DNA数据存储

毛秀海 李 凡 左小磊*
①(上海交通大学医学院分子医学研究院 上海 200127)
②(上海交通大学医学院附属仁济医院 上海 200127)

摘 要：分子数据存储作为一种稳定性强、存储密度高的数据存储方式，表现出巨大的潜力。它有望解决当今日益增长的巨大信息量与存储能力之间差距不断扩大的问题。作为一种典型的分子数据存储方式，DNA数据存储可以作为一种替代性、变革性的存储介质，用于突破现用存储方式的物理极限，满足不断增加的数据存储需求。该综述将对DNA数据存储的历史、工作流程、及当前的发展状态进行概述，同时讨论现今DNA数据存储存在的问题、挑战及发展趋势。

关键词：分子数据存储；DNA数据存储；编码；解码；读取
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2020)06-1303-10
DOI: 10.11999/JEIT190852

DNA Data Storage

MAO Xiuhai LI Fan ZUO Xiaolei
①(*Institute of Molecular Medicine, School of Medicine, Shanghai Jiao Tong University, Shanghai* 200127, *China*)
②(*Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai* 200127, *China*)

Abstract: Molecular data storage has great potential as durable and high-density data-storage media, which will deal with the growing gap between produced information and the data storage ability. With storing data in molecular form, DNA can provide alternative substrates for storage to overcome the physical limits for existing medias. This review provides an overview of the history, process and the current status of the DNA data storage, and presents the problems of current data storage technology.

Key words: Molecular data storage; DNA data storage; Encoding; Decoding; Read

1 引言

当今世界已经完全进入大数据时代，一切生活相关的活动都涉及数据的存储和处理[1–3]。现代数据的指数型增长状况，已超过了现有存储器件容量的增长速度。然而现有的存储介质，例如磁性存储(例如磁带或硬盘驱动器)、光学存储(例如蓝光)和固态存储(例如闪存)，已不能满足日益增长的存储容量的需求，并成为人类不得不面对的难题。存储介质的评价主要参照以下性能指标：存储密度(每单位物理容量的比特数)、保存时间(数据可保存并可读取的最长时间)、访问速度(访问数据的延迟和带宽)和数据成本(包括每次读取时的成本)。存储密度、耐用性和数据读写的能源成本是数据储存的主要参考因素，其目的是储存大量的数据并实现低成本读写、长时间的保存。主流的数据存储通常是通过改变材料的性能来实现的：如闪存和相变存储器中的电学特性、蓝光光盘中的光学特性或硬盘驱动器和磁带中的磁性特性。虽然这些技术已经取得了飞速的发展，但它们的存储密度都已接近极限。相比之下，分子数据存储使用尽可能少的原子来存储一个比特的信息，实现高密度存储，使其更具吸引力。

针对该需求，科研工作者已开发诸多种类的分子或原子水平的数据存储形式[4]，其中以DNA分子为存储介质是最具吸引力与潜力的。自上世纪60年代以来，由于DNA分子具有高存储密度、高稳定性、易复制性等特征，学界已就DNA作为存储媒

收稿日期：2019-11-01；改回日期：2020-05-18；网络出版：2020-05-21
*通信作者： 左小磊 zuoxiaolei@sjtu.edu.cn
基金项目：中国科学技术部国家重点研发计划(2018YFA0902600)，国家自然科学基金(21804019, 21804088)，上海市浦江人才计划(19PJ1407300)
Foundation Items: The Ministry of Science and Technology of China (2018YFA0902600), The National Natural Science Foundation of China (21804019, 21804088), Shanghai Pujiang Program (19PJ1407300)

介的可行性展开讨论。近十年以来，随着DNA分子的合成手段以及新一代测序技术的进步和普及，以DNA分子为存储介质的研发及其相关技术得到了学术界和产业界越来越多的关注[5–11]。例如，DNA的存储密度可以高达10^{18} Byte/mm^3[12]，约比目前的最佳存储介质高6个数量级[3,12–14]。此外，与当前计算机存储系统相比，基于DNA存储系统的能源效率提高约100万倍。更重要的是，由于DNA测序仪在生命科学和医学领域的应用日益广泛，使得DNA信息读取能力也在一直进步与提高，并也将促进DNA数据存储的快速发展[15]。

与传统介质相比，DNA还展现了其他优势[16]。首先，DNA作为信息存储方式，显著提高了数据存储的保存时间。在远离光线、湿度和适宜温度的条件下，DNA可以保存几百到几千年[17]。因此，我们仍能从数千年前的化石中提取DNA序列信息。而商业磁带和光盘等存储介质通常只能保存几十年[18,19]。除此之外，DNA具有易于复制的独特优势，使用PCR技术，可让我们在短时间内低成本地复制大量数据，在更新维护即将过期的数据等方面，展现了巨大的优势[20]。例如，美国国会图书馆定期会将相当大容量的信息资源转移到新一代磁带上，并且磁带驱动器只能向后兼容有限数量的前几代磁带。这样极大限制了转移数据量和速率，同时磁带驱动器还不能满足磁带的发展速率。采用DNA存储技术，将极大提高其效率。同时，我们还可以利用DNA杂交及反应过程对数据执行图像相似性搜索等一系列智能操作。

DNA数据存储的基本过程包括将数字信息编译成DNA序列(编码)，合成实际DNA分子序列，组织这些DNA序列形成数据库并长期存储，检索和选择性地访问(随机访问)DNA序列，读取分子(测序)，转换成数据(解码)。本文将描述DNA数据存储编码和解码的技术现状，并重点介绍其随机访问和保存等方面的机理。此外，还讨论了DNA数据存储面临的挑战和发展趋势[21]。

2 DNA数据存储发展历史

使用DNA进行数据存储的理念可以追溯到20世纪60年代中期，当时Norbert Wiener和Mikhail Neiman讨论了基因记忆的概念[22,23]。然而那时DNA测序和合成技术仍在起步阶段，限制了DNA存储的发展。直至20年后，Davis[24]提出“Microvenus”实验，并实验性地证明了DNA数据存储的概念。在该实验中，Joe Davis利用DNA存储技术，把古日耳曼符文“female Earth”编码成一幅35 bit的图像。1999年，这个概念再次被证明[25]：他们将一段信息被写入DNA微粒中。与此同时，受到Davis工作的启发，其他研究组[14,24–30]也相继开展了一些基于DNA分子的活细胞存储工作。

直到21世纪10年代早期，Church等人[31]和Goldman等人[32]分别重新兴起了DNA数据存储的概念，并在DNA存储领域取得了重大突破。他们利用DNA分子实现数百kbit的数据存储，使得DNA数据存储在可预见的时间范围内成为可能。DNA存储在容量方面有明显的指数级提高，在仅仅6年的时间里大约提高了3个数量级。之后，DNA合成领域也得到了飞速发展：基于亚磷酰胺的DNA合成体系，在几十年中已逐渐完善；酶促DNA合成也作为一个新兴的研究领域，已经成功地用于数据存储。对于DNA数据读取，大多数研究使用的是Illumina推广的一种商业测序方法——合成测序法。最近，多个小组开展了纳米孔测序方法。尽管现在数据量不大，但是其显示了巨大的潜力。

如前所述，早期的DNA数据存储工作大多涉及体内克隆和存储元件[33]。最近，体内DNA数据存储方面研究致力于合成生物学方向，即在活体内将新信息记录到基因组的某些区域。然而，与体外DNA数据存储不同，由于修饰或者添加额外的DNA到活细胞中的操作复杂性，以及细胞体积较大导致存储密度的降低，使体内DNA数据存储将难以替代主流数据存储。但是，其将在记录细胞历史和环境的信息等方面，具有独特的优势。这样的系统可以被定义为分子标记磁带，它将记录细胞内动态的、按时间顺序排列的分子事件，并将相关信息存储在DNA分子中。

3 DNA数据存储实现

3.1 DNA数据存储整体框架

如图1所示DNA数据存储流程包括信息写入、信息保存、信息检索和信息读取4个部分，具体如下：

信息写入(Writing)：该部分主要包括DNA编码和DNA合成两个部分。DNA编码是通过计算机算法将比特串映射成DNA序列，即含有A, G, C, T的序列。DNA序列的长度是任意的，但也是有限的，因此信息比特串通常会被分成小的块，然后重新组装成原始数据。在重新组装中，需要在每个块中包含一个索引[17,31]或者在不同的DNA序列中存储重叠的块[32]。Heckel等人[34]从理论角度描述了索引方法下的存储容量，并证明了基于索引的编码方案是最优的。根据特定算法，加入一些冗余信息之后，转换后的DNA编码序列进一步加工成为DNA信息编码。然后进过第2步：DNA合成，将信息数据写入到一系列DNA分子中，同时生成每个序列的许

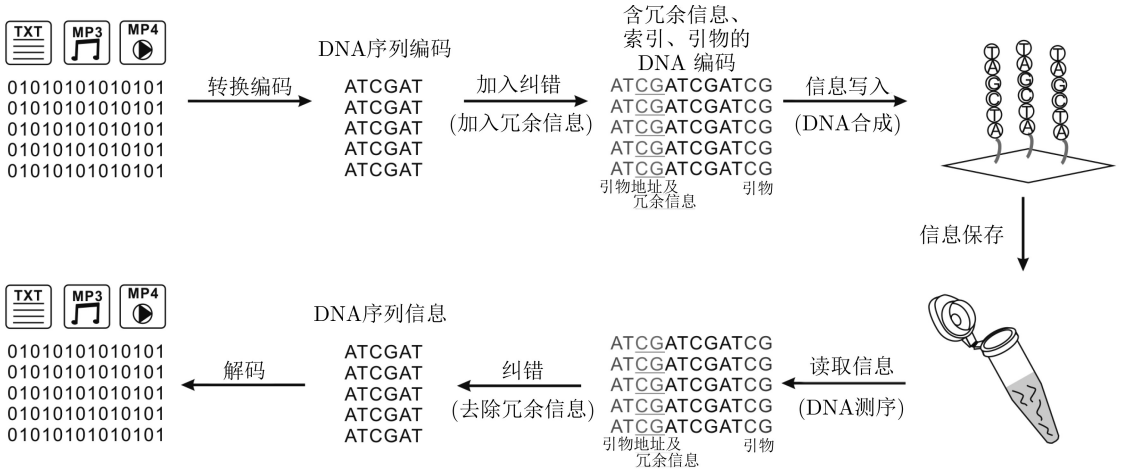

图 1 DNA数据存储整体框架图

多物理副本。面对巨大的信息量，需要合成大量的DNA序列。在该方面，基于阵列合成表现出极大的优势：其可以并行、同时合成大量不同的序列[35]。

信息存储(Store)：合成后的DNA需要储存，并形成一个存储系统。Organick等人[12]估计，一个单独的DNA样本库可以存储10^{12} Byte的数据。同时，基于这些DNA样本库，可以构建出一个大型存储系统[36]。

信息读取(Read)：信息读取包括DNA信息检索及测序和DNA信息解码两个部分。当接收到数据读取请求时，相应的DNA样本库需要被物理检索和采样。我们需要做到随机访问，能够从大量的数据中读取特定的数据项。虽然这一特性在主流数字存储媒体中很容易支持，但因为同一分子池中数据项之间缺乏物理组织，分子存储实现该工作更具挑战性。DNA数据存储中的随机存取可以通过选择性的过程来实现，例如根据探针检索，将含有目标数据项的磁珠提取，或者在编码过程中使用与数据项相关的引物进行PCR[36,37]。选定一个DNA样本后，下一步是对它进行测序，产生一组DNA序列信息。经过纠错和解码，这些序列信息被翻译成真实的数据。

3.2 DNA数据编码与解码

为了将数据存储在DNA中，必须首先通过翻译代码将其转换为DNA序列。在设计DNA编码算法时，有几个标准很重要。首先，编码必须有效利用DNA。尽管合成DNA的成本变得越来越便宜，但长链DNA的合成仍然相对昂贵。DNA的信息容量最多是每个核苷酸2 bit[38]，例如A = 00, C = 01, G = 10和T =11。但是，此理论容量为受几个因素的限制。首先，不同A·T: G·C比率的DNA序列含有不同的解链温度(Tm值)，在同一条件下进行PCR操作，其效率会受到很大影响。此外，DNA序列内部由于相互作用产生折叠，导致测序过程中产生较高的错误率。因为这些因素，每个核苷酸并非都可以放置在每个位置，从而限制了存储容量[39]。理论上讲，DNA存储可以看作是在DNA链信道上传输信息，通过对链进行测序、解码来接收信息。如上所述，由于各种类型的错误，该信道是含有噪音的。Shannon[40]提出的信息理论定义了噪音信道的容量概念，并提供了一种数学模型进行计算容量。此信道容量为可靠传输信息的速率提供了严格的上限。在经典的信息理论中，噪声是独立分布的。与经典的信息理论不同，DNA数据存储中的错误模式在很大程度上取决于输入序列。基于此，Erlich等人[38]在综合各方面因素之后，推导出DNA存储设备总的Shannon信息容量大约为，每核苷酸1.57 bit。

其次，设计DNA编码算法另外还需要考虑两个重要方面：(1)对信息传递需要进行纠错设计；(2)实现简单而直接的数据检索。很多因素都会导致DNA合成和测序容易出错，如前面所述的一些。有相关DNA数据存储的论文报告显示：每个碱基每个位置大约1%的错误。更准确地说，在用Caruthers的化学方法通过阵列合成DNA，并使用Illumina仪器对其进行测序时，对于DNA链中给定的位置，在合成、测序过程中会出现大约1%的错误率[37,38]。Yazdi[37]和Bornholt[36]等人指出大多数错误是由于测序引起。与此相比，最新的纳米孔测序，其误差约为10%[12]。随后，Heckel等人[41]通过分析之前的研究结果，进一步对编码通道进行了表征，发现错误主要来自合成和测序：DNA操作、PCR和存储可能会导致DNA碱基的丢失，从而导致某些序列在重组时会不成比例地变化。如果在存储应用中，向用户公开这种级别的错误将是灾难性的。因此，最重要的是覆盖错误，即修正原始存储媒体上的代码。有趣的是，现代磁性介质也存在大

约1%的错误率。简而言之，所有介质类型的可靠数据存储和检索都需要进行纠错处理。利用信息论或编码理论，通过开发编码方案，可以使得我们在含有噪音的媒介和传播渠道上进行可靠地传输数据。更为重要的是，DNA存储通道还可能出现碱基插入和删除的错误，这使得编码更有挑战性。此外，在进行DNA数据存储时，另一个重要方面是，由于无法合成任意长度的DNA链，因此不可能创建包含所有数据的一条链。相反，必须将数据分成多个较小的片段，每个数据片段编码成DNA序列的一部分。与此同时，基于编码设计，解码器必须知道所有片段的顺序，并允许检索所有数据。

基于此，DNA数据编码和解码必须考虑纠错码的设计和数据检索编码的设计，如下所述：

3.2.1 纠错码的设计

纠错码归纳起来就是添加冗余信息，这用于提高在出现错误或丢失数据的情况下检索到原始信息的正确率。冗余越多，结果存储过程对错误(或损失)的容忍度就越大。冗余可以有两种基本类型。物理冗余来自于给定DNA序列的多个物理副本。逻辑冗余来自于在DNA序列中编码数据时嵌入额外的信息。使用阵列合成DNA，自然会产生大量相同DNA序列的物理拷贝。虽然物理冗余有助于容忍衰减，并允许在合成和测序中出现一些错误，但它并不足以保证高保真率。例如，尽管有很大程度的物理冗余，但是Church等人[31]没有任何形式的逻辑冗余，也没有实现零比特错误。有几种逻辑冗余设计方法，我们将在下面进行介绍。

多个研究项目开发了DNA数据存储编码，以应对读写错误和随着时间推移时DNA发生降解等问题[12,17,34,36–38]。当比特被编码成DNA序列时，它们会经过一系列的转换和检查。由于比特的数目比目前科技合成的单一DNA序列所能容纳的比特数目大，因此比特被分成更小的序列。大多数DNA数据存储工作中，都会将索引添加到每个更小的序列中，以确定其在原始文件中的相对位置，如图1所示。这些序列可能经过一个逻辑异或操作，即用已知种子产生的伪随机数来确保DNA链是不同的，然后通过编码(如Reed Solomon编码)来增加冗余，以纠正错误，如图1。然后，不同的编码器可以使用不同的规则将比特转换为序列。常见的转换规则是避免重复碱基(均聚物)，并在两端添加引物靶位点。

第1种现代纠错码出现在20世纪40年代[40]。所有的纠错码都会给要在通道上存储或传输的原始数据增加冗余。接收方可以使用这些额外的冗余数据来检查接收到的消息是否一致，如果不一致，则可能重构原始数据。要添加的冗余数据量可以根据通道的噪声特性、使用的代码和成功解码的期望概率而变化。

第2种采用计算机技术的纠错方法是使用Reed-Solomon码[42]。Reed Solomon编码可以追溯到20世纪60年代，它们通常用于DNA数据存储，但也被用于其他各种应用中，如光盘(如小型光盘、数字视频和蓝光光盘)、2D可视代码(如快速响应(QR)代码)和数据传输(如WiMax)。Reed Solomon码的基本思想是将原始数据映射成一组符号，即编码数据的基本单元。随后，根据符号映射过程中的线性方程组系数，符号解映射至原始数据。这些代码可以纠正两个问题：丢失的原始符号(称为擦除)和损坏的原始符号(称为错误)。实现纠正错误和擦除，需要增加不同数量的冗余和计算工作量。

对于DNA数据存储，其他学者也提出了相关的错误纠正和缓解机制。Goldman等人[32]所使用一种重叠代码，其本身并不是一种错误纠正码，但它可以使一段数据在4种不同的DNA序列中以不同的偏移量重复。例如，一段数据可能出现在第1个DNA序列的前1/4部分，另一个序列的第2个1/4部分，另一个序列的第3个1/4部分或者最后的1/4部分。该策略，主要的目标是在不同的序列中以不同的方式定位那条信息，以避免由于合成和排序而可能发生的与链上的相对位置相关的系统错误。结果是，序列根据它们的重叠重新组装。

Bornholt等人[36]提出了一个优化的解决方案：不在多个位置重复相同的信息，而是通过异或运算将多个信息汇总为一个或多个附加信息。这种机制减少了所需的额外开销，但仍然不如基于Reed solomon的代码有效。Grass等人[17]提出基于Reed Solomon代码，使用内部代码和外部代码。随后，Erlich等人[38]提出喷泉码，用于DNA数据存储。基于喷泉码编码，k段原始数据被编译成生成k'段无限数量的编码符号。随后基于此，可以从任意k段或少于k段编码数据中恢复出原始数据。虽然喷泉码是处理擦除的最佳方法，但它们需要额外的措施来检查和纠正其他错误。由于这些代码是针对擦除校正的，因此它们是在低错误率情况下，最合适的DNA数据存储编码方法。

最近，有两项研究提出使用简并碱基[43,44]。基本原理是给预设的混合碱基设定另外的符号。例如，除了基本符号A, T, C, G外，该方法还可以使用50% As和50% Ts作为附加符号。这意味着，在序列的某个位置，可以找到A或T的概率都是50%。

这些方法增加了逻辑冗余，因为组合符号为每个位置提供了不止4个选择。这些方法用逻辑密度的改进换取额外的物理冗余。这也是解析符号所必需的。因为如果没有额外的物理副本，就不可能确定给定位置上各种碱基的比例。在一个50: 50 A和T混合作为一个附加符号的例子中，解析这个符号的最小理论物理冗余是2，使这个混合位点能够由一个副本中的A和另一个副本中的T表示；当两个碱基各占50%时，每个位置的最大理论逻辑密度为$\log_2(10)$或更少。

3.2.2 数据检索：随机访问

扩大DNA数据存储需要一种选择性读取数据片段的方法，这在计算机科学领域被称为随机存取。由于性能和成本的原因，对池中的所有DNA进行排序来检索所需的数据项是不切实际的。幸运的是，选择性提取DNA片段在分子生物学工作中很常见，可以用于随机存取。目前比较流行的两种方法是PCR扩增和磁珠提取。

利用PCR进行DNA数据的存储工作如下：在写的过程中，系统为不同的数据片段分配唯一的引物对，并在合成DNA序列时加入这些引物信息。当用户请求一段数据时，系统会找到相应的引物，扩增包含所需数据的DNA序列，并对结果池的样本进行排序。如表1所示，两项独立研究提出并演示了这样一个基于PCR的系统：将数字数据映射到DNA序列时，将数据标识符映射到引物。该方法的一个关键挑战是设计不与有效载荷冲突的引物，并在多个数据项同时被请求时启用多重PCR。

Baum[48]提出了利用磁珠提取构建联想搜索记忆的理论体系。其思想是用与分子探针查询杂交的标识符标记数据项。Stewart等人最近通过实验证明了一种直接在DNA信息中进行图像相似性搜索的方法。该方法通过对DNA池中图像集合的特征向量进行编码，然后搜索与查询图像相似的图像(例如，查找与双筒望远镜输入图像相似的所有图像)。这是一个例子，说明在DNA中，除了简单的直接获取数据外，还可以做更多的事情。

即使采用随机存取的方法，也不可能将所有数据都收集到一个DNA池中。非常的复杂混合物会有很长的扩散时间，导致提取的特异性降低。Organick等人估计，为基于PCR的随机访问提供的DNA池规模可达tb级数据，这个规模相当大，但不足以实现分子数据存储极限。为了超越这个限制，可能需要创建一个物理上隔离的池库，这些池是按需检索的。这需要以一种不牺牲太多密度的方式来实现。该方向是目前比较活跃的研究领域。

3.2.3 DNA数据存储工作的比较

在表1中，我们提供了迄今为止在DNA数据存储方面值得注意的实际编码工作的比较。对于每一个研究工作，本文列举了编码DNA的数据总量、合成和测序方法、用于测序的覆盖范围(相当于物理冗余可用于解码)、用于重组的方法、工作中使用的DNA链的长度、整体逻辑密度、有效载荷的逻辑密度以及是否可以随机访问。

DNA的自然碱基可以达到的最大理论逻辑密度是2 bit，因为在一个位置上可能的4个碱基最多可以代表2 bit。然而，大多数策略以较低的密度结束，通常在1左右。这是由于启动随机访问功能(只包括在整个逻辑密度值中)添加的引物序列所需的开销，以及编码过程中为方便错误纠正过程而添加

表 1　体外DNA数据存储比较研究

文献	数据容量	合成方法	测序方法	物理冗余(覆盖率)	重新组装	链长(碱基数)	逻辑密度(bit/碱基)	逻辑密度(有效载荷)	是否能随机访问
文献[31]	650 kB	亚磷酰胺(沉积)	合成测序	3000×	索引序列连接	115	0.60	0.83	否
文献[32]	630 kB	亚磷酰胺(沉积)	合成测序	51×	重叠序列连接	117	0.19	0.29	否
文献[17]	80 kB	亚磷酰胺(电化学)	合成测序	372×	索引序列连接	158	0.86	1.16	否
文献[37,45]	3 kB	亚磷酰胺(沉积)	纳米孔测序	200×	索引序列连接	880～1000	1.71	1.74	是
文献[38]	2 MB	亚磷酰胺(沉积)	合成测序	10.5×	种子序列连接	152	1.18	1.55	否
文献[46]	22 MB	亚磷酰胺(沉积)	合成测序	160×	索引序列连接	230	0.89	1.08	否
文献[36]	150 kB	亚磷酰胺(电化学)	合成测序	40×	索引序列连接	117	0.57	0.85	是
文献[12]	200 MB	亚磷酰胺(沉积)	合成测序	5×	索引序列连接	150～200	0.81	1.10	是
文献[43]	8.5 MB	亚磷酰胺(沉积)	合成测序	164×	索引序列连接	194	1.94	2.64	否
文献[44]	854 kB	亚磷酰胺(柱子)	合成测序	250×	索引序列连接	85	1.78	3.37	否
文献[12]	33 kB	亚磷酰胺(沉积)	纳米孔测序	36×	索引序列连接	150	0.81	1.10	是
文献[47]	18 B	酶(柱基)	纳米孔测序	175×	无(单体)	150～200	1.57	1.57	否

的冗余(包括在整个逻辑密度值和仅有效载荷逻辑密度值中)。纠错码需要这种冗余来确保没有一个DNA序列是恢复数据所必需的。可靠数据恢复所需的物理冗余级别是由测序准备协议、原始测序错误率(可以通过多次读取的平均来降低)和所提供的逻辑冗余级别(允许容忍原始错误和信息片段的丢失)决定的。较低的物理冗余导致更大序列丢失的可能性，因此需要更高的逻辑冗余来实现高概率的恢复。值得注意的是，逻辑密度并不是物理信息存储密度的直接度量。这是因为实现更高逻辑密度的方法可能需要更高的物理冗余，因此可能导致整体物理密度降低。然而，更高的逻辑冗余意味着需要合成更多独特的DNA序列，因此可能导致更高的成本和更低的读写吞吐量。

3.3 DNA合成用于数据存储

到目前为止，大多数关于DNA数据存储都是在磷酸基寡核苷酸合成方法上来写入DNA[49]。表1显示，最近的研究工作展示了存储的最大数据量，这证明了流程的成熟度。该方法通过循环添加可逆性阻断的单核苷酸以防止不必要的均聚物的形成，一次构建一个核苷酸链。去除阻塞基团可用酸溶液或光致反应(使用光致基团)来完成。每个合成循环都将选择的单体加入到现有的聚合物中，通过氧化来增强键合，用溶剂洗去多余的单体，去除上一个添加的单体中的阻塞基团，然后调用下一个合成循环或结束合成过程。最常见的合成错误源于移除最后添加的单体中的阻塞基，使插入和删除成为合成中常见的错误类型。

DNA合成可以通过控制机制来选择将哪些碱基添加到哪些链中。这使得在固体衬底的不同位置合成不同的序列成为可能，通常被称为基于阵列的合成[35]。其中有3种最常用的：光学阵列(光激活的pH值变化或对光不稳的)；电子阵列(通过选择性地解锁序列，同时添加相同的碱基至所有解锁的序列)；基于沉积的阵列(通过选择性的沉积被添加的碱基)。

实现更高的DNA合成吞吐量依赖于增加并行性。这可以通过一种或两种方法的组合来实现：增加用于生长DNA的固体基质的面积以适应数量增加的合成点，或者使这些合成点变小。为了减少光斑的尺寸，上述过程必须进一步小型化，这就产生了物理上的挑战：光波必须缩放，或者光干涉必须用来瞄准单个的更小的光斑；电子设备必须被制造出来单独操纵较小的点，液滴必须沉积在更小的区域。这些方法中的任何一种都会不可避免地导致更多错误，导致每个DNA序列的拷贝更少。这对于生物技术应用来说是有问题的，因为生物技术需要大量的DNA，而且缺陷率很低，但是对于DNA数据存储来说是可以接受的，因为DNA数据存储时，错误纠正码允许更低的物理冗余和更高的错误率。

与磷胺化学中使用的危险化学品相反，酶合成策略是基于水相，条件温和。在这种情况下，DNA聚合酶，如末端脱氧核苷酸转移酶(TdTs)，在没有模板的情况下以可控的方式结合碱基。该方法比以磷酰胺为基础的合成更快速、更清洁。酶合成的一个主要挑战是控制单碱基的添加，因为TdT酶倾向于催化每个循环添加多个碱基。此外，酶的合成可以在更短的周期内以更低的错误率产生更长的链。从最近储存的大量信息可以看出(表1)，酶合成仍然是一个新兴的领域，一种很有前途的合成长链的廉价方法。Lee等人[47]最近演示了，使用酶合成在DNA中成功地编写短消息，并使用纳米孔测序器进行读取。该方法探索了在相同碱基的不同运行之间的转换中编码信息，巧妙地避开了TdT单碱基添加和更高的纳米孔测序错误的挑战。

在表1中，DNA数据存储的一个重要的比较点是链长：虽然有的研究工作使用了较长的链长度[43]，但大多数保持在大约150到230个核苷酸链。原因是较长的链长度是极具挑战性的：酸基去胶方法由于反复的酸洗而导致去嘌呤，光学方法由于图像漂移而导致去嘌呤。由于亚磷酰胺合成方式不容易达到更长的长度，因此通过重组过程(非常类似于基因装配)获得更长的序列，这增加了写入过程的成本和时间。对于前面描述的大多数编码方案，当将合成的DNA序列的长度从100个核苷酸增加到400个核苷酸时，成本会大幅减少，如果超过这个长度，很快就会导致收益递减(5%或更少)。

3.4 DNA测序用于数据读出

目前最广泛使用的DNA测序平台是由Illumina公司推广的，它是基于图像处理和合成测序的概念[50]。将单链DNA序列连接到底物表面，通过PCR扩增成小簇的物理拷贝，然后将带有荧光标记的互补碱基逐个连接到单个序列上。将荧光标记产生的空间荧光图案捕获到图像中，在进行处理后，荧光斑点的颜色报告序列中的各个碱基。接着，用化学方法去除荧光标记，留下互补的碱基，并在待识别的序列中建立下一个碱基。将这种技术扩展到更高的吞吐量将取决于更精确的光学设置和图像处理方面的改进[51]。

另一种获得发展势头的DNA测序方法是纳米孔技术[52]。纳米孔技术的基石是捕获DNA分子，然后通过一个电压钳位的纳米级孔隙。该过程这将

导致孔隙电流的微小波动，而这些波动取决于通过的DNA链序列。在DNA数据存储方面，纳米孔测序相对于其他竞争方法的主要优势是数据的实时读出。也就是说，序列数据基本上可以实时地从设备中流出，这可能一种新型的数据访问应用程序成为可能。使用纳米孔设备进行DNA存储的主要挑战是降低较高的错误率，此外还可以合成或重组相对较长的DNA链，利用纳米孔平台的延长读取长度来增加测序的通量[52]。越来越多的技术、设备采用纳米孔技术来读取DNA数据，特别是来自牛津纳米孔技术的便携式设备MinION。但是，MinION目前的读取吞吐量比Illumina机器低。另外，现在有一个台式大小的纳米孔测序机(PromethION)，它可能更适合大规模的数据检索应用。尽管产生的错误率比Illumina测序仪高，MinION可以在更高的覆盖率测序下，准确地恢复数据(更多的读取相同的序列)和推断一个共识序列。

实现更高的数据读取率将来自更高的并行度和更快的排序周期。这意味着更快的化学反应，更密集的感应区域和更大的流动池。这可能是一个的光学读数排序问题，因为每个区域需要有足够的分离空间，以避免荧光信号的重叠。数据存储的高排序吞吐量需要非常大，这可能使这种排序方法不切实际。相比之下，基于纳米孔的测序可能会提供更密集的读出传感器，因为孔径基本上可以与DNA链的数量级相同。因此，纳米孔测序似乎提供了一个更好的可扩展性路径，同时，通过适当的错误纠正来降低错误率。

表1显示，到目前为止，大多数演示都使用了Illumina公司的测序仪。基于该方法，使用测序覆盖度作为代价，即测序仪观察到的给定测序的物理分子数(从5×到3000×不等)，应对解码过程中排序错误的问题。

4　DNA数据存储面临的主要挑战及发展趋势

从短期和中期来看，访问延迟将继续保持较高的位值(分钟到小时级别)。但是，只要带宽(吞吐量的数据读写)高，体外DNA数据存储可以与商业数据存储应用共存，或者将其取代。这是因为归档存储可以容忍更高的延迟，并且可以从更小的占用空间和更低的静态数据能耗中获益。

当前DNA数据存储的总体写入吞吐量可能是每秒千字节的级别。本文估计，在10年内，与主流云存档存储系统竞争的系统将需要提供每秒十亿字节的读写吞吐量。这是一个6个数量级的合成缺口和大约2～3个数量级的测序缺口。在成本缺口方面，2016年磁带存储的成本约为$ 16/Tb[53]，并以每年约10%的速度下降。每阵列的DNA合成成本大约为0.0001美元，相当于$ 8/Tb，比磁带高7～8个数量级。虽然吞吐量和成本差距似乎令人生畏，正如本文所述，DNA数据存储的要求不同于生命科学：为了准确度可以牺牲速度，同时可以显著降低物理冗余，降低错误校正码的使用。这使得合成法和测序法的规模和性能都有了进一步的提高。本文预计这将带来相应的成本削减，因为成本将分摊到更大的合成基质和更大的DNA批次。与此相关的是，由于数据存储所需的每个序列的拷贝数比生命科学所需的拷贝数低几个数量级，因此通过更多的并行合成和更小的光斑大小来提高吞吐量，也将导致试剂使用量的比例节省。

最后，一个重要的考虑是DNA分子的物理存储和保存。尽管已经有证据表明，DNA是可以保存几千年(有时是几十万年)，但DNA的降解速度可能比这要快得多。这取决于它所处的环境(例如，高温、高湿和紫外线可能会导致它的降解)[54]。为了解决这个问题，不同的研究小组提出了各种各样的方法来为DNA保存提供合适的条件。化学解决方案包括脱水和/或冻干、添加物(例如，生物碱或海藻糖)或用二氧化硅[17]等保护材料进行化学封装。制备化学溶液和添加剂的速度更快，而封装提供了更长的货架寿命和更好的保护，在较高的湿度(50%)的环境。储存DNA的容器有各种各样的材料和形式，比如滤纸(比如来自Whatman)、密封的不锈钢小胶囊(比如来自Imagene)和塑料井盖(比如来自Biomatrica)。这些项目是为生物样品量身定制，并优化纯度。因此，密度和成本是折中的。用于DNA数据存储的物理库需要开辟一条完全自动化和拓展性的道路，而不需要显著降低密度。这仍然是一个很大程度上开放的研究课题。要使这些系统能够用于大规模的档案存储，这些系统的自动化面临多重挑战。这样的环境通常在人为干扰最小的情况下运行。在合成和测序之外的大多数DNA操作仍在实验室环境中由人类进行。最近，文献[55,56]首次公开了全自动DNA数据存储系统的演示。微流体技术的最新进展令人鼓舞，我们希望它们能用于DNA数据存储的自动化。

数字革命改变了人类与数据的关系，引领社会进入信息时代。我们正在生成的不断扩展的数据类型和数据量超出了我们当前的存储能力。数字数据存储的新形式需要跟上时代的步伐。DNA数据存储密度已经接近极限，其作为一种具有巨大发展潜

力的存储方式，有望替代目前磁带和磁盘等主流格式。经过漫长的自然进化演进过程，基于其优异的化学稳定性和数据传递精准性，DNA被自然选择为遗传信息的载体，其也将被用作信息时代数据的载体。与此同时，最初为生命科学应用开发的DNA合成、测序和检索技术可以在数据存储系统中重新利用。随着对DNA数据 存储的研究不断取得进展，我们期待在DNA数据存储领域技术创新，将逐步弥补目前存储方式的不足。

参 考 文 献

[1] GANTZ J and REINSEL D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far East[R]. IDC iView, 2012: 1–16.

[2] EXTANCE A. How DNA could store all the world's data[J]. *Nature*, 2016, 537(7618): 22–24. doi: 10.1038/537022a.

[3] ZHIRNOV V, ZADEGAN R M, SANDHU G S, *et al*. Nucleic acid memory[J]. *Nature Materials*, 2016, 15(4): 366–370. doi: 10.1038/nmat4594.

[4] COLQUHOUN H and LUTZ J F. Information-containing macromolecules[J]. *Nature Chemistry*, 2014, 6(6): 455–456. doi: 10.1038/nchem.1958.

[5] 王君珂, 印珏, 牛人杰, 等. DNA计算与DNA纳米技术[J]. 电子与信息学报, 2020, 42(6): 1313–1325. doi: 10.11999/JEIT190826.

WANG Junke, YIN Jue, NIU Renjie, *et al*. DNA computing and DNA nanotechnology[J]. *Journal of Electronics & Information Technology*, 2020, 42(6): 1313–1325. doi: 10.11999/JEIT190826.

[6] 许进, 强小利, 张凯, 等. 基于探针图的并行型图顶点着色DNA计算模型(英文)[J]. 工程, 2018, 4(1): 61–77. doi: 10.1016/j.eng.2018.02.011.

XU Jin, QIANG Xiaoli, ZHANG Kai, *et al*. A DNA computing model for the graph vertex coloring problem based on a probe graph[J]. *Engineering*, 2018, 4(1): 61–77. doi: 10.1016/j.eng.2018.02.011.

[7] 蓝雯飞, 邢志宝, 黄俊, 等. DNA自组装计算模型求解二部图完美匹配问题[J]. 计算机研究与发展, 2016, 53(11): 2583–2593. doi: 10.7544/issn1000-1239.2016.20150312.

LAN Wenfei, XING Zhibao, HUANG Jun, *et al*. The DNA self-assembly computing model for solving perfect matching problem of bipartite graph[J]. *Journal of Computer Research and Development*, 2016, 53(11): 2583–2593. doi: 10.7544/issn1000-1239.2016.20150312.

[8] 朱维军, 周清雷, 张钦宪. 基于DNA计算的线性时序逻辑模型检测方法[J]. 计算机学报, 2016, 39(12): 2578–2597. doi: 10.11897/SP.J.1016.2016.02578.

ZHU Weijun, ZHOU Qinglei, and ZHANG Qinxian. A LTL model checking approach based on DNA computing[J]. *Chinese Journal of Computers*, 2016, 39(12): 2578–2597. doi: 10.11897/SP.J.1016.2016.02578.

[9] 夏宏, 张实君. 基于分子计算的逻辑模型构建[J]. 科技通报, 2016, 32(5): 11–15. doi: 10.3969/j.issn.1001-7119.2016.05.003.

XIA Hong and ZHANG Shijun. Constructing the logical model based on molecular computing[J]. *Bulletin of Science and Technology*, 2016, 32(5): 11–15. doi: 10.3969/j.issn.1001-7119.2016.05.003.

[10] 周旭, 周炎涛, 欧阳艾嘉, 等. 一种最大团问题的tile自组装高效模型[J]. 计算机研究与发展, 2014, 51(6): 1253–1262. doi: 10.7544/issn1000-1239.2014.20120904.

ZHOU Xu, ZHOU Yantao, OUYANG Aijia, *et al*. An efficient tile assembly model for maximum clique problem[J]. *Journal of Computer Research and Development*, 2014, 51(6): 1253–1262. doi: 10.7544/issn1000-1239.2014.20120904.

[11] 周旭, 周炎涛, 李肯立, 等. 基于tile自组装模型的最大匹配问题算法研究[J]. 电子学报, 2015, 43(2): 262–268. doi: 10.3969/j.issn.0372-2112.2015.02.009.

ZHOU Xu, ZHOU Yantao, LI Kenli, *et al*. Efficient maximum matching problem algorithms in the tile assembly model[J]. *Acta Electronica Sinica*, 2015, 43(2): 262–268. doi: 10.3969/j.issn.0372-2112.2015.02.009.

[12] ORGANICK L, ANG S D, CHEN Y J, *et al*. Random access in large-scale DNA data storage[J]. *Nature Biotechnology*, 2018, 36(3): 242–248. doi: 10.1038/nbt.4079.

[13] RUTTEN M G T A, VAANDRAGER F W, ELEMANS J A A W, *et al*. Encoding information into polymers[J]. *Nature Reviews Chemistry*, 2018, 2(11): 365–381. doi: 10.1038/s41570-018-0051-5.

[14] DNA to the rescue for data storage[J]. *Chemical & Engineering News*, 2015, 93(35): 40-41.

[15] 陈为刚, 黄刚, 李炳志, 等. 音视频文件的DNA信息存储[J]. 中国科学: 生命科学, 2020, 50(1): 81–85. doi: 10.1360/SSV-2019-0211.

CHEN Weigang, HUANG Gang, LI Bingzhi, *et al*. DNA information storage for audio and video files[J]. *Scientia Sinica Vitae*, 2020, 50(1): 81–85. doi: 10.1360/SSV-2019-0211.

[16] GREENGARD S. Cracking the code on DNA storage[J]. *Communications of the ACM*, 2017, 60(7): 16–18. doi: 10.1145/3088493.

[17] 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 Edition*, 2015, 54(8): 2552–2555. doi: 10.1002/anie.201411378.

[18] LUNT B M. How long is long-term data storage?[C].

Archiving Conference, Society for Imaging Science and Technology, 2011: 29–33.

[19] SHRIVASTAVA S and BADLANI R. Data storage in DNA[J]. *International Journal of Electrical Energy*, 2014, 2(2): 119–124.

[20] GREENBERG A, HAMILTON J, MALTZ D A, *et al*. The cost of a cloud: Research problems in data center networks[J]. *ACM SIGCOMM Computer Communication Review*, 2008, 39(1): 68–73. doi: 10.1145/1496091.1496103.

[21] SHETH R U and WANG H H. DNA-based memory devices for recording cellular events[J]. *Nature Reviews Genetics*, 2018, 19(11): 718–732. doi: 10.1038/s41576-018-0052-8.

[22] WIENER N. Interview: Machines smarter than men[J]. *US News World Report*, 1964, 56: 84–86.

[23] NEIMAN M S. On the molecular memory systems and the directed mutations[J]. *Radiotekhnika*, 1965, 6: 1–8.

[24] DAVIS J. Microvenus[J]. *Art Journal*, 1996, 55(1): 70–74. doi: 10.1080/00043249.1996.10791743.

[25] CLELLAND C T, RISCA V, and BANCROFT C. Hiding messages in DNA microdots[J]. *Nature*, 1999, 399(6736): 533–534. doi: 10.1038/21092.

[26] BANCROFT C, BOWLER T, BLOOM B, *et al*. Long-term storage of information in DNA[J]. *Science*, 2001, 293(5536): 1763–1765.

[27] AILENBERG M and ROTSTEIN O D. An improved huffman coding method for archiving text, images, and music characters in DNA[J]. *BioTechniques*, 2009, 47(3): 747–754. doi: 10.2144/000113218.

[28] WONG P C, WONG K K, and FOOTE H. Organic data memory using the DNA approach[J]. *Communications of the ACM*, 2003, 46(1): 95–98. doi: 10.1145/602421.602426.

[29] ARITA M and OHASHI Y. Secret signatures inside genomic DNA[J]. *Biotechnology Progress*, 2004, 20(5): 1605–1607. doi: 10.1021/bp049917i.

[30] YACHIE N, SEKIYAMA K, SUGAHARA J, *et al*. Alignment-based approach for durable data storage into living organisms[J]. *Biotechnology Progress*, 2007, 23(2): 501–505. doi: 10.1021/bp060261y.

[31] CHURCH G M, GAO Yuan, and KOSURI S. Next-generation digital information storage in DNA[J]. *Science*, 2012, 337(6102): 1628. doi: 10.1126/science.1226355.

[32] GOLDMAN N, BERTONE P, CHEN Siyuan, *et al*. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA[J]. *Nature*, 2013, 494(7435): 77–80. doi: 10.1038/nature11875.

[33] GIBSON D G, GLASS J I, LARTIGUE C, *et al*. Creation of a bacterial cell controlled by a chemically synthesized genome[J]. *Science*, 2010, 329(5987): 52–56. doi: 10.1126/science.1190719.

[34] HECKEL R, SHOMORONY I, RAMCHANDRAN K, *et al*. Fundamental limits of DNA storage systems[C]. 2017 IEEE International Symposium on Information Theory, Aachen, Germany, 2017: 3130–3134.

[35] KOSURI S and CHURCH G M. Large-scale *de novo* DNA synthesis: Technologies and applications[J]. *Nature Methods*, 2014, 11(5): 499–507. doi: 10.1038/nmeth.2918.

[36] BORNHOLT J, LOPEZ R, CARMEAN D M, *et al*. A DNA-based archival storage system[J]. *ACM SIGPLAN Notices*, 2016, 50(4): 637–649.

[37] YAZDI S M H T, YUAN Yongbo, MA Jian, *et al*. A rewritable, random-access DNA-based storage system[J]. *Scientific Reports*, 2015, 5: 14138. doi: 10.1038/srep14138.

[38] ERLICH Y and ZIELINSKI D. DNA fountain enables a robust and efficient storage architecture[J]. *Science*, 2017, 355(6328): 950–954. doi: 10.1126/science.aaj2038.

[39] 谭丽, 孙季丰, 郭礼华. 基于memetic算法的DNA序列数据压缩方法[J]. 电子与信息学报, 2014, 36(1): 121–127.
TAN Li, SUN Jifeng, and GUO Lihua. DNA sequence data compression method based on memetic algorithm[J]. *Journal of Electronics & Information Technology*, 2014, 36(1): 121–127.

[40] SHANNON C E. A mathematical theory of communication[J]. *The Bell System Technical Journal*, 1948, 27(3): 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[41] HECKEL R, MIKUTIS G, and GRASS R N. A characterization of the DNA data storage channel[J]. *Scientific Reports*, 2019, 9(1): 9663. doi: 10.1038/s41598-019-45832-6.

[42] REED I S and SOLOMON G. Polynomial codes over certain finite fields[J]. *Journal of the Society for Industrial and Applied Mathematics*, 1960, 8(2): 300–304. doi: 10.1137/0108018.

[43] ANAVY L, VAKNIN I, ATAR O, *et al*. Improved DNA based storage capacity and fidelity using composite DNA letters[J]. *bioRxiv*, 2018. doi: 10.1101/433524.

[44] CHOI Y, RYU T, LEE A C, *et al*. Addition of degenerate bases to DNA-based data storage for increased information capacity[J]. *bioRxiv*, 2018. doi: 10.1101/367052.

[45] YAZDI S M H T, GABRYS R, and MILENKOVIC O. Portable and error-free DNA-based data storage[J]. *Scientific Reports*, 2017, 7: 5011. doi: 10.1038/s41598-017-05188-1.

[46] BLAWAT M, GAEDKE K, HÜTTER I, *et al*. Forward error correction for DNA data storage[J]. *Procedia Computer Science*, 2016, 80: 1011–1022. doi: 10.1016/j.procs.2016.05.398.

[47] LEE H H, KALHOR R, GOELA N, *et al*. Enzymatic DNA synthesis for digital information storage[J]. *bioRxiv*, 2018.

doi: 10.1101/348987.

[48] BAUM E. Building an associative memory vastly larger than the brain[J]. *Science*, 1995, 268(5210): 583–585. doi: 10.1126/science.7725109.

[49] CARUTHERS M H. The chemical synthesis of DNA/RNA: Our gift to science[J]. *Journal of Biological Chemistry*, 2013, 288(2): 1420–1427. doi: 10.1074/jbc.X112.442855.

[50] GOODWIN S, MCPHERSON J D, and MCCOMBIE W R. Coming of age: Ten years of next-generation sequencing technologies[J]. *Nature Reviews Genetics*, 2016, 17(6): 333–351. doi: 10.1038/nrg.2016.49.

[51] SHENDURE J, BALASUBRAMANIAN S, CHURCH G M, *et al*. DNA sequencing at 40: Past, present and future[J]. *Nature*, 2017, 550(7676): 345–353. doi: 10.1038/nature24286.

[52] DEAMER D, AKESON M, and BRANTON D. Three decades of nanopore sequencing[J]. *Nature Biotechnology*, 2016, 34(5): 518–524. doi: 10.1038/nbt.3423.

[53] FONTANA JR R E and DECAD G M. Moore's law realities for recording systems and memory storage components: HDD, tape, NAND, and optical[J]. *AIP Advances*, 2018, 8(5): 056506. doi: 10.1063/1.5007621.

[54] BONNET J, COLOTTE M, COUDY D, *et al*. Chain and conformation stability of solid-state DNA: Implications for room temperature storage[J]. *Nucleic Acids Research*, 2010, 38(5): 1531–1546. doi: 10.1093/nar/gkp1060.

[55] PRAKADAN S M, SHALEK A K, and WEITZ D A. Scaling by shrinking: Empowering single-cell 'omics' with microfluidic devices[J]. *Nature Reviews Genetics*, 2017, 18(6): 345–361. doi: 10.1038/nrg.2017.15.

[56] NEWMAN S, STEPHENSON A P, WILLSEY M, *et al*. High density DNA data storage library via dehydration with digital microfluidic retrieval[J]. *Nature Communications*, 2019, 10(1): 1706. doi: 10.1038/s41467-019-09517-y.

毛秀海：男，1986年生，副研究员，研究方向为DNA纳米技术.

李 凡：男，1983年生，副研究员，研究方向为分子医学及DNA纳米技术.

左小磊：男，1980年生，研究员，研究方向为DAN电化学传感器、3D DNA探针和癌症早期诊断.