大 学 化 学
Univ. Chem. **2025,** *40* (8), 211

•科普• doi: 10.12461/PKU.DXHX202410079 www.dxhx.pku.edu.cn

敦煌壁画的“DNA变身”

孙乐源[1]，解笑瑜[2]，陈方方[1,*]

[1]西北大学化学与材料科学学院，化学国家级实验示范中心，西安 710127

[2]西安交通大学医学部药学院，西安 710061

摘要：DNA作为一种密度极高且持久性强的信息存储载体，在长期保存文化遗产数据方面具备巨大的潜力。本文通过形象的语言探讨利用DNA存储技术将敦煌壁画转化为DNA序列，利用双层错误校正码和de Bruijn图算法确保数据的稳定性和恢复性。并利用“细胞磁盘”作为存储介质，有望实现DNA数据的高密度和长期存储，激发读者了解前沿科技的兴趣。

关键词：DNA存储；壁画保护；前景与挑战

中图分类号：G64；O6

“DNA Transformation” of Dunhuang Murals

Leyuan Sun [1], Xiaoyu Xie [2], Fangfang Chen [1,*]

[1] National Chemistry Experimental Teaching Demonstration Center, College of Chemistry and Materials Science, Northwest University, Xi’an 710127, China.

[2] School of Pharmacy, Health Science Center, Xi’an Jiaotong University, Xi’an 710061, China.

Abstract: DNA, as a highly dense and durable information storage medium, holds significant potential for the long-term preservation of cultural heritage data. This paper explores the process of converting data from the Dunhuang murals into DNA sequences using DNA storage technology, highlighting the application of dual-layer error correction codes and de Bruijn graph algorithms to ensure data stability and recoverability. Furthermore, the use of “cell disk” as a storage medium is expected to enable high-density and long-term DNA data storage. The goal is to inspire readers to engage with and explore cutting-edge technologies.

Key Words: DNA storage; Mural protection; Prospects and challenges

敦煌壁画作为中华民族的艺术瑰宝，承载着深厚的历史文化底蕴。然而，随着时间的流逝和环境的侵蚀，让这些珍贵的艺术瑰宝面临着褪色和损毁的威胁。如何让这些壁画再次“活”起来，成为人类文明的永恒见证？李想是一位文物保护工作者，结束了一天的忙碌后，她在思考如何拓宽文化遗产保护方式的过程中进入梦乡。

梦境中，李想听到一声温和的问候：“前辈，您好！欢迎您来到一千年后的未来。” 她打量四周，发现自己置身于一个充满科技感的空间，周围展示着的敦煌壁画不仅与原壁画别无二致，甚至在细节上更加清晰，让她感到无比震撼。“这是怎么做到的？”李想不禁问道。小机器人小一解释道：“这些壁画之所以能够如此完美地呈现，得益于一项革命性的技术——DNA存储。”小一继续介绍，

收稿：2024-10-21；录用：2025-01-02；网络发表：2025-05-20

*通讯作者，Email: chenff@nwu.edu.cn

基金资助：国家自然科学基金面上项目(82373832)；陕西基础科学研究院项目(23JHQ048)

“您也知道，历经千年的风雨侵蚀和人为破坏，敦煌壁画的颜料层开始剥落，色彩逐渐褪色，一些细节甚至已经无法辨认，敦煌壁画面临着严重的保护挑战。”

“面对这样的挑战，一个大胆的设想油然而生：如果能够把敦煌壁画的图案和色彩转化为特殊的代码，存储在一种几乎与地球同寿的介质中，那么壁画就能抵御时间的侵蚀，保存千年甚至万年。这听起来像是科幻小说里的情节，但科学家们已经将这一设想变为现实。”机器人小一继续说道，“DNA存储技术通过将数字信息编码为DNA的碱基序列，利用DNA作为信息载体来实现信息的长期存储。DNA分子的稳定性和高密度存储特性，使得它成为理想的存储介质。在这个时代，我们已经能够精确地将壁画的每一个细节转化为DNA序列，并在需要时重新‘唤醒’它们，恢复壁画原貌。”

接下来就由我来为您简单介绍，并带您简单参观我们的文保研究所。”机器人小一如是说。

1 初见DNA壁画

“时代的车轮滚滚向前，如您所知，信息存储技术也经历了翻天覆地的变革。从最初的石刻、泥板，到后来的纸张、胶片，再到你们那个时代的磁带、光盘和硬盘，每一次技术的飞跃都极大地推动了文明的传承与发展。然而，面对日益增长的数据存储需求，传统的存储介质开始显得捉襟见肘。在这样的背景下，DNA存储概念的提出并非空想。DNA分子由四种碱基(腺嘌呤A、胸腺嘧啶T、鸟嘌呤G、胞嘧啶C)组成，通过不同的排列组合，构成了生物遗传信息的基础。科学家们发现，这种四元的碱基配对系统，与计算机中的二进制系统有着天然的相似性。通过特定的编码规则，可以将二进制数据转换为DNA的碱基序列，再通过化学或者酶促合成技术合成具有特定序列的DNA链，借此来实现信息的存储。当需要读取数据时，通过扩增数据编码的DNA目标序列，并通过数字解码恢复原始信息。这种存储方式不仅具有极高的存储密度，而且具有很好的稳定性，理论上可以在常温下保存数千年而不丢失信息。我们目前采用的是由天津大学元英进教授团队在2022年提出的DNA存储算法，不仅可以将敦煌壁画的图像信息通过特殊的编码转换成为DNA序列[1]，还能大大提高这一技术的容错率。”机器人小一讲解道。李想颔首，她确实听说过这种技术，“这项技术是具体如何操作的呢？”

机器人小一娓娓道来：“DNA存储技术在存储容量和数据完整性方面为敦煌壁画的长期保存提供了新的解决方案。简单来说，我们首先使用高精度扫描技术捕捉壁画的每一个细节，并将其转化为数字信息，完成壁画的数字化数据准备。如图1流程所示，我们将10幅字节大小为6.8 MB的敦煌壁画图像数据，转变为DNA碱基序列。然后通过大量的寡核苷酸合成，生成了包含210,000种独特类型的DNA寡核苷酸的‘主池’数据库，每种寡核苷酸代表图像文件的一个部分。最后，这些寡核苷酸

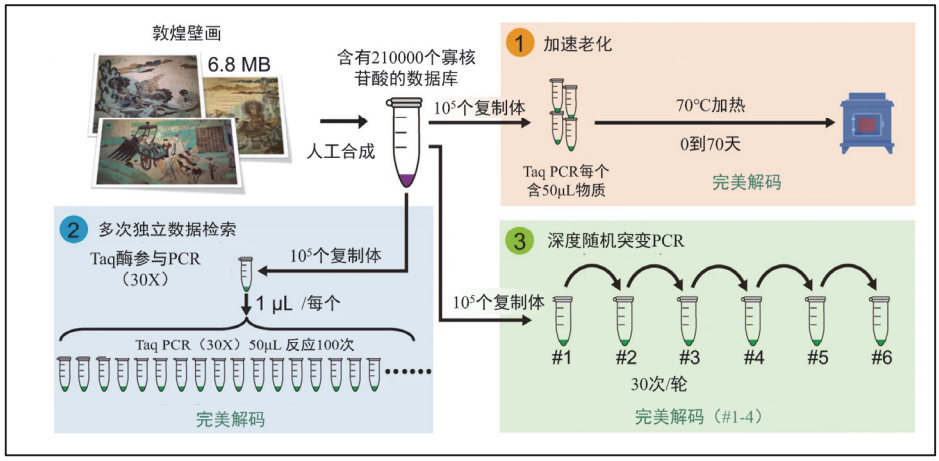

图1 敦煌壁画的DNA存储[1]

被合成为合适的DNA序列，就存储了完整的敦煌壁画数据。我们将这些DNA序列在70 ℃下保存了70天去模拟在恶劣环境下进行DNA存储的可行性，尽管在这个过程中DNA受到了严重的损伤，通过特定的算法仍能够将数据恢复。此外，还通过故意引入非特异性扩增和深度错误倾向扩增来引入错误以验证是否能进行准确的壁画DNA存储，均取得了满意的结果。而后，我们把这些DNA序列拆解重排，并小心翼翼地嵌入到微生物的基因中。DNA序列就像是指令，指导微生物如何复制和保存这些信息。”机器人小一继续说，“随着微生物的繁殖，这些信息得以代代相传，实现了长期存储，并能够在适当的时机被重新读取和应用。”

2 DNA存储数据转换原理

2.1 数据的采集

“DNA存储的流程与传统存储技术相似，都是从数据采集开始，然后将数据存储在目标介质中。一般来说，DNA存储的数据采集分为转码和DNA定向合成两个阶段。转码过程首先将待存储的数据(如文本、图片、音频文件等)转化为二进制数据，然后将数字信息转化编码为DNA序列。这不仅仅是复制粘贴那么简单，而是一场跨越时空的数据大迁徙。首先，我们用最先进的技术，为壁画进行一次超高分辨率的‘体检’。将壁画的每一个细节——图像信息(颜色、纹理和形状)、颜料信息(种类和各种颜料的分布区域)和病害信息(裂缝、褪色、腐蚀情况)——都捕捉下来，转化为数字信号。然后把它分解成一个个像素点，每个点都记录着红、绿、蓝(RGB)三种颜色的信息。但是，DNA存储技术是个四进制的系统，它只认得A、T、C、G这四种碱基。所以，我们需要把这些RGB颜色值转换成DNA能理解的语言。这就像是把壁画的颜色信息翻译成一种密码，只不过这个密码是用DNA编写的。我们先来个简化操作，把敦煌壁画的RGB颜色值转换成灰度值。这样，原本复杂的色彩信息就变得简洁明了，数据量也大大减少。然后，我们把这些灰度值转换成二进制字符串，就像是把壁画的颜色信息编成了一串代码。接下来，就是把二进制字符串翻译成DNA的四进制编码。这个过程就是把一串数字密码转换成一种只有DNA才能解读的语言。我们把00翻译成A，01翻译成C，10翻译成G，11翻译成T (见图2)。这样，每两个二进制数字就能对应一个DNA的碱基。最后，我们把这些DNA的碱基序列排列起来，就得到了一幅DNA图谱，里面储存着壁画的详细信息。这就是DNA存储的第一步：数据的采集。通过这个过程，我们不仅把壁画的颜色信息保存了下来，还为它穿上了一件DNA的外衣。”小机器人小一兴致勃勃地介绍道。

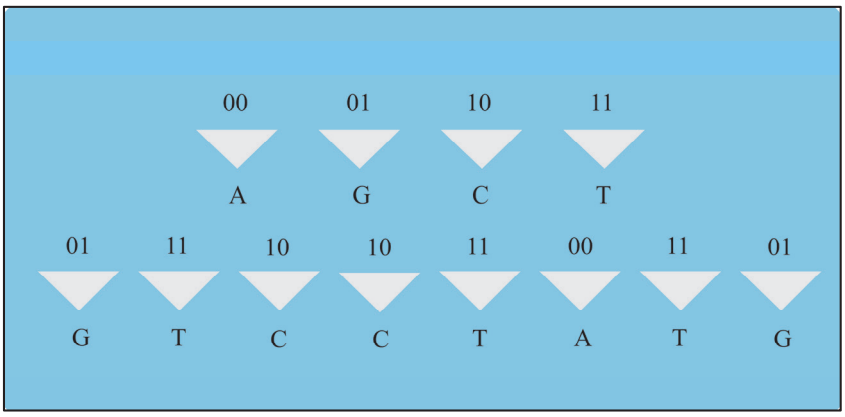

图2 二进制数据的DNA转码[2]

“在DNA存储技术的第二阶段，即DNA的定向合成中，将数字化的壁画DNA图谱转化为具体的DNA片段。这一过程是实现DNA数据存储的关键，同时也影响着整个技术的成本效益。首先，我们要合成的是寡核苷酸链，这是一些DNA小片段。以前，我们用的是柱式合成法，就像是在多孔的玻

璃柱上，一步步搭建碱基的积木。这种方法虽然降低了成本，但效率慢得像是蜗牛爬行，引物过长，而且产出的DNA纯度也不尽人意。但随着科技的进步，我们迎来了酶促合成方法的时代。这种办法有效地解决了柱式合成法产物纯度低、合成周期长的问题，同时进一步降低了合成成本，大大提高了DNA的合成效率。接下来，我们需要将这些寡核苷酸链拼接成更长的DNA片段。最初，我们用的是基于碱基互补配对原则的拼接技术，就像是用连接酶这个巧妙的针线，将DNA的片段一针一线地缝合起来。随着时间的推移，科学家们又陆续推出了一系列基于等温聚合酶、限制内切酶、外切酶的等温延伸方法，不仅降低了操作难度，也减小了生产规模对这一技术的限制，让拼接过程更加高效，更加精准。对于那些更长的DNA片段，我们还有一个秘密武器——细胞内的拼接技术。将DNA的片段放入细胞这个天然的工作室，让它们在细胞的帮助下，自然地拼接成我们所需要的长链DNA。有了这项技术，我们就能够将千年的文化遗产，以DNA的形式，完整永久地保存下来。”机器人小一优雅地微笑道。

李想听完不由地发问：“我了解到，DNA分子在自然环境中容易受到温度、湿度、紫外线和氧化等因素的影响，导致分子结构破坏和信息丢失。在50 °C和50%湿度的条件下，DNA分子最多只能保存两周[3]。你们是如何克服这些问题成功实现壁画的数据转换呢？”

机器人小一自豪地回答：“这是一个很好的问题。确实，DNA存储技术在早期面临了许多挑战，尤其是在数据的稳定性和持久性方面。为了避免DNA在合成、扩增、测序和保存过程中由于断链、重排、插入、缺失问题导致数据损失，就像是我们在玩拼图游戏时，拼图块可能会丢失或者放错位置。人们借鉴了信息领域的压缩技术，试图在技术层面减少这些错误的发生。早期，科学家们尝试了一种‘双层错误校正码’的编码解码系统。这套系统由外码和内码组成，外码就像是保镖，给DNA穿上了一件防护服，保护DNA在组装过程中不掉链子；内码则像是内部监控，确保DNA内部的信息准确无误。通常情况下，由于应用了较为精准的删除码(霍夫曼编码、喷泉码(Fountain)或者RS码)等来解决掉落链的问题，外码的读取难度远远低于内码。相应地，由于在分子内出现的错误具有多样性，研究人员为处理这种错误的内码设定了一套独特的‘数据信誉’程式，从而避免复制过程中引入过多错误片段导致数据中含有噪声，进而提高数据的纯净度和准确性。遗憾的是，由于这种系统设计复杂，且具有‘数据信誉’特征，它无法处理DNA链断裂和重排的情况，在一定程度上遏制了DNA存储技术的发展。”机器人小一继续解答，“然后，我们的科学家们——元英进教授团队受前人研究成果启发[4]，开发了一种全新的DNA链组装算法。他们利用de Bruijn图这种神奇的数学工具，将DNA序列拆分成小单元，并嵌入冗余码，就像是在DNA的每个小片段上都贴上了标签。然后，通过贪婪路径搜索，就像是在迷宫中找到了一条最佳路径，将这些小片段重新组装起来。这种方法就像是DNA的超级胶水，即使DNA链断裂或者重排，我们也能准确地读取数据。实验证明，这种算法可以从70 °C高温下经过70天时间孵育的DNA溶液中，精确地检索出6.8 MB的壁画数据，而且不需要任何特殊保护。这一技术的突破，不仅让DNA存储技术向前迈进了一大步，也为文化遗产保护等领域带来了无限可能。”

2.2 数据的存储

“现在我们使用的DNA存储技术发展到哪个阶段了？这些DNA数据是如何长期保存的？”李想对DNA存储技术的未来发展表现出浓厚的兴趣。机器人小一随即调出相关资料，开始详细解释：“目前，我们所采用的最新技术是‘细胞磁盘’，这是一种创新的DNA数据体内存储系统(图3)[5]。”

“这项技术的核心在于利用具有自我复制能力的酵母细胞作为存储信息的‘室’，类似于传统计算机硬盘中的‘磁盘块’，高达10^5个‘室’组成了‘细胞磁盘’。”机器人小一继续说，“通过将大量的‘细胞磁盘’组织起来，形成了一个高密度的存储系统。每个酵母细胞的基因组都插入了一个定制的基于CRISPR-Cas9的‘锁与钥匙’模块，这使得我们可以精确地从‘细胞磁盘’中选择性地检索、擦除或重写目标细胞‘室’中的数据。‘细胞磁盘’与DNA合成和测序相配合，构成了一个完整的基于“细胞磁盘”的DNA数据存储系统，从而实现了可随机读取和重写的数据存储。”

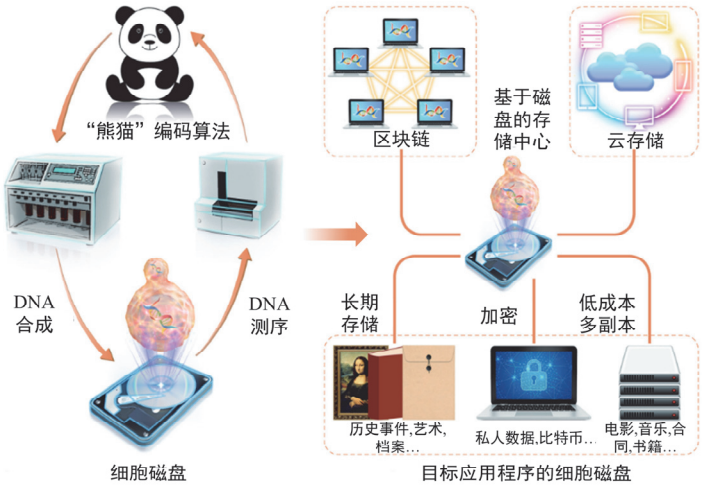

图3　细胞磁盘应用示意图[5]

李想继续追问：“这种‘细胞磁盘’技术是如何从早期的存储技术发展而来的？”

机器人小一耐心地解释道：“DNA存储技术的发展经历了多个阶段。最初，科学家们探索了DNA体外存储的方法。如为了达到较好的保存效果，通常将干燥的DNA样品保存在干燥的环境中，并通过加入添加剂使得DNA能够保存更长的时间。此外，受到古代化石的启发，人们甚至开创了利用二氧化硅封装DNA的方法，以实现更稳定的存储[6]。这些方法虽然能够实现DNA的保存，但存在稳定性和活性保持的挑战。随后，研究者们转向了体内存储技术，通过将数据编码的DNA序列整合到微生物的基因组中，利用生物体的自然复制机制，从而实现长期、高保真的数据保持。”

“早期的体内存储尝试将DNA信息存储在质粒中，1996年有研究学者在质粒中存储了小维纳斯女神的图像[7]，并成功将其转入大肠杆菌进行存储。”机器人小一解释道，“2010年，人工合成的基因组首次被转入酵母细胞，并能够复制和传代[8]。此外，通过利用重组酶技术，我们可以将DNA数据整合进染色体，实现数据的体内存储。例如，2021年的一项研究中，Chen等利用转化偶联重组技术，在酿酒酵母中构建了含有两张图片和一条视频信息的人工染色体，并成功将这些数据传递了100代，证明了该数据保存方法的有效性[9]。”

“然而，基于重组酶的DNA数据存储在特定的位置就需要一个特定的重组酶，因此存在一些限制，比如存储密度不高，与宿主基因组的接口有限，最终导致数据写入效率比较低。”机器人小一提到，“为了克服这些限制，CRISPR-Cas系统被用于将DNA数据整合进染色体的过程(见图4a)。2017年，Shipman等利用CRISPR-Cas系统在大肠杆菌中存储了一些黑白图像和一张飞驰骏马动图编码成

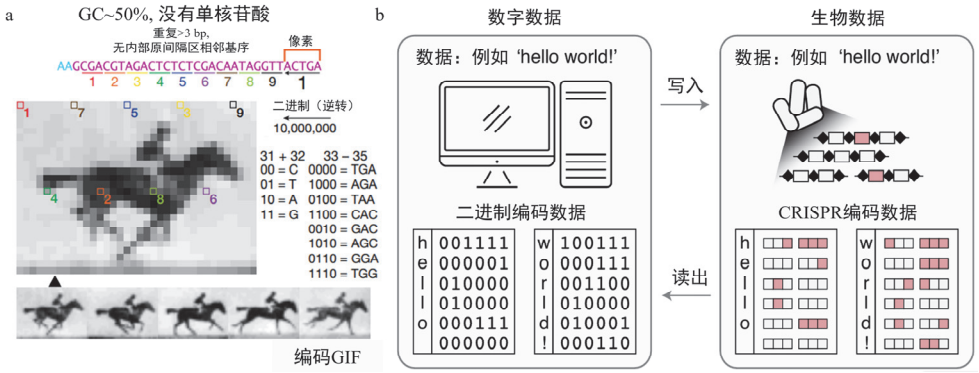

图4　基于CRISPR-Cas系统的DNA数据存储[10,11]

的大约2.6 kB大小的短片信息，并在多代繁殖后成功恢复了总体精确度为90%的数据[10]。2021年，Yim等进一步利用电化学方法调控CRISPR-Cas系统，在生物体内自动存储数据，他们将‘hello world’编码进大肠杆菌的DNA，并在80代后仍保持了90%以上的正确率(见图4b)。他们还将大肠杆菌混入土壤微生物中，对混合物进行测序，发现仍然可以恢复存储的信息[11]。”

“科学家们不断探索新的存储技术和方法。”机器人小一总结道，“从最初的体外存储，到利用微生物基因组的体内存储，再到现在的‘细胞磁盘块’技术，每一步都代表了DNA存储技术在稳定性、密度和可访问性方面的进步。”

3 DNA存储的挑战

李想对DNA存储技术的潜力表现出了浓厚的兴趣：“尽管传统的磁性/光学/固态存储介质的写入和读取成本低且易于使用，但由于其寿命较短，密度低，储存能力会受到密度的限制，相应的吞吐量也比较低。而且这些传统介质在长期存储中会遇到消磁、磨损等问题，需要频繁更新和维护。与传统存储介质相比，DNA存储在寿命、存储密度和耐久性等方面具有显著优势。我想知道，DNA数据的长期稳定存储中会存在哪些需要克服的问题呢？”

机器人小一详细解释道：“确实，DNA存储技术在多个方面优于传统存储方式。它不仅存储密度高，而且能够在适宜的条件下保持极长的寿命，理论上可达数十万年。这使得DNA存储成为长期保存数据的理想选择，正如您所见，千年后的敦煌壁画依然能够通过DNA存储技术得以复原。”

“但是，”小一继续说道，“即便DNA存储技术的优势显而易见，它也依旧存在着一系列亟待解决的问题。首先，DNA数据存储系统依赖于测序技术，但这一过程可能会对DNA造成损伤。DNA存储需要干燥和低温的环境，而PCR扩增，作为测序的一个必要步骤，却需要加热，这与DNA的最佳保存条件相矛盾。此外，反复的测序可能会对DNA造成不可逆的伤害，而添加物的增加也可能干扰测序结果。这些因素都可能导致长期存储和多次读取后，DNA存储的数据出现严重缺损，因此，提高存储稳定性是一个亟待解决的问题。其次，DNA存储技术可能引发信息安全问题。如果DNA存储技术被用于军事等领域，人体可能成为携带DNA数据的‘云硬盘’，这虽然方便，但也带来了安全隐患。DNA的载体丰富且不易检测，这使得信息安全问题变得更加复杂。未来，随着DNA存储技术的普及，每个人都可能成为超大容量的DNA数据载体，这可能引发法律和伦理问题，比如数据所有权、隐私保护和生物安全等。”

李想继续说道：“考虑到这些挑战，你们是否已经在考虑如何拓展DNA存储技术的应用范围？”

机器人小一认真地回答：“您的洞察非常深刻。我们确实在探索DNA存储技术的新应用。目前，细胞或细菌作为存储介质，虽然尺寸较大，但存储信息的总量相对较低，且存在遗传不稳定性带来的数据丢失和错误风险[9]。未来的研究可以集中在提高单个细胞或细菌的数据存储容量，以及建立一个能够覆盖读取、写入、增加、删除、修改和查询的一体化体内DNA存储系统。此外，酵母基因组合成计划已经展示了从kbp到Mbp级染色体的从头合成和精简能力。未来的技术可以考虑如何将数据信息存储到这些合成染色体中，以提高存储容量。同时，通过进一步优化CRISPR-Cas系统，我们可以实现大片段基因组在细胞内的任意增加、删除和修改，以及整个细胞层面存储信息的增删和修改。”

这时，李想提出了一个创新的想法：“我还有一个建议，考虑到DNA在多次复制中可能出现错误，我们是否可以在体内DNA的编码设计中加入纠错码，以降低DNA信息存储在长期传代中的错误率，从而提高数据的稳定性。此外，还可以利用微生物孢子等特殊形态存储数据，也许能更好地保护DNA免受环境影响，确保体内存储信息的稳定性。”随着讨论的深入，李想和机器人小一共同探讨了DNA存储技术的未来发展方向。

美好的时光总是过得飞快，转眼就到了告别的时刻，李想送上了她真挚的祝愿：“今天的学习

让我切实感受到了未来科技的斐然成就。作为科技道路上的同行者，我衷心希望你们能继续在DNA存储领域取得突破，为人类带来更大的福祉。”机器人小一坚定地点了点头。一阵熟悉的铃声突然将李想从梦中唤醒。她睁开眼，发现自己仍躺在温暖的床上，原来那场跨越千年的对话只是一场梦。然而，这场梦却给她留下了深刻的启示和无尽的鼓舞。她坚信，只要持续努力、不断创新，并且与团队紧密合作，DNA存储这一曾经遥不可及的梦想，终将在不久的将来成为现实。而她，正与无数科技工作者一起，共同书写着人类科技进步的辉煌篇章。

参 考 文 献

[1] Song, L.; Geng, F.; Gong, Z. Y.; Chen, X.; Tang, J.; Gong, C.; Zhou, L.; Xia, R.; Han, M. Z.; Xu, J. Y.; *et al*. *Nat. Commun.* **2022,** *13*, 5361.

[2] DNA是否可以存储除了遗传信息以外的复杂信息? 目前有哪些技术难点? [2022-05-19]. https://www.zhihu.com/question/532951888/answer/2492305770

[3] Bruskov, V. I.; Malakhova, L. V.; Masalimov, Z. K.; Chernikov, A. V. *Nucleic Acids Res*. **2002,** *30*, 1354.

[4] Zerbino, D. R.; Birney, E. *Genome Res*. **2008,** *18*, 821.

[5] Hou, Z. H.; Qiang, W.; Wang, X. X.; Chen, X. X.; Hu, X.; Han, X. Y.; Shen, W. L.; Zhang, B.; Xing, P.; Shi, W. P.; *et al. Adv. Sci.* **2024,** *11*, 2305921.

[6] Kohll, A. X.; Antkowiak, P. L.; Chen, W. D.; Nguyen, B. H.; Stark, W. J.; Ceze, L.; Strauss, K.; Grass, R. N. *Chem. Commun.* **2020,** *56*, 3613.

[7] Davis, J. *Art J.* **1996,** *55*, 70.

[8] Gibson, D. G.; Glass, J. I.; Lartigue, C.; Noskov, V. N.; Chuang, R. Y.; Algire, M. A.; Benders, G. A.; Montague, M. G.; Ma, L.; Moodie, M. M.; *et al*. *Science* **2010,** *329*, 52.

[9] Chen, W. G.; Han, M. Z.; Zhou, J. T.; Ge, Q.; Wang, P. P.; Zhang, X. C.; Zhu, A. Y.; Song, L. F.; Yuan, Y. J. *Natl. Sci. Rev.* **2021**, *8*, nwab028.

[10] Shipman, S. L.; Nivala, J. N.; Macklis, J. D.; Church, G. M. *Nature* **2017,** *547*, 345.

[11] Yim, S. S.; McBee, R. M.; Song, A. M.; Huang, Y. M.; Sheth, R. U.; Wang, H. H. *Nat. Chem. Biol.* **2021,** *17*, 246.