遗传 Hereditas (Beijing) 2021 年 10 月, 43(10): 924－929
www.chinagene.cn



# 前沿信息技术在生物医学大数据中的应用及展望

郑广勇[1]，曾涛[1]，李亦学[1,2,3,4]

1. 中国科学院上海营养与健康研究所，中国科学院计算生物学重点实验室，生物医学大数据中心，上海 200031
2. 广州国家实验室，广州 510320
3. 国科大杭州高等研究院，中国科学院大学，杭州 310013
4. 复旦大学遗传与发育协同创新中心，上海 200438

**摘要：** 近年来，随着以高通量组学检测技术为代表的生物技术(biological technology, BT)的发展，生物医学研究领域开始进入大数据时代。面对高维度、跨层次、多模态生物医学大数据，科学研究需要数据密集型科研新范式。云计算、区块链、人工智能等前沿信息技术(information technology, IT)的蓬勃发展为这种新型研究范式的实践提供了技术手段。本文对云计算、区块链、人工智能等前沿信息技术在生物医学大数据中的应用进行了描述，并对数据密集型科研新范式支撑环境的构建提出了前瞻展望,以期建立融合 BT&IT 技术的新型研究方案和科研新范式，最终推动生物医学研究跨越式发展。

**关键词：** 组学；云计算；区块链；人工智能；数据密集型科研新范式

# Application and prospect of cutting-edge information technology in biomedical big data

Guangyong Zheng[1], Tao Zeng[1], Yixue Li[1,2,3,4]

*1. CAS Key Laboratory of Computational Biology, Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
*2. Guangzhou Laboratory, Guangzhou 510320, China*
*3. Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310013, China*
*4. Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai 200438, China*

**Abstract:** In recent years, with the development of various high-throughput omics based biological technologies (BT), biomedical research began to enter the era of big data. In the face of high-dimensional, multi-domain and multi-modal biomedical big data, scientific research requires a new paradigm of data intensive scientific research. The vigorous development of cutting-edge information technologies (IT) such as cloud computing, blockchain and artificial intelligence provides technical means for the practice of this new research paradigm. Here,we describe the application of such

**收稿日期：** 2021−05−31；**修回日期：** 2021−09−16
**基金项目：** 中国科学院战略性先导科技专项课题(编号：XDB38050200)资助[Surpported by the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDB38050200)]
**作者简介：** 郑广勇，副研究员，研究方向：生物信息学。E-mail: gyzheng@picb.ac.cn
**通讯作者：** 李亦学，教授，研究方向：生物信息学。E-mail: yxli@sibs.ac.cn
DOI: 10.16288/j.yczz.21-192
**网络出版时间：** 2021/9/27 12:05:04
URI: https://kns.cnki.net/kcms/detail/11.1913.R.20210926.1702.002.html

cutting-edge information technologies in biomedical big data, and propose a forward-looking prospect for the construction of a new paradigm supporting environment for data intensive scientific research. We expect to establish a new research scheme and new scientific research paradigm integrating BT & IT technology, which can finally promote the great leap forward development of biomedical research.

**Keywords:** omics; cloud computing; blockchain; artificial intelligence; new paradigm of data intensive scientific research

自 2001 年“人类基因组计划”完成，生物医学研究开始进入“后基因组时代”。伴随着对基因组、转录组、蛋白组及代谢组等组学的深入研究，人们在微观的分子层面对生命科学有了系统化的认知[1]。近 10 年来，随着各种高通量组学技术的快速发展，基因组、表观遗传组、转录组、蛋白质组、代谢组、微生物组、相互作用组等组学数据正以前所未有的速度进行累积，如何高效分析解读这些组学背后的科学规律，从而在微观层面更加全面地认识生物体的分子机理，成为生物医学研究领域的一个重要课题[2]。特别值得注意的是，为了深入测量并描述生物体的行为和功能，表型组学近年应运而生。表型组是指生物体从微观(分子、细胞)到宏观(器官、组织、生物体)，从胚胎发育到出生、生长、衰老及死亡过程中，由基因与环境以及二者互相作用产生的所有形态、功能、行为等方面的生物学性状集合[3]。从表型组的定义可以知道，其涵盖了时间(生物体从出生到死亡的过程)和空间(分子、细胞、器官、组织、生物体)两个方面的信息。在表型组研究中，对生物体的物理表型(体质、影像)、化学表型(基因、蛋白质、转录组、代谢物、免疫因子等)以及生物表型(如肺功能、心功能和认知功能等)进行从宏观到微观的测量和分析，从而系统反映生物体在时间和空间两个维度上的动态变化过程[4]。由于表型组数据涵盖两个维度信息，刻画了从分子到生物体不同层次特性，同时包含文本、图片、影像等不同模式的数据，因而具有高维度、跨层次、多模态的特征。各类组学技术的蓬勃发展推动了生物医学领域研究进入数据密集型科研新范式时期，从而为领域的发展带来了挑战和机遇。在大数据时代，面对数据密集型科研新范式，生物信息学研究人员需要在传统的计算生物学方法中引入云计算、区块链、人工智能等前沿信息技术(information technology, IT)，支撑这种科研新范式的实践，进而高效解读海量不同维度、不同层次的生物医学领域数据，实现领域大数据的汇聚研究[5]。在此基础上，如果能够构建数据密集型科研新范式的支撑系统，则可以帮助科学家和临床医生从系统的层面上通过数据密集型的计算分析和计算实验，深度挖掘和发现大数据背后的价值，理解多维数据背后的科学规律，从而有力支持生物医学问题的基础研究和转化研究工作。本文将首先对云计算、区块链、人工智能等前沿信息技术在生物医学大数据中的应用进行描述，然后对数据密集型科研新范式支撑环境的构建提出展望。

## 1　云计算技术在生物医学大数据中的应用

云计算(cloud computing)是分布式计算的一种，指的是通过网络“云”将巨大的数据计算处理程序分解成无数个小程序，然后通过多部服务器组成的系统进行处理和分析这些小程序得到结果并返回给用户。与传统的本地计算技术相比，云计算技术具有以下优点：

(1)扩展性好：相比于传统的服务器计算，云计算能够快速地对应用进行动态扩展。云计算可根据用户不同的应用搭配不同的计算资源和存储资源，进行细粒度的资源部署，从而提高资源的使用效率。

(2)兼容性高：目前市场上大多数 IT 资源、软、硬件都支持虚拟化，因此云计算的兼容性非常强，能够对不同性能的机器进行统一管理配置，从而提高服务效率。

(3)可靠性高：由于云计算对各种计算资源进行统一的管理配置，因此单点服务器故障不会影响整个系统对外提供服务，因而比传统的本地服务器计算具有更高的可靠性。

(4)性价比高：将资源放在虚拟资源池中统一管理一定程度上优化了物理资源，用户不再需要昂贵的、存储空间大的主机，而是可选择相对廉价的计

算资源统一组成云并拥有不逊于大型主机的性能，因此具有良好的性价比。

面对生物医学大数据的快速增长，云计算的优点使其成为生物医疗领域计算生物学工作的必然选择。目前，云计算技术已经在许多生物医学基础研究和应用研究中进行使用，并取得了良好的效果(表 1)。Fischer 等[6]构建了基于云计算技术的全外显子测序数据分析流程，为罕见遗传疾病的机理研究提供了有效支撑。Samuel 等[7]搭建了一个跨平台访问的云计算资源池，为微生物组学数据分析提供了便利。Ben 等[8]构建了一款基于云计算技术的 SNP (single nucleotide polymorphism)识别工具，该工具可以高效地从人类基因组测序数据中识别 SNP 信息。Guo 等[9]使用云计算技术，构建了高效的宏基因组测序数据从头拼接软件，为宏基因组测序数据的解读提供了解决方案。美国国立生物技术研究中心 NCBI (National Center for Biotechnology Information)推出了基于 Google 云和亚马逊云的 BLAST+版本(https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=CloudBlast)，有效满足了超大规模的生物医学领域数据的序列比对的需求。美国 Broad 研究中心推出了基于 Google 云的 GATK4 套装软件(https://gatk.broadinstitute.org/)，从而为从大规模的基因组测序数据中识别胚系突变(germline mutation)和体细胞突变(somatic mutation)提供了解决方案。 近年，笔者基于云计算技术建立了智慧多组学数据分析系统(https://aicloud.biosino.org/casmap)。该系统能够对基因组、转录组、表观遗传组、微生物组、代谢组等多种生命组学大数据进行自动化分析。该系统与传统的分析系统相比，具有以下优点：(1)方便的数据分析，系统为多种组学数据的分析流程提供了可视化的操作界面，取代繁琐的命令行模式，零编程经验用户也可以通过简单的鼠标操作完成专业的组学数据分析。用户在系统中可以一键运行各种组学分析流程，并获得分析结果报告，并可以把报告中图表用于后续的论文发表。(2)可靠的数据挖掘，系统的后端存储了 500GB+的生命科学专业注释数据，帮助用户在开展组学数据分析时获得更为可靠的结果。系统中的分析流程全部根据高影响因子的 SCI 论文分析过程进行研发，确保数据挖掘的先进性，精准解读数据背后的意义。(3)高效的数据处理，系统基于云计算技术进行开发，能够弹性地应对用户的少量、中量、海量数据分析需求，极大的提高了分析效率，减少了数据处理时间，帮助用户高效快速地对各类生命组学数据进行深入解读。

## 2 区块链技术在生物医学大数据中的应用

区块链作为近年来的一项新兴技术，它具有去中心化、可追溯、不可伪造、公开透明等属性。区块链本质上是一个分布式数据库，采用去中心化和去信任的途径构建可信任的网络。狭义来讲，区块链是一种按照时间顺序将数据区块以顺序相连的方式组合成的一种链式数据结构，并以密码学方式保证不可篡改和不可伪造的分布式账本。广义来讲，区块链技术是利用块链式数据结构来验证与存储数据、用分布式节点共识算法来生成和更新数据、利用密码学的方式保证数据传输和访问的安全、利用智能合约来编程和操作数据的一种全新的分布式基础架构与计算范式。区块链是由节点参与的分布式

**表 1 云计算技术在生物医学大数据中的典型应用**

**Table 1 Typical application of cloud computing technology in biomedical big data**

| 云计算网站 | 网站功能描述 | 网站访问路径 |
|---|---|---|
| SIMPLEX | 全外显子数据分析平台 | https://icbi.i-med.ac.at/exome/ |
| CloVR | 微生物组学数据分析平台 | http://clovr.org/ |
| Crossbow | SNP 识别系统 | http://bowtie-bio.sourceforge.net/crossbow |
| DIME | 宏基因组数据拼接软件 | 无 |
| ElasticBLAST | 大规模序列搜索平台 | https://blast.ncbi.nlm.nih.gov/doc/elastic-blast |
| GATK | 基因组序列分析平台 | https://gatk.broadinstitute.org/ |
| CASMAP | 多组学数据分析平台 | https://aicloud.biosino.org/casmap |

数据库系统，众节点形成点对点的网络，没有中心化设备和管理机构，它不需要第三方信任背书。

目前，阻碍生物医学大数据广泛应用的一个主要问题是数据孤岛化。由于利益分配机制不明、隐私泄露风险、伦理法规约束等，大部分医疗领域的研究人员在实验数据和临床数据共享方面往往犹豫不决。因此，在生物医疗领域迫切需要建立安全、互利的数据共享机制，从而使数据通过流通与汇聚释放价值，推进生物医药产业的创新发展[10]。区块链技术的去中心化、可追溯、不可伪造、公开透明等属性赋予该技术应用于生物医疗领域数据管理共享的能力。Fan 等[11]通过区块链技术构建了电子病历共享系统，该系统在包含病人隐私的同时提供了病历的脱敏共享，有效促进了医疗信息流通。Jin 等[12]基于区块链技术搭建了个人基因组共享系统，为基因组数据共享提供了一个技术案例。美国哈佛大学的 Church 等[13]提出通过区块链技术来管理和共享个人基因组数据，将大大加速基因组研究和产业应用，具有良好的科学价值和社会经济价值。近年，笔者和合作伙伴一起构建了基于区块链技术和隐私安全计算技术的智能数据共享分析系统(https://platform.sdap.biosino.org/)，为医疗领域的数据共享提供了一站式的解决方案。该分析系统具有以下技术优点：(1)使用区块链技术构建灵活的多方数据共享模块，智能合约触发数据的确权和授权过程，分布式账本对数据的加载和消费情况进行记录和追溯，解决数据共享的信任问题；(2)使用隐私计算技术构建安全的多方数据分析模块，数据分析在沙箱内进行，不分享原始数据，分享数据的价值；(3)使用部分中心化+多节点分布式的技术构建智能多方数据存储模块，分布式的存储方案不仅保证原始数据的安全，同时也避免了大规模数据在不同用户间传输过程，大大提高了数据分析效率。

## 3　人工智能技术在生物医学大数据中的应用

人工智能是研发模拟、延伸、扩展人类智慧的理论、方法及技术的一门新兴学科，近年成为信息科学发展的一个重要研究方向[14,15]。利用人工智能技术建立符合生物医学大数据特征的数据库、算法及计算环境，正广泛深入生命科学的各个领域。人工智能技术广泛且深入的融入生物医学研究是目前生命科学发展的一个重要趋势。一方面，人工智能能够从海量的生物异质大数据中发现人类大脑无法分析、无法理解的数据结构，捕捉到人类无法意识到的生物学特征。另一方面，人工智能所应用的计算方法既可以模拟人类思考的特点，也可以完全摆脱人类的传统思考模式。利用这样的类脑方法来研究生命科学，可以更有效地处理生命现象的极端复杂性，使得研究更接近生命的本质。所以，人工智能技术可以帮助生物医学领域研究实现关键的实质性突破，革新生物医学研究的现有范式，拓展生物医学研究的范围，有助于阐明生物医学领域大量悬而未决的基本问题。

目前，人工智能技术已在生物医学研究的多个方向进行了应用和探索，在众多复杂的研究场景中都有新的发现：

(1)在分子细胞机理研究方面，基于人工智能技术中的深度学习方法可以建立高效的分子相互作用预测模型，进而帮助科学家解读复杂的生物过程背后的分子规律。例如，近年来，深度学习模型的快速发展与广泛应用有助于刻画细胞内基因的时空表达和顺式–反式调控[16]，蛋白–蛋白相互作用[17]，蛋白–代谢小分子相互作用[18]，细胞间的通讯[19]等生物过程机理。

(2)在生命组学数据分析方面，基于自然语言和人工智能逻辑的组学数据分析平台 DrBioRight，为下一代组学分析范式提供了五个特征示范[20]：(i)准确识别不具有专门技术性知识的用户所提出的分析请求；(ii)帮助用户探索和理解与任务相关的组学数据和分析结果；(iii)通过稳定用户群的贡献保持对组学数据和分析方法的及时更新；(iv)经由用户对分析质量的反馈不断修正和更新平台性能；(v)与智能移动平台和社交媒体实现良好匹配，为分析流程增加更多的灵活性。

(3)在生物医学知识图谱发展方面，基于监督的深度学习策略，关系抽取模型能够在不依赖于人工标注数据的情况下应用到各种生物医学关系抽取场景当中，可从千万篇科研文献中挖掘理解药物、靶点、病毒、副作用等等生物医学实体之间相互作用

规律的生物医学实体关系网络，进而通过抽取出的提示性信息指导实验验证；例如通过查找文献支持来验证针对“非典”或“中东呼吸综合征”的老药新用策略的可行性，及其针对“新冠病毒”的有效性[21]。

(4)在生物模型算法发展方面，scDEC 使用一组生成对抗网络将高维单细胞数据映射到低维隐空间，在低维空间进行聚类分析，再使用另一组生成对抗网络将低维数据映射回高维空间，从而为在单细胞数据分析提供集数据降维、生成与细胞聚类于一体的智能算法[22]。基于卷积神经网络算法的人工智能模型可在大量临床影像数据基础上进行学习训练临床诊断模型，从而辅助临床医生实现对患者的高准确率诊断[23]。人工智能技术与计算物理、量子化学、分子动力学等技术的结合，将有助于提高药物发现与发展这一关键环节的效率与成功率，从而降低新药研发成本，为新药研发带来了新的发展动力[24]。

## 4 结语与展望

现代生物医学研究的目标之一是在分子、细胞、组织、器官等层面上解析生物体外在表型所对应的内在组成形式及其相互作用规律。由于生命体系的高度复杂和精准调控特性，以生物化学、分子生物学等学科为代表的现代生物医学研究发展了几十年后，遇到了重大的瓶颈。现代生物医学研究的重点突破，需要对研究技术和研究模式进行根本性的变革。近年来，随着以高通量组学检测技术为代表的生物技术(biological technology, BT)的成熟与发展，以及以云计算、区块链、人工智能为代表的前沿信息技术的发展，建立融合 BT&IT 技术的新型研究方案和科研新范式，将是打破现代生物医学研究瓶颈，推动生物医学研究跨越式发展的必由之路。

面对数据密集型科研新范式的需求，构建一个融合 BT&IT 技术，界面友好、安全可靠、用户充分可及的生物医学大数据操作系统，进而建立密集型科研新范式的应用支撑环境，可以非常有效地帮助生命科学研究人员方便地实现生物医学大数据的获取、交互共享、智能化调度、多维深度展示、高性能计算和深度挖掘分析等各类科学实验活动，进而加速生物医学大数据整合，融汇和贯通各类高维多层次复杂数据，推动数据共享和充分利用，实现生物医学大数据的汇聚研究，推动生物医学研究获得革命性进展。

## 参考文献(References):


[1] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Gene*, 2016, 17(6): 333–351. [DOI]

[2] Nimrod R, Ron S. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res*, 2018, 46(20): 10546–10562. [DOI]

[3] Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. *Nat Rev Genet*, 2010, 11(12): 855–866. [DOI]

[4] Brown SDM, Holmes CC, Mallon AM, Meehan TF, Smedley D, Wells S. High-throughput mouse phenomics for characterizing mammalian gene function. *Nat Rev Genet*, 2018, 19(6): 357–370. [DOI]

[5] Milicchio F, Rose R, Bian J, Min J, Prosperi M. Visual programming for next-generation sequencing data analytics. *BioData Min*, 2016, 9:16. [DOI]

[6] Fischer M, Snajder R, Pabinger S, Dander A, Schossig A, Zschocke J, Trajanoski Z, Stocker G. SIMPLEX: cloud-enabled pipeline for the comprehensive analysis of exome sequencing data. *PLoS ONE*, 2012, 7(8): e41948. [DOI]

[7] Angiuoli SV, Matalka M, Gussman A, Galens K, Vangala M, Riley DR, Arze C, White JR, White O, Fricke WF. CloVR: a virtual machine for automated and portable sequence analysis from the desktop usingcloud computing. *BMC Bioinformatics*, 2011, 12:356. [DOI]

[8] Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. Searching for SNPs with cloud computing. *Genome Biol*, 2009, 10(11): R134. [DOI]

[9] Guo X, Yu N, Ding XJ, Wang JX, Pan Y. DIME: a novel framework for de novo metagenomic sequence assembly. *J Comput Biol*, 2015, 22(2): 159–177. [DOI]

[10] Byrd JB, Greene AC, Prasad DV, Jiang XQ, Greene CS. Responsible, practical genomic data sharing that accelerates research. *Nat Rev Genet*, 2020, 21(10): 615–629. [DOI]

[11] Fan K, Wang S, Ren Y, Li H, Yang Y. MedBlock: efficient and secure medical data sharing via blockchain. *J Med Syst*, 2018, 42(8): 136. [DOI]

[12] Jin XL, Zhang M, Zhou ZY, Yu XY. Application of a blockchain platform to manage and secure personal genomic data: a case study of LifeCODE.ai in China. *J*

*Med Internet Res*, 2019, 21(9): e13587. [DOI]

[13] Zhavoronkov A, Church G. The advent of human life data economics. *Trends Mol Med*, 2019. 25(7): 566–570. [DOI]

[14] Wu F, Lu CW, Zhu MJ, Chen H, Zhu J, Yu K, Li L, Li M, Chen QF, Li X, Cao XD, Wang ZY, Zha ZJ, Zhuang YT, Pan YH. Towards a new generation of artificial intelligence in China. *Nat Mach Intell*, 2020, 2(6): 312–316. [DOI]

[15] Zhao XT, Yang YD, Qu HZ, Fang XD. Applications of machine learning in clinical decision support in the omic era. *Hereditas(Beijing)*, 2018, 40(9): 693–703.
赵学彤, 杨亚东, 渠鸿竹, 方向东. 组学时代下机器学习方法在临床决策支持中的应用. 遗传, 2018, 40(9): 693–703. [DOI]

[16] Alexandra Maslova, Ricardo N. Ramirez, Ke Ma, Hugo Schmutz, Chendi Wang, Curtis Fox, Bernard Ng, Christophe Benoist, Sara Mostafavi. Deep learning of immune cell differentiation. *Proc Natl Acad Sci USA*, 2020, 117(41): 25655–25666. [DOI]

[17] Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, George M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods*, 2019, 16(12): 1315–1322. [DOI]

[18] Morton JT, Aksenov AA, Nothias LF, Foulds JR, Quinn RA, Badri MH, Swenson TL, Van Goethem MW, Northen TR, Vazquez-Baeza Y, Wang M, Bokulich NA, Watters A, Song SJ, Bonneau R, Dorrestein PC, Knight R. Learning representations of microbe–metabolite interactions. *Nat Methods*, 2019, 16(12): 1306–1314. [DOI]

[19] Shao X, Lu XY, Liao J, Chen HJ, Fan XH. New avenues for systematically inferring cell-cell communication: through single-cell transcriptomics data. *Protein Cell*, 2020, 11(12): 866–880. [DOI]

[20] Li J, Chen H, Wang YM, May Chen MJ, Liang H. Next-generation analytics for omics data. *Cancer Cell*. 2021, 39(1): 3–6. [DOI]

[21] Hong LX, Lin JJ, Li SY, Wan FP, Yang H, Jiang T, Zhao D, Zeng JY. A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories. *Nat Mach Intell*, 2020, 2(6): 347–355. [DOI]

[22] Liu Q, Chen SQ, Jiang R, Wong WH. Simultaneous deep generative modeling and clustering of single cell genomic data. *Nat Mach Intell*, 2021, 3(6): 536–544. [DOI]

[23] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 2017, 542(7639): 115–118. [DOI]

[24] Jing YK, Bian YM, Hu ZH, Wang LR, Xie XQ. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. *AAPS J*, 2018, 20(3): 58. [DOI]

(责任编委: 方向东)