

引文格式：

刘杨奕, 张轶, 刘凯. DNA 数据存储中信息处理技术的研究进展与挑战 [J]. 集成技术, 2024, 13(3): 25-38.

Liu YY, Zhang Y, Liu K. Research progress and challenges of data processing technology for DNA-based information storage [J]. Journal of Integration Technology, 2024, 13(3): 25-38.

DNA 数据存储中信息处理技术的研究进展与挑战

刘杨奕¹ 张轶² 刘凯^{2*}

¹(中国科学院大学材料科学与光电技术学院 北京 100049)

²(清华大学化学系稀土新材料教育部工程研究中心 北京 100084)

摘要 作为新一代信息存储介质，DNA 具有高信息密度和长期保存能力，有望解决全球数据存储介质耗竭的问题。但目前 DNA 信息存储技术的发展主要围绕信息“冷存储”开展，这使得存储过程中出现修改、更新、删除、销毁等需求时束手无策。该文从“冷存储”技术的现状出发，通过归纳总结 DNA 信息存储介质难以实现“热存储”应用的原因，解析用于信息处理功能的一系列“热存储”技术，包括加密销毁、重写再生、擦除恢复、运算记录等，详细论证 DNA 介质用作信息处理载体的可行性与有效性，分析各技术之间的关联性和挑战性，以期为 DNA 存储技术低能耗、高精度、高效率、高安全性的应用奠定基础，并推动新一代智能型信息存储介质和信息处理系统的发展。

关键词 脱氧核糖核酸；信息存储；高密度存储；信息处理；信息安全

中图分类号 TP 384；Q 523 **文献标志码** A **doi**: 10.12146/j.issn.2095-3135.20231031002

Research Progress and Challenges of Data Processing Technology for DNA-Based Information Storage

LIU Yangyi¹ ZHANG Yi² LIU Kai^{2*}

¹(College of Materials Science and Opto-Electronic Technology, University of Chinese Academy of Sciences, Beijing 100049, China)

²(Engineering Research Center of Advanced Rare Earth Materials (Ministry of Education), Department of Chemistry, Tsinghua University, Beijing 100084, China)

*Corresponding Author: kailiu@tsinghua.edu.cn

Abstract DNA molecules exhibit highly promising properties of high storage density and extended lifespan as a medium for next generation digital data storage. Thus, it is expected to act as an alternative to address the global issue of insufficient data storage materials. However, the current advances of DNA-based information storage are mainly focused on “the cold storage of information”, so that it is difficult to realize quick data

收稿日期：2023-10-31 修回日期：2023-12-27

基金项目：国家重点研发计划项目(2018YFA0902600, 2021YFF1200300)

作者简介：刘杨奕，助理研究员，研究方向为 DNA 信息编辑与活细胞存储技术；张轶，助理研究员，研究方向为 DNA 信息编码算法与高稳定 DNA 信息保存技术；刘凯(通讯作者)，教授，研究方向为生物分子信息材料及存储应用，E-mail: kailiu@tsinghua.edu.cn.

processing in DNA, such as rewriting, updating, deleting, and erasure. Based on the current situation of “cold storage” technology, the reasons why DNA information storage media is difficult to achieve “hot storage” applications has been summarized. In that context, the data processing including destruction, encryption, rewriting, regeneration, decay, recovery, and arithmetic recording can be realized. The feasibility of using DNA media as an information processing carrier has been comprehensively demonstrated, and corresponding advantages and disadvantages were emphasized. This review aims to highlight the importance of DNA storage technology with the potential low energy consumption, high accuracy, high efficiency, and high security. Additionally, it will show the perspective regarding the integration of DNA characteristics for the next generation of intelligent information storage and processing systems.

Keywords DNA; information storage; high density storage; information processing; information security

Funding This work is supported by National Key Research and Development Program of China (2018YFA0902600, 2021YFF1200300)

1 引 言

全球数据的指数级增长已经超过了当前技术的存储容量，因此需要创新现有的存储策略。DNA 作为一种保存遗传信息的天然介质，在存储信息时具有超高存储物理密度和超长保存期限等优点，是极具前景的下一代数字信息存储材料^[1]。30 年来，DNA 信息存储技术的发展十分迅速。2001 年，Bancroft 等^[2]首次实现了信息向 DNA 序列的编码，并将《双城记》开篇的两句文本通过聚合酶链式反应进行了复制。10 年之后，信息存储技术的概念被进一步细化。2011 年，Roy 等^[3]首次提出基于 DNA 碱基的编解码算法概念，而 Church 等^[4]则于 2012 年利用新一代 DNA 合成和测序技术将一本含有 53 426 个单词和 11 个图片的书籍编码至 DNA 序列中，正式宣告基于 DNA 的新一代信息存储介质的诞生。随着信息技术和生物技术的融合发展，2018 年，Organick 等^[5]开发了类似于现阶段物理存储芯片的 DNA 信息存储系统，可以实现 200 MB 信息的随机读取，标志着 DNA 已发展为可用于数字信息稳定、可靠存储的材料。目前，将信息写

入 DNA 并完成信息读取的整体流程包括以下几个步骤：首先，利用编码算法将数字信息转换为二进制数据，并将二进制数据编码为核苷酸序列^[6-7]；其次，DNA 通过人工合成储存在体外或体内^[8-10]；最后，存储的信息通过 DNA 测序后解码为原始文件^[11-12]。此外，一些科研人员还开发了 DNA 分子存储像素或矢量数据的信息存储技术，这一存储方式不涉及二进制转码，直接由像素点信息映射至 DNA 序列^[13]。

上述的信息存储过程充分利用 DNA 分子的信息编码能力，与传统信息存储介质相比，这种信息存储方法所需的维护成本很低，维护 10⁹ GB 规模数据的平均耗电功率小于 0.1 W。但前提是这些数据需要以“冷数据”的形式保存，信息不能被访问。尽管科研人员已经开发了各种基于 DNA 存储介质的信息编码和读取方法，但将信息在 DNA 中灵活处理的热存储技术仍未被完全开发。这意味着信息自写入 DNA 之后，直到被读取之前，一直处于无法被编辑的状态^[14]。在信息编码、写入、存储、检索和读取策略蓬勃兴盛的同时，能实现 DNA 介质可随机编写信息的 DNA 热存储技术的发展仍较为缓慢，主要原

因有以下几个方面。首先, 信息序列的索引技术不完善, 索引字段的设计需要极高的特异性, 检索特定文件所需的聚合酶链式反应引物序列需要与数据本身分开存储, 这使得信息随机读取和编辑的难度大、效率低^[15]。其次, 信息序列在随机读取和备份扩增的过程中依赖于体内外高保真 DNA 聚合酶对信息存储序列的高浓度扩增, 这一过程极大地增加了 DNA 序列突变和信息失效的可能性, 使 DNA 作为存储介质存在安全隐患^[16]。最后, DNA 在存储信息的过程中, 仍需要遵守 DNA 合成和测序技术的基本原则, 在利用生物体大分子相互作用实现信息处理的过程中, 也需要遵守生命遗传规律, 因此, 在设计编解码算法和信息识别工具时, 对技术要求很高, 极大地限制了数据处理技术的发展^[17]。综上所述, 针对 DNA 存储的信息处理技术发展缓慢的问题, 以及为解决存储中信息丢失、拷贝错误、读取失败、数据泄露、内容无法更新等难题, 需要开发相应的热存储技术, 以弥补 DNA 仅能作为冷存储介质应用的缺陷。

在基因工程技术、合成生物学技术、人工智能、DNA 纳米技术等学科的发展下, DNA 数据存储领域在过去 30 年间取得了显著进展, 并且可以预见的是, DNA 信息存储与处理技术仍有不断探索的空间。本综述将介绍几种 DNA 信息存储领域中正在蓬勃发展的信息处理技术, 从信息加密、再生、恢复、运算等角度阐述前沿技术对 DNA 信息存储技术发展的支撑作用, 使 DNA

作为信息存储介质在低耗能、高信息携带能力的基础上, 具有更高的耐用性和更广的适用性。本文旨在为今后 DNA 信息存储技术的发展提供新思路, 将 DNA 与物理存储介质对标, 探讨 DNA 由新型数字信息存储介质发展为在线数据处理介质的可行性, 讨论 DNA 数字信息存储中的信息处理技术的关键挑战和机遇。

2 信息在 DNA 中的稳定存储

信息的稳定存储是数据处理的基础。由于研究领域尚未发展出统一的编解码算法和标准的存储方式, 因此, 信息在 DNA 中的存储形式因应用角度的不同而各有不同(图 1)。DNA 在动植物体内具有良好保存的能力, 特别是在骨骼等控制良好的环境中, 具有更高的稳定性, 储存时间可长达数十万年^[18]。但在体外存储的条件下, DNA 溶液或干粉的半衰期仅为数月至数年^[19-20]。目前, DNA 的长期储存需要定制的保护材料, 而在成型的 DNA 信息存储材料中, DNA 载量仅为百分之几, 无法充分发挥 DNA 的信息携带能力^[21]。而随着 DNA 存储兼并算法的开发^[22]和非天然碱基基元的扩充^[23], 信息在 DNA 分子中的实际存储密度已经远超 DNA 的理论编码密度(香农容量 2.0 bits/nt)。尽管通过改进编码算法或降低逻辑冗余可以提升 DNA 的信息携带能力, 但在很大程度上会损失信息存储的保真度^[24]。进一步考虑 DNA 合成的成本, 可以通过平衡 DNA 冗余度与



图 1 DNA 信息存储方法与信息处理技术

Fig. 1 DNA information storage method and information processing technology

存储读取错误率之间的关系,降低编码密度,进而控制 DNA 存储的成本^[25]。由此可见,为了信息在 DNA 中的稳定存储,科研人员已经在存储的安全性、密度、成本等方面进行了衡量。为了进一步发展信息处理技术,需要对现存的 DNA 信息存储平台进行分析,通过比较体外存储和体内存储的优势和劣势,充分探讨存储形式对信息保存、信息编码、信息处理的潜在影响。

2.1 体外存储

以“DNA 硬盘”形式存储在细胞外的数字信息需要考虑 DNA 序列容易受到环境因素(如温度、湿度、紫外线等)的影响而发生突变或降解等问题^[26-27]。为了解决这一问题,研究者开发了各类 DNA 保护技术,如 DNA 甲基化^[23]、DNA 矿化^[28-29]等。一方面,这些方法提高了 DNA 的稳定性,保证 DNA 不被环境中的核酸酶降解;另一方面,使得含有信息的 DNA 材料具有极强的形状可塑性,可以与硅或其他保护介质结合,制作成 DNA 信息存储微流控液滴^[30]、信息存储芯片^[31-32]、信息存储纳米颗粒^[33-34]、纳米折纸^[35-36]等(图 2(a)~(c))。这一类结合纳米技术制备的 DNA 信息存储材料不仅可以极大地提升 DNA 信息长期“冷存储”的稳定性,还可以通过表面修饰或者纳米材料的理化性质,使用机械设备定量检测材料的光、电、磁信号,对携带信息进行识别筛选^[15],对缩略信息进行提取^[37],对完整信息进行简便高通量读取^[38-39]和重复读取^[40-41],实现简易的信息处理应用。进一步地,二氧化硅封装 DNA 信息也可以嵌入日常使用的物品中,随着物品的保存和运输而隐秘地存储,在物理层面实现信息的隐写,提高了信息存储的安全可靠性^[42]。另外,在体外存储时,无须考虑信息拷贝数等 DNA 生物遗传规律,可以设计多碱基简并算法或多序列简并算法,减少序列冗余度,提高信息写入读取效率^[22,43]。体外存储对算法和序列长度的包容度极大地促进了高通量信息的高密度

存储^[44]。

2.2 体内存储

以“DNA 光盘”或“DNA 磁带”形式存储在活细胞中的数字信息可以利用细胞的复制扩增实现高保真复制,保存成本低,数据稳定性好。为了保证数字信息有逻辑地映射编码至 DNA 序列,且符合生物遗传规律随基因组的复制而复制,科研人员针对体内信息存储开发了特异的编解码算法和长链 DNA 的合成方法。Song 等^[45]根据德布鲁因图开发了一种高密度的活细胞信息存储算法,存储物理密度可达 265 PB/g。而利用计算机信道编码逻辑发展的 DNA 信息编解码算法则更适合多模态信息的活细胞存储,如利用 Base64 算法和霍夫曼算法开发的数字信息质粒存储系统可以实现对文本、表格和图片信息的高密度压缩存储,并适用于活细胞复制和正常生理代谢(图 2(d))^[46-47]。另外,由于活细胞中可用的生物大分子工具丰富,因此,可以通过外源信号刺激直接将数字信息标记在细胞内的 DNA 上(图 2(e))^[48]。这种将数字信息直接写入到活细胞 DNA 的信息存储方法脱离了传统意义上的 DNA 信息存储模式,直接建立了外界环境和胞内 DNA 分子的操纵关系,为信息在 DNA 中的高效处理奠定了基石。

3 信息在 DNA 中的高效处理

作为一种具有高密度和高准确性的信息存储方式, DNA 存储的冷存储功能和应用已被广泛研究,亟需开发用于信息处理的热存储技术。为了对 DNA 数字信息存储中的信息处理技术进行充分讨论,本文根据信息处理的目的划分了信息的加密销毁、重写再生、擦除恢复和运算记录 4 类技术。其中,“加密销毁”是使非目标信息接收人无法破解、使用、备份存储信息 DNA 的技术;“重写再生”是使已存储信息的 DNA 修

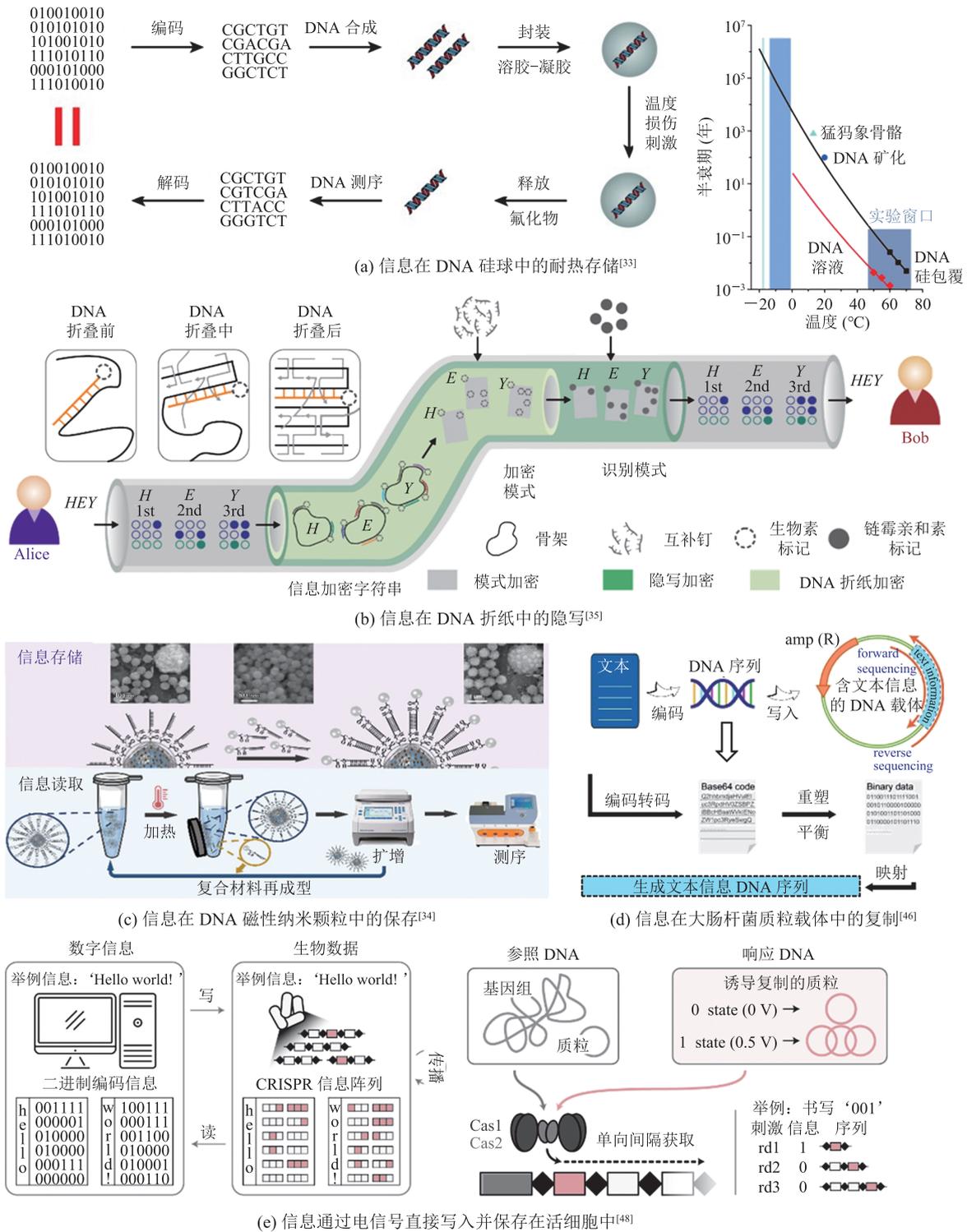


图 2 信息在 DNA 中稳定存储的形式与应用

Fig. 2 Forms and applications of stable storage of information in DNA

改或写入新信息的技术;“擦除恢复”是使信息损坏后重新进行准确读取的技术;“运算记

录”是利用 DNA 分子进行逻辑运算,并记录处理结果的技术。上述 4 种子类在基础技术上具有

一定的相似性,但各类技术的应用场景不同,对于 DNA 作为信息存储介质的实际使用来说,有多层次的指导作用。另外,不同技术面临着不同的信息安全性问题,这是实现 DNA 信息存储从“冷存储”到“热存储”的关键核心问题。接下来将阐述信息在 DNA 中高效处理的基本原则,并分析其对信息存储技术发展的影响。具体分析讨论如下。

3.1 加密销毁

信息存储技术在保证 DNA 存储的高密度和高准确性的同时,还应该考虑信息安全性的问题。为确保信息的安全性,首先需要存储的数字信息进行加密或销毁处理。这种处理方式可防止未经授权的人员获取敏感信息,同时保证信息在存储或传输的过程中不被篡改或泄露。加密技术需要绝对的信息高保真性,而销毁技术需要绝对的信息破坏不可修复性,两种技术相辅相成,可确保 DNA 信息存储技术的安全性。

在现代信息加密技术的基础上,科研人员建立了一系列针对 DNA 信息存储的数据加密算法^[49]。此外,为了应对 DNA 序列可能发生的突变或降解情况, DNA 数据的加密算法还需要相应的保护机制,以保证数据的完整性和安全性^[50]。另外,在将信息编码并写入生物体的基因组中时,可以利用细胞膜这一生理屏障保护并加密信息^[51]。而生物的基因组携带有天然的基因信息,通过测序读取的方法,很难分辨基因信息和数字信息,因此利用活细胞进行 DNA 信息存储,保密性极高^[52]。2020 年,Grass 等^[53]发现遗传短串联重复序列包含足够的冗余度来生成强加密密钥,并且只需要 DNA 测序,即可同时读取密钥和数据。2023 年,Zheng 等^[54]发现可以利用核酸分子的结构性质,通过焓熵协同调节信息加密(图 3(a))。除了利用天然生物大分子工具外,2021 年,Fan 等^[55]还开发了一种镜像核酸加密信息的方法,使信息只能通过镜像聚合酶进行复制,扩展了 DNA 信

息存储的形式和基元。多种信息加密方法保证 DNA 存储信息在编码、写入、读取等多阶段可以被安全处理。

除了数据加密技术外,为了保证数据的安全性,科研人员还开发了将特定信息完全销毁的生物工具。2018 年,Li 等^[56]利用常间回文重复序列从集关联蛋白系统(clustered regularly interspaced palindromic repeats/CRISPR-associated proteins system, CRISPR/Cas)的序列识别作用,将混合了真假信息的 DNA 序列进行酶反应筛选,利用混淆信息传递保护正确信息。2020 年, Kim 等^[57]通过化学标记,使得真假信息可以被温度区分,并且在 95 °C 只用 5 min 即可永久删除信息,为提升 DNA 信息存储的安全性提供了新思路(图 3(b))。

综上,研究人员在算法和核酸序列操纵两个维度挖掘信息加密和销毁的新办法,确保信息在处理过程中不被篡改或泄露,这些方法的应用不仅保证了 DNA 存储的高密度和高准确性,同时也确保了信息的安全性,增强了 DNA 存储技术应用的可靠性。因此,发展 DNA 信息加密与销毁技术不仅是 DNA 信息存储和处理技术中的关键研究方向,更是保证信息安全的关键步骤。

3.2 重写再生

在 DNA 存储中,信息的重写再生是实现高效处理的关键环节之一。通过重写再生的方法,原始信息可以被随机访问,信息的拥有者可以修改存储的信息或添加新的信息。这一过程需要严格控制误差,确保信息的准确性和可重复性。为了实现已存储信息的随机访问和重写,研究人员在信息再生技术上进行了不断拓展。

针对海量“冷存储”数据,在改写信息时需要通过计算机检索索引序列。2015 年,Tabatabaei Yazdi 等^[58]通过重叠聚合酶链式反应的方法,将新信息的编码序列覆盖原序列,获得含有少量非目标信息的新数据库,此方法在不影

响原始数据的基础上提升了 DNA 携带信息的能力。2017 年, Chandrasekaran 等^[59]则根据 DNA 互补配对的能量差异, 通过序列可逆的互补配对规则, 完成了原始信息的寻址和重写, 这种信息重写方法也是开发 DNA 逻辑门运算和记录的重要思路。同年, Shipman 等^[13]在活细胞中利用基因编辑 CRISPR 序列的重复间隔短片段, 通过基因编辑工具的靶向识别, 实现了信息在大肠杆菌中的修改和再生。2022 年, Liu 等^[60]则利用 CRISPR/Cas 工具, 针对完整的信息进行活细胞写入和精准编辑, 实现了多模态信息在活细胞内的改写再生(图 3(c))。相似地, 2023 年,

Sadremomtaz 等^[61]利用 CRISPR/Cas 工具将体外存储的长 DNA 片段进行碱基替换, 并开发了一种信息存储模块置换的存储方式, 使得信息在 DNA 中的存储顺序可以被重新设计(图 3(d))。

另外, 为了解决信息在 DNA 中写入和读取速度较慢的问题, 近年来, 科研人员开发了一些新的信息再生技术, 依靠此类技术, 不再需要人工合成 DNA 序列, 可以直接在细胞中完成信息存储, 再利用生物芯片或纳米技术进行快速、高效的 DNA 序列读取。2022 年, Liu 等^[62]通过建立 DNA 数据库和数据检索系统, 使用户对存储在 DNA 中的数据进行更方便的查找和使用, 并

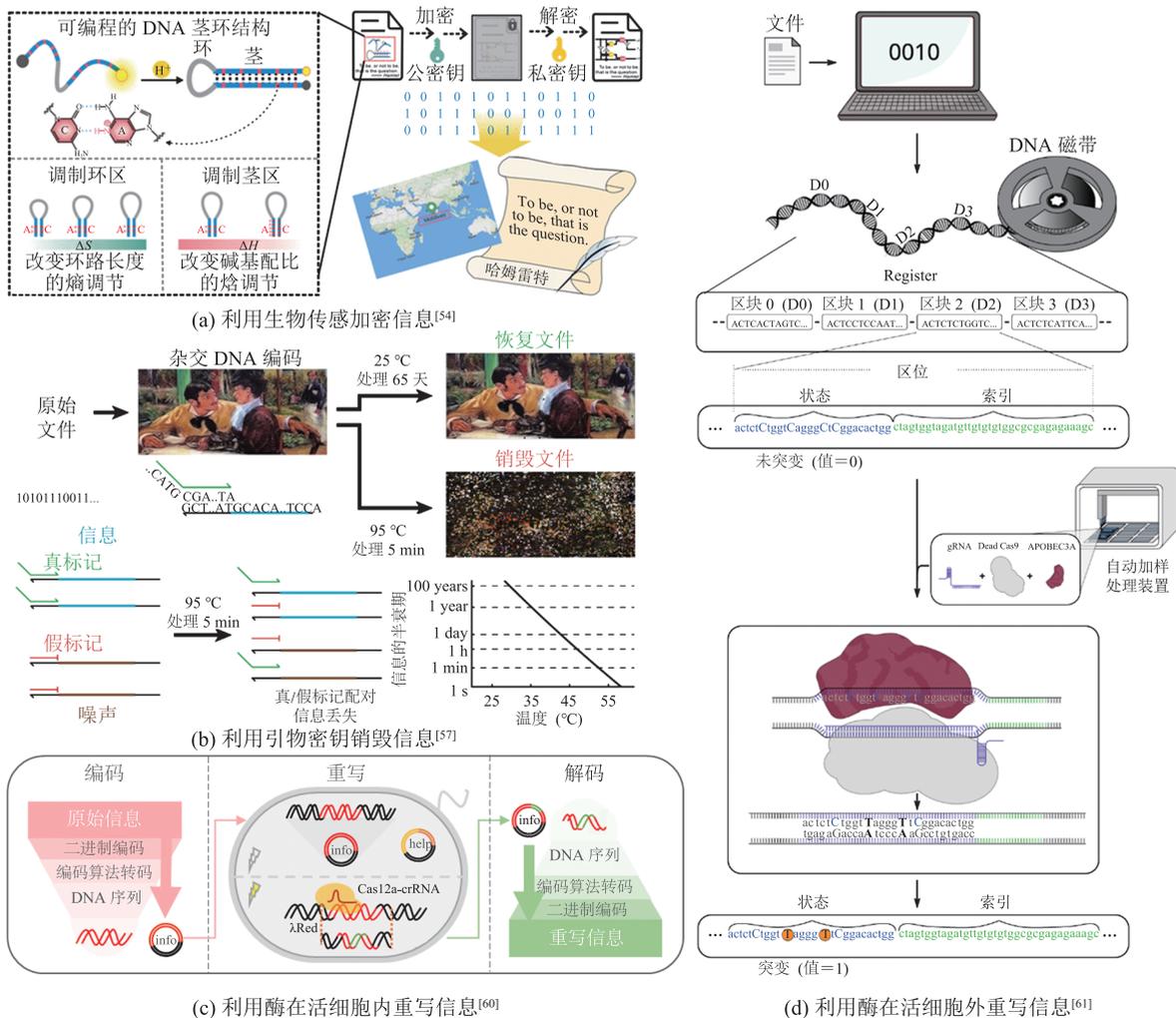


图 3 用于信息加密、销毁、重写的 DNA 信息处理技术

Fig. 3 The DNA based information processing techniques for information encryption, damage and regeneration

基于已有数据搭建新数据库。2023 年, Xu 等^[63]设计了 DNA 区块发夹结构, 通过引物串联, 实现特征寡核苷酸的有序排列, 进而建立了以短链 DNA 结构为信息基元的“活字印刷”技术。同样的, 2023 年, Yan 等^[64]利用 30 nt 以内的寡核苷酸链存储信息, 再通过桥连互补与连接酶延长 DNA 链, 生成高密度、长序列的信息存储 DNA, 该信息存储链可以直接使用纳米孔测序读取。

随着生物技术的不断发展, 未来的重写再生技术将成为 DNA 信息存储介质用于信息处理的重要环节, 信息处理过程将更加精确和高效, 为开发 DNA 大数据信息热存储技术奠定基础。

3.3 擦除恢复

除上述两种信息处理方法外, 在特定场景中, 存储在 DNA 中的数据也存在需要删除或恢复的情况。擦除恢复技术可用于暂时删除 DNA 序列中的信息, 并在需要时再对存储的原始信息进行恢复, 以便于在必要时保护敏感信息; 也可用于恢复由存储不当而损坏的信息, 弥补数据长期存储过程中的意外丢失, 提高 DNA 信息存储技术的可靠性。该技术同样需要高保真性和安全性, 以确保信息处理过程的严密可靠。

在编码算法方面, 为了将任意数字文件存储在 DNA 中, 其二进制数据使用纠错码转换为 DNA 序列, 可以充分规避 DNA 在合成、存储和测序中必然会出现的错误^[65-66]。另外, 针对 DNA 存储中数据可能出现丢失的情况, 2022 年, Pan 等^[67]设计了一种双向 DNA 存储算法, 可以通过机器学习根据剩余的 DNA 序列反向推理缺失的序列, 恢复丢失的信息。同年, Meiser 等^[68]则从连接酶、聚合酶的角度出发, 筛选用于信息存储高修复能力的复合酶(图 4(a)~(b))。长时间存储的信息如果不能通过标准解码算法读取, 那么基于信息恢复技术仍能保证存储在 DNA 中的信息被有效恢复。随着生物合成技术和分子生物

学技术的发展, 未来的擦除恢复技术可能会更加精确和高效, 如在短时间(毫秒级)内完成信息的删除或修改操作, 同时保证信息的高保真性和安全性。

3.4 运算记录

运算记录技术可用于跟踪和处理 DNA 序列中的信息, 并发展 DNA 信息存储介质的逻辑运算功能。科研人员通过利用外界工具与 DNA 分子之间的相互作用, 将信息以人工信号的方式经 DNA 逻辑门信号记录在 DNA 序列中。这一信息处理技术将 DNA 信息存储技术的应用形式进行了拓展, 无须利用 DNA 合成技术即可实现如结绳记事般的简单信息处理和记录功能。首先, 可以利用 DNA 的互补配对特性完成简单的逻辑门运算, 如 2020 年, Wang 等^[69]利用 DNA 分子在体外存储数字信息, 他们发现, 脱氧核糖核苷酸单链间的碱基互补配对过程有热量的释放和吸收, 这一能量差别可以被改造为一种逻辑信号开关。2021 年, Bee 等^[70]利用 DNA 杂交技术, 开发了快速筛选相似图片的 DNA 信息快速运算处理方法(图 4(c))。随着合成生物学的发展, 蛋白与 DNA 被开发为基因电路中的元器件, 人工编程的基因线路具有在细胞内传递信号、完成分子运算的能力, 但由于功能元件的开发仍处于初级阶段, 因此要实现媲美物理存储介质运算能力的人造基因线路仍十分困难。噬菌体重组酶可以催化 DNA 在特定基因组位点重组, 2012 年, Bonnet 等^[71]利用噬菌体重组酶驱动 DNA 序列整合重组, 实现了分子级的 DNA 运算与结果记录(图 4(d))。进一步通过多种信号响应蛋白的应用, 2017 年, Sheth 等^[72]实现了活细胞内生物活动事件的记录, 为数字信息与 DNA 信息存储的快速映射与记录建立了基本模型。而在细胞内, 可利用的核酸蛋白工具种类丰富, 2022 年, Bhattarai-Kline 等^[73]利用 CRISPR/Cas 工具在细胞中实现细胞活动的记录, CRISPR/Cas 整合酶对条

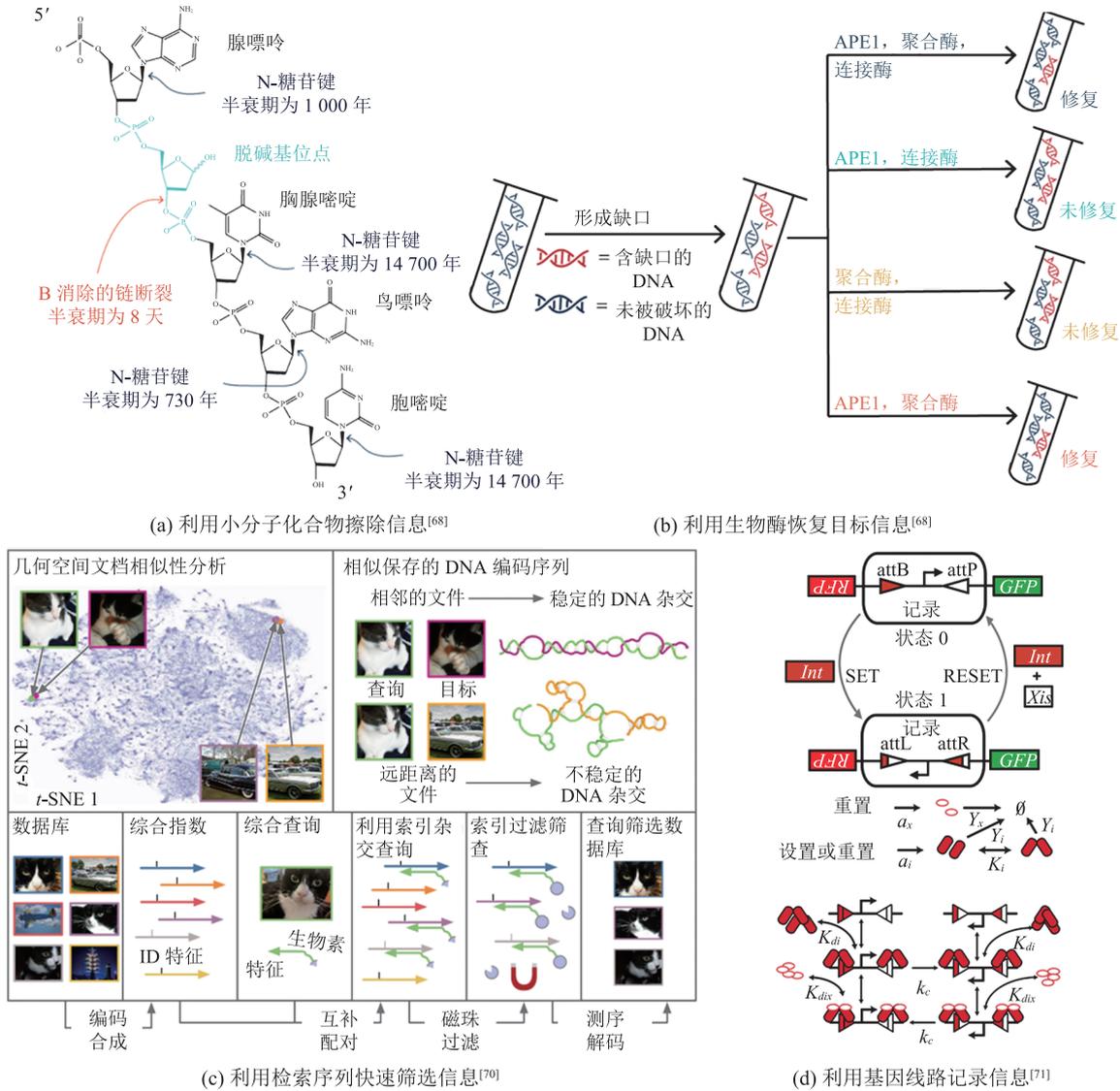


图 4 用于信息擦除、恢复、运算、记录的 DNA 信息处理技术

Fig. 4 The DNA based information processing techniques for information decay, recovery, computing and record

形码的单向整合可以通过简单的逻辑规则重建物理记录, 无须额外的信息处理过程即可从 DNA 序列直接读取出蛋白表达的顺序。2023 年, Lim 等^[74]也利用生物蛋白逻辑开关, 将图像通过光遗传电路记录在 DNA 中, 无须 DNA 序列的额外合成即可在细胞水平完成图像信息的记录。尽管本节中提及了信息的运算记录, 但这一记录过程并不等同于信息在 DNA 分子水平的存储过程, 因为本节中涉及的案例都不具备数字信息与 DNA 序列映射转化的编解码过程, 没有使用

碱基单元对记录信息进行高密度存储^[75]。不过, 随着生物技术的不断发展, 未来的运算记录技术可能会更加全面和高效, DNA 信息存储技术将发展成为能够媲美物理存储计算的高通量信息存储处理技术, 实现 DNA 信息存储的全面“热存储”应用。

4 DNA 信息存储与处理技术的瓶颈与潜力

一方面, 信息在 DNA 中的高效处理技术确

保 DNA 作为一种信息存储介质除了有着切实有效的安全保存方法外,在信息合成、写入、存储、使用、销毁的全过程中,还具有严密的数据保护方法。另一方面,结合纳米材料学和合成生物学,也证明了 DNA 是一种极具潜力的热存储材料,多种信息处理途径确保 DNA 是一种安全风险极低的信息存储材料。但目前的 DNA 信息存储与处理技术还具有一定的瓶颈和缺陷,未来的发展应该从以下几个方面进行考量。

4.1 瓶颈

为实现 DNA 存储的可操作性处理,需要解决 DNA 合成效率的问题,以及不同处理技术和存储形式之间的兼容性问题^[76]。与编解码算法和 DNA 测序技术相比,在实现 DNA 信息存储这一应用的全过程中, DNA 合成技术的发展相对滞后,这使得信息在写入和处理的步骤中受到限制。另外,由于 DNA 信息存储技术尚未形成统一的行业执行标准,因此,发展的信息处理技术无法广泛适配于所有的 DNA 信息存储平台,这阻碍了 DNA 信息存储与处理技术走向商业化应用的进程。

一方面, DNA 合成技术是信息存储和处理技术的基石,只有保证低成本、高效率、高通量地合成 DNA,才能使信息快速编码至 DNA 分子中,并充分发挥 DNA 分子的存储能力。目前的 DNA 合成技术主要有化学法和生物酶法两种。其中,化学法以固相合成寡核苷酸为基础,通过修饰的亚磷酰胺碱基将核苷酸依次合成在固相载体上,合成的 DNA 链长度在 120 个碱基以内;生物酶法基于末端转移酶的 DNA 聚合活性,将游离的脱氧核糖核苷酸在不依赖于模板的情况下进行聚合,但目前仅处于理论研究阶段,暂未大规模应用。这两种合成方法受限于化学反应和酶反应的时间,在长链 DNA 制备时,需要通过增加合成循环反应的次数实现 DNA 链的延长,经数十小时才完成一定规模 DNA 信息存储序列的

合成。与现在的物理存储方法相比, DNA 信息存储技术在写入和处理过程中都无法快速地得到编码序列,因此,基于 DNA 信息存储的信息处理技术在速度和成本上都远不及硬盘存储。所以,在发展 DNA 数字信息存储的信息处理技术时,降低 DNA 合成技术的成本,并提升合成速度,也是未来技术开发的关键。

另一方面,发展可广泛应用的 DNA 数字信息存储与处理技术需要制定行业标准。这需要开发一系列通用的、标准化的工具和方法,使得不同的实验室科研人员和公司企业使用者可以按照同样的方式来操作 DNA 存储介质。例如,可以开发通用的 DNA 合成、处理和测序试剂盒,以及通用的编解码算法等,使得不同的人员在使用不同的平台和系统时,都可以按照同样的方式来进行 DNA 存储和读取。另外,为实现 DNA 存储处理技术的广泛应用,需要制定统一的编码和存储标准,参考计算机存储信息的原理,并结合 DNA 合成、测序等技术的实际应用情况进行。例如,可以制定统一的标准来规定如何将二进制信息转换为 DNA 序列,以及如何将 DNA 序列转换回二进制信息等。此外,对于 DNA 存储的读取和合成技术来说,还需要制定相应的标准来规范操作流程,以确保数据的准确性和可靠性。

4.2 潜力

DNA 信息存储是一种具有广泛应用前景的数据存储方式,在热存储方向上的发展虽然相对滞后,但是随着新技术的不断融合和应用,以及科研人员的不懈努力, DNA 存储将会克服现有的困难,并发挥其在密度、稳定性和生物相容性等方面的优势。利用生物分子的识别作用,通过从头设计细胞的基因线路和组织的信号传导,使胞内搭载的 DNA 存储区域可以按照语言生成模型运算和生成信息,简化固态 DNA 合成速度慢、写入难度大的问题。进一步地,与蛋白质工程或材料技术相融合,构建多基元、多维度、多

功能的 DNA 信息存储与处理平台^[77-78]。

同时, 也应该注意到, 在发展 DNA 存储技术的过程中, 需要充分考虑其可持续性和对环境的影响, 确保 DNA 存储技术的健康发展。因此, 随着技术的不断进步和应用场景的不断扩大, DNA 存储将会在未来引领数据存储市场的发展, 并成为一种具有广泛应用前景的数据存储方式。尽管在信息读写效率、便携程度、应用成本上仍难以对标物理存储介质, 但随着生物技术和信息技术的融合发展, 热存储的标准化和可交互应用将大大提升 DNA 信息存储的优势, 成为现有物理存储介质的替代存储方式。随着生物技术的不断发展, 相信未来还会有更多、更有效的方法出现, 并为信息安全保驾护航。

参 考 文 献

- [1] Wang SP, Mao XH, Wang F, et al. Data storage using DNA [J]. *Advanced Materials*, 2023, 36(6): e2307499.
- [2] Bancroft C, Bowler T, Bloom B, et al. Long-term storage of information in DNA [J]. *Science*, 2001, 293(5536): 1763-1765.
- [3] Roy B, Rakshit G, Singha P, et al. An improved symmetric key cryptography with DNA based strong cipher [C] // *Proceedings of the 2011 International Conference on Devices and Communications*, 2011: 1-5.
- [4] Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA [J]. *Science*, 2012, 337(6102): 1628.
- [5] Organick L, Ang SD, Chen YJ, et al. Random access in large-scale DNA data storage [J]. *Nature Biotechnology*, 2018, 36(3): 242-248.
- [6] Bornholt J, Lopez R, Carmean DM, et al. A DNA-based archival storage system [C] // *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*, 2016: 637-649.
- [7] Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture [J]. *Science*, 2017, 355(6328): 950-954.
- [8] Lee HH, Kalhor R, Goela N, et al. Terminator-free template-independent enzymatic DNA synthesis for digital information storage [J]. *Nature Communications*, 2019, 10(1): 2383.
- [9] Lee H, Wiegand DJ, Griswold K, et al. Photon-directed multiplexed enzymatic DNA synthesis for molecular digital data storage [J]. *Nature Communications*, 2020, 11: 5246.
- [10] Xu CT, Zhao C, Ma B, et al. Uncertainties in synthetic DNA-based data storage [J]. *Nucleic Acids Research*, 2021, 49(10): 5451-5469.
- [11] Fritz MHY, Leinonen R, Cochrane G, et al. Efficient storage of high throughput DNA sequencing data using reference-based compression [J]. *Genome Research*, 2011, 21(5): 734-740.
- [12] Meiser LC, Antkowiak PL, Koch J, et al. Reading and writing digital data in DNA [J]. *Nature Protocols*, 2020, 15(1): 86-101.
- [13] Shipman SL, Nivala J, Macklis JD, et al. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria [J]. *Nature*, 2017, 547(7663): 345-349.
- [14] Bennet D, Vo-Dinh T, Zenhausern F. Current and emerging opportunities in biological medium-based computing and digital data storage [J]. *Nano Select*, 2021, 3(5): 883-902.
- [15] Banal JL, Shepherd TR, Berleant J, et al. Random access DNA memory using Boolean search in an archival file storage system [J]. *Nature Materials*, 2021, 20(9): 1272-1280.
- [16] Tomek KJ, Volkel K, Simpson A, et al. Driving the scalability of DNA-based information storage systems [J]. *ACS Synthetic Biology*, 2019, 8(6): 1241-1248.
- [17] Meiser LC, Nguyen BH, Chen YJ, et al. Synthetic DNA applications in information technology [J]. *Nature Communications*, 2022, 13(1): 352.
- [18] van der Valk T, Pečnerová P, Diez-del-Molino D, et al. Million-year-old DNA sheds light on the genomic history of mammoths [J]. *Nature*, 2021, 591(7849): 265-269.

- [19] Matange K, Tuck JM, Keung AJ. DNA stability: a central design consideration for DNA data storage systems [J]. *Nature Communications*, 2021, 12(1): 1358.
- [20] Antkowiak PL, Koch J, Nguyen BH, et al. Integrating DNA encapsulates and digital microfluidics for automated data storage in DNA [J]. *Small*, 2022, 18(15): e2107381.
- [21] Fei ZJ, Gupta N, Li MJ, et al. Toward highly effective loading of DNA in hydrogels for high-density and long-term information storage [J]. *Science Advances*, 2023, 9(19): eadg9933.
- [22] Choi Y, Ryu T, Lee AC, et al. High information capacity DNA-based data storage with augmented encoding characters using degenerate bases [J]. *Scientific Reports*, 2019, 9(1): 6582.
- [23] Mayer C, McInroy GR, Murat P, et al. An epigenetics-inspired DNA-based data storage system [J]. *Angewandte Chemie International Edition*, 2016, 55(37): 11144-11148.
- [24] Cao B, Zhang XK, Cui S, et al. Adaptive coding for DNA storage with high storage density and low coverage [J]. *NPJ Systems Biology and Applications*, 2022, 8(1): 23.
- [25] Gimpel AL, Stark WJ, Heckel R, et al. A digital twin for DNA data storage based on comprehensive quantification of errors and biases [J]. *Nature Communications*, 2023, 14(1): 6026.
- [26] Bonnet J, Colotte M, Coudy D, et al. Chain and conformation stability of solid-state DNA: implications for room temperature storage [J]. *Nucleic Acids Research*, 2010, 38(5): 1531-1546.
- [27] Organick L, Nguyen BH, McAmis R, et al. An empirical comparison of preservation methods for synthetic DNA data storage [J]. *Small Methods*, 2021, 5(5): e2001094.
- [28] Chen WD, Kohl AX, Nguyen BH, et al. Combining data longevity with high storage capacity—layer-by-layer DNA encapsulated in magnetic nanoparticles [J]. *Advanced Functional Materials*, 2019, 29(28): 1901672.
- [29] Ma JY, Wang KX, Ma M, et al. Magnetic microsphere/silica nanoparticle composite structures for switchable DNA storage [J]. *ACS Applied Nano Materials*, 2022, 5(10): 15619-15628.
- [30] Yu MS, Lim D, Kim J, et al. Processing DNA storage through programmable assembly in a droplet-based fluidics system [J]. *Advanced Science*, 2023, 10(32): e2303197.
- [31] Nguyen BH, Takahashi CN, Gupta G, et al. Scaling DNA data storage with nanoscale electrode wells [J]. *Science Advances*, 2021, 7(48): eabi6714.
- [32] Geng CY, Liu SQ, Jiang XY. A nanoparticle-coated microfluidic chip for automated, non-destructive extraction of encapsulated DNA in data storage [J]. *Chemical Science*, 2023, 14(15): 3973-3981.
- [33] Grass RN, Heckel R, Puddu M, et al. Robust chemical preservation of digital information on DNA in silica with error-correcting codes [J]. *Angewandte Chemie International Edition*, 2015, 54(8): 2552-2555.
- [34] Newman S, Stephenson AP, Willsey M, et al. High density DNA data storage library via dehydration with digital microfluidic retrieval [J]. *Nature Communications*, 2019, 10(1): 1706.
- [35] Seeman NC, Sleiman HF. DNA nanotechnology [J]. *Nature Reviews Materials*, 2017, 3(1): 17068.
- [36] Zhang YN, Wang F, Chao J, et al. DNA origami cryptography for secure communication [J]. *Nature Communications*, 2019, 10(1): 5469.
- [37] Tomek KJ, Volkel K, Indermaur EW, et al. Promiscuous molecules for smarter file operations in DNA-based data storage [J]. *Nature Communications*, 2021, 12(1): 3518.
- [38] Lopez R, Chen YJ, Dumas Ang S, et al. DNA assembly for nanopore data storage readout [J]. *Nature Communications*, 2019, 10(1): 2933.
- [39] Xu CT, Ma B, Gao ZL, et al. Electrochemical DNA synthesis and sequencing on a single electrode with scalability for integrated data storage [J]. *Science Advances*, 2021, 7(46): eabk0100.
- [40] Bogels BWA, Nguyen BH, Ward D, et al. DNA storage in thermoresponsive microcapsules for repeated random multiplexed data access [J]. *Nature Nanotechnology*, 2023, 18(8): 912-921.

- [41] Lau B, Chandak S, Roy S, et al. Magnetic DNA random access memory with nanopore readouts and exponentially-scaled combinatorial addressing [J]. *Scientific Reports*, 2023, 13(1): 8514.
- [42] Koch J, Gantenbein S, Masania K, et al. A DNA-of-things storage architecture to create materials with embedded memory [J]. *Nature Biotechnology*, 2020, 38(1): 39-43.
- [43] Anavy L, Vaknin I, Atar O, et al. Data storage in DNA with fewer synthesis cycles using composite DNA letters [J]. *Nature Biotechnology*, 2019, 37(10): 1229-1236.
- [44] Goldman N, Bertone P, Chen SY, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA [J]. *Nature*, 2013, 494(7435): 77-80.
- [45] Song LF, Geng F, Gong ZY, et al. Robust data storage in DNA by de Bruijn graph-based de novo strand assembly [J]. *Nature Communications*, 2022, 13(1): 5361.
- [46] Zhang Y, Kong LL, Wang F, et al. Information stored in nanoscale: encoding data in a single DNA strand with Base64 [J]. *Nano Today*, 2020, 33: 100871.
- [47] Ren YB, Zhang Y, Liu YW, et al. DNA-based concatenated encoding system for high-reliability and high-density data storage [J]. *Small Methods*, 2022, 6(4): e2101335.
- [48] Yim SS, McBee RM, Song AM, et al. Robust direct digital-to-biological data storage in living cells [J]. *Nature Chemical Biology*, 2021, 17: 246-253.
- [49] Zhang Y, Ren YB, Liu YY, et al. Preservation and encryption in DNA digital data storage [J]. *ChemPlusChem*, 2022, 87(9): e202200183.
- [50] Hao M, Qiao HY, Gao YM, et al. A mixed culture of bacterial cells enables an economic DNA storage on a large scale [J]. *Communications Biology*, 2020, 3(1): 416.
- [51] Chen WG, Han MZ, Zhou JT, et al. An artificial chromosome for data storage [J]. *National Science Review*, 2021, 8(5): nwab028.
- [52] Sun FJ, Dong YM, Ni M, et al. Mobile and self-sustained data storage in an extremophile genomic DNA [J]. *Advanced Science*, 2023, 10(10): e2206201.
- [53] Grass RN, Heckel R, Dessimoz C, et al. Genomic encryption of digital data stored in synthetic DNA [J]. *Angewandte Chemie International Edition*, 2020, 59(22): 8476-8480.
- [54] Zheng LL, Li JZ, Wen M, et al. Enthalpy and entropy synergistic regulation-based programmable DNA motifs for biosensing and information encryption [J]. *Science Advances*, 2023, 9(20): eadf5868.
- [55] Fan CY, Deng Q, Zhu TF. Bioorthogonal information storage in L-DNA with a high-fidelity mirror-image *Pfu* DNA polymerase [J]. *Nature Biotechnology*, 2021, 39(12): 1548-1555.
- [56] Li SY, Liu JK, Zhao GP, et al. CADs: CRISPR/Cas12a-assisted DNA steganography for securing the storage and transfer of DNA-encoded information [J]. *ACS Synthetic Biology*, 2018, 7(4): 1174-1178.
- [57] Kim J, Bae JH, Baym M, et al. Metastable hybridization-based DNA information storage to allow rapid and permanent erasure [J]. *Nature Communications*, 2020, 11(1): 5008.
- [58] Tabatabaei Yazdi SMH, Yuan YB, Ma J, et al. A rewritable, random-access DNA-based storage system [J]. *Scientific Reports*, 2015, 5: 14138.
- [59] Chandrasekaran AR, Levchenko O, Patel DS, et al. Addressable configurations of DNA nanostructures for rewritable memory [J]. *Nucleic Acids Research*, 2017, 45(19): 11459-11465.
- [60] Liu YY, Ren YB, Li JJ, et al. *In vivo* processing of digital information molecularly with targeted specificity and robust reliability [J]. *Science Advances*, 2022, 8(31): eabo7415.
- [61] Sadremomtaz A, Glass RF, Guerrero JE, et al. Digital data storage on DNA tape using CRISPR base editors [J]. *Nature Communications*, 2023, 14(1): 6472.
- [62] Liu F, Li JS, Zhang TZ, et al. Engineered spore-forming *Bacillus* as a microbial vessel for long-term DNA data storage [J]. *ACS Synthetic Biology*, 2022, 11(11): 3583-3591.

- [63] Xu CT, Ma B, Dong X, et al. Assembly of reusable DNA blocks for data storage using the principle of movable type printing [J]. *ACS Applied Materials & Interfaces*, 2023, 15(20): 24097-24108.
- [64] Yan YQ, Pinnamaneni N, Chalapati S, et al. Scaling logical density of DNA storage with enzymatically-ligated composite motifs [J]. *Scientific Reports*, 2023, 13(1): 15978.
- [65] Welzel M, Schwarz PM, Lochel HF, et al. DNA-Aeon provides flexible arithmetic coding for constraint adherence and error correction in DNA storage [J]. *Nature Communications*, 2023, 14(1): 628.
- [66] Park SJ, Kim S, Jeong J, et al. Reducing cost in DNA-based data storage by sequence analysis-aided soft information decoding of variable-length reads [J]. *Bioinformatics*, 2023, 39(9): btad548.
- [67] Pan C, Tabatabaei SK, Tabatabaei Yazdi SMH, et al. Rewritable two-dimensional DNA-based data storage with machine learning reconstruction [J]. *Nature Communications*, 2022, 13(1): 2984.
- [68] Meiser LC, Gimpel AL, Deshpande T, et al. Information decay and enzymatic information recovery for DNA data storage [J]. *Communications Biology*, 2022, 5(1): 1117.
- [69] Wang F, Lv H, Li Q, et al. Implementing digital computing with DNA-based switching circuits [J]. *Nature Communications*, 2020, 11(1): 121.
- [70] Bee C, Chen YJ, Queen M, et al. Molecular-level similarity search brings computing to DNA data storage [J]. *Nature Communications*, 2021, 12(1): 4764.
- [71] Bonnet J, Subsoontorn P, Endy D. Rewritable digital data storage in live cells via engineered control of recombination directionality [J]. *Proceedings of the National Academy of Sciences*, 2012, 109(23): 8884-8889.
- [72] Sheth RU, Yim SS, Wu FL, et al. Multiplex recording of cellular events over time on CRISPR biological tape [J]. *Science*, 2017, 358(6369): 1457-1461.
- [73] Bhattarai-Kline S, Lear SK, Fishman CB, et al. Recording gene expression order in DNA by CRISPR addition of retron barcodes [J]. *Nature*, 2022, 608(7921): 217-225.
- [74] Lim CK, Yeoh JW, Kunartama AA, et al. A biological camera that captures and stores images directly into DNA [J]. *Nature Communications*, 2023, 14(1): 3921.
- [75] Doricchi A, Platnich CM, Gimpel A, et al. Emerging approaches to DNA data storage: challenges and prospects [J]. *ACS Nano*, 2022, 16(11): 17552-17571.
- [76] Ruttan MGTA, Vaandrager FW, Elemans JAAW, et al. Encoding information into polymers [J]. *Nature Reviews Chemistry*, 2018, 2(11): 365-381.
- [77] Ren Y, Zhang Y, Liu Y, et al. Highly reliable and efficient encoding systems for hexadecimal polypeptide-based data storage [J]. *Fundamental Research*, 2023, 3(2), 298-304.
- [78] Liu YW, Zhao KL, Ren YB, et al. Highly plasticized lanthanide luminescence for information storage and encryption applications [J]. *Advanced Science*, 2022, 9(7): e2105108.