

水稻 mRNA 多聚腺苷化信号位点的序列分析

陆颖^{①②} 高晨曦^{①②③} 韩斌^{①③*}

(^① 中国科学院上海生命科学研究院国家基因研究中心, 上海 200233; ^② 中国科学院研究生院, 北京 100039; ^③ 中国科学院上海生命科学研究院植物生理生态研究所, 上海 200032. * 联系人, E-mail: bhan@ncgr.ac.cn)

摘要 大部分真核生物 mRNA 的加工涉及到 mRNA 前体的分裂/多聚腺苷化, 在 3'-末端形成多聚腺苷酸(poly(A))尾巴. 为了研究水稻 mRNA 前体多聚腺苷化过程所必需的 poly(A)加尾信号、下游调控元件和 poly(A)位点等序列的特点, 用 3'-末端带有 poly(A) 的 EST 与全长 cDNA 序列进行比较, 建立了一个来自 9953 个基因的、覆盖 12969 条 poly(A)位点两侧各 40 碱基序列的数据库. 结果发现, 只有 7.9% 的 mRNA 使用 AAUAAA 作为加尾信号, 超过 60% 使用 AAUAAA 的 1~2 个碱基变化的序列作为加尾信号, 11.5% 的 mRNA 使用 AAUGAA 及其单碱基变化的加尾信号. 在约 25% 的 mRNA 的 3'-末端存在多个 poly(A)位点. 在 90% 的 mRNA 前体的加尾信号下游都能检测到富含 U/GU 的调控元件, 尤其在以 AAUAAA 的多碱基变化序列作为加尾信号的 mRNA 前体中, 半数以上都能在 poly(A)位点两侧检测到下游调控元件. 而且, 这些调控元件的位置对 poly(A)位点的选择有限制作用. 总之, 虽然水稻 mRNA 加尾信号的保守性较低, 但大量下游调控元件的存在保证了多聚腺苷化过程的正常进行.

关键词 水稻 表达序列 RNA 剪接 多聚腺苷化

大部分真核生物 mRNA 前体(pre-mRNA)在剪接之前, 都是先在其 3'-末端的某个位置切开, 再进行多聚腺苷化(polyadenylation), 形成多聚腺苷酸(poly(A))尾巴^[1,2]. 在此过程中, 分裂/多聚腺苷酸特异性因子(cleavage/polyadenylation specificity factor, CPSF)直接识别 poly(A)位点上游的 AAUAAA 序列, 即 poly(A)加尾信号(polyadenylation signal, PAS); 刺激分裂因子(cleavage stimulatory factor, CstF)主要识别下游的富含 GU 或 U 的序列, 即下游调控元件(downstream element, DUE)^[3,4]. 在哺乳动物中, PAS 是存在于 poly(A)位点上游 10~30 nt 处非常保守的 AAUAAA 序列^[5], 而位于 poly(A)位点下游 20~40 nt 处的是富含 U 的 DUE^[6,7]. 间隔一定距离存在的 PAS 和 DUE 是 CPSF 和 CstF 结合并形成稳定的 RNA-CPSF-CstF 复合物所必需的^[4], 所以对 PAS 和 DUE 序列的研究对理解多聚腺苷化过程而言非常重要.

在植物和酵母(*S. cerevisiae*)中, PAS 的序列有较大的变化^[6]. 从最初的 EST 研究中发现, 酵母、水稻(*O. sativa*)和拟南芥(*Arabidopsis*)的 mRNA 使用 AAUAAA 作为 PAS 的比例分别只有 13.2%, 6.2%^[8] 和 10%^[9]. 目前, 对 PAS 序列的认识主要来自对动物的研究, 对植物 PAS 的统计分析因其序列的复杂性而进行得较少. 近年来, 随着水稻基因组测序的完成, EST 数量的增加, 尤其是 3 万多条水稻全长 cDNA 序

列的发表^[10], 使我们有可能利用基因组的精确序列和 cDNA 序列中 3'-非翻译区(3'-UTR) 的序列全面分析水稻 mRNA 的 PAS 和 DUE 以及它们对 poly(A)位点的影响. 由于 PAS, DUE 和多聚腺苷化位点(poly(A)位点)都集中于 poly(A)位点两侧-40~+40 nt 的范围内, 所以这个区域的序列对植物 mRNA 的多聚腺苷化过程有着重要的意义. 鉴于近 50% 植物基因使用 AAUAAA 的多碱基变化序列作为 PAS^[11], 我们将其他物种研究 PAS 的方法^[5,12]与概率统计学方法^[13,14]结合起来, 对水稻 mRNA 的 PAS 进行全面的分析. 同时, 对 PAS 下游调控序列进行了研究, 以助于解释含有不同 PAS 序列的水稻 mRNA 是如何有效进行多聚腺苷化的.

1 材料与方法

(i) poly(A)位点的确定及其上、下游序列数据库的建立. 根据已经报道过的方法^[5,12,15], 对 EST 进行筛选和比较. 首先, 来自 TIGR (the Institute for Genomic Research, 2005 年 2 月)的 206961 条水稻 EST 中, 将 3'-端含有 10 个以上腺嘌呤(A)或者在 5'-端含有 10 个以上胸腺嘧啶(T)的 EST 挑选出来(那些 5'-端含有多个 T 的 EST, 只记录它们的反向互补链), 共 60964 条, 推测它们可能含有(poly(A))尾巴. 那些连续的 A 与上游序列的连接处就被认为是 mRNA 的 poly(A)位点. 然后, 将这些位点 3'-端连续 A 去掉. 接着, 通过 BLAST 分析(e 值 $<1 \times 10^{-30}$)^[16], 将这些 EST

与水稻全基因组序列(来自 RGP, 2005 年 1 月)进行比较, 只有匹配度超过 95% (Identity 0.95)的 EST 才被用于接下来的数据分析. 随后, 根据 EST 与基因组序列的比较结果, 确定 EST 的 3'-末端在基因组上的位置, 将该位置前后各 10 个碱基(nt)的基因组序列取出. 如果这段序列中存在连续 6 个 A, 或在某 10 个连续碱基中有 8 个(或以上)的 A, 那么就认为它是从 mRNA 内部起始(internal priming issue)的 EST, 将其除去. 如果有 2 条以上的 EST/cDNA 支持同一个 poly(A)位点, 那么这个 EST 也被保留.

水稻全长 cDNA 序列来自 KOME (the Knowledge-based Oryza Molecular Biological Encyclopedia, 2005 年 1 月)^[10]. 通过 BLAST 分析, 将挑选出来的 EST 与 cDNA 序列进行比较, 只有符合以下条件的 EST 才被认为与 cDNA 序列有共同的基因来源, EST 的 3'-末端所对应的才是真正的 poly(A)位点: (1) 匹配度超过 95%; (2) EST 与全长 cDNA 序列有不少于连续 50 nt 的匹配; (3) EST 和全基因组序列匹配的 5'-末端与 EST 和 cDNA 序列匹配的 5'-末端相比, 差异不超过 25 nt; (4) EST 和全基因组序列匹配的 3'-末端与 EST 和 cDNA 序列匹配的 3'-末端相比, 差异不超过 5 nt; 如果 EST 和全基因组序列匹配的 3'-末端比 EST 和 cDNA 序列匹配的 3'-末端位于更下游处, 那么该 EST 的 3'-末端也被认为是一个真正的 poly(A)位点.

把可能来自于嵌合的、可变剪切的和未剪切的 RNA 以及末端序列存在较大误差的 EST 去除以后, 共有 32594 条 EST 对应于 9953 条 cDNA 序列, 每条 EST 的 3'-末端都代表一个 poly(A)位点. 根据 poly(A)位点在全长 cDNA 序列中的位置, 将该位点上游 50 nt 的全长 cDNA 的 UTR 序列取出. 如果 EST 的 3'-末端位于 cDNA 序列 3'-末端的下游, 就从基因组中将该位点上游 50 nt 的序列取出. 这些 50 nt 的 UTR 序列组成 3'-上游序列数据库. 类似的, 从 poly(A)位点下游的基因组中取出 50 nt 的序列组成下游序列数据库. 同一 poly(A)位点上游和下游序列相连就是该 mRNA 前体的 poly(A)位点两侧序列. 对于那些对应于同一 cDNA 序列的多条 EST, 如果有 2 个或 2 个以上 poly(A)位点的间距小于 25 nt, 那么只有被最多 EST 支持的那个 poly(A)位点才被记录. 这样, 共有来自 9953 个基因的 12969 条不同 poly(A)位点的上游和下游序列最终被用于其后的分析.

(ii) PAS 的确定和统计. 根据人及其他物种 PAS 的分析方法^[5,12], 结合对特定序列中有过度表现的序列单元(over-represented words)进行预测^[13], 来确定水稻 mRNA 中的 PAS. 具体包括以下两步: (1) 对于 12969 条表征不同 poly(A)位点的上游序列, 在 poly(A)位点上游-40~-1 nt 区域, 搜索最保守的 AAUAAA 信号及其 17 个单碱基变化信号(single nucleotide variants of AAUAAA signal, P1PAS), 如果在这个区域更接近 poly(A)位点的地方存在任何一个上述信号, 那么该 PAS 的位置和序列就被记录下来; (2) 在分析那些不含 AAUAAA 或 P1PAS 的上游序列时, 先将所有 AAUAAA 的 2~3 个碱基变化的六碱基序列(2~3 nucleotide variants of AAUAAA signal, 分别简称为 P2PAS 和 P3PAS)列出, 根据各六碱基序列在基因组随机序列和在 poly(A)位点上游序列内出现的频率计算它们的正态值(standardized score, S 值), 计算公式详见参考文献[13]和[14]. 然后, 记录每个 P2PAS 或 P3PAS 在 12969 条 poly(A)位点上游出现位置的平均值和标准差(SD), 挑选出现位置 SD<9.0 的 P2PAS 和 P3PAS. 最后, 在不含 AAUAAA 或 P1PAS 的序列中, 按挑选出的 P2PAS 和 P3PAS 的 S 值大小, 依次搜索-40~-1 nt 的序列. 当某条序列中含有被检测的 PAS 时, 这条序列将被移出, 不再被其后检测的 PAS 重复统计. 据此, 各个 P2PAS 或 P3PAS 出现的频率和位置都被记录下来.

(iii) poly(A)位点下游调控元件的检测. 如近年来的报道所述, DUE 主要由富含 U 或 GU 的五碱基序列组成, 它包括 URE (U-rich element, UUUUU 及单碱基变化的序列)^[4,17,18]和 GUE (GU-rich element, GUUGU, UGUGU, GUGUU 和这些序列中 1 个碱基发生变化的序列)^[4]. 我们对 12969 条 poly(A)位点下游序列的+1~+40 nt 区域和 poly(A)位点上游至 PAS 之间的区域进行了搜索, 并将 DUE 的序列和位置进行了统计.

2 结果与讨论

2.1 水稻 mRNA 的 poly(A)位点的分析

共有 9953 条全长 cDNA 序列拥有至少 1 个 EST 支持的 poly(A)位点. 其中, 7455 个(74.9%)有一个 poly(A)位点, 2498 个(25.1%)有多个 poly(A)位点(表 1). 据此推算, 每个水稻基因平均有 1.30 个 poly(A)

表 1 带有多个 poly(A)位点的 mRNA 的数量

每个 mRNA 带的 poly(A)位点数量	mRNAs 的数量	百分数(%)
1	7455	74.9
2	2060	20.7
3	358	3.6
≥ 4	80	0.8
合计	9953	100

位点. 通过分析-1 和 0 nt 位置的碱基组成, 发现有 36% 的 mRNA 在 poly(A)位点处的碱基是 UA, 而 CA 的比例是 31%, 与已报道的植物和人类 mRNA 的 poly(A)位点处的碱基特征相符^[11,19].

在 2498 个多 poly(A)位点的基因里共检测到 5514 个 poly(A)位点. 从 poly(A)位点至基因终止密码子的距离平均为 200~400 nt(图 1). 在含有 3 个 poly(A)位点的 mRNA 中, 这 3 个位点的平均距离分别为 212, 302 和 391 nt. 而对于含有 4 个 poly(A)位点的 mRNA 而言, 它们的平均距离分别为 228, 309, 388 和 470 nt. 这意味着, 离终止密码子最近的 poly(A)位点平均位于距离终止密码子 200 nt 的 UTR 区域中, 再往下游每增加 80~90 nt, 就可能出现一个新的 poly(A)位点. 与人类 mRNA 每隔 600 nt 就可能存在一个 poly(A)位点的情况相比, 水稻 mRNA 的 poly(A)位点的位置更为集中, UTR 长度比人类的小得多^[5].

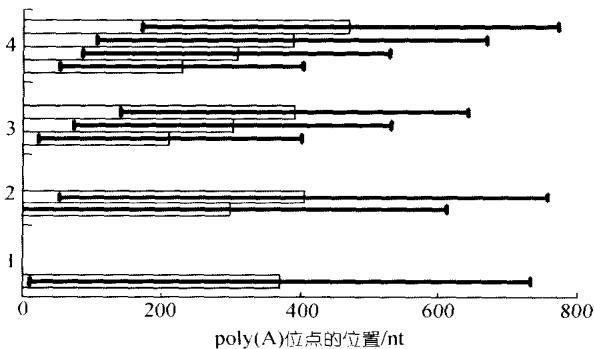


图 1 带有多个 poly(A)位点的 mRNA 中各 poly(A)位点的平均位置

纵坐标的 1~4 分别表示带有 1, 2, 3 和 4 个 poly(A)位点的 mRNA. 在带有 4 个 poly(A)位点的 mRNA 中, 从下往上依次是第 1, 第 2, 第 3 和第 4 个 poly(A)位点, 其他的依次类推. 柱状图显示 poly(A)位点至终止密码子的平均距离和标准差. 横坐标的原点(0)表示终止密码子的位置

2.2 poly(A)位点上游存在的 AAUAAA 和 P1PAS 信号的分析

为了了解水稻中那些保守的 AAUAAA 和 P1PAS

信号的存在情况, 我们搜索了 12969 条 poly(A)位点上游-40~-1 nt 的序列, 结果发现, 虽然 AAUAAA 是最保守的 PAS, 但所占比例只有 7.9%, 而各个 P1PAS 的比例都低于 4%(表 2). 这比以前报道过的水稻中有 6.2% 的 mRNA 以 AAUAAA 作为 PAS 的比例要高^[8], 可能是我们检测了更多的 EST 和 cDNA 序列的缘故. AUUAAA 是人和动物中使用频率仅次于 AAUAAA 的 PAS^[5]. 但在水稻中, AUUAAA 的比例只有 1.9%, 甚至低于 AAUAUA, AAUGAA, AAUAAU, AA-GAAA 等信号. 被 1 个 C 取代的 P1PAS 信号(如 AACAAA 和 CAUAAA)所占的比例明显也低于 AAUAAA 被 A, U 或 G 取代的信号. 就所有 P1PAS 的情况而言, 有 33.5% 的水稻 mRNA 使用 P1PAS 作为它的 PAS, 这个比例大大高于以 AAUAAA 作为 PAS 的 mRNA 数量. 我们的统计结果显示, AAUAAA 和 P1PAS 的最后一个碱基至 poly(A)位点的平均距离大约是 18.7 nt, 标准差都小于 9.0 nt, 表明 PAS 的位置是相对集中的.

表 2 AAUAAA 和 P1PAS 在 poly(A)位点上游出现的频率和位置

PAS	带有 PAS 的 mRNA 数量	频率%	平均位置±SD/nt
AAUAAA	1022	7.9	-19.2 ± 6.7
AAUAUA	480	3.7	-19.1 ± 7.3
AAUGAA	457	3.5	-18.7 ± 7.2
AAUAAU	422	3.3	-18.0 ± 8.2
AAGAAA	320	2.5	-18.9 ± 7.8
UAUAAA	305	2.4	-19.1 ± 7.1
AAUUAA	301	2.3	-18.0 ± 8.9
AAUAAG	287	2.2	-18.3 ± 7.6
AUUAAA	246	1.9	-17.5 ± 9.0
AAUACA	236	1.8	-18.3 ± 7.7
AAUCAA	233	1.8	-18.5 ± 7.5
AGUAAA	194	1.5	-18.2 ± 7.5
AACAAA	179	1.4	-18.0 ± 8.4
GAUAAA	143	1.1	-18.3 ± 7.9
AAUAGA	170	1.3	-20.0 ± 7.2
CAUAAA	132	1.0	-18.4 ± 8.5
AAUAAC	143	1.1	-17.5 ± 8.2
ACUAAA	95	0.7	-16.9 ± 8.7
合计	5365	41.4	-18.7 ± 7.8

2.3 水稻 mRNA 中 P2PAS 和 P3PAS 的预测和分析

在水稻 mRNA 中, 有多达 59% 的 mRNA 在 3'-末端没有 AAUAAA 及 P1PAS 序列. 而从以前的研究

结果来看,有近50%的植物基因可能使用P2PAS或P3PAS^[11].为了检测水稻mRNA中P2PAS和P3PAS的存在,我们选取*S*值>1.96(*P*值<0.05)、*SD*<9.0 nt的90个P2PAS和P3PAS,按其*S*值大小,依次搜索那些不含AAUAAA和P1PAS的上游序列.结果显示,在4885条(37.6%)上游序列中检测到有P2PAS或P3PAS存在,它们的平均位置为-19.3±8.7 nt(平均值±标准差),与AAUAAA和P1PAS的位置十分相似.在这些检测到的P2PAS或P3PAS中,只有10个的频率大于1%.根据这些PAS的碱基组成,以其最接近的P1PAS序列为标准,进行分类(表3).比如,AAUGGA, AAUGAU, AAUGCA和AUUGAA可以看作是AAUGAA中的1个碱基发生变化,归入AAUGAA型的PAS.又如, AUGAAA和AAAUGA等2个序列与AAUGAA有连续5个碱基的重复,故也归入AAUGAA型的PAS.另外, AAUGGU, AUGGAA和UGGAAA与AAAUGGAAA非常相似,且AAAUGGAAA已经被实验证实在植物中具有PAS的功能^[20],在本研究分析的水稻mRNA中有1.3%含有这类PAS.所以将这3个PAS归入AUGGAA型的PAS.

最后,一共得到6类PAS,它们存在于3669条(28.3%)上游序列中,而其他的P2PAS和P3PAS只在1216条(9.3%)序列中出现.如图2所示,AAUGAA类型的PAS在mRNA的3'-末端出现的频率最高,而AAUGAA本身的出现频率在P1PAS中也处于第2位.1993年,在对玉米醇溶蛋白mRNA的研究中,发现AAUGAA序列具有PAS的作用^[21],如果将AAUGAA变成AAUAAA,将有一半mRNA的剪切和加尾将无法进行^[22].在AAUGAA型的P2PAS中,有些PAS(如AAUGGA, AAUGAU, AAUGCA和

AUUGAA)的频率甚至超过了ACUAAA, AAUAAC, CAUAAA和GAUAAA等P1PAS.对这些六碱基序列的研究结果表明,含有AAUG, AUGA和UGAA的序列偏好出现在mRNA的3'-末端,暗示着含有UG序列的AAUGAA型PAS能够被CPSF的有效识别.另外, AAUAAU型、UAUAAA型、AAUAAA型和AUUAAA型的比例也较高.在所有P2PAS和P3PAS中,UAUAUA出现频率最高(1.8%).已有研究证实,3'-末端的UAUAUA具有调控mRNA多聚腺苷化的作用^[23,24].

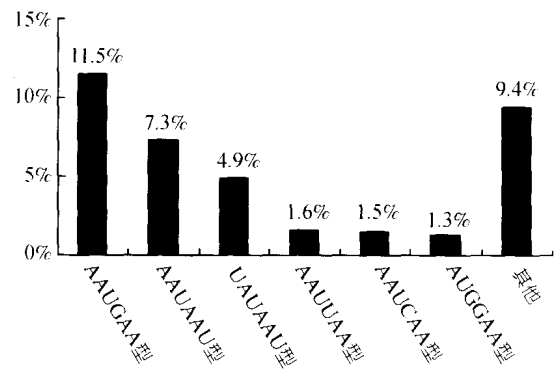


图2 在12,969条poly(A)位点上游序列中6类P2PAS和P3PAS的比例
P2PAS和P3PAS的具体分类见表3

统计结果显示,水稻PAS序列的保守性较低,但平均位置都集中poly(A)位点上游约19 nt处.如果考虑AAUAAA信号和P1PAS,水稻中有80%的mRNA使用AAUAAA或其1~3个碱基变化的序列作为PAS,含有P1PAS和P2PAS的mRNA序列超过60%(图3).大多数mRNA使用AAUAAA或与其相似的序列作为PAS,多聚腺苷化也是依赖于AAUAAA(或与其相似

表3 P2PAS和P3PAS的分类及其在poly(A)位点上游出现的频率

类型	P2PAS和P3PAS	出现频率(%)
AAUAAU	UAUAAU, AUUAAU, AGUAAU, AAUCAU, AAUACU, AAGAAU, CAUAAU, AAUAGU, AAUAUG, AAUCUA, AAUAUC	7.34
AAUCAA	AUCAA, AAUCAG, AUCAAA, AUCAAU	1.55
AAUGAA	AAUGGA, AAUGAU, AAUGCA, AUGAAA, AUUGAA, ACUGAA, UAUGAA, AAUGAG, GAUGAA, AGUGAA, CAUGAA, AAAGAA, AAAUGA	11.54
AUGGAA	AAUGGU, AUGGAA, UGGAAA	1.29
AAUAAA	AGUAAA, AAUUGA, AAGUAA	1.64
UAUAAA	UAUCAA, UAUAAAG, UAUAUA, UAUACA, UAUAUG	4.93
其他	AUGUAA, UGUGAA, AAUGCU, GAUAUG, AUGCAA, AAUGUG, AAAGAU, etc.	9.29

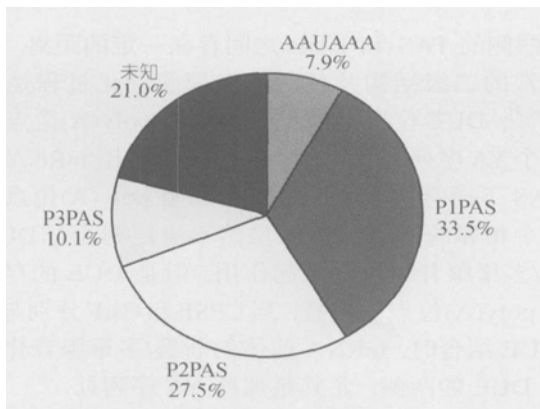


图 3 水稻 mRNA 中分别以 AAUAAA 及其一碱基变化、二碱基变化和三碱基变化的六碱基序列作为 PAS 的比例

的 PAS)与 CSPF 结合的过程。

2.4 poly(A)位点附近的富含 U 或 GU 的调控序列的分析

虽然不是所有 DUE 对 mRNA 前体的多聚腺苷化都有作用^[22], 但鉴于某些 URE 在多聚腺苷化过程的作用已被证实^[17,25], 因而我们对 PAS 至 poly(A)位点之间的区域以及 poly(A)位点下游+1~+40 nt 的区域的 DUE 进行了搜索。结果发现, 在 PAS 至 poly(A)位点之间, 有 63%的 mRNA 前体存在 DUE; 在+1~+40 nt, 则有 90%含有 DUE。其中, 59%在 poly(A)位点两侧都有 DUE 存在, URE 所占的比例比 GUE 高得多。然而, 在含有不同类型 PAS 的 mRNA 前体中, 检测到 DUE

的比例却明显不同。如图 4 所示, 在超过 60%的带有 P2PAS 和 P3PAS 的 mRNA 前体中, DUE 存在于 poly(A)位点的两侧, 这个比例明显高于带有 AAUAAA 和 P1PAS 的 mRNA 前体。在 poly(A)位点下游, 最近的 DUE 的第一个碱基至 poly(A)位点和 PAS 的平均距离分别为 15.5 ± 10.5 nt 和 34.0 ± 19.2 nt。而 DUE 的最佳位置是 poly(A)位点下游 10~30 nt 处^[4,11]。另外, 在对许多 mRNA 前体序列的分析中, 我们经常看到有连续的 DUE 出现, 这些五碱基元件之间彼此相隔不超过 2 个碱基, 并广泛分布于 poly(A)位点的下游区域(图 5)。从人类精子细胞研究得到的

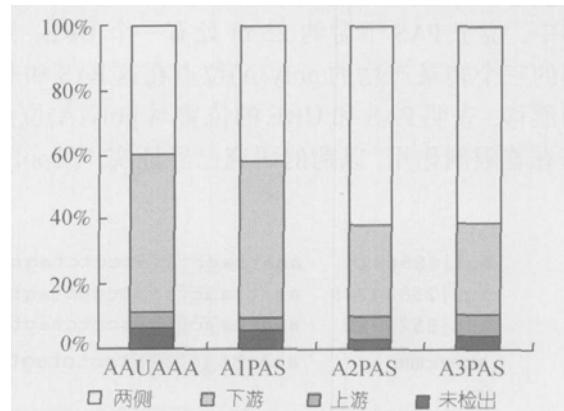


图 4 带有不同类型 PAS 的 mRNA 前体中 DUE 的分布柱状图中从上至下依次是在 poly(A)位点两侧(两侧)、只在 poly(A)位点上游(上游)、只在 poly(A)位点下游(下游)检测出有 DUE, 以及和未检测出 DUE(未检出)的 mRNA 前体的比例。横坐标显示是在 DUE 上游 PAS 的类型

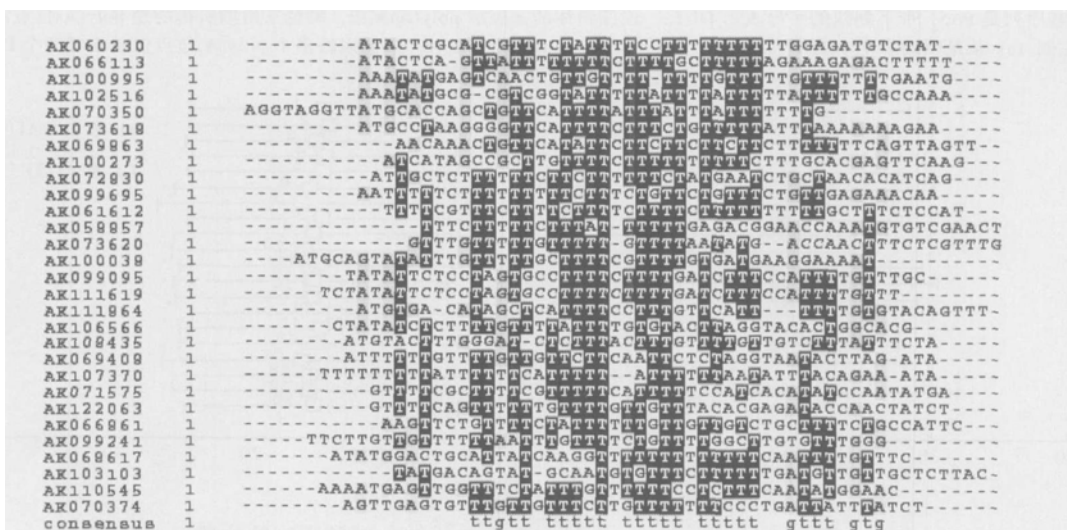


图 5 poly(A)位点下游 DUE 的序列

在多重序列比较中, 黑色字体的碱基是相同的碱基(DNA 序列中为 T), 最下方显示的是比较后得到的一致序列(consensus), 表示相邻 TTTTT 或类似的 DUE 之间相隔 1 个碱基。每条序列对应的上游 cDNA 序列的登录号标在图左侧。位置 1 表示各序列的第一个碱基就是 poly(A)位点下游的第一个碱基

模型说,不同类型的 CstF 与 URE(或 GUE)形成很强的结合时,可以促使 CSPF 与 poly(A)位点上游的非 AAUAAA 序列结合,从而启动多聚腺苷化过程^[3]. 所以我们推测,水稻中非 AAUAAA 的 PAS 的使用可能会影响 mRNA 前体的断裂/多聚腺苷化的效率^[19], 但大量 DUE 的存在,加之 PAS 上游的其他调控序列的存在,使 mRNA 前体的断裂/多聚腺苷化过程依然能够顺利启动.

2.5 PAS 和 DUE 对 poly(A)位点位置的限制作用

不同基因的转录产物可能使用某个 PAS 下游的不同 poly(A)位点. 从图 6(a)可以看出,在一个 mRNA 前体中,位于 PAS 下游约 25 nt 处有一个 URE, 这个基因的三个转录产物的 poly(A)位点在这 PAS 和 URE 之间漂移,表明 PAS 和 URE 的位置对 poly(A)位点的选择有着限制作用. 以前的研究已经证实,在 poly(A)

位点两侧的 PAS 和 DUE 之间存在一定的距离,对形成正常的二级结构及其后的多聚腺苷化过程是有利的^[26~28]. DUE 位置的变化有可能使 poly(A)位点移向另一个 YA 序列^[17,27]. 从图 6(b)可以看出, mRNA 前体的 PAS 下游存在多个 URE, 而 3 个 poly(A)位点都位于两个相邻的 DUE 中间. 虽然不清楚哪几个 DUE 在断裂/多聚腺苷化过程中起作用,但是 DUE 的存在影响了 poly(A)位点的位置. 当 CPSF 和 CstF 分别与 PAS 和 DUE 结合时, mRNA 前体的断裂/多聚腺苷化就发生在 DUE 的两侧,尤其是那些 YA 序列处.

2.6 poly(A)位点两侧-40~+40 nt 序列碱基组成的特点

在 poly(A)位点两侧从-40~+40 nt 的序列中存在着 PAS, DUE 和 poly(A)位点等与多聚腺苷化相关的重要序列. 通过计算,任意六碱基序列在-40~-1 nt,

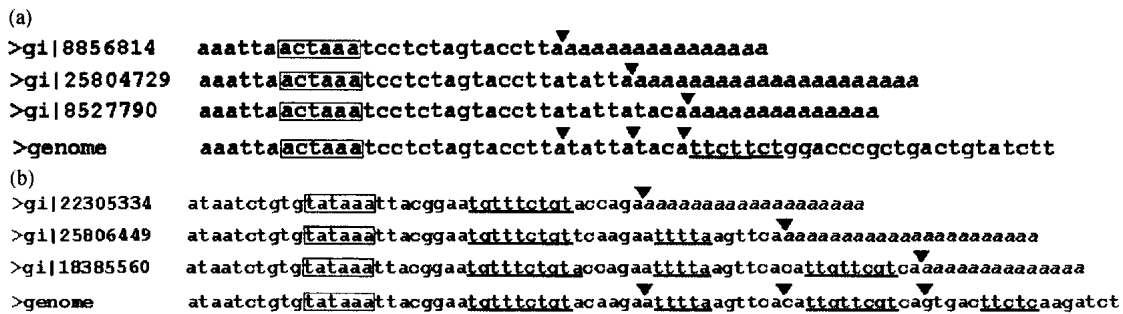
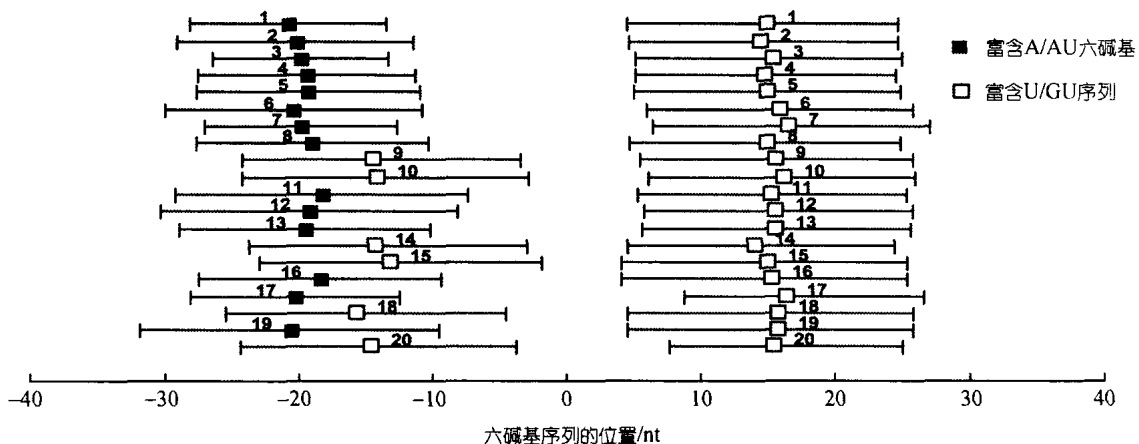


图 6 PAS, DUE 和 poly(A)位点之间的位置关系

带方框的六碱基序列是 PAS, 带下划线的字母表示 DUE, 连续斜体的 a 表示 poly(A)尾巴, 黑色三角形所指的是 poly(A)位点, EST 的登录号标在图左侧. (a) 某基因的多个 poly(A)位点位于同一个 PAS 和 DUE 之间; (b) 某基因的多个 poly(A)位点位于相邻两个 DUE 之间



在-40~0 nt 范围内, 实心方框表示富含 A/AU 的六碱基序列的最后一个碱基的平均位置和 SD, 空心方框表示的是富含 U/GU 序列的最后一个碱基的平均位置和 SD, 具体的六碱基序列见表 4. 在 0~+40 nt 范围内的空心方框表示的是富含 U/GU 序列的第一个碱基的平均位置和 SD, 具体序列见表 4. 方框上方的数字表示其 S 值的序号. 横坐标的原点(0)表示 poly(A)位点的位置

表 4 在-40~+40 nt 区域具有最大 S 值的六碱基序列

poly(A)位点上游的六碱基序列					poly(A)位点下游的六碱基序列				
S 值的名次	六碱基序列	S 值	最后一个碱基的位置/nt	位置的 SD/nt	S 值的名次	六碱基序列	S 值	第一个碱基的位置/nt	位置的 SD/nt
1	GAAUAA	30.2	-20.8	7.4	1	UUGUGU	38.15	14.96	9.76
2	UGAAUA	29.6	-20.2	8.8	2	UGUGUU	37.99	14.57	10.16
3	AAUAAA	26.9	-19.8	6.6	3	UCUGUU	36.52	15.40	9.55
4	AUGAAU	25.8	-19.4	8.1	4	UUCUGU	34.23	14.87	9.72
5	UGAAUG	25.7	-19.3	8.3	5	UGUUUG	34.19	14.98	9.81
6	UGUAAU	24.8	-20.4	9.6	6	UGUUCU	33.96	15.90	9.87
7	AAUGAA	24.1	-19.8	7.2	7	UGCUGU	32.97	16.59	10.40
8	UGAAAU	22.4	-19.0	8.7	8	UGUUGU	32.54	15.01	9.83
9	UUUGUG	22.1	-14.4	10.9	9	UUUGUU	31.44	15.65	10.21
10	UUUGUU	22.1	-14.0	11.1	10	CUGUUU	30.53	16.17	9.73
11	GUUAAU	21.8	-18.3	9.4	11	UGUCUG	30.43	15.31	10.04
12	UGGAAU	21.8	-19.2	9.0	12	GUUUGU	29.97	15.68	10.03
13	AUAAUG	21.2	-19.5	7.8	13	UUGUUU	29.68	15.56	10.12
14	UUGUGU	21.2	-14.2	11.2	14	CUGUGU	29.64	14.05	10.38
15	CUGUUA	21.1	-13.2	11.3	15	UCUGUG	28.52	15.05	10.24
16	UAAUGU	21.1	-18.4	9.2	16	UUUUGU	28.21	15.25	10.01
17	GUAAUG	20.7	-20.3	8.6	17	UUUGUG	27.96	16.48	10.01
18	UGUUGU	20.5	-15.7	11.2	18	UGUUUC	26.93	15.72	9.98
19	CAAUAA	20.5	-20.7	6.8	19	UGUGCU	26.46	15.71	10.01
20	UCUGUU	20.4	-14.5	10.8	20	UUGUUC	26.43	15.46	9.55

+1~+40 nt 两个区域的 S 值, 可以从碱基组成的偏好性上研究这两个区域的序列特点. 如表 4 所示, 在-40~-1 nt, 20 个 S 值最大的六碱基序列中, 既包括富含 A/AU 的序列(如 GAAUAA, AAUAAA), 它们以含有 AAUG, AUGA 和 UGAA 的序列居多, 也包括富含 U/UG 的序列(如 UUUGUG, UUUGUU). 在+1~+40 nt 区域, S 值最大的六碱基序列都是富含 U/UG 的(表 4). 就这些六碱基序列的位置而言, 在 poly(A)位点上游, 富含 A/AU 的序列平均位于-19~-21nt, 位置 SD 大多小于 9.0 nt, 这正是 PAS 集中出现的位置; 富含 U/UG 的序列平均位于-13~-15 nt, 是 PAS 和 poly(A)位点之间 DUE 的位置(图 7). 富含 U/UG 序列位置 SD 大于 9.5 nt, 意味着 DUE 存在于 poly(A)位点两侧一个比较大的范围内. 这些都验证了我们前面的统计结果. 总之, 水稻 mRNA 前体 poly(A)位点的-40~+40 nt 区域序列的特点是, 在位置相对集中但序列并不保守的 PAS 下游, 存在着大量 DUE, 这些 DUE 有可能促进以非 AAUAAA 序列作为 PAS 的 mRNA 前体在断裂/多聚腺苷化时, CPSF 与 PAS 的有效结合.

致谢 本工作为国家高技术研究发展计划(批准号: 2002AA2Z1003)和中国科学院和上海市科学技术委员会(批准号: 038019315)资助项目.

参 考 文 献

- 1 Moore C L, Sharp P A. Accurate cleavage and polyadenylation of exogenous RNA substrate. *Cell*, 1985, 41: 845-855
- 2 Wichkens M. How the message got its tail addition of poly(A) in the nucleus. *Trends Biochem Sci*, 1990, 15: 277-281
- 3 MacDonald C C, Redondo J L. Reexamining the polyadenylation signal: Were we wrong about AAUAAA? *Mol Cell Endocrinol*, 2002, 190: 1-8
- 4 Zarudnaya M I, Kolomiets I M, Potyahaylo A L, et al. Downstream elements of mammalian pre-mRNA polyadenylation signals: Primary, secondary and higher-order structures. *Nucleic Acids Res*, 2003, 31: 1375-1386
- 5 Beaudoin E, Freier S, Wyatt J R, et al. Patterns of variant polyadenylation signal usage in human genes. *Genome Res*, 2000, 10: 1001-1010
- 6 Proudfoot N. Poly(A) signals. *Cell*, 1991, 64: 671-674
- 7 Colgan D F, Manley J L. Mechanism and regulation of mRNA

- polyadenylation. *Genes*, 1997, 11: 2755–2766
- 8 Graber J H, Cantor C R, Mohr S C, et al. *In silico* detection of control signals: mRNA 3'-end-processing sequences in diverse species. *Proc Natl Acad Sci USA*, 1999, 96: 14055–14060
- 9 Loke J C, Stahlberg E A, Strenski D G, et al. Compilation of mRNA polyadenylation signals in *Arabidopsis* revealed a new signal element and potential secondary structures. *Plant Physiol*, 2005, 138: 1457–1468
- 10 Kikuchi S, Satoh K, Nagata T, et al. Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science*, 2003, 301: 376–379
- 11 Hunt A G. Messenger RNA 3' end formation in plants. *Annu Rev Plant Physiol Plant Mol Biol*, 1994, 45: 47–60
- 12 Tian B, Hu J, Zhang H, et al. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res*, 2005, 33: 201–212
- 13 Marino-Ramirez L, Spouge J L, Kanga G C, et al. Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res*, 2004, 32: 949–958
- 14 Schbath S. An efficient statistic to detect over- and under-represented words in DNA sequences. *J Comput Biol*, 1997, 4: 189–192.
- 15 Beaudoin E, Gautheret D. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res*, 2001, 11: 1520–1526.
- 16 Altschul S F, Madden T L, Schaffer A A, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res*, 1997, 25: 3389–3402
- 17 Chou Z F, Chen F, Wilusz J. Sequence and position requirements for uridylyte-rich downstream elements of polyadenylation signals. *Nucleic Acids Res*, 1994, 22: 2525–2531
- 18 Beyer K, Dandekar T, Keller W. RNA ligands selected by cleavage stimulation factor contain distinct sequence motifs that function as downstream elements in 3'-end processing of pre-mRNA. *J Biol Chem*, 1997, 272: 26769–26779
- 19 Sheets M D, Ogg S C, Wickens M P. Point mutations in AAUAAA and the poly (A) addition site: Effects on the accuracy and efficiency of cleavage and polyadenylation *in vitro*. *Nucleic Acids Res*, 1990, 18: 5799–5805
- 20 Li Q, Hunt A G. A near-upstream element in a plant polyadenylation signal consists of more than six nucleotides. *Plant Mol Biol*, 1995, 28: 927–934
- 21 Wu L, Ueda T, Messing J. 3'-end processing of the maize 27 kDa zein mRNA. *Plant J*, 1993, 4: 535–544.
- 22 Wu L, Ueda T, Messing J. The formation of mRNA 3'-ends in plants. *Plant J*, 1995, 8: 323–329
- 23 Andrews E M, DiMaio D. Hierarchy of polyadenylation site usage by bovine *papillomavirus* in transformed mouse cells. *J Virol*, 1993, 67: 7705–7710
- 24 Aranda A, Perez-Ortin J E, Moore C, et al. Transcription termination downstream of the *Saccharomyces cerevisiae* FBP1 [changed from FPB1] poly(A) site does not depend on efficient 3' end processing. *RNA*, 1998, 4: 870
- 25 Chen J S, Nordstrom J L. Bipartite structure of the downstream element of the mouse beta globin (major) poly(A) signal. *Nucleic Acids Res*, 1992, 20: 2565–2572
- 26 Chen F, MacDonald C C, Wilusz J. Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Res*, 1995, 23: 2614–2620
- 27 Wahle E. 3'-end cleavage and polyadenylation of mRNA precursors. *Biochim Biophys Acta*, 1995, 1261: 183–194
- 28 Brown P H, Tiley L S, Cullen B R. Effect of RNA secondary structure on polyadenylation site selection. *Genes Dev*, 1991, 5: 1277–1284

(2006-01-17 收稿, 2006-02-16 接受)