

Yeast 基因下游二级结构与多聚腺苷作用信号

张 静^{1)*} 石秀凡³⁾ 刘次全^{2, 3)}

(¹⁾ 云南大学应用统计中心, 昆明 650091; ²⁾ 云南大学现代生物中心, 昆明 650091;

³⁾ 中国科学院昆明动物研究所, 昆明 650223)

摘要 形成真核生物 mRNA 3' 末端的多聚腺苷 (poly (A)) 作用涉及前体 mRNA 下游的三个元件: 效率元件 (EE)、定位元件 (PE) 以及实际的剪切和 poly (A) 作用位点, 实验研究提出了一些 EE 和 PE 的碱基序列组成. 对 180 个 Yeast 基因下游 (终止密码子后 200 个碱基) 二级结构进行的详细分析显示, 约 86% 的 EE、89% 的 PE 与二级结构中碱基非配对的环 (发夹环、膨胀环、内环或多分支环) 区或连接单链区有关. 这个结果提示, 反式因子对 EE 和 PE 的识别和作用在一定程度上有赖于 EE 和 PE 的二级结构特征. 借助 mRNA 二级结构可以提高对 EE 和 PE 位点预测的准确性.

关键词 Yeast, 多聚腺苷作用信号, mRNA 二级结构

学科分类号 Q617

在真核生物中, 从 DNA 转录得到的前体 (precursor) mRNA, 需要进行一些加工过程才能成为指导蛋白质合成的成熟 mRNA. 前体 mRNA 的加工包括内含子剪接、5' 端加帽、3' 端加尾等过程. 由于 mRNA 3' 末端的 poly (A) 在 mRNA 的调控和翻译中都有很重要的作用^[1], 所以对其形成过程给予了较多的关注. 实验表明, mRNA 3' 末端的形成涉及两个作用: 一个是在前体 mRNA 的某个特定定位点处的核苷内剪切; 再一个是 poly (A) 尾巴接到剪切位点上游的聚合作用^[2]. 剪切和多聚腺苷作用是以三个顺式元件作为作用信号: a. 类似增强子 (enhancer-like) 的元件, 亦称效率元件 (efficiency element), 序列特征一般为 UAUAUA, UAUGUA, UUUAUA 等; b. poly (A) 位点的定位元件 (position element), 序列特征一般为 AAUAAA, AAAAAAAAA, UUAAGAAC, AAGAA 等, 其中 AAUAAA 是最强的信号, 可能就是一致序列; c. 实际的剪切和聚合作用位点本身. 此外, 还有很多复杂的反式因子参与作用, 其中至少有两个剪切因子, 一个剪切激发因子 (在 Yeast 中为 poly (A) 增加因子), 一个专一性 (specificity) 因子, 以及一个 poly (A) 聚合酶. 专一性因子识别多聚腺苷作用的定位信号^[3~5].

Yeast 的基因组序列已完全测定, 这为从理论上系统研究其信息结构提供了方便. Helden 等对 yeast 的 6 217 个基因终止密码子以后的下游序列进行的统计分析, 揭示了一些高度表达的寡聚核苷酸可能就是多聚腺苷作用信号, 其中得分最高的信

号序列是 UAUAUA, 其次是其各种单核苷的变异形式 (UAUGUA, UACAUA, UAAUA 等) 和一个字符的移动形式 (AUUAUA), 这与实验得出的效率元件一致. 此外还获得了另一些高表达的词, 如 U 丰富和 A 丰富的信号, A 丰富的序列即是可能的定位元件^[6].

然而我们知道, RNA 在细胞内一般以折叠的形式存在, 前体 mRNA 也应如此. 我们曾对前体 mRNA 内含子的剪接位点在 mRNA 二级结构中的位置进行分析, 得到了剪接位点的一些结构特征, 并且这些结构特征与实验证实的剪接机制基本吻合^[7,8]. 这说明利用二级结构来研究 RNA 的生物作用机制不失为一个有效的途径. 同样, 探讨多聚腺苷作用信号在 mRNA 折叠结构中的结构特征, 将有助于进一步了解 mRNA 3' 端的形成机制. 本文对部分 yeast 基因下游序列 (终止密码子后) 的二级结构进行预测, 并分析各种多聚腺苷作用信号在这些二级结构中的位置, 结果表明多聚腺苷作用信号在这些二级结构中确实具有明显的结构特征, 并且这些结构特征与多聚腺苷作用的机制相符合.

1 材料和方法

1.1 本文的研究所使用的样本取自 Helden 等进行统计分析使用的样本 (网址: [http://www. ucmb. ulb. ac. be/bioinformatics/ rsatools/](http://www.ucmb. ulb. ac. be/bioinformatics/rsatools/)), 样本的序

* 通讯联系人.

Tel: 0871-5036207, E-mail: jzhang213@263.net

收稿日期: 2000-09-14, 接受日期: 2000-12-12

列是 yeast 基因的终止密码子后的 200 个碱基, 之所以选取终止密码子后的 200 个碱基, 是因为一些实验研究显示, 发生多聚腺苷作用的序列一般位于终止密码子下游的 200 个碱基内, 而且统计分析也显示, yeast 基因组中两个基因之间的区域平均是 326 个碱基, 因此 163 应是基因下游的平均长度^[6]. 为了能更清楚地说明问题, 所选样本需含有实验证实的效率元件序列片段, 一般为六碱基序列片段, 如 UAUUAUA, UUUAUA, UAUGUA 等, 同时在这些序列片段后 (间隔 10 个碱基以上), 还要有 A 丰富的序列片段, 如 AAUAAA, AAGAA 等, 这是可能的定位元件. 在此条件下, 我们从 Helden 的 6 217 个样本序列中随机选取了 180 个.

1.2 在每个样本中找出可能的效率元件 (简记为 EE), 先查找 UAUUAUA, 如果没有此种序列片段, 则找其单碱基变异的形式: UUUAUA, UAUGUA 或 UAUUUA 等序列片段. 此外还要在 EE 位点后找出可能的定位元件 (简记为 PE), 先查找 AAUAAA, 若

无此种序列片段, 则查找 AAAAAAAAA, TTAAGAAC, AAGAA 或 AATAATGA 等富含 A 的序列片段.

1.3 利用 Mathews 等的 RNAstructure3.2 程序对每个样本序列进行二级结构预测, 能量计算用的是 Turner 等于 1998 年提出的能量模型 (温度为 37℃), 二级结构的能量是全局最小的^[9]. 确定 1.2 节中找出的可能的 EE 和 PE 在二级结构中所处的位置.

2 结 果

仔细分析 180 个二级结构后, 可以看出如下结果:

2.1 根据 1.2 节的方法确定的“效率元件” (EE) 中有 86% 涉及环 (发夹环、内环、膨胀环或多分支环) 区或连接单链区, 亦即效率元件的六个碱基中至少有两个或全部位于二级结构的碱基非配对区, 其中比较多的是发夹环. 如图 1~ 图 3 中所示.

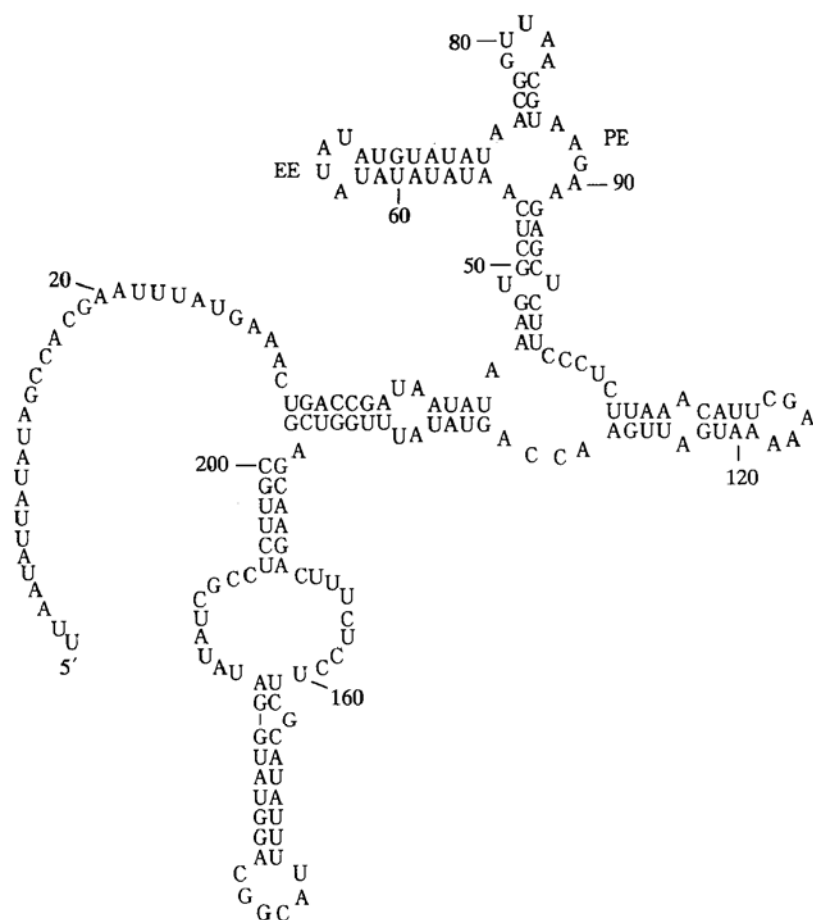


Fig. 1 Secondary structure of YAL051W sequence

YAL051W is the serial number of a downstream sequence of yeast genes in Helden's web site, similarly hereinafter. UUAUAUA (62~ 67) is predicted as the efficiency element (EE) from the sequence, it just forms a hairpin loop. AAGAA (87~ 91) is predicted as the position element (PE), it is located in a branched loop region.

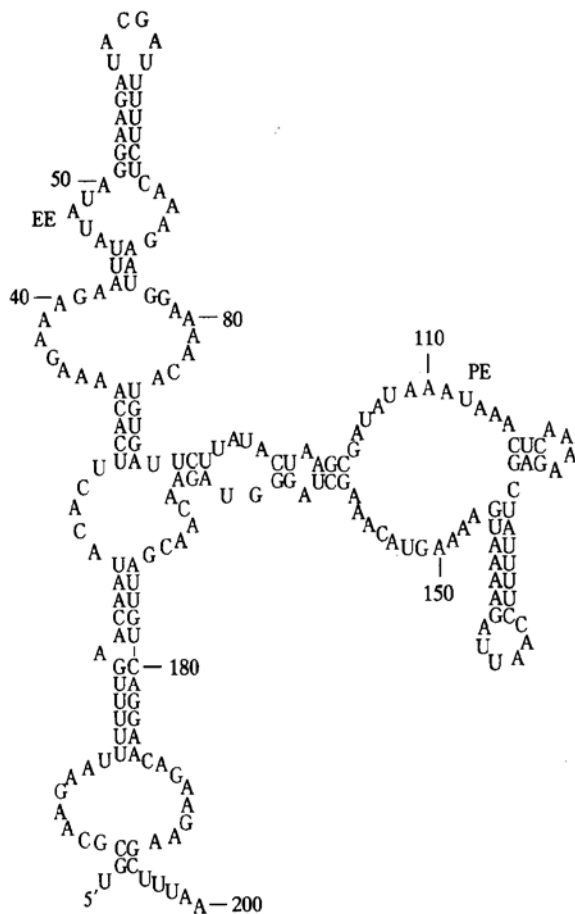


Fig. 2 Secondary structure of YBR098W sequence

UAUAUA (45~ 50) is predicted as the efficiency element (EE) from the sequence, it just forms a hairpin loop. AAUAAA (110~ 115) is predicted as the position element (PE), it is located in a branched loop region. The space distance between EE and PE is short in the secondary structure, although the two elements are separated by 60 bases in sequence.

2.2 根据 1.2 节的方法确定的“定位元件”(PE)中大约有 89% 涉及环区或连接单链区, 其中绝大多数位于多分支环处, 位于发夹环区的较少. 特别地, AAUAAA 多位于多分支环区或连接单链区, 如图 2 中所示.

2.3 通过折叠, 效率元件(EE)和定位元件(PE)的空间距离非常靠近.

3 讨论

已经知道, 许多反式因子参与进行多聚腺苷作用的过程, 这些因子是如何找到其作用位点并完成其功能的呢? 一般认为是顺式元件的碱基序列在起作用, 但是注意到有些样本中这类碱基序列并不是唯一的, 例如图 1 的样本中就有两个连续的 UAUAUA 片段 (此例中的两个 UAUAUA 连续出现, 有些样本中两个 UAUAUA 之间还可能间隔多

个碱基). 因此还不能认为碱基序列即是反式因子识别和作用的唯一信号, 而空间结构特征应该是另一个值得考虑的重要因素. 我们所获得的结果也正说明了这一点. RNA 二级结构中各种环的结构特征比较配对的双链区来说具有碱基突出并有较大的柔性等特点, 其几何形态容易被识别, 从化学性质上看, 非配对的单链区一般不太稳定, 所以易发生作用. 事实上, 对 *E. coli* 的实验已显示, 大部分 mRNA 3' 末端以茎-环结构为特征; 在多聚腺苷作用的过程单链区对于剪切是很有效的^[10].

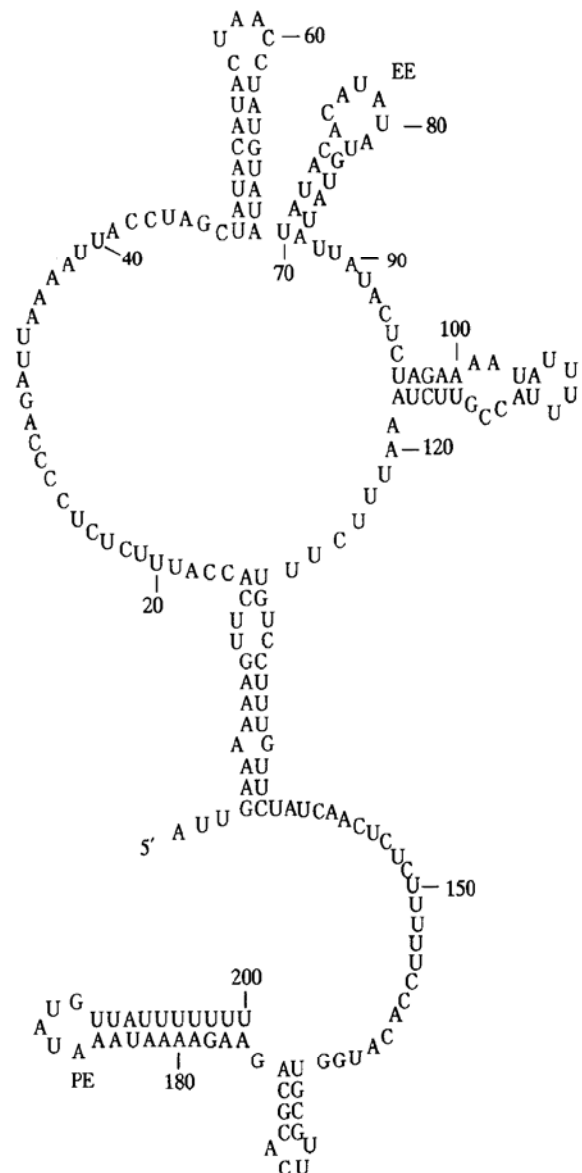


Fig. 3 Secondary structure of YAL018C sequence

UAUAUG (78~ 83) seems to have more possibility to become an efficiency element (EE) than UAUAUA (62~ 67). AAUAAA (181~ 186) is predicted as the position element (PE) from the sequence, but AGAAAA (97~ 102) may be more possible to become PE according to the secondary structure.

在折叠结构中, 效率元件 (EE) 和定位元件 (PE) 的距离在空间上一般都很靠近, 例如图 1 和图 2 中的 EE 和 PE 都很靠近, 特别是图 2 中的情形, 虽然 EE 和 PE 在序列上相距 60 个碱基. 事实上, 一般说来 RNA 折叠后大多是这样的结果, 其实 poly (A) 的实际作用位点在空间距离上与效率元件 (EE) 和定位元件 (PE) 也会很靠近, 只不过其信号序列还没有确定, 所以本文不具体讨论. 各种顺式元件在空间距离上的这种特点有利于各种因子共同作用, 以在尽可能短的时间内完成多聚腺苷作用的过程.

实验表明, 作为效率元件, UAUUA 是最强的信号, 其次 UUUAUA, UAUGUU 和 UAUUUA 也都是比较强的信号^[5], 一般说来, UAUUA 的单碱基变异序列都可能成为效率元件, 所以如果仅从碱基序列上看, 在一个基因的下游序列中, 如果既有 UAUUA 序列片段, 同时又有其碱基发生变异的片段, 就可能会有多个“信号序列”, 究竟哪个成为实际的效率元件呢? 实验表明, 一般是 UAUUA, 只有当 UAUUA 失效时, 其他的序列片段才可能成为效率元件^[2]. UAUUA 何时会失效? 从二级结构上看, 我们认为当它完全位于茎区 (碱基配对区) 时, 有可能失效, 因为茎区比较稳定, 不易发生作用. 例如图 3 中, 66~71 的 UAUUA 位于茎区, 如果它作为效率元件, 作用效率可能较低, 而位于发夹环区的 UAUAUG (78~83) 作为效率元件也许更合理一些. 如果考虑这样的因素, 2.1 节中“效率元件”位于环区或连接单链区的比例将提高到大约 92%. 对于定位元件也有同样的问题, 如图 3, 当我们结合二级结构特征考虑时, “定位元件”位于环区或连接单链区的比例也可提高到 93%.

综合上述, 在 mRNA 3' 末端 poly (A) 形成的过程中, 反式因子对顺式元件的识别和作用很可能

是依靠顺式元件在前体 mRNA 下游二级结构中的结构特征来完成的. 因此在预测信号位点时, 结合 mRNA 的二级结构进行考虑, 有可能提高预测的准确性.

参 考 文 献

- 1 Beelman C, Parker P. Degradation of mRNA in eukaryotes. *Cell*, 1995, **81** (2): 179~183
- 2 Brambilla A, Mainieri D, Agostoni-Carbone M L. A simple signal element mediates transcription termination and mRNA 3' end formation in the DEG1 gene of *Saccharomyces cerevisiae*. *Mol Gen Genet*, 1997, **254** (6): 681~688
- 3 Butler J S, Platt T. RNA processing generates the mature 3' end of yeast CYC1 messenger RNA *in vitro*. *Science*, 1988, **242** (4883): 1270~1274
- 4 Irniger S, Braus G H. Saturation mutagenesis of a polyadenylation signal reveals a hexanucleotide element essential for mRNA 3' end formation in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA*, 1994, **91** (1): 257~261
- 5 Guo Z, Russo P, Yun D-F, *et al.* Redundant 3' end-forming signals for the yeast CYC1 mRNA. *Proc Natl Acad Sci USA*, 1995, **92** (10): 4211~4214
- 6 Helden J, Olmo M, Pérez-Ortín J. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucl Acids Res*, 2000, **28** (4): 1000~1010
- 7 张 静, 刘次全. 内含子二级结构与剪接位点. *生物物理学报*, 1996, **12** (3): 477~481
Zhang J, Liu C Q. *Acta Biophysica Sinica*, 1996, **12** (3): 477~481
- 8 张 静, 刘次全. 真核生物 mRNA 二级结构与内含子剪接. *生物化学与生物物理进展*, 1999, **26** (4): 362~366
Zhang J, Liu C Q. *Prog Biochem Biophys*, 1999, **26** (4): 362~366
- 9 Mathews D H, Andre T C, Kim J, *et al.* In updated recursive algorithm for RNA secondary structure prediction with improved free energy parameters? In: *Molecular modeling of nucleic Acids*. ACS Symposium Series 682, American Chemical Society, Washington D C, 1998. 246~257
- 10 Yehudai Resheff S, Schuster G. Characterization of the *E. coli* poly (A) polymerase: nucleotide specificity, RNA-binding affinities and RNA structure dependence. *Nucl Acids Res*, 2000, **28** (5): 1139~1144

Secondary Structure of Yeast Genomic Downstream Region and Polyadenylation Signals

ZHANG Jing*

(The Center of Applied Statistics, Yunnan University, Kunming 650091, China)

SHI Xiu-Fan

(Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology,
The Chinese Academy of Sciences, Kunming 650223, China)

LIU Ci-Quan

(Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology,
The Chinese Academy of Sciences, Kunming 650223, China;
Modern Biological Center, Yunnan University, Kunming 650091, China)

Abstract Polyadenylation of 3'-forming in eukaryote concerns three elements located in precursor mRNA downstream region: efficiency element (EE), position element (PE) and the actual site for cleavage and polyadenylation. Several base sequences of EE and PE have been proposed by many experiments. The secondary structures of 180 yeast genomic downstream regions (200 bases downstream the stop codon) have been analysed in detail. It is showed that about 86% of EE sites and 89% of PE sites are related to the regions of non-paired loop (hairpin loop, bulge loop, interior loop or multi-branched loop) or to connecting single strand. This result suggests that the identifications and actions of EE and PE by *trans*-factors, to certain extent, are dependent on the structural features of EE and PE in the secondary structure. According to the secondary structure of mRNA, the prediction accuracy of EE and PE sites may be improved.

Key words yeast, polyadenylation signal, mRNA secondary structure

* Corresponding author. Tel: 86-871-5036207, E-mail: jzhang213@263.net

Received: September 14, 2000 Accepted: December 12, 2000