科学通报 2010 年 第 55 卷 第 14 期：1335 ~ 1340
专题：生物信息学 论 文 www.scichina.com csb.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

miRNA 基因和编码基因启动子区核小体定位分析

刘宏德, 张德金, 谢建明, 袁志栋, 马昕, 卢志远, 龚乐君, 孙啸*

东南大学生物电子学国家重点实验室, 南京 210096
* 联系人, E-mail: xsun@seu.edu.cn

2009-04-22 收稿, 2009-08-28 接受
国家自然科学基金资助项目(60671018, 30800209)

摘要 研究了把基因启动子区的核小体定位对于分析基因的转录调控具有重要意义. 利用核小体定位的预测技术——弯曲度谱, 分析了编码基因和 miRNA 基因启动子周围核小体定位的特征. 基因的转录起始位点处, 有一个核小体缺失区域, 且在下游约 200 bp 处, 有较强的核小体定位信号. 独立转录的内含子 miRNA 基因与基因间区 miRNA 基因, 在启动子区具有相似的核小体定位特征, 在上游 0 ~ –400 bp 间, 有一个较宽的核小体缺失区域, 在该区域分布有较多的转录因子结合位点; 而依赖编码基因转录的内含子 miRNA 基因, 其启动子与蛋白编码基因启动子具有相似的核小体定位特征, 在转录起始位点上游–200 ~ –400 bp 和–400 ~ –600 bp 处, 各有一个较强的核小体定位. 这些结果表明, 独立转录的 miRNA 基因(包括基因间区 miRNA 和独立转录内含子 miRNA)和蛋白编码基因, 在启动子区可能具有不同的核小体定位特征. 核小体定位不仅参与编码基因的转录调节, 也影响 miRNA 基因的转录.

关键词
核小体定位
启动子
miRNA 基因

75% ~ 90%的真核基因组 DNA 包裹在组蛋白八联体上形成核小体[1]. 核小体定位是指 DNA 双螺旋相对于组蛋白核的位置. 这种定位作用封闭了位于核小体上的蛋白结合位点[2,3], 进而阻止蛋白质与 DNA 的结合, 以此达到调节基因转录的作用.

基因启动子结构和功能的分析对于研究转录调节、构建基因间相互作用网络等都具有重要意义. 目前, 对于编码基因启动子的研究较多, 认识较深入. miRNA 由于在细胞的发育分化、疾病(癌症)的发生发展中的有重要作用, 而备受关注[4,5], 但由于 miRNA 前体(pri-RNA)的不稳定性和 miRNA 的组织特异表达量低等原因, 使得对 miRNA 基因自身的转录机制尚不明确[6]. 而从核小体定位的角度分析 miRNA 基因的启动子将有助于理解 miRNA 基因的转录调节机制[6,7].

本文利用作者开发的核小体预测技术——弯曲度谱, 分别分析了编码基因、基因间 miRNA 基因和内含子 miRNA 基因的启动子区域核小体分布的特征. 结果显示独立转录的 miRNA 基因与编码基因在启动子区核小体定位特征上有所差异. 这些分析对于理解 miRNA 基因的转录过程具有重要作用.

1 数据和方法

(ⅰ) 数据. 用一段 DNA 序列(人类 20 号染色体: 83.5~86.1 kb)来检验弯曲度谱的预测能力, 并将结果与 Segal 等的模型结果进行比较(http://genie.weizmann.ac.il/soft-ware/nucleo_prediction.html, version 3.0)[8], 该段序列的核小体位置实验检测数据来自文献[9].

672 条蛋白编码基因启动子 DNA 序列来自人类 20 号染色体, 序列取自 UCSC (http://genome.ucsc.edu/).

英文版见: Liu H D, Zhang D J, Xie J M, et al. Analysis of nucleosome positioning in promoters of miRNA genes and protein-coding genes. Chinese Sci Bull, 2010, 55, doi: 10.1007/s11434-009-3730-2

每条序列长度为 1400 bp, 转录起始位点(TSS)上游(5′端)1000 bp, 下游(3′端)400 bp. 117 种人类 miRNA 基因的转录起始位点信息来自文献[6]. 通过 UCSC, 以 TSS 为对齐点, 分别提取了 miRNA 基因 TSS 上游 1000 bp 和下游 400 bp 的 DNA 序列进行分析. 表 1 列出了 miRNA 的名称, 其中, 基因间区 miRNA 有 43 种; 根据是否具有独立的启动子, 将 74 种内含子 miRNA 分为两类, 即依赖编码基因转录的内含子 miRNA 和独立转录的内含子 miRNA, 两者的数量分别为 51 种和 23 种. 本文中, 将独立转录的内含子 miRNA 基因和基因间区 miRNA 基因合称为“独立转录 miRNA 基因”.

(ⅱ) 核小体预测模型. 核小体预测使用作者此前开发的弯曲度谱[10]. 在以前的研究中, 发现核小体 DNA 双螺旋在两端(~50 bp)比在中间(~47 bp)具有更大的弯曲度, 即具有一种模式, 称作核小体弯曲度特征模式; 利用此模式, 建立了弯曲度谱. 本文利用核小体晶体结构信息, 重构了弯曲度特征信号(图 1), 以提高预测准确度.

利用弯曲度特征预测核小体的方法为:

(1) 用 eq.(1)计算待测 DNA 序列的弯曲度信号[11]

$$C = \nu^0 (n_2 - n_1)^{-1} \sum_{j=n_1}^{n_2} (\rho_j - i\tau_j) \exp\left(\frac{2\pi ij}{\nu^0}\right), \qquad (1)$$

其中 C 的模为弯曲度, ν^0 为 DNA 的平均周期(10.4 bp), ρ 和 τ 分别表征 16 种二苷相对于 B DNA 结构旋转和倾斜的程度, $n_2 - n_1$ 为计算时所取 DNA 片段的长度, 本文中为 11 bp, 计算步长为 1 bp.

(2) 预测核小体. 核小体预测以两条信号(DNA 序列的弯曲度信号和核小体 DNA 的弯曲特征信号)的卷积运算实现, 卷积结果称作弯曲度谱(curvature profile). 如果弯曲度信号中有一段信号与核小体 DNA 弯曲度特征信号相似, 则在弯曲度谱的相应位置会出现一个波峰, 基于此, 便可预测核小体位置. 同时, 我们提供了该方法的在线预测工具(http://www.gri.org.cn/icons).

(ⅲ) 启动子区域核小体定位特征分析. 基因启动子区域的核小体分布特征通过以下过程来计算: 首先计算每条 DNA 序列的弯曲度谱, 然后以 TSS 为对齐点, 加和所有的弯曲度谱并做平滑. 为了研究启动子区核小体定位与转录因子结合位点(TFBS)之间

表 1 3 类 miRNA 的名称 a)

依赖编码基因转录的内含子 miRNA (总计 51 种)	hsa-mir-548b hsa-mir-550-2 hsa-mir-553 hsa-mir-554 hsa-mir-559 hsa-mir-561 hsa-mir-566 hsa-mir-571 hsa-mir-578 hsa-mir-580 hsa-mir-589 hsa-mir-590 hsa-mir-609	hsa-mir-616 hsa-mir-618 hsa-mir-619 hsa-mir-624 hsa-mir-627 hsa-mir-629 hsa-mir-636 hsa-mir-637 hsa-mir-641 hsa-mir-642 hsa-mir-643 hsa-let-7g hsa-mir-103-1	hsa-mir-140 hsa-mir-148b hsa-mir-149 hsa-mir-152 hsa-mir-16-2 hsa-mir-185 hsa-mir-186 hsa-mir-191 hsa-mir-22 hsa-mir-25 hsa-mir-26b hsa-mir-301 hsa-mir-326	hsa-mir-330 hsa-mir-378 hsa-mir-423 hsa-mir-449 hsa-mir-574 hsa-mir-584 hsa-mir-615 hsa-mir-647 hsa-mir-657 hsa-mir-661 hsa-mir-7-1 hsa-mir-611
独立转录的内含子 miRNA(总计 23 种)	hsa-mir-548c hsa-mir-550-1 hsa-mir-604 hsa-mir-634 hsa-mir-635 hsa-mir-639	hsa-let-7c hsa-let-7e hsa-mir-125b-2 hsa-mir-128b hsa-mir-149 hsa-mir-153-1	hsa-mir-20a hsa-mir-30c-1 hsa-mir-339 hsa-mir-33b hsa-mir-340 hsa-mir-450-1	hsa-mir-564 hsa-mir-632 hsa-mir-658 hsa-mir-9-1 hsa-mir-98
基因间 miRNA (总计 43 种)	hsa-mir-200b hsa-mir-101-1 hsa-mir-92b hsa-mir-135b hsa-mir-607 hsa-mir-146b hsa-mir-210 hsa-mir-129-2 hsa-mir-612 hsa-mir-100 hsa-mir-200c	hsa-let-7i hsa-mir-345 hsa-mir-193b hsa-mir-484 hsa-mir-138-2 hsa-mir-195 hsa-mir-365-2 hsa-mir-10a hsa-mir-21 hsa-mir-371 hsa-mir-10b	hsa-mir-130b hsa-mir-563 hsa-mir-138-1 hsa-mir-565 hsa-mir-572 hsa-mir-9-2 hsa-mir-146a hsa-mir-219-1 hsa-mir-30c-2 hsa-mir-30a hsa-mir-148a	hsa-mir-594 hsa-mir-320 hsa-mir-30b hsa-let-7f-1 hsa-mir-222 hsa-mir-374 hsa-mir-18b hsa-mir-505 hsa-mir-648 hsa-mir-196a-2

a) 独立转录的内含子 miRNA 基因和基因间区 miRNA 基因合称为“独立转录 miRNA 基因”, 其数量为 66 种; 来源于文献[6]

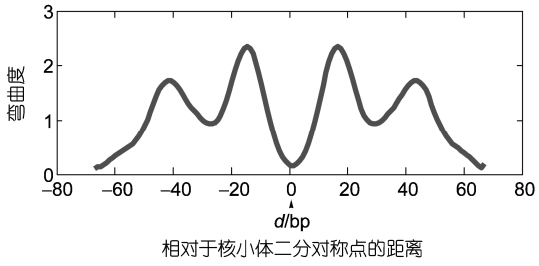

图 1　重构的核小体 DNA 弯曲度特征信号，箭头所指为核小体 DNA 的对称位置(Dyad axis)

的关系，本文用 JASPAR[12]计算了人类 64 种转录因子在 miRNA 基因启动子上结合位点的分布特征，转录因子名称见表 2. JASPAR 提供了每种转录因子的结合位点权重矩阵(PWM). 对于每条 miRNA 基因启动子序列，首先用 64 种转录因子的 PWM 分别进行扫描，对每个位点，加和所有的扫描得分并做归一化处理；然后计算所有序列每个位点的得分，得到转录因子结合位点的分布特征.

2　结果与讨论

2.1　弯曲度谱对一段 DNA 序列的核小体预测

图 2 显示了弯曲度谱和 Segal 等的模型[8]对一段 DNA 片段(人类 20 号染色体: 83.5~86.1 kb)的核小体预测结果. 在两个预测中(黑色粗线和黑色细线)，波峰表示有核小体定位，波谷表示不稳定核小体或者没有核小体定位. 实验检测的结果用灰线表示，该段序列包含 10 个核小体(椭圆表示)[9]. 两者预测结果见表 3，弯曲度谱具有较高的阳性准确率，其他指标两者相当. 总体来看，弯曲度谱具有较好的核小体预测能力.

2.2　基因启动子区核小体定位分析

图 3 显示了编码基因 TSS 周围核小体的分布，图中对随机序列的预测信号呈一条较平坦的曲线，而基因启动子序列的核小体定位有明显的分布特征: 在 TSS 处有一个强烈的核小体缺失区域(nuclesome-free region, NFR)；在紧接着的下游约 200 bp，有核小体定位; TSS 上游 0~−200 bp 处，有一个较弱的核小体定位信号. 这些特征与文献报道的结论一致[6~9,13,14].

TSS 处包含转录起始复合物的结合位点，基因转录时，这些位点的染色质必须处于开放状态，即不能有核小体占位. 因此，TSS 处的核小体缺失是转录过程中染色质所必须具备的特征.

利用弯曲度谱，本文计算了 miRNA 基因的启动子区核小体定位特征(图 4). 图 4(a)是基因间 miRNA 基因启动子特征. 根据是否具有独立启动子，将内含子 miRNA 基因分为依赖编码基因转录(依赖转录) 和独立转录两类(见表 1)，图 4(b)和(c)分别显示这两类启动子区的核小体定位特征. 结果表明: 3 类 miRNA

表 2　人类 64 种转录因子名称

MIZF ; NFYA; ESR1; NR3C1; HNF4A ; NF-kappaB ; TBP ; Cebpa ; REST ; BRAC1 ; TFAP2A ; E2F1 ; ELK1 ; GABPA ; ELK4 ; SPI1 ; SPIB ; ETS1 ; FOXF2 ; FOXD1 ; FOXC1 ; FOXL1 ; FOXI1 ; SOX9 ; SRY ; PBX1 ; NKX3-1 ; Pdx1 ; LHX3 ; TLX1-NFIC; MEF2A; SRF; NR2F1; PPARG-RXRA; PPARG; RORA_1; RORA_2; RXRA-VDR; NR1H2-RXRA; TP53; Pax6; REL; NFKB1; RELA; STAT1; TEAD; IRF1; IRF2; MZF1_1-4; MZF1_5-13; RREB1; SP1; YY1; ZNF354; GATA2; GATA3; NHLH1; Myf; TAL1-TCF3; MAX; MYC-MAX; USF1; CREB1; NFIL3; HLF

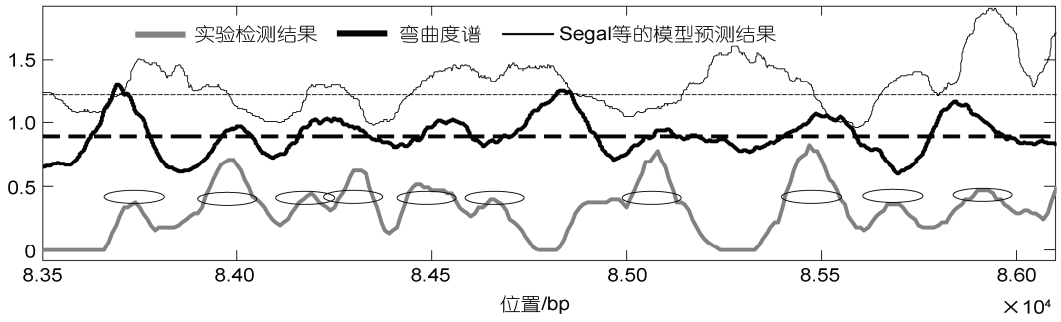

图 2　对一段 DNA 序列(人类 20 号染色体: 83.5~86.1 kb)的核小体预测

黑色粗线弯曲度谱；黑色细线 Segal 等的模型的预测；灰线，实验测定的该段序列的核小体定位信号[9]，椭圆表示实验检测的核小体. 粗虚线弯曲度谱的预测阈值线；细虚线, Segal 等的模型的预测阈值线

表 3 弯曲度谱和 Segal 等的模型的预测结果比较 a)

	TP	FP	FN	阳性准确率(%)	预测准确率(%)	灵敏度(%)
弯曲度谱	7	1	3	87.5	70	70
Segal 等的模型[8]	7	2	3	77.78	70	70

a) 预测位置和实际位置偏移小于 30 bp 定义为真阳性(TP), 偏移超过 30 bp 定义为假阳性(FP), 两者位置边界相距超过 30 bp 定义为假阴性(FN), 阳性准确率=100×TP/(TP+FP), 预测准确率=100×TP/实际的核小体总数=100×TP/10, 灵敏度=100×TP/(TP+FN); 实验数据显示该段序列包含 10 个"核小体"[9]

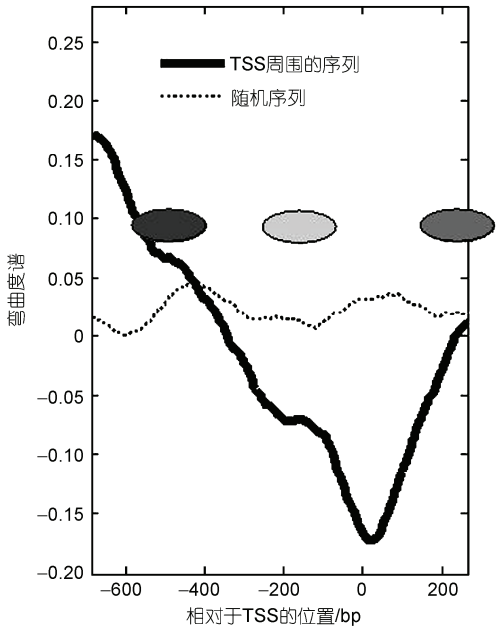

图 3 编码基因转录起始位点(TSS)周围核小体定位特征
实线是对编码基因启动子序列的计算结果, 虚线是对等长度的随机序列的结果(随机序列共 672 条, 每条长度为 1400 bp). 椭圆表示定位的核小体, 颜色的深度与核小体的定位概率(稳定性)成正比, 横坐标为相对于 TSS 的位置, 纵坐标为弯曲度谱信号强度

基因(基因间 miRNA、独立转录内含子 miRNA、依赖转录内含子 miRNA)均在 TSS 处存在核小体缺失区域, 且在 TSS 下游约 200 bp 处有一个核小体定位. miRNA 基因的这些特征与编码基因相似(图 3 和 4). 这说明 TSS 附近的染色质的开放是编码基因和 miRNA 基因共有的特征, 是基因转录的基本条件. 已有的研究证实: RNA 聚合酶 II(Pol II)不仅参与编码基因的转录, 也参与 miRNA 基因的转录. 尽管有些 miRNA 基因利用 Pol III 转录, 但两类聚合酶具有相同的启动子元件[15,16]. TSS 处的核小体缺失正是这种相似性的反映, 同时, 这个结果也与实验检测结果吻合[6].

基因间 miRNA、独立转录内含子 miRNA 和依赖转录内含子 miRNA 的启动子区核小体定位的差异主要体现以下方面(见表 4): (1)基因间 miRNA(图 4(a))和独立转录内含子 miRNA(图 4(c))的启动子在 TSS 上游 0~−400 bp 范围有一个的较宽核小体缺失区域(NFR). 而依赖转录的内含子 miRNA 启动子无此特征(图 4(b)), 其 TSS 上游的 NFR 小于 200 bp. (2)依赖转录的内含子 miRNA 启动子在 TSS 上游−200~−400 bp

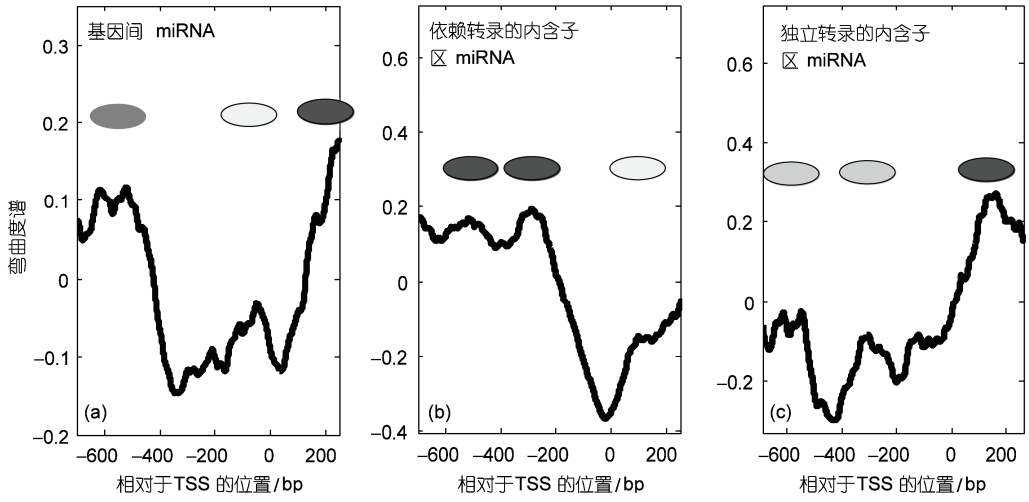

图 4 miRNA 基因启动子核小体定位特征
(a) 基因间 miRNA 基因启动子; (b) 依赖编码基因转录的内含子 miRNA 基因启动子; (c) 独立转录的内含子 miRNA 基因启动子. 椭圆表示定位的核小体, 颜色的深度与核小体的定位概率(稳定性)成正比, 横坐标为相对于 TSS 的位置, 纵坐标为弯曲度谱信号强度

表 4 独立转录 miRNA 基因与编码基因的启动子区核小体定位特征 a)

	上游			TSS	下游
	−400~−600 bp	−200~−400 bp	0~−200 bp	0	~200 bp
编码基因启动子	NP	NP(位置不确定)	NP(较强)	NFR	NP
独立转录 miRNA 基因启动子 b)	NP	NFR	NP(很弱)	NFR	NP

a) 核小体定位; NFR: 核小体缺失区域; b) 包括独立转录内含子 miRNA 基因和基因间 miRNA 基因的启动子

间有一个较强的核小体定位信号, 这与编码基因启动子的特征相似; 基因间 miRNA 启动子在此区域没有定位的核小体; 独立转录内含子 miRNA 启动子在此区域的定位信号非常弱.

比较图 4(b)和图 3 发现, 依赖转录的内含子 miRNA 启动子的核小体定位与编码基因的更为相似. 由于该类 miRNA 基因的转录依赖其所在的编码基因的启动子, 所以实际上两者共享启动子. 而基因间 miRNA 基因和独立转录内含子 miRNA 基因(统称为“独立转录 miRNA 基因”)启动子与编码基因的启动子具有不同的核小体定位特征(尤其在上游 0~−400 bp, 见表 4). 这表明独立转录 miRNA 基因与编码基因在转录上有所差异. 由于核小体定位与蛋白(转录因子)的可接近性有关, 因此, 这种差异说明转录因子在两类启动子上的结合位点分布不同. 需要说明的是, 由于本文涉及的 miRNA 的启动子是通过染色质结构解析而得的[6], 因此本文的结论可能带有一定的偏向性.

如前所述, 独立转录 miRNA 基因启动子在 TSS 上游 0 ~ −400 bp 范围内有一个较宽的核小体缺失区域, 而编码基因和依赖转录的内含子 miRNA 基因启动子在此范围的核小体缺失区域较窄(<200 bp), 而在 0 ~ −200 bp 内有较强的核小体定位信号(依赖转录的内含子 miRNA 基因启动子在此区的核小体定位稍偏向上游). 由于核小体定位会封闭蛋白的结合位点, 阻碍蛋白在 DNA 上的结合, 因此, 核小体定位信号很强的区域必然没有或者很少有 TFBS, 相反, 核小体定位很弱或空缺的区域, 染色质处于开放状态, 适合蛋白的结合, 所以核小体空缺区域应该分布有较多(密)TFBS[6~9,13,14]. 由此推测: 独立转录 miRNA 基因的 TFBS 较多地分布在 TSS 上游 0~−400 bp 范围内. 图 5 证实了这种推测. 图 5 的阴影曲线表示独立转录 miRNA 基因的启动子区核小体定位的分布, 在 TSS 上游 0~−400 bp 的范围内, 核小体定位信号很弱, 其中明显的核小体缺失区域以灰色框标注; 图 5 的黑色曲线是 TFBS 在独立转录 miRNA 基因的启动子区的分布特征, 峰的高度(归一化)表示 TFBS 的分布密度. 从图 5 可见: 在核小体缺失区域或弱核小体定位区域, 转录因子具有较高的密度, 这有力地支持了核小体缺失区域即为转录因子可能结合位点区域的论断, 也说明核小体定位不仅参与编码基因的转录调节, 也影响 miRNA 基因的转录.

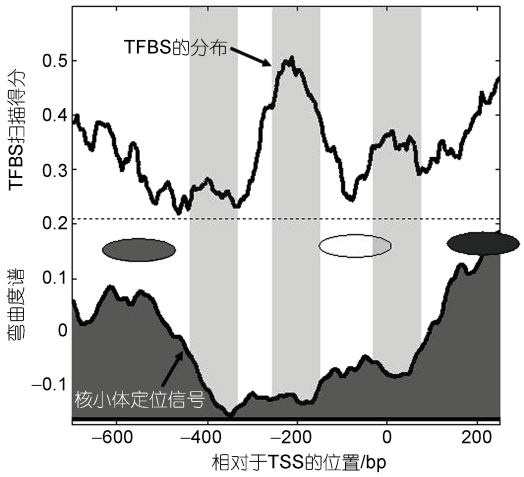

图 5 独立转录 miRNA 基因启动子的核小体定位特征(图下部阴影曲线)和转录因子位点扫描得分分布(图上部黑色实线)之间的关系, 扫描的转录因子见表 2. 椭圆表示定位的核小体, 颜色的深度与核小体稳定性成正比. 横坐标为相对于 TSS 的位置

3 结论

本文利用弯曲度谱分析了基因启动子区核小体定位特征, 在基因启动子 TSS 处, 有一个典型的核小体缺失区域; 同时, 发现独立转录的 miRNA 基因启动子在 TSS 上游 0~−400 bp 范围内有一较宽的核小体缺失区域, 这与编码基因和依赖转录的内含子 miRNA 基因启动子有明显的差异. 本文的研究对于分析 miRNA 基因的转录具有重要意义. 遗憾的是, 由于本文的预测方法是基于 DNA 序列的, 因此, 较

难探查核小体定位的组织特异性, 这也是基于序列依赖性预测核小体位置方法的普遍缺陷. 也许利用组织特异的组蛋白修饰信息是研究组织特异核小体定位的一个途径.

参考文献

1 Lewin B. Gene VIII. New Jerseg: Prentice Hall, 2004
2 Segal E, Mittendorf Y F, Chen L, et al. A genomic code for nucleosome positioning. Nature, 2006, 442: 772—778
3 Henikoff S. Nucleosome destabilization in the epigenetic regulation of gene expression. Nat Rev Genet, 2008, 9: 15—26
4 Lagos-Quintana M, Rauhut R, Lendeckel W, et al. Identification of novel genes coding for small expressed RNAs. Science, 2001, 294: 853—858
5 Pasquinelli A E, Hunter S, Bracht J. MicroRNAs: A developing story. Curr Opin Gene Dev, 2005, 15: 200—205
6 Ozsolak F, Poling L L, Wang Z X, et al. Chromatin structure analyses identify miRNA promoters. Genes Develop, 2008, 22: 3172—3183
7 Abeel T, Saeys Y, Bonnet E, et al. Generic eukaryotic core promoter prediction using structural features of DNA. Genome Res, 2008, 18: 310—323
8 Kaplan N, Moore I K, Mittendorf Y F, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature, 2009, 458: 362—366
9 Schones D E, Cui K, Cuddapah S, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell, 2008, 132: 887—898
10 Liu H D, Wu J S, Xie J M, et al. Characteristics of nucleosome core DNA and their applications in predicting nucleosome positions. Biophys J, 2008, 94: 4597—4604
11 Widlund H R, Kuduvalli P N, Bengtsson M, et al. Nucleosome structural features and intrinsic properties of the TATAAACGCC repeat sequence. J Biol Chem, 1999, 274: 31847—31852
12 Wasserman W W, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet, 2004, 5: 276—287
13 Ozsolak F, Song J S, Liu X S, et al. High-throughput mapping of the chromatin structure of human promoters. Nat Biotechnol, 2007, 25: 244—248
14 Ioshikhes I, Albert I, Zanton S J, et al. Nucleosome positions predicted through comparative genomics. Nat Genet, 2006, 38: 1210—1215
15 Yoontae L, Kim M, Han J, et al. MicroRNA genes are transcribed by RNA polymerase II. EMBO J, 2004, 23: 4051—4060
16 Johnson S S, Zhang C, Fromm J, et al. Mammalian Maf1 is a negative regulator of transcription by all three nuclear RNA polymerases. Mol Cell, 2007, 26: 367—379