第 13 卷 第 3 期
2024 年 5 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 13 No. 3
May 2024

引文格式：
王钰, 许佳欣, 林明香, 等. DNA 信息存储核心技术及其发展 [J]. 集成技术, 2024, 13(3): 102-115.
Wang Y, Xu JX, Lin MX, et al. Core technology and development of DNA information storage [J]. Journal of Integration Technology, 2024, 13(3): 102-115.

DNA 信息存储核心技术及其发展

王 钰[1,2] 许佳欣[2,3] 林明香[4] 崔君婷[2] 戴俊彪[2,5] 王 洋[1*] 黄小罗[2*]

[1](中国科学院深圳先进技术研究院 云计算研究中心 深圳 518055)
[2](中国科学院深圳先进技术研究院 深圳合成生物学创新研究院 广东省合成基因组学重点实验室 深圳市合成基因组学重点实验室 深圳 518055)
[3](深圳市人民医院呼吸与危重症医学科 呼吸疾病研究所 深圳 518020)
[4](中国科学院深圳先进技术研究院 深圳 518055)
[5](中国农业科学院深圳农业基因组研究所(岭南现代农业科学与技术广东省实验室深圳分中心) 深圳 518120)

摘 要 在过去的几十年里，互联网技术的发展和普及推动人类进入了数字信息时代，互联网已成为人类生活的重要组成部分。随着数字化生活方式的到来，人们每时每刻都在产生大规模的数字信息，如何将这些信息进行便捷有效的存储是个必须面对的问题。针对数据存储面临的种种问题，该文从现有的存储方式和存储介质出发，对当前存储领域进行深入研究，分析了 DNA 作为未来大数据存储介质的优势，以及 DNA 存储的核心技术和潜在的应用前景。另外，该文通过对 DNA 信息存储的核心技术进行剖析和讨论，提出了未来 DNA 信息存储发展的趋势和见解，以期对 DNA 信息存储发展提供新的思路。

关键词 信息存储；DNA 存储；存储介质
中图分类号 TP 333；Q 819 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20231120001

收稿日期：2023-11-20 **修回日期**：2024-04-08
基金项目：国家重点研发计划项目(2021YFF1201700)；国家档案局科技项目(2022-X-006)；广东省合成基因组学重点实验室项目(2023B1212060054)；深圳合成基因组学重点实验室项目(ZDSYS201802061806209)；深圳市科技计划项目(RCYX20221008092950122)
作者简介：王钰，硕士，研究方向为 DNA 信息存储；许佳欣，博士，研究方向为芯片测序技术与应用；林明香，硕士，研究方向为档案 DNA 信息存储；崔君婷，工程师，研究方向为软件工程；戴俊彪，研究员，博士生导师，研究方向为合成基因组学与合成生物使能技术；王洋(通讯作者)，研究员，博士生导师，研究方向为云计算和大数据分析处理，E-mail：yang.wang1@siat.ac.cn；黄小罗(通讯作者)，高级工程师，博士生导师，研究方向为 DNA 数据存储与合成生物使能技术，E-mail：huangxl@siat.ac.cn。

Core Technology and Development of DNA Information Storage

WANG Yu[1,2]　XU Jiaxin[2,3]　LIN Mingxiang[4]　CUI Junting[2]　DAI Junbiao[2,5]
WANG Yang[1*]　HUANG Xiaoluo[2*]

[1](*Research Center for Cloud Computing*, *Shenzhen Institute of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)

[2](*Shenzhen Key Laboratory of Synthetic Genomics*, *Guangdong Provincial Key Laboratory of Synthetic Genomics*, *Shenzhen Institute of Synthetic Biology*, *Shenzhen Institute of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)

[3](*Institute of Respiratory Diseases*, *Department of Pulmonary and Critical Care Medicine*, *Shenzhen People's Hospital*, *Shenzhen* 518020, *China*)

[4](*Shenzhen Institute of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)

[5](*Shenzhen Branch*, *Guangdong Laboratory for Lingnan Modern Agriculture*, *Agricultural Genomics Institute at Shenzhen*, *Chinese Academy of Agricultural Sciences*, *Shenzhen* 518120, *China*)

[*]*Corresponding Authors*: *yang.wang*1*@siat.ac.cn*; *huangxl@siat.ac.cn*

Abstract　Over the past few decades, the rapid development and widespread adoption of internet technology have propelled humanity into the digital information age, the internet has evolved into a crucial component of human life. With the emergence of the digital lifestyle, individuals are continuously generating massive amounts of digital information. Effective and convenient storage of this information is regarded as a significant challenge that needs to be overcome. Starting with introducing the existing storage methods and media, this article analyzes the current state of the storage field, followed by delving into the advantages, core technologies, and the potential applications of DNA as a big data storage medium in the coming days. Furthermore, this report proposes the future development trends and gives insights into DNA-based information storage, with aiming to offer new thoughts for the advancement of DNA-based data storage technology.

Keywords　information storage; DNA storage; storage medium

Funding　This work is supported by National Key Research and Development Program of China (2021YFF1201700), Scientific and Technological Program by State Archive Administration (2022-X-006), Guangdong Provincial Key Laboratory of Synthetic Genomics (2023B1212060054), Shenzhen Key Laboratory of Synthetic Genomics (ZDSYS201802061806209), Shenzhen Science and Technology Program (RCYX20221008092950122)

1　引　言

人类文明的进步往往带动科学技术的发展。早期，人类记录信息的方式十分简单，在石头上、竹片上、骨头上就能记录所有需要的信息。但是随着社会的发展，人类开启了虚拟的网络世界，进入了数字化时代，需要记录存储的数据量呈指数级增长。数据产生的速度之快，现有的主流存储设备很快将难以应对，现有的主流存储方式包括硬盘、磁盘、光盘等，存储密度远远不能满足这些海量数据的需求。

DNA 作为生物界中被用来存储遗传信息的主要载体，具有存储密度高、稳定性强、存储周期长等特性。亿万年来，生物的生长发育调控和表观性状等信息都被存储在 DNA 中，在对生物遗传学的研究过程中，人类逐渐揭开了

DNA 的神秘面纱。随着生物工程技术的发展，人类已经能成功地读取和修改 DNA 中的信息，这也为 DNA 作为信息存储介质提供了基础支撑。近年来，随着 DNA 合成和测序技术的进步，一方面使得 DNA 实现超大规模的数据存取成为可能，另一方面控制了合成和测序的成本。与现有的存储介质相比，DNA 的信息存储密度高达 4.6×10^{17} bytes/mm^3，比其他存储介质高几个数量级[1-2]。此外，利用 DNA 进行数据存储还有低能耗的优点。同时，DNA 因其纳米结构而在生物传感器、药物递送和生物计算等方面均有应用。

DNA 数据存储主要利用 DNA 的 4 个编码碱基 A、T、C、G 对信息进行编码存储。DNA 信息存储的一般步骤是使用 DNA 信息编码算法对数据进行编码，将数据信息转换成 DNA 序列信息。对于保真度需求较高的数据可以添加纠错码，把 DNA 序列分割成 DNA 序列片段，并使用纠错算法生成纠错碱基，将纠错碱基加入 DNA 序列片段中。然后通过 DNA 序列合成技术将 DNA 序列片段进行合成，合成后的产物又能存储在细胞内或其他存储载体上。信息的读取和恢复是通过 DNA 测序技术将存储的 DNA 提取测序，然后借助纠错算法和 DNA 解码算法进行恢复。本文将通过对现有存储领域的存储方式和存储介质进行分析比较，并结合 DNA 存储技术的详细介绍，讨论 DNA 作为数据存储介质在未来的发展趋势和展望。

2 数据存储面临的问题及 DNA 数据存储优势

首先，数据存储需要各种存储介质作为数据载体，其中就包括光盘、磁盘、硬盘等设备。但是由于这些设备的材料和结构不同，在长期的存储过程中会出现老化和损坏等问题，所以，存储周期越长、越容易保存的介质越适用。其次，大规模的数据量会使得数据维护的成本提高，如纸张、光盘、胶片等存储介质在自然环境中极易被损坏，需要付出更多的维护成本为这些设备提供适宜的保存环境。再次，数据存储需要较高的保密性和安全性，网络和电子设备中的数据正在遭受网络黑客和病毒的威胁。最后，数字化发展使得信息交流的频率越来越高，这就要求数据的传输和备份速度达到较高的水平。以上就是数据存储所面临的一些主要问题，包括介质保存周期、信息维护成本、保密安全和信息传输等方面。

国际数据公司的研究报告显示，到 2025 年，全球产生的数据量会达到 175 ZB[3]（图 1），上述的传统介质已经无法满足这一数据增长带来的存储需求。DNA 是天然的信息载体，与传统的信息存储介质相比，具有极高的信息存储密度，理论密度达到 460 EB/g，意味着只要 1 kg 的 DNA 就能存储当前全世界所有的信息[4]。2018 年，Organick 等[2]将 35 个文件 200 MB 的数据存入 DNA 中。此外，在适宜的环境下，DNA 具有超长的存储周期，其半衰期可达几千年，与现有的存储介质相比，有极大优势。DNA 存储在体外能保存数十万年，如从古生物中仍能提取出该生物的 DNA。另外，Grass 等[5]通过实验推算，

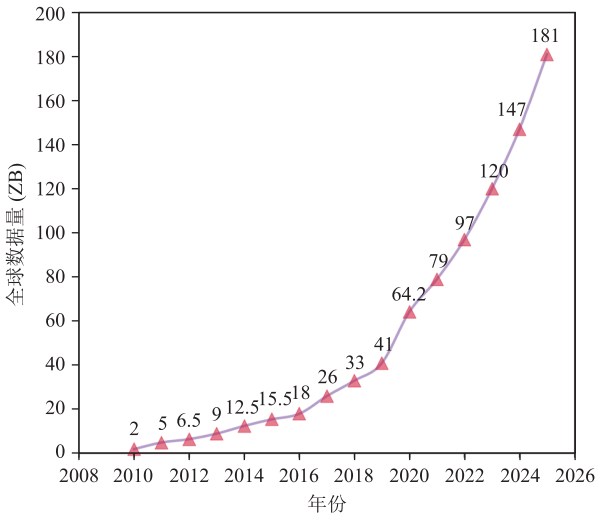

图 1 全球数据增长趋势[3]

Fig. 1 Global data growth trends[3]

将 DNA 封装在二氧化硅固体中于 9 ℃ 能保存 2000 年，在－18 ℃ 能保存 200 万年之久。作为生物信息载体，DNA 还具有体内存储的优势，可以通过转基因技术将编码信息导入生物体中，在需要的时候可以通过引物探针提取。此外，DNA 还具有能耗低、维护成本低的特点。据研究，在 DNA 中存储 10^9 GB 的数据，功率不超过 0.1 W[6]。DNA 信息存储与传统存储方式的性能及优势对比如表 1 所示。

3　传统数据存储介质

3.1　磁存储

1898 年，Valdemar 发明了电报，首次证明了磁介质可以用来记录信息。直到 1953 年，IBM 公司才推出了第一台磁带机 IBM 727 和第一个硬盘驱动器 IBM 305，开启了磁盘存储的时代。1973 年，IBM 研制出温彻斯特(Winchester)硬盘，是现代硬盘的原型。最初的 RAMAC 由 50 片 24 in(1 in＝0.025 4 m)的盘片组成，重达 1 t，可存储 5 MB 信息。随着技术发展，硬盘体积不断缩小，容量大幅提升。2021 年，希捷推出单盘容量高达 20 TB 的企业级硬盘[11]。但传统硬盘的存储密度已接近极限(约 1 TB/in^2)。磁存储技术发展至今，产生了许多的记录方式，如叠瓦式磁记录、二维磁记录、微波辅助磁记录及热辅助磁记录等。研究表明，全球互联网每年传输的流量数据约有 1 ZB，其中 80%～90% 的数据被作为冷数据存储，不会被访问。作为现今最通用的冷数据存储介质，硬盘在今天依旧流行，但其缺点是耗能大[12]。与硬盘相比，磁带具有更低的能耗和存储成本。早在 20 世纪 50 年代，磁带就被作为纸带和打孔卡片的替代品，用作计算机的存储设备[13]。尽管磁带的存储能力不如硬盘，但是作为最早的数字信息存储介质之一，磁带的存储能力也从原来的 GB 级(10^9 bytes)发展到了 TB 级(10^{12} bytes)[7]。综合磁带的成本、能耗、存储能力和存储周期等因素，磁带作为冷数据存储的重要介质仍被广泛使用。另外，磁盘作为计算机中重要的冷数据存储介质，基于磁盘的文件系统能够高效利用磁盘空间和组织文件[14]。

3.2　光存储

1972 年，荷兰飞利浦公司首先开发出激光视盘，并于 1978 年投入市场。1982 年，飞利浦公司和日本索尼公司推出了第一张激光数字唱片[15]，并于 1984 年制定了针对计算机存储的黄皮书标准。随着 1988 年只读光盘文件系统标准 ISO 9660 的发布[16]，光盘不仅用于唱片，也作为大容量数据存储工具流行起来。最早问世的光盘的使用波长为 780 nm，数值孔径为 0.45 mm，存储容量一般为 650～750 MB。之后，高密度数字视频光盘使用波长为 650 nm 的激光，数值孔径为 0.6 mm，将光盘单面单层容量提升到了 4.7 GB(DVD-9 为单面双层，容量为 8.5 GB)。蓝光光碟使用波长为 405 nm 的激光，将数值孔径提高到 0.85 mm，其单面单层容量可达 25 GB。

表 1　传统介质与 DNA 参数比较[7-10]

Table 1　Comparison of traditional media and DNA parameters[7-10]

介质	保存周期	存储密度 (bits/cm^3)	每 GB 数据使用功率 (W)
DNA	＞1 000 年	～10^{19}	＜0.1
硬盘	＜10 年	～10^{13}	0.04
闪存	～10 年	～10^{16}	0.01～0.04
DRAM	～64 ms	～10^{13}	0.04
只读 CD	～100 年	～10^{12}	＜0.1
磁带	～30 年	～10^{12}	＜0.1

光存储技术使用激光照射存储介质，使其某些性质(反射率、反射光极化方向等)发生改变，以记录信息，读取时再通过介质局部点对激光的不同反应识别信息。通常把采用非磁性介质的光存储技术称为第一代光存储技术，其特点是写入的内容不可擦除；而采用磁性材料作为介质的光存储技术被称为第二代光存储技术，即磁光存储技术，主要特点是可擦除重写。然而，受限于二维平面的存储点尺寸和制作工艺，继续降低波长、增大数值孔径或增加刻录层数变得越来越困难，传统光存储技术的存储密度已临近上限，从二维向多维发展是光存储技术未来发展的必然趋势。当前可复用维度有介质的三维空间、波长、光强及偏振等[17-21]，相应技术有“全息光存储技术”“多波长多阶光存储技术”“双光束超分辨率光存储技术”“五维光存储技术”等。

3.3 半导体存储

1967 年，IBM 公司的研究员 Robert Dennard 提交了 DRAM 的专利；1970 年，美国 Intel 公司推出第一款商用 DRAM 芯片 Intel 1103。此后，DRAM 以符合摩尔定律的速度开始了快速的产品迭代。根据集成电路工艺的不同，半导体存储器可分为双极型半导体存储器和金氧半场效晶体管半导体存储器。前者具有高速的优点，后者具有集成度高、制作简单、成本低、功耗低等优点，因此，金氧半场效晶体管半导体存储器的应用更为广泛。1984 年，日本东芝公司的桀冈富士雄博士提出了闪存存储器概念。1988 年，美国 Intel 公司提出 NOR Flash 技术，并推出了第一款商用闪存芯片。1989 年，日本东芝公司提出 NAND Flash，大幅降低了制作成本。作为现在主流的两种闪存技术，NAND Flash 因读写速度快、容量大、价格低廉、寿命长等优点而广泛应用于存储卡、U 盘、固态硬盘等存储设备，而 NOR Flash 支持芯片内执行的特点使其在物联网、可穿戴设备等领域占有一席之地。半导体存储 DRAM 和 NAND Flash 都被广泛应用于固态硬盘的存储，固态硬盘拥有与磁盘硬盘完全不同的存储系统，相同的是，通过文件存储系统都能够高效地发挥介质的存储能力[22]。

4 DNA 存储

4.1 DNA 存储时间轴

自 DNA 发现以来，DNA 信息存储经历了数次里程碑式的发展，并在近二三十年内迅速发展。1869 年，DNA 第一次被人类发现，但直到 1944 年才被细菌学家艾弗里首次证实是生物体内的遗传物质[23]。人类经过不断的探索，发现了 DNA 的双螺旋结构，并且能够对 DNA 进行测序和体外合成。1996 年，Davis[24]通过实验验证了 DNA 信息存储的可能性，成功将 35 bits 的图像编码后存入 DNA 中，并转入大肠杆菌，完成了信息的读取。2001 年，Bancroft 等[25]成功将“IT WAS THE BEST OF TIMES IT WAS THE WORST OF TIMES,”和“IT WAS THE AGE OF FOOLISHNESS IT WAS THE EPOCH OF BELIEF.”两句话编码成 DNA 序列，并通过聚合酶链式反应扩增和测序解码将两句话恢复出来。随后，2012 年，Church 等[1]将总计 659 kB 的数据存储在 DNA 中；2013 年，Goldman 等[26]将 739 kB 的数据存储在 DNA 中，这两项工作使 DNA 存储技术达到了初步应用的水平，实现了重大的技术突破，也使得 DNA 数据存储被更多研究人员关注。随后，DNA 数据存储进入了快速发展时期，DNA 存储技术体系也逐渐清晰。DNA 存储流程中一系列生化反应过程所带来的数据传输失真和不稳定导致数据的错误率较高。2015 年，Grass 等[5]使用 RS 码将 0.08 MB 的数据进行存储。2016 年，Blawat 等[27]使用前向纠错码实现了 22 MB 数据的存储和纠错。紧接着，2017 年，Erlich 等[28]将喷泉码引入 DNA 数

据存储中，不仅提高了数据存储密度，还提供了更加稳定的纠错系统。2018 年，Organick 等[2]将 35 个文件(200 MB 的数据)存入 DNA 中，并通过实验验证了一个能够实现数据随机存取的大规模 DNA 存储系统。图 2 列举了从 DNA 作为遗传物质被发现直到 2020 年以前，在 DNA 存储领域的一些重要事件。2020 年以后，国内外在 DNA 信息存储领域的研究成果众多，包括编码算法、纠错算法、存储系统、加密算法、实际应用等方面[29-33]。随着 DNA 数据存储技术的发展和配套的生物工程技术的进步，DNA 有望成为未来重要的信息存储介质，DNA 存储技术也将走进人类的日常生活。

4.2　DNA 信息存储体系

DNA 信息存储体系主要包括信息编解码、信息读写和 DNA 保存等几个重要部分(如图 3 所示)。DNA 信息存储的具体实施步骤如下：首先，对信息进行编码，转换成 4 个碱基组成的碱基序列；其次，利用 DNA 合成技术将编码序列进行合成；最后，将合成好的序列进行信息存储。另外，信息的读取操作是通过 DNA 测序完成的，通过测序仪可完成高通量、高精度的序列读取过程。除了这些通用步骤外，还需要进行数据的随机访问和纠错，以及信息的解码等操作。

4.2.1 信息编解码

DNA 数据存储以 DNA 链为载体，将数据编码为 A、T、C、G 4 个碱基组成的有一定长度和顺序的序列。在信息科学中，可以用信息熵衡量一个随机变量存储的信息量，同样，在 DNA 信息存储中，每个碱基包含的信息量也可以用信息熵表示。根据公式(1)可知，每个碱基最多可表示两个 bit 的信息(当 log 的底数为 2 时，可表示以二进制为单位，$p(i)$为碱基 i 的概率)[34]。在 DNA 测序过程中，如果 DNA 的 GC 含量相对较大或较小，则都会导致测序结果的

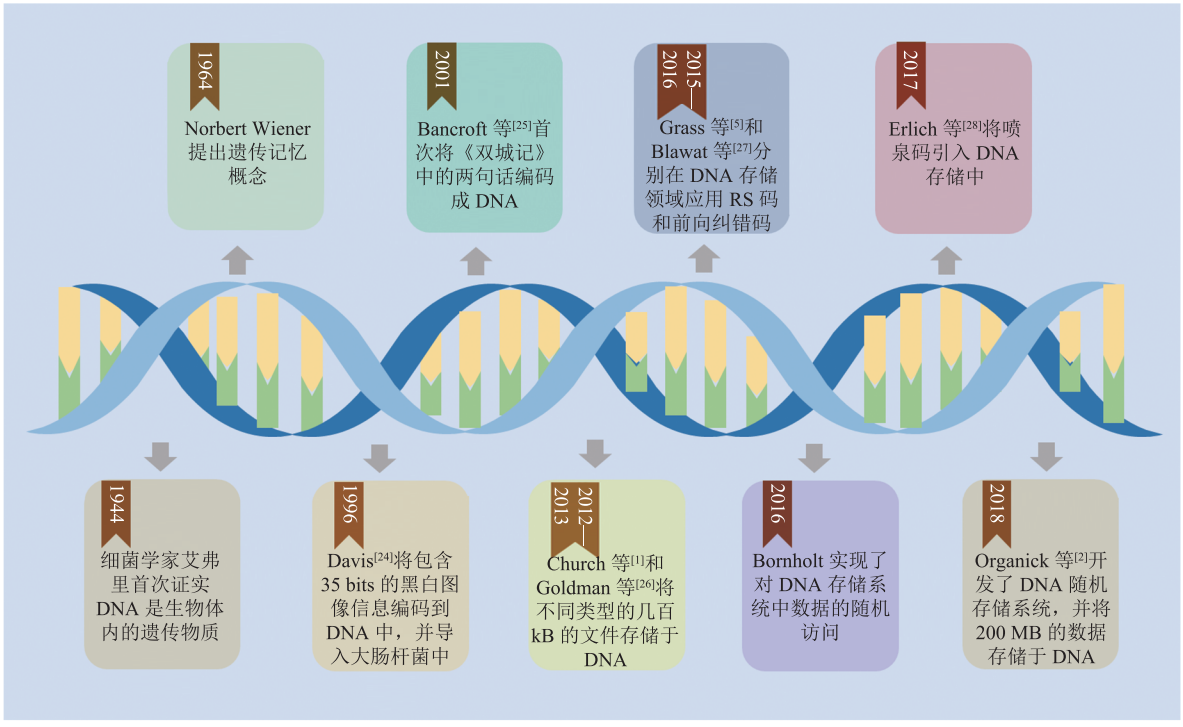

图 2　DNA 信息存储发展历史

Fig. 2　The development history of DNA-based information storage

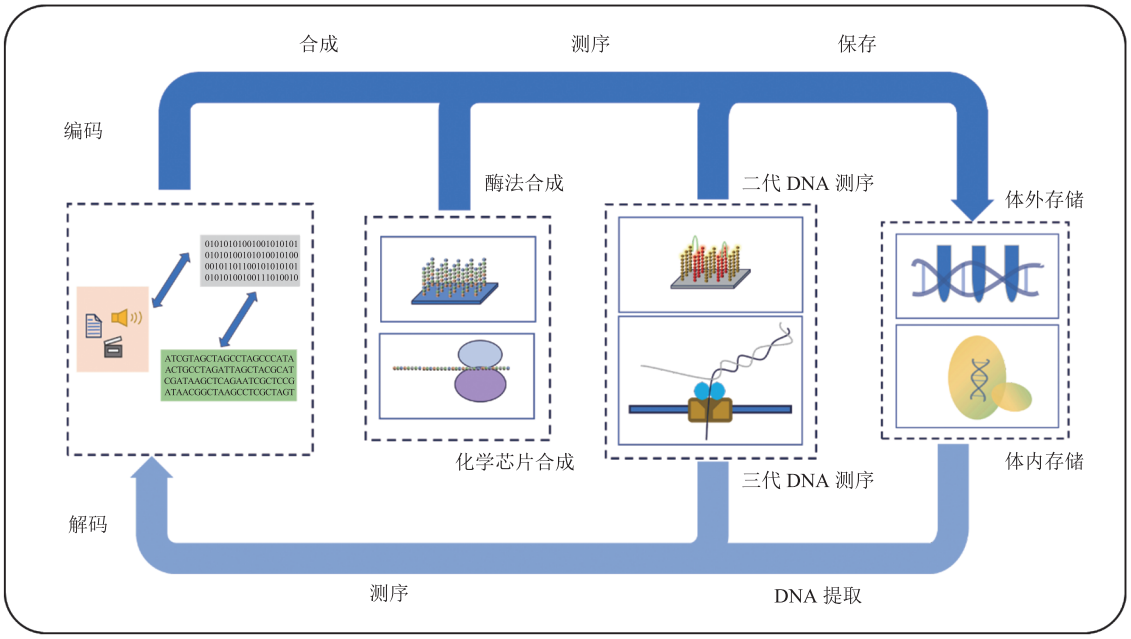

图 3 DNA 信息存储流程

Fig. 3 The process of DNA-based information storage

错误增加，所以，在进行碱基编码时，通常会控制 GC 含量，并使其接近 50%，以降低错误发生的概率。

$$H=\sum_i^{\mathrm{A,T,C,G}} P(i)\log_2(i)\leqslant \log_2\sum_i^{\mathrm{A,T,C,G}} p(i)\frac{1}{p(i)}=\log_2 4=2\ \mathrm{bits} \tag{1}$$

现代通用计算机上的编码都采用二进制形式，以高低不同的两种电位表示 0 和 1 两个数字。从 DNA 存储发展至今，碱基编码大致分为如下几种编码方式：双碱基编码、三碱基编码、四碱基编码和其他编码。双碱基编码于 2012 年由 Church 等[1]首次使用，用两个碱基表示一位二进制数(A 和 G 表示 0，C 和 T 表示 1)。双碱基编码能够很轻松地将碱基编码和各种二进制文件系统进行兼容，使格式转换和操作更灵活简单，同时，在 GC 含量和均聚物的控制方面也更方便。但相应地，碱基的信息利用率不高，信息密度低。三碱基编码由 Goldman 等[26]于 2013 年提出，该编码方式又称为三进制霍夫曼编码，利用了霍夫曼压缩编码的特性，提高了碱基的信息熵。碱基编码属于信源编码，在信源编码中，信息熵又等于平均码长。平均码长的计算方法如公式(2)所示。

$$L=\sum_i^{\mathrm{A,T,C,G}} p(i)l(i) \tag{2}$$

其中，$p(i)$为 i 的概率分布；$l(i)$为 i 的码长。从公式(2)可知，要使平均码长 L 最小，就需要将最大的码长分配给最小概率的随机事件。霍夫曼编码就是根据这一原理设计的，将出现频率最高的字母用最短的码字编码，以达到信息熵最大的效果。另外，最常用的编码方法是四进制编码：首先将 A、T、C、G 用 0、1、2、3 表示，然后映射成二进制 00、01、10、11。四碱基编码的优点是理论上编码效率能达到 2 bits/nt，使单碱基的信息熵达到理论极限。但是由于这种编码的理论效率太高，其对序列的可控性也较差，很容易出现 GC 含量异常和均聚物较多的情况，所以提高了错误率。除了以上 3 种常用的编码方法外，还有其他编码方法，如华大基因提出的阴阳编码

方法，将信息通过“阴”和“阳”两种编码方式分别编码，并将最后的编码结果进行结合，形成最终的 DNA 编码序列[35]。

4.2.2 错误控制

在 DNA 数据存储系统内部，信息的转换和传递过程会有额外的错误产生，为了提高内部信息传输的保真度，需要进行错误控制，在 DNA 数据存储中又被称为纠错编码。纠错编码的本质是在信息编码之后引入多余的冗余信息。冗余信息能够提高信息的稳定性，使其不易被篡改，但其添加又会导致成本和信息处理的计算开销增加。现有的纠错编码主要有直接冗余编码、喷泉码、RS 码和卷积码等，如图 4 所示。其中，直接冗余编码以 Goldman 等[26]的编码方式为主，该方法首先将长序列分成长度为 100 bp 的片段，然后从起始位置开始每隔 25 bp 进行切分，相邻片段之间有 75 bp 的重叠区域，也就是长序列被分割之后大约产生了 4 倍的冗余。该方法不添加信息编码外的冗余，而是直接提高信息码自身的冗余，虽然有一定的纠错能力和组装优势，但是编码效率和信息密度太低。另外，有不少纠错编码方法都会通过添加 RS 码进行纠错，而 RS 码是一种前向纠错码，具有可以检验和纠正多个随机码元错误的优点，且当码长越长时，RS 码的纠错性能越强。Blawat 等[27]首次在 DNA 编码时使用了 RS 码，添加的 RS 码为 RS(255,223,33)。其中，码长为 255；信息码码长为 223；最小距离为 33。喷泉码在处理删除错误方面具有极大优势。2017 年，Erlich 等[28]将喷泉码引入到 DNA 存储中。他们将输入信息转换成 67 088 个 32 bytes 的片段，每个水滴为 38 bytes。其中，4 bytes 是产生的随机数种子，32 bytes 是有效载荷，2 bytes 为 RS 纠错码。最后，他们在 0 错误率的情况下完全恢复了被保存的数据。其他的复合编码则是多种算法组合的纠错方法。以 Press 等[36]的 HEDGES 为例，该方法将哈希编码和贪婪穷举搜索的解码联用，基于该编码方式能够在人为引入 1.2% 的错误的情况下将数据正确解码。以上算法都仅能应对单碱基错误或者少数位点错误，难以解决大规模错误的情况。为此，Song 等[32]开发了 DBGPS 算法，该算法基于德布鲁因图(de Brujin graph，DBG)和贪婪路径搜索，能够处理各种大规模错误情形，如链的断裂和重排等。DBGPS 分为两个步骤：第一步是对测序数据进行 k-mer 计数，根据 k-mer 之间的联系构建 DBG，通过 k-mer 的覆盖度对构建的 DBG 进行简化处理；第二步是在 DBG 中执行贪婪路径搜索，并选择出候选路径，之后再对候选路径进行内嵌的循环冗余校验码校对，通过校验的候选路径将被保留输出，未通过校验的候选路径则继续进行组装和校验。

4.2.3 信息写入

通过信息编码将数据信息编码成特定的 DNA 序列后，就需要通过 DNA 合成技术将编码好的 DNA 序列进行从头合成，从而实现信息的写入。DNA 合成的首次尝试可以追溯到 20 世纪 50 年代，Michelson 等[37]首次合成了寡聚二核苷酸。直到 20 世纪 80 年代，亚磷酰胺法才被发明，并沿用至今。亚磷酰胺法主要分为 4 个步骤：(1)去除亚磷酰胺上的 DMT(二甲氧基三苯基甲基)基团；(2)将下一个待结合的亚磷酰胺结合到去除 DMT 基团的亚磷酰胺上；(3)将暴露的 5′-OH 乙酰化，防止引入错误；(4)对磷酸三酯进行氧化。通过这 4 个步骤的循环，将碱基一个个加到 DNA 链上，实现 DNA 的从头合成。亚磷酰胺法合成虽然是现在最主流的方法，但仍然存在多种缺陷，并因此限制了该方法的发展，如合成反应的产率随合成长度的增加而急剧降低，在延长效率达到 99% 时，合成 120 bp 的寡核苷酸的理论产率为 30%，但是当合成 200 bp 的寡核苷酸时，产率仅有 13%。化学法合成的 DNA 长度被限制在 200 bp 以内[38]，且亚磷酰胺法的

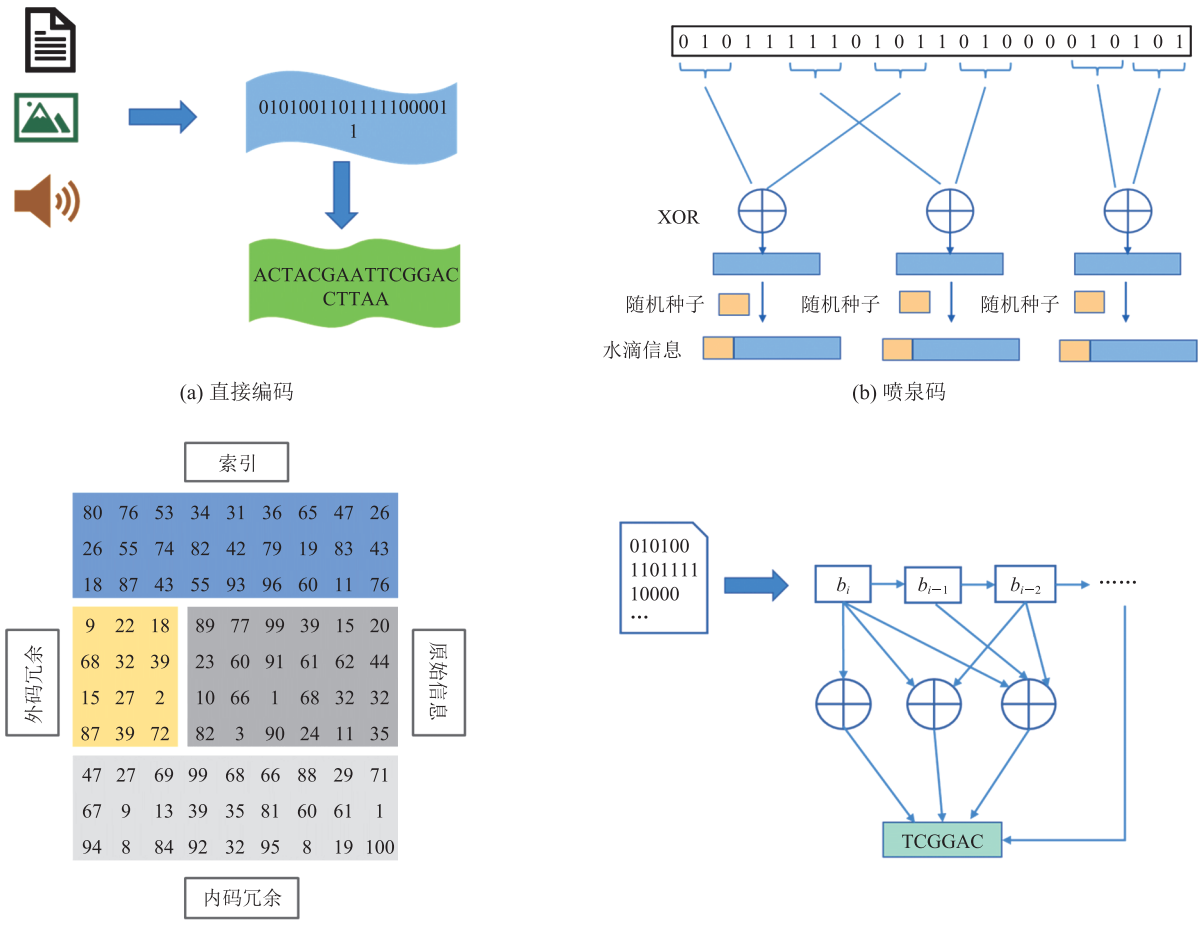

图 4 DNA 信息存储中的纠错编码方式

Fig. 4 Error correction coding methods in DNA information encoding

稳定性差，需要大量有害的有机溶剂，无法合成多重复序列。除了亚磷酰胺法外，还有酶促法。通过酶促反应合成 DNA 既高效又准确。酶促法是不依赖模板的寡聚核苷酸合成方法，通过 TdT（末端脱氧核苷酰转移酶）在 5′到 3′的方向上延长寡聚核苷酸。为了进行精准的延长，该方法采用了可逆终止机制，即在 3′的位置用合成中断终止子或者保护基团修饰的核苷三磷酸（NTP），以此确保每个反应只添加单个核苷酸[39]。与亚磷酰胺法相比，酶促法能够合成更长的寡聚核苷酸链。但是，随着链长度的延长，DNA 链会形成稳定的二级结构[39-40]，对合成和测序都有不利影响，因此，酶促法一般用来合成 3 kB 以内的 DNA 链。2012—2022 年，在 DNA 信息存储的相关工作中，所使用的合成、测序和存储等技术信息如表 2 所示。

现今，高通量的商业合成大多基于芯片，根据使用的保护基团的性质又可分为喷墨打印合成、电化学合成、光刻合成和热合成等。喷墨打印合成将核苷单体打印在芯片上，使其沉积在特定位点，以进行 DNA 合成。喷墨打印的优点是通量高、速度快，大多商业公司的喷墨打印通量都在几十万条序列，合成长度在 300 nt 以内。在通电条件下，电化学合成利用电极周围发生的氧化还原反应改变溶液酸碱度，从而去除核苷酸的保护基团，以进入下一步合成反应。电化学合成

表 2　DNA 信息存储研究工作及技术比较

Table 2　The comparison of research work and technology in DNA information storage

作者	年份	数据量	合成方法	测序平台	测序深度	逻辑密度 (bits/nt)	添加纠错	存储方式
Church 等[1]	2012	650 kB	亚磷酰胺法	Illumina	3 000×	0.60	否	活体外
Goldman 等[26]	2013	630 kB	亚磷酰胺法	Illumina	51×	0.19	是	活体外
Grass 等[5]	2015	80 kB	亚磷酰胺法	Illumina	372×	0.86	是	活体外
Bornholt 等[41]	2016	150 kB	亚磷酰胺法	Illumina	40×	0.57	否	活体外
Blawat 等[27]	2016	22 MB	亚磷酰胺法	Illumina	160×	0.89	是	活体外
Erlich 等[28]	2017	2 MB	亚磷酰胺法	Illumina	10.5×	1.18	是	活体外
Chauhan 等[42]	2021	3.6 kB	亚磷酰胺法	Nanopore	294×	1.72	是	活体外
Organick 等[2]	2018	200 MB	亚磷酰胺法	Illumina & Nanopre	5×	0.81	是	活体外
Nguyen 等[43]	2018	10.5 kB	亚磷酰胺法	Illumina	22×	1.32	否	活体内
Lee 等[39]	2019	18 bytes	酶促法	Nanopre	175×	1.57	否	活体外
Press 等[36]	2020	-	亚磷酰胺法	Illumina	50×	1.20	是	活体外和活体内
Song 等[32]	2022	6.8 MB	亚磷酰胺法	Illumina	28×	1.30	是	活体外

的优点是通量高，缺点是电极间的酸碱度会相互影响，造成合成的错误率较高。光刻合成则利用激光在芯片上的光照催化保护基团酶解或直接脱去特定位点的核苷酸的保护基团，触发合成的继续。光刻合成的限制因素主要包括光的衍射和散射，严重影响了其效率和正确率。热合成利用半导体技术在芯片上集成大量的温控位点，通过控制位点温度激活合成反应。热合成高度依赖控制系统，通过控制系统还能实现纯化和矫正操作，大大提高了合成的产率。

近年来，DNA 合成技术不仅在方法上取得了进步，在合成规模和效率上也有很大提高。早期的 DNA 合成基于固相的亚磷酰胺合成法，虽然精度高，但是通量低、成本高。随后发展出的基于微阵列芯片的电化学合成技术能够并行进行 10^6 个合成反应，极大地提高了合成速度和通量，大大降低了合成所需的成本。Palluk 等[44]提出的酶促合成可实现 10～20 s 一个碱基的合成速度，是亚磷酰胺合成法的几十倍，成本也有望降低几个数量级。由于 DNA 信息存储还没有建立类似于计算机的完全自动化设备，所以，目前的合成速度不能准确地度量数据的吞吐量。不同的 DNA 合成公司使用的合成平台不同，有酶促合成、Gibson 组装和滚环扩增等，合成的 DNA 长度有十几倍的差距，通量也因合成速率的不同而差距较大。

4.2.4 信息读取

要将信息从 DNA 中读取出来，就需要借助 DNA 测序技术。1977 年，Sanger 等[45]使用双脱氧链终止法完成了 DNA 的首次测序实验。该方法使用 DNA 聚合酶扩增模板链，在反应体系中需要添加脱氧核苷酸三磷酸(dNTP)和 4 种带有荧光标记的双脱氧核苷酸磷酸(ddNTP)，双脱氧核苷酸磷酸一旦结合到扩增链上，就会终止伸长反应。通过调整 dNTP 和 ddNTP 的浓度，可使反应体系中产生成百上千种不同长度的产物，利用凝胶电泳可将不同长度的产物分离，最后，通过放射自显影图谱可得到待测的 DNA 序列信息。第一代的 Sanger 测序具有较高的准确率，人类基因组计划就是使用的 Sanger 测序，直到现在，Sanger 测序还在被使用，且仍是准确率最高的测序方法。在 20 世纪 80—90 年代，一些研究团队开发了大规模并行的测序方法，这些测序方法大大提高了测序的通量(从 100 kB 到 GB 和

TB 级别），并且能够实现边合成边测序，这些测序方法被统称为二代测序。具有代表性的二代测序平台有 Roche 454 FLX 平台、Illumina 平台和 SOLID 测序平台[46-48]。虽然二代测序在通量上有了极大提高，但是测序的序列长度依然是该方法的短板——通常，二代测序的读长只有 50～300 bp。为了提高测序的读长长度，三代单分子测序技术应运而生。三代测序主要分为两类，一类是 Pacific Biosciences 公司开发的 SMRT 测序技术[49]，另一类是 Oxford Nanopore Technologies 公司的 Nanopore 测序技术。其中，SMRT 测序将每个合成反应置于 ZMW（零模波导孔）中进行，在保证测序通量的同时，也提高了测序的读长。SMRT 测序的读长长度在 10～25 kB。Nanopore 测序[50]则是利用不同碱基所带电荷的不同，使待测序序列通过带有电位膜的纳米孔，在不同碱基穿过纳米孔时，记录电信号的变化，最后转换成碱基信息。Nanopore 测序能够测序长 DNA 序列，在理论上是不受 DNA 序列长度限制的，但由于现有文库的限制，该方法目前的最长平均测序长度为 23.8 kB，其缺点为准确率不高。Nanopore 的测序通量在 10^9～10^{13} bp，错误率可以控制在 10% 左右，1 kB 的成本比 Illumina 低了一个数量级[51]。

5 DNA 存储发展趋势和挑战

DNA 数据存储作为一项新兴的技术，之所以在近十年发展迅速，主要原因是相关的 DNA 合成和测序技术的发展成熟。DNA 合成技术的发展使大规模的编码序列能够在短时间内以较低的成本进行信息写入，从而使得 DNA 存储的数据量提升了数个数量级，从概念模型一下跳跃到应用层面。同时，DNA 测序技术的成熟将使得测序的成本进一步压缩，不仅能对大规模数据进行快速测序，还能允许大量冗余的存在，从而间接提高了 DNA 存储系统稳定性的可操作空间。目前，DNA 信息存储的成本较高，远高于传统数据存储介质，未来需要开发能够降低 DNA 信息存储成本的技术，如 DNA 酶促合成技术。目前，DNA 酶促合成技术还未成熟，仍具有巨大潜力，未来，酶促合成速度有望达到 20 s 一个碱基，甚至更快，合成速度将会大幅提高，成本也将呈数量级下降，届时，DNA 数据存储将有更大的发挥空间，能够实现真正的商用。但同时，现有的合成技术仍存在不足：一方面，对 DNA 序列的要求较高，需要 GC 含量在 40%～60%，且不能含有较长的单碱基重复；另一方面，合成的 DNA 长度有限制。二代测序中需要进行大量的桥式聚合酶链式反应，在聚合酶链式反应的过程中，计算的 GC 含量将会影响解链速率和二级结构形成等，从而使测序中出现较高的错误率。

DNA 测序技术作为 DNA 存储框架中的信息读取重要组成模块，也有望引领 DNA 存储的变革与进步。速度快和准确率高是实现大规模 DNA 信息读取的必备条件。目前，DNA 存储中常用的读取技术为二代 Illuminate 测序，虽然通量高、成本较低，但是存在大量的浪费，不能有效控制成本。另外，Nanopore 测序一直是被广泛看好的三代测序技术，其优点是能够处理超长读长的序列，缺点是错误率较高。

高效的信息编码技术能够在一定程度上克服合成和测序带来的不利影响。现有的 DNA 编码算法从原来的单纯实现信息编码到需要考虑 GC 含量、均聚物长度、信息密度等因素，编码系统也逐渐成熟。但是现有的编码算法种类较多、标准不一，且都基于二进制的编码方式，无法充分利用 DNA 序列自身的编码特征，编码算法缺乏创新性突破。近年来，DNA 数据存储领域纠错算法的开发成为该领域的研究热点，针对不同问题需求的算法层出不穷，但这些算法往往只在某

个方面表现良好，尚没有通用性。当 DNA 合成和测序等一系列生化技术发展成熟，取得质的飞跃之后，DNA 信息存储领域也将迎来 DNA 信息存储的大数据时代，海量的 DNA 数据将会产生和被处理。此时，对 DNA 数据的管理需求将成为 DNA 信息存储的主要发展方向和动力，类比计算机的大数据时代涌现了一大批新兴的技术，如大数据处理、云计算、分布式存储等，DNA 信息存储相关的大数据存储和计算也将蓬勃发展。

6　总　结

与传统的介质存储相比，DNA 信息存储具有多种天然的优势和巨大的应用前景和潜力，虽然目前仍然存在成本高、操作复杂、实现难度高等问题，但是随着技术的进步，这些问题正在逐步得到解决。从高保真的末端转移酶合成 DNA 的能力可以预见，DNA 合成将实现高通量、高精度和低成本的转变。近年来，人工智能和大语言模型的发展成为研究的热点，也给许多领域注入新的动力，DNA 测序技术与人工智能结合也是当前研究的重要课题，旨在通过 AI 算法提高 DNA 测序的准确度。另外，将机器学习和深度学习等方法应用到信息编码中，也有望提高信息编码的稳定性和纠错能力，以及实现特定类型文件的特定高效编解码。

未来，DNA 存储要想取得进一步的突破，仍需要依赖其他技术的突破。开发各种高保真的 DNA 功能酶，使 DNA 合成、精确编辑效率大大提高，是对未来 DNA 合成技术的需求。高通量、高精度、长读长是测序技术的发展目标。另外，DNA 信息存储软硬件系统的研发也是需要解决的问题之一，未来可开发大规模数据存储的 DNA 存储操作系统和配套的硬件设备，以实现 DNA 信息存储的高度自动化。从 DNA 存储的原理、核心技术和发展趋势等方面进行分析，DNA 依然是未来最具潜力的存储介质之一。DNA 存储领域的各个技术难点涉及生物技术、机械工程、通信工程、软件工程等众多领域，相信在这些领域专家的共同努力下，DNA 信息存储将在未来实现更低成本、高效、便捷的存储，从而超越传统的存储介质。

参 考 文 献

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

[2] Organick L, Ang SD, Chen YJ, et al. Random access in large-scale DNA data storage [J]. Nature Biotechnology, 2018, 36(3): 242-248.

[3] Reinsel D, Gantz J, Rydning J. Data age 2025: the digitization of the world—from edge to core [EB/OL]. (2020-01-29)[2024-04-25]. https://www.i-scoop.eu/big-data-action-value-context/data-age-2025-datasphere/.

[4] Vitak S. Technology alliance boosts efforts to store data in DNA [J]. Nature, 2021.

[5] 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.

[6] 董一名, 孙法家, 武瑞君, 等. DNA 数字信息存储的研究进展 [J]. 合成生物学, 2021, 2 (3): 323-334. Dong YM, Sun FJ, Wu RJ, et al. Research progress on DNA molecules for digital information storage [J]. Synthetic Biology Journal, 2021, 2(3): 323-334.

[7] Dee RH. Magnetic tape for data storage: an enduring technology [J]. Proceedings of the IEEE, 2008, 96(11): 1775-1785.

[8] Kallepalli DLN, Alshehri AM, Marquez DT, et al. Ultra-high density optical data storage in common transparent plastics [J]. Scientific Reports, 2016, 6(1): 26163.

[9] Katayama K, Chinda Y, Shimizu O, et al. Long term stabilities of magnetic tape for data storage in

office environment [J]. Journal of Applied Physics, 2015, 117(17): 17E305.

[10] Zhirnov V, Zadegan RM, Sandhu GS, et al. Nucleic acid memory [J]. Nature Materials, 2016, 15(4): 366-370.

[11] 魏家琦, 柳洋, 赵巍胜. 先进大容量存储技术 [J]. 物理, 2021, 50(12): 812-822.
Wei JQ, Liu Y, Zhao WS. Advanced mass storage technologies [J]. Physics, 2021, 50(12): 812-822.

[12] Bhushan B. Historical evolution of magnetic data storage devices and related conferences [J]. Microsystem Technologies, 2018, 24: 4423-4436.

[13] Harris JP, Phillips WB, Wells JF, et al. Innovations in the design of magnetic tape subsystems [J]. IBM Journal of Research and Development, 1981, 25(5): 691-700.

[14] Qian YJ, Cheng W, Zeng LF, et al. MetaWBC: POSIX-compliant metadata write-back caching for distributed file systems [C] // Proceedings of the SC22: International Conference for High Performance Computing, Networking, Storage and Analysis, 2022: 1-20.

[15] 郑穆. 光存储技术发展趋势 [J]. 电子技术与软件工程, 2018, (4): 188-189.
Zheng M. Development trends of optical storage technology [J]. Electronic Technology & Software Engineering, 2018, (4): 188-189.

[16] Kovarick A. The importance of ISO 9660 [J]. The Electronic Library, 1990, 8(2): 86.

[17] 陈韦良, 张静宇. 大容量光存储的维度扩展 [J]. 光电工程, 2019, 46(3): 67-78.
Chen WL, Zhang JY. Dimension expansion of high-capacity optical data storage [J]. Opto-Electronic Engineering, 2019, 46(3): 67-78.

[18] Ando E, Miyazaki J, Morimoto K, et al. J-aggregation of photochromic spiropyran in Langmuir-Blodgett films [J]. Thin Solid Films, 1985, 133(1-4): 21-28.

[19] Alasfar S, Ishikawa M, Kawata Y, et al. Polarization-multiplexed optical memory with urethane–urea copolymers [J]. Applied Optics, 1999, 38(29): 6201-6204.

[20] Royon A, Bourhis K, Bellec M, et al. Silver clusters embedded in glass as a perennial high capacity optical recording medium [J]. Advanced Materials, 2010, 22(46): 5282-5286.

[21] Parthenopoulos DA, Rentzepis PM. Three-dimensional optical storage memory [J]. Science, 1989, 245(4920): 843-845.

[22] Cheng W, Luo M, Zeng LF, et al. Lifespan-based garbage collection to improve SSD's reliability and performance [J]. Journal of Parallel and Distributed Computing, 2022, 164: 28-39.

[23] Avery OT, Macleod CM, Mccarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III [J]. Journal of Experimental Medicine, 1944, 79(2): 137-158.

[24] Davis J. Microvenus [J]. Art Journal, 1996, 55(1): 70-74.

[25] Bancroft C, Bowler T, Bloom B, et al. Long-term storage of information in DNA [J]. Science, 2001, 293(5536): 1763-1765.

[26] 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.

[27] Blawat M, Gaedke K, Hütter I, et al. Forward error correction for DNA data storage [J]. Procedia Computer Science, 2016, 80: 1011-1022.

[28] Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture [J]. Science, 2017, 355(6328): 950-954.

[29] Hou ZH, Qiang W, Wang XX, et al. "Cell Disk" DNA storage system capable of random reading and rewriting [J]. Advanced Science, 2024, 11(15): 2305921.

[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): 2303197.

[31] Lu MW, Wang Y, Qiang W, et al. Towards high-density storage of text and images into DNA by the "Xiao-Pang" codec system [J]. Science China Life Sciences, 2023, 66(6): 1447-1450.

[32] 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.

[33] Huang XL, Cui JT, Qiang W, et al. Storage-D: a user-friendly platform that enables practical and personalized DNA data storage [J]. iMeta, 2024, 3(2): e168.

[34] Csiszár I, Körner J. Information theory: coding theorems for discrete memoryless systems [M]. Orlando: Cambridge University Press, 2011.

[35] Ping Z, Chen SH, Zhou GY, et al. Towards practical and robust DNA-based data archiving using the yin–yang codec system [J]. Nature Computational Science, 2022, 2(4): 234-242.

[36] Press WH, Hawkins JA, Jones Jr SK, et al. HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints [J]. Proceedings of the National Academy of Sciences, 2020, 117(31): 18489-18496.

[37] Michelson AM, Todd AR. Nucleotides part XXXII. Synthesis of a dithymidine dinucleotide containing a 3′: 5′-internucleotidic linkage [J]. Journal of the Chemical Society (Resumed), 1955: 2632-2638.

[38] Eisenstein M. Enzymatic DNA synthesis enters new phase [J]. Nature Biotechnology, 2020, 38(10): 1113-1115.

[39] 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.

[40] Horgan A, Macevicz SC. Increasing long-sequence yields in template-free enzymatic synthesis of polynucleotides: United States, US20220403436A1 [P]. 2022-12-22[2024-04-08]. https://patents.google.com/patent/US20220403436A1/en.

[41] 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.

[42] Chauhan D, Saxena A, Sharma S. Portable and error-free DNA-based data storage [C] // Proceedings of the 2021 4th International Conference on Recent Developments in Control, Automation & Power Engineering, 2021: 418-421.

[43] Nguyen HH, Park J, Park SJ, et al. Long-term stability and integrity of plasmid-based DNA data storage [J]. Polymers, 2018, 10(1): 28.

[44] Palluk S, Arlow DH, De Rond T, et al. *De novo* DNA synthesis using polymerase-nucleotide conjugates [J]. Nature Biotechnology, 2018, 36(7): 645-650.

[45] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors [J]. Proceedings of the National Academy of Sciences, 1977, 74(12): 5463-5467.

[46] Shokralla S, Spall JL, Gibson JF, et al. Next-generation sequencing technologies for environmental DNA research [J]. Molecular Ecology, 2012, 21(8): 1794-1805.

[47] Goodwin S, Mcpherson JD, Mccombie WR. Coming of age: ten years of next-generation sequencing technologies [J]. Nature Reviews Genetics, 2016, 17(6): 333-351.

[48] Escalona M, Rocha S, Posada D. A comparison of tools for the simulation of genomic next-generation sequencing data [J]. Nature Reviews Genetics, 2016, 17(8): 459-469.

[49] Coupland P, Chandra T, Quail M, et al. Direct sequencing of small genomes on the Pacific Biosciences RS without library preparation [J]. Biotechniques, 2012, 53(6): 365-372.

[50] Quick J, Quinlan AR, Loman NJ. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer [J]. GigaScience, 2014, 3(1): 22.

[51] Dong YM, Sun FJ, Ping Z, et al. DNA storage: research landscape and future prospects [J]. National Science Review, 2020, 7(6): 1092-1107.