第 13 卷 第 3 期
2024 年 5 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 13 No. 3
May 2024

引文格式：
殷晓荷, 张舒颖, 张瑞峰, 等. 非天然碱基基因密码扩展 DNA 数据存储的机遇与挑战 [J]. 集成技术, 2024, 13(3): 39-53.
The opportunities and challenges of expanding DNA data storage with non-natural base gene code [J]. Journal of Integration Technology, 2024, 13(3): 39-53.

非天然碱基基因密码扩展 DNA 数据存储的机遇与挑战

殷晓荷[#] 张舒颖[#] 张瑞峰 尚林春 李凌君[*]

（河南师范大学化学化工学院 新乡 453007）

摘 要 目前，传统的存储技术主要将硅基材料作为存储介质，但全球现有的硅资源却无法满足日益增长的数据存储需求。随着数据时代的发展，存储技术创新面临新的挑战。在自然界，DNA 分子储存着丰富的遗传信息。从化学生物学的角度分析，将 DNA 分子作为介质进行数据信息存储有望为存储技术的创新提供一个新机遇，而非天然碱基核苷酸可以扩充遗传字母表，增加 DNA 存储容量，但在其实际应用方面，目前还有很多问题待解决。该文综述了 DNA 存储技术的研究进展，对当前 DNA 存储现状、待解决的技术难题与发展前景等进行了分析；并在此基础上，介绍了非天然碱基对（UBPs）作为合成生物学的一个新方向在 DNA 信息存储领域的潜在优势和技术挑战。

关键词 DNA 分子存储；化学生物学；非天然碱基核苷酸
中图分类号 Q 523；TP 333 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20231031001

The Opportunities and Challenges of Expanding DNA Data Storage with Non-natural Base Gene Code

YIN Xiaohe[#] ZHANG Shuying[#] ZHANG Ruifeng SHANG Linchun LI Lingjun[*]

(*School of Chemistry and Chemical Engineering of Henan Normal University*, *Xinxiang* 453007, *China*)

[*]*Corresponding Author*: *lingjunlee@htu.edu.cn*

[#]*Equal Contribution*

Abstract At present, traditional storage technologies mainly use silicon-based materials as storage media, but the existing silicon resources in the world cannot meet the growing data storage needs of data. With the development of the data era, innovation in storage technology has faced challenges. DNA molecules store rich genetic information, and from the perspective of chemical biology, DNA molecules can be used as a

收稿日期：2023-10-31 **修回日期**：2024-01-15
基金项目：国家自然科学基金项目（22077027, U23A20106）
作者简介：殷晓荷（共同第一作者），硕士研究生，研究方向为非天然碱基核苷酸的 DNA 存储；张舒颖（共同第一作者），硕士研究生，研究方向为非天然碱基核苷酸的 DNA 存储；张瑞峰，博士研究生，研究方向为非天然碱基核苷酸的合成；尚林春，博士研究生，研究方向为计算化学；李凌君（通讯作者），教授，研究方向为非天然碱基核苷酸的设计与合成，E-mail：lingjunlee@htu.edu.cn。

medium for data information storage. This provides a new opportunity for storage technology. Unnatural base nucleotides can expand the genetic alphabet and increase the storage capacity, but there are still many issues to be resolved in their practical applications. This article reviews the progress of DNA storage technology, analysing the current state of DNA storage, unresolved technical challenges, and development prospects. Furthermore, it introduces unnatural base pairs (UBPs) as a new direction in synthetic biology, highlighting their potential advantages and technical challenges in the field of DNA information storage.

Keywords DNA molecular storage; chemical biology; non-natural base nucleotides

Funding This work is supported by National Natural Science Foundation of China (22077027, U23A20106)

1 引 言

人类文明的延续离不开信息储存，从古时的结绳记事到现代的磁光电存储技术，信息存储技术已经有了质的飞跃。但社会高新发展的同时，也给存储技术提出了新的“要求”，当前存储迫切需要一种更高密度、更为稳定的存储技术来应对大数据时代。传统的计算机二进制语言只需要 0 和 1 两个符号就可以编码所有的信息，同理，利用生物体内的 4 种碱基，其实也可以编码信息。DNA 存储技术是将 DNA 分子作为信息存储的介质，通过一定的编码规则将数据转化为 DNA 序列，再将其存储于 DNA 分子中，最后从 DNA 分子中通过 DNA 测序等手段获取信息的一种新型存储技术。通常，DNA 存储技术可分为信息编码、DNA 合成、DNA 测序及解码 4 个步骤[1]。近十几年来，有关 DNA 存储技术方面的研究有了很大的进展，其存储密度、DNA 合成技术及测序技术都取得了明显的进步，但不可否认的是，目前，DNA 存储技术还存在几大技术难题。依赖于有机化学的 DNA 合成方法不仅合成效率低，而且会限制存储在 DNA 分子内数据的大小；还有更关键的合成成本问题，如果成本太高，则势必会影响 DNA 存储技术的商业化发展；另外，在 DNA 存储的过程中，其读取效率也有待提高。本文介绍了 DNA 存储技术的几个技术难题，分析了目前 DNA 存储研究的重点与难点，在此基础上，介绍了非天然碱基核苷酸拓展 DNA 存储的潜力，并展望了未来 DNA 存储技术与非天然碱基核苷酸的交叉研究新机遇。

2 DNA 存储的研究进展

根据国际数据公司预测，到 2025 年，全球的数据产出量将达到 175 ZB，而 1 ZB≈1.18×10^{21} B[2]。伴随着数据的爆炸式增长，数据存储也面临着新的挑战，因此，研究探索其他的数据存储方式显得尤为重要。DNA 存储作为一种新型的数据存储方式，不仅比传统数据存储技术的存储密度高，还有着其他先天优势，如易获取、免维护、可长期保存等。

将 DNA 分子作为信息存储介质的想法早在 20 世纪 70 年代就被提出，由 Davis[3]于 1996 年首次成功构建出了真正的 DNA 存储——成功将数字信号 0 和 1 对应到了 DNA 的 4 个碱基中。之后几年，关于 DNA 存储的研究也有不少的尝试，但在实际应用中几乎没有什么意义，直到进入 21 世纪，DNA 存储才有了新的进展，成为了一项主流研究[4]。

2012 年 8 月，Church 等[5]在 *Science* 上介绍

其开发了一种利用新一代 DNA 合成和测序技术进行 DNA 存储编码的方案，可以用来编码任意数字信息。在这项研究工作中，Church 团队将 A/C 与二进制中的“0”对应，G/T 与二进制中的“1”对应，发展了这种“2 对 1”的对应关系。他们以这种方式对序列进行设计，不仅增强了灵活性，还避免了难读取、高含量 GC 区的写入及重复序列等问题，成功将 53 426 个单词、11 个 JPG 图像和 1 个 JavaScript 程序存入 DNA 中，成就了合成生物学领域的“丘奇之书”，这也宣告了 DNA 数据存储时代的到来。考虑到 DNA 存储技术的实用性，Goldman 等[6]于 2013 年在 *Nature* 上介绍了一种可扩展的方法，成功将由硬盘存储的、总计 739 KB 的计算机文件进行编码、合成、排序，最后又重建出该文件，准确率达到了 100%，可比之前存储更多信息，实现了高容量存储，为数据信息存储技术提供了现实性，使 DNA 存储技术又向前迈进了一步。到 2017 年，华盛顿大学和微软亚洲研究院将大于 2.0×10^8 B 的数据信息编码入 DNA 分子中[7]。2018 年我国清华大学的戴俊彪课题组创建了一种利用生物体进行存储的“数据-DNA”编码方法[8]。2017 年，纽约基因组中心和哥伦比亚大学合作研发出了“DNA 喷泉算法”，提高了数据的存储密度。到 2018 年，华盛顿大学将超过 200 MB 的数据存入 DNA 分子中，并实现了 DNA 数字存储的随机访问[9]。2022 年，我国深圳华大生命科学研究院开发了一种“阴阳”双编码方法[10]，为 DNA 存储的多类型提供了重要工具。

经过数十年的研究，DNA 存储技术得到了长足发展。但在 DNA 数据编码方法、DNA 合成方法、DNA 测序方法等方面，目前还存在着新挑战；在 DNA 信息存储密度上仍然有进一步的提升空间。非天然碱基基因密码作为一个能有效扩展遗传字母表的合成生物学新方向(图 1)，有望为 DNA 存储技术领域的挑战和问题解决提供新机遇。

3　合成 DNA 的稳定性较差

理论上，DNA 存储技术非常完美，可以摆脱当前面临的海量数据窘境，但从现实情况出发，目前，DNA 存储无法大规模应用于市场上，它还有几大问题尚未解决(图 2)。

3.1　合成成本高

目前，通过化学合成法合成长链 DNA 片段仍存在着一些技术上的困难。现阶段，通过化学合成法只能得到 200 nt 以内的 DNA 序列。“写”DNA 的价格也非常可观，就目前来说，合成寡核苷酸的平均成本约为 0.001 美元/碱基，因此，存储 1 TB 数据需要约 10 亿美元。以第二代测序为例，单个样本的数据量动辄达到 TB 级。而与 DNA 存储相比，磁带存储的成本则要低很多，当存储数据规模与 DNA 存储相同(即存储 1 TB 数据)时，磁带存储的成本仅为 16 美元[11]。因此，DNA 数据存储进入实用阶段的阻

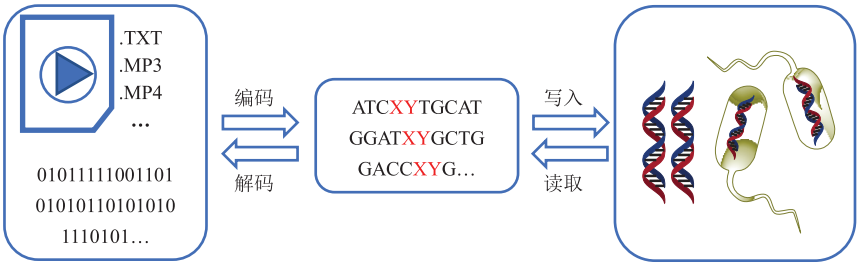


图 1　非天然碱基核苷酸 (X/Y) 密码与 DNA 数据存储

Fig. 1　Unnatural base nucleotide (X/Y) codes and DNA data storage

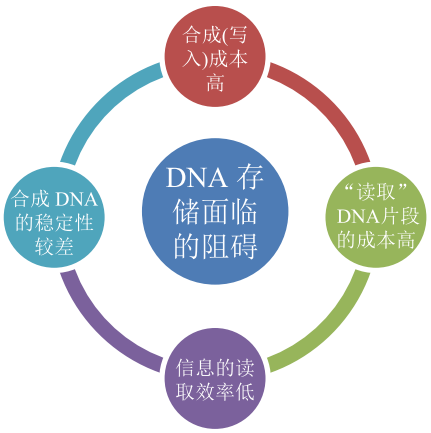


图 2 DNA 存储面临的阻碍

Fig. 2 The barriers to DNA storage

碍之一便是昂贵的 DNA 合成费用。所以，与传统存储介质相比，这一缺陷相对削弱了 DNA 存储优势。而第三代 DNA 合成技术基于酶合成，现阶段，该技术仍处于发展初期，先前报道过的酶合成方法还不能用于高通量 DNA 合成，其低通量方法也还未能步入实际应用中。

3.2 信息的读取效率低

信息解码也被称为信息的读取，主要依赖于 DNA 测序技术，自 1977 年首代 DNA 测序技术(Sanger 法)发明以来，测序技术就取得了很大的进步，在成本上减少了 10 万倍，技术方面也发展到了第三代测序。其中，一代测序的准确性较高，但耗时长，成本高；二代测序的成本、时间虽然都下降了，但在序列读取长度方面，与第一代测序技术相比，则要短很多，只能满足冷数据的读取需求；三代测序不仅增加了序列读取长度，提高了读取速率，还消除了对聚合酶链式反应(polymerase chain reaction，PCR)扩增的依赖[4]，但在测序过程中一次测序长度的错误率较高，需要通过重复测序进行纠错，变相增加了测序成本。目前，DNA 测序主要应用于短片段信息存储，依赖于二代测序，即高通量 DNA 测序技术，从建库到读取等一系列过程下来，需数天时间，数据无法实时读取[12-16]。随着数据信息量的爆炸式增长，与其他测序技术相比，三代测序虽然错误率高，但在 DNA 存储方面的读取速度上占有很大优势，随着未来技术的发展，其测序的正确率有望得到提高，将很可能更广泛地应用于 DNA 数据存储[17-21]。

3.3 合成 DNA 的稳定性较差

DNA 分子上的信息擦除、防伪等操作受限于生化反应的精确度，因此，DNA 分子上的信息虽然具有动态稳定性，但其信息稳定性仍无法达到完全准确。现阶段，加速老化模型被使用在许多研究中，但该模型的本质是外推，因此可能会导致偏离长期的实际稳定性，进而会影响数百年，甚至数百万年的稳定性预测[22]。DNA 分子具有温度敏感特性，因此，冷藏系统对 DNA 分子的稳定性和准确性尤为重要。在加速老化模型中，测量降解的方法往往不够敏感，因此依赖于大的降解效应或需要扩增步骤(如通过 PCR)，这可能会导致结果被扭曲或出现偏差。

4 非天然碱基基因密码

A、T、G、C 是自然界存储和提取生命遗传信息的基因密码。通过形成 A-T(U)和 G-C 两组碱基对，天然碱基在聚合酶的催化下完成 DNA 复制和 RNA 转录，并通过核糖体翻译蛋白。早在 20 世纪 60 年代，科学家就提出利用非天然碱基对(unnatural base pairs，UBPs)扩增自然界的基因密码[23]。半个多世纪以来，不同结构的非天然碱基先后被化学生物学家设计合成，它们可以通过非标准氢键、疏水相互作用和其他形状互补模式相互识别，进而进行有效的复制转录和翻译，甚至构筑出包含人工基因密码的半合成生命体[24]。

4.1 非天然碱基家族的发展

基于氢键的非天然碱基对与天然碱基对一样，靠氢键进行彼此间的识别。在基于氢键的非

天然碱基对中，比较具有代表性的是 isoG-isoC 碱基对(结构见图 3)。1990 年，Benner 团队通过交换碱基中氢键供受体位置的方法，利用非标准氢键模式成功地创制了 G-C 碱基对的结构类似物 isoG-isoC 非天然碱基对。但 isoG 在生理 pH 下的互变异构化和 isoC 在碱性 pH 下的不稳定性降低了该非天然碱基对的识别效率。为了克服这些缺陷，Galindo-Murillo 等[25]进一步创制了 P-Z 非天然碱基对(结构见图 3)。X 射线晶体学显示，P-Z 非天然碱基对在 a 型和 b 型双链 DNA 中均表现出与标准沃森-克里克模型天然碱基对结构的高度相似性，并且可在 Taq DNA 聚合酶催化下进行高保真度的 DNA 复制。2007 年，Yang 等[26]在 P-Z 非天然碱基对的基础上又成功创制了 B-S 非天然碱基对，并利用这两对非天然碱基对合成了 8 个字母(4 个天然碱基字母，4 个非天然碱基字母)的 DNA 和 RNA，创造了靶向特定细胞的非天然碱基核酸适配体。此外，P 碱基可形成两个氢键与天然的 C 碱基发生错配。来自日本理化研究所的 Hirao 和他的同事由此引入了阻断氢键的空间位阻概念，以降低碱基间错配的可能性。基于这一概念，他们设计出了 x-y 非天然碱基对和 s-y 非天然碱基对(结构见图 3)，其中，x 和 s 对 T 的 4-位羰基起到了立体阻碍作用。但由于无法排除 y 碱基与 A 碱基的错配，降低了 y 与 s 或 x 识别配对的选择性，因而导致该类非天然碱基对并不能进行高效复制[27-28]。

1998 年，Morales 等[29]报道了通过疏水相互作用结合的非氢键 A-T 类似物，Q-F(图 3)和 Z-F 的人工核苷酸。Z-F 非天然碱基对缺乏天然嘌呤

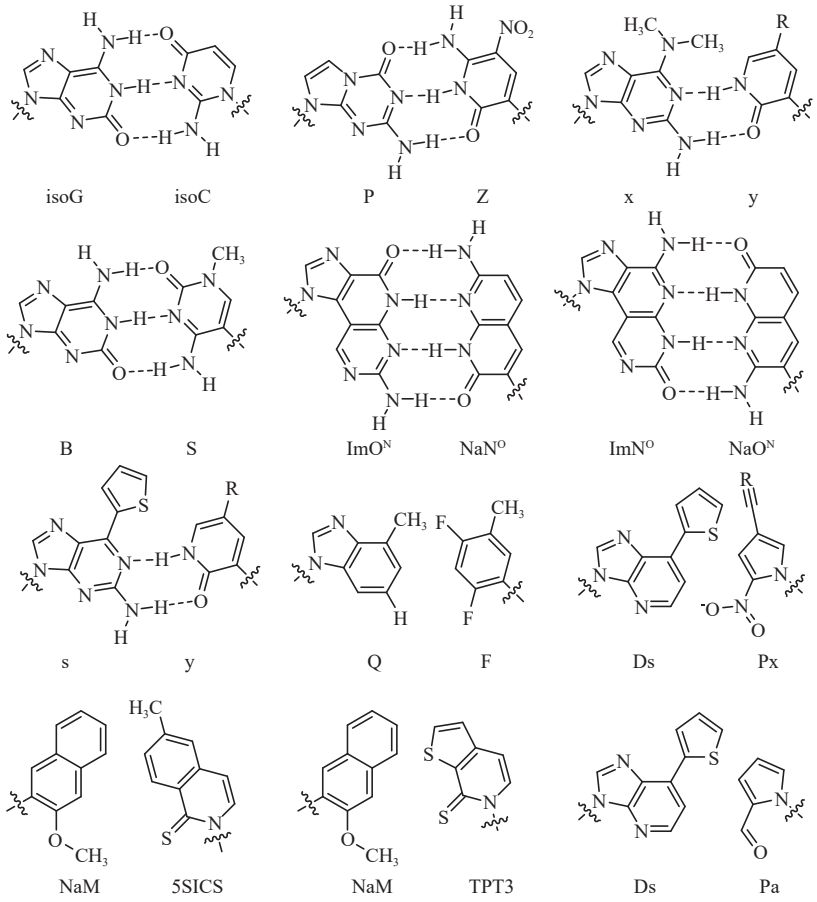


图 3　常见的非天然碱基对

Fig. 3　Common unnatural base pairs

结构 3-位 N 原子氢键受体和天然嘧啶碱基的 2-位羰基氢键受体，因而减少了它与天然碱基的识别作用，但是这些具有类似于天然碱基结构的疏水杂环分子自身可以被有效识别。比如，F 可以在 DNA 聚合酶催化下特异性地引入到模板链中 Z 的对位，表现出与天然 A-T 碱基对类似的识别复制能力[30-31]。该团队的研究凸显了 DNA 复制过程中碱基对之间形状互补的重要性。

在设计出 x-y 非天然碱基对和 s-y 非天然碱基对之后，Mitsui 等[32]又将 Morales 等[29]发现的形状互补原理应用到非天然碱基对设计中，发现通过将 Q 与 F 碱基的 5 元环模拟物配对可改善 Q-F 非天然碱基对的形状互补性，从而设计出了具有吡咯骨架和醛基的 Pa(吡咯-2-羧醛，图 3)。聚合酶反应动力学研究结果显示：Q-Pa 非天然碱基对在复制和转录效率方面比 Q-F 非天然碱基对更优。该研究团队还设计了疏水性的 Ds 碱基(图 3)和具有改良形状互补性的 Px 碱基，并使用修饰的三磷酸盐减少 Px 碱基与 A 碱基的错配[30-31,33]。此外，Px 碱基可以通过修饰与功能基团进行连接，从而广泛用于 qPCR、DNA 适配体和 RNA 标记等 DNA 技术和生物医学领域[30-31,34-35]。

在非天然碱基基因密码领域，另外一系列突出工作来自于斯克里普斯研究所的 Romesberg 团队。该团队报道了一种疏水性自配对的非天然碱基对(PICS-PICS)，它具有高的识别效率、选择性和结构稳定性[36-39]。PICS-PICS 可以在大肠杆菌聚合酶催化下插入到模板链 PICS 的对面，实现特异性识别配对[30-31]。但 PICS-PICS 非天然碱基对在引入到 DNA 双链后不能有效地进行链延伸反应。为了发现性能更为优良的此类非天然碱基对，Wu 等[40]采用了一种随机筛选的方法，通过 DNA 聚合酶反应动力学实验从 3 600 对非天然碱基组合中发现了复制和延伸效率最优的 MMO2-5SICS 非天然碱基对[40]；并在此基础上，通过进一步的构效关系优化，获得了 NaM-5SICS 非天然碱基对[39,41]。NaM-5SICS 非天然碱基对能够进行有效的 PCR 扩增。2014 年，Malyshev 等[42]将 NaM-5SICS 用于在活细胞中扩增基因字母。他们将包含 NaM-5SICS 非天然碱基对的质粒转入大肠杆菌细胞，并通过事先表达在大肠杆菌表面的核苷酸转运蛋白将 MMO2-5SICS 和 NaM-5SICS 输入到细胞质中。研究发现，大肠杆菌能够自发地利用 MMO2-5SICS 和 NaM-5SICS 完成质粒中 5SICS 和 NaM 人工基因字母的复制和细胞传代。该工作代表了地球上首个包含 6 碱基基因字母生命体的成功构筑，被 *Science* 评为当年的全球科学十大突破。Kimoto 等[30]、Malyshev 等[42]、Dhami 等[43]、Dien 等[44]通过进一步的结构优化创造了首个复制识别保真度达到天然 A-T 碱基水平(>99.99%)的 TPT3-NaM 非天然碱基对。利用 TPT3-NaM 非天然碱基对，Zhang 等[45]在 2017 年成功实现了在大肠杆菌中利用非天然碱基对编码非天然氨基酸蛋白的工作，从而实现了非天然碱基在生命体中存储遗传信息和提取遗传信息的全链条可能性[45]。2021年，Ptacin 等[46]利用非天然碱基，对半合成生命体完成了对治疗黑色素瘤蛋白质药物的定点精准修饰[46]。

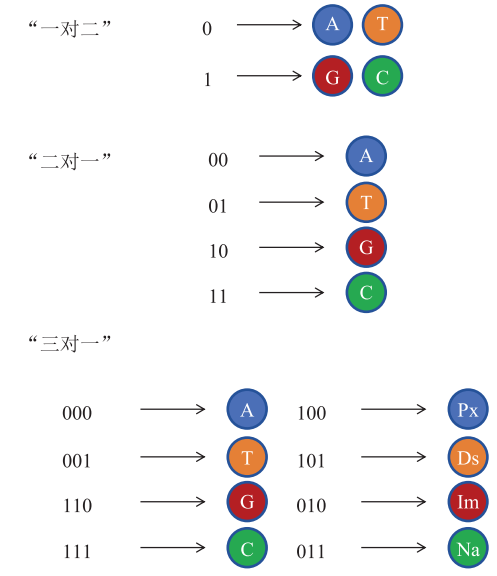
近年来，非天然碱基对基因密码领域的研究显示：类似于 4 个天然的遗传字母，性能优良的非天然碱基对不仅可以扩展遗传字母表，而且在多生物技术领域具有应用潜力，如非天然碱基对可高保真度地进行 PCR 扩增和体外转录，以及通过 SELEX 筛选获得高保真度的适配体等。2014 年，首个包含非天然碱基基因密码的生命体构筑成功，非天然碱基核苷酸已经被证明能够整合到正常的细胞中，并顺利地进行识别、复制和扩增。人工碱基基因密码开启了合成生物学研究的一个崭新篇章，同时也为核酸数据存储技术领域注入了新技术。

4.2 非天然碱基基因编码二进制转码

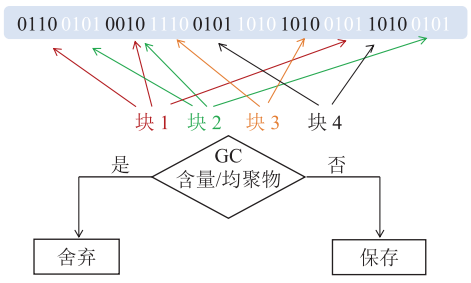
在 DNA 信息存储过程中，数字信息首先需

要通过转码方案转编成 ATCG 序列，然后，将这些序列合成为寡核苷酸或 DNA 片段，实现长期储存。在检索数据环节，采用 DNA 测序方法从合成的 DNA 中获得原始 ATCG 序列后，逆向转码为数字信息，最后，进行信息的读取。1996 年，Davis[3]进行了一次开创性的尝试，将一个图标转换为一串二进制数字，编码成一个 28 bp 的合成 DNA 分子，随后成功测序并检索图标[3]。2012 年，Church 等[47]提出了一种“一对二”二进制转码方法，并通过该转码方法将大量数字数据成功存储在 DNA 中。这种“一对二”二进制转码方法在基因字母表和美国信息交换标准码(American Standard Code for Information Interchange，ASCII)字符之间实现了直接映射，且具备转码交换的灵活性，但该方法牺牲了基因字母存储信息的密度。为了克服这个问题，2017 年，Erlich 等[48]开发喷泉码(Fountain code)，采用基本的“二对一”转码表，将[00,01,10,11]分别映射到[A,T,G,C]。如图 4(b)所示，他们首先根据预先设计的伪随机数字序列将原始二进制信息分割成小块；其次，根据转码表将带有随机种子的选定块按位相加，并转码到 4 个天然的 DNA 字母，从而创建一个新的数据块；最后，验证步骤的目的是防止单核苷酸的重复和异常 GC 含量。该编码方案中的核苷酸相互关联，具有网格状拓扑结构，随机选择和有效性验证机制确保长的单核苷酸均聚物不会出现在编码序列中。但这种方案的编码和解码复杂程度与数据大小并不是线性相关的；解码过程可能非常复杂，需要更多的资源和更长的计算时间。

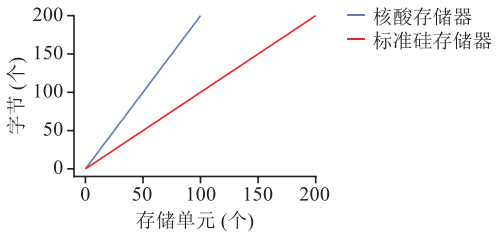
2020 年，Saptarshi 等[49]在上述转码算法的基础上，首次提出将非天然碱基应用于 DNA 信息存储的设想，并将文献报道的 Ds/Px/Im/Na 非天然碱基扩展的基因密码表用于二进制转码，从而提出“三对一”的转码方法(图 4(a))。通过这种映射，可利用单个核苷酸的碱基表示出从 000 到 111 的 8 个二进制数。理论上，与传统硅基存储系统相比，这种基于 8 个碱基的 DNA 存储系统在数据存储密度上具有显著优势。随后，他们将这种“三对一”的转码方法应用于对 ASCII 的



(a) “一对二”“二对一”和“三对一”二进制转码表



(b) “二对一”转码增加数据安全性和可靠性



(c) 核酸存储器与标准硅存储器的存储单元与对应字节的比较

图 4　DNA 存储转码方式与存储特点

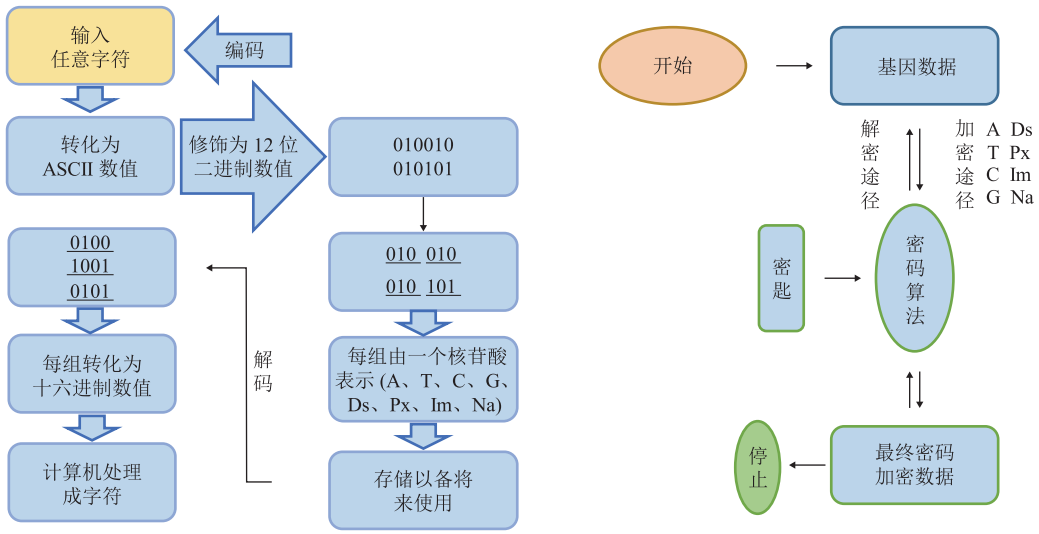
Fig. 4　Transcoding methods and characteristics of DNA storage

编码转码(图 5(a))。首先，该方案通过检测输入的字符，将其转化为对应的二进制 ASCII 值，然后在原始 ASCII 的 8 位二进制数值中进行修饰，数值前加上所需修饰的数值，将其转变为 12 位二进制数值；随后，将修饰后的 12 位二进制数值分成 4 组 3 位的二进制数值，每组数值对应一个编码的核苷酸，形成了 4 个代表不同数值的核苷酸，这样就完成了一个存储过程，即为编码过程。在随后的解码过程中，当每收到一组编码的核苷酸时，就将其转化为 12 位的二进制编码，然后按 4 位为一组分为 3 组，并集成为一个 3 组十六进制代码，并交付给计算机系统进行数据解读。该编码方案不仅保持了较高的数据密度，而且通过二进制和十六进制两种编码系统表示编码后的数据，展示了将 3 位二进制数据同化成一个单一的 DNA 字母的可能性。

4.3 基于非天然碱基基因密码的加密信息算法

虽然 Saptarshi 等[49]展示了包含 4 个非天然碱基的 8 碱基基因字母表在转化二进制及重塑 ASCII 字符中应用的理论价值，但是在真实的 DNA 数据存储中，天然的基因密码并不能真正做到完全随机编码。A、T、G、C 编码信息的能力严格受到 DNA 序列中 GC 含量和由相同碱基连续排列组成的字串(Homopolymer run-length)最大长度的限制。鉴于此原因，在 Church 等[47]建立了基本的 A、T、G、C 天然碱基编码二进制信息的基本框架之后，大量的研究聚焦于建立 DNA 存储信息的新算法，在用于规避 GC 含量和均聚物长度限制的同时，提升 DNA 存储信息的存储密度、纠错能力、检错能力和鲁棒性。

在此背景下，2022 年，Biswas 等[50]将所设计的“三对一”转码方法应用于融合了 Vernam 和 Vigenére 密码的加密算法，并展示了这种算法可对抗均聚物的运行长度和 GC 含量限制，而不会对每个核苷酸的实际数据密度造成很大的影响。具体的算法框架如图 5(b)所示，首先将二进制信息依据表 1 所示的 3-字节转码方法转码为 8 碱基(ATGCDINP)的 DNA 存储信息。在此过程中，利用综合的信息加密方法建立加密文本，并利用这些包含 8 碱基的加密文本建立有效的数据加密钥匙和相对应的解密方法，从而构建基于 8 碱基的加密和解密途径 。该团队进而评价融合



(a) 基于非天然碱基扩展基因密码的 ASCII 编码流程图

(b) 加密信息算法流程图

图 5 DNA 存储流程图

Fig. 5 DNA storage flowchart

了加密和解密途径的 8 碱基 DNA 存储算法中，对抗均聚物的运行长度和 GC 含量限制，并展示了该算法在文本数据和图像数据存储方面的应用。最后从平均报道数据密度[average reported data density(bits/nucleotide)]、考虑 GC 含量和均聚物长度限制方面与已建立的 20 种算法进行了比对(表 1)[50]。

5　非天然碱基基因密码用于 DNA 信息存储所面临的挑战

科学家虽然从理论计算角度展示了扩展基因密码，如包含 DINP 的 8 碱基系统在二进制转码、编码 ASCII 字符及更加复杂的信息编码系统中，其应用的可能性和潜在的巨大优势，包括提升单碱基编码信息的密度至 2.99～3 bits/base，有效规避 GC 含量与由相同碱基连续排列组成的字串最大长度限制等问题。但是这种建立在理论计算上的研究与实践应用之间仍然存在诸多方面的差距。为了推动非天然碱基在证实的 DNA 信息存储领域的应用系统的建立，结合本研究团队在非天然碱基创制及生物技术领域的长期研究，本文尝试总结该领域存在的 3 方面重要挑战。

5.1　寻找更为普适性的非天然碱基测序方法

读取存储在 DNA 分子中的信息需要开发便捷廉价的测序方法。与天然碱基的测序方法相比，非天然碱基的测序方法十分不完善。以扩展核酸记忆算法中提到的 Ds-Px(Thienyl Imidazo

表 1　非天然碱基用于信息加密算法与已有算法的比对[50]

Table 1　Unnatural bases are used to compare information encryption algorithms with existing ones[50]

研究小组	平均报告数据密度 (字节/核苷酸)	是否说明均聚物运行长度	是否说明 GC 含量
Church et al. (2012)	0.830 0	否	否
Jimenez-Sanchez (2013)	2.000 0	否	否
Goldman et al. (2013)	0.330 0	是	否
Grass et al. (2015)	1.440 0	否	否
Yazdi et al. (2015)	1.580 0	否	是
Bornholt et al. (2016)	0.880 0	是 (在一定程度上)	否
Blawat et al. (2016)	0.920 0	是	否
Erlich and Zielinski (2017)	1.570 0	是	是
Yazdi et al. (2017)	1.720 0	是	是
Suyehira et al. (2017)	1.330 0	是	否
Song et al. (2018)	1.900 0	是	是
Immink and Cai (2018)	1.999 7	是	是
for $k = 5$, $m = 6$	-	-	-
Immink and Cai (2018)	-	-	-
Organick (2018)	0.810 0	否	否
Wang (2019)	1.917 0	是	是
Lopez (2019)	1.570 0	否	否
Choi (2019)	1.780 0	否	否
Anavy (2019)	1.940 0	否	否
Ceze et al. (2019) and Lee(2019)	1.570 0	是	否
Biswas (2019, 2020)	3.000 0	否	否
Nguyen et al. (2020)	1.995 7	是	是
方案结果	＞2.99	是	是

Pyridine and a nitro propynyl pyrrole)和 Im-Na(an Imidazo pyrimidine and a naphthyridine)非天然碱基为例，其测序方法并不能和建立在 Watson-Crick 碱基识别模式上的天然碱基 Sanger 测序方法完全兼容。Ds-Px 的测序方法是利用在天然碱基 Sanger 测序过程中缺乏于 Ds 或 Px 识别的双脱氧核苷酸荧光信号，发展而来的 Sanger Gap 测序，利用测序数据中的 Gap 位置识别 Ds 或 Px。这种方法的缺陷是无法区分 Ds 和 Px，且不能使用商业化的测序方法(需要在测序过程中额外添加双脱氧的 dDsTP 和 dPxTP)[51]。2012 年，Yamashige 等[51]利用 Ds-Px 配对偏好性发展了与天然碱基 Sanger 测序方法更为兼容的“诱导突变”测序方法[51]。其还利用在 PCR 扩增反应中加入具有诱导突变功能的非天然碱基核苷酸，通过两轮扩增可以得到在原始 Ds-Px 位点分别突变 A-T 和 T-A 的 PCR 产物，进而通过两轮 Sanger 测序分别定位出 Ds 和 Px 的位置。这种方法解决了 Sanger Gap 测序不能区分 Ds 和 Px，以及与商业测序不兼容的问题，并成功应用于适配体(aptamer)中随机非天然碱基的测序。但是该方法自身的序列偏好性和诱变信号复杂性的问题仍需进一步解决。

表 2 列出了目前具有代表性的非天然碱基的测序方法。其中，本文研究团队发展的针对 TPT3-NaM 人工碱基的“桥梁碱基”测序方法可以特异性地将 DNA 中多个 TPT3-NaM 碱基分别转换为 C-G 碱基和 A-T 碱基，并且展示了极高的序列普适性，可以读通 4 096 条随机序列中的 4 096 条序列(读通率为 100%)。这些性能为发展 TPT3-NaM 作为 DNA 存储字母提供了重要的测序技术基础。但是，一些非天然碱基，如 Im-Na 等，目前仍然缺乏可以识别其在随机序列中位置的方法。

5.2 非天然碱基的最高含量及相同非天然碱基均聚物长度的最大限度

在基于天然碱基的 DNA 数据存储算法构建过程中，大量的注意力集中在解决 GC 含量约束和均聚物长度约束两个阻力方面。理论上，虽然非天然碱基的加入可以有效稀释 DNA 序列中的 GC 含量，并显著减少均聚物长度的限制。但是，非天然碱基自身的物理化学性质和非天然碱基与 DNA 聚合酶的特殊识别方式使得非天然碱基在 DNA 序列中的最大含量及最大的连续排列组成的字串长度也不尽相同。表 3 列出了目前具有代表性的非天然碱基在 PCR 扩增和测序反应中检测到的连续排列组成的字串最大长度。此外，非天然碱基在一段 DNA 序列中的最高含量是一个更为复杂的问题。不同的人工碱基利用不同的碱基识别方式，如 P-Z 碱基、Im-Na 碱基使

表 2 具有高 DNA 复制能力的非天然碱基及其测序方法

Table 2 An unnatural base with high DNA replication capacity and its sequencing method

非天然碱基	一代测序方法	适用范围	诱导后测序方法	适用范围	四代测序方法	适用范围
Ds-Px	Sanger Gap[51]	已知位点的验证/识别保真度检测	诱导突变[52]	未知位点的检测/适配体测序	-	-
Im-Na	Sanger Gap[51]	已知位点的验证/识别保真度检测	-	-	-	-
P-Z	限制性酶切[53]	识别保真度检测	诱导突变[54]	适配体测序	纳米孔[55]	多碱基已知位点的验证
TPT3-NaM	Sanger terminal[56]	已知位点的验证/识别保真度检测	桥梁碱基[57]	多碱基位点的验证/未知位点的检测/识别忠实性检测	纳米孔[58]	单碱基已知位点的验证/识别忠实性检测

用不同于 A-T 和 G-C 碱基的氢键识别方式，Ds-Px 和 TPT3-NaM 碱基则采用基于碱基几何形状的疏水识别方式。与天然碱基的 Watson-Crick 氢键识别模式不同，非天然碱基的引入必然改变标准的 DNA 双螺旋结构[46]，并且引入的非天然碱基数量越多，对 DNA 标准双螺旋结构的影响越剧烈。Hiriao 团队曾经报道了在 DNA 序列中保持有效复制能力的两对 Ds-Px 非天然碱基的最小间隔为 5 个天然碱基[59]。目前，由于实验数据的缺乏，对于大多数非天然碱基而言，在一段给定长度的 DNA 序列中，GC 最高含量仍然无法给出。非天然碱基的最高含量和最大长度限制数据的不足带来了非天然碱基编码 DNA 数据存储信息的更多不确定性。为了解决这方面的问题，未来仍然需要进行大量的实验性研究。

表 3　非天然碱基核苷酸连续排列最大限度

Table 3　Maximum continuous arrangement of unnatural base nucleotides

非天然碱基	连续排列最大限度
Ds-Px	不能相邻或最小间隔为 5 个天然碱基[59]
Im-Na	未知[60]
P-Z	3[54]
TPT3(5SICS)-NaM	3[61]

5.3　非天然碱基应用于 DNA 信息存储的成本问题

与传统的存储技术相比，DNA 数据存储的费用昂贵，阻碍其进入实际应用阶段。虽然非天然碱基核苷可以与天然碱基核苷采用相似的亚磷酰化固相合成方法顺利引入到 DNA 序列中，但是这些非天然碱基核苷的合成成本仍然是一个需要考虑的问题。与天然碱基核苷酸可以通过生物发酵的方法合成不同，目前报道的所有能够有效复制的非天然碱基核苷均需要通过多步繁琐的化学合成来制备，其合成成本远高于天然碱基。例如：商业化的腺嘌呤碱基脱氧核糖核苷（DNA 中 A 碱基的核苷形式）的合成价格约为 5 元/g，而非天然碱基核苷酸，如 NaM 碱基脱氧核糖核酸，其商品化价格约为 5 000 元/g。因此，优化非天然碱基的合成途径，降低其合成成本，是建立 DNA 信息存储技术所需的一个研究方向。此外，DNA 信息存储的另外一个成本问题来自于 DNA 测序。如上所述，目前，非天然碱基的测序技术仍然远远落后于天然碱基的测序技术。虽然利用非天然碱基独特的理化性质也可以实现对 DNA 中非天然碱基的准确定位，但是从成本和技术通用性角度考虑，创制与天然碱基测序技术兼容的非天然碱基二代测序和三代测序方法将更为经济实用。

6　未来及展望

全球数据的爆炸式增长几乎已经超过了现有的存储容量，因此迫切需要一种新的存储策略。DNA 作为遗传信息的天然介质，成为了下一代存储介质的有力候选者，是未来数据存储的理想载体。DNA 存储数据具有很多优势，如存储密度大等。首先，假设 DNA 能够像大肠杆菌那样进行包装，那么，全世界的数据信息都可以储存在质量为 1 kg、空间如粉笔盒大小的一堆 DNA 中；其次，DNA 存储耗能极少，要存储同样大小的信息，DNA 存储的耗能量相当于闪盘的亿分之一；最后，DNA 存储还具有卓越的耐用性，一般物理存储设备的使用寿命往往不到 10 年，而 DNA 则可将遗传信息保存 100 年以上，如果是在低于零下 18 ℃ 的低温环境中，则甚至可保存上万年，数十万年。但是，现有 DNA 存储的系统建立是在利用天然 ATGC 基因字母或者天然类似物编码的基础上，这不仅在源头上限制了用于信息编码的源代码数量和算法设计的空间，而且不可避免地受到自然原则对遗传密码使用的底层逻辑约束，如 GC 含量的约束和均聚物长度的约束。

非天然碱基基因密码的诞生从生命中心法则

的底层扩展了可编码生命遗传信息的字母表。非天然碱基基因密码的独特之处在于它们能够正交遗传天然碱基字母，并能够进行遗传信息的复制和存储。理论算法显示，非天然碱基可应用于DNA 存储，并有望将非天然碱基和 DNA 存储的优点相结合，在扩展遗传字母表、扩大 DNA 多样性的同时，提高存储密度、降低存储能耗量、增加数据保密性能。但是，如上所述，目前的DNA 存储成本仍然很高，而目前合成非天然碱基核苷酸的成本也比较高，这在一定程度上限制了非天然碱基在 DNA 存储中的使用。目前，绝大多数非天然碱基在一段 DNA 序列中的最高可用含量还无法确定，这也给非天然碱基编码 DNA 数据信息带来了不确定性。此外，降低非天然碱基核苷酸存储时的读取成本，发展廉价的非天然碱基测序技术也是迫切需要解决的关键技术[62]。

值得一提的是，随着非天然碱基基因密码研究的深入，越来越多的性能优良的非天然碱基分子被创制，其配套的合成测序技术也随之被发展出来。此外，随着天然 DNA 合成、测序技术的日趋完善，这些技术也日益展现出与非天然碱基基因密码兼容空间。因此，本文认为通过更多科学家在 DNA 存储框架下对非天然碱基技术的配对性优化与创新研究，利用非天然碱基编码 DNA、进行信息储存将是一个具有广阔应用前景的领域。尤其是，利用非天然碱基编码 DNA，有望打破现有 DNA 信息存储对 GC 含量和均聚物长度的约束限制，提高 DNA 存储密度；同时，亦可利用非天然碱基读取方式与天然碱基读取方式不同的特点，从物理源头增加 DNA 数据的保密性能。

参 考 文 献

[1] 董一名, 孙法家, 武瑞君, 等. DNA 数字信息存储的研究进展 [J]. 合成生物学, 2021, 2(3): 323-334.
Dong YM, Sun FJ, Wu RJ, et al. Research progress on DNA molecules for digital information storage [J]. Synthetic Biology Journal, 2021, 2(3): 323-334.

[2] Reinsel D, Gants J, Rydning J. Data age 2025: the evolution of data to life-critical [Z/OL]. https://assets.ey.com/content/dam/ey-sites/ey com/en_gl/topics/workforce/Seagate-WP-DataAge2025-March-2017.pdf.

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

[4] Dong YM, Sun FJ, Ping Z, et al. DNA storage: research landscape and future prospects [J]. National Science Review, 2022, 7(6): 1092-1107.

[5] Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA [J]. Science, 2012, 337(6102): 1628.

[6] Goldman N, Bertone P, Chen SY, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA [J]. Nature, 2013, 494(7435): 77-80.

[7] Bornholt J, Lopez R, Carmean DM, et al. Toward a DNA-based archival storage system [J]. IEEE Micro, 2017, 37(3): 98-104.

[8] 戴俊彪, 吴庆余, 乃哥麦提·伊加提, 等. 将数据进行生物存储并还原的方法: 中国, CN201610786435.2 [P]. 2018-03-13[2023-10-31]. https://kns.cnki.net/kcms2/article/abstract?v=o3lE5Q-17mDIUMYgCHsWCJkIc_ww_bQ-SadSwQWdmTMOpGvPHKTN0pUQ7zij2onDg3PzCbpi2Y4nUtGXtmDc3lqE7TbC2IsLKS7v0cTvq_3W7luIccSVp5OnkkQ50xiEURBDE6WYPUw=&uniplatform=NZKPT&language=CHS.
Dai JB, Wu QY, Negomaiti I, et al. The method of biological storage and restoration of data: China, CN201610786435.2 [P]. 2018-03-13[2023-10-31]. https://kns.cnki.net/kcms2/article/abstract?v=o3lE5Q-17mDIUMYgCHsWCJkIc_ww_bQ-SadSwQWdmTMOpGvPHKTN0pUQ7zij2onDg3PzCbpi2Y4nUtGXtmDc3lqE7TbC2IsLKS7v0cTvq_3W7luIccSVp5OnkkQ50xiEURBDE6WYPUw=&uniplatform=NZKPT&language=CHS.

[9] Organick L, Ang SD, Chen YJ, et al. Random access in large-scale DNA data storage [J]. Nature

Biotechnology, 2018, 36(3): 242-248.

[10] Ping Z, Chen SH, Zhou GY, et al. Towards practical and robust DNA-based data archiving using the yin–yang codec system [J]. Nature Computational Science, 2022, 2(4): 234-242.

[11] Bioglio V, Grangetto M, Gaeta R, et al. On the fly Gaussian elimination for LT codes [J]. IEEE Communications Letters, 2009, 13(12): 953-955.

[12] Wetterstrand KA. DNA sequencing costs: data [J/OL]. National Human Genome Research Institute, 2013. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

[13] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies [J]. Nature Reviews Genetics, 2016, 17(6): 333-351.

[14] Escalona M, Rocha S, Posada D. A comparison of tools for the simulation of genomic next-generation sequencing data [J]. Nature Reviews Genetics, 2016, 17(8): 459-469.

[15] Mardis ER. A decade's perspective on DNA sequencing technology [J]. Nature, 2011, 470(7333): 198-203.

[16] Pushkarev D, Neff NF, Quake SR. Single-molecule sequencing of an individual human genome [J]. Nature Biotechnology, 2009, 27(9): 847-850.

[17] Coupland P, Chandra T, Quail M, et al. Direct sequencing of small genomes on the pacific biosciences RS without library preparation [J]. Biotechniques, 2012, 53(6): 365-372.

[18] Quail MA, Smith M, Coupland P, et al. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers [J]. BMC Genomics, 2012, 13: 341.

[19] Quick J, Quinlan AR, Loman NJ. A reference bacterial genome dataset generated on the MinION^{TM} portable single-molecule nanopore sequencer [J]. GigaScience, 2014, 3(1): 22.

[20] Jain M, Fiddes IT, Miga KH, et al. Improved data analysis for the MinION nanopore sequencer [J]. Nature Methods, 2015, 12(4): 351-356.

[21] Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science [J]. Nature Reviews Genetics, 2016, 17(3): 175-188.

[22] Matange K, Tuck JM, Keung AJ. DNA stability: a central design consideration for DNA data storage systems [J]. Nature Communications, 2021, 12(1): 1358.

[23] Rich A. On the problems on evolution and biochemical information transfer [M]. New York: Academic Press, 1962: 103-126.

[24] Hashimoto K, Fischer EC, Romesberg FE. Efforts toward further integration of an unnatural base pair into the biology of a semisynthetic organism [J]. Journal of the American Chemical Society, 2021, 143(23): 8603-8607.

[25] Galindo-Murillo R, Barroso-Flores J. Hydrophobic unnatural base pairs show a Watson-Crick pairing in micro-second molecular dynamics simulations [J]. Journal of Biomolecular Structure and Dynamics, 2020, 38(14): 4098-4106.

[26] Yang ZY, Sismour AM, Sheng PP, et al. Enzymatic incorporation of a third nucleobase pair [J]. Nucleic Acids Research, 2007, 35(13): 4238-4249.

[27] Ohtsuki T, Kimoto M, Ishikawa M, et al. Unnatural base pairs for specific transcription [J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(9): 4922-4925.

[28] Hirao I, Ohtsuki T, Fujiwara T, et al. An unnatural base pair for incorporating amino acid analogs into proteins [J]. Nature Biotechnology, 2002, 20(2): 177-182.

[29] Morales JC, Kool ET. Efficient replication between non-hydrogen-bonded nucleoside shape analogs [J]. Nature Structural Biology, 1998, 5(11): 950-954.

[30] Kimoto M, Hirao I. Genetic alphabet expansion technology by creating unnatural base pairs [J]. Chemical Society Reviews, 2020, 49(21): 7602-7626.

[31] Hirao I, Kimoto M, Yamashige R. Natural versus artificial creation of base pairs in DNA: origin of nucleobases from the perspectives of unnatural base pair studies [J]. Accounts of Chemical Research, 2012, 45(12): 2055-2065.

[32] Mitsui T, Kitamura A, Kimoto M, et al. An

unnatural hydrophobic base pair with shape complementarity between pyrrole-2-carbaldehyde and 9-methylimidazo[(4,5)-b]pyridine [J]. Journal of the American Chemical Society, 2003, 125(18): 5298-5307.

[33] Hirao I, Kimoto M, Mitsui T, et al. An unnatural hydrophobic base pair system: site-specific incorporation of nucleotide analogs into DNA and RNA [J]. Nature Methods, 2006, 3(9): 729-735.

[34] Kimoto M, Sato A, Kawai R, et al. Site-specific incorporation of functional components into RNA by transcription using unnatural base pair systems [J]. Nucleic Acids Symposium Series, 2009, (53): 73-74.

[35] Yamashige R, Kimoto M, Okumura R, et al. Visual detection of amplified DNA by polymerase chain reaction using a genetic alphabet expansion system [J]. Journal of the American Chemical Society, 2018, 140(43): 14038-14041.

[36] Ogawa AK, Wu YQ, McMinn DL, et al. Efforts toward the expansion of the genetic alphabet: information storage and replication with unnatural hydrophobic base pairs [J]. Journal of the American Chemical Society, 2000, 122(14): 3274-3287.

[37] Ogawa AK, Wu YQ, Berger M, et al. Rational design of an unnatural base pair with increased kinetic selectivity [J]. Journal of the American Chemical Society, 2000, 122(36): 8803-8804.

[38] Leconte AM, Hwang GT, Matsuda S, et al. Discovery, characterization, and optimization of an unnatural base pair for expansion of the genetic alphabet [J]. Journal of the American Chemical Society, 2008, 130(7): 2336-2343.

[39] Malyshev DA, Seo YJ, Ordoukhanian P, et al. PCR with an expanded genetic alphabet [J]. Journal of the American Chemical Society, 2009, 131(41): 14620-14621.

[40] Wu YQ, Ogawa AK, Berger M, et al. Efforts toward expansion of the genetic alphabet: optimization of interbase hydrophobic interactions [J]. Journal of the American Chemical Society, 2000, 122(32): 7621-7632.

[41] Malyshev DA, Pfaff DA, Ippoliti SI, et al. Solution structure, mechanism of replication, and optimization of an unnatural base pair [J]. Chemistry A European Journal, 2010, 16(42): 12650-12659.

[42] Malyshev DA, Dhami K, Lavergne T, et al. A semi-synthetic organism with an expanded genetic alphabet [J]. Nature, 2014, 509(7500): 385-388.

[43] Dhami K, Malyshev DA, Ordoukhanian P, et al. Systematic exploration of a class of hydrophobic unnatural base pairs yields multiple new candidates for the expansion of the genetic alphabet [J]. Nucleic Acids Research, 2014, 42(16): 10235-10244.

[44] Dien VT, Holcomb M, Feldman AW, et al. Progress toward a semi-synthetic organism with an unrestricted expanded genetic alphabet [J]. Journal of the American Chemical Society, 2018, 140(47): 16115-16123.

[45] Zhang Y, Ptacin JL, Fischer EC, et al. A semi-synthetic organism that stores and retrieves increased genetic information [J]. Nature, 2017, 551(7682): 644-647.

[46] Ptacin JL, Caffaro CE, Ma L, et al. An engineered IL-2 reprogrammed for anti-tumor therapy using a semi-synthetic organism [J]. Nature Communications, 2021, 12(1): 4785.

[47] Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA [J]. Science, 2012, 337(6102): 1628.

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

[49] Saptarshi B, Subhrapratim N, Jamuna KS, et al. Extended nucleic acid memory as the future of data storage technology [J]. International Journal of Nano and Biomaterials, 2020, 9(1-2): 2-17.

[50] Biswas S, Dey S, Nath P, et al. Cipher constrained encoding for constraint optimization in extended nucleic acid memory [J]. Computational Biology and Chemistry, 2022, 99: 107696.

[51] Yamashige R, Michiko K, Takezawa Y, et al. Highly specific unnatural base pair systems as a third base pair for PCR amplification [J]. Nucleic

Acids Research, 2012, 40(6): 2793-2806.

[52] Hamashima K, Soong YT, Matsunaga KI, et al. DNA sequencing method including unnatural bases for DNA aptamer generation by genetic alphabet expansion [J]. ACS Synthetic Biology, 2019, 8(6): 1401-1410.

[53] Yang ZY, Chen F, Alvarado BJ, et al. Amplification, mutation, and sequencing of a six-letter synthetic genetic system [J]. Journal of the American Chemical Society, 2011, 133(38): 15105-15112.

[54] Zhang LQ, Yang ZY, Sefah K, et al. Evolution of functional six-nucleotide DNA [J]. Journal of the American Chemical Society, 2015, 137(21): 6734-6737.

[55] Kawabe H, Thomas CA, Hoshika S, et al. Enzymatic synthesis and nanopore sequencing of 12-letter supernumerary DNA [J]. Nature Communications, 2023, 14(1): 6820.

[56] Malyshev DA, Seo YJ, Ordoukhanian P, et al. PCR with an expanded genetic alphabet [J]. Journal of the American Chemical Society, 2009, 131(41): 14620-14621.

[57] Wang HL, Zhu WY, Wang C, et al. Locating, tracing and sequencing multiple expanded genetic letters in complex DNA context via a bridge-base approach [J]. Nucleic Acids Research, 2023, 51(9): 52.

[58] Ledbetter MP, Craig JM, Karadeema RJ, et al. Nanopore sequencing of an expanded genetic alphabet reveals high-fidelity replication of a predominantly hydrophobic unnatural base pair [J]. Journal of the American Chemical Society, 2020, 142(5): 2110-2114.

[59] Matsunaga KI, Kimoto M, Hirao I. High-affinity DNA aptamer generation targeting von willebrand factor a1-domain by genetic alphabet expansion for systematic evolution of ligands by exponential enrichment using two types of libraries composed of five different bases [J]. Journal of the American Chemical Society, 2017, 139(1): 324-334.

[60] Saito-Tarashima N, Minakawa N. Unnatural base pairs for synthetic biology [J]. Chemical and Pharmaceutical Bulletin, 2018, 66(2): 132-138.

[61] Malyshev DA, Dhami K, Quach HT, et al. Efficient and sequence-independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet [J]. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109(30): 12005-12010.

[62] 黄小罗, 戴俊彪. 人工 DNA 合成技术: DNA 数据存储的基石 [J]. 合成生物学, 2021, 2(3): 335-353.
Huang XL, Dai JB. DNA synthesis technology: foundation of DNA data storage [J]. Synthetic Biology Journal, 2021, 2(3): 335-353.