

DNA 序列作为信息隐藏载体的研究^{*}

郑国清¹, 刘九芬², 黄达人³, 徐安龙¹

- (1. 中山大学生物化学系, 广东 广州 510275;
2. 中国人民解放军信息工程大学信息研究系, 河南 郑州 450002;
3. 中山大学科学计算与计算机应用系, 广东 广州 510275)

摘要: 研究了 DNA 序列能否成为信息隐藏的载体。通过对 DNA 序列进行分析证实了: 由于核苷酸序列中有强的随机噪声, DNA 序列可以作为信息隐藏的载体; 进而通过对 DNA 序列特征进行分析, 提出了一个嵌入对策: 秘密消息嵌入非编码区的高复杂度区域有很好的安全性。该文的研究对提出以 DNA 序列为载体的信息隐藏算法具有重要的指导作用。

关键词: DNA 序列; 信息隐藏; 载体

中图分类号: TP391 **文献标识码:** A **文章编号:** 0529-6579 (2005) 01-0013-04

人类基因组计划 (human genome project, HGP) 是人类历史上可以与曼哈顿原子弹计划和阿波罗登月计划相提并论的第 3 大科学计划。它的逐步实施以及分子生物学相关学科的发展, 在短短几年内必将影响生物、医学、农业乃至军事的众多领域。在此背景下, 本文研究以 DNA 序列为载体的信息隐藏。该问题的研究目前尚属空白。

信息隐藏^[1,2]是近年来多媒体信号处理领域提出的一种信息安全的新方法。它通过把秘密信息隐藏在可公开的媒体信息里, 达到证实该媒体信息的所有权归属和完整性或传递秘密信息的目的。近年来随着网络和数字媒体技术的飞速发展, 人们在网络上正在从事越来越多的活动。因此信息安全成为国内外倍受关注的焦点研究课题。以密码学为基础的信息加密技术, 是以往主要的信息安全手段, 今后仍将发挥重要的作用, 但在实际中这是远远不够的。例如, 一份加密之后的 E mail 如果在毒贩和其他尚未被怀疑的人之间, 或者国家工作人员与敌对势力之间传递, 显然, 它包含着某种深意。因此信息安全不仅包含加密, 还包括以隐藏信息为根本的信息安全。

与以多媒体为载体的信息隐藏一样, 基于 DNA 序列的信息隐藏能够传递秘密信息, 不一样的是, 它可以保护医学、分子生物学、遗传学等领域的知识产权。另外, 攻击者不会对 DNA 序列进

行常规的信号处理, 例如, 有损压缩、滤波、加噪声、上/下抽样、A/D 和 D/A 转换等, 也不会进行伸缩、平移、裁剪等几何攻击, 从而基于 DNA 序列的信息隐藏比较稳健。因此研究以 DNA 序列为载体的信息隐藏具有重要的意义。

随着 HGP 的实施, 越来越多的动植物、微生物基因组序列得以测定, 基因序列数据出现了爆炸式增长, 达到每 8 个月翻一番。目前, 平均每 1 min 就有一个序列增加到核酸序列数据库中。而且相互交流、共同协作, 构成了生物信息学界的一股新风气。世界上生物信息学家通过国际互联网共享已有的数据、资源, 相互交流各自提出的分析方法, 生物序列作为 E mail 的内容并不稀奇。这些为 DNA 序列作为信息隐藏的载体提供了可能。

那么作为遗传物资载体的 DNA, 它的序列究竟是否能够成为隐藏信息的载体呢? 本文试图回答这个问题; 然后根据 DNA 序列的特征, 本文探讨嵌入对策问题。

1 相关知识

在探讨 DNA 序列能否成为信息隐藏的载体之前, 必须知道 DNA 序列的相关知识^[3,4]。

一般认为, 生物可分为原核生物和真核生物。真细菌、衣原体、支原体属于原核生物, 酵母、动物属于真核生物。二者的区别主要是真核生物的

* 收稿日期: 2004-03-18

基金项目: 国家自然科学基金资助项目 (60473022, 60133020); 河南省高校杰出人才科研创新工程资助项目 (2003KJCX008); 中国人民解放军信息工程大学博士启动基金资助项目

作者简介: 郑国清 (1964年生), 男, 博士后; E-mail: zgqx@public2.zz.ha.cn.

DNA 由磷脂双分子层包裹，形成细胞核，而原核生物的 DNA 没有核膜包绕，只是聚集在称为拟核的区域。

DNA 是一种大分子，由 2 股长链以螺旋式构成，这种螺旋结构是在 1953 年由 Watson 和 Crick 提出，他们因此获 1962 年的 Nobel 奖。

DNA 分子上的一个个有生物功能、长短不一的片段是基因。基因是所有生命的遗传单位，由若干按一定顺序排列的核苷酸组成。核苷酸由磷酸基团、脱氧核糖及碱基构成，有 4 种不同的碱基，即：腺嘌呤、鸟嘌呤、胞嘧啶及胸腺嘧啶，分别用 A, G, C, T 表示。核苷酸按其所含碱基的不同也分为 4 种。

在 DNA 双链之间，位于内侧的碱基以氢键相结合，A 与 T 始终配对存在，形成 2 个氢键；G 与 C 始终配对出现，形成 3 个氢键。

生命的遗传信息隐藏在碱基的排列顺序之中，而遗传信息的表达却采用多肽（蛋白质）的形式，这样就必须将核酸的 4 字符（A, G, C, T）语言，翻译成多肽序列的 21 字符语言（20 种氨基酸+终止密码子），研究证实，3 个特定排列的碱基对应 1 个氨基酸，所有的对应关系总结成遗传密码。密码子（三联体密码）由 A, G, C, T 中每 3 个字母重复排列而成，共有 $4^3=64$ 个。每一密码子对应一种氨基酸；但由于只有 20 种氨基酸，故这种对应是多对一的，是高度简并的。除了线粒体及很小一部分原核生物外，自然界所有的生命形式都共用这本密码。

在确定了三联体码在 DNA 上线性串接的结合方式后，发现了为蛋白质编码的基因结构。这些编码蛋白质的基因部分在 DNA 上所处的位置，称为 DNA 的编码区，也称外显子。基因序列中处在外显子之间的序列片段，称为内含子。编码区在高等生物和人类的基因组中的比例不超过 5%。例如：人类基因组中编码区只占 1.1%，内含子占了 24%，而基因与基因之间的间隔序列却占了 75%。

真核基因具有一些关键特征（图 1），这些特征包括内含子、外显子、编码序列（coded sequence, CDS）、未翻译区（untranslated region, UTR）等。而原核基因通常缺少内含子，结构相对简单。UTR 是位于 CDS 两侧的序列，未被翻译成蛋白质。真核生物中，由外显子剪接而得到编码区，而内含子尽管也被转录，但在 mRNA（信使核糖核酸）剪接过程中被剪切掉。

本文所谓的 DNA 序列就是 DNA 经过测序而得到的由 4 个字母（A, G, C, T）可重复排列而组

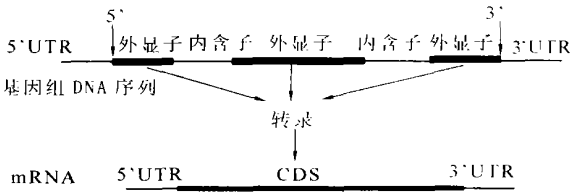


图 1 真核基因的一些关键特征
Fig 1 Some key features of eukaryotic gene

成的一维链。

2 载体研究

在现实中，并不是所有的数据都能用作秘密通信的载体，因为嵌入过程中对载体的修改应该对任何不参与通信的一方是不可视的。这就要求载体必须含有足够多的冗余数据，这些冗余数据可以被秘密信息代替。文本数据含有很少的可用来进行秘密通信的冗余信息。那么对于由 A, C, G, T 组成的一维字符序列 DNA 序列片段来说，它能否充当秘密通信的载体呢？

我们对 DNA 序列进行分析。单拷贝 DNA 序列是在随机突变（碱基的随机突变、缺失和插入，片断的随机重组等）和自然选择的联合作用下形成的。由于强大中性（既非选择有利也非选择有害）突变压力，它们和随机序列靠得很近^[5]。

表 1 DNA 序列的信息参数 D_1
Tab 1 Information parameter D_1 of the DNA sequences

位置	病毒	原核	<i>E. coli</i>	真核 (除 Mam)	哺乳类	免疫球蛋白
5'	0.089	0.049	0.036	0.111	0.033	0.037
3'	0.058	0.056	0.023	0.106	0.058	0.034
内含子				0.124	0.064	0.061
编码区	0.026	0.045	0.017	0.037	0.026	0.027

由表 1^[5] 可看到，一般高等生物序列的熵和随机序列的偏离仅为 1%~2% 的数量级。表中 D_1 描述 DNA 序列中碱基分布相对于随机等概率分布的偏离，*E. coli* 表示大肠杆菌，Mam 表示哺乳类。

$$H = - \sum_i p_i \log_2 p_i,$$
$$D_1 = H_{\max} - H = 2 + \sum_i p_i \log_2 p_i$$

其中， p_i ($i=A, G, C, T$) 是 DNA 序列中碱基 i 出现的概率。

但是，DNA 是一种遗传语言，它包含了丰富的信息内容。因此核苷酸序列中有强的随机噪声，信号和有序结构是夹杂于无序背景之中。

DNA 序列中的随机噪声成分可被用来作为秘密信息的掩饰。事实上, 发送者在嵌入处理之前可对秘密消息进行加密处理, 得到一个外表具有相当随机特性的消息, 并将它嵌入到 DNA 序列中, 替代 DNA 序列的某些随机性。很显然, 将传统密码学与信息隐藏技术结合到一起, 增加了整个通信过程的安全性, 因为攻击者很难监测到嵌入在载体中的密文。即使攻击者试图去提取秘密消息, 也将没有证据表明提取出来的信息是否只是一些随机的比特。这样, 攻击者就无法确定提取出来的信息是有意义的还是自然随机性的一部分。

综上, DNA 序列可作为隐藏信息的载体。

3 嵌入对策

虽然攻破一个信息伪装系统的工作由检测隐藏的信息、提取隐藏的信息和破坏掉隐藏的信息 3 部分组成, 但是如果一个攻击者能够证明一个秘密信息的存在性, 则该系统就已经不安全了。显然, 秘密信息的嵌入对策对信息隐藏算法的安全性是至关重要的。DNA 序列片段是由 A, C, G, T 组成的一维字符序列, 当然嵌入的信息也要转换成 A, C, G, T 组成的一维字符序列, 然后嵌入到 DNA 序列片段。那么对 DNA 序列片段而言, 秘密信息嵌在哪里, 攻击者就不能检测到秘密信息的存在? 下面先讨论 DNA 序列的特征。

3.1 编码区特征

虽然编码区仅占真核基因不到 5%, 但由于编码区密码子使用非随机性, 有偏好性, 许多物种密码子第 3 个碱基倾向使用 G 或 C 而不是 A 或 T, 从而产生编码区规则性^[5-9]:

- (1) 双密码子计数 (即连续密码子对出现频率的计数) 编码区和非编码区有差异。
- (2) 编码区有周期特征。这里的“周期性”指同一核苷酸在相距 3, 6, 9, ... 碱基对位置上多次出现的趋势。
- (3) 密码子第 3 位 G、C 出现频率, 编码区较非编码区有较高的值。

3.2 非编码区的特征

虽然高等生物和人类的基因组中非编码区占绝大部分, 但它没有编码区密码子使用偏好性, 从而没有编码区的上述特征。但它有以下特点^[6]:

- (1) 真核基因中存在大量的长短不一的重复序列, 这些重复序列覆盖了非编码区的部分区域, 它们几乎不会覆盖启动子或外显子编码区。
- 重复序列大体上分为两类: 一是串联重复序列, 二是散在重复序列, 而其中最常见的是可移位

元素 (转座子)。前者包括微卫星序列 (2~5 bp, 重复 100 次左右)、小卫星序列 (15 bp 重复单元组成 500~30 000 bp 的区段) 和卫星序列 (5~10 bp 或更长的重复单元组成很长的序列, 长度可达 10⁵ bp)。卫星序列的重复次数在群体中是不变的 (但在进化上相近的物种间已显示出区别), 小卫星和微卫星则是可变的 (在群体中有个体差异)。可移位重复序列有短散置组分 (SINE) 和长散置组分 (LINE)。SINE 中最著名的是 Alu 序列, 300 bp 的重复单元 100 万拷贝散乱地分布在人类基因组中。LINE 与 SINE 类似, 但重复单元更长, 可长达 7 000 bp。

核苷酸序列包含一些具有某种组成偏好的区域, 例如, 大量 AAAAAAAAAA、TTTTTTTTTTTTTT 或 ATATATATATAT 这样的片段, 它们包含较少信息被视为低复杂度区域; 相反, 杂乱包含 A、C、G、T 四种字母的一串, 一般被视为高复杂度区域。

注意: 由于核酸序列字母数 (4) 较少, 即使在没有组织的随机序列中, 长度短的重复子序列也会由于统计的原因以一定概率出现。例如: 长度为 2 的子序列在长度 16 的核酸序列中平均出现一次, 若核酸序列长度大于 16, 就可能重复出现; 同样, 长度为 3 的子序列在长度 64 的核酸序列中平均出现一次, 若核酸序列长度大于 64, 就可能重复出现; 长度为 4 的子序列在长度 256 的核酸序列中平均出现一次, 若核酸序列长度大于 256, 就可能重复出现; 长度为 5 的子序列在长度 1 024 的核酸序列中平均出现一次, 若核酸序列长度大于 1 024, 就可能重复出现。这种重复不算。

(2) UTR 位于 CDS 两侧的序列, 未被翻译成蛋白质。3' UTR 无论是对基因还是对序列衍生的物种都是高度特异的。

3.3 单核酸的多态性

不同群体和个体在生物学性状以及在对疾病的易感性/抗性上有差别, 这是由于基因组中存在差异。这种差异很多表现为单个碱基上的变异, 也就是单核苷酸的多态性 (single nucleotide polymorphism, SNP)。SNP 在基因组中分布相当广泛, 没有两个人的基因组完全相同。对于不同个体基因簇的遗传学研究结果反映了不同个体之间核苷酸序列的高频突变。

综上可以看到, 修改编码区比较困难; 显然修改重复序列也比较困难, 而且有些重复序列非常保守; 注意到: 3' UTR 对基因是高度特异的, 故 3' - UTR 也不在被选之列; 并且由于准备在嵌入处理之前对秘密消息进行加密处理, 得到一个外表具有相

当的随机特性的消息，去取代 DNA 序列的某些随机性，因此，秘密消息嵌入或替代非编码区的高复杂度区域有很好的安全性，而且 SNP 会进一步提高系统的安全性。如此，嵌入秘密信息的 DNA 序列片段符合生物学意义，不容易引起攻击者的怀疑。

4 结 论

本文研究 DNA 序列成为信息隐藏载体的可能性。对 DNA 序列进行分析的结果表明：核苷酸序列中有强的随机噪声，信号和有序结构是夹杂于无序背景之中。由于 DNA 序列中的随机噪声成分可被用来作为秘密信息的掩饰，因此 DNA 序列可以作为秘密通信的载体。

进而通过对 DNA 序列特征进行分析，提出了一个基于 DNA 序列的嵌入对策：秘密消息替代非编码区的高复杂度区域既有很好的安全性，又符合

生物学意义，而且 SNP 还会进一步提高系统的安全性。

本文的研究对提出以 DNA 序列为载体的信息隐藏算法具有重要的指导作用。

参考文献:

[1] PETITCOLAS F A P, ANDERSON R J, KUHN M G. In - formation hiding —— A survey[J] . Proc IEEE, 1999, 87 (7): 1062— 1078.

[2] BENDER W, GRUHL D, HWANG R, et al. Applications for data hiding[J] . IBM Systems J, 2000, 39(3— 4): 547— 568.

[3] 张新生, 王梓坤. 生命信息遗传中的若干数学问题[J] . 科学通报, 2000, 45(2): 113— 118.

[4] 王哲. 生物信息学概论[M] . 西安: 第四军医大学出版社, 2002.

[5] 罗辽复. 生命进化的物理观[M] . 上海: 上海科学技术出版社, 2000.

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

A Research of DNA Sequences as Covers of Information Hiding

ZHENG Guo qing¹, LIU Jiu fen², HUANG Da ren³, XU An long¹

(1. Department of Biochemics, Sun Yat sen University, Guangzhou 510275, China;

2. Department of Information Research, PLA Information Engineering University, Zhengzhou 450002, China;

3. Department of Scientific Computation and Computer Applications,
Sun Yat sen University, Guangzhou 510275, China)

Abstract: Whether DNA sequences can be covers of information hiding is researched. First, the DNA sequences are analyzed and the feasibility of DNA sequences as covers of information hiding is verified because of the strong stochastic noises in the nucleotide sequences. Secondly, based on the characteristics of DNA sequences, an embedding strategy is proposed by which secret messages embedded in the high complexity regions of non coding regions should have better security. The results are the key issues to design information hiding algorithms based on DNA sequences.

Key words: DNA sequences; information hiding; cover