酵母基因中转录正调控内含子序列特征的统计分析*

张 静[1)**] 石秀凡[2)]

([1)]云南大学应用统计中心，昆明 650091；[2)]中国科学院昆明动物研究所，细胞与分子进化重点实验室，昆明 650223)

摘要 大量实验研究显示，真核基因的许多内含子具有调控转录的功能，但是对此问题尚缺乏全面的研究．观察一些酵母基因的转录频率与基因的内含子序列后，发现一个值得注意的现象：转录频率高的基因内含子序列一般都比较长，而转录频率低的基因内含子一般都比较短．这提示高效转录基因的较长内含子中可能含有某些增强基因转录的特征性结构．于是选取两组酵母基因的内含子进行详细研究，第一组的基因具有较高的转录频率 (> 30)，第二组的基因转录频率较低 (≤10)．对寡核苷酸 (主要是四核苷酸、五核苷酸) 的出现频率进行统计比较分析，探测到一批寡核苷酸，它们在第一组内含子中出现的频率显著高于在第二组内含子中出现的频率，同时也显著高于与第一组内含子相邻的外显子中的出现频率．其中一些寡核苷酸与实验研究得到的转录调控元件相同．从这批寡核苷酸在内含子和外显子序列中的分布看，高效转录基因内含子的序列结构确实有利于基因的转录．

关键词 酵母，内含子，基因转录频率，寡核苷酸出现频率

学科分类号 Q61

真核基因组中含有大量的内含子 (intron) 序列，众多的研究表明，这些曾经被认为是“垃圾”的内含子也具有调控基因表达的功能[1]．迄今为止，研究者对内含子调控功能关注较多的是，导致同一基因在不同发育阶段产生不同蛋白质产物的 mRNA 选择性剪接 (alternative splicing)．近几年来的大量实验研究表明，真核基因的某些内含子还具有调控基因转录的功能，起着转录增强子 (enhancer) 或抑制子 (repressor) 的作用[2~5]．然而，目前对于这项功能的研究还主要是一些零散的实验，要揭示内含子调控转录的机理，还需要大量实验和理论的研究．

酵母 (yeast) 基因组是较早测序并且研究得较多的基因组．我们在观察一些酵母基因转录频率 (转录频率是指单位时间内产生的转录物的数量) 的数据后发现，转录频率高的基因内含子序列大多比转录频率低的基因内含子序列长．这似乎提示转录频率高的基因内含子可能含有某些增强转录效率的信号序列，在转录过程中起着启动子 (promoter) 或增强子 (enhancer) 的作用，而转录频率较低的基因内含子则缺乏这些序列．对高效转录基因内含子和低效转录基因内含子中寡核苷酸出现的频率进行对比分析，结果发现一些寡核苷酸在高效转录基因内含子中出现的频率明显高于低效转录基因内含子中的出现频率．由此我们推测这些寡核苷酸可能与转录的正调控有关．当然，仅有这个结果还不足以说明高效转录基因内含子的正调控作用，因为外显子序列也有可能起作用．因此我们又将高效转录基因的内含子与相邻外显子序列的寡核苷酸使用情况作比较，结果表明，以上得到的寡核苷酸中，绝大多数在外显子中的出现频率更是显著低于在高效转录基因内含子中的出现频率．从这些寡核苷酸在两组内含子序列和外显子序列中的分布来看，高效转录基因的内含子确实具有某些有利于基因转录的序列特征．

1 方 法

1.1 选取样本

我们从酵母内含子数据库 (YIDB, http://www.imb-jena.de/RNA.html) 中选出转录频率比较高 (> 30) 的基因，共计 76 个基因，并且从 EMBL 数据库中取出它们的内含子序列，共有 77 个 (ypl198w 有两个内含子) (表 1)．按同样方法再选出转录频率较低 (≤10) 的基因，共计 77 个基因，77 个内含子 (表 2)．为了叙述方便，本文中我们将表 1 和表 2 中的内含子分别称为“第一组内含子”和“第二组内含子”．同时将与第一组内含子相邻的外显子序列也取出，共计 77 个．

* 教育部科研重点基金资助项目 (00246)．

** 通讯联系人．

Tel: 0871-5036207, E-mail: zhangjing@ynu.edu.cn

收稿日期：2002-09-28，接受日期：2002-12-30

Table 1 The yeast genes with higher transcription frequencies (> 30) and the lengths of the corresponding introns

ORF name	Tran. freq. [1)]	Intron length	ORF name	Tran. freq. [1)]	Intron length
ybl027w	60.5	384	yjl136c	156.6	460
ybl072c	160	308	yjl177w	72	317
ybl087c	147	504	yjl189w	122.9	386
ybl092w	108.5	333	yjr094wa	122.1	275
ybr048w	38.3	511	yjr145c	81.2	260
ybr084w	64.4	506	ykl006w	106.2	398
ybr181c	87.2	352	ykl180w	110.7	306
ybr189w	76.2	413	ykr057w	152.9	322
ybr190w	202.4	388	ykr094w	69	368
ydl075w	104.4	421	ylr048w	57.8	359
ydl082w	82.6	365	ylr061w	102.1	389
ydl083c	109.5	432	ylr185w	85.2	359
ydl130w	120	301	ylr287ca	116.6	430
ydl136w	82.6	405	ylr344w	31.2	447
ydl191w	113.1	491	ylr406c	80.1	349
ydr064w	83.7	539	ylr448w	77.4	384
ydr025w	71.3	339	yml024w	110.7	398
ydr450w	95.4	435	yml073c	53.9	415
ydr471w	79.9	384	ymr142c	59.7	402
ydr500c	87.6	389	ymr194w	37.6	463
yer074w	63.4	466	ynl069c	75.5	449
yer117w	133.6	525	ynl096c	51.4	345
yer131w	36.1	471	ynl162w	96.2	512
yfl039c	45.5	308	ynl301c	60.7	432
ygl030w	89.9	230	ynl302c	121.4	551
ygl031c	175.3	456	yol120c	91.3	447
ygl076c	70.6	468	yol121c	64.4	390
ygl103w	126.3	511	yor096w	83.8	401
ygl189c	208.4	368	yor182c	119.4	411
ygr027w	30.7	312	yor234c	107.8	527
ygr034w	122.1	354	yor293w	124	437
ygr118w	173.7	320	yor312c	73.7	407
ygr148c	163.7	379	ypl079w	201.6	421
ygr214w	35.3	455	ypl090c	80.9	594
yhl001w	91.5	398	ypl143w	115.6	525
yhr010w	122.9	561	ypl198w	34.3	409, 407
yil018w	164.4	400	ypr043w	92.9	403
yil148w	82.1	434	ypr132w	118.1	365

[1)] Transcription frequency, i.e., the number of transcripts produced in unit time[6], similarly hereinafter.

Table 2 The yeast genes with lower transcription frequencies (≤10) and the lengths of the corresponding introns

ORF name	Tran. freq.	Intron length	ORF name	Tran. freq.	Intron length
yal001c	1. 1	90	yhr041c	4. 3	101
yal030w	2. 3	113	yhr077c	1. 4	113
ybl018c	3. 4	75	yhr097c	0. 8	124
ybl026w	4. 9	128	yhr101c	0. 2	87
ybl040c	6. 1	97	yhr123w	2	91
ybl050w	2	116	yil004c	3. 9	131
ybl059w	1. 2	69	yjl001w	10. 7	116
ybr090c	8. 8	357	yjl024c	0. 7	77
ybr119w	1. 6	89	yjl041w	2. 9	118
ybr230c	3. 8	97	yjl206c	6. 1	143
ycr097w	3. 4	136	yjr021c	0. 3	80
ydl012c	3. 5	86	ykl007w	1. 7	141
ydl029w	7. 1	123	ylr078c	4. 1	89
ydl064w	8. 2	110	ylr128w	0. 7	94
ydl079c	0. 3	292	ylr202c	1. 5	116
ydl108w	1. 7	81	ylr275w	2. 3	90
ydl189w	2. 1	146	ylr306w	0. 4	134
ydl219w	10. 1	71	ylr426w	3. 3	71
ydr059c	0. 4	80	ylr464w	0. 8	279
ydr092w	8. 2	268	yml025c	1	99
ydr129c	6. 1	111	yml067c	8. 7	93
ydr305c	1. 3	89	yml094w	4. 4	83
ydr397c	8	92	yml124c	6. 2	298
ydr424c	6. 1	96	ymr033w	2. 9	86
yel012w	0. 7	123	ymr225c	2. 6	147
yer003c	8	93	ynl050c	3. 6	91
yer007ca	10	103	ynl130c	3. 7	441
yer093ca	0. 9	75	ynl147w	8. 6	120
yer133w	9. 2	525	ynl246w	2. 1	95
yer179w	0. 3	92	ynl265c	1. 1	105
ygl087c	1. 8	87	ynl312w	5. 9	108
ygl137w	5	200	yor221c	0. 9	201
ygl232w	5. 5	58	yor318c	0. 2	347
ygr001c	2. 9	93	ypl031c	2. 1	102
ygr029w	6. 4	83	ypl129w	7. 4	105
ygr183c	8. 4	213	ypl175w	0. 4	84
yhr001wa	3. 3	63	ypl241c	0. 8	80
yhr012w	4. 2	119	ypr111w	0. 3	59
yhr016c	0. 5	168			

1.2 算法

由于转录因子的作用位点一般在四到十几个核苷酸之间，所以本文主要对四核苷酸、五核苷酸的情况进行统计分析. 虽然对六核苷酸的情况也作了统计，但限于篇幅，本文未列其详细结果. 事实上，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可获得一些较长寡核苷酸的结果.

用$n_1(\mathrm{b})$和$n_2(\mathrm{b})$分别表示寡核苷酸 (b) 在第一组内含子和第二组内含子中出现的次数；n_1 和 n_2 分别表示确定长度 (4 或 5) 的所有寡核苷酸在第一组内含子和第二组内含子中出现的总数，即

$$n_1 = 2 \times \sum_{1}^{k_1}(L_{1i} - l + 1),$$

$$n_2 = 2 \times \sum_{1}^{k_2}(L_{2i} - l + 1)$$

其中 L_{1i}和L_{2i}分别表示第一组内含子和第二组内含子中第 i 个内含子的长度；l 表示寡核苷酸的长度；k_1 和 k_2 分别代表第一组内含子和第二组内含子的个数，本文中 $k_1 = k_2 = 77$. 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 DNA 的两条链上都有活性[7, 8]. 寡核苷酸 (b) 在两组内含子中的出现频率分别为 $n_1(\mathrm{b})/n_1$ 和 $n_2(\mathrm{b})/n_2$，记为 $f_1(\mathrm{b})$和$f_2(\mathrm{b})$.

本文的主要目的是分析第一组内含子中增强基因转录的序列信息，其基本特征应是某些寡核苷酸的频繁使用，因此我们从分析寡核苷酸的出现频率入手，并采用单边假设检验法进行分析，即假设 $H_0: f_1(\mathrm{b}) \leqslant f_2(\mathrm{b})$，其备择假设为$Ha: f_1(\mathrm{b}) > f_2(\mathrm{b})$. 寡核苷酸 (b) 在两组内含子中出现频率的标准差为:

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

计算 u 值:

$$u = \begin{cases} \dfrac{n_1(\mathrm{b})/n_1 - n_2(\mathrm{b})/n_2}{s}, & \text{如果 } n_1(\mathrm{b}) \geqslant 30 \text{ 且 } n_2(\mathrm{b}) \geqslant 30 \\ \dfrac{n_1(\mathrm{b})/n_1 - n_2(\mathrm{b})/n_2 - 0.5/n_2}{s}, & \text{如果 } n_1(\mathrm{b}) \geqslant 30,\ n_2(\mathrm{b}) < 30 \\ \dfrac{n_1(\mathrm{b})/n_1 - n_2(\mathrm{b})/n_2 - 0.5/n_1 - 0.5/n_2}{s}, & \text{如果 } n_1(\mathrm{b}) < 30 \text{ 且 } n_2(\mathrm{b}) < 30 \end{cases}$$

其中$n_1(\mathrm{b})$、$n_2(\mathrm{b})$小于 30 时，进行了连续性矫正. 如果$n_1(\mathrm{b})$和 $n_2(\mathrm{b})$都小于 5，则直接用二项分布计算概率. 取显著水平 $\alpha = 0.05$，当 $u > 1.64$，即 $P < 0.05$ 时，拒绝 H_0 而接受 $H\mathrm{a}$. 即认为寡核苷酸 (b) 在第一组内含子中的出现频率显著高于在第二组内含子中的出现频率. u 值越大，差异越显著.

在以上分析中，将第二组内含子的序列换成外显子序列，即得第一组内含子与外显子比较的结果. 为了便于区分，以下将两组内含子比较所得的 u 值记为u_1，而将第一组内含子与外显子比较的 u 值记为 u_2.

1.3 寡核苷酸的抽取

抽取出 u_1 和 u_2 均大于 1.64 的寡核苷酸，分析它们在两组内含子和外显子中的分布情况.

2 结 果

2.1 四核苷酸的出现频率

将第一组内含子中四核苷酸的出现频率分别与第二组内含子和外显子的作比较，按照 u_1 和 u_2 均大于 1.64 的标准抽取四核苷酸，256 种四核苷酸中只有为数不多的一些被抽出 (表 3). 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′-GATA-3′ 和 5′-TATC-3′ 出现的数目相同.

表 3 中的四核苷酸可分为两族，一族全部由 A、T 组成，另一族含有 G、C. 第一族的 u_1 和 u_2 值都比较高，尤其是 u_2 值. 而且，除了 TTAA 和 ATAT 外，其他四核苷酸在第一组内含子中都存在，TTAA 和 ATAT 也仅在一个内含子中空缺. 但是这一族四核苷酸在第二组内含子和外显子中的分布就不如在第一组内含子中那样广泛，特别是 ATAT，它只在第二组内含子的 49 个、外显子的 18 个中出现. 含有 G、C 的四核苷酸中，GATA 是已被实验证实了的与转录调控密切相关的元件. 其余几个与一些实验报道的转录调控元件的部分序

列相同或相似. 例如, 从 Paltaulf 的实验得知 CATGTGAA 是转录因子的结合位点, 我们的分析分别探测到其中的四核苷酸 CATG 和 TGAA.

Table 3 The tetranucleotides extracted by frequency comparison and the related messages

sequence	u_1	u_2	ms_{I}	occ_{I}	ms_{II}	occ_{II}	ms_{E}	occ_{E}
TAAA(TTTA)	2.46	18.05	77	686	68	185	60	141
AAAT(ATTT)	4.48	17.70	77	918	69	217	68	271
AATT	4.90	11.67	77	790	45	174	64	385
ATTA(TAAT)	2.74	15.84	77	566	60	145	56	127
TTAT(ATAA)	2.22	17.85	77	623	64	170	59	115
TTAA	4.11	10.01	76	540	44	118	60	236
ATAT	3.20	23.17	76	724	49	184	18	54
TATT(AATA)	3.76	19.03	77	750	70	182	63	151
TTCA(TGAA)	2.87	3.99	77	555	63	140	76	400
CATT(AATG)	1.86	9.0	77	463	60	127	65	208
CATG	1.64	6.21	51	210	21	54	34	92
GAAT(ATTC)	3.82	5.60	77	430	54	92	70	260
GAAA(TTTC)	1.83	4.51	77	649	65	184	77	462
GATA(TATC)	2.47	8.85	76	343	44	84	62	134

The tetranucleotides in brackets denote the reverse complements. u_1(u_2) denotes u value of the frequency difference of the tetranucleotide between the first set of introns and the second set of introns (exons). ms_{I} (ms_{II}) denotes the matching sequences, i.e., the number of introns in the first set (the second set) which contain at least one occurrence of the pattern. ms_{E} denotes the number of matching sequences in exons. occ_{I}, occ_{II} and occ_{E} denote the occurrence numbers of the tetranucleotide in the first set of introns, the second set of introns and exons respectively. Similarly in **Table 4.**

2.2 五核苷酸的出现频率

五核苷酸共有 1 024 种形式, 其中 u_1 和 u_2 均大于 1.64 的五核苷酸见表 4. 与表 3 对照可见, 它们中的大多数是表 3 中四核苷酸的延伸. 其中几个四核苷酸的延伸值得注意, 例如 GATA, 其 5′侧面延伸的碱基为 G 和 C, 形成 5′-SGATA-3′(S 代

Table 4 The pentanucleotides extracted by frequency comparison and the related messages

sequence	u_1	u_2	ms_{I}	occ_{I}	ms_{II}	occ_{II}	ms_{E}	occ_{E}
AAATT(AATTT)	4.35	10.42	76	327	33	56	48	93
ATTAT(ATAAT)	2.08	10.75	72	214	35	49	27	33
ATTAA(TTAAT)	1.89	7.80	63	164	26	37	31	42
ATATT(AATAT)	4.17	14.54	73	290	33	49	13	19
AAAAT(ATTTT)	1.86	13.95	74	370	51	95	39	60
AAATA (TATTT)	1.72	13.23	75	291	43	74	26	36
TTAAA(TTTAA)	3.21	10.01	67	213	31	39	26	41
ATTTA(TAAAT)	2.32	9.78	66	203	31	44	30	39
TATTA(TAATA)	2.64	11.13	69	196	27	40	12	21
GAATA(TATTC)	2.16	7.44	68	145	27	30	28	36
GAATT(AATTC)	2.66	2.68	61	149	19	27	45	95
TGAAT(ATTCA)	2.06	4.81	67	165	29	36	46	78
AGAAT(ATTCT)	2.18	2.06	65	128	21	25	51	87
TGAAA(TTTCA)	2.34	4.93	72	241	40	54	58	128
GAAAT(ATTTC)	2.10	5.58	74	196	30	44	42	88
CAAAT(ATTTG)	1.75	3.58	65	149	25	34	47	83
AATGA(TCATT)	1.70	7.98	70	191	31	46	39	54
GATAA(TTATC)	3.42	5.88	63	105	11	12	23	30
CGATA(TATCG)	1.70	4.31	45	66	12	12	17	21
GGATA(TATCC)	1.82	3.97	41	60	9	10	19	20
AGCAT(ATGCT)	1.65	1.70	44	61	10	11	29	39
AACTG(CAGTT)	1.67	2.65	51	81	14	16	33	45
ACTAT(ATAGT)	1.92	7.56	57	106	17	21	11	16

表 G 或 C) 的模体 (motif); GATA 3′侧面延伸的碱基为 A, 而 GATAA 是一个典型的转录元件核心. GAAT 两端的延伸均为 A 或 T. 由 A、T 构成的五核苷酸的 u_1 和 u_2 值仍然比较高, 尤其是 u_2 值; 它们在第一组内含子中的分布比在第二组内含子和外显子中的分布都广泛得多. CAAAT 与实验获得的元件 CAAT、CAAAAT 都很接近[9], 事实上, 六核苷酸的分析探测到了 CAAAAT (或 ATTTTG); GAAAT 也是实验得到的转录调控元件的一部分[10]. 这两个五核苷酸本身也很相似, 共有序列为 SAAAT.

2.3 寡核苷酸在序列中的分布特征

表 3 和表 4 说明其中的寡核苷酸在第一组内含子中的出现频率显著高于在第二组内含子和外显子中的出现频率, 因此它们可能与转录的正调控有关, 但是这并不意味着每个寡核苷酸个体就是加快转录效率的元件, 还需要在序列中作系统分析. 于是, 我们在所分析的序列中将表 3 和表 4 中的寡核苷酸全部显示出来, 由于重叠或连接, 它们在序列中形成很多较长的寡核苷酸片段. 不过在所分析的三组 DNA 序列 (第一组、第二组内含子和外显子) 中, 重叠或连接所得的寡核苷酸片段的宽度差异较大. 第一组内含子中有比较多的长寡核苷酸, 有些长达 30 多个碱基; 而在第二组内含子和外显子中, 重叠或连接所得的寡核苷酸一般都不太长. 从所显示出的核苷酸在序列中的分布密度看, 在第一组内含子中的平均密度最高, 为 53%, 在第二组内含子中为 44%, 在外显子中仅为 28%. 在第一组内含子的序列中, 有许多由 A、T 组成的长寡核苷酸, 同时有较多含 G、C 的四或五核苷酸分布在序列中, 有的嵌入在长寡核苷酸中 (图 1). 在第二组内含子序列中, 也有较多由 A、T 组成的寡核苷酸, 但是含 G、C 的四或五核苷酸则比较稀少. 在外显子序列中, 含 G、C 的四或五核苷酸略多一些, 但是由 A、T 构成的寡核苷酸则很稀少.

5′-gtatgtcaAGAATATTATATgTTTAgtaccagaTGAAAcaAGAATGAgtTGAAATTCAAAATGcgcgGAAAATcCATTATgatcaaggGAAAT
TGAATATGATAAATGAATTATTCAGAAAgtcacAATTAccagattgGAATGAACTGtcaagcTGAAAagcaactaaggATTAcCATTAcgtgTTT
ATgATTTTgttaGAAAAATATTGAAAacttGATATTTCTTTCcgctagTTTCAtcttTTTCAATTAcTTAAccggatcAAAATTTTtTTTActaa
caacATAACTATTTTTATTAcag-3′

Fig. 1 The intron sequence of ykl180w

The tetranucleotides and pentanucleotides in **Table 3** and **Table 4** are capitalized. The overlapping or connecting of them form wider oligonucleotides.

3 讨 论

大量实验研究表明, 转录调控是许多因子协同作用的过程[11], 这也就是说 DNA 上必须有多个转录因子的结合位点同时存在, 而且转录活性与结合位点的多寡似乎成正相关. 例如, Carey 等[12] 对酵母激活蛋白 GAL4 结合位点的研究显示, 带有 2 个和 5 个结合位点模板的转录活性分别是 1 个结合位点的 3.6 倍和 12.5 倍. 对基因上游调控转录机理的研究表明, 在多数情况下, 一定数量及长度的 TATA 元件是调控转录的基础, 此外, 许多含 G、C 的元件如 GATA、CCAAT 等也是转录调控的关键因素. 而且结合位点越多, 转录的效率也会越高. 实验表明, TATA 盒不单是在基因上游起作用, 在基因下游的 TATA 盒也有调控转录的作用[13], 内含子中的 GATA 也可成为转录因子的作用位点[3]. 这些现象说明, 基因内的转录调控机制与基因上游的调控机制有很大的相似性. 据此, 我们来考察所分析的三组序列增强基因转录的可能性.

本文结果表明, A、T 构成的寡核苷酸在外显子序列中出现得少且短, 虽然有一些含 G、C 的寡核苷酸 (这里所指的寡核苷酸均指表 3、表 4 中列出的, 下同), 但是它们在序列中所占的比例还是比较低, 仅为 28% (包括 A、T 型). 第二组内含子中, A、T 构成的寡核苷酸从分布密度上看比外显子的高一些, 但由于这类内含子序列比较短, 所以 A、T 构成的寡核苷酸的绝对数量和长度都有限, G、C 构成的寡核苷酸也少. 因此外显子和第二组内含子都难以起到提高转录效率的作用. 只有碱基序列普遍较长的第一组内含子, A、T 构成的寡核苷酸很丰富, 并且多数较长, 含 G、C 的寡核苷酸也比较多, 它们中多数还是实验证实了的转录元件. 这样的序列结构应该有利于转录因子的协同作用, 提高基因的转录效率.

在进行四核苷酸的比较时, 还有 6 对四核苷酸

的 u_1 值大于 1.64，但是 u_2 值小于 1.64，它们是 GAGA (TCTC)、AGAC (GTCT)、GACA (TGTC)、ACCA (TGGT)、CGTG (CACG) 和 GTCG (CGAC). 这说明它们在第一组内含子中的出现频率显著高于在第二组内含子中的出现频率，而在外显子中的出现频率与在第一组内含子中的出现频率差异不大. 与表 3 中的四核苷酸相比，它们的 G、C 含量较高，其原因部分可归结为内含子与外显子序列的碱基成分的差别. 在第一组内含子中，f_A= 0.34, f_T= 0.33, f_G= 0.17, f_C= 0.16; 而在外显子中，f_A= f_T= 0.29, f_G= f_C= 0.21. 两种序列中，A、T 的含量均比 G、C 高，但是外显子中 G、C 含量相对高一些，这正是酵母基因组的特点. 实验表明 GAGA (TCTC) 与转录调控有关[14]. 若将 GACA 看成 GAGA 的变形，再注意到表 3 中的 GATA 和 GAAA，可得到一致序列 GANA (N 代表任意碱基)，这也是一个值得注意的结果. 虽然以上这些含 G、C 较多的四核苷酸在外显子中的分布较多，可能由于没有足够的 A、T 构成的寡核苷酸序列，外显子还是难以起到调控转录的作用. 但是，它们如果出现在内含子中，就有可能起到增强转录的作用.

综上所述，第一组内含子 (亦即转录频率较高的基因内含子) 确实具有有利于基因转录的两个结构特征：a. 含有大量 A、T 组成的类似 TATA 元件的寡核苷酸，其中一些还比较长；b. 具有较多包含 G、C 的潜在的转录因子结合位点，其中一些是已被实验证实的调控元件，如 GATA、CATG、TGAA 等.

最后说明一点，在所分析的两组内含子中，虽然都含有大量全由 A 或 T 构成的寡核苷酸，这是内含子的一个普遍特征，但是它们的 u_1 和 u_2 值均小于 1.64，这似乎说明内含子的这个结构特征与转录调控无关，同时也说明了本文所用分析方法的有效性.

致谢 本文在完成过程中得到云南大学现代生物中心曹槐教授和柳维波助理研究员的帮助. 特此致谢!

参 考 文 献

1 Mattick J S. Introns: evolution and function. Curr Opin Genet Dev, 1994, **4** (6): 823~ 831

2 Brinster R L, Allen J M, Behringer R R, *et al*. Introns increase transcriptional efficiency in transgenic mice. Proc Natl Acad Sci USA, 1988, **85** (3): 836~ 840

3 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

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

5 Clement J Q, Wilkinson 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

6 Holstege F C, Jennings E G, Wyrick J J, *et al*. Dissecting the regulatory circuitry of a eukaryotic genome. Cell, 1998, **95** (6): 717~ 728

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

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

9 Kohler J, Schafer-Preuss S, Buttgereit D. Related enhancers in the intron of the betal tubulin gene of *Drosophila melanogaster* are essential for maternal and CNS-specific expression during embryogenesis. Nucleic Acids Res, 1996, **24** (13): 2543~ 2550

10 Moreno-Herrero F, Herrero P, Colchero J. Analysis by atomic force microscopy of Med8 binding to *cis*-acting regulatory elements of the SUC2 and HXK2 genes of *Saccharomyces cerevisiae*. FEBS Lett, 1999, **459** (3): 427~ 432

11 薛 文, 王 进, 黄启来, 等. 真核基因转录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

12 Carey M, Lin Y S, Green M R, *et al*. A mechanism for synergistic activation of a mammalian gene by GAL4 derivatives. Nature, 1990, **345** (6273): 361~ 364

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

14 Handen J S, Rosenberg H F. Intronic enhancer activity of the eosinophil-derived neurotoxin (RNS2) and eosinophil cationic protein (RNS3) genes is mediated by an NFAT-1 consensus binding sequence. J Biol Chem, 1997, **272** (3): 1665~ 1669

Statistical Analysis of Sequence Features of Introns With Positive Transcriptional Regulation in Yeast Genes*

ZHANG Jing**
(*The Center of Applied Statistics, Yunnan University, Kunming* 650091, China)
SHI Xiu-Fan
(*Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, The Chinese Academy of Sciences, Kunming* 650223, China)

Abstract A great deal of experimental studies have shown that many introns of eukaryotic genes function as regulators of transcription. However, comprehensive studies of this problem have not yet been conducted. After checking the transcription frequencies of some *Saccharomyces cerevisiae* (yeast) genes and their introns, a remarkable phenomenon was discovered that generally the introns of the genes with higher transcription frequencies are longer, and the introns of the genes with lower transcription frequencies are shorter. This suggests that the longer introns of genes with higher transcription frequencies may contain some characteristic sequence structures, which could enhance the transcription of genes. Therefore, two sets of introns of yeast genes were chosen for further study. The transcription frequencies of the first set of genes are higher (> 30), and those of the second set of genes are lower ($\leqslant 10$). Some oligonucleotides are detected by statistically comparative analyses of the occurrence frequencies of oligonucleotides (mainly tetranucleotides and pentanucleotides), whose occurrence frequencies in the first set of introns are significantly higher than those in the second set of introns, and are also significantly higher than those in the exons flanking the introns of the first set. Some of these extracted oligonucleotides are the same as the regulatory elements of transcription revealed by experimental analyses. Besides, the distributions of these extracted oligonucleotides in the two sets of introns and the exons show that the sequence structures of the first set of introns are favorable for transcription of genes.

Key words yeast, intron, transcription frequency, occurrence frequency of oligonucleotide

* This work was supported by a grant from The Ministry of Education of China for Key Program in Science Researches (00246).
** Corresponding author. Tel: 86-871-5036207, E-mail: zhangjing@ynu.edu.cn
Received: September 28, 2002 Accepted: December 30, 2002