

文章编号: 0529-6579 (1999) 01-0025-28

一种从基因数据库中识别反义 snoRNA 基因的新方法^{*}

屈良鹄, 周 惠

(中山大学生物工程研究中心, 广州 510275)

摘 要: 报道一种从国际分子生物学数据库中识别反义 snoRNA 基因的计算机新方法. 介绍新方法的生物学原理及在反义 SnoRNA 新基因发现中的应用.

关键词: 基因识别, snoRNA, 基因库, 计算机方法

分类号: Q 754, TP 39 **文献标识码:** A

基因的计算机识别和鉴定是生物信息学一个重要领域. 随着人类基因组计划的进展, 国际分子生物学数据库中的 DNA 序列与日俱增, 大批新基因源源不断地被报道. 目前, 大部分新基因的发现首先都是通过计算机“Gene finder”等 DNA 分析程序来进行, 这些 DNA 分析方法对于鉴定蛋白质基因十分有效. 但是, 在 Genbank 中, 有可能存在大量的编码 RNA 的基因序列, 如各种反义 RNA, snoRNA 以及许多尚未被发现的调控 RNA 分子的基因. 现有计算机的基因发现程序并不能有效地识别这些基因. 例如在已全部测定的酵母基因组中^[1], 可能编码蛋白质的 6 275 个 ORF 都被全部确定, 而可能编码 RNA 的基因 (除 rRNA, tRNA, snRNA 等典型的 RNA 基因外) 却发现很少. 事实上, 根据实验证据和理论推测, 细胞中应存在着数目巨大的各种 RNA 基因. 对 RNA 类基因的计算机识别和鉴定, 已成基因分子生物学和生物信息学的重要任务.

本文报道了一种根据反义 snoRNA 基因结构与功能关系以及生物物种 rRNA 结构进化规律而设计的一种新的计算机方法及其在 snoRNA 新基因发现中的应用.

1 原 理

新方法原理主要基于① 已知的反义 snoRNA 的序列特征; ② 已知的反义 snoRNA 结构与功能的关系; ③ 不同生物 rRNA 的结构进化规律及甲基化位点的保守性. 现已发现的 snoRNA 基因可以分成 3 种类型: H/ACA box snoRNA, boxC/D snoRNA (其中大部分为反义 snoRNA) 和 MRP RNA^[2]. 虽然, 不同种类的反义 snoRNA 在一级结构上几乎完全不同 (除保守的 boxC 和 boxD 外), 但是, 它们都包含 1 段与 18S 或 28S rRNA 互补的特征序列, 该序

^{*} 基金项目: 国家自然科学基金重点项目 (39730300)

收稿日期: 1998-03-20 第一作者简介: 屈良鹄, 男, 45 岁, 教授

列长度可达 10~21 个核苷酸. 该特征序列也是反义 snoRNA 的功能元素, 它与位于其 3' 端的 4 个核苷酸 CUGA (boxD') 共同指导 rRNA 互补序列中 1 个特定核苷酸的核糖甲基化 (图 1)^[3].

从图 1 可以看到: 1 个反义 snoRNA 的特征序列与它的功能位点是严格相关的. 虽然不同的反义 snoRNA 特征序列的长度不一定相同, 但是, 任何 1 个反义 snoRNA 指导 1 个特定 rRNA 甲基化位点的功能序列都可以用 $N_{11-21} \cdots N_5N_4N_3N_2X_1CUGA$ 来表示, 其中, N_5 与 rRNA 中甲基化核苷酸配对, X 为不能确定的未知核苷酸. 反过来, 也可通过 1 个确定的 rRNA 甲基化位点推出与之对应的反义 snoRNA 的功能序列, 即反义 snoRNA 的特征序列及其尾随的 CUGA 盒.

基于这一原理, 就可以按图索骥, 从 rRNA 中已经确定的 2'-O-核糖甲基化位点出发, 推出其相关的 snoRNA 功能序列. 由于该序列最短长度已达 16 个核苷酸 ($N_{11-21} + X + CUGA$), 因此可用于对国际数据库直接的计算机比较分析, 从中获得包含 snoRNA 功能序列的 DNA 片段, 然后再用计算机察看其是否具有 snoRNA 基因的 boxC/D 和末端配对序列等共有特征, 鉴定出新的 snoRNA 基因序列.

根据 Maden^[4] 报道, 人、爪蟾和单细胞酵母 rRNA 中大部分核苷酸甲基化位点都已被确定. 比较分析指出, 这些化学修饰核苷酸都位于较为保守的 rRNA 核心序列中, 从高等生物到酵母有很高的相似性. 因此, 从一个保守的甲基化位点推出的序列, 可以被用来发现不同生物中同源的 (或同功) 的 snoRNA 基因. 这对于那些基因组计划即将完成, 但 rRNA 甲基化位点尚未被确定的生物 (如果蝇、线虫、拟南芥等) 反义 snoRNA 基因的鉴定是十分必要的.

2 方 法

(1) 按照已确定的人、爪蟾和酵母 rRNA 甲基化位点, 推出相应的反义 snoRNA 最短的功能序列. 例如, 按照人 28S rRNA 中已知的 Cm2328 甲基化核苷酸, 可以推测出其对应的反义 snoRNA 最短的功能序列为 CCACCAAGATCXCTGA.

(2) 通过 Internet 联网进入国际分子生物学数据库, 采用库中提供的 FASTA 程序和现有参数配置, 将上述序列输入 FASTA 程序, 然后对核酸数据库中几个重要分库 (常规库、最新库和 EST 库) 的全部 DNA 序列进行比较, 其结果可以以 E-mail 形式返回.

(3) 以 PCgene 软件包分析从国际数据库获得的数据, 察看 snoRNA 基因功能序列的上下游 100 个核苷酸中是否分别存在 boxC 或 boxD 以及 4 个核苷酸长的末端配对序列. 符合条件的序列, 作为 snoRNA 新基因的候选者, 作进一步的实验验证.

采用上述方法可以直接对欧洲 EMBL 核酸数据库和美国的酵母基因组数据库进行分析, 图 2, 3 为分析的部分结果. 从图 2 可以看到根据人 28S rRNA Am2328 推测的功能序列可以被用来识别脊椎动物 (人, 鸡) 和植物中 1 种新的 snoRNA 基因 (现已被命名为 U24). 值得指出的是, 植物拟南芥中存在 U24 snoRNA 的同源分子, 表明该甲基化位点在人和植物之间是十分保守的. 图 3 显示了对酵母基因组数据库分析的结果, 发现了 1 个与

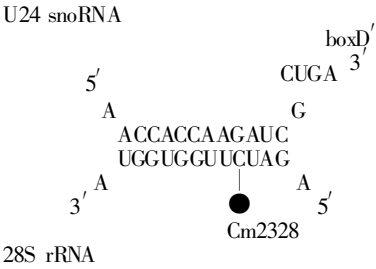


图 1 人 U24 反义 snoRNA 的结构与功能

Fig 1 Structure and function of human U24 antisense snoRNA

EM - PL: ATM3E9 A1022223 Arabidopsis thaliana DNA chromos (79 972 nt)
CM2328 CCACCAAGATCNCCTGA
 boxC
 ::: :
EM - PL: ACTTTGTTCTTTCAGT TGATGA TTGGTTTACCACCAAGATCTCTGAGGCCCTTAGTTTTCT
 76 310 76 320 76 330 76 340 76 350 76 360
EM - PL: ACTGTTTATCTACCAACTCATATTGTGATACTCATTTAAATTAGTTCATCCATAAATCT
 76 370 76 380 76 390 76 400 76 410 76 420

EM - HUM1: HSL7A X52138 Human L7a gene for large ribosoma (5 506 nt)
CM2328 CCACCAAGATCNCCTGA
 boxC
 ::: :
EM - HUM: CTATCTGAGAGATGG TGATGA CATTTTAAACCACCAAGATCGCTGATGACCAACACCCCT
 3 250 3 260 3 270 3 280 3 290 3 300
EM - HUM: TUCTAGTGCCCCAGACATGAACCTTGACATGGAATTTGAGCCTCACTCGGTCTCACCCCTT
 3 310 3 320 3 330 3 340 3 350 3 360

EM - OV: GGRPL7A D14522 Chicken gene for ribosomal protein (4 619 nt)
CM2328 CCACCAAGATCNCCTGA
 boxC
 ::: :
EM - OV: CTATCTGAATGGTAG TGCTGA CATAcataaccaccaagatcgctgactgcactcttgcctc
 1 670 1 680 1 690 1 700 1 710 1 720
EM - OV: TTGCATGCTCTTACTGFGTGGATTGGGAACGGTTTTCAGCTTAAAGTCIGTATTTGTAGGA
 1 730 1 740 1 750 1 760 1 770 1 780

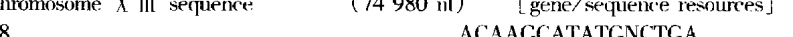
图2 欧洲 EMBL 数据库中 snoRNA 新基因 28—Cm2328 的识别 (boxC 已注明)

Fig 2 Identification of novel snoRNA gene 28- Cm2328 from EMBL DNA data base

酵母 18S rRNA 中第 28 位腺苷酸甲基化相关的 snoRNA 新基因 (现已被命名为 Z4).

CHR13 chromosome X III sequence (74 980 nt) [gene/sequence resources]

18S AM28 ACAAGCATATGCTGA

CHR13 TGCAGAAGATGAACAAATTACTCAAATAG ACAAGCATATGTCTGATCTTCATGATTGAA

298 140 298 150 298 160 298 170 298 180 298 190

CHR13 AAGAGCAATTAGTGTCTGATGCAATTTAACTAGAAAATTTATCTGAGAGAGATGAGAGA

298 200 298 210 298 220 298 230 298 240 298 250

图3 酵母基因组数据库中 snoRNA 新基因 18-Am28 的识别 (boxC/D 序列及末端配对已注明)

Fig. 3 Identification of novel snoRNA gene 18—Am28 from the complete *S. cerevisiae* genome library

3 讨论

过去, 我们建立了计算机分析基因内含子序列的方法, 从国际分子生物学数据库中发现了哺乳动物中十几种新的内含子 snoRNA 基因^[3]. 但是, 该方法需要构建专门的基因内含子数据库 (包括不断对其进行更新); 而且仅局限于对内含子 snoRNA 的分析. 新的方法采用以反义 snoRNA 功能序列直接对国际数据库扫描分析, 大大提高了分析效率和扩大了分析范围, 能够及时地对数据库中最新数据进行跟踪分析, 因此, 具有明显的优越性. 采用上述方法, 我们已获得来自于各种真核生物 100 多种反义 snoRNA 新基因侯选序列, 并在人、酵母和水稻中发现了几种独特的 snoRNA 基因组织^[5~7].

根据反义 snoRNA 结构与功能理论, 真核生物中所有的反义 snoRNA 都可以用上述方法来分析识别。但是, 由于生物进化的多样性, 一些反义 snoRNA 基因可能具有特殊的结构(如存在不标准的特征序列), 从而, 给这些特殊的反义 snoRNA 基因的计算机识别带来困

难. 因此, 进一步研究反义 snoRNA 基因的进化规律, 将能够为设计识别这些特殊的反义 snoRNA 基因的计算机方法提供理论基础.

参考文献:

- [1] GOFFEAU A, BARRELL B G, BUSSEY H, et al. Life with 6 000 genes. *Science*, 1996, 274: 546~567
- [2] MAXWELL E S, FOURNIER M J. The small nucleolar RNAs. *Ann Rev Biochem*, 1995, 35: 897~934
- [3] NICOLOSO M, QU L H, MICHOT B, et al. Intron-encoded antisense small nucleolar RNAs: the characterization of nine novel species points to their direct role as guides for the 2'-O-ribose methylation of rRNAs. *Mol Biol*, 1996, 260: 178~195
- [4] MADEN B E. The numerous modified nucleotides in eukaryotic ribosomal RNA. *Prog Nucl Acids Res Mol Biol*, 1990, 39: 241~303
- [5] 屈良鹄, 钟翎, 施苏华, 等. 水稻 Hsp70 基因第一个内含子编码 2 种核仁小分子 RNA. *自然科学进展*, 1997, 7 (2): 234~241
- [6] 陆勇军, 钟翎, 周惠, 等. Z3 酵母中一种新的反义 SnoRNA. *科学通报*, 1997, 43 (5): 506~510
- [7] ZHOU H, QU L H. EMBL accession number: Aj224022-Aj224031. 1998

A New Method for Identifying Antisense snoRNA Genes in DNA Databanks

QU Lianghu^{*}, ZHOU Hui

Abstract: We have developed a new computer screening method for identifying antisense snoRNA genes in DNA databanks. Based on the relationship between the structure and function of antisense snoRNA, a snoRNA function sequence model, with a minimal length of 16 nucleotide (N_{11~21}+X+CUGA) was established. Most of antisense snoRNA gene function sequences, due to the 2'-O-ribose methylation sites determined in human, *X. laevis* and yeast rRNAs, were derived. The sequences were then used to screen published DNA databases such as EMBL database and the complete *S. cerevisiae* genome library by performing FASTA program through Internet. More than 100 novel snoRNA gene sequences have been discovered by this new process, demonstrating the efficiency and high performance of the new method which provides us a powerful tool in snoRNA gene discovery and function study.

Keywords: gene identification, snoRNA, gene bank, computer method

^{*} Biotechnology Research Center, Zhongshan University, Guangzhou 510275, China