华大基因 BGI

国家基因库 CHINA NATIONAL GENEBANK

华大基因-国家基因库系列

生物信息数据库建设、使用与管理指南

张 勇 严志祥 魏晓锋 朱 赢 操利超◎主编

科学出版社

华大基因-国家基因库系列

生物信息数据库建设、使用与管理指南

张　勇　严志祥　魏晓锋　朱　赢　操利超　主　编

科学出版社
北　京

内 容 简 介

本书以“生物信息数据库建设”为中心，首先全面地介绍了国内外常见的生物信息数据库，并就生物“大数据”时代生物信息数据库建设过程中存在的问题进行了探讨。先从数据来源、数据类型、生物数据分析方法、生物信息分析平台等方面探讨如何对生物数据库的数据进行管理，包括数据采集、数据存储、备份和恢复等。接下来介绍数据库设计的方法、原则及常见生物信息数据库的特点，并以 NCBI 为例介绍了生物信息数据库如何进行数据提交和检索等。最后阐述在生物信息数据库构建过程中，应结合生物数据本身的特点，不仅要考虑到生物数据的重要性、安全性和准确性，还从伦理学和知识产权保护的角度进行探讨，而这是生物领域研究工作者容易忽略但十分重要的部分，应当予以重视。

本书对我国生物信息数据库建设的规范化和标准化具有重要参考价值，并有利于促进我国生物行业的健康快速发展。

图书在版编目(CIP)数据

生物信息数据库建设、使用与管理指南 / 张勇等主编 . —北京：科学出版社，2014.10

（华大基因—国家基因库系列）

ISBN 978-7-03-041800-5

Ⅰ. ①生… Ⅱ. ①张… Ⅲ. ①生物-信息-数据库 Ⅳ. ①Q-39

中国版本图书馆 CIP 数据核字(2014)第 219590 号

责任编辑：夏 梁 孙 青 / 责任校对：钟 洋

责任印制：钱玉芬 / 封面设计：北京铭轩堂广告设计有限公司

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

中国科学院印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2014 年 10 月第 一 版 开本：787×1092 1/16

2014 年 10 月第一次印刷 印张：19

字数：430 000

定价：150.00 元

（如有印装质量问题，我社负责调换）

《生物信息数据库建设、使用与管理指南》编委会

前　言

生命科学是一门研究自然生命特性的重要学科。一直以来，无数人都为生命的奥秘痴迷不已。早期的生命科学更多关注生物物质的化学本质，随着对生命科学的认识越来越深入，研究关注点得以进一步细分。20 世纪 50 年代，遗传物质 DNA 双螺旋结构的发现开创了分子水平的生物学新时代，而现代生物学中最重要的基本规律之一——中心法则，则更清晰地描述出 DNA、RNA、蛋白质之间的关系，极大地推动了现代生物学的发展，也清晰地阐述了基因组、转录组、蛋白质组之间的信息流向关系，可以说为“组学”的研究发展奠定了坚实的基础。

20 世纪 80 年代，美国科学家率先提出人类基因组计划，并于 1990 年启动。按照计划，科学家们准备解码人体内的 4 万个基因，绘制人类基因图谱。人类基因组计划与曼哈顿原子弹计划和阿波罗登月计划并称为三大科学计划，被誉为生命科学的“登月计划”。这一科学大工程由美国、英国、法国、德国、日本和中国六个国家共同参与，项目耗资 30 亿美元。中国有幸成为六个国家之一，可以说自人类基因组计划起，我国就开始参与其中，并且在后续的基因组学科高速发展中不断贡献力量。这一重大项目的顺利参与和实施也标志着我国生命科学研究，特别是基因组学研究开始跻身国际前列。

伴随人类基因组计划项目的实施，越来越多的基因组完成测序，随之产出了数以亿计的碱基序列。由于数据量巨大，为了更好地分析，必须引入信息学的技术和相关手段，因此，一门新的学科——生物信息学得以迅速发展，利用数学、信息学、统计学、计算机科学等方法来研究生物学问题。今天，生物信息学成为生命科学中的重要组成部分，生命科学也开始由过去的实验科学转变为开始重视信息科学的使用。

基因组计划实施的最基本的工作即“测序”，“测序”技术可以说是数据产出的方法。20 世纪 90 年代至 21 世纪初，基因数据产出主要依靠 Sanger 法（第一代测序技术）。2007 年，第二代测序技术的出现使得生命科学数据的增长一下呈现了爆炸趋势。这种爆炸式发展远远超越了经典的摩尔定律。各国研究人员都渐渐意识到，这种前所未有、突破了 IT 技术发展的速度，未来将使得数据的存储、管理、应用成为巨大挑战。因此，我们要思考关键性问题、挑战，对当前情况做好总结，对未来做好展望、布局。这一巨大挑战背后也孕育着巨大机遇。

“大数据”正在成为我们工作与生活中越来越重要的一个要素。由于测序技术的飞速发展，生命科学也开始进入“大数据”时代。2013 年 3 月 6～8 日，国家自然科学基金委员会信息科学部、管理科学部、政策局会同生命科学部、地球科学部在上海联合举办主题为“‘大数据’技术与应用中的挑战性科学问题”的第 89 期“双清论坛”。我有幸做了题为“基因组学——大科学、大数据、大挑战”的大会报告，会间与信息科学、计算机科学专

家等进行了深入交流。大家一方面感叹生命科学的数据爆炸超出想象，另一方面也发现现有的信息技术、计算机算法等与生命科学还有待更深入融合，行业间的交流沟通还是远远不足的。

生命科学领域最知名的生物信息数据库有NCBI、EBI、DDBJ(组成INSDC协会)。它们形成了一个生命信息数据联盟，共享数据，这种格局使得我国对于生命科学信息数据的把控力较弱，存在风险。为了更好地保护生物信息资源，支撑我国基础科学研究乃至推动产业落地，2011年1月，国家发展和改革委员会批复同意依托深圳华大基因研究院组建运营深圳国家基因库，下设生物样本资源库和基因信息数据库两大部分。同年10月，发改委、财政部、工信部、卫生部批复同意建设方案。现在，基因信息数据库已经成为我国最大最全的数据库，并成功实现了支撑基础科学研究和推进产业落地的发展模式。

2011年，从数据的存储、管理、挖掘、应用的角度，我提出了“生物谷歌”的概念，试图打造一个“整合全球生物信息，使人人皆可访问并从中受益”的体系。这是全球范围内第一次提出这样的想法。2012年12月13~15日，我在第一届深圳国际生物科技创新论坛暨展览会上报告了“生物谷歌”的概念，被媒体广泛报道。后来，我站在大数据的发展趋势及未来大数据的应用角度(尤其是医疗领域)又对未来进行了展望和畅想。

为了更好地确立规范，积极推动生物信息数据库在我国的发展，我们启动了一系列相关的标准制定工作。2013年12月，国家标准化管理委员会批复同意深圳华大基因研究院与中国食品发酵研究院、深圳市标准技术研究院组成ISO/TC276国内技术对口工作组，以工作组的形式承担国内技术对口单位的任务，成为国际相关标准制定的中国唯一对口单位。2014年1月，我们起草的《生物基因信息数据库建设与管理规范》作为深圳市地方标准发布实施。2014年5月，我参加了在柏林召开的ISO/TC276国际会议，代表中国发言并争取相应权益，强调凭借我们在生物信息领域的优势，中国完全有实力与各国一道引领相关国际标准的制定，各国代表也充分尊重和认可了我们的实力。这让我们更加感到生物信息数据库及相关领域的重要性和工作推进的迫切性。

畅想未来，一方面站在国家战略布局的角度，我国现在不断加强对于遗传资源、数据的保护和管理是十分必要的；另一方面，我们需要通过制定标准、规范，以更加开放的心态与国际加强合作，共享经验；最后，应该有更多的人参与到生物信息数据的产出、存储、挖掘、应用的研究和工作中去。回顾信息技术、互联网的发展历史，我们今天的生命科学生物信息的数据爆炸和当时的互联网信息爆炸何其相似。在未来的10~15年，非常有可能涌现出一批科研院所、企业实现颠覆性的研究和工作。“大数据”将给我们带来前所未有的机遇，生命科学正在向一个全新的领域迈进。

生物信息数据库(广义到生命科学大数据)的涵盖范围、涉及领域非常广。本书以“生物信息数据库”为重点对象，内容涵盖：国内外现状和未来问题挑战；生物信息数据的来源、类型、分析平台、质量标准、管理体系、备份恢复和安全管理；生物信息数据库的设计与开发；生物信息数据库的使用；伦理学问题；知识产权保护。由于篇幅原因，本书仅仅针对与“生物信息数据库”相关的部分内容，我们期待与大家分享工作中的经验体会，与各位同仁共同探讨和交流，促进我国生物信息数据库规范有序地发展，提升行业领域的整体水平，为未来的更快发展提供必要支撑。

最后感谢我的同事严志祥组织本书的编写，感谢魏晓锋、朱赢和操利超具体把握本书的技术内容；感谢陈凤珍、肖萍和袁翠红在资料汇集和文档校对中付出的努力，感谢李博、李玲、李霞、刘克、沈维燕、除超萍、游丽金等对本书编写提出的宝贵意见和建议；特别感谢国家基因库周欣和程乐对本书出版提供的大力支持；还要感谢科学出版社生物分社为本书的出版所做的大量工作，特别是王静分社长和夏梁编辑所付出的辛劳。

本书成稿仓促，虽经努力，文中仍不免有疏漏和错误之处，希望得到大家的批评和指正。我们也将进行持续的改进，不断提升现有内容质量。

张　勇

2014 年秋于深圳

目　录

第三篇 建立生物信息数据库的实用指南

第四篇 生物信息数据库的使用管理规范

第五篇　伦理学问题

第六篇　生物信息数据库知识产权保护

第一篇　生物信息数据库概述

本篇首先从整体上对生物信息数据库的发展历程等进行简单叙述。然后在分类的基础上，对一些代表性的生物信息数据库进行了详细介绍，包括数据库的建立、数据的内容以及数据提交和下载等方面。本篇的最后对国内外生物信息数据库的发展状况进行了简单介绍，并就在生物信息数据库的建设过程中存在的问题及解决思路进行了探讨。

1 生物信息数据库

1.1 生物信息数据库概述

最近五六十年,随着各种实验和测序技术的不断发展与完善,生物分子原始数据量急剧膨胀,积累了海量的分子生物学数据。同时,随着越来越多生物物种的基因组及相关序列测定的完成,获得数据已经不是生物研究的重心,现在生物学研究的重心是:从海量生物数据中挖掘有用的信息,并研究其所包含的生物学意义。

但是在面对如此大量的、以指数方式增长的数据时,传统的方法已经来不及迅速消化这些新数据,无法对这些数据进行有效的管理和维护。为了能更好地利用这些数据,有必要采用更有效的方法来管理和维护这些数据。为此已经产生了成百上千个生物信息数据库,来对这些信息进行收集、分类处理。《核酸研究》(*Nucleic Acids Research*,NAR)杂志在2014年的*Database Issue*中公布:NAR在线分子生物学数据库收集网站已经收录了1552个分子生物信息数据库(http://www.oxfordjournals.org/nar/database/a/)。这些数据库都是由专门的机构建立和维护的,他们负责收集、组织、管理和发布生物分子数据,向生物学研究人员提供大量有用的信息。现在的数据库不仅提供数据检索,还提供数据分析等服务,以满足生物学研究人员研究和应用的需要。

近年来因特网的飞速发展,使得在全球范围内有效管理飞速增长的生物信息数据成为可能。GenBank、EMBL(European Molecular Biology Laboratory)、DDBJ(DNA Data Bank of Japan)三大数据库于1988年建立了合作关系,共同成立了国际核酸序列数据库协会(International Nucleotide Sequence Database Collection,INSDC,http://www.insdc.org/),每天将更新的数据进行交换和共享,以保证各数据库中数据信息的完整与同步。每个机构负责收集来自不同地理分布的数据(EMBL负责欧洲,GenBank负责美洲,DDBJ负责亚洲等),对其进行分析及注释,然后将来自各地的所有信息汇总在一起,三个数据库共同享有并向世界开放。从理论上说,这三个数据库所拥有的DNA序列数据是完全相同的。

生物信息数据库逐渐发展到今天,已经形成了明显的特征。

1)数据量呈指数增长趋势,数据库的更新速度不断加快

生物分子数据库最突出的特征是数据量呈指数增长,随着数据量的爆炸性增长,数据库的更新速度也不断加快,许多数据库几乎是每天更新。

2)数据库使用频率增长更快

人们越来越感觉到生物分子数据的重要性,也认识到其中的价值,因此,各种数据库的使用人数和使用频率都在不断增加。

3)数据库的复杂程度不断增加

数据库中除了基本数据之外,还包括大量的注释、参考文献、与其他数据库的链接等

相关信息。例如,在 SwissProt 数据库中,注释项涉及蛋白质的功能、结构域和活性位点、二级结构、四级结构、翻译后修饰、与其他蛋白质的相似性、与该蛋白质关联的疾病、序列变化等。

4) 数据库网络化

《核酸研究》杂志 2013 年收录了 1512 个分子生物信息数据库,这些数据库大都可以通过因特网来访问,并且公共数据库之间相互链接,使用户可以迅速得到大量相关的生物分子信息。有的系统还将多个生物分子数据库整合在一起,形成具有专业用途的二次数据库和复合数据库等。

5) 面向应用

首先,各个数据库服务器除了提供生物数据之外,还提供许多工具,如核酸数据库提供的序列搜索、比对分析工具等,生物大分子结构数据库提供的结构比对程序、结构模拟程序等。其次,还在原始数据库的基础上开发了许多具有特殊生物学意义和专门用途的二次数据库、蛋白质分类数据库、蛋白质二级结构数据库等。

6) 先进的软硬件配置

从计算机硬件方面来看,许多数据库使用高性能的大型服务器,高效地管理数据和运行服务程序。从系统软件方面来看,许多数据库使用数据库管理系统。例如,欧洲生物信息学研究所(European Bioinformatics Institute,EBI)用 Oracle 数据库软件管理、维护核酸数据库 EMBL;而基因组数据库(genome database,GDB)则由 Sybase 数据库系统来管理和维护,即便是安装镜像,也需要有 Sybase 系统的支撑。

随着数据库复杂程度的不断增加,世界上已经建立了几百个二次数据库。值得一提的是现在一次数据库和二次数据库之间的界限已经越来越模糊,许多一次数据库在发展过程中已经兼具二次数据库的特点,如 SCOP(structural classification of proteins)、CATH(class,architecture,topology and homologous superfamily)等数据库都已经具有一些二次数据库的特点。

1.2 生物信息数据库的分类

现在的生物信息数据库数量巨大、种类繁多,归纳起来大致可以分为:核酸序列和结构数据库、蛋白质序列和结构数据库、常见的基因组数据库、综合数据库和其他主要数据库等。数据库的分类可参照表 1-1,一些代表数据库及网址见表 1-2。

表 1-1 数据库分类表

数据库分类	代表数据库
核酸序列数据库	GenBank、ENA、DDBJ
核酸结构数据库	NDB
蛋白质序列数据库	PIR、SwissProt、UniProt、OWL、PROSITE
蛋白质结构数据库	PDB、PDBj、PDBv、BMRB、wwPDB、SCOP、CATH、Pfam
蛋白质相互作用数据库	HPRD、BIND、IntAct、MINT、DIP、MIPS

续表

数据库分类	代表数据库	
常见的基因组数据库	常见模式生物的基因组数据库	Colibri、AceDB、SGD、ZFIN、FlyBase、MGI、TAIR
	微生物数据库	GenoList
	真核生物数据库	炎黄数据库、水稻数据库、熊猫数据库
疾病数据库	OMIM、ClinVar、dbSNP、LSDB、HGMD、单基因遗传病网	
综合数据库	NCBI、UCSC、EMBL、Ensembl、miRBase、dbRES、CBIS、HHMD、中国动物主题数据库	
代谢及功能注释数据库	KEGG、HMDB、DGA、HAMAP	
其他主要数据库	HAPMAP、BOLD、ENCODE、GO、DBD、European Ribosomal RNA database、GigaDB、CLiMB	

1.2.1 核酸序列数据库

最早的 DNA 序列数据库于 1982 年在欧洲分子生物学实验室诞生，随即开始了一个数据库爆炸的时代。国际上代表性的核酸序列数据库为：美国国家生物技术信息中心（National Center of Biotechnology Information，NCBI）管理的核酸序列数据库（GenBank）、欧洲生物信息学研究所（EBI）管理的核酸序列数据库（ENA）及日本国家遗传学研究所（National Institute of Genetics，NIG）管理的 DNA 数据库（DNA database of Japan，DDBJ）。三者的关系如图 1-1 所示。

图 1-1 GenBank、ENA、DDBJ 关系图

表 1-2 各代表数据库及网址

代表数据库	网址
Ensembl	http://asia.ensembl.org/index.html
AceDB	http://www.acedb.org/
SGD	http://www.yeastgenome.org/
UCSC	http://genome.ucsc.edu/
GenBank	http://www.ncbi.nlm.nih.gov/genbank/
DDBJ	http://www.ddbj.nig.ac.jp/
BioSino	http://www.biosino.org/pages/database.htm
PIR	http://pir.georgetown.edu/
TrEMBL	http://www.expasy.org/
NDB	http://ndbserver.rutgers.edu/
PDB	http://www.rcsb.org/pdb/home/home.do(http://www.wwpdb.org/)
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
CATH	http://www.cathdb.info/
KEGG	http://www.genome.jp/kegg/
Colibri	http://genolist.pasteur.fr/Colibri/
PROSITE	http://PROSITE.expasy.org/
DSSP	http://swift.cmbi.ru.nl/gv/dssp/
HSSP	http://swift.cmbi.ru.nl/gv/hssp/
NCBI	http://www.ncbi.nlm.nih.gov/
OWL	http://www.bioinf.man.ac.uk/dbbrowser/OWL/index.php
YH database	http://yh.genomics.org.cn/
BGI Rise Database	http://rise2.genomics.org.cn/

中国自主开发的核酸序列数据库——BioSino 数据库，由中国科学院上海生命科学研究院生物信息中心维护。

1.2.1.1 GenBank

GenBank 于 1979 年开始建设，1982 年投入使用，由美国国家生物技术信息中心（NCBI）构建、维护和管理（http://www.ncbi.nlm.nih.gov/genbank），图 1-2 为 GenBank 主页。GenBank 包含了所有已知的核酸序列和蛋白质序列，以及与它们相关的文献著作和生物学注释。GenBank 的主要数据来源为：由序列发现者提交的序列、由测序中心提交的大量表达序列标签（expressed sequence tag，EST）序列和其他测序数据、其他数据机构协作交换数据而来的数据及美国专利商标局提供的已发表专利的序列数据。

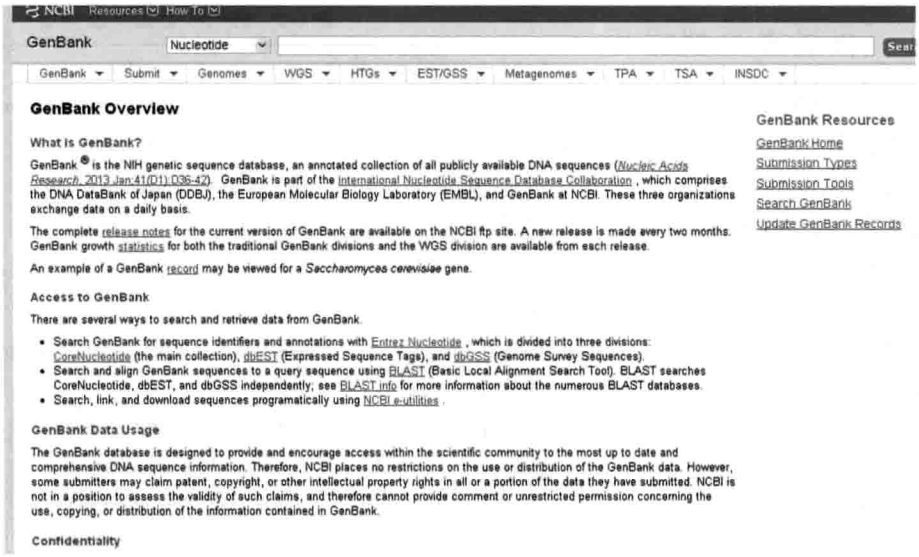

图 1-2 GenBank 数据库主页

GenBank 里的数据按来源大约有 55 000 个物种。每个记录代表了一个单独的、连续的、带有注释的 DNA 或 RNA 片段；每条 GenBank 数据记录包含了对序列的简要描述、学名、物种分类名称、参考文献、序列特征表以及序列本身的信息。序列特征表里包含对序列生物学特征注释，如编码区、蛋白质翻译、转录单位、重复区域、变异和修饰位点等。GenBank 中的数据文件，按照惯例被分成独立的类别，根据主要的分类群分为：细菌类（BCT）、病毒类（VRLA）、灵长类（PRI）、啮齿类（ROD），以及新增的 EST 数据、基因组测序数据、高通量基因组（high-throughput genomic，HTG）、环境样品（environment sample，ENV）等 16 类。

测序工作者可以把自己工作中获得的新序列提交给 NCBI，添加到 GenBank 数据库。BankIt 是 NCBI 提供的一个在线提交序列的工具（http://www.ncbi.nlm.nih.gov/BankIt），由一系列表单构成，使用比较简便。用户提交序列后，会从电子邮件收到自动生

成的数据条目、GenBank 的新序列编号,以及完成注释后的完整的数据记录。BankIt 工具适用于独立测序工作者提交少量序列,特别是只有一条或很少几条记录需要提交时,应该选择 BankIt。BankIt 不适合大量序列的提交,也不适合提交很长的序列,EST 序列和 GSS 序列也不应该用 BankIt 提交。

Sequin 是 NCBI 提供的独立的多平台提交程序(http://www.ncbi.nlm.nih.gov/Sequin),大量的序列提交可以由 Sequin 程序完成。Sequin 程序能方便地编辑和处理复杂注释,并包含一系列内建的检查函数来保证序列的质量。它还可用于提交来自系统进化、种群和突变研究的序列,可以加入比对的数据。在不同操作系统下运行的 Sequin 程序都可以在 FTP 上找到并下载使用。

NCBI 以传统的纯文本文件格式在 FTP 服务器上发布 GenBank 的数据,并且每两个月更新一次。GenBank 的数据类型见表 1-3,用户可以从 NCBI 的 FTP 服务器上免费下载完整的 GenBank 数据库数据集,或下载积累的新数据。NCBI 还提供广泛的数据查询、序列相似性搜索以及其他分析服务,用户可以从 NCBI 的主页上找到这些服务。

表 1-3 GenBank 数据类型

缩写	代表数据类型
PRI	primate sequences
ROD	rodent sequences
MAN	other mammalian sequences
VRT	other vertebrate sequences
INV	invertebrate sequences
PLN	plant, fungal and algal sequences
BCT	bacterial sequences
VRL	viral sequences
PHG	bacteriophage sequences
SYN	synthetic sequences
UNA	unannotated sequences
EST	EST sequences(expressed sequence tags)
PAT	patent sequences
STS	STS sequences(sequence tagged sites)
GSS	GSS sequences(genose survey sequences)
HTG	HTG sequences(high throughput genosic sequences)
HTC	HTC sequences(high throughput cDNA sequences)
ENV	Environmental sampling sequences
CON	Constructed sequences

NCBI 的数据库检索查询系统是 Entrez(http://www.ncbi.nlm.nih.gov/sites/gquery),使用 Entrez 检索系统可以访问 GenBank 中的序列记录。Entrez 是一个基于 Web 的数据库检索系统,可以检索 35 个数据库。利用 Entrez 系统,用户不仅可以方便地检索 GenBank 的核酸数据,还可以检索来自 GenBank 和其他数据库的 DNA 和蛋白质序列数据、基因组图谱数据、种群序列数据、来源于分子模型数据库(molecular modeling database, MMDB)的蛋白质结构数据,以及由 PubMed 获得 MEDLINE 的文献数据。Entrez 中搜索到的数据可以以多种格式输出,也可以逐个下载或打包下载。

BLAST(basic local alignment search tool)序列相似性搜索是 GenBank 数据最基本和使用最多的分析方式。NCBI 提供 BLAST 系列程序集,用于检测一条查询序列与数据库所有序列的相似性。BLAST 搜索可以在网站上运行,也可以在 FTP 上下载独立的程序集在本地运行。