文章编号:1671-4229(2024)03-0001-07

深度学习在 DNA 存储读段重建的应用

姚翔宇[1]，刘希晨[2]，咎乡镇[1]，许　鹏[1]，刘文斌[1]*
(1. 广州大学 计算科技研究院，广东 广州　510006；2. 广州商学院 信息技术与工程学院，广东 广州　511363)

摘　要：DNA 存储技术是一项着眼于未来的具有划时代意义存储技术，它将数字信息编码为核苷酸序列，然后通过化学合成将序列写入 DNA 分子，最后通过 DNA 测序技术读取信息。相较于电子存储技术，其在信息密度、数据安全性以及保存年限等方面具有极大的优势。然而在 DNA 存储技术的数据读取端，测序读段存在着大量碱基替换、插入和删除错误，因此需进行读段重建来恢复原始数据。读段重建方法要求高成功率以及一定的时效性，以实现文件可靠存取并提高 DNA 存储技术的读写效率。文章介绍了现有的基于深度学习的读段重建模型，通过对模型架构、重建理念以及纠错能力等方面的比较指出了目前研究的局限性，并展望了未来深度学习在读段重建中可能的研究方向。

关键词：DNA 存储技术；碱基错误；读段重建；深度学习
中图分类号：TN911　　**文献标志码**：A

The application of deep learning in reads reconstruction for DNA storage

YAO Xiang-yu[1], *LIU Xi-chen*[2], *ZAN Xiang-zhen*[1], *XU Peng*[1], *LIU Wen-bin*[1]*
(1. Institution of Computational Science and Technology, Guangzhou University, Guangzhou 510006, China;
2. School of Information Technology and Engineering, Guangzhou College of Commerce, Guangzhou 511363, China)

Abstract: DNA storage technology is a promising new type of data storage technology, it encodes digital information into nucleotide sequences, then writes the sequences into DNA molecules through chemical synthesis, and finally reads the information through DNA sequencing technology. Compared to electronic storage technology, it has great advantages in information density, data security, and lifespan. However, it outputs a large number of sequencing reads with base errors (substitution, insertion and deletion), therefore, to restore the original data from erroneous reads, reads reconstruction is usually performed. Accurate read reconstruction requires high success rate and time efficiency to achieve reliable file access and accelerate the reading and writing process of DNA storage technology. This paper introduces the existing deep learning-based reads reconstruction models. By comparing architecture, basic concepts and error correction abilities, we point out the limitations of these methods and discuss the prospects for future research directions.

Key words: DNA storage; base error; reads reconstruction; deep learning

收稿日期：2023－12－11；**修回日期**：2024－04－09
基金项目：国家自然科学基金资助项目(62072128，62002079，62102104)
作者简介：姚翔宇(1994—)，男，博士研究生. E-mail：yxy@e.gzhu.edu.cn
* 通信作者. E-mail：wbliu6910@gzhu.edu.cn
引文格式：姚翔宇，刘希晨，咎乡镇，等. 深度学习在 DNA 存储读段重建的应用[J]. 广州大学学报(自然科学版)，2024，23(3)：1-7.

随着全球数据量呈指数级增长,传统存储介质如光盘、硬盘等存储密度已接近极限,能源消耗巨大,无法满足未来海量数据的存储需求且不符合信息技术的可持续发展要求[1-2]。近年来,研究者们发现利用人工合成的脱氧核糖核苷酸(DNA)作为存储介质,具有存储密度高、存储时间长、易获取且免维护的优点[3-5]。此外,1克DNA能够存储大约2拍字节,相当于大约300万张CD。目前,国内外关于DNA存储技术的研究已取得初步的进展。2021年微软开发出首个纳米级DNA存储器,能够在每个平方厘米的区域上,同时合成2 650条碱基序列。同年12月,东南大学团队成功将该校校训"止于至善"存入一段DNA序列,相关成果已发表在Science Advance期刊上[6]。2022年10月,天津大学合成生物学团队将10幅精选敦煌壁画存入DNA分子中,并通过加速老化等实验,发现这些壁画信息在常温下可保存千年,在9.4 ℃下可保存两万年。同年,国家"十四五"规划将DNA存储列为与新一代移动通信技术、量子信息、第三代半导体等并列的新兴技术。

DNA存储技术主要包括DNA编码、DNA合成、PCR扩增和DNA测序,其中,DNA合成和测序分别对应数据的写入和读取[7]。数据读取端的输出为若干存在碱基替换、插入和删除错误的读段,错误主要来自于DNA合成以及测序技术。有研究显示,第三代高通量测序技术的错误率高达10%~15%,且88%以上的测序读段长度不正确[8]。因此,读段重建是利用DNA存储技术实现文件可靠存取的关键。

针对DNA存储信道的特点,研究者们开发了多种纠错码以及纠错算法用于读段重建。纠错码基于冗余编码思想,即在数据中添加冗余信息,以便在数据传输过程中检测和纠正错误。DNA存储中,纠错码的实现方式主要有RS码[9-10]、BCH码[11]、汉明码[12]以及LDPC码[13-14]。目前,纠错码只适用于错误率较低的存储环境,当错误率高于5%时,其纠错能力无法满足DNA存储技术的要求。此外,应用纠错码会明显降低DNA分子的信息存储密度。因此,为实现高错误率环境下的文件可靠存取,研究者们陆续开发出多种纠错算法,如HEDGES算法[15]、德布莱因图算法[16]、多序列比对算法[17]以及调制编解码算法[18]。纠错算法的纠错能力高于纠错码,一般可纠正5%~15%的碱基错误,其中,调制编解码算法的纠错能力达到40%。纠错算法虽然能实现文件的可靠存取,但是其往往编码复杂,且解码需要大量时间和空间开销,严重影响文件的读写效率。

近年来,研究者们发现DNA存储的错误分布并非随机,而是呈现出位置特异性以及碱基特异性。2021年,Sabary等[19]开发出SQLOC,可对合成DNA分子中的错误进行建模。同年,Chaykin等[20]开发出DNA存储模拟器,可在特定模式下给DNA序列注入错误。此外,DNA存储要求将二进制信息编码为DNA序列,而相应的编码方式使测序读段具有了结构和层次。因此,测序读段存在两种模式,即碱基错误模式以及序列结构层次模式。

随着人工智能技术的不断发展,深度学习已逐渐被应用到各种不同的任务和领域中。深度学习通过学习样本数据的内在规律和表示层次来获取特征信息,从而完成特定的任务。在DNA存储中,可通过学习测序读段的碱基错误模式或结构层次模式来将测序读段映射为正确读段。相较于纠错码以及纠错算法,深度学习方法将极大提高DNA存储技术的读写效率。目前,深度学习在读段重建中的研究尚处于初级阶段,纠错能力较弱,约1%。未来,DNA存储技术将朝着快速高效的方向发展,深度学习也必将应用于DNA存储的各个方面。随着模型不断优化以及新模型的提出,深度学习与DNA存储技术的高度融合将进一步推动DNA分子成为新一代绿色存储介质。

1 读段重建问题定义

DNA存储技术通常会对携带信息的DNA分子进行PCR扩增,因此,每条测序读段都有多个拷贝。根据是否使用多拷贝信息可将读段重建分为单读段重建和多读段重建,如图1所示。DNA存储中读段重建问题也可数学形式化如下。

DNA信道首先将输入的二进制信息m编码

成由 DNA 码字所构成的序列,然后将序列分割成 n 个长度为 l 的子序列 $x_1,x_2,\cdots x_n$,其中,$x_i \in \sum^l$,$\sum = \{A,C,G,T\}$。每个子序列通过 DNA 的插入 - 删除 - 替换信道后,会得到 $t>0$ 个独立拷贝 $y_1,y_2,\cdots y_t$。最后,解码这些拷贝序列得到二进制信息 $\hat{m}$。

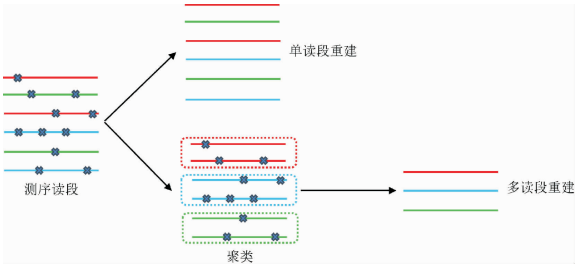

图 1　读段重建示意图

Fig. 1　Reads reconstruction sketch map

读段重建算法通常进行一个序列到序列的映射,可描述为 $R:(\sum^*)^t \rightarrow \sum^*$,算法的输入为拷贝序列 $y_1,y_2,\cdots y_t$,输出为 $\hat{x}$。读段重建问题就是最小化原始序列 x 与 $\hat{x}$ 之间的编辑距离或汉明距离[21]。

2　相关模型

测序仪一般通过碱基判别器将其产生的电信号翻译成碱基,从而得到测序读段。碱基判别器的准确度会影响测序读段的质量,因此,设计高精度的判别器将有助于读段重建。由于高通量测修仪要求具有时效性,碱基判别器一般基于深度学习技术进行开发。本节首先介绍目前精度最高的碱基判别器 Fast-Bonito[22],然后介绍已有的 3 种不同的读段重建模型。

2.1　碱基判别器

Fast-Bonita 采用语音识别中 QuartzNet[23]的网络结构,如图 2 所示。其由多个瓶颈卷积模块和 CTC(连续时序分类)解码器组成。瓶颈卷积可减少模型的参数量,结合残差连接能够搭建更深的网络结构,从而使网络具有更好的拟合效果,最后通过 CTC 解码器得到预测结果。

目前碱基判别器的开发仍处于初级阶段,网络结构大多借鉴语音识别模型。然而不同于语音数据,测序仪所产生的电信号不具备连续关系,未来可结合测序信号的特点来开发模型。其次,测序技术正朝着高通量方向发展,已有的碱基判别器无法同时保证时间和精度,如何权衡时间及精度是未来研究的重点。

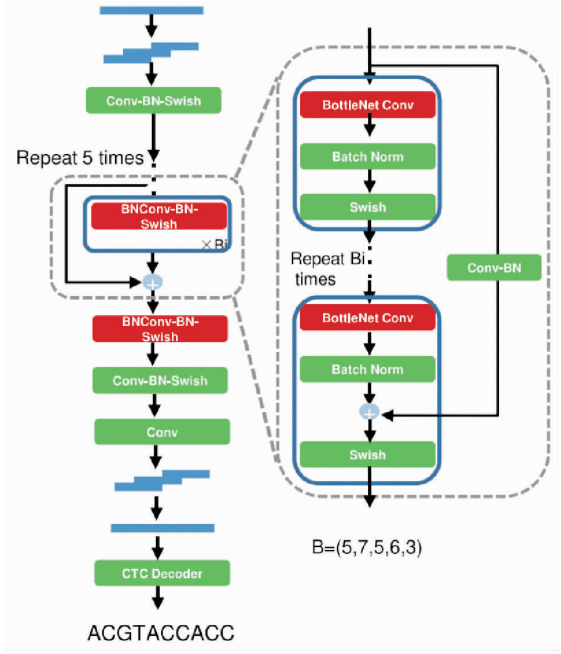

图 2　Fast-Bonita 网络结构

Fig. 2　Architecture of Fast-Bonita

碱基判别器虽无法纠正 DNA 存储中的错误,但可通过提高电信号转化的准确度来减少 DNA 存储的错误,从而提高读段重建成功率。此外,碱基判别器的网络结构以及数据处理方式值得读段重建模型进行参考借鉴。

2.2　读段重建模型

2.2.1　DNAformer

DNAformer 通过学习碱基错误模式进行读段重建,网络结构如图 3 所示[24]。基本思想是用 one-hot 编码将测序读段转变为矩阵,再将同源读段(相同原始序列的拷贝)进行矩阵相加。相加后,值较大的位置碱基错误较少,反之错误较多。然后用多个卷积核对矩阵进行深度可分离卷积操作,方便模型捕捉到由插入/删除错误引起的不同程度的碱基位移,并将卷积结果拼接起来输入编码器提取特征。最后通过解码器得到预测结果。

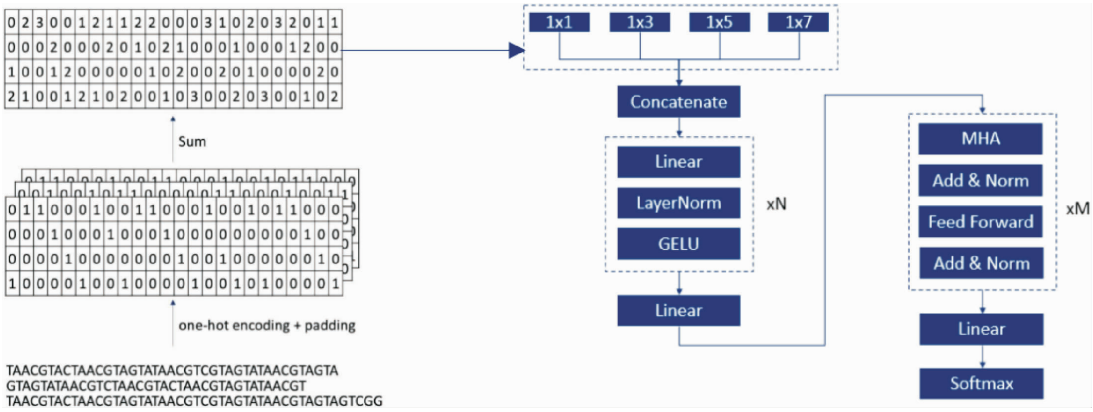

图3 DNAformer网络结构

Fig. 3 Architecture of DNAformer

DNAformer的设计是基于对DNA存储中主要错误类型为碱基替换而非碱基插入/删除的认识。因此,one-hot编码后,可根据同源读段的相加结果来判断各个位置发生错误的概率并进行纠正。然而,插入/删除错误会使碱基发生位移,当错误率较高(>2%)时,该模型的预测精度会显著降低。

2.2.2 RRCC-DNN

RRCC-DNN同样致力于学习读段的碱基错误模式,网络结构如图4所示,包括3部分:注意力模块、编码器以及解码器[25]。其基本思想与DNAformer十分相似,即利用同源读段相加来判断各个位置的错误并纠正。

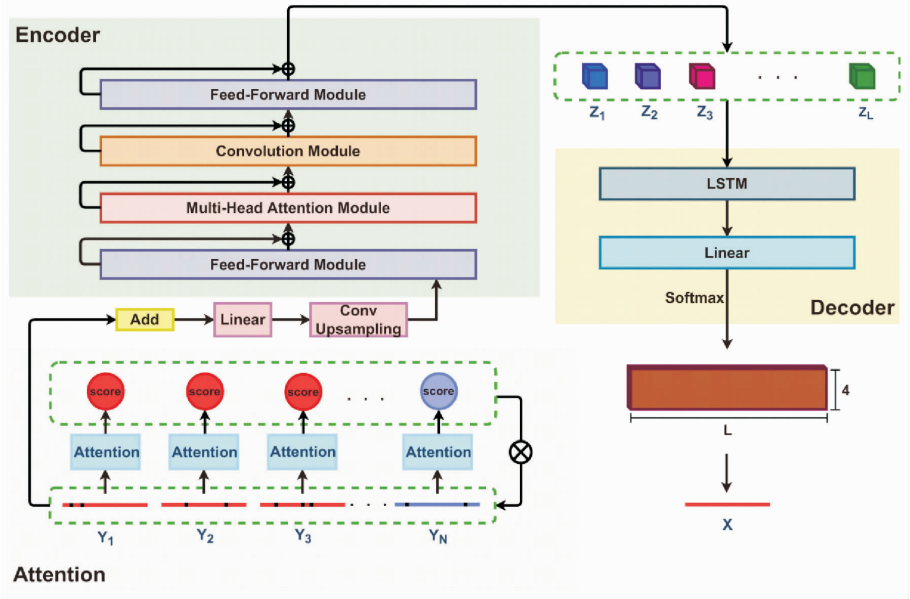

图4 RRCC-DNN网络结构

Fig. 4 Architecture of RRCC-DNN

不同于DNAformer,RRCC-DNN是根据读段的注意力权重进行相加,即读段矩阵乘以其注意力分数再相加,过程如图5所示。首先用one-hot编码将读段转变为矩阵,再用两个连续的一维卷积将其特征通道数降为1,不同大小的两个卷积核使模型可捕捉到由插入/删除错误引起的不同程度的碱基位移。最后利用公式(1)和(2)计算读段的注意力分数。

$$e_i = v^{\mathrm{T}} f(W\tilde{y}_i + b) + k, \tag{1}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{i=1}^{N} e_i}, \tag{2}$$

其中,$\tilde{y}_i$为卷积后得到的向量,参数W和b用于将

向量映射到低维空间以减少网络参数,f 为激活函数,v^T 和 k 为线性层参数,e_i 为读段的注意力分数,α_i 为标准化后的注意力分数。

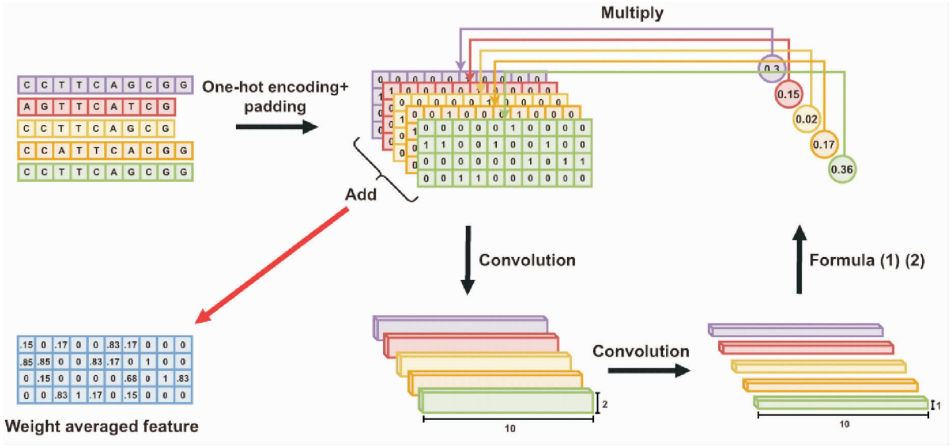

图 5　RRCC-DNN 注意力模块

Fig. 5　Attention module of RRCC-DNN

RRCC-DNN 利用注意力机制为错误率较低的读段赋予较高的权重,而错误率高的读段或者错误聚类的非同源读段则赋予较低的权重,从而聚焦于对读段重建更为关键的信息,并降低对其他信息的关注度,甚至过滤掉无关信息,以此来提高模型的预测精度。但插入/删除错误引起的碱基位移仍会对其预测精度造成巨大影响。

2.2.3　SRR

与前面两种模型不同,SRR[21] 利用 Transformer[26] 学习序列的结构层次来进行读段重建,流程如图 6 所示。基本思想是首先将文件中每字节信息编码成长度为 4 的 DNA 码,并且将该文件所使用的 DNA 码也一并存储。测序后,将读段分为 4 类,即 LL(大于标准长度)、SL(小于标准长度)、CLBC(长度正确但存在错误 DNA 码)和 CLGC(长度正确且 DNA 码正确),其中,CLGC 为默认的正确读段并在训练中作为标签使用。其次,利用 DNA 存储模拟器给 CLGC 注入错误,从而又产生 4 种读段并丢弃第四类读段,其他 3 种读段作为训练集训练 3 个模型,即 Model LL,Model SL 和 Model CLBC,用于 3 种读段的重建。最后,利用训练好的 3 个模型来重建包含错误的 3 种测序读段。

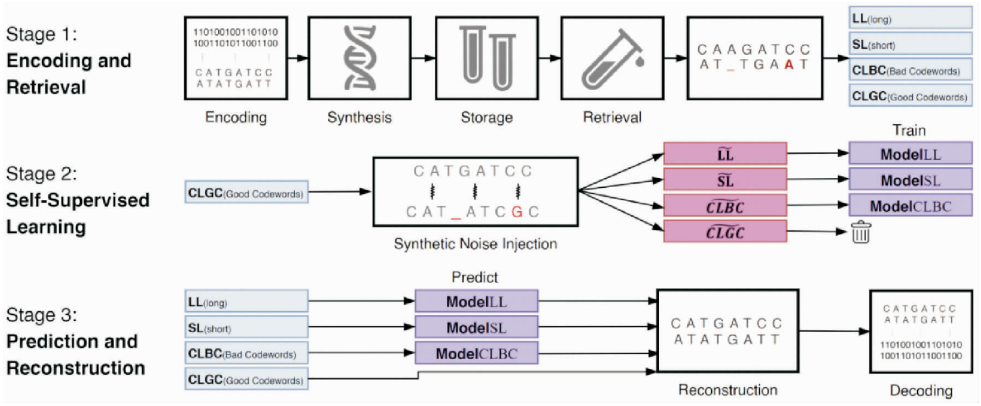

图 6　SRR 流程图

Fig. 6　Workflow of SRR

SRR 属于自监督学习模型,它利用测序读段本身的结构和属性,自动生成标签,从而进行模型的训练,学习到对读段重建有价值的表征。同时,SRR 模型对单个读段进行重建,无需使用同源读段的信息,因此,也无需对同源读段进行聚类,从而进一步提高了 DNA 存储的解码效率。然而文中所使

用的数据集错误率较低(<1%),在高错误率的存储环境下,该模型的成功率还有待验证。

3 模型比较

表1为上述3种不同的读段重建模型各自的网络架构、学习方式、重建方式、对DNA存储的错误容忍度以及纠错成功率的比较。从表1可见,3种读段重建方法都采用Transformer及其改进版本。其次,监督学习方式通过学习碱基的错误模式来进行读段重建,而自监督学习通过学习序列的结构层次来进行读段重建。此外,已有的这些方法主要依靠卷积或训练插/删特定模型来处理插入和删除错误,且总体纠错能力较弱。

表1 模型比较

Table 1 Comparison of modules %

模型	网络架构	学习方式	重建方式	插/删错误	错误容忍度	成功率
DNAformer	Transformer	监督	碱基错误	卷积	1.0	99.95
RRCC-DNN	Bert + LSTM	监督	碱基错误	卷积	1.0	99.86
SRR	Transformer	自监督	序列结构层次	多模型	0.3	92.50

由此可见,目前基于深度学习的读段重建研究存在局限性。①模型选择。目前的研究主要利用Transformer进行读段重建,然而在DNA存储中,测序读段是信息的载体,与语音数据、文本数据、视频数据等数据不同,碱基之间的连续关系较弱,Transformer很难对其预测。②错误纠正能力较差。模型的纠错能力主要受限于读段中的插/删错误,其处理此两类错误的策略主要分为两种:卷积和训练插/删特定模型。卷积层善于提取局部特征,而插/删错误则会影响后续全部碱基。此外,插/删特定模型根据测序读段的长度划分3个训练集,分别训练3个网络结构相同的模型,以此来提高模型处理3种错误的能力。这种简单的划分方式可能会造成训练数据分布混乱,最后导致模型对3种错误都欠拟合。

4 总结与展望

深度学习技术在DNA存储读段重建的主要优势是读/写高效性以及无需冗余信息。传统读段重建方法主要通过在序列中添加冗余信息用于检错和纠错,这使得编码和解码过程需要大量时间和空间开销,并且显著增加了DNA合成以及测序成本。目前,阻碍DNA存储技术商业化最主要的问题就是高成本,存储200 MB数据大约耗资80万美元。因此,深度学习的应用有望解决DNA存储技术在发展中所面临的主要困境,并推动其走向商业化道路。

然而目前深度学习在读段重建的研究仍处于初级阶段,主要是利用Transformer进行重建,并且纠错能力较弱,约1%,无法满足DNA存储技术的实际需求。DNA存储需要满足一定的生化约束条件(均聚物、GC含量等),其碱基之前具备一定关系,模型设计时,可以考虑这一点。此外,DNA存储的错误多分布于均聚物中,而且A、C两种碱基占主要部分,因此,在编码阶段应尽量避免均聚物以及合理控制A、C两种碱基的含量。未来可尝试建立测序读段的图表示模型,DNA存储的碱基错误可视作图像的噪声,并利用图像处理模型来进行读段重建。此外,对于多读段重建方法,聚类精度会影响模型的预测精度,利用深度学习技术开发高效的DNA序列聚类算法将有助于提高读段重建的成功率。最后,随着DNA合成以及测序技术实现高通量,DNA存储技术对读写效率以及纠错能力的要求也愈发提高,因此,如何充分发挥深度学习模型的强大能力,来推动DNA存储技术的应用将成为未来的研究热点。

参考文献:

[1] 姚翔宇,咎乡镇,谢恋,等. DNA存储技术的复杂度概述[J]. 广州大学学报(自然科学版),2021,20(1):12-22.
[2] 毛秀海, 李凡, 左小磊. DNA数据存储[J]. 电子与信息学报, 2020, 42(6): 1303-1312.
[3] 咎乡镇, 姚翔宇, 许鹏, 等. DNA存储文件系统研究进展[J]. 电子与信息学报, 2023, 45(6): 1911-1920.

[4] 昝乡镇, 姚翔宇, 许鹏, 等. 一种高效的前向纠错码桶分配 DNA 存储解码方法[J]. 电子与信息学报, 2022, 44(10): 3650-3656.
[5] 昝乡镇,姚翔宇,许鹏,等. DNA 存储中的纠错方法综述[J]. 广州大学学报(自然科学版), 2021,20(2):13-22.
[6] Xu C T, Ma B, Gao Z L, et al. Electrochemical DNA synthesis and sequencing on a single electrode with scalability for integrated data storage[J]. Science Advance,2021(7):46-58.
[7] Church G M, Gao Y, Kosuri S. Next-generation digital information storage in DNA[J]. Science, 2012, 337(6102):1628-1641.
[8] Chen Y J, Takahashi C N, Organick L, et al. Quantifying molecular bias in DNA data storage[J]. Nature Communication, 2020(11):3264-3288.
[9] Meiser L C, Antkowiak P L, Koch J, et al. Reading and writing digital data in DNA[J]. Nature Protocols, 2019,15(1): 86-101.
[10] 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.
[11] Blawat M, Gaedke K, Huetter I, et al. Forward error correction for DNA data storage[J]. Procedia Computer Science, 2016, 80(1): 1011-1022.
[12] Takahashi C N, Nguyen B H, Strauss K, et al. Demonstration of end-to-end automation of DNA data storage[J]. Scientific Reports, 2019, 9(1): 4998-5009.
[13] Deng L, Wang Y X, Noor-A-Rahim M, et al. Optimized code design for constrained DNA data storage with asymmetric errors[J]. IEEE Access, 2019, 7(1): 84107-84121.
[14] Lu X Z, Jeong J, Kim J W, et al. Error rate-based log-likelihood ratio processing for low-density parity-check codes in DNA storage[J]. IEEE Access, 2020, 8(1): 162892-162902.
[15] Press W H, Hawkins J A, Jones S K, et al. HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints[J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(31): 18489-18496.
[16] Song L, Geng F, Gong Z Y, et al. Robust data storage in DNA by De Bruijn graph-based denovo strand assembly[J]. Nature Communications, 2022, 13(1): 5361-5382.
[17] Xie R, Zan X, Chu L, et al. Study of the error correction capability of multiple sequence alignment algorithm (MAFFT) in DNA storage[J]. BMC Bioinformatics, 2023, 24(1): 12859-12868.
[18] Zan X, Xie R, Yao X, et al. A robust and efficient DNA storage architecture based on modulation encoding and decoding [J]. Journal of Chemical Information and Modeling, 2023, 63(12):3967-3976.
[19] Sabary O, Orlev Y, Shafir R, et al. SOLQC: Synthetic oligo library quality control tool[J]. Bioinformatics, 2021, 37(5): 720-722.
[20] Chaykin G, Furman N, Sabary O, et al. DNA storage simulator[EB/OL]. (2021-07-31)[2023-11-18]. https://github.com/gadihh/DNASimulator.
[21] Nahum Y, Ben-Tolila E, Anavy L. Single-read reconstruction for DNA data storage using transformers[EB/OL]. (2021-10-10)[2023-11-18]. https://arxiv.org/abs/2109.05478.
[22] Xu Z M, Mai Y T, Liu D H, et al. Fast-Bonito: A faster deep learning based basecaller for nanopore sequencing[J]. Artificial Intelligence in the Life Sciences, 2021, 1(1):100011-100016.
[23] Kriman S, Beliaev S, Ginsburg B, et al. IEEE International Conference on Acoustics, 2020-03-12[C]. Barcelona: Institute of Electrical and Electronics Engineers, 2020.
[24] Bar-Lev D, Orr I, Sabary O, et al. Deep DNA storage: Scalable and robust DNA storage via coding theory and deep learning[EB/OL]. (2021-08-31)[2023-11-18]. http://doi.org/10.48550/arXiv.2109.00031.
[25] Qin Y, Zhu F, Xi B. Robust multi-read reconstruction from contaminated clusters using deep neural network for DNA storage[EB/OL]. (2022-10-20)[2023-11-18]. http://doi.org/10.48550/ arXiv.2210.11106.
[26] Ashish V, Shazeer N, Parmar N, et al. Conference on Neural Information Processing Systems, 2017-05-13[C]. Long Beach: MIT Press, 2017.

【责任编辑: 陈　钢】