PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2021,48(5):494~504
www.pibb.ac.cn

DNA数据存储技术原理及其研究进展

滕 越[1)*] 杨 姗[1)] 李金玉[1)] 崔玉军[1)] 刘芮存[1)] 王升启[2)*]

([1)] 病原微生物生物安全国家重点实验室，军事科学院军事医学研究院微生物流行病研究所，北京 100071；
[2)] 军事科学院军事医学研究院辐射医学研究所，北京 100850）

摘要 信息生产与数据存储能力之间的差距日益扩大，急需分子数据存储等高密度持久性信息保存替代方案，基于脱氧核糖核酸（DNA）的数据存储因在信息保留时间、物理密度和体积编码容量等方面优于多数传统存储介质，而广受关注. 本文概述了DNA数据存储技术的基本原理，总结了体外DNA存储数据库与体内分子存储器系统的研究进展，讨论了基于DNA分子的数据存储系统所涉及的各种影响因素以及面临的挑战.

关键词 DNA，数据存储，细胞存储器，核酸数据库，分子计算机
中图分类号 TP333，Q33 **DOI**：10.16476/j.pibb.2020.0224

造纸术出现之前，人类利用岩石和动物骨头，以及黏土制成的材料记录信息. 随着材料及技术的迭代，现代存储系统依靠磁性、电子或光学介质保存数据[1-3]. 由于数据存储需求的不断增加，现代数据储存系统因高昂的基础设施成本和运行功耗而不堪重负[4-5]. 现代社会迫切需要耐久、可扩展和经济的替代存储介质. 脱氧核糖核酸（DNA）作为一种有前途的存储介质，因其耐久性、大存储容量以及高体积密度而引起了人们极大的兴趣[6-8].

DNA数据存储的基本过程包括将数字信息编码成DNA序列（编码），将序列写入DNA分子（合成），通过物理调整并组织成文库进行长期存储、检索和选择性访问（随机访问），读取分子（测序）并将其转换回数字数据（解码）. 使用DNA进行数据存储的基本概念可以追溯到20世纪60年代中期，Wiener和Neiman[9-10]提出了“遗传存储器”（genetic memory）的概念，但当时DNA测序和合成技术还处于起步阶段. 直到20多年后，DNA数据存储的概念才首次被Joe Davis以生物艺术作品“Microvenus”的形式进行验证[11]. Davis[12]为“女性地球”编码了一个35比特的古代日耳曼符文图案. 这一概念于1999年再次被证明可用于将秘密信息通过密写术隐藏在滴加于纸张上的DNA微点中[13]. 该微点实验是第一个在存储或恢复过程中不涉及活细胞操作的DNA数据存储实验. 从Davis开始，其他后续工作大多都将数据存储在活细胞内[12-19]，这种存储方法既实用又富有前瞻性，合成DNA通常被克隆到复制型载体中以方便测序和选择正确合成的序列. 在2010年，Church等[20]和Goldman等[21]的一个突破性工作则重新定义了体外DNA数据存储的概念. 研究使用基于经过数十年完善的亚磷酰胺DNA合成方法作为数据的写入方式保存了数千字节数据，并采用合成测序进行数据读取[22]. 然而，多数早期DNA数据存储工作（如水印或隐写）涉及体内克隆和存储组件[12-19, 22-23]. 体内DNA数据存储是利用合成生物学方法记录生物体基因组特定区域内新产生的信息. 与传统存储介质相比，细胞尺寸较大且总存储密度较低，因此体内DNA数据存储不太可能成为一般主流数字数据存储的可行替代方案. 尽管对活细胞内的天然DNA进行修饰和添加的过程比较复杂，但体内DNA数据记录和存储可实现新的应用，例如记录关于细胞历史和环境的信息[24-28]. 该类系统可被视为分子记录纸带，记录细胞内动态时间序列的分子事件，并保存数据以供后续读取.

* 通讯联系人.
滕越. Tel：010-63869835，E-mail：yueteng@sklpb.org
王升启. Tel：010-66932251，E-mail：sqwang@bmi.ac.cn
收稿日期：2020-07-07，接受日期：2020-10-16

与传统介质相比，DNA具有下列重要优势. 首先，使用DNA进行数据存储可提供高达10^{18} byte/mm^{3}的存储密度，超出目前最高存储密度介质约6个数量级[3, 5, 7]. 高存储密度还有助于以较低的能量成本长时间保存分子中的数据. 其次，DNA易于复制，可使用PCR实现以极低的时间和资源成本复制大量数据. 再次，存储在DNA中的数据可利用DNA杂交过程对其进行相似性搜索等操作[8]. 例如，美国国会图书馆花费很大比例的资源将数据转移到新一代磁带中，而磁带驱动器只与过去一代或几代磁带兼容，因此传统存储介质的复制时间与数据数量和副本数量成正比，然而DNA存储可以通过PCR在一定时间内制作大量拷贝. 最后，可从几千年前的化石中读出DNA序列表明DNA作为数据存储系统在自然界长期保存[5]. 在避光干燥以及合理的温度条件下，DNA可以保存几百年到几千年，而商业磁带和光盘等档案存储介质的典型寿命则仅有几十年. 此外，基于DNA的数据存储同样受益于生物技术和生命科学中DNA写入和读取技术的快速发展. 本综述主要介绍了DNA数据存储技术的基础原理，概述了体内与体外DNA存储系统，并重点讨论主流应用所面临的挑战[1, 20].

1　DNA数据存储技术原理概述

DNA数据存储通常包括信息编码、信息储存、信息检索和信息读取等主要步骤（图1）. 第一步，要对DNA中写入的数据进行信息编码，即通过计算机算法将比特序列映射到DNA序列. 然后合成编码后的DNA序列，生成每个序列的多个物理拷贝. DNA序列可任意排列但长度有限，因此比特序列被分解成较小的块，之后可将其重新组装成原始数据. 为此需要在每个块中加入一个索引[29]或在DNA序列中存储相互重叠的数据块[21]. Heckel等[30]描述了索引方式下的理论存储容量，并证明基于索引的编码是最优方案. 任何实用的存储信息量都需要合成大量不同DNA序列，此类工作更适于采用基于阵列的合成，因其能够以并行方式合成许多独特的序列[31]. 第二步，合成的DNA需要以合适的方式进行信息储存. Organick等[7]估计单个物理隔离的DNA池可以存储约10^{12}字节数据，而扩展至大型存储系统需要由此类DNA池组成的池库. 第三步，对相应的合成DNA池进行物理检索和采样，即信息检索. 为了避免读取池中所有数据，需要如计算机设计中的随机访问能力，或从庞大的数据集中选择要读取特定数据项的能力. 虽然这在主流数字存储介质中易于实现，但由于同一分子池中缺乏跨数据项的物理组织结构，这种操作在分子存储中较难实现. DNA数据存储中的随机存取可以通过选择性过程实现，如使用映射到数据项的探针进行磁珠提取或在编码过程期间使用与数据项关联的引物进行PCR[7, 29]. 最后一步，选定DNA样本后，下一步是对其进行测序，产生一组测序仪测序片段，并将其以高保真度解码回原始数字数据即信息读取，其成功与否取决于整个过程中的测序覆盖率和错误率.

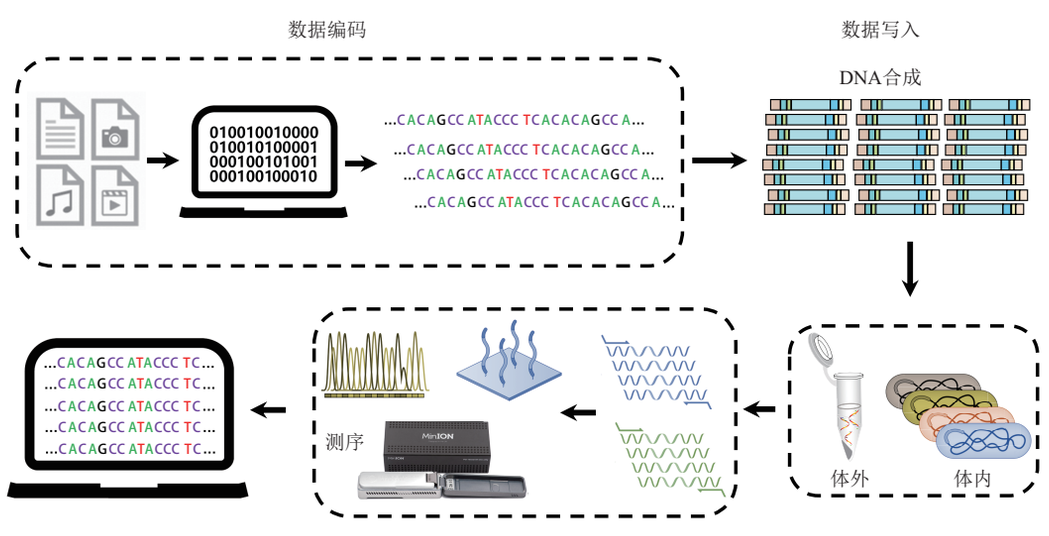

Fig. 1　The overall framework of DNA data storage

图1　DNA数据存储整体框架图

2 体外DNA数据库

自Baum[32]设想用合成DNA构建大规模联想存储以来，已对多种DNA存储设计方案进行实验验证[33-34]. 在此回顾了最近构建的大型DNA数据库的基本流程（图2），重点介绍存储容量、编码密度、纠错、随机访问和可重写性等关键指标（表1）. 2012年，Church等[20]使用合成寡核苷酸对包含一本书、11张图片和一个计算机程序的数据进行编码. 他们将数据分割成块以避免长序列，并对每个碱基仅编码一位数据以减少不必要的序列模式. 在每个寡核苷酸中加入一个地址以对序列进行有序组装. 这种方法必须对整个寡核苷酸池进行测序和解码才能对文件进行检索，而无法进行随机访问. 经过扩增和共有序列比对，527万位数据被恢复，而只有10位数据发生错误. Goldman等[21]将不同文件类型（ASCII文本、PDF文件、JPEG图片和MP3音频）存储在具有纠错功能的DNA存储器中. 使用定制的Huffman代码，作者将每个字节转译成一系列三进制数，其可通过简单的旋转编码在不生成均聚物的情况下转换成核苷酸. 由此产生的长字符串被分割成具有4倍编码冗余的重叠片段. 相邻片段反向互补以减少系统误差，并在序列重构过程中采用多数投票算法进行纠错. 每个数据编码寡核苷酸包含一个二元地址（用于识别文件和文件内的位置信息）以及奇偶校验（用于纠错）. 作者采用这种方法准确重建了所有文件，但有两个25-nt区域需要人工干预才能恢复. 这些需要干预的区域包含形成自反向互补模式的重复序列因而无法合成，作者建议将输入随机化以避免编码过程中出现此类重复序列. 为了说明存储的鲁棒性和成本效率，作者对测序片段以低于1/10覆盖率进行二次采样以模拟低测序覆盖率，仍然完美地重建了数据. 利用独立的寡核苷酸池和可调的数据冗余，Bornholt等[29]采用优化的编码密度存储不同类型数据. 在其键值架构中，每个键定义文件编码链的存储池和引物的分配，从而在各存储池中实现基于PCR的高效随机访问. 作者通过有效载荷之间的异或操作引入块级冗余，并将结果编码成新链. 编码冗余可依据数据块重要性进行微调. 与Goldman[21]采用的编码相比，此方法将信息编码密度提高了一倍，并在较小的干预下实现了完整的数据恢复.

Yazdi等[35]提出一个支持无损随机访问和内容重写的DNA数据库. 他们设计了互不干扰引物，并将前缀同步代码应用于数据块，使序列交叉杂交最小化，并使用GBlocks或OE-PCR实现内容编辑，但此类基于字典的编码仅限于存储文本[35]. 在另一项研究中，Yazdi等[36]改进了编码策略并用易出错的纳米孔测序仪恢复了两个压缩图像. 约束编码可减少均聚物并平衡码字GC含量. 使用长码字（1 000 bp）能够实现高效编码，而且用数学方法构建的地址支持鲁棒性随机访问. 作者针对测序容错设计了一个流程，其中集成了共有序列比对，用于码字预估. 识别出高质量的测序片段后，可采用各类多序列比对算法生成不同的共有序列. 然后通过具有GC平衡约束的多数投票算法产生共有序列集，并对其通过BWA比对和误差校正进一步优化. 这种比对策略将多数插入和替换错误转化为易于通过均聚物检查进行纠正的删除错误. 尽管MinION测序仪的错误率较高，作者使用1.1×10^{23}字节/g的编码密度实现了无损数据恢复.

Grass等[5]对集成了纠错码的DNA存储文件进行长期化学保存，并用串联RS代码增加冗余以解决单碱基错误和完整序列丢失. 其内码可纠正每个序列中3个以上的任意碱基错误，外码可进一步纠正8.5%的序列错误或17%的完整序列丢失. 该编码方案提供了稳健的容错能力，但无法实现随机存取. 可将寡核苷酸封装在硅胶颗粒中保护DNA免受潮湿和氧化影响以便长期储存. 在为期一周的加速老化实验中，作者模拟DNA在4个半衰期内的衰变，并通过氟化物蚀刻释放寡核苷酸，并完全恢复了储存的数据. Blawat等[37]对22 MB压缩电影进行鲁棒性存储和无损恢复，实现了DNA存储容量的飞跃. 基于对实验数据的系统分析，作者提出可通过定制前向纠错方案减少DNA合成、PCR扩增和测序过程中的各类错误. 在编码过程中，每个数据字节被映射到5个核苷酸构成的数据块（DNA符号）. 通过将双位元组映射到DNA符号中不同位置的核苷酸可使碱基替换引起的错误降到最低. 使用交替重复DNA符号对二进制数位进行编码可检测出InDel错误并有助于消除自反补序列. 每个寡核苷酸包含地址、数据有效载荷和错误检测代码三部分. 复杂BCH编码用于地址保护，RS块码用于对连续有效载荷冗余进行编码以及对每个寡核苷酸的奇偶校验进行循环冗余检查. 这种编码方案在实际的DNA存储系统中可实现非常小的残差概率.

Erlich和Zielinski[38]提出一种基于喷泉码的简

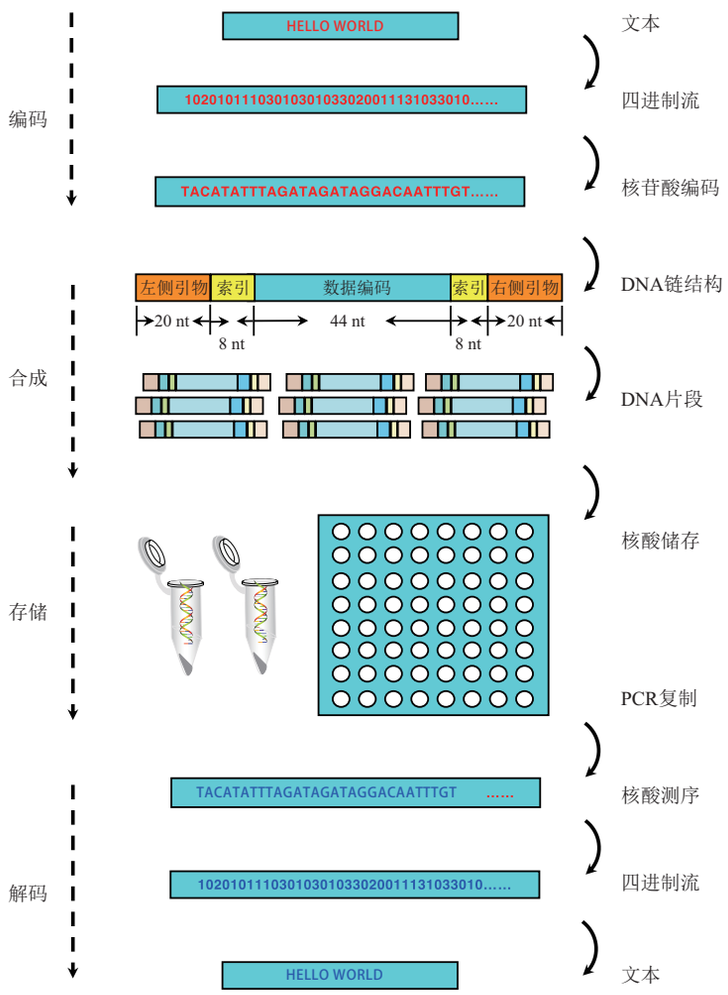

Fig. 2　In vitro DNA storage process

图2　体外DNA存储的过程

洁DNA存储策略，并对2.14 MB压缩数据（包括操作系统、软件、电影和PDF/文本/图像文件）进行存储，其密度相当于核苷酸信息容量的86%. 作者将二进制数据分为互不重叠片段，每次使用时先用特定分布函数生成随机数，再根据随机数选择若干片段. 每组选择的片段被编码为一个“液滴”，其中包含作为有效负载片段的逐位异或计算结果，以及用于片段识别的伪随机编号生成器种子和用于纠错的RS代码. 整个液滴形成过程用Luby变换描述，重复该变换以生成足够数量的二元液滴池. 采用从｛00，01，10，11｝到｛A，C，G，T｝的直接映射将液滴转换为寡核苷酸，并在此过程中去除具有均聚物或GC含量不理想的寡核苷酸，重复此转化过程直至发现有效寡核苷酸. 喷泉码的性质允许通过收集足够的液滴并反转Luby变换以重建数据. 该编码方案通过生成不同数量的寡核苷酸实现高度可调的冗余，同时不会使算法设计复杂化. 去除错误寡核苷酸并保留分析高质量测序片段保证了解码的高度鲁棒性. 作者进一步测试了采用串行PCR扩增和在稀释溶液中的存储，证实在最大物理密度2.15×10^{17}字节/g条件下可实现几乎无限的数据检索和无损恢复.

Organick等[7]用鲁棒性随机访问扩展了DNA存储容量，仅用5×的测序覆盖率实现了>200 MB编码压缩数据（包括35个不同大小/类型的文件）的无损解码. 作者设计了一个流程来进化并优化大量正交引物，从均聚物、自身互补性、GC含量等方面对其进行评分和筛选. 数据编码过程中对二进制数位进行伪随机化和片段分割，RS外码用于引入冗余，内码用于将数位转换为核苷酸，然后为每

个寡核苷酸分配引物以实现基于PCR的随机访问. 由于覆盖率较低，其解码器被设计成最大限度地利用包括噪声在内的可用测序片段，并根据相似度对其进行迭代聚类. 为了估计原始序列，通过轨迹重建从每个聚类中生成一个共有序列，然后重新使用内部/外部代码和随机化以重建数据. 为了进一步测试误差容限，作者将两个文件的寡核苷酸组装成长序列，使用易出错的MinION对其进行测序，尽管测序片段覆盖率较低，但成功地恢复了数据. 在撰写本文时，该研究组人员在DNA中成功储存了＞400 MB数字数据.

表1　DNA数据存储的代表方案

Table1　Representative scheme for DNA data storage

研究	总体数据	合成	测序	覆盖度	重组	链长度	每个核苷酸字节数	仅有有效载荷	随机访问
Church等[20]	650 kB	磷酰胺（沉积）	合成测序	3 000×	索引	115	0.60	0.83	否
Goldman等[21]	630 kB	磷酰胺（沉积）	合成测序	51×	部分重叠	117	0.19	0.29	否
Grass等[5]	80 kB	磷酰胺（电化学）	合成测序	372×	索引	158	0.86	1.16	否
Bornholt等[29]	150 kB	磷酰胺（电化学）	合成测序	40×	索引	117	0.57	0.85	是
Erlich和Zielinsky[38]	2 MB	磷酰胺（沉积）	合成测序	10.5×	种子	152	1.18	1.55	否
Blawat等[37]	22 MB	磷酰胺（沉积）	合成测序	160×	索引	230	0.89	1.08	否
Organick等[7]	200 MB	磷酰胺（沉积）	合成测序	5×	索引	150~200	0.81	1.10	是
Anavy等[39]	8.5 MB	磷酰胺（沉积）	合成测序	1 640×	索引	194	1.94	2.64	否
Choi等[40]	854 B	磷酰胺（柱）	合成测序	250×	索引	85	1.78	3.37	否
Yadzi等[35-36]	3 kB	磷酰胺（柱）	纳米孔	200×	索引	880~1 000	1.71	1.74	是
Organick等[7]	33 kB	磷酰胺（沉积）	纳米孔	36×	索引	150	0.81	1.10	是
Lee等[23]	18 B	酶法（柱）	纳米孔测序	175×	NA	150~200	1.57	1.527	否

3　体内DNA存储

基因组编辑可用于在生物体内设计基于DNA的存储系统，体内DNA的数据存储方法如图3所示. 早期关于细胞存储的研究利用反转重组将二进制状态记录到基因组DNA中. Bonnet等[41]通过反复反转和重建大肠杆菌基因组DNA片段实现可重写数字寄存器. 为了实现两种状态之间的可控切换，作者利用带有切除酶辅助因子的噬菌体整合酶来调节重组反应的方向性. 整合酶的单独表达能够反转具有预设识别位点的DNA片段，而整合酶和切除酶的共表达能够使片段回复到其原始方向. 这种基于DNA的锁存器存储方式支持在体内进行反复状态切换，并可在100代细胞中保持二进制状态. Siuti等[42]进一步整合各种基于DNA反转的功能模块并在大肠杆菌细胞中构建了集成逻辑/存储系统. 虽然可利用重组酶识别位点的特定关联或重叠来设计更大规模的存储器阵列和细胞状态运算器[43-44]，但基于重组酶的系统必须依赖高度正交的酶和长DNA片段编码单数据位，这不仅限制了存储器的可扩展性，还未充分利用DNA的信息容量. Farzadfard和Lu[45]设计了“SCRIBE”系统来实现体内可扩展的模拟存储器. 该系统利用ssDNA模板与重组酶共表达，并利用转录信号或光等调控输入诱导基因组DNA定位突变. 互补模块可用于反复重写，不同模板ssDNAs可用于在独立位点上进行多路记录以提供可编程性和可扩展性. 模拟信息（如瞬时信号的幅度和持续时间）以跨细胞群分布的累积突变的形式被记录.

为了设计鲁棒性细胞存储，科学家们还从活细胞的免疫机制，特别是CRISPR中获得了启示. Shipman等[25]利用CRISPR阵列（图3a）中的定向间隔序列采集来记录细菌基因组中的时间信息并显著改善了存储容量. 该方案可将任意信息编码成预设长度的合成寡核苷酸并将其通过电穿孔导入过表达Cas1-Cas2蛋白复合物的细胞，之后这些合成的寡核苷酸被作为原型间隔序列整合到扩展的CRISPR阵列中. 由于新的间隔序列总是被添加到阵列前端，整合后间隔序列的排序可以反映采集事件的时间史. 作者注意到在原型间隔序列中添加5'-PAM可以提高采集频率，并可在整合过程中确定间隔序列方向. 此系统可以多种方式记录，如序列内容、时间顺序和集成间隔序列方向. 因为间隔序列采集是在单细胞水平上随机发生的，存储内容读

取依赖于细胞群体CRISPR阵列分析. Shipman等[46]重点研究了存储的可扩展性，并在其后续研究中优化了原型间隔序列设计和数据重建策略. 作者将帧像素编码为人工合成的原型间隔序列，并通过连续电穿孔将一部简短的GIF电影存储到细菌基因组中，随后通过CRISPR阵列测序以高精度恢复了存储的数据. Sheth等[26]开发了“TRACE”框架并通过整合细胞内产生的DNA片段来记录CRISPR阵列中的生物信号. DNA间隔序列（触发DNA）的表达受到胞内信号的时间特征调节，因此可将时间生物信号转化为触发DNA的丰度变化. 为了记录代表信号缺失的时间间隔，以背景速率将参考间隔序列加入到基因组CRISPR阵列中. 作者通过分析细胞群CRISPR阵列中采集间隔序列的频率和排序，重建细胞信号的动态时间史，并使用条形码CRISPR阵列实现了高精度的多路复用时间记录.

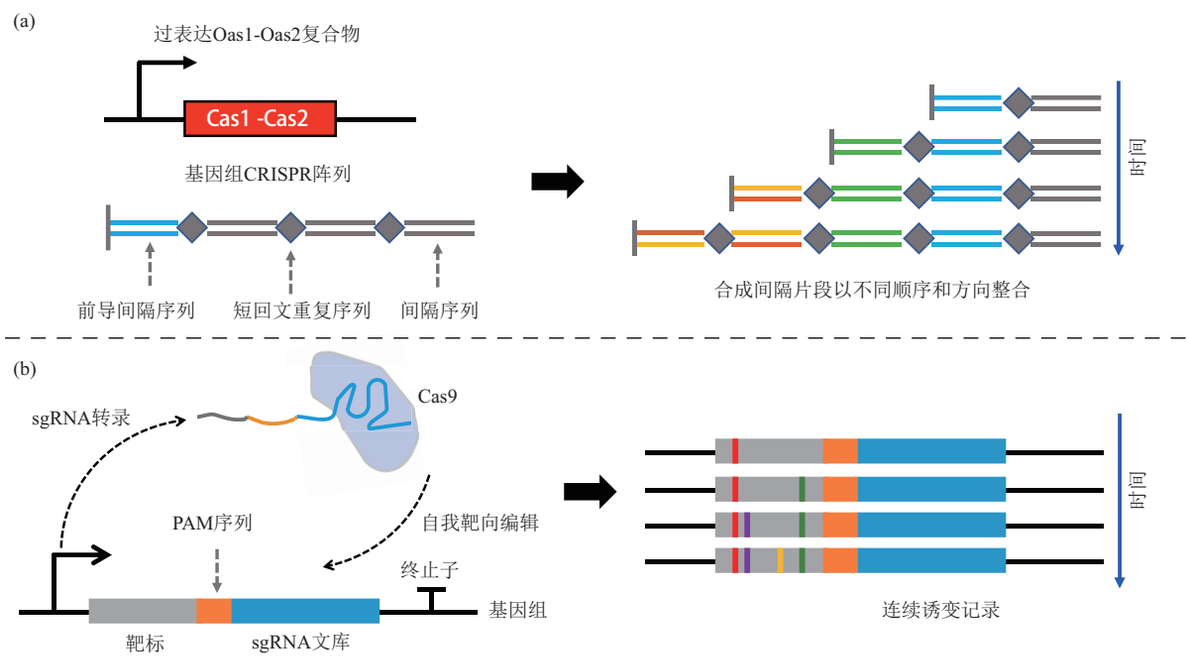

Fig. 3　A data storage process *in vivo* DNA

图3　DNA体内的数据存储过程

定向间隔序列采集并不是实现基于CRISPR存储的唯一方法. Perli等[24]利用自靶向CRISPR/Cas（图3b）设计可编程和多路复用的存储器架构，用于在人体细胞中进行连续纵向记录. 作者在sgRNA编码位引入PAM序列（protospacer adjacent motif），不断将Cas核酸酶活性引向其在基因组上的sgRNA编码区，从而实现自靶向sgRNA（stgRNA）位点累积突变. 通过将stgRNA或Cas9的表达与不同的诱导因子相偶联，可以在体内连续记录细胞活动的持续时间和强度等事件，还可在独立的stgRNA位点实现多路复用记录，之后通过分析细胞群体中sgRNA位点的进化模式来重建记录. 此方法设计新颖但也存在一些局限性，如长时间重复自靶向事件将截短stgRNA并导致靶向特异性降低或PAM丢失. 因此长期记录必须依赖于较长的stgRNA，这可能会使序列设计复杂化并降低可扩展性. 为了提高存储容量和数据解释能力，作者建议使用CRISPR碱基编辑等技术在stgRNA上引入更明确的突变. Frieda等[47]提出“MEMOIR”系统，其通过CRISPR/Cas介导的突变在条形码便笺上记录细胞状态，可根据从原位单细胞读取的数据分析便签删除模式跟踪细胞谱系和动态细胞事件历史，而无需中断细胞活动进行测序. Tang和Liu[27]开发了“CAMERA”的框架，通过CRISPR介导的多拷贝质粒操作实现可重写模拟记录器. 第一种策略是把模拟信息如信号振幅或持续时间转化为两个互斥质粒之间拷贝数比的变化. 这两个记录质粒被设计为具有几乎相同的序列，使其可稳定共存于细胞中且可通过诱导CRISPR/Cas活性选择性地切割其中一个质粒. 这种质粒补偿系统可以忠实记录多个模拟

信号并支持反复擦除/重写. 第二种策略是设计含有CRIPSR碱基编辑器的写入质粒，并以质粒上碱基突变为信号进行记录. 虽然引入正交诱导调节因子限制了记录的多样性，由于每个细胞包含大量记录质粒，CAMERA框架能够以小细胞群（10~100个细胞）实现灵敏可靠的记录[48-50].

4 DNA数据存储技的应用障碍与挑战

尽管DNA数据存储技术不断发展，但其在工程实践中依然面临障碍.

首先，在信息编码和解码过程中容易产生错误，即DNA合成和测序过程容易出错. 关于DNA数据存储的论文显示，每个碱基在每个位置的错误率约为1%[27, 30-31]. 即对于DNA链中的给定位置，当合成并回测序列时，约1%的测序片段将在该位置出现错误. 此估值适用于采用化学法在阵列中合成DNA，并使用Illumina进行合成测序. Yazdi等[35]和Bornholt等[29]观察到多数错误是由测序造成的. Organick等[7]发现纳米孔测序误差率约为10%. Heckel等分析以前三项研究的结果并进一步描述了编码通道的特征，表明错误大多源于合成和测序，而DNA操作、PCR和存储则可能会导致擦除，即某些序列在混合物中比例降低[5, 21, 38, 51].

存储应用的终端客户无法承受这种程度的错误风险，因此在原始存储介质中附加纠错代码至关重要. 现代磁介质的原始误差率同样为1%左右[52]. 简言之，可靠的数据存储和检索需要针对所有类型介质（以及诸如无线电之类的通信信道）的纠错方法. 计算机科学中有一个领域叫做信息理论（编码理论），其主要涉及开发编码方案，在有噪声的介质和通信渠道上可靠地传递数字数据. 与其他存储通道只有替换错误不同，DNA通道也可出现碱基插入和缺失，增加了编码的复杂程度. 在中短期内访问延迟（读取时间）可能会继续保持在较高的水平（几分钟到几小时），但只要带宽（数据写入和读取通量）较高，体外DNA数据存储就可以与商业介质共存或取代其在档案数据存储中的应用. 这是因为档案存储可以承受更高的延迟，并将从更小的资源占用和更低的静息数据能量成本中获益.

其次，DNA数据存储的整体写入通量应在每秒千字节左右. 估计未来10年能够与主流云档案存储竞争的系统需要达到每秒千兆字节的读写通量. 目前的合成能力与其有6个数量级的差距，测序能力则有2~3个数量级的差距. 在成本方面，2016年磁带存储成本约为16美元/TB，并以每年约10%的速度下降[53]. DNA合成成本通常是保密的，但据业内著名分析师估计，阵列法DNA合成成本约为每碱基0.000 1美元，相当于8亿美元/TB，比磁带高7~8个数量级[54]. 虽然通量和成本差距令人望而生畏，但预计相应的成本会不断降低，因为可将成本在更多数量的合成底物和更大批量的DNA中进行分摊. 由于数据存储所需每个序列的拷贝数比生命科学低几个数量级，通过更多的平行合成和更小的生长点尺寸来提高通量也将以相应比例降低试剂使用成本.

再次，是DNA分子的物理存储和保存. 尽管已有证据表明可读取数千年（或数十万年）前的DNA[55]，但DNA的降解速度可能比这快得多，这取决于其所处的条件（如高温、高湿和暴露于紫外线可能导致其降解）[6, 56]. 为了解决这个问题，不同的研究小组提出了多种方法，为DNA保存提供适宜的条件[57-58]. 化学方法包括脱水和/或冻干、添加剂（如BiomatricaDNAStable或海藻糖）或用保护材料（如二氧化硅）进行化学封装[5]. 化学溶液和添加剂的制备更加迅速，而封装在较高湿度（50%）环境中可提供更长和更好的保护. DNA储存容器有多种材料和形式，如滤纸（Whatman）、密闭不锈钢微胶囊（Imagene）和塑料孔板（Biomatrica）. 这些容器都是为生物样品量身定制，并对其纯度进行了优化，因而存储密度和成本可能会受到一定影响. DNA数据物理存储库需要在完全自动化和可扩展的模式下运行，同时又不显著降低存储密度，这在很大程度上仍然是一个有待研究的课题. 要实现系统自动化并使其能够用于大规模档案存储，尚存在诸多挑战. 存储环境通常需要将人为干扰最小化，而多数DNA合成和测序之外的操作仍由实验人员在实验室环境中进行. 最近，Takahashi等首次公开展示全自动DNA数据存储系统，微流体的最新进展同样令人鼓舞[40, 58-61]，期望这些技术能用于DNA数据存储的自动化.

5 研究展望

与传统存储介质相比，DNA具有出色的体积密度、寿命和能量利用系数，可长期保存数据且对环境影响较小. 存储在DNA中的信息可以大规模并行方式进行复制和处理，而不需像传统存储架构那样依赖复杂的电路布线和严格的空间布局. 这些特征使DNA成为在分子水平上构建高度密集、耐

用和多功能存储的必选介质，但基于DNA的存储系统同样面临限制和挑战. 基于DNA的存储在读/写速度方面无法与电子存储相抗衡. 即使能够实现完全自动化，在合成、测序、实验室样品制备等过程中也不可避免地会产生延迟. 因此目前基于DNA的存储主要用于数据存档. 然而以经济可行的规模建立DNA存档受限于常规化学合成的慢速、不可靠、价格昂贵，而这些缺点在过去40年中几乎没有改进[61]. 有几种可能的途径用于解决DNA合成中的这些障碍，如最近研究表明一种称为末端脱氧核苷酸转移酶的特殊聚合酶可无需模板高保真合成定制DNA[62]. 随着学界和商业界的不断努力，酶法或许会彻底改变DNA合成现状，实现更大的通量、更高的保真度和更低数量级的成本[39]. 尽管可以合成DNA之外更简单、更便宜的分子来设计分子存储系统，但检索和操作非标准合成分子中的编码信息可能仍会受到现有技术和仪器的限制[63].

相比之下，基于DNA的存储系统则利用了生命科学中快速发展的工具. 鉴于其技术预备性、可实现规模性、鲁棒性以及可编程碱基配对支持分子计算和数据库操作的可能性，我们认为DNA是迄今为止构建实用分子存储系统最可行的材料. 创新核苷酸信息映射策略可以减少常规合成的复杂度. 尽管编码方案不同，目前多数体外DNA存储依赖大量独特序列的从头合成. 虽然编码密度是一个关键考量，但从头合成的方法不可扩展，随着数据规模的增加，成本会变得很高. 另一种方法是设计和预制短DNA片段文库作为码字词典，通过重新排列片段并经由有效的酶促反应将它们串联成可寻址的有效载荷链来表示任意数据. 通过精心优化的码字设计，可最大限度减少错误交叉干扰并提高序列多样性以实现数据编码的灵活性. 这种模块化方法或可大幅提高扩展性和成本效率，为DNA存储的商业化应用铺平道路.

目前DNA存储与纠错方案对合成和测序错误甚至完整序列丢失具有容错能力，若编码和纠错策略能够进一步优化，基于DNA的存储就可采用低成本低保真技术得以实现. 我们认为这一发展路径有很大的改进空间，但尚需进行如DNA存储误差模型全面评估等大量工作. DNA与生物相关，需要详细的法规来指导技术商业化并维护数字和生物安全. 虽然挑战依然存在，但合成DNA存储系统的未来依然光明，并可能会对全球数据管理和医疗保健等领域产生深远影响. 在学术界和工业界的共同努力下，相信在可预见的未来会有很多方法构建低成本且实用的DNA存储.

参　考　文　献

[1] Sheth R U, Wang H H. DNA-based memory devices for recording cellular events. Nature Reviews Genetics, 2018, **19**(11): 718-732

[2] Goda K, Kitsuregawa M. The history of storage systems. Proceedings of the IEEE, 2012, **100**(Special Centennial Issue): 1433-1440

[3] Rutten M G, Vaandrager F W, Elemans J A, *et al*. Encoding information into polymers. Nature Reviews Chemistry, 2018, **2**(11): 365-381

[4] Greengard S. Cracking the Code on DNA Storage. USA: ACM New York, 2017

[5] Grass R N, Heckel R, Puddu M, *et al*. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angewandte Chemie International Edition, 2015, **54**(8): 2552-2555

[6] Zhirnov V, Zadegan R M, Sandhu G S, *et al*. Nucleic acid memory. Nature Materials, 2016, **15**(4): 366-370

[7] Organick L, Ang S D, Chen Y-J, *et al*. Random access in large-scale DNA data storage. Nature Biotechnology, 2018, **36**(3): 242-248

[8] Sakakibara Y, Mi Y. DNA computing and molecular programming// Lecture Notes in Computer Science, 2011: 6518

[9] Wiener N. Machines smarter than men? Interview with Dr. Norbert Wiener, noted scientist. US News World Rep, 1964, **56**: 84-86

[10] Neiman M S. On the molecular memory systems and the directed mutations. Radiotekhnika, 1965, **6**: 1-8

[11] Dawkins R. The blind watchmaker. Journal of Animal Ecology, 1986, **16**(3): 423-424

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

[13] Clelland C T, Risca V, Bancroft C. Hiding messages in DNA microdots. Nature, 1999, **399**(6736): 533-534

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

[15] Wong P C, Wong K-K, Foote H. Organic data memory using the DNA approach. Communications of the ACM, 2003, **46**(1): 95-98

[16] Arita M, Ohashi Y. Secret signatures inside genomic DNA. Biotechnology Progress, 2004, **20**(5): 1605-1607

[17] Yachie N, Sekiyama K, Sugahara J, *et al*. Alignment-based approach for durable data storage into living organisms. Biotechnology Progress, 2007, **23**(2): 501-505

[18] Portney N G, Wu Y, Quezada L K, *et al*. Length-based encoding of binary data in DNA. Langmuir, 2008, **24**(5): 1613-1616

[19] Ailenberg M, Rotstein O D. An improved Huffman coding method for archiving text, images, and music characters in DNA. Biotechniques, 2009, **47**(3): 747-754

[20] Church G M, Gao Y, Kosuri S. Next-generation digital information storage in DNA. Science, 2012, **337**(6102): 1628

[21] 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

[22] Gibson D G, Glass J I, Lartigue C, *et al*. Creation of a bacterial cell controlled by a chemically synthesized genome. Science, 2010, **329**(5987): 52-56

[23] Lee H H , Kalhor R , Goela N , *et al*. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. Nature Communications, 2019, **10**(1): 1-12

[24] Perli S D, Cui C H, Lu T K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. Science, 2016, **353**(6304): 1115-1115

[25] Shipman S L, Nivala J, Macklis J D, *et al*. Molecular recordings by directed CRISPR spacer acquisition. Science, 2016, **353**(6298): 463-463

[26] Sheth R U, Yim S S, Wu F L, *et al*. Multiplex recording of cellular events over time on CRISPR biological tape. Science, 2017, **358**(6369): 1457-1461

[27] Tang W, Liu D R. Rewritable multi-event analog recording in bacterial and mammalian cells. Science, 2018, **360**(6385): eaap8992

[28] Glaser J I, Zamft B M, Marblestone A H, *et al*. Statistical analysis of molecular signal recording. Plos Comput Biol, 2013, **9**(7): e1003145

[29] Bornholt J, Lopez R, Carmean D M, *et al*. A DNA-based archival storage system//Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems, 2016: 637-649

[30] Heckel R, Shomorony I, Ramchandran K, *et al*. Fundamental limits of DNA storage systems//2017 IEEE International Symposium on Information Theory (ISIT), IEEE, 2017: 3130-3134

[31] Kosuri S, Church G M. Large-scale de novo DNA synthesis: technologies and applications. Nature Methods, 2014, **11**(5): 499-507

[32] Baum E B. Building an associative memory vastly larger than the brain. Science, 1995, **268**(5210): 583-585

[33] Reif J H, Labean T H, Pirrung M, *et al*. Experimental construction of very large scale DNA databases with associative search capability//International Workshop on DNA-Based Computers. Berlin: Springer, 2001: 231-247

[34] Stewart K, Chen Y-J, Ward D, *et al*. A content-addressable DNA database with learned sequence encodings // International Conference on DNA Computing and Molecular Programming. Cham: Springer, 2018: 55-70

[35] Yazdi S H T, Yuan Y, Ma J, *et al*. A rewritable, random-access DNA-based storage system. Scientific Reports, 2015, **5**:14138

[36] Yazdi S H T, Gabrys R, Milenkovic O. Portable and error-free DNA-based data storage. Scientific Reports, 2017, **7**(1): 1-6

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

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

[39] Anavy L, Vaknin I, Atar O, *et al*. Improved DNA based storage capacity and fidelity using composite DNA letters. bioRxiv, 2018: 433524. https://doi.org/10.1101/433524

[40] Choi K, Ng A H, Fobel R, *et al*. Digital microfluidics. Annual Review of Analytical Chemistry, 2012, **5**(1):413-440

[41] Bonnet J, Subsoontorn P, Endy D. Rewritable digital data storage in live cells via engineered control of recombination directionality. Proc Natl Acad Sci USA, 2012, **109**(23): 8884-8889

[42] Siuti P, Yazbek J, Lu T K. Synthetic circuits integrating logic and memory in living cells. Nature Biotechnology, 2013, **31**(5): 448-452

[43] Yang L, Nielsen A A, Fernandez-Rodriguez J, *et al*. Permanent genetic memory with> 1-byte capacity. Nature Methods, 2014, **11**(12): 1261-1266

[44] Roquet N, Soleimany A P, Ferris A C, *et al*. Synthetic recombinase-based state machines in living cells. Science, 2016, **353**(6297): 363

[45] Farzadfard F, Lu T K. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. Science, 2014, **346**(6211): 1256272

[46] 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. Nature, 2017, **547**(7663): 345-349

[47] Frieda K L, Linton J M, Hormoz S, *et al*. Synthetic recording and in situ readout of lineage information in single cells. Nature, 2017, **541**(7635): 107-111

[48] Nielsen A A, Voigt C A. Multi-input CRISPR/C as genetic circuits that interface host regulatory networks. Molecular Systems Biology, 2014, **10**(11): 763

[49] Kim H, Bojar D, Fussenegger M. A CRISPR/Cas9-based central processing unit to program complex logic computation in human cells. Proc Natl Acad Sci USA, 2019, **116**(15): 7214-7219

[50] Farzadfard F, Gharaei N, Higashikuni Y, *et al*. Single-nucleotide-resolution computing and memory in living cells. Molecular Cell, 2019, **75**(4): 769-780. e764

[51] Heckel R, Mikutis G, Grass R N. A characterization of the DNA data storage channel. Scientific Reports, 2019, **9**(1): 1-12

[52] Albrecht T R, Arora H, Ayanoor-Vitikkate V, *et al*. Bit-patterned magnetic recording: theory, media fabrication, and recording performance. IEEE Transactions on Magnetics, 2015, **51**(5): 1-42

[53] Song X, Reif J. Nucleic acid databases and molecular-scale computing. ACS Nano, 2019, **13**(6): 6256-6268

[54] Allentoft M E, Collins M, Harker D, *et al*. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proceedings of the Royal Society B: Biological Sciences, 2012, **279**(1748): 4724-4733

[55] Bonnet J, Colotte M, Coudy D, *et al*. Chain and conformation stability of solid-state DNA: implications for room temperature storage. Nucleic Acids Research, 2010, **38**(5): 1531-1546

[56] Ivanova N V, Kuzmina M L. Protocols for dry DNA storage and shipment at room temperature. Molecular Ecology Resources, 2013, **13**(5): 890-898

[57] Howlett S E, Castillo H S, Gioeni L J, *et al*. Evaluation of

DNAstable™ for DNA storage at ambient temperature. Forensic Science International: Genetics, 2014, **8**(1): 170-178

[58] Takahashi C N, Nguyen B H, Strauss K, *et al*. Demonstration of end-to-end automation of DNA data storage. Scientific Reports, 2019, **9**(1): 1-5

[59] Prakadan S M, Shalek A K, Weitz D A. Scaling by shrinking: empowering single-cell'omics' with microfluidic devices. Nature Reviews Genetics, 2017, **18**(6): 345-361

[60] Newman S, Stephenson A P, Willsey M, *et al*. High density DNA data storage library *via* dehydration with digital microfluidic retrieval. Nature Communications, 2019, **10**(1): 1-6

[61] Carmean D, Ceze L, Seelig G, *et al*. DNA data storage and hybrid molecular-electronic computing. Proceedings of the IEEE, 2018, **107**(1): 63-72

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

[63] Service R F. DNA printers poised to jump from paragraphs to pages. Science, 2018, **362**: 143

Principle and Progress of DNA Data Storage

TENG Yue[1)*], YANG Shan[1)], LI Jin-Yu[1)], CUI Yu-Jun[1)], LIU Rui-Cun[1)], WANG Sheng-Qi[2)*]

([1)]*State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Academy of Military Medical Sciences, Beijing* 100071, China;

[2)]*Beijing Institute of Radiation Medicine, Beijing* 100850, China)

Abstract The gap between information production and data storage capacity is continuously growing, so there is an urgent need for new methods of high-density, persistent data storage. With advances in DNA synthesis and sequencing, attempts have been made to use synthetic DNA for data storage and information exchange. DNA storage has many advantages compared with hard disk information storage, including high information density specificity (data bits per gram) and long storage time. Using different algorithmic strategies, text, images, audio, and movies have been encoded into synthetic DNA for storage. In addition to big data storage applications, DNA may be valuable in the exchange of classified information. This review summarizes the basic principles of DNA data storage, introduces research progress into DNA storage *in vitro*, and analyzes works concerning data size, logic density, and DNA synthesis. We also describe *in vivo* molecular memory systems, including the adoption of CRISPR to design the DNA storage system. Finally, we discuss various influencing factors and challenges of data storage systems based on DNA.

Key words DNA, data storage, cell memory, nucleic acid database, molecular computer

DOI: 10.16476/j.pibb.2020.0224

* Corresponding author.

TENG Yue. Tel：86-10-63869835，E-mail：yueteng@sklpb.org

WANG Sheng-Qi. Tel：86-10-66932251，E-mail：sqwang@bmi.ac.cn

Received：July 7，2020 Accepted：October 16，2020