DNA或将成为下一代信息存储技术

刘曙光　黄山学院

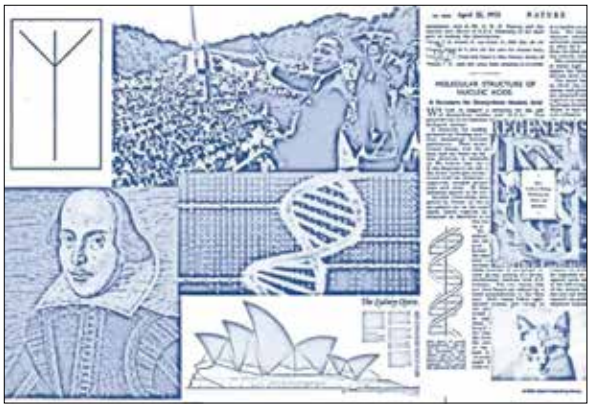

图1 存入DNA的照片、影像、文本、论文

互联网和人工智能等信息技术的快速发展使得人类社会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30ZB增长至2025年的163ZB，该趋势将很快导致数据量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、数据易受环境因素影响、生产设备耗能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（DNA）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。此外，DNA编码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 存储技术是生物技术与信息处理技术共同发展的结果，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。哈佛大学的 Church 研究团队于 2012 年将 650kB 数据存于 DNA，2017 年又将视频文件存入大肠杆菌 DNA；欧洲生物信息实验室于 2013 年利用 DNA 分子实现了 20MB 的数据存储；2016 年，微软研究院和华盛顿大学将包括 154 首莎士比亚十四行诗、一张照片、一篇 PDF 版本的科学论文、马丁·路德·金演讲“我有一个梦想”的片段和压缩算法的 text 文本存入 DNA（见如图 1）；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 215PB 信息存储到 1g DNA 分子内；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了各行各业的关注。

1 DNA的结构及特点

早在1869年，Miescher在细胞核中就发现了一些化合物的混合体，并命名为“核素”（Nuclein）。不久，Miescher和Aitmann正式提出“核酸”（Nucleic acid）这一术语，它包括脱氧核酸（deoxyribonucleic acid，DNA）和核糖核酸（ribnucleic acid，RNA）两大类。1944年，Avery等人首先证实了“DNA是遗传物质”。1953年，Watson和Crick发现了DNA双螺旋的结构，如图2所示。1962年两人因发现DNA双螺旋的结构获得了诺贝尔生理学或医学奖，开启了分子生物学时代，使遗传的研究深入到分子层次，“生命之谜”被打开，人们可以清楚地了解遗传信息的构成和传递的途径。

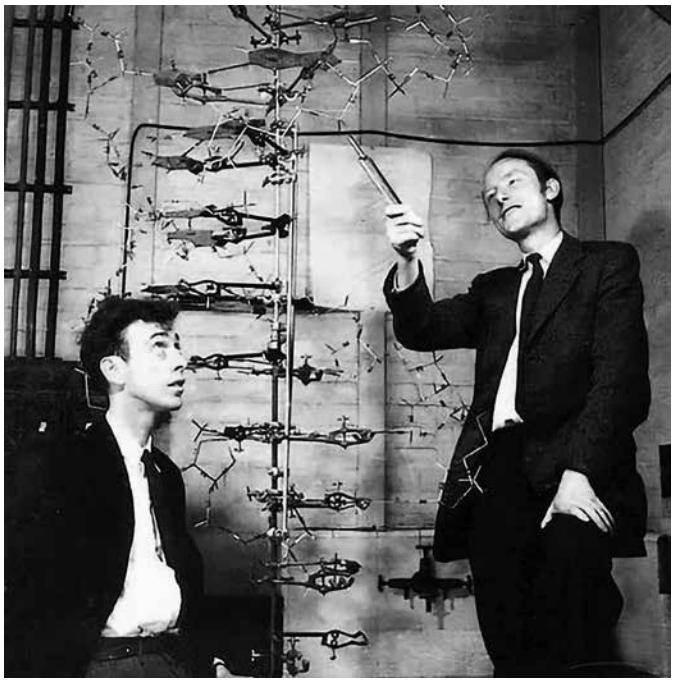

图2 Watson和Crick发现了DNA双螺旋结构

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，图3所示为一条DNA双链。

双螺旋结构的DNA拥有更多可利用空间：单位质量（1g）的DNA约有1021个碱基，可存储455EB信息，此信息量约为全球一年所产生信息总

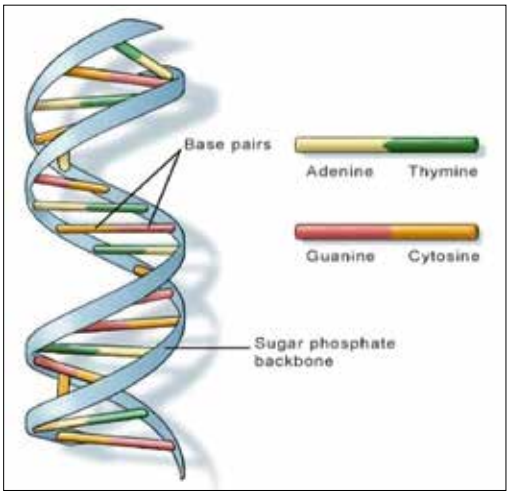

图3 DNA模型

量的1/4；单位体积（1cm^3）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106倍，是闪存的103倍。

表 1 传统存储设备与 DNA 存储性能参数对比

存储设备	DNA	硬盘	闪存	内存
保存时长（years）	>100	10	10	<64ms
数据密度（bits per cm^3）	10^{19}	10^{13}	10^{16}	10^{13}
功耗（watts per gigabyte）	$<10^{-10}$	0.04	0.01～0.04	0.1～0.4
读写速度（μs per bit）	>1h	7000μs	0.005μs	0.06μs

从表 1 还可看到，DNA 存储时长至少为硬盘、闪存的 10 倍。研究人员能对 11 万年前的北极熊基因组、1.8 亿年前的植物化石基因组进行测序。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 6830000 年只降解 1bp。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 DNA数据存储的原理

2.1 DNA存储的技术框架

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技术框架如图4所示。

2.2 DNA编码写入

DNA编码写入功能主要由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→转为碱基序列的编码过程，整个编码过程的示例如图5所示。

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。

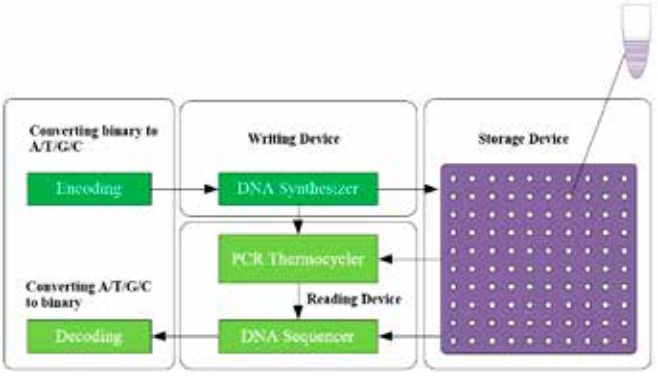

图4 DNA存储技术框架

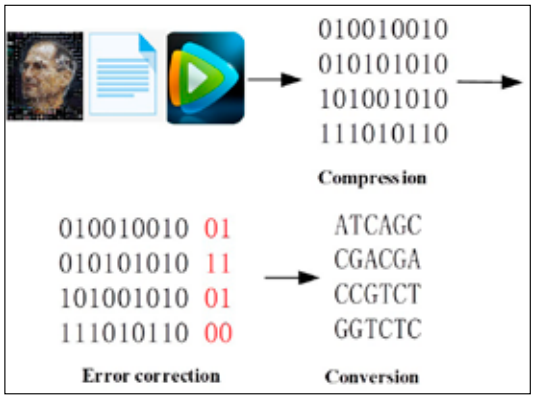

图5 DNA编码写入流程图

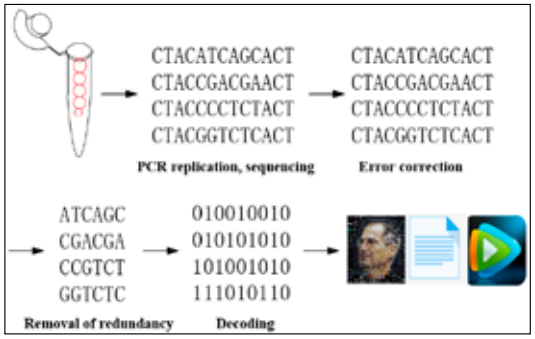

图6 DNA解码读取流程图

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，同时受生物活动的影响，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。

2.3 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 6 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。

2.4 纠错

在DNA存储信息的过程中,无论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。DNA常用的纠错方法有汉明码纠错、RS码纠错、LDPC码纠错。

3 DNA存储存在的问题

现阶段DNA存储还存在很多需要解决的技术问题。首先，目前人工合成DNA的成本过高且费时。磁介质0、1之间的转换只需通过加磁消磁即可实现，光介质可以通过刻录机将数据写在光盘上，这些比较容易实现。而将数据“写”入DNA则困难得多，虽然已经有自动合成仪可将碱基连接起来形成 DNA 序列，但一般只能合成短链DNA，难以做到“即时写”，且DNA存储系统是通过增加冗余度提高容错能力的，这更增加了成本和时间；其次，DNA的测序还远不够完美，目前的测序技术只能批量读取数据，即使只从存

储系统中访问一个字节的信息，系统也必须对整个DNA池进行测序和解码，导致检索文件耗时过长。虽然可以通过PCR技术精确复制需要提取字符串的副本以加快读取速度，但相对于其他的存储技术依然没有优势，造成无法“即时写”也无法“即时读”；同时DNA存储技术在编码之后不能改变或重写，在读取或恢复数据时会不可避免地存在一些错误，这相对于其他存储介质也是一个较大的缺点，因此在DNA存储中，微小的错误可能会产生很大的影响，造成存储信息不能被读取或难以理解。所以就目前来说，DNA 存储技术用途有限，要取代当前的存储技术还有很多问题需要解决。

4 结束语

生命的信息存在于DNA分子之中，构成DNA的4种碱基的不同排列方式，存储了地球上所有生命的信息，因此DNA分子是一种容量巨大的信息存储工具。随着现代社会数字化信息的不断积累，数据的存储需求越来越高，现在使用的磁介质（磁带、磁盘、硬盘等）和光介质（CD、DVD等）在存储量上将很难达到要求。为了满足人们未来对数据存储的需求，寻找具有更好存储性能的新材料、新技术成为一个重要的问题。近年来，随着DNA合成技术（数据写入）和DNA测序技术（数据读取）的突破性发展，DNA存储已成为下一代存储技术热点。2019年7月1日，著名科普杂志《科学美国人》公布了2019年十大突破性技术榜单，DNA存储技术榜上有名。

传统半导体存储方式或许在未来一段时间内仍将占据数据存储方式的主流，而包括DNA存储、蛋白质存储和小分子代谢物存储等形式的碳基生物存储方式在技术层面、尤其是成本方面还有很长的路要走。但是，纵观数据存储历史可以发现，随着人类社会数据量的不断增加，存储介质也在持续不断地迭代变化。人类历史的脚步不会停下，碳基生物存储、尤其是DNA存储是我们不得不关注的重要前沿技术方向。

参考文献

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

[2] GODA K, KITSUREGAWA M. The History of Storage Systems. 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? Biotech, 2018, 8(5): 239–247.

[4] WILLIAMS E D, AYRES R U, HELLER M. The 1.7 Kilogram Microchip of Semi–conductor Devices. Environmental Science & Technology, 2004, 38(6): 1915–1916.

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

[6] BORNHOLT J, LOPEZ E, CARMEAN D M, et al. A DNA–Based Archival Storage System. 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. International Conference on Computer Aided System Engineering. IEEE, 2015.

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

[9] SHIPMAN S L, NIVALA J, MACKLIS J D, et al. CRISPR–Casencoding of a digital movie into the genomes of a population of living bacteria. Nature, 2017, 547(7663): 345–349.

[10] ERLICH Y, ZIELINSKI D. DNA Fountain enables a robust and efficient storage architecture. 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. Nature, 2013, 494(7435): 77–80.

[12] BORNHOLT J, LOPEZ R, CARMEAN D M, et al. Toward a DNA–based Archival Storage system. IEEE Micro, 2016, pp. 637–649.

[13] CASTILLO m. from Hard Drives to Flash Drives to DNA Drives. 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. Nucleic Acids Reach 2009, 38(5):1531–1546.

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