一种高效的前向纠错码桶分配DNA存储解码方法

昝乡镇[①] 姚翔宇[①] 许 鹏[①] 陈智华[①] 石晓龙[①] 李树栋[②] 刘文斌*[①]

[①](广州大学计算科技研究院 广州 510006)

[②](广州大学网络空间技术先进研究院 广州 510006)

摘 要：与传统存储方式相比，脱氧核糖核酸(DNA)存储的难点是测序序列中的插入和删除错误给信息解码过程带来了巨大挑战。针对具有1位纠错能力的前向纠错编码DNA存储，该文提出一种桶式分配策略提高解码的精度和效率。首先，搜索每个分组中所有测序读长的可识别DNA码，根据1位纠错能力确定其对应的合法编码；其次，根据每个可识别DNA码在测序读长的位置确定相应编码的最佳编码位置(即桶)；最后，按照众数投票确定每个桶中的最终编码。仿真结果表明在0.10和0.05错误率条件下，平均解码准确率在20X测序深度时可达94%以上；在0.15错误率条件下，平均解码准确率在60X测序深度时可达90%以上。

关键词：存储解码方法；脱氧核糖核酸存储；插入错误；删除错误；替换错误

中图分类号：TN918.3 **文献标识码**：A **文章编号**：1009-5896(2022)10-3650-07

DOI: 10.11999/JEIT210697

An Efficient Bueket-allocation Decoding Method Based on Forward Error Correction Codes for Deoxyribo Nucleicecid Storage

ZAN Xiangzhen[①] YAO Xiangyu[①] XU Peng[①] CHEN Zhihua[①]

SHI Xiaolong[①] LI Shudong[②] LIU Wenbin[①]

[①](*Institution of Computational Science and Technology, Guangzhou University, Guangzhou* 510006, *China*)

[②](*Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou* 510006, *China*)

Abstract: Compared with traditional storage, the difficulty of DeoxyriboNucleic Acid (DNA) data storage is that insertion and deletion errors in sequenced reads pose a great challenge to data recovery. For forward error-correcting coded DNA storage with one-base error-correcting capability, a bucket allocation strategy is proposed to improve the decoding accuracy and efficiency. Firstly, all identifiable DNA codes of reads in each cluster are searched and the corresponding valid codes according to the one-base error-correcting capability are determined; Then, for each identifiable DNA code, appropriate coding position (i.e. bucket) according is allocated to its position in a read; Finally, the consensus code for each bucket is determined using majority voting strategy. Simulation results show that the proposed method can correct more than 94% errors at the coverage of 20X when error rate is 5% or 10%, and correct more than 90% errors at the coverage of 60X when error rate is 15%.

Key words: Storage decoding method; DeoxyriboNucleic Acid (DNA) storage; Insertions error; Deletions error; Substitutions error

1 引言

云计算、大数据等技术的发展，人类存储数据的需求呈现出指数级增长的趋势。据国际数据公司预测[1]，2025年全球数据总量预计达到175 ZB。传统存储介质技术在满足未来数据存储需求方面逐渐暴露出一系列缺点[2, 3]，比如有效存储时间短、数据易损坏以及维护成本高……与此同时，携带有遗传信息的脱氧核糖核酸(DeoxyriboNucleic Acid, DNA)，因其具有超高的存储密度、低维护成本以及数据保存持久等特点[4–6]，有望是一种极具潜力的存储介质，解决海量数据存储面临的困境。

DNA存储主要涉及合成、聚合酶链反应 (Polymerase Chain Reaction, PCR)扩增、测序等生物过程。由于技术局限，这些过程会导致DNA存储发生一系列复杂组合错误，从而给数据的可靠恢复带来了挑战[7,8]。据估计，二代测序技术和阵列合

收稿日期：2021-07-13；改回日期：2021-09-30；网络出版：2021-10-26

*通信作者： 刘文斌 wbliu6910@gzhu.edu.cn

基金项目：国家自然科学基金(62072128, 61876047, 62002079)

Foundation Items: The National Natural Science Foundation of China (62072128, 61876047, 62002079)

成技术的错误发生概率为1%～2%[9]，三代测序技术将达10%～15%[10]。与传统数字存储相比，DNA存储序列的碱基错误不仅有替换错误，还包括插入错误。这些错误的复杂交织远远超出了传统纠错码(Error Correcting Codes, ECC)[11–14]的纠错能力。同时，碱基插入还会导致DNA序列的长度与标准长度不一致。有研究表明，三代纳米孔测序技术会产生大约88%的非标准长度序列[15]。以往利用RS纠错码[16,17]或喷泉码[18,19]的DNA存储方法中，通常会舍弃掉这些测序读段，从而导致大量浪费，并增加了测序过程的时间和费用。因此，研究面向DNA碱基插入、删除和替换组合错误的高效信息恢复方法是未来DNA存储亟待解决的一个重要问题。

Blawat等人[13]对每个字节设计了两套DNA编码，然后交替使用第1类和第2类DNA编码来编码原始信息。当插入错误发生时，将会打破两类DNA编码交替出现的规律，由此可以定位发生插入/删除的位置。Press等人[20]通过哈希操作将二进制序列的每一比特，都与其附近比特、序列索引产生强关联后，再进行编码。在解码的时候，通过贪心穷举搜索策略评估每一比特满足强关联的程度，进而完成碱基插入、删除和替换等错误的纠正。但是该方案需要较高的冗余度，且解码过程复杂。Xue等人[21]通过添加一些冗余位，将二进制序列拆分成两个子串，使其满足莱文斯坦码(Levenshtein code)的数据形式，然后利用莱文斯坦码可以纠正1 bit插入/替换的性质，完成DNA序列中1位碱基错误的纠正。天津大学Song等人[22]通过构建多拷贝读段的德布莱英图(de Bruijn graph)，将一致性序列的确定问题转换为图中的最大权路径搜索问题，从而过滤掉插入、删除、替换导致的低频子串路径。本研究团队[23]最近提出一种基于前向纠错码的英文文本3层纠错方法，基本思想是通过前向纠错码对DNA读长进行初步纠错，然后对转化的字符序列进行多序列比对纠错，最后通过单词拼写进一步纠正错误。该方法在错误率0.05情况下，20X测序深度纠错准确率达90%以上。但在错误率0.10情况下，60X测序深度仅达到64%。因此，难以适应高错误率的情况。

本文在前向纠错码的基础上，通过在序列编码中采用“索引+CRC哈希+索引”模式，提高测序读长的聚类精度；然后，提出了基于可识别DNA码的桶式分配策略的纠错算法。仿真结果表明在0.1和0.05错误率条件下，平均解码准确率在20X测序深度时可达94%以上；在0.15错误率条件下，平均解码准确率在60X测序深度时可达90%以上。

2　方法

2.1　编码方法

表1为本文使用的英文文本前向纠错编码表[23]。该编码表包含26个常规DNA编码序列(表1中白色底纹部分)以及4个特殊DNA编码序列(表1中阴影部分)。常规DNA编码序列用于编码英文字母、标点符号和数字等字符。特殊DNA编码序列包括大写键、标点符号键、数字键和空格键。除空格键外，其他特殊键用于标记位于其后的下一个DNA编码序列编码何种字符(大写字母、数字字符或标点符号字符)，例如编码序列“CTTGTC ACACAC”表示编码的字符为数字字符“6”。需要说明的是，前一个编码序列不是特殊键的DNA编码序列编码的为小写英文字母。

该编码表具有如下特点。首先，编码表的设计遵循了生物序列的约束，比如任意两个DNA编码序列拼接不会产生长度大于2的均聚物、鸟嘌呤和胞嘧啶 (Guanine and Cytosine, GC)含量保持平衡且分布均匀，有利于减少DNA分子在DNA存储过程中的错误。其次，该编码表任意两个DNA编码序列的汉明距离都至少为3，因此具有1位碱基替换纠错的能力。此外，该编码表435对序列的平均编辑(插入、删除、替换)距离为3.85， 仅有12对编码间的编辑距离为2。因此，可以近似认为该编码表具有一位的纠错能力。

2.2　解码方法

2.2.1　分组策略

DNA存储解码的第1步是对测序读长(reads)

表 1　编码表

DNA编码	字母	标点符号	数字	DNA编码	字母	标点符号	数字
TAACCG	a	@	4	ACACAC	l		6
TAAGGC	p	￥		ACTCTG	i	"	5
ATCACG	e	$	2	TCAGAG	j	%	
ATGAGC	y	,	8	ACAGGT	x	}{	
ATGGAG	g	*		ACTGCA	f	~	9
TACCAC	k	/		AGACCT	s	+	0
ATCCGT	b	()		TCTCGT	h	-	
ATGCCA	v	:		TCGAAC	c	?	1
TAGCGA	r	'	3	ACGACT	o	!	
TAGGCT	t	[]		CATTCG	z	&	
ACATCG	w	;		CTACAG	q	=	
ACTTGC	n	.		CAACGT	d	__	
TCTACG	m	Enter		CAGACA	u	#	7
TGCATA		大写键		GTATGA		标点符号键	
CTTGTC		数字键		CGGTAT		空格键	

分组或聚类，即将属于同一编码序列的测序读长划分为一组，为后续基于多序列的一致性序列推断奠定基础。由于测序读长中的各种错误，分组的一个目的是精度高，尽可能减少将其他序列的测序读长错误加入分组；另一个是召回率高，即尽可能将属于同一个序列的测序读长判别出来并分为一组。在机器学习领域，这两个指标往往难以同时满足，需要进行一定的折中。

本文采用图1所示的序列设计，在存储序列前后各加一个该序列的索引，再加一个索引的循环冗余校验码 (Cyclic Redundancy Check, CRC)。图1中索引值和CRC校验值的编码按照两个连续比特编码成一个DNA碱基[24]。分组原则是：(1)如果索引1与索引2相同，则按索引1分组；(2)如果索引1与索引2不同，则选择与CRC校验值一致的索引分组。以上两个条件不满足就丢弃该序列。

本质上这一方法属于基于测序读长索引值直接分组的方法，其时间复杂度为$O(N)$(N为测序读长的总数)。另一种分组方法是直接基于序列比对的相似性分组方法，其时间复杂度为$O(N^2n^2)$(n 为测序长度)。相比于前者，后者的召回率相对较高，但时间复杂度大。特别是DNA存储中N为百万级别数量时，所需时间将难以想象。

2.2.2 桶式纠错策略

图2(a)给出了一条测序读长可能发生错误情况的示意图。从编码单元的角度看，按照其受影响的程度可以分为3类：(1)第1类、完全正确，没有受到错误影响(深绿色)；(2)仅受到1位插入/删除/替换的影响(浅绿色)；(3)大于1位插入/删除/替换的影响(白色)。由于本文所用的30个前向纠错编码具有1位纠错能力，本文对一个测序读长观测到的长度为5,6或7的子串有如下两个假设：

(1)如果一个6碱基子串是合法编码，它以很高的概率属于第1类编码；

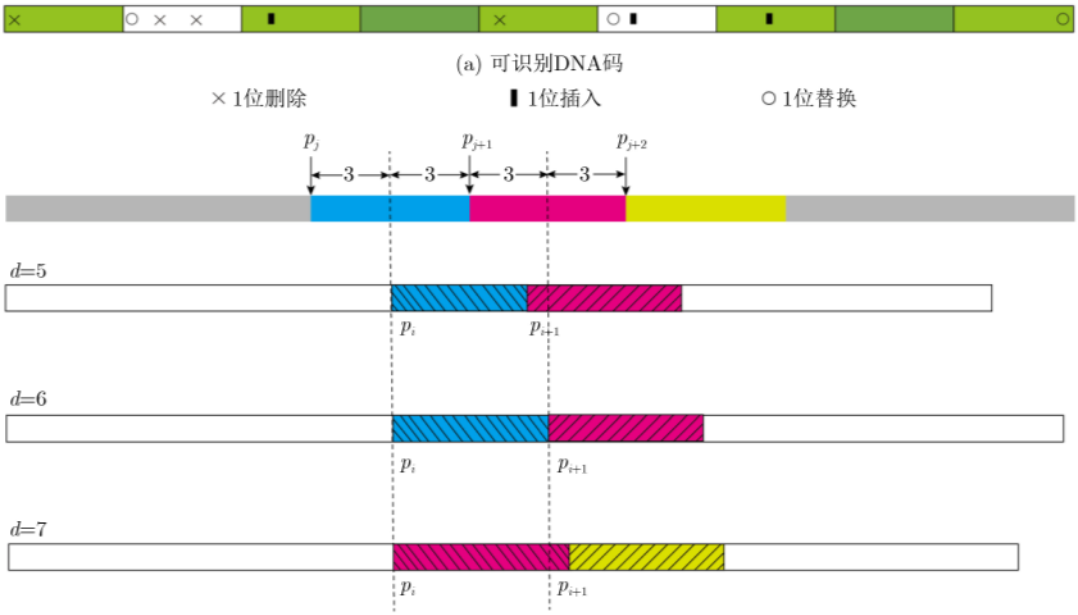
(2)如果一个子串与合法编码的编辑距离为1，它以较高的概率属于第2类编码。

本文将在一个测序读长中出现的上述两种字串称为可识别DNA码。基于上述认识，本文提出一个基于可识别DNA码桶分配策略的纠错算法，基本思想是搜索一个分组中的所有测序读长的可识别DNA码，根据其在测序读长的位置分配合适的编码位置(即桶)，最后根据每个桶中的编码投票确定最终的编码。

可识别DNA码的搜索方法如下：

(1)搜索长度为5, 6和7的DNA子串并计算其与编码表的最小编辑距离；

图 1 DNA存储序列结构示意图



(b) 可识别DNA码分配示意图

图 2 桶式分配纠错策略示意图

(2)如果存在可识别DNA码，则确定第1个碱基位置，从该可识别DNA码后面的碱基重复(1)；

(3)如果不存在可识别DNA码，则从当前位置前进一个碱基，重复(1)。

重复上述步骤直到扫描完整个测序读长，即可得到其中的所有可识别DNA码。需要说明的是，上述过程每个可识别DNA码将根据其最小编辑距离赋值不同的权重。当最小编辑距离为0，则权重设置为1，否则设置为0.1。为了提高计算效率，本文提前编制好一个长度为5, 6, 7的所有DNA串与30个合法编码的最小编辑距离表，上述搜索过程的最小编辑距离直接查表即可得到，避免了重复计算。

本文编码中，每个合法编码的位置形成一个间隔为6的整数序列，如1, 7, 13, ⋯, 175(共有$n=30$个编码)。由于插入删除的影响，导致测序读长中可识别DNA码的位置往往会偏离其原始位置。可识别DNA码的分配问题可以简单描述为：给定一个合法编码位置序列$p_j(j=1,2,\cdots,n)$和一个可识别DNA码的位置序列$p_i(i=1,2,\cdots,m)$，寻找p_i到p_j的一个最小代价$f=\sum_{k=1}^{r}|p_i^k-p_j^k|$的匹配(其中$r=\min(m,n)$)。显然，最小代价匹配应尽可能将每个可识别DNA码的p_i分配到距其最近的一个p_j。但是，当两个相邻可识别DNA码连在一起时，可能会发生分配冲突。具体说，就是当p_i位于p_j和p_{j+1}中间时，p_i既可以分配到p_j，也可以分配到p_{j+1}。此时，需要考虑下一个可识别DNA码的位置p_{i+1}。根据p_i和p_{i+1}间的距离$d=p_{i+1}-p_i$。分为以下3种情况(图2(b))：

(1)当$d=5$时，p_{i+1}到p_{j+1}距离为2，p_{i+1}应分配到p_{j+1}，从而p_i应分配到p_j；

(2)当$d=6$时，p_{i+1}在p_{j+1}和p_{j+2}中间，p_i的分配不影响p_{i+1}，本文将p_i应分配到p_j；

(3)当$d=7$时，p_{i+1}到p_{j+2}距离为2，p_{i+1}应分配到p_{j+2}，为了保持p_i和p_{i+1}的相邻状态，应该将p_i应分配到p_{j+1}。

基于上述认识，本文提出如下局部最优可识别DNA码分配算法：

(1)如果当前p_i不在p_j和p_{j+1}中间，根据最短距离分配，否则转(2)；

(2)如果p_j未分配，按照上面的3种情况进行分配；否则将p_i分配给p_{j+1}；

(3)重复(1)，直到p_r分配完成。

如果本文将每个合法编码位置当作一个桶，对一个序列分组中所有DNA测序读长进行解码的过程可以描述为：将测序读长中的可识别DNA码按照其对应的合法编码放入对应的桶，最后根据每个桶中编码权重进行投票，即可确定该序列的可能编码。

3　实验结果

本文仿真实验采用《老人与海》和《罗伯特·路易斯·斯蒂文森评传》的英文文本，总文件大小为324 kB。编码英文文本的DNA存储行的长度为208碱基，其中数据域长度为180个碱基，索引1和索引2均为10碱基，CRC校验值为8碱基。最终形成11637条DNA序列。仿真实验的错误率分别为0.05, 0.1和0.15。每种错误率下，插入:替换:删除的比例设置为1∶2∶2, 1∶1∶1以及2∶2∶1 3种情况。测序深度分别为20, 30, 40, 50及60。每组参数重复1000次。仿真实验的配置为Intel(R) Xeon(R) Silver 4210 CPU @ 2.20 GHz处理器、30 GB内存的服务器，软件环境为CentOS Linux release 7.6系统。

3.1　分组策略性能分析

图3分别给出了“索引”、“索引+CRC”、“索引+CRC+索引”3种分组策略的平均精度和平均召回率。从图3(a)可以看出，后两种分组的平均精度基本接近100%，且明显高于简单索引分组策略。这主要是因为CRC或索引的校验作用明显提高了索引分组的精度。此外，简单索引分组的精度随错误率增加而降低。

图3(b)的平均召回率表明“索引+CRC+索引”分组的召回率明显高于“索引+CRC”。这主要是因为只要有一个索引通过CRC检验，即可以以很高的概率保证分组的正确性，因而提高了召回率。和简单索引相比，“索引+CRC+索引”的召回率随错误率增加而逐渐降低。

以错误率0.15为例，“索引+CRC+索引”分组的平均精度约为100%，平均召回率约为11%；简单索引的平均精度为33%，平均召回率约为20%。前者的召回率虽然约为后者的1/2，但是基本都是正确分组。而后者约有67%来自其他分组的错误测序读长，这将造成后面一个桶中会有大量不正确可识别DNA码，最终导致投票失败。因此，“索引+CRC+索引”分组策略既保障了精度又适当提升了召回率，为后面桶式纠错奠定了关键的基础。

3.2　纠错策略性能分析

图4(a)是不同测序深度情况下的解码平均准确率。可以看出：(1)当错误率为0.05和0.10，本文方法在测序深度20时平均准确率就达到94%以上。但是随着测序深度增加，平均准确率基本不变。这可能与DNA存储测序的分布不均性有关。这里存储测序的分布不均匀主要包括两个方面：一是序列中

碱基错误分布的不均匀(仿真数据里表现为序列中碱基错误随机发生的不均匀性)；二是测序序列拷贝数分布的不均匀(仿真数据里表现为编码序列的抽样分布不均匀)。DNA存储测序的不均性，导致了无论测序深度多高，总是有些序列因为拷贝数太少以及序列随机错误发生的不均匀性，导致不能准确解码，进而影响了平均准确率。当错误率增加到0.15，平均准确率极具下降，测序深度20时的平均准确率约为70%。随着测序深度的增加，在测序深度60时就达到90%。(2)插入删除比例对纠错性能有一定影响，当错误率为0.05和0.10，删除比例较大时的平均精度较高。这可能是低错误率删除错误对合法编码的影响小于插入的破坏程度。当错误率为0.15时，删除比例较大时的平均精度较低。这说明高错误率删除错误对合法编码的影响大于插入的破坏程度。例如，合法编码“ATGAGC”，两位碱基缺失后的编码可能为“ATGA”，“ATGC”，“AGAC”，…，任意两个位置的碱基缺失，都造成了合法编码本身信息的破坏，每种破坏均有可能导致纠正错误(注：高错误率下插入错误或删除错误导致发生错误的合法DNA码普遍出错的碱基数量大于等于2)。而合法编码两位碱基插入错误的引入，比如“ATGAGCXX”，“XATGAGCX”，“XXATGAGC”，“ATXGXAGC”，… (X为插入碱基)，并不会破坏合法编码本身的信息。此外在合法编码所有两位插入错误的种类中，存在少量种类是可以正确识别并纠正的，比如“ATGAGCXX”，“XATGAGCX”和“XXATGAGC”。这表明，缺失两个碱基的合法DNA编码序列纠正失败的概率，要大于插入两个碱基的合法DNA编码纠正失败的概率。这就导致给定一高错误率，删除错误比例较大的情况下可识别DNA码的识别准确率低于插入比例较大情况下的可识别DNA码的识别准确率。

图4(b)是不同测序深度的情况下的解码平均运行时间。可以看出：(1)随着测序深度的增加，算法运行时间近似线性增加；(2)在给定测序深度下，算法运行时间随错误率增加而减少。这主要是由于错误率对分组测序读长数量的影响导致。图3(b)显示在0.05, 0.10和0.15时，分组读长数量占测序深度的百分比大致分别为0.60, 0.27和0.11，基本与相应错误率下的时间比一致。

3.3 与其他方法的比较

表2给出了不同方法的纠错策略、插入/删除纠错、覆盖率、错误率、准确率和存储模型。与文献[25,26]的文本存储工作相比，本文提出的方法可以对包括插入、删除和替换在内的3种错误进行纠正，而其他两种方法则没有这种能力，这限制了它

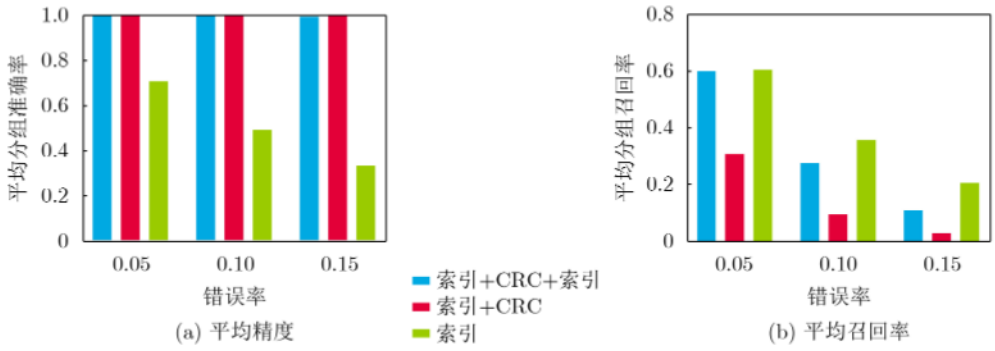


图 3 3种分组策略性能比较

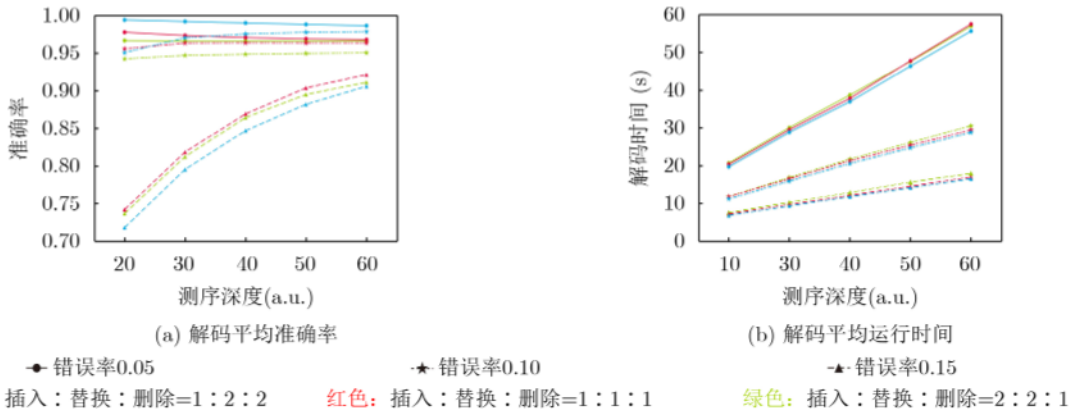


图 4 解码过程的平均正确率与平均运行时间

表 2　与其他方法的比较

文献	纠错策略	是否插入/删除?	测序深度(X)	错误率(%)	数据恢复准确率	存储方法
文献[20]	HEDGES	是	50	3	1.000	体外
文献[21]	莱文斯坦码	是	NA	1	0.900	体外
文献[22]	德布莱茵图	是	60	10	0.920	体外
文献[23]	3层纠错机制	是	20	5	0.905	体外
文献[25]	否	否	2740	≤2	0.998	体外
文献[26]	否	否	NA	≤2	1.000	体内
本文方法	桶式分配纠错	是	20	5	0.970	体外
			20	10	0.940	
			60	15	0.906	

们只能在噪声非常低(≤2%)的情况下应用且需要较高的测序深度。本文方法在20X测序深度下，在错误率10%的条件下能恢复94%以上的数据。而文献[25]恢复错误率约为2%的数据所需要的测序深度约为2700X，这对于未来的大数据存储是不可行的。

与文献[20–23]中的4种一般DNA存储方法相比，本文方法比文献[21]的方法更强大，文献[21]只能纠正1位碱基的插入。在错误率为3%和50X测序深度的情况下，本文方法和文献[20]中的方法几乎相同，可以达到99%以上的平均准确率。和文献[22]相比，在错误率为10%的情况下，60X测序深度下本文方法的平均准确率为94%以上，高于文献[22]92%的平均准确率。此外，在错误率为10%和20X测序深度下，本文方法的平均准确率依然达到94%以上，远远高于文献[22]50%的平均准确率。和文献[23]相比，在错误率为5%和20X测序深度的情况下，本文方法的平均准确率可以达到97%以上，高于文献[23]的平均准确率。此外，本文方法在错误率15%和60X测序深度的情况下，平均准确率可以达到90%以上，远高于文献[23]64%的平均准确率。

4　结束语

如何解决序列中的组合插入、删除和替换错误，是DNA存储信息可靠性的基础。DNA存储的解码过程主要包括两个方面：测序读长的分组和基于组内测序读长的一致性序列的恢复。为了提高DNA存储信息的恢复精度，本文主要在以上两个方面进行了如下的研究：(1)提出了“索引+CRC哈希+索引”的序列索引编码方法，仿真结果表明该索引编码的分组精度可以达到99%以上，并保证较高的召回率。(2)在文本字符前向纠错编码的基础上，提出一种基于可识别DNA码的桶分配纠错算法。影响本文纠错方法精度的因素主要有3个：一是可识别DNA码的检索与纠错；二是可识别DNA码桶的分配；三是基于可识别DNA码权重分配的多数投票。仿真结果表明在0.10和0.05错误率条件下，平均解码准确率在20X测序深度时可达94%以上；在0.15错误率条件下，平均解码准确率在60X测序深度时可达90%以上。此外，在给定错误率的情况下，本文提出的解码算法为线性时间复杂度$O(N)$。因此，适合于未来面向大数据的DNA存储应用。最后，如何解决可识别DNA码的最优分配，进一步提高分配的准确率将是未来研究的一个主要的方向。

参 考 文 献

[1] REINSEL D, GANTZ J, and RYDNING J. The digital of the world from edge to core[EB/OL]. http://book.itep.ru/depository/dig_economy/idc-seagate-dataage-whitepaper.pdf, 2020.

[2] WILLIAMS E D, AYRES R U, and HELLER M. The 1.7 kilogram microchip: Energy and material use in the production of semiconductor devices[J]. *Environmental Science & Technology*, 2002, 36(24): 5504–5510. doi: 10.1021/es025643o.

[3] GODA K and KITSUREGAWA M. The history of storage systems[J]. *Proceedings of the IEEE*, 2012, 100: 1433–1440. doi: 10.1109/JPROC.2012.2189787.

[4] 许鹏, 方刚, 石晓龙, 等. DNA存储及其研究进展[J]. 电子与信息学报, 2020, 42(6): 1326–1331. doi: 10.11999/JEIT190863.
XU Peng, FANG Gang, SHI Xiaolong, *et al*. DNA storage and its research progress[J]. *Journal of Electronics & Information Technology*, 2020, 42(6): 1326–1331. doi: 10.11999/JEIT190863.

[5] 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.
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.

[6] CEZE L, NIVALA J, and STRAUSS K. Molecular digital

data storage using DNA[J]. *Nature Reviews Genetics*, 2019, 20(8): 456–466. doi: 10.1038/s41576-019-0125-3.

[7] GAO Yanmin, CHEN Xin, QIAO Hongyan, *et al*. Low-bias manipulation of DNA oligo pool for robust data storage[J]. *ACS Synthetic Biology*, 2020, 9(12): 3344–3352. doi: 10.1021/acssynbio.0c00419.

[8] DONG Yiming, SUN Fajia, PING Zhi, *et al*. DNA storage: Research landscape and future prospects[J]. *National Science Review*, 2020, 7(6): 1092–1107. doi: 10.1093/nsr/nwaa007.

[9] 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.

[10] STANCU M C, VAN ROOSMALEN M J, RENKENS I, *et al*. Mapping and phasing of structural variation in patient genomes using nanopore sequencing[J]. *Nature Communications*, 2017, 8(1): 1326. doi: 10.1038/s41467-017-01343-4.

[11] 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. doi: 10.1038/s41598-019-41228-8.

[12] KUMAR U K and UMASHANKAR B S. Improved hamming code for error detection and correction[C]. 2007 2nd International Symposium on Wireless Pervasive Computing, San Juan, USA, 2007: 1. doi: 10.1109/ISWPC.2007.342654.

[13] BLAWAT M, GAEDKE K, HÜTTER I, *et al*. Forward error correction for DNA data storage[J]. *Procedia Computer Science*, 2016, 80: 1011–1022. doi: 10.1016/j.procs.2016.05.398.

[14] LU Xiaozhou, 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: 162892–162902. doi: 10.1109/ACCESS.2020.3021700.

[15] 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.

[16] ANTKOWIAK P L, LIETARD J, DARESTANI M Z, *et al*. Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction[J]. *Nature Communications*, 2020, 11(1): 5345. doi: 10.1038/s41467-020-19148-3.

[17] 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.

[18] 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.

[19] JEONG J, PARK S J, KIM J W, *et al*. Cooperative sequence clustering and decoding for DNA storage system with fountain codes[J]. *Bioinformatics*, 2021, 37(19): 3136–3143. doi: 10.1093/bioinformatics/btab246.

[20] PRESS W H, HAWKINS J A, JONES JR 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. doi: 10.1073/pnas.2004821117.

[21] XUE Tianbo and LAU F C M. Notice of violation of IEEE publication principles: Construction of GC-balanced DNA with deletion/insertion/mutation error correction for DNA storage system[J]. *IEEE Access*, 2020, 8: 140972–140980. doi: 10.1109/ACCESS.2020.3012688.

[22] SONG Lifu, GENG Feng, GONG Ziyi, *et al*. . Robust data storage in DNA by de Bruijn graph-based decoding[J]. *bioRxiv*, 2022, 13(1): 5361. doi: 10.1101/2020.12.20.423642.

[23] ZAN Xiangzhen, YAO Xiangyu, XU Peng, *et al*. A hierarchical error correction strategy for text DNA storage[J]. *Interdisciplinary Sciences*: *Computational Life Sciences*, 2022, 14(1): 141–150. doi: 10.1007/s12539-021-00476-x.

[24] BORNHOLT J, LOPEZ R, CARMEAN D M, *et al*. A DNA-based archival storage system[J]. *ACM SIGPLAN Notices*, 2016, 51(4): 637–649. doi: 10.1145/2954679.2872397.

[25] ZHONG Yunpeng, QI Shanshan, SHENG Fuxu, *et al*. A new digital information storing and reading system based on synthetic DNA[J]. *Science China Life Sciences*, 2018, 61(6): 733–735. doi: 10.1007/s11427-017-9131-7.

[26] LEE U J, HWANG S, KIM K E, *et al*. DNA data storage in Perl[J]. *Biotechnology and Bioprocess Engineering*, 2020, 25(4): 607–615. doi: 10.1007/s12257-020-0022-9.

咎乡镇：男，博士生，研究方向为DNA存储、生物信息学.
姚翔宇：男，硕士生，研究方向为DNA存储、生物信息学.
许　鹏：男，副教授，研究方向为DNA存储、生物信息学.
陈智华：女，副教授，研究方向为DNA存储、生物信息学.
石晓龙：男，教授，研究方向为DNA存储、生物信息学.
李树栋：男，副教授，研究方向为DNA存储、网络安全.
刘文斌：男，教授，研究方向为DNA存储、生物信息学.

责任编辑：余　蓉