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：**
袁涛, 曲强, 姜青山. 基于混沌系统和喷泉码的 DNA 加密编码方法 [J]. 集成技术, 2024, 13(3): 4-24.
Yuan T, Qu Q, Jiang QS. An encrypted DNA encoding method based on chaotic system and fountain code [J]. Journal of Integration Technology, 2024, 13(3): 4-24.

# 基于混沌系统和喷泉码的 DNA 加密编码方法

袁 涛[1,2] 曲 强[2*] 姜青山[2]

[1]（南方科技大学 深圳 518055）
[2]（中国科学院深圳先进技术研究院 深圳 518055）

**摘 要** 在这个海量数据时代，DNA 是一种很好的新信息存储媒介。与传统的物理存储介质相比，它具有能耗低、存储密度高、存储寿命长等固有的优点。随着 DNA 存储技术的快速发展，如何保障新技术下的信息安全至关重要。为此，该文结合加密领域研究和 DNA 编码领域研究，提出了一种基于混沌系统和喷泉码的 DNA 加密编码方法，利用混沌系统加密原理，在 DNA 喷泉码编码过程中进行加密，在保留 DNA 喷泉码特性的同时，保障了编码信息的安全性。该方法可用于任意类型数据，可实现高信息密度和任意约束条件的 DNA 编码。同时，通过仿真实验证明，该方法可以有效抵抗多种密码学攻击，并对 DNA 存储过程产生的数据错误有一定纠错能力。

**关键词** DNA 存储；加密；喷泉码；信息安全
**中图分类号** TP 301；Q 819 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20231101001

# An Encrypted DNA Encoding Method Based on Chaotic System and Fountain Code

YUAN Tao[1,2] QU Qiang[2*] JIANG Qingshan[2]

[1]( *Southern University of Science and Technology*, *Shenzhen* 518055, *China* )
[2]( *Shenzhen Institute of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China* )
[*]*Corresponding Author*: *qiang@siat.ac.cn*

**Abstract** In this era of massive data, DNA serves as a promising new medium for information storage. Compared to traditional physical storage media, it possesses inherent advantages such as low energy consumption, high storage density, and long storage lifespan. With the rapid development of DNA storage, ensuring information security under new technologies becomes crucial. In this regard, this paper combines research in the fields of encryption and DNA coding, proposing a DNA encryption coding method based on chaotic systems and fountain codes. The encryption principle of chaotic systems is utilized during the DNA

**收稿日期**：2023-11-01 **修回日期**：2024-03-01
**基金项目**：国家重点研发计划项目（2020YFA0909100, 2021YFF1200100, 2021YFF1200104）
**作者简介**：袁涛，硕士研究生，研究方向为 DNA 存储；曲强（通讯作者），研究员，研究方向为区块链、DNA 存储，E-mail：qiang@siat.ac.cn；姜青山，研究员，研究方向为数据挖掘、DNA 存储。

fountain code encoding process, preserving the characteristics of DNA fountain codes while ensuring the security of encoded information. This method is applicable to any types of data, achieving high information density and DNA encoding under arbitrary constraints. Furthermore, through simulation experiments, it is demonstrated that this method can effectively resist various cryptographic attacks and possesses error-correction capabilities for data errors generated during the DNA storage process.

**Keywords**　DNA storage; encryption; fountain code; information security

**Funding**　This work is supported by the National Key Research and Development Program of China (2020YFA0909100, 2021YFF1200100, 2021YFF1200104)

## 1　引　言

随着信息技术的高速发展，全球每天产生的数据量将呈指数级增长。国际数据公司预计，2025 年，全球数据规模将达到 175 ZB[1]。DNA 是能解决海量数据存储问题的存储介质之一，其具有高密度、长耐久性和低维护成本的特性[2-4]。用 DNA 存储数据的研究成为当前计算机和生物交叉领域的研究热点[5]。作为相关领域，DNA 加密也受到越来越多研究者的关注。

近年来，基于 DNA 的加密方法主要分为两种类型。一种方法是利用 DNA 分子的生化特性进行加密。2014 年，Yang 等[6]提出了一种利用 DNA 自组装结构的 DNA 加密方法——利用 DNA 链识别和链置换模拟数据按位异或操作，从而以一次性密码本的形式完成信息的加密和解密。2016 年，Zakeri 等[7]通过大量的短串联重复 DNA 序列生成高强度密钥进行加密，其将少量密文及密钥编码成 DNA 分子后，隐藏在多个无关 DNA 分子中间，通过特定测序实现解密。2019 年，Zhang 等[8]提出了用于安全通信的 DNA 折纸密码技术。该团队利用 M13 病毒蛋白骨架的折叠，形成了纳米级自组装的盲文样式，而 DNA 折纸的固有纳米级寻址性还允许蛋白质结合隐写术，进一步保护了信息的机密性。2022 年，Zhu 等[9]提出了一种可操作的 DNA 链置换加密方法。他们在每次加密中调整生化反应的参数，并根据 DNA 链的浓度变化获得密钥。另一种方法是通过 DNA 碱基排序和编码/解码设计加密算法。图像加密领域的相关研究[10-13]通常将明文和伪随机序列编码为 DNA 序列，然后用碱基计算规则操作两个 DNA 序列，最后转化为二进制密文，将 DNA 序列作为中间结果。而在 DNA 存储领域，2020 年，Grass 等[14]将代表身份的序列编码为用于高级加密标准（advanced encryption standard，AES）加密的密钥，使用 AES 加密文本，再进行 DNA 编码，同时引入 Reed-Solomon（RS）码等纠错技术，确保了密文的可靠性。2021 年，Peng 等[15]提出了一种基于混淆映射和 DNA 存储技术的一次性密码本算法，利用编码映射和混淆编码表实现了数据和生物信息的转换。2023 年，Zan 等[16]提出适用于 DNA 存储的大规模数据图像加密方法，在保证安全性的同时，增大了应用数据的规模。同年，Yao 等[17]提出了一种基于基因杂交和基因突变的面向 DNA 存储的图像加密算法和一种基于前向纠错码的图像 DNA 加密存储算法[18]，两者都具有较高的鲁棒性，可在图像密文部分丢失的情况下解密为高度相似的明文图像。

由于当前技术水平限制，DNA 序列在合

成、存储和测序过程中可能发生序列丢失和碱基错误问题，因此，将数据转化为 DNA 序列时应满足一定的约束条件和信息密度要求，且 DNA 存储可能面临应用数据规模大和类型多的情景，在这些问题上，上述 DNA 加密方法不能统一解决。为此，本文提出了一种基于混沌系统和 DNA 喷泉码的加密编码方法(DNA chaos-fountain encoding，DCFE)，利用超混沌系统加密原理，在喷泉码编码过程中进行加密，以保证信息安全，同时实现高信息密度和具备纠错特性的 DNA 编码，可用于任意数据规模和类型。

## 2 相关工作

### 2.1 DNA 喷泉码

Erlich 等[19]是第一个将喷泉码应用于 DNA 存储的研究团队，实现了理论密度接近香农极限的 DNA 编码，具有一定的纠错能力，可以满足 DNA 存储所需的生物约束条件。喷泉码是一种无码率擦除码，源自擦除信道传输，因其具备与喷泉相似的特性而得名。在喷泉码传输中，发送端的编码器将 $n$ 个源符号转换为无限个编码符号，并将它们分组发送，与码率无关。接收端的解码器可以通过接收一定数量的编码符号而成功解码消息，不管这些符号来自消息的哪一部分，都可以高概率地恢复源符号。

Erlich 使用了 Luby 提出的基本喷泉码的 LT 码(Luby transform code)[20]。LT 码编码的主要原理为通过度分布函数(固定概率分布随机函数)生成度值，根据度值生成多个随机数，选择对应的信息符号进行异或操作得到编码符号。其理论上可生成无限数量的编码符号，具体算法伪代码如表 1 所示。

Shokrollahi[21]提出 LT 码的一种优化——Raptor 码，其本质上是低密度奇偶校验(low density parity check，LDPC)码与 LT 码的结合，

**表 1 LT 码编码算法**

**Table 1 LT code encode algorithm**

| Algorithm 1: LT 码编码算法 |
|---|
| Input: $n$ 个信息符号，编码符号数量 $N$<br>Output: $N$ 个编码符号 |
| 1: 选择度分布函数<br>2: while 编码符号数量未达到 $N$ do<br>3: 根据度分布函数生成一个随机数 $R$<br>4: 生成 $R$ 个 1 到 $n$ 的范围内的不同随机数<br>5: for 每个随机数 do<br>6: 从 $n$ 个信息符号中选择随机数对应的符号<br>7: end for<br>8: 将选择的 $R$ 个信息符号异或得到 1 个编码符号<br>9: end while |

可以解决 LT 码的不足，是喷泉码在 DNA 存储领域的进一步研究目标。其用于 DNA 编码中可实现高信息密度(每一碱基对应的二进制位数)、强纠错能力和满足任意约束的特性。

DNA 存储中的 Raptor 码编码[22]过程如图 1 所示，主要由 3 个部分组成：预编码、LT 编码和 DNA 映射。在预编码部分，使用 LDPC 码将纠错位添加到信息符号中，以生成中间符号。然后对这些中间符号进行 LT 编码，生成结果块。最后，将结果块进行 DNA 映射(即二进制到 DNA 4 种碱基 ACGT 的转化)，筛选符合自定义条件的 DNA 链。其中，因为中间符号包含纠错位，所以，在解码时无须恢复全部信息符号，使解码复杂度接近线性关系，同时降低了 LT 编码过程中的度分布要求，保证了编码结果的平均度值的恒定。

RFC(Raptor forward error correction)5053[23]是 Raptor 码的更高性能实现，Schwarz 等[24]完成了其在 DNA 存储中的运用，其与 Raptor 码的主要差异为预编码过程，后续的 LT 编码和 DNA 映射原理相同。

RFC 5053 的预编码过程如下：以 $K$ 个源符号作为一个源块，将 $K$ 个源符号前添加($S+H$)位零元素，组成扩展源符号 $\boldsymbol{D}$，并作为输入符号($S$ 和 $H$ 值由固定表给出)，将输入扩展源符号 $\boldsymbol{D}$ 与编码约束矩阵 $\boldsymbol{A}$ 的逆矩阵相乘，可以得到 $L$ 个

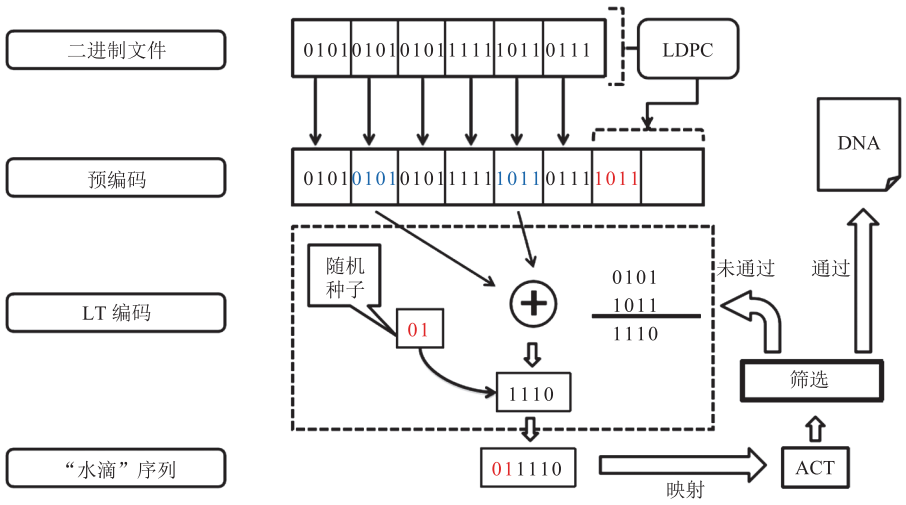



**图 1　DNA Raptor 码编码过程**

**Fig. 1　DNA Raptor code encoding process**

预编码的中间符号 $\boldsymbol{C}$，如公式(1)所示。

$$\boldsymbol{C}=\boldsymbol{A}^{-1}\boldsymbol{D} \tag{1}$$

该约束矩阵 $\boldsymbol{A}$ 如图 2 所示。

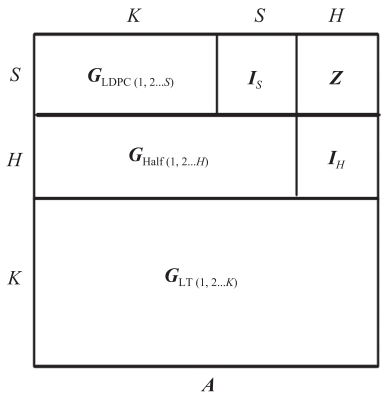



**图 2　约束矩阵 $\boldsymbol{A}$**

**Fig. 2　Constraint matrix $\boldsymbol{A}$**

图 2 中，$\boldsymbol{G}_{\mathrm{LDPC}}$ 为 LDPC 符号的 $S\times K$ 维生成矩阵；$\boldsymbol{G}_{\mathrm{Half}}$ 为 Half 符号(格雷码)的 $H\times(K+S)$ 维生成矩阵；$\boldsymbol{I}_S$ 为 $S\times S$ 维单位矩阵；$\boldsymbol{I}_H$ 为 $H\times H$ 维单位矩阵；$\boldsymbol{Z}$ 为 $S\times H$ 维零矩阵；$\boldsymbol{G}_{\mathrm{LT}}$ 为 LT 编码符号的 $K\times L$ 维生成矩阵。

## 2.2　混沌系统

自 1984 年 Matthews[25]提出了基于混沌的加密方案以来，混沌密码学逐渐发展为密码学的新分支。在安全性方面，与传统的加密系统相比，高维混沌系统已被证明具备更好的表现，特别是在图像加密领域。近年来，新提出的基于混沌系统的加密算法[26-28]都已体现出对各类攻击的优良抵抗能力。混沌系统加密的关键之一是混沌方程的确定，由混沌方程得出的正李雅普诺夫指数的数量和大小能够初步衡量混沌系统的复杂性，正李雅普诺夫指数的数量越多，其值越大，系统的复杂性越高，随机性也越强[29]。

经典 Chen 等[30]混沌系统和在其基础上改进的 Liu 等[31]混沌系统的公式分别如公式(2)和公式(3)所示。

$$\begin{cases}\dot{x}=a\cdot x+k\cdot y-y\cdot z\\ \dot{y}=-b\cdot y-z+x\cdot z\\ \dot{z}=-x-c\cdot z+x\cdot y\end{cases} \tag{2}$$

其中，$a$、$b$、$c$、$k$ 为常数参数；$x$、$y$、$z$ 为状态参数。系统是三维混沌系统，常数参数设置为 $a=4.6$，$b=12$，$c=5$，$k=1$。

$$\begin{cases}\dot{x}=a\cdot(y-x)+g\cdot y\cdot z\\ \dot{y}=c\cdot x+d\cdot y-x\cdot z-w\\ \dot{z}=-b\cdot z+x\cdot y\\ \dot{w}=r\cdot w+h\cdot y\end{cases} \tag{3}$$

其中，$a$、$b$、$c$、$d$、$g$、$h$、$r$ 为常数参数；$x$、$y$、$z$、$w$ 为状态参数。系统是四维超混沌系统，

常数参数设置为 $a=14$，$b=43$，$c=-1$，$d=16$，$g=4$，$h=4.9$，$r=-0.07$。新状态变量 $w$ 是一个线性状态反馈控制器，用于调整(放大或缩小)原始状态变量 $y$ 的变化。

## 3 五维超混沌伪随机序列生成器

本文构建了一种新的、用于加密的五维超混沌系统，与经典的三、四维超混沌系统相比，其在满足超混沌系统定义的前提下具备更大的李雅普诺夫指数，代表着更高的混沌系统复杂性。同时，维数扩增带来更多的状态参数，更高的安全性。对该五维超混沌系统添加运算扰动和数值转化操作，得到可生成大规模密钥流的五维超混沌伪随机序列生成器，在 DNA 喷泉码编码过程中生成密钥流对数据进行加密。

### 3.1 新五维超混沌系统构建

基于经典 Chen 等[30]混沌系统和 Liu 等[31]混沌系统，本文构建了一种新的、用于加密的五维超混沌系统，其具备更大的李雅普诺夫指数和维数，如公式(4)所示：

$$\begin{cases}\dot{x}=a\cdot(y-x)+g\cdot y\cdot z+q\cdot u\\ \dot{y}=d\cdot y+c\cdot x-x\cdot z-k\cdot w\\ \dot{z}=-b\cdot z+x\cdot y+q\cdot w\\ \dot{w}=r\cdot w+h\cdot y-z\\ \dot{u}=m\cdot u+p\cdot x\end{cases} \tag{4}$$

其中，$a$、$b$、$c$、$d$、$g$、$h$、$r$、$m$、$k$、$p$ 和 $q$ 为常数参数；$x$、$y$、$z$、$w$ 和 $u$ 为状态参数。主要的思想是向 Liu 等[31]混沌系统中添加线性状态反馈控制器，调整各状态变量间的关系，使得所有的状态都与控制器相关，增加系统的李雅普诺夫指数数量，以获得一个更好的混沌系统。

### 3.2 混沌系统性能评估

混沌系统的性能可以从李雅普诺夫指数、李雅普诺夫维数、混沌吸引子、耗散性、平衡点 5 个方面进行评估。

(1)李雅普诺夫指数。李雅普诺夫指数是表征相空间中相邻轨迹分离率的度量。李雅普诺夫指数的数量≥2 的混沌系统为超混沌系统。李雅普诺夫指数越大，表明系统具有越强的混沌行为，轨迹表现出更大的复杂性和不可预测性。李雅普诺夫指数的计算公式如下：

$$\lambda_{\mathrm{L}i}=\lim_{T\to\infty}\frac{1}{T}\ln\left\|\frac{\mathrm{d}\boldsymbol{y}(T)}{\mathrm{d}\boldsymbol{y}_0}\right\| \tag{5}$$

其中，$\lambda_{\mathrm{L}i}$ 为第 $i$ 个李雅普诺夫指数；$\boldsymbol{y}(T)$为时间 $T$ 时的状态向量；$\boldsymbol{y}_0$ 为初始状态向量。

(2)李雅普诺夫维数。混沌系统的结构不同于常规系统，其表现出分形性质，可利用李雅普诺夫维数对系统的分形维数进行量化。将一个 $M$ 维的李雅普诺夫维数记作 $D_{\mathrm{L}}$，其计算公式为

$$D_{\mathrm{L}}=l+\frac{1}{\left|\lambda_{\mathrm{L}l+1}\right|}\sum_{i=1}^{l}\lambda_{\mathrm{L}i} \tag{6}$$

其中，$D_{\mathrm{L}}$ 为李雅普诺夫维数；$l$ 为使得 $\sum_{i=1}^{l}\lambda_{\mathrm{L}i}>0$ 的最大整数；$\lambda_{\mathrm{L}i}$ 为第 $i$ 个李雅普诺夫指数。

(3)混沌吸引子。在混沌系统中，随着时间的推移，系统轨迹的演化趋于收敛于某一点。如果这个点与初始状态无关，则将它称为吸引子。通过各参数的相位图来展现混沌吸引子。

(4)耗散性。混沌系统不仅具有收敛性，还具有耗散性。耗散的性质决定了一个混沌系统最终是否收敛。五维超混沌系统的耗散性计算如公式(7)所示：

$$\Delta V=\frac{\partial\dot{x}}{\partial x}+\frac{\partial\dot{y}}{\partial y}+\frac{\partial\dot{z}}{\partial z}+\frac{\partial\dot{w}}{\partial w}+\frac{\partial\dot{u}}{\partial u} \tag{7}$$

其中，当 $\Delta V<0$ 时，系统的轨迹最终将收敛到单个点或一个集合，而混沌运动将被限制在一个吸引子内，这表明该系统具有耗散性，其收敛速率为 $\mathrm{e}^{\Delta Vt}$。

(5)平衡点。五维超混沌系统的平衡点 $\boldsymbol{O}$ 通过公式(8)求解得到，用于评价系统的稳定性。得到平衡点 $\boldsymbol{O}$ 后，计算相应的雅可比矩阵，求出

矩阵特征值，通过特征值中是否存在正的特征值来判断系统在平衡点 $\boldsymbol{O}$ 处是否稳定。

$$\begin{cases} a\cdot(y-x)+g\cdot y\cdot z+q\cdot u=0 \\ d\cdot y+c\cdot x-x\cdot z-k\cdot w=0 \\ -b\cdot z+x\cdot y+q\cdot w=0 \\ r\cdot w+h\cdot y-z=0 \\ m\cdot u+p\cdot x=0 \end{cases} \tag{8}$$

### 3.3　新五维超混沌系统参数确立

本文提出的新五维超混沌系统的具体常数参数，是在保证系统混沌特性的基础上通过搜索得到的最优参数值。

基于经典 Chen 等[30]和 Liu 等[31]混沌系统参数值，可确定部分参数（$a$、$b$、$c$、$d$、$g$、$h$、$r$）的可用初始值，对应参数的最优值从该值开始搜索。由于增加的线性状态反馈控制器 $w$ 和 $u$ 的主要作用是调整（放大或缩小）原始状态变量 $x$、$y$、$z$ 的变化，其相关的对应参数（$m$、$k$、$p$、$q$）的最优值从 0 开始搜索。

通过大量计算得出，当 $a\in[10,27]$，$b\in[35,60]$，$c\in[-5,3]$，$d\in[10,24]$，$g\in[-2,10]$，$h\in[-10,10]$，$r\in[-5,5]$，$m\in[-5,5]$，$k\in[-10,10]$，$p\in[-10,10]$，$q\in[-10,10]$时，系统能较好地保持混沌特性（满足混沌系统公式可解的条件，属于超混沌系统，具有正确的李雅普诺夫指数）。其中，前 5 项为决定混沌系统状态的主要参数，后 6 项为增强混沌系统性能的状态反馈参数。

通过粒子群优化算法对本文提出的系统常数参数进行优化，找出最佳参数值。粒子群优化算法的基本原理为初始化一组粒子 $R=\{x_1, x_2, \cdots, x_J\}$，其中，假设粒子数为 $J$，并建立目标函数 $f(x_i)$。每个粒子对应一个目标函数 $f(x_i)$，每个粒子有速度 $\boldsymbol{v}_i(t)$和位置 $\boldsymbol{x}_i(t)$两个属性，通过更新粒子的速度和位置进行迭代，并记录粒子的个体极值 Pbest 和群极值 Gbest，寻找能使目标函数达到最佳的粒子。其中，速度和位置的更新分别如公式（9）和公式（10）所示：

$$\boldsymbol{v}_i(t+1)=\omega\boldsymbol{v}_i(t)+ c_1r_1\left(\text{Pbest}_i(t)-\boldsymbol{x}_i(t)\right)+c_2r_2\left(\text{Gbest}(t)-\boldsymbol{x}_i(t)\right) \tag{9}$$

其中，$\omega$ 为惯性权重；$r_1$ 和 $r_2\in[0,1]$，为随机常数；$c_1$ 和 $c_2$ 为常数。更新迭代的速度和位置具有限制，需预先设置，如 $\boldsymbol{v}_i(t)\in[v_{\min}, v_{\max}]$，$\boldsymbol{x}_i(t)\in[x_{\min}, x_{\max}]$。

$$\boldsymbol{x}_i(t+1)=\boldsymbol{x}_i(t)+\boldsymbol{v}_i(t+1) \tag{10}$$

对应于本文提出的五维超混沌系统，位置 $\boldsymbol{x}_i(t)$即为每个混沌公式的常数参数，目标函数即为对应的李雅普诺夫指数最大值，相应的伪代码见表 2。

限定参数范围，在 10 个粒子 100 次迭代下，确立最佳参数为 $a=14.609\,5$，$b=46.740\,3$，$c=0.222\,0$，$d=17.036\,7$，$g=5.764\,6$，$h=4.893\,2$，$r=0.053\,3$，$m=-0.861\,4$，$k=2.906\,9$，$p=-5.561\,8$，$q=5.705\,7$。该参数下具备最大的李雅普诺夫指数，同时满足其他混沌系统评估标准。

**表 2　粒子群优化算法**

**Table 2　Particle swarm optimization algorithm**

Algorithm 2: 粒子群优化算法

Input: 粒子数量 $J$，迭代次数 $\beta$
Output: 最佳个体常数参数

```
初始化粒子群(J 组混沌系统公式)
for 每个粒子 do
    初始化混沌系统常数参数和速度
    计算粒子的适应度
    更新个体最佳常数参数
end for
更新全局最佳常数参数
while 未达到迭代次数 β do
    for 每个粒子 do
        更新常数参数和速度
        计算粒子的适应度
        更新个体最佳常数参数
    end for
    更新全局最佳常数参数
end while
```

最佳参数下对应的李雅普诺夫指数如下：$\lambda_{L1}=7.766$，$\lambda_{L2}=2.534$，$\lambda_{L3}=-3.230$，$\lambda_{L4}=-15.589$，$\lambda_{L5}=-39.035$。在这 5 个李雅普诺夫指数中，有 2 个为正（正值的数量大于 1 即为超混沌系统），3 个为负，总数量为 5，所以这是一个五维的超混

沌系统。表 3 展示了几种不同混沌系统的李雅普诺夫指数。

代入实际值到公式(6)，经计算，李雅普诺夫维数为

$$D_{\mathrm{L}}=3+\frac{\lambda_{\mathrm{L1}}+\lambda_{\mathrm{L2}}+\lambda_{\mathrm{L3}}}{|\lambda_{\mathrm{L4}}|}=3.453\,5 \tag{11}$$

结果表明，该系统是混沌的($D_{\mathrm{L}}$ 不是一个整数)。

计算得到的不同相位平面混沌吸引子如图 3 所示。各个混沌吸引子都表现出复杂的动力学行为。

代入实际值到公式(7)，得耗散性值：

$$\begin{aligned}\Delta V&=\frac{\partial\dot{x}}{\partial x}+\frac{\partial\dot{y}}{\partial y}+\frac{\partial\dot{z}}{\partial z}+\frac{\partial\dot{w}}{\partial w}+\frac{\partial\dot{u}}{\partial u}\\&=-a+d-b+r+m\\&=-45.121\,2<0\end{aligned} \tag{12}$$

由于 $\Delta V<0$，所以该系统具有耗散性，其收

**表 3　不同混沌系统的李雅普诺夫指数**

**Table 3　The Lyapunov exponents of different chaotic systems**

| 混沌系统 | $\lambda_{\mathrm{L1}}$ | $\lambda_{\mathrm{L2}}$ | $\lambda_{\mathrm{L3}}$ | $\lambda_{\mathrm{L4}}$ | $\lambda_{\mathrm{L5}}$ |
|---|---|---|---|---|---|
| Lorenz[32] | 1.497 | 0 | −22.460 | | |
| Chen 等[30] | 2.019 | 0 | −12.022 | | |
| Liu 等[31] | 4.998 | 0.152 | −0.026 | −45.816 | |
| Gao 等[33] | 3.012 | 1.943 | | | |
| 本文 | 7.766 | 2.534 | −3.230 | −15.589 | −39.035 |

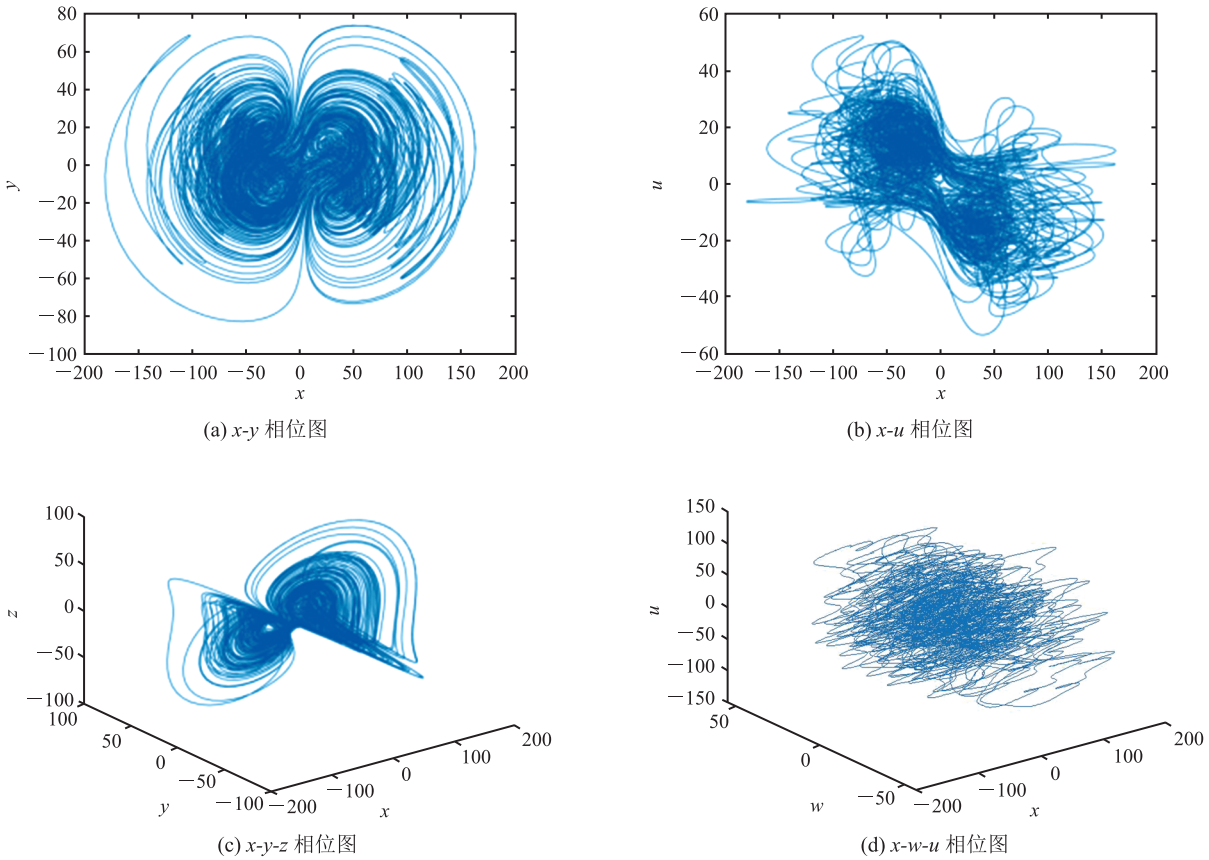



**图 3　各个平面混沌吸引子图**

**Fig. 3　Chaotic attractor diagrams in various planes**

敛速率为 $e^{-45.121\,2t}$。

基于公式(8)，代入实际值计算得到平衡点 $\boldsymbol{O}(0,0,0,0,0)$，相应的雅可比矩阵为

$$\mathrm{Jaco}=\begin{bmatrix} -a & a+g\cdot z & g\cdot y & 0 & q \\ c-z & d & -x & -k & 0 \\ y & x & -b & q & 0 \\ 0 & h & -1 & r & 0 \\ p & 0 & 0 & 0 & m \end{bmatrix} \tag{13}$$

$$=\begin{bmatrix} -14.609\,5 & 14.609\,5 & 0 & 0 & 5.705\,7 \\ 0.222\,0 & 17.036\,7 & 0 & -2.906\,9 & 0 \\ 0 & 0 & -46.740\,3 & 5.705\,7 & 0 \\ 0 & 4.893\,2 & -1 & 0.053\,3 & 0 \\ -5.561\,8 & 0 & 0 & 0 & -0.861\,4 \end{bmatrix}$$

矩阵的特征方程如公式(14)所示($\boldsymbol{I}$ 为单位矩阵)：

$$\det(\lambda \boldsymbol{I}-\mathrm{Jaco})=0 \tag{14}$$

将参数代入，并进行简化，得到:

$$(\lambda+46.618\,6)(\lambda+11.822\,1)(\lambda+3.748\,2)(\lambda-0.804\,5)(\lambda-16.263\,2)=0 \tag{15}$$

在结果中，有 3 个负值和 2 个正值，即 −46.618 6、−11.822 1、−3.748 2 和 0.804 5、16.263 2。由于计算得到的特征值中存在正的特征值，因此可以得出系统在平衡点 $\boldsymbol{O}$ 处是不稳定的。

### 3.4　超混沌伪随机序列生成器

为了真正利用混沌系统进行数据加密，需要构造一个伪随机序列生成器，将混沌系统的多个状态变量转换为用于加密的伪随机序列。其输入参数为 5 个状态变量值的初始值，将这 5 个值作为加密的密钥。伪随机序列生成的流程如下。

(1)初始化：设置混沌系统的状态变量的初始值和序列长度，不断进行混沌系统运算。

(2)引入扰动：每运算 500 次，基于 $z_t$ 值来调整 $x_t$ 和 $y_t$，进行第一次扰动操作，如公式(16)。

$$\begin{cases} x_t=x_t+1, y_t=y_t-1, \\ \quad \text{if } z_t\geqslant 0, t\equiv 1 \bmod 500 \\ x_t=x_t-1, y_t=y_t+1, \\ \quad \text{if } z_t\leqslant 0, t\equiv 1 \bmod 500 \end{cases} \tag{16}$$

此外，每运算 1 000 次，基于 $z_t$ 值来调整 $w_t$ 和 $u_t$，进行第二次扰动操作，如公式(17)。

$$\begin{cases} w_t=2w_t, u_t=u_t/2, \\ \quad \text{if } z_t\geqslant 0, t\equiv 1 \bmod 1\,000 \\ w_t=w_t/2, u_t=2u_t, \\ \quad \text{if } z_t\leqslant 0, t\equiv 1 \bmod 1\,000 \end{cases} \tag{17}$$

其中，$x_t$、$y_t$、$z_t$、$w_t$、$u_t$ 为混沌系统产生的状态变量；$t$ 为混沌系统运算的次数。

(3)状态变量选择：从固定间隔后的混沌系统运算结果中选择特定数量的状态变量。本文选择保留每 50 个数值中的前 30 个数值，舍弃后 20 个数值。

(4)将实数转换为整数：提取 5 个实数的小数部分，然后应用缩放因子处理大数，得到均匀缩放的数字序列，如公式(18)：

$$\begin{cases} x_t=(x_t-\mathrm{floor}(x_t))\cdot 10^{12} \\ y_t=(y_t-\mathrm{floor}(y_t))\cdot 10^{12} \\ z_t=(z_t-\mathrm{floor}(z_t))\cdot 10^{12} \\ w_t=(w_t-\mathrm{floor}(w_t))\cdot 10^{12} \\ u_t=(u_t-\mathrm{floor}(u_t))\cdot 10^{12} \end{cases} \tag{18}$$

其中，floor( )为向下取整函数。

(5)异或操作：对 $x_t$、$y_t$、$z_t$、$w_t$、$u_t$ 进行异或操作，得到一个 $\mathrm{key}_t$ 值，加入最终的伪随机数序列。取 $x_t$ 的第 1～3 位，$y_t$ 的第 4～9 位，$z_t$ 的第 1～6 位，$w_t$ 的第 7～9 位，$u_t$ 的第 10～12 位，其他位置为 0，$\mathrm{key}_t$ 值计算公式为

$$\mathrm{key}_t=x_{t,1\sim 3}\oplus y_{t,4\sim 9}\oplus z_{t,1\sim 6}\oplus w_{t,7\sim 9}\oplus u_{t,10\sim 12} \tag{19}$$

利用五维超混沌系统的运算和附加的扰动操作，本文构建了一个具有增强随机性和有限范围的伪随机数序列。伪随机序列生成算法的伪代码见表 4。

## 4　DNA 加密编码方法

通过结合超混沌系统的强伪随机性和 Raptor 码在 DNA 存储领域的高性能编码能力，本文提出了一种为 DNA 存储提供安全和高效编码的新加密编码方法——DNA chaos-fountain encoding

(DCFE)。DCFE 方法有以下优势。

**表 4 伪随机序列生成算法**

**Table 4 Pseudorandom sequence generation algorithm**

Algorithm 3: 伪随机序列生成算法

Input: 混沌系统的状态变量的初始值：$x$、$y$、$z$、$w$、$u$；序列长度 len
Output: 伪随机序列：$key_1$ ... $key_n$

```
初始化混沌系统常数参数：a、b、c、d、e、h、r、m、k、p、q;
while 密钥流长度未达到 len do
    固定间隔进行扰动
    根据混沌系统公式运算指定次数
    选择指定数量状态变量
        for 每组状态变量 do
            将实数转换为定长整数
            5 个整数部分异或得到一个 key_i 值
            该 key_i 值加入伪随机序列
        end for
end while
```

(1) 强安全性：本文通过超混沌系统构建了一个强随机性的、用于加密的伪随机序列，同时结合喷泉码的随机化特性进行编码，保证了存储的数据的安全性。与常规的加密方法相比，本方法的密钥空间更大，密文抗攻击能力更强，更能保障各类型数据在 DNA 存储中的安全。

(2) 自纠错：Raptor 码的使用提升了 DNA 数据的自纠错能力。Raptor 码可用于解决 DNA 序列丢失错误，配合 RS 码(用于纠正 DNA 序列中的替换错误)，确保了所存储数据的完整性和准确性。

(3) 任意约束：可自定义编码的约束条件，通过筛选生成满足条件的 DNA 链。生物约束指 DNA 分子在编码时应满足的碱基序列要求，否则，DNA 链在合成、存储和测序过程易发生错误(碱基突变或丢失)。其中，均聚物(DNA 链中连续的重复碱基数量)和 GC 含量(DNA 4 种碱基 A、C、G、T 中的 G 和 C)是编码时主要考虑的约束条件，在合成、存储和测序时，显著影响 DNA 序列产生错误的概率。其中，均聚物长度不能超过 3；一条 DNA 链中，GC 总含量应在 40%～60%。符合相应约束的 DNA 编码可减少 DNA 存储过程产生的数据错误，降低因此产生的开销，提升数据准确性。此外，重复子序列、汉明距离和最小自由能等也是可供选择的约束条件，但影响相对较小，而开销较大。

(4) 通用性：DNA 加密编码可用于任意规模和类型的数据。

(5) 高信息密度：实现二进制数据到 DNA 碱基序列的高密度转化，降低了 DNA 存储的成本开销。

经过编码和加密，DCFE 方法将目标文件转化为多条固定长度的 DNA 序列，结合人工合成 DNA 技术，实现用 DNA 作为媒介存储文件；相应地，要读取文件时，将存储的 DNA 链通过检索测序后得到多条固定长度的 DNA 序列，使用 DCFE 方法进行解码和解密，得到原始文件。本文中拟采用的 DNA 链组成结构如图 4 所示。基于 DNA 分子的生物特性和合成技术要求，一般将一条 DNA 链的长度控制在 100～300 nt (nucleotide，核苷酸)之间，通常包含前端引物(5′引物)、序号、数据、RS、后端引物(3′引物)几部分。具体地，DNA 链的中序号为 LT 编码步骤时使用的随机数生成种子(即编码时的序号)，大小为 16 nt；数据为加密编码算法生成的编码数据，大小为 120 nt；RS 码用于纠正 DNA 序列的替换错误，大小为 8 nt；前、后端引物不参与加密和编码，单独设计，大小均为 24 nt。其中，数据和 RS 码的长度可根据需求自行调整。


**图 4 DNA 链组成结构**

**Fig. 4 DNA strand composition structure**

### 4.1 加密和编码

DCFE 方法的加密和编码的整体流程如图 5 所示。

第一步，超混沌伪随机序列生成，步骤如下。

(1) 根据密钥 *KEYSA*(一组状态变量)确定

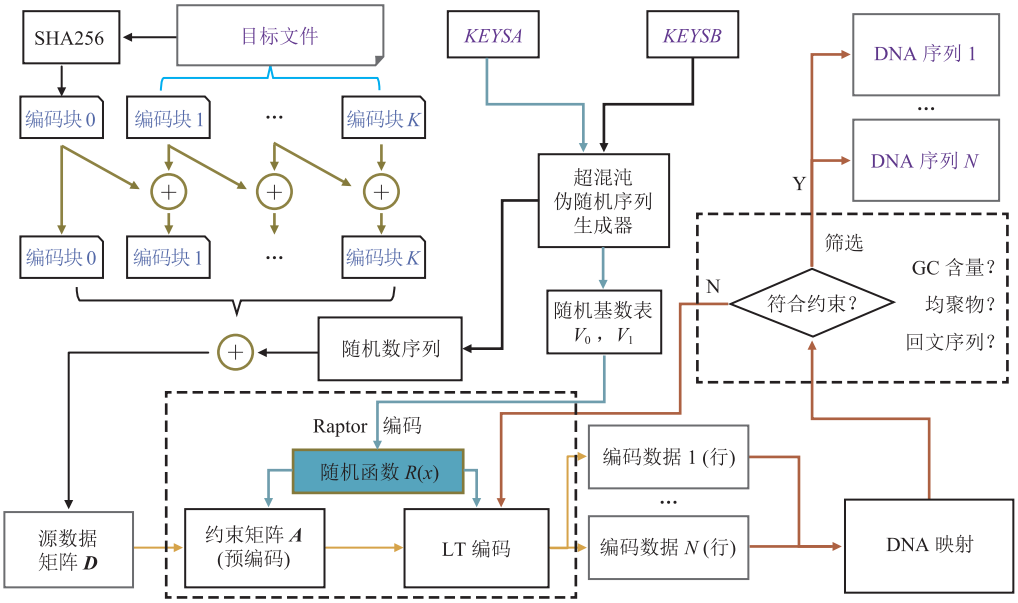



**图 5　加密和编码的整体流程**

**Fig. 5　The overall process of encryption and encoding**

超混沌系统的初始参数，进行初次超混沌伪随机序列生成运算，得到 512 个 4 byte unsigned integer，将其作为 Raptor 码中随机数生成器 Rand[$X$,idx,mod]的参照表 $V_0$ 和 $V_1$（各有 256 个 4 byte unsigned integer）数据。Rand[$X$,idx,mod]定义如下：

$$\text{Rand}[X,\text{idx},\text{mod}]=\left(V_0\left[(X+\text{idx})\%256\right]\oplus V_1\left[\left(\text{floor}(X/256)+\text{idx}\right)\%256\right]\right)\%\text{mod} \tag{20}$$

其中，$X$ 为输入值；idx 为一个索引变量；mod 为应用于结果的模。将 $V_0$ 和 $V_1$ 的值用按位的 $X$ 异或运算进行组合，然后应用模运算得到最终结果，即 0 和(mod−1)之间的整数。

(2) 根据密钥 *KEYSB*，通过相同伪随机序列生成器生成后续加密使用的随机数序列。

第二步，将目标文件进行 Raptor 加密编码，步骤如下。

(1) 分块。对目标文件进行分块，每块大小为 30 byte，以块为单位进行编号，编号记作 id，大小为 4 byte，不参与加密。同时使目标文件通过 SHA256 计算全局哈希值，取最后 30 byte 值作为初始哈希密钥，并存入第一个编码块，目标文件产生的分块从第二个编码块开始。

(2) 加密。将每一个编码块与哈希密钥异或，并将结果作为新的哈希密钥。通过构造的超混沌伪随机序列生成器生成 30 个随机数，将随机数与哈希异或的编码块数据异或，完成初次加密。全体编码块构成源数据矩阵 $\boldsymbol{D}$。

(3) 中间符号生成。统计分块数量 $K$，通过矩阵 $\boldsymbol{A}$ 生成中间符号 $\boldsymbol{C}$，矩阵 $\boldsymbol{A}$ 如图 2 所示，其 $\boldsymbol{G}_{\text{LT}}$ 部分是通过随机数生成器 Rand[$X$,idx,mod]生成的。以 10 KB 文件为例，分块大小为 30 byte，计算得 $K$=342，查表得对应的 $S$=31，$H$=10，则 $L$=383（不同 $K$ 对应的参数记录在特定表中）。通过公式(1)，最终生成 383 个中间符号。

(4) LT 编码。中间符号通过 LT 编码生成编码数据。LT 编码的度值和随机数由随机数生成器 Rand[$X$,idx,mod]生成，每个中间符号对应一行编码数据。在 LT 编码过程中，可通过提前构建解码矩阵 $\boldsymbol{G}'$（由每行编码数据对应的随机数组成）减少最终解码需要的数据行数，当满足 $\boldsymbol{G}'$ 可

逆时，可提前结束编码，降低冗余，即预解码操作。通过预解码操作降低冗余，最终生成 350 行编码数据。通过设置冗余，可额外生成对应数量的编码数据。

第三步，将生成的编码数据转化为 DNA 序列，步骤如下。

(1)映射。根据每个编码数据 id 对 8 取模的结果，决定 1 byte 二进制数据到 DNA 碱基的映射规则，得到 4 nt 的 DNA 序列。DNA 映射规则如表 5 所示。每行编码数据生成 120 nt 的 DNA 序列。此外，在序列前增添 4 byte 的随机数种子(即块 id)，在序列后增添由 30 byte 的编码块产生的 2 byte RS 纠错码，并将这 6 byte 数据按映射规则转化为 DNA 碱基。通过映射，最终得到 144 nt 的 DNA 序列。

(2)筛选。对每个 DNA 序列进行生物约束检测，即均聚物和 GC 含量检测。根据检测结果，保留通过的 DNA 序列，去除不符合约束的 DNA 序列，从已生成的 24 byte 的中间符号处重新开始进行上述步骤，直到所有编码数据转化为合格的 DNA 序列。

**表 5 DNA 映射规则**

**Table 5 DNA mapping rules**

| 规则 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 00 | A | A | T | T | G | G | C | C |
| 01 | C | G | C | G | T | A | T | A |
| 10 | G | C | G | C | A | T | A | T |
| 11 | T | T | A | A | C | C | G | G |

### 4.2 解密和解码

DCFE 方法解密和解码的整体流程与加密编码的流程一一对应，如图 6 所示。第一步都是进行超混沌伪随机序列生成，第二、三步操作反向进行，其解密与加密过程相对称，使用相同随机数进行解密。

第一步，超混沌伪随机序列生成，步骤如下。

进行超混沌运算。和加密编码的步骤一样，根据密钥确定超混沌系统的初始参数，进行初次超混沌伪随机序列生成运算，得到 Raptor 码中随机数生成器 Rand[$X$,idx,mod]的参照表 $V_0$ 和 $V_1$ 的各 256 个 4 byte unsigned integer 数据。解密需要以同样的随机数进行。因为都是根据固定数值和固定公式运算得到的伪随机数，所以加密和解密

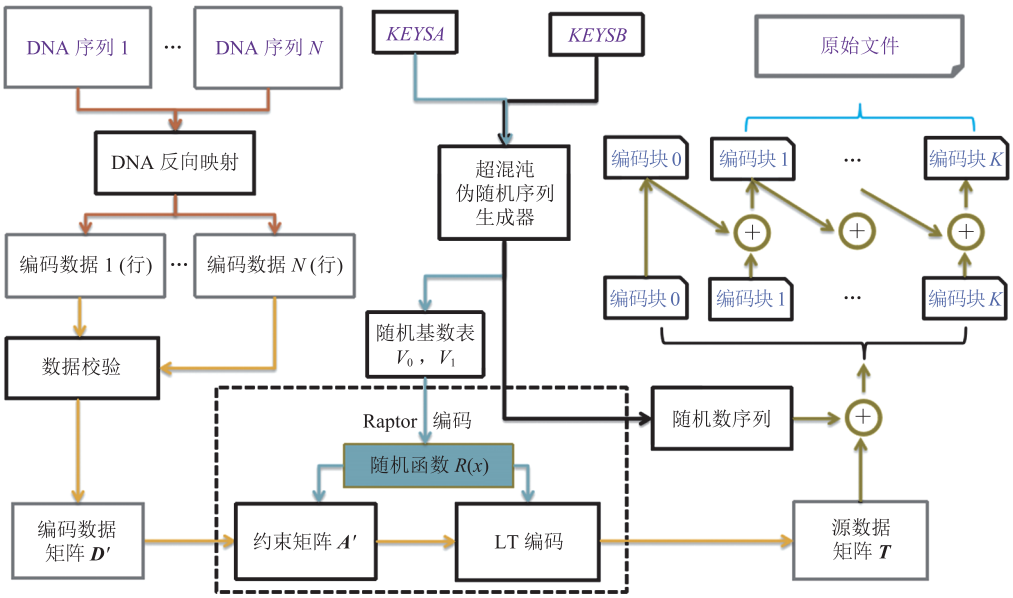



**图 6 解密和解码的整体流程**

**Fig. 6 The overall process of decryption and decoding**

过程的随机数是一致的。

第二步，DNA 反向映射，步骤如下。

每个合格的 DNA 序列都被反向映射其对应的二进制数据，按照相同的 DNA 映射规则逆向进行。每一条 DNA 链通过构造的超混沌伪随机序列生成器生成 30 个随机数，根据每个随机数后 3 位对 8 取模的结果，决定 4 nt DNA 碱基到二进制数据的映射规则，得到 1 byte 的二进制数据。DNA 映射规则如表 5 所示。144 nt 的 DNA 序列最终转化为 36 byte 的二进制数据，包含 4 byte 的随机数种子(即块 id)、30 byte 的编码块及其对应的 2 byte RS 纠错码。

第三步，Raptor 解密解码，步骤如下。

(1)校验。根据 2 byte RS 纠错码，对每条 DNA 链转化的 36 byte 二进制数据进行初步校验，得到校验后的 30 byte 编码块数据和 4 byte 对应块 id。

(2)中间符号还原。从全部编码块中选择略大于编码时分块总数数量的编码块，构成编码数据矩阵 $\boldsymbol{D'}$。当源文件的大小为 10 KB 时，对应 342 个编码块，通过矩阵 $\boldsymbol{A'}$ 求中间符号 $\boldsymbol{C}$，如公式(21)所示，$\boldsymbol{A'}$ 结构类似 $\boldsymbol{A}$，其 $\boldsymbol{G}_{\mathrm{LT}}$ 部分为由相同随机数生成器 Rand[$X$,idx,mod]生成的 $N\times L$ 维 $\boldsymbol{G'}_{\mathrm{LT}}$ 矩阵。最终生成 391 个中间符号。

$$\boldsymbol{C}=\boldsymbol{A'}^{-1}\boldsymbol{D'} \tag{21}$$

(3)LT 解码。通过公式(22)可将中间符号还原为源数据 $\boldsymbol{T}$，矩阵 $\boldsymbol{G'}$ 的每一行由更新参照表后的随机数生成器 Rand[$X$,idx,mod]生成，对应 LT 编码时的随机数值。

$$\boldsymbol{T}=\boldsymbol{G'C} \tag{22}$$

(4)解密。源数据 $\boldsymbol{T}$ 的每一行对应一个编码块，找到第一个编码块，通过第一步生成的对应随机数与编码块数据异或，两次异或还原数据，实现解密，并记作初始哈希密钥。再依次以相同方式解密，即先与随机数异或，再与哈希密钥异或，同时更新哈希密钥，最终解密全部数据。

## 5　实验结果分析

本文提出的 DCFE 方法可将任意类型数据进行加密和编码，最终转化为 DNA 序列，用于 DNA 存储。其中，DNA 序列的冗余(满足 100% 解码条件下的额外 DNA 链数量)、约束规则可自定义。本文的第 5.1～5.5 节实验用于证明算法加密的安全性，设置冗余为 20%，满足 GC 总含量在 40%～60%、均聚物长度小于 4 的两项约束规则，通过密钥空间、密钥敏感性、密文相关性、密文信息熵和抗差分攻击分析 5 项安全评估标准进行分析；第 6 项实验用于验证五维超混沌伪随机序列生成器的性能，通过 NIST SP800-22 测试进行分析；第 7 项实验用于证明方法的纠错能力，通过不同错误率下的解码模拟进行分析；最后一项实验分析对当前 DNA 存储领域的各类方法作了比较和总结。

以 10 KB 的文本数据为例，将明文通过 DCFE 方法进行加密和编码得到 DNA 密文，如图 7 所示。10 KB 数据对应的分块数量为 342，理论对应 342 条 DNA 链，每条 DNA 链的长度为 144 nt。Raptor 码的特性要求参与解码的 DNA 链数量略大于分块数量 342 才可成功解码，因此，设置初步编码结果为 350 条 DNA 链。通过预解码操作，即不断编码生成 DNA 链并尝试解码，直到成功解码且生成的 DNA 链数量为 350，取该 350 条 DNA 链作为初步结果，可达到 0.26 的冗余度和 1.59 bits/nt 的信息密度(理论上限为 2 bits/nt)。此外，可通过增加 DNA 链数量提升数据纠错能力，因此选择 20% 的冗余，额外生成 20% 的 DNA 链，最终生成 411 条 DNA 链，达到 0.512 的冗余度和 1.32 bits/nt 的信息密度。

本文针对不同类型和大小的数据进行测试，其结果如表 6 所示，证明了 DCFE 方法的通用性。

### 5.1　密钥空间分析

密钥空间的大小决定了该方案能否抵抗暴

力攻击。本文提出方法的密钥空间主要由混沌系统的初始参数组成，共用到 2 组状态变量作为密钥，每个状态变量取十进制的 12 个数位，则计算得密钥空间大小为

$$\left(\left(10^{12}\right)^{5}\right)^{2}=10^{120}\approx 2^{400} \tag{23}$$

通过计算可知，该算法的密钥空间远大于理论安全密钥空间值 $2^{128}$ [34]，表明本文所提出的方法能够抵抗暴力攻击。

### 5.2 密钥敏感性分析

密钥的敏感性指密钥的微小变化能够引起加密后密文的巨大变化。如果一个加密算法表现出这一特性，则表明如果没有正确的密钥，就不能恢复原始明文。为了证明这一点，本文对提出的方法采取了以下步骤进行实验。

(1) 确定原始 *KEYS*1 (两组状态变量作为一个整体)。

(2) 修改原始密钥 *KEYS*1 中的一个数位，生成对比密钥 *KEYS*2。

(3) 使用 *KEYS*1 和 *KEYS*2 作为加密密钥，通过 DCFE 方法对相同的明文进行加密。

(4) 比较从两个加密过程中得到的密文序列。

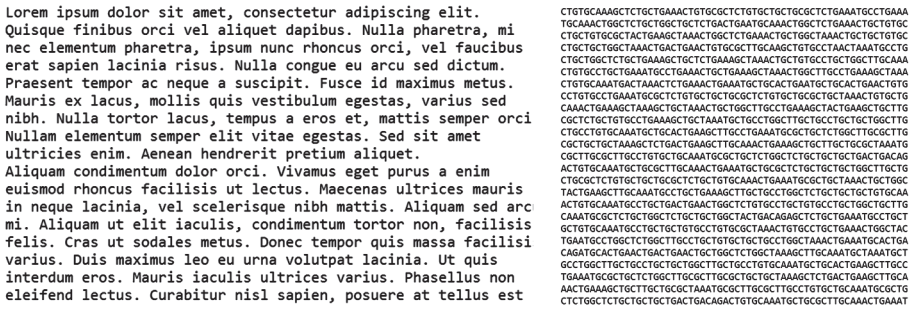
选取密文序列中随机同位置的 50 个字符进行比较，取其中 3 次测试结果，如图 8 所示，数值 0、1、2、3 分别表示一种碱基，展示了仅一数位不同密钥对应加密密文的差异。

进一步通过比特变化率 (number of bit change rate，NBCR) 对密钥敏感性进行测试，将每个碱基按固定规则转化为 0～1 bit，NBCR 的计算公式如下：

$$\mathrm{NBCR}=\frac{\mathrm{Ham}\left(B1,B2\right)}{Len}\times 100\% \tag{24}$$

其中，$B1$、$B2$ 为待比较的两个比特流，$\mathrm{Ham}(B1,B2)$ 为 $B1$ 和 $B2$ 的汉明距离；$Len$ 为比特流长度。NBCR 的理论值为 50%，越接近理论值，密钥敏感性越强。

DCFE 方法密钥为超混沌系统的两组状态变量，合计 10 项密钥参数，设置初始密钥，再分




(a) 明文 (b) DNA 密文

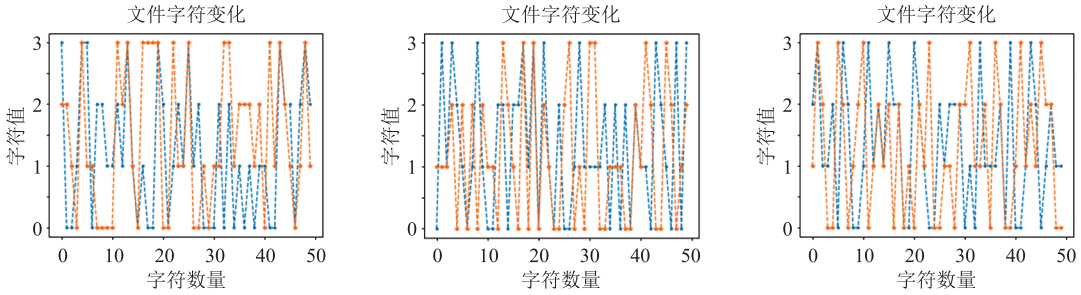
**图 7 加密示例 (部分明文及其对应密文)**

**Fig. 7 Encryption examples (partial plaintext and corresponding ciphertext)**

**表 6 不同数据加密和编码结果**

**Table 6 Different data encryption and encoding results**

| 加密前大小 (MB) | 加密后大小 (nt) | 类型 | 名称 | 耗时 (s) |
|---|---|---|---|---|
| 0.010 | 144×411 | 文本 | Chapter.txt | 0.31 |
| 0.980 | 144×41 372 | 文本 | A Tale of Two Cities.pdf | 37.44 |
| 0.252 | 144×10 531 | 图片 | Lena.bmp | 10.32 |
| 3.890 | 144×163 267 | 视频 | I have a dream.mp4 | 151.82 |




(a) 第 1 次测试　(b) 第 2 次测试　(c) 第 3 次测试

**图 8　密文字符变化对比**

**Fig. 8　Comparison of changes in ciphertext characters**

别对密钥的其中一项进行改变，翻转其中一位(每项参数由 12 位十进制数组成)，得到 10 组与初始密钥仅一位差异的不同密钥，分别加密相同文件，计算对应的 NBCR，如图 9 所示。其中，*A*()为 *KEYSA* 中状态变量；*B*()为 *KEYSB* 中状态变量。

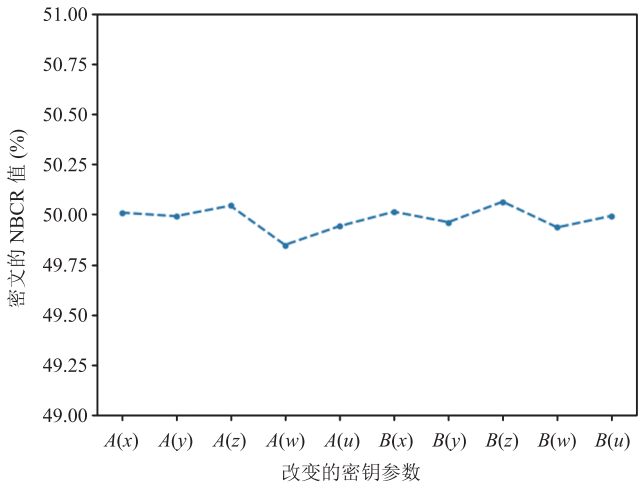



**图 9　密钥敏感性测试**

**Fig. 9　The key sensitivity testing**

从结果中可以看出，密钥中的任意一个位的变化，其对应密文的 NBCR 始终接近 50%，表明加密后密文的巨大差异，证明了本文加密方法是密钥敏感的。

### 5.3　密文相关性分析

加密的意义之一在于降低数据之间的相关性，相关性越低，密文越能抵抗统计攻击。通常，由于图像之间的相关性比文本更强，因此，以图像为例，测试本文方法产生的密文相关性，验证加密效果。

一般从图像的水平、垂直、对角 3 个方向分别选取像素点进行相关性分析，以图片 Lena (512×512 个像素点)为例，其加解密结果如图 10 所示(其中，将加密生成的 DNA 序列转化为对应数量的像素点构成图 10(b))，相关性如图 11 所示。其中，明文在各方向上的像素值近似线性相关，而密文在各方向上的像素值则是近似均匀随机分布的，证明加密算法消除了原始图像的相关性。

相关性系数 $\rho(Y,Z)$ 的计算如公式(25)～(28)所示，对多张图片随机取 10 000 个像素点进行计算，同时将加密后的 DNA 序列转化为二进制数据后，取 10 000 个字节进行计算，结果如表 7 所示。

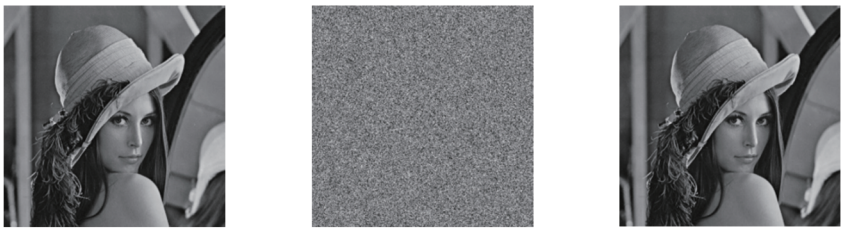
$$\rho(Y,Z)=\frac{\operatorname{cov}(Y,Z)}{\sqrt{D(Y)D(Z)}} \tag{25}$$

$$\operatorname{cov}(Y,Z)=\frac{1}{\mathrm{Num}}\sum_{j=1}^{\mathrm{Num}}\left(y_j-E(y)\right)\left(z_j-E(Z)\right) \tag{26}$$

$$E(Y)=\frac{1}{\mathrm{Num}}\sum_{j=1}^{\mathrm{Num}}y_j \tag{27}$$

$$D(Y)=\frac{1}{\mathrm{Num}}\sum_{j=1}^{\mathrm{Num}}\left(y_j-E(Y)\right)^2 \tag{28}$$

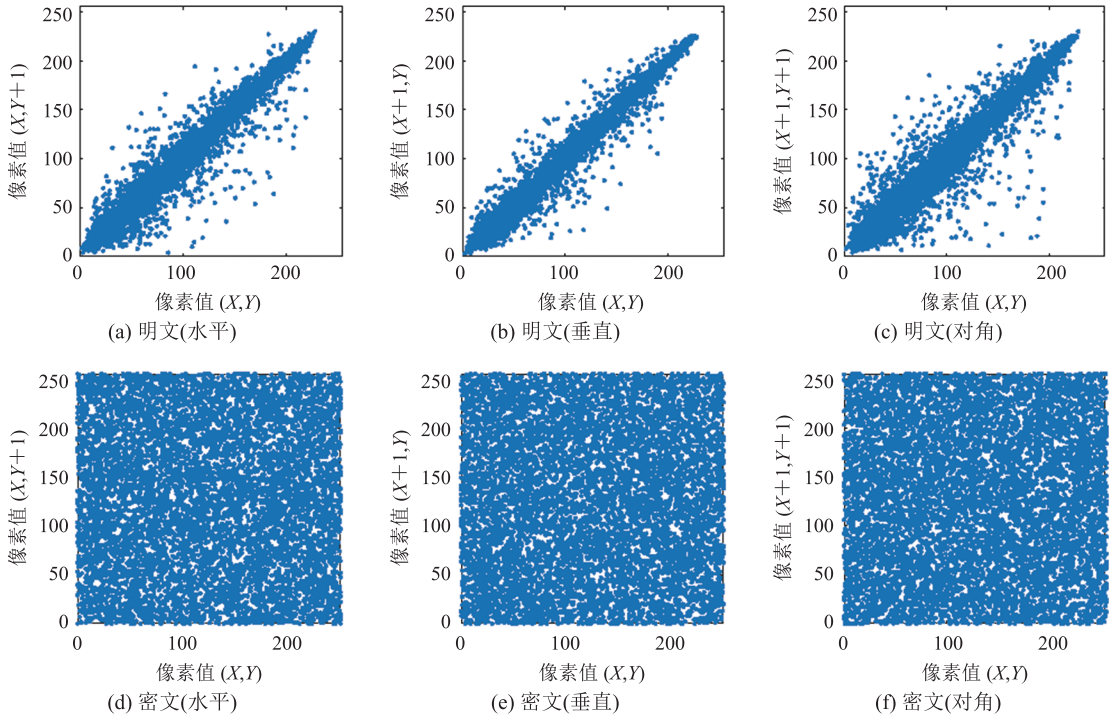
其中，$Y$、$Z$ 分别为两张图片的像素点集；$y_j$ 为 $Y$ 中的第 $j$ 个像素点值；Num 为像素点总数量。其中，公式(25)中的 $\rho(Y,Z)$ 针对图像某一方向像素



(a) 原始图片 (b) 加密结果 (c) 解密结果

**图 10 图片加解密对比**

**Fig. 10 Comparison of image encryption and decryption**




(a) 明文(水平) (b) 明文(垂直) (c) 明文(对角)

(d) 密文(水平) (e) 密文(垂直) (f) 密文(对角)

**图 11 明文与密文在各方向上的相关性**

**Fig. 11 The correlation of plaintext and ciphertext in all directions**

**表 7 不同图片的相关性系数**

**Table 7 The correlation coefficient of different images**

| 名称 | 明文相关性系数 | | | 密文相关性系数 | | |
|---|---|---|---|---|---|---|
| | 水平 | 垂直 | 对角 | 水平 | 垂直 | 对角 |
| Lena.bmp | 0.974 2 | 0.986 1 | 0.960 6 | 0.006 6 | 0.002 1 | 0.003 8 |
| Baboon.bmp | 0.977 5 | 0.968 6 | 0.942 2 | −0.004 2 | 0.007 3 | −0.003 6 |
| Peppers.bmp | 0.966 8 | 0.983 1 | 0.986 6 | 0.003 4 | −0.005 6 | 0.002 3 |
| Cameraman.bmp | 0.985 3 | 0.987 0 | 0.976 5 | 0.001 2 | 0.002 2 | 0.001 9 |
| Lena.png | 0.876 9 | 0.958 4 | 0.943 3 | 0.008 4 | −0.003 6 | −0.002 4 |
| Baboon.png | 0.932 1 | 0.990 4 | 0.937 1 | 0.003 2 | 0.006 6 | 0.007 8 |

点进行计算，密文理论值为 0。常规图像加密方法对压缩格式 png 和 jpeg 等类型图像会先进行解压缩，转化为原生像素格式，再进行加密，从而增大数据量，提高存储成本。本文针对压缩格式直接加密，取一字节对应一像素进行相关性系数计算，结果证明，不仅原生格式图像加密后各方向不相关，压缩格式图像直接加密得到的密文也不相关。

不同方法计算得出的 Lena 图像相关性系数如表 8 所示，第二、三项为常规图像加密方法，后三项为当前 DNA 存储领域的图像加密方法。结果证明，DCFE 方法产生的密文相关性在精度上接近当前研究进展。

**表 8　不同方法的 Lena 图像密文相关性系数**

**Table 8　The correlation coefficients of Lena image cipher with different methods**

| 方法 | 水平 | 垂直 | 对角 |
|---|---|---|---|
| DCFE | 0.006 6 | 0.002 1 | 0.003 8 |
| Xu 等[26] | −0.001 5 | 0.004 1 | 0.006 9 |
| Wang 等[13] | 0.002 3 | −0.002 0 | −0.007 3 |
| Zan 等[16] | −0.011 3 | −0.010 1 | 0.007 9 |
| Yao 等[17] | 0.000 3 | −0.000 8 | 0.001 0 |
| 姚翔宇等[18] | −0.000 5 | −0.007 7 | −0.000 4 |

## 5.4　密文信息熵分析

信息熵用于衡量密文中的不确定度水平，由公式(29)给出。当测试的信息熵结果接近理论最大值时，说明被加密算法加密后的密文在统计上是独立的，即密文每个部分相互独立，彼此不依赖。

$$H(x)=-\sum_{i=1}^{\mathrm{num}}\left(p(x_i)\cdot\log_2\left(p(x_i)\right)\right) \tag{29}$$

其中，num 为符号 $x$ 的可取值数量；$p(x_i)$ 为符号 $x_i$($x$ 的一种取值)出现的概率。

由于 DNA 存储数据最终以碱基形式存在，因此其 num＝4，信息熵理论最大值为 2。而二进制数据以 byte 为单位，按规则转化为 DNA 序列，计算 1 byte 二进制密文对应的信息熵时取 num＝$2^8$，理论最大信息熵为 8。加密算法通常以 byte 为单位计算密文信息熵，因此，本文将密文 DNA 序列转化为二进制数据，以 byte 为单位计算信息熵，结果如表 9 所示。

不同方法计算得出的密文信息熵如表 10 所示，表中第二、三项为常规图像加密方法，后四项为当前 DNA 存储领域的加密方法。表中倒数第二项的理论最大信息熵为 11，其他项的理论最大信息熵为 8。

根据表 9 和表 10 可以得出结论，密文在统计上是相互独立的，可以抵抗统计攻击。

**表 9　不同文件的密文信息熵**

**Table 9　The cipher information entropy of different files**

| 名称 | 信息熵 |
|---|---|
| Chapter.txt | 7.982 4 |
| A Tale of Two Cities.pdf | 7.966 3 |
| Lena.bmp | 7.986 2 |
| I have a dream.mp4 | 7.954 7 |

**表 10　不同方法的密文信息熵**

**Table 10　The cipher information entropy of different methods**

| 方法 | 信息熵 |
|---|---|
| DCFE | 7.986 2 |
| Xu 等[26] | 7.999 2 |
| Wang 等[13] | 7.996 0 |
| Peng 等[15] | 7.937 6 |
| Zan 等[16] | 7.950 1 |
| Yao 等[17] | 10.994 3 |
| 姚翔宇等[18] | 7.998 3 |

## 5.5　抗差分攻击分析

差分攻击指利用输入明文之间的差异(差分)推断密钥或破解密码，通过轻微改变明文，用同样密钥的加密算法加密之后，如果得到的新密文和原明文对应的原密文之间的差距十分大，则说明该加密算法具有抵抗差分攻击的能力。平均像

素改变率(normalized pixel change rate，NPCR)和平均像素改变强度(unified average changing intensity，UACI)是衡量图像抵抗差分攻击能力的经典指标。

本文选择经典图像 Lena 进行测试，改变其中一个像素点数据，得到新图像，分别取对应加密后的原 DNA 序列密文，以 4 nt 为单位转化为 1 像素，计算对应的 NPCR 和 UACI，如公式(30)～(32)所示。

$$\mathrm{NPCR}=\frac{\sum_{i=1}^{n}\mathrm{diff}\left(x_i,y_i\right)}{n} \tag{30}$$

$$\mathrm{diff}\left(x_i,y_i\right)=\begin{cases}1, & x_i\neq y_i\\ 0, & x_i=y_i\end{cases} \tag{31}$$

$$\mathrm{UACI}=\frac{\sum_{i=1}^{n}\frac{|x_i-y_i|}{255}\times 100\%}{n} \tag{32}$$

不同方法计算得出的密文 NPCR 和 UACI 如表 11 所示，NPCR 的理论期望值为 1，UACI 的理论期望值为 0.334 6。由表 11 可知，本文方法具有较强的抗差分攻击能力。

**表 11 不同方法的密文 NPCR 和 UACI**

**Table 11 The NPCR and UACI of cipher using different methods**

| 方法 | NPCR | UACI |
|---|---|---|
| DCFE | 0.994 2 | 0.335 0 |
| Xu 等[26] | 0.996 1 | 0.334 6 |
| Wang 等[13] | 0.996 1 | 0.335 0 |
| Peng 等[15] | 0.901 4 | 0.016 7 |
| Zan 等[16] | ≈1 | ≈0.5 |
| Yao 等[17] | 0.996 1 | 0.571 3 |
| 姚翔宇等[18] | 0.996 1 | 0.291 9 |

### 5.6 NIST SP800-22 测试

NIST SP800-22 测试是检验伪随机序列随机性的重要标准。它总共包含 16 个测试，如果生成的伪随机序列通过了所有测试，则表明该序列具有良好的随机性。对本文提出的五维超混沌伪随机序列生成器的性能进行评估，生成 100 条不同的伪随机序列进行测试。结果如表 12 所示，可以看出，该五维超混沌伪随机序列生成器生成

**表 12 NIST SP800-22 测试**

**Table 12 NIST SP800-22 test**

| 测试项目 | *P*-value (>0.01) | 通过比例 (≥0.96) | 结果 |
|---|---|---|---|
| 频率 | 0.955 8 | 1.00 | Pass |
| 块频率 | 0.941 1 | 0.98 | Pass |
| 累积和(正向) | 0.241 6 | 1.00 | Pass |
| 累积和(反向) | 0.289 8 | 1.00 | Pass |
| 游程 | 0.334 5 | 1.00 | Pass |
| 最长游程 | 0.683 3 | 0.98 | Pass |
| 秩 | 0.779 1 | 1.00 | Pass |
| 快速傅里叶变换 | 0.334 5 | 0.99 | Pass |
| 非重叠模板匹配 | 0.249 2 | 1.00 | Pass |
| 重叠模板匹配 | 0.942 5 | 1.00 | Pass |
| 通用统计 | 0.304 1 | 0.99 | Pass |
| 近似熵 | 0.678 6 | 1.00 | Pass |
| 随机游动 | 0.804 3 | 1.00 | Pass |
| 串行 (*P*-value 1) | 0.864 6 | 0.99 | Pass |
| 串行 (*P*-value 2) | 0.289 6 | 0.99 | Pass |
| 线性复杂度 | 0.154 7 | 1.00 | Pass |

的伪随机序列通过了全部测试，证明了其生成的伪随机序列具有极好的随机性，进而保证了加密算法的安全性。

### 5.7　纠错能力分析

在 DNA 存储过程中，由于合成和测序生物技术的缺陷，序列丢失和碱基错误是最常见的两类错误。序列丢失意味着一些 DNA 链可能在存储过程中丢失，如偏置 PCR 和测序过程。碱基错误指 DNA 合成、存储和测序过程中发生的碱基替换、擦除(插入和缺失)。

本文提出的 DCFE 方法由于结合了 DNA Raptor 码，能够以少量的冗余解决任意序列丢失问题，并添加了 RS 码，用于解决特定数量的替换和擦除错误。因此，该方法通过少量冗余实现了用于 DNA 存储的可自纠错加密编码，且冗余越大，相应的纠错性能越强。

DCFE 方法的纠错能力主要源于 DNA Raptor 码和 RS 码。RS 码的纠错能力由其码长决定，其纠错能力为码长的一半，在 DCFE 方法中，每条 DNA 链中的 RS 码长为 8 nt，即可纠正该链的最多 4 位碱基错误。DNA Raptor 码的纠错能力取决于冗余，其只能解决序列丢失错误，但可在 RS 纠错失败的情况下舍弃整条 DNA 链，将其视为序列丢失，从而实现纠错。图 12 为 DNA Raptor 码的纠错能力在不同冗余下的表现，此处冗余特指在满足 100% 成功解码的最低 DNA 链数量基础上增加的 DNA 链数量，其增加的 DNA 链数量越大，DNA Raptor 码的纠错能力越强。实验数据为 10 KB 文本，测试在不同序列丢失率下 100 次解码的成功次数。

在当前的 DNA 存储领域加密方法中，基于 DNA 生化特性进行加密的方法未考虑纠错，而基于 DNA 碱基排序及编解码设计的加密方法，如 Zan 等[16]、Yao 等[17]和姚翔宇等[18]考虑到了图像的鲁棒性，即在部分密文数据丢失后，将其还原为可供识别的相似图像，却未能将发生的数据错误进行纠正，得到完整正确的原始数据。本文提出的 DCFE 方法能够在一定冗余下(20%)100% 纠正 15% 的序列丢失错误和每条 DNA 链的 4 位碱基错误，实现自纠错的 DNA 加密编码。

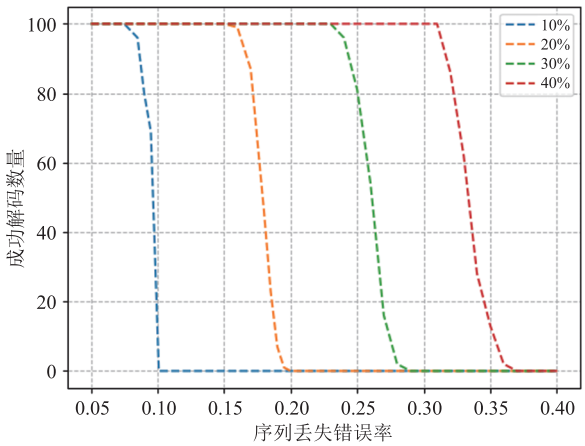



**图 12　不同冗余下的 DNA Raptor 码的纠错能力**

**Fig. 12　Error correction performance of DNA Raptor codes under different redundancies**

### 5.8　DNA 存储的各类加密方法性能比较

作为新兴研究方向，近年来，适用于 DNA 存储领域的加密方法研究相对较少。在传统加密算法上进行改进是一种研究方向，尤其是基于图像加密算法的改进(DNA 编解码和碱基计算规则被广泛用于图像加密的中间过程)。用于 DNA 存储的加密方法需要将密文转化为固定长度的碱基序列，以 DNA 链形式存储，因此必须考虑相关的生物约束和纠错能力，避免 DNA 存储过程产生的错误导致解密失败或丢失部分明文信息，使原始数据无法还原。参考 Zan 等[16]对不同方法的比较，表 13 展示了当前 DNA 存储领域各类加密方法的性能评估。

根据表 13 和已有的实验分析，与其他已有方法相比，本文提出的 DCFE 方法具有以下优势：(1)基于混沌系统和喷泉码特性，加密和编码过程是动态的，具有强随机性，保障数据安全；(2)通过筛选方式获取满足约束条件的结果，不仅可满足 DNA 存储过程最常用的生物约

表 13 DNA 存储的各类加密方法性能

Table 13 The performance of various encryption methods for DNA storage

| 方法 | 动态编码 | 动态加密 | 生物约束 | 大规模加密 | 纠错能力 | 信息密度 (bits/nt) | 加密数据类型 |
|---|---|---|---|---|---|---|---|
| Yang 等[6] | √ | * | * | × | × | 0.006 | 任意 |
| Zhang 等[8] | × | * | * | × | × | 0.001 | 任意 |
| Zhu 等[9] | √ | √ | * | × | × | 2.000 | 任意 |
| Peng 等[15] | √ | √ | √ | × | × | 1.650 | 任意 |
| Zan 等[16] | √ | √ | √ | √ | × | 1.000 | 图像 |
| Yao 等[17] | √ | √ | √ | × | × | 1.000 | 图像 |
| 姚翔宇等[18] | √ | √ | √ | × | × | 1.000 | 图像 |
| DCFE | √ | √ | √ | √ | √ | 1.320 | 任意 |

注：√ 代表可接受的支持水平，× 代表最低限度的支持水平，* 代表部分支持

束，且可自定义新约束条件用于筛选；(3)可用于任何规模和类型的数据加密，适用范围广泛；(4)具备纠错能力，能纠正 DNA 存储过程中发生的部分序列丢失和替换错误，保证解密解码后获得完整且正确的原始数据；(5)具有相对较高的信息密度，降低 DNA 存储成本。

其中，前 3 项方法基于生化技术进行加密(DNA 组装、DNA 折纸和 DNA 链置换)；后 5 项方法基于 DNA 碱基排序及编解码设计的加密方法。动态加密和编码指该过程有较高随机性，包含多个随机步骤。生物约束指 GC 含量和均聚物约束，这是 DNA 存储最常用的两种生物约束。纠错能力指纠正 DNA 存储过程中的错误，以达到 100% 还原明文的能力。

## 6 结 论

随着 DNA 存储技术的高速发展，与之相关的数据安全问题越发重要。由于 DNA 存储独特的读写和存储方式，设计与这种新技术兼容的加密方法时需要保证数据具备一定的自纠错能力，避免因 DNA 序列在合成、存储和测序过程中产生的错误导致密文无法正确还原为明文。基于对混沌系统加密原理和 DNA 喷泉码编码的研究，本文构建了新的五维超混沌系统伪随机序列生成器，并将其用于 DNA Raptor 码的编解码过程，提出二者联合的 DCFE 方法用于 DNA 加密存储。与当前 DNA 存储领域的加密方法相比，DCFE 方法在安全性、可满足约束条件、适用数据类型和规模、纠错能力和信息密度等方面均有良好表现。但由于时间和精力的限制，DCFE 方法产生的 DNA 链在信息密度和纠错能力上未能同时达到当前 DNA 存储领域纯编码方法研究中的最优水平，后期工作将针对此不足开展，优化其相关表现。

## 参 考 文 献

[1] 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.
Zan XZ, Yao XY, Xu P, et al. A survey on error correcting algorithms in DNA storage [J]. Journal of Guangzhou University (Natural Science Edition), 2021, 20(2): 13-22.

[2] Bonnet J, Colotte M, Coudy D, et al. Chain and conformation stability of solid-state DNA: implications for room temperature storage [J]. Nucleic Acids Research, 2010, 38(5): 1531-1546.

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

[4] 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.

[5] Hakami HA, Chaczko Z, Kale A. Review of big data storage based on DNA computing [C] // Proceedings of the 2015 Asia-Pacific Conference on Computer Aided System Engineering, 2015: 113-117.

[6] Yang J, Ma JJ, Liu S, et al. A molecular cryptography model based on structures of DNA self-assembly [J]. Chinese Science Bulletin, 2014, 59: 1192-1198.

[7] Zakeri B, Carr PA, Lu TK. Multiplexed sequence encoding: a framework for DNA communication [J]. PLoS One, 2016, 11(4): e0152774.

[8] Zhang YN, Wang F, Chao J, et al. DNA origami cryptography for secure communication [J]. Nature Communications, 2019, 10(1): 5469.

[9] Zhu EQ, Luo XH, Liu CJ, et al. An operational DNA strand displacement encryption approach [J]. Nanomaterials, 2022, 12(5): 877.

[10] Wu XJ, Kan HB, Kurths J. A new color image encryption scheme based on DNA sequences and multiple improved 1D chaotic maps [J]. Applied Soft Computing, 2015, 37: 24-39.

[11] Wu JH, Liao XF, Yang B. Image encryption using 2D Hénon-Sine map and DNA approach [J]. Signal Processing, 2018, 153: 11-23.

[12] Wu XJ, Wang KS, Wang XY, et al. Color image DNA encryption using NCA map-based CML and one-time keys [J]. Signal Processing, 2018, 148: 272-287.

[13] Wang X, Su Y. Image encryption based on compressed sensing and DNA encoding [J]. Signal Processing: Image Communication, 2021, 95: 116246.

[14] Grass RN, Heckel R, Dessimoz C, et al. Genomic encryption of digital data stored in synthetic DNA [J]. Angewandte Chemie International Edition, 2020, 59(22): 8476-8480.

[15] Peng WP, Cui S, Song C. One-time-pad cipher algorithm based on confusion mapping and DNA storage technology [J]. PLoS One, 2021, 16(1): e0245506.

[16] Zan XZ, Chu L, Xie RZ, et al. An image cryptography method by highly error-prone DNA storage channel [J]. Frontiers in Bioengineering and Biotechnology, 2023, 11: 1173763.

[17] Yao XY, Xie RZ, Zan XZ, et al. A novel image encryption scheme for DNA storage systems based on DNA hybridization and gene mutation [J]. Interdisciplinary Sciences: Computational Life Sciences, 2023, 15(3): 419-432.

[18] 姚翔宇, 苏燕青, 昝乡镇, 等. 一种基于前向纠错码的图像 DNA 加密存储算法 [J]. 信息安全学报, 2023, 8(6): 28-36.
Yao XY, Su YQ, Zan XZ, et al. An image encryption and storage algorithm based on forward error correction DNA codes [J]. Journal of Cyber Security, 2023, 8(6): 28-36.

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

[20] Luby M. LT codes [C] // Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002: 271.

[21] Shokrollahi A. Raptor codes [J]. IEEE Transactions on Information Theory, 2006, 52(6): 2551-2567.

[22] 张淑芳, 彭康. 采用 Raptor 码的 DNA 信息存储技术 [J]. 激光与光电子学进展, 2020, 57(15): 151701.
Zhang SF, Peng K. DNA information storage technology based on Raptor code [J]. Laser & Optoelectronics Progress, 2020, 57(15): 151701.

[23] Shokralli A, Watson M, Luby TSM. RFC 5053 Raptor Forward Error Correction scheme for object Delivery [Z/OL]. https://www.rfc-editor.org/info/rfc5053.

[24] Schwarz PM, Freisleben B. NOREC4DNA: using near-optimal rateless erasure codes for DNA

storage [J]. BMC Bioinformatics, 2021, 22(1): 406.

[25] Matthews R. On the derivation of a “Chaotic” encryption algorithm [J]. Cryptologia, 1984, 13(1): 29-42.

[26] Xu QY, Sun KH, Cao C, et al. A fast image encryption algorithm based on compressive sensing and hyperchaotic map [J]. Optics and Lasers in Engineering, 2019, 121: 203-214.

[27] Masood F, Masood J, Zhang LJ, et al. A new color image encryption technique using DNA computing and chaos-based substitution box [J]. Soft Computing, 2022, 26(16): 7461-7477.

[28] Zhu HG, Ge JX, Qi WT, et al. Dynamic analysis and image encryption application of a sinusoidal-polynomial composite chaotic system [J]. Mathematics and Computers in Simulation, 2022, 198: 188-210.

[29] 罗利军, 李银山, 李彤, 等. 李雅普诺夫指数谱的研究与仿真 [J]. 计算机仿真, 2005, 22(12): 285-288.
Luo LJ, Li YS, Li T, et al. Research and simulation of Lyapunov’s exponents [J]. Computer Simulation, 2005, 22(12): 285-288.

[30] Chen ZQ, Yang Y, Yuan ZZ. A single three-wing or four-wing chaotic attractor generated from a three-dimensional smooth quadratic autonomous system [J]. Chaos, Solitons & Fractals, 2008, 38(4): 1187-1196.

[31] Liu J, Tong XJ, Liu Y, et al. A joint encryption and error correction scheme based on chaos and LDPC [J]. Nonlinear Dynamics, 2018, 93(3): 1149-1163.

[32] Lorenz EN. Deterministic nonperiodic flow [J]. Journal of the Atmospheric Sciences, 1963, 20(2): 130-141.

[33] Gao S, Wu R, Wang X, et al. A 3D model encryption scheme based on a cascaded chaotic system [J]. Signal Processing, 2023, 202: 108745.

[34] Dong YH, Zhao G, Ma YJ, et al. A novel image encryption scheme based on pseudo-random coupled map lattices with hybrid elementary cellular automata [J]. Information Sciences, 2022, 593: 121-154.