DNA 信息隐藏方法的安全性分析和保密增强方法

卢明欣[1]，傅晓彤[1]，秦 磊[2]，肖国镇[1]

(1. 西安电子科技大学 综合业务网理论与关键技术国家重点实验室，陕西 西安 710071；
2. 皇后大学 癌症研究所，金斯顿 安大略省 加拿大 K7L3N6)

摘要：针对一种 DNA 微点信息隐藏方法，提出了用 PCR 扩增、DNA 测序等技术攻击的方法，论证了由于寻找匹配引物的困难性和测序中电泳技术以及纯化技术的限制，上述攻击不能奏效. 论证了可以用可能词作为 PCR 引物进行有效的攻击，提出要以引物和编码方式为密钥或者用传统加密技术对要隐藏的信息预先进行加密以防止该种攻击.

关键词：信息隐藏；安全性；DNA

中图分类号：TN918.1 **文献标识码**：A **文章编号**：1001-2400(2006)03-0448-05

A study of the security of a steganography method and security enhancement methods

LU Ming-xin[1], *FU Xiao-tong*[1], *QIN Lei*[2], *XIAO Guozhen*[1]

(1. State Key Lab. of Integrated Service Networks, Xidian Univ., Xi'an 710071, China; 2. Queen's Univ., Cancer Research Institute, 10 Stuart St. Kingston, ON K7L3N6, Canada)

Abstract: Based on PCR and DNA sequencing technology, several attacks against a DNA steganography method are proposed and discussed. Due to the limitation of finding complementary primes, electrophoretic technology and separation technology, attacks above do not work. The attact by using possible words as primes is also discussed, and security enhancement methods to prevent such attacks are proposed.

Key Words: steganography; security; DNA

随着 DNA 计算的发展，新兴的 DNA 密码学也得到了发展，有可能和传统的密码学、量子密码学成为密码学的三大分支. DNA 密码是在 1994 年 Adleman 提出 DNA 计算(也就是俗称的生物计算机)之后才开始得到关注，目前还处于探索阶段. 从实现技术和安全依据上看，传统的密码借助于电子计算设备，通过数学运算进行加密和解密，其安全性依赖于各种数学困难问题(如大整数分解问题、离散对数问题等)；量子密码是利用量子通信技术实现的，其安全性主要依赖于物理学定律——海森堡测不准定理，该定理也可理解为一个物理学上的困难问题即量子测量的困难问题[1]；DNA 密码是利用现代基因工程技术实现的，其安全性主要依赖于各种生物学上的困难问题.

DNA 密码目前已经有了一些初步的成果，如 Reif 等科学家认为，每克 DNA 就含有大约 10^{21} 个核苷酸，如果按照四进制编码，可看作是大约 10^{20} 字节[2]. 这样超高容量的存储密度非常适合存储一次一密乱码本. 在此基础上，他提出了利用 DNA 实现的一次一密方法. 不过，在现有的生物技术条件下，Reif 的方案实现代价非常高. 此外，Celland 等人用 DNA 微点实现了信息隐藏，把著名的"June 6 invasion: Normandy"隐藏到 DNA 微点中[3]. 遗憾的是，Celland 等人的论文中几乎没有涉及安全性论述，只是在文中提出希望后人能对这个系统的安全性进行数学和生物学方面的分析. 笔者介绍了 Celland 等人提出的 DNA 信息隐藏方法(该

收稿日期：2005-10-19

基金项目：国家重点基础研究发展规划项目(973 项目)(G1999035804)

作者简介：卢明欣(1974-)，男，西安电子科技大学博士研究生.

方法也可看作是一个私钥加密系统),分析了该系统的安全性并提出保密增强的方法,是对该信息隐藏方法的补充性研究. 其论述也在一定程度上反映出 DNA 密码在加密和解密过程,密文的形式以及安全依据等方面与数学密码的不同.

1　背景知识

DNA 的学句是脱氧核糖核酸. 从构成上看,DNA 是由核苷酸组成的一种生物大分子. 核苷酸含有 4 种不同的碱基:腺嘌呤(A)、鸟嘌呤(G)、胞嘧啶(C)和胸腺嘧啶(T). 相应地,核苷酸也按所含碱基的不同分成 4 种,核苷酸排列成链状. DNA 分子由两条长链组成,在氢键的作用下两条链连接在一起,呈现出螺旋式结构. 双链连接的方式是按照碱基互补配对的原则,即 A 与 T 始终配对存在;G 与 C 始终配对存在. DNA 双螺旋结构是由 James D. Watson 和 Francis H. Crick 发现的,所以称为 Watson-Crick 对.

PCR 技术是非常重要的生物学技术,也是 DNA 在计算、信息储存、加密、信息隐藏等领域中应用的重要工具. DNA 体积微小,双螺旋结构的直径约 2 nm,螺距约 3.5 nm. 对体积微小且数量极少的特定 DNA 片段进行操作非常困难,通过扩增技术,把少量的特定 DNA 大量复制后操作就容易多了. PCR 技术就是一种快速的特定 DNA 片断扩增技术. PCR 是基于 Watson-Crick 互补配对特性实现的. 它通过分别与双链目的 DNA 序列两个 3′端互补的寡核苷酸引物,由 Taq DNA 聚合酶从 5′→3′进行一系列 DNA 聚合反应,扩增出所需的目的 DNA. 在这里,拥有已知的或对某一物种通用的引物,是进行 PCR 扩增的关键. 该技术十分灵敏,理论上每一个目的 DNA 分子经 20 轮扩增后,数量可达 10^6,从而实现短时间内大量扩增 DNA 序列.

2　DNA 信息隐藏方案介绍

1999 年,Celland 等人成功地把著名的"June 6 invasion: Normandy"隐藏在 DNA 微点中,从而实现了利用 DNA 作为载体的信息隐藏[3]. 他们的方法如下:

(1) 确定编码方式. 他们没有采用传统的二进制编码方式,而是把核苷酸看作是四进制编码,用 3 位核苷酸表示 1 个字母. 譬如字母 A 用核苷酸序列 CGA 表示,字母 B 用核苷酸序列 CCA 表示,…….

(2) 制作消息序列. 把需要传递的消息按上面的编码方式编成相应的 DNA 序列,如 AB 用 CCGCCA 表示. 编码结束以后,人工合成相应的有 69 个核苷酸的 DNA 序列,并在 DNA 序列前后各链接上有 20 个核苷酸的 5′和 3′引物. 这样,需要隐藏的 DNA 消息序列就准备好了.

(3) 信息隐藏. 用超声波把人类基因序列粉碎成长度为 50-100 的核苷酸双链,并变性成单链,作为冗余的 DNA 使用,再把含有信息的 DNA 序列混杂到冗余的 DNA 序列中,喷到信纸上形成无色的微点,就可通过普通的非保密途径传送了.

(4) 信息读取. 接收方和发送方的共享秘密是编码方式和引物. 接收方收到含有消息 DNA 微点的信纸后,提取出微点中的 DNA. 由于接收方预先通过安全的途径得到了引物,所以他可以用已有的引物对含有消息的 DNA 序列进行 PCR 扩增,通过测序得出消息 DNA 序列,然后根据预先约定的编码方式恢复出消息(明文).

图 1 是 Celland 的信息隐藏方法介绍,图 1(a)是合成的消息序列,图 1(b)是编码方式,图 1(c)是 PCR 扩增结果,图 1(d)是 PCR 扩增后,通过测序得到的消息序列以及对应的明文. 原图对于未用到的字母,如 F,P,Z 省略了对应的 DNA 码字.

3　安全性讨论

传统密码学领域尚有许多问题远未解决,同传统密码的研究相比,生物学家的研究更加不完善. 他们的工作主要还停留在实验阶段,缺乏完善的理论,更缺乏有效的方法来对生物学困难问题进行衡量. 这里就以生物学上共识性的困难问题作为依据,对 Celland 等人所提方案的安全性进行论证. 当然,DNA 密码系统的安全性很复杂,除了生物学上的困难问题外,数学工具也是构成系统安全的基石之一.

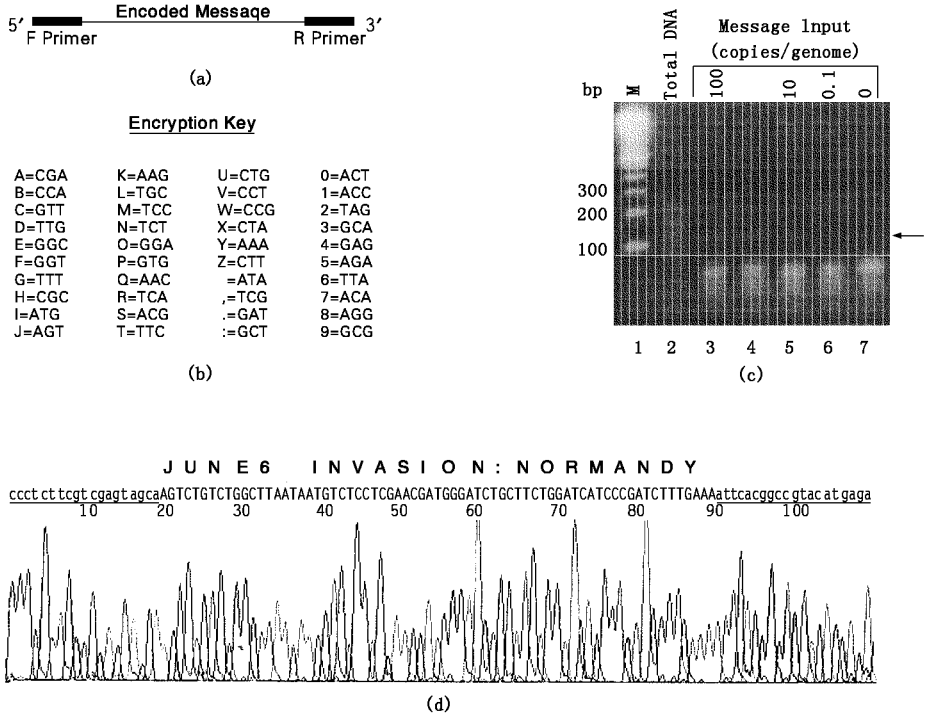

图 1 信息隐藏与提取

文献[3]中提出的 DNA 信息隐藏方法具有 3 层安全性,要攻破本系统就要突破全部 3 层安全性. 第 1 层安全性是 DNA 微点无色无味,喷涂在物品表面不易被发现. 除非对附带有 DNA 微点的物品进行专门检查,否则难以发现 DNA 微点的存在. 但是,这种安全性并不比使用隐写墨水安全多少,用攻破隐写墨水的技术就足可以找出这些微点. 鉴于隐写墨水在历史上已经被轻松破译,可认为第 1 层安全性是很脆弱的. 第 2 层安全性是因为在现有的生物技术条件下,要从未知的 DNA 混合液中分离出未知的特定 DNA 序列并测序是困难的,这是从生物学方面保障了该信息隐藏方案的安全性. 第 3 层安全性是由数学编码方法提供的.

首先讨论第 2 层安全性——生物学的安全性. 生物学仍然是以实验事实为主的,与以定理和公式为主的、精确的数学是大不相同的. 所以,这里进行安全分析也主要是以分析实验事实而不是公式推导为依据. 假设攻击者找到了这些 DNA 微点,要想得到消息序列,有如下几种可能的攻击方式:

攻击方法 1:用引物把特定的序列扩增出来以测序(这相当于没有密钥而尝试随机寻找密钥的穷举攻击法). 指定的消息接收者有对应于消息序列的引物,所以他可对特定的消息序列进行 PCR 扩增并测序从而得到消息. 这类似于拥有解密密钥可方便解密的情况. 对于一个攻击者,如果他没有这些引物,他就只能是随便找些引物尝试进行扩增,这相当于没有解密密钥随机选择一个密钥进行尝试的情况. 首先考虑随机寻找引物进行扩增的方法. 考虑到扩增的效果,比较理想的引物长度是 20 个核苷酸. 用这么长的引物扩增的结果是最稳定的. 一般情况下,如果在发生错误匹配,使得匹配的引物长度变为 17 个核苷酸,那么有 3 个核苷酸错误匹配并且均在远离待扩增序列端的情况下,还是能够得到可接受的结果. 那么,分离并扩增所有这些 DNA 序列所需要的引物数量为 4^{34} 对. 合成少数引物并且期望能够偶然成功的概率是极其微小的. 如果有人想合成所有这些引物,所需要的财力物力是无法承受的. 并且,即使得到了所有这些引物,如果不知道哪些是真正需要的,就只能把这些引物进行逐个尝试. 然而,DNA 混合物并不像数字那样可被反复利用而不产生变化,PCR 扩增也不是像数学运算那样精确. 这种尝试如果多次进行,会对含有消息序列的 DNA 混合物造成污染,严重影响 PCR 扩增的进行. 要把所有的引物都尝试一遍并且还期望能得到正确的结果也是不可能的.

攻击方法 2:对微点中的所有序列测序,然后进行数学分析找出消息序列. 这种方法相当于传统的信息隐藏方法中对所有数据进行检测. 但是,传统的信息隐藏方法所有数据都是可方便读出的,而 DNA 信息隐藏方法的数据难以读出,所以也就难以检测. 这是因为,DNA 微点中的序列是未知的混合序列,使用现有的测序方法都无法进行有效的测序. 现在主要有两种测序方法:Maxam-Gilbert 方法(化学法降解法)[4]和 Sanger 的方法(双脱

氧链末端终止法或酶法)[5]. 这两种方法的检测序列的关键一个步骤都是通过电泳(凝胶电泳或者毛细管电泳)读取序列. 如果要测的是混合序列,进行电泳的 A,T,C,G 4 个泳道里就都会有信号,这样就难以区分序列组成,也就无法完成测序. 所以,这两种方法都无法对混合序列进行有效的测序. 即使有人根据这些信号进行数学分析,考虑到消息序列和冗余序列都具有一定的随机性,并且消息序列的数量比较少,泳道中由消息序列表达的信号不会很强,也就不太可能分析出有价值的信号. 此外,Sanger 的方法需要用引物退火到 DNA 模板上才能完成测序反应. 如果对待测序列完全未知,就无法合成测序所必须的引物. 这也是 Sanger 的方法不能对未知混合 DNA 序列进行测序的一个原因. 目前几乎所有的测序方法都是基于 Sanger 的方法,包括最精确的质谱测序法(MS)[6]. 近年来,也提出了一些其他的方法作为 Sanger 方法的替代,包括使用酶消化核酸片断的 MS 方法[7,8],以及用 DNA 微阵列进行测序的方法[9,10]. 但是这些方法都还处于探索阶段,并且都有自身的缺点. 例如 MS 方法就要求极纯的 DNA 样本. 如果试图把 DNA 混合物纯化以后再测序,是极其困难的. 文献[11]中指出,对于仅仅是序列中核苷酸排列不同的 DNA 序列,即使用先进的磁珠法进行纯化,使用长度为 20 mer的带有磁珠的探针来分离长度为 40 mer的目标 DNA 探针时,也最多能分离纯化出 1%的 DNA 序列. 在实际的操作中,所能纯化出的 DNA 序列是远小于 1%的.

其次,讨论第 3 层安全性——数学方面的安全性. 如果很多年以后,第 1 和第 2 层安全性都被突破了,那么攻击者的破译工作就是从一些数字序列中找出消息序列. 对于消息的发送者和接收者,如果引物不重复使用,那么,就要保存数量巨大的引物,并且在数据读取的时候还要知道如何选择正确的引物,这可能也需要在发送消息时发送额外的信息. 这样操作,所需要的工作量巨大,类似于一次一密. 如果引物重复使用,虽然可使得实现更容易,但是该系统就可用下面的简单方法来破译.

攻击方法 3:在两次加密的过程中,对于消息序列,引物是固定的,而 5′和 3′引物之间的消息序列是变化的;对于冗余序列,如果重复使用,整个序列就是不变的,如果不重复使用,整个序列就都不同. 通过这些不同的特征,借助于电子计算机的分析,就可比较容易地找出消息序列来. 此后,针对文献[3]中提出的编码方案,只要能够得到足够多的密文数量,就可根据语言的统计特性,按照英文字母频度进行攻击. 如英文中出现频度最高的是字母 E,对于大量文献统计的结果是出现概率约为 0.127[12]. 对应的,就是消息序列中 GGT 出现的频度最高. 采用这种初等密码分析方法,就可很容易地攻破文献[3]中提出的编码方案.

4　保密增强

针对该信息隐藏方案,这里提出需要注意如下问题并采取相应措施以增强其保密性.

在文献[3]中提到,加密钥是编码方式. 在上面的论述中可看出,引物才是真正的密钥. 如果引物泄漏了,该系统的编码方式只具有简单的安全性. 一般来说,一个信息隐藏系统的编码方式是不太重要的,采用普通的编码方式(如电子计算机中普遍使用的二进制码)就可以了. 那么,该系统的编码方式的保密性是否也是毫不重要,甚至可以公开呢?答案是否定的. 这是因为编码方式的泄漏,会导致如下的安全漏洞. 任何一个密码系统都是有一定的应用背景的,相应地,特定的关键词也就会大量出现. 比如,针对文献[3]中隐藏的消息,虽然攻击者不知道消息的具体内容,但是可以用把关键词作为 PCR 扩增引物的方法进行攻击. 具体过程如下:

(1) 攻击者猜到消息里很可能会有 June, July, invasion, Normandy, Calais 之类的可能词.

(2) 如果编码方式泄漏了,就可以把上述可能词转换成 DNA 序列,然后作为引物对微点中的 DNA 序列进行 PCR 扩增. 比如, June6 就可编码为 AGTCTGTCTGGCTTA, Normandy 可编码为 TCTGGATCATCCCGATCTTTGAAA.

(3) 把所有这些可能词都编码成 DNA 序列,把这些序列和这些序列的互补以及转置序列作为引物放到 DNA 混合物中,无关的可能词不会引起 PCR 反应. 而"June"和"Normandy"这两个可能词编码成的引物会引起 PCR 反应,把"June 6 invasion: Normandy"这句话完整地扩增出来. 此处,"June6"编码序列 AGTCTGTCTGGCTTA 是引物 1,"Normandy"编码的互补反转序列 TTTCAAAGATCGGGATGATCCGA 作为引物 2.

具体过程见图 2. 图 2 中上面一条 DNA 长链是消息序列,下面一条 DNA 长链是消息序列的互补序列.

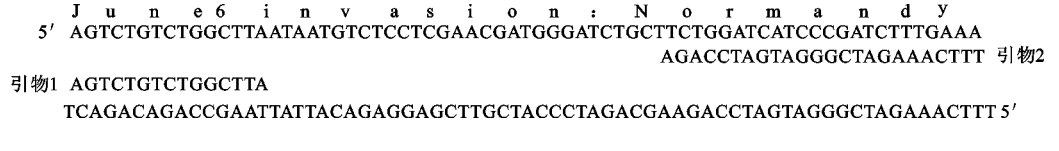

图 2 用可能词进行 PCR 扩增

所以,文献[3]中的加密钥应该是由引物和编码方式共同组成.鉴于对每组引物更换编码方式比较困难,这里提出进行隐藏前,对明文预先采取一定的数学变换,比如预选先用 DES 之类的加密算法加密,这样可使得攻击者难以找到关键词作为引物.此时,加密钥就由引物和一个普通的数学加密系统的密钥组成.这种方法的好处是对于相同的明文,改变数学加密钥就可改变对应的 DNA 消息序列,使用更方便.

5 结 束 语

讨论了用 PCR 扩增、DNA 测序等技术对文献[3]提出的 DNA 信息隐藏方法进行攻击的方法,提出由于寻找匹配引物的困难性和测序中电泳技术以及纯化技术的限制,上述攻击不能奏效.论证了可能词可作为 PCR 引物成功攻击该方法,提出要以引物和编码方式为密钥或者用传统加密方法对要隐藏的信息先进行加密.文中提出对明文预先用传统方法加密的保密增强方法,可有效防止可能词作为 PCR 引物进行的攻击.

参考文献:

[1] Bennett C H, Brassard G, Ekert A K. Quantum Cryptography[J]. Scientific American, 1992, 267(1): 50-57.

[2] Gehani A, LaBean T H, Reif J H. DNA—Based Cyrptography[A]. 5th DIMACS Workshop on DNA Based Computers [C]. Cambridge: MIT Univ., 1999.

[3] Celland C T, Risca V, Bancroft C. Hiding Messages in DNA Microdots[J]. Nature, 1999, 399(6 736): 533-534.

[4] Gilbert W. DNA Sequencing and Gene Structure[J]. Science, 1981, 214(4 527): 1 305-1 312.

[5] Sanger F. Determination of Nucleotide Sequences in DNA[J]. Science, 1981, 214(4 526): 1 205-1 210.

[6] Edwards J R, Itagaki Y, Ju J. DNA Sequencing Using Biotinylated Dideoxynucleotides and Mass Spectrometry[J]. Nucleic Acid Research, 2001, 29(21): e104.

[7] Pieles U, Zurcher W, Scharl M. Matrix-assisted Laser Desorption Ionization Time-of-flight Mass Spectrometry: a Powerful Tool for the Mass and Sequence Analysis of Natural and Modified Oligonucleotides[J]. Nucleic Acids Research, 1993, 21(14): 3 191-3 196.

[8] Wu H, Chan C, Aboleneen H. Sequencing Regular and Labeled Oligonucleotides Using Enzymatic Digestion and Ionspray Mass Spectrometry[J]. Anal Biochem, 1998, 263(2): 129-138.

[9] Drmanac R. DNA Sequence Determinatioin by Hybridization: a Strategy for Efficient Large-scale Sequencing[J]. Science, 1993, 260(5 114): 1 649-1 652.

[10] Drmanac R. Sequencing by Hybridization(SBH): Advantages, Achievements, and Opportunities[J]. Adv Biochem Eng Biotechnol, 2002, 77(1): 75-101.

[11] Julia K, Gifford D K. The Efficiency of Sequence-Specific Separation of DNA Mixtures for Biologic Computing[A]. 3rd Annual DIMACS Workshop on DNA-Based Computers[C]. Philadelphia: Pennsylvania, 1997.

[12] Stinsin D R. Cryptography: Theory and Practice[M]. [s.l.]: CRC Press, 1995.

(编辑:齐淑娟)