基于纳米孔数据读取的 DNA 存储技术

林艺生[1]，武瑞君[2]，钱珑[3]，张成[1]

[1]（北京大学计算机学院，北京 100871）

[2]（中国生物技术发展中心，北京 100039）

[3]（北京大学定量生物学中心，北京 100871）

摘要：现代社会正处于数据爆炸的时代，全球对数据存储的需求已经远远超过了已有的存储能力。作为天然的遗传信息载体，DNA 还能作为信息材料，实现一种稳定、高效、低功耗的数据存储。目前的 DNA 存储过程主要分为六个环节：编码、写入、保存、检索、读取、解码。其中，而纳米孔测序技术被广泛用于读取 DNA 中所存储的信息。本文系统性介绍了纳米孔分子检测技术的原理和发展历史，及其在 DNA 存储的应用。此外，本文总结了机器学习在纳米孔检测技术的应用，着重介绍了结合机器学习的新型纳米孔检测技术。这为纳米孔检测技术的发展提供了新方向，也为新型实用化 DNA 存储系统的发展奠定了基础。

关键词 纳米孔测序；纳米孔检测；DNA 存储；机器学习

中图分类号：Q 819　文献标志码 A　doi: 10.12146/j.issn.2095-3135.20231030003

DNA storage based on nanopore information readout

LIN Yisheng[1], WU Ruijun[2], QIAN Long[3], ZHANG Cheng[1]

[1] (School of Compute of Peking University, Beijing, 100871, China)

[2] (China National Center for Biotechnology Development Beijing, 100039, China)

[3] (Center for Quantitative Biology, Peking University, Beijing, 100871, China.)

Abstract: In the era of data explosion, the global demand for data storage has far exceeded the existing storage capacity. DNA, the natural carrier of genetic information, provides a stable, efficient, and sustainable data storage solution. The process of DNA storage is divided into six main parts: encoding, writing, preservation, access, reading, and decoding; and nanopores are widely used to read information stored in DNA because they are capable of detecting a multitude of signals including DNA sequence and its modifications and structural variants. In this review, we systematically introduce the principle and research history of nanopore-based DNA signal detection, and the applications of nanopore sequencing in DNA storage. In addition, we summarize the applications of machine learning in nanopore detection which presents a new direction for the development of nanopore technology and lays the foundation for building a better DNA storage system.

Key words: Nanopore sequencing; Nanopore detection; DNA storage; Machine learning

Funding: This work was supported by: National Key Research and Development Program of

来稿日期：2023-10-30 修回日期：2024-02-25

基金项目：国家重点研发项目(2021YFF1200100)，国家自然科学基金(62073133, 62273008, 12090054)，之江实验室(2022RD0AB03)，预研项目(31511090301)

作者简介：林艺生，研究生，研究方向为 DNA 存储、纳米孔测序分析、生物信息学；武瑞君，副研究员，研究方向为生物技术与生物医药领域战略；钱珑（通讯作者），助理研究员，研究方向为复杂生物系统的进化理论、计算合成生物学、DNA 分子信息存储，E-mail：long.qian@pku.edu.cn；张成（通讯作者），副研究员，研究方向为生物计算（DNA 计算和分子电路）、DNA 存储、纳米孔测序、生物信息技术、纳米智能机器，E-mail：zhangcheng369@pku.edu.cn

China (2021YFF1200100 to C.Z.), National Natural Science Foundation of China (62073133, 62273008 and 12090054 to J. Y., C. Z. and L.Q.), Zhejiang Lab (NO. 2022RD0AB03 to C.Z.), Prestudy project (31511090301 to C. Z)

1 引言

人类文明的发展离不开信息的传递与交互。通过获取和输出不同的信息，人类认识并改变世界。目前，我们正处于一个数据爆炸的时代，数字化技术和互联网技术的蓬勃发展使得大量的数据可以被收集、存储和分析。根据国际数据公司（International Data Corporation）预测，2025 年，全球数据量将达到 188ZB[1]，这些海量数据的保存将需要大量的存储介质。然而，在可预见的未来，传统存储介质将不可避免地陷入资源枯竭的困境[2]。因此，寻找新的存储介质已经迫在眉睫。DNA 作为遗传信息的载体，有着许多传统存储介质无法比拟的优点：超高的信息密度（455 EB/g[3]）、超强的稳定性（半衰期超过 500 年[4]，[5]）、维护能耗相对较低、可编程性和可寻址性[6]。因此，大量的 DNA 存储研究工作喷涌而出。

与传统的存储技术一样，DNA 存储过程同样也分为六部分：编码、写入、保存、检索、读取、解码[7]。其中，写入和读取，通常对应着合成和测序。目前，固相芯片合成技术和高通量测序可以迅速、低价地大规模合成和读取 DNA[8]，[9]。但高通量测序技术长度十分有限（< 200 bp）[10]，因此，样品的处理和大量短序列组装造成了较长的数据读取延迟。因此，需要一种实时读取且精确可靠的 DNA 测序技术来实现分子存储的高效读取。

受到分子识别和跨膜运输的生物过程启发，科学家们开发了纳米孔检测技术，作为单分子检测的超灵敏分析工具[11]，[13]。通常情况下，单分子在外加电位的影响下进入纳米孔，在孔中结合或反应，从而引起电流变化。通过分析电流在幅度、持续时间、频率的变化，纳米孔被用于检测 DNA[14]，[18]、RNA[19]、肽[20]、蛋白质[21]，[22]、代谢物和蛋白质-DNA 复合物[23]等。其中，对 DNA 测序的应用最为成熟，并在商用中取得巨大成功。事实上，牛津纳米孔科技（Oxford Nanopore Technologies, ONT）已经于 2014 年开发出了商用测序仪 MinION[24]。目前，纳米孔测序技术已被广泛地应用于 DNA 存储系统，以实现海量分子数据的精确访问和读取。

本综述将对以下方面进行系统性阐述：（1）纳米孔检测技术的发展；（2）纳米孔检测技术的应用；（3）基于纳米孔检测的 DNA 存储；（4）机器学习在纳米孔检测的应用。我们希望通过本文的综述，推进纳米孔检测技术和 DNA 存储技术的发展。

2 纳米孔检测技术的发展

纳米孔检测技术是一项起源于 1976 年的电化学单分子检测技术[11]，通过实时检测纳米孔的电流，可以分析出孔内分子的功能、结构、动力学等信息。相对于其他单分子检测技术，纳米孔检测可以在不破坏原有分子结构的情况下无标记地检测分子[26]，并且十分便携（ONT 的 MinION 测序仪仅 90g）[24]。

2.1 纳米孔检测的一般原理

纳米孔检测仪的工作原理类似于 Coulter 计数器，单个带电分子在电场力的作用下通过镶嵌在膜上的纳米孔，产生瞬时阻断电流[27]。通常情况下，纳米孔蛋白被固定在具有电阻性的聚合物膜中，并被浸泡在电解质溶液中。当施加恒定电压时，电解质离子会通过纳米孔并在膜两侧产生电流。带负电荷的单链 DNA 或 RNA 分子会从负极一侧（cis）向正极一侧（trans）经过纳米孔。转移速度由一种马达蛋白（motor protein）控制，它会以一定的方式使核酸分子通过纳米孔[29]。核酸分子的转移过程中，离子电流的变化与纳米孔中存在核苷酸序列相对应，因此可以使用特定的算法将离子电流解码成核苷酸序列，从而实现单个分子的实时测序。除了控制转移速度以外，马达蛋白还具有解旋酶的作用，能够使双链 DNA（double-stranded DNA，dsDNA）或 RNA-DNA 双链解旋成单链分子通过纳米孔。

通过多个纳米孔组成阵列，可以实现大规模测序。以 ONT 的 MinION 为例（图 1 所示），每一个纳米孔集成于微支架（microscaffold），多个微支架组成微支架阵列（array of microscaffolds），可使多个纳米孔在运输和使用过程中保持稳定。每个微支架与其自身电极

相对应，该电极连接至传感器阵列芯片（sensor chip）中的通道上。每个纳米孔通道均由定制的专用集成电路（application-specific integrated circuit，ASIC）单独控制和测量，支持同时进行多个纳米孔实验。

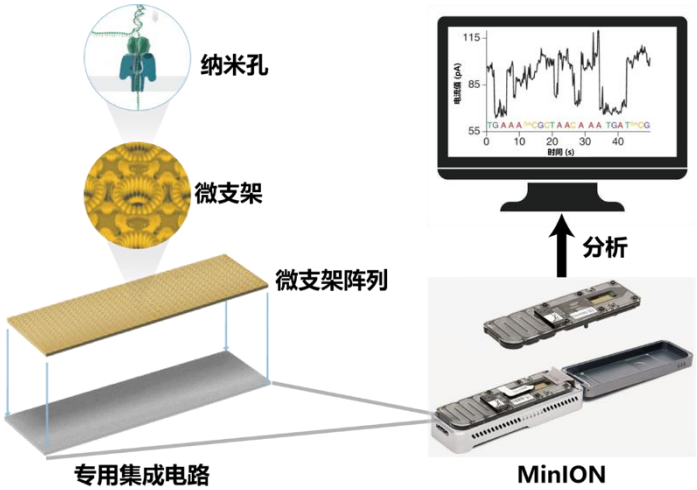


图 **1**　纳米孔测序原理[28](Copyright © 2008 - 2024 Oxford Nanopore Technologies plc.)

Fig. 1　Principle of nanopore sequencing

2.2 纳米孔的种类

目前，纳米孔主要有两大类：生物纳米孔和固态纳米孔。生物纳米孔由蛋白质亚基、肽在脂质双分子层或多聚物膜中自组装形成，能够识别直径约为 1~10 纳米的分子[11]，[12]，[26]。固态纳米孔是在超薄的人造膜（如 SiN_x）中制造而成，这使得纳米孔的直径可以扩大至数百纳米，因此可以分析大型生物分子和复合物。用于制造固态纳米孔的技术，如电子/离子铣削[30]，[31]和激光蚀刻[32]，[33]，可以在纳米尺度上操纵纳米孔的大小，但是相对于生物纳米孔，固态纳米孔只能操纵表面结构。然而生物纳米孔的化学修饰可以进一步增强纳米孔与特定分子之间的相互作用，提高整体灵敏度和选择性，使得纳米孔可以在溶液中可控地捕获、识别、运输各种分子[12]，[34]，[36]。但生物纳米孔也有缺点：缺乏调整蛋白质孔径和几何结构的自由，以及在一定的 pH 值、温度和离子浓度下才具有有限的稳定性。

生物纳米孔主要有 α-HL（α-hemolysin）[37]、MspA（*Mycobacterium smegmatis* porin A）[38]、Ael（Aerolysin）[39]、phi29 motor[40]、ClyA（Cytolysin A）[41]、OmpG（Outer membrane protein G）[42]等。就孔径（蛋白质孔隙在天然形态下最窄收缩直径）而言，α-HL（~1.4 nm）、MspA（~1.2 nm）、OmpG（~1.3 nm）和 Ael（~1.0 nm）适用于单链 DNA 的研究，但是不适用于较大的分析物，如双链 DNA 和折叠蛋白质。而 CylA（~3.3 nm）和 phi29 motor（~3.6 nm）就足够通过双链 DNA 分子。另一个重要的因素是纳米孔的长度，这对检测灵敏度至关重要。在 DNA 链的检测中，孔中核苷酸的离子电流信号在纳米孔中移动时受邻近核苷酸的影响，如果孔道长度过长，会导致电流信号很难解码成核苷酸。例如，双链 DNA 中，两个碱基之间的距离约为 0.34 nm，在具有 5 nm 长的 β-管的 α-HL 孔中，将同时存在 15 个左右的碱基，会有 4^{15} 种的碱基组合，导致难以解码。而 MspA 只有 0.6 nm 的收缩区，是最适用于 DNA 测序的纳米孔之一。

固态纳米孔的制作材料主要有 SiN_x、SiO_2、Si-C、Al_2O_3、石墨烯等。相对于生物纳米孔，固态纳米孔在稳定性、电流噪声等方面有着显著优势，但是受限于如今的半导体工艺制造水平，固态纳米孔的制造还较为复杂和昂贵[30]，[31]。总之，每种纳米孔都有着不同的特性，应该针对不同的任务应用合适的纳米孔。

3 纳米孔检测技术的应用

纳米孔检测技术最初是为了实时检测离子和小分子[12]，[43]，随后，纳米孔检测的发展集

中于 DNA 测序[11]，[14]-[16]。然而，纳米孔的应用远不止于测序，它已经被用于研究许多生物化学系统中分子的特性。首先，纳米孔的一个关键优势在于它能够以较高速率、顺序地连续捕获大量单分子，这使得纳米孔能够在短时间内检测许多分子，从而实现实时检测。其次，纳米孔将分析物的结构和化学性质转化为可测量的离子电流信号，并可以通过一定的算法实现电流信号到单分子的映射[44]。综合以上两个特点，纳米孔检测技术可用于研究多种分子的特征，并且是无标签识别，不会影响分子的原有特征。此外，纳米孔能够识别缺乏合适信号放大标签或其信息隐藏在噪声中的单分子种类[45]。因此，纳米孔可以很好地用于精准医学所需的分子诊断应用，实现核酸、蛋白质或代谢物等分析物和其他生物标志物的识别。

3.1 核酸识别

纳米孔测序的概念出现在 20 世纪 80 年代，并通过纳米孔和相关马达蛋白的一系列技术进步得以实现[46]。α-HL 是第一个可以检测到 RNA 和 DNA 阻塞电流的纳米孔[11]。野生型 α-HL 可以识别 DNA 的四种碱基，这是迈向单核苷酸分辨率的关键一步。但是由于 α-HL 的孔道较长，无法识别复杂序列。在图 2(a)中，Lieberman 等人通过引入 phi29 DNA 聚合酶，极大减缓了 DNA 的过孔速度，从而实现了复杂序列的识别[18]。相对于之前缺乏控制的 DNA 过孔实验，马达蛋白的引入不仅减缓了过孔速度，而且减少了过孔时 DNA 链的动力学波动，从而提高了数据质量[29]。

除了通常的 DNA、RNA 测序以外，目前的纳米孔测序技术还能够检测非天然核酸。非天然核酸（xeno-nucleic acids，XNAs）是一种人工合成的核酸类似物，具有不同于天然核酸 DNA 和 RNA 的糖骨架。在图 2(b)中，Yan 等人利用 NIPSS（nanopore-induced phase shift sequencing）实现了 FANA（2-deoxy-2-fluoroarabinoic acid）的直接测序方法[35]。通过将 FANA 与 DNA 驱动链连接，他们发现利用 phi29 DNA 聚合酶的 NIPSS 可以实现 FANA 的直接测序。Yan 的工作证明了纳米孔可以清楚地检测 DNA、RNA 和 XNA。

3.2 蛋白质识别

蛋白质纳米孔测序技术相较于 DNA 测序面临更多挑战，主要由于蛋白质的三维结构复杂性、多样化的电荷分布以及更庞大的氨基酸种类。尽管如此，科研人员已经取得了一系列进展来克服这些难题。Yusko 等人利用双层固态纳米孔确定了单个蛋白质的近似形状、体积、电荷、旋转扩散系数和偶极矩[21]。此外，配体如生物素[47]、蛋白结构域[48]或抗体[49]可以直接附着在生物纳米孔上，即使在复杂溶液（如血清）中也能识别特定氨基酸或蛋白质。在图 2(c)中，Fahie 等人利用生物素修饰的 OmpG 生物纳米孔特异性识别两种抗生素抗体[50]。此外，Bell 等人还可以利用经过蛋白质特异性结合物修饰的 DNA 载体来识别蛋白质[51]。针对蛋白质线性化及电渗效应驱动的过孔问题，一种策略是借助高浓度变性缓冲液构建平台，使得蛋白质变直并产生足够的电渗动力，推动长度从 250 至 750 个氨基酸不等的蛋白质穿越 α-HL[52]。另一种创新手段是在 CytK 纳米孔内部设计带负电荷的环结构，诱导电渗流以促使蛋白质有效穿过纳米孔[53]。

近年来，科研人员们在精确识别所有 20 种天然氨基酸方面取得了重大突破。早在 2020 年，野生型 Ael 纳米孔已被证实可以识别其中 13 种氨基酸[54]。最近的两项研究进一步实现了对全部 20 种氨基酸的明确鉴定，并且还进行了短肽的序列测定。改造过的 MspA 异八聚体纳米孔不仅能鉴别所有 20 种氨基酸，还能识别四种翻译后修饰（Protein translational modifications，PTMs）[55]。另一项研究则展示了含苯丙氨酸的肽探针与葫芦脲复合物如何通过 α-HL 成功区分 20 种氨基酸[56]。

这些显著的科学进展不仅证明了纳米孔检测技术在解决蛋白质三维结构复杂性、电荷分布多样性以及氨基酸种类繁多等难题上的可行性，而且为未来提升信息存储密度开辟了新的可能。随着对所有 20 种天然氨基酸及多种 PTMs 识别能力的不断提升，我们可以预见蛋白

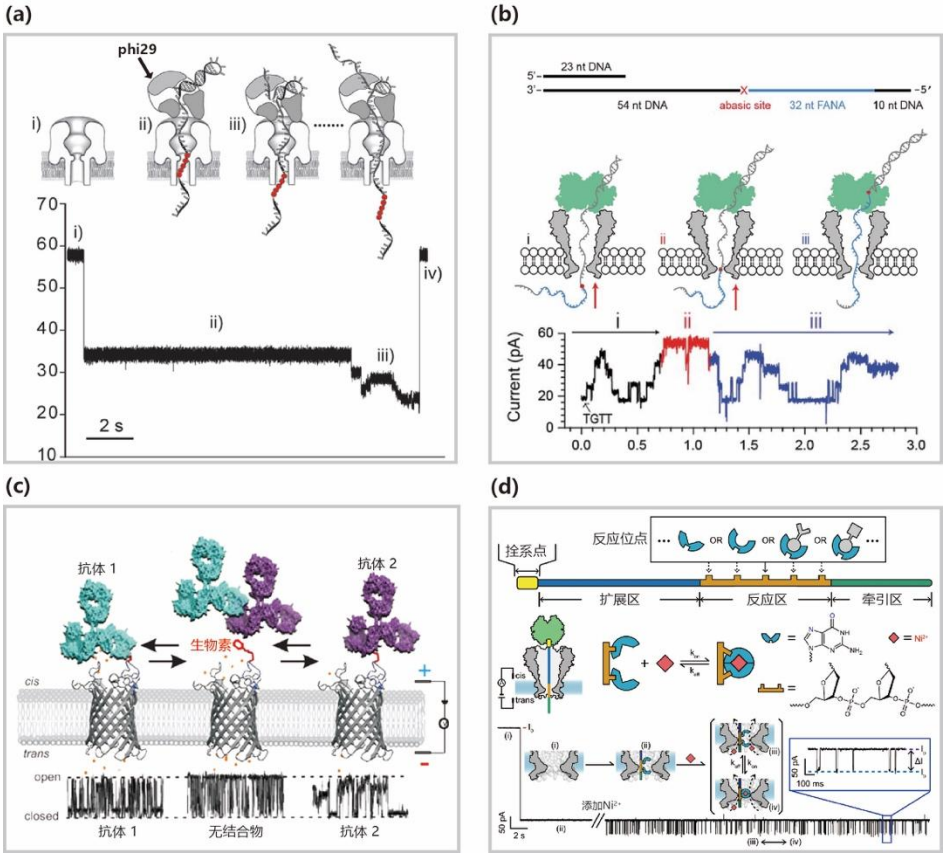
质分子将成为 DNA 存储之外的新型信息载体。理论上，蛋白质序列的多样性远超 DNA 序列，这使得每一种氨基酸组合都可能编码更多信息，从而大幅提高信息存储容量。此外，蛋白质特定的修饰状态可以进一步扩展编码空间，实现更复杂和多层次的信息存储。

3.3 无机分子识别

除了检测生物分子外，纳米孔还可用于检测无机化学分子。Jia 等人于 2021 年提出了一种基于纳米孔测序技术的可编程随机纳米传感器（programmable nano-reactors for stochastic sensing，PNRSS），可以在单分子水平上实时监测化学反应[36]。通过 PNRSS 可以直接观察到金属离子、简单有机化合物（如乳酸）和核苷类似物等广泛的单分子反应。如图 2(d)所示，PNRSS 可以识别 Ni^{2+}离子。此外，他们利用机器学习的方法提高了 PNRSS 的传感分辨率，能够实现多种分子的同时识别。

3.3 生物标记物检测与定量

与实验室制作的样品相比，医学标本具有更大的复杂性和异质性，且需要极高的精度、特异性和敏感性。得益于纳米孔的灵敏性和可编程性，纳米孔有望作为新型的体液检测仪。由于 α-HL 纳米孔可以用于选择性检测溶液中杂交的 microRNA 分子与寡核苷酸探针，Wang 等人从纯化的肺癌患者血浆样品中成功定量检测 microRNA-155 生物标志物[57]。此外，Rozevsky 等人开发了 RT-qNP（reverse transcription quantitative nanopore sensing），一种无需纯化与扩增的 RNA 定量方法[58]。RT-qNR 在人类细胞系中准确量化了癌细胞转移相关基因 $MACC_1$ 和 $S_{100}A_4$，而且还实现了 SARS-CoV-2 的无 PCR 检测。



(a) phi29 DNA 聚合酶可以使得 DNA 的过孔速度减慢，红色片段为所测序片段，对应 iii 阶段的电流（Copyright © 2010 American Chemical Society）[18] (b) NIPSS 对 FANA 直接测序，蓝色片段是 FANA 区域，对应 iii 阶段的电流[35] (c) OmpG 纳米孔结合生物素，在血清中识别两种抗体（Copyright © 2015 American Chemical Society）[50] (d) PNRSS 分为三部分：扩展区、反应区、牵引区，通过反应区的化学结构捕获 Ni^{2+}离子，并加以识别[36]

图 2　纳米孔检测技术的应用

Fig.2　The applications of nanopore detection

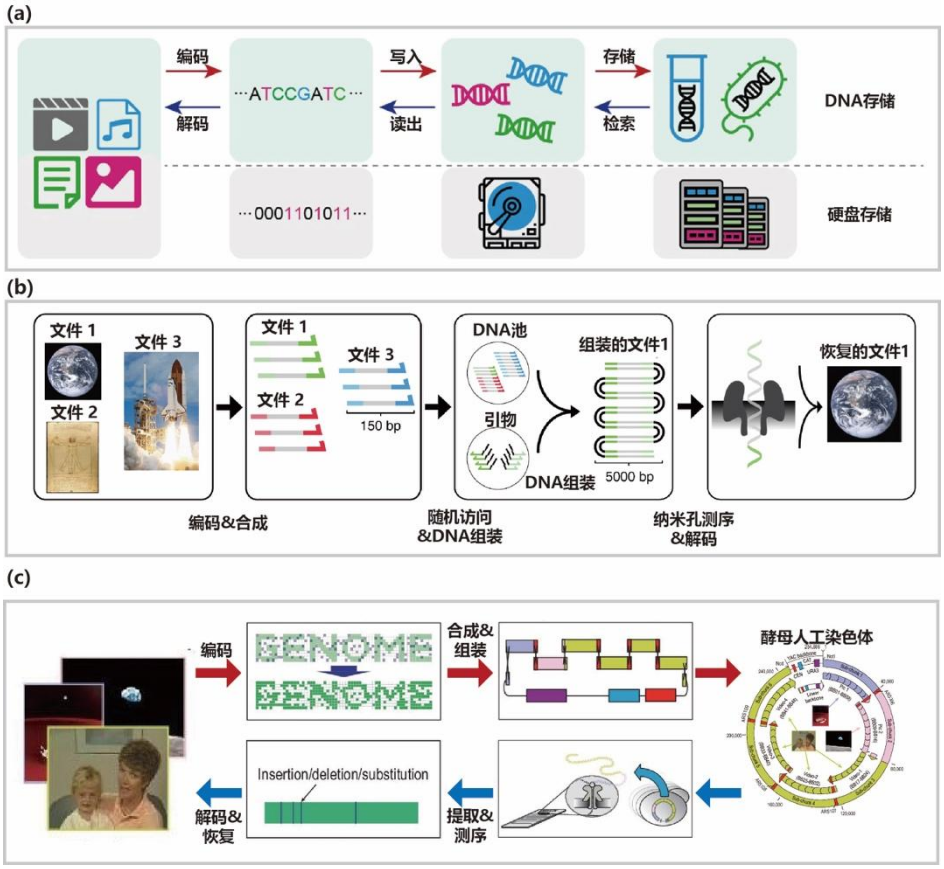
4 基于纳米孔数据读取的 DNA 存储技术

虽然携带数据信息的 DNA 链在保存过程中可能受到外界环境因素（如紫外线、极端温度变化、细菌和病毒的生物污染）的干扰和破坏，但是，DNA 具有超高的信息密度和超长的寿命，因此依然有望成为下一代新型存储介质。DNA 保护技术和保护设备的发展将大大提高 DNA 存储的存储时间和存储质量[59]- [61]。目前，纳米孔可以识别 DNA/RNA 的序列以及序列上的修饰，而序列以及修饰均可被认为是信息位，因此，DNA 存储主要分为两类：（1）基于序列的 DNA 存储；（2）基于结构/修饰的 DNA 存储。

4.1 基于序列的 DNA 存储

如图 3(a)所示，DNA 存储的流程通常分为六部分[7]：（1）编码：将数据通过特定算法编码成 DNA 序列；（2）写入：使用 DNA 合成技术合成编码信息的 DNA 链；（3）保存：在体内/体外恰当地保存 DNA；（4）检索：随机访问特定 DNA 链；（5）读取：利用测序技术读取所访问 DNA 链上的信息；（6）解码：从测序数据解码信息。目前，纳米孔检测技术被普遍用于读取步骤。

在 2018 年，Organick 等人编码并存储了 35 个不同的文件（总共 200MB），共使用了 1300 万个寡核苷酸，并且使用纳米孔测序，无错误地随机访问了每个文件[62]。此外，他们还设计了新的解码算法来最大化读出信息，从而降低无错误解码所需的测序覆盖率。2019 年，Lopez 等人使用便携式纳米孔测序设备 MinION 成功解码了存储在 DNA 中 1.6MB 的信息[63]。如图 3(b)所示，他们结合了 DNA 自组装和纳米孔测序，大大提高了纳米孔测序的吞吐量，并且这种组装技术适用于任何短 DNA 扩增子的纳米孔测序。



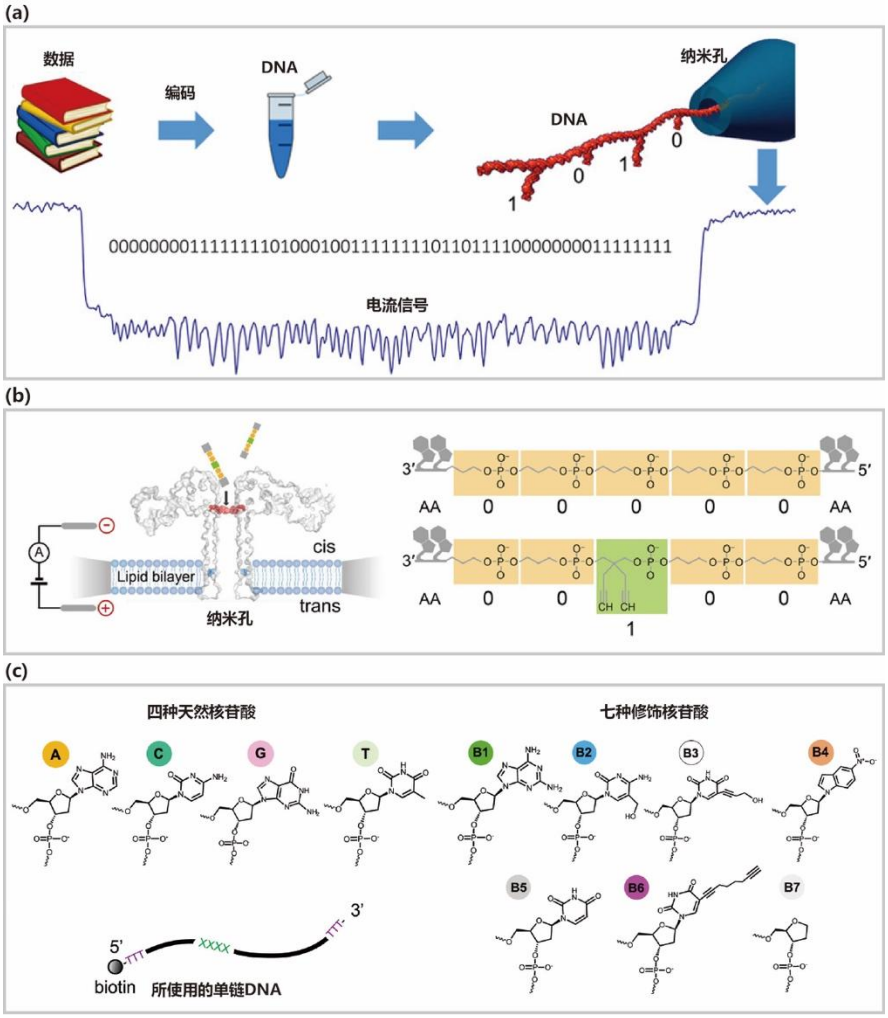
(a) DNA 存储流程与传统硬盘存储类比[7] (b) Lopez 等人的 DNA 存储流程[63] (c) 酵母人工染色体实现 DNA 存储的流程[64]

图 3　基于合成的 DNA 存储系统

Fig.3　Synthetic-based DNA storage systems

不仅仅限于 DNA 体外存储，纳米孔测序还能用于 DNA 体内存储。在图 3(c)中， Chen 等人设计并合成了一条酵母人工染色体（25.4 万个碱基），实现了 DNA 体内存储[64]。他们结合 LDPC（low-density parity-check）和伪随机序列设计编码算法，可以在一定程度上容忍测序时碱基的插入/删除错误，十分适用于纳米孔测序。通过这一编码算法，在 10.79%的错误率下，仍然实现了数据的可靠恢复。此外，他们还将该人工染色体繁衍了 100 代证明了其稳定性。2023 年，Sun 等人利用细菌基因组构建便携式 DNA 数据存储系统[65]。结合 Reed-Solomon 码和 RaptorQ 码，他们设计 MEPCAL（Mixed Error Processing Coding for Arbitrary Length）编解码算法，将 5.56KB 的压缩数据编码成了 50540 bp 的 DNA 序列。这些序列在合成以后，通过他们所开发的 RSGE（recombinase-based site-specific genome engineering）工具箱，迭代整合进大肠杆菌和蓝晶盐单胞菌的基因组中。最后利用纳米孔测序实现了无损信息恢复。从这两项工作中，可以看出体内存储的可行性。

4.2 基于结构/修饰的 DNA 存储



(a) DNA 发夹架构实现 DNA 存储，8bp 长度对应 0，16bp 对应 1（Copyright © 2018 American Chemical Society）[68] (b) 生物聚合物存储，两种不同的单体分别对应 0 和 1（黄色为 0，绿色为 1）[69] (c) 扩展 DNA 存储系统，4 种传统碱基和 7 种修饰碱基组成更大的分子字母表[70]

图 **4** 基于结构/修饰的 DNA 存储

Fig. 4 Structure/modification based DNA storage systems

纳米孔测序技术可以检测 DNA 链上的修饰以及 DNA 的二级结构，为 DNA 存储开拓了一个新方向。可编程的 DNA 纳米结构可以作为一种检索数据的方式[66]，[67]。如图 4(a)所示，2018 年 Chen 等人研发了一种利用 DNA 发夹结构（DNA hairpins）的 DNA 存储方案[68]。

在这项工作中，不同长度的 DNA 发夹被认为是不同的码元，并通过固态纳米孔识别不同的发夹结构。利用内径约为 5nm 的石英纳米孔可以识别长度为 8 bp 和 16 bp 的 DNA 发夹。因此，将 8 bp 对应比特 0，16 bp 对应比特 1，将 56 个发夹连接到 7228bp 的 DNA 链上，可以存储 56 bit 的信息。同样是利用 DNA 发夹结构，2020 年 Bell 等人基于 DNA 折纸（DNA origami）原理设计了一个 DNA 纳米结构库[51]。库中每个成员都包含了一个独一无二的条形码，其中每一位由多个 DNA 哑铃发夹结构的存在或不存在表示。通过固态纳米孔，他们实现了 3 bit 长度的条形码识别，准确率达到 94%。

除了利用 DNA 的纳米结构和传统的碱基，还有利用生物聚合物的序列的 DNA 存储技术。2020 年，Cao 等人利用专门定制的生物聚合物序列作为比特信息存储载体[69]。如图 4(b)所示，生物聚合物序列是由两种不同的单体（n-propyl phosphate 和[2,2-diynyl]-propyl phosphate）和天然核苷酸组成的生物杂交大分子，其中两种单体分别对应 0 和 1。他们利用 α-HL 纳米孔成功识别了单比特分辨率的生物聚合物。此外，使用深度学习的方法实现了 4 bit 的高精度编解码。

基于传统的四种碱基 DNA 存储系统，2022 年，Tabatabaei 等人引入了 7 种化学修饰的核苷酸，将分子字母表的大小从 4 扩大到了 11，信息密度提高了 1.73 倍[70]。如图 4(c)所示，除了 4 种传统碱基 A、C、G、T，还引入了 7 种化学修饰碱基 B1、B2……B7。在扩大分子字母表后，他们利用 MspA 纳米孔和链霉亲和素减缓过孔速度，成功识别了 4 bp（$4 \times \log_2 11 \approx 14$ bit）的信息。进一步地，他们还利用 ONT 的 GridION 测序仪，结合深度学习技术同时识别 66 种不同的信息。

4.3 其他 DNA 存储读取技术

DNA 存储的读取技术除了纳米孔测序之外，还有 Sanger 测序[71]、二代测序[72]、单分子实时测序[73]。

Sanger 测序[71]是一种基于链终止法的 DNA 测序技术。它通过在 PCR 过程中引入不同长度的链终止核苷酸，使得每个 DNA 片段具有不同的长度并在凝胶电泳中分离，从而确定碱基顺序。Sanger 测序准确率高，但速度慢、成本高且难以实现大规模并行化操作，对于大数据量的 DNA 存储应用来说效率较低。

二代测序（Next-generation sequencing, NGS）[72]，如 Illumina 公司的边合成边测序技术，利用大规模并行化的原理快速生成大量短序列数据，然后通过算法拼接成完整序列。NGS 技术具有高通量、低成本的特点，适合基因组研究和部分生物医学应用；但在 DNA 存储领域，由于其读长短、需要复杂的序列组装过程，因此在大数据的读取中存在较高的延迟，需要在算法和读取技术上继续探索这一问题的解决方案。

单分子实时测序（Single Molecule Real-Time Sequencing, SMRT）[73]，能够对单个 DNA 分子进行长时间连续观测，并记录其合成过程中的荧光信号变化来确定碱基序列，尤其擅长于读取长序列。SMRT 的显著优势在于能通过 HiFi 模式获得串联的多轮测序结果，从而对 DNA 测序达到 99.9%的准确率，且其错误模式没有明显偏好性，易于修复，这对于复杂区域的解码或 DNA 存储中连续数据的读取极其有利；然而，该技术的通量较低，读长较纳米孔测序更短，单位成本相较于其他技术更加高昂。SMRT 有望通过提高通量，继续在需要长序列分析的 DNA 存储场景下发挥作用。

5 机器学习在纳米孔数据处理的应用

近年来，机器学习（Machine Learning）在生物信号处理中发挥着巨大作用[74]。深度学习（Deep Learning）是机器学习领域中一个研究方向，通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布特征表示。深度神经网络（Deep Neural Networks，DNN）、卷积神经网络（Convolutional Neural Network，CNN）和递归神经网络（Recurrent Neural Network，RNN）等深度学习模型已经在解决诸如视觉识别（visual recognition）、语

音识别（speech recognition）和自然语言处理（natural language processing，NLP）等很多问题方面都表现出非常好的性能[75]。而纳米孔测序会产生大量的原始测序数据，因此许多针对纳米孔测序进行数据分析的深度学习模型喷涌而出，这些模型能够精确地进行测序数据的处理，如 Basecalling、错误校正、修饰碱基检测等。

5.1 Basecalling

Basecalling 是将纳米孔测序的原始信号解码成核苷酸序列，这对数据精确性和碱基修饰的检测至关重要。ONT 占有了最大多数的纳米孔测序市场，因此以下围绕 ONT 测序数据的 Basecalling 工具进行综述。Basecalling 的发展历史主要分为四个阶段[76]，[80]：（1）2016 年，利用隐马尔可夫模型（Hidden Markov Model，HMM）和递归神经网络对分割后的电流数据进行 basecalling；（2）2017 年，直接对原始电流信号进行 basecalling；（3）2018 年，使用一种 flip-flop 的模型识别单个核苷酸；(4) 2019 年，根据需求数据训练个性化 basecalling 模型。目前，ONT 开发了一系列的 basecalling 工具作为“技术演示”的软件（例如，Nanonet，Scrappie 和 Flappie），这些工具随后被投入到正式可用的软件包中（例如，Albacore 和 Guppy）。

ONT 的测序仪可以进行每秒数千次的电流测量。DNA 或 RNA 分子过孔时会有多个连续的核苷酸（即 k-mer）在孔内，导致该时刻的电流由多个核苷酸同时决定。原始电流可以根据电流的变化分割成电流信号段，每个段对应着一个 k-mer。一个电流信号段包含了多个原始信号值，这一段信号的平均值、方差和持续时间，这些数据共同构成事件（event）数据。每个事件对相邻核苷酸的依赖性是与马尔科夫链十分相似的，因此 HMM 自然被应用于早期的 basecalling，例如，ONT 的 Nanocall[81]。ONT 随后的 Nanonet 和 DeepNano[82]训练 RNN 从事件数据中预测 k-mer 从而提高 basecalling 准确率。其中，Nanonet 还是一个双向递归神经网络（bidirectional RNN），结合上下游的信息更加精准地进行 basecalling。

然而，在原始电流信号到事件的转换中，会丢失部分信息，有可能降低 basecalling 的准确性。ONT 的开源软件 Scrappie（同时用于 Albacore 和 Guppy）和第三方软件 Chiron[83]率先实现了直接将原始电流信号转换为 DNA 序列的神经网络模型。接着，ONT 发布了 Flappie，该模型使用了 flip-flop 模型直接从原始信号中识别单个碱基，实现了单碱基分辨率的 basecalling。此外，如图 5(a)所示，Causalcall 软件使用时间卷积网络（temporal convolutional network）和连接时间分类（connectionist temporal classification，CTC）解码器来识别远程序列特征，改进 basecalling[84]。与固化的 basecalling 模型相比，ONT 引入了 Taiyaki（用于 Guppy）来实现个性化 basecalling，例如针对特殊序列/物种训练个性化模型。Taiyaki 使用自然语言处理的先进技术来处理原始电流信号的高度复杂性和长期依赖性。不仅仅止于 basecalling，Taiyaki 还可以训练模型来识别修饰的碱基，例如 5mC 和 6mA。目前，ONT 的 Guppy 是最为广泛使用的 basecalling 工具，并且可以利用显卡加速运算，相较于之前的 basecalling 工具，在准确性和速度上有着无可比拟的优势[76]。ONT 在原有的 Guppy 基础上开发了新一代架构的 basecalling 软件 Dorado。相较于 Guppy，Dorado 在保持计算资源成本相对可控的基础上，显著提升了 3.8 倍的性能[96]。不仅如此，Dorado 还增强了对修饰碱基的检测能力，能够在进行 basecalling 的同时，以较小的额外计算成本实现对 5-甲基胞嘧啶（5mC）及 5-羟甲基胞嘧啶（5hmC）等修饰碱基的高灵敏度检测，从而为表观遗传学研究提供了更为强大的支持。

5.2 序列纠错

尽管纳米孔测序的平均准确率正在提高，但部分序列的准确率仍然非常低，如 ONT 的 1D 和 2D/1D^2 短序列的错误率远高于 NGS 生成的数据。因此，需要对测序结果进行序列纠错（error correction），以保证下游分析的正确性。目前主要有两种基于传统算法的序列纠错方案[85]，[86]：自我纠错（self-correction）和混合纠错（hybrid-correction）。自我纠错采用基于图的算法，在同一来源的不同分子之间产生一致性序列，与从相同分子产生的纳米孔

测序结果进行对比从而纠错（例如 Canu[87]和 LoRMA[88]）。而混合纠错利用高准确率的短序列，通过基于对齐的算法（例如 LSC[89]和 Nanocorr[90]）、基于图的算法（例如 LorDEC[91]）、基于双端对齐/图的算法（例如 HALC[92]）来纠正长序列。混合纠错工具结合足够的短读段覆盖度，能够将长读段的错误率降低到与短读段相似的水平（约为 1%-4%）[85]，[86]，而自我纠错则能将错误率降至约 3%-6%[86]，这可能是由于 ONT 数据中存在非随机的系统性误差所致。

除了传统算法，还有科研人员利用机器学习算法进行序列纠错，例如 NanoReviser[93]。Wang 等人通过 CNN 和双向长短时记忆网络（Bi-LSTM），充分利用了原始电信号及碱基识别器提供的碱基识别序列中的信息，在经过训练后 NanoReviser 能使 *E. coli* 数据集和人类数据集的测序错误率均降低了 5%以上，并且在加入甲基化注释后，NanoReviser 使 *E. coli* 数据集的错误率降低了 7%。

结合传统算法和机器学习算法，对 basecalling 的结果进一步进行序列纠错，可以明显降低序列信息的错误率，从而提高从 DNA 存储到基因组学各个领域的数据解析质量。

5.3 基因组从头组装

纳米孔的长序列测序有着较高错误率，尽管如此，随着技术的不断进步和算法优化，科研人员已经开发出多种策略来克服纳米孔测序错误率高的问题。

Canu[87]不仅依赖于传统的 overlap–layout–consensus 框架，还创新性地运用了自适应错误修正算法，在组装过程中实时识别并纠正碱基识别错误，从而显著提升了长序列数据在基因组从头组装中的应用效果。Miniasm[94]则以其极简快速的特点著称，尤其适用于小且简单的基因组从头组装，通过全对全序列映射实现超快构建初步遗传图谱，尽管初始组装准确性可能受限，但后续可通过诸如 Nanopolish[95]等工具进行 polishing 和序列质量提升。

此外，基于机器学习方法，如 ONT 公司的 Medaka，能够有效利用神经网络模型分析原始电流信号以提高碱基调用精度，并在此基础上改善组装结果。此类方法通过深度学习算法捕捉复杂的信号模式，实现了对纳米孔测序错误的智能预测和校正，从而有助于进一步降低组装误差。

5.4 DNA 和 RNA 修饰检测

纳米孔有着极强的灵敏性，在通过正常碱基和修饰碱基时，两者的电流会有细微区别，因此通过机器学习的方法可以区分两者电流，直接检测一些 DNA 和 RNA 的修饰。近年来，已经开发了一些 DNA 和 RNA 修饰检测工具。Nanorow（集成于 ONT 的 Tombo）能够从 ONT 数据检测三种 DNA 修饰：5mC，6mA 和 4mC[97]。随后，出现了许多 DNA 修饰检测工具：Nanopolish（5mC）[98]、signalAlign（5mC，5hmC 和 6mA）[99]、DeepMod（5mC 和 6mA）[100]、DeepSignal（5mC 和 6mA）[101]和 NanoMod（5mC 和 6mA）[102]。如图 5(b)和图 5(c)所示，在两次基准测试中[103]，[104]，Nanpolish，Megalodon 和 DeepSignal 在单分子水平上具有高精度的单核苷酸分辨率 5mC 检测。此外，ONT 的 Dorado 也有修饰检测的功能，在性能上远高于 Guppy[96]。Nanopore 测序技术与另一单分子测序技术 Pacbio SMRT 相比，ONT 在检测 5mC 有优势，在检测 6mA 时精度较低[80]，[98]，[105]。

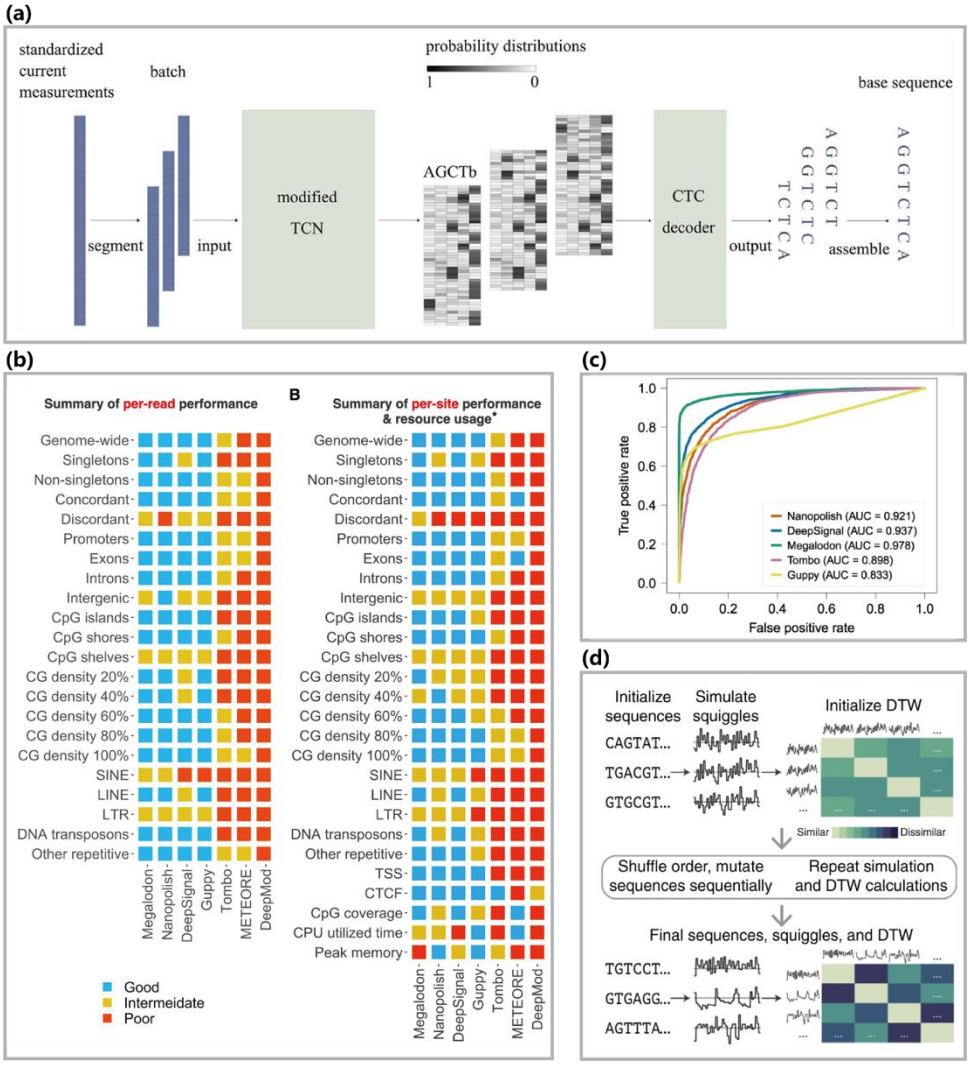
目前，ONT 的直接 RNA 修饰测序已经取得了重大进展，能够获得可靠的测序数据。检测 RNA 修饰的工具也随之出现，主要分为两种方法：（1）基于错误分布统计，例如 EpiNano（m6A）[106]和 ELIGOS[107]（m6A 和 5moU）；（2）基于电流信号，例如 Tombo（m6A 和 m5C）[97]和 MINES（m6A）[108]。但是，在单分子水平上用单核苷酸分辨率检测 RNA 修饰尚未得到证实。

5.5 DNA 存储

机器学习技术可以应用于 DNA 存储的多个方面，例如，加快 DNA 存储编码速度，高效准确地随机访问，从而促进大规模 DNA 存储系统的商业化进程。最直观的，Basecalling

最开始就是利用机器学习的技术，如 HMM、RNN、CNN，结合计算机硬件速度的发展，可以提高碱基识别的准确率和效率。此外，机器学习技术还可以扩展纳米孔可检测的分子类型，提高 DNA 存储的信息密度。

机器学习不止可以在 basecalling 加快 DNA 存储读取速度，还可以跳过 basecalling 直接分析原始信号。2020 年，Doroschak 等人提出了 Porcupine，一种终端用户分子标记系统，仅需无核酸酶水和一台 MinION 设备就可以在几秒内检测 DNA 分子标签[109]。Porcupine 分子标签由分子比特（molbits）组成，共有 96 种，因此可以编码 96 bit 的信息。为了迅速检测标签，Porcupine 直接使用纳米孔测序的原始电流信号来区分分子比特，无需 basecalling，从而加快识别速度。如图 5(d)所示，为了提高分子比特的分类准确率，Porcupine 使用 ONT 的 Scrappie 模拟分子比特的电流信号，通过遗传算法使得 96 种分子比特的电流信号尽可能正交，提高后续分子比特的识别准确率。通过预训练的卷积神经网络模型，Porcupine 成功存储了 32 bit 的分子标签（64 bit 的纠错码，可纠 9 位错），并且最快可以在 7 秒内成功识别标签，实现了实时性。Porcupine 用它的可容错性和实时性证明了 DNA 存储和纳米孔检测技术结合后潜在的商用价值。



(a) Casualcall 神经网络架构[84] (b) 修饰检测工具 per-read 水平和 per-site 水平基准测试[103] (c) 修饰检测工具 ROC 对比[104] (d) Porcupine 设计序列算法流程[109]

图 **5** 机器学习在纳米孔检测的应用

Fig. 5 The applications of machine learning in nanopore technology

机器学习还可以用于计算机模拟来扩大 DNA 存储系统的容量。通常来说，在 DNA 存

储系统中，DNA 引物（DNA primer）利用 DNA 杂交的原理实现随机访问。因此在大规模的 DNA 存储系统中，需要相当多的 DNA 引物。然而，为了降低错误访问的概率，必须减少引物序列之间的相似性，这会导致引物设计十分复杂且困难，从而严重阻碍了大规模 DNA 存储的发展。因此，精确控制和预测 DNA 杂交过程对于设计大规模 DNA 存储系统至关重要。2021 年，Buterez 提出了一种应用于预测 DNA 杂交任务的机器学习方法。他使用 nupack 生成了一个 250 万大小的杂交数据集，评估了多个机器学习模型（RNN、CNN、CNN lite、RoBERTa），发现机器学习模型能够更准确地预测 DNA 杂交，还能减少运行时间[110]。

6 总结

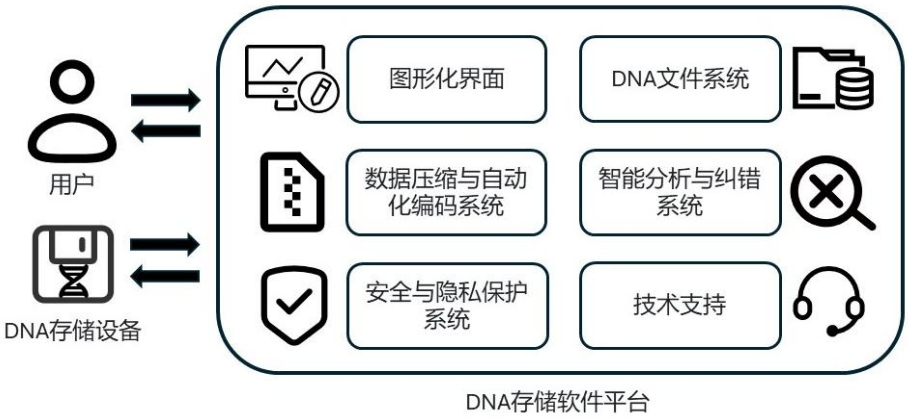
DNA 存储技术作为一项极具前瞻性和革命性的信息存储方案，其利用生物分子的特性，尤其是 DNA 分子的巨大存储容量和潜在的超长寿命，为解决未来数据爆炸式增长带来的存储问题提供了可能。然而，从实验室研究走向实际应用的过程中，仍存在诸多技术挑战。

首先，传统的测序方法，如 Sanger 测序[71]或二代测序技术（Next-generation sequencing，NGS）[72]，在速度、成本及操作便携性上无法满足 DNA 存储的实际应用需求。纳米孔检测技术在此背景下应运而生，并展现出显著优势。国内外商业巨头都积极研发并推出了具有竞争力的纳米孔测序仪产品。例如，ONT 公司推出的 MinION 便携式测序仪，通过与 Apple 公司搭载 M3 芯片的新一代 iMac 或 MacBook 等设备连接，能够实现现场实时的数据读取和分析，推动“任何人在任何地方测序”[111]。国内如齐碳科技已推出首款国产纳米孔基因测序仪 QNome-9604，这些都预示着 DNA 存储领域在全球范围内正迎来快速发展。

尽管如此，当前纳米孔测序技术在推动 DNA 存储时仍面临一些瓶颈。高通量测序对于处理大规模 DNA 存储数据至关重要，因此，虽然 MinION 在展示测序原理和便携性方面表现突出，ONT 推出的 GridION、PromethION 2/24/28 等高通量测序平台备受关注；面对海量数据的读取需求，更高通量的测序仪将发挥更大作用。此外，R9.4 版本的测序准确率已达到约 85–94%，测序速度可达 450 base/s，然而对长均聚物的测序能力仍有待提高，尤其是在信号识别上的局限性[76]，[84]。随着 R10 系列纳米孔的推出，特别是 R10.4 版本及其引入的 duplex sequencing 技术，有望提升对长片段和复杂序列的准确性，但这还需要进一步验证和完善[112]，[113]。

伴随着人工智能技术的快速发展，机器学习算法的应用正逐步颠覆纳米孔检测技术和 DNA 存储的传统框架。通过深度学习和模式识别技术，不仅可以优化生物分子检测的多样性和复杂性，而且能够开发出更为精确的 basecalling 算法，提升 DNA 存储编解码效率。更重要的是，机器学习有望通过对蛋白质结构（如 AlphaFold[114]）的解析和优化生物纳米孔设计，从源头上改进纳米孔检测的性能，从而大幅推进 DNA 存储技术的发展。DNA 存储技术尚未普及的一个主要原因是缺乏直观易用的用户接口和工具链，要求操作者具备一定的生物实验技能和数据分析能力。为了优化 DNA 存储技术，在图 6 中本综述提出一种用户友好的 DNA 存储软件平台，用户通过该平台与高度集成化的 DNA 存储设备进行交互。

展望未来十年，随着纳米孔测序技术的不断革新、机器学习在该领域的深入渗透以及更先进的用户友好型的 DNA 存储系统构建，DNA 存储技术有可能实现重大突破，真正迈向商业化应用阶段，相信在在不久的将来，我们将有机会见证全新的 DNA 存储时代的到来。



用户通过与 DNA 存储软件平台进行简单易用的交互，从而实现对 DNA 存储设备的操作。图形化界面为用户提供可视化接口；DNA 文件系统进行 DNA 数据的组织与管理；数据压缩与自动化编码系统通过高效的数据编码算法，将数据转为 DNA 存储的格式；智能分析与纠错系统集成先进的错误检测和校正机制，提高数据恢复的准确度，并提供修复建议；安全与隐私保护系统确保 DNA 存储的数据安全性；技术支持解答用户在使用过程中遇到的问题

图 **6　用户友好的 DNA 存储软件平台**

Fig. 6　A user-friendly DNA storage software platform

参考文献

[1] Reinsel D, Gantz J, Rydning J. The digital of the world from edge to core [EB/OL]. (2020-05-21) [2023-10-24]. https:/wwwseagate.com/files/www-content/our-story/trends/files/dataage-idc-report-final.pdf

[2] Williams E D, Ayres R U, Heller M. The 1.7 kilogram microchip: Energy and material use in the production of semiconductor devices [J]. Environmental science & technology, 2002, 36(24): 5504-5510.

[3] Church G M, Gao Y, Kosuri S. Next-generation digital information storage in DNA [J]. Science, 2012, 337(6102): 1628-1628.

[4] Allentoft M E, Collins M, Harker D, et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils [J]. Proceedings of the Royal Society B: Biological Sciences, 2012, 279(1748): 4724-4733.

[5] Bhat W A. Bridging data-capacity gap in big data storage [J]. Future Generation Computer Systems, 2018, 87: 538-548.

[6] Jaeger L, Chworos A. The architectonics of programmable RNA and DNA nanostructures [J]. Current opinion in structural biology, 2006, 16(4): 531-543.

[7] Hao Y, Li Q, Fan C, et al. Data storage based on DNA [J]. Small Structures, 2021, 2(2): 2000046.

[8] Merrifield B. Solid phase synthesis [J]. Science, 1986, 232(4748): 341-347.

[9] Belitsky J M, Nguyen D H, Wurtz N R, et al. Solid-phase synthesis of DNA binding polyamides on oxime resin [J]. Bioorganic & medicinal chemistry, 2002, 10(8): 2767-2774.

[10] Reuter J A, Spacek D V, Snyder M P. High-throughput sequencing technologies [J].

Molecular cell, 2015, 58(4): 586-597.

[11] Kasianowicz J J, Brandin E, Branton D, et al. Characterization of individual polynucl eotide molecules using a membrane channel [J]. Proceedings of the National Academy of Sciences, 1996, 93(24): 13770-13773.

[12] Kasianowicz J, Walker B, Krishnasastry M, et al. Genetically engineered pores as met al ion biosensors [J]. MRS Online Proceedings Library (OPL), 1993, 330: 217.

[13] Bayley H, Cremer P S. Stochastic sensors inspired by biology [J]. Nature, 2001, 413 (6852): 226-230.

[14] Clarke J, Wu H C, Jayasinghe L, et al. Continuous base identification for single-mole cule nanopore DNA sequencing [J]. Nature nanotechnology, 2009, 4(4): 265-270.

[15] Cherf G M, Lieberman K R, Rashid H, et al. Automated forward and reverse ratcheti ng of DNA in a nanopore at 5-Å precision [J]. Nature biotechnology, 2012, 30(4): 34 4-348.

[16] Manrao E A, Derrington I M, Laszlo A H, et al. Reading DNA at single-nucleotide r esolution with a mutant MspA nanopore and phi29 DNA polymerase [J]. Nature biote chnology, 2012, 30(4): 349-353.

[17] Cao C, Ying Y L, Hu Z L, et al. Discrimination of oligonucleotides of different leng ths with a wild-type aerolysin nanopore [J]. Nature nanotechnology, 2016, 11(8): 713-718.

[18] Lieberman K R, Cherf G M, Doody M J, et al. Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase [J]. Journal of the A merican Chemical Society, 2010, 132(50): 17961-17972.

[19] Wanunu M, Dadosh T, Ray V, et al. Rapid electronic detection of probe-specific micr oRNAs using thin nanopore sensors [J]. Nature nanotechnology, 2010, 5(11): 807-814.

[20] Movileanu L, Schmittschmitt J P, Scholtz J M, et al. Interactions of peptides with a protein pore [J]. Biophysical journal, 2005, 89(2): 1030-1045.

[21] Yusko E C, Bruhn B R, Eggenberger O M, et al. Real-time shape approximation and fingerprinting of single proteins using a nanopore [J]. Nature nanotechnology, 2017, 12(4): 360-367.

[22] Bell N A W, Keyser U F. Specific protein detection using designed DNA carriers and nanopores [J]. Journal of the American Chemical Society, 2015, 137(5): 2035-2041.

[23] Squires A, Atas E, Meller A. Nanopore sensing of individual transcription factors bou nd to DNA [J]. Scientific reports, 2015, 5(1): 11643.

[24] Jain M, Olsen H E, Paten B, et al. The Oxford Nanopore MinION: delivery of nano pore sequencing to the genomics community [J]. Genome biology, 2016, 17: 1-11.

[25] Neher E, Sakmann B. Single-channel currents recorded from membrane of denervated frog muscle fibres [J]. Nature, 1976, 260(5554): 799-802.

[26] Ying Y L, Cao C, Hu Y X, et al. A single biomolecule interface for advancing the s ensitivity, selectivity and accuracy of sensors [J]. National Science Review, 2018, 5(4): 450-452.

[27] Arroyo J O, Kukura P. Non-fluorescent schemes for single-molecule detection, imagin g and spectroscopy [J]. Nature Photonics, 2016, 10(1): 11-17.

[28] Nanopore Technologies. Flow cells and nanopores [EB/OL]. [2024-02-01]. https://nano poretech.com/how-it-works/flow-cells-and-nanopores.

[29] Wang Y, Zhao Y, Bollas A, et al. Nanopore sequencing technology, bioinformatics an d applications [J]. Nature biotechnology, 2021, 39(11): 1348-1365.

[30] Li J, Stein D, McMullan C, et al. Ion-beam sculpting at nanometre length scales [J]. Nature, 2001, 412(6843): 166-169.

[31] Storm A J, Chen J H, Ling X S, et al. Fabrication of solid-state nanopores with sing le-nanometre precision [J]. Nature materials, 2003, 2(8): 537-540.

[32] Gilboa T, Zrehen A, Girsault A, et al. Optically-monitored nanopore fabrication using a focused laser beam [J]. Scientific reports, 2018, 8(1): 9765.

[33] Yamazaki H, Hu R, Zhao Q, et al. Photothermally assisted thinning of silicon nitride membranes for ultrathin asymmetric nanopores [J]. ACS nano, 2018, 12(12): 12472-12 481.

[34] Meller A. Dynamics of polynucleotide transport through nanometre-scale pores [J]. Jou rnal of physics: condensed matter, 2003, 15(17): R581.

[35] Yan S, Li X, Zhang P, et al. Direct sequencing of 2′-deoxy-2′-fluoroarabinonucleic aci d (FANA) using nanopore-induced phase-shift sequencing (NIPSS) [J]. Chemical scien ce, 2019, 10(10): 3110-3117.

[36] Jia W, Hu C, Wang Y, et al. Programmable nano-reactors for stochastic sensing [J]. Nature Communications, 2021, 12(1): 5811.

[37] Song L, Hobaugh M R, Shustak C, et al. Structure of staphylococcal α-hemolysin, a heptameric transmembrane pore [J]. Science, 1996, 274(5294): 1859-1865.

[38] Niederweis M. Mycobacterial porins–new channel proteins in unique outer membranes [J]. Molecular microbiology, 2003, 49(5): 1167-1177.

[39] Parker M W, Buckley J T, Postma J P M, et al. Structure of the Aeromonas toxin pr oaerolysin in its water-soluble and membrane-channel states [J]. Nature, 1994, 367(646 0): 292-295.

[40] Valpuesta J M, Carrascosa J L. Structure of viral connectors and their function in bac teriophage assembly and DNA packaging [J]. Quarterly reviews of biophysics, 1994, 2 7(2): 107-155.

[41] Mueller M, Grauschopf U, Maier T, et al. The structure of a cytolytic α-helical toxin pore reveals its assembly mechanism [J]. Nature, 2009, 459(7247): 726-730.

[42] Subbarao G V, van den Berg B. Crystal structure of the monomeric porin OmpG [J]. Journal of molecular biology, 2006, 360(4): 750-759.

[43] Mindell J A, Zhan H, Huynh P D, et al. Reaction of diphtheria toxin channels with sulfhydryl-specific reagents: observation of chemical reactions at the single molecule l evel [J]. Proceedings of the National Academy of Sciences, 1994, 91(12): 5272-5276.

[44] Boersma A J, Bayley H. Continuous stochastic detection of amino acid enantiomers w ith a protein nanopore [J]. Angewandte Chemie, 2012, 124(38): 9744-9747.

[45] Li M Y, Ying Y L, Li S, et al. Unveiling the heterogenous dephosphorylation of DN A using an aerolysin nanopore [J]. ACS nano, 2020, 14(10): 12571-12578.

[46] Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing [J]. Nature biotechnology, 2016, 34(5): 518-524.

[47] Movileanu L, Howorka S, Braha O, et al. Detecting protein analytes that modulate tra nsmembrane movement of a polymer chain within a single protein pore [J]. Nature bi otechnology, 2000, 18(10): 1091-1095.

[48] Thakur A K, Movileanu L. Real-time measurement of protein–protein interactions at si ngle-molecule resolution using a biological nanopore [J]. Nature biotechnology, 2019, 37(1): 96-101.

[49] Wei R, Tampé R, Rant U. Stochastic sensing of proteins with receptor-modified solid-state nanopores [J]. Biophysical Journal, 2012, 102(3): 429a.

[50] Fahie M A, Yang B, Mullis M, et al. Selective detection of protein homologues in se rum using an OmpG nanopore [J]. Analytical chemistry, 2015, 87(21): 11143-11149.

[51] Bell N A W, Keyser U F. Digitally encoded DNA nanostructures for multiplexed, sin gle-molecule protein sensing with nanopores [J]. Nature nanotechnology, 2016, 11(7): 645-651.

[52] Ouldali H, Sarthak K, Ensslen T, et al. Electrical recognition of the twenty proteinoge nic amino acids using an aerolysin nanopore [J]. Nature biotechnology, 2020, 38(2): 1 76-181.

[53] Sauciuc A, Morozzo della Rocca B, Tadema M J, et al. Translocation of linearized fu ll-length proteins through an engineered nanopore under opposing electrophoretic force [J]. Nature Biotechnology, 2023: 1-7.

[54] Ouldali H, Sarthak K, Ensslen T, et al. Electrical recognition of the twenty proteinoge nic amino acids using an aerolysin nanopore [J]. Nature biotechnology, 2020, 38(2): 1 76-181.

[55] Wang K, Zhang S, Zhou X, et al. Unambiguous discrimination of all 20 proteinogeni c amino acids and their modifications by nanopore [J]. Nature Methods, 2023: 1-10.

[56] Zhang Y, Yi Y, Li Z, et al. Peptide sequencing based on host–guest interaction-assiste d nanopore sensing [J]. Nature Methods, 2024, 21(1): 102-109.

[57] Wang Y, Zheng D, Tan Q, et al. Nanopore-based detection of circulating microRNAs in lung cancer patients [J]. Nature nanotechnology, 2011, 6(10): 668-674.

[58] Rozevsky Y, Gilboa T, van Kooten X F, et al. Quantification of mRNA expression using single-molecule nanopore sensing [J]. ACS nano, 2020, 14(10): 13964-13974.

[59] Guo J, Amini S, Lei Q, et al. Robust and long-term cellular protein and enzymatic activity preservation in biomineralized mammalian cells [J]. ACS nano, 2022, 16(2): 2164-2175.

[60] Baoutina A, Bhat S, Partis L, et al. Storage stability of solutions of DNA standards [J]. Analytical chemistry, 2019, 91(19): 12268-12274.

[61] Kohll A X, Antkowiak P L, Chen W D, et al. Stabilizing synthetic DNA for long-term data storage with earth alkaline salts [J]. Chemical communications, 2020, 56(25): 3613-3616.

[62] Organick L, Ang S D, Chen Y J, et al. Random access in large-scale DNA data storage [J]. Nature biotechnology, 2018, 36(3): 242-248.

[63] Lopez R, Chen Y J, Dumas Ang S, et al. DNA assembly for nanopore data storage readout [J]. Nature communications, 2019, 10(1): 2933.

[64] Chen W, Han M, Zhou J, et al. An artificial chromosome for data storage [J]. National Science Review, 2021, 8(5): nwab028.

[65] Sun F, Dong Y, Ni M, et al. Mobile and Self‐Sustained Data Storage in an Extremophile Genomic DNA [J]. Advanced Science, 2023: 2206201.

[66] Liu H, Wang J, Song S, et al. A DNA-based system for selecting and displaying the combined result of two input variables [J]. Nature communications, 2015, 6(1): 10089.

[67] Ge Z, Gu H, Li Q, et al. Concept and development of framework nucleic acids [J]. Journal of the American Chemical Society, 2018, 140(51): 17808-17819.

[68] Chen K, Kong J, Zhu J, et al. Digital data storage using DNA nanostructures and solid-state nanopores [J]. Nano letters, 2018, 19(2): 1210-1215.

[69] Cao C, Krapp L F, Al Ouahabi A, et al. Aerolysin nanopores decode digital information stored in tailored macromolecular analytes [J]. Science Advances, 2020, 6(50): eabc2661.

[70] Tabatabaei S K, Pham B, Pan C, et al. Expanding the molecular alphabet of DNA-based data storage systems with neural network nanopore readout processing [J]. Nano letters, 2022, 22(5): 1905-1914.

[71] Sanger F. Determination of nucleotide sequences in DNA [J]. Science, 1981, 214(4526): 1205-1210.

[72] Bentley D R, Balasubramanian S, Swerdlow H P, et al. Accurate whole human genome sequencing using reversible terminator chemistry [J]. nature, 2008, 456(7218): 53-5

9.

[73] Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules [J]. Science, 2009, 323(5910): 133-138.

[74] 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.

[75] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. nature, 2015, 521(7553): 436-444.

[76] Wick R R, Judd L M, Holt K E. Performance of neural network basecalling tools for Oxford Nanopore sequencing [J]. Genome biology, 2019, 20: 1-10.

[77] Rang F J, Kloosterman W P, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy [J]. Genome biology, 2018, 19(1): 90.

[78] Magi A, Semeraro R, Mingrino A, et al. Nanopore sequencing data analysis: state of the art, applications and challenges [J]. Briefings in bioinformatics, 2018, 19(6): 1256-1272.

[79] Senol Cali D, Kim J S, Ghose S, et al. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions [J]. Briefings in bioinformatics, 2019, 20(4): 1542-1559.

[80] Amarasinghe S L, Su S, Dong X, et al. Opportunities and challenges in long-read sequencing data analysis [J]. Genome biology, 2020, 21(1): 1-16.

[81] David M, Dursi L J, Yao D, et al. Nanocall: an open source basecaller for Oxford Nanopore sequencing data [J]. Bioinformatics, 2017, 33(1): 49-55.

[82] Boža V, Brejová B, Vinař T. DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads [J]. PloS one, 2017, 12(6): e0178751.

[83] Teng H, Cao M D, Hall M B, et al. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning [J]. GigaScience, 2018, 7(5): giy037.

[84] Zeng J, Cai H, Peng H, et al. Causalcall: Nanopore basecalling using a temporal convolutional network [J]. Frontiers in Genetics, 2020, 10: 1332.

[85] Fu S, Wang A, Au K F. A comparative evaluation of hybrid error correction methods for error-prone long reads [J]. Genome biology, 2019, 20(1): 1-17.

[86] Lima L, Marchet C, Caboche S, et al. Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data [J]. Briefings in bioinformatics, 2020, 21(4): 1164-1181.

[87] Koren S, Walenz B P, Berlin K, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation [J]. Genome research, 2017, 27(5): 722-736.

[88] Salmela L, Walve R, Rivals E, et al. Accurate self-correction of errors in long reads

using de Bruijn graphs [J]. Bioinformatics, 2017, 33(6): 799-806.

[89] Au K F, Underwood J G, Lee L, et al. Improving PacBio long read accuracy by short read alignment [J]. 2012.

[90] Goodwin S, Gurtowski J, Ethe-Sayers S, et al. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome [J]. Genome research, 2015, 25(11): 1750-1756.

[91] Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction [J]. Bioinformatics, 2014, 30(24): 3506-3514.

[92] Bao E, Lan L. HALC: High throughput algorithm for long read error correction [J]. BMC bioinformatics, 2017, 18: 1-12.

[93] Wang L, Qu L, Yang L, et al. NanoReviser: an error-correction tool for nanopore sequencing based on a deep learning algorithm [J]. Frontiers in Genetics, 2020, 11: 900.

[94] Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences [J]. Bioinformatics, 2016, 32(14): 2103-2110.

[95] Loman N J, Quick J, Simpson J T. A complete bacterial genome assembled de novo using only nanopore sequencing data [J]. Nature methods, 2015, 12(8): 733-735.

[96] Dittforth S, Ozturk D, Mueller M. Benchmarking the Oxford Nanopore Technologies basecallers on AWS [EB/OL]. (2023-05-18) [2023-06-15]. https://aws.amazon.com/blogs/hpc/benchmarking-the-oxford-nanopore-technologies-basecallers-on-aws/

[97] Stoiber M, Quick J, Egan R, et al. De novo identification of DNA modifications enabled by genome-guided nanopore signal processing [J]. BioRxiv, 2016: 094672.

[98] Simpson J T, Workman R E, Zuzarte P C, et al. Detecting DNA cytosine methylation using nanopore sequencing [J]. Nature methods, 2017, 14(4): 407-410.

[99] Rand A C, Jain M, Eizenga J M, et al. Mapping DNA methylation with high-throughput nanopore sequencing [J]. Nature methods, 2017, 14(4): 411-413.

[100] Liu Q, Fang L, Yu G, et al. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data [J]. Nature communications, 2019, 10(1): 2449.

[101] Ni P, Huang N, Zhang Z, et al. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning [J]. Bioinformatics, 2019, 35(22): 4586-4595.

[102] Liu Q, Georgieva D C, Egli D, et al. NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data [J]. BMC genomics, 2019, 20(1): 31-42.

[103] Yuen Z W S, Srivastava A, Daniel R, et al. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing [J]. Nature communications, 2

021, 12(1): 3438.

[104] Liu Y, Rosikiewicz W, Pan Z, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation [J]. Genome biology, 2021, 22(1): 1-33.

[105] Fang G, Munera D, Friedman D I, et al. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing [J]. Nature biotechnology, 2012, 30(12): 1232-1239.

[106] Liu H, Begik O, Lucas M C, et al. Accurate detection of m6A RNA modifications in native RNA sequences [J]. Nature communications, 2019, 10(1): 4079.

[107] Jenjaroenpun P, Wongsurawat T, Wadley T D, et al. Decoding the epitranscriptional landscape from native RNA sequences [J]. Nucleic acids research, 2021, 49(2): e7-e7.

[108] Lorenz D A, Sathe S, Einstein J M, et al. Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution [J]. Rna, 2020, 26(1): 19-28.

[109] Doroschak K, Zhang K, Queen M, et al. Rapid and robust assembly and decoding of molecular tags with DNA-based nanopore signatures [J]. Nature communications, 2020, 11(1): 5454.

[110] Buterez D. Scaling up DNA digital data storage by efficiently predicting DNA hybridisation using deep learning [J]. Scientific Reports, 2021, 11(1): 20517.

[111] Sanghera, G. Oxford Nanopore meets Apple's M3 Silicon Chip: Hailing a new era for distributed genome analysis [EB/OL]. (2023-11-08) [2024-02-10]. https://nanoporetech.com/about-us/news/blog-oxford-nanopore-meets-apples-m3-silicon-chip-hailing-new-era-distributed-genome.

[112] Tytgat O, Gansemans Y, Weymaere J, et al. Nanopore sequencing of a forensic STR multiplex reveals loci suitable for single-contributor STR profiling [J]. Genes, 2020, 11(4): 381.

[113] Huang Y T, Liu P Y, Shih P W. Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing [J]. Genome biology, 2021, 22(1): 1-17.

[114] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021, 596(7873): 583-589.

作者投稿函需随投稿推荐评审专家三位：

推荐评审专家： 1. 欧阳颀，教授，北京大学定量生物学中心，E-mail: qi@pku.edu.cn,，电话：010-62756943；

2. 陆祖宏，教授，东南大学生物电子学国家重点实验室，E-mail: zhlu@seu.edu.cn；

3. 于军，研究员，中国科学院北京基因组研究所，E-mail: junyu@big.ac.cn。