

文章编号:1003-2053(2026)01-0101-07

# 数据驱动的生命科学研究范式:形成与挑战

杨宁,李建会

(北京师范大学哲学学院,北京 100875)

**摘要:**当代生命科学研究模式正在经历重大转型。伴随生物技术与信息技术间的持续融合,生命科学的研究信念、方法和文化三方面的演化共同推动着生命科学研究范式的转换;研究信念从还原论转向整体论,并借助网络思维实现二者间的辩证融合;深度学习崛起后,计算模拟和建模方法在数据、算法和算力的强耦合下实现效率与精度的双飞跃;科学组织与管理模式逐渐数据中心化、平台化与生态化,唤醒对开放科学文化的倡导与支持。数据驱动下的生命科学研究新范式在加速科学发现进程、扩展科学发现范畴的同时,也带来诸多生物技术和伦理方面的挑战。

**关键词:**数据驱动;生命科学;研究范式;开放科学

**中图分类号:**B80-0;G301

**文献标识码:**A

2003年,人类基因组计划宣告完成,生命科学研究重点从特定时间点的静态“蓝图”转向随时间、环境变化的动态网络。伴随着研究数据爆炸而日益凸显的,是传统的假设驱动方法在数据处理和系统建模方面遭遇的维度灾难和效率困境。生命科学亟需新的认识方法。2013年,《科学》杂志的一篇社论指出,包含序列、结构和图像的公开数据库激增,使生命科学家无需在实验室工作台上“湿手”就能获得基础性发现<sup>[1]</sup>。大数据开始成为生命科学发现的新基础和新起点。2021年,《自然》杂志发文解析蛋白质三维结构预测系统 AlphaFold2<sup>[2]</sup>,系统阐明基于深度学习模型的计算预测方法何以在历史上首次媲美传统实验方法的准确度。利用大数据和 AI 模型探索生物机制、疾病机理和领域内更多未知空间的科学策略迅速崛起并占据主流。

20年间,生命科学研究模式发生重大变革,对科学的组织、规划、执行和管理方式产生重大影响,甚至重塑了数据、模型和知识等科学发现要素在哲学上的概念化,并由此引发对知识可靠性的担忧。深入探讨数据驱动的生命科学研究范式的形成过程与未来挑战,既是进一步分析新范式的认识论和方法论的重要铺垫,也是为该领域未来健康高速发展和持续高能创新夯基垒台。

## 1 生物大数据与科研模式变革

自分子生物学中心法则确立以来,生物数据的内涵随其获取方式的变化经历了两次关键转型。20世纪七八十年代,基因工程落地,计算机的引入为生物信息学的发展提供了重要工具,推动了早期生物算法和生物数据库的开发,以及自动测序、微阵列(基因芯片)和质谱等生物技术的发展,为基因组学革命奠定坚实的基础。90年代,针对多种模式生物基因组的研究计划相继实施。其中,“人类基因组计划”既是一项重要的基础研究,也是一项技术壮举,标志着生物技术与信息技术之间的全面融合。人类基因组测序完成后的10年里,伴随高通量测序的迭代和测序成本的下降(速度甚至超过摩尔定律),原本围绕中心法则建立的首批组学也逐渐走向更多元化的细胞内分子与通路测绘,衍生出代谢组、糖组、脂质组和微生物组等领域。这一阶段,生物数据的内涵从以往手工实验产生的小规模数据转型为测序产生的大规模组学数据,统称为生物大数据。生命科学由此进入一个生物学、计算科学和工程学相互融合的新纪元,在生命组学尺度上展开一幅全新的“大科学、大工程”研究图景<sup>[3]</sup>。

收稿日期:2024-12-02;修回日期:2025-06-14

基金项目:国家社会科学基金重大项目(22&ZD045)

作者简介:杨宁(1994-),女,博士研究生,E-mail:694558893@qq.com。李建会(1964-),男,教授、博士生导师。

过去10年,多组学研究揭示出更多与基因功能、基因调控机制和蛋白质分子互作有关的生物网络知识,医学和药学领域开始追求前所未有的个性化和理性化<sup>[4]</sup>;医疗理念从基于“统一标准”的循证医学发展为关注个体遗传和环境差异的精准医学,产生大量电子健康记录、医学图像和生活方式档案等;基于人工智能(AI)模型对复杂生物现象的模拟,药物发现与设计的虚拟实验产生大量蛋白质结构数据和药物反应数据。至此,生物大数据的内涵从单一组学类型扩展为生命健康复合大数据体系,覆盖来自基础研究、临床医药、健康管理和环境评估等方面的多模态数据。

与上述生物数据演变密切关联的,是数据在科学实践中日渐要素化和中心化的价值放大过程,数据驱动的研究方法推动生命科学研究范式的重大变革<sup>[5]</sup>。例如,传统的蛋白质结构解析依赖X射线衍射和核磁共振等物理实验方法,须经过耗时耗力的昂贵试错过程来验证一个假设;而在通用数据库和深度学习算法的有力支撑下,AlphaFold2可以在短时间内从蛋白质序列、结构数据径直推理具有原子精度的三维结构图像,并且历史上第一次,单链蛋白质的预测准确性能够与实验法相媲美<sup>[2]</sup>。可见,新范式以大数据和AI模型为核心要素,利用知识表征和“机器猜想”执行高通量、高精度的虚拟实验,加速研究进程。

近期,AI驱动的第五范式被提出和热议<sup>[6-8]</sup>。有学者认为,第五范式基于算法分析建模实现的深度学习自动化和智能寻源工具化,已然与数据管理和统计方法分析所构建的数据范式框架不可通约<sup>[7]</sup>。然而,针对生命科学领域当前范式的阐释,无论是从经验层面勾勒其技术生态<sup>[8]</sup>,还是从哲学视角论证其构成要素<sup>[5]</sup>,尚都建立在与实验、经验和计算范式的对比之上,与数据驱动范式间的真正差异还不够清晰。

无论如何,在生物技术、数据和数据处理、解析技术等科学进展的交汇点上,从对单个分子到整个系统的测量、映射、控制和创造能力的进步,及其对生命本质理解的加深,都切实反映着生命科学研究范式的变革。通过回顾学科发展历程,我们发现,是具体科学实践中的信念演化、方法革新和文化嬗变共同孕育了这场数据驱动下的重要革命。

## 2 数据驱动新范式的形成背景

### 2.1 信念基础:复杂系统

一门学科的发展总是脱离不了所处时期的理性背景,这种认知框架往往影响着共同体的思想和行动。20世纪以重新发现孟德尔遗传定律拉开序幕,数理科学以强大的逻辑、形式化方法和建模工具推动生物学从一门描述性学科转型为假设驱动的实验学科。世纪中叶,DNA双螺旋结构的发现和“中心法则”的提出确立了“基因”作为生命蓝图的决定者,也将机械还原论对生物学研究的统治推至顶峰。彼时,分子生物学家大多是主张“生命是分子机器”的极端因果决定论者<sup>[9]</sup>,坚信对“基因”的准确把握可以帮助消除癌症等生存威胁。因此,对“序列-结构-功能”这一因果决定链条的信仰引发了对生物大分子(以蛋白质和核酸为主)三维构型的狂热研究。

然而,还原论视阙下孤立又静止的机械拆解抹杀了有机体发生发展的种群和环境背景,导致重要的整体性问题——进化、形态和生物行为——久未获得有效解答。有机体自组织和自适应中的表观复杂性和历时动态性无法通过机械自然观得到说明,生物系统在适应和涌现等特性中的进化力与创造力也远远超出还原论的指导范畴,深入理解自然和生命系统迫切需要一种全局的、“向上”的生物观取代纯粹“向下”的还原。20世纪的最后20年,关于“部分如何相互作用产生高阶结构和属性”的研究盛行,通过不同层次的生物信息来阐释和理解整体系统的方法逐步占据主流。当对生物体的探究被重置于与“进化”和“环境”三者合一的非线性世界,还原论对生物学的统治戛然而止。

基于组学数据的累积,世纪之交的生物演化研究开始系统对比不同物种的基因序列和基因表达,并逐步认识到生命是具有层次性的信息系统:基因是执行生命功能的信息,调控分子互作的生物网络则规定了基因的执行方式和效果。换句话说,21世纪的生命观将生命信息的处理过程看作生物网络、模块和系统的内部连接模式,对其内部机制的探究必须首先建立表征分子互作的数学模型。因此,生命科学研究模式开始转向对信息的组织、存取和选择性使用,利用系统分析框架处理多组学数据集并映射生物体的行为和形态,帮助刻画了更清晰、完整

的生物发育和演化图景<sup>[10]</sup>。

在整体论视域下,组学和系统生物学的研究内容开始显现重要交织,例如,转录组学研究利用基于基因表达测量获得的系统结构推断基因调控网络,系统生物学研究利用通路富集分析获得的基因表达模式推断生物系统的属性<sup>[11]</sup>。从方法论上,系统分析手段可以说是物理学、计算机科学、数学和工程学的技术、方法和知识在生命科学领域的荟萃,利用基于实验、计算和模型的生物仿真操纵来把握不确定性、涌现性等复杂系统属性。其方法要点在于对目标系统进行不同层面的干预,然后监测和收集相关通路的反应,整合数据并建立数学模型来预测系统的可能行为。

复杂系统的网络思维为具体的模拟解析提供了有效的途径:利用数字化和统计技术,分子被抽象为节点,分子间的连接被简化为数学属性,诸如大小、密度、聚类、稳健性和中心性,以及前馈和反馈等概念<sup>[12]</sup>。网络思维将生物功能定位在各要素间的连接方式和特征上,强调系统的动力学和组合效应(如级联反应),彻底改变了对分子构效关系和疾病机理的研究模式。以癌症研究为例,传统分子生物学研究模式只能逐个探寻单一位点、基因或通路的功能,当前,通过整合组学、生物网络和细胞信号通路等大数据,动态网络生物标记(DNB)法已经可以揭示癌症形成的多因素机制,更好地指导治疗和预防策略<sup>[13]</sup>。更重要的是,这提示我们,网络表征通过可视化的抽象与压缩实现对复杂现象的系统性说明。通过跨领域的人才和方法,生命科学研究以一种辨证融贯的研究框架调和了长期存在于生物学中的还原论和整体论间的紧张对立。

## 2.2 方法要素:多位一体

在“科学始于数据”的时代,大数据并非天然等同于高价值。显化数据资源的认知价值还有赖于適切、有效的数据分析方法,依托数据、算法和算力的“三位一体”获得新的洞见。

实验生物学家在20世纪里收集、分类和比较实验数据集的工作可被视为数学、计算机科学和生物学融合发展的前奏,为生物大数据的存取和处理技术提供了快速生长的土壤。21世纪以来,生物信息学的进展将这种融合推向高潮:系统论、信息论和控制论为生命科学提供新的视角,研究日益系统化、信息化和工程化;不同生物尺度上的测量技术进步飞快,探究生物过程及其机制的效率和精度实现飞跃;

生物大数据的来源从多组学拓展到与健康相关的全链路,对复杂生物系统的全景式理解成为可能。由此,生命科学研究开始以数据为中心,实践策略是对有据可查、被妥善保存且可靠维护的高质量数据进行去语境化和重新语境化<sup>[14]</sup>,在这样的数据旅迁中采用生物信息学方法解读潜在的有意义模式/关系,然后对等同于假说的该模式/关系进行验证、评价和选择。生物大数据涉及组学序列、时空图谱和显微影像等类型,体外仿真的时空维度和体系尺度往往随数据的非结构化程度攀升。因此,生物大数据的转化应用场景越复杂,大数据模式识别和自动化推理就越需要更先进的生物模拟和建模技术。对多源异构科学数据进行分类、整合并最大化挖掘生物信息的实践诉求最终推动了科学计算方式的革命。

首先是算力革新。分布式计算、云计算和异构计算等高性能计算技术的发展为生物大数据的管理、存取和分类整合提供了解决方案,能够满足对大规模数据的高速、可扩展的需求。此外,数据压缩和索引技术的进步也大大缩小了存储空间的需求,提高了数据的传输和访问效率。计算成本的下降使得在大规模上存储、操作和解读这些数据愈加可行,早期的平台化系统通过联动多个数据中心实现了对跨尺度(从基因到生态)、多模态(从组学到生态学)生物数据的整合,同时提供了分析、建模和可视化等丰富的数据处理场景<sup>[15]</sup>。这类高度整合的生物信息学平台以规范的操作流程在诸多跨学科或地域的资源上协调部署各类研究目标,帮助建立了结构化、可重复和可验证的实验设计与执行。

其次是算法革新。在大数据和大算力的支持下,深度神经网络掀起了基础科学研究中的机器学习算法热潮。AI所带来的高维复杂函数的逼近能力和高维复杂数据的处理能力系统攻克了计算辅助科研的瓶颈。以AlphaFold2为代表的系统处理大规模科学数据的方案在多学科领域实现颠覆性突破,在科学表征、数据建模和模型求解方面的效率和精度得到跨量级的提升。当前,生命科学聚焦先导编辑技术、新型测序技术、空间转录组技术、蛋白质结构预测、人类全基因组泛癌分析和重疾血液生物标志物等主题<sup>[16]</sup>。生物大数据与AI的融合成为相关研究实践的中坚力量,赋能基因调控规律解析、药物设计和生物育种与制造等重大科学问题取得一系列突破,使得基于人机协作的数据驱动型发现模式越来越得到科学界的认可和采纳。

数智技术作为新质生产力也带来科学知识生产关系的变革。在传统科学实践中充当技术媒介的研究工具如今已经发生明显的主体性前移,开始出现“AI科学家”和“自动化实验室”等具有主动性和创造性的新的研究主体形象,替代人类完成数字孪生、反事实推理等高维模拟任务和高危、高强度的实验室劳作。在此基础上,如果能够将机器学习和专家知识理性结合起来,将有望得到一个人机协同的双智能推理与决策框架,这或能应对机器学习算法的不透明性问题,最终突破还原论和经典计算范式的思想枷锁<sup>[8]</sup>,在基础科学研究中发挥超出预计的创新价值。

### 2.3 文化支撑:开放协作

基于技术进展的方法革新为科学革命提供了重要基础,随后,开放共享观念、跨学科综合素养和跨境、跨机构协作等科学文化进展成为数据驱动研究模式高速发展的催化剂。

20 世纪的生物学成果几乎被少数学者和机构控制,其流通依赖付费订阅的纸质出版物。其中,模式生物研究数据尤其与个人专利和名望缠结,仅在一些由学科带头人领导的研究小团体内部施行着缓慢的通讯(newsletter)交流方式<sup>[17]</sup>。这种个人主义严重制约着同行评审和复制检验对研究成果有效性和可信度的评估,直到 20 世纪 90 年代才出现挑战上述病态传播模式的开放存取(Open Access)运动,争取免费或低成本地及时同步研究动态,打破昂贵纸媒筑起的信息壁垒。此后,PubMed Central 档案库开始通过电子形式储存生物学和医学文献,成为领域内鼓励科研成果自由流通和使用的早期代表。

世纪之交,多个跨国科学组织和国际科学计划继续普及关于科学成果的开放存取理念,强调科学知识是公共行为体和公共资金协助下的产物。例如,“人类基因组计划”的百慕大原则要求基因序列数据应在生成后 24 小时内上传到公共数据库<sup>[18]</sup>,推动了全球范围内的基因组数据共享。组学数据爆炸后,生物资源管理走向针对特定类型的开放数据库,诸如蛋白质组学数据库(PRIDE)、转录组学数据库(GEO)和基因表达数据库(GTEx)等,以及多方面、大规模的国家样本库,诸如英国生物样本库(UK Biobank)以及欧洲生物样本库和生物分子资源研究基础设施(BBMRI-ERIC),这些开放资源直接催生了基于组学大数据的转化应用研究,诸如著名的英国“十万人基因组计划”(2012 年)和美国

“精准医学计划”(2015 年)等。

此间,公共数据库的认知角色发生微妙变化。一方面,通过接收尚未或永不发表在学术刊物上的数据,数据库挑战了传统学术期刊的科学权威地位。另一方面,数据库不再是存放研究历史的静态资料室,而是经由数据管护员进行结构化和去语境化处理的流动中转站,辅助生命科学研究走向以大数据为中心、以计算预测为动力的知识生产模式。同时,科研组织形式也在发生变化:新的研究方法和工具体系需要大规模的跨领域个体/机构协作,强调研究者定位、获取和整合数据的能力以及数据共享的系统化和流程化,呼吁一种平台化的科研模式,譬如,通过科学计算工作流实现研究流程的可维护性、可扩展性和成果的可视化和规模化,最终走向更大范围上软件基础设施的协同开发和共享。

随着科学数据与 AI 深入融合,国际组织和政府积极制定推进开放科学的共同标准和指导原则,倡导可及、共享、透明、再利用和与社会互动等科学文化。例如,联合国教科文组织正式提出倡导科学工作方法系统性改革的“开放科学”(Open Science)愿景<sup>[19]</sup>:通过结合各种运动和实践,使多语种科学知识对每个个体开放、可获取和可重复使用,加强科学合作和信息共享,甚至向传统科学界以外的社会参与者开放科学知识的创造、评估和交流过程。其中,开放数据(Open Data)的“FAIR”原则<sup>[20]</sup>——Findable(可发现的)、Accessible(可及的)、Interoperable(可互操作的)和 Reusable(可复用的)——帮助促进了良好的数据管理和信息策划、流通,成为科学研究全球协作的基础。

总之,通过追求开放科学和跨界协作,数据驱动的研究范式得以深度演化。然而,当前开放科学仍在应对来自科学奖励的公平分配、数据开放与隐私保护间的冲突权衡和知识产权管理(如数据技术带来的新型垄断)等方面的挑战,需要更多的政策性扶持来激励科研人员的创新热情和能力。

## 3 数据驱动新范式的价值与挑战

### 3.1 认知价值

数据驱动发现能够有效应对生物计算的复杂性,最大化生物大数据的认知价值,从而建立一套以数据为中心的认识体系:

第一,生物系统的复杂性、演化性和涌现性所带

来的计算难题是数据驱动新范式的主要用武之地。生物系统的动态发展和层级分化等固有属性决定了相关数据的生产和处理必须是历时性的和跨尺度的,巨量的多模态表征引发“维度灾难”。新范式综合了数据科学和计算机工程的特点,汇集算法、统计、建模和仿真技术,以极快的速度处理复杂关系,例如,关联和交叉引用不同来源的数据可以提高科学发现的可靠性和预测精度,有助于确定未来的研究方向,从而帮助实验设计与优化。

第二,面对当代科研实践的跨境、跨团队协作需求,数据驱动新范式的计算与连接潜力展现出高度的供需适配性。生命科学各研究群体的方法、材料和目标不同,数据生产上的背景知识差异导致研究间的概念分歧、技术困难和沟通障碍<sup>[21]</sup>。大数据在科学推理中的可移动性使其在不同语境下能够被重新诠释,且其复用已经超出了生成时的预期。数据的流动性和可塑性有效缩小了上述裂隙,提高了生物数据在经验洞察中的证据价值,满足了多元的认识论需求。

第三,数据驱动新范式拓宽了对科学知识的哲学理解,挑战了传统以理论为中心的科学观,倡导以数据为中心的认知框架。随着生命健康大数据成为知识生产的基础,在此基础上的生物建模、系统仿真和实验干预/操纵等研究手段相应成为一类中间成果。为了确保科学推理和科学解释的有效性,数据及其处理方式需接受严格审查。与假设驱动的实验生物学相比,数据驱动新范式以截然不同的研究初心形成了开放式研究的新认知体系,突破传统决定论的局限,帮助发现超越假设或者现有理论的新知识<sup>[22]</sup>。

### 3.2 未来挑战

#### (1) 技术挑战

生物大数据使用和建模策略依赖特定的问题、约束、风险和机遇,也存在使用多种模型来拟合数据或分类信息的情况。因此,需要充分关注诸如数据质量、数据集标准和模型适配性等可能损害知识发现可靠性的因素。

生命科学研究一直缺乏高质量的系统性生物数据,不同问题域间的可及数据量差异较大,一些前沿领域尤甚。例如,药物靶点信息的累积远超药物作用的下流效应信息(药物引起的生理、生化功能或形态的变化),但后者也可能大幅改变下游基因、蛋白质和代谢物的表达,甚至引发恶性疾病<sup>[23]</sup>。并

且,目前仍缺乏专门的生物算法和工具,在组学序列的组装和注释、生物规律分析方面往往遭遇算法精度、效率和适用范围等限制。

此外,机器学习可解释性问题也引起质疑。除输入数据外,模型的解释力还取决于它响应输入数据的哪些特征以及如何响应,因为扩大样本以提升模型性能的策略并非一劳永逸,还可能出现数据泄漏问题。例如,数据相对充足的蛋白质结构研究通常选用参数化程度高的深度学习模型,在训练和评估中,样本集里拥有进化关联的两个条目可能分别进入了训练集和测试集,导致算法只是记住了相似数据而非学习到变量间的真实关联,模型泛化能力因而被高估,甚至重新增加过拟合风险<sup>[24]</sup>。

在利用算法模型分析高维生物数据时,模型预测与实验结果的一致性并不能充分验证模型的有效性,因为调整部分参数或约束以匹配有限的实验结果并不困难。也就是说,领域知识仍然至关重要,它有助于选择适当的模型结构,确保模型行为正确,并指导开发有生物学意义的特征/预测因子。例如,大语言模型已广泛应用于生物医学领域的临床笔记总结和医疗质量评估,但基于数字指标(如均方差、灵敏度、特异性等)的调整和评估可能导致次优或误导结论。如若缺乏领域知识,分类错误(即假阳性/I型错误和假阴性/II型错误)可能在没有充分考虑临床和经济后果差异的情况下被同等对待,从而引发不良的医疗结果。

#### (2) 伦理挑战

万物互联后,各类数据在科学和商业之间来回转移,人们经历了从积极参与数据分享到奋力争取个人信息不被收集(profiling)的态度转折。生命科学进步正在引发重大的伦理问题,例如,胚胎筛选、选择和编辑可能导致人类特征被人为选择,遗传和健康隐私可能在数据共享和模型训练的过程中被不当采集或非法传播。

首先,需要通过负责任的数据使用、技术部署、算法设计和结果评估来解决研究过程中的安全问题。生物医学领域的伦理要求涉及数据分析的技术精度、准确性和统计性能<sup>[25]</sup>,因为信息偏差和推理谬误势必干扰该领域的应用标准,而与分析目标相契合的算法能够削弱个人信念和意图所引入的偏见,特别是对数据集或算法进行设置或调整的人员。此外,研究者需意识到数据准确性、算法复杂性和模型可解释性的影响,避免轻易将结果泛化至研究范

围之外。这要求科研环路内外的利益相关者理解数据意义和操作影响,增强透明度和信任度。

其次,在公共卫生领域,基于数据挖掘的假设生成正在打破医疗健康研究对公共和私人界限的传统划分,需要更具针对性的伦理准则。医疗健康数据可能来自临床、实验、社会调查和登记等多种活动,使用语境的变化会产生截然不同的效应,数据复用或数据库的组合很可能造成隐私侵犯。例如,当来自医疗服务规划和保险公司精算的数据被复用于涉及残疾、遗传病、吸毒、青少年犯罪等敏感问题分析时,即便执行了知情同意和匿名处理,信息的交叉组合也可能导致对患病风险和弱势/特殊群体的识别、歧视和侮辱<sup>[20]</sup>。对此,还应征得个人对数据出让和信息传播的同意,规范生物样本数据传输和流动,以安全和合乎道德的方式共享和交换信息。

## 4 结 语

人类社会迈入数智时代的同时,科学信念、方法和文化三方面的进展和互作促成了生命科学研究范式的变革,并表现出强烈的数据中心主义:基于比较和计算的理论实验地位跃升,面向知识生产的数据处理成为科学实践的核心,AI成为高分辨率表征、大数据分析和复杂系统建模等关键问题中的利器;科研组织模式从以个体或机构为中心的封闭“小作坊”,转向项目集权的、基于大模型的开放协作大平台;跨界协作成为一大特色,既包括将领域知识融入智能模型的跨学科协作,也涉及学术和非学术机构之间在技术创新和数据累积方面优势互补的产研协作。

伴随机器学习和生物信息学的快速进步,生物大数据有望进一步为生命科学发现提供新的机遇。当前的进展已允许根据个体的遗传倾向、实时的分子测量,甚至通过采集与解码神经信号来量身定制疾病诊疗和健康管理方法。未来,或可依靠“专家知识+机器猜想”策略建构重大生命科学项目的专用算法与模型,为生物模拟提供更问题导向的计算方案和更具解释力的技术架构,加速科学发现的进程。

在更深刻的战略意义上,我国在基础生命科学理论、底层关键共性技术以及生物信息资源积累和开放等方面与国际水平仍有差距,需要积极组建以国家为主导的有组织科研体系,合理设置科研成果

考评和激励制度,同时加强科研人员的信息技术素养。通过资源、设施和人才方面的前瞻性思考和规划性布局,数据驱动的生命科学研究范式有望整体性提升我国生命科学领域的创新策源能力,真正拥抱生物经济时代的到来。

## 参考文献:

- [1] Service R F. Biology's dry future[J]. Science, 2013, 342(6155): 186 – 189.
- [2] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583 – 589.
- [3] 贺福初. 大发现时代的“生命组学”(代序)[J]. 中国科学:生命科学, 2013, 43(1): 1 – 15. He F C. “Life omics” in the era of great discovery (Preface)[J]. Scientia Sinica Vitae, 2013, 43(1): 1 – 15.
- [4] 吴家睿. 21 世纪生物医学的三个主要发展趋势[J]. 生命科学, 2022, 34(11): 1327 – 1335. Wu J R. Three major development trends of 21st – century biomedicine[J]. Chinese Bulletin of Life Sciences, 2022, 34(11): 1327 – 1335.
- [5] 李建会, 李长华. 大数据驱动下的生命科学研究范式变革[J]. 哲学动态, 2023, 45(10): 97 – 105, 127 – 128. Li J H, Li C H. The paradigm shift of big data – driven life? Science research [J]. Philosophical Trends, 2023, 45(10): 97 – 105, 127 – 128.
- [6] 李国杰. 智能化科研(AI4R): 第五科研范式[J]. 中国科学院院刊, 2024, 39(1): 1 – 9. Li G J. AI4R: the fifth scientific research paradigm[J]. Bulletin of Chinese Academy of Sciences, 2024, 39(1): 1 – 9.
- [7] 黄时进. 新一代人工智能驱动科学发现的新范式[J]. 社会科学, 2024(3): 54 – 65. Huang S J. A new paradigm driven by the new generation of artificial intelligence for scientific discovery[J]. Journal of Social Sciences, 2024(3): 54 – 65.
- [8] 李鑫, 于汉超. 人工智能驱动的生命科学研究新范式[J]. 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[J]. Bulletin of Chinese Academy of Sciences, 2024, 39(1): 50 – 58.
- [9] Weinberg R. Point: Hypotheses first[J]. Nature, 2010, 464(7289): 678 – 678.
- [10] Levine A. Seventy years and two paradigm shifts: The changing faces of biology[EB/OL]. <https://www.ias.edu/publications/institute-letter/institute-letter-fall-2019>. 2019 – 09 – 12 / 2024 – 11 – 10.

- [11] Conesa A, Mortazavi A. The common ground of genomics and systems biology [J]. *BMC Systems Biology*, 2014, 8(2): 1–10.
- [12] Stevens H. Networks: Representations and tools in post-genomics [A]. Richardson S S, Stevens H. *Postgenomics: Perspectives on Biology after the Genome* [C]. Durham, NC: Duke University Press, 2015: 103–125.
- [13] Li Y, Chen L. Big biological data: Challenges and opportunities [J]. *Genomics, Proteomics & Bioinformatics*, 2014, 12(5): 187–189.
- [14] Leonelli S. What difference does quantity make? On the epistemology of big data in biology [J]. *Big Data & Society*, 2014, 1(1): 2053951714534395.
- [15] Oinn T, Greenwood M, Addis M, et al. Taverna: lessons in creating a workflow environment for the life sciences [J]. *Concurrency and Computation: Practice and Experience*, 2006, 18(10): 1067–1100.
- [16] 杨帆. 2023 研究前沿 [R]. 北京:中国科学院科技战略咨询研究院, 2023. Yang F. 2023 Research fronts [R]. China: Institute of Science Strategy and Management, Chinese Academy of Sciences, 2023.
- [17] Strasser B J. Data-driven sciences: From wonder cabinets to electronic databases [J]. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 2012, 43(1): 85–87.
- [18] Mifsud J, Gavrilovici C. Big data in healthcare and the life sciences [A]. Myskja B, Helgesson G, Hofmann B, et al. *Ethics and Integrity in Health and Life Sciences Research* [C]. Leeds: Emerald Publishing Limited, 2018: 63–83.
- [19] Unesco N. UNESCO recommendation on open science [EB/OL]. <https://www.unesco.org/en/open-science/about>. 2023–09–21 / 2024–11–10.
- [20] Wilkinson M D, Dumontier M, Aalbersberg I J J, et al. The fair guiding principles for scientific data management and stewardship [J]. *Scientific Data*, 2016, 3(1): 1–9.
- [21] Leonelli S. Scientific research and big data [EB/OL]. <https://plato.stanford.edu/archives/sum2020/entries/science-big-data>. 2020–05–29 / 2024–11–10.
- [22] 吴家睿. 后基因组时代的生命观与科学观 [J]. *科学*, 2021, 73(2): 38–43, 4. Wu J R. Thinking of life and science in the post-genome era [J]. *Science*, 2021, 73(2): 38–43, 4.
- [23] Wishart D S, Feunang Y D, Guo A C, et al. DrugBank 5.0: A major update to the drugbank database for 2018 [J]. *Nucleic Acids Research*, 2018, 46(D1): D1074–D1082.
- [24] Greener J G, Kandathil S M, Moffat L, et al. A guide to machine learning for biologists [J]. *Nature Reviews Molecular Cell Biology*, 2022, 23(1): 40–55.
- [25] Leon-Sanz P. Key points for an ethical evaluation of healthcare big data [J]. *Processes*, 2019, 7(8): 493–504.

## The rise and challenges of the data – driven research paradigm in Life Sciences

YANG Ning, LI Jian – hui

(School of Philosophy, Beijing Normal University, Beijing 100875, China)

**Abstract:** The contemporary life science research paradigm is undergoing a major transformation. Along with the continuous integration between biotechnology and information technology, the evolution of research beliefs, methods, and cultures have jointly promoted the paradigm shift of life science research; research beliefs have shifted from reductionism to holism, and the dialectical integration of the two has been realized with the help of network thinking; the rise of deep learning has led to the realization of double leaps in efficiency and precision of computational simulation and modeling methods under the strong coupling of data, algorithms, and computing power; and the organization and management modes have gradually become data – centric, platform – based, and ecological, awakening the advocacy and support of an open scientific culture. The new paradigm of data – driven life science research has accelerated the process of scientific discovery and expanded the scope of scientific discovery, and at the same time brought about many biotechnological and ethical challenges.

**Key words:** data – driven; Life Sciences, research paradigm; Open Science