酵母基因内含子中二聚体寡核苷酸转录调控位点的统计分析*

胡 俊[1,2] 张 静[1]**

（[1]云南大学应用统计中心，昆明 650091；[2]云南农业大学基础信息工程学院，昆明 650201）

摘要 对高效和低效转录酵母基因内含子序列中寡核苷酸的出现频率进行对照分析，结果显示高效和低效内含子序列的结构有差异，而且高效转录内含子序列含有较多潜在的转录因子结合位点. 观察实验获得的转录调控位点，发现许多调控位点不是相邻接的寡核苷酸，而是由一对保守寡核苷酸构成，这对寡核苷酸被一段长度固定的非保守区域间隔开. 于是对此形式的二聚体寡核苷酸（dyad）在高效和低效内含子序列中出现的频率进行统计比较分析，抽提出在高效内含子组出现的频率显著高于在低效内含子组出现频率的二聚体寡核苷酸，分析这些二聚体寡核苷酸在两组内含子序列中的分布特征，并对照实验结果，这些二聚体寡核苷酸可能与基因转录的正调控有关.

关键词 酵母，内含子，二聚体寡核苷酸，频率分析，转录调控位点

学科分类号 Q61

真核生物基因组含有大量的非编码区，基因表达的信息主要蕴藏在这些非编码区中. 对非编码序列的信息结构和生物功能的研究已成为后基因组时代的主要任务之一. 内含子作为非编码序列的重要成员，已有大量实验研究表明某些内含子具有调控基因转录的功能，起着启动子（promoter）和增强子（enhancer）的作用[1,2]，有的则具有抑制子（repressor）的功能[3]. 张静等[4,5]对高效和低效转录内含子序列统计分析的结果显示，高效和低效转录内含子序列的结构有差异，高效转录内含子序列中含有较多潜在的转录因子结合位点，而且从内含子在基因中的实际位置看，高效转录基因中有较多的内含子接近基因的5′端，有些甚至位于5′非翻译区，这些结果提示，高效转录内含子序列对基因转录的调控可能与基因上游的转录调控有关联[6]. 要证实这些结论并揭示内含子的调控转录机制，首先需要探明内含子中可能的转录调控位点. 由于转录因子的DNA结合结构域与互相邻接的特异性核苷酸形成直接的接触，多数调控位点都是邻接核苷酸形成的短链（6～8 bp）[7]，以前的工作中，我们在高效转录内含子序列中探测到了这种形式的潜在转录因子结合位点. 但观察数据库SCPD和TRANSFAC收集的转录因子结合位点后，发现除了邻接核苷酸形成的结合位点，许多转录因子如Gal4p的结合位点CGGnnnnnnnnnnnnCCG由一对保守的三核苷酸构成，这对三核苷酸被一段长度固定的间隔区（spacer）分隔开，原因是这些转录因子常以二聚体形式和两个半结合位点结合起作用，因此它们的结合位点形成二聚体寡核苷酸（dyad）结构，这种形式的结合位点对酵母中的锌双核聚合体蛋白具有普遍性. 为了对内含子所蕴含的转录调控元件的结构信息有一个全面、充分的认识，我们分析了二聚体寡核苷酸在高效和低效转录内含子序列中的使用情况.

本文对高效和低效转录内含子序列中二聚体寡核苷酸的出现频率进行统计比较分析，提取到一批二聚体寡核苷酸，它们在高效内含子组中出现的频率显著高于在低效内含子组中出现的频率，其中一些与实验获得的转录因子结合位点相同或相似. 同时在高效内含子序列中用期望频率法[7,8]进行分析，探测到一些过表达的二聚体寡核苷酸，结果表明，提取到的二聚体寡核苷酸绝大多数与期望频率法抽提的二聚体寡核苷酸相同.

1 样本和方法

1.1 样本

1.1.1 从YIDB数据库中选取转录频率较高（>30 mRNAs/h）的基因，然后从EMBL数据库选取它们的内含子序列，共77个，简称高效内含子

* 国家自然科学基金资助项目（30360027）和云南大学理（工）科校级科研项目资助（2002T009XX）.

** 通讯联系人.

Tel：0871-6541419，E-mail：zhangjing@ynu.edu.cn

收稿日期：2003-11-05，接受日期：2003-12-24

组. 再用同样的方法选出转录频率较低（≤10 mRNAs/h）的内含子序列（77 个），简称低效内含子组. 见文献［4］的表 1 和表 2.

1.1.2 随机选取一些酵母基因序列片段作为分析的对照序列（共计 159 个），为了保证用于对照的寡核苷酸数目的随机性，提高分析结果的可靠性，每个序列片段的长度都取为 800 bp. 这些序列包含外显子和内含子，形成外显子-内含子-外显子的核苷酸序列片段.

1.2 方法

1.2.1 二聚体寡核苷酸：二聚体寡核苷酸（dyad）是指一对长度相同的保守寡核苷酸，这对寡核苷酸被一段数目固定的可变碱基间隔开，其一般形式为 $D = w_1 \cdot n_s \cdot w_2$. 这里 w_1 和 w_2 分别是二聚体寡核苷酸的第一和第二个寡核苷酸，其长度一般为 3 bp，n_s 是长度为 s 的任意核苷酸，s 是间隔区的长度，即 w_1 和 w_2 之间的碱基数目[7]，SCPD 数据库收录的转录因子结合位点的间隔区长度多数在 0 ~ 11 bp 之间，为了对照已知的转录因子结合位点，本文中的 $s \in \{0, 1, \Lambda, 11\}$；当 $s = 0$ 时二聚体寡核苷酸即为六核苷酸.

1.2.2 方法 1（比较方法）：用 $n_1(\mathrm{D})$ 和 $n_2(\mathrm{D})$ 分别表示二聚体寡核苷酸 D 在高效内含子组和低效内含子组中出现的次数（$s = 0, 1, \ldots, 11$ 的总和），T_1 和 T_2 表示所有二聚体寡核苷酸在高效和低效内含子组中出现的总次数：

$$T_1 = \sum_{s=0}^{11} \sum_{i=1}^{77} [2 \times (L_{1i} - 2k - s + 1)]$$

$$T_2 = \sum_{s=0}^{11} \sum_{i=1}^{77} [2 \times (L_{2i} - 2k - s + 1)]$$

这里因子 2 表示沿着正链和反转互补链的 5′→3′方向计算总数；L_{1i} 和 L_{2i} 分别是高效和低效内含子组中第 i 个内含子的长度，k 是寡核苷酸 w_1 或 w_2 的长度，本文中 $k = 3$；s 是间隔区的长度. 二聚体寡核苷酸 D 在高效内含子组中出现的频率 $f_1(\mathrm{D}) = n_1(\mathrm{D})/T_1$，在低效内含子组出现的频率 $f_2(\mathrm{D}) = n_2(\mathrm{D})/T_2$.

本文的目的是探测内含子序列中与转录正调控有关的二聚体寡核苷酸，对内含子序列结构特征的研究表明，它们应该在高效内含子组中出现的频率较大，而在低效内含子组出现的频率较小[4]. 为了统计二聚体寡核苷酸在两组内含子中使用频率的差异，我们检验假设 H_0：$f_1(\mathrm{D}) \leqslant f_2(\mathrm{D})$，备择假设 H_1：$f_1(\mathrm{D}) > f_2(\mathrm{D})$.

二聚体寡核苷酸 D 在高效和低效内含子组中出现频率差异的标准差表示为

$$S = \sqrt{\frac{n_1(\mathrm{D}) + n_2(\mathrm{D})}{T_1 + T_2}\left(1 - \frac{n_1(\mathrm{D}) + n_2(\mathrm{D})}{T_1 + T_2}\right)\left(\frac{1}{T_1} + \frac{1}{T_2}\right)}$$

构造统计量

$$u = \begin{cases} \dfrac{n_1(\mathrm{D})/T_1 - n_2(\mathrm{D})/T_2}{S}, \\ \text{如果 } n_1(\mathrm{D}) \geqslant 30,\ n_2(\mathrm{D}) \geqslant 30 \\ \dfrac{n_1(\mathrm{D})/T_1 - n_2(\mathrm{D})/T_2 - 0.5/T_2}{S}, \\ \text{如果 } n_1(\mathrm{D}) \geqslant 30,\ n_2(\mathrm{D}) < 30 \\ \dfrac{n_1(\mathrm{D})/T_1 - n_2(\mathrm{D})/T_2 - 0.5/T_1 - 0.5/T_2}{S}, \\ \text{如果 } n_1(\mathrm{D}) < 30,\ n_2(\mathrm{D}) < 30 \end{cases}$$

取显著水平 $\alpha = 0.001$，则当 $u > 3.10$ 时拒绝 H_0，接受备择假设 H_1，即认为二聚体寡核苷酸 D 在高效内含子组出现的频率显著高于在低效内含子组中出现的频率.

1.2.3 方法 2（期望频率法）：van Helden 等曾用此方法分析酵母基因上游的转录调控位点，本文中我们同时也用此方法对二聚体寡核苷酸进行分析，以考察上述比较方法的有效性.

用 $n_{\mathrm{occ}}(\mathrm{w})$ 表示寡核苷酸 w 在对照序列中出现的次数，则 w 在对照序列中出现的频率为

$$f(\mathrm{w}) = n_{\mathrm{occ}}(\mathrm{w}) / \sum\nolimits_j [n_{\mathrm{occ}}(\mathrm{w}_j)]$$

而二聚体寡核苷酸 D 在高效内含子组中出现的期望频率计算为

$$f_{\exp}(\mathrm{D}) = f(\mathrm{w}_1) \times f(\mathrm{w}_2)$$

则二聚体寡核苷酸 D 以期望频率 $f_{\exp}(\mathrm{D})$ 在高效内含子组至少出现 n 次的概率为

$$P(\mathrm{D} \geqslant n) = \sum_{i=n}^{T_1} C_{T_1}^{i} [f_{\exp}(\mathrm{D})]^i \times [1 - f_{\exp}(\mathrm{D})]^{T_1 - i}$$

设定阈值 L，将每个二聚体寡核苷酸的 $P(\mathrm{D} \geqslant n)$（简称 P 值）与 L 比较，$P < L$ 的二聚体寡核苷酸在高效内含子组中出现的频率显著高于在对照序列中出现的频率，是过表达的二聚体寡核苷酸[7]，因此有效抽提二聚体寡核苷酸的关键是阈值 L，这个阈值依赖于 w_1、w_2 的长度 k 及间隔区的长度 s. 因为共有 $4^3 = 64$ 种三核苷酸形式，而 64 种三核苷酸形成 $64^2 = 4\,096$ 种二聚体寡核苷酸，减去回文二聚体寡核苷酸数目[8]，还有 2 080 种相异二聚体寡核苷

酸，s 取 12 个值. 于是设定阈值 $L = 1/(2\ 080 \times 12) = 4.00 \times 10^{-5}$，提取 $P < L$ 的二聚体寡核苷酸，共有 430 种二聚体寡核苷酸被提取.（由于篇幅限制未列出提取结果）

2 结 果

抽取 $u > 3.10$ 的二聚体寡核苷酸，有 76 种二聚体寡核苷酸被抽提出，表 1 中列出了这些二聚体寡核苷酸在两组内含子中出现的次数和 u 值，粗体字表示的二聚体寡核苷酸 P 值小于 L，表 2 列出了与 SCPD 数据库中的转录因子结合位点相同或相似的二聚体寡核苷酸. 根据提取的二聚体寡核苷酸的相似性和碱基成分对它们进行分类，第 1 类全由 A、T 组成，它们的 u 值普遍较高，在高效内含子组中出现次数较多，其中的 TAT $\{n\}_0$AAT 即 TATAAT 的划线部分与 TATAA 相同，而 TATAA 是上游核心启动子的 TATA 盒（TATA-box），TTT $\{n\}_0$AAA是 TBP 的结合位点TTTAAATAAGT 的一部分. 第 2 类二聚体寡核苷酸的共有序列是 SAW $\{n\}_s$ WWT（AWW $\{n\}_s$WTS），（S 表示 C 或 G，W 表示 A 或 T），当间隔区长度 $s = 0$ 时，这类二聚体寡核苷酸含有较多 C\GATA 及 G\CAAA、GAAT，而 GATA 是转录因子结合位点的核心元件. 此外 CAAAAT（ATTTTG）是实验证实的转录调控元件，具有增强子的活性[5]，对照 SCPD 数据库中的转录因子结合位点，当 $s = 0$ 时 GAAAAT 与转录因子 ECB 结合位点的共有序列 GGAAAAD（D 表示 A、G 或 T）相似，$s = 1$ 时，ATT $\{n\}_1$TTC 是 YY*n*ATTGTTY 的一部分（Y 表示 C 或 T），而 YY*n*ATTGTTY 是 ROX1 结合位点的共有序列. 第 3 类的共有序列是 AWT $\{n\}_s$SAA（TTS $\{n\}_s$AWT），其中 ATT $\{n\}_1$GAA 与 STE12 的结合位点 TGTTTTC-ATTTGAAACA 部分相同. 这三类二聚体寡核苷酸的 P 值都小于 L. 第 4 类二聚体寡核苷酸的一个三核苷酸为 CC\GA\T（A\TC\GG），另一个三核苷酸全由 A 或 T 组成. 当 $s = 2$ 时 CCA $\{n\}_2$AAT 和 CCA $\{n\}_2$TAT是转录因子 MIG1 结合位点的共有序列 CCCCR*nn*WWWWW（R 表示 A 或 G）的一部分；$s = 4$ 时 CCA $\{n\}_4$TAT 与 CCCCR*nn*WWWWW 相似；$s = 3$ 时 CGT $\{n\}_3$AAA 与 MATa1pha2 结合位点的共有序列 CRTGTWWWW 相同. 第 5 类二聚体寡核苷酸的共有序列是 WSW $\{n\}_s$WWT（AWW $\{n\}_s$WSW），其中的六核苷酸 TGA $\{n\}_0$ATT，即 TGAATT 与 GCN4 结合位点的共有序列 TGANTN 相同，这里 N 表示任意碱基. 第 6 类二聚体寡核苷酸中的CTC $\{n\}_2$AAT 划线部分与转录因子 SCB 结合位点的共有序列 CNCGAAA 相同. 第 7 类二聚体寡核苷酸都含有三核苷酸 GAT（ATC），它们的 P 值都小于 L，ATC $\{n\}_3$AAA、ATC $\{n\}_3$AAT 与 NBF 结合位点的共有序列 ATGYGRAWW 相似.

另外有一些与数据库中的调控位点相同或相似的二聚体寡核苷酸 GAA $\{n\}_s$CCA（TGG $\{n\}_s$TTC），TGA $\{n\}_s$TTT（AAA $\{n\}_s$TCA），TGA $\{n\}_s$TTA（TAA $\{n\}_s$TCA），TGA $\{n\}_s$ATG（CAT $\{n\}_s$TCA），GGA $\{n\}_s$AAT（ATT $\{n\}_s$TCC），表中未列出，它们的 u 值也比较大，分别是 2.49，2.44，2.35，2.81，2.55，在转录效率较高的内含子组出现的频率也比较显著地高于在转录效率较低的内含子组中出现的频率.

表 1 中的二聚体寡核苷酸在转录效率较高的内含子组出现的频率显著高于在转录效率较低的内含子组中出现的频率，将这些二聚体寡核苷酸放到两组内含子序列中进行分析，为了保持研究的系统性，我们仍以高效内含子组中的 ykl180w 的内含子序列为例. 从图 1 可以看出，表 1 中一些二聚体寡核苷酸的 w_1、w_2 分别会与另一个二聚体寡核苷酸的 w_1 或 w_2 在序列中重叠后形成长二聚体寡核苷酸（二聚体寡核苷酸的长度指 w_1 和 w_2 的长度之和），如AATTATTCA，有一些 w_1 和 w_2 则与相邻二聚体寡核苷酸的间隔区重叠. 它们相互连接形成长寡核苷酸片段. 但是在两组内含子序列中抽提到的二聚体寡核苷酸的分布情况存在差异，它们在高效内含子序列中形成较多长二聚体寡核苷酸和长寡核苷酸片段，其中含有实验证实了的转录因子结合位点，而低效内含子序列中形成的长寡核苷酸不多. 在高效内含子序列中表 1 中的二聚体寡核苷酸平均分布密度为 72.4%，在低效内含子序列中为 59%，从形成的长寡核苷酸片段看，高效内含子序列中长寡核苷酸的平均宽度为 9.23，低效内含子序列中为 7.38.

Table 1 The list of dyads with $u > 3.10$

Dyad		n_1	n_2	u
AAA $\{n\}_s$ATT	（**AAT $\{n\}_s$TTT**）	1 373	260	7.19
AAA $\{n\}_s$TAT	（**ATA $\{n\}_s$TTT**）	1 117	215	6.31
AAA $\{n\}_s$TTT		1 750	254	11.4
AAT $\{n\}_s$AAA	（**TTT $\{n\}_s$ATT**）	1 384	324	4.38
AAT $\{n\}_s$AAT	（**ATT $\{n\}_s$ATT**）	1 001	194	5.9
AAT $\{n\}_s$ATT		1 024	196	6.1
AAT $\{n\}_s$TAT	（**ATA $\{n\}_s$ATT**）	816	176	4.25
ATA $\{n\}_s$AAT	（**ATT $\{n\}_s$TAT**）	838	175	4.65
ATT $\{n\}_s$AAA	（**TTT $\{n\}_s$AAT**）	1 196	276	4.27
ATT $\{n\}_s$AAT		978	214	4.49
TTT $\{n\}_s$AAA		1 390	266	7.11
ATT $\{n\}_s$ATA	（**TAT $\{n\}_s$AAT**）	875	209	3.25
Consensus（1）	WWW $\{n\}_s$WWW	（WWW $\{n\}_s$WWW）		
GAA $\{n\}_s$TTT	（**AAA $\{n\}_s$TTC**）	850	176	4.77
GAT $\{n\}_s$ATT	（**AAT $\{n\}_s$ATC**）	460	91	3.85
CAT $\{n\}_s$ATT	（**AAT $\{n\}_s$ATG**）	552	123	3.21
CAA $\{n\}_s$AAT	（**ATT $\{n\}_s$TTG**）	603	132	3.52
GAA $\{n\}_s$AAT	（**ATT $\{n\}_s$TTC**）	704	150	4.07
GAA $\{n\}_s$ATT	（**AAT $\{n\}_s$TTC**）	642	129	4.41
CAT $\{n\}_s$AAT	（**ATT $\{n\}_s$ATG**）	527	115	3.32
GAA $\{n\}_s$TAT	（**ATA $\{n\}_s$TTC**）	541	119	3.29
Consensus（2）	SAW $\{n\}_s$WWT	（AWW $\{n\}_s$WTS）		
AAT $\{n\}_s$CAA	（**TTG $\{n\}_s$ATT**）	590	122	3.98
AAT $\{n\}_s$GAA	（**TTC $\{n\}_s$ATT**）	719	156	3.93
ATT $\{n\}_s$GAA	（**TTC $\{n\}_s$AAT**）	622	130	4
Consensus（3）	AWT $\{n\}_s$SAA	（TTS $\{n\}_s$AWT）		
CCA $\{n\}_s$AAT	（ATT $\{n\}_s$TGG）	320	63	3.24
CGT $\{n\}_s$AAA	（TTT $\{n\}_s$ACG）	275	48	3.65
CCA $\{n\}_s$TAT	（ATA $\{n\}_s$TGG）	284	53	3.35
Consensus（4）	CSW $\{n\}_s$WAW	（WTW $\{n\}_s$WSG）		
TGA $\{n\}_s$AAT	（**ATT $\{n\}_s$TCA**）	623	127	4.22
TGA $\{n\}_s$ATT	（**AAT $\{n\}_s$TCA**）	570	116	4.05
TGA $\{n\}_s$TAT	（**ATA $\{n\}_s$TCA**）	491	108	3.14
AGA $\{n\}_s$ATT	（AAT $\{n\}_s$TCT）	463	79	4.88
TCT $\{n\}_s$AAT	（ATT $\{n\}_s$AGA）	469	79	5
TCT $\{n\}_s$ATT	（AAT $\{n\}_s$AGA）	482	106	3.11
TCA $\{n\}_s$AAT	（**ATT $\{n\}_s$TGA**）	533	118	3.21
ACA $\{n\}_s$ATT	（**AAT $\{n\}_s$TGT**）	517	108	3.65
Consensus（5）	WSW $\{n\}_s$WWT	（AWW $\{n\}_s$WSW）		
GTC $\{n\}_s$TTT	（AAA $\{n\}_s$GAC）	242	45	3.11
GAC $\{n\}_s$ATT	（AAT $\{n\}_s$GTC）	204	32	3.59
CAC $\{n\}_s$TAT	（**ATA $\{n\}_s$GTG**）	246	46	3.11
CTC $\{n\}_s$AAT	（**ATT $\{n\}_s$GAG**）	289	54	3.37
Consensus（6）	SWC $\{n\}_s$WWT	（AWW $\{n\}_s$GWS）		
TTT $\{n\}_s$GAT	（**ATC $\{n\}_s$AAA**）	570	127	3.26
ATT $\{n\}_s$GAT	（**ATC $\{n\}_s$AAT**）	416	85	3.43
Consensus（7）	WTT $\{n\}_s$GAT	（ATC $\{n\}_s$AAW）		

The dyads in brackets denote the reverse complements. n_1 and n_2 denotes the occurrence numbers of the dyad in the set of introns with higher transcription frequencies and the set of introns with lower transcription frequencies respectively. The dyads with $P < L$ are boldfaced.

Fig. 1 The intron sequence of ykl180w

The dyads in Table 1 are capitalized. Most of them form wider oligonucleotides and wider spaced dyads. All underlined dyads (each of them is indicated by a pair of number) are similar to the known binding sites. Boldface parts are just the regulatory elements revealed by experimental analysis.

Table 2 Dyads matching the known binding site in SCPD

Transcription factor	Known sites	Dyad	
TBP	TATAA	**TATAAT**	**ATTATA**
STE12	TGTTTTC ATTTGAAACA	ATT $\{n\}_1$GAA	TTC $\{n\}_1$AAT
ROX1	YYn ATTGTTY	ATT $\{n\}_1$TTC	GAA $\{n\}_1$AAT
MIG1	CC CCRnnWWWWW	CCA $\{n\}_2$AAT CCA $\{n\}_2$TAT	ATT $\{n\}_2$TGG ATA $\{n\}_2$TGG
	CC CCRnnWWWWW	CCA $\{n\}_3$AAT	ATT $\{n\}_3$TGG
	CC CCRnnWWWWW	CCA $\{n\}_4$TAT	ATA $\{n\}_4$TGG
ECB	G GAAAAD	GAAAAT	ATTTTC
MATalpha2	CRTGTWWWW	CGT $\{n\}_2$AAA	TTT $\{n\}_2$ACG
		CAT $\{n\}_2$ATT CAT $\{n\}_2$AAT	AAT $\{n\}_2$ATG ATT $\{n\}_2$ATG
	CRTGTWWWW	CGT $\{n\}_3$AAA	TTT $\{n\}_3$ACG
		CAT $\{n\}_3$ATT CAT $\{n\}_3$AAT	AAT $\{n\}_3$ATG ATT $\{n\}_3$ATG
SCB	CNCGAAA	**CTC*nn*AAT**	**ATT*nn*GAG**
GCN4	TGANTN	TGAATT	AATTCA

This table shows the motifs that match the known binding sites. In the column of known sites, the segments that match dyads in Table 1 are underlined. The last column is the reverse complements.

3 讨 论

从抽取的结果看，表 1 中的二聚体寡核苷酸在高效内含子组中使用频率比较高，间隔区长度 $s=0$ 时，二聚体寡核苷酸中含有一定数量的 TATA 元件和 GATA，它们是启动和增强基因转录的关键因素，实验研究表明 TATA 盒不但在基因的上游起作用，在下游也有调控转录的作用[9]，内含子中的 GATA 也可能成为转录因子的作用位点[1]. 当 s 取其他值时表 1 中的二聚体寡核苷酸有些与实验获得的转录调控位点相同或相似. 考察抽取的二聚体寡核苷酸在高效内含子序列中的分布，由于高效内含子的序列普遍较长，而且这些二聚体寡核苷酸在高效内含子组中出现次数多，因此重叠连接形成较多的长二聚体寡核苷酸和长寡核苷酸，其中含有实验获得的转录调控位点. 这些结果表明，表 1 中的二聚体寡核苷酸很可能与转录正调控有关.

将表 1 中的二聚体寡核苷酸和用期望频率法抽取的结果进行对比，表 1 的二聚体寡核苷酸有 60 个用期望频率法也抽取到了，它们在高效内含子组中是过表达的. 其余 16 个在高效内含子中出现的平均次数为 342 次，明显少于过表达二聚体寡核苷酸出现的平均次数 741 次，但它们在高效内含子组中出现频率显著高于低效内含子组. 本文的结果说明方法 1 和方法 2 都能识别高效内含子组中的一些特殊模体（motif），但是两种方法提取的二聚体寡核苷酸也存在差异，例如期望频率法提取到了一些和已知调控位点相同或相似的二聚体寡核苷酸 TGA $\{n\}_s$TTT（AAA $\{n\}_s$TCA），TGA $\{n\}_s$TTA（TAA $\{n\}_s$TCA），TGA $\{n\}_s$ATG（CAT $\{n\}_s$TCA），GGA $\{n\}_s$AAT（ATT $\{n\}_s$TCC），但它们的 $u<3.10$，用方法 1 未能提取到. 事实上期望频率法选择的对

照序列包含了外显子、高效内含子和低效内含子，期望频率f_{exp}（D）是二聚体寡核苷酸在三者中出现的频率平均，基于期望频率提取的过表达二聚体寡核苷酸既有潜在的转录调控元件，同时也有包含外显子和内含子特征的模体. 方法 1 对高效和低效转录内含子序列中二聚体寡核苷酸的出现频率进行对照分析，抽提出的二聚体寡核苷酸在高效内含子组中出现频率显著高于低效内含子组，属于高效内含子的特征，因而可能与转录正调控有关.

文中所用的两种方法都没有抽提到 CGG $\{n\}_s$CCG，它在高效内含子组出现 14 次，s = 0，2，4，5，6，10，11；在低效内含子组出现 6 次，s = 0，2，7. 两组内含子中 CGG $\{n\}_s$CCG 的含量都很少，间隔区长度 s 都没有取遍 0 ~ 11 之间的所有值，这可能与内含子的碱基成分有关，高效内含子组中 A、T、G、C 的含量分别为 34%，33%，17%，16%，低效内含子组中分别为 33%，33%，17%，17%，两组内含子 A-T 的含量都高于 G-C，这是内含子的一个普遍特征. 实际上 CGG $\{n\}_{11}$CCG，CGG $\{n\}_{10}$CCG，CGG $\{n\}_6$CCG 在高效内含子组都出现了 2 次，低效内含子组中不含有这些模体，它们都是实验获得的二聚体寡核苷酸结构的转录因子结合位点.

参 考 文 献

1 Katharina H S, Cox T C, May B K. Identification and characterization of a conserved erythroid-specific enhancer located in intron 8 of the human 5-aminolevulinate synthase 2 gene. J Biol Chem, 1998, **273** (27), 16798 ~ 16809

2 Bhattacheryyz N, Banerjee D. Transcriptional regulatory sequences within the first intron of the chicken apolipoprotein A Ⅰ (apoA Ⅰ) gene. Gene, 1999, **234** (2): 371 ~ 380

3 Clement J Q, Wilkison M F. Rapid induction of nuclear transcripts and inhibition of intron decay in response to the polymerase Ⅱ inhibitor DRB. J Mol Biol, 2000, **299** (5): 1179 ~ 1191

4 张 静，石秀凡. 酵母基因中转录正调控内含子序列特征的统计分析. 生物化学与生物物理进展，2003，**30** (2)：231 ~ 238
Zhang J, Shi X F. Prog Biochem Biophys, 2003, **30** (2): 213 ~ 318

5 Zhang J, Hu J, Shi X F, *et al*. Detection of potential positive regulatory motifs of transcription in yeast introns by comparative analysis of oligonucleotide frequencies. Comput Biol Chem, 2003, **27** (4 ~ 5): 497 ~ 506

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 van Helden J, Rios A F, Collado-Vides J. Discovering regulatory element in non-coding by sequences analysis of spaced dyads. Nucleic Acids Res, 2000, **28** (8): 1808 ~ 1818

8 van Helden J, Andre B, Collado-Vides J. Extracting regulatory site from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol, 1998, **281** (8): 827 ~ 842

9 Kutach A K, Kadonaga J T. The downstream promoter element DPE appears to be widely used as the TATA box in *Drosophila* core promoter. Mol Cell Biol, 2000, **20** (13): 4754 ~ 4764

Statistical Analysis of Dyad-like Transcriptional Regulatory Sites in The Introns of Yeast Genes*

HU Jun[1,2], ZHANG Jing[1)**]

([1)] *The Center of Applied Statistics, Yunnan University, Kunming* 650091, China;
[2)] *The Fundamental Information Engineering College, Yunnan Agricultural University, Kunming* 650201, China)

Abstract A comparative analysis of the occurrence frequency of oligonucleotides in two sets of yeast genes with higher and lower transcription frequencies respectively has shown that the sequence structures of the two sets of introns are different. There are more potential binding sites in the introns of genes with higher transcription frequencies. After observing regulatory sites obtained by experimental analysis, many transcriptional regulatory sites in yeast consist of a pair of highly conserved oligonucleotides, spaced by a non-conserved region of fixed width (dyad). Therefore, dyad-like transcriptional regulatory sites are analyzed. Some dyads are extracted by statistical comparative analysis of the occurrence frequencies, whose occurrence frequencies in the set of introns with higher transcription frequencies are higher significantly than those in the set of introns with lower transcription frequencies. Analyzing the distribution of the extracted dyads in two sets of introns, and comparing with the regulatory sites revealed by experiments, these dyads are probably related to positive transcriptional regulation.

Key words yeast, intron, dyad, frequency analysis, transcriptional regulatory site

* This work was supported by grants from The National Natural Science Foundation of China (30360027) and The Natural Science Foundation of Yunnan University (2002T009XX).

** Corresponding author. Tel: 86-871-6541419, E-mail: zhangjing@ynu.edu.cn

Received: November 5, 2003 Accepted: December 24, 2003