中国科学: 生命科学 2025 年 第 55 卷 第 10 期: 2031 ~ 2042
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
评 述 中国知名大学及研究院所专栏 天津大学合成生物学20周年专辑

DNA信息存储的可靠性研究进展

岳雪清[1†], 郑至仪[1†], 曹瑞盈[1], 周鹏华[1], 陈鑫[1,2*]

1. 天津大学应用数学中心, 天津 300072
2. 合成生物技术全国重点实验室, 天津大学, 天津 300072
† 同等贡献
* 联系人, E-mail: chen_xin@tju.edu.cn

收稿日期: 2025-08-20; 接受日期: 2025-09-12; 网络版发表日期: 2025-09-26
国家重点研发计划(批准号: 2020YFA0712100, 2024YFF1500500)和国家自然科学基金(批准号: 62071327)资助

摘要 全球数据量在互联网、人工智能与大模型的推动下呈指数级增长, 传统硅基存储在密度、能耗、成本与寿命方面已逼近物理与经济极限. DNA作为新型信息存储介质, 具备超高存储密度、超长保存寿命和低维护能耗等优势, 成为未来大规模数据存储的重要候选方案. 近年来, DNA信息存储的整体可靠性显著提升: 众多实验实现了数据的零误码重构, 通过优化封装与存储条件, 保存稳定性可推至理论半衰期数万年以上. 本文从编码策略、生化过程与解码机制三方面系统综述了DNA存储可靠性研究进展, 涵盖喷泉码、HEDGES码、阴阳码等新型编码方法, 适用于DNA存储的合成、测序与保存技术, 以及应对大规模数据、复杂噪声与数据无序性的编解码优化. 还探讨了深度学习、仿真工具与系统集成在提升可靠性方面的潜力, 并展望了DNA存储未来的应用前景.

关键词 DNA存储, 数据编码, DNA合成, DNA测序, 数据解码

我们正处在一个前所未有的信息时代, 随着互联网、人工智能及大模型的快速发展, 全球每年产生的数据量已达到泽字节级别, 并且毫无放缓的迹象[1]. 然而, 传统硅基存储技术在制造成本、能耗、寿命与存储密度等方面日益逼近物理与经济极限. 例如, 硬盘存储密度受限于物理约束仅约为2 Tb/in^2, 企业级硬盘持续运行的平均功耗超9 W/盘, 大规模部署时能耗负担显著; 磁带虽成本低且可在理想条件下保存20~30年, 但其读取速度慢且易受环境波动影响数据完整性. 这些瓶颈推动研究者探索新型存储介质, 其中, DNA因具备超高理论存储密度(约460 EB/g)、超长保存寿命(可达数千年)、低维护能耗和极小物理体积等优势而备受关注. 自2012年, Church等人[2]首次报道DNA信息存储以来, 该领域的研究已覆盖数据编码、DNA合成与测序、介质存储及解码等关键环节. 然而, 要实现其在实际应用中的广泛推广, 可靠性是需要关注的重要方面.

DNA存储的可靠性是指在数据编码、合成、存储、测序和解码全流程中, 确保数据完整性、可恢复性和鲁棒性的能力, 以应对合成与测序中的生物化学错误(如插入、删除、替换)、存储介质降解以及数据无序性等挑战. 可靠性通过定量指标统一衡量, 包括单

引用格式: 岳雪清, 郑至仪, 曹瑞盈, 等. DNA信息存储的可靠性研究进展. 中国科学: 生命科学, 2025, 55: 2031–2042
Yue X Q, Zheng Z Y, Cao R Y, et al. Research progress of the reliability of DNA data storage (in Chinese). Sci Sin Vitae, 2025, 55: 2031–2042, doi: 10.1360/SSV-2025-0200

碱基错误率、信息密度(bit/nt)、合成与测序的错误率、解码恢复率及所需测序覆盖度等指标. 这些指标共同评估系统在面对各种错误和挑战时的表现, 确保数据在长期保存和检索中的完整性.

随着合成与测序精度提升及编码策略优化, DNA存储的整体可靠性已显著提高, 多项实验在受控条件下实现了全流程误码率[3]低于10^{-6}, 并在数百MB规模数据集上完成零误码重构[4]. 在此背景下, 本文将围绕DNA存储全流程中的三个核心环节: 编码的可靠性、生化过程的可靠性以及解码的可靠性, 系统综述国内外相关研究进展. 首先, 在编码层面, 重点评述GC含量约束、均聚物长度限制等生化可行性约束对错误率与有效信息密度的影响, 以及从传统固定长度编码到新兴喷泉码、HEDGES码、阴阳码等方法在冗余设计、错误检测与纠正方面的优化策略. 其次, 在生化过程方面, 分析DNA合成与高通量测序中的插入、缺失、替换等典型错误类型及其来源, 总结通过工艺优化、反应体系改良和存储介质封装技术提升数据稳定性的最新成果, 并比较不同技术平台在精度与速度上的差异. 最后, 在解码环节, 探讨从传统比对算法到理论研究的演进, 重点关注其在应对大规模数据、复杂噪声与数据无序性等问题上的能力. 通过对上述三个环节的系统梳理, 文章旨在系统性地梳理和分析当前DNA信息存储领域在可靠性方面的研究进展, 全面探讨DNA存储各个环节中可能出现的错误和挑战.

本文从编码、生化过程、解码三个方面系统综述DNA存储的可靠性研究(图1), 第一部分聚焦编码可靠性, 分析生物约束策略、冗余与纠错算法以及AI在编码可靠性中的潜力与优势; 第二部分探讨生化过程可靠性, 涵盖合成、保存与测序技术; 第三部分剖析解码可靠性, 从编解码方案、理论基础以及模拟工具三个角度进行分析.

1 编码的可靠性

DNA编码技术是DNA信息存储系统的核心环节, 其任务是将数字数据(二进制序列)转换为符合DNA合成与测序生物化学约束的碱基序列(A, T, G, C), 确保

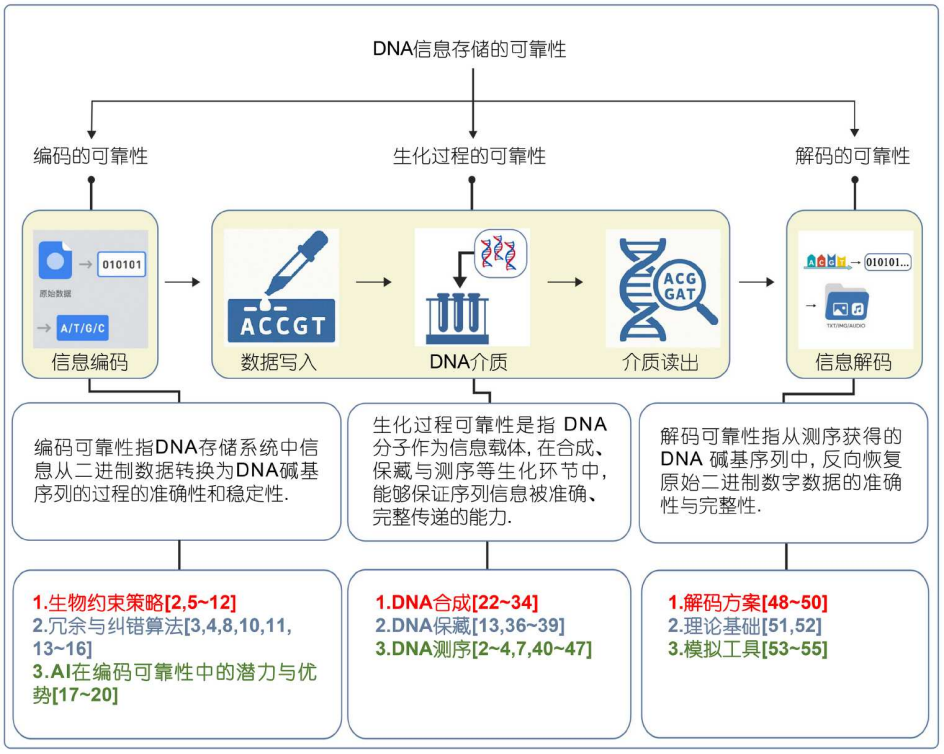

图 1 DNA信息存储关键环节的可靠性
Figure 1 Reliability of key processes in DNA data storage

在合成、存储和测序过程中能够有效抵抗替换、插入、删除(IDS)等错误, 保障数据在编码阶段的完整性和可恢复性.

编码可靠性指DNA存储系统中信息从二进制数据转换为DNA碱基序列过程的准确性和稳定性, 可通过错误率、恢复率和信息密度(bit/nt)等指标衡量. 编码过程通常涉及两个紧密关联的步骤: 二进制到DNA序列的转换(包括生物约束策略的设计)和引入冗余算法(以增强纠错能力). 这些步骤的顺序可根据具体设计灵活调整, 通常先对数字信息进行分组和嵌入纠错, 然后采用约束编码转换为DNA序列, 以优化整体效率. 此外, 人工智能(AI)技术的应用为优化编码策略和提升可靠性提供了新路径. 本节从以下三个小节展开讨论: 生物约束策略、冗余与纠错算法, 以及AI在编码可靠性中的潜力与优势. 代表性方法的特征对比见表1.

1.1 生物约束策略

在DNA编码的第一步, 即二进制数据到DNA序列的转换过程中, 生物约束策略的设计至关重要. 这一策略旨在生成符合DNA合成、保存和测序有生物化学特性的碱基序列, 从而降低错误发生概率并提升整体可靠性. DNA作为生物分子, 受限于其物理化学性质, 例如, GC含量(鸟嘌呤-胞嘧啶比例)和均聚物长度(连续相同碱基的序列), 这些因素直接影响合成效率和测序准确性. 如果忽略这些约束, 序列可能导致聚合酶链反应(PCR)失败、二级结构形成或测序偏差, 从而增加替换、插入或删除错误的发生率.

Church等人[2]在早期DNA存储实验中, 将GC含量控制在40%~60%的范围内, 以避免极端GC序列引起的合成困难和测序偏差. 该策略在后续工作中得到扩展, 例如, Ping等人[5]在2019年提出的Yin-Yang Codec系统通过“阴阳”规则编码, 确保生成的序列GC平衡, 同时最小化二级结构的形成风险. 该系统在实验中实现了99.94%的恢复率, 证明了GC约束在提升可靠性方面的作用. 同样, Welzel等人[6]提出的DNA-Aeon编码允许用户自定义GC含量阈值, 支持40%~60%的窗口化控制, 进一步适应不同合成平台的需求. Goldman等人[7]在向实用高容量DNA存储迈进的研究中, 也强调了GC平衡的重要性, 他们编码了739 kB数据并实现了100%准确恢复, 突显了这一约束在扩展性方面的作用.

另一个关键约束是均聚物长度限制. 长均聚物(如连续5个A或T)易导致聚合酶滑动, 引发插入或删除错误. Press等人[8]提出的HEDGES编码方案明确限制均聚物长度不超过3~4个碱基, 并在实验中实现了对高密度数据的完美恢复, 避免了合成过程中的滑移错误. 2020年, Antkowiak等人[9]使用光刻合成技术, 进一步强调了这一约束, 通过避免长均聚物和不期望基序(如限制性酶切位点), 降低错误率至可控水平. 该研究存储了莫扎特乐谱文件, 并在高错误合成条件下实现了完整恢复, 突显了均聚物限制在实际应用中的必要性.

此外, 一些策略还包括避免不期望基序和二级结构. Hawkins等人[10]在设计中整合了最小化发夹结构的算法, 确保序列在测序时保持线性形式, 从而减少删除错误. Zhao等人[11]提出的Composite Hedges Nanopores编码扩展了这一理念, 其中使用退化八字母表来生成序列, 不仅满足GC和均聚物约束, 还将信息密度从0.59 bit/nt提高至1.17 bit/nt. 该系统在纳米孔测序实验中处理了15.9%的插入删除错误, 证明了多约束策略在高错误环境下的鲁棒性. Anavy等人[12]引入复合DNA字母, 通过混合四种核苷酸的比例编码, 隐含地满足了GC平衡和避免长均聚物的约束, 同时减少了20%的合成周期, 编码了6.4 MB数据. 这些生物约束策略的目的是在编码源头最小化潜在错误, 确保序列在后续合成和测序阶段的生物兼容性, 从而为整体可靠性奠定基础.

1.2 冗余与纠错算法

在完成二进制到DNA序列的转换后, 编码过程的第二步引入冗余算法, 以增强纠错能力. 这一步骤通过添加额外信息(如校验位或冗余块)来应对合成、存储和测序中的错误, 包括替换、插入和删除. 冗余算法的设计目标是平衡信息密度与恢复率, 确保即使在高错误率下也能实现数据完整性. 常见的指标包括恢复率(通常要求接近100%)和信息密度(bit/nt), 这些算法往往基于信息论原理, 如香农容量, 以接近理论上限. 需要注意的是, 一个编码的冗余率与所应对的错误率和错误类型密切相关, 因此, 在比较不同研究工作的冗余率时, 应考虑其前提条件, 例如, 是否针对同等水平的错误率(如10%错误率), 还是应对的错误类型(如主要针对替换错误或插入删除错误)各有不同, 这有助于

表 1 DNA编码方法对比

Table 1 Comparison of DNA encoding schemes

编码方法	存储机制	信息密度(bit/nt)	误差校正和检测	生物约束及访问机制	作者(国家)
固定映射	体外合成寡核苷酸	0.6	共识测序, 处理替换	避免极端GC含量、重复序列和二级结构; 地址块+PCR, 支持随机访问	Church等人[2](美国)
阴阳码	体外: 200 nt寡核苷酸池; 体内:酵母细胞中~54 kb DNA片段	体外: 1.965; 体内: 1.59	阴阳码, RS码纠错	GC含量40%~60%, 均聚物长度≤4 nt; 体外通过PCR引物实现数据访问, 体内通过全基因组测序访问	Ping等人[5](中国、美国)
DNA-Aeon	体外合成寡核苷酸	–	算术编码+Raptor码, CRC+聚类	GC, motif可自定义码本, 均聚物长度≤3 nt; 支持随机访问	Welzel等人[6](德国)
Huffman编码	–	–	四重冗余+概率	避免均聚物≥2 nt; 顺序访问	Goldman等人[7](英国、美国)
HEDGES	体外合成寡核苷酸	实际可变编码率: 0.166~0.75	HEDGES(哈希+RS码)	GC平衡, 禁止均聚物>4 nt; 顺序访问	Press等人[8](美国)
伪随机化结合RS 码	体外合成寡核苷酸	0.806	RS码+聚类	随机化避免极端序列; 顺序访问	Antkowiak等人[9](瑞士、奥地利、美国、德国)
FREE 条形码	体外合成寡核苷酸	–	FREE码	GC平衡, 均聚物长度≤2 nt; 条码索引+聚类	Hawkins等人[10](美国)
Composite Hedges Nanopores编码	体外合成寡核苷酸	∑8(8字母): 1.17; ∑4(4字母): 0.59	Composite Hedges码+RS码	无特殊生物约束; 兼容纳米孔测序, 纳米孔+快速解码	Zhao等人[11](中国)
复合DNA字母	体外合成寡核苷酸	3.32(模拟)	复合碱基编码+纠错码	依赖混合比例避免极端序列; 顺序访问	Anavy等人[12](以色列)
双层RS码级联	体外硅基质封装DNA	0.84	RS码	避免同聚物和高GC含量; PCR+测序, 支持随机访问	Grass等人[13](瑞士)
FEC(分层纠错编码, RS码与BCH码级联)	体外合成寡核苷酸	1.6	RS码+BCH码+CRC	均聚物长度≤3 nt, 避免自反向互补序列; 顺序访问	Blawat等人[14](德国)
长块RS码+引物库随机访问(外码RS码, 内码旋转编码)	体外合成寡核苷酸	0.81	RS码+聚类	避免极端GC含量, 规避长均聚物; 引物设计+随机访问PCR	Organick等人[4](美国)
DNA Fountain码	体外合成寡核苷酸	1.57	喷泉码(LT码+RS码)	GC 45%~55%, 均聚物长度≤3 nt; PCR+测序, 支持随机访问	Erlich等人[3](美国)
Trellis BMA编码	体外合成寡核苷酸	–	Trellis BMA 编码	无特殊生物约束; 编码+索引+聚类	Srinivasavaradhan等人[15](美国)
ECC测序	体外合成寡核苷酸	–	双基流图+信息论纠错	无特殊生物约束; 顺序访问	Chen等人[16](中国)
Turbo Autoencode神经网络编码	体外合成寡核苷酸	–	Turbo自编码器纠错	GC含量40%~60%; 均聚物长度≤3 nt; 顺序访问	Welzel等人[17](德国)
DNA-QLC	体外合成寡核苷酸	–	Levenshtein码	GC含量50%; 均聚物长度≤2 nt; 顺序访问	Zheng等人[18](中国)

更客观地评估算法的性能.

Reed-Solomon(RS)码是最早应用于DNA存储的冗余算法之一. 2015年, Grass等人[13]使用RS码将83 kB数据编码为4991个DNA片段, 并在硅胶封装中模拟千年存储, 实现了无错误恢复, 冗余率约为1.13(针对约1%~2%错误率). Blawat等人[14]扩展了RS码的应用, 存储22 MB数据并处理插入、删除和交换错误, 残余错误概率降至硬盘级别, 冗余率约为1.13~1.2, 针对更高复杂错误类型. 该研究强调了RS码在应对DNA通道模型(包括合成和扩增错误)中的优势, 但也指出其对局部相关性和丢失率变异的敏感性. Organick等人[4]在大规模随机访问DNA存储中, 使用RS码编码200 MB数

据, 支持单个文件提取, 并通过优化测序冗余实现了零错误恢复, 冗余率1, 应对大规模数据下的不均衡丢失, 进一步证明了RS码在扩展性方面的潜力.

喷泉码(fountain codes)代表了更先进的无速率冗余策略. Erlich和Zielinski[3]提出的DNA Fountain码引入了Luby变换(LT)码变体, 生成略多于输入数据的DNA寡核苷酸, 冗余率1.07, 针对约5%~10%丢失率, 实现了215 PB/g的物理密度和完美恢复. 该算法通过异或操作创建重叠数据块, 允许从任意子集恢复信息, 特别适合处理寡核苷酸丢失. 实验中, 它成功存储了2.15 MB文件, 包括操作系统和电影, 信息密度达1.57 bit/nt, 接近香农容量的86%. Srinivasavaradhan等人[15]提出的Trellis BMA算法, 结合BCJR推理和BMA共识方法, 在IDS通道上实现线性复杂度纠错, 并在实验数据集中降低了错误率, 冗余率视错误类型调整, 针对高IDS错误. 该算法利用多迹重构数据集, 强调了冗余在实际DNA读出的作用.

针对插入删除(indel)错误的专用算法进一步提升了可靠性. Press等人[8]通过哈希编码和贪婪exhaustive搜索, 将indel转换为替换错误, 并与RS外码结合, 实现10%错误下的无错误恢复, 冗余率约1.2~1.67, 专为indel错误设计. 该方案在模拟中支持PB级数据, 信息密度接近理论上限. Hawkins等人[10]使用填充/截断右端编辑(FREE)码, 纠正长度变化的indel错误, 并在高通量测序中实现了10^6个单错误纠错的16-mer条码列表, 冗余率较低, 但针对特定短序列错误. Zhao等人[11]结合了HEDGES和纳米孔优化, 处理15.9% indel和7.8%替换错误, 仅需4~8倍物理冗余即可完整恢复文本和图像数据, 冗余率4~8, 针对高indel环境. Chen等人[16]提出的ECC sequencing, 使用双基流图和信息论纠错, 结合荧光SBS化学, 实现98.1%原始准确率下的200 bp无错误序列, 冗余率约1.1~2, 针对替换主导错误. 该方法通过三组正交退化序列嵌入冗余, 显著提高了读长和准确性.

这些冗余算法的应用证明了其在提升编码可靠性的关键作用. 通过引入可控冗余, 它们不仅纠正随机错误, 还适应DNA通道的特定噪声模型, 确保高恢复率和信息密度.

1.3 AI在编码可靠性中的潜力与优势

近年来, 人工智能(AI)技术在DNA编码可靠性中的应用已成为热点趋势, 为优化生物约束和冗余算法提供了强大工具. AI的优势在于其数据驱动的学习能力, 能够处理复杂、非线性约束和噪声模型, 超越传统方法. 机器学习模型, 如神经网络和遗传算法, 可自动生成优化序列, 预测错误模式, 并动态调整冗余水平, 从而提升整体恢复率和信息密度.

在优化生物约束和冗余算法方面, AI通过生成对抗网络(GAN) 模拟DNA序列的生物兼容性, 并结合纠错机制处理存储错误. Welzel等人[17]在Turbo Autoencoders中进一步利用PyTorch框架, 开发了Autoturbo-DNA自编码器, 针对DNA存储通道的插入、删除和替换错误, 生成符合用户定义约束(如GC含量和均聚物长度)的序列. 该方法通过训练期间的噪声生成, 优化序列的生物兼容性, 显著提高了编码效率. DNA-QLC方法[18]利用量化ResNet VAE(QRes-VAE)模型进行高效图像压缩, 并通过Levenshtein(LC)码实现DNA序列的鲁棒错误纠正. 该方案的编码密度比DNA Fountain高出2.4倍, 在2%的错误率下实现SSIM值为0.917的可靠图像重建. AI驱动的压缩与组合约束满足的结合使DNA-QLC成为可编程加密和数据存储的优越替代方案. Bar-Lev等人[19]提出的DNAformer管道, 整合深度神经网络(DNN)与张量积纠错码(ECC)和安全边际机制, 在3.1 MB数据上实现3200倍速度提升和40%准确率改善, 码率达1.6 bits/base, 验证了AI在高噪声环境下的鲁棒性. Ruan等人[20]的DSI-ResCNN框架, 结合残差卷积神经网络(ResCNN)和DNA序列丢失控制(SDC)模块, 增强了DNA图像存储的容错能力, 通过深度学习优化序列重建, 显著改善了高错误环境下的图像保真度.

作为新兴趋势, AI有望推动DNA编码从规则驱动向智能优化转型, 显著提升可靠性. 然而, 其应用需谨慎验证, 以确保生物兼容性和计算可行性.

2 生化过程的可靠性

在生化过程中, 可靠性贯穿于合成、保存与测序的全流程. 当前, 合成阶段多采用微阵列化学合成生产短链寡核苷酸, 通常小于200 nt, 以满足高通量需求, 此过程受单步偶联效率限制, 易产生插入、缺失、替换和丢失等错误. 在保存阶段, DNA分子易受脱嘌呤、链断裂、氧化等化学降解影响, 其速率受存储形态、

温湿度和光照等环境因素制约. 物理封装、化学修饰与纳米材料介质可显著延缓降解过程并提升环境适应性. 在测序阶段, DNA存储主要依赖高通量二代测序实现批量读取, 但其读取过程存在GC含量偏倚及碱基识别错误, 容易引入替换、插入和缺失等噪声. 这些累积的错误和降解会影响数据恢复能力, 因此, 生化过程可靠性是指DNA分子作为信息载体, 在合成、保存与测序等生化环节中, 能够保证序列信息被准确、完整传递的能力.

2.1 DNA合成

DNA合成技术自20世纪50年代[21]起快速发展, 形成三代技术路线: 第一代为基于柱的固相亚磷酰胺化学合成[22]; 第二代为高通量微阵列合成, 根据"脱保护"技术原理不同, 可分为光化学合成法[23~25]、喷墨打印法[26~28]、电化学合成法[29~31]等方法, 以上两代都属于DNA的化学合成法; 第三代为酶促合成技术, 常用TdT酶.

2.1.1 化学合成

1980年, Beaucage和Caruthers[22]提出固相亚磷酰胺三酯合成法, 形成脱保护、偶联、加帽、氧化的四步循环. 其偶联效率达99%~99.8%, 但随着链长增长, 合成正确率显著下降, 如200 bp长链正确率仅约37% ($0.995^{200}\approx0.37$). 目前, 该方法已实现商业化, 柱式合成仪可并行处理48~1536个反应柱, 合成DNA链长度范围6~200个碱基, 单个反应柱的寡核苷酸产量一般5~5000 nmol, 单个循环耗时2~6.5分钟.

固相合成DNA方法因合成准确率较高, 至今仍广泛应用于对纯度要求较高的科研与临床场景. 然而, 该方法每个反应柱仅能合成大量相同序列的DNA分子, 而DNA存储更侧重于合成数量庞大、互不相同的序列, 且每种序列的用量较小. 柱式合成在通量和成本上难以满足这一需求, 而微阵列合成技术则凭借其并行合成成千上万种寡核苷酸的能力, 更契合DNA存储对高通量、多样性和低成本的要求.

光化学合成法[23]是最早的技术路径之一. 研究表明, 受光的衍射、散射以及光解效率不均等因素影响[24], 该方法的错误率较高, 其中缺失错误率可高达4.65%~13.6%, 替换和插入错误也分别达到了0.56%~3.6%和0.17%~4.6%. 为降低成本和提升灵活性, 无掩模光化学合成[25]与喷墨打印技术[26,27]应运而生. 喷墨法通过精确控制"墨水"喷射实现序列合成, 虽然在短链合成中表现优异, 例如, 有研究通过优化溶剂体系, 在合成30个核苷酸时实现了高达99.01%的步进偶联效率[28], 但随着链长增加, 液滴偏移和交叉污染的风险剧增, 导致错误累积, 限制了更长序列的高保真合成. 为了进一步提升合成密度和精度, 电化学合成法[29]被提出. 该技术利用微电极阵列精确控制酸的生成, 从而实现定点脱保护. 通过将反应区域限制在微米[30]甚至纳米尺度[31], 合成密度得以提升至2500万条/cm^2. 然而, 高密度也带来了新的挑战, 即酸的扩散可能导致邻近位点的非预期脱保护, 从而引发错误. 在先进的系统中, 合成180 nt长链的错误率约为1.8%~4.3%[31].

尽管微阵列合成的错误率高于柱式合成, 但DNA存储技术不需要依赖于完全无误的DNA合成, 通过引入里德-所罗门(RS)纠错码, 即使在原始合成DNA错误率较高的情况下, 也成功实现了100%的数据恢复[31]. 由于用于存储的DNA不需要像生物应用那样追求单分子的绝对完美, 这降低了对高精度合成技术的依赖, 从而加快了DNA存储技术的实际应用进程.

2.1.2 酶促合成

酶促DNA合成技术在水相中进行反应, 更具生物相容性, 所用试剂毒性更低、条件更温和, 具备合成更长DNA链(可达约8000 nt)的潜力. 该技术分为模板依赖型(如CDC法)和非依赖型(以TdT酶为主), 后者可在无模板下定向添加碱基.

DNA Script[32]将喷墨打印沉积方法与基于酶的合成相结合推出全球首台酶促桌面打印机"SYNTAX", 可在6~13小时内合成96条寡核苷酸(60~120 nt), 单步偶联效率99.4%. Lee等人[33]通过DMD光控释放金属离子激活TdT酶, 在阵列上实现12个位点的DNA并行合成. Li等人[34]利用工程化E-ZaTdT酶实现八重并行操作的DNA酶促合成, 单步偶联效率最高达99.4%, 平均合成准确率最高为99.35%, 最长可合成18个碱基的寡核苷酸, 单比特写读延迟约4.4分钟, 展现出较高的精度与效率. 虽然酶促合成单步偶联效率高于化学合成, 且正朝着并行化方向发展, 但是目前酶促DNA合成技术在合成通量上还达不到DNA存储应用所需量级.

2.2 DNA保存

在DNA被应用于数据存储之后，如何确保其在现实环境中的长期稳定性, 成为研究的关键问题. 尽管从DNA分子理论上可在理想条件下保存数十万年，其在实际使用中却易受到多种环境因素协同作用而发生降解[35]. 例如, 高温可显著加速水解与氧化反应, 每升高10℃, 水解速率增加2~3倍; 湿度超过70%时, 磷酸二酯键断裂风险大幅上升; 紫外辐射则会诱导嘧啶二聚体形成，导致碱基错误. 此外，在极端pH条件下，小于4或大于9时, 可破坏双螺旋结构, 金属离子(如Fe^{3+})催化活性氧生成, 引发碱基氧化; 同时, 微生物污染和细胞裂解后残留的核酸酶也会持续性降解DNA链. 面对这些复杂的降解机制，近年来针对DNA介质保存的研究不断深入, 逐步将其从理论可行推向工程可控.

为延长DNA介质的稳定性，研究者提出了一些保存策略, 如物理封装、材料创新和生物防护. 在物理封装方面，Grass等人[13]采用溶胶-凝胶法将DNA分子嵌入氨基功能化的硅石纳米多孔结构中，通过硅石惰性材料的物理化学屏障作用(孔径2~5 nm的三维网状结构)，有效隔离外界机械应力与化学干扰，显著延缓DNA降解. 材料创新路径中, Lange等人[36]提出的苏糖核酸技术, 通过将分子骨架替换为苏糖, 显著提高了抗核酸酶水解能力，体外实验显示其抗酶性能接近100%. 在生物防护方面, Liu等人[37]设计的芽孢杆菌孢子封装系统利用小酸可溶性蛋白与DNA结合形成稳定复合体, 即便在高温(70℃)和氧化应激条件下, 其数据错误率仍低于5%. 这些方法不仅有效缓解了环境应激下的降解风险，也为DNA存储提供了坚实的工程基础.

针对保存策略的可行性，多个研究进一步通过严苛条件模拟与误差控制算法验证其长期稳定性. 例如, 工程化芽孢杆菌孢子[37]在80℃与强氧化环境中仍保持99%以上的数据准确性，抗紫外能力亦达96%以上; 树枝状胶体封装体系[38]在70℃加速老化实验中推算其在4℃下的半衰期约为6000年，−18℃条件下可达200万年，且错误率维持在10^{-3}~10^{-4}. Takahashi等人[39]的一项中子辐射实验显示, 干燥DNA在典型存储设施中受高能粒子影响极低，其年错误率低于10^{-12}, 远优于磁带等传统存储介质. Grass等人[13]在二氧化硅封装DNA的热胁迫与信息恢复实验中指出，高温会导致碱基错误和序列错误增加，但这些错误仍在Reed-Solomon纠错能力范围内, 原始信息能够被完全恢复.

这些成果表明，通过策略性封装与结构修饰, DNA介质在应对温度、氧化、辐射等多种极端环境条件下仍能保持极高的稳定性与数据保真能力，表明DNA介质在长期保存和信息保真方面已表现出高度可靠性.

2.3 DNA测序

在DNA存储研究中, DNA测序技术应用最广泛的是二代测序技术，最常用Illumina测序，原因在于其高通量、低成本、短片段(100~300 bp)读长与DNA存储的短链高通量合成策略高度匹配. 在部分探索性研究中，第三代单分子测序技术(如纳米孔测序、PacBio SMRT)被用于读取长链DNA、进行实时测序或实现对纠错算法进行“压力测试”等特殊需求.

1977年，Sanger等人[40]使用双脱氧链终止法完成了DNA的首次测序实验, 这开启了DNA第一代测序的时代，桑格测序的读长通常在约500~1000 bp之间，具有极高的测序准确率(99%), 一般能在单次反应中处理一个或少量样本.

Illumina测序采用可逆终止子边合成边测序策略, 将文库分子固定在流动池表面，通过桥式扩增生成高密度簇，实现同步循环合成与荧光信号采集. 其平台通常能达到小于0.1%的平均单碱基错误率，并支持大规模并行测序，读长常见为2×150 bp，也可延伸至2×300 bp[41,42]. Illumina测序的错误模式具有规律性[43]: 替换错误最常见, 尤其在读长末端, 由于光信号衰减与碱基累积掺入不均导致判读不确定性增加; 均聚物与高GC区可能引起聚合酶效率波动和簇同步性差异, 局部错误率上升. 插入/缺失错误虽低, 但在特定序列中仍可能出现. 为降低误差，Illumina分析常结合Phred质量值、错误模型校正和冗余覆盖共识序列[44]生成.

2012年, Church等人[2]用HiSeq100nt双末端测序拼接115 nt数据块，约3000×覆盖，在527万比特电子书中仅有10比特错误，首次验证深度测序可纠正随机错误. 2013年，Goldman等人[7]使用HiSeq2000对117 nt DNA双末端测序7960万个读长对, 少量手动干预即可100%重建原始文件，确立HiSeq在高保真、大规模DNA存储验证中的地位. 随着需求向复杂编码和大规模存储发展，Illumina平台多样化. Erlich和Zielinski[3]使用

MiSeq 150循环双末端测序3200万读长验证“DNA喷泉”编码, 实现了接近理论容量极限的编码验证; Wetzl等人[45]使用Illumina MiSeq2×150 bp双末端测序1000万读长验证防水静电纺丝聚合物纤维中DNA存储器的非破坏性检索, 实现了99%以上的数据恢复率.

DNA存储领域的第三代单分子测序技术主要采用纳米孔测序. 该技术通过监测单链DNA分子跨越纳米孔时引发的离子电流扰动, 实现碱基序列的直接解析, 无需PCR扩增即可实现实时读取. 与二代测序相比, 纳米孔测序能够读取更长的DNA序列; 同时, 测序设备体积小巧、成本较低, 便于现场快速检测. 纳米孔测序的特性使其在DNA数据存储中展现出独特用途. 纳米孔测序的高插入-缺失错误率特性被用于纠错码鲁棒性的压力测试, 验证了编码方案在极端错误环境下的稳健性[4]; 纳米孔测序无需进行温控PCR操作, 与等温Cas9切割流程兼容, 可实现全等温数据读取[46]; 纳米孔测序能够直接识别非天然碱基对, 突破了传统合成测序在碱基种类上的限制, 为DNA存储扩展编码字母表提供技术支撑[47].

测序错误率的高低直接决定了测序深度与覆盖率的需求, 高错误率需要更高覆盖度以通过多次读取平均化随机错误, 会增加系统成本和测序时间.

DNA存储的生化过程在编码、合成和测序三个阶段均会产生一定的错误. 错误率会影响数据解码的准确性, 高错误率会导致编码后的DNA序列在读取时产生偏差, 增加纠错码的负担; 若错误率超过纠错码的容错能力, 则可能导致数据块无法正确解码, 从而降低整个存储系统的数据完整性和可恢复性, 即降低整个系统的可靠性.

3 解码的可靠性

DNA存储的解码需要从海量测序读段中可靠地重建原始数字信息. 其三大挑战为: 数据规模巨大、信道噪声复杂、数据无序性. 现代测序技术产生数十亿计的DNA读段, 对传统算法的处理能力构成巨大压力, 要求解码系统必须具备高效率和可扩展性; DNA在合成、存储和测序过程中引入的信道噪声会随机改变碱基序列和读段长度, 合成与测序过程错误的不对称性, 合成环节更易产生缺失和替换错误, 而测序环节则更常出现插入与删除等错误, 使得重建原始信息成为一项复杂的模式匹配与纠错任务; DNA分子以混杂文库形式存在, 数据无序性要求解码流程中的数据重组步骤必须精确地恢复读段的逻辑顺序.

在此背景下, 解码可靠性指从测序获得的DNA碱基序列中, 反向恢复原始二进制数字数据的准确性与完整性. 为应对这些挑战, 研究者们通过创新性的编解码方案、坚实的理论基础以及高效的模拟工具, 不断提升解码效率和精度, 使得解码可靠性得以显著增强.

随着DNA存储技术的不断成熟, 针对DNA存储的独特生化特性, 研究者们提出了多种创新的编码与解码方案, 显著提升了DNA存储系统的可靠性和实用性. 如第一部分所论及, 传统纠错码、针对插入删除场景设计的专用码, 还有可感知生化约束的编码策略等, 现已成为DNA编码设计领域的重要组成部分. 另外, 2021年, 天津大学元英进团队[48]合成了254886 bp的酵母人工染色体, 高效存储37.8 kB数据. 通过酵母稳定复制和便携纳米孔测序, 实现快速无错恢复. 采用伪随机序列与LDPC码纠正高达10%的插入删除错误, 测序覆盖度仅16.8×, 展现大规模存储与解码潜力. 2025年, 吴华明团队[49]开发HELIX系统, 成功将两张60 MB图像编码为13万条183 bp序列, 并在5.8×覆盖度下实现高保真恢复, 展示了针对图像类数据的结构化冗余优化能力. Ding等人[50]提出的Derrick系统首次将软判决解码引入DNA存储, 结合每个位点的置信度或概率信息, 将低置信度的“错误”位置标记为“擦除”, 有效提升了Reed-Solomon码的纠错能力, 在不增加冗余的情况下显著降低解码失败率. 上述方法通过对复杂错误模式的深度适应, 为DNA存储系统的高效且准确解码提供了重要技术支撑.

近年来, 研究人员开始从理论层面深入探索DNA存储的性能极限, 这为设计更可靠的编解码方案奠定了基础. Zrihan等人[51]通过定义与合成周期数相关的容量函数, 刻画了不同字母表规模和编码密度下的理论性能上界, 进一步提出可达容量的高效编码器与合成成本优化模型, 为DNA存储编解码方案的设计提供了理论依据. Cao与Chen[52]基于实际PCR测序数据, 提出DNA存储通道符合对数正态分布, 建立了对应的最小测序覆盖深度估计方法. 实验显示, 在每条设计链需至少 2 个读段才能成功检索、编码率为0.8的条件下, 解码成功所需最大覆盖深度为6.12~12.85, 显著高于均

匀通道预测值, 强调解码策略应贴合真实信道特性以保障数据恢复可靠性. 这些理论成果为编解码方案的设计提供了坚实的指导, 确保了技术发展的方向正确性.

另外, 一系列模拟工具也被相继提出, 这些工具能在实际合成DNA之前, 模拟整个数据存储和恢复过程, 以分析和预测可能出现的错误. 代表性工具如Storalator[53], MESA[54]和DeepSimulator[55], 分别聚焦于不同环节: Storalator支持从合成到测序再到重建的全过程流程模拟, 适用于快速测试算法效率; MESA则强调GC含量、K-mer结构、均聚物等序列特性对错误率的影响, 可辅助序列优化设计; 而DeepSimulator通过深度学习生成电流信号, 真实再现纳米孔测序过程, 是纳米孔信号处理算法开发的重要工具. 这些模拟器通过减少实际实验的试错成本, 为DNA存储技术的发展和应用提供了重要的技术支持.

解码阶段的可靠性不仅取决于单一算法的精度, 还受覆盖度、错误模式分布、索引设计、聚类算法鲁棒性以及纠错码配置等多因素共同影响. 研究者们通过创新性的编解码方案、坚实的理论基础以及高效的模拟工具, 共同应对了海量数据、复杂噪声和数据无序性等挑战, 从而为原始数字信息的完整和正确重建奠定了基础.

在整个DNA存储系统中, 解码可靠性起着至关重要的作用. 编码、合成、保存和测序环节都会引入不可避免的噪声与误差, 而解码作为最后一道关口, 决定了这些误差能否被有效消除, 原始信息能否被无损恢复.

4 总结与展望

随着互联网、人工智能和大数据技术的飞速发展, 全球数据量呈指数级增长, 对存储技术提出了前所未有的挑战. DNA存储作为一种新兴的信息存储介质, 以其超高存储密度、极长保存寿命和低维护成本, 展现出独特的应用潜力. 本文系统回顾了DNA存储领域的编码可靠性、生化过程可靠性及解码可靠性等关键技术进展.

当前研究表明, DNA存储在全流程的可靠性已经取得显著进展. 通过精心设计的编码策略与冗余纠错码, 单碱基错误率可控制在10^{-3}~10^{-4}的水平, 使得插入、删除和替换等生物化学错误在解码阶段可被有效修正. 合成与测序环节的高保真技术进一步提升了数据完整性, 其中, Illumina平台的碱基错误率通常在0.1%~1%之间, 结合适当的测序覆盖度(20~50×), 能够实现几乎无损的信息恢复; 纳米孔测序虽误差略高, 但通过重复测序和覆盖优化同样保证了高恢复率. 存储介质方面, 干燥或封装的DNA在常规环境下表现出极高稳定性, 理论半衰期可达数千年, 即使在中子辐射等苛刻条件下, 信息丢失也可忽略不计. 这些实验结果表明, 结合全流程的编码、合成、保存、测序与解码策略, DNA存储系统能够实现大于99.9%的信息恢复率, 具备高度鲁棒性和可恢复性, 从而在应对合成与测序误差、存储介质降解以及数据无序性等挑战时表现出可靠的数据保持能力, 为其在长期、大规模数据存储中的应用奠定了坚实基础.

随着合成、测序及介质保存技术的不断进步, 编解码技术的深入研究, DNA存储的可靠性正逐步从理论可行向工程可控转变. 未来在可靠性方面的发展可以从以下几个方向展望:

(1) 编解码算法智能化: 利用深度学习模型根据特定生化平台的错误特征自动生成最优编码, 实现对冗余水平的动态精调. 通过统一的深度学习模型直接处理原始测序数据, 实现无缝的信号解读、错误纠正、序列聚类与最终解码.

(2) 生化与物理技术协同: 技术的进步将推动酶促合成、高保真测序及先进封装技术的协同发展, 从物理层面大幅降低操作和环境因素对数据完整性的影响.

(3) 面向应用的可靠性分级: 对于国家档案等需永久保存的"冷数据"采用高冗余高强封装, 而容错性高的流媒体备份等数据可降低冗余以节约成本, 这种按需可靠性策略将拓展DNA存储的应用, 使其成为实用、可扩展且长期可靠的方案.

DNA存储作为信息存储领域的前沿方向, 正处于技术飞跃的关键阶段. 通过多学科交叉融合的持续创新, 未来有望实现超大规模、低成本、长寿命的生物分子级数据存储, 满足人类对数据存储容量和可靠性的极致追求, 推动信息技术迈向全新纪元.

参考文献

1 Reinsel J, Gantz J, Rydning J. The digitization of the world from edge to core. Framingham: International Data Corporation, 2018, 16: 1–28
2 Church G M, Gao Y, Kosuri S. Next-generation digital information storage in DNA. Science, 2012, 337: 1628
3 Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture. Science, 2017, 355: 950–954
4 Organick L, Ang S D, Chen Y J, et al. Random access in large-scale DNA data storage. Nat Biotechnol, 2018, 36: 242–248
5 Ping Z, Chen S, Zhou G, et al. Towards practical and robust DNA-based data archiving using the yin-yang codec system. Nat Comput Sci, 2022, 2: 234–242
6 Welzel M, Schwarz P M, Löchel H F, et al. DNA-Aeon provides flexible arithmetic coding for constraint adherence and error correction in DNA storage. Nat Commun, 2023, 14: 628
7 Goldman N, Bertone P, Chen S, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature, 2013, 494: 77–80
8 Press W H, Hawkins J A, Jones Jr S K, et al. HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints. Proc Natl Acad Sci USA, 2020, 117: 18489–18496
9 Antkowiak P L, Lietard J, Darestani M Z, et al. Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. Nat Commun, 2020, 11: 5345
10 Hawkins J A, Jones Jr S K, Finkelstein I J, et al. Indel-correcting DNA barcodes for high-throughput sequencing. Proc Natl Acad Sci USA, 2018, 115: E6217
11 Zhao X, Li J, Fan Q, et al. Composite Hedges Nanopores codec system for rapid and portable DNA data readout with high INDEL-Correction. Nat Commun, 2024, 15: 9395
12 Anavy L, Vaknin I, Atar O, et al. Data storage in DNA with fewer synthesis cycles using composite DNA letters. Nat Biotechnol, 2019, 37: 1229–1236
13 Grass R N, Heckel R, Puddu M, et al. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angew Chem Int Ed, 2015, 54: 2552–2555
14 Blawat M, Gaedke K, Hütter I, et al. Forward error correction for DNA data storage. Procedia Comput Sci, 2016, 80: 1011–1022
15 Srinivasavaradhan S R, Gopi S, Pfister H D, et al. Trellis BMA: coded trace reconstruction on IDS channels for DNA storage. In: 2021 IEEE International Symposium on Information Theory (ISIT). New York: IEEE, 2021. 2453–2458
16 Chen Z, Zhou W, Qiao S, et al. Highly accurate fluorogenic DNA sequencing with information theory-based error correction. Nat Biotechnol, 2017, 35: 1170–1178
17 Welzel M, Dreßler H, Heider D. Turbo autoencoders for the DNA data storage channel with Autoturbo-DNA. iScience, 2024, 27: 109575
18 Zheng Y, Cao B, Zhang X, et al. DNA-QLC: an efficient and reliable image encoding scheme for DNA storage. BMC Genomics, 2024, 25: 266
19 Bar-Lev D, Orr I, Sabary O, et al. Scalable and robust DNA-based storage via coding theory and deep learning. Nat Mach Intell, 2025, 7: 639–649
20 Ruan C, Yang L, Han R, et al. DSI-ResCNN: a framework enhancing the error-tolerance capacity of DNA storage for images. IEEE Access, 2025, 13: 50777–50793
21 Vitak S. Technology alliance boosts efforts to store data in DNA. Nature, 2021, doi: 10.1038/d41586-021-00534-w
22 Beaucage S L, Caruthers M H. Deoxynucleoside phosphoramidites—a new class of key intermediates for deoxypolynucleotide synthesis. Tetrahedron Lett, 1981, 22: 1859–1862
23 Fodor S P A, Read J L, Pirrung M C, et al. Light-directed, spatially addressable parallel chemical synthesis. Science, 1991, 251: 767–773
24 Lietard J, Leger A, Erlich Y, et al. Chemical and photochemical error rates in light-directed synthesis of complex DNA libraries. Nucleic Acids Res, 2021, 49: 6687–6701
25 Singh-Gasson S, Green R D, Yue Y, et al. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat Biotechnol, 1999, 17: 974–978
26 Blanchard A P, Kaiser R J, Hood L E. High-density oligonucleotide arrays. Biosens Bioelectron, 1996, 11: 687–690
27 Ma Y, Zhang Z, Jia B, et al. Automated high-throughput DNA synthesis and assembly. Heliyon, 2024, 10: e26967
28 Su X, Li X, Zhang Y, et al. Three-phase contact line confined dense nanoparticle array for high-capacity DNA synthesis. Chem Eng Sci, 2023,

281: 119135
29 Egeland R D. Electrochemically directed synthesis of oligonucleotides for DNA microarray fabrication. Nucleic Acids Res, 2005, 33: e125
30 Maurer K, Cooper J, Caraballo M, et al. Electrochemically generated acid and its containment to 100 micron reaction areas for the production of DNA microarrays. PLoS One, 2006, 1: e34
31 Nguyen B H, Takahashi C N, Gupta G, et al. Scaling DNA data storage with nanoscale electrode wells. Sci Adv, 2021, 7: eabi6714
32 Carter S R. Benchtop DNA Synthesis Devices: Capabilities, Biosecurity Implications, And Governance. Washington: Nuclear Threat Initiative, 2022
33 Lee H, Wiegand D J, Griswold K, et al. Photon-directed multiplexed enzymatic DNA synthesis for molecular digital data storage. Nat Commun, 2020, 11: 5246
34 Li K, Lu X, Liao J, et al. DNA-DISK: automated end-to-end data storage via enzymatic single-nucleotide DNA synthesis and sequencing on digital microfluidics. Proc Natl Acad Sci USA, 2024, 121: e2410164121
35 Zhirnov V, Zadegan R M, Sandhu G S, et al. Nucleic acid memory. Nat Mater, 2016, 15: 366–370
36 Lange M J, Burke D H, Chaput J C. Activation of innate immune responses by a CpG oligonucleotide sequence composed entirely of threose nucleic acid. Nucleic Acid Ther, 2019, 29: 51–59
37 Liu F, Li J, Zhang T, et al. Engineered spore-forming *Bacillus* as a microbial vessel for long-term DNA data storage. ACS Synth Biol, 2022, 11: 3583–3591
38 Lin K N, Volkel K, Cao C, et al. A primordial DNA store and compute engine. Nat Nanotechnol, 2024, 19: 1654–1664
39 Takahashi C N, Ward D P, Cazzaniga C, et al. Evaluating the risk of data loss due to particle radiation damage in a DNA data storage system. Nat Commun, 2024, 15: 8067
40 Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA, 1977, 74: 5463–5467
41 Bentley D R, Balasubramanian S, Swerdlow H P, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 2008, 456: 53–59
42 Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol, 2008, 26: 1135–1145
43 Pfeiffer F, Gröber C, Blank M, et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Sci Rep, 2018, 8: 10950
44 Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 2011, 27: 2987–2993
45 Wetzl C, Soukarie D, Yeregui Elosua J, et al. Integrating DNA-based memory in water-resistant electrospun polymer fibers for nondestructive data retrieval. ACS Appl Mater Interfaces, 2025, 17: 46089–46098
46 Imburgia C, Organick L, Zhang K, et al. Random access and semantic search in DNA data storage enabled by Cas9 and machine-guided design. Nat Commun, 2025, 16: 6388
47 Huang X, Hou Z, Qiang W, et al. Towards next-generation DNA encryption via an expanded genetic system. Natl Sci Rev, 2025, 12: nwae469
48 Chen W, Han M, Zhou J, et al. An artificial chromosome for data storage. Natl Sci Rev, 2021, 8: nwab028
49 Qu G, Yan Z, Chen X, et al. DNA data storage for biomedical images using HELIX. Nat Comput Sci, 2025, 5: 397–404
50 Ding L, Wu S, Hou Z, et al. Improving error-correcting capability in DNA digital storage via soft-decision decoding. Natl Sci Rev, 2024, 11: nwad229
51 Zrihan A, Yaakobi E, Yakhini Z. Studying the cycle complexity of DNA synthesis. In: 2024 IEEE Information Theory Workshop (ITW). New York: IEEE, 2024. 633–638
52 Cao R, Chen X. Optimizing sequencing coverage depth in DNA storage: insights from DNA storage data. arXiv, 2025, 2501.06801
53 Chaykin G, Sabary O, Furman N, et al. DNA-storalator: a computational simulator for DNA data storage. BMC BioInf, 2025, 26: 204
54 Schwarz M, Welzel M, Kabdullayeva T, et al. MESA: automated assessment of synthetic DNA fragments and simulation of DNA synthesis, storage, sequencing and PCR errors. Bioinformatics, 2020, 36: 3322–3326
55 Li Y, Wang S, Bi C, et al. DeepSimulator1.5: a more powerful, quicker and lighter simulator for nanopore sequencing. Bioinformatics, 2020, 36: 2578–2580

Research progress of the reliability of DNA data storage

YUE XueQing[1†], ZHENG ZhiYi[1†], CAO RuiYing[1], ZHOU PengHua[1] & CHEN Xin[1,2*]

[1] *Center for Applied Mathematics, Tianjin University, Tianjin 300072, China*
[2] *State Key Laboratory of Synthetic Biology, Tianjin University, Tianjin 300072, China*
† *Contributed equally to this work*
* *Corresponding author, E-mail: chen_xin@tju.edu.cn*

Driven by the rapid growth of the Internet, artificial intelligence, and large-scale models, global data volume is increasing exponentially, while traditional silicon-based storage is approaching its physical and economic limits in terms of density, energy consumption, cost, and lifespan. As a novel information storage medium, DNA offers ultra-high storage density, extremely long preservation lifetime, and low maintenance energy requirements, making it a promising candidate for large-scale data storage in the future. In recent years, the overall reliability of DNA data storage has been significantly improved, with numerous experiments achieving zero-error reconstruction. By optimizing encapsulation and storage conditions, the stability of DNA storage can theoretically reach half-lives of tens of thousands of years. This article systematically reviews the progress of reliability research in DNA data storage from three perspectives: coding strategies, biochemical processes, and decoding mechanisms. It covers novel coding schemes such as fountain codes, HEDGES codes, and Yin-Yang codes, synthesis and sequencing technologies suitable for DNA storage, and encoding/decoding optimizations addressing high error rates and data disorder. Furthermore, it discusses the potential of deep learning, simulation tools, and system integration in enhancing reliability, and provides a perspective on the future applications of DNA storage.

DNA storage, data encoding, DNA synthesis, DNA sequencing, data decoding

doi: 10.1360/SSV-2025-0200