# 基于 DNA 序列的信息隐藏*

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：研究了以 DNA 序列为载体的信息隐藏，提出了一个信息隐藏算法。该算法首先将秘密信息经预处理、编码、调制变成一维随机 DNA 序列；然后由一个随机数序列决定隐藏信息在载体的高复杂性区域的位置；最后在隐藏信息的位置，秘密信息字符替代载体字符。由该算法实现的信息隐藏既具有一定的稳健性和安全性，又符合生物学意义。实验结果也证明了该算法的有效性。

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）02-0005-05

人类基因组计划（human genome project，HGP）是人类历史上的第 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。信息安全是当前国内外倍受关注的焦点研究课题。信息隐藏[1,2]作为信息安全领域的一种新技术，随着网络技术的发展受到了人们广泛的重视和研究。在这些背景下，本文研究以 DNA 序列为载体的信息隐藏。这种信息隐藏不仅可以传递秘密信息，还可以保护医学、分子生物学、遗传学等领域的知识产权。该问题的研究目前尚属空白。

随着 HGP 的实施，基因序列数据出现了爆炸式增长。但是，并不是所有的数据都能用作秘密通信的载体。对 DNA 序列及其特征进行分析的结果表明[3]：DNA 序列可以作为秘密通信的载体；秘密消息替代 DNA 序列非编码区的高复杂度区域既有很好的安全性，又符合生物意义。本文在文 [3] 的基础上，提出了一种以 DNA 序列为载体的信息隐藏算法。由该算法实现的信息隐藏具有一定的稳健性和安全性。

## 1 信息隐藏算法

提出一种以 DNA 序列为载体的信息隐藏算法。该算法首先将秘密信息经预处理、编码、调制变成一维随机 DNA 序列；然后由一个随机数序列决定隐藏信息在载体的高复杂性区域的位置；最后在隐藏信息的位置，秘密信息字符替代载体字符。

### 1.1 嵌入信息的预处理

在数据嵌入之前，我们先记录待嵌入数据的长度，并将它也隐藏于载体中，用于控制嵌入和检测过程。

DNA 序列片段是由 A，C，G，T 组成的一维字符序列，嵌入的信息显然也要转换成 A，C，G，T 组成的一维字符序列。嵌入数据的长度同要传送的秘密信息一起经置换（表 1）变成由 A，C，G，T 组成的一维字符序列。

用 A，C，G，T 可重复排列的三联体或四联体等编码一个字符。利用英语字母分布的不均匀性（参见文 [4] 附录中英文字母的分布频率），尽量使 A，C，G，T 分布均匀，并尽量使密文同明文分离，构造置换表如表 1。表 1 用 A，C，G，T 的三联体编码一个字符。例如，字符 A 由三联体 GGT 表示。

### 1.2 编码

用 0(00)、1(01)、2(10)和 3(11)4 个数码对 4 种碱基进行二进制数字编码，共有 4！＝24 种编码组合。由于二进制数字 0 与 1 互补，从而 1(01)与 2(11)、0(00)与 3(11)互补。而 4 个碱基中 C 与 G 互补、T 与 A 互补，故在 24 种数字编码中，有 8 种编码—0123↔ACGT，AGCT，CATG，CTAG，GATC，

* 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：zgzx@Public2.zz.ha.cn

GTAC，TCGA，TGCA，满足碱基互补法则。

表 1 置换表

Tab. 1 Replacement table

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| AAA | | AAG | J | AGA | P | AGG | T |
| AAT | Z | AAC | G | AGT | N | AGC | L |
| GAA | B | GAG | K | GGA | Q | GGG | |
| GAT | O | GAC | D | GGT | A | GGC | V |
| TAA | C | TAG | I | TGA | R | TGG | W |
| TAT | G | TAC | E | TGT | U | TGC | E |
| CAA | F | CAG | T | CGA | T | CGG | X |
| CAT | E | CAC | M | CGT | A | CGC | Y |
| ATA | O | ATG | [ ] | ACA | [ ] ] | ACG | 3 |
| ATT | | ATC | [ , ] | ACT | S | ACC | 4 |
| GTA | S | GTG | C | GCA | N | GCG | 5 |
| GTT | E | GTC | [ . ] | GCT | R | GCC | 6 |
| TTA | | TTG | [ ! ] | TCA | N | TCG | 7 |
| TTT | | TTC | [ [ ] | TCT | 0 | TCC | 8 |
| CTA | H | CTG | [ : ] | CCA | 1 | CCG | 9 |
| CTT | A | CTC | [ — ] | CCT | 2 | CCC | |

1.3 嵌入信息的生成

方法 1：秘密信息经过预处理后，再经过编码变成二进制数。然后由发送者选择密钥，作为随机序列发生器的初始状态，产生一个二进制序列。同秘密信息的二进制位流进行位异或运算，最后用上述满足碱基互补法则的 8 种的一种把二进制序列转化成由 A，C，G，T 组成的一维随机字符序列。

方法 2：如果秘密信息比较长，可以先对其进行无失真压缩，变成二进制位流。然后再由用户选择密钥，产生一个随机二进制序列，同秘密信息的二进制位流进行位异或运算。其它同上。

值得一提的是，对秘密信息进行无失真压缩，对提高隐藏系统的安全性很有好处。

1.4 载体的来源

基因组信息主要来源于核苷酸序列数据库。国际上最重要的公共核苷酸序列数据库有：GenBank（美国，http://www.ncbi.nlm.nih.gov/）、EMBL（欧洲，http://www.embl-heidelberg.de/）、DDBJ（日本，http://www.ddbj.nig.ac.jp/），其中 GenBank 由美国国家生物技术信息中心 NCBI、EMBL 由欧洲生物信息学研究所 EBI、DDBJ（DNA Data Bank of Japan）由信息生物学中心 CIB 维护和发布。这 3 个数据库已建立交换协议，每日同时更新核酸序列资料。对用户而言，在任意一个数据库中查询数据基本上是等价的。根据嵌入数据量的多少，在公共数据库中下载基因组序列，分别从中适当截取一段或多段非编码区序列（3′-UTR untranslated region 不在被选之列），作为载体。

1.5 信息的嵌入

（1）将秘密信息长度和秘密信息按 1.1－1.3 变成 A，C，G，T 组成的一维随机字符序列。

（2）选择载体，屏蔽载体的低复杂性区域[5]，剩下载体的高复杂性区域。

（3）使用伪随机数发生器扩展秘密信息[6]。由发送者选择密钥 $S$，作为随机数发生器的初始状态，产生一个随机数序列 $k_1, k_2, \cdots, k_m$（$m$ 为嵌入一维字符序列的长度）。隐藏信息在载体的高复杂性区域的位置为：$j_1 = k_1, j_i = j_{i-1} + k_i$。从而伪随机地决定了 2 个嵌入位置的距离。

（4）在隐藏信息的位置，秘密信息字符替代载体字符。需要说明的是，若嵌入信息位置离屏蔽的区域太近，以致于该位置被秘密字符替代后，当提取信息屏蔽载体的低复杂性区域时该位置也被屏蔽，那么该位置载体字符不被秘密字符替换。因此随机数序列 $k_1, k_2, \cdots, k_m$ 长度大于嵌入一维字符序列的长度。嵌入了秘密信息的 DNA 序列即为隐藏对象。

1.6 信息的提取

（1）选择待检测的 DNA 序列。

（2）屏蔽载体的低复杂性区域，剩下载体的高复杂性区域。

（3）检测嵌入的信息长度。接收者根据剩下载体的高复杂性区域的长度，判断信息隐藏的可能最大长度。接收者根据该值的位数 $L$、已知的密钥 $S$ 和随机发生器的信息，重构长度为 $3L$ 的 $k_i$，进一步获得嵌入的信息长度元素的索引序列，在待检测的 DNA 序列的相应位置抽出字符，经 1.1－1.3 的逆过程，得到嵌入信息的长度值。否则，该 DNA 序列没有隐藏信息，中断检测。

（4）由密钥 $S$ 和嵌入的信息长度类似上步获得整个嵌入位置的索引序列 $j_i$。在待检测的 DNA 序列的相应位置抽出字符，经 1.1－1.3 的逆过程，得到嵌入的秘密信息。同样地，若嵌入信息位置离屏蔽的区域太近，那么该位置字符不提取。

## 2 算法的稳健性和安全性分析

对一个信息隐藏系统来说，攻击者的主要目的是能够证明秘密信息的存在。

由嵌入对策的分析可知[2]，本文算法实现的信息隐藏的 DNA 序列片段符合生物学意义，不容易引起攻击者的怀疑。而且由于使用了伪随机发生器以相当随机的方式扩展了秘密信息，因此 DNA 序列的修改的部分和没有修改的部分具有大致相同的

统计特性，从而其它人不能找到信息存在的统计证据。

攻击者可能会试图寻找用来做秘密通信的载体序列。攻击者惟一可做的就是到数据库搜索。重复序列常常会搅乱其分析。因为如果两个序列的重复片段对齐，可能给出得分很高的比较结果。如果攻击者在数据库搜索之前屏蔽掉重复序列，由于我们仅在随机位置修改了高复杂度区域，导致有比较好的比对结果。但一个 DNA 经自动测序得到的序列数据会出错[7]。原因可能是多方面的，如克隆（cloning）出错，若使用了不恰当的引物，或在多聚酶链式反应（polymerase chain reaction，PCR）中使用了低效率的酶；或碱基识别程序在测序电泳胶板上找出 4 个测序反应的荧光吸收峰时出错。为了尽可能找到所有吸收峰，提高测序准确性，程序以一定间隔监测碱基的吸收峰。但如果胶板的物理性质或其它条件发生变化，影响光线通透性，碱基吸收峰可能变得不规则。尽管检测软件有一定容错能力，还是会出现一定偏差，有些本来没有的碱基被读出，而应该读出的碱基却不能读出，有些不能判断该峰值代表哪个碱基，结果就表现为错误的插入、缺失或出现模糊基；或 DNA 序列拼接时出错，序列拼接时是一项技术程度很高的工作。因此攻击者可能会认为，比对不上的碱基可能是由测序错误或SNP（single nocleotide polymorphism）造成。[3] 即使攻击者试图去提取秘密消息，他提取出来的信息只是一些随机的字符，他没有证据表明提取出来的信息是有意义的。

对一个主动攻击者来说，他在试图不对隐藏对象做大的改动的前提下，从隐藏对象删除被嵌入的信息。但攻击者在没有检测到秘密信息存在的情况下，一般不会对 DNA 序列进行常规的信号处理和几何攻击。另外，本文算法将秘密信息直接编码到 DNA 序列内容中去，而不是将信息直接编码到文本格式中去（比如，调整字符间距或行间距），从而可对抗重新调整文本格式的攻击；本文算法也不是将信息直接编码到文本信息存储的格式（如 GenBank 格式、FASTA 格式等）中，从而可对抗通过一种格式转化到另一种格式对嵌入信息的损坏。因此本文算法比较稳健。

如果一个攻击者能够伪造消息或者以一个通信方的名义启动和运行信息隐藏的协议，则该攻击者是一个恶意的攻击者。本文算法利用能惟一标识发送者的密钥来隐藏和加密信息，只有正确密钥的拥有者才能检测、提取隐藏的信息，因此本文算法可对抗恶意的攻击者。

## 3　实　验

为证明本文算法的有效性，进行了实验。为简单起见，嵌入的秘密信息很短。秘密信息为：JUNE6 INVASION：NORMANDY，长度为 23。

那么，实际嵌入的信息长度为 3×23＝69。要嵌入的信息长度和秘密信息经预处理变为：GCC CCG AAG TGT AGT CAT GCC ATG TAG GCA GGC GGT GTA TAG GAT TCA CTG TCA ATA T GA CAC GGT AGT GAC CGC。然后，经由 0123/CTAG 编码成二进制位流。输入密钥，经 1.3 变为：TGCTAGGG CGGTTAGCAAGTAATGGAAAGCTGCTGGCTACTTGCTC CAGTTATGGCTAACCTCGGCACCCGGGGT。

载体来自人的第 6 号染色体的 MHC（主要组织相容性复合体，major histocompatibility complex，MHC）基因家族中的非编码区，在该基因家族中的位置为：3 968 162—3 978 162，长度为 10 000 个字符。记为：Liujf10。MHC 整条序列从 http://www.sanger.ac.uk/HGP/Chr6/MHC.shtml 下载，从 HLA－F to KNSL2（HSET），全长 4 647 455 bp。由 Repeat Masker[5] 屏蔽载体的低复杂性区域，剩下载体的高复杂性区域。屏蔽结果（见表 2）。

发送者选择密钥 $S(111)$，产生一个随机数序列 $k_1, k_2, \cdots, k_m$。隐藏信息在载体的高复杂性区域的位置为：

$A[85]$ ＝ {11, 100, 155, 168, 240, 327, 393, 491, 564, 596, 628, 636, 703, 748, 841, 891, 933, 994, 1052, 1057, 1091, 1114, 1169, 1235, 1263, 1285, 1372, 1437, 1454, 1478, 1537, 1562, 1603, 1675, 1731, 1804, 1809, 1858, 1927, 1965, 1977, 1996, 2022, 2038, 2061, 2107, 2163, 2165, 2184, 2265, 2281, 2352, 2417, 2513, 2588, 2594, 2663, 2758, 2788, 2865, 2947, 3032, 3041, 3082, 3097, 3105, 3129, 3164, 3239, 3290, 3334, 3366, 3395, 3424, 3515, 3588 , 3633, 3722, 3785, 3877, 3929, 3962, 3990}

在给出屏蔽结果的载体中，上述嵌入信息位置只有 1052 紧邻屏蔽的区域，该位置载体字符不被秘密字符替换。在其它隐藏信息的位置，秘密信息字符替代载体字符得到隐藏对象。

提取信息时，首先选择待检测的 DNA 序列，然后屏蔽载体的低复杂性区域，剩下载体的高复杂性区域。对 Liujf10 来说，剩下载体的高复杂性区域的长度为 4071，因此隐藏信息的长度不超过 4 位数，从而接收者根据该值、已知的密钥 $S$ 和随机发生器的信息，重构长度为 12 的 $k_i$，进一步获得嵌入的信息长度元素的索引序列，在待检测的 DNA 序列的相应位置抽出字符，经 1.1—1.3 的逆

表 2 屏蔽结果

Tab. 2 The filtrated results

| SW score | query sequence | position in query begin | position in query end | matching repeat | repeat class/family | position in repeat begin | position in repeat end |
|---|---|---|---|---|---|---|---|
| 2320 | liujf10 | 1552 | 1953 | L1MA7 | LINE/L1 | 5424 | 5832 |
| 941 | liujf10 | 1955 | 2269 | MER4C | LTR/ERV1 | 695 | 276 |
| 219 | liujf10 | 2343 | 2385 | L2 | LINE/L2 | 2611 | 2653 |
| 400 | liujf10 | 2415 | 2551 | MIR | SINE/MIR | 104 | 247 |
| 22 | liujf10 | 2552 | 2573 | AT_ rich | Low_ complexity | 1 | 22 |
| 1567 | liujf10 | 2574 | 2855 | L1PB3 | LINE/L1 | 6150 | 5882 |
| 1610 | liujf10 | 2856 | 3203 | L1M4 | LINE/L1 | 5166 | 4817 |
| 728 | liujf10 | 3162 | 3333 | L1MD2 | LINE/L1 | 6183 | 6358 |
| 1212 | liujf10 | 3496 | 3799 | AluJo | SINE/Alu | 314 | 2 |
| 723 | liujf10 | 3817 | 4069 | L1MCa | LINE/L1 | 1193 | 932 |
| 1640 | liujf10 | 4072 | 4405 | MER70B | LTR/ERVL | 578 | 271 |
| 3356 | liujf10 | 4537 | 4825 | PRIMA4_ LTR | LTR/ERV1 | 594 | 313 |
| 2297 | liujf10 | 4826 | 5140 | AluSx | SINE/Alu | 1 | 308 |
| 3356 | liujf10 | 5141 | 5431 | PRIMA4_ LTR | LTR/ERV1 | 313 | 1 |
| 335 | liujf10 | 6045 | 6206 | MLT1J2 | LTR/MaLR | 289 | 450 |
| 21 | liujf10 | 6890 | 6910 | AT_ rich | Low_ complexity | 1 | 21 |
| 1641 | liujf10 | 7459 | 7825 | MER44D | DNA/MER2_ type | 609 | 1 |
| 2006 | liujf10 | 8524 | 8822 | LTR43 | LTR/ERV1 | 302 | 602 |
| 27 | liujf10 | 8823 | 8863 | AT_ rich | Low_ complexity | 1 | 41 |
| 1950 | liujf10 | 9042 | 9498 | MLT1E3 | LTR/MaLR | 624 | 139 |
| 2825 | liujf10 | 9555 | 10000 | MER9 | LTR/ERVK | 1 | 442 |

过程，得到嵌入信息的长度值 69。其它按 1.6 的第 4 步可以准确从隐藏对象中提取秘密信息。

另外，又选择了其它 7 个载体进行实验。其中 3 个长度都为 10 000 个字符的 MHC 中的非编码区的序列，位置分别为：2952143－2962143；89720－99720；1166077－1176077。另 4 个为酵母基因组中第 10 号染色体的非编码区，从 http：/NCBI/genebank/下载，编号分别为：NOT：A_13744－21526；NOT：A_196174－201455；NOT：A_209764－217143；NOT：E_7553－13720。位置分别为：13744－21526；196174－201455；209764－217143；7553－13720。按 1.6 都可以准确从它们的伪装对象中提取秘密信息。实验结果证实了本文算法的有效性。

## 4 结 语

本文研究以 DNA 序列为载体的信息隐藏。在文 [3] 的基础上，提出了一种基于 DNA 序列的信息隐藏算法。该算法具有一定的稳健性和安全性。实验结果也证明了该算法的有效性。

研究在 DNA 分子中隐藏认证信息是我们下一步的工作。如今各种各样的假冒伪劣商品随处可见，给国家、企业、个人带来了重大损失。如何保护企业、消费者的合法利益，这是保障市场经济正常发展的重要条件。在 DNA 分子中隐藏认证信息，不仅可以用来保护医学、分子生物学、遗传学等领域的知识产权，还可以认定非数字产品的所有权。因此该研究将具有重要意义。

参考文献：

[1] PEITCOLAS F A P, ANDERSON R J, KUHN M G. Information 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, 2003, 39(3－4): 547－568

[3] 郑国清，刘九芬，黄达人，等. DNA 序列作为信息隐藏载体的研究[J]，中山大学学报（自然科学版），2005，44（1）：13－16.

[4] THOMAS F S. Basic cryptanalysis. Headquarters Department of the Army, Field Manual NO 34－40－2. ftp://ftpp.ox.ac.uk/cryptanalysis/basjc－cryptanalysis.ps.tar.gz. 2000.

[5] SMIT A F A. The repeat masker server. http://repeatmasker.genome.washington.edu/cgi－bin/RepeatMasker, 1996.

[6] MOLLER S, PFITZMANN A, STIRAND I. Computer based steganography: how it works and way therefore any restrictions on cryptography are nonsense, at best[C]. Information Hiding: First International Workshop, Proceedings, 1996, 1174: 7－21.

[7] BAXEVANIS A D, OUELLETTE B F F. 生物信息学基因和蛋白质分析的实用指南[M]. 李衍达，等译. 北京：清华大学出版社，2000.

（下转第 13 页）

1967, 73: 360—363.
[6] BAUM L E, PETRIE T. Statistical inference for probabilitic functions of finite state Markov chains[J]. Annmath Stat, 1996, 37: 1554—1563.
[7] BAKER J K. The dragon system —— an overview[J]. IEEE Trans Acount Speech Signal Prpcessing, 1975, 23(1): 24—29.
[8] JELINEK F. A fast sequential decoding algorithm using a stack[J]. IBM J Res Develop, 1969, 13: 675—685.
[9] KROGH A, BROWN M, MIAN I S, et al. Hidden Markov models in computational biology: applications to protein modeling[J]. Journal of Molecular Biology, 1994, 235: 1501—1531.
[10] Washington University in St Louis. The pfam database of protein families and HMMs [EB/OL]. http://pfam.wustl.edu/2004—04—12.
[11] ALEXEY G, MURZI N. Structural classification of proteins [EB/OL]. http://scop.mrc-lmb.cam.ac.uk/scop/2004—04—12.

## Multiple Alignment Analysis Based on Hidden Markov Models

*LUO Ze-ju*[1], *ZHU Si-ming*[1], *HE Miao*[2]
(1. Department of Mathematics, Sun Yat-sen University, Guangzhou 510275, China;
2. School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China)

**Abstract**: Hidden Markov models (HMMs) are statistical models of sequential data that have been used successfully in many machine learning applications, especially in protein family recognition in the last few years. This is due to the vast increase of data in biology that makes two fields (computational science and biology) associate with each other. To study multiple sequence alignment and profile HMMs, one essential algorithm of HMMs and how to build an HMMs are discussed. By $E$ values and training scores an experiment for recognition and classification of protein family is given.
**Key words**: hidden Markov models (HMMs); multiple sequence alignment; protein classification

---

(上接第 8 页)

## Information Hiding Based on DNA Sequences

*ZHENG Guo-qing*[1], *LIU Jiu-fen*[2], *HUANG Da-ren*[3], *XU An-long*[1]
(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 Computing and Computer Applications, Sun Yat-sen University, Guangzhou 510275, China)

**Abstract**: Information hiding in DNA sequence is discussed, and an information hiding algorithm is proposed. In this algorithm secret messages are transformed into stochastic DNA sequences by beforehand managing, coding and modulating, and then the locations of secret messages embedded in the high complexity regions are decided by a stochastic sequence. Finally, the covered characters at embedded locations are translated from the location of secret message along with hidden information. This information hiding with the algorithm is demonstrated by robust, secure and of biologically significant. The efficiency of the algorithm is demonstrated by experiment results also.
**Key words**: DNA sequences; information hiding; cover