音视频文件的DNA信息存储

陈为刚^{1,3*}, 黄刚¹, 李炳志^{2,3}, 尹焯⁴, 元英进^{2,3}

1. 天津大学微电子学院, 天津 300072;

2. 天津大学化工学院, 天津 300072;

3. 教育部合成生物学前沿科学中心, 天津大学, 天津 300072;

4. 深圳华大基因股份有限公司, 深圳 518000

* 联系人, E-mail: chenwg@tju.edu.cn

收稿日期: 2019-09-12; 接受日期: 2019-10-09; 网络版发表日期: 2019-11-11

国家自然科学基金(批准号: 21621004, 21750001)和天津大学自主创新基金资助

摘要 随着全球数据量爆炸式增长, 传统的数据存储介质面临极大挑战. DNA凭借其存储密度高、可长期保存等优点, 成为很有潜力的新型数据存储介质. 数字影音档案数据量大, 要求保存时间久, 适合使用DNA作为存储介质. 本文以存储历史影音档案文件为例, 编码合成了272340条、长度为169碱基的寡核苷酸, 存储数据量超过3.0 MB, 包括两段音视频片段与一份中文文本; 设计了融合寡核苷酸合成池、扩增引物、数据块、不同文件等要素的多层次读取方法. 设计的编码处理与读取方法实现了面向DNA介质的高密度存储.

关键词 数据存储, DNA信息存储, 合成寡核苷酸池

随着物联网、科学研究以及社交网络的发展, 全球数据爆发式增长, 如何长期存储这些数据已成为人类面临的重要问题^[1,2]. 传统存储媒介例如传统机械硬盘、光盘、固态硬盘(solid state disk, SSD)、磁带等, 存储密度无法与数据增长相适应, 存储环境条件要求也较高, 介质使用时间一般不超过10年, 往往需要复杂的数据更新机制. 因此, 保障数据长期保存的连续性与可靠性的复杂度与成本均很高. 近年来, 随着合成生物学的发展, 尤其是基因组合成与组装技术的发展^[3~5], 研究者提出了利用DNA序列的不同碱基存储数字信息的新方法, 即DNA信息存储. DNA信息存储密度高, 介质贮存时间长, 特别适合大规模海量数据的长期存储, 已成为一种非常有潜力的新型数据存储方式^[6~10].

本文选择音视频档案文件, 利用合成的DNA介质, 进行了音视频文件的编码、合成、存储、扩增、读取等系统研究, 开展了全流程实验验证. 针对两段有历史价值的音视频文件以及一份中文文本, 设计合成了272340条寡核苷酸, 实现数据存储; 借助聚合酶链式反应(polymerase chain reaction, PCR)选择性扩增、逻辑块识别, 支持逻辑块粒度的单独读取, 实现了27万寡核苷酸池中的单逻辑块读取, 支持小文本文件以及跨引物池的大视频文件的读出; 并与文献^[8~10]的数据存储密度与读取复杂度进行了比较.

1 DNA信息存储流程与读取方法

为验证音视频数据的DNA信息存储方法, 以两段

引用格式: 陈为刚, 黄刚, 李炳志, 等. 音视频文件的DNA信息存储. 中国科学: 生命科学, 2020, 50: 81–85

Chen W G, Huang G, Li B Z, et al. DNA information storage for audio and video files (in Chinese). Sci Sin Vitae, 2020, 50: 81–85, doi: 10.1360/SSV-2019-0211

具有重要历史意义的音视频以及一份中文文本为例, 验证利用DNA开展信息存储的关键技术环节。

实现方法的基本思路是: 利用标准化的逻辑块与高通量合成寡核苷酸池(oligo pool, OP)实现音视频数据的存储与随机读取。采用不同合成池与扩增引物区分不同数据, 支持文件、逻辑数据块的多层次读取; 采用分组纠错编码技术, 实现逻辑块的可靠恢复; 采用文件复用技术, 实现数据与逻辑块的大小匹配, 提高逻辑数据块利用率。

设计的DNA存储系统的整体流程如图1所示。其中, 存储端的处理如图1A所示, 具体包括:

步骤1: 文件复用与编码。两段历史影像包括视频和音频。其中, 视频采用H.264/AVC(advanced video codec)编码, 音频采用AAC(advanced audio codec)编码, 两段音视频的数据量分别为1.62和1.39 MB。另外, 为充分利用存储资源, 第二段视频文件复用了中文文本文件, 大小为26.4 KB。存储文件的详细信息见网络版补充材料S1。

步骤2: 纠错编码与映射。首先, 将音视频数据和文本转化为二进制比特流, 分成35568字节大小的数据分组, 得到89个分组。然后, 每一数据分组采用里德-所罗门(Reed-Solomon, RS)码与低密度奇偶校验(low-density parity-check, LDPC)码级联编码得到大小为73440字节的编码块; 每一编码块以24字节为单元添加块内编号, 转化为30字节, 对应120 nt序列。然后, 120 nt序列添加9 nt序列用于区分不同数据块, 得到129 nt长度序列。利用上述流程, 35568字节的原始数据块转化为3060条长度为129 nt的序列; 89个分组共计得到272340条序列。

步骤3: 添加扩增引物。129 nt长度序列两端的扩增引物长度分别为20 nt, 添加测序引物后, 总长度为169 nt。本文采用4组不同引物。

步骤4: 序列合成。采用高通量的芯片方法合成寡核苷酸序列实现数据存储, 需要合成的寡核苷酸单链数量为272340条, 每条寡核苷酸长度为169 nt。

步骤5: 合成池存放。合成的寡核苷酸以冻干粉形式低温保存。

多层次数据读取的处理步骤如图1B所示, 主要包含以下步骤。

步骤1: 读取索引建立与参数选择, 包括寡核苷酸合成池、扩增引物(PCR primer, PP)、数据块(data

block, DB)的选择。用户发起文件读取请求后, 确定需要读取的合成池OP以及每个合成池使用的引物PP。例如, 用户发起读取视频文件2的需求, 系统确定需要使用合成池2与合成池3, $OP=\{2, 3\}$ 。合成池2采用引物3选择性扩增, 也即 $PP=3$, 合成池3采用引物4, 也即 $PP=4$ 。视频文件2的数据块包括数据块1~40, 也即 $DB=\{1, 2, \dots, 40\}$ 。又例如, 要读取文本文件, $PP=4$, $DB=\{41\}$ 。

步骤2: PCR选择性扩增与测序。根据读取请求, 选取合适的合成池与扩增引物, 取适量的样品进行PCR扩增建库。具体建库流程、PCR扩增体系与处理流程见网络版补充材料S2。

步骤3: 二代高通量测序。扩增之后的样品进行二代高通量测序, 采用Illumina公司测序设备, 双端读取150碱基(paired-end 150)的方式。得到的高通量测序数据说明见网络版补充材料S3。

步骤4: 测序数据处理、译码、恢复文件。具体流程如图1B所示。实现由二代高通量测序数据恢复为原始视频文件并解码播放。先依次进行引物检测、识别每个读段的数据块地址、按照数据块实现纠错, 然后恢复原始数据, 实现音视频的解码与播放或者文本输出。

2 实验方法

根据上述的DNA信息存储的方法, 本文将3份文件映射为272340条、长度为169 nt的寡核苷酸序列。由于每个92 K合成池最多约可以合成92000条序列, 为合成方便, 将所有序列分成3个引物池进行合成。从Twist Bioscience订购了3个92 K的合成寡核苷酸池。3个寡核苷酸池实际包含的不同寡核苷酸数量分别为91000条、91000条、90340条, 合成寡核苷酸的总质量分别为962, 1066, 1140 ng。数据读取验证实验中, 将合成样品溶解分别得到高浓度样品, 其浓度分别为83.5, 83.5, 82.9 $\mu\text{g}/\mu\text{L}$; 进一步得到低浓度样品, 其浓度为8.35, 8.35, 8.29 $\mu\text{g}/\mu\text{L}$ 。具体样品处理方法见网络版补充材料S2。随后, 开展了读取验证实验。从3个合成寡核苷酸池分别取等量样品合并得到一个混合合成池, 分别取3 μL 高浓度与低浓度混合液开展PCR选择性扩增实验, 然后进行二代高通量测序, 得到8个样品的测序数据文件(标记为NGS-7到NGS-14), 测序方法与测序

图 1 基于DNA的音视频文件存储全流程. A: 音视频文件、文本数据转化为二进制码流并复用, 进行分组纠错编码与映射, 添加扩增引物, 生成272340条长度为169 nt的单链序列; 分成3个合成池进行合成; B: 读取端, 根据读取请求, 采用合适引物对引物池进行选择扩增并高通量测序, 然后利用测序数据恢复音视频文件或文本文件

Figure 1 The full procedure using DNA as storage media for audio and video files. A: Audio and video files plus a Chinese text file were first converted into bit stream, multiplexed, block coded using error-correction codes, mapped to uniform sequences, added amplification primers and then mapped to 272340 oligos with a length of 169 nt; then, oligos are synthesized using three independent synthesis pools; B: during the reading, according to the access request, proper primers were selected, oligos in the pool were selectively amplified and sequenced using next-generation sequencing (NGS) technology, and finally efficient algorithms were used to recover the original audio and video files or the text file

数据详细说明见网络版补充材料S3.

3 结果与讨论

实验中, 对两个浓度的样品, 分别采用4对引物扩增, 获得8个样品. 针对每一样品的测序数据, 利用不同部分的数据开展了多次独立文件恢复验证, 文件恢复详细情况见表1.

从表1可看出, 在采用高浓度样品的读取实验中,

平均寡核苷酸分子拷贝数为10000, 视频文件1需5×深度测序数据即可无差错恢复逻辑块1~24; 需20×深度无差错恢复逻辑块25~48, 组合后获得完整的视频文件1. 恢复视频文件2与中文文本文件, 采用扩增引物3和引物4, 测序深度需要20×. 由于不同扩增引物的扩增效率、扩增实验的实验操作存在差别, 不同文件需要的测序数据深度存在差别, 这是后续可深入量化研究的内容.

在低浓度样品的读取实验中, 平均寡核苷酸分子

表 1 不同文件的数据恢复情况

Table 1 Data recovery statistics for different files

文件	存储媒介	引物对	逻辑块	高浓度样品(平均10000分子拷贝)			低浓度样品(平均1000分子拷贝)		
				测序文件	测序深度	文件读取 平行验证	测序文件	测序深度	文件读取 平行验证
视频1	合成池, 272340条	1	1~24	NGS-11	5×	168/168 ^{a)} , 成功	NGS-7	20×	28/28, 成功
					10×	84/84, 成功		30×	19/19, 成功
					30×	27/27, 成功		40×	14/14, 成功
		2	25~48	NGS-12	20×	36/36, 成功	NGS-8	50×	前5个逻辑块成 功, 其余19块失败
30×					24/24, 成功	全部, 544×			
视频2		3	1~20	NGS-13	20×	34/34, 成功	NGS-9	全部, 726×	失败
					30×	23/23, 成功			
		4	21~40	NGS-15	20×	46/46, 成功	NGS-10	全部, 887×	失败
	30×				31/31, 成功				
中文 文本	41				20×	46/46, 成功			
		30×	31/31, 成功						

a) M/N表示利用测序数据开展了N次平行实验, 成功M次

拷贝数为1000, 视频文件1, 引物1对应的24个逻辑块、引物2对应的前5个逻辑块(编号25~29), 可成功无差错恢复。根据样品浓度计算, 前24个数据块实现了平均每条寡核苷酸分子拷贝数约为1000的可靠恢复, 低于文献[8]报道的平均1300分子拷贝。在该分子拷贝数下, 数据存储密度可达到125 PB/g。同时, 文献[8]中实现了72000条序列组成的单合成池读取; 本文是从27万分子合成池中选择性成功读取88740条序列, 考虑选择性读取会造成分子分布不均衡, 本文方法实现难度更大。综合考虑分子拷贝数与读取方式两个因素, 本文方法与文献[8]比较具有一定优势。文献[9]重点展示了DNA信息存储的随机读取功能, 未给出样品用量, 无法得到其存储密度; 其方案采用码长非常长(最大到65536符号)的RS码, 长RS码的译码处理复杂度非常高, 实用性存在问题; 因此, 在译码复杂度方面, 本文方法与文献[9]相比具有优势, 可利用单台计算机实现整个读取流程验证。文献[10]实现了约两万字的中文文本的存储, 存储数据量为43.1 KB, 从存储数据量、存储类型以及存储密度方面, 本文均具有较大优势。本实验中的其他几份样品, 由于分子扩增的不均衡, 寡核苷酸分子丢失率

很高, 丢失比例高达70%以上, 无法恢复数据, 因此, 高效均衡的大合成池的扩增方法将是未来研究的重点。

综合以上, 本文实现了音视频文件DNA信息存储的全部流程, 还实现了不同层次的随机读取方法, 可支持合成池、扩增引物、逻辑块的单独读取, 为设计标准化DNA信息存储系统提供有价值参考。

4 结论

DNA存储的密度高, 存储介质可用时间久, 有望解决快速增长的数据存储需求。本文以音视频与文本数据为例, 开展了DNA信息存储实验。利用高通量合成方法合成了272340条、长度为169 nt的寡核苷酸, 存储两段具有历史价值的音视频文件与一份中文文本, 存储数据量大于3.0 MB。进一步, 依据合成池存储的特点, 验证了不同层次读取方法, 具有很好的可扩展性。在未来, 将深入研究寡核苷酸高通量合成方法、存储读取的高通量测序方法、测序引物设计与扩增、高效高可靠信息编码等过程的参数优化与系统设计, 提高DNA数据存储的有效性与可靠性。

参考文献

1 Xiao Q H. Long-term Preservation of Electronic Records: Theory and Practice (in Chinese). Beijing: Social Sciences Academic Press, 2014 [肖

秋会. 电子文件长期保存: 理论与实践. 北京: 社会科学文献出版社, 2014]

- 2 CCSDS 650. 0-M-2, Reference Model for an Open Archival Information System (OAIS). Washington: Magenta Book, 2012, 1
- 3 Kosuri S, Church G M. Large-scale *de novo* DNA synthesis: Technologies and applications. *Nat Methods*, 2014, 11: 499–507
- 4 Xie Z X, Li B Z, Mitchell L A, et al. “Perfect” designer chromosome V and behavior of a ring derivative. *Science*, 2017, 355: eaaf4704
- 5 Wu Y, Li B Z, Zhao M, et al. Bug mapping and fitness testing of chemically synthesized chromosome X. *Science*, 2017, 355: eaaf4706
- 6 Church G M, Gao Y, Kosuri S. Next-generation digital information storage in DNA. *Science*, 2012, 337: 1628
- 7 Goldman N, Bertone P, Chen S, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 2013, 494: 77–80
- 8 Erlich Y, Zielinski D. DNA fountain enables a robust and efficient storage architecture. *Science*, 2017, 355: 950–954
- 9 Organick L, Ang S D, Chen Y J, et al. Random access in large-scale DNA data storage. *Nat Biotechnol*, 2018, 36: 242–248
- 10 Li Y M, Zhong Y P, Qi S S, et al. A new digital information storing and reading system based on synthetic DNA (in Chinese). *Sci Sin Vitae*, 2018, 48: 102–104 [李彦敏, 钟云鹏, 祁姗姗, 等. 基于合成DNA的数字信息储存和读取系统. 中国科学: 生命科学, 2018, 48: 102–104]

DNA information storage for audio and video files

CHEN WeiGang^{1,3}, HUANG Gang¹, LI BingZhi^{2,3}, YIN Ye⁴ & YUAN YingJin^{2,3}

¹ School of Microelectronics, Tianjin University, Tianjin 300072, China;

² School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China;

³ Frontiers Science Center for Synthetic Biology (MOE), Tianjin University, Tianjin 300072, China;

⁴ BGI Genomics, Shenzhen 518000, China

With the explosive growth of the global data volume, traditional storage media have faced great challenge in satisfying the increasing storage requirement. Deoxyribonucleic acid (DNA) has become a new storage medium with great potential due to its ultra-high storage density and ultra-long storage durability. DNA may serve as the most suitable medium for digital archives of audio and video (AV) files because of its large storage volume and long durability. In this paper, we stored historical AV files to show the storage potential of DNA media. We encoded two AV clips and a Chinese text with the overall volume larger than 3.0 MB into pooled 272340 oligos with a length of 169 nt. Simultaneously, we presented a hierarchical flexible access scheme integrating different synthesis pool, primers, data block and files. The proposed encoding and access method verifies the ultra-high density storage using synthetic pools.

data storage, DNA information storage, synthetic oligonucleotide (oligo) pool

doi: [10.1360/SSV-2019-0211](https://doi.org/10.1360/SSV-2019-0211)