

doi:10.3969/j.issn.1672-5565.2014.04.05

鸡基因组差异表达基因 翻译起始密码子上下游序列的比较研究

王樟凤,卓亮亮,黄子琦,柴学文,饶友生*

(南昌师范学院生物技术研究所,江西 南昌,330032)

摘要:翻译起始调控是基因表达调控的一个关键步骤之一。本文以鸡为研究材料,比较研究了鸡基因组高表达基因和低表达基因翻译起始密码子上下游的碱基序列差异,旨在寻找影响鸡基因表达水平的特异性调控位点。全部3 020个单剪接基因完整的mRNA序列及有详细注释的5' UTRs序列从Ensembl下载。编写计算机程序,读取每个基因mRNA起始密码子上下游各位点的碱基。研究发现,起始密码子上游-3、-2位点可能是鸡基因组基因表达起始密码子正确识别的关键位点。起始密码子上下游的碱基组成分析发现,高表达基因和低表达基因起始密码子的上游均倾向使用(G+C),高表达基因的使用偏倚尤为强烈。序列差异比较发现,高表达基因在-9、-6、-3、+4位点显著偏向G,在-1、-2、-4、-5位点显著偏向C。低表达基因起始密码子上游使用A、U的频率显著高于低表达基因。在-19位点强烈偏向A,在+1、+11、+14位点强烈偏向U。

关键词:鸡;基因表达调控;起始密码子;序列差异

中图分类号:Q953.3 **文献标志码:**A **文章编号:**1672-5565(2014)-04-263-06

Comparison of the sequences around the translation initiation codon between differentially expressed genes in chicken

WANG Zhangfeng, ZHUO Liangliang, HUANG Ziqi, CHAI Xuewen, RAO Yousheng*

(Department of Biological Technology, Nanchang Normal University, Nanchang Jiangxi 330046, China)

Abstract: The control of translation initiation is one of the most fundamental processes in the regulation of gene expression. Here, we investigated the differences of the sequences around the translation initiation codon between highly and lowly expressed genes, and searched for the specific sites regulating gene expression in chicken. Total 3 020 nuclear genes with complete CDSs and no evidence of multiple-splicing forms were included in this study. The complete mRNA sequences and 5' UTRs sequences corresponding to all annotated genes in chicken were downloaded from Ensembl. Sequences around translation initiation codon of the genes were extracted by a personal program. We found that the -3 and -2 sites, upstream of the initiation codon, are more likely the key sites for correct recognition of the translation initiation codon in the chicken genome. Analyses of base composition around the initiation codons indicated that the upstream of the initiation codons tends to use G+ C, especially for highly expressed genes. We also found strong bias of G at position -9, -6, -3 and +4 and strong bias of C at position -1, -2, -4 and -5 in the highly expressed genes. The A and U occurrence frequency of upstream of the initiation codon in the lowly expressed genes is significantly high than that in the highly expressed genes. A strong bias of A was found at position -19, and another strong bias of U was found at position +1, +11, and 14 in the lowly expressed genes. We argued that these sites are more likely the important regulation sites of gene expression, and deserve more further attention.

Keywords: Chicken; Gene expression Regulation; initiation codon; Sequence difference

收稿日期:2014-08-07;修回日期:2014-11-04.

基金项目:江西省教育厅科学技术研究项目(GJJ 13781, GJJ14781)。

作者简介:王樟凤,女,学士,讲师,研究方向:动物分子遗传与育种;E-mail: tmxk1965@ aliyun.com.

* 通信作者:饶友生,男,博士,教授,研究方向:动物分子遗传与育种;E-mail: rys8323571@ aliyun.com.

蛋白质的翻译包括氨基酸的活化、肽链合成的起始、延伸、终止及新生肽链的修饰加工等。肽链合成的起始是基因表达调控的一个关键步骤。在原核生物中,核糖体 30S 小亚基首先与 mRNA 模板结合,然后与 fMet-tRNA^{Met} 结合,继而与 50S 核糖体大亚基结合生成 70S. mRNA. fMet-tRNA^{Met} 起始复合物启动肽链合成的起始。肽链合成正确起始的关键是核糖体 30S 小亚基具有专一性识别和选择 mRNA 起始密码子的特性。Shine 和 Dalgarno 等的研究发现,原核生物 mRNA 的 5' 端非翻译区 (5' Untranslated Region, 5' UTR) 存在一段高度保守的核苷酸序列,也即所谓 SD 序列 (5'-AGGAGGU-3', Shine-Dalgarno sequence), 该序列能与 30S 小亚基的 16S rRNA 3' 端富嘧啶的序列 "5'-GAUACCUCCUUA-3'" 高度互补配对,从而保证了原核生物肽链合成的正确起始。真核生物蛋白质合成的起始机制和原核生物基本相似,其差异主要是核糖体较大 (80S), mRNA 5' 端具有 m⁷GpppNp 的“帽子”结构,3' 端具有多聚腺苷酸的“尾巴”(Poly (A) tail), fMet-tRNA^{Met} 不甲酰化,起始复合物包含更多的起始因子等^[1]。研究者们认为, mRNA 5' 端的 m⁷GpppNp “帽子”结构是正确起始的必要元件,但是,在真核细胞 mRNA 5' UTR 的顺式调控元件 (Cis-acting RNA elements) 中常常包含有上游起始密码子 (Upstream start codons, uAUGs, 位于正确起始密码子的上游) 和上游阅读框 (Upstream open reading frames, uORFs), 如果肽链合成起始于 uAUGs 就必然会导致 mRNA 的错误翻译。因此, mRNA 5' 端的 m⁷GpppNp “帽子”结构并不足以保证肽链合成的准确起始^[2]。Kozak 研究了 211 个真核细胞基因 mRNA 起始密码子上下游的核苷酸序列,发现了一个类似于原核生物 SD 序列的共有序列 (Consensus sequence) “GCCGCC (A/G) CCAUGG” (AUG 为正确的起始密码子, A/G 代表该位点为 A 或 G), 并认为该共有序列保证了翻译的准确起始^[3]。据此, Kozak 提出了一个“滑动扫描”模型,该模型认为真核细胞的 40S 小亚基与多个起始因子结合后与 mRNA 5' 端的 m⁷GpppNp “帽子”结合,继而沿着 mRNA 从 5' 端朝 3' 端滑行,当遭遇起始密码子 AUG 并且 AUG 周围存在共有序列时 (即所谓 Good context, 该序列可能和 40S 小亚基的 18S rRNA 特异性配对), 肽链的合成才能准确起始^[4-6]。但随后的研究发现,由于该研究的样本过小, Kozak 的所谓共有序列并不是真核细胞所共有的,不同的物种在起始密码子 AUG 周围的“喜好序列” (Preferred sequence) 各不相同。在单子叶植物中,该“喜好序

列”表现为“GCGGC(A/C)(A/G)(A/C)CAUGGGCC”, 双子叶植物为“AAAAAAA (A/C) AAUGGCU”^[7]; 真菌为“AAAAAAAAAUGTC”; 果蝇为“ACAACCAAAAUGGC”等^[8]。这表明不同的物种,其蛋白质翻译起始密码子的分子识别机制可能不同。最近, Nakagawa 等对 47 种真核生物的 mRNA 进行了研究 (包括 22 中后生动物如人、鼠、鸡、线虫、果蝇等, 8 种植物, 9 种真菌和 8 种寄生虫), 研究发现,起始密码子上下游的序列呈现丰富的多样性,在起始密码子上游-3 位置偏向于 A/G, -2 位置偏向于 A, 而+5 位置偏向于 C^[9]。作者认为,上述特定位点碱基组成的极显著偏好可能决定了起始密码子的正确识别。本文以鸡为研究材料,比较研究了鸡基因组高表达基因和低表达基因翻译起始密码子上下游的序列差异,期望能够找寻到影响鸡基因表达水平的特异性调控位点。本研究对于鸡转基因研究中外源基因的改造和内源基因的表达有重要的指导意义。

1 材料与方法

1.1 基因序列

鸡基因序列从 NCBI RefSeq database (ftp://ftp.ncbi.nlm.nih.gov/genomes/Gallus_gallus/) 数据库下载,基因的蛋白质编码序列 (coding sequences, CDSs)、完整的 mRNA 序列或 cDNA 序列以及有详细注释的 5' UTRs 序列从 Ensembl 下载。对序列数据的收集制定以下严格标准:(1) CDS 序列完整的且为单剪接基因;(2) CDS 以 ATG 起始,序列长度 > 300 bp, 碱基总数是 3 的倍数而且 CDS 的内部不含有终止密码子;(3) 基因的表达产物为已知的蛋白质分子,其 5' UTR 的 5' 末端和 3' 末端有明确的注释或已通过独立的实验验证。

1.2 基因表达数据

基因表达数据从 NCBI FTP website (<http://www.ncbi.nlm.nih.gov/sites/entrez>) 下载。使用 BLASTN 对每个基因的 EST 和该基因的 mRNA/cDNA 序列进行比对。如果比对长度大于 > 400 bp 且序列一致性 > 96%, 则认为该 EST 为该基因的表达产物;如果比对长度为 100~400 bp 且序列一致性 > 98%, 该 EST 也被接受为该基因的表达产物。全部鸡基因表达数据包括 18 种组织表达谱,分别为血液、脑、法氏囊、盲肠、连接组织、胚胎组织、软骨板、胚胎发育早期性腺组织、胚胎发育早期头部组织、心脏、翅、肝脏、肌肉、卵巢、胰腺、脾脏、睾丸、胸腺。对每个基因在每种组织中表达的 EST 数目进行均一化处理。基因的表达水平定义为全部 18 中组织中 EST 的总和。

1.3 数据处理与分析

编写计算机程序,读取每个基因 mRNA (或 cDNA) 起始密码子上下游各位点的碱基。起始密码子上游从 AUG 的 A (不包括 A) 往上读起,各位点分别定义为-1、-2、-3... -24;起始密码子下游从 AUG 的 G (不包括 G) 往下读起,各位点分别定义为+1、+2、+3... +24。统计研究样本全部基因 mRNA 起始密码子上下游各位点的碱基组成(即各位点 A、U、G、C 四种碱基所占的比例),绘制碱基分布图谱。

2 结果分析

鸡基因组蛋白质编码基因约为 2~3 万个。由于多剪接基因不同剪接体 5'UTR 的长度及起始密码子上下游的序列组成可能存在显著不同,本研究只选取单剪接基因。在全部 10 289 个单剪接基因中,多数基因的 mRNA 序列或 cDNA 序列没有明确的 5'UTR 注

释或者没有经过独立的实验验证,排除这些基因,最终共有 3 020 个基因纳入统计分析。把全部 3 020 个基因按基因的表达水平(全部 18 种组织 EST 的总和)排序,EST 的总和前 20% 的基因(共 755 个)定义为高表达基因,EST 的总和后 20% 的基因(共 755 个)定义为低表达基因。编写计算机程序,读起各基因起始密码子上下游各位点的碱基。

表 1 为高表达基因和低表达基因上游 24 个位点(-1、-2、-3... -24)和下游 24 个位点(+1、+2、+3... +24)A、U、G、C 的使用频率。如果四种碱基随机使用,那么各碱基的使用频率应该趋近 0.25。很显然,A、U、G、C 四种碱基的使用并不是随机的。在-3 位点,高表达基因和低表达基因均倾向使用 A,使用频率分别为 0.436 5 和 0.436 4。在+1 位点,高表达基因和低表达基因均倾向使用 G,使用频率分别为 0.528 4 和 0.474 1。在-2 位点,高表达基因和低表达基因均较少使用 U 和 C。

表 1 鸡高表达基因和低表达基因翻译起始密码子上下游 24 个位点的碱基使用频率

Table 1 The nucleotide acid occurrence frequency around the translation initiation codon in highly and lowly expressed genes in chicken

| 碱基 位点 | 腺嘌呤(A) | | 尿嘧啶(U) | | 鸟嘌呤(G) | | 胞嘧啶(C) | |
|----------|---------|---------|---------|---------|---------|---------|---------|---------|
| | 高表达 | 低表达 | 高表达 | 低表达 | 高表达 | 低表达 | 高表达 | 低表达 |
| -24 | 0.154 6 | 0.238 4 | 0.165 3 | 0.219 3 | 0.356 8 | 0.275 0 | 0.323 1 | 0.267 0 |
| -23 | 0.157 8 | 0.217 8 | 0.165 4 | 0.222 5 | 0.374 8 | 0.283 6 | 0.301 9 | 0.275 8 |
| -22 | 0.159 3 | 0.227 0 | 0.180 4 | 0.192 8 | 0.387 9 | 0.297 0 | 0.272 1 | 0.283 0 |
| -21 | 0.124 0 | 0.250 0 | 0.191 3 | 0.191 3 | 0.394 6 | 0.308 6 | 0.289 9 | 0.250 0 |
| -20 | 0.141 1 | 0.237 3 | 0.172 3 | 0.218 9 | 0.362 5 | 0.266 4 | 0.323 9 | 0.277 1 |
| -19 | 0.164 9 | 0.347 5 | 0.154 6 | 0.171 2 | 0.378 4 | 0.227 5 | 0.301 9 | 0.253 7 |
| -18 | 0.142 0 | 0.241 3 | 0.165 4 | 0.230 8 | 0.377 7 | 0.283 3 | 0.314 7 | 0.244 3 |
| -17 | 0.150 9 | 0.228 0 | 0.169 8 | 0.217 5 | 0.341 0 | 0.263 7 | 0.338 1 | 0.290 6 |
| -16 | 0.143 0 | 0.233 7 | 0.187 8 | 0.224 8 | 0.341 0 | 0.241 1 | 0.328 0 | 0.300 2 |
| -15 | 0.136 1 | 0.210 8 | 0.177 6 | 0.194 7 | 0.342 4 | 0.295 7 | 0.343 8 | 0.298 6 |
| -14 | 0.151 7 | 0.249 2 | 0.170 2 | 0.208 4 | 0.375 8 | 0.266 7 | 0.302 1 | 0.275 5 |
| -13 | 0.139 6 | 0.203 1 | 0.156 5 | 0.217 7 | 0.354 0 | 0.262 6 | 0.349 7 | 0.316 4 |
| -12 | 0.174 6 | 0.224 6 | 0.177 4 | 0.204 5 | 0.339 4 | 0.260 3 | 0.308 4 | 0.310 4 |
| -11 | 0.191 3 | 0.253 5 | 0.155 0 | 0.215 0 | 0.353 3 | 0.257 8 | 0.300 2 | 0.273 5 |
| -10 | 0.183 8 | 0.232 2 | 0.150 4 | 0.213 8 | 0.335 6 | 0.284 7 | 0.330 0 | 0.269 1 |
| -9 | 0.156 2 | 0.263 3 | 0.147 9 | 0.189 0 | 0.435 6 | 0.294 1 | 0.260 0 | 0.253 5 |
| -8 | 0.143 0 | 0.219 7 | 0.137 5 | 0.225 3 | 0.368 6 | 0.272 6 | 0.350 7 | 0.282 3 |
| -7 | 0.150 8 | 0.244 5 | 0.172 8 | 0.207 4 | 0.294 9 | 0.256 8 | 0.381 3 | 0.291 2 |
| -6 | 0.172 9 | 0.221 6 | 0.145 9 | 0.180 5 | 0.498 6 | 0.358 4 | 0.182 4 | 0.239 3 |
| -5 | 0.149 3 | 0.225 8 | 0.165 5 | 0.224 4 | 0.274 5 | 0.269 3 | 0.410 4 | 0.280 2 |
| -4 | 0.191 9 | 0.244 5 | 0.083 2 | 0.148 6 | 0.233 5 | 0.258 1 | 0.491 2 | 0.348 6 |
| -3 | 0.436 5 | 0.436 4 | 0.044 0 | 0.076 3 | 0.465 9 | 0.372 1 | 0.053 4 | 0.115 1 |
| -2 | 0.300 1 | 0.326 2 | 0.115 5 | 0.139 8 | 0.147 4 | 0.189 0 | 0.436 9 | 0.344 8 |
| -1 | 0.164 2 | 0.210 5 | 0.071 5 | 0.112 5 | 0.283 4 | 0.305 9 | 0.480 7 | 0.370 8 |

续(表1)

| 碱基 位点 | 腺嘌呤(A) | | 尿嘧啶(U) | | 鸟嘌呤(G) | | 胞嘧啶(C) | |
|----------|---------|---------|---------|---------|---------|---------|---------|---------|
| | 高表达 | 低表达 | 高表达 | 低表达 | 高表达 | 低表达 | 高表达 | 低表达 |
| 起始密码 | AUG | | AUG | | AUG | | AUG | |
| +1 | 0.181 4 | 0.222 5 | 0.149 6 | 0.137 7 | 0.528 4 | 0.474 1 | 0.140 3 | 0.165 5 |
| +2 | 0.185 4 | 0.315 2 | 0.125 8 | 0.176 1 | 0.178 8 | 0.162 9 | 0.509 9 | 0.345 6 |
| +3 | 0.117 8 | 0.219 8 | 0.182 7 | 0.217 2 | 0.476 8 | 0.335 0 | 0.222 5 | 0.227 8 |
| +4 | 0.250 3 | 0.254 3 | 0.147 0 | 0.188 0 | 0.402 6 | 0.317 8 | 0.200 0 | 0.239 7 |
| +5 | 0.283 4 | 0.311 2 | 0.184 1 | 0.218 5 | 0.203 9 | 0.176 1 | 0.328 4 | 0.294 0 |
| +6 | 0.125 8 | 0.201 3 | 0.127 1 | 0.185 4 | 0.403 9 | 0.335 0 | 0.343 0 | 0.278 1 |
| +7 | 0.249 0 | 0.287 4 | 0.141 7 | 0.172 1 | 0.337 7 | 0.300 6 | 0.271 5 | 0.239 7 |
| +8 | 0.264 9 | 0.286 0 | 0.245 0 | 0.264 9 | 0.221 1 | 0.205 2 | 0.268 8 | 0.243 7 |
| +9 | 0.158 9 | 0.184 1 | 0.161 5 | 0.200 0 | 0.385 4 | 0.312 5 | 0.294 0 | 0.303 3 |
| +10 | 0.237 0 | 0.274 1 | 0.131 1 | 0.165 5 | 0.369 5 | 0.308 6 | 0.262 2 | 0.251 6 |
| +11 | 0.267 5 | 0.267 5 | 0.247 6 | 0.243 7 | 0.254 3 | 0.234 4 | 0.230 4 | 0.254 3 |
| +12 | 0.136 4 | 0.198 6 | 0.145 6 | 0.190 7 | 0.427 8 | 0.332 4 | 0.290 0 | 0.278 1 |
| +13 | 0.230 4 | 0.245 0 | 0.136 4 | 0.182 7 | 0.386 7 | 0.315 2 | 0.246 3 | 0.256 9 |
| +14 | 0.276 8 | 0.267 5 | 0.251 6 | 0.247 6 | 0.223 8 | 0.254 3 | 0.247 6 | 0.230 4 |
| +15 | 0.157 6 | 0.181 4 | 0.161 5 | 0.233 1 | 0.390 7 | 0.307 2 | 0.290 0 | 0.278 1 |
| +16 | 0.247 6 | 0.308 6 | 0.161 5 | 0.177 4 | 0.357 6 | 0.262 2 | 0.233 1 | 0.251 6 |
| +17 | 0.278 1 | 0.274 1 | 0.254 3 | 0.288 7 | 0.223 8 | 0.194 7 | 0.243 7 | 0.242 3 |
| +18 | 0.112 5 | 0.185 4 | 0.156 2 | 0.176 1 | 0.376 1 | 0.303 3 | 0.354 9 | 0.335 0 |
| +19 | 0.229 1 | 0.276 8 | 0.153 6 | 0.202 6 | 0.332 4 | 0.267 5 | 0.284 7 | 0.252 9 |
| +20 | 0.246 3 | 0.242 3 | 0.270 1 | 0.299 3 | 0.226 4 | 0.219 8 | 0.256 9 | 0.238 4 |
| +21 | 0.118 1 | 0.196 0 | 0.160 6 | 0.194 7 | 0.415 6 | 0.323 1 | 0.305 4 | 0.286 0 |
| +22 | 0.223 1 | 0.254 3 | 0.131 4 | 0.192 0 | 0.381 1 | 0.312 5 | 0.264 2 | 0.241 0 |
| +23 | 0.291 6 | 0.246 3 | 0.255 6 | 0.267 5 | 0.213 0 | 0.213 2 | 0.239 6 | 0.272 8 |
| +24 | 0.110 8 | 0.184 1 | 0.146 8 | 0.214 5 | 0.415 2 | 0.275 4 | 0.327 1 | 0.325 8 |

图1反映了(A+U)和(G+C)碱基在起始密码子上下游的使用情况。高表达基因(图1a)和低表达基因(图1b)在起始密码子的上游均倾向使用(G+C),高表达基因(A+U)和(G+C)的使用差异尤为显著。统计发现,高表达基因和低表达基因起始密码子上游的GC含量分别为 $0.674\ 8 \pm 0.018\ 4$ 和 $0.556\ 8 \pm 0.015\ 6$ 。高表达基因起始密码子的下游也倾向使用(G+C)碱基,其GC含量为 $0.616\ 2 \pm 0.039\ 2$,但变异系数较大。低表达基因起始密码子的下游(A+U)和(G+C)的使用无显著差别。

图2反映了高表达基因和低表达基因起始密码子上下游的碱基使用差异。如图2a所示,在起始密码子的上游,低表达基因使用A的频率普遍高于高表达基因。在-19位点,差异尤为显著,低表达基因

使用A的频率为0.347 5,高表达基因为0.164 9。图2b显示,在起始密码子上游的全部24个位点,低表达基因使用U的频率均高于高表达基因;在起始密码子的下游,除了+1、+11、+14三个位点外,低表达基因使用U的频率也均高于高表达基因。从图2c、图2d可发现,在起始密码子的上游,高表达基因使用G、C的频率普遍高于低表达基因。在-9位点、-6位点、-3位点、+4位点,高表达基因使用G的频率分别为0.435 6、0.498 6、0.465 9、0.476 8,低表达基因使用G的频率分别为0.294 1、0.358 4、0.372 1、0.335 0。在-1、-2、-4、-5位点,高表达基因使用C的频率分别为0.480 7、0.436 9、0.491 2、0.410 4,低表达基因使用C的频率分别为0.370 8、0.344 8、0.348 6、0.280 2。其它位点C的使用频率差异不显著。

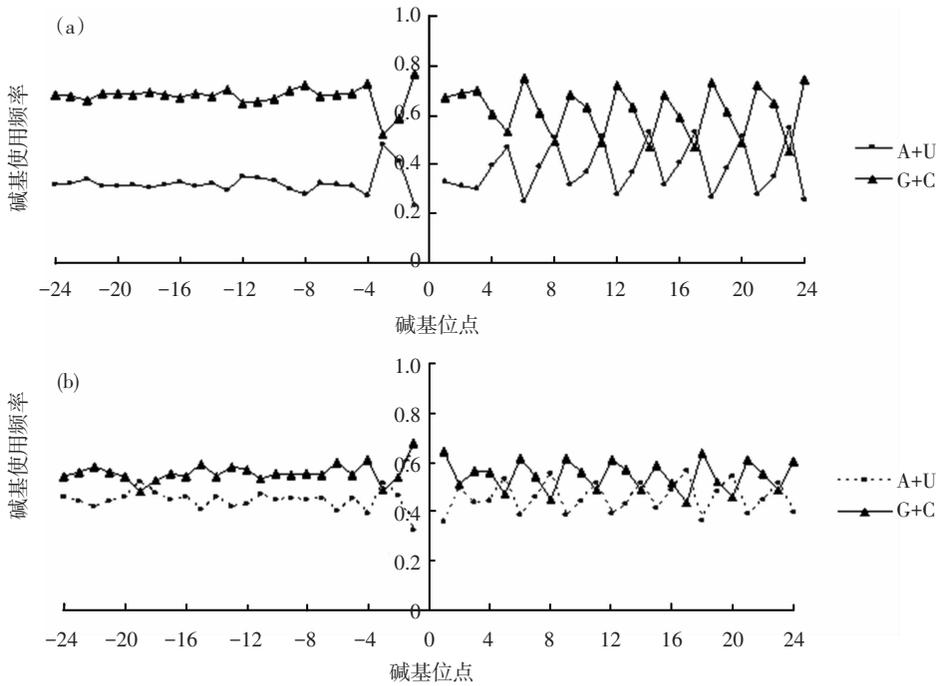


图 1 高表达基因、低表达基因起始密码子上下游(A+U)和(G+C)碱基使用频率散点图
Fig.1 The scat plot of A+U and G+C occurrence frequency around the translation initiation codon in highly and lowly expressed gene in chicken

注：“0”代表起始密码子 AUG。起始密码子上游从 AUG 的 A(不包括 A)往上读起,各位点分别定义为-1、-2、-3... -24;起始密码子下游从 AUG 的 G(不包括 G)往下读起,各位点分别定义为+1、+2、+3... +24。(a)高表达基因起始密码子上下游(A+U)、(G+C)碱基使用频率散点图;(b)低表达基因起始密码子上下游(A+U)、(G+C)碱基使用频率散点图。

Notes: "0" represents the translation initiation codon AUG. The upstream sequence was read from A of AUG codon (not including A). Each point was defined as -1, -2, -3... -24, respectively. The downstream sequence was read from G of AUG codon (not including G). Each point was defined as +1, +2, +3... +24, respectively. (a)The scat plot of A+U and G+C occurrence frequency around the translation initiation codon in highly expressed gene; (b)The scat plot of A+U and G+C occurrence frequency around the translation initiation codon in lowly expressed gene.

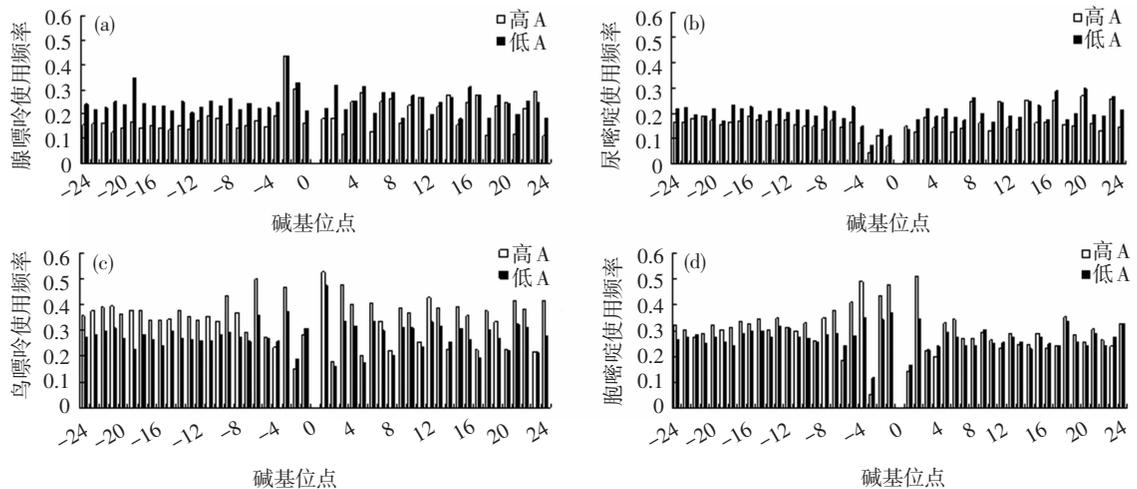


图 2 鸡高表达基因和低表达基因翻译起始密码子上下游各碱基的使用差异
Fig.2 Comparison of the base occurrence frequency around the translation initiation codon between highly and lowly expressed gene in chicken

注:(a)鸡高表达基因和低表达基因碱基 A 的使用比较;(b)鸡高表达基因和低表达基因碱基 U 的使用比较;(c)鸡高表达基因和低表达基因碱基 G 的使用比较;(d)鸡高表达基因和低表达基因碱基 C 的使用比较。

Notes: (a)Difference of A occurrence frequency around the translation initiation codon between highly and lowly expressed gene ; (b)Difference of U occurrence frequency around the translation initiation codon between highly and lowly expressed gene; (c)Difference of G occurrence frequency around the translation initiation codon between highly and lowly expressed gene; (d)Difference of C occurrence frequency around the translation initiation codon between highly and lowly expressed gene.

3 讨论

真核细胞在翻译水平的调控主要是 mRNA 稳定性调控及 mRNA 翻译起始调控。尽管真核细胞的 mRNA 5' 端具有特定的 m⁷GpppNp 的“帽子”结构,但该结构并不能保证肽链合成的准确起始,这是因为在真核细胞 mRNA 5' UTR 的顺式调控元件中常常包含有上游起始密码子和上游阅读框,如果肽链合成起始于 uAUGs 就必然会导致 mRNA 的错误翻译。研究者们发现,原核生物 5' UTR 的 SD 序列是其蛋白质合成正确起始的关键。尽管在真核细胞中是否存在类似于原核细胞的共有序列(Consensus sequence)的问题上还有许多争议,但研究者们普遍认为,mRNA 翻译起始密码子(AUG)上下游的碱基序列对真核细胞蛋白质翻译的准确起始有重要的调控作用^[10-11]。

本研究以鸡为研究材料,比较研究了鸡基因组高表达基因和低表达基因翻译起始密码子上下游的序列差异,期望能够找寻到影响鸡基因表达水平的特异性位点。研究发现,在-3 位点,高表达基因和低表达基因均倾向使用 A,在+1 位点,高表达基因和低表达基因均倾向使用 G,在-2 位点,高表达基因和低表达基因均较少使用 U 和 C。本研究在-3 位点、-2 位点的结果和 Nakagawa 等的研究结果相同^[9],笔者推测,-3、-2 位点可能是鸡基因组基因表达起始密码子 AUG 正确识别的关键位点。mRNA 翻译起始密码子上下游的碱基组成分析发现,高表达基因和低表达基因起始密码子的上游均倾向使用(G+C),高表达基因的使用偏倚尤为强烈,其平均 GC 含量达到 0.674 8±0.018 4;高表达基因起始密码子的下游同样也倾向使用(G+C)碱基。Rao 等最近的研究发现,鸡蛋白质编码基因 5' UTR 的 GC 含量和基因的表达丰度、表达幅度及最高表达水平呈显著正相关^[12]。作者认为,5' UTR 的 GC 含量和基因表达水平的高度正相关可能是由于该区域 CpG 岛(CpG island)的二核苷酸 CpG 中的胞嘧啶没有甲基化或甲基化程度低,阻止了在自然条件下容易发生的胞嘧啶脱氨基导致的胞嘧啶→腺嘌呤的突变所致。尽管本研究探究的只是翻译起始密码子上下游的 48 个碱基序列,而不是全部 5' UTR 序列全长,但也部分佐证了上述研究所结果。

高表达基因和低表达基因起始密码子上下游的碱基序列差异分析发现,高表达基因起始密码子上

游使用 G、C 的频率显著高于低表达基因。在-9、-6、-3、+4 位点显著偏向 G,在-1、-2、-4、-5 位点显著偏向 C。低表达基因起始密码子上游使用 A、U 的频率显著高于低表达基因。在-19 位点强烈偏向 A,在+1、+11、+14 位点强烈偏向 U。上述位点可能是鸡基因表达水平的重要调控位点,值得深入探究。

参考文献(References)

- [1] KOZAK M. How do eucaryotic ribosomes select initiation regions in messenger RNA [J]. *Cell*, 1978, 15: 1109-1123.
- [2] PESOLE G, MIGNONE F, GISSI C, et al. Structural and functional features of eukaryotic mRNA untranslated regions [J]. *Gene*, 2001, 276: 73-81.
- [3] RESCH A M, OGURTSOV A Y, ROGOZIN I B, et al. Evolution of alternative and constitutive regions of mammalian 5'UTRs [J]. *BMC Genomics*, 2009, 10: 162.
- [4] KOZAK M. An analysis of 50-noncoding sequences from 699 vertebrate messenger RNAs [J]. *Nucleic Acids Res*, 1985, 15: 8125-8148.
- [5] KOZAK M. Compilation and analysis of sequence upstream from the translational start site in eukaryotic mRNAs [J]. *Nucleic Acids Res*, 1984, 12: 857-872.
- [6] KOZAK M. Pushing the limits of the scanning mechanism for initiation of translation [J]. *Gene*, 2006, 299: 1-34.
- [7] JOSHI C, ZHOU H, HUANG X, et al. Context sequences of translation initiation codon in plants [J]. *Plant Mol Biol*, 1997, 35: 993-1001.
- [8] CAVENER D, RAY S. Eukaryotic start and stop translation sites [J]. *Nucleic Acids Res*, 1991, 19: 3185-3192.
- [9] NAKAGAWA S, NIIMURA Y, GOJOBORI T, et al. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes [J]. *Nucleic Acids Res*, 2008, 36(3): 861-71.
- [10] ZUR H, TULLER T. New universal rules of eukaryotic translation initiation fidelity [J]. *PLoS Comput Biol*, 2013, 9(7): e1003136.
- [11] KIM Y, LEE G, JEON E, et al. The immediate upstream region of the 5'-UTR from the AUG start codon has a pronounced effect on the translational efficiency in *Arabidopsis thaliana* [J]. *Nucleic Acids Res*, 2014, 42(1): 485-98.
- [12] RAO Y S, CHAI X W, WANG Z F, et al. Impact of GC content on gene expression pattern in chicken [J]. *Genet Sel Evol*, 2013, 45: 9.