引用格式:牛莹,张勋才,韩栋,等. 一种基于 DNA 序列运算的信息隐藏方案[J]. 轻工学报,2016,31(1):61 – 66.
中图分类号:TP309;TP18 **文献标识码**:A
DOI:10.3969/j.issn.2096 – 1553.2016.1.011
文章编号:2096 – 1553(2016)01 – 0061 – 06

一种基于 DNA 序列运算的信息隐藏方案

An information hiding scheme based on DNA sequence operation

牛莹,张勋才,韩栋,王燕,崔光照,王子成
NIU Ying, ZHANG Xun-cai, HAN Dong, WANG Yan, CUI Guang-zhao, WANG Zi-cheng

郑州轻工业学院 电气信息工程学院,河南 郑州 450002
College of Electric Information Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, *China*

关键词:
信息隐藏;DNA 密码;
DNA 序列

Key words:
information hiding;
DNA cryptography;
DNA sequence

摘要:为减少 DNA 加密算法中的生物操作,提出一种基于 DNA 序列运算的信息隐藏方案. 该方案将海量 DNA 序列作为天然的 DNA 密码本,结合 DNA 数字编码规则,将待加密的信息与参考序列进行异或运算后转换为 DNA 序列,再通过混入冗余序列来实现信息的隐藏. 对该方案性能和安全性的分析与验证结果表明,该方案不需要生物操作,成本低,易于传输,有较强的安全性.

收稿日期:2014 – 12 – 31
基金项目:国家自然科学基金项目(61472371,61472372,61076103);河南省基础与前沿技术研究计划项目(142300413214);河南省高等学校青年骨干教师资助计划项目(2013GGJS—106);河南省高校科技创新人才支持计划资助项目(15HASTIT019)

作者简介:牛莹(1982—),女,河南省洛阳市人,郑州轻工业学院讲师,硕士,主要研究方向为智能信息处理与控制.
通信作者:崔光照(1957—),男,河南省洛宁县人,郑州轻工业学院教授,博士,主要研究方向为生物计算与信息安全.

Abstract: In order to reduce the biological operation of mass DNA cryptography algorithm, an information hiding sequence was proposed based on DNA sequence operation. Using DNA sequence database as a natural DNA password, combining with the DNA digital encoding rules, the information to be encrypted and reference sequences was converted to DNA sequence by an XOR operation. Then, by mixing with redundant sequences information hiding was achieved. The performance and security analysis of the scheme results showed that the sequence did not need the biological operation, had the character of low cost, was easy to transport and had strong security.

0 引言

随着社会信息化程度的不断提高,危害信息安全的事件与日俱增,如数据泄露、“棱镜门”事件等,这些充分说明信息安全是任何组织和个人都必须重视的问题,甚至关涉国家安全战略[1]. 信息安全已经成为并将继续成为信息时代的重大课题,其中,密码技术是核心信息安全技术. 如今,计算机的飞速发展使得基于数学困难问题的传统密码学面临新的挑战[2-3]. 在非传统密码理论与技术研究领域,视觉密码、量子密码、DNA 密码已成为密码学研究的新方向[4-6].

DNA 密码学的发展是随着 DNA 计算[7]研究的不断深入产生的. DNA 计算引入了崭新的数据结构和计算方法,提供了分子级的并行处理能力,对传统的信息安全提出了挑战[7-8]. 1996 年,D. BONEH 等[8]首次用 DNA 计算来破解传统的加密标准 DES. 1999 年,C. T. CELLAND 等[9]利用 DNA 作为信息载体实现了信息的隐藏,并把二战中著名的“June 6 invasion: Normandy”信息隐藏到 DNA 微点中. 2003 年,B. SHIMANOVSKY 等[10]利用冗余密码子将信息隐藏在 mRNA 序列中,并提出利用算术编码进行信息嵌入的隐写方法. 2004 年,A. GEHANI 等[11]借助 DAN 作为信息载体,利用生化技术在 DNA 分子上实现了“一次一密”的传统加密算法. 2005 年,K. TANAKA 等[12]利用 DNA 解决了密钥分配问题. 2009 年,崔光照等[13]利用 DNA 合成技术、PCR 扩增技术及 DNA 数字编码技术,结合传统密码学提出了一种基于 DNA 技术的加密方法. 2010 年,H. SHIU 等[14]提出了插入法、互补配对法和替代法 3 种数据隐藏方案. 2012 年,C. GUO 等[15]提出了一种基于 DNA 序列的信息隐藏方案,该方案具有较低的修改率和较高的嵌入容量. 2013 年,G. C. L. GOFF 等[16]将 DNA 微粒技术与热缩片结合,把 DNA 聚合物固定在聚乙烯热缩片上,形成尺寸在 100 μm 内的三维 DNA 水凝胶微粒阵列,实现了三维(微粒阵列)加密模型. 2014 年,张勋才等[17]给出了一种基于 RNA 二级结构的信息隐藏方案,把编码为 RNA 序列的明文嵌入到参考 RNA 序列,以自由能作为约束条件,通过预测软件和设置条件实现信息的隐藏和提取. 同年,X. W. FANG 等[18]提出了一种基于 DNA 芯片的信息隐藏方案,将信息分为普通信息 Mo 和秘密信息 Ms,其中 Ms 被嵌入到芯片中,当发送信息后,只有预期的收件人能够读取 Ms,其他人员在没有密钥的情况下,只能读取 Mo. 对于秘密信息和普通信息而言,相应的 DNA 微阵列可以是相同的,因此统计隐写分析无法检测一个给定的 DNA 芯片是否包含秘密信息.

总之,DNA 密码是利用 DNA 的生物特性,借助于生化反应来实现信息的加密与解密,故其安全性不依赖于数学困难问题,具有一些潜在的、独特的优良特性. 但是,现有的 DNA 加密方法需要借助于生物操作在实验室完成,成本较高且耗时;加密后的信息也无法通过现有的计算机网络进行有效传输,仅适合在一些特殊

领域里应用. 为此,本文拟从 DNA 数据库中大量的 DNA 序列出发,充分利用 DNA 数据库中海量序列的优势来设计信息隐藏方案,以确保信息安全,减少耗时的生物操作,降低成本.

1 DNA 序列数据库

DNA 序列数据库是所有已知核酸信息集合的一个数据资料库,它包含核酸的核苷酸序列,单核苷酸多态性、结构、性质、相关描述等内容. 它的建立不但极大地方便了生物研究人员的科研工作,而且还具有更深层的生物学意义. 目前国际上比较重要的核酸(含蛋白质)一级数据库有美国国家生物技术信息中心的 GenBank,欧洲生物信息学研究所的 EMBL 和日本国立遗传学研究院的 DDBJ,这些数据库是同步更新的. 其中,美国 GenBank 数据库中最常用的是序列文件,其基本单位是序列条目,由核苷酸碱基排列顺序和注释两部分组成. 目前,该数据库文件可以从生物信息资源中心通过计算机网络获得. 序列在数据库中的 ID 号被称为序列代码,它具有唯一性和永久性,用户可以通过序列查询系统从数据库中检索和获取序列数据及相关信息.

DNA 序列数据库的规模正在以指数方式增长,平均不到 9 个月就增加 1 倍. 1998 年 1 月,EMBL 中收录了 15 500 个物种的序列,其序列数目已超过百万,其中 50% 以上为模式生物的序列[19]. 随着测序技术的快速发展,DNA 序列数据库中序列信息的增长速度变得更快,到目前为止能够公开获取的 DNA 序列超过 1.63 亿条[14]. 如此巨大规模的 DNA 序列,相当于一个天然的密码本.

2 隐藏方案设计

本文利用 DNA 数据库中大量的 DNA 序列作为密码本,选取某一条作为参考序列,通过相应的序列提取原则并结合 DNA 数字编码规则,设计相应的算法来实现信息的隐藏.

2.1 DNA 序列提取原则

DNA 序列数据库中的 DNA 序列数量庞大,且具有永久性和唯一性,为我们提供了巨大的信息处理资源,如同“一次一密”中的密码本. 只要确定选择的序列 ID 号,这条序列的信息就被唯一锁定. 因此,只要根据信息隐藏要求,结合算法需要,随机从数据库中选取合适的 DNA 序列即可.

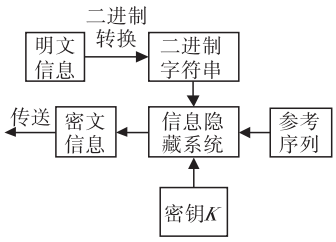
2.2 DNA 数字编码规则

在计算机信息中,最基本的编码方式是二进制编码,即用 0 或 1 以及它们的组合状态来对任一事件进行编码. 数字编码易于进行数值运算和数据处理,便于存储和查询. 而 DNA 分子用 A,T,G,C 这 4 种碱基进行字符编码来实现信息的存储,虽然有其优势,但不便于进行数值运算和处理. 为便于处理,这里用 00,01,10,11 对 DNA 序列中的 4 种碱基进行数字编码[20]. 显然这种编码格式共有 24 种组合方式. 如果从生物学角度考虑 A 与 T 配对,G 与 C 配对,则转换后若 $A=00$,则 T 只能选择 11,那么最终只能在这 24 种组合方式中保留8 种编码规则[21]. 从数学的角度出发,转化后若 $A=00$,则 T 可以从 01,10,11 中任意选择其中的一个.

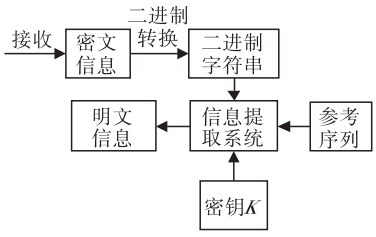
2.3 信息隐藏模型

信息隐藏模型见图 1. 模型中二进制字符串是将明文信息转换成对应的 ASCII 码的二进制表示;参考序列为基因数据库中的 DNA 序列;密钥 $K=(R,p,r,q)$,其中 R 是 DNA 数字编码规则,p 为事先商定的参考 DNA 序列 ID 号,r 是一个随机数,q 为引物信息. r 的选取所遵循的原则:从 r 开始选取的碱基都必须在序列 p 中,否则重新生成. 例如:AC168908 序列含 218 028 个碱基对,如果从中选取 100 个碱基,则$r=218\ 020$,显然此时只能从中取出 9 个碱

基,剩余的 91 个碱基无法继续获取,故 r 无效,需要重新生成.



a) 信息隐藏



b) 信息提取

图 1 信息隐藏模型

Fig. 1 Model of information hiding

2.4 信息隐藏算法

基于前面的隐藏模型,信息隐藏及提取流程见图 2.

信息隐藏算法设计如下.

输入:密钥 $K=(R, p, r, q)$,明文信息 W,ASCII 码标准 I 表.

输出:隐藏秘密信息的 DNA 序列 Ms.

步骤 1 将待隐藏的明文信息 W 转换成对应的 ASCII 码二进制字符串 B_1.

步骤 2 确定密钥 $K=(R, p, r, q)$.

步骤 3 根据 p 代表的参考 DNA 序列,开始信息隐藏:1)将二进制字符串 B_1 中的数据每 2 位 1 组,分为 n 组;2)由随机数 r 在参考序列 p 中找到开始进行信息隐藏的碱基,从该碱基开始依次取出 n 个碱基,组成新的 DNA 序列 S,然后通过 DNA 数字编码规则 R 将该序列转换为二进制字符串 B_2;3)将 B_1 和 B_2 进行二进制异或运算得到二进制字符串 B_3,再通过规则 R 将 B_3 转换为序列 M.

步骤 4 将引物信息 q 加到序列 M 两端生成密文序列 Ms,并加入冗余序列,公开传送.

信息提取过程是信息隐藏过程的逆过程,其算法设计如下.

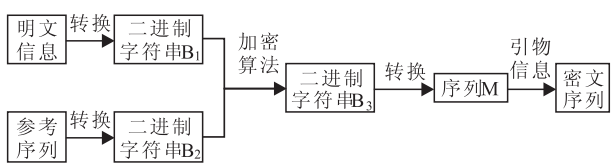
输入:密文信息 Ms,密钥 K,ASCII 码标准 I 表.

输出:明文信息 W.

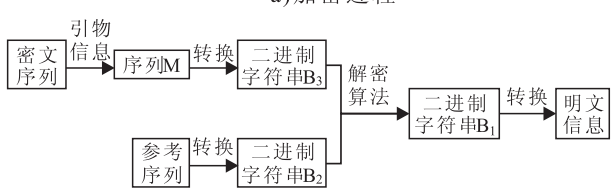
步骤 1 接收密文信息,通过引物信息 q,规则 R 将密文序列转换为二进制字符串 B_3.

步骤 2 由密钥 K 指导下载参考 DNA 序列,开始信息提取:1)由随机数 r 在参考序列中找到开始进行信息隐藏的碱基,从该碱基开始依次取出与密文序列相同个数的碱基,组成新的 DNA 序列 S,然后通过规则 R 将该序列转换为二进制字符串 B_2;2)将 B_3 和 B_2 进行二进制异或运算,得到二进制字符串 B_1.

步骤 3 根据 ASCII 码标准 I 表,将二进制字符串 B_1 恢复成相对应的明文信息 W.



a) 加密过程



b) 解密过程

图 2 信息隐藏及提取流程图

Fig. 2 Flow diagram of information hiding and extraction

2.5 算法仿真

假设要隐藏的信息为 Zhengzhou,将其转换为二进制字符串的密文序列见表 1. 选择的密钥 K 如下.

R:$A=00$, $C=11$, $G=01$, $T=10$;

p:M90100;

r:1 029;

q:前引物 5'-TCATGCTCAGCATTGGTCGTATGG-3',后引物 5'-CCCACGGTTGATAGGTTGATGCTC-3'

表 1 密文序列生成表

Table 1 Generation of cipher text

明文信息	B_1	序列 S	B_2	B_3	序列 M
Z	01011010	ATGA	00011000	01000010	TCAC
h	01101000	GCAG	10110010	11011010	ATGG
e	01100101	TTGT	01011001	00111100	GCCG
n	01101110	TCCA	01111100	00010010	CGGG
g	01100111	GACA	10001100	11101011	CCAC
z	01111010	AGCA	00101100	01010110	GGTC
h	01101000	GGCT	10101101	11000101	ATTT
o	01101111	AATA	00000100	01101011	TCTA
u	01110101	CTGA	11011000	10101101	TACT

事先商定的参考 DNA 序列为人环氧化酶-2的 mRNA 序列,该序列在 GenBank 数据库中的 ID 号为 M90100,包含 3 387 个碱基对.该序列的碱基信息如下(部分碱基被省略,其中小写的碱基即为算法中用到的碱基,第 1 个碱基 a 所在位置数即为 1 029):

5'-GTCCAGG … atgagcagttgttccagacaagcaggctaatactga…AAAAAAAAAAG-3'

随机确定该序列中起始碱基位置数为 1 029,则生成的密钥 $K=(R,$ M90100,1 029, $q)$. 接收者通过公共信道获得密文后,经过简单的处理便可恢复明文信息. 运行信息隐藏算法后得到的密文信息序列为 5'-TCATGCTCAGCATTGGTCGTATGGTCACATGGGCCGCGGGCCACGGTCATTTTCTAT-ACTCCCACGGTTGATAGGTTGATGCTC-3'.

3 分析与讨论

目前,对于 DNA 密码系统的安全性分析理论尚未成熟[22]. 这里仅从性能(容量、负载、嵌入容量)[14] 和安全性两方面对本方案进行讨论.

3.1 性能(容量、负载和嵌入容量)分析

根据文献[14]中关于性能的评价指标,通过容量、负载和嵌入容量对该隐藏方案进行性能评价. 容量是指嵌入秘密信息后 DNA 序列的总长度,负载是指取出参考 DNA 序列后剩余的序列长度,嵌入容量是指每一个碱基嵌入的比特数. 表 2 为本方案的嵌入方法与文献[14]中 3 种方法的性能比较,其中,记引物 DNA 序列的长度为 $|S|$,秘密信息 DNA 序列的长度为 $|M|$. 从表 2 可以看出,该方案与文献[14]中的插入法类似,但是该方法有更好的隐蔽性.

表 2 本方案的嵌入方法与文献[14]中 3 种方法的性能比较

Table 2 Comparisons among three schemes in the literature[14] and the proposed scheme

方法	容量	负载	嵌入容量
插入法	$\lvert S\rvert+\frac{\lvert M\rvert}{2}$	$\frac{\lvert M\rvert}{2}$	$\frac{\lvert M\rvert}{\lvert S\rvert+\frac{\lvert M\rvert}{2}}$
互补对法	$\lvert S\rvert+\lvert M\rvert(k+35)$	$\lvert M\rvert(k+35)$	$\frac{\lvert M\rvert}{\lvert S\rvert+\lvert M\rvert(k+35)}$
替换法	$\lvert S\rvert$	0	$\frac{\lvert M\rvert}{\lvert S\rvert}$
本方案	$\lvert S\rvert+\lvert M\rvert$	$\lvert M\rvert$	$\frac{\lvert M\rvert}{\lvert S\rvert+\lvert M\rvert}$

3.2 安全性分析

首先,攻击者可能会试图寻找用来进行信息隐藏的参考 DNA 序列,但只能通过搜索 DNA 序列数据库来找出发送者所采用的 DNA 序列. 到目前为止,可以公开获取的 DNA 序列超过 1.63 亿条,攻击者要想从如此巨大规模的数据库中确定 1 条 DNA 序列是非常困难的,这种暴力破解成功的概率只有 1.63×10^{-8}.

其次,如果攻击者对密文 DNA 序列进行暴力破解,序列中的每个碱基都有 5 种可能:隐藏后的 00,01,10,11 或者没有隐藏信息,那么攻击者进行序列攻击的破解概率是 $1/5^n$.

序列攻击破解概率曲线如图 3 所示. 由

图 3 可以看出,随着序列长度的增加,进行序列攻击破解的概率逐渐趋于 0.

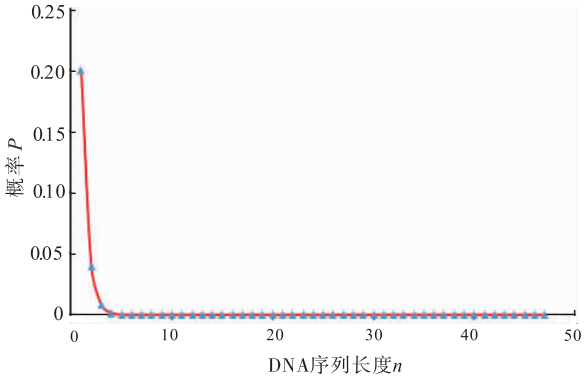


图 3 序列攻击破解概率曲线

Fig. 3 Probability curve of sequence attack

攻击者要想进一步提高破译密文的概率,还必须确定密钥 K 中的随机数 r,从而确定参考 DNA 序列中开始隐藏信息的位置. 一条 DNA 序列可能有几千到几十亿个碱基对,文中采用的参考 DNA 序列含 3 387 个碱基对,可想而知,攻击者获得正确随机数 r 的几率也是很小的.

最后,将明文信息隐藏后得到的密文 DNA 序列呈现杂乱状态,这些杂乱的 DAN 序列要经过两次转换后才能最终获得明文信息. 第 1 次转换是通过 DNA 数字编码规则实现的,第 2 次转换是通过算法实现的. 最终,转换前跟转换后呈现出一对多映射的假象. 本文中密文 DNA 序列含 84 个碱基对,而转换为明文信息的有效序列含 36 个碱基对. 如果攻击者锁定了有效碱基的位置,那么破译的概率是 4^{-36}. 实际情况下,对于攻击者,在没有引物信息 p 的前提下无法确定有效碱基. 因此对于攻击者而言,即使他获得了密文 DNA 序列信息,成功破译密文的概率仍然很低.

总之,该方案的实现减少了耗时的生物操作,降低了所需成本. 加密后的信息也可以通过现有的计算机网络进行有效传输.

4 结语

本文以 DNA 序列数据库中大量的 DNA 序列为出发点,研究了基于 DNA 序列的信息隐藏方案,设计了相应的信息隐藏算法,并从方案的性能和安全性两方面进行了分析. 结果表明:本方案不需要生物操作,成本低,易于传输,有较强的安全性,有助于 DNA 资源的有效利用,是对 DNA 密码学发展的有益探索.

参考文献:

[1] 王延峰,韩琴琴,韩栋,等. 基于核酸的信息安全技术研究现状及发展建议[J]. 中国科学院院刊,2014,29(1):83.

[2] WANG X Y,YU H B. How to break MD5 and other hash functions[J]. Lectrue notes in computer science,2005,3494:19.

[3] WANG X Y,YIN Y Q L,YU H B. Finding collisions in the full SHA-1[J]. Lectrue notes in computer science,2005,3621:17.

[4] NAOR M,SHAMIR A. Visual cryptography[J]. Lectrue notes in computer science,1995,950:1.

[5] EKERT A K. Quantum cryptography based on Bell's theorem[J]. Physical review letters,1991,67(6):661.

[6] 肖国镇,卢明欣,秦磊,等. 密码学的新领域——DNA 密码[J]. 科学通报,2006,51(10):1139.

[7] ADLEMAN L M. Molecular computation of solutions to combinatorial problems[J]. Science,1994,266(5187):1021.

[8] BONEH D,DUNWORTH C,LIPTON R J. Breaking DES using a molecular computer[J]. DNA based computers,1996,27:37.

[9] CELLAND C T,RISCA V,BANCROFT C. Hiding messages in DNA microdots[J]. Nature,1999,399(6736):533.

[10] SHIMANOVSKY B,FENG J,POTKONJAK M. Hidingdata in DNA[C]. Lectrue notes in computer science,2003,2578:373.

[11] GEHANI A,LABEAN T,REIF J. DNA-based cryptography[J]. Lectrue notes in computer science,2004,2950:167.

(下转第 88 页)

dependent molecular arrangement and topography of ultrathin ionic liquid films on a silica surface[J]. Chem Commun,2013,49(71):7803.

[24] KAMBOJ R, BHARMORIA P, CHAUHAN V, et al. Effect of cationic head group on micellization behavior of new amide-functionalized surface active ionic liquids[J]. Phys Chem Chem Phys,2014,16(47):26040.

[25] TAKAOKA G H, TAKEUCHI M, RYUTO H, et al. Production and irradiation of ionic liquid cluster ions[J]. Nucl Instrum Meth B, 2013, 307:257.

[26] TAKAOKA G H, HAMAGUCHI T, TAKEUCHI M, et al. Surface modification using ionic liquid ion beams[J]. Nucl Instrum Meth B,2014,341:32.

[27] FU Y C, SU Y Z, WU D Y, et al. Supramolecular aggregation of inorganic molecules at Au(111) electrodes under a strong ionic atmosphere[J]. J Am Chem Soc,2009, 131(41):14728.

[28] PAN G B, FREYLAND W. 2D phase transition of PF_6^- adlayers at the electrified ionic liquid/Au(111) interface[J]. Chem Phys Lett,2006,427:96.

[29] ATKIN R, ABEDIN S Z E, HAYES R, et al. AFM and STM studies on the surface interaction of[BMP]TFSA and[EMIm]TFSA ionic liquids with Au(111)[J]. J Phys Chem C,2009,113(30):13266.

[30] ATKIN R, BORISENKO N, DRüSCHLER M, et al. An in situ STM/AFM and impedance spectroscopy study of the extremely pure 1-butyl-1-methylpyrrolidinium tris(pentafluoroethyl)trifluorophosphate/Au(111) interface: potential dependent solvation layers and the herringbone reconstruction[J]. Phys Chem Chem Phys,2011(13):6849.

[31] SEGURA J J, ELBOURNE A, WANLESS E J, et al. Adsorbed and near surface structure of ionic liquids at a solid interface[J]. Phys Chem Chem Phys,2013,15(9):3320.

[32] SU Y Z, FU Y C, YAN J W, et al. Double layer of Au(100)/ionic liquid interface and its stability in imidazolium-based ionic liquids[J]. Angew Chem, 2009, 48(28):5148.

[33] SU Y Z, YAN J W, LI M G, et al. Electric double layer of Au(100)/imidazolium-based ionic liquids interface: effect of cationsize[J]. J Phys Chem C,2013, 117(1):205.

[34] BUCHNER F, FORSTER-TONIGOLD K, UHL B, et al. Toward the microscopic identification of anions and cations at the ionic liquid|Ag(111) interface: a combined experimental and theoretical investigation[J]. ACS Nano, 2013,7(9):7773.

[35] KAISEI K, KOBAYASHI K, MATSUSHIGE K, et al. Fabrication of ionic liquid thin film by nano-inkjet printing method using atomic force microscope cantilever tip[J]. Ultramicroscopy,2010,110:733.

(上接第 66 页)

[12] TANAKA K, OKAMOTO A, SAITO I. Public-key system using DNA as a one-way function for key distribution[J]. Biosystems,2005,81(1):25.

[13] 崔光照,秦利敏,王延峰,等. 基于 DNA 技术的加密方案[J]. 计算机工程与应用,2009,45(8):104.

[14] SHIU H J, NG K L, FANG J F, et al. Data hiding methods based upon DNA sequences[J]. Information of science, 2010,180(11):2196.

[15] GUO C, CHANG C C, WANG Z H. A new data hiding scheme based on DNA sequence[J]. International journal of innovative computing information & control, 2012, 8(1):1.

[16] GOFF G C L, BLUM L J, MARQUETTE C A. Shrinking hydrogel-DNA spots generates 3D microdots arrays[J]. Macromolecular bioscience,2013,13(2):227.

[17] 张勋才,韩琴琴,王燕,等. 一种基于 RNA 二级结构的信息隐藏方案[J]. 郑州轻工业学院学报(自然科学版),2014,29(1):1.

[18] FANG XW, LAI XJ. DNA-chip-based information hiding scheme[J]. Communications in computer and information science,2014,472:123.

[19] ATTWOOD T K, PARRY-SMITH D J. 生物信息学概论[M]. 罗静初译. 北京:北京大学出版社,2002:83.

[20] 饶妮妮. 一种基于重组 DNA 技术的密码方案[J]. 电子学报,2004,32(7):1216.

[21] 陈惟昌,陈志华. 遗传密码和 DNA 序列的高维空间数字编码[J]. 生物物理学报,2000,16(4):760.

[22] AMOS M, PAUN G, ROZENBERG G, et al. Topics in the theory of DNA computing[J]. Theoretical computer science,2002,287(1):3.