doi: 10.3978/j.issn.1000-4432.2021.01.24
View this article at: http://dx.doi.org/10.3978/j.issn.1000-4432.2021.01.24

人工智能和区块链技术在生物样本库信息化建设的应用展望

王东妮[1]，东野枚枚[1]，张栩琳[1]，杨子英[1] 综述 林浩添[1,2] 审校

(1. 中山大学中山眼科中心，中山大学眼科学国家重点实验室，广州 510060；
2. 中山大学精准医学科学中心，广州 510080)

[摘 要] 近年来，使用人工智能(artificial intelligence，AI)技术对临床大数据及图像进行分析，对疾病做出智能诊断、预测并提出诊疗决策，AI正逐步成为辅助临床及科研的先进技术。生物样本库作为收集临床信息和样本供科研使用的平台，是临床与科研的桥梁，也是临床信息与科研数据的集成平台。影响生物样本库使用效率及合理共享的因素有信息化建设水平不均衡、获取的临床及检验信息不完全、各库之间信息不对称等。本文对AI和区块链技术在生物样本库建设中的具体应用场景进行探讨，展望大数据时代智能生物样本库信息化建设的核心方向。

[关键词] 人工智能；生物样本库；信息化建设；区块链

Prospect of application of artificial intelligence and block chain in the information construction of Biobank

WANG Dongni[1], DONGYE Meimei[1], ZHANG Xulin[1], YANG Ziying[1], LIN Haotian[1,2]

*(1. State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060;
2. Center for Precision Medicine, Sun Yat-sen University, Guangzhou 510080, China)*

Abstract In recent years, artificial intelligence (AI) technology has been applied to analyze clinical big data and images and then make intelligent diagnosis, prediction and treatment decisions. It is gradually becoming an advanced technology to assist clinical and scientific research. Biobank is a platform for collecting clinical information and samples for scientific research, serving as a bridge between clinical and scientific research. It is also an integrated platform of clinical information and scientific research data. However, there are some challenges. First, clinical and laboratory information obtained is incomplete. Additionally, the information among different databases is asymmetric, which seriously impedes the information sharing among different Biobanks. In this article, the specific application scenarios of AI technology and blockchain in the construction of a Biobank were discussed, aiming to pinpoint the core direction of the information construction of an intelligent Biobank in the era of big data.

Keywords artificial intelligence; Biobank; informatization construction; blockchain

收稿日期 (Date of reception)：2020–06–30

通信作者 (Corresponding author)：林浩添，Email: gddlht@aliyun.com

基金项目 (Foundation item)：广东省科技计划项目 (2019B030316012). This work was supported by the Science and Technology Planning Projects of Guangdong Province (2019B030316012), China.

疾病生物样本库是建设数量最多且普遍存在于各级医疗机构的生物样本库，样本由患有相关疾病的患者捐献。医院的电子病历系统(electronic medical record，EMR)、医院信息系统(hospital information system，HIS)、实验室信息系统(laboratory information system，LIS)、影像归档和通信系统(picture archiving and communication systems，PACS)等已广泛应用于各级医疗机构[1]，存储着大量的临床信息。将这些系统通过接口连接到生物样本库信息系统是生物样本库信息化建设的关键步骤，也是注释样本属性的数据提取方式，但各管理系统中储存的数据类型种类繁多，包含结构化数据、以自然语言描述的非结构化数据、影像数据、检查报告等，单纯抓取原始数据到生物样本库信息系统只会重复储存，造成数据冗余，浪费人力、物力和财力[2]，生物样本库急需利用智能化的数据处理方式来应对这一难题。由于机器学习技术的进步，人工智能(artificial intelligence，AI)在医学上的应用引起了广泛关注[3-4]。自然语言处理技术在信息检索中的应用[5]、深度学习技术在自然语言处理和在图像识别中的应用以及区块链技术均有望成为解决生物样本库信息化建设与信息共享的核心技术。

1 生物样本库发展概况

临床生物样本蕴藏着许多与疾病相关的信息，是不可复制的科学研究资源。20世纪90年代以来，欧美等发达国家纷纷建立了大规模人群样本库，如拥有70万例样本的美国国家癌症研究所建立的国家级肿瘤生物样本库(Cooperative Human Tissue Network，CHTN)[6]、招募了50万名40~69岁志愿者并记录其医疗健康数据的英国生物样本库(United Kingdom Biobank，UK Biobank)[7]、囊括了欧洲30多个国家的200多个机构的泛欧洲生物样本库与生物分子资源研究平台(Biobanking and Biomolecular Resources Research Infrastructure，BBMRI)[8]。为了保护我国各民族基因组并供永久性研究，中国科学院在1994年建立了中华民族永生细胞库，是目前国内规模最大的各民族永生细胞库[9]。自此，国内各类生物样本库应运而生，在早期的生物样本库建设过程中，建设者都更加注重样本数量建设，生物样本得到快速积累，但由于对样本信息疏于管理，导致信息错漏，对样本的应用及共享产生了极大的限制。如何深入挖掘样本信息，加速科学的共享与利用，是当前生物样本库建设的重要方向。

2 语义分析在样本库基础建设领域的应用

国内生物样本库立足于我国丰富的遗传资源、多样化的疾病类型，建设与发展模式渐趋成熟，在疾病防控、精准医疗、早筛早诊中的作用日益增加。标准化的样本和数据管理作为生物样本库的重要一环，是获取高质量样本和数据的基础，也是促进转化医学和精准医学发展的基石。然而，庞大的临床资源却由于数据结构化程度低、电子信息化建设水平参差不齐，导致利用率较低[2]。如何有效地整合、挖掘现有临床资源，是生物样本数据库建设的基础问题。

随着AI的发展、深度学习模型的开发和优化，语义分析有望成为解决上述问题的钥匙。语义分析是AI的一个分支，将自然语言转化为计算机能够理解的语言，通过如循环神经网络(Recurrent Neural Networks，RNNs)、长短时记忆模型(Long Short-Term Memory，LSTMs)及其他模型训练机器学习、“理解”，并以自然语言给出分析结果[10]。目前，语义分析的信息处理已从表层特征向深层语义分析转变，并在多个领域内应用。在舆情分析方面，通过抓取社交媒体的相关信息，语义识别应用于包括欺诈交易识别等的犯罪活动检测[11]；在生物医学方面，自然语言处理和关系提取已应用于文献整合、构建疾病的全蛋白质谱及基因序列标记[12]等。

在生物样本库信息化建设中，整理样本捐献者的临床信息是必不可少的环节，如患者的基本信息、门诊信息和住院信息可以从HIS系统里获取；患者的检验信息可以从LIS系统里获取；患者的影像信息可以从PACS系统中获取；患者的病历信息可以从EMR中获取。语义分析的文本信息提取可以帮助研究者提取与样本相关的关键信息，即通过对文本信息的抽取，精炼庞杂的临床数据，为样本带上多个“标签”，方便研究者进行样本的筛选和统计分析；文本分类和聚类可以实现样本的自动分类，方便研究者进行大型队列研究和数据分类；智能检索可以协助研究者在临床信息数据池中挖掘和提取有效信息，在将信息结构化处理后，建立语义化描述疾病资源

相关特征的模式，进一步提高临床数据的可用性、共享性。

3 图像识别技术在样本信息挖掘工作中的应用

高度信息化建设的生物样本库会全面保留样本捐献者的基本信息、临床信息、样本信息、科研数据等，其中就包含X线、CT、MRI、裂隙灯照片、眼底照片、病理图片、细胞与组织照片等图像数据。从这些非结构化的图像中提取有效信息，往往依赖于科研人员的临床经验，准确性与一致性都得不到保证。

图像识别是指在计算机系统的辅助下对图像进行处理与分析，识别并提取目标区域的技术[13]。在实际工作中，HIS，LIS系统中大部分是结构化数据，比较方便获取，但PACS系统内的数据和电子病历数据为图片和文本数据，想从中提取信息需要研究者逐个查看并整理关键信息。将AI技术与PACS系统集成开发[14]，将AI诊断结果反馈在样本库系统中，可以对患者的临床诊断进行验证，减少误诊漏诊，使样本使用者在选择样本时对患者的诊断进行二次核查，提高科研的严谨性。基于深度学习的图像识别技术在放射学、超声学、病理学、皮肤科学、眼科学等一些需要影像数据分析的医学学科中成果繁多[15-19]。特别是在眼科学领域发展迅猛。中山大学中山眼科中心AI团队研发了通过收集、分析患者的裂隙灯图片，开发了集筛查、危险度评估和辅助治疗为一体的先天性白内障智能诊断与决策系统CC-Cruiser[20]。并以该系统为核心完成了全球首个AI多中心随机对照临床研究，提出了医学AI临床应用评判标准，推动了AI临床转化和落地应用的进程[21]。

近日，该团队研发的一种基于解剖学和病理学特征的医学图像密集标注技术Visionome问世，该技术比传统图片分类标注方法多产生12倍标签，可准确识别多种眼前段病变，准确率高达93.75%，且在20种未经过学习的眼病大规模筛查场景中准确率达84.00%，实现了AI跨专科、多病种应用[22]。Visionome所产生的标签正是生物样本库呈待结构化的图像数据。与常规的AI诊断不同的是，生物样本信息数据库内本身就包含患者确切的临床诊断，提取图像对应的诊断结果，可以作为重要的参考标准提高Visionome识别各类标签的准确性，将这些标签存储在生物样本库系统内作为对样本属性的注释，更加细化的区分了样本分析前变量，有望成为未来科学研究的新模式。

4 区块链技术在生物样本共享模式的应用探索

生物样本库旨在为基础科研和临床医学研究提供合适的样本及数据，“只存不用”、“样本私有化”、无法实现样本资源的应用和共享，只会发展成“私库”或“垃圾库”，令生物样本库失去其存在的意义[23]。我国生物样本库在共享方面普遍存在的问题主要有：1)“私库”比较泛滥，样本拥有者共享意愿低；2)缺乏完善的共享机制平台；3)样本基本信息及其关联信息没有统一标准，不利于数据结构化和共享；4)存在知情同意、隐私泄露、“生物剽窃”等伦理问题和法律问题。如何通过技术手段解决样本共享问题，是实现生物样本价值最大化的根本途径。

区块链和AI同属于近年来炙手可热的新兴技术，但区块链不属于AI技术，它们之间是相辅相成的关系。区块链能够为数据安全、数据管理、数据共享提供强大的技术保障，同时为数据来源的真实性和生物样本的伦理问题提供更好的解决方案。进一步而言，区块链下的生物样本信息集成数据库，或能成为大数据时代中生物信息数据挖掘和AI应用的先决条件。

4.1 基于区块链结构下的数据安全建立信任关系

我国生物样本的主要获取渠道包括医学检验、病理检验、手术诊疗等，这一系列行为过程所涉及的隐私保护和伦理问题贯穿生物样本库建设始终，存在捐献者知情同意、数据保密、捐献者和样本库间的相互信任、样本库商业化运营、国际合作等核心伦理挑战[24]。区块链技术作为一个由多方共同维护、去中心化的分布式记账技术，核心在于通过对等网络协议、共识算法、非对称加密、哈希等关键技术解决数据传递与交换过程中的信任问题。区块链的链式结构在于将不同的数据区块按时间戳顺序相连来进行数据存储与验证；区块链网络中的每个节点都可以共享数据，并且同步条件下的所有副本都与其他节点完全相同；访问者需要获得唯一的私钥解密公钥进行区块内容的访问[25-26]。因此，即使黑客获得私钥

企图篡改单一数据区块，将无法使攻击生效，黑客必须同时攻击与该数据区块相连的所有节点中的所有副本，由此产生的技术难度极大，目前仍无法实现。

区块链的可溯源、不可篡改、高冗余、安全透明及成本低廉等属性，可有效解决生物样本数据泄露、捐献者隐私保护和伦理问题，使人们愿意信任和乐意共享数据。在区块链技术能带来各方信任的基础上，建立相关问责制、样本和数据流向公开和捐献者自主决策的治理体系，将适应并保护所有利益相关者的需求和权利，包括捐献者、研究人员及样本库基金赞助者[27-28]。Mamo等[28]率先做出尝试，创立了一个“动态同意”的门户网站Dwarna，作为生物样本库不同利益相关者的枢纽，Dwarna连接生物样本库管理者、研究人员、捐献者和公众。参与者可在研究过程中根据自己的意愿进行同意/撤销同意的操作。而同意变更的记录将保存在区块链中，区块链会为其附加一个时间戳。通过在区块链中托管同意变更，使研究过程更为透明。

4.2 实现智能样本流程管理和生物信息共享

智能合约是基于可信和不可篡改的数据，自动化验证和执行预先定义好的规则和条款。智能合约允许在没有第三方的情况下进行可信交易，并具有可追踪且不可逆转的特性。这对于生物样本库的信息化管理具有高度适用性，通过智能合约控制链流程，有助于实现生物样本从采集到出入库的全流程智能化管理[29]。

生物样本携带的基因信息对疾病预防有重要指导作用。对个体生物信息进行纵向对比，能追踪个人身体健康的变化；若进行横向对比，可进行大数据挖掘，这些数据的价值不言而喻。但现实中，这些数据往往存储在孤立的医疗或科研机构里，机构与机构之间无法进行数据流通，个体本身也无法真正实现对数据的拥有权和使用权。Nebula Genomics公司推出了一项业务，消费者花费999美元(项目代币)测试自己的基因信息，并使用区块链技术保障其数据和交易记录的安全性。消费者可以对自己的数据进行管理，自主决定把数据有偿或无偿分享给他人。国外私人企业先于公立机构利用区块链技术在生物样本信息的共享上做出了尝试，对于第三方样本库的运营提供一定的参考意义。

除了个人生物信息，生物样本库之间也可以利用区块链技术实现安全的信息交换。Evangelatos等[30]开发出一个生态系统，在生物银行和免费/自由开源软件(free/libre open source software，FLOSS)之间利用区块链技术实现数据接口，保护信息共享空间免受搭便车问题的影响，并在不妨碍其运营框架的情况下保证其可持续性。

4.3 区块链模式下生物样本信息共享的发展趋势

随着互联网技术的发展，生物样本库将发展为生物银行，其运营目标是通过线上数据共享，线下实现生物样本的分享，以实现资源的合理利用和价值提升。大数据时代下的生物样本库需要发展新的数据管理技术来为日常运营、信息共享提供有力的支撑，令生物样本库真正成为分享型样本库，能支持样本存储的核心业务，支撑样本分享和数据共享的业务模式。Dwarna门户网站、Nebula Genomics公司、Nikolaos Evangelatos团队等展示了区块链面向生物银行的应用，设计实现这样的系统的可能性。但在不同国家不同国情不同研究领域的样本库，需要针对具体情况设计个性化的信息化管理系统。生物样本库的信息化系统应秉承“分类适用”的理念，与领域和应用紧密结合，故其架构也与应用相对应，可能是去中心的，也可能是弱中心或多中心的。信息化是业务发展和改革的基础，很多时候也是改革的先锋，甚至引领应用创新。区块链技术的应用前景甚好，但需要医疗行业规范和医疗数据知识产权规范等宏观设计与规范执行，在我国的探索依然任重道远，但我们相信，与区块链促进了金融技术的演进一样，生物样本库将伴随新的共享模式焕发出新的生命力。

5 小结

本文分析了AI技术在生物样本库信息化建设中可能的应用场景，通过使用语义识别、图像识别技术辅助科研人员快速检索到更加符合条件的目的样本，利用区块链技术促进样本的应用共享。但AI技术本身还处在发展阶段，虽可以节省科研数据收集时间，却存在许多技术上的瓶颈，不能充分提取原始数据的有效信息。此外，生物样本库还处在标准化建设的初级阶段，尚未建立

统一的标准数据集，导致AI技术在生物样本库建设中的应用存在异质性，不利于广泛应用。未来AI技术和生物样本库标准化建设的共同发展可以促使生物样本库信息化建设的统一，促进数据与样本的共享和合理使用。

参考文献

1. Van Driest SL, Wells QS, Stallings S, et al. Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records[J]. JAMA, 2016, 315(1): 47-57.
2. Asche CV, Kim J, Kulkarni AS, et al. Assessment of association of increased heart rates to cardiovascular events among healthy subjects in the united states: analysis of a primary care electronic medical records database[J]. ISRN Cardiol, 2011, 2011: 924343.
3. Lee JG, Jun S, Cho YW, et al. Deep learning in medical imaging: general overview[J]. Korean J Radiol, 2017, 18(4): 570-584.
4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence[J]. Nat Med, 2019, 25(1): 44-56.
5. Wiles KR, Washington MK. Implementation of an error-reporting module within a biorepository IT application to enhance operations[J]. Biopreserv Biobank, 2014, 12(6): 365-373.
6. UK Biobank data on 500,000 people paves way to precision medicine[J]. Nature, 2018, 562(7726): 163-164.
7. Holub P, Swertz M, Reihs R, et al. BBMRI-ERIC directory: 515 biobanks with over 60 million biological samples[J]. Biopreserv Biobank, 2016, 14(6): 559-562.
8. 褚嘉祐, 徐玖瑾, 傅松滨, 等. 中华民族永生细胞库的建立[J]. 国际遗传学杂志, 2008(4): 241-247.
   CHU Jiayou, XU Jiujin, FU Songbin, et al. The establishment of the immortalize cell bank of different Chinese ethnic groups[J]. International Journal of Genetics, 2008(4): 241-247.
9. Zhou L, Zhang D. NLPIR: A theoretical framework for applying natural language processing to information retrieval[J]. J Am Soc Inf Sci Technol, 2003, 54(2): 115-123.
10. Velupillai S, Mowery D, South BR, et al. Recent advances in clinical natural language processing in support of semantic analysis[J]. Yearb Med Inform, 2015, 10(1): 183-193.
11 Hill R, Akhgar B, Saathoff G, et al. Application of big data for national security[M]. Oxford: Butterworth-Heinemann, 2015.
12. Kobeissy F, Wang KK, Alawieh A, et al. Leveraging biomedical and healthcare data[M]. Academic Press, 2018.
13. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data[J]. Radiology, 2016, 278(2): 563-577.
14. 王磊, 郑云肆, 王培军. PACS与人工智能诊断系统的接口研究与实现[J]. 中国数字医学, 2020, 15(1): 22-24.
   WANG Lei, ZHENG Yunlu, WANG Peijun. The interface research and implementation on PACS and artificial intelligence diagnostic system[J]. China Digital Medicine, 2020, 15(1): 22-24.
15. Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology[J]. Nat Rev Cancer, 2018, 18(8): 500-510.
16 Huang Q, Zhang F, Li X. Machine learning in ultrasound computer-aided diagnostic systems: a survey[J]. Biomed Res Int, 2018, 2018: 5137904.
17. Wong ST. Is pathology prepared for the adoption of artificial intelligence?[J]. Cancer Cytopathol, 2018, 126(6): 373-375.
18. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542(7639): 115-118.
19. Long E, Lin H, Liu Z, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts[J]. Nature Biomedical Engineering, 2017, 1(2): 0024
20. Lin H, Li R, Liu Z, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial[J]. EClinicalMedicine, 2019, 9: 52-59.
21. Li W, Yang Y, Zhang K, et al. Dense anatomical annotation of slit-lamp images improves the performance of deep learning for the diagnosis of ophthalmic disorders[J]. Nat Biomed Eng, 2020, 4(8): 767-777.
22. 陈思静, 吴茂锋, 李佩娟. 生物样本库的建设与发展[J]. 生物化工, 2019, 5(4): 164-166.
   CHEN Sijing, WU Maofeng, LI Peijuan. Construction and development of biobank[J]. Shengwu Huagong, 2019, 5(4): 164-166.
23. 高梦婕, 王化群. 基于区块链的可搜索医疗数据共享方案[J]. 南京邮电大学学报(自然科学版), 2019, 39(6): 94-103.
   GAO Mengjie, WANG Huaqun. Blockchain-based searchable medical data sharing scheme[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science), 2019, 39(6): 94-103.
24. 单芳, 桑爱民, 薛琴, 等. 生物样本库研究的隐私保护问题及伦理反思[J]. 中国卫生事业管理, 2020, 37(1): 43-46.
   SHAN Fang, SANG Aimin, XUE Qin, et al. Ethical issues and reflections on privacy protection in biobank research[J]. Chinese Health Service Management, 2020, 37(1): 43-46.
25. 刘彦松, 夏琦, 李柱, 等. 基于区块链的链上数据安全共享体系研究[J]. 大数据, 2020, 6(5): 92-105.

LIU Yansong, XIA Qi, LI Zhu, et al. Research on secure data sharing system based on blockchain[J]. Big Data Research, 2020, 6(5): 92-105.

26. Hansson MG. Ethics and biobanks[J]. Br J Cancer, 2009, 100(1): 8-12.
27. Hawkins AK, O'Doherty K. Biobank governance: a lesson in trust[J]. New Genet Soc, 2010, 29(3): 311-327.
28. Mamo N, Martin GM, Desira M, et al. Dwarna: a blockchain solution for dynamic consent in biobanking[J]. Eur J Hum Genet, 2020, 28(5): 609-626.
29. 娄颜超, 陈要伟. 区块链技术在医疗领域中的应用[J]. 电子技术与软件工程, 2020(3): 188-189.

LOU Yanchao, CHEN Yaowei. Application of blockchain technology in the medical field[J]. Electronic Technology and Software Engineering, 2020(3): 188-189.

30. Evangelatos N, Upadya SP, Venne J, et al. Digital transformation and governance innovation for public biobanks and free/libre open source software using a blockchain technology[J]. OMICS, 2020, 24(5): 278-285.

本文引用：王东妮, 东野枚枚, 张栩琳, 杨子英, 林浩添. 人工智能和区块链技术在生物样本库信息化建设的应用展望[J]. 眼科学报, 2021, 36(1): 91-96. doi: 10.3978/j.issn.1000-4432.2021.01.24

Cite this article as: WANG Dongni, DONGYE Meimei, ZHANG Xulin, YANG Ziying, LIN Haotian. Prospect of application of artificial intelligence and block chain in the information construction of Biobank[J]. Yan Ke Xue Bao, 2021, 36(1): 91-96. doi: 10.3978/j.issn.1000-4432.2021.01.24