ⅱ.①张… Ⅲ. ①数据处理–应用–生物学–研究 Ⅳ.①Q–39

中国版本图书馆 CIP 数据核字(2015)第 260636 号

责任编辑：罗 静 田明霞 / 责任校对：陈玉凤

责任印制：徐晓晨 / 封面设计：刘新新

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京京华虎彩印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 1 月第 一 版 开本：720×1000 B5

2016 年 1 月第一次印刷 印张：9

字数：181 000

定价：75.00 元

(如有印装质量问题，我社负责调换)

《大数据：生物学变革新契机》
编写委员会

主　编　张　旭

副主编　于建荣

编写成员（按姓氏汉语拼音排序）

陈大明　陈润生　陈兴委　范月蕾

韩敬东　洪胜君　黄　河　江洪波

黎　浩　李亦学　李祯祺　刘　雷

马俊才　毛开云　苏　燕　孙晓濛

王　绛　吴　刚　熊　燕　徐　萍

于建荣　张文生　张　旭

前　言

“大数据”这一名词自 2012 年在奥巴马国情咨文中被重点提及后，近几年来发展迅猛，已经在很多行业得以体现。大数据带来的信息风暴正在变革我们的生活、工作和思维，开启一次时代的重要转型。生命科学作为新世纪最活跃的学科，也正在经历一场数据革命，全世界每年产生的生物数据总量高达 EB 级，生物大数据已成为“大数据”重要的组成部分。国际著名商业咨询机构 BCC Research 的分析报告《Next Generation Sequencing：Emerging Clinical Applications and Global Markets》指出：“2013 年，全球新一代测序和数据分析市场总额为 5.1 亿美元，至 2018 年，这一市场总额将增长至 76 亿美元，复合年增长率达到 71.6%。”生物大数据蕴涵着巨大的产业价值，今后将成为与能源、矿产一样的战略资源。

虽然大数据已成为一个热词，但鲜有人对大数据为生物变革带来的机遇进行翔实的探讨和分析。究竟大数据会对生物技术与产业的发展带来哪些变革，生物大数据的开发与利用又包含着哪些关键技术，全球哪些国家正在生物领域对大数据展开布局，我国发展生物大数据的现状又是怎样，对于这些疑问，《大数据：生物学变革新契机》一书将用详细的数据材料与较为丰富的内容，从不同的角度对生物大数据广泛的影响力予以阐述。

目前，大数据建设已引起生物学界与产业界的广泛重视，大数据的重要性在生物研究与产业发展的方方面面得以体现，如在最近被炒得很热的精准医学领域，大数据作为精

准医学发展的基础，是提升个人遗传密码数据整合与分析能力的关键。因此占领大数据高地，对于各国生物技术的发展，以及国际竞争力的提升都有着重要的意义。2013 年，美国政府为全面推动生物医学大数据基础研究发展，启动了 Big Data to Knowledge 计划，大力推动和改善与生物医学大数据相关的收集、组织和分析工具及技术。2014 年 3 月，在伦敦举行的高性能计算机技术和大数据会议上，英国大学与科学国务大臣 David Willetts 宣布，英国医学研究理事会（MRC）将投资 3200 万英镑资助首批 5 大项目，来加强医学生物信息学的能力、产能和核心基础设施建设。这项“医学生物信息学计划”预计总投资 5000 万英镑，将通过建立耦合复杂生物数据和健康记录的新方法，来解决关键的医学难题。

发达国家对生物大数据“野心勃勃”，我国在生物大数据的发展方面却尚不尽如人意。当前，中国已是生物数据产出大国，但目前还远不是生物大数据存储和应用的强国。面向生物大数据的国家级技术研究中心尚未建立，技术研发宏观规划和引导缺乏，专业人才队伍储备不足，核心分析技术创新不多，数据共享规则制定不完善等都是制约我国生物大数据发展的“瓶颈”。发达国家生物数据基础设施建设起步较早，早在 20 世纪 80 年代起，美国、欧洲和日本相继开始建设世界级生物数据中心——美国国家生物技术信息中心（NCBI）、欧洲生物信息研究所（EBI）和日本 DNA 数据库（DDBJ）。目前，这三大生物数据中心掌握并管理着全世界主要生物数据和知识资源，处于数据垄断地位。相对于大数据基础设施建设，我国人才队伍建设任务更加紧迫，因为专业人才培养周期较长，无法取得立竿见影的效果。除了硬件和人才建设外，适宜的政策法规和社会文化软环境对生物大数据发展也相当重要。数据

共享规则不完善，缺乏对数据权益的保护，不利于形成良性发展的数据共享生态系统。即使如此，我国在生物大数据领域也具有遗传资源多样、临床数据丰富、基础研究人员众多等诸多优势，更重要的是，大数据的内在哲学观点及其认识论本质是整体论，与东方传统哲学中的整体和集体思想暗合，而且中国社会易于集中、易于统一组织的传统社会特质与文明特性将十分利于生物医学大数据的建设发展。这为我国未来大力发展生物大数据奠定了坚实的基础。

国家之间的竞争是战略层面的竞争，是国家意志的较量，在美、欧、日已形成大数据发展战略形势下，我国尽早形成完整明确的生物大数据国家战略，对今后的学术、技术和产业发展至关重要。无论是从产业发展潜力还是国家战略安全角度，中国作为日益强大的世界大国，都必须在大数据领域有所作为。如何对整个生物大数据领域作长期的全局规划，形成有特色的生物大数据体系，是新时期摆在我国大数据建设面前的关键问题。在这样的大背景下，《大数据：生物学变革新契机》应运而生，相信其对相关的专业人士以及对该领域有兴趣的普通读者都有重要的参考价值。

陈润生

2015 年 11 月

目　录

第1章

大数据时代已经来临

继蒸汽技术革命和电力技术革命之后，以原子能、电子计算机、空间技术和生物工程的发明和应用为主要标志的信息技术革命（即第三次科技革命）不仅极大地推动了人类社会经济、政治、文化领域的变革，而且影响了人类生活方式和思维方式。技术创新和数字设备的普及为人们带来“数据的产业革命”。对日益扩大的、多种多样的且极富关联的数字数据的分析，将揭示关于集体行为的潜在联系，并有可能改进决策的方式。大数据的开发，关键在于将不完善的、复杂的数据转换成可操作的信息，这要利用先进的计算工具揭示大型数据集合内部及数据集合间尚未被发现的趋势与相关性。大数据作为第三次科技革命重要组成部分对人类社会的重要影响正在不断涌现。

1.1 大数据的发展历程

信息技术革命的步幅基本以 20 年为一个周期。20 世纪 50 年代，信息技术革命开始进入架构化时代；

本章作者：李亦学　上海生物信息技术研究中心

随着数字资源的不断发展，数字化于 20 世纪 70 年代开始深入人心；从 20 世纪 90 年代开始，基于“信息高速公路”的构建，人类社会进入了网络化的时代；近 5 年来，以移动互联网、云计算和物联网为标志的智慧化时代已经到来，而这三者均与大数据有着千丝万缕、密不可分的关系。

大数据的应用和技术是在互联网的快速发展中诞生的，起点可追溯到 2000 年前后。当时，互联网的网页数量呈爆发式增长，每天新增约 700 万个；至 2000 年年底，全球网页总数达 40 亿个，对于用户来说，检索信息变得不再便捷。以谷歌为主的多家公司率先建立了覆盖数十亿网页的索引库，开始提供较为精确的搜索服务，大大提升了人们使用互联网的效率，这就是大数据应用的起点。在那时，搜索引擎要存储和处理的数据不仅具备前所未有的数据量，而且主要以非结构化数据形式存在，导致传统技术无法应对。为此，谷歌提出了一套以分布式为特征的全新技术体系，即后来陆续公开的分布式文件系统（Google file system，GFS）、分布式并行计算（MapReduce）和分布式数据库（BigTable）等技术，以较低的成本实现了之前无法解决的大数据问题。这些技术奠定了当前大数据技术的基础，可以将其认为是大数据技术的发展源头[1]。

随着互联网行业的迅速兴起，这种创新的海量异构数据处理技术在电子商务、智能推荐、社交网络等方面得到广泛应用，获得了极大的商业成功。此时，全社会开始重新审视数据所带来的巨大价值。尤其是以金融、电信等为代表的、拥有大量数据的行业开始尝试这种新的理念与技术，并取得一系列初步成效。与此同时，业界不断对谷歌提出的技术体系进行扩展与完善，使之能够在更多的场景下获得应用。

早在 1980 年，著名未来学家 Alvin Toffler 就在《第三次浪潮》一书中前瞻性地指出过大数据时代即将到来[2]，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。20 世纪 90 年代，数据仓库之父 Bill Inmon 开始关注大数据的发展。时至 2007 年 1 月，已故的图灵奖得主 Jim Gray 在他最后一次演讲中描绘了数据密集型科研“第四范式（the fourth paradigm）”的愿景[3]：随着数据量的高速增长，计算机将不限于模拟仿真的功能，还可以进行分析总结，并得出理论。时隔一年，*Nature*

于 2008 年 9 月出版了大数据专刊“Big Data：Science in the Petabyte Era”[4]，阐述了现代科学面临的一个巨大挑战——如何处理已有的海量数据。2011 年 2 月，*Science* 推出了一期关于数据处理的专刊“Dealing with data”[5]，又一次从互联网技术、互联网经济学、超级计算、环境科学、生物医药等多个方面介绍了海量数据所带来的技术挑战。同年，麦肯锡（McKinsey）、世界经济论坛（World Economic Forum，WEF）等知名机构对这种数据驱动的创新进行了研究总结，随即在全世界兴起了一股大数据热潮。

大数据的发展主要分为 3 个阶段[6]。

（1）初始萌芽期（20 世纪 90 年代至 2003 年）

随着数据库技术与数据挖掘理论的逐步成熟，一系列商业化的智能工具与知识管理技术得以被人们应用。此时，对于大数据的研究主要集中于算法模型、模式识别等热点方向，侧重于数据挖掘和机器学习等基础信息技术。

在这段时间内，大数据的提升体现在“量”和“质”两方面，这为数据时代的到来奠定了坚实的基础。从数据的“量”上来看，如果从人类文明出现开始计算，那么直到 2003 年，人类总共才生成 5 EB（ExaByte）左右的数据。这是由于计算机出现后，伴随着数字化与网络化的完善，数据产生的规模与速度才开始急剧上升。由于数据规模呈指数上升态势，因此仅过去几年所产生的数据就比以往 4 万余年的数据总量还要丰富。从数据的“质”出发，很大一部分数据源于人体或环境等，通过数据流将原本看似无关的多种数据维度或属性关联起来，并通过数据的流动与共享来实现大数据的社会经济价值。

（2）快速突破期（2004～2009 年）

非结构化数据的爆发带动大数据技术的快速突破。2004 年 Facebook 的创立是非结构化数据爆发时期的标志性事件，社交网络的流行直接导致大量非结构化数据的涌现，而这种局面是传统处理方法难以应对的。在此阶段，大数据研究的热点方向趋于云计算、MapReduce、开源分布式系统基础架构（Hadoop）和人工智能等。

据统计，在全球范围内可用的数字数据数量从 2005 年的 150 EB 增加至 2010 年的 1200 EB。数据总量预计每年增长 40%左右，这一增

长率意味着数字数据的存储预计将从 2007 年到 2020 年增长 44 倍[7]。国际数据公司（IDC）预计 2020 年全球数据使用量将达到 40 ZB（ZettaByte），需要约 429 亿个 1 TB 的硬盘进行存储，届时中国产生的数据量将占到全球总量的 21%[8]。正是如此海量异构的数据驱动传统计算机算法的进步，才最终导致了人工智能的突破性进展。互联网的稳步发展为机器模型的训练提供了足够多的样本集合，而这种以深度学习为代表的数据驱动算法不仅是对传统计算机算法的颠覆，而且为人工智能带来了实质性的突破。最具代表性的案例莫过于美国国家标准与技术研究所（National Institute of Standards and Technology，NIST）于 2005 年对全球的机器翻译系统的评测，不是专门从事机器翻译的 Google 竟然以一骑绝尘的显著优势获得各项评比的第一名，其优势就在于使用了近乎于其他系统万倍左右的数据量。

（3）稳健发展期（2010 年至今）

随着手机、平板电脑等智能终端的应用日益广泛与频繁，数据的碎片化、分布式、流媒体特征更加明显，移动数据急剧增长。传统的大数据“三核心”（即 GFS、MapReduce 和 BigTable）遭遇数据运算和处理能力方面的瓶颈，而 2010 年谷歌为应对这种趋势而开发的 Percolator、Dremel 和 Pregel 日趋成为新“三核心”。与此同时，非关系型数据库（NoSQL）再次自我革新，开始转向兼具关系型易查询和非关系型高扩展性的新型云数据库（NewSQL），如谷歌的 Spanner、亚马逊的 RDS、微软的 SQL Azure 等，大数据的核心技术仍在快速发展。

急剧增长的大数据样本与颠覆性的大数据技术相互促进，实现发展的正循环。随着需要处理的数据量的增大与数据类型的增加，所需的技术支持需求愈发明显。而以云计算、人工智能等为代表的大数据技术应运而生，这些技术的数据诉求可以通过大数据样本的迅速增长得以满足，并随着数据处理量的增大而产生更加精准的结果。

综上所述，60 年前，数字计算机使得信息可读；20 年前，Internet 使得信息可获得；10 年前，搜索引擎爬虫将互联网变成一个数据库；现在，Google 及类似公司处理海量语料库如同一个人类社会实验室。数据量的指数级增长不但改变了人们的生活方式和企业的运营规模，而且改变了科研范式[9]。

1.2　大数据的定义

针对大数据，目前存在多种不同的理解和定义。通常来说，“大数据”（big data）是指无法在一定时间内用通常的软件工具进行捕获、管理的巨型数据集。2011 年，麦肯锡全球研究院（McKinsey Global Institute，MGI）于《大数据：下一个创新、竞争和生产力的前沿（Big data：The next frontier for innovation, competition, and productivity）》报告中率先提出了大数据的概念[10]，即“大数据是指无法在一定时间内用传统数据库软件工具对其内容进行采集、存储、管理和分析的数据集合”。此后，经过高德纳咨询公司（Gartner Group）的炒作周期（hype cycle）曲线和 Viktor Mayer-Schönberger 的《大数据时代：生活、工作与思维的大变革（Big data：A revolution that will transform how we live, work, and think）》[11]的宣传与推广，大数据迅速为人所熟知，并开始“风靡全球”。2012 年，高德纳咨询公司将其对大数据的定义进行了修改，阐明“大数据是一种数量大、增速快且复杂多样的信息资产，需要通过新式的处理手段从中形成更强的决策能力、洞察力与优化处理方法”[12,13]。美国国家科学基金会（National Science Foundation，NSF）则将大数据定义为“由科学仪器、传感设备、互联网交易、电子邮件、音视频软件、网络点击流等多种数据源生成的大规模、多元化、复杂、长期的分布式数据集”。按照 NIST 发布的研究报告的定义，大数据是用来描述在网络的、数字的、遍布传感器的、信息驱动的世界中呈现出的数据泛滥的常用词语。这种大量数据资源为解决以前不可能解决的问题带来了可能性。

综上所述，维基百科将大数据定义为“所涉及的数据量规模巨大到无法通过人工，在合理时间内达到截取、管理、处理、并整理成为人类所能解读的形式的信息”。

1.3　大数据的 4V 特征

麦塔集团（META Group，现为高德纳咨询公司）的研究人员 Doug Laney 于 2001 年的研究报告中指出数据增长的挑战和机遇具有 3 个方

向：即数量性（volume，指数据的大小）、速度性（velocity，指数据输入与输出的速度）与多变性（variety，指数据类型与结构的复杂多样），并称之为“3V”或“3Vs”[14]。此外，还有一些机构对大数据的含义进行了补充，加入第四个 V（veracity，真实性）作为大数据的第四特点[15]，如 IBM；也有一些公司将价值（value）称为大数据的第四属性，如甲骨文（Oracle）公司和中国移动研究院。IDC 同样将大数据的四大特征定义为海量的数据规模、快速的数据流转与动态的数据体系、多样的数据类型和巨大的数据价值。

海量化：用于聚合分析的数据规模异常庞大。目前，全球每两天创造的数据规模等同于从人类文明起始至 2003 年间产生的数据量总和。全球的数据总量正以前所未有的速度增长，通过设备与网络每天都会产出上百万兆字节的数据。数据规模的大幅增长远远超过了硬件的发展速度，从而导致了数据储存与处理的危机。导致数据规模激增的原因首先是互联网络的广泛应用，使用网络的人、企业、机构迅速增多，从而让数据的获取与分享变得相对容易；其次是各种传感器数据获取能力的大幅提高，使得人们获取的数据越来越接近原始事物本身，描述同一事物的数据量激增。

多样化：数据来源广泛，且形态多元。从数据格式出发，可以分为文本、图片、音频与视频等；从数据关系来看，能够分为结构化、半结构化、非结构化数据。随着互联网络与传感器等技术的飞速发展，非结构化数据大量涌现。由于非结构化数据没有统一的结构属性，难以用表结构来表示，而且在记录数据数值的同时还需要存储数据的结构，从而增加了数据存储、处理的难度。据不完全统计，目前全球的非结构化数据已占数据总量的 75%以上，且非结构化数据的增长速度比结构化数据快 10～50 倍。随着非结构化数据的比重越来越大，其中蕴含的经济、社会方面的价值日益受到人们关注，同时也对传统的数据分析处理算法和软硬件环境提出了挑战。

快速化：不仅数据的增长速度不断加快，而且要求数据访问、处理与交付等处理速度快。目前，数字数据储量每 3 年就会翻 1 倍，人类存储信息的速度比世界经济的增长速度快 4 倍。随着数据发布共享的不断普及，个人甚至成为了数据产生的主体之一，快速增长的数据

量要求数据处理的速度作出相应的提升，才能使得大数据有“用武之地”。否则，日益激增的数据不但不能为解决问题带来优势，反而会成为快速解决问题的负担。同时，数据不是静止不动的，而是在网络中不断流动的，且其数据价值通常会随着时间的推移而迅速下降。如果数据没有得到有效处理，也就失去了其价值和意义。通过对大数据的快速处理，能够迅速对经济、社会等各方面情况作出深入了解，并提供决策支持，从而及时制订出合理而准确的应对策略，这将成为提高企业乃至国家竞争力的关键。

价值化：价值潜藏在大数据背后，也是其终极意义所在。据统计，美国社交网站脸书（Facebook）有 10 亿用户，通过对其用户信息进行分析后，广告商即可根据结果实现精准投放。据资料报道，仅在 2012 年，运用大数据的世界贸易额就已经达到 60 亿美元。随着社会信息化程度的日益提高，数据存储规模不断增大，数据的来源与类型越发多元化。数据正成为一种新型的资产，是形成竞争力的重要基础，并已经成为竞争力提升的关键点。

1.3.1　生物大数据的内涵及范畴

在大数据的时代，生物医学领域将会是最为重要的一个大数据应用行业。生物医学领域最大的变革是人们即将进入个性化或精准医疗时代，而支撑个性化医疗和个性化用药的基础正是生物医学领域的大数据。每个人的基因组大小为 30 亿个碱基，个体之间的遗传差异信息为 300 万个碱基位点。不考虑其他分子水平的信息，仅基因组水平的个体化数据就是上百万到上亿字符的信息。如此众多的个体化分子水平差异数据，让人们可以对每个个体或每类疾病表型进行精确分型，从而实现全方位的分子检测和个性化医疗。特别值得注意的是，这些信息不仅仅停留在基础研究阶段，有相当一部分已经进入临床应用。生物大数据的爆发，对信息管理提出了严峻的技术挑战，同时也意味着巨大的商业机会。

随着大数据的日益“流行”，生物大数据也获得了人们越来越多的关注。目前，全世界每年的生物数据产生总量已经高达 EB 级。如此海量的数据为人们深入了解生物学过程、疾病机制等多方面提供了前