引用格式：江海平，高纯纯，刘文豪，等．数据驱动的生命科学研究进展．中国科学院院刊，2024, 39(5): 862-871, doi: 10.16418/j.issn.1000-3045.20240225003.
Jiang H P, Gao C C, Liu W H, et al. Advances in data-driven life sciences research. Bulletin of Chinese Academy of Sciences, 2024, 39(5): 862-871, doi: 10.16418/j.issn.1000-3045.20240225003. (in Chinese)

数据驱动的生命科学研究进展

江海平[1,2]　高纯纯[3]　刘文豪[1,2]　杨运桂[3]　李　鑫[1,2*]

1 中国科学院动物研究所　北京　100101
2 北京干细胞与再生医学研究院　北京　100101
3 国家生物信息中心　北京　100101

摘要　生命科学发展日新月异，伴随着大量实验技术的更新，生物大数据逐渐产生并在生命科学研究中扮演着日益重要的角色。首先，生物大数据具有多样性和复杂性，包括基因组数据、表观基因组数据、蛋白质组数据等多种类型。这些数据为研究人员提供了更全面的信息，有助于揭示生命现象背后的规律。其次，数据驱动的生命科学新发展和应用涵盖了基因编辑、精准医疗、药物研发等诸多领域，为人类健康和生命质量提供了前所未有的可能性。然而，生命科学研究大数据时代也面临着包括数据存储、数据共享、隐私保护等多方面的问题，以及如何将海量数据转化为可靠的科学发现等挑战。文章简要概括了生物数据推动生命科学的发展规律，梳理了生物大数据组成、特点及来源，阐述并讨论了数据驱动的生命科学研究新范式下的共性问题和我国面临的挑战。

关键词　科学研究范式，大数据，生命科学

DOI　10.16418/j.issn.1000-3045.20240225003
CSTR　32128.14.CASbulletin.20240225003

1 生物数据推动生命科学发展阶段的演变

在过去的几个世纪中，生命科学一直处于快速发展和演变的阶段，从最初对生命现象的简单观察和描述，到如今分子生物学、基因组学和系统生物学等领域的兴起，生命科学研究范式持续演变[1,2]。这种研究范式的变化深受生物数据类型和规模的发展所推动，并带来了生命科学发展演进的3个阶段（图1）——每个阶段都在前一个阶段的基础上递进，不断涌现新的技术和方法来快速推动生命科学研究的不断进步。

*通信作者
资助项目：中国科学院稳定支持基础研究领域青年团队计划（YSBR-076）
修改稿收到日期：2024年5月8日

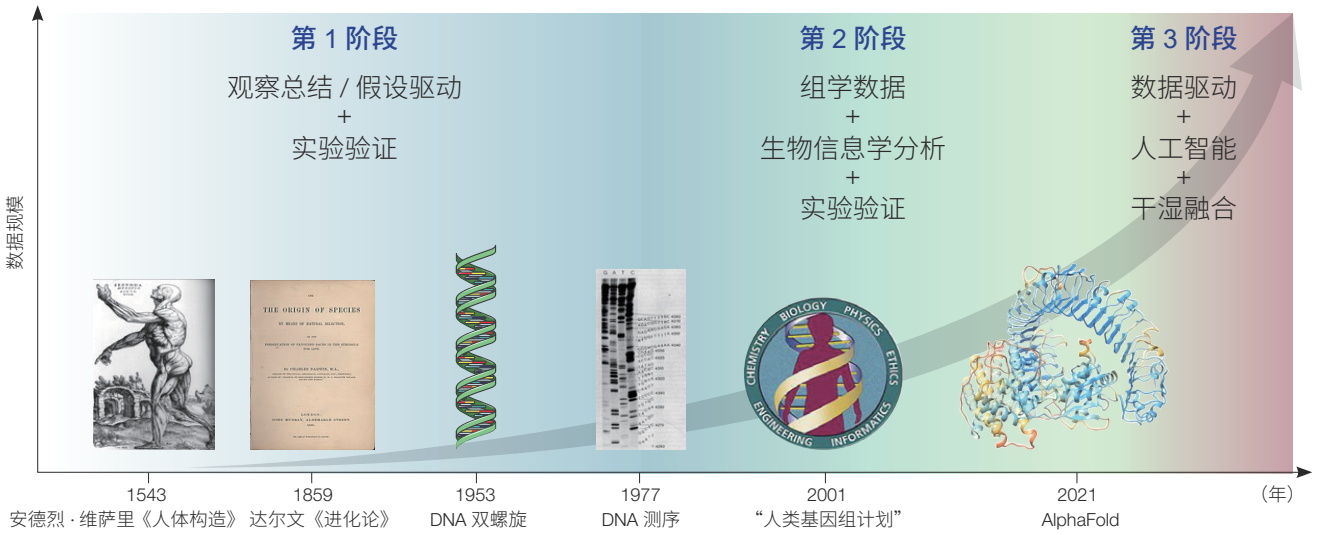

图 1　生物数据发展和生命科学阶段性发展

Figure 1　Developing biological data and evolving stages of life science research

第 1 阶段（16 世纪—20 世纪下半叶）：以观察总结和假设驱动为主，实验数据作为辅助支持和验证依据。在早期，生物学家主要依靠手工实验和观察描述获取数据，并从中提炼归纳出一些假说。但这些数据通常是表面的、局部的、有限的，产生的假说也是宏观和粗略的，无法对生命的深层机制进行解析。其原因在于认知水平和技术的限制导致无法获得和解析更深层次的生物学数据。这一时期的生命科学研究的典型代表有：16 世纪的安德烈 · 维萨里[3]通过动物和人体的解剖数据全面认识机体结构；19 世纪，达尔文[[4]]通过环球考察采集和分析大量标本数据提出进化论等。其后，随着物理学、化学等学科的发展，以及实验技术和分析方法的快速进步，尤其是DNA 双螺旋结构的发现和中心法则的提出[5]，将生命科学研究引入分子生物学时代。生物学家可以将复杂的生命系统拆解为微观的分子和细胞组分并逐个进行研究，以获得对生物系统单一维度、深层次的描述数据。研究人员通常采用被动分析的方法，即根据事先提出的假设来遍历和解释实验数据，此时形成的是对生命系统深入却零散、片面的认知。

第 2 阶段（20 世纪下半叶—21 世纪初）：以组学数据为基础，结合生物信息学分析和实验验证。测序技术的出现[6]和“人类基因组计划”的实施[7]将生命科学引入了高通量生物研究时代。基因组学、转录组学、表观组学、糖组学等多种组学技术呈现了细胞在不同层面的整体生命图景。生物学家能够在早期发育、癌症、衰老、疾病等多个生命过程中进行高通量、大规模的数据采集。此时，他们不再局限于验证特定的假设，而是通过多种组学数据来探索未知领域。多组学数据的分析需要更复杂的计算工具和算法，包括生物信息学、统计学等。这些工具和方法帮助研究人员从海量数据中发现隐藏的模式和关联，从而获得更全面、更深入的生物学知识。另外，使用生物信息学对组学数据分析获得的知识还需要使用湿实验进行验证。尽管这一阶段能够对生物学数据进行低维度的描述和解释，却难以对复杂的生命系统进行高维度模拟，以实现对生命的全面系统解析。

第 3 阶段（21 世纪初至今）：以生物大数据驱动，使用人工智能和干湿融合对生命系统进行解析与重构。生命系统呈现分子、细胞、组织、个体等多层次的结构，并且这些层次之间高度互联、动态调控，形成了一个复杂的系统；而由此获得的数据也具有多层

次、动态变化的特点。此外，随着生命科学研究的不断深入，海量的多组学数据、文献资料和其他生物学数据持续涌现和积累，从而导致数据规模和复杂性进一步增加。这种多类型、多维度且体量巨大的生物学数据被称为生物大数据。然而，传统的数据分析方法已经无法满足处理这一复杂性的需求。针对不同层次、不同维度、不同类型的生物大数据进行有效整合、汇集和深入分析，以揭示其中蕴含的高维度生物规律，成为当今生命科学研究面临的挑战之一。人工智能，尤其是神经网络技术，因其擅长从低维度的大规模数据中提取高维度隐匿规律的优势成为解决这一挑战的有效工具。例如，AlphaFold能够预测蛋白质的三维结构[8]，GeneCompass等工具能模拟基因调控网络[9]。这些工具和技术证明了使用人工智能可以挖掘生物大数据中数据之间的关联，抽提生命的内在结构，从而更全面地理解生命现象的本质和规律，揭示生物体内部复杂的互动关系和调控机制。然而，当前人工智能技术仍然仅能有效整合、分析某一层面的生物数据（如转录组）。要实现对复杂互联的生命系统进行全面、系统和深刻的认知，需要积累更多的系统性生物大数据，并运用人工智能技术对多模态的生物大数据进行有效整合，以实现对生命系统整体图景的认知。而且，人工智能指导的自动化机器人已经实现了在化学和材料学上自主设计、规划和执行真实世界的实验，从而显著提高了科学发现的速度和数量，并改善了实验结果的可复制性和可靠性[10,11]。未来使用生物大数据训练的人工智能结合自动化机器人，将可能建立干湿融合的自进化研究新范式，以实现对更复杂的生命系统进行更高效和更深入的解析。

综上，生物学数据推动生命科学发展经历了从观察总结和假设驱动为主、组学数据为基础到生物大数据驱动的3个递进阶段。在这个过程中，生物学数据呈现规模递增、类型丰富和层次加深的特点，也推动了对生命本质的认知从对生命系统宏观总结、生命元件深入认知、生命系统全面低维度描述到生命系统解析和重构的不断深入。

2 数据驱动生命科学研究的内涵和特点

数据驱动生命科学研究的内涵体现在其对研究范式、方法论和认知模式的深刻影响上。① 强调了以数据为核心的研究方法，将数据的采集和分析置于中心位置。这意味着研究者不再仅依赖于个别案例或局部现象，而是通过收集大规模、多样化的生物学数据来推动研究的发展。② 数据驱动的生命科学研究具有跨学科性和整合性的特点。随着技术的发展和数据的积累，生命科学的研究越来越需要跨越不同学科领域，如生物学、计算机科学、统计学等，进行数据的整合和分析。③ 数据驱动的生命科学研究着重于量化生物现象，并试图将其系统化地理解。传统的生物学研究往往是基于定性观察和描述，而数据驱动的方法则更加注重通过数据收集、处理和分析，建立生物系统的量化模型。这种量化和系统化的方法使得研究者能够更全面地理解生命系统的复杂性，并从中发现隐藏的规律和关联。④ 数据驱动的生命科学研究强调实验数据与数字化建模的结合。通过收集大量的实验数据，并运用数学模型和计算方法进行数字化建模，进行高通量、高准确度地预测和筛选，从而可以高效验证和修正生物学理论，并提出新的假设和预测。这种湿实验与数字化建模结合的研究方式使得生命科学研究更加系统和深入，推动了生物学知识的不断进步。

数据驱动生命科学研究的特征具有3项显著性特点。① 生物学数据具有多样性和丰富性的特点。生物数据涵盖了生物系统的各个层次和多个方面——从基因组序列到蛋白质结构，再到细胞功能和生物表型，生物学数据包含了丰富的信息，为研究者提供了深入探索生命现象的基础。② 生物学数据具有高维度和大规模的特点。随着技术的进步，生物学数据的维度和规模不断增加。例如，基因组学和转录组学等高通量

测序技术的出现，使得研究者能够同时研究成千上万个基因或基因表达物，从而获得高维度的数据。这种高维度和大规模的数据为研究者提供了更全面的视角，使他们能够发现更复杂的生物学规律。③ 生物学数据往往具有动态性和时空特征。生物系统具有在不同时间和空间尺度上的变化。例如，转录组数据可以反映基因在不同发育阶段或不同环境条件下的表达变化，蛋白质互作网络数据可以揭示细胞内信号传导的动态过程。这种动态性和时空特征使得研究者能够更深入地理解生命系统的复杂性，并探索其调控机制和功能。

3 生物大数据组成和特点

大数据（Big Data）通常代表了大量、多样、不断变化且快速聚合属性的巨型数据集，并且这些属性过于复杂或“大”，无法通过传统手段处理[12]。而生物大数据在广义上被定义为来源于或用于生物的海量数据。目前，比较常见的生物大数据类型包括：① 研究类型数据，如基因组、蛋白质组、转录组、糖组等多种组学测序数据，以及成像数据、药物研发和临床试验数据等；② 电子健康数据，如电子医疗档案、可移动/穿戴设备采集的实时监控数据等；③ 生物样本库，如生物多样性资源库、临床样本库等；④ 知识成果，如生物相关的文献、专利、标准等。

生物大数据除了具备“大数据”的特点外，还具有明显的生物学数据自身特性，即大数据量（volume）、多样化（variety）、高速（velocity）和有价值（value）的“4V”特点[13]（图2）。生物学研究技术和手段的快速发展推动了生物大数据的高速发展，使生物学研究从表面的点观测进入全面和更深层次的图像和数据解析。

大数据量。容量是大数据中涉及的数据量的绝对大小。国际癌症组织建立了癌症基因组图谱计划（TCGA），目前已收录的来自各种癌症的组学数据已突破2.5 PB[14]①。自2015年，中国科学院北京基因组研究所（国家生物信息中心）建立了国内首个组学原始数据汇交、存储、管理与共享系统GSA（组学原始数据归档库），目前数据量已突破42 PB[15]。数据库的数据量上升速度之快完美地凸显了生物大数据的蓬勃发展。

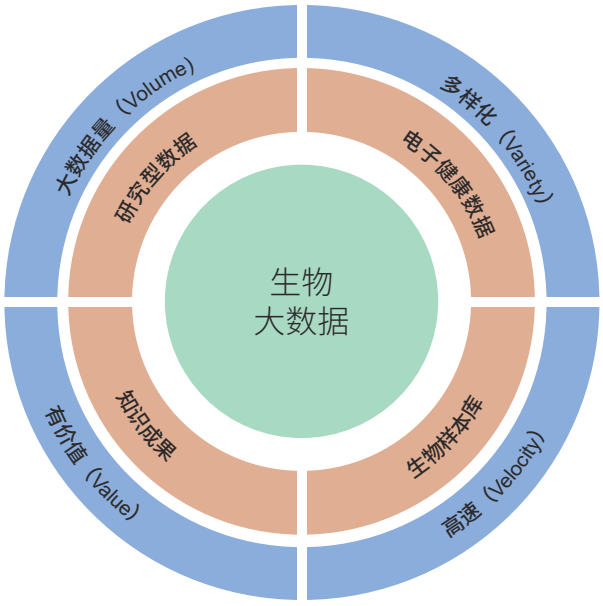

图2　生物大数据的组成和特点

Figure 2　Components and characteristics of biological big data

多样化。多样化代表所收集数据的多样性，组学技术的进步和电子医疗的出现，产生了不同来源、不同格式和不同用途的大量数据，扩大了可用和需要处理的数据类型和数据源的范围[16]。对于生物学样本的研究，经历了从文本数据、图像数据、芯片数据到高通量测序数据的变化，扩充了生物学的研究材料。

高速。速度是由输入和处理数据的速度定义的，指的是数据创建、处理和分析的速度和频率[17]。近年来，为应对生物大数据的急剧增长，人工智能方法被用于生物大数据的解析。

① PB是计算机存储容量单位，1 PB=1024 TB=2^{50} Byte（字节）。

有价值。价值表示所收集的数据在临床研究的结果变化、行为改变和工作流程改进方面的有用性[13]。所有研究性生物大数据的产出，都在特定的方面加深了生物学的认识，推动了生物学研究的发展，体现了生物大数据不可忽视的价值性。例如，临床的影像学数据高效、精准地帮助医生判断患者的病灶和原因，测序数据的解析全面地阐释了表型的根本原因等。

4 技术发展推动生物大数据的产生

生物技术和信息技术的融合推动了生命科学从“假说驱动”向“数据驱动”的转变，促进了生物大数据的爆发式增长、精准解析和生命科学的巨大进步。自从“人类基因组计划”实施以来，测序技术得到了快速发展，引发了基因组、转录组、表观遗传组、蛋白质组、代谢组、糖组等多种组学数据急剧增加，同时也催生了生物技术与信息技术的融合，推动生命科学研究进入数据型科学发现的时代[18]。

在生命科学的发展过程中，得益于测序技术的快速发展，组学类型的生物大数据增长尤为凸显。自1977年Sanger第一代测序技术出现以来，第二代高通量测序技术、第三代单分子全长测序技术和第四代纳米孔测序技术相继涌现，广泛应用于生物学各个领域，推动了生命科学研究的巨大进步。Sanger测序技术被用于细菌和噬菌体基因组的测序，但其1次只能分析1个测序反应[5,19]，产量有限、时间花费长且成本高昂，导致“人类基因组计划”耗时10多年才完成。自2004年以来，“下一代测序”（next-generation sequencing）技术的发展实现了高通量平行测序，大幅增加了测序数据的输出量[20]。第二代测序技术支持基因组、转录组和表观遗传组等多种组学测序，单次测序可以产生4亿条读段、120 GB数据。第三代测序技术又被称为“长读段”测序，可以检测全基因组重复和结构变异检测，实时靶向读取DNA分子[21]。最新的第三代测序仪，平均读长可达10—15 kb，产生约36.5万个读段[22]。第四代测序技术是基于纳米孔系统的DNA测序技术，装置小巧可达手持尺寸，超过100 kb的DNA可以穿过纳米孔，通过许多通道，以相对较低的成本获得数十到数百Gb的序列[23]。测序技术的快速发展对基础研究、临床诊断治疗等具有重要意义。随着精准医疗概念的提出，电子健康记录开始发展。尽管存在不适当访问等潜在风险，但电子健康记录的便携性、准确性和即时性为精准医疗策略、医疗体系完善和智能疗法筛选等提供了重要支持[24]。

在生命科学研究中，信息技术和生物技术的规模化应用丰富了生物样本库的建设。伴随着生物大数据的急剧增长，美国国立生物技术信息中心（NCBI）数据库[25]、欧洲生物信息学研究所（EBI）数据库[26]、日本DNA数据库（DDBJ）[27]和中国国家基因组数据中心[15]等大数据库中的数据类型不断丰富，包括从多组学测序原始数据到表达信息矩阵，数据量从TB向PB甚至更高不断增加，从而为生命科学领域的研究提供了丰富的数据资源。此外，生物大数据的发展也推动了知识成果的积累，促进了生物学数据相关文献不断提升和生物技术专利的快速更新迭代，极大地推动了生物领域的研究，有望给生物学和生物医学研究领域带来革命性的变化。

5 大数据时代下生命科学研究面临的挑战及解决方案

面对生物大数据驱动生命科学研究新范式的发展趋势，研究人员面临着来自不同来源的多维度大数据的挑战。这些大数据包括庞大的结构化和非结构化的信息集合。如何有效地从如此庞大的原始数据中提取信息对于推动科学发明、工业进步和经济发展至关重要。随着新型生物技术的发展，具有多模态、多维度、分布分散、关联隐匿、多层次交汇等特点的生物大数据逐渐形成。如何建立适合生命科学的数据处理和分析流程，构建共享可及且高速传输的数据库，有

效整合数据，为生命科学AI Ready（人工智能就绪）的实现提供完整、安全、真实和契合的高质量数据，将促进新的科学发现并拓展生命科学的探索范围。

5.1 生物大数据处理的挑战

大量的数据在收集整合过程中，因不同实验室和研究人员之间的差异及技术平台差异等因素都可能引起批次效应。批次效应会导致数据变异性增加，真阳性生物信号和假阴性信号的膨胀。当批次效应被误认为感兴趣的结果（假阳性）时，可能会引发更严重的后果[28]。针对批次效应，如今较为公认的方法包括：ComBat包，通过经验贝叶斯估计器来校正数据的批次效应；Seurat包，通过建立锚定的方法将不同批次之间相似的细胞集成单细胞簇[29]。

除了批次效应的存在，数据也可能出现缺失的情况，会导致建模偏差增加或模型准确性降低的问题。针对不同的缺失情况，有着不同的插补解决方案。最简单的插补方法是将信息替换为数据全局特征的值（平均值或中位数等），但是简单的插补会导致标准误差太小，未考虑不确定性[30]。多重插补方法是处理缺失值最常用的方法[31]，即多次对缺失值进行插补，并结合结果以考虑观察到的变异性并减少推断误差。

大量生物学数据的出现，不可避免地会出现批次效应和缺失。针对这些问题优化统一前期数据处理的流程，并开发更加合理的处理批次效应和插补缺失值的方法，以使分析结果更加的可靠，避免出现假阳性的结果。但这些方法只能限制批次效应和减少数据缺失的影响，最终仍需要制定统一的实验和数据标准。

5.2 生物大数据分析的挑战

大数据的出现不仅为深入研究生物系统提供了前所未有的机会，也为数据挖掘和分析提出了新的挑战。大数据分析的首要需求是找到兼顾成本和时间的解决方案。建立有效的生物信息工作流程系统和分析工具对生物数据的分析至关重要。机器学习和深度学习已成为从生物大数据生成处理信息的最先进技术[32]，这些技术在Cloud、Hadoop、apache Spark等大数据平台上执行时，可以有效地从此类生物大数据中提取信息。针对多组学数据异构化的性质，使用具有并行计算的分布式系统的算法适合大数据分析。如MapReduce[33]可以在由数千台计算机组成的大型集群上使用各种并行和分布式算法。

针对生命科学数据的高维度、异质性和复杂性等特征，应着力发展生物大数据的先进分析方法和工具，以加快大数据分析速度、减少分析成本、降低分析的技术壁垒。建立标准的大数据分析流程，以期能够得到准确、可复现和可解释的分析结果。数据驱动的研究新范式的发展对数据分析的方法、工具和算力等资源提出了新的挑战，需要加快建设新一代数据分析基础建设，以做好迎接新范式的准备。

5.3 生物大数据共享可及的挑战

在全国乃至全球范围内，生物数据的共享可及是大数据研究的重要组成部分。① 需要建立数据库用于储存原始或分析结果数据，以实现数据公开和可共享。国际上已经建立了多个用于储存生命科学数据的数据库。例如，NCBI建立的GenBank数据库是世界上最大的基因组数据库之一[34]。另外，蛋白质数据银行（PDB）是一个著名的大分子结构信息数据库，储存了包括蛋白质、核酸等多种生物大分子的信息[35]。我国国家基因库生命大数据平台（CNGBdb）[36]已归档了3 721个研究项目，多组学数据量达6 612 TB，支撑了全球近300个科研单位的科研数据汇交和共享。② 需要高效的程序以使数据能够快速且完整的提供给研究人员。Fasq是一个高效的数据传输软件，它能够在30 s内传输24 GB的数据[37]。然而，它需要大量的互联网连接带宽，数据传输的成本非常昂贵。Smart HDFS（Hadoop分布式文件系统）是一种异步多管道文件传输协议[38]，它使用全局和局部优化技术来选择更高性能的数据节点，从而提升数据传输的性能。

尽管我国已经建立起如国家基因库生命大数据平

台等的大型数据库，但其存储仍存在着规范性不强、存储量不高、数据格式不统一、数据可用性不足和存在大量的使用壁垒等问题。因此，我国生命科学领域需要更好地统筹协调和资源整合，加强科学数据资源的整合与共享，建立规范化的数据存储流程，构建高存储容量、低使用壁垒的数据库，以满足数据驱动下的新范式的需求。面对数据传输的挑战，我国还应该加强数据供给模式的改革，提升数据传输的硬件设施，设计和优化传输程序，以提供更加快速的传输速度为重点，并建立相关协议对数据访问进行管理，进而保护数据的真实性。

5.4 建立大数据+生命科学研究新范式

将生物大数据处理成AI Ready状态对于数据驱动的生命科学研究至关重要。这一过程为人工智能系统的训练和优化提供了基础，并为人工智能系统提供了丰富的信息资源，有助于提高其理解世界的能力，增强预测和决策的准确性，实现个性化服务和定制化产品，同时推动创新和发现。面对生命现象中复杂的非线性关系和难以预测的特征，大数据驱动下的人工智能[7]技术展现出强大的能力，并已在生命科学领域的多个方面展现出颠覆性的应用潜力。例如，Geneformer[39]在基于3 000万个单细胞转录组的大规模语料库进行了预训练，以实现上下文特异性预测；跨物种生命基础大模型GeneCompass[9]在超过1.2亿个单细胞的训练数据集上实现了对基因表达调控规律的全景式学习理解等多个生命科学问题的分析。

然而，在我国在实现AI Ready过程中，核心技术仍相对匮乏，需大力发展自主原创的算法、模型和工具等。针对生命科学的AI Ready过程中大数据的多模态和多维度等特征，急需发展针对性的先进计算与分析方法。未来应开发更加适合生物大数据分析的硬件、软件和新计算介质，并在生命科学和人工智能技术的融合过程中，探索新的人工智能-生物交互模式。充分利用人工智能+生物大数据，同时与湿实验结合，将建立干湿融合的生命科学研究新范式。

6 总结和未来展望

数据驱动的生命科学作为生物科学领域的重要趋势，正面临着海量生物大数据的包括数据存储、传输、处理和分析等多个方面的挑战。然而，通过不断开发新的技术和方法，尤其是人工智能技术的发展，能够更高效地整合和分析生物大数据，从而挖掘生物学内在规律，深入理解生物系统的复杂性。

未来，为实现对复杂生命系统更完美的模拟和解构，需从数据质量、处理算法、场景化等多方面进行优化。① 应生产和获取高质量系统性的生物大数据。当前的生物学数据虽然规模大、类型多，但数据来源各异、离散度高、偏差大，整体数据质量水平不高。而且生命系统是多层级的复杂系统，要将不同层级打通，需要如胚胎发育、疾病、癌症、衰老等生命过程的多维度、多模态、时空对齐的高质量、系统性生物大数据，为人工智能提供可靠的数据基础，减少噪声和偏差的影响。② 需开发生命适配的人工智能算法。生物大数据具有多维度、多层次、非结构化和动态变化的特点，当前人工智能算法难以有效处理。未来需要针对生物数据特点开发生命适配的人工智能算法，来更好捕捉复杂生命网络中的结构和规律。③ 增强模型的解释性，揭示潜在的生物学机制也是未来重要的研究方向。④ 整合生物学数据、利用人工智能技术以及自动化的高通量实验和数据获取技术。有望实现干湿融合的自进化模式，为生命科学研究带来革命性范式革新。

参考文献

1 Kuhn T S. The Structure of Scientific Revolutions. Chicago: University of Chicago Press, 1962.

2 李鑫, 于汉超. 人工智能驱动的生命科学研究新范式. 中国科学院院刊, 2024, 39(1): 50-58.
Li X, Yu H C. A new paradigm of life science research

driven by artificial intelligence. Bulletin of Chinese Academy of Sciences, 2024, 39(1): 50-58. (in Chinese)

3 Vesalius A B. De Humani Corporis Fabrica. Basel: Andreas Oporinus, 1543.

4 Darwin C, Kebler L. On the Origin of Species. London: John Murray, 1859.

5 Watson J D, Crick F H C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature, 1953, 171: 737-738.

6 Maxam A M, Gilbert W. A new method for sequencing DNA. PNAS, 1977, 74(2): 560-564.

7 Lander E S, Linton L M, Birren B, et al. Initial sequencing and analysis of the human genome. Nature, 2001, 409: 860-921.

8 Borkakoti N, Thornton J M. AlphaFold2 protein structure prediction: Implications for drug discovery. Current Opinion in Structural Biology, 2023, 78: 102526.

9 Yang X, Liu G, Feng G, et al. GeneCompass: Deciphering Universal Gene Regulatory Mechanisms with Knowledge-Informed Cross-Species Foundation Model. (2023-09-26). https://www. biorxiv. org/content/10.1101/2023.09.26.559542v1.

10 Burger B, Maffettone P M, Gusev V V, et al. A mobile robotic chemist. Nature, 2020, 583: 237-241.

11 Merchant A, Batzner S, Schoenholz S S, et al. Scaling deep learning for materials discovery. Nature, 2023, 624: 80-85.

12 Panesar A. Machine Learning and AI for Healthcare. Coventry: Apress, 2019.

13 Baro E, Degoul S, Beuscart R, et al. Toward a literature-driven definition of big data in healthcare. BioMed Research International, 2015, 2015: 639021.

14 Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Contemporary Oncology (Pozn), 2015, 19(1A): A68-A77.

15 CNCB-NGDC Members and Partners. Database resources of the national genomics data center, China National Center for Bioinformation in 2022. Nucleic Acids Research, 2022, 50 (D1): D27-D38.

16 Cheng C Y, Soh Z D, Majithia S, et al. Big data in ophthalmology. Asia-Pacific Journal of Ophthalmology, 2020, 9(4): 291-298.

17 Ristevski B, Chen M. Big data analytics in medicine and healthcare. Journal of Integrative Bioinformatics, 2018, 15 (3): 20170030.

18 Shen L, Bai J W, Wang J, et al. The fourth scientific discovery paradigm for precision medicine and healthcare: Challenges ahead. Precision Clinical Medicine, 2021, 4(2): 80-84.

19 Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors. PNAS, 1977, 74(12): 5463-5467.

20 Mardis E R. Next-generation sequencing platforms. Annual Review of Analytical Chemistry, 2013, 6: 287-303.

21 van Dijk E L, Jaszczyszyn Y, Naquin D, et al. The third revolution in sequencing technology. Trends in Genetics: TIG, 2018, 34(9): 666-681.

22 Slatko B E, Gardner A F, Ausubel F M. Overview of next-generation sequencing technologies. Current Protocols in Molecular Biology, 2018, 122(1): e59.

23 Cao M D, Ganesamoorthy D, Elliott A G, et al. Streaming algorithms for identification pathogens and antibiotic resistance potential from real-time MinION™ sequencing. GigaScience, 2016, 5(1): s13742-16-0137-2.

24 Mueller C, Herrmann P, Cichos S, et al. Automated electronic health record to electronic data capture transfer in clinical studies in the German health care system: Feasibility study and gap analysis. Journal of Medical Internet Research, 2023, 25: e47958.

25 Sayers E W, Bolton E E, Brister J R, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 2022, 50(D1): D20-D26.

26 Thakur M, Bateman A, Brooksbank C, et al. EMBL's European Bioinformatics Institute (EMBL-EBI) in 2022. Nucleic Acids Research, 2023, 51(D1): D9-D17.

27 Fukuda A, Kodama Y, Mashima J, et al. DDBJ update: Streamlining submission and access of human data. Nucleic Acids Research, 2021, 49(D1): D71-D75.

28 Leek J T, Scharpf R B, Bravo H C, et al. Tackling the

widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics, 2010, 11(10): 733-739.

29 Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. Cell, 2019, 177(7): 1888-1902.

30 Alvi M A, Wilson R H, Salto-Tellez M. Rare cancers: The greatest inequality in cancer research and oncology treatment. British Journal of Cancer, 2017, 117(9): 1255-1257.

31 Carvalho D M, Richardson P J, Olaciregui N, et al. Repurposing Vandetanib plus Everolimus for the treatment of *ACVR1*-mutant diffuse intrinsic pontine glioma. Cancer Discovery, 2022, 12(2): 416-431.

32 Angermueller C, Pärnamaa T, Parts L, et al. Deep learning for computational biology. Molecular Systems Biology, 2016, 12(7): 878.

33 Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Communications of the ACM, 2008, 51(1): 107-113.

34 Benson D A, Karsch-Mizrachi I, Clark K, et al. GenBank. Nucleic Acids Research, 2012, 40(D1): D48-D53.

35 Berman H M, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Research, 2000, 28(1): 235-242.

36 Chen F Z, You L J, Yang F, et al. CNGBdb: China National GeneBank DataBase. Yi Chuan, 2020, 42(8): 799-809.

37 Marx V. Biology: The big challenges of big data. Nature, 2013, 498: 255-260.

38 Zhang H, Wang L Q, Huang H. SMARTH: Enabling multi-pipeline data transfer in HDFS// 2014 43rd International Conference on Parallel Processing. Minneapolis: IEEE, 2014: 30-39.

39 Theodoris C V, Xiao L, Chopra A, et al. Transfer learning enables predictions in network biology. Nature, 2023, 618: 616-624.

Advances in data-driven life sciences research

JIANG Haiping[1,2] GAO Chunchun[3] LIU Wenhao[1,2] YANG Yungui[3] LI Xin[1,2*]

（1 Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China;

2 Beijing Institute for Stem Cell and Regenerative Medicine, Chinese Academy of Sciences, Beijing 100101, China;

3 China National Center for Bioinformation, Beijing 100101, China）

Abstract The field of life sciences is rapidly evolving, driven by advancements in experimental techniques and vast biological big data which gradually arise and play an increasingly important role in life science research. First of all, biological big data has diversity and complexity, including genomic data, epigenomic data, proteomic data and other types. These data provide researchers with more comprehensive information and help reveal the laws behind life phenomena. Second, new data-driven developments and applications in life sciences cover many fields such as gene editing, precision medicine, drug development, etc., providing unprecedented possibilities for human health and quality of life. However, the era of big data for life science research also faces challenges in various aspects including data storage, sharing, and privacy protection, as well as how to transform massive data into reliable scientific discoveries. This paper provides a brief overview of the law of development of biological data in driving life sciences, sorts out the composition and characteristics of biological big data and its sources, as well as elaborates and discusses the common problems and challenges faced by our country under the new paradigm of data-driven life science research.

Keywords scientific paradigm, big-data, life science

江海平 中国科学院动物研究所博士后。主要研究领域：衰老、癌症和人工智能。E-mail: jianghaiping@ioz.ac.cn

JIANG Haiping Ph. D. Postdoctoral Researcher of Institute of Zoology, Chinese Academy of Sciences (CAS). His research focuses on aging, cancer and artificial intelligence. E-mail: jianghaiping@ioz.ac.cn

李 鑫 中国科学院动物研究所研究员。主要研究领域：干细胞与再生、衰老及癌症，人工智能与生物计算。
E-mail: xinli@ioz.ac.cn

LI Xin Ph. D. Professor of Institute of Zoology, Chinese Academy of Sciences (CAS). His research focuses on stem cells and regeneration, aging, and cancer metastasis, artificial intelligence and computational biology. E-mail: xinli@ioz.ac.cn

■责任编辑：岳凌生

*Corresponding author