

人管家基因启动子序列中的组合 转录调控元件分析 *

李慧敏^{1, 4)**} 石 玉²⁾ 杨志刚³⁾ 江绍萍^{1)**}

⁽¹⁾ 云南民族大学数学与计算机科学学院, 昆明 650500; ⁽²⁾ 郑州师范学院数学与统计学院, 郑州 450044;

⁽³⁾ 四川大学计算机学院, 成都 610065; ⁽⁴⁾ 云南大学生物资源保护和利用重点实验室, 昆明 650091)

摘要 研究表明, 第一内含子可能参与基因转录调控. 利用统计方法提取人管家基因上游至第一内含子序列中潜在的组合转录调控模体, 分析模体间的距离、区域分布等特征, 探讨内含子参与基因转录调控的可能性及其参与方式. 在管家基因中共获得 960 对潜在转录调控模体对, 其中 57% 与实验已知的具有转录相互作用的因子对吻合, 共涉及 12 组因子对. 分析发现, 绝大多数模体对(> 80%) 偏向于上游区域及“上游-内含子”区域, 进一步支持了内含子参与基因转录调控的假设, 并据此推测内含子与上游序列之间具有转录协同作用, 模体在基因转录起始位点(TSS)附近较为集中, 模体对的两个模体之间距离较近, 60%左右距离在 200 bp 以内, 特别地, 65%的模体对特征距离在 100 bp 以内, 短距离间隔有利于转录因子间的协同作用. 这些结果将有助于对人基因转录调控机制及内含子功能的深入认识.

关键词 人管家基因, 组合转录调控, 模体对, 内含子, 区域偏好, 距离分布

学科分类号 Q6, Q7, Q987

DOI: 10.3724/SP.J.1206.2014.00069

转录调控是真核生物基因表达调控的重要环节, 也是当代功能基因组学研究的前沿和热点. 尽管实验和理论分析都对真核基因的转录调控做过很多研究, 但对其转录调控的机制还不完全清楚. 对真核基因转录调控所涉及范围缺乏充分的认识是造成这个结果的重要原因之一. 目前, 对真核基因转录调控的研究主要集中在基因转录起始位点(transcription start site, TSS)上游或邻近区域^[1-2], 然而, 实验结果表明基因其他区域(如内含子、5'-UTR 区域)也含有转录调控元件.

内含子(intron)是从前体 mRNA 到成熟 mRNA 加工过程中被切除掉的非编码序列, 在真核基因中所占比例很高. 很长一段时间内, 内含子被认为是基因组中的“垃圾”(junk)片段. 然而, 随着对内含子研究的不断深入, 人们发现这些所谓的“垃圾”片段具有重要的生物学功能^[3]. 内含子中不仅含有多种非编码 RNA, 还包含众多与基因表达有关的调控元件. 某些基因的内含子中也含有转录调控元件, 它们可以作为增强子(enhancer)或抑制子

(silencer)发挥作用, 甚至有些内含子可以作为基因启动子(promoter)^[4-5].

对酵母等简单模式真核生物内含子参与转录调控已有较多的认识^[6-7]. 研究表明, 酵母基因的转录频率很大程度受内含子序列影响, 并在高效转录基因内含子中检测到一批转录调控元件^[6]; 对酵母高效转录基因上游和内含子中转录调控元件的分析表明, 内含子和上游之间具有转录协同作用^[7]. 酵母基因组的 6 000 多个基因中, 仅有 239 个具有内含子. 而哺乳动物大部分基因具有内含子, 并且基因序列和内含子形式更加复杂. 这些内含子是否也参与基因转录调控? 如果参与, 其作用机制如何?

* 国家自然科学基金(31360277, 31460297 和 11461083), 云南省教育厅科学研究项目(2013Z042), 河南省教育厅科学技术项目(14A110020)及云南大学理工类科研基金项目(2012CG016)资助.

** 通讯联系人.

李慧敏. Tel: 13888183685, E-mail: lihuimin_1980@126.com

江绍萍. Tel: 13888096496, E-mail: jiang2005124@163.com

收稿日期: 2014-06-03, 接受日期: 2014-08-11

这些问题都亟需解决。

多项实验研究表明,哺乳动物某些基因内含子尤其第一内含子中也含有转录调控元件^[8-12]。小鼠 *C-fos* 基因的第一内含子中包含 TBP、TRE 和 CRE 族转录因子结合位点^[8]; 小鼠 *Plp1* 基因和人 *HSD3B2* 基因第一内含子中均发现了 YY1 因子的结合位点^[9-10]; 人 *APP* 基因第一内含子中包含转录因子 E4BP4/NFIL3 的结合序列^[11]。前期我们曾对典型转录调控元件 TATA box、GC box 和 CAAT box 在人基因中的出现情况进行分析,发现这几个元件在第一内含子中具有较高的频率,进一步支持了第一内含子参与转录调控的假设^[13]。

管家基因(housekeeping genes, HK genes)是生物体内几乎所有细胞中都存在的基因,对其涉及内含子的转录调控元件分析,将有助于我们理解内含子在基因转录调控中的作用。本文拟通过对人 HK 基因启动子序列(含第一内含子)中共出现的元件(亦称模体, motif)进行细致分析,找出其中可能存在的组合调控规律。目前预测组合调控元件大多基于模体对共出现概率^[14],而其中单个模体的识别是模体对共出现分析的第一步。一般计算上寻找单个模体常用的方法之一是基于字符串分析,即统计共调控基因中模体出现次数,将其与期望次数进行比较,找出具有统计显著性的模体。采用不同方法计算期望次数时,获得的结果也有差异。为使结果更准确,本文综合考虑常见的几种方法,将在不同方法下都显著出现的模体作为潜在转录调控模体。已经证实以此为基础预测组合调控元件时取得了较好效果^[15]。此外,在预测组合调控元件时, Yu 等^[14]考虑了观测到的模体对距离偏离背景模体对距离问题,但他们假设背景距离服从均匀分布。而本文所用背景距离分布是真实观测到的背景序列的距离分布,同时假设相互作用的模体间距离小于背景距离,这样更符合生物学意义。

1 数据与方法

1.1 基因序列数据

根据 HuGEIndex(<http://www.biotechnologycenter.org/hio/>)给出的人 HK 基因 ID,从 NCBI(<http://www.ncbi.nlm.nih.gov/>)数据库中提取它们对应的基因序列。由于研究主要表明第一内含子中含有转录调控元件,因此本文将基因 TSS 上游 1 000 bp 至第一个内含子结束的序列作为启动子序列。为避免重复定义,对于具有可变剪切的基因,只取其中一

条序列作为启动子序列。剔除无内含子和线粒体中的基因,共获得 414 条 HK 基因的启动子序列。同时取出 NCBI 数据库中给出的人基因的全部启动子序列(启动子序列按上述定义)作为对照集。由于涉及内含子的转录调控是我们关注的重点,本文所有启动子序列都是针对有内含子的基因而言的。

1.2 提取潜在转录调控模体

我们分别采用 *U*-Score, *S*-Score 和 *Z^m*-Score 三种统计量提取启动子序列中过出现的模体(over-represented motif),然后将在三种方法下都出现的模体作为潜在转录调控元件,进行下一步分析。具体统计量如下:

a. *U*-Score 统计量。 S_1 和 S_2 分别表示人 HK 基因和人全部基因的启动子序列(下同)。长度为 l 的 DNA 序列(l -mer) w 在 S_1 和 S_2 中的出现次数分别为 $n_1(w)$ 和 $n_2(w)$, n_1 和 n_2 分别表示所有形式的 l -mer 在 S_1 和 S_2 中出现的总次数, w 在 S_1 和 S_2 中出现的频率分别为 $n_1(w)/n_1$ 和 $n_2(w)/n_2$ 。计算

$$U(w) = \frac{n_1(w)/n_1 - n_2(w)/n_2}{\sqrt{\frac{n_1(w)+n_2(w)}{n_1+n_2} \left(1 - \frac{n_1(w)+n_2(w)}{n_1+n_2}\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (1)$$

当 $U(w) > 1.64$ 时($P < 0.05$),认为 w 在 S_1 中过出现。

b. *S*-Score 统计量。如果 w 在 S_1 中过出现,它应该比较普遍地出现在 S_1 中,而在随机序列中较少。为此,将 S_1 中的每条序列保持长度和碱基含量不变,随机打乱(shuffled)作为一组随机序列。为减少打乱过程中随机涨落造成的误差,将 S_1 中序列打乱 1 000 次。令 $N(w)$ 表示 w 在 S_1 中出现的次数, $N_i(w)$ ($i=1, 2, \dots, 1000$) 表示第 i 组随机序列中 w 出现的次数, $\hat{N}(w)$ 和 $\hat{V}(w)$ 分别表示 $N(w)$ 的均值和方差, $\hat{N}(w) = \sum_{i=1}^{1000} N_i(w)/1000$, $\hat{V}(w) = \sum_{i=1}^{1000} (N_i(w) - \hat{N}(w))^2 / (1000-1)$ 。计算

$$S(w) = \frac{N(w) - \hat{N}(w)}{\sqrt{\hat{V}(w)}}, \quad (2)$$

当 $S(w) > 1.64$ 时($P < 0.05$),认为 w 在 S_1 中过出现。

c. *Z^m*-Score 统计量。由于 DNA 序列碱基之间的关联性,假设启动子序列是一个 m 阶 Markov 链,状态空间 $\Omega = \{A, C, G, T\}$ 。 μ 和 π 分别表示 Markov 链的释放概率和转移概率。则 $w = w_1 w_2 \dots w_l$ 在 S_1 中概率为 $P(w) = \mu(w_1) \dots \pi(w_l - m, w_{l-m+1}) \dots \pi(w_{l-1}, w_l)$,

其中 $\mu(w_1w_2\cdots w_m)$ 和 $\pi(w_{i-m}\cdots w_{i-1}, w_i)$ ($i=m+1, \cdots, l$) 分别由 $N(w_1w_2\cdots w_m)/n$ 和 $N(w_{i-m}\cdots w_i)/N(w_{i-m}\cdots w_{i-1})$ 估计, n 为 S_1 中序列总长度, $N(\cdot)$ 表示 $(\cdot)$ 在 S_1 中出现的次数. $\tilde{N}_m(w)$ 和 $\tilde{V}_m(w)$ 分别为 Markov 模型下 w 出现次数 $N(w)$ 的期望和方差, 具体估计方法见文献[16]. 计算

$$Z^n(w) = \frac{N(w) - \tilde{N}_m(w)}{\sqrt{\tilde{V}_m(w)}} \quad (3)$$

当 $Z^n(w) > 1.64$ 时 ($P < 0.05$), 认为 w 在 S_1 中出现.

1.3 提取组合转录调控模体对

由于转录调控模块中不同模体之间的相对位置(在模体对中模体之间的相对位置可体现为两模体之间的间隔距离)也具有一定的规律性^[17], 因此分两步提取潜在转录调控模体对.

a. 将出现过模体两两组合, 计算模体对 M1:M2 在 S_1 中的共出现率 P_{occ} :

$$P_{occ} = 1 - \sum_{k=0}^{t-1} C_g^k C_{G-g}^{T-k} / C_G^T \quad (4)$$

其中, g 和 G 分别表示 S_1 和 S_2 中基因的条数, t 和 T 分别表示 S_1 和 S_2 中包含 M1:M2 的基因条数. 取 P_{occ} 的 95% 分位点为阈值 b , 当 $P_{occ} < b$ 时, 认为 M1:M2 在 S_1 中的出现频率显著高于其在 S_2 中的出现频率, 称 M1:M2 为共出现模体对 (co-occurred motif pair).

b. 利用 Mann-Whitney U-test 检验共出现模体对 M1:M2 在 S_1 和 S_2 中的距离差异. Mann-Whitney U-test 的统计量 U 为

$$U = n_1n_2 + n_1(n_1+1)/2 - R_1 \quad (5)$$

其中 n_1 和 n_2 分别表示 M1:M2 在 S_1 和 S_2 中出现的次数, R_1 是 M1:M2 在 S_1 中距离的秩和. 作如下变换:

$$Z = \frac{U - n_1n_2/2}{\sqrt{n_1n_2(n_1+n_2+1)/12}} \quad (6)$$

假设相互作用的模体之间具有较近的距离间隔, 进行单边假设检验, 当 $Z < -1.64$, 即 $P_{dis} < 0.05$ 时, 认为 M1:M2 在 S_1 中距离显著小于其在 S_2 中的距离. 结合 M1:M2 在 S_1 中的共出现, 称 M1:M2 在 S_1 中过出现, 即是潜在的组合转录调控模体对 (combinatorial transcriptional regulatory motif pairs).

由于转录因子在 DNA 双链具有活性^[18], 本文统计都是在双链上进行. 即模体对 M1:M2, M1':M2, M1:M2', M1':M2' 具有相同的 P_{occ} 和 P_{dis} 值,

其中 M1' 和 M2' 分别是 M1 和 M2 的反转互补序列.

2 结 果

2.1 探测的组合转录调控模体对

研究表明, 真核生物基因中转录因子结合位点的核心片段通常较短 (6~10 bp)^[19], 本文中主要考察 6 核苷酸, 长的结合位点可由 6 核苷酸连接或重叠而成. 6 核苷酸共有 $4^6=4096$ 种形式, 利用 1.2 中方法抽提出 218 个潜在转录调控模体, 组成 $218 \times 218 = 47524$ 个候选模体对, 由方法 1.3, 共得到 960 对在 HK 基因中潜在的转录调控模体对. 将这些模体对与 TRANSCOMPEL 数据库以及文献[1, 20] 中预测的相互作用因子对进行对照, 发现 57.29% 与已知结果吻合, 共涉及 12 对转录因子的相互作用. 其中 6 对有转录因子 SP1 参与, 分别是 NF1:SP1、SP1:HNF4、P53:SP1、AP1:SP1、STAT1:SP1 和 NF-KAPPAB:SP1. SP1 是 CpG 岛的关键结合因子, 而未甲基化的 CpG 岛是基因高表达的必要条件, HK 基因中 CpG 岛大部分未甲基化^[21], 推测 SP1 与 HK 基因的高表达有关. 其余 6 对分别是 YY1:NF1、E2F:E2F、AP2ALPHA:EGR1、AP2ALPHA:AREB6、AP2:ER 和 HNF4:COUP. 分析发现, 除 SP1:HNF4、YY1:NF1 以及 HNF4:COUP 的因子之间是拮抗 (antagonism) 作用外, 其余 9 对转录因子之间都是协同 (synergism) 作用. 由于 P53:SP1、AP1:SP1、STAT1:SP1、NF-KAPPAB:SP1、AP2ALPHA:AREB6、AP2:ER 和 HNF4:COUP 的结合模体对较少 (< 20), 而 AP2ALPHA:EGR1 所对应模体对全部是 NF1:SP1 的结合序列, 因此 NF1:SP1、SP1:HNF4、YY1:NF1、E2F:E2F 结合模体对信息可以反映潜在组合转录调控元件的特征, 以下分析主要针对这几组转录因子对及它们对应的模体对 (表 1).

2.2 相互作用模体对的区域分布

将启动子分为上游、外显子和内含子 3 个区域, 模体对 M1:M2 可能出现在 6 种区域 (图 1a): a. 2 个模体都在上游 (U-U); b. 1 个在上游, 另 1 个在外显子 (U-E); c. 1 个在上游, 另 1 个在内含子 (U-I); d. 2 个都在外显子 (E-E); e. 1 个在外显子, 1 个在内含子 (E-I); f. 2 个都在内含子 (I-I). 利用公式 $Z_D = (m - Qp_0) / \sqrt{Qp_0(1-p_0)}$ 来确定 M1:M2 在区域 D 的出现情况. 其中, m 是 M1:M2 在 D 出现的次数, Qp_0 为 M1:M2 在 D 的期望出现次数, p_0 为 D 内 M1:M2 在背景序列中出现的频率, Q 是

Table 1 Extracted motif pairs and corresponding TF pairs

TF pairs	Motif pairs (Mps)				
NF1:SP1	CGGCGC:GGGCGG, CGCGCC:GGGGCG, GAGGGG:GGCGCG, CCGGCC:GCCCCG, CGGGGC:GCGCCG, CCGGCC:GCGGGG, CGGGGC:GCGCCG, CCGGCC:GCGGGG, CCGGCC				

*Reversed and complementary motif pairs were not listed.

M1:M2 在 HK 基因中出现的次数. M1:M2 在某区域中 $Z_0 > 1.64$ (即 $P < 0.05$) 时, 认为 M1:M2 在该区域具有偏好性, 即更倾向于在区域 D 组合.

图 1b 显示了探测的模体对在各区域的出现情况. 从图中可以看出, 将近 80% 的模体对偏好 U-U 区域, 约 90% 的模体对偏好 U-I 区域, 只有少数模体对偏好 U-E (4.16%)、E-I (2.92%) 和 I-I 区域 (0.73%), 没有发现偏好 E-E 区域的模体对. 此外, 约有 75% 的模体对偏好一个以上区域, 同时偏好 U-U 和 U-I 区域的模体对为 73%, 而 40 个偏好 U-E 区域的模体对中, 有 32 个同时也偏好 U-U 区域和 U-I 区域. 分别对各组调控模体对的出现区域进行分析 (表 2), 发现每组模体对都偏好 U-U 区域和 U-I 区域. 其中 NF1:SP1 最偏好 U-I 区域, SP1:

HNF4, YY1:NF1 最偏好 U-U 区域, 而 E2F:E2F 两个区域的偏好性差别不大. 已知基因上游是转录调控的重要区域, 而模体对偏好 U-I 区域则反映了上游与内含子之间在转录调控水平上具有相互作用. 特别是, NF1 和 SP1 之间具有协同作用^[22], 而 SP1 和 HNF4, YY1 和 NF1 之间为拮抗作用, 因此推测上游与内含子之间具有转录协同作用, 可以增加基因转录效率, 与 HK 基因的高表达一致. SP1: HNF4 和 YY1:NF1 模体对也偏好 U-E 和 E-I 区域, 表明上游和外显子, 外显子和内含子之间也可能存在转录相互作用, 但这种作用多为拮抗作用. 此外, 模体对在 E-E 和 I-I 区域都不具偏向性, 提示在人 HK 基因中, 外显子和内含子作为独立启动子的可能性较小.

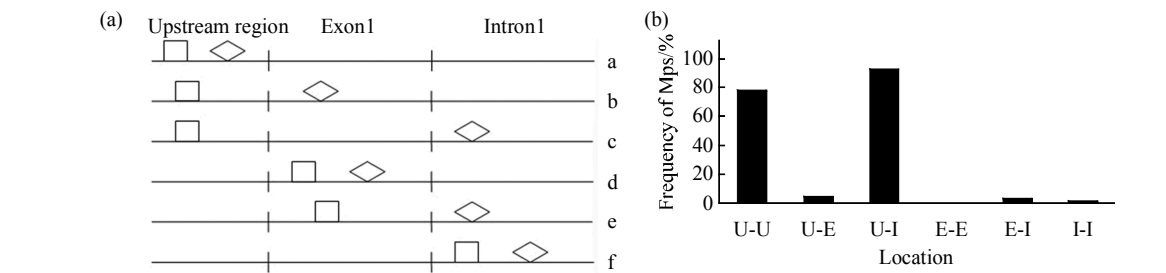


Fig. 1 Frequency of motif pairs in different locations

Table 2 The Z-value of motif pairs in different locations

TF pairs	Z_{U-U}	Z_{U-E}	Z_{U-I}	Z_{E-E}	Z_{E-I}	Z_{I-I}	Loc-num
NF1:SP1	69.21	2.96	73.3	-49.55	0.27	-61.92	3
SP1:HNF4	38.74	6	36.24	-16.35	8.86	-45.3	4
YY1:NF1	28.21	2.61	26.04	-14.24	2.23	-30.83	4
E2F:E2F	9.59	-1.1	9.78	-9.35	-4.12	-3.79	2

*Loc-num: the numbers of preferred locations corresponding to TF pairs.

2.3 相互作用模体对的距离特征

对模体对中两个模体的位置进行分析, 发现 95%的模体位于基因上游-1 000 bp 到下游 3 050 bp 以内, 模体出现最频繁的几个区间为-150~-100 bp、-50~0 bp, 100~150 bp 和 200~250 bp(图 2a).

由于这些区域都位于-200~300 bp 之间, 我们细致分析了模体在-200~300 bp 之间的情况(图 2b). 结果发现, 模体在这些区间的一些位置出现了峰值, 如-125、-65、-54、-12、69、136 和 235 bp 等. 其中, 出现次数最多的位置在上游-65 bp, 模体在

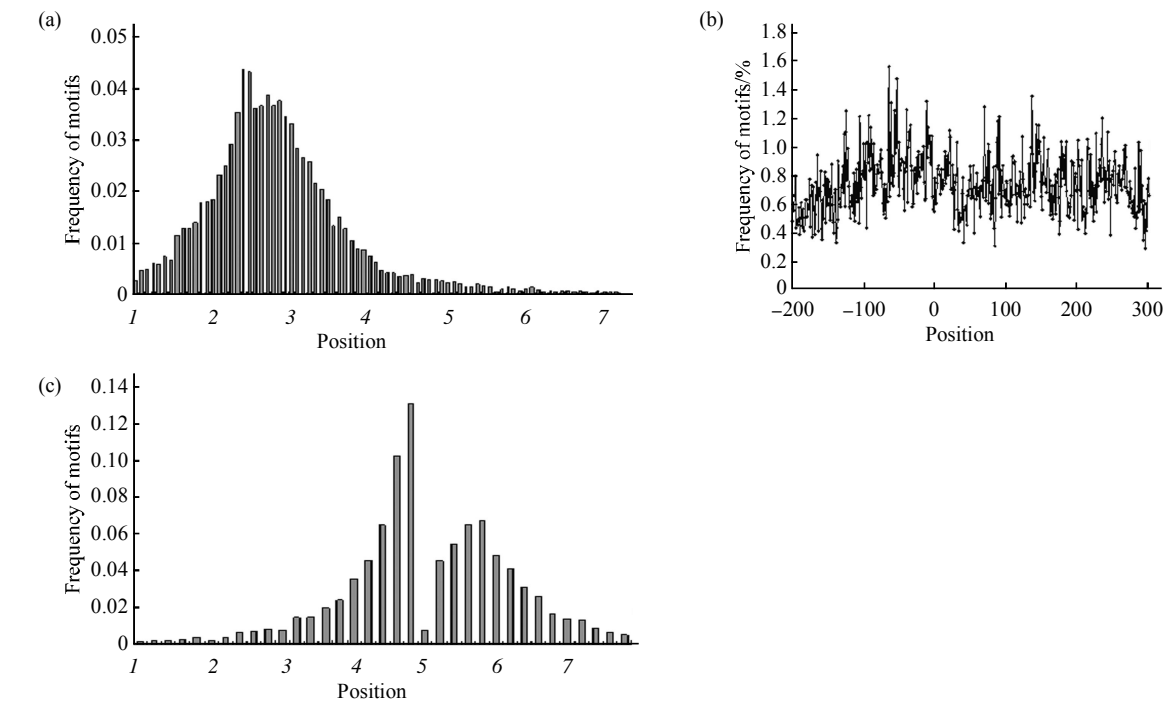


Fig. 2 Positions of motifs relative to TSS

(a) Positions of all extracted motifs relative to TSS. 1: -1 000~-950 bp; 2: -350~-300 bp; 3: 300~350 bp; 4: 950~1 000 bp; 5: 1 600~1 650 bp; 6: 2 250~2 300 bp; 7: 2 900~2 950 bp. (b) Positions of motifs relative to TSS in region -200~-300 bp. (c) Positions of motifs in U-I location. "0" denote the TSS. The windows is 50 bp in (a) and (c). 1: -1 000~-950 bp; 2: -750~-700 bp; 3: -500~-450 bp; 4: -250~-200 bp; 5: 0~50 bp; 6: 250~300 bp; 7: 500~550 bp.

这个位置共出现 3 546 次, 而 -54 bp 和 69 bp 位置出现模体的次数也比较高, 分别为 3 213 次和 2 390 次. 这些位置都距离基因 TSS 较近, 有利于转录因子的结合.

为考察内含子中转录调控元件的分布特征, 我们对 U-I 区域中的模体到 TSS 距离做进一步分析, 结果见图 2c. 有 95% 的模体位于基因上游 1 000 bp 到下游 750 bp 以内, 特别是内含子中的模体, 绝大部分位于 TSS 下游 500 bp 以内, 在 TSS 附近模体出现次数最多. 提示内含子中的模体与上游区域中模体之间发生作用的一个必要条件是模体距离 TSS 不能太远.

利用 $Z_d = (q_d - Qf_d) / \sqrt{Qf_d(1-f_d)}$ 来估计模体对 M1:M2 的两个模体之间的间隔距离(简称模体对距离)偏好 d 的程度. 其中 Q 仍是 M1:M2 在 HK 基因

中出现的次数, q_d 和 Qf_d 分别表示 M1:M2 在 HK 基因启动子中距离为 d 的实际次数和期望次数, f_d 为 M1:M2 在背景序列中距离为 d 的频率. 若在特定的距离 d 内, 当 $Z_d > 1.64$ 时 ($\log(Z_d) > 0.21$, $P < 0.05$), 认为 M1:M2 偏好在此距离范围内出现. Z_d 越大, 表明这种偏好性越显著. 图 3a 显示了过出现模体对的距离分布特征. 从图中可以看出, 大部分 (> 80%) 模体对距离小于 500 bp, 并且有超过 40% 的模体对距离小于 100 bp, 而只有不到 20% 的模体对距离大于 500 bp, 这表明相互作用模体之间距离较近. 每组组合调控模体对的距离分布和模体对的总体距离分布特征类似, 但也有一些差别. 例如, 模体对 NF1:SP1 和 SP1:HNF4 的距离小于 200 bp 的比率更高, 分别是 65.15% 和 63.01%.

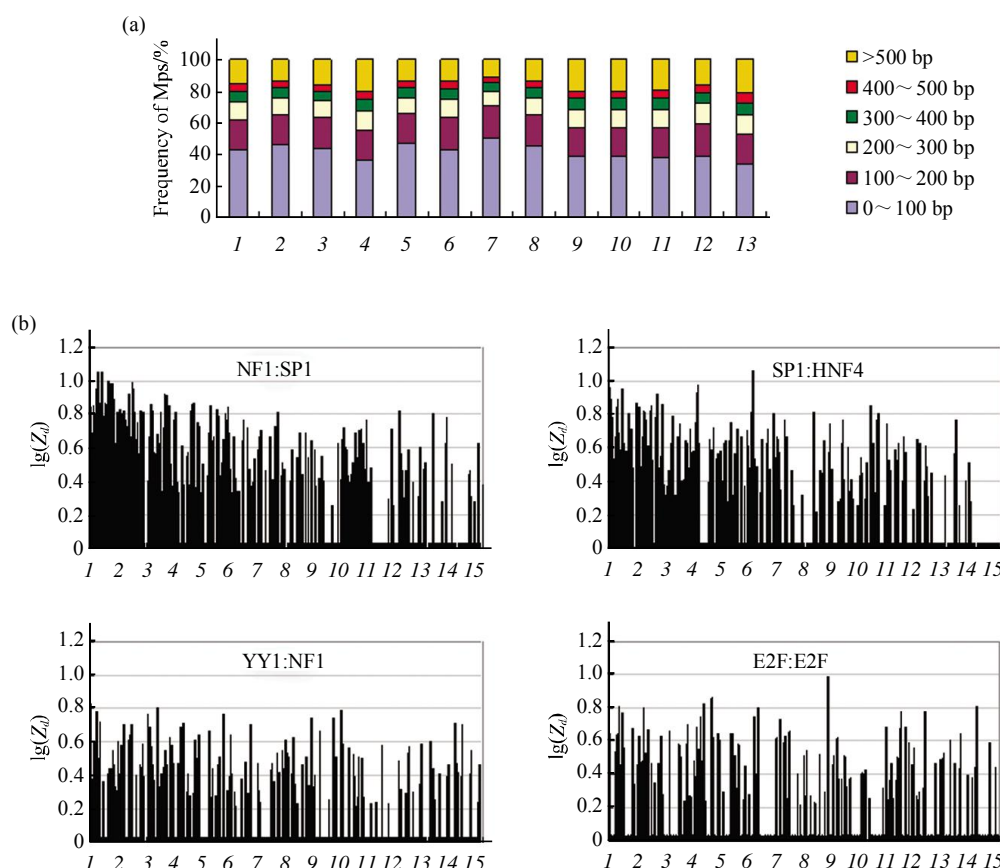


Fig. 3 Distance distributions of motif pairs

(a) Distance distributions of all extracted motif pairs. 1: All; 2: NF1:SP1; 3: SP1:HNF4; 4: YY1:NF1; 5: P53:SP1; 6: E2F:E2F; 7: AP2ALPHA:EGR1; 8: AP1:SP1; 9: STAT1:SP1; 10: NF_KAPPA:SP1; 11: AP2ALPHA:AREB6; 12: AP2:ER; 13: HNF4:COUP. (b) Distance preferences of motif pairs corresponding to 4 TF pairs in distance region 0~500 bp. 1: ~0 bp; 2: ~35 bp; 3: ~70 bp; 4: ~105 bp; 5: ~140 bp; 6: ~175 bp; 7: ~210 bp; 8: ~245 bp; 9: ~280 bp; 10: ~315 bp; 11: ~350 bp; 12: ~385 bp; 13: ~420 bp; 14: ~455 bp; 15: ~490 bp.

由于模体对距离主要位于 500 bp 以下, 我们特别对 4 组组合调控模体对在 500 bp 以内的距离偏好性进行分析(图 3b). 可以看出, NF1:SP1 和 SP1:HNF4 在 500 bp 以内的大部分距离具有偏好性 ($\lg(Z_d) > 0.21$). 而 NF1:SP1 在距离 66 bp 以下更为偏好. 进一步分析发现, 协同作用的模体对距离比拮抗作用模体对距离更近. 如, NF1:SP1 在距离 9 bp 和距离 15 bp 处取得了峰值, 而模体对 SP1:HNF4 和 YY1:NF1 分别在距离 183 bp 和 442 bp 处取得峰值. 如果 $Z_d > 3.10$ ($\lg(Z_d) > 0.49$), 此时认为调控模体对具有较强的距离偏好, 并称距离 d 为特征距离(characteristic distance), 结果共得到 1 030 个特征距离. 其中, 有 670 个距离小于 100 bp, 占所有特征距离的 65%. 这些特征表明, 协同作用模体需要较近的距离.

3 讨 论

真核基因的转录一般要有多个转录因子共同调控, 表现在 DNA 序列上即为这些转录因子的结合位点在序列上共同出现. 因此, 理解基因转录调控机理的一个重要工作是分析这些结合位点的组合调控规律. HK 基因具有高表达、共调控特征, 用统计方法更易于找出它们启动子序列中潜在的组合转录调控元件. 我们利用 3 种不同的统计量来寻找组合调控元件, 使得分析结果更可靠.

首先探测到一批出现过模体对, 从表 1 中模体对的碱基组成来看, 大部分模体对富含 CG. 而实际上, 在 HK 基因启动子序列中, AT 含量较高, 这说明探测到的模体是非随机的. 这些模体对大部分与已知具有相互作用的转录因子对相吻合, 是潜在的组合转录调控模体对, 表明转录因子在 HK 基因中更倾向于结合 CG-rich 模体. 从模体对所对应转录因子对之间的相互作用来看, 超过 3/4 的模体对之间是转录协同作用, 具有拮抗作用的模体对只有不到 1/4. 已知协同作用可以提高基因的转录速率^[23], 这些结果表明抽提出的组合调控模体对以及它们转录因子之间的组合调控关系与 HK 基因的高表达相关. 当然, 提取出的部分模体对还未找到相应的实验对照(限于篇幅, 未列出), 如模体对 CGGCGC:GGCCTC (P_{oc} 和 P_{dis} 值分别为 1.15×10^{-8} 和 1.22×10^{-7}), 可能是因为目前关于组合调控的实验分析结果毕竟还比较少. 不过本文的结果可以为实验分析提供一些参考.

内含子是否参与转录调控也是我们研究的一个

重要目标. 本文关于模体对的出现区域分析表明, 除 U-U 区域富含模体对外, 模体对在 U-I 区域也呈现了明显的区域偏好性, 表明 HK 基因中的组合调控关系更容易发生在上游区域, 以及上游和内含子之间. 已知上游区域是转录调控的核心区域, 因此模体对偏好上游启动子区域容易理解. 而 U-I 区域中富含模体对, 进一步支持了内含子参与基因转录调控的假设. 本文还发现协同作用的模体对更偏好 U-I 区域, 并且相对拮抗作用模体对, 协同作用模体对出现次数较多, 可以推测内含子与上游调控元件之间的组合调控是导致 HK 基因高表达的一个关键原因. 从模体对在序列分布上的特征来看, 模体对中模体在基因 TSS 附近较为集中, 有利于调控基因的转录过程, 特别是内含子中模体多数位于 TSS 下游 500 bp 以内. HK 基因第一外显子普遍较短, 这表明内含子中模体与上游模体间隔不远; 另一方面, 模体对的两个模体之间距离较近, 60%左右距离在 200 bp 以内, 并且, 65%的模体对特征距离在 100 bp 以内. 由于相互作用的两个转录因子结合位点不能太远, 因此这种距离特征也使得上游和内含子之间的协同作用得以进行. 此外, 区域偏好性分析结果也表明, 上游和外显子、外显子和内含子之间也存在转录相互作用, 主要表现为拮抗作用, 出现次数较少, 在 HK 基因转录调控过程中并不占主要作用.

我们前期对人组织特异性基因中的组合调控关系进行了细致分析^[15], 发现 HK 基因和组织特异性基因中的组合调控模体对特征具有明显不同, 主要表现在: a. HK 基因中的组合调控模体对主要富含 CG, 而组织特异性基因中模体对则主要富含 AT, 表明两者之间转录因子结合序列不同; b. 除 U-U 区域外, HK 基因中组合调控模体对主要偏好 U-I 区域, 而组织特异性基因中模体对更倾向于 I-I 区域, 表明内含子在两者中参与组合调控方式不同. 在前者为与上游区域协同作用, 而后者中则是内含子中的调控元件相互作用. 这些结果都表明了 HK 基因特有的组合调控关系.

本文不仅获得了 HK 基因中组合转录调控元件的重要特征, 并且再次印证了内含子可以参与基因的转录调控, 同时推测了内含子参与转录调控的主要方式, 有助于理解人基因转录调控的机制和内含子的功能. 当然, 我们的分析仅仅是针对小部分基因而言的, 因为 HK 基因具有共调控特性. 以后我们将扩大基因范围, 从全基因组水平来研究基因的

转录调控, 期望更深入地理解基因转录调控的机理及内含子在基因转录调控过程中的作用. 此外, 本文虽然仅以两个模体构成的模体对来分析组合调控关系, 但是在序列中由一些模体对的交叠可推出更多模体的组合调控关系^[24], 即组合调控模块(module), 这将是我們下一步研究的重点.

参 考 文 献

- [1] Hu Z, Gallo S M. Identification of interacting transcription factors regulating tissue gene expression in human. *BMC Genomics*, 2010, **11**: 49
- [2] Cheng C, Yan K K, Hwang W, *et al.* Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput Biol*, 2011, **7**(11): e1002190
- [3] Chorev M, Carmel L. The function of introns. *Front Genet*, 2012, **3**: 55
- [4] Morello L, Breviaro D. Plant spliceosomal introns: not only cut and paste. *Curr Genomics*, 2008, **9**(4): 227-238
- [5] Ying S, Kojima T, Kawada A, *et al.* An intronic enhancer driven by NF- κ B contributes to transcriptional regulation of peptidylarginine deiminase type I gene in human keratinocytes. *J Invest Dermatol*, 2010, **130**(11): 2543-2552
- [6] 张 静, 石秀凡, 杨恒芬. 酵母内含子在基因序列中的分布对基因转录效率的影响. *生物化学与生物物理进展*, 2003, **30**(6): 945-949
Zhang J, Shi X F, Yang H F. *Prog Biochem Biophys*, 2003, **30**(6): 945-949
- [7] Hu J, Li H, Zhang J. Analysis of transcriptional synergy between upstream regions and introns in ribosomal protein genes of yeast. *Comput Biol Chem*, 2010, **34**(2): 106-114
- [8] Wang E T, Sandberg R, Luo S, *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature*, 2008, **456**(7221): 470-476
- [9] Coulon V, Chebli K, Cavelier P, *et al.* A novel mouse c-fos intronic promoter that responds to CREB and AP-1 is developmentally regulated *in vivo*. *PLoS One*, 2010, **5**(6): e11235
- [10] Zolova O E, Wight P A. YY1 negatively regulates mouse myelin proteolipid protein (Plp1) gene expression in oligodendroglial cells. *Asn Neuro*, 2011, **3**(4): pii e00067
- [11] Foti D M, Reichardt J K. YY1 binding within the human HSD3B2 gene intron 1 is required for maximal basal promoter activity: identification of YY1 as the 3beta1-A factor. *J Mol Endocrinol*, 2004, **33**(1): 99-119
- [12] Shakes L A, Du H, Wolf H M, *et al.* Using BAC transgenesis in zebrafish to identify regulatory sequences of the amyloid precursor protein gene in humans. *BMC Genomics*, 2012, **13**: 451
- [13] Li H, Chen D, Zhang J. Analysis of intron sequence features associated with transcriptional regulation in human genes. *PLoS One*, 2012, **7**(10): e46784
- [14] Yu X, Lin J, Masuda T, *et al.* Genome-wide prediction and characterization of interactions between transcription factors in *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 2006, **34**(3): 917-927
- [15] Li H, Chen D, Zhang J. Statistical analysis of combinatorial transcriptional regulatory motifs in human intron-containing promoter sequences. *Comput Biol Chem*, 2013, **43**(2): 35-45
- [16] Schbath S. An efficient statistic to detect over and under-presented words in DNA sequences. *J Comp Biol*, 1997, **4**(2): 189-192
- [17] Remenyi A, Scholer H R, Wilmanns M. Combinatorial control of gene expression. *Nat Struct Mol Biol*, 2004, **11**(9): 812-815
- [18] van Helden J, André B, Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol*, 1998, **281**(5): 827-842
- [19] Wingender E, Dietze P, Karas H, *et al.* TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res*, 1996, **24**(1): 238-241
- [20] Yu X, Lin J, Zack D J, *et al.* Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res*, 2006, **34**(17): 4925-4936
- [21] Marianne A, Paul K. Functional characterization of transcriptional regulatory elements in the upstream region and intron 1 of the human S6 ribosomal protein gene. *J Biochem*, 1998, **336**(2): 327-335
- [22] Darne C H, Morel L, Claessens F, *et al.* Ubiquitous transcription factors NF1 and Sp1 are involved in the androgen activation of the mouse vas deferens protein promoter. *Mol Cellular Endocrinol*, 1997, **132**(1-2): 13-23
- [23] 薛 文, 王 进, 黄启来, 等. 真核基因转激活的多位点协同调控. *生物化学与生物物理进展*, 2002, **29**(4): 510-513
Xue W, Wang J, Huang Q L, *et al.* *Prog Biochem Biophys*, 2002, **29**(4): 510-513
- [24] Kato M, Hata N, Banerjee N, *et al.* Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biology*, 2004, **5**(8): R56

Analysis of Combinatorial Transcriptional Regulatory Elements in Promoters of Human Housekeeping Genes*

LI Hui-Min^{1,4)**}, SHI Yu²⁾, YANG Zhi-Gang³⁾, JIANG Shao-Ping^{1)**}

¹⁾ School of Mathematics and Computer Science, Yunnan Minzu University, Kunming 650500, China);

²⁾ Department of Mathematics and Statistics, Zhengzhou Normal University, Zhengzhou 450044, China;

³⁾ College of Computer Science, Sichuan University, Chengdu 610065, China;

⁴⁾ Laboratory for Conservation and Utilization of Bio-resources, Yunnan University, Kunming 650091, China)

Abstract A lot of researches suggest that there are transcriptional regulatory elements in the first introns of genes. However, the ways about introns participating in transcriptional regulation are not still understood. In this paper, we extracted combinatorial transcriptional regulatory motifs in regions from upstream sequences to the first introns in human housekeeping (HK) genes by statistical methods, analyzed the characteristics of distances and locations of motifs, and discussed the potential and ways of introns involving in transcriptional regulation. 960 potential transcriptional regulatory motif pairs are extracted in HK genes, and 57% of them are in accordance with 12 known interacting transcription factor pairs. The location preferences of the over-represented motif pairs have been studied. The results show that about 80% and 90% of the motif pairs in HK genes occur in upstream-upstream and upstream-intron regions, respectively. This results further support the conjecture of the introns could participate in transcriptional regulation of genes, and it was speculated that there are transcriptional synergism between introns and upstream promoter. Extracting motifs are mainly focus on vicinity of transcriptional start sites (TSSs), and the space distance between two motifs in one motif pairs is short. The preference analyses on distance of the motif pairs show that 60% of the motif pairs in HK genes are shorter than 200 bp, and 65% of characteristic distances are within 100 bp. This is conducive for synergism of transcription factors. These results will be helpful for understanding the mechanisms of the transcriptional regulation for HK genes in human.

Key words human housekeeping genes, combinatorial transcriptional regulation, motif pairs, introns, location preference, distance distribution

DOI: 10.3724/SP.J.1206.2014.00069

*This work was supported by grants from The National Natural Science Foundation of China (31360277, 31460297, 11461083), Yunnan Province Education Department Scientific Research Fund Project (2013Z042), The Natural Science Foundation of Henan Education Department (14A110020) and The Natural Science Foundation of Yunnan University(2012CG016).

**Corresponding author.

LI Hui-Min. Tel: 13888183685, E-mail: lihuimin_1980@126.com

JIANG Shao-Ping. Tel: 13888096496, E-mail: jiang2005124@163.com

Received: June 3, 2014 Accepted: August 11, 2014