国家出版基金项目
NATIONAL PUBLICATION FOUNDATION

“十三五”国家重点出版物出版规划项目

软物质前沿科学丛书

生物分子大数据分析

Big Data Analysis of Biomolecules

赵蕴杰/著

科学出版社
龍門書局

“十三五”国家重点出版物出版规划项目

软物质前沿科学丛书

生物分子大数据分析

Big Data Analysis of Biomolecules

赵蕴杰 著

科学出版社
龍門書局

北京

内 容 简 介

本书主要介绍了物理模型和大数据分析方法在生物分子中的应用，重点介绍了动态网络、直接耦合分析和机器学习等大数据分析方法。通过阅读本书，读者不仅可以了解网络和机器学习等问题的基础知识，还可以通过小分子结合靶点分析、生物分子相互作用预测和小分子代谢物分类等研究实例，了解相关算法的使用，帮助读者结合自身研究选择合适的大数据分析方法。

本书适合对生物问题感兴趣的物理、数学、化学、生物和计算机专业的研究生阅读，对从事生物物理、生物信息学和分子生物学等研究领域的读者也有参考价值。

图书在版编目(CIP)数据

生物分子大数据分析/赵蕴杰著. —北京：龙门书局, 2019.9
（软物质前沿科学丛书）
“十三五”国家重点出版物出版规划项目 国家出版基金项目
ISBN 978-7-5088-5633-9

Ⅰ. ①生… Ⅱ. ①赵… Ⅲ. ①数据处理–应用–分子生物学 Ⅳ. ①Q7-39

中国版本图书馆 CIP 数据核字(2019) 第 182338 号

责任编辑：钱 俊 陈艳峰 / 责任校对：杨 然
责任印制：吴兆东 / 封面设计：无极书装

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

北京虎彩文化传播有限公司 印刷
科学出版社发行 各地新华书店经销
*
2019 年 9 月第 一 版 开本：720 × 1000 B5
2019 年 9 月第一次印刷 印张：9 1/2
字数：188 000

定价：98.00 元

(如有印装质量问题，我社负责调换)

软物质前沿科学丛书编委会

作 者 简 介

赵蕴杰，理论物理博士，华中师范大学物理科学与技术学院副教授。研究方向为生物物理学，在 *Nature Immunology*，*Nucleic Acids Research*，*Bioinformatics* 等刊物上发表学术论文 30 余篇，主持和参与国家自然科学基金、湖北省自然科学基金等多项课题。2016 年入选湖北省“楚天学者”人才计划，2019 年入选华中师范大学“桂子青年学者”学者名师支持计划。主要教授普通物理、生物信息学、机器学习与生物物理、分子生物物理等课程。受邀担任 *Scientific Reports* 的 Editorial Board Member 和 *Physical Reviews* 等国际学术杂志审稿人。

丛 书 序

社会文明的进步、历史的断代，通常以人类掌握的技术工具材料来刻画，如远古的石器时代、商周的青铜器时代、在冶炼青铜的基础上逐渐掌握了冶炼铁的技术之后的铁器时代，这些时代的名称反映了人类最初学会使用的主要是硬物质. 同样，20 世纪的物理学家一开始也是致力于研究硬物质，像金属、半导体以及陶瓷，掌握这些材料使大规模集成电路技术成为可能，并开创了信息时代. 进入 21 世纪，人们自然要问，什么材料代表当今时代的特征？什么是物理学最有发展前途的新研究领域？

1991 年，诺贝尔物理学奖得主德热纳最先给出回答：这个领域就是其得奖演讲的题目 ——“软物质”. 以《欧洲物理杂志》B 分册的划分，它也被称为软凝聚态物质，所辖学科依次为液晶、聚合物、双亲分子、生物膜、胶体、黏胶及颗粒等.

2004 年，以 1977 年诺贝尔物理学奖得主、固体物理学家 P.W. 安德森为首的 80 余位著名物理学家曾以 “关联物质新领域” 为题召开研讨会，将凝聚态物理分为硬物质物理与软物质物理，认为软物质 (包括生物体系) 面临新的问题和挑战，需要发展新的物理学.

2005 年，*Science* 提出了 125 个世界性科学前沿问题，其中 13 个直接与软物质交叉学科有关. “自组织的发展程度” 更是被列入前 25 个最重要的世界性课题中的第 18 位，“玻璃化转变和玻璃的本质” 也被认为是最具有挑战性的基础物理问题以及当今凝聚态物理的一个重大研究前沿.

进入新世纪，软物质在国外受到高度重视，如 2015 年，爱丁堡大学软物质领域学者 Michael Cates 教授被选为剑桥大学卢卡斯讲座教授. 大家知道，这个讲座是时代研究热门领域的方向标，牛顿、霍金都任过这个最著名的卢卡斯讲座教授. 发达国家多数大学的物理系和研究机构已纷纷建立软物质物理的研究方向.

虽然在软物质研究的早期历史上，享誉世界的大科学家如爱因斯坦、朗缪尔、弗洛里等都做出过开创性贡献，荣获诺贝尔物理奖或化学奖. 但软物质物理学发展更为迅猛还是自德热纳 1991 年正式命名 “软物质” 以来，软物质物理不仅大大拓展了物理学的研究对象，还对物理学基础研究尤其是与非平衡现象 (如生命现象) 密切相关的物理学提出了重大挑战. 软物质泛指处于固体和理想流体之间的复杂的凝聚态物质，主要共同点是其基本单元之间的相互作用比较弱 (约为室温热能量级)，因而易受温度影响，熵效应显著，且易形成有序结构. 因此具有显著热波动、多个亚稳状态、介观尺度自组装结构、熵驱动的顺序无序相变、宏观的灵活性等特征. 简单地说，这些体系都体现了“小刺激，大反应”和强非线性的特性. 这些特性

并非仅仅由纳观组织或原子或分子的水平结构决定，更多是由介观多级自组装结构决定. 处于这种状态的常见物质体系包括胶体、液晶、高分子及超分子、泡沫、乳液、凝胶、颗粒物质、玻璃、生物体系等. 软物质不仅广泛存在于自然界，而且由于其丰富、奇特的物理学性质，在人类的生活和生产活动中也得到广泛应用，常见的有液晶、柔性电子、塑料、橡胶、颜料、墨水、牙膏、清洁剂、护肤品、食品添加剂等. 由于其巨大的实用性以及迷人的物理性质，软物质自 19 世纪中后期进入科学家视野以来，就不断吸引着来自物理、化学、力学、生物学、材料科学、医学、数学等不同学科领域的大批研究者. 近二十年来更是快速发展成为一个高度交叉的庞大的研究方向，在基础科学和实际应用方面都有重大意义.

为推动我国软物质研究，为国民经济作出应有贡献，在国家自然科学基金委员会中国科学院学科发展战略研究合作项目 “软凝聚态物理学的若干前沿问题” (2013.7~2015.6) 资助下，本丛书主编组织了我国高校与研究院所上百位分布在数学、物理、化学、生命科学、力学等领域的长期从事软物质研究的科技工作者，参与本项目的研究工作. 在充分调研的基础上，通过多次召开软物质科研论坛与研讨会，完成了一份 80 万字研究报告，全面系统地展现了软凝聚态物理学的发展历史、国内外研究现状，凝练出该交叉学科的重要研究方向，为我国科技管理部门部署软物质物理研究提供一份既翔实又前瞻的路线图.

作为战略报告的推广成果，参加本项目的部分专家在《物理学报》出版了软凝聚态物理学术专辑，共计 30 篇综述. 同时，本项目还受到科学出版社关注，双方达成了 “软物质前沿科学丛书” 的出版计划. 这将是国内第一套系统总结该领域理论、实验和方法的专业丛书，对从事相关领域的研究人员将起到重要参考作用. 因此，我们与科学出版社商讨了合作事项，成立了丛书编委会，并对丛书做了初步规划. 编委会邀请了 30 多位不同背景的软物质领域的国内外专家共同完成这一系列专著. 这套丛书将为读者提供软物质研究从基础到前沿的各个领域的最新进展，涵盖软物质研究的主要方面，包括理论建模、先进的探测和加工技术等.

由于我们对于软物质这一发展中的交叉科学的了解不很全面，不可能做到计划的 “一劳永逸”，而且缺乏组织出版一个进行时学科的丛书的实践经验，为此，我们要特别感谢科学出版社钱俊编辑，他跟踪了我们咨询项目启动到完成的全过程，并参与本丛书的策划.

我们欢迎更多相关同行撰写著作加入本丛书，为推动软物质科学在国内的发展做出贡献.

主　编　　欧阳钟灿

执行主编　　刘向阳

2017 年 8 月

序　　言

生物物理学是一门传统的学科，目前已经有很多很好的教材和专著讲述生物物理的基本理论和方法。随着生物大数据时代的到来，新的研究方向和研究方法不断出现，学生和研究人员都非常需要比较系统地掌握这方面的知识，特别是生物大数据的分析方法，因此需要有教材或参考书及时反映生物物理学相关领域这些新的发展。赵蕴杰博士编写的《生物分子大数据分析》一书很好地满足了这方面的需求。该书针对生物大数据，介绍了几种近年来被研究者广泛关注和大量使用的重要的分析方法，包括动态网络、直接耦合分析和机器学习等。特别是通过作者自己的研究工作，书中详细地介绍了这些方法在生物物理学以及相关研究领域中的应用，不仅能让读者接触到研究前沿，而且能让读者掌握这些方法并用于自己的研究，非常有实用价值。赵蕴杰博士在书中介绍的研究方向和研究方法上都有比较深入的探索，相信该书对想要了解和进入生物大数据分析相关领域的读者会有很大的帮助。

肖　奕

华中科技大学物理学院生物物理研究所

2019 年春节于武汉

前　言

生物物理学是推动科学技术发展的基础理论学科，大数据分析的迅猛发展在我们研究生物物理学以及探索新理论、寻找新方法中起着非常重要的促进作用。2014年 Mike May 在*Science*上发表了一篇题为*Big biological impacts from big data*的论文，提出大数据和大数据分析对生物学和医学的发展有重要的促进作用。

21 世纪的我们生活在一个大数据的时代，海量的图片、文字、地理信息和消费记录等大数据与我们的生活息息相关。1 分钟热门微博的转发量可以超过 10 万，1 分钟中国在线移动支付金额有近 4 亿元，1 分钟中国约有 7.6 万件快递被收发。大数据与大数据分析在金融、通信、军事和科学研究等诸多领域有着越来越重要的贡献。

生物大数据也在急剧增加。第一个人类基因组的测序工作历时 13 年并投入了 30 亿美元才获得 30 亿核苷酸序列。现在人类基因组测序仅需 1000 美元，每周可以产生 320 多个基因组数据。随着生物实验技术的不断改进与完善，生物序列数据、二级结构数据、三级结构数据、代谢物数据和小分子药物数据等海量数据急剧增加。生物大数据有明显的复杂性特征。药物设计的研发工作往往需要联合基因组数据、蛋白组数据、细胞信号传导、临床研究和环境科学等复杂的研究数据。因此，目前急需能快速整合分析生物大数据的新理论与新方法。近些年，生物大数据与大数据分析已经取得了相当丰硕的成果。复杂网络分析、机器学习和深度学习等大数据分析方法对理解生物学功能、解释疾病机理和预防疾病等问题有重要的帮助。

生命体由大量的蛋白质、核酸等生物大分子以及小分子组成，这些生物分子以及之间的相互作用和化学反应通路构成了复杂的生物网络系统。本书从生物分子中几种典型的大数据类型出发，结合分子动力学模拟、复杂网络分析、多序列比对、机器学习和深度学习等理论模型，对生物分子的结构特征和结合靶点预测等问题进行研究；探讨了生物分子序列共进化和三级结构空间相互作用的关系；同时还讨论分析了深度学习模型在代谢物分子分类等问题中的应用。

我的学生和相应领域的研究专家参与了本书的数据搜集、撰写和校对工作，分别是王慧雯 (华中师范大学博士研究生，复杂网络分析)、简弋人 (美国乔治·华盛顿大学博士研究生，RNA 相互作用预测)、刘志超 (美国乔治·华盛顿大学博士研究生，生物代谢物分析)、曾洲豪 (美国乔治·华盛顿大学博士研究生，深度学习)、王凯丽 (华中师范大学硕士研究生)、王晓因 (华中师范大学硕士研究生)、邱嘉迪 (华中师范大学本科生) 和刘文硕 (华中师范大学本科生)。感谢华中科技大学物理

学院肖奕教授为本书作序，感谢美国乔治·华盛顿大学物理系曾辰教授为本书提供了大量宝贵的修改意见。本书的出版得到了华中师范大学物理科学与技术学院和科学出版社的大力支持。感谢华中师范大学物理科学与技术学院贾亚教授、科学出版社提供的帮助和关心支持。本书的写作占用了大量的业余时间，感谢家人的理解和支持。

由于水平有限，书中难免有不当之处，还望读者指正，请将建议发到如下邮箱：yjzhaowh@mail.ccnu.edu.cn。

赵蕴杰

2019 年春于武汉

目　　录

第1章 绪 论

生物分子包括蛋白质、核酸和其他生物体内的各类有机分子，是所有生命有机体的核心分子，有遗传信息存储、能量存储、催化反应、生物代谢、分子运输、病毒防御和结构支持等重要的生物学功能。我们处在生物分子数据信息急剧增长和分析方法不断涌现的大数据时代[1−8]。生物实验大数据和生物信息学大数据分析方法使我们可以从分子水平理解生物分子的结构特征、生物学功能和调控机制，并将其应用于揭示疾病的发病机理和相关疾病的诊断治疗。

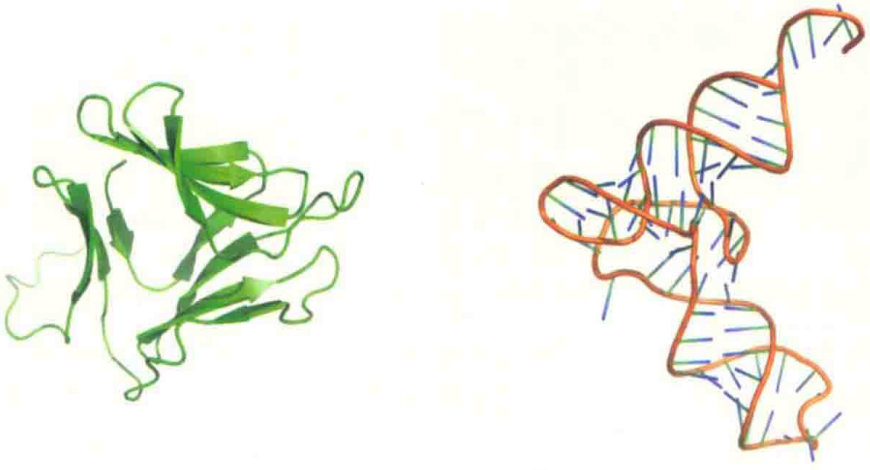

图 1.1 蛋白质结构与 RNA 结构示意图。左图为植物凝集素蛋白 (PDB code：1KJ1)，右图为 tRNA(PDB code：5UD5)

1.1 迅速增长的生物数据

近 20 年来，生物分子实验数据发展的一个显著特点是数据量的急剧膨胀，迅速形成和产生了拥有海量数据的数据信息库，提供了各种需要的实验信息。生物分子数据库包括多种类型：有序列数据、二级结构数据、三级结构数据、代谢物数据、分子结合靶点数据、药物结合口袋数据、小分子药物数据和基因组数据等不同类型的数据库。

随着高通量测序技术的出现和不断完善，遍布世界各地的大学和实验室等研究机构每天都在测定和产生不同物种的序列，源源不断的序列信息更新到数据库中，生物分子序列信息的增长量十分惊人。例如，华大基因通过基因测序已完成超过

300 万例 HPV 检测，179 万例耳聋基因检测和 340 万例无创产前检测；中国国家基因库于 2011 年开始建设，涵盖了人类微生物资源、海洋多样性资源、疾病资源、植物资源、人类遗传资源、动物资源和地球微生物资源等亿万样本资源，形成了有一定规模的生命大数据库，支撑服务于我国临床检测、疾病防治、生物农业、物种多样性保护等生命经济的各个领域；美国国家生物技术信息中心建立的序列数据库 GeneBank(https://www.ncbi.nlm.nih.gov/genbank) 在 2000 年底约有 100 亿个碱基对，2010 年底，GeneBank 中的数据涨到约 1200 亿个碱基对，2018 年 GeneBank 中的数据则涨到约 2850 亿个碱基对，GeneBank 中的序列数据随着时间有大幅度的增加 (图 1.2)[9−11]。

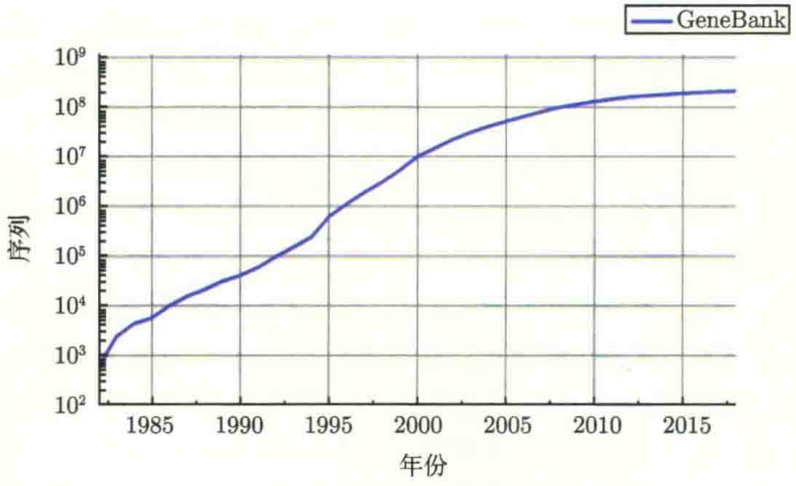

图 1.2　GeneBank 数据库中的序列增长情况 (截至 2018 年，数据来自 https://www.ncbi.nlm.nih.gov/genbank)

NONCODE 数据库 (http://www.noncode.org/) 在 GeneBank 的海量数据中筛选出了特定长度的非编码 RNA 基因，是整合分析非编码 RNA 的综合数据平台 [12−14]。NONCODE 数据库主要有：非编码 RNA 的序列信息 (对相应序列进行了注释)；长链非编码 RNA 信息 (lncRNA 基因)；RNA-seq 的相关数据；已发表的非编码 RNA 相关论文；遗传变异数据库 (dbSNP)；微阵列数据和全基因组关联研究 (GWAS) 等。NONCODE 数据库从非编码 RNA 的大数据中理解并解释长链非编码 RNA 基因与疾病的关系，对现代生物学和医学研究有重要的帮助。目前 NONCODE 数据库已经更新到了第 5 版，包含人、小鼠、牛、大鼠、黑猩猩、大猩猩、红毛猩猩、恒河猴、负鼠、鸭嘴兽、鸡、斑马鱼、果蝇、线虫、酵母、拟南芥和猪等 17 个物种，共有 548 640 个长链非编码 RNA 转录物和 354 855 个长链非编码 RNA 基因 (详见表 1.1)。

表 1.1 NONCODE 数据库中不同物种序列的分布情况

(数据来自 http://www.noncode.org/)

物种	lncRNA 转录物数目	lncRNA 基因数目
人	172 216	96 308
小鼠	131 697	87 774
牛	23 515	22 227
大鼠	24 879	22 127
黑猩猩	18 004	12 790
大猩猩	18 539	15 095
红毛猩猩	15 178	13 106
恒河猴	9 128	6 010
负鼠	27 167	17 795
鸭嘴兽	11 210	9 163
鸡	12 850	9 527
斑马鱼	4 852	3 503
果蝇	42 848	15 543
线虫	3 154	2 552
酵母	55	52
拟南芥	3 763	3 472
猪	29 585	17 811
总共	548 640	354 855

生物分子的三维结构是从分子层次理解和阐明生物学规律的基础。因此，解析生物分子的三维结构是研究结构–功能关系，理解其作用机制的重要步骤。目前，测定生物分子三维结构的实验方法主要有：X 射线晶体衍射分析、核磁共振波谱分析和冷冻电镜实验。其中，X 射线晶体衍射分析是应用最为广泛的三维结构解析技术，主要步骤为：首先，将纯化的生物分子结晶；然后，搜集和处理 X 射线衍射实验数据；最后，确定相位并修正结构。然而，X 射线晶体衍射分析方法需要较好的晶体结构，对蛋白质或核酸等生物分子的结晶要求较高，随机聚合沉淀等结晶是该技术的主要瓶颈。如何筛选和优化生物分子形成有序晶体，限制了 X 射线晶体衍射技术在生物分子上的应用。

核磁共振波谱分析是另一个普遍应用的三维结构测定方法，在医学诊断等研究中也有相当广泛的应用。核磁共振波谱分析测定分子三维结构的精度取决于实验数据的质量，目前已可以达到相当于 X 射线衍射晶体 2Å的实验结构精度。除此以外，核磁共振波谱分析还可以研究生物分子的动力学性质、折叠过程和结构变化等问题。但是，核磁共振波谱分析仅适用于分子量较小的蛋白质、核糖核酸 (RNA) 等生物分子的结构测定，对于拓扑结构较为复杂的复合体和大分子则只能测定其部分结构域，较难做出精确解析。

冷冻电镜技术是近几年发展比较迅速的一种结构测定方法，可以用来解析尺

寸较大的复合体三维结构[15–17]。冷冻电镜不需要生物分子形成晶体结构，仅需少量的生物样品就可以通过快速冷冻获得生物大分子的结构特征。近年来，冷冻电镜技术硬件和软件的快速发展极大地提高了冷冻电镜的应用范围，可以得到近原子分辨率的生物大分子结构。另外，图像采集质量的提高、基于深度学习数据处理等技术的出现提高了冷冻电镜的实验精度，高分辨率生物大分子结构的出现对理解生物学调控机理有重要的贡献。

随着 X 射线晶体衍射、核磁共振波谱分析和冷冻电镜实验技术的不断完善，PDB 结构数据库 (www.rcsb.org) 中的生物分子结构数据得到显著增长：2008 年 PDB 结构数据库中生物分子结构的年增加量为 7075 个结构数据，2017 年数据库中的生物分子结构的年增加量则为 13 049 个结构数据 (图 1.3)[18,19]。近十年，越来越多的科研工作者正在使用 PDB 数据库中的结构数据来分析生物分子的性质、特征和理解其生物学机理 (图 1.4)。

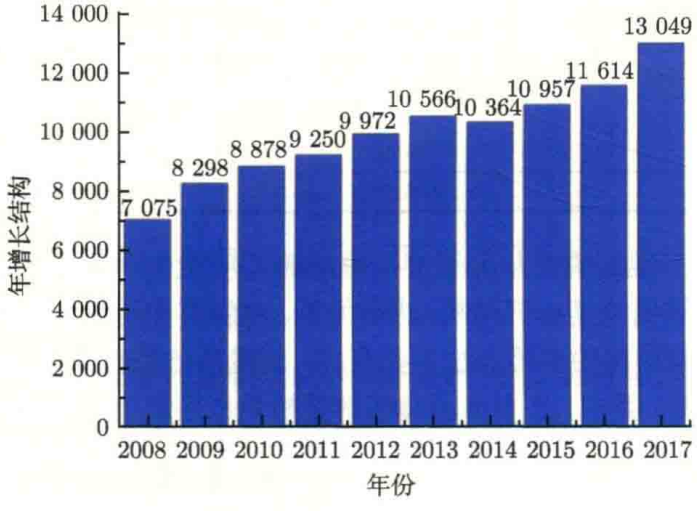

图 1.3　PDB 结构数据库的年增长结构数据 (数据来自 www.rcsb.org)

蛋白质结构分类数据库 (Structural Classification of Proteins-extended, SCOPe) 是伯克利加州大学实验室在 PDB 结构的基础上搭建的蛋白质结构分类数据库[20]。目前，该数据库中有 alpha 结构蛋白 289 类，beta 结构蛋白 178 类，alpha/beta 结构蛋白 148 类，alpha+beta 结构蛋白 388 类，多结构域蛋白 71 类，膜蛋白 60 类，小蛋白 98 类等不同类型的蛋白质结构 (截至 2018 年 10 月)。CATH 是伦敦大学 20 世纪 90 年代中期开发和维护的蛋白质结构分类数据库，提供蛋白质结构域分类信息 (http://www.cathdb.info)[21]。CATH 从 PDB 结构数据库中识别蛋白质三维结构的蛋白质结构域，并按照拓扑结构和同源相似性进行了分类，将 95 000 000 个蛋白质结构域分成了 6119 个结构域家族。

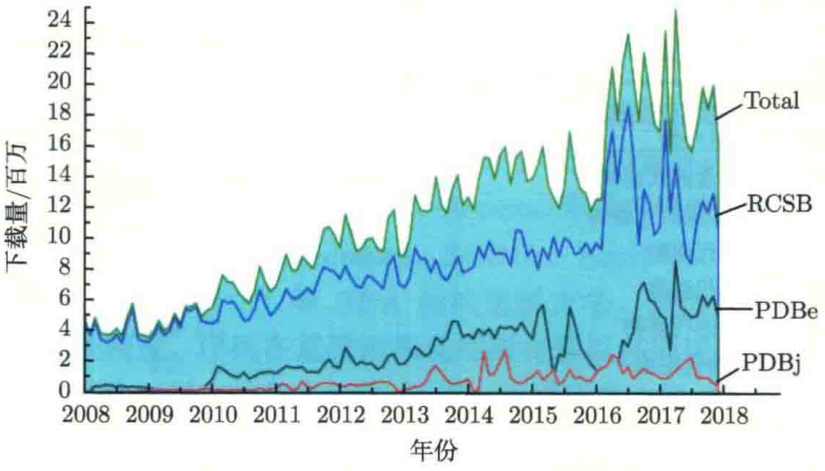

图 1.4 PDB 结构数据库的年下载量 (数据来自 www.rcsb.org)

KEGG(https://www.kegg.jp) 是从分子层次了解生物系统功能的数据库 [22]。该数据库通过基因组大规模测序和高通量实验技术等实验方法搜集分子水平信号，从而理解细胞、生物有机体和生态系统等不同层次的规律。其中，KEGG 代谢数据库汇集了生物系统中小分子、生物聚合物和其他化学物质，也搜集了生物代谢、膜转运、信号传递、细胞周期等生化过程信息。

HMDB(http://www.hmdb.ca) 是专门针对人类代谢物和代谢信息的数据库 [23]。该数据库收集整理了相关书籍、期刊文献和其他代谢数据库中的信息，有质谱 (Mass Spectra，MS) 和核磁共振 (Nuclear Magnetic Resonance，NMR) 等实验技术对尿液和血液等样本分析的结果。参考质谱和核磁共振实验结果，该数据库还有化合物描述、结构信息、物化数据、疾病相关性、通路信息、酶数据、基因序列数据、SNP 和突变数据等。目前，该数据库包含 114 100 个代谢物条目 (详见表 1.2)，包括水溶性和脂溶性代谢物以及被视为丰富 (>1 km) 或相对罕见 (<1 nm) 的代谢物，与其他数据库 (KEGG、PubChem、ChEBI、PDB、UniProt 和 GenBank) 有较为完善的交叉引用链接，对生物化学和代谢组学等相关领域的研究学者有重要的帮助。

DrugBank(http://www.drugbank.ca) 是由加拿大卫生研究院发展的药物数据库，搜集整理了小分子药物、药物靶点、药物靶点相互作用和药物机理等信息 [24]。2006 年建立以来，DrugBank 数据库中的药物数目和药物相互作用数量等数据都有了较大增长，现已收录药物 10 562 个，其中已批准的小分子药物有 3254 种 (详见表 1.3)。目前，DrugBank 数据库更新收录了数百种药物对代谢物水平、基因表达水平和蛋白质表达水平调节的信息，增加了老药新用试验和新药临床试验的数据。DrugBank 对改善药物有效性、药物耐药性和药物安全等问题有显著