DNA数据存储技术研究进展

张淑芳[1] 彭 康[1] 宋香明[1] 张子昱[2] 王汉杰[3]
(天津大学电气自动化与信息工程学院 天津 300072)[1] (天津大学国际工程师学院 天津 300072)[2]
(天津大学生命科学学院 天津 300072)[3]

摘 要 随着计算机技术和网络技术的飞速发展,由此产生的海量数据给传统数据存储方式带来了巨大挑战,因此研究人员开始致力于寻找新一代存储方案。脱氧核糖核酸(Deoxyribonucleic Acid,DNA)作为天然的遗传信息存储介质,具有存储容量大、能耗低和寿命长等优点,有效克服了传统硬盘和计算机存储等方式的不足,故DNA数据存储技术成为信息技术和生物技术交叉领域的研究热点。文中综述了DNA数据存储技术的研究进展,首先对DNA及其存储的理论框架进行了介绍;然后详细阐述了DNA数据存储中的编码技术:二进制数据的压缩编码算法、纠错算法以及二进制数据到DNA 4种碱基的转换方法;最后对现阶段已有的DNA存储方案进行了分析,并对DNA数据存储研究存在的挑战进行了讨论。

关键词 数据存储,DNA,压缩编码,纠错算法,存储密度
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.06.002

Research Progress on DNA Data Storage Technology

ZHANG Shu-fang[1] PENG Kang[1] SONG Xiang-ming[1] ZHANG Zi-yu[2] WANG Han-jie[3]
(School of Electrical and Information Engineering,Tianjin University,Tianjin 300072,China)[1]
(Tianjin International Engineering Institute,Tianjin 300072,China)[2]
(School of Life Sciences,Tianjin University,Tianjin 300072,China)[3]

Abstract With the rapid development of computer technology and network technology,the massive generated data have brought great challenges to traditional data storage methods,so researchers begin to focus on finding a new generation of storage scheme. As a natural genetic information storage medium,Deoxyribonucleic acid (DNA) has advantages of large storage capacity,low energy consumption and long life,which effectively overcome the shortcomings of traditional storage methods,such as hard disk and computer storage. The DNA data storage method has become a research hotspot in the intersection field of information and biotechnology. This paper reviewed the research progress on DNA data storage technology. Firstly,DNA and its theoretical framework of storage are introduced. Then,the coding technologies in DNA data storage are elaborated,which includes compression coding algorithm of binary data,error correction algorithm and conversion method from binary data to four bases of DNA. Finally,the existing DNA storage schemes are analyzed,and the challenges in DNA data storage research are discussed.

Keywords Data storage,DNA,Compression coding,Error correction algorithm,Storage density

1 引言

互联网和人工智能等信息技术的快速发展使得信息量呈指数级增长。据统计,全球数据信息总量将由2018年的30 ZB增长至2025年的163 ZB[1],该趋势将很快超过现有硬盘等存储介质的承受能力。现阶段人们大量使用的便携式硬盘、USB闪存和集成电路等存储体系已逐渐暴露出存储期限短[2]、数据易受环境因素影响[3]、生产设备耗能[3-5]以及污染环境等不足,因此,亟需寻找一种新的数据存储介质。

脱氧核糖核酸作为已知最密集、稳定的数据存储介质之一,具有密度大、能耗低、无磨损和寿命长等潜在优势[6]。此外,DNA与信息存储有众多相似之处:1)均按一定顺序编码存储信息;2)均用符号注明信息段的起始点与终止点;3)均引入纠错码确保信息的完整性。基于以上特点,DNA数据存储应运而生。

DNA数据存储技术是生物技术与信息处理技术共同发

到稿日期:2018-12-23 返修日期:2019-02-24
张淑芳(1979-),女,博士,副教授,主要研究方向为数字图像处理及视频编码,E-mail:shufangzhang@tju.edu.cn(通信作者);**彭 康**(1995-),女,硕士,主要研究方向为生物医学图像处理;**宋香明**(1993-),男,硕士,主要研究方向为生物医学图像处理;**张子昱**(1995-),男,硕士,主要研究方向为生物医学图像处理;**王汉杰**(1984-),男,博士,副教授,主要研究方向为纳米生物医学工程。

展的结果，它开辟了一种新的存储模式，其发展对于节省存储能源及推进大数据存储发展有着重要作用。

DNA 数据存储近年来逐渐成为全球研究的热点[7]。哈佛大学的 Church 研究团队[8]于 2012 年将 650 kB 数据存于 DNA 后，2017 年又将视频文件存入大肠杆菌 DNA[9]；哥伦比亚大学和纽约基因组中心的研究人员于 2017 年提出了一种最大化 DNA 存储技术，利用该技术可将 2.15 亿千兆字节信息存储到 1g DNA 分子内[10]；欧洲生物信息实验室于 2013 年利用 DNA 分子实现了 20 MB 的数据存储[11]；2016 年，微软研究院和华盛顿大学联合将 200 MB 数据存入 DNA[12]，同时微软已计划于 2020 年在数据中心建立基于 DNA 的数据存储系统；2018 年，Catalo 公司与英国剑桥顾问公司共同建造了一个校车大小的机器，计划有朝一日将电影或文档信息存于 DNA 中并用该机器保存 DNA，此外该公司预计在 2019 年推出首个 DNA 数据存储商业服务。DNA 信息存储领域目前已得到各行各业的关注[13]。

本文以 DNA 信息存储为主线，介绍 DNA 基本理论及信息存储框架，从压缩、纠错、转换 3 方面详细说明存储框架中的编码技术，并通过研究对比国内外主流 DNA 信息存储方案，讨论目前该领域存在的主要问题。

2 DNA 数据存储框架

2.1 DNA 简介

DNA 分子是一个以 4 种脱氧核苷酸为单位连接成的长链，这 4 种脱氧核苷酸分别含有 A(腺嘌呤)、T(胸腺嘧啶)、C(胞嘧啶)、G(鸟嘌呤)4 种碱基，这 4 种碱基两两配对，构成 DNA 双链，这种碱基对形式可视为二进制代码的一种形式，如图 1 所示为一条 DNA 双链。

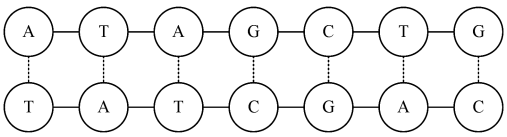

图 1 DNA 模型

Fig. 1 DNA model

双螺旋结构的 DNA 拥有更多可利用空间[14]：单位质量的 DNA 约有 10^{21} 个碱基，可存储 455 EB 信息[15]，此信息量为全球一年信息总量的 1/4；单位体积的 DNA 可存储的信息为整个互联网的 33 倍。表 1 对比了传统存储设备与 DNA 存储各方面的性能参数。由表 1 可知，DNA 单位体积的存储密度是硬盘和存储器的 10^6 倍，是闪存的 10^3 倍。

表 1 传统存储设备与 DNA 存储的性能参数

Table 1 Performance parameters of traditional storage devices and DNA storage

存储设备	DNA	硬盘	闪存	存储器
存储时长	>100 a	10 a	10 a	<64 ms
存储密度/(bits/cm^3)	10^{19}	10^{13}	10^{16}	10^{13}
用电量/(W/GB)	$<10^{-10}$	0.04	0.01～0.04	0.1～0.4
访问时间	>1 h	7000 μs	0.005 μs	0.06 μs

从表 1 还可得到，DNA 存储时长至少为硬盘、闪存的 10 倍。研究人员能对 70 万年前的基因组[16]、11 万年前的北极熊基因组[17]、1.8 亿年前的植物化石基因组[18]进行测序。DNA 作为数据存储设备比 DVD、磁带等存储设备具有更长的使用保质期。同时，它还可以通过聚合酶链反应(PCR，一种可对特定 DNA 片段进行放大扩增的生物技术)较容易地实现扩增以获取所需数量的拷贝副本。DNA 作为最稳定的储存设备之一，对于外部环境，如高温、震荡等具有极强的抗干扰能力，研究表明 DNA 在 −5 ℃时每 6 830 000 年只降解 1 bp[19]。由于肉眼不可见，DNA 可隐藏于一般遗传物质中，安全性远高于普通存储设备。

2.2 DNA 存储框架

DNA 存储即利用 DNA 的 A，T，C，G 4 个碱基对信息编码，结合生化技术，按碱基序列顺序通过人工合成技术合成 DNA，写入信息实现存储；读取信息时，利用 PCR 技术对存储链进行复制扩增以备份，再对扩增得到的 DNA 片段进行测序、解码，恢复原始信息。DNA 作为存储设备对信息进行保存及读取的整体流程如图 2 所示。利用 DNA 存储数据的主要框架包括 3 部分：编码写入部分、存放部分及解码读取部分。

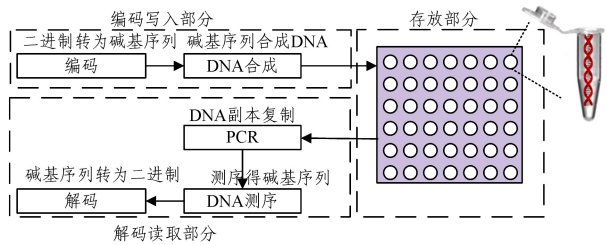

图 2 DNA 存储流程图

Fig. 2 Flow chart of DNA storage

2.2.1 DNA 编码写入

DNA 编码写入部分主要由 DNA 编码及 DNA 合成组成。

DNA 编码指通过一定的对应关系或规则将需要存储的文件信息转化为 DNA 碱基序列(即含有 A，G，C，T 的序列)，进而实现后期的合成及存储。不同 DNA 模型适用于不同的信息类型，有的模型仅适用于文本信息，有的仅适用于图片信息，也有的对任何信息均可实现转化。虽然模型方法间存在一定差异，但是 DNA 编码的主要过程基本一致，都经历压缩—引入纠错[20]—转为碱基序列的编码过程，整个编码示例如图 3 所示。

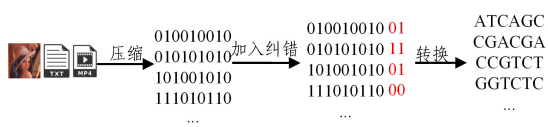

图 3 DNA 编码流程

Fig. 3 Flowchart of DNA coding

DNA 合成是将碱基序列中的碱基逐个连接形成 DNA 链的过程。由于细胞的排外性及受生物活动的影响[21]，在利用 DNA 存储信息时研究人员一般采用体外人工合成的方式合成 DNA 链。整个生物模型如图 4 所示。

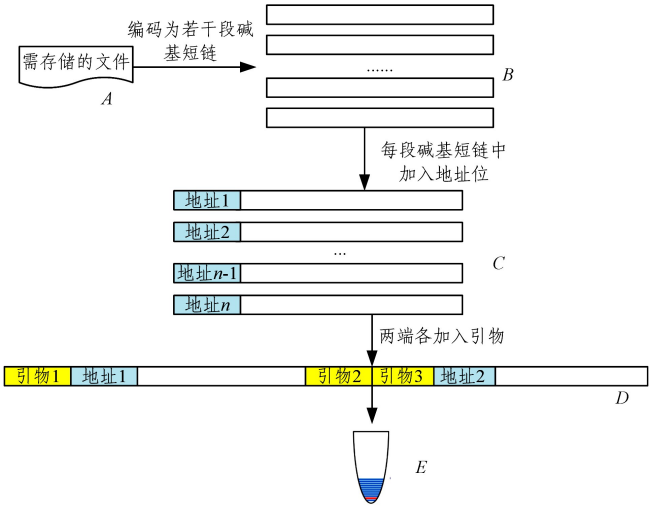

图 4　DNA 存储生物模型

Fig. 4　Biological model of DNA storage

因为合成 DNA 长链在时间、错误率、技术难点等方面均高于短链，故研究人员通常将碱基序列分为若干段短链，如图 4 中的 A 至 B 过程；为了标明某段短链在信息中所处的位置，引入地址位，通过它可以快速定位、查找、拼接各段信息，如图 4 中 B 至 C 过程；最后将每段短链拼接成 DNA 长链，拼接前需在序列两端加入引物（一种具有特定核苷酸序列的大分子，能在核苷酸聚合作用开始时刺激 DNA 合成），即获得存储信息的 DNA 链[14]。

合成过程中，DNA 序列的均聚物（由一种单体聚合而成的聚合物，如 DNA 链 AAAAA 就为均聚物）及 GC 含量过高均会对合成造成影响，故编码时还需注意碱基序列均聚物及 GC 含量问题。

2.2.2　DNA 解码读取

DNA 解码读取的关键技术为解码技术。解码由 DNA 链获取存储的信息，是 DNA 编码的逆过程。整个 DNA 解码的读取过程如图 5 所示。解码前需进行 PCR 复制，扩增得到多个 DNA 片段副本，再对副本进行 DNA 测序（DNA 测序技术可分析特定 DNA 片段的碱基序列，即能获取 DNA 的 A，G，C，T 的排列方式）。获取碱基序列后对序列纠错、去冗余、解码，以读取原始数据。

图 5　DNA 解码流程

Fig. 5　Flow chart of DNA decoding

3　DNA 存储编码技术

DNA 编码是 DNA 存储的关键技术，编码结果直接影响存储性能的优劣。整个 DNA 存储编码过程由压缩、纠错和转换 3 部分组成。

3.1　压缩

为了最大化地利用 DNA 存储空间，将信息存入 DNA 前须对信息去除冗余以达到压缩的目的。在 DNA 编码中，常见的压缩方法有哈夫曼编码、喷泉码和 LZMA 等。

3.1.1　哈夫曼编码

哈夫曼编码是一种广泛应用于数据文件压缩的编码方法，其压缩率可达 20%～90%。它的编码核心是对输入文件基于一个变长编码表进行编码，该编码表通过评估各符号出现的概率配以不同长度的码字（出现频率越高，码字越短，频率越低，码字越长），以此降低编码后字符的期望值及平均码长，用最少的码字编码整个输入文件，从而实现无损压缩[22-23]。DNA 编码中，用哈夫曼编码对目标文件进行压缩扫描，可获取压缩的二进制码流，提高 DNA 存储密度[11-12,24]。具体压缩编码过程如下。

Step1　对目标文件扫描计数，统计所有字符出现的频率，确定各字符权重值；

Step2　按出现频率的大小对字符排序；

Step3　分别将出现频率最低的两个字符编码为 0 和 1，再将这两个字符的频率相加作为新字符的出现频率，将新字符与剩余字符重新排序，完成一次压缩；

Step4　对新的排序结果重复 Step3 直至全部字符均完成编码；

Step5　由最后一级逐级向前记录各字符编码的码字，完成压缩。

图 6 给出含 7 种信源符号的目标文件哈夫曼编码压缩示例过程。

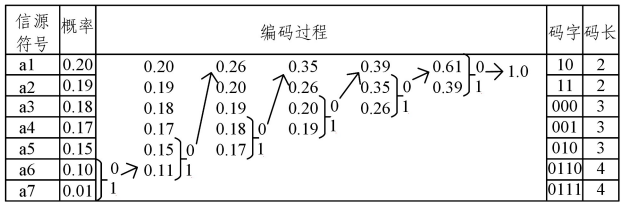

图 6　哈夫曼编码示例

Fig. 6　Example of Huffman coding

哈夫曼编码适用于输入文件的各字符非等概率出现的情况，但用该方法在每次编码前均需对字符进行概率统计，压缩耗时较长。

3.1.2　喷泉码

喷泉码是一种抹除码，它可以从给定的 m 组源符号中生成无限多个编码符号序列，当接收端收到任意 k 个编码序列即可高概率地恢复这 m 组源信息。整个喷泉码编码过程如图 7 所示，具体步骤如下[25-27]。

Step1　将源文件等分为 m 组，在 1 至 m 内按某一度分布 Ω 随机选取一个整数 d，称 m 为码长，称 d 为编码的度；

Step2　在 m 个组中均匀地随机选取 d 个不同的包，让

这 d 个包进行异或运算，得到一个编码码字，发送给接收端；

Step3 重复 Step2 直至接收端发送终止信号停止编码。

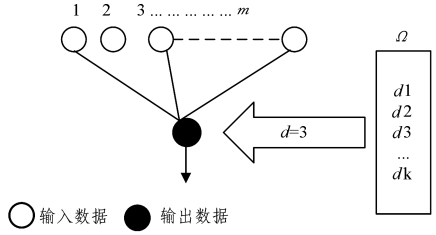

图 7 喷泉码编码过程

Fig. 7 Coding process of fountain code

喷泉码具有独立随机性，且编译码复杂程度低，能以较小的译码开销来高概率地恢复信息，将其应用于并行存储的 DNA 中可极大地提高存储效率。

文献[10]在 DNA 编码时就利用喷泉码压缩信息，图 8 为其 DNA 喷泉码的应用流程图，具体可分为以下几步：1)将目标文件等分成 4 组并确定度分布函数 Ω；2)利用 Luby 变换生成随机种子，通过随机种子选出 d 个段，对段内信息进行异或运算，获取液滴信息，不断重复该过程直至获得足够的液滴信息。压缩存储后最终将存储密度提升至 1.57 bits/nt。

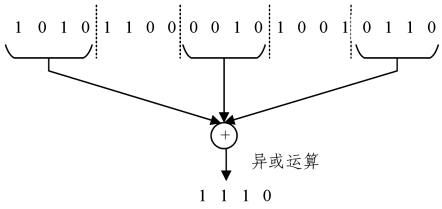

图 8 DNA 喷泉码

Fig. 8 DNA fountain code

3.1.3 LZMA

LZMA(Lempel-Ziv-Markov chain-Algorithm)是 lgor Pavlov 于 2001 年提出的一种基于 Deflate 和 LZ77 法的改良优化算法。它运用二叉树、散列表、基数树等方法对源文件序列进行字典查找，并对字典中的重复序列进行单独压缩。由于引入了 Range Coder 熵编码，LZMA 在具有高压缩率的同时兼顾了压缩速度，能较快地实现压缩[28]。LZMA 算法的性能如表 2 所列[29]。

表 2 LZMA 算法的压缩性能

Table 2 Compression performance of LZMA algorithm

LZMA 算法的压缩性能	参考值	实际值
压缩比	—	170%
可调节的字典大小/MB	≤256	默认为 8
压缩速度/(MB/s)	1	0.4
解压速度/(MB/s)	0.5～1	0.2

LZMA 算法在 DNA 序列压缩领域应用广泛[30-32]。Yim 研究团队于 2012 年对一张 BMP 图片的二进制码流利用 LZMA 算法压缩后存储于 DNA 中[33]，获得了较理想的存储效果。

LZMA 算法充分利用数据的结构特点，简单、可行地实现了压缩[34]。但该方法并不适用于高通量数据，且压缩过程的耗时较长。

3.2 纠错

在 DNA 存储信息的过程中，无论是 DNA 编码、DNA 合成还是 DNA 解码，均有可能出现错误，导致最终读取的信息与原始信息间出现偏差。为了尽量避免这种情况给存储带来的干扰，在 DNA 存储过程中可引入相应的纠错机制来提高存储的准确性[20]。常用的纠错方法有汉明码纠错、RS 码纠错、LDPC 码纠错。

3.2.1 汉明码纠错

汉明码(Hamming Code)是一种线性分组编码。当传输信道的性能良好时，汉明码是很好的低冗余开销纠错选择，它通过在传输的信息流中加入校验码进行纠错，r 位校验码可检测纠正 k 位信息流中的一位错误，信息码流与校验码位数之和称为传输码元总位数，记为 n，三者间满足：

$$2^r-1\geqslant n=r+k \tag{1}$$

汉明(n,k)码即为每 n 位传输码流中含有 k 位信息码流和引入的 $n-k$ 位校验码。不同位置的校验码利用重叠的奇偶校验纠错不同的信息码元。编码中，校验码的位置与其负责校验的码元间是有规律的。表 3 给出了(7,4)汉明码校验码和信息码元位置及对应的校验关系，其中校验码元记为 $p1$，$p2$ 和 $p3$，信息码元记为 $d1$，$d2$，$d3$ 和 $d4$，3 个校验码元分别为不同码元提供偶校验。由表 3 可知，$p1$ 为 $p1$，$d1$，$d2$，$d4$(即 1，3，5，7 位)提供校验。

表 3 (7,4)汉明码校验表

Table 3 Checklist of (7,4) Hamming code

位数	1	2	3	4	5	6	7
码元	$p1$	$p2$	$d1$	$p3$	$d2$	$d3$	$d4$
$p1$	Yes	No	Yes	No	Yes	No	Yes
$p2$	No	Yes	Yes	No	No	Yes	Yes
$p3$	No	No	No	Yes	Yes	Yes	Yes

汉明码纠错步骤如下：

1)确定校验码位数 r。每个校验位需符合相应的奇偶校验规定。

2)将长度为 k 的源文件与校验码一起编码为 $k+r$ 的新码流。

3)对接收端接收到的码流进行 r 个奇偶校验，若所有的校验结果均正确则传输信息无误；若发现错误，则由校验结果可唯一确定错误位。

汉明码编译码方式简单，被广泛应用于硬盘及蓝牙技术中，是一种较好的低冗余纠错编码[35-36]。宋香明[37]将汉明码用于 DNA 数据存储纠错环节，读取阶段可无错误恢复信息。

3.2.2 RS 码纠错

RS 码是一种典型的线性循环码，即源文件编码后的码流 $T(x)$向左或向右移动后仍为有限组码组中的一组，它可对随机错误、突发错误及二者的组合进行纠错。RS 码的运算都定义在伽罗华域 GF(2^m)中，每个码元均可视为域中的一个元素。设 $q=2^m$，则有：

$$GF(2^m)=[0,\alpha^0,\alpha^1,\cdots,\alpha^{q-2}] \tag{2}$$

当 $m=4$ 时，多项式 $p(x)=x^4+x+1$，假设 $p(x)$的解为 α，则有 $\alpha^4+\alpha+1=0$，在伽罗华域中加法为异或运算，进一步有 $\alpha^4=\alpha+1$，由此可得 GF(16)中的各元素，如表 4 所列。

表 4　GF(16)元素表

Table 4　Element table of GF(16)

GF(16)元素	二进制	十进制
0	0000	0
α^0	0001	1
α^1	0010	2
α^2	0100	4
α^3	1000	8
α^4	0011	3
α^5	0110	6
α^6	1100	12
α^7	1011	11
α^8	0101	5
α^9	1010	10
α^{10}	0111	7
α^{11}	1110	14
α^{12}	1111	15
α^{13}	1101	13
α^{14}	1001	9

RS 码生成多项式的公式为：

$$g(x)=(x+\alpha)(x+\alpha^2)\cdots(x+\alpha^{2i}) \tag{3}$$

因此 RS 码纠错的基本思路为：选择一个合适的生成多项式 $g(x)$，使该多项式为每个信息段计算获得的码字多项式 $c(x)$的倍式，即$\frac{c(x)}{g(x)}=0$。当接收端接收到的码字多项式 $c'(x)$与 $g(x)$的余式不为 0 时，则传输过程中出现错误，利用有限域的封闭性可间接获取出错位置和出错值；若余式为 0，则传输未出错。

RS 码能用较小的冗余恢复更多的数据信息，因此其在 DNA 信息存储领域也有着广泛应用[10,38]。但由于涉及有限域和伽罗华域，其计算量较大[39]。

3.2.3　LDPC 码纠错

低密度奇偶校验码(Low Density Parity Check Code，LDPC)是一种具有稀疏校验阵的分组纠错码，它主要通过奇偶校验码的方式进行纠错。奇偶校验码是通过增加冗余信息使码字中 1 的个数恒为奇数或偶数的编码方法。

LDPC 码主要由校验矩阵 $\boldsymbol{H}$ 唯一定义。为便于描述，通常引入一个 Tanner 图来定义 LDPC 码。图 9 为 $\boldsymbol{H}$ 矩阵对应的一个 Tanner 图，它主要由校验节点(check node)和变量节点(variable node)组成，每个校验节点的边数称为度数。校验节点与校验矩阵 $\boldsymbol{H}$ 的每一行相对应，变量节点与 $\boldsymbol{H}$ 的每一列相对应。当矩阵 $\boldsymbol{H}$ 中第 i 行第 j 列为 1 时，代表第 i 个校验码元与第 j 个变量码元关联，在 Tanner 中表现为第 i 个校验节点与第 j 个变量节点相连；当在 Tanner 图中，与某校验节点相连的所有变量节点包含奇数个 1 时，该校验节点为 1，否则为 0。

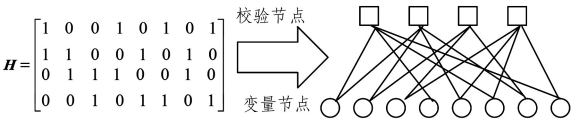

图 9　LDPC 码的 Tanner 图

Fig. 9　Tanner figure of LDPC code

由于 LDPC 编译码简单、可行、易操作，性能逼近香农限，纠错计算量小，解码时间快，且具有缺损补偿功能等优势，因此被广泛应用于所有信道中[33,40-41]。

在 DNA 信息存储中也有研究团队[33]将 LDPC 用于纠错环节，以防止在 DNA 合成及测序中出现随机错误，提高文件读取的准确性。

3.3　转换模型

DNA 链是由 A，T，C 和 G 4 个碱基组成的，由于计算机中的数据都是以二进制(即 0，1)形式存在，因此将信息存储至 DNA 中实质上就是将信息的二进制码流编码为碱基序列存入 DNA。DNA 编码模型的作用是将信息的码流编码为 DNA 碱基序列。根据 DNA 的组成及结构，常见的 DNA 存储编码模型有 3 种：二进制模型、三进制模型和四进制模型。

3.3.1　二进制模型

DNA 存储中二进制模型通常是将 A，T，C，G 4 个碱基中的任意两个定义为 0，另外两个定义为 1，即整个碱基序列只有 0 和 1 两种状态。2012 年，Church 等[8]提出的 DNA 存储方案使用的就是二进制模型，编码过程中他们按 A 或 G 等于 0，C 或 T 等于 1 将信息码流转为碱基序列，方便后续合成 DNA，其总体方案如图 10 所示。例如信息段 1000110 按该模型即编码为 TCAAGTC 碱基列。

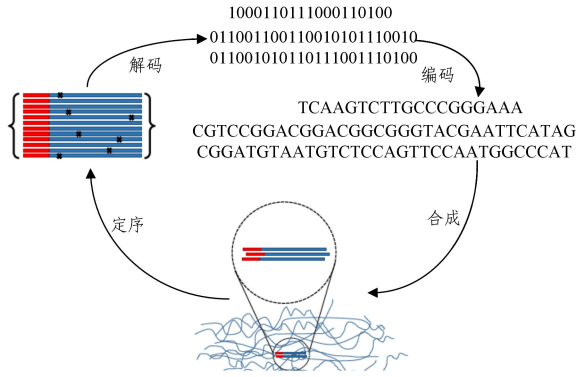

图 10　Church 等的存储框架示意图

Fig. 10　Storage framework diagram of Church et al.

这种二进制模型相对简单，并且能够较好地避免 DNA 中出现 GC 含量不均衡、均聚物较多等情况，可降低后期合成 DNA 的难度。但就编码效率而言，相同长度的碱基序列二进制模型能存储的信息量较少，编码效率不高。

3.3.2　三进制模型

三进制编码模型指整个碱基序列只有 3 种状态：0，1，2。Goldman[11]研究团队采用三进制编码模型将信息存储至 DNA。他们首先将文件内容转为三进制码流，接着对码流中的 0，1，2 按表 5 中的对应关系编码为碱基序列进行存储。

表 5　三进制模型编码表

Table 5　Coding table of ternary model

前一位碱基	紧接的待编码符号		
	0	1	2
A	C	G	T
C	G	T	A
G	T	A	C
T	A	C	G

由表 5 可知，三进制编码模型主要是通过前一个碱基来确定后一个碱基[42]，并没有像二进制模型一样在碱基与数据间建立某种映射关系。相比二进制模型，用三进制模型能存储更多的信息，但是三进制模型也未充分利用 DNA 的存储能力。

3.3.3 四进制模型

DNA 由 4 个核苷酸组成，故可视为一个天然的四进制编码模型，相当于把碱基中的 A,T,C,G 看作 0,1,2,3。对于读入 DNA 的二进制码流，只要将其转为四进制就可以编码为碱基序列。表 6 是常见的四进制模型映射关系表，但该映射关系并不唯一，将 A,T,C,G 分别对应至二进制数据的 00,01,10,11，共有 $A_4^4=24$ 种组合方案，理论上这 24 种方案彼此是等价的。

表 6 四进制模型编码表

Table 6 Coding table of quaternary model

二进制数据	00	01	10	11
对应碱基	A	T	C	G

四进制模型相对于其他两种模型而言存储能力最强，理论上每位碱基可以编码 2bit 数据，一定程度上提高了存储效率，降低了 DNA 存储成本，故在 DNA 数据存储领域应用广泛[10,12,26]。但这种模型易出现 GC 含量过高、均聚物较多等不利于 DNA 合成的情况，对后续的 DNA 存储操作有一定影响。

4 DNA 存储的发展及挑战

4.1 DNA 存储的发展

早在 1988 年，Graig Venter 研究所[42]的研究人员将他们的名字共 35 位信息成功编码至 DNA 中，初步实现了 DNA 存储。但受 DNA 合成、测序等生物技术以及研究 DNA 存储领域的科研工作者较少等因素的影响，该领域当时并未有较大进展[24,43-45]。随着科技的发展，各项技术逐渐完善[46]，随之产生的信息数据量也与日俱增，致使越来越多的研究人员、机构、企业等纷纷加入到 DNA 存储的研究领域[47-51]，DNA 数据存储技术得到较快发展[52-54]。现阶段 DNA 存储技术框架已基本成型，均为编码＋纠错＋地址的模式。四进制转换模型已成为 DNA 存储的主流转换模型。近几年提出的 DNA 存储方案均带有纠错机制，在读取信息时出现错误的概率越来越小，大多可做到无错误恢复；并且引入的冗余也越来越少，编码效率得到了极大提高，同时降低了存储成本。此外，也有科学工作者对活体 DNA 存储进行了研究[9,55]。

目前提出的 DNA 存储方案越来越多，现阶段较典型的存储方案有：2012 年 Church 等[8]用二进制进行转换存储了 659 kB 信息，该方案虽然降低了 DNA 合成的难度，但引入的冗余过多；2013 年 Goldman 等[11]利用哈夫曼编码、四倍重叠法、三进制编码实现信息的压缩、纠错、转换，将 739 kB 的内容存入 DNA 中；2015 年 Grass 等[38]将 RS 纠错应用于 DNA 存储，无错误地读取了存储于 DNA 中的 83 kB 信息；2016 年 Bornholt 等[12]实现了具有随机访问和内容重写功能的 DNA 存储；同年 Blawat 等[26]运用前向纠错技术确保了 DNA 读取的准确性；2017 年 Erlich 等[10]基于喷泉码压缩数据，只引入了 20.71%的冗余，大大降低了 DNA 存储的成本。表 7 给出了这几种 DNA 存储编码方案的参数对比，其中，储数据量为实验中成功编码并存储于 DNA 中，且进行生物实验的信息量；存储密度为引入纠错位、地址位等冗余后，每个核苷酸中存储的信息位数；冗余为合成 DNA 的碱基序列中未含信息的碱基数在碱基序列中的占比；成本为用 DNA 每存储一位信息的花销；无错误恢复表示解码得到的文件与原文件是否一致。

表 7 2012－2017 年 DNA 主要存储方案的性能参数

Table 7 Performance parameters of DNA storage schemes from 2012 to 2017

	Church 方法	Goldman 方法	Grass 方法	Bornholt 方法	Blawat 方法	Erlich 方法
存储数据量/MB	0.65	0.63	0.08	0.15	22	2.11
存储密度/(bit/nt)	0.83	0.33	1.14	0.88	0.92	1.57
冗余/%	17.00	79.11	35.96	44.30	42.50	20.71
纠错机制	无	四倍重叠移步＋多数票决	RS 编码纠错	无	前向纠错	RS 编码纠错
成本/(元/bit)	1.81	4.55	1.32	1.70	1.63	0.96
无错误恢复	否	否	是	否	是	是

本研究团队在 DNA 存储领域进行了大量工作，针对现有 DNA 信息存储方案存储密度低，不能很好地发挥 DNA 分子的信息存储潜力的问题，提出了一种具有高存储密度的 DNA 信息存储编码方案[56]，并将四进制哈夫曼 DNA 编码与汉明纠错码相结合，建立了一套完整的 DNA 存储模型，该模型仅引入了 20%的数据冗余量，存储密度高达 5.66 bit/nt，且能无错误恢复存储的信息，适用于图片、音频、文本等多种格式的文件[37]。

现阶段，DNA 存储的主要研究集中于以下 4 个方面。

1)编解码方式[57]。将 DNA 存储与现阶段主流的信源信道编解码方式结合，寻找更适合 DNA 的编码方法，尽可能充分地利用 DNA 存储空间，引入较少的冗余。

2)纠错机制。现阶段 DNA 存储中纠错方法大部分采用 RS 编码纠错，可寻找其他更适合、冗余更小的 DNA 存储纠错方法。

3)生物技术。DNA 存储由于生物成本过高并未大规模投入使用，但随着生物技术的迅猛发展，DNA 合成和测序的成本每年均以指数形式快速下降[58]；同时纳米孔技术的发展也让 DNA 测序成本从 2002 年的218750元/兆降至 2016 年的 4.41 元/兆[59]。此外，手持式单分子 DNA 测序仪的发明使 DNA 测序更简便[60]。研究低能耗、便携式 DNA 合成及测序技术能在一定程度上降低 DNA 存储的成本，推进 DNA 存储的发展。

4)随机存取。现阶段 DNA 存储技术主要适用于不需频繁读取但需较长存储期限、较高安全性能的信息存储，如一些医疗信息、法律文档[14]等。如果想扩大 DNA 的存储应用范围，除了降低成本还需实现性能更优、更便捷的随机存取功能，使 DNA 能够像硬盘、磁带那样随时随地写入或读取信息。

4.2 DNA 数据存储发展面临的挑战

虽然近几年 DNA 存储得到了较大的发展，但现阶段 DNA 存储并未实现大规模使用，主要原因有：

1)成本高。目前 DNA 合成一个碱基的费用约为 1.04 元(根据具体合成要求可能会有一定偏差[61]，用 DNA 存储 1TB 内容的成本约为硬盘的七千万倍。

2)耗时长。将信息存入 DNA 中需要历经编码和 DNA

合成，信息读取需历经 DNA 测序和解码，而 DNA 合成及测序都需花费时间，并且碱基序列越长，合成和测序耗费的时间越多，因此用 DNA 进行存储并不能像硬盘、磁带等存储设备一样做到及时存取。

3）技术难点多。在 DNA 存储过程中，如何利用现有的或更好的编解码技术将更多的信息编码至 DNA 中是一个技术难点；另一方面，在 DNA 存储实现过程中提高生物实验的操作精度从而减少误差也是一个技术难点。

结束语　DNA 数据存储技术是近十几年发展起来的研究热点，虽然目前还存在着众多技术难点待突破研究，但不可否认 DNA 的存储能力、稳定性和抗干扰性等特点确实比现阶段的存储设备更高。本文对 DNA 存储框架的全过程进行了描述，详细阐述了 DNA 存储中的编码技术，综述了该领域近年来的研究成果，总结了现阶段主流的 4 个研究方向，并对其中存在的挑战及问题进行了概括。2020 年，预计全世界产生的数据将突破 40ZB，DNA 作为新一代存储介质有着无限的潜力。

参 考 文 献

[1] ZHIRNOV V, ZADEGAN R M, SANDHU G S, et al. Nucleic Acid Memory[J]. Nature Materials, 2016, 15(4): 366-370.

[2] GODA K, KITSUREGAWA M. The History of Storage Systems[J]. Proceedings of the IEEE, 2012, 100(13): 1433-1440.

[3] PANDA D, MOLLA K A, BAIG M J, et al. DNA as a digital information storage device: hope or hype? [J]. Biotech, 2018, 8(5): 239-247.

[4] 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, 2004, 38(6): 1915-1916.

[5] EXTANCE A. How DNA could store all the world's data[J]. Nature, 2016, 537(7618): 22-24.

[6] BORNHOLT J, LOPEZ R, CARMEAN D M, et al. A DNA-Based Archival Storage System[C]// International Conference on Architectural Support for Programming Languages & Operating Systems. 2016.

[7] HAKAMI H A, CHACZKO Z, KALE A. Review of Big Data Storage Based on DNA Computing[C]// Computer Aided System Engineering. IEEE, 2015.

[8] CHURCH G M, GAO Y, KOSURI S. Next-Generation Digital Information Storage in DNA[J]. Science, 2012, 337(6102): 1628.

[9] SHIPMAN S L, NIVALA J, MACKLIS J D, 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.

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

[11] GOLDMAN N, BERTONE P, CHEN S, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA[J]. Nature, 2013, 494(7435): 77-80.

[12] BORNHOLT J, LOPEZ R, CARMEAN D M, et al. Toward a DNA-Based Archival Storage System[J]. IEEE Micro, 2016, PP(99): 637-649.

[13] CASTILLO M. From Hard Drives to Flash Drives to DNA Drives[J]. American Journal of Neuroradiology, 2014, 35(1): 1-2.

[14] BONNET J, COLOTTE M, COUDY D, et al. Chain and conformation stability of solid-state DNA: Implications of room temperature storage[J]. Nucleic Acids Research 2009; 38(5): 1531-1546.

[15] WANG S W. DNA storage with error correction mechanism [D]. Changsha: National University of Defense Technology, 2014. (in Chinese)
王诗薇. 带纠错机制的 DNA 存储[D]. 长沙：国防科学技术大学，2014.

[16] ORLANDO L, GLNOLHAC A, ZHANG G, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse[J]. Nature, 2013, 499(7456): 74-78.

[17] MILLER W, SCHUSTER S C, WELCH A J, et al. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change[J]. Proceedings of The National Academy of Sciences of The United States of America, 2012, 109(36): E2382-E2390.

[18] KIM S, SOLTIS D E, SOLTIS P S, et al. DNA sequences from Miocene fossils: an ndhF sequence of Magnolia latahensis (Magnoliaceae) and an rbcL sequence of Persea psedocarolinensis (Lauraceae)[J]. American Journal of Botany, 2004, 91(4): 615-620.

[19] 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 Biological Sciences, 2012, 279(1748): 4724-4733.

[20] HUFFMAN D A. A method for the construction of minimum-redundancy codes[J]. Resonance, 2006, 11(2): 91-99.

[21] GIBSON D G, VENTER J C. Creation of a bacterial cell controlled by a chemically synthesized genome. [J]. Science, 2010, 329(5987): 52-56.

[22] SAVITRI P A I, ADIWIJAYA, MURDIANSYAH D T, et al. Digital medical image compression algorithm using adaptive Huffman coding and graph based quantization based on IWT-SVD[C]// International Conference on Information & Communication Technology. 2016.

[23] PADMAVATI S, MESHARAM V. DCT combined with fractal quadtree decomposition and Huffman coding for image compression[C]// International Conference on Condition Assessment Techniques in Electrical Systems. 2016.

[24] AILENBERG M, ROTSTEIN O. An improved Huffman coding method for archiving text, images, and music characters in DNA [J]. BioTechniques, 2009, 47(3): 747-754.

[25] CAIRE G, SHAMAI S, SHOKROLLAHI A, et al. Universal variable-length data compression of binary sources using fountain codes[C]// Information Theory Workshop. IEEE, 2016: 123-128.

[26] BLAWAT M, GAEDKE K, HüTTER I, et al. Forward Error Correction for DNA Data Storage [J]. Procedia Computer Science, 2016, 80(5): 1011-1022.

[27] BYERS J W, LUBY M, MITZENMACHER M. A digital fountain approach to asynchronous reliable multicast[J]. IEEE Jour-

nal on Selected Areas in Communications, 2002, 20(8): 1528-1540.

[28] BING L I, ZHANG L, LIU Y. FPGA hardware implementation of the LZMA compression algorithm[J]. Journal of Beijing University of Aeronautics & Astronautics, 2015, 41(3): 375-382.

[29] LAN C, XU J, ZENG W, et al. Compound image compression using lossless and lossy LZMA in HEVC[C]// IEEE International Conference on Multimedia & Expo. IEEE, 2015.

[30] JI Z, ZHOU J R, ZHU Z X. Bioinformatics Features Based DNA Sequence DATA Compression Algorithm[J]. Acta Electronica Sinica, 2011, 39(5): 991-995. (in Chinese)
纪震, 周家锐, 朱泽轩. 基于生物信息学特征的 DNA 序列数据压缩算法[J]. 电子学报, 2011, 39(5): 991-995.

[31] PINHO A J, PRATAS D, FERREIRA P J S G. Bacteria DNA sequence compression using a mixture of finite-context models [C]// Statistical Signal Processing Workshop. 2011.

[32] CAO M D, DIX T I, ALLISON L, et al. A Simple Statistical Algorithm for Biological Sequence Compression[C]// Data Compression Conference. 2007.

[33] YIM K Y, YU C S, LI J W, et al. The Essential Component in DNA-Based Information Storage System: Robust Error-Tolerating Module[J]. Frontiers in Bioengineering & Biotechnology, 2014, 2(2): 49-53.

[34] SALOMON D. Data compression: the complete reference[M]// Data Compression: The Complete Reference. New York: Springer-Verlag, 2000.

[35] SHEN Y F, PAN L. Principle and Method of the Error Detection and Correction of Ternary Hamming Codes [J]. Chinese Journal of Computers, 2015, 38(8): 1648-1655. (in Chinese)
沈云付, 潘磊. 三值汉明码检错纠错原理和方法[J]. 计算机学报, 2015, 38(8): 1648-1655.

[36] SINGH A K. Error detection and correction by hamming code [C]// International Conference on Global Trends in Signal Processing. 2017.

[37] SONG X M. Research on DNA Information Storage Method Based on Huffman Coding [D]. Tianjin: Tianjin University, 2018. (in Chinese)
宋香明. 基于 Huffman 编码的 DNA 信息存储方法研究[D]. 天津: 天津大学, 2018.

[38] GRASS R N, HECKEL R, PUDDU M, et al. Robust chemical preservation of digital information on DNA in silica with error-correcting codes[J]. Angewandte Chemie International Ed in English, 2015, 54(8): 2552-2555.

[39] WANG L, WEI Z, YANG W, et al. Multiple channel error-correction algorithms for LCC decoding of Reed-Solomon codes and its high-speed architecture design[J]. IET Communications, 2017, 11(9): 1407-1415.

[40] ALWAN M H, SINGH M, MAHDI H F. Performance comparison of turbo codes with LDPC codes and with BCH codes for forward error correcting codes[C]// Research & Development. 2016.

[41] BALDI M, MATURO N, RICCIUTELLI G, et al. On the error detection capability of combined LDPC and CRC codes for space telecommand transmissions [C] // Computers & Communication. 2016.

[42] CHRISTY, BOGARD, ERIC, et al. DNA media storage[J]. Progress in Natural Science, 2008, 18(5): 603-609.

[43] WONG P C, WONG K K, FOOTE H. Organic data memory using the DNA approach [J]. Communications of the ACM, 2003, 46(1): 95-98.

[44] KASHIWAMURA S, YAMAMOTO M, KAMEDA A, et al. Potential for enlarging DNA memory: the validity of experimental operations of scaled-up nested primer molecular memory[J]. Biosystems, 2005, 80(1): 99-112.

[45] NOZOMU Y, KAZUHIDE S, JUNICHI S, et al. Alignment-Based Approach for Durable Data Storage into Living Organisms [J]. Biotechnology Progress, 2007, 23(2): 501-505.

[46] SCHUSTER S C. Next-generation sequencing transforms today's biology[J]. Nature Methods, 2008, 5(1): 16-18.

[47] SONG Y, KIM S, HELLER M J, et al. DNA multi-bit non-volatile memory and bit-shifting operations using addressable electrode arrays and electric field-induced hybridization[J]. Nature Communications, 2018, 9(1): 1-8.

[48] HEAVEN D. Now we can store video in living DNA[J]. New Scientist, 2017, 235(3134): 11-14.

[49] YAZDI S M H T, GABRYS R, MILENKOVIC O. Portable and Error-Free DNA-Based Data Storage [J]. Scientific Reports, 2017, 7(1): 1-4.

[50] BLAWAT M, GAEDKE K, HÜTTER I, et al. Forward Error Correction for DNA Data Storage [J]. Procedia Computer Science, 2016, 80(5): 1011-1022.

[51] JIANG L, QIU W, ALDIRINI F, et al. Feasibility study of molecular memory device based on DNA using methylation to store information[J]. Journal of Applied Physics, 2016, 120(2): 96-100.

[52] YAZDI S M H T, KIAH H M, GARCIA E R, et al. DNA-Based Storage: Trends and Methods[J]. IEEE Transactions on Molecular, Biological and Multi-Scale Communications, 2015, 1(3): 230-248.

[53] TABATABAEI Y S M H, YUAN Y, MA J, et al. A Rewritable, Random-Access DNA-Based Storage System[J]. Scientific Reports, 2015, 5(9): 1-10.

[54] FATIMA A, UL H I, HAIDER A, et al. Trends to store digital data in DNA: an overview[J]. Molecular Biology Reports, 2018, 45(5): 1479-1490.

[55] JAIN S, FARNOUD F, SCHWARTZ M, et al. Duplication-correcting codes for data storage in the DNA of living organisms [J]. IEEE Transactions on Information Theory, 2016, PP(99): 1.

[56] 张淑芳, 宋香明. 一种高存储密度的 DNA 信息存储编码方案: 201811445344.8[P]. 2018.

[57] SILVA P Y D, GANEGODA G U. New Trends of Digital Data Storage in DNA[J]. Biomed Research International, 2016, 2016 (5536): 1-14.

[58] CARR P A, CHURCH G M. Genome engineering[J]. Nature Biotechnology, 2009, 27(12): 1151-1162.

[59] SHENDURE J, LIEBERMAN A E. The expanding scope of DNA sequencing[J]. Nature Biotechnology, 2012, 30(11): 1084

[60] PENNISI E. Genome sequencing. Search for pore-fection[J]. Science, 2012, 336(6081): 534-537.

[61] KOSURI S, CHURCH G M. Large-scale de novo DNA synthesis: technologies and applications[J]. Nature Methods, 2014, 11(5): 499-507.