DOI:10.11784/tdxbz202506038

面向 DNA 存储低串扰信息检索的引物设计方法

张淑芳[1,2]，杨华卿[1]，王鹏浩[1]，罗　茗[1]
（1. 天津大学电气自动化与信息工程学院，天津 300072；2. 合成生物技术全国重点实验室（天津大学），天津 300072）

摘　要：在 DNA 信息存储中，从寡核苷酸池中准确有效地检索信息是提升其实用性的关键. 目前，基于 PCR 扩增的信息检索方法因具备技术成熟、操作简单和成本低等优势被广泛应用. 该方法通过引物与目标文件 DNA 序列上引物结合位点的特异性结合并扩增实现信息检索，若引物间的正交性不强，极易造成引物与非目标 DNA 序列的错误结合从而导致信息检索出错. 为了提升 DNA 存储信息检索的准确性，提出一种面向 DNA 存储低串扰信息检索的引物设计方法，该方法主要包括随机引物生成、引物间加权汉明距离设计和强正交性引物筛选 3 个部分. 首先随机生成满足 C_{GC}、均聚物长度、熔化温度和吉布斯自由能等基本引物设计规则的引物；然后设计加权汉明距离以确保对引物间正交性的衡量更加精准；最终构建以引物为节点、以加权汉明距离为路径的全耦合网络，并基于网络路径优化的策略筛选出具备强正交性的引物用于 DNA 存储信息检索，以降低信息检索出错的概率. 实验结果表明，通过该方法设计的引物与传统方法相比具有更强的正交性，并且采用该方法设计的引物进行 DNA 存储信息检索可将误检可能性降低 1/5 左右，有效提升了 DNA 存储信息检索的准确性，有助于推动 DNA 存储的实用化进程.

关键词：DNA 存储；信息检索；引物；引物正交性
中图分类号：TP391；Q811.4　　**文献标志码**：A　　**文章编号**：0493-2137（2026）06-0565-08

Primer Design Method for Low-Interference Information Retrieval in DNA Storage

Zhang Shufang[1,2]，Yang Huaqing[1]，Wang Penghao[1]，Luo Ming[1]
（1. School of Electrical and Information Engineering，Tianjin University，Tianjin 300072，China；
2. State Key Laboratory of Synthetic Biology，Tianjin University，Tianjin 300072，China）

Abstract：The accurate and efficient retrieval of data from an oligonucleotide pool is crucial for enhancing the practicality of DNA information storage systems. Currently，PCR amplication is widely used as an information retrieval method due to its advantages，such as its advanced technological nature，ease of operation，and low cost. This method retrieves information by specifically binding primers to primer binding sites in the target DNA sequences，followed by amplification. However，a lack of primer orthogonality most likely leads to nonspecific binding，resulting in errors. To improve the accuracy of DNA storage information retrieval，this study proposed a novel method for designing primers that enable low-interference information retrieval. It consisted of three main components：random primer generation，the design of weighted Hamming distance for primers，and screening of primers with strong orthogonality. First，primers that met basic primer design rules such as C_{GC}，homopolymer length，melting temperature，and Gibbs free energy were randomly generated. Then，the weighted Hamming distance was designed to ensure a more accurate measurement of primer orthogonality. Finally，a fully-coupled network with primers as nodes and weighted Hamming distances as paths was constructed. From this complex，primers with strong orthogonality were selected using a network path optimization strategy，thereby reducing the likelihood of errors. The

收稿日期：2025-06-27；**修回日期**：2025-07-29.
作者简介：张淑芳（1979—　），女，博士，副教授.
通信作者：张淑芳，shufangzhang@tju.edu.cn.
基金项目：天津市科技计划资助项目（22JCYBJC01390）；合成生物技术全国重点实验室自主创新基金资助项目（HCZC-202610A）.
Supported by the Science and Technology Support Program of Tianjin，China（No. 22JCYBJC01390），the Independent Innovation Fund of State Key Laboratory of Sythetic Biology（No. HCZC-202610A）.

experimental results confirmed that these primers were more orthogonal than those obtained using traditional methods, reducing the risk of false retrieval by about one-fifth. Thus, they effectively improved the accuracy of DNA storage information retrieval and advanced the practical applications of DNA information storage systems.

Keywords: DNA storage; information retrieval; primer; primer orthogonality

DNA 存储作为一种新型信息存储技术,因具备存储密度高、存储时间长、维护成本低等优势而受到广泛关注[1-4]. 与硅基存储不同,DNA 存储将信息存储在 DNA 分子中. 在硅基存储中可借助于寻址实现对目标信息的访问,然而,DNA 存储中分子处于无序混杂状态[5-7],无法做到类似于硅基存储中的寻址. 因此,若存储少量数据,可通过对所有 DNA 分子进行测序解码以获取目标信息;但若存储海量数据,该方法的信息检索成本将显著提升.

国内外学者对 DNA 存储信息的随机访问开展了广泛研究. 其中,基于聚合酶链式反应(polymerase chain reaction, PCR)扩增的 DNA 存储信息检索方法因操作简单、可实现性强等优点而成为 DNA 存储主流信息检索方法[8-11]. 该方法为不同文件分配专属引物,并在文件的 DNA 序列上添加引物结合位点. 在检索时使用目标文件的引物进行 PCR 扩增,这些引物仅与目标文件 DNA 序列上的位点相结合并进行指数型复制,从而实现对目标文件的访问. 该方法最初由 Yazdi 等[12]提出,通过设计不同引物实现对文件的选择性访问. 2018 年,Organick 等[13]在此基础上优化了引物库设计,在设计引物时加入了生化约束条件,以降低 DNA 存储信息检索的错误率. 2021 年,Song 等[14]在 DNA 序列中加入了 3 级引物结合位点,以实现 DNA 存储中的大规模文件检索. 2024 年,张淑芳等[15]提出了基于引物索引矩阵的文件高效随机检索方法,通过构建引物索引矩阵将引物检索效率提升为常数级时间复杂度. 同年,Wang 等[16]提出了一种面向 PCR 扩增信息检索的高效 DNA 编码算法,该方法降低了引物 3'端与 DNA 序列之间的非特异性配对概率,从而提升 DNA 存储信息检索的准确性.

此外,Newman 等[17]在 2019 年设计了微流控芯片,通过该设备可实现对目标 DNA 分子的提取并最终完成对目标文件的访问. 2021 年,Piantanida 等[18]提出了一种可对文件进行布尔逻辑检索的 DNA 存储架构,将 DNA 分子固化到二氧化硅微球表面并进行封装,最后添加 3 条索引序列用于文件检索. 同年,Bee 等[19]提出了一种相似性文件检索方法,利用杂交探针和磁珠亲和纯化技术获取与目标图像相似的全部文件索引,然后基于这些索引从数据库中访问多个相似图像.

经上述研究,基于 PCR 扩增的 DNA 存储信息检索方法在检索准确率、检索规模以及检索效率方面都有所提升,但仍然会出现检索到非目标 DNA 序列的情况,这是由于引物间的正交性不强,并且直接采用汉明距离作为引物间正交性的指标与实际存在偏差. 为了解决上述问题,本文提出了一种面向 DNA 存储低串扰信息检索的引物设计方法,采用先生成再筛选的策略,首先生成若干符合基本引物设计原则的随机引物,然后计算随机引物间的加权汉明距离,以随机引物为节点、以引物间的加权汉明距离为路径长度构建全耦合网络,再对网络的节点和路径进行裁剪,最终筛选出正交性强的随机引物,并实现 DNA 存储的低串扰信息检索.

1 面向 DNA 存储信息检索的强正交性引物设计方法

为了实现基于 PCR 扩增的 DNA 存储信息检索,需设计正交性强的引物,将这些引物分配给不同的待存储文件,在待存储文件的 DNA 序列中加入相应的引物结合位点,从而可通过特定引物检索目标文件. 存储过程如图 1 左上部分所示,先对待存储文件进行 DNA 编码,然后在 DNA 序列两端添加引物结合位点,合成后便可进行存储.

如图 1 右上部分所示,在进行 DNA 存储信息检索时,需先将目标文件对应的引物加入寡核苷酸池中进行 PCR 扩增. 在此过程中,引物会与目标文件 DNA 序列上的位点相结合,并完成目标文件 DNA 序列的指数型复制. PCR 扩增结束后,对产物进行测序,以读取 DNA 序列上的碱基排布. 对测序结果进行解码还原后,可实现目标文件信息的检索.

面向 DNA 存储低串扰信息检索的强正交性引物设计方法如图 1 下侧所示,主要分为 3 个步骤:首先,生成随机引物,并为其编号;其次,计算上述随机引物间的加权汉明距离;再次,以随机引物为节点、以加权汉明距离为路径长度构建全耦合网络,并通过网络裁剪的方式筛选符合条件的强正交性引物.

1.1 随机引物生成

在基于 PCR 扩增的 DNA 存储信息检索中,引

物通常是长度范围为 15 ~ 30 nt 的碱基序列，其中 nt 为核苷酸(nucleotide)的缩写. 且引物应满足 G 和 C 的含量(C_{GC})、均聚物长度、熔化温度和吉布斯自由能(ΔG)等设计规则，其计算式和计算复杂度如表 1 所示.

引物的 C_{GC} 应在 0.4 ~ 0.6 之间，引物中不应出现长度超过 3 的均聚物，引物熔化温度应在 55 ~ 75 ℃ 之间，且为了避免引物自身形成二级结构，引物的 ΔG 值应小于 14.6 kJ/mol. 本文在随机引物生成阶段生成长度为 l(15 nt≤l≤30 nt)的碱基序列 s，然后对碱基序列 s 进行引物设计规则检测以筛选出符合规则的引物 p. 只有满足所有规则的碱基序列才能当选为引物，一旦碱基序列不满足某条规则即被丢弃. 为降低计算复杂度，本文在进行引物筛选时，先通过计算复杂度低的检测将不满足引物设计规则的碱基序列筛除，再执行计算复杂度高的规则检测. 计算复杂度由获得该项指标所需的运算步数决定. 由于计算复杂度从低到高的排序为 C_{GC}、均聚物长度、熔化温度和吉布斯自由能，所以规则检测顺序与之相同.

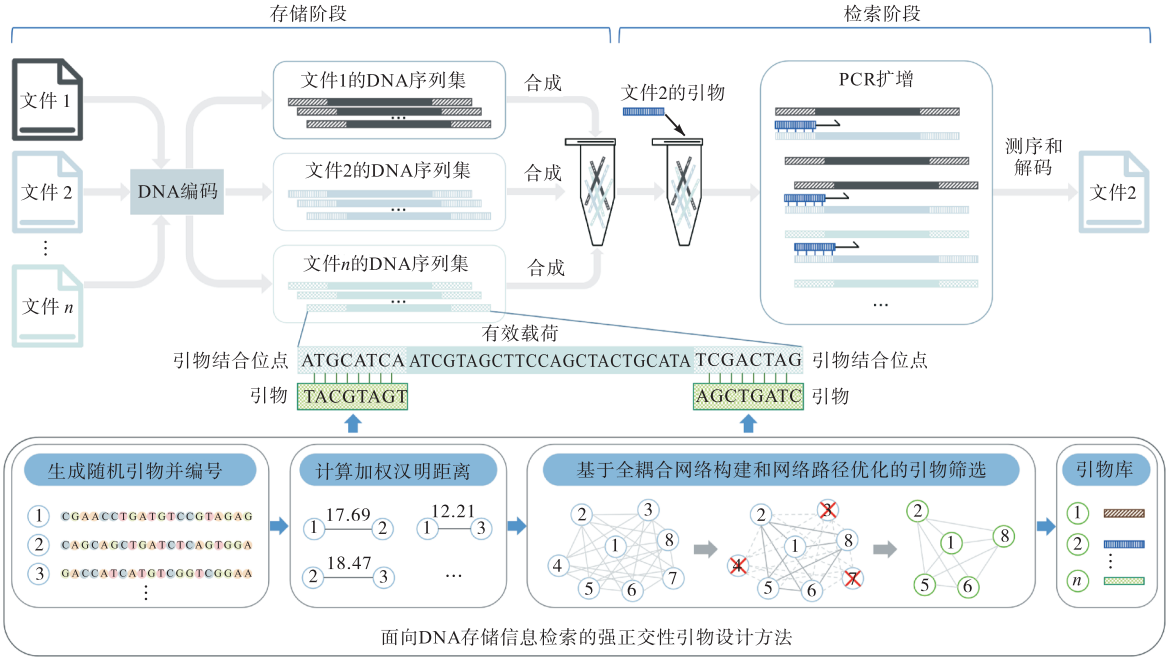


图 1　基于强正交性引物设计的 DNA 存储信息检索示意

Fig.1　Schematic of information retrieval in DNA storage based on primer design with strong orthogonality

表 1　引物设计规则、规则计算公式和复杂度

Tab.1　Primer design rule，rule calculation formulas and complexity

规则名称	规则描述	计算式	计算复杂度
C_{GC}	$0.4 \leqslant C_{GC} \leqslant 0.6$	$C_{GC}=(Q_G+Q_C)/L_s$，Q_G 和 Q_C 分别为碱基序列 s 中 G 和 C 的个数，L_s 为碱基序列 s 的长度	5 步
均聚物长度	均聚物长度≤3	均聚物长度 = $Q_{4A}+Q_{4C}+Q_{4G}+Q_{4T}$，$Q_{4A}$、$Q_{4C}$、$Q_{4G}$ 和 Q_{4T} 分别为碱基序列 s 中 AAAA、CCCC、GGGG 和 TTTT 的个数	7 步
熔化温度	55 ℃ < t_m < 75 ℃	$t_m=4(Q_G+Q_C)+2(Q_A+Q_T)$，$Q_G$、$Q_C$、$Q_A$ 和 Q_T 分别为碱基序列 s 中 G、C、A 和 T 的个数	9 步
ΔG	ΔG < 14.6 kJ/mol	$\Delta G=\Delta H-T\Delta S$，$\Delta H$ 为焓变，T 为热力学温度，ΔS 为熵变	较高

1.2　引物间加权汉明距离的计算

目前大部分方法采用通过汉明距离作为衡量引物间正交性的指标，但由于该指标没有考虑碱基位置对引物间正交性的影响. 因此，本文提出了利用加权汉明距离来衡量引物间的正交性.

为计算引物间的加权汉明距离，本文定义了碱基异或，这种概念是从二进制异或推演而来，如图 2 所示. 以 A 与 A 异或为例，可先将其转化为二进制 00 与 00 异或，逐位异或结果为 00，再将其映射为碱基 A. 按照此方法逐个推演其他碱基组合的异或结果，最终可以获得图 2(d)所示的碱基异或关系表.

在利用碱基异或关系表计算两个引物 p_1 和 p_2 间的汉明距离 D_h 时，需首先对两个引物进行碱基异或获得一个新的碱基序列 p_{eor}，再统计上述碱基序列中

C、G、T 的个数,两个引物间的汉明距离计算过程分别为

$$p_{\text{eor}} = p_1 \oplus p_2 \tag{1}$$

$$D_{\text{h}} = m_{\text{eor,C}} + m_{\text{eor,G}} + m_{\text{eor,T}} \tag{2}$$

式中 $m_{\text{eor,C}}$、$m_{\text{eor,G}}$ 和 $m_{\text{eor,T}}$ 分别为碱基序列 p_{eor} 中 C、G 和 T 的个数.

异或$_2$	0	1
0	0	1
1	1	0

(a) 二进制异或关系表

映射	A	C	G	T
	00	01	10	11

(b) 二进制与碱基映射关系表

异或$_m$	00	01	10	11
00	00	01	10	11
01	01	00	11	10
10	10	11	00	01
11	11	10	01	00

(c) 两位二进制异或关系表

异或$_4$	A	C	G	T
A	A	C	G	T
C	C	A	T	G
G	G	T	A	C
T	T	G	C	A

(d) 碱基异或关系表

图 2 二进制异或到碱基异或推演过程

Fig.2 Derivation process from binary XOR to base XOR

为了获取不同碱基位置对引物间正交性影响的权值,本文使用 primer3 软件对 DNA 存储的信息检索过程进行仿真,并计算出引物与 DNA 序列间的 ΔG 值代表二者的结合稳定程度. 在上述仿真过程中,计算了正确引物(与 DNA 序列上的引物结合位点满足碱基互补配对原则)以及单汉明距离引物(与正确引物的汉明距离为 1)与 DNA 序列间的 ΔG 值. 将正确引物与 DNA 序列间的 ΔG 值记为 ΔG_0,则有

$$\Delta G_0 = \Delta H_0 - T\Delta S_0 \tag{3}$$

式中:ΔH_0 为正确引物与 DNA 序列间的焓变;ΔS_0 为正确引物与 DNA 序列间的熵变.

单汉明距离引物与 DNA 序列间的 ΔG 值记为 ΔG_x,则有

$$\Delta G_x = \Delta H_x - T\Delta S_x \tag{4}$$

式中:ΔH_x 为单汉明距离引物与 DNA 序列间的焓变;ΔS_x 为单汉明距离引物与 DNA 序列间的熵变;x 代表单汉明距离引物与正确引物存在不同碱基的位置,例如,当 x 等于 1 时,代表该单汉明距离引物与正确引物只有第 1 个碱基不同,其余碱基均相同.

由于正确引物与 DNA 序列上的引物结合位点能够完全互补配对,而单汉明距离引物与其存在 1 个无法互补配对的碱基,所以前者的结合稳定性更强,即 $\Delta G_0 > \Delta G_x$. 二者的差值 W_x 为

$$W_x = \Delta G_0 - \Delta G_x \tag{5}$$

式(5)反映了碱基位置 x 对引物与 DNA 序列结合稳定程度的影响.

对 W_x 进行归一化处理(除以其最大值 $W_{x,\max}$),获得归一化参数 W_x'($0 \leqslant W_x' \leqslant 1$),其计算式为

$$W_x' = W_x / W_{x,\max} \tag{6}$$

利用这些归一化参数对汉明距离进行加权,通过求取归一化参数 W_x' 与汉明距离 D_{h} 的积以获得加权汉明距离 D_{h}',其计算式为

$$D_{\text{h}}' = W_x' D_{\text{h}} \tag{7}$$

1.3 基于全耦合网络构建和网络路径优化的引物筛选

为了进一步筛选出强正交性引物,本文采用基于全耦合网络构建和网络路径优化的引物筛选策略. 当随机生成引物的数量达到预定的起始引物个数时,可进行全耦合网络构建. 在构建全耦合网络时,将引物视为节点,将引物间的加权汉明距离视为节点间的路径长度,该过程主要分为 3 个步骤:

步骤 1 以引物 1 作为第 1 个节点,根据该引物与其他引物间的加权汉明距离,建立与其他各节点的路径;

步骤 2 根据引物 2 与其他引物间的加权汉明距离,建立该引物节点与其他各节点的路径. 在此过程中,网络空间形态逐渐确定;

步骤 3 重复上述步骤,直至网络中任意两个引物节点间均有路径相连,最终形成全局耦合网络.

全耦合网络的构建方法伪代码如下.

```
输入:P_List:包含所有引物的列表
      WH_Matrix:包含引物间加权汉明距离的矩阵
输出:GCN:全耦合网络
函数构建 GCN(P_List, WH_Matrix)
    引物数量 = P_List 的长度
    GCN = 新建矩阵(引物数量,引物数量)
    对于 i 从 1 到引物数量-1:
        距离 = WH_Matrix[0][i]
        GCN[0][i] = 距离
        GCN[i][0] = 距离
    对于 j 从 1 到引物数量-1:
        对于 k 从 j + 1 到引物数量-1:
            距离 = WH_Matrix[j][k]
            GCN[j][k] = 距离
            GCN[k][j] = 距离
    返回 GCN.
```

全耦合网络构建完成后,便可对其进行网络路径

优化，以裁剪掉正交性差的引物节点，保留正交性强的引物用于 DNA 存储信息检索. 本文在进行网络优化时引入了引物强度 S_i 的概念，其计算方法为

$$S_i=\sum_j D'_{\mathrm{h},ij} \qquad 1\leqslant j\leqslant n, j\neq i \tag{8}$$

式中：$D'_{\mathrm{h},ij}$ 表示引物 i 与引物 j 间的加权汉明距离；n 表示当前网络中引物节点的总个数. 对于网络中的每一个引物节点，首先按照式(8)计算其引物强度，并按照引物强度从高到低对引物节点进行排序；然后裁剪引物强度最低的引物节点，若存在引物强度相等的情况，则再根据其最小路径值确定被裁剪的引物节点，每次裁剪掉 1 个引物节点，直至网络中剩下引物节点的个数等于终止引物个数，可满足 DNA 存储信息检索的需求.

2　实验结果与分析

2.1　加权汉明距离与汉明距离的对比

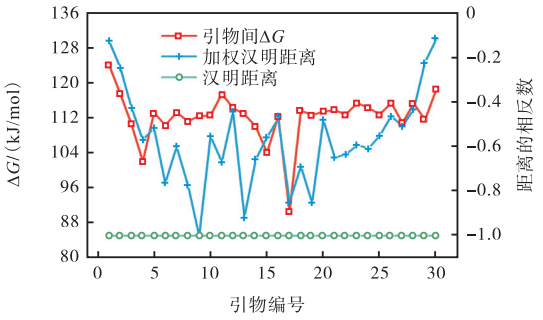
引物间正交性是反映引物间是否会发生串扰的属性，引物间距离(包括汉明距离和加权汉明距离等)和引物间 ΔG 值均为反映引物间正交性的指标.

为了验证加权汉明距离的有效性，本文随机生成了 30 组汉明距离为 1 的引物，分别计算引物间的汉明距离、加权汉明距离及 ΔG 值. 3 组重复实验的结果如图 3 所示，横轴为引物编号，左纵轴为引物间 ΔG 值，是正方形点线的纵坐标，右纵轴为引物间距离的相反数，是加号点线和圆形点线的纵坐标. 引物间距离越大，引物间 ΔG 值越小，因此两者的波动趋势相反，为了更直观地展示两者的联系，采用引物间距离的相反数作为右纵轴.

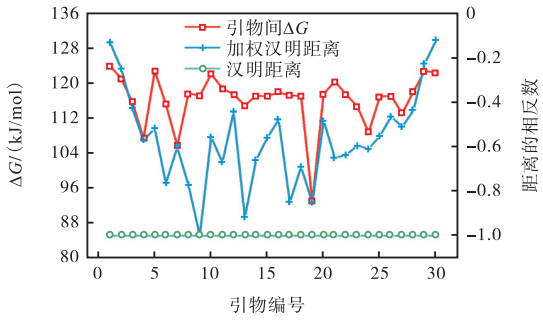
图 3 的 3 组重复实验都表明引物间加权汉明距离的相反数波动趋势与引物间 ΔG 值的波动趋势更接近，而引物间汉明距离的相反数波动趋势与引物间 ΔG 值的波动趋势存在较大差异. 这是由于加权汉明距离考虑了碱基差异位置不同对引物间正交性的影响，而汉明距离并未考虑这项因素. 因此，与直接采用汉明距离相比，本文采用加权汉明距离作为引物间正交性的指标更为恰当.

2.2　引物正交性分析

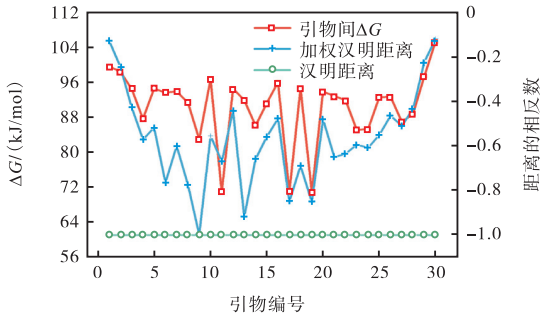
类似于通信码字间的正交性，DNA 存储引物间的正交性越强，DNA 存储信息检索发生串扰的可能性就越低. 为了更好地衡量引物间的正交性，本文定义了平均引物距离 $\bar{D}'_{\mathrm{h}}$ 与平均引物强度 $\bar{S}$ 两个指标，并基于这两个指标开展了不同引物长度、不同起始引物个数、不同终止引物个数下的引物正交性分析. 平



(a) 第 1 组重复实验



(b) 第 2 组重复实验



(c) 第 3 组重复实验

图 3　引物间的汉明距离、加权汉明距离与 ΔG 值

Fig.3　Hamming distance, weighted Hamming distance and ΔG values of primers

均引物距离 $\bar{D}'_{\mathrm{h}}$ 与平均引物强度 $\bar{S}$ 计算式分别为

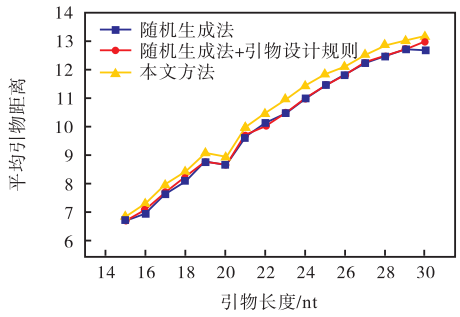
$$\bar{D}'_{\mathrm{h}}=\frac{\sum D'_{\mathrm{h},ij}}{N(N-1)} \qquad 1\leqslant i\leqslant N, 1\leqslant j\leqslant N, i\neq j \tag{9}$$

$$\bar{S}=\frac{\sum_i S_i}{N} \qquad 1\leqslant i\leqslant N \tag{10}$$

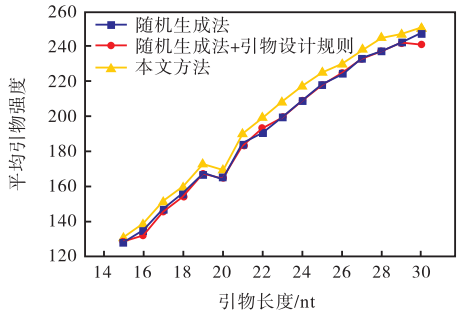
式中 N 为引物的总个数.

为了分析不同引物长度下本文生成引物的正交性，分别利用随机生成方法、考虑引物设计规则的随机生成法与本文方法生成了长度为 15 ~ 30 nt 的引物，并计算出各组引物的平均引物距离和平均引物强度，其结果如图 4 所示. 随着引物长度的增长，平均引物距离和平均引物强度整体呈上升趋势，这意味着引物间正交性增强. 在图 4 中，相比于其他两种方

法，本文方法对应的折线位于最上方，这说明在不同的引物长度下均为本文方法生成的引物具有更强的正交性.



(a) 不同引物长度下的平均引物距离



(b) 不同引物长度下的平均引物强度

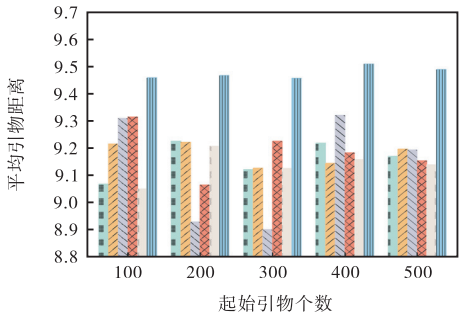
图 4 引物长度对引物正交性的影响

Fig.4 Effect of primer length on primer orthogonality

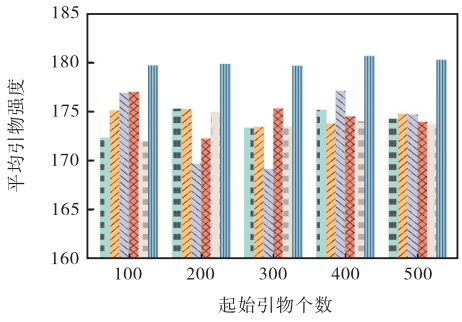
起始引物个数是本文引物设计方法的一个重要参数，它是进行基于网络路径优化的引物筛选前的引物个数. 为了验证不同起始引物个数下引物设计方法的有效性，本文分别采用随机生成法，考虑 C_{GC}、均聚物长度、t_m、ΔG 的随机生成法，以及本文方法进行引物设计，起始引物个数分别为 100、200、300、400、500. 对于上述 30 组引物，分别计算其平均引物距离和平均引物强度，并绘制出如图 5 所示的柱状图.

图 5(a)显示的是 30 组引物的平均引物距离，图 5(b)显示的是 30 组引物的平均引物强度. 与其他方法相比，随机生成法设计出的引物具有最低的平均引物距离和平均引物强度，这是由于该方法没有考虑任何引物设计规则，因此设计出的引物正交性较差. 相反，本文引物设计方法生成的引物具有更大的平均引物距离和平均引物强度，这是由于本文不仅考虑了 C_{GC}、均聚物长度、t_m 和 ΔG 等设计规则，而且通过引物筛选将正交性差的引物筛除，从而提升剩余引物间的正交性.

终止引物个数是指经过基于网络路径优化的引物筛选后剩余引物的个数，为了分析不同终止引物个数下本文方法生成引物的正交性，在起始引物个数为



(a) 不同起始引物个数下的平均引物距离



(b) 不同起始引物个数下的平均引物强度

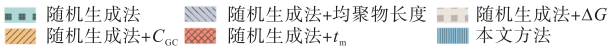


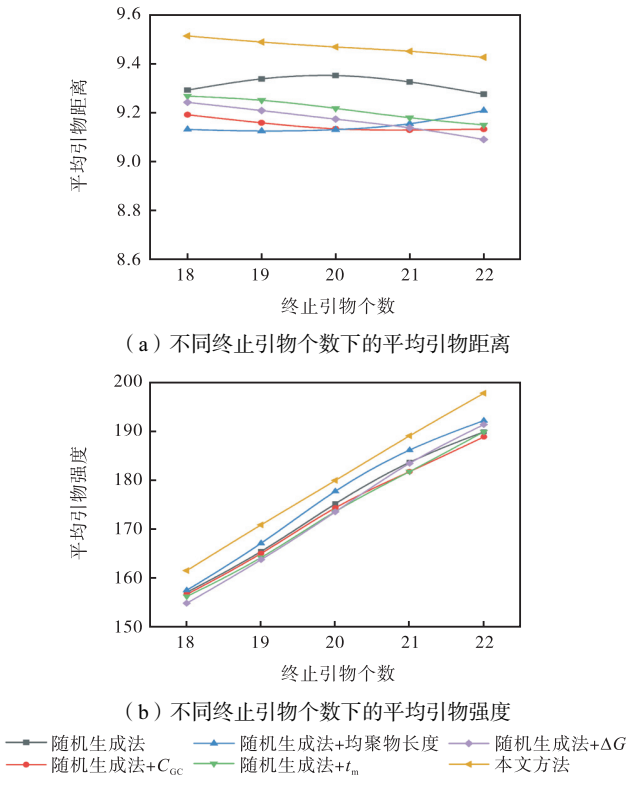
图 5 起始引物个数对引物正交性的影响

Fig.5 Effect of the number of starting primers on primer orthogonality

100 时，分别利用本文方法与 5 种对照方法筛选出 18、19、20、21、22 个终止引物，并均设置 5 个对照组. 计算出上述各组引物的平均引物距离和平均引物强度，其结果如图 6 所示. 图 6(a)显示的是平均引物距离随终止引物个数变化的曲线，图 6(b)显示的是平均引物强度随终止引物个数变化的曲线，其中，实心左三角曲线代表本文方法，其他 5 条曲线分别对应 5 种方法. 在两个曲线图中，均为本文方法对应的曲线位于最上方，这说明本文方法设计的引物具有更强的正交性. 另外，在图 6(a)中，随着终止引物个数的增大，本文方法生成引物的平均引物距离略有下降，这说明终止引物个数越少，本文方法生成引物的正交性越强.

2.3 DNA 存储信息检索仿真

用于 DNA 存储信息检索的 PCR 扩增是一个生化过程，其依赖于分子层面的热运动. 引物之所以能与目标 DNA 序列上的特定位点相结合，是因为它们完全满足碱基互补配对原则，具有很高的 ΔG 值，结合稳定程度高. 由于引物与其他 DNA 序列上的引物结合位点不是完全互补配对的，所以 ΔG 值相对较低，但这并不意味着引物不会与其他 DNA 序列上的引物结合位点结合，只是结合的相对较少. 当引物与



（a）不同终止引物个数下的平均引物距离

（b）不同终止引物个数下的平均引物强度

图 6　终止引物个数对引物正交性的影响

Fig.6　Effect of the number of termination primers on primer orthogonality

其他 DNA 序列的 Δ*G* 值越高时，发生误检的可能性就越高；反之，当引物与其他 DNA 序列的 Δ*G* 值越低时，发生误检的可能性也越低.

本文利用 primer3 软件对 DNA 存储信息检索过程进行模拟，并计算出引物与非目标 DNA 序列间的 Δ*G* 值，绘制出图 7 所示的热力图. 其中，横轴为引物的分组编号，取值范围为 1 ~ 10；纵轴为 DNA 序列的分组编号，编号由左侧的字母和右侧的数字组成，字母 a、b、c、d、e 分别代表长度为 100 nt、150 nt、200 nt、250 nt、300 nt，数字 1 和 2 分别代表引物设计方法为本文方法和随机生成法. 颜色条代表引物与非目标 DNA 序列间的 Δ*G* 值，颜色越深，Δ*G* 值越高. 由图 7 可见，a1、b1、c1、d1、e1 组的颜色条较浅，a2、b2、c2、d2、e2 组的颜色条较深，这说明前者的 Δ*G* 值低于后者，即与随机生成法生成的引物相比，本文方法生成的强正交性引物与非目标 DNA 序列发生错配的可能性更低.

为了验证本文所提引物设计方法与现有 DNA 编码方案结合时均具有较低的误检可能性，分别将文献[8-11，20]中提出的编码方法与本文的引物设计方法相结合，并采用 primer3 软件对信息检索过程进行仿真，计算出引物与非目标 DNA 序列结合的平均 Δ*G* 值与平均 t_m 值，结果如表 2 所示.

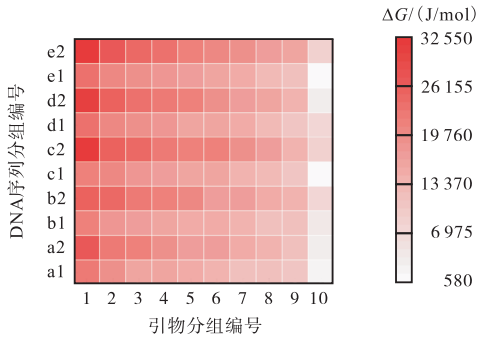


图 7　引物与非目标 DNA 序列间的Δ*G* 值

Fig.7　Δ*G* values between primers and non-target DNA sequences

表 2　不同方法组合下引物与非目标 DNA 序列间的平均 Δ*G* 值与平均 t_m 值

Tab.2　Average Δ*G* values and average t_m values of primers and non-target DNA sequences under different method combinations

DNA 编码方案	引物设计方法	平均 Δ*G* 值/(J/mol)	平均 t_m 值/℃
文献[8]	本文方法	16 274.88	−10.10
	传统方法	18 067.22	4.38
文献[9]	本文方法	15 132.23	−10.26
	传统方法	17 518.99	6.13
文献[10]	本文方法	16 508.26	−8.51
	传统方法	17 599.03	2.27
文献[11]	本文方法	15 512.89	−8.42
	传统方法	18 757.71	5.19
文献[20]	本文方法	14 825.79	−13.12
	传统方法	18 120.65	4.27

由表 2 可知，在与上述经典 DNA 编码方案结合时，与传统方法设计的引物相比，本文方法所设计的引物与非目标 DNA 序列之间的平均 Δ*G* 值及平均 t_m 值均更低，这说明本文引物设计方法在与各种经典 DNA 编码方案结合时均具有较低的误检可能性.

3　结　语

本文提出了一种实现 DNA 存储低串扰信息检索的引物设计方法，其采用加权汉明距离衡量引物间的正交性，弥补了直接用汉明距离时无法考虑引物中不同碱基位置差异对引物间正交性影响的缺陷. 此外，采用基于网络路径优化的引物筛选策略，可将弱正交性引物剔除，从而提升剩余引物间的正交性. 实验结果表明，与汉明距离相比，本文所提加权汉明距离与引物间正交性具有更高的适配性；采用本文方法设计的引物与传统方法相比具有更强的正交性；将所设计引物用于 DNA 存储信息检索时可降低信息误检的可能性，有助于推动 DNA 存储的实用化进程.

参考文献：

[1] Xu Q，Lu Z H，Bi K. DNA-LSIED：DNA lossy storage for images by encryption and corrective denoising method[J]. Signal，Image and Video Processing，2025，19：11.

[2] Zheng Y F，Cao B，Zhang X K，et al. DNA-QLC：An efficient and reliable image encoding scheme for DNA storage[J]. BMC Genomics，2024，25：266.

[3] 刘彦军，杨越飞，胡迎新. 基于非天然核酸的高密度 DNA 存储编码方法[J]. 生物信息学，2026，24(1)：70-84.
Liu Yanjun，Yang Yuefei，Hu Yingxin. High-density DNA storage encoding method based on unnatural nucleic acids[J]. Chinese Journal of Bioinformatics，2026，24(1)：70-84(in Chinese).

[4] 张宣梁，李青婷，王 飞. DNA 存储系统中的数据写入[J]. 合成生物学，2024，5(5)：1125-1141.
Zhang Xuanliang，Li Qingting，Wang Fei. Data writing in DNA storage systems[J]. Synthetic Biology Journal，2024，5(5)：1125-1141(in Chinese).

[5] Wang K，Cao B，Ma T，et al. Storing images in DNA via base128 encoding[J]. Journal of Chemical Information and Modeling，2024，64(5)：1719-1729.

[6] Seo S，Tandon A，Lee K W，et al. Information density enhancement using lossy compression in DNA data storage[J]. Advanced Materials，2025，37：2403071.

[7] Rasool A，Hong J W，Hong Z L，et al. An effective DNA-based file storage system for practical archiving and retrieval of medical MRI data[J]. Small Methods，2024，8(10)：2301585.

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

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

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

[12] Yazdi S M H T，Yuan Y B，Ma J，et al. A rewritable，random-access DNA-based storage system[J]. Scientific Reports，2015，5(1)：14138.

[13] 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.

[14] Song X，Shah S，Reif J. Multidimensional data organization and random access in large-scale DNA storage systems[J]. Theoretical Computer Science，2021，894：190-202.

[15] 张淑芳，李予辉，李炳志. DNA 存储场景下基于引物索引矩阵的文件高效随机检索方法[J]. 电子与信息学报，2024，46(6)：2568-2577.
Zhang Shufang，Li Yuhui，Li Bingzhi. Efficient file random access method based on primer index matrix in DNA storage scenarios[J]. Journal of Electronics and Information Technology，2024，46(6)：2568-2577(in Chinese).

[16] Wang Q，Zhang S F，Li Y H. Efficient DNA coding algorithm for polymerase chain reaction amplification information retrieval[J]. International Journal of Molecular Sciences，2024，25(12)：6449.

[17] Newman S，Stephenson A P，Willsey M，et al. High density DNA data storage library via dehydration with digital microfluidic retrieval[J]. Nature Communications，2019，10(1)：1706.

[18] Piantanida L，Hughes W L. A PCR-free approach to random access in DNA[J]. Nature Materials，2021，20(9)：1172-1178.

[19] Bee C，Chen Y J，Queen M，et al. Molecular-level similarity search brings computing to DNA data storage[J]. Nature Communications，2021，12(1)：4764.

[20] Ping Z，Chen S H，Zhou G Y，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.

(责任编辑：孙立华)