DOI: 10.13376/j.cbls/2024003
文章编号：1004-0374(2024)01-0011-10

毛开云，副研究员，中国科学院上海营养与健康研究所生命科学信息中心产业与技术情报部副主任，主要从事生物领域的产业与技术情报研究、专利信息分析和知识产权分析评议工作。2016 年获评全国专利信息实务人才 (国家知识产权局)。先后主持和参与科技部、国家卫健委食品司、国家知识产权局、中国科学院、上海市科委等来源的课题，主编《细胞治疗：技术与产业》等著作。

江洪波，博士，研究员，硕士生导师，现任中国科学院上海营养与健康研究所产业与技术情报部主任。2014 年入选国家知识产权局“全国专利信息领军人才”，2016 年入选“中国科学院特聘研究员”计划特聘骨干人才。主要研究方向为产业与技术情报、竞争情报、科技查新。先后承担科技部、商务部、工信部、生态环境部、国家开发银行、上海市科委、上海市经信委、上海市商务委、上海市知识产权局等委托的决策咨询课题研究工作，以及多家企业委托的产业研究和知识产权战略课题。

2023年计算生物学科技发展态势

毛开云[1]，江　源[1]，袁银池[1]，张　华[2]，周丽萍[3]，江洪波[1,4]*
(1 中国科学院上海生命科学信息中心，中国科学院上海营养与健康研究所，上海 200031；2 上海市生物医药科技发展中心，上海 201203；3 上海生物医药公共技术服务有限公司，上海 201203；4 中国科学院大学，北京 100049)

摘　要：计算生物学借助大量生物数据的模拟与分析，探寻生物体及生态系统的结构和功能，从而加深对生物体的认识与理解。得益于算法的持续优化和计算机性能的提高，计算生物学逐步克服了大量数据处理和分析的难题。2023 年，计算生物学在基因组学、蛋白质结构解析与预测、药物研发、疾病诊断与预测等多个应用领域取得突破性进展。随着技术的进步和数据的积累，计算生物学在未来的发展前景非常广阔，但仍存在数据质量问题、算法和模型复杂度、实验验证的难度、多学科交叉融合的挑战，以及伦理和社会问题等诸多难点需突破。

关键词：计算生物学；深度学习；人工智能；蛋白质结构
中图分类号：Q-03；Q-33　　**文献标志码：**A

收稿日期：2024-01-05；**修回日期：**2024-01-18
基金项目：上海市2023年度“科技创新行动计划”软科学研究项目“美国生物经济行政令背景下上海生物经济发展战略研究”(23692126400)
***通信作者：**E-mail: hbjiang@sinh.ac.cn

The technological development trends of computational biology in 2023

MAO Kai-Yun[1], JIANG Yuan[1], YUAN Yin-Chi[1], ZHANG Hua[2], ZHOU Li-Ping[3], JIANG Hong-Bo[1,4]*

(1 Shanghai Information Center for Life Sciences, Shanghai Institute of Nutrition and Health, Chinese Academy of Science, Shanghai 200031, China; 2 Shanghai Center of Biomedicine Development, Shanghai 201203, China; 3 Shanghai Biopharma Service Co.,Ltd , Shanghai 201203, China; 4 University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Computational biology simulates and analyzes a large amount of biological data to discover the structure and function of organisms and ecosystems, deepening our understanding and comprehension of organisms. Thanks to the continuous optimization of algorithms and the improvement of computer performance, computational biology has gradually overcome the challenges of massive data processing and analysis. In 2023, computational biology has made breakthrough progress in multiple application fields such as genomics, protein structure analysis and prediction, drug development, disease diagnosis and prediction. With the advancement of technology and the accumulation of data, computational biology has a very broad development prospect in the future. However, there are still many difficulties that need to be overcome, such as data quality issues, algorithm and model complexity, difficulty in experimental verification, challenges in interdisciplinary integration, ethical and social issues, and so on.

Key words: computational biology; deep learning; artificial intelligence; protein structure

计算生物学 (Computational Biology) 作为一门源于 20 世纪的交叉学科，紧密伴随着计算机科学和生物学的发展而日益凸显其重要性，其核心使命是通过大量生物数据模拟和分析，发现生物体和生态系统的结构和功能，从而深化对生命现象的理解。借助计算机的强大数据处理能力，计算生物学使得生物系统的探究得以全面且深入，进而更好地揭示生命运行的基本规律。

计算生物学的发展历程并非坦途。初期，计算机技术的限制给数据处理和分析带来极大的挑战。然而，随着算法的不断改进和计算机性能的提升，计算生物学逐渐克服了这些难题，开始在基因组学、蛋白质结构解析与预测、药物研发、疾病诊断与预测等多个领域取得突破性成果。如今，计算生物学已逐渐成为生命科学领域不可或缺的一部分，为生物学家提供了有力的研究工具，助力他们探索生命的奥秘。

1 计算算法和模型赋能生命科学研究

随着计算机技术的飞速发展，人们得以借助生物信息学深入研究生物数据的内在规律，从而为生物学、医学等领域带来前所未有的洞察力。

1.1 利用计算方法与模型深入探究生物体结构和生态系统功能

生物学是一门研究生物体的结构、功能、生长、起源、进化和分布的广阔领域的学科，涵盖了多个分支。生物学的不同分支可以通过实验和计算进行研究，而计算算法和模型有助于对生物体如何工作的理解从亚细胞扩展到整个生物体水平，从而可以深入探究生物体的层次结构和生态系统的复杂功能，预测生物过程的未来发展，甚至参与新药设计和基因编辑等实际应用。

在分子生物学领域，计算算法和模型的应用已经非常广泛。例如，通过建立蛋白质折叠模型，科学家可以预测蛋白质的三维结构，这对于理解蛋白质的功能和设计新药物具有重要意义。此外，基因组学的研究中也大量应用了计算生物学的方法，从基因序列的比对、基因表达模式的分析到全基因组关联研究等，都离不开计算算法和模型的支撑。

在生态学领域，计算算法和模型的应用也日益增多。例如，通过建立生态模型，可以模拟生态系统的动态变化，预测物种的分布和数量变化，为环境保护和生态修复提供科学依据。此外，借助大数据和机器学习技术，还可以对生态系统中各种复杂的关系进行深入挖掘和分析，为生态系统管理和生态工程提供有力支持。

总之，计算算法和模型已经成为生物学研究的重要工具和方法。随着技术的不断进步和应用领域的不断拓展，计算生物学的发展前景将更加广阔。

1.2 计算生物学为生命科学带来“干湿结合的数据闭环”新模式

在海量数据的开发利用上，生物学传统实验和

分析手段已显得力不从心。计算生物学作为一跨学科的综合性方法，涵盖多个专业领域，已成为解决相关问题的重要手段，为科研和实践提供了坚实保障。

在科研领域，计算生物学展现出对传统实验的替代甚至超越能力。相较于操作水平、实验器具、观察水平等精度有限的传统生物实验，基于计算机的计算生物学成本更低、速度更快，且在理论上拥有无限的计算精度和高度可复制性。通过将过往经验内化于人工智能 (artificial intelligence, AI) 模型中，计算生物学能够自动化、规模化和并行化地提出假设，让科研人员摆脱对少数天才的依赖，降低下游开发的门槛，从而对整个行业格局产生重大影响。

计算生物学更为生命科学提供了全新的研究思路 —— 干湿结合的数据闭环模式，以此开启了“假设 - 验证 - 优化假设”的新路径，使研发效率得到显著提升。新的生物学研究方式应以理论推测为出发点，再回归实验中验证理论假设。计算生物学正是在这一理念的指引下蓬勃发展，通过干湿循环实验，不仅提高了 AI 预测模型的精度，还为湿实验提供了高参考价值的假设，实现了两者之间的良性迭代加速。

2 计算生物学研究进展

从生物序列分析到生物网络分析，计算生物学中的算法和模型现在得到了更广泛的应用，可以使研究者更加全面地理解和研究生物学问题。在组学研究方面，生物计算模型可以帮助研究基因组和蛋白质的变异规律、复杂性和功能。在药物研究方面，借助生物计算模型，研究人员可以更加精准预测药物的作用机制、缩短新药研发周期。在临床研究方面，生物计算模型为疾病预测、诊断辅助、治疗优化和数据保护等提供了技术支持。

2.1 计算生物学推动组学研究取得突破

随着高通量组学平台的发展，生物医学研究大多采取了多组学技术结合的方法，不同组学来源 (如遗传学、蛋白质组学和代谢组学) 的数据可以通过基于机器学习 (machine learning, ML) 的预测算法进行整合，以揭示系统生物学的复杂工作。2023 年，计算生物学在基因组学、蛋白质组学、转录组学等研究上取得重要进展。

基因组和蛋白质组映射的主要目的是理解基因表达调控机制、解析生物过程、预测蛋白质功能以及寻找潜在的药物靶点。通过研究基因组和蛋白质组映射，科学家们可以深入了解生物体的生物学功能，为疾病诊断、治疗和预防提供重要线索。基因组和蛋白质组映射的研究涉及到多个层面，包括基因序列与蛋白质序列的比对、蛋白质结构预测、基因表达调控分析等。蛋白质与基因组比对对于注释非模式生物中的基因至关重要。2023 年 1 月，哈佛医学院和丹娜 · 法伯癌症研究所的研究人员研发了一种名为 miniprot 的新型对齐器，该工具能将蛋白质序列映射至完整基因组。miniprot 融合了 k-mer sketch 算法与基于单指令多数据流 (single instruction multiple data, SIMD) 的动态编程技术，其速度较现有工具提升了数十倍。在真实数据测试中，miniprot 展现了令人满意的准确度。2023 年 2 月，美国佛罗里达大学的研究人员提出一种预测蛋白质 -DNA 复合物结构的新型计算方法 MELD-DNA。该方法通过贝叶斯推理将分子动力学模拟与一般知识或实验信息相结合。MELD-DNA 可具备对多种结合模式进行采样的能力，能够从候选序列中筛选出最具优势的结合模式，揭示 DNA 序列间的定性结合偏好，并在预测蛋白质 -DNA 复合物方面表现出较高的准确性 [2]。

基因突变是生物体遗传信息发生变化的过程，这种变化可能导致生物体的表型、生理功能，乃至疾病风险发生改变。随着高通量测序技术的发展，越来越多的基因突变数据被揭示，如何从这些庞大的数据中挖掘有价值的信息成为了生物信息分析的重要任务。目前计算算法和机器学习方法已经广泛应用于基因突变的研究中。2023 年 3 月，复旦大学与耶鲁大学、麻省理工学院和哈佛大学布罗德研究所等合作构建了基于 EN-TEx 资源的多组织个人表观基因组与遗传变异影响模型，EN-Tex 能为研究人员提供丰富的数据和模型来帮助进行更为准确的个体化基因组学研究 [3]。2023 年 4 月，美国格拉斯通研究所和加州大学的研究人员发现人类加速进化区发生的结构变异导致人类基因组折叠方式不同于其他灵长类动物，研究团队使用机器学习模型来预测 DNA 折叠模式，并将其应用于人类和黑猩猩的 DNA 序列分析，从而确定了在人类中以不同方式折叠的基因组区域 [4]。2023 年 7 月，德克萨斯大学与哥伦比亚大学的研究人员利用人工智能将全身 X 射线图像数据和来自 3 万多名英国生物银行参与者的相关基因组数据相结合，阐明了人类骨骼比例的遗传基础 [5]。

蛋白质组学研究方面，计算生物学助力研究者

对蛋白质结构、功能和相互作用进行深入研究。通过机器学习算法，研究者可以快速地预测蛋白质结构，揭示蛋白质功能及调控机制。2023 年 6 月，我国华大智造研究团队推出了一款名为 EvoPlay 的算法模型，该模型将传统的强化学习应用于蛋白质设计领域，不仅提升了传统机器学习指导的定向进化 (machine-learning-guided directed evolution, MLDE) 的采样效率，还能结合最新的蛋白质结构解析模型 (alphaFold2) 直接设计出具有目标结构的氨基酸序列。EvoPlay 算法既适用于传统定向进化，也可融入“从头设计”的理性设计框架 [6]。2023 年 10 月，瑞士联邦理工学院的研究人员通过定量蛋白质组学测量的亚单位比率，构建了一种计算算法，用以检测蛋白质复合物的变化。他们将此算法应用于癌症细胞系及患者活检的检测，成功地在更具侵略性的癌症中发现了组蛋白去乙酰化酶 2 抗体 (HDAC2) 表观遗传复合物的显著重构。所提出的算法可作为“R 包”使用，并通过从自下而上的蛋白质组数据集中提取功能相关信息来推断蛋白质复合物状态的变化 [7]。

迁移学习通过运用在庞大通用数据集上预训练的深度学习模型，并对有限特定任务数据进行微调，为自然语言理解和计算机视觉等领域带来革新。2023 年 5 月，丹娜 · 法伯癌症研究所与哈佛大学联手开发了一种上下文感知的、基于注意力的深度学习模型 Geneformer，在涵盖约 3 000 万个单细胞转录组的大规模语料库中进行预训练，以实现在网络生物学有限数据环境下的上下文特定预测。Geneformer 对网络动力学有了基本认识，以完全自监督的方式在模型注意力权重中编码网络层次。通过对 Geneformer 进行微调，可实现广泛的下游应用，从而加速发现关键的网络调节器和候选治疗目标 [8]。

2.2 新的算法和模型加快药物设计与开发

药物研发进程的优化缩短了从候选药物至上市产品的周期，药物开发过程需要深入地研究、分析和临床试验，同时满足监管合规性要求，这导致开发过程漫长而复杂，成本不确定性高。计算生物学在药物设计中发挥重要作用，通过计算机辅助药物筛选和优化，可以降低药物研发的时间和成本。2023 年，计算生物学在药物设计与开发方面取得重要进展。

相互作用是药物研发过程中至关重要的一环，可以帮助研究人员确定哪些药物可以与特定的生物靶点相互作用并发挥预期疗效。随着计算生物学的发展，越来越多的计算方法被应用于预测药物与靶点的相互作用。2023 年 2 月，阿斯利康和英国谢菲尔德大学合作开发一种基于双线性注意网络的模型 (DrugBAN)，旨在预测药物与靶点相互作用。尽管 DrugBAN 在解释预测配体方面表现出色，但其对于蛋白质序列的可解释性预测相对较弱。为此，作者提出将 3D 蛋白质信息整合到建模框架中，以提高药物 - 靶点相互作用预测模型的可解释性 [9]。2023 年 8 月，腾讯 AI Lab 推出了一种名为 DeepAIR 的深度学习框架，该框架整合了适应性免疫耐受 (adaptive immune resistance, AIR) 的序列和三维结构特征，旨在实现集成序列和结构信息的 AIR 抗原结合分析，从而预测 AIR 与抗原之间的结合。研究结果显示，DeepAIR 在预测 AIR 抗原结合反应性方面表现出卓越的性能，并且优于 SOTA 预测器 [10]。

计算生物学在药物设计中的应用日益广泛，主要应用包括分子模拟与建模、定量结构 - 活性关系 (QSAR) 研究、计算机辅助药物设计等方面，计算机辅助药物设计是计算生物学在药物研发领域的核心应用之一。通过药效团搜索、分子对接、分子动力学模拟等技术，研究人员可以快速找到与靶点相互作用的候选药物分子。2023 年 5 月，瑞士洛桑联邦理工学院研究团队成功研发了一种基于蛋白质表面特征指纹图谱的机器学习方法，该方法能够从零开始设计新型蛋白质。这些人工设计的蛋白质在与癌症免疫治疗靶标 (如 PD-1、PD-L1、CTLA-4) 或新冠病毒靶标 (S 蛋白) 结合亲和力方面达到了与天然产生的抗体相当的水平。以蛋白质表面为核心的设计方法捕捉到了蛋白质分子间识别的物理和化学决定因素，为蛋白质 - 蛋白质相互作用 (PPI) 的从头设计提供了全新途径。这一方法可轻易拓展至多种具有重要疾病治疗价值的蛋白质靶点，并可直接生成蛋白质结合物 [11]。2023 年 5 月，百度美国研究院与斯微 (上海) 生物科技股份有限公司、俄勒冈州立大学和罗切斯特大学共同研发出一种名为 LinearDesign 的 mRNA 序列优化算法。该算法运用自然语言处理中网格解析 (lattice parsing) 技术，对 mRNA 疫苗序列进行优化，从而提升疫苗的稳定性和有效性。值得关注的是，LinearDesign 能够在短短 11 min 内锁定最稳定的新冠 mRNA 疫苗序列，从而极大地加快原本缓慢且成本高昂的疫苗设计流程 [12]。2023 年 8 月，华为与复旦大学联合提出了分子优化模型 Q-Drug (药物的量子启发优化算法)，该框架利用量子启发算法来优化离散二元域变量上

的分子，将药物设计带入量子空间，为基于量子计算概念的更好的分子设计技术提供了新的可能性[13]。2023 年 10 月，复旦大学马剑鹏团队成功研发了一种名为 OPUS-DSD 的新型智能计算方法[14]。该方法在 cryoDRGN1.0 的基础上进行了整合与优化，包括采用 3D 卷积架构和隐空间的先验知识，以提高隐空间的平滑性。OPUS-DSD 算法不仅能够成功解析冷冻电子显微镜 (Cryo-EM) 结构解析技术中传统方法无法分辨的生物大分子 (如蛋白质、核酸或蛋白质 / 核酸复合物等) 结构，还能高效精确地分辨出生物大分子柔性结构域在受测样品中的构象分布。这一方法有助于构建高精度生物大分子结构模型，从而解决药物设计中因目标蛋白结构不准确而导致新药研发失败的问题。

2.3 AI技术提升疾病诊断与预测准确率

高度灵活且可重复使用的人工智能 (AI) 模型的快速发展，有望给医学领域带来全新变革。由于难以获取大型、多样化的医疗数据集，以及医疗领域具有复杂性的特征，快速发展的根基模型并未广泛渗透到医疗人工智能行业之中。当前的医疗人工智能模型仍大多采用特定任务的方法，其所训练出的模型不太灵活，仅限于执行由训练数据集及其标签预定义的任务。以 ChatGPT 为代表的大模型的日益成熟，让医疗 AI 的研究者看到了打造通才型 (全能型) 医疗 AI 的希望。2023 年 4 月来自斯坦福大学、哈佛大学、多伦多大学和耶鲁大学医学院的研究团队提出一种通用医疗人工智能 (generalist medical AI, GMAI) 基础模型[15]，它可以使用少量数据或没有指定标记的数据来执行不同的任务。通过在大型、多样化的数据集上进行自我监督，GMAI 将灵活地解释不同的医疗模式组合，包括来自影像、电子健康记录、实验结果、基因组学、图表或医疗文本等多种形式的数据。关于通用医疗 AI 临床应用潜力和局限的研究表明，通用医疗 AI 模型前景整体向好，“生成放射学报告”“手术过程特征增强提取”“辅助临床决策”“生成文档”“聊天机器人”“生成蛋白质”等 6 大具体医疗场景有望早日落地，但依然面临着诸多挑战，有待进一步解决和完善。

医疗人工智能在推动医疗保健领域方面具有巨大潜力，如支持循证医学实践、实现个性化患者治疗、降低成本以提升医疗保健提供者和患者的体验。为充分发挥这一潜力，有必要对医疗人工智能模型在处理大规模、异构患者数据上的性能进行系统化和定量的评估。为此，IHU Strasbourg、丹娜 · 法伯癌症研究所、Intel 等多个研究机构联合组建的研究团队推出了一款名为 MedPerf 的开放平台，旨在在医疗领域为 AI 模型提供基准测试。MedPerf 专注于通过安全地将 AI 模型分发至各机构 (如医疗机构) 以实现 AI 模型联合评估的平台。通过引入模型的数据过程，各设施得以在高效且受人工监督的条件下评估和验证人工智能模型的性能，同时充分保障隐私[16]。

此外，2023 年计算生物学在疾病诊断和预测方面取得多项重要进展。2 月，加拿大麦吉尔大学使用高度多重成像质谱细胞术 (imaging mass cytometry, IMC) 揭示了与肺腺癌 (lung adenocarcinoma, LUAD) 患者临床结果相关的肿瘤免疫微环境 (tumour immune microenvironment, TIME) 的空间分辨特征，使用深度神经网络模型探索了空间分辨率特征与临床结果的关系，证明了可以使用基于人工智能的系统从原始 IMC 图像中提取的特征来预测各种临床结果[17]。2023 年 3 月，中山大学的研究人员通过最新的算法、先进的计算能力和大数据建立的基于人工智能的诊断模型将能够自动检测膀胱癌患者的淋巴结转移，特别是微转移。同时，建立的膀胱癌淋巴结转移诊断模型可作为淋巴结转移检测的可靠辅助诊断工具[18]。6 月，美国加利福尼亚大学圣迭戈分校团队研发了一种新型“贴片式”心脏超声成像仪，可对心脏射血功能相关的重要参数进行实时监测，该设备融入了深度学习模型，可以从连续图像记录中自动提取左心室容积，产生关键心脏性能指标的波形，如搏出量、心输出量和射血分数，在各种环境中显著提高对心脏性能进行动态监测的准确性[19]。9 月，香港大学 Ed X.Wu 团队通过双采集深度学习 3D 超分辨率技术，突破了低成本超低场 (ULF) 磁共振成像 (MRI) 技术的极限，提高其图像分辨率[20]。9 月，英国伦敦大学学院和莫菲尔德眼科医院的研究团队提出了一个视网膜图像基础模型 RETFound[21]，利用自监督学习在超过 160 万张未标注的视网膜图像上训练而成，在眼部疾病诊断 / 预后及系统性疾病的预测等任务中，都具有极佳的性能。10 月，斯坦福大学 Vinit Mahajan 团队等将微量液体活检蛋白质组学、单细胞转录组学和人工智能相结合，生成了一个“蛋白质组学时钟”，从而根据眼睛的蛋白质谱来预测人的年龄和疾病[22]。12 月，斯坦福大学 Tony Wyss-Coray 团队开发了一个基于机器学习的人工智能算法 LASSO，其可以更好地预测衰老相关的疾

病和死亡风险[23]。

3 计算生物学市场发展

全球计算生物学市场根据应用分类可细分为细胞和生物模拟、药物发现和疾病建模、临床试验等部分。细胞和生物模拟部分进一步细分为计算基因组学、计算蛋白质组学、药物基因组学等。

3.1 市场规模增长

目前计算生物学的价值主要集中在科研领域，如提升生物实验效率及精度、补充实验依据，应用在计算推演生物性质及原理、搭建预测及判断模型、对生物体进行控制改造上[24]。在国外，各初创公司的业务已广泛覆盖细胞和生物模拟等各类场景，特定建模的计算生物学软件实现了对外商用。Mordor Intelligence 发布的市场研究报告数据显示[25]，全球计算生物学市场正呈现出强劲的增长势头，规模预计将从 2023 年的 68 亿美元增长到 2028 年的 127.2 亿美元，预测期间的复合年增长率为 13.33%。其中，

表1 2023年计算生物学领域主要产品的进展详情

公司名称	产品进展
英矽智能(Insilico Medicine)	已开发药物发现平台Pharma.AI，包括靶点识别PandaOmics、化合物设计Chemistry42和临床试验结果预测InClinico三个人工智能引擎。2月8日，美国食品药品管理局(FDA)授予Pharma.AI平台发现的INS018_055孤儿药认定，用于特发性肺纤维化(IPF)的治疗；3月13日，将ChatGPT应用到英矽智能的靶点发现平台PandaOmics中。
百图生科(北京)智能技术有限公司	3月23日，发布生命科学大模型驱动的AIGP(AI Generated Protein)平台，主要有三大类功能：“F2P—根据形状、结构、功能、理化性质等需求生成蛋白质”；“P2P—根据给定蛋白质，生成对应的互作蛋白质”；“C2P—根据给定细胞的组学数据，发现调控细胞功能的靶点蛋白质，并设计相应的调控蛋白质”。7月6日，百图生科与清华大学合作开发出蛋白质语言模型xTrimoPGLM，模型参数量高达千亿级(100 B)，融合了蛋白质理解、蛋白质生成两类任务的预训练方法。
杭州碳硅智慧科技发展有限公司	3月24日，发布新药研发平台DrugFlow1.0，涵盖靶标发现、苗头化合物发现和先导化合物优化等环节，集成了靶标发现、活性预测、成药性预测、分子生成优化、虚拟筛选、AI建模等模块。
谷歌云(Google Cloud)	4月13日，谷歌云推出医学大模型Med-PaLM 2，该模型是第一个在MedQA (美国医学执照考试)测试集中达到“专家”水平的AI大模型。5月16日，推出Target and Lead Identification Suite工具，有助于预测和理解蛋白质的结构，还推出了Multiomics Suite工具，以便于摄取、存储、分析和共享基因组数据。
北京水木分子生物科技有限公司	8月18日，联合清华大学智能产业研究院(AIR)开源了可商用的多模态生物医药百亿参数大模型BioMedGPT-10B，旨在提升药物设计和优化、临床试验设计等药物研发过程的效率，在自然语言、分子和蛋白质的跨模态问答任务上实现SOTA (state-of-the-art)的精度；还共同开源了面向生物医药领域的免费可商用Llama 2大语言模型的BioMedGPT-LM-7B；此外，AIR-智源健康计算联合研究中心开源了可商用的小分子药物基础模型DrugFM，旨在为生物医药领域提供大模型底座。
腾讯	9月8日，药物发现平台云深智药iDrug开发的算法框架tFold，优化了蛋白质结构预测。
微软	9月12日，提出了基于离散扩散模型的蛋白质序列生成模型EvoDiff，包含6.4亿个参数。该模型通过融合进化规模数据与扩散模型的调控能力，在序列空间中生成可调节的蛋白质。
百度	10月9日，百度飞桨螺旋桨联合百图生科研发了文心生物计算大模型HelixFold-Single，专注于蛋白质结构预测。该模型开源并提供在线服务，突破了依赖多序列比对(MSA)检索的主流模型如AlphaFold2的速度限制。
Google DeepMind	10月31日，DeepMind及其衍生公司Isomorphic Labs联合推出了新一代AlphaFold模型。该模型的预测对象不再局限于蛋白质折叠，能够针对小分子配体、蛋白质、核酸及翻译后修饰的生物分子等，提供原子级精度的结构预测，还能预测配体与蛋白质之间的相互作用。Isomorphic Labs公司正将此模型用于药物设计。
北京深势科技有限公司	11月21日，成功研发了Uni-Mol Docking V2，通过固定配体构象中的刚性参数来调控偏折自由度，增强了对配体结构中键长、键角及手性关系预测的质量，从而实现对蛋白质-配体对接的高精度预测。

北美地区是最大的市场，美国是行业内领先国家，政府每年用于发展计算生物学的平均支出估计为1.4 亿美元，由于美国药物发现和开发工作的高支出，预计将保持主导地位；欧洲地区是全球第二大市场，德国、英国和法国占据重要比例；亚太地区是增长最快的市场，中国和日本是药物支出的主要国家，该地区合同研究组织 (CRO) 的持续增长，有助于降低药物研发成本，从而推动市场规模扩大。同时，较大的化学仿制药和生物类似药生产规模，进一步刺激了计算生物学领域的市场需求。

3.2　产品开发取得突破

目前，计算生物学领域涌现了许多令人瞩目的产品和技术。在出现大量优势自研算法后，软件平台所占比重将有明显上升。国外已开始通过打包订阅、按照使用量计费等方式对外商用其计算生物学服务。而国内生物计算的应用主要集中在药物发现场景，包括虚拟筛选、蛋白质结构预测、基于靶点的化合物性质预测、分子生成等领域，大多选择内部应用，大多数已开源平台尚未达到可收费水平，能够提供特定建模的计算生物学软件将成为短期内商业化的重要发力点。表 1 呈现了 2023 年计算生物学领域产品取得的新进展。

3.3　领域投融资较热

计算生物学属于工具性质的学科，市面上尚不存在严格意义上的计算生物学公司，而多以 AI 制药、组学、精准医疗等名义出现。在 IT桔子数据库，行业限定为“医疗健康 - 生物技术和制药、医疗信息化”，标签限定为“智能医疗、AI 医疗、智慧医疗、人工智能、AI+ 药物研发、机器学习或深度学习”等计算生物学相关概念，截至 2023 年 12 月 25 日，检索得到 1 030 条投资事件。

由图 1 可以发现，计算生物学领域的投融资金额和数量在 2017 年及之后的 5 年里快速增长，由于 2019 年新冠疫情的出现，2021 年计算生物学公司的投资额激增，几乎是 2019 年和 2020 年募集资金总和，近两年投资热度逐渐降低，融资事件和金额均有所下降。国外的投融资额平均值更高，国内占到了总投资事件的 70.58%，而国内外的投资金额相当。计算生物学领域的公司融资轮次均偏向早期，B 轮之前 (不含 B 轮) 的交易数量就占到了总投资事件的 64.56%。

国内的投资事件主要分布在北京、上海、杭州和深圳等城市，国外主要分布在美国、加拿大、英国、以色列和印度等地区。2023 年，全球计算生物学领域发生 82 起投资事件，融资金额约为 93 亿人民币，深势科技、Causaly、腾迈医药、Superluminal Medicines、分子之心和本导基因等企业受到了资本市场的青睐 (表 2)。此外，10 月 10 日，赛诺菲与百图生科达成战略合作共同开发用于生物治疗药物发现的领先模型，此前赛诺菲与多家 AI 制药等公司展开了全方位合作，收购方或合作方如 Amunix 和 Owkin、Exscientia、英矽智能、Atomwise。

4　未来展望

计算生物学可以通过高效精准的计算推演，为上层应用提供支持，如基于蛋白质功能及相互作用预测、化合物性质预测、基因点位预测等，从而加速 AI 制药、疾病研究、物种改造等领域的发展。

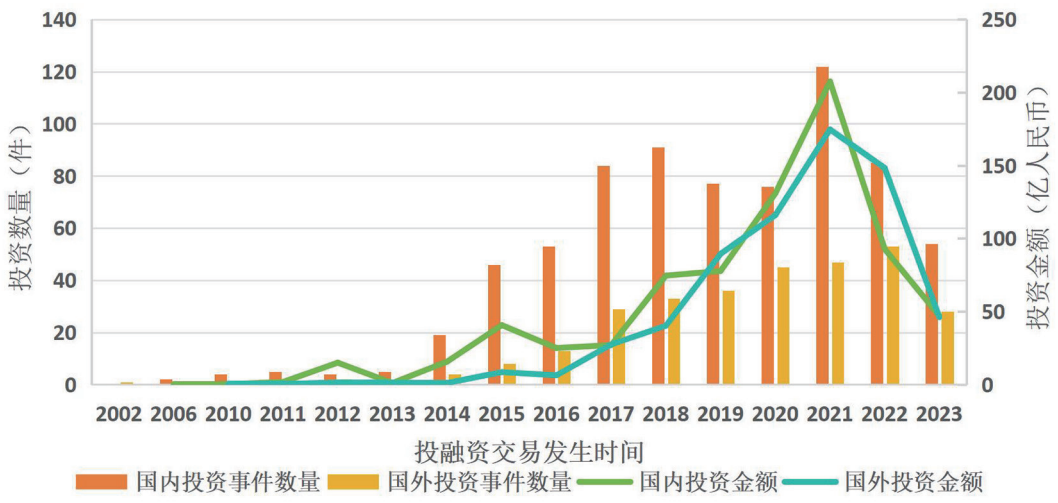

图1 计算生物学应用的投资情况

表2 2023年计算生物学领域重要融资事件

融资金额(RMB)	公司名称	所属地区	所处赛道	融资时间	轮次	投资方
7亿	北京深势科技有限公司	北京	药物模拟研发平台	2023-08-18	C轮	和玉资本、正心谷资本、众源资本、Evergreen Scitech Delta
约4亿	Causaly Ltd.	英国	生物医学研究人工智能平台	2023-07-19	B轮	Index Ventures、Pentech Ventures、ICONIQ Growth、Marathon Venture Capital、EBRD、Visionaries Club
超2亿	上海腾迈医药科技有限公司	上海	药物分子发现平台	2023-03-02	A轮	启明创投、斯道资本(富达亚洲)、奥博资本、F-Prime Capital Partners
超2亿	Superluminal Medicines Inc.	美国	AI生成药物服务商	2023-09-05	种子轮	Insight Partners、NVIDIA英伟达、RA Capital Management、Gaingels
数亿	北京分子之心科技有限公司	北京	AI蛋白质设计平台	2023-02-20	战略投资	联想创投、凯赛生物
2亿	上海本导基因技术有限公司	上海	基因治疗创新药物研发	2023-06-27	B轮	山蓝资本、通德资本、龙磐投资、鹏来资本、春和资本
约2亿	Engine Biosciences Pte Ltd.	美国	人工智能新药研发平台	2023-11-03	A轮	Polaris Partners、EDBI、Invus、Seeds Capital、ClavystBio、Coronet Ventures
1亿	上海智峪生物科技有限公司	上海	AI合成生物技术开发应用服务商	2023-05-25	A轮	惠每资本、清池资本、宏沣投资、钱塘创投、杭州和达投资
1亿	杭州璞睿生命科技有限公司	浙江杭州	生命大数据服务商	2023-12-06	Pre-A轮	启明创投
超六千万	Aureka Biotechnologies Inc.	美国	人工智能生物医药开发商	2023-06-25	种子轮	险峰旗云、纽尔利资本
超五千万	Novorex Inc.	韩国	小分子药物研发商	2023-11-30	A轮	KB Investment、BNH Investment、UTC Investment、Woori Venture Partners、Company K Partners、Aon Investment、Quad Asset Management、Hana Ventures、Technology Assurance Fund
未透露	深圳阿尔法分子科技有限责任公司	广东深圳	AI创新药物研发商	2023-08-24	A轮	新恒利达资本、上海生物医药基金、华仟资本
千万级	北京水木分子生物科技有限公司	北京	生物医药基础大模型研发商	2023-08-23	种子轮	汇芯投资

计算生物学在未来的发展仍面临诸多技术难点和挑战，包括数据质量问题、算法和模型复杂度、实验验证的难度、多学科交叉融合的挑战、数据隐私、信息安全、公平性和透明度等伦理和社会问题，需要不断加强跨学科合作、推进技术创新、完善伦理规范等方面的工作，以推动计算生物学的健康发展。计算生物学的未来研究将在以下几个方面开展。

一是算法优化和创新。为了更好地应对生物信息学中的复杂计算问题，算法研究人员需要不断优化现有算法，提高计算效率。同时，创新算法的提出也将为计算生物学领域带来新的突破，这包括基于图论、机器学习、量子计算等的新型算法。

二是大数据和高性能计算技术的融合。随着基因测序、代谢组学等生物实验技术的发展，生物数据量呈现出爆炸式增长。处理这些大规模生物数据，需要借助高性能计算技术，如 GPU、云计算和集群计算等。未来，计算生物学将与大数据技术紧密结合，实现生物数据的快速处理和分析。

三是跨学科合作。计算生物学涉及到生物学、计算机科学、数学、物理学等多个学科，跨学科合作对于解决生物信息学问题具有重要意义。通过多学科知识的整合和交流，有望突破现有的技术瓶颈，推动计算生物学的发展。

四是人工智能和计算生物学相结合。近年来，人工智能技术取得了显著进展，如深度学习、自然语言处理等。将这些技术应用于生物信息学领域，有望提高生物数据挖掘的准确性和效率。例如，通过人工智能技术进行蛋白质结构预测、药物筛选等任务，将大大缩短研究周期，降低研究成本。

五是注重计算生物学在医学和农业等领域的应用。计算生物学在基因组学、蛋白质组学等领域的研究成果已经开始应用于医学诊断、治疗和药物研发。未来，计算生物学技术将进一步应用于农业、环境保护等领域，为人类的生活和发展提供有力支持。

[参 考 文 献]

[1] Li H. Protein-to-genome alignment with miniprot. Bioinformatics, 2023, 39: 1-6

[2] Esmaeeli R, Bauza A, Perez A, et al. Structural predictions of protein-DNA binding: MELD-DNA. Nucleic Acids Res, 2023, 51: 1625-36

[3] Rozowsky J, Gao J, Borsari B, et al. The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models. Cell, 2023, 186: 1493-511

[4] Keough KC, Whalen S, Inoue F, et al. Three-dimensional genome re-wiring in loci with human accelerated regions. Science, 2023, 380: eabm1696

[5] Kun E, Javan EM, Smith O, et al. The genetic architecture and evolution of the human skeletal form. Science, 2023, 381: eadf8009

[6] Wang Y, Tang H, Huang L, et al. Self-play reinforcement learning guides protein engineering. Nat Mach Intell, 2023, 5: 845-86

[7] Buljan M, Banaei-Esfahani A, Blattmann P, et al. A computational framework for the inference of protein complex remodeling from whole-proteome measurements. Nat Methods, 2023, 20: 1523-9

[8] Theodoris CV, Xiao L, Chopra A, et al. Transfer learning enables predictions in network biology. Nature, 2023, 618: 616-24

[9] Bai P, Miljković F, John B, et al. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. Nat Mach Intell, 2023, 5: 126-36

[10] Zhao Y, He B, Xu F, et al. DeepAIR: a deep learning framework for effective integration of sequence and 3D structure to enable adaptive immune receptor analysis. Sci Adv, 2023, 9: eabo5128

[11] Gainza P, Wehrle S, Van Hall-Beauvais A, et al. *De novo* design of protein interactions with learned surface fingerprints. Nature, 2023, 617: 176-84

[12] Zhang H, Zhang L, Lin A, et al. Algorithm for optimized mRNA design improves stability and immunogenicity. Nature, 2023, 621: 396-403

[13] Xiong Z, Cui X, Lin X, et al. Q-Drug: a framework to bring drug design into quantum space using deep learning. arXiv preprint arXiv:2308.13171, 2023

[14] Luo Z, Ni F, Wang Q, et al. OPUS-DSD: deep structural disentanglement for cryo-EM single-particle analysis. Nat Methods, 2023, 20: 1729-38

[15] Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature, 2023, 616: 259-65

[16] Karargyris A, Umeton R, Sheller MJ, et al. Federated benchmarking of medical artificial intelligence with MedPerf. Nat Mach Intell, 2023, 5: 799-810

[17] Sorin M, Rezanejad M, Karimi E, et al. Single-cell spatial landscapes of the lung tumour immune microenvironment. Nature, 2023, 614: 548-54

[18] Wu S, Hong G, Xu A, et al. Artificial intelligence-based model for lymph node metastases detection on whole slide images in bladder cancer: a retrospective, multicentre, diagnostic study. Lancet Oncol, 2023, 24: 360-70

[19] Hu H, Huang H, Li M, et al. A wearable cardiac ultrasound imager. Nature, 2023, 613: 667-75

[20] Man C, Lau V, Su S, et al. Deep learning enabled fast 3D brain MRI at 0.055 tesla. Sci Adv, 2023, 9: eadi9327

[21] Zhou YK, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. Nature, 2023, 622: 156-63

[22] Wolf J, Rasmussen DK, Sun YJ, et al. Liquid-biopsy proteomics combined with AI identifies cellular drivers of eye aging and disease *in vivo.* Cell, 2023, 186: 4868-84.

e12
[23] Oh HS, Rutledge J, Nachun D, et al. Organ aging signatures in the plasma proteome track health and disease. Nature, 2023, 624: 164-72
[24] 量子位智库. 计算生物学深度产业报告[EB/OL]. https://www.qbitai.com/2022/08/ 36776.html
[25] Mordor Intelligence. 计算生物学市场规模和份额分析-增长趋势和预测(2023-2028)[EB/OL]. [2023-12-25]. https://www.mordorintelligence.com/zh-CN/industry-reports/computational-biology-market