

利用 5' LongSAGE 标签分析蜜蜂肌球蛋白调节性轻链基因 *MLC-2* 选择性转录起始位点

周沫玲^{1,2} 孙亮先² 陈培山² 谢宪兵² 黄周英² 倪辉^{1*}

(1. 集美大学 生物工程学院 福建 厦门 361021; 2. 泉州师范学院 福建 泉州 362000)

摘要:使用西方蜜蜂雄蜂头部 5' LongSAGE 文库中的肌球蛋白调节性轻链基因 *MLC-2* 的标签序列,在蜜蜂全基因组序列上定位了该基因的转录起始位点(TSS),并进而预测了其启动子的结构。结果显示:蜜蜂 *MLC-2* 基因存在 17 个 TSS,其中 16 个定位于 *MLC-2* 基因开放阅读框上游 100 bp 的范围内。*MLC-2* 基因的各 TSS 的使用效率不同,其中有 3 个优势 TSS,由之起始的转录本分别占该基因总转录本数量的 9.76%、54.47% 和 11.38%。从碱基组成来看,蜜蜂 *MLC-2* 基因 TSS 的第 1 个碱基为 A、G、T、C 的概率分别为 58.8%、29.4%、0 和 11.8%。*MLC-2* 基因的启动子结构预测结果表明,在 TSS 上游的 300 bp 区域内有 3 个核心启动子元件。研究结果对研究蜜蜂 *MLC-2* 基因的表达调控与确定其全长 cDNA 序列具有重要意义。

关键词:5' LongSAGE; 标签; *MLC-2* 基因; 西方蜜蜂; 转录起始位点; 启动子

中图分类号:S893.69 文献标识码:A 文章编号:1000-2286(2010)01-0149-04

Identification of Alternative Transcription Start Sites of *MLC-2* Gene in *Apis mellifera* by 5' LongSAGE

ZHOU Mo-ling^{1,2}, SUN Liang-xian², CHEN Pei-shan², XIE Xian-bing², HUANG Zhou-ying², NI Hui^{1*}

(1. College of Bioengineering, Jimei University, Xiamen 361021, China; 2. Quanzhou Normal University, Quanzhou 362000, China)

Abstract:The transcription initiation start site (TSS) of a predicted honeybee *MLC-2* gene (myosin regulatory light chain 2) was localized on the genome sequence of *Apis mellifera* with the head of drone's 5' LongSAGE tags mapping to genome sequences in this study, and the structure of promoter was predicted. The results showed that there were 17 TSSs that were assigned to *MLC-2* gene model and 16 tags that mapped in a 100 bp-region upstream of reference start codon. Interestingly, the authors found that those TSSs displayed different initiation efficiency and the *MLC-2* mRNA transcripts were started from 3 dominant TSSs, which initiated 9.76%, 54.47% and 11.38% respectively of the total transcripts of *MLC-2* gene in adult drone head. When base composition of the first base of the 17 alternative TSSs was considered, the authors found that A, G, T and C accounted for 58.8%, 29.4%, 0 and 11.8%, respectively. Then, the promoter structure of *MLC-2* gene was computational analyzed and 3 core promoter elements were predicted in DNA region of 300 bp upstream of TSSs. The precise identification of TSSs was fundamental for the establishment of full-length cDNA sequence and expression regulation of honeybee *MLC-2* gene.

Key words:5' LongSAGE; tag; *MLC-2* gene; *Apis mellifera*; transcription start site (TSS); promoter

收稿日期:2009-10-30 修回日期:2009-11-15

基金项目:国家自然科学基金项目(30300183)和福建省教育厅省属高校项目(2008F5051)

作者简介:周沫玲(1972-),女,硕士生,主要从事食品科学研究;* 通讯作者:倪辉, E-mail:nihui@zmu.edu.cn。

西方蜜蜂 (*Apis mellifera*) 是一种有益的社会性昆虫, 在全球农业和生态中起着重要作用^[1-2]。同时, 西方蜜蜂也是研究社会行为学和基因表达调控的重要动物模型^[3-6]。2006 年西方蜜蜂全基因组序列的释放, 标志着蜜蜂研究已进入了功能基因组时代^[7]。

在完成蜜蜂基因组的测序后, 对基因组进行功能注释, 即鉴定基因组序列中所含有的基因及其调控序列是蜜蜂基因组计划的重要任务。基因的表达与调控是功能基因组学的重要研究内容, 转录起始是表达调控的关键时期。对于编码蛋白质的基因而言, 理论上, 只要能在基因组序列上准确定位基因的转录起始位点 (TSS), 就可以分析该基因的启动子序列和蛋白编码序列。Wei 等所报道 5' LongSAGE 技术可高通量的在基因组中定位 mRNA 的 TSS 序列^[8]。在之前的研究中, 我们使用 5' LongSAGE 技术分析雄蜂头部的转录组, 并用所获得的 5' LongSAGE 标签对蜜蜂基因组的 TSS 进行了作图, 结果发现蜜蜂基因的转录存在选择性起始现象^[9]。为了阐明蜜蜂基因的选择性 TSS, 本研究选取了蜜蜂肌球蛋白轻链调节蛋白基因 *MLC-2* (similar to myosin regulatory light chain 2) 为例, 用来源于该基因的 17 条 5' LongSAGE 标签序列对基因组进行了作图, 并进而分析了该基因的启动子结构。

1 材料与方 法

1.1 来自蜜蜂 *MLC-2* 基因的 5' LongSAGE 标签序列

所使用的 *MLC-2* 基因的 17 条独立 5' LongSAGE 标签序列来源于国家人类基因组研究中心和泉州师范学院模式生物研究中心合作构建的意大利蜜蜂 (*Apis mellifera ligustica*) 雄蜂头部的 5' LongSAGE 文库, 标签序列如下:

5' AAGTTCACCTTCTATGATTC3' 5' ACAAACCAACTGTAGACAT3' 5' ACAGTTGGTTCGTTCTGAC3'
5' ACTTAGACAAGTTCACCTTC3' 5' AGACAAGTTCACCTTCTATG3' 5' AGTTGGTTCGTTCTGACCA3'
5' ATCCTCTTACAAACCAACT3' 5' ATCTCGTACCATCCTCTTA3' 5' ATGATTCGGACAGTTGGTT3'
5' ATTTATCGACACGCAAAGT3' 5' CTATGATTCGGACAGTTGG3' 5' CTTACAAACCAACTGTAGA3'
5' GACAGTTGGTTCGTTCTGA3' 5' GATTCGGACAGTTGGTTCG3' 5' GGACAGTTGGTTCGTTCTG3'
5' GGTTCGTTCTGACCATCTC3' 5' GTTGGTTCGTTCTGACCAT3'

1.2 蜜蜂 *MLC-2* 基因 TSS 的基因组定位和起始效率分析

在 NCBI (<http://www.ncbi.nlm.nih.gov>) 网站上使用 BLAST 将蜜蜂 *MLC-2* 基因的 17 条独立 5' LongSAGE 标签序列定位到蜜蜂基因组序列上 (Amel_4.0), 获取该基因的 TSS, 并根据文库中标签的检测到的频率分析 *MLC-2* 基因各 TSS 的起始效率。

1.3 启动子的分析

从 NCBI 网站下载包含 *MLC-2* 基因的 Contig 序列。根据 *MLC-2* 基因 5' LongSAGE 标签序列在 Contig 中的定位结果, 用 DNASTar 软件截取该基因 TSS 上游 3 kb 至下游 100 bp 的序列, 在果蝇基因组计划 (BDGP) 所提供的软件进行启动子预测 (http://www.fruitfly.org/seq_tools/promoter.html)。

2 结果与分析

2.1 蜜蜂 *MLC-2* 基因的结构

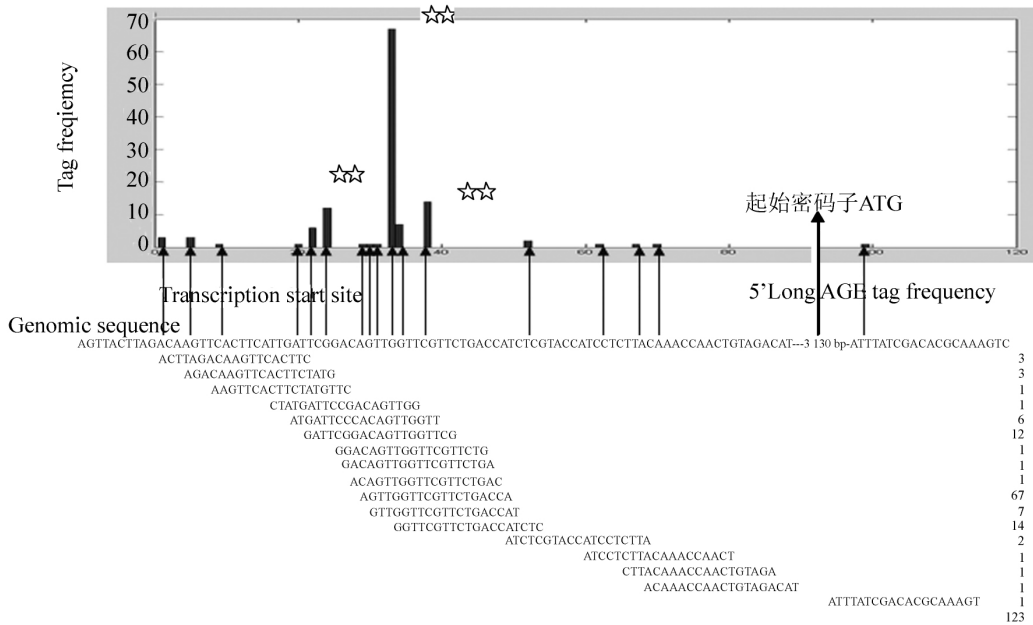
由基因预测软件在蜜蜂基因组里进行基因组功能注释可知, 蜜蜂 *MLC-2* 基因 (Gene ID:409881) 位于 12 号染色体 Contig NW_001253031 (正链) 的 1 043 590 ~ 1 040 008 nt, 基因编码序列 (CDS) 长 3 583 bp, 含 4 个内含子和 5 个外显子, 其预测 mRNA 长 1 029 bp (ID:XM_393371), 预测编码 1 个 213 AA 的蛋白, 此产物蛋白为果蝇 *MLC-2* 蛋白的同源物。

2.2 蜜蜂 *MLC-2* 基因的 TSS

定位结果如图 1 所示: 在成年雄蜂头部中, *MLC-2* 基因至少可以在启动子的 16 个位点上起始转录, 其中有 3 个 TSS 优势起始位点, 由之成功起始的 mRNA 分别占该基因总 mRNA 的 9.76%、54.47% 和 11.38%。从碱基组成来看, 蜜蜂 *MLC-2* 基因 TSS 的第 1 个碱基为 A、G、T、C 的概率分别为 58.8%、29.4%、0 和 11.8%。在 3 个优势 TSS 中, 1 个标签的起始碱基是 A, 共检测到 67 次, 而另两个标签的第 1 个碱基是 G, 分别检测到 12 次和 14 次。

2.3 蜜蜂 *MLC-2* 基因的启动子结构

对 *MLC-2* TSS 上游 3 kb 到下游 100 kb 的序列进行在线预测, 结果如图 2 所示: 在所分析的 DNA



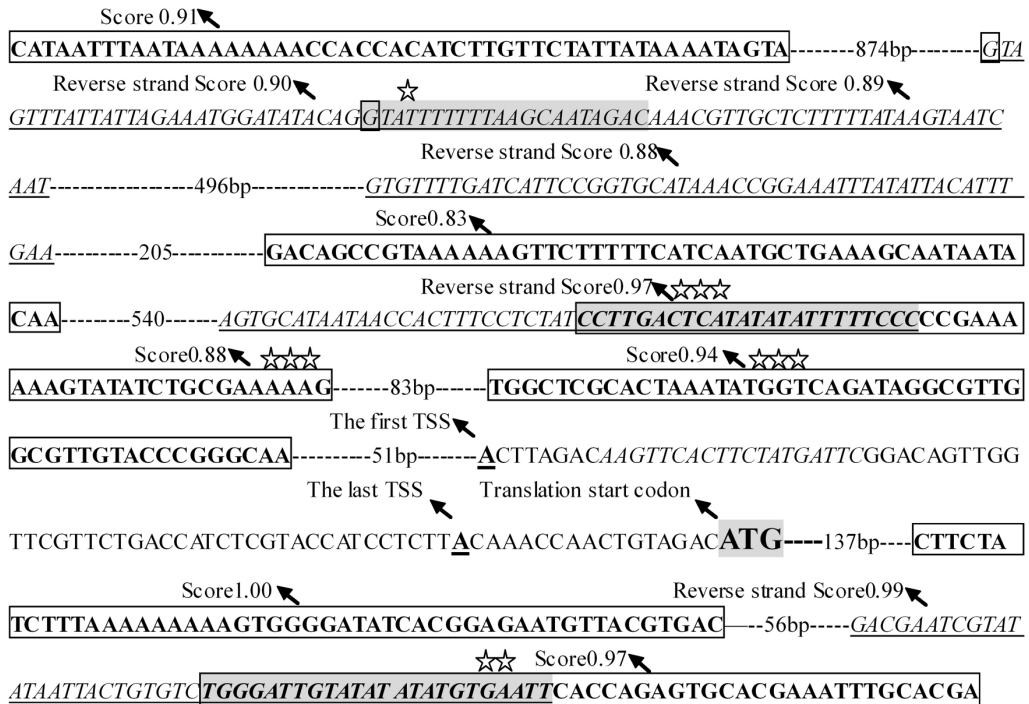
箭头表示预测基因的转录起始位点;星号表示优势 TSS。

Arrows mean start sites and stars mean good TSS.

图 1 用 5' LongSAGE 标签确定的蜜蜂 *MLC-2* 基因 mRNA 起始位点及起始频率

Fig. 1 The mRNA start sites and frequency of honeybee *MLC-2* gene

序列的“+”链上发现了 6 个潜在的启动子元件,其中 2 个元件位于所有转录起始位点上游 300 bp 内,其分值为 0.88 和 0.94,可能是该基因的核心启动子元件。在“+”链上,距离 TSS 上游 2 582 ~ 2 632 bp



图中以边框阴影字体标注的为在“+”链上的启动子元件,以下划线斜体表示“-”链上的启动子元件。“箭头”所指为所预测的启动子元件或 TSS 或翻译起始密码子;“星号”的个数表示该启动子序列的核心程度。

Border means the promoter element of plus strand, underscore and italic print means the promoter element of negative strand, arrow means promoter element of forecasting, TSS or translation initiation codon, the number of star means the core frequency of promoter sequence.

图 2 蜜蜂 *MLC-2* 基因启动子序列预测结果

Fig. 2 Forecasting of promoter sequence of honeybee *MLC-2* gene

也发现存在 1 个启动子元件,其分值为 0.91,而在翻译起始密码子下游 137~327 bp 也发现了 2 个启动子元件,其分值分别为 1.00 和 0.97。由于这 3 个预测的启动子元件距离 TSS 较远,因此认为它们并非关键的调控元件。在“-”链 TSS 上游 218~1 704 bp 发现了 4 个启动子元件,其分值分别为 0.90、0.89、0.88 和 0.97,其中距离第 1 个 TSS 上游 211 bp 处的启动子元件的分值为 0.97,所以可能属于核心启动子元件。在“-”链距离 TSS 下游 251~301 bp 预测到 1 个启动子元件,其分值 0.99。此外,从图 2 可知,所预测的 *MLC-2* 基因正链和负链的启动子序列存在部分重叠现象。本研究在“+”链和“-”链上均预测到启动子元件,并且 DNA 2 条链上的元件之间存在重叠现象。该结果表明蜜蜂 *MLC-2* 基因在转录起始阶段需要多个转录因子与启动子区域 2 条 DNA 链上的多个顺式调控元件作用。

3 讨论

5' LongSAGE 技术利用 CAP-Trapper 原理获得含有 5' 帽子的 cDNA,然后通过一系列加接头、酶切、链接、克隆和高通量测序等操作,最终得到转录组中各 mRNA 分子的 5' 端标签序列,而各标签的检测次数可代表转录组中各 mRNA 分子丰度^[10]。对于已经完成基因组测序的物种而言,使用 5' LongSAGE 标签对基因组作图,可以高通量地获得精确的 TSS 信息^[8]。

在哺乳动物中, *MLC-2* 基因编码肌球蛋白的调节性轻链,而磷酸化的 *MLC-2* 刺激肌球蛋白头部的 ATP 酶活性,增强肌动蛋白与肌球蛋白相互作用,从而能调节肌动蛋白细胞骨架的稳定性^[11]。本研究使用西方蜜蜂雄蜂头部 5' LongSAGE 文库中的肌球蛋白调节性轻链基因 *MLC-2* 的标签序列,在蜜蜂全基因组序列上定位了该基因的转录起始位点(TSS),并进而预测了其启动子的结构。因为 5' LongSAGE 标签序列是来自与全长 mRNA 分子的 5' 末端的 20 nt 片断,因此可代表 mRNA 的 TSS 序列;而所检测到的标签次数则可代表基因以该 TSS 起始转录的频率。在定位于蜜蜂 *MLC-2* 基因 CDS 附近的 17 个独立 5' LongSAGE 标签序列中(图 1),有 16 个标签定位于 *MLC-2* 基因的起始密码子 ATG 上游 90 bp 范围之内,其中标签“AAGTTCACTTCTATGATTC”所定位的 TSS 与预测 mRNA 的 5' 端相同。该结果从转录水平上证明了蜜蜂基因组计划所预测 *MLC-2* 基因确实存在。图 1 中有 1 个 5' LongSAGE 标签被定位到起始密码子下游的 3 130 bp 处。因为预测基因 *MLC-2* 长度为 3 583 bp,而该标签序列位于与 *MLC-2* 基因正链上,因此该标签可能不是来自 *MLC-2* 基因的转录本,而可能来自于与 *MLC-2* 基因相邻的一个尚未发现的新基因。由于真核生物蛋白质编码基因的启动子元件一般位于其 TSS 的上游 300 bp 至下游 50 bp 之内,只有为数不多的编码蛋白基因的启动子分布于 TSS 上游 3 kb 的位置,而超出 TSS 下游 100 bp 的范围内很少存在启动子元件。因此,在确定了 *MLC-2* 基因的 TSS 后,本研究选择了从 TSS 上游 3 kb 到下游 100 bp 的序列进行在线预测启动子预测,得出如图 2 所示的结果。

本研究证明了蜜蜂 *MLC-2* 基因存在选择性起始转录的现象,并对该基因的转录调控序列进行了分析。该研究结果为利用 5' LongSAGE 标签序列数据来分析启动子和确定转录起始位点提供了范例,为今后研究蜜蜂 *MLC-2* 基因的表达调控与确定其全长 cDNA 序列打下了基础。

参考文献:

- [1] Winston M L. The biology of the honey bee [M]. Cambridge: Harvard Univ Press, 1987.
- [2] Morse R A, Calderone N W. The value of honey bee pollination in the United States [J]. Bee Culture, 2000, 128: 1-15.
- [3] Fahrbach S E, Robinson G E. Behavioral development in the honey bee: Toward the study of learning under natural conditions [J]. Learn Mem, 1995, 2: 199-224.
- [4] Robinson G E. From society to genes with the honey bee [J]. Amer Sci, 1998, 86(5): 456-462.
- [5] Page R E, Gadau J, Beye M. The emergence of hymenopteran genetics [J]. Genetics, 2002, 160: 375-379.
- [6] Robinson G E, Grozinger C M, Whitfield C W. Sociogenomics: Social life in molecular terms [J]. Nat Rev Genet, 2005, 6: 257-270.
- [7] G M W. Insights into social insects from the genome of the honeybee *Apis mellifera* [J]. Nature, 2006, 443: 931-949.
- [8] Wei C L, Ng P, Chiu K P, et al. 5' Long serial analysis of gene expression (LongSAGE) and 3' Long SAGE for transcriptome characterization and genome annotation [J]. Proc Natl Acad Sci USA, 2004, 101(32): 11701-11706.
- [9] Zheng H J, Sun L X, Peng W J, et al. Global identification of transcription start sites in the genome of *Apis mellifera* by using 5' LongSAGE [J]. Genome Biol (Submitted), 2002.
- [10] Seki M, Carninci P, Nishiyama Y, et al. High-efficiency cloning of *Arabidopsis* full-length cDNA by biotinylated CAP trapper [J]. Plant J, 1998, 15: 707-720.
- [11] Komatsu S, Ikebe M. ZIP kinase is responsible for the phosphorylation of myosin II and necessary for cell motility in mammalian fibroblasts [J]. Journal of Cell Biology, 2004, 165(2): 243-254.