DNA存储及其研究进展

许 鹏[①②] 方 刚[①] 石晓龙[①] 刘文斌*[①]

[①](广州大学计算科技研究院 广州 510006)

[②](黔南民族师范学院计算机与信息学院 都匀 558000)

摘 要：DNA存储是一种以生物大分子DNA作为信息载体的一种新的存储技术。与传统的电子信息存储相比，DNA存储具有容量大、密度高、低能耗等优点。随着DNA合成、测序技术的发展以及大数据时代对数据存储需求的指数增长，近年来DNA存储在存储容量、密度以及可靠性等方面都取得了巨大的进展。该文主要介绍了DNA存储的发展历史、DNA存储的基本流程、DNA存储在数据库、文档存储以及体内存储的研究进展。最后，总结了DNA存储未来面临的挑战以及发展方向。

关键词：DNA存储；DNA数据库；体内存储；体外存储

中图分类号：TP301 **文献标识码**：A **文章编号**：1009-5896(2020)06-1326-06

DOI: 10.11999/JEIT190863

DNA Storage and Its Research Progress

XU Peng[①②] FANG Gang[①] SHI Xiaolong[①] LIU Wenbin[①]

[①](*Institute of Computing Science and Technology, Guangzhou University, Guangzhou* 510006, *China*)

[②](*School of Computer Science of Information Technology, Qiannan Normal University for Nationalities, Duyun* 558000, *China*)

Abstract: DNA storage is a new kind of technique to store information by the biological molecules DNA. Compared with traditional electronic storage medium, DNA storage has the advantages such as massive storage capacity, high storage density and low energy consumption. The developments in DNA synthetic and sequencing technique, and the exponential requirement for big data storage have pushed the research of DNA storage achieving great progress on storage capacity, storage density and reliability. The development history of DNA storage, its general workflow, and the development in DNA database, documental storage and in vivo storage are introduced. Finally, the challenge of DNA storage and its potential future research direction are pointed out.

Key words: DNA storage; DNA database; In vivo storage; In vitro storage

1 引言

作为遗传信息的载体，DNA本身就是一种天然的优良存储介质。DNA存储了从微生物到人类的亿万生命的遗传信息，并保证生命现象的稳定遗传。例如，1 μm大小的大肠杆菌可以利用纳米尺度的DNA与蛋白质器件执行信息处理的功能，而功耗仅为10～13 W。近年来，生命系统在信息存储与处理方面，具有并行性、高存储密度及低能耗等优点引起来了越来越多科学家的关注。

早在1959年，诺贝尔奖获得者Feynman[1]就首次提出了分子尺度计算机的概念。1994年，图灵奖获得者Adleman[2]首次以DNA分子作为“数据”的载体，以生物酶及生化操作作为“算子”解决了一个哈密尔顿最短路径问题，开启了生物计算的新纪元。普林斯顿大学的Baum[3]在Science上发表文章“Building an associative memory vastly larger than the brain”，正式提出构建基于DNA的大容量数据库存储体系。DNA存储最早可以追溯到1964年，Wiener和Neiman[4,5]提出了分子“遗传记忆(genetic memory)”。1996年，Davis[6]首次将一段编码35 bit的黑白图像的DNA序列(5’-CCCCCCAACGCGCGCG-3’)“写入”细胞载体存

收稿日期：2019-11-01；改回日期：2020-04-23；网络出版：2020-05-13

*通信作者： 刘文斌 wbliu6910@126.com

基金项目：国家重点研发计划(2019YFA0706402)，国家自然科学基金(61572367, 61573017, 61972107, 61972109)

Foundation Items: The National Key R&D Program of China (2019YFA0706402), The National Natural Science Foundation of China (61572367, 61573017, 61972107, 61972109)

储。1999年，Bancroft等人[7]成功恢复了隐藏在DNA序列中的一条23个字符的消息。2000年，Leier等人[8]实现了可以应用于DNA“条形码”的加密和解密方法。2001年，杜克大学的Reif等人[9]首次构建了一个小型可以随机访问的DNA数据库。直到2012年，美国哈佛大学Church[10]将一本50000字的图书存储在DNA中，开启了DNA存储的热潮。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。2018年，美国发布的《半导体合成生物学路线图》预测基于DNA分子的数据存储将有望解决海量数据存储，数据中心规模与能耗方面的挑战(https://www.src.org/program/grc/semisynbio/semisynbio-consortium-roadmap/)。2019年7月，《科学美国人》将DNA存储列为年度十大突破性技术之一。IT巨头微软已计划于2020年在数据中心建立基于DNA的数据存储系统。探索新一代DNA信息存储新体系，对大数据时代信息技术可持续发展具有重要的战略意义，本文主要介绍DNA存储的基本流程、信道模型的复杂性、主要进展及面临的挑战。

2　DNA存储及其优点

自然界中，由A, T, C, G, 4个核酸碱基组成的脱氧核糖核酸(DeoxyriboNucleic Acid, DNA)承载了所有生物体的遗传信息。DNA存储是利用脱氧核糖核酸(DNA)生物大分子作为介质，按照一定的编码策略将文本、图片、声音和视频等信息转化为相应的DNA序列，借助生物合成技术合成相应的DNA分子在体内或体外加以贮存，利用DNA分子的特异性杂交技术，如基于聚合酶链式反应PCR或磁珠分离技术访问数据。DNA存储的一般流程如图1所示，主要包括以下6个步骤：

(1) 编码：将0, 1二进制信息编码为由A, T, C, G组成的DNA序列；

(2) 合成：利用各种高通量技术合成编码信息的DNA序列；

(3) 存储：选择合适的载体(体内/体外)将合成的DNA序列进行存储；

(4) 检索：利用DNA碱基配对的特异性杂交，同特定的引物序列提取DNA分子；

(5) 测序：对提取到的DNA分子进行测序得到DNA序列；

(6) 解码：根据解码规则将DNA序列中的信息复原。

相较于传统存储介质，DNA在数据存储方面具有以下4个优点：

(1) 存储密度高，DNA可以达到～10^7 GB/mm^3，比传统存储介质提高了7个数量级。

(2) 保存寿命极长，DNA数据在没有特别人工干预的情况下能保存千年之久。

(3) 维护成本极低，DNA数字存储所需要的占地，资源，能源均远远小于传统存储介质。

(4) 数据备份十分便捷，PCR为DNA的快速复制扩增提供了技术保障。

这些优势有望解决大数据时代电子信息技术对海量数据的有效存储和管理面临的困境。

3　DNA存储信道模型的复杂性

DNA存储过程主要涉及合成、聚合酶链式反应(Polymerase Chain Reaction, PCR)扩增及测序等技术，可以将DNA存储过程理解为一个信道模型，该信道模型主要存在由3种技术引起的3种错误[11,12]。对于一个DNA序列，DNA合成过程将会产生几百到上千个拷贝，DNA合成过程将会发生替换(substitutions)、插入(insertions)和删除(deletions)等错误，导致同一序列的每个拷贝将会出现各种不同的3种错误。PCR技术主要用于合成后进一步扩增以及信息读取时对少量样本的扩增，PCR扩增过程也会引入替换错误。测序过程主要会引起替换错误，插入和删除的概率大概在10^{-6}左右。此外，单链DNA分子存储过程也会发生碱基退化变异错误。

丢失是DNA存储过程的另一个重要的问题。在DNA合成过程中，一个序列可能由于各种原在出现合成终止导致丢失。在PCR过程中，由于扩增引物的序列偏好可能导致某些序列出现扩增异常而丢失。测序过程，由于各个序列拷贝分布的不均匀性，可能导致某些序列没有测序导致丢失。

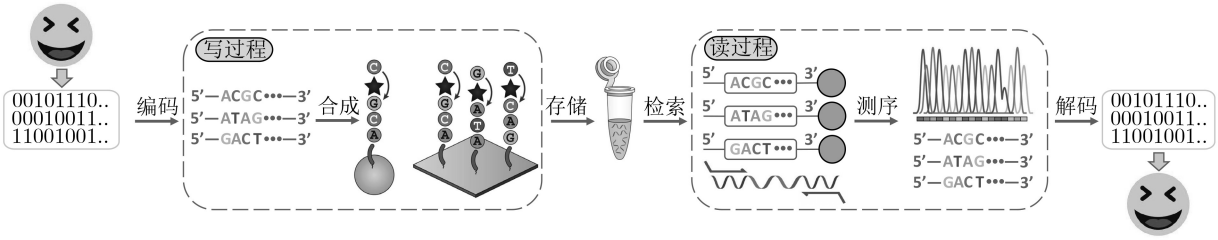


图 1 DNA存储的主要流程

DNA存储信道还存在单链DNA分子分布的不平衡性，主要包括合成过程的不均衡性、PCR扩增过程导致的不均衡性以及测序过程的不均衡性组成。这些不均衡性层层叠加影响，最终导致了测序文件中序列分布的多样性，增加了解码过程的复杂性。

上述错误、丢失及分布复杂性增加了测序文件解码的复杂性。为了解决这些信道“噪声”，常用的解决方案主要包括以下3个方面：

(1) 编码设计：基于DNA分子的特性，任何01二进制信息最终翻译的DNA序列应该保证GC含量在50%左右，GC含量过大或过小都容易引起合成、PCR及测序过程的错误。引物或地址DNA序列需要有足够大的汉明距离。DNA序列应该尽可能避免产生各种不期望的2级结构。最近，文献[13]人工合成了另外4种核苷酸，突破性地创造出具有8个字母的DNA分子。碱基数的增加将进一步增加DNA编码的灵活性、鲁棒性以及编码空间。

(2) 物理冗余：主要指通过PCR扩增增加每个DNA序列的拷贝数或者保证序列中的片段重复性，使得同一片段信息出现在不同的DNA序列中。

(3) 逻辑冗余：主要指通过各种校验码(如LDPC)或纠错码(如RS吗、喷泉码)等，解决各种错误或序列丢失问题。

4 DNA数据库

1995年，Baum[3]提出了基于PCR技术和磁珠分离技术的联想搜索和随机访问的设想。2001年，Bancroft等人[14]设计了一种基于检索链(图2左)和存贮链(图2右)的数据存储方案(如图2)。信息存储链左右两端有一对公共的前向和后向引物序列用于PCR扩增。在前向引物与信息序列之间有一个索引序列用于对信息的检索。检索链主要存储对信息链内容的索引地址。

2002年，杜克大学的Reif等人[9]提出了一种基于数据块的DNA存储结构(如图3所示)，每个数据块相当于数据库中的一个字段。当一个数据库由k个块构成，每个块有n个不同的编码信息时，从每个块中取一个特定的编码线性连接后就可以形成一个容量为n^k的数据库。他们通过实验构建了一个规模为12^7的DNA数据库，这一数据库可以真正实现Baum设想的联想搜索。

显然这种数据库的容量由每个块的编码容量和块的个数决定，大容量数据库需要大量特异性的编码序列集合。Garzon等人[15]系统地阐述了DNA计算和DNA存储中的编码问题，随后一大批学者致力于编码问题的研究。笔者的团队[16–18]在模板编码的基础上，提出了模板框的编码理论，可以应用于这种块数据库的编码设计。

Yamamoto等人[19–21]提出了一种基于嵌套PCR (Nested PCR)的存储方式，每个信息的地址由一个类似于Reif提出的块结构构成，信息的检索可以通过其对应的地址块递归做PCR扩增，他们构建了一个存储容量为16.8 M的数据库。2015年，Yazdi等人[22]提出了一种互不相关(mutually uncorrelated) DNA地址码，构建了一种可重写可随机访问的存储系统。2018年，Stewart等人[23]提出了一种基于关联搜索的图像DNA存储数据库。他们首先将图像用特定个数的元特征表示，然后通过神经网络学习每个元特征图像的编码，使得相似的元特征具有相似的DNA特征编码。

5 档案文件存储

2012年，Church研究小组[10]用A/C分别代表奇位/偶位0，T/G分别代表奇位/偶位1，将650 kB数据存入DNA中，使DNA存储数据容量比之前提高了1000倍，从而点燃了DNA存储研究之火。2013年，Goldman等人[24]首先利用Huffman编码将文本或二进制文件利用霍夫曼编码转化为由编码的三进制。然后，根据当前编码碱基，用其他3个碱基分别代表0, 1, 2转化为DNA编码。此外，为了克服合成及测序错误，他们采取了如图4(a)所示的

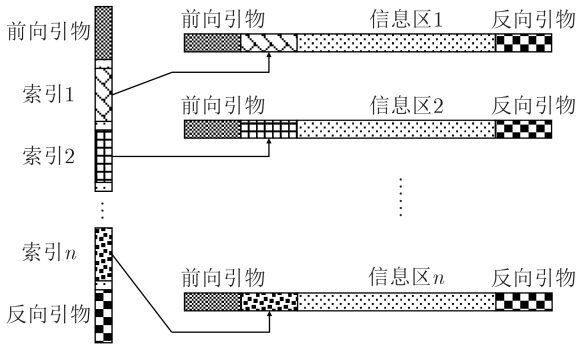


图 2 检索链和信息存贮链示意图

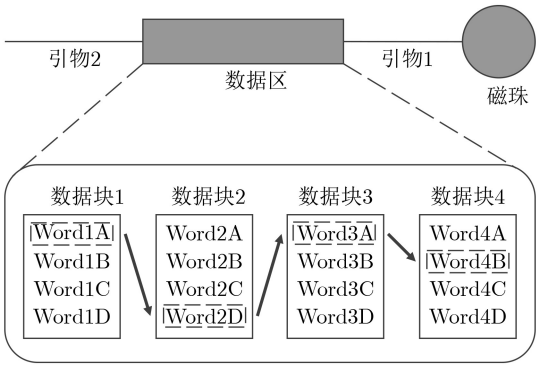


图 3 基于磁珠的DNA数据库示意图

4倍信息块冗余编码策略，实现了729 kB的存储总量。2016年，华盛顿大学和微软的研究人员[25]提出了的异或逻辑编码策略，成功实现了约151 kB数据的DNA存储。如图4(b)所示，2个信息串A, B异或产生一个新的校验串C(即A⊕B=C)，任何1个串的信息可以通过其它2个串得到。这一方法将Goldman方法中的信息冗余由的4倍降为1.5倍。2017年，纽约基因组中心和哥伦比亚大学计算机系的Erlich等人[26]提出了DNA喷泉码的编码方法，不仅提高了DNA存储的纠错能力，同时也大大提高了单个碱基的编码位容量(约1.98 bit/nt)。他们合成了72000条200 nt的DNA序列，总计存储了2.15 MB信息。2017年，Yazdi等人[27]结合纠错码和纳米孔测序仪技术，开发了一个便携式的具有随机访问的存储系统。2018年，微软和华盛顿大学的研究小组[28]采用对长序列编码的随机化以及Reed Solomon的纠错策略，用1.34×10^7条长150 bp的DNA链实现了35个文件总计200 MB的存储。2019年，我国天津大学的陈为刚[29]在研究音视频文件的存储中，通过引物池和数据分块机制实现了文件和数据块的随机查找。为了防止各种错误，他们采用采用里德-所罗门(Reed-Solomon, RS)码与低密度奇偶校验(Low-Density Parity-Check, LDPC)码级联编码的纠错方式。苏黎世联邦理工学院Meiser等人[30]提出了一种基于数据块的编码纠错机制，通过内码解决丢失问题，通过外码解决3种错误。以色列理工学院的Anavy等人[31]提出了复合DNA码降低编码的冗余性，他们综合了基于二进制序列的喷泉码纠错，以及翻译后DNA序列的RS纠错码。

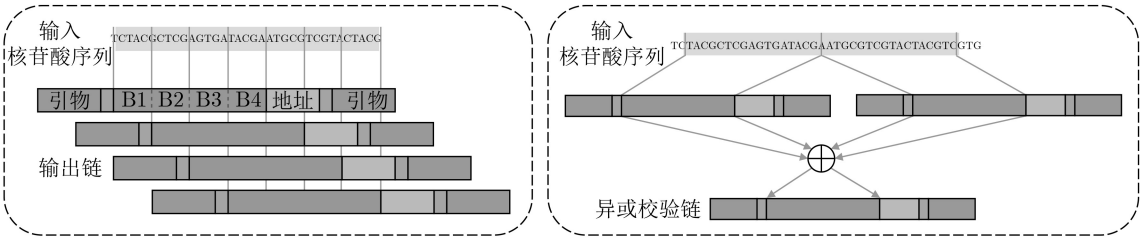
DNA序列一般由数据区、索引编码区以及两端的扩增引物区组成。DNA存储的“写”(即合成)、“读”(即测序)过程往往容易产生核酸碱基的替换(substitutions)、插入(insertions)和删除(deletions)。此外，在数据的存储及检索过程的PCR扩增过程中，也可能产生扩增不平衡导致信息丢失现象。因此，编码技术对于DNA存储系统的稳定性、可靠性及效率至关重要。已有的各种存储研究都采用序列随机化、纠错码以及信息冗余等技术来降低“读”、“写”、“存”过程的错误影响。随着存储容量的增大，构建DNA存储系统所需的码字(codeword)也随之增大。Lenz等人[32]研究了DNA编码数量的上下限问题。Anavy等人[31]提出了DNA组合码的编码方法降低编码的冗余性。最近，Benner率领的研究团队[13]人工合成了另外4种核苷酸，突破性地创造出具有8个字母的DNA分子。碱基数的增加将进一步增加DNA编码的灵活性、鲁棒性以及编码空间。

6　活体存储

与体外DNA存储相比，体内存储便于数据的随时复制，但在数据密度上存在一定劣势。此外，活体存储的缺点是生物体中的DNA存在着变异、删除和插入的风险。Yachie等人[33]提出了基于纠错码的变异检测方法。2010年，Gibson等人[34]将人工合成了一个支原体基因组(～1.8 Mbp)转入酵母细胞，这是将人工信息完整存储在细胞的一次壮举。尽管合成基因组中有4个基因被4 kbp的DNA序列替换，存储于细胞中的DNA仍然可以随着细胞的复制代代相传。2017年，文献[35]将784 Byte和494 Byte的4色和21色的图片，以及一部2.6 kB的无声短片，通过CRISPR-Cas基因编辑技术存储于细菌活体内，数据还原率达到90%左右。2019年，麻省理工学院的科学家[36]利用基因编辑技术实现了对小分子、光照等生物信号的读/写，该技术可以用来研究细胞的动态响应过程，类似于细胞行为记录仪。

7　DNA存储面临的挑战

首先，从大规模工业化应用的角度，目前的合成及测序成本还太高，特别是合成的费用大约是测序的4个数量级。从存储的角度长序列片段的存储效率更高，但是目前合成序列长度一般为100～300碱基，超过此长度的合成成本将急剧增加，同时合成和测序的错误率也会随之增加。因此，合成和测序技术的进步是DNA存储走向实际应用的技术基础。



(a) 4倍信息块冗余策略　　(b) 2个信息串异或产生新校验串

图 4 2种数据链冗余设计策略

其次，从DNA存储信道模型的角度，编码理论及方法是DNA存储的核心理论问题，高效鲁棒的编码将有望克服目前DNA存储合成、PCR扩增及测序技术的不足，而且可能降低存储的费用。目前的编码基本的流程是将加入纠错冗余信息的二进制字符串直接翻译为DNA序列，这一编码方式将是基于二进制流的纠错，如何结合DNA生化特性，直接研究基于DNA信息流上的组合DNA编码理论将有望解决DNA存储信息到的高可信信息存储问题。

由于DNA存储信息的复杂性，高效可靠的DNA存储需要有坚实的DNA编码理论支撑。目前的DNA存储的编码主要是利用电子信息通讯技术理论建立的各种纠错码的应用。DNA存储需要结合合成、PCR扩增及测序的生化特性，深入研究DNA存储信道的模型及高效的纠错码理论和方法。

最后，从信息应用的角度，目前的档案数据存储是一种“死数据”，仍然需要将信息解码为二进制信息由电子计算机做进一步处理分析。基于DNA层次上的随机访问及信息检索技术还处于初步阶段。因此，这一存储模式严重制约了DNA存储在频繁访问大数据领域的应用。

8 结束语

从1995年Baum的最初设想到200 MB存储量，DNA存储领域在系统设计、编码方法、合成测序技术等方面都取得了巨大的进展。本文主要介绍了DNA存储的基本原理、DNA存储信道模型的复杂性、DNA存储近年来在DNA数据库、档案文档及体内存储的研究进展，并且指出了DNA存储面临的挑战。我们相信，未来的十年，随着硬件技术、编码理论以及存储应用领域的扩大，DNA存储必将会在一些数据存储应用方面进入商业应用。最后，经过二十年的发展，DNA计算在计算模型、编码方法及生化算子等方面已经取得了丰富的成果。因此，未来DNA计算的高度并行计算能力与DNA存储的海量数据系统的融合，将会进一步推动DNA存储在关联搜索、信息统计及聚类分析等方面的发展。

参 考 文 献

[1] FEYNMAN R P. There's plenty of room at the bottom[J]. *Resonance*, 2011, 16(9): 890. doi: 10.1007/s12045-011-0109-x.

[2] ADLEMAN L M. Molecular computation of solutions to combinatorial problems[J]. *Science*, 1994, 266(5187): 1021–1024. doi: 10.1126/science.7973651.

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

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

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

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

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

[8] LEIER A, RICHTER C, BANZHAF W, *et al*. Cryptography with DNA binary strands[J]. *Biosystems*, 2000, 57(1): 13–22. doi: 10.1016/S0303-2647(00)00083-6.

[9] REIF J H, LABEAN T H, PIRRUNG M, *et al*. Experimental construction of very large scale DNA databases with associative search capability[C]. The 7th International Workshop on DNA-Based Computers, Tampa, USA, 2002: 231–247. doi: 10.1007/3-540-48017-X_22.

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

[11] 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. doi: 10.1109/ISIT.2017.8007106.

[12] 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.

[13] HOSHIKA S, LEAL N A, KIM M J, *et al*. Hachimoji DNA and RNA: A genetic system with eight building blocks[J]. *Science*, 2019, 363(6429): 884–887. doi: 10.1126/science.aat0971.

[14] BANCROFT C, BOWLER T, BLOOM B, *et al*. Long-term storage of information in DNA[J]. *Science*, 2001, 293(5536): 1763–1765. doi: 10.1126/science.293.5536.1763c.

[15] GARZON M H and DEATON R J. Codeword design and information encoding in DNA ensembles[J]. *Natural Computing*, 2004, 3(3): 253–292. doi: 10.1023/B:NACO.0000036818.27537.c9.

[16] 王向红, 刘文斌, 朱翔鸥, 等. DNA计算中的单模板编码方法改进研究[J]. 电子学报, 2009, 37(12): 2720–2724. doi: 10.3321/–j.issn:0372-2112.2009.12.021.
WANG Xianghong, LIU Wenbin, ZHU Xiangou, *et al*. Improving the single template method in DNA computing[J]. *Acta Electronica Sinica*, 2009, 37(12): 2720–2724. doi: 10.3321/–j.issn:0372-2112.2009.12.021.

[17] 刘文斌, 朱翔鸥, 王向红, 等. 一种优化DNA计算模板性能的新方法[J]. 电子与信息学报, 2008, 30(5): 1131–1135. doi:

10.3724/SP.J.1146.2006.01640.

LIU Wenbin, ZHU Xiangou, WANG Xianghong, *et al*. A new method to optimize the template set in DNA computing[J]. *Journal of Electronics & Information Technology*, 2008, 30(5): 1131–1135. doi: 10.3724/SP.J.1146.2006.01640.

[18] 刘文斌, 陈丽春, 白宝钢, 等. DNA计算中的模板框优化方法研究[J]. 电子学报, 2007, 35(8): 1490–1494. doi: 10.3321/j.issn:0372-2112.2007.08.014.

LIU Wenbin, CHEN Lichun, BAI Baogang, *et al*. Research on optimizing the template frame in DNA computing[J]. *Acta Electronica Sinica*, 2007, 35(8): 1490–1494. doi: 10.3321/j.issn:0372-2112.2007.08.014.

[19] 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. doi: 10.1016/j.biosystems.2004.10.007.

[20] KASHIWAMURA S, YAMAMOTO M, KAMEDA A, *et al*. Experimental challenge of scaled-up hierarchical DNA memory expressing a 10, 000-address space[C]. Preliminary Proceeding of 11th International Meeting on DNA based Computers, London, UK, 2005.

[21] YAMAMOTO M, KASHIWAMURA S, OHUCHI A, *et al*. Large-scale DNA memory based on the nested PCR[J]. *Natural Computing*, 2008, 7(3): 335–346. doi: 10.1007/s11047-008-9076-x.

[22] 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.

[23] STEWART K, CHEN Y J, WARD D, *et al*. A content-addressable DNA database with learned sequence encodings[C]. The 24th International Conference on DNA Computing and Molecular Programming, Jinan, China, 2018: 55–70. doi: 10.1007/978-3-030-00030-1_4.

[24] 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.

[25] BORNHOLT J, LOPEZ R, CARMEAN D M, *et al*. A DNA-based archival storage system[J]. *ACM SIGARCH Computer Architecture News*, 2016, 44(2): 637–649. doi: 10.1145/2980024.2872397.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] MEISER L C, ANTKOWIAK P L, KOCH J, *et al*. Reading and writing digital data in DNA[J]. *Nature Protocols*, 2020, 15(1): 86–101. doi: 10.1038/s41596-019-0244-5.

[31] Anavy, L., Vaknin, I., Atar, O. *et al*. Data storage in DNA with fewer synthesis cycles using composite DNA letters[J]. Nat Biotechnol 2019, 37, 1229–1236. doi: https://doi.org/10.1038/s41587-019-0240-x.

[32] LENZ A, SIEGEL P H, WACHTER-ZEH A, *et al*. Coding over Sets for DNA storage[C]. 2018 IEEE International Symposium on Information Theory, Vail, USA, 2018: 2411–2415. doi: 10.1109/ISIT.2018.8437544.

[33] YACHIE N, OHASHI Y, and TOMITA M. Stabilizing synthetic data in the DNA of living organisms[J]. *Systems and Synthetic Biology*, 2008, 2(1/2): 19–25. doi: 10.1007/s11693-008-9020-5.

[34] 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.

[35] 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. doi: 10.1038/nature23017.

[36] FARZADFARD F, GHARAEI N, HIGASHIKUNI Y, *et al*. Single-nucleotide-resolution computing and memory in living cells[J]. *Molecular Cell*, 2019, 75(4): 769–780.E4. doi: 10.1016/j.molcel.2019.07.011.

许　鹏：男，1986年生，博士后，研究方向为生物信息学.
方　刚：男，1969年生，教授，研究方向为生物信息学.
石晓龙：男，1975年生，教授，研究方向为生物信息学.
刘文斌：男，1969年生，教授，研究方向为生物信息学.