酵母基因上游序列中潜在的转录正调控位点分析 *

王秀荷 张 静 **

(云南大学统计系，应用统计中心，昆明 650091)

摘要 前期研究表明，高效转录酵母基因内含子在序列长度、寡核苷酸使用、以及位置分布等方面都有着区别于低转录内含子的特征. 进一步观察发现：上游基因间区域的序列长度与基因转录频率也有与内含子序列相同的现象，转录频率高的上游基因间序列一般都比转录频率低的长. 对高效转录和低效转录上游基因间序列的寡核苷酸使用频率进行统计比较分析，抽提出高转录基因上游区可能的转录正调控元件. 与酵母的所有非编码序列比较，这些可能的正调控元件基本上也是过表达的 (over-represented)，其中多数和实验所得的一些位点特征相吻合. 这些元件富含 G、C，这与内含子中可能的正调控元件在碱基组成上有一定的互补性. 从这些特征看，高效转录基因上游的序列结构确实有利于基因的转录.

关键词 酵母，基因上游，转录频率，调控位点，寡核苷酸频率分析

学科分类号 Q61

前期研究中，我们曾对高效转录酵母基因内含子与低效转录酵母基因内含子进行了比较分析，从它们的寡核苷酸使用以及位置分布看，高转录酵母基因内含子具有有利于基因转录的特征性结构[1~4]. 由于基因上游是整个转录调控的核心，该区域中的调控信号除了决定基因是否表达、何时表达外，是否还有影响基因转录效率的因素，这是我们所关心的问题. 对上述样本基因的初步观察发现了与内含子相似的现象，即高效转录酵母基因的上游基因间的序列 (以下我们简称基因上游) 长度一般比低效转录基因的长 (图 1). 约 78%的高转录基因的上游长度在 500 bp 以上，而只有不到 27%的低转录基因上游长度达到了 500 bp. 这似乎提示高转录基因上游序列中含有较多潜在的有利于基因转录的转录因子结合位点，基因上游的组织也影响着基因的转录效率. 为了揭示这些具有内含子的酵母基因的高转录效率调控机制，本文拟对这些酵母基因的上游序列结构特征 (主要是寡核苷酸的使用) 做些细致的分析.

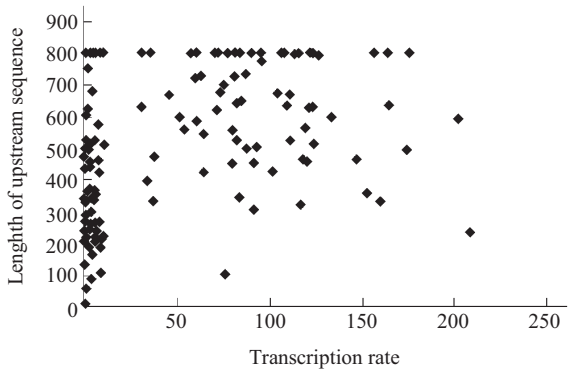

Fig.1 The scatter plots of lengths of upstream sequences *vs*. transcription rates of genes

尽管对酵母基因上游区调控位点的研究已有不少工作， 但是一般都是针对一些特定基因族进行的[5]. 在酵母基因组中，含有内含子的基因不多，大约为 230 个. 我们所研究的具有内含子的高转录 (转录速率 ≥ 30 mRNA/h) 基因中，几乎都是编码核糖体蛋白的. 相反，低转录(转录速率 ≤10 mRNA/h) 的那组中，没有发现核糖体蛋白基因. 因此我们的分析结果反映的正好是核糖体蛋白基因的调控特征. 虽然一些研究已揭示了参与核糖体蛋白基因调控的转录因子(如 RAP1，ABF1，TAF，GCN4，ADR1 等) 及其协同调控特征[6~8]，但是对核糖体蛋白基因所有调控位点的结构还没有系统的认识. 本文则提供了比较详细的调控这类基因转录的潜在位点信息. 这些信息一方面有助于对高转录基因的转录调控机制的理解，另一方面可为其他真核生物表达水平的预测提供一定的线索.

*国家自然科学基金资助项目(30360027).

** 通讯联系人.

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

收稿日期：2005-04-08, 接受日期：2005-06-29

1 材料和方法

1.1 材料

为了保持研究的系统性，我们仍用文献[1]中所列的样本(只是去除了内含子位于上游的几个)，转录频率按照酵母内含子数据库 (YIDB，http://www.imb-jena.de/RNA.html) 所提供的. 从 NCBI 酵母基因组数据库(http://www.ncbi.nlm.nih.gov)中取出包括上游序列的酵母基因，共 67 个转录频率较高 (>30 mRNAs/h) 的基因和 71 个转录频率较低 (≤10 mRNAs/h) 的基因. 根据 van Helden 等[5]的分析，酵母基因上游的转录调控位点一般位于翻译起始点上游 800 bp 区域内，但是我们得到的这两组基因中，很多基因的翻译起始点上游 800 bp 区域中含有其他基因的编码区，因此我们只考虑上游的2个基因之间的区域. 事实上，从图 1 也可看出，转录调控的信息似乎也主要位于上游基因间的区域中.

上游调控区中还含有一般的基础调控元件 (basal regulatory element) (即高转录和低转录都共有的转录元件)，这些元件只用上面的样本比较是无法抽取的. 所以我们也取出酵母基因的所有非编码序列，作为参照序列以抽取一般的调控元件.

1.2 提取上游区的潜在调控位点

为了和前期工作保持一致[1, 2]，本文也主要是对四核苷酸、五核苷酸的情况进行了统计分析. 当然，将抽提出的短寡核苷酸放在序列中考察，从它们的重叠或连接即可获得一些较长寡核苷酸的结果.

分析方法也沿用我们前期工作中所提出的，即对寡核苷酸的频率进行统计比较分析. 只是在本文中，对周期为 1 和 2 的寡核苷酸的计数作了调整，也就是将连续出现 n 次 (n > 4 或 5) 的某个核苷酸只算作 1 个 4- (或 5-) 核苷酸. 而在前期研究中，在对 4- (或 5-) 核苷酸统计时，对连续出现的 n 次 (n > 4 或 5) 某个核苷酸片段，是按 n-3 (或 n-4) 个 4- (或 5-) 核苷酸来计算的. 例如，在前期对 4- 核苷酸的统计中，对于连续出现的 8 个 AAAAAAAA，统计值为 5，而改进后我们只把它算做 1 个 4- 核苷酸；同理，在遇到诸如 ATATATAT 此类的序列时，我们也记为 1 个 ATAT 4- 核苷酸. 这样比较符合通行的做法. 结果表明在我们的分析中，两种计数法所得到的结果几乎没有差别，这主要归因于我们所使用的频率比较方法的特点，即只保留特异性的过表达寡核苷酸片段.

将高转录基因上游区和低转录基因上游区寡核苷酸的使用频率进行比较，以抽提可能的正调控元件. 用 $n_1(b)$和 $n_2(b)$分别表示寡核苷酸(b)在高转录基因上游和低转录上游中出现的次数，n_1 和 n_2 分别表示确定长度(4 或 5)的所有寡核苷酸在高转录上游和低转录上游中出现的总数，即

$$n_1 = 2\times\Sigma n_1(b),$$
$$n_2 = 2\times\Sigma n_2(b)$$

其中因子 2 表示对寡核苷酸的统计是在 DNA 双链上 (方向均为 5′→3′) 进行的，因为大多数转录因子在两条链上都有活性[5,9]. 寡核苷酸(b)在两组上游区中的出现频率分别为 $n_1(b)$ / n_1 和 $n_2(b)$ / n_2，记为 $f_1(b)$和 $f_2(b)$. 采用单边假设检验法进行分析，即零假设 $H_0: f_1(b) \leqslant f_2(b)$，其备择假设 $Ha: f_1(b) > f_2(b)$. 寡核苷酸(b)在两组上游区中出现频率差异的标准差为：

$$s=\sqrt{\frac{n_1(b)+n_2(b)}{n_1+n_2}\left(1-\frac{n_1(b)+n_2(b)}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

计算 u 值：

$$u=\frac{n_1(b)/n_1-n_2(b)/n_2-\delta_1\times0.5/n_1-\delta_2\times0.5/n_2}{s}$$

$$\delta_i=\begin{cases}0, \text{ if } n_i(b)\geqslant30\\ 1, \text{ if } n_i(b)<30\end{cases}, \ i=1,2$$

其中 $n_1(b)$，$n_2(b)$小于 30 时，进行了连续性矫正. 如果 $n_1(b)$和 $n_2(b)$都小于 5，则直接用二项分布计算概率. 取显著水平 α= 0.05，当 u > 1.64，即 P < 0.05 时，拒绝 H_0 而接受 Ha. 即认为寡核苷酸(b)在高转录基因上游中的出现频率显著高于在低转录基因上游中的出现频率. u 值越大，差异越显著.

用相同的方法对高转录基因上游和非编码区寡核苷酸的使用频率进行比较，抽取高转录基因上游区中可能的一般调控元件.

为了便于区分，以下将高转录和低转录比较所得的 u 值就记为 u，而将高转录和非编码比较所得的 u 值记为 u_1.

抽取出 u 和 u_1 均大于 1.64 的寡核苷酸，分析它们在两组上游中的分布情况.

2 结　　果

2.1 四核苷酸的情形

将高转录基因上游中四核苷酸的出现频率分别与低转录基因上游中四核苷酸的出现频率和非编码区四核苷酸的出现频率相比较，按照 u > 1.64 的阈

值, 抽提出高转录基因上游可能的正调控四核苷酸和高转录基因上游可能的一般调控四核苷酸 (表 1). 由于对寡核苷酸出现数目的统计是沿 DNA 的两条链进行的，所以每个寡核苷酸出现的数目与其反转互补的寡核苷酸数目相同. 例如，5′ GCTA 3′和 5′ TAGC 3′出现的数目相同.

Table 1 The tetranucleotides extracted by frequency comparison

Tetranucleotide	u	occ_1	ms_1	occ_2	ms_2
AGGC(GCCT)	3.57	120	65	56	58
ACCT(AGGT)	2.06	114	67	74	59
ACAT(ATGT)	3.77	272	66	135	59
ACGG(CCGT)	3.92	67	67	44	38
GGGT(ACCC)	2.43	98	65	50	45
ATGG(CCAT)	2.67	133	67	80	60
TTCC(GGAA)	1.91	195	66	134	55
CTCC(GGAG)	3.12	97	65	54	52
GCCC(GGGC)	4.62	69	66	34	33
GGCC	5.35	83	62	33	29
GGTA(TACC)	2.74	117	67	54	50
CCTA(TAGG)	3.80	105	66	50	49
CTAG	3.89	136	66	70	60
CCAG(CTGG)	3.05	111	66	60	54
CCCA(TGGG)	6.85	132	66	47	44
TCCA(TGGA)	2.01	128	66	73	58
GCCG(CGGC)	3.21	72	61	39	34
GCGC	3.45	84	64	41	40
CCAA(TTGG)	2.28	174	66	103	50
CGGA(TCCG)	4.24	78	62	48	39
GTAC	2.90	135	66	75	52
CACC(GGTG)	3.89	99	66	55	53
AAAG AGCA ACAG ACCA GAAA GCAA GGTC GACC GATC CGTA CATT CTAA TACG CTTT CTTC TTAG TGGT TGCT TCAA TCTG TTGT TTGC TTGA CCGG AGCC GACC ATCC GTGG					

The tetranucleotides in brackets denote the reverse complements. u denotes u value of the frequency difference of the tetranucleotide between highly-transcribed and lowly-transcribed upstream sequences. ms_1 (ms_2) denote the matching sequences. occ_1(occ_2) denote the occurrence number of the tetranucleotide in the highly-transcribed (lowly-transcribed) up-stream sequences. Similarly in Table 2.

表 1 中上半部分列出了高转录和低转录比较所得结果，即可能的转录正调控元件，下半部分列出的是高转录基因样本与非编码序列比较所得的结果. 在上半部分的四核苷酸中，除了 CGGA (TCCG)，GTAC 和 CACC(GGTG)外，也都在高转录基因样本与非编码序列比较时被探测到，这说明它们也是非编码序列中的过表达寡核苷酸. 通过与 TRANSFAC 数据库[10]中酵母的转录调控位点比较，这些四核苷酸中大多数能与之匹配，如 AGGT 是被 Vignais 等[11]实验所支持的很常见的调控元件.

从碱基组成上看，两类基因的上游序列中 G + C 含量均较低，其中高转录基因和低转录基因的 G + C 含量分别为 38%和 36%，它们的 A+T 含量相对就要高得多，分别为 62%和 64%. 尽管碱基组成与内含子相差不大，两组内含子中 G + C 含量分别为 33%和 35% (A+T 含量为 67%和 65%). 但从我们的抽提结果来看，两类基因的上游和内含子在正调控位点的组成上则存在着较大的差异，不论是四核苷酸还是五核苷酸，高转录基因上游中的潜在正调控位点富含 G、C 两种碱基，而高转录基因内含子中的潜在正调控位点富含 A、T 两种碱基[1]. 相对来说，高转录基因样本与非编码序列比较所得的结果中 A+T 含量略高一些 (表中下半部分)，而非编码序列的 G+C 含量为 35% (A+T 为 65%). 这些结果说明我们提取出来的寡核苷酸在高转录基因上游中并非随机出现的.

2.2 五核苷酸的情形

五核苷酸共有 1 024 种形式，取阈值为 1.64 时，高转录与低转录基因上游的比较共抽提出 116 个，与非编码区比较得到 248 个五核苷酸. 表 2 中列出了 u >2.33 的结果. 同样，表 2 上半部分为高转录和低转录比较所得结果，下半部分的是高转录基因样本与非编码序列比较所得的结果. 上半部分列出的所有五核苷酸在高转录基因样本与非编码序列比较时也都被探测到. 这些五核苷酸中的大多数也都能与 TRANSFAC 数据库中酵母的转录调控位点相匹配，如 ACCCA、CACCC、ATGTA 等就是较为常见的调控元件. 与表 1 对照可见，它们中的很多是表 1 中四核苷酸的延伸. 例如，我们在上游区找到的五核苷酸 ACCCA，它既可以看成是四核苷酸 ACCC 的延伸，也可以看成是 CCCA 四核苷酸的延伸.

Table 2 The pentanucleotides extracted by frequency comparison

Pentanucleotide	u	occ_1	ms_1	occ_2	ms_2
AAATG(CATTT)	3.38	77	51	45	37
AGGTA(TACCT)	2.84	35	31	16	11
ACACC(GGTGT)	2.71	40	39	13	8
ACCCA(TGGGT)	3.74	55	50	12	11
CAAAC(GTTTG)	2.60	70	52	27	23
CACCC(GGGTG)	2.75	36	33	10	6
CGTAC(GTACG)	2.44	44	42	10	8
CCATG(CATGG)	2.76	26	19	9	8
TATTG(CAATA)	3.78	80	55	45	32
TCTAG(CTAGA)	2.99	43	37	16	10
ACATT(AATGT)	3.29	91	56	50	26
ACCGT(ACGGT)	2.99	25	23	11	8
ATGTA(TACAT)	3.02	80	54	44	31
CCCAT(ATGGG)	2.52	49	44	12	11
CCTAG(CTAGG)	2.67	24	22	10	5

GGATC AGGGC GATGG ATGTC GTCCA TCCGG CCGGA GCGCA CGCCG ACCCG GGTAC
TAGGA GGGCA TGGAA TCGGA GGAAA CACCA TCTGG CCCAA TGGTA TTTGG AGGTG
ACTAG GCCAT AAATA TATGT TACCG CTCCT TAGGT CCAGT CACGG AGGTG CCACT

2.3 寡核苷酸在序列中的分布以及与 TRANSFAC 位点的比较

对基因上游调控转录机理的研究表明，在多数情况下，许多含 G、C 的元件都是转录调控的关键因素，它们的存在对提高转录调控的效率有一定的作用. 以上所抽提出的寡核苷酸，特别是表中上半部分，富含 G、C 碱基，所以它们作为转录正调控元件的可能性是比较大的. 但是这并不意味着我们提取到的每个寡核苷酸个体就是转录效率的调控元件，实际的调控 motif 往往是由较长的序列构成的，而且调控位点之间是协同作用的. 为了了解所抽提出的寡核苷酸在序列中的分布情况，我们在所分析的序列中将表 1 和表 2 所列的寡核苷酸全部显示出来，由它们的重叠或连接即可获得一些较长寡核苷酸的信息. 我们发现三组 DNA 序列 (高转录上游序列、低转录上游序列和非编码序列) 重叠或连接所得的寡核苷酸片段的宽度差异较大. 高转录上游序列中有较多的长寡核苷酸，有些长达 30 多个碱基，而低转录上游序列和非编码序列中所得寡核苷酸片段长度一般都不太长. 并且从所显示出的核苷酸在序列中的分布密度看，在高转录上游序列中的平均密度最高，约为 45%，而低转录上游序列和非编码序列中的平均密度都较低，分别只有 36%和 21%.

此外，我们根据 TRANSFAC 提供的几个核糖体蛋白基因转录因子 RAP1， ABF1， TAF， GCN4， ADR1 和 REB1 的调控位点距阵(matrix)，对高转录基因序列进行了调控位点搜索. 这些位点中约 80%被我们的频率比较方法探测到，亦即含有我们抽提出的寡核苷酸. 图 2 显示了 yhl001w 基因上游序列的情况. 我们用 TRANSFAC 矩阵探测到了 9 个结合位点序列(图 2 中带下划线的部分)，

其中有几个为两个或三个因子位点的重叠，所有这些序列片段中均含有用我们的方法抽提出的寡核苷酸 (图 2 中大写的部分). 事实上除了上述 6 个转录因子，核糖体蛋白基因还有其他一些转录因子，如 Fhl1，但还未见其位点距阵的实验报道. 所以图 2 中那几个未带下划线的大写寡核苷酸还不能说是没有实验支持. 就已有实验支持的结果看，它们实际上可作为实验分析的线索.

5′ GCCCttcaacttatatagCACCTtTACATatctttaatttttacgaagagCATTTtattcaacttttGTACCTttaaagAC
REB1 RAP1 ABF1/TAF
GGttttcaGCCCAcACATTTtttgttatttttACACCacacagtaaaactttcttgctttttttcttgagaCTCCccgaccattgttcg
TAF/ADR1/RAP1 TAF/ADR1 GCN4/REB1
ataTTCCAAttttcgcaatcttcagaaaaGGGTAtCCTAACATTctaaattaataataaagaagagAGGCtacaagCATT
ABF1 REB1
Tattagtgtgaag CATTTaaagtaacttcaggaaataaacGCGCaaataaaCCAAaa 3′
TAF/GCN REB1/RAP1

Fig.2 The upstream sequence of yhl001w

The tetranucleotides and pentanucleotides in **Table 1** and **Table 2** are capitalized. The underlined parts are transcription factor binding sites hit by TRANSFAC matrices. Several sites are overlapping.

3 讨 论

我们知道基因上游是整个转录调控的核心，该区域中的调控信号决定基因是否表达、何时表达以及表达水平等. 本文的分析结果进一步表明该区域中含有决定基因转录效率信息. 在本文中我们对高转录和低转录上游基因间序列的寡核苷酸使用频率进行统计比较分析，抽提出高转录基因上游区可能的转录正调控元件. 与酵母的所有非编码序列比较，这些可能的正调控元件基本上也是过表达的，而且这些元件和实验所得的一些位点特征相吻合. 值得注意的是，这些上游可能的正调控元件富含 G、C，这与内含子中的情况相反，我们在内含子中抽提出的可能的正调控元件富含 A、T [1,2]，尽管上游与内含子的碱基成分相差不大. 高转录基因的内含子比较靠近基因上游区的现象[3]，曾使我们推测内含子与基因上游在转录调控上可能有密切关联. 从组合调控的角度看，上游调控元件与内含子调控元件在碱基组成上的互补性似乎也支持这个推测.

对基因上游调控转录机理的研究表明，TATA 元件是绝大多数基因转录所必需的，它的缺失会大大降低转录水平. 然而核糖体蛋白基因的转录是个例外，他们大多缺乏 TATA 元件，而由其他因子如 RAP1，TAF 等行使激活转录的功能[7]. 我们的分析也未抽提出 TATA 元件，无论是在上游还是在内含子中. 虽然一些序列从表面上看也含有 TATA 的片段(如图 2 中靠近 5′端处)，但这些片段只是随机出现的非普遍的情况，不具有生物学的 TATA-box 功能. 这也说明我们所用方法的有效性. 这些含内含子编码核糖体蛋白基因的序列结构确实有利于基因的转录. 余下的工作是要揭示蕴含在这些结构特征中组合调控的规律性.

参 考 文 献

1 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): 231～238

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

3 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

4 胡 俊, 张 静. 酵母基因内含子中二聚体寡核苷酸转录调控为点的统计分析. 生物化学与生物物理进展, 2004, **31**(5): 449～454
Hu J, Zhang J. Prog Biochem Biophys, 2004, **31** (5): 449～454

5 van Helden J, Andre 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

6 Lieb J D, Liu X L, Botstein D, *et al*. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nature Genetics, 2001, **28** (4): 327～334

7 Mencía M, Moqtaderi Z, Gelsberg J V, *et al*. Activator-specific recruitment of TFIID and regulation of ribosomal protein genes in yeast. Mol Cell, 2002, **9** (4): 823～833

8 Yu Q, Qiu R, Foland T B, *et al*. Rap1p and other transcriptional regulators can function in defining distinct domains of gene expression. Nucleic Acids Res, 2003, **31** (4): 1224～1233

9 Himes R, Tagoh H, Goonetilleke N, *et al*. A highly conserved intronic element in the c-fms (M-CSF receptor) gene comtrols macrophage-specific and regulated expression. J Leukocyte Biol, 2001, **70** (5): 812～820

10 Wingender E, Chen X, Fricke E, *et al*. The TRANSFAC system on gene expression regulation. Nucl Acids Res, 2001, **29** (1): 281～283

11 Vignais M L, Huet J, Buhler J M, *et al*. Contacts between the factor TUF and RPG sequences. J Biol Chem, 1990, **265** (24):14669～14674

Analysis of Potential Sites in Upstream Regions for Positive Transcriptional Regulation of Yeast Gene*

WANG Xiu-He, ZHANG Jing**

(*Department of Statistics, Center of Applied Statistics, Yunnan University, Kunming* 650091, China)

Abstract It has been demonstrated that there are differences between introns of highly-transcribed genes and those of lowly-transcribed genes in sequence length, position preference and oligonucleotide usage. Further observation showed a similar phenomenon: the lengths of upstream intergenic sequences of highly-transcribed genes are generally longer than those of lowly-transcribed genes. Based on the statistical comparative analysis on the occurrence frequencies of oligonucleotides in the upstream intergenic regions of the two sets of genes, some potential sites were extracted in the upstream regions of highly-transcribed genes which are likely to enhance the transcription of genes. These regulatory elements turned out to be also over-represented by comparing the upstream sequences of highly-transcribed genes to all non-coding sequences of yeast genes. Most of these elements are agreement with transcription factor binding sites obtained from experimental analyses. And these elements are G,C rich, this seems to be supplement to those potential sites in introns extracted before, which are A,T rich, in base composition. Such sequence structures of highly-transcribed genes are favorable to the transcription of genes.

Key words yeast, upstream region of gene, transcription rate, regulatory site, oligonucleotide frequency analysis

---

*This work was supported by a grant from The National Natural Science Foundation of China (30360027).

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

Received: April 8, 2005 Accepted: June 29, 2005