酵母内含子在基因序列中的分布对基因转录效率的影响*

张 静[1,2)]** 石秀凡[3)] 杨恒芬[1)]

（[1)]云南大学应用统计中心，昆明 650091；[2)]云南财贸学院计算机科学系，昆明 650221；
[3)]中国科学院昆明动物研究所，中国科学院细胞与分子进化重点实验室，昆明 650223）

摘要 对酵母中高效转录和低效转录基因内含子序列寡核苷酸使用情况的对照分析，显示两类内含子的序列结构有差异，并且高效转录基因内含子序列含有较多潜在的转录因子结合位点，由此推测内含子可能参与基因转录的调控. 这个结论有待更多的数据证实. 对内含子和外显子在两组基因序列中的分布（长度、位置等）进行详细比较分析后显示，高效转录基因内含子和低效转录基因内含子的长度有比较明显的界限. 两组基因中外显子长度的均值虽然有些差异，却没有明显的界限. 基因序列长度与外显子长度的情况相似. 虽然内含子的相对位置在两类基因中都很靠近5′端，但是从实际位置看，高效转录基因中比较多的内含子很靠近基因的5′端，有些则位于5′-UTR区域. 这些结果提示，基因的转录效率与内含子的长度有关，与外显子及基因序列的长度无关，内含子的位置也可能影响转录效率，内含子对基因转录的调控可能与基因上游的转录调控有关联，或者是上游调控的延续.

关键词 酵母，基因转录频率，内含子长度，内含子位置

学科分类号 Q61

虽然很多实验研究表明，真核基因的某些内含子具有调控基因转录的功能[1~3]，然而对其调控机制和此类内含子的特征还缺乏全面和系统的认识. 在前面的工作中，我们对高效转录和低效转录的酵母基因内含子序列的寡核苷酸使用情况进行过统计对照分析，结果表明，两组基因内含子的序列结构有较大的差异，同时还探测到一批寡核苷酸，它们在第一组（高效转录基因）内含子中出现的频率显著高于在第二组（低效转录基因）内含子以及外显子中出现的频率，而且它们中的一些正是实验研究揭示的转录因子结合位点. 由此我们推测这些高效转录基因内含子的序列结构可能有利于基因的转录，对基因转录有正调控的作用[4]. 当然，这还只是内含子序列结构特征的反映. 要证实这个结论，还有很多问题需要研究. 首先要考虑的是基因序列中的其他部分（比如外显子）是否也具有调控作用. 虽然在文献［4］中也将内含子和外显子做过比较，但主要是就内含子和外显子序列组成考虑的，还有必要对其他方面进行研究. 其次，如果内含子确实调控了基因的转录，那么其调控机制如何更值得研究.

Sakurai 等[5]曾对一些真核基因内含子的位置分布做过分析，结果表明，单内含子基因的内含子偏向基因的5′端，酵母基因中的偏向尤为显著，并认为这种偏向可能与转录有关. 不过，他们考虑的仅是内含子的相对位置. 事实上，如果考虑内含子的实际位置（或称绝对位置），并结合基因的转录效率进行分析，结果并不是如此简单. 本文对两类内含子在基因序列中的分布（长度、位置等）情况进行详细的比较分析. 结果显示：两组基因内含子的长度有较明显的界限，而外显子和基因序列却没有. 从内含子在基因中的实际位置看，高效转录基因的内含子较之低效转录基因内含子更倾向基因的5′端，甚至是5′-UTR区域. 这些结果进一步支持了酵母内含子参与基因转录调控的推测，也提示内含子的转录调控与基因上游的转录调控存在一定程度的关联性或连续性.

1 样本和方法

1.1 样本

为了保持研究结果的系统性，本文仍用文献［4］中的两组基因样本进行分析. 需要说明的是，起初我们以30和10作为“高频率”和“低频率”的阈值只是凭直觉来确定的，后经查验得知，高频组（第一组）中除ybr084w和yfl039c外，其余74

* 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：2003-05-14，接受日期：2003-06-28

个均是编码核糖体蛋白的基因. 而低频组（第二组）已确定生物功能的基因中未有核糖体蛋白基因. 迄今知道酵母有137个核糖体蛋白基因[6]. 由此可见，我们选取的样本与生物学功能是吻合的.

1.2 分析方法

1.2.1 内含子、外显子及基因长度的比较分析：根据EMBL数据库提供的DNA序列数据，统计出两组基因中每个样本的内含子和外显子碱基数，即内含子和外显子的长度. 如果一个基因中有两个内含子，则内含子长度为两个内含子的长度之和. 外显子长度指的是所有外显子的长度之和. 内含子相对长度为内含子长度与基因长度（内含子与外显子长度之和）的比值.

然后将两组样本基因转录频率与内含子、外显子及基因长度的关系用散点图显示，比较两组点的分布情况，并分析两组样本内含子、外显子及基因长度的均值和标准差.

1.2.2 内含子位置的比较分析：内含子在基因中的绝对位置以内含子（第一个内含子）到起始密码子的距离（即第一个外显子的长度 E_1）来测定. 如果内含子位于5′-UTR区域，内含子在基因中的绝对位置定义为负值，其绝对值为内含子的第一个碱基到起始密码子的距离. 内含子在基因中的相对位置定义为 $E_1/(E_1+E_2)$ 或 $E_1/(E_1+E_2+E_3)$（图1）.

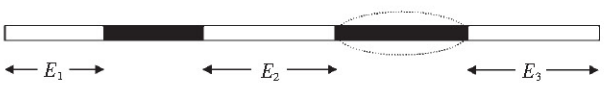

Fig. 1 The relative position of intron in gene is defined with the lengths of exons: $E_1/(E_1+E_2)$ (one intron) or $E_1/(E_1+E_2+E_3)$ (two introns)

□: exon; ■: intron.

这样定义的内含子相对位置与Sakurai等[5]的定义相同，主要是为了与其结果相对照. 如果内含子位于5′-UTR区域，内含子在基因中的相对位置定义为内含子绝对位置与外显子长度的比.

统计两组样本的内含子绝对位置和相对位置在各区间段上的分布情况，并以直方图形式进行对照分析.

2 结 果

2.1 内含子、外显子及基因长度

将内含子的长度与基因转录频率的关系用散点图表示出来（图2）. 可以看出，除少数几个样本外，两组内含子的长度比较明显地以300 bp作为分界（图2a），而且第二组内含子的长度主要集中分布在200 bp以下. 两组内含子相对长度的分界值大约为30%（图2b）. 两组基因外显子的长度没有较明显的分界线（图3），只是第二组基因外显子长度比第一组的变化幅度较大一些. 第一组外显子的长度比较集中，除了一个长度为1 128 bp外，其余都不超过1 000 bp. 而第二组外显子中有不少长度超过1 000 bp，其中有几个长度还在2 000 bp以上. 基因长度（内含子与外显子长度之和）的情况与外显子的情况相似，两组基因的长度没有明显的界限（图4）. 第一组基因的长度比较集中，而第二组基因的长度变化较大一些，这主要是由相应的外显子长度变化较大所致.

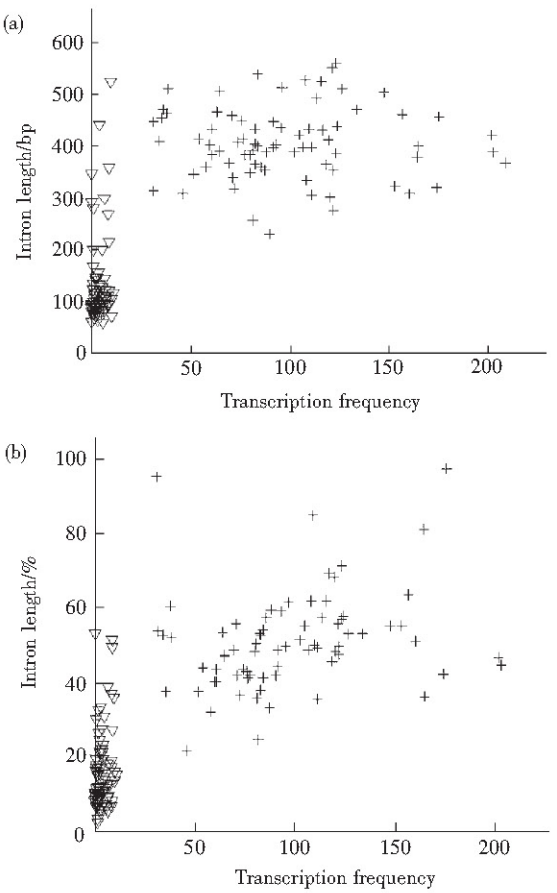

Fig. 2 The scatter plots of intron lengths vs. transcription frequencies of genes

+ and ▽ represent Set Ⅰ genes and Set Ⅱ genes respectively, similarly hereinafter. The length in (a) is the actual length of intron(s), i.e., the base number of intron(s). The length in (b) is the relative length of intron(s), i.e., the base number of intron(s) / the base number of intron(s) and exon (s).

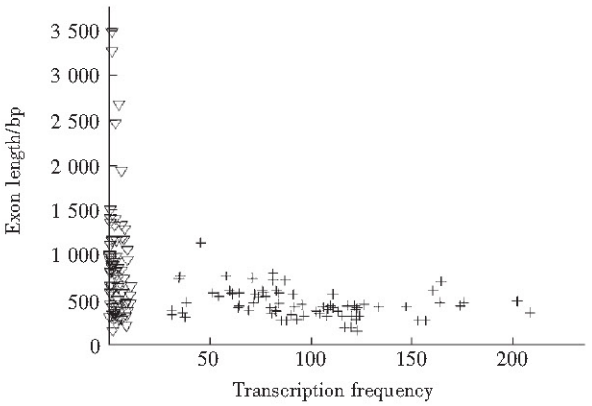

Fig. 3 The scatter plots of exon lengths vs. transcription frequencies of genes

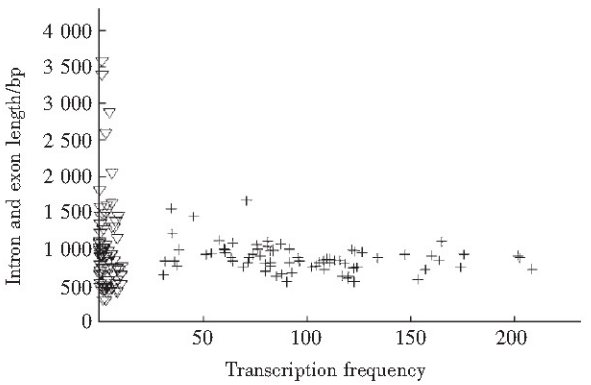

Fig. 4 The scatter plot of gene lengths（length of intron and exon）vs. transcription frequencies of genes

表1列出了两组样本的内含子、外显子及基因长度的均值和方差情况. 从中可以清楚地看出第一组内含子长度和相对长度的均值都远大于第二组内含子的相应值，而外显子的情况则相反，两者互补的结果是内含子与外显子之和相差不大. 但是无论哪项指标，第二组样本的标准差都很大，第一组样本的较小. 这也说明，高效转录基因的内含子、外显子乃至基因的长度都比较接近. 第二组内含子和外显子长度变化幅度较大的原因是有几个非常大的值，除去这几个特异值后计算均值和标准差（表中括号内的数值），标准差虽然降低了很多，但是与第一组样本相比，还是比较大.

Table 1 The means and standard deviations of the lengths of introns，exons and genes in two sets of samples

		Intron length /bp	Intron relative length /%	Exon length /bp	Intron & exon length /bp
Set Ⅰ	$\bar{x}$	412	49	460	878
	s	92	10	163	201
Set Ⅱ	$\bar{x}$	132（109）	16	856（712）	994（827）
	s	85（41）	11	625（338）	631（445）

The numbers in brackets are the values computed except several samples with abnormal lengths.

2.2 内含子在基因中的位置

内含子在两组基因中绝对位置和相对位置的分布情况见图5. 从相对位置看，两组样本的内含子分布情况差不多，都非常倾向基因的5′端，这与Sakurai等[5]的结果完全相同. 然而绝对位置的情况却有所不同，从图5a明显可以看出，第一组基因中与起始密码子的距离不超过20 bp的内含子个数比第二组基因的多，然而与起始密码子距离20 bp以上的情况却相反，第二组基因内含子的数目要比第一组的多. 第一组内含子中有7个位于5′-UTR区，而第二组中却没有位于5′-UTR区的内含子. 位于5′-UTR区7个内含子的绝对位置均小于－320 bp（即绝对值在320 bp以上）；除了ybl072c的相对位置为－0.5，其余6个相差不大，在－1.3至－0.9之间，而且其中4个样本的相对位置均为－1. 这几个样本中，内含子3′端与外显子5′端之间的碱基数目都不多，最多的只有15 bp. 所以它们相对位置的绝对值在1附近，意味着内含子长度与外显子长度相差不大. 事实上，第一组基因的绝大多数样本都有这个特点，亦即内含子长度与外显子长度相差不大.

3 讨 论

从直观感觉上看，一般会认为基因的长度与其转录频率成反比. 然而，从**2.2**的结果看，两组基因序列长度的均值相差不大，特别是将第二组基因

中几个特别长的基因去除后，所得基因长度的均值还比第一组基因的小，只是标准差较大. 所以基因的长度不是决定基因转录快慢的原因.

从基因序列的内容看，内含子和外显子是两个主要部分. 就外显子而言，虽然两组基因的外显子长度均值相差较大，但是从图3上看，它们没有明显的界限，第二组外显子长度均值大的原因是这组外显子长度分布不均匀（标准差较大），而第一组外显子的长度分布比较均匀. 从外显子的序列组成看，将外显子拼接形成编码序列，根据酵母数据库（http://genome-www. stanford. edu/Saccharomyces/）提供的密码子使用的几个指数（Codon Bias，Codon Adaptation Index，Frequency of Optimal Codons），第一组基因编码序列的各项指数都比较高，而第二组基因编码序列的几项指数都较低. 但是两组编码序列所表现出的这些性质差异应该是在翻译的调控中起作用. 事实上，一些研究表明，编码序列密码子的偏倚使用对翻译速度有直接影响[7]. 在转录过程中，由于基因序列一般都含有内含子片段，被内含子分割的外显子片段虽然在碱基使用上有其特异性，但是上述指数特征不一定能表现出来. 由此看来，外显子调控转录的可能性也不大.

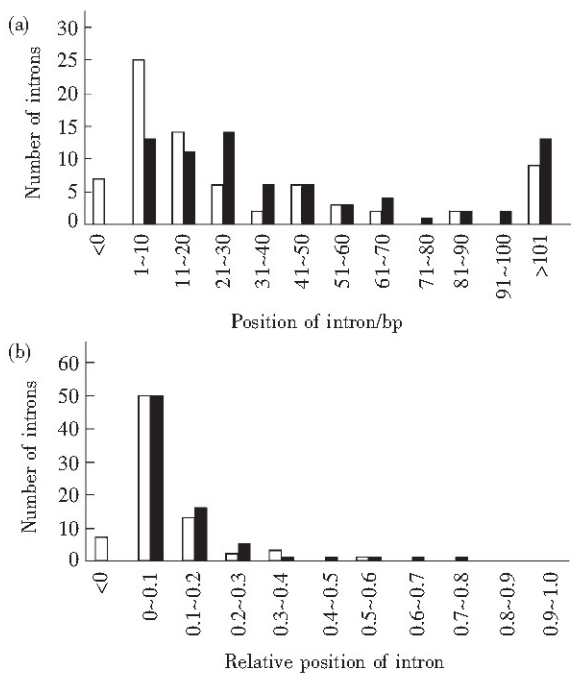

Fig. 5　The distribution of position of introns in genes

(a) The distribution of the actual position, i. e., the distance between the 5′-end of the first intron and the start codon. (b) The distribution of the relative position, i. e., the actual position / the length of exons. □: Set Ⅰ genes; ■: Set Ⅱ genes.

再看内含子的情况，表1显示，无论是相对长度还是绝对长度，第一组内含子的平均长度都远大于第二组内含子的平均长度，而且第一组的标准差和相对均值也比第二组的小得多. 图1更清楚地表明两组基因内含子的长度分布有着很显著的差异. 我们先前对两组内含子寡核苷酸使用情况的分析已显示，两组内含子的序列特征存在差异[4]，第一组内含子序列中含有许多潜在的转录因子结合位点，第二组内含子中的潜在位点却比较稀少. 这也提示，高效转录基因内含子较长的原因，可能是需要这些内含子提供充足的位点供转录因子结合.

综合以上分析，可以确信酵母基因的转录频率在很大程度上受内含子调控. 余下的问题是内含子的调控是如何起作用的. 对此，我们来考察内含子在基因中所处的位置，虽然两组基因内含子的相对位置都非常靠近基因的5′端（图5b），但是绝对位置却有差异（图5a）. 内含子越靠近基因的5′端，它与上游调控位点的距离越近，从而增加了两者协同作用的可能性. 而实验研究已证实，转录因子间的协同作用能极大地提高基因的转录效率[8,9]. 从这一方面看，第一组基因具有较高转录频率的部分原因很可能是内含子与上游调控序列的关联作用. 当然要彻底弄清内含子调控基因转录的机制，还需要做更多深入细致的研究.

参　考　文　献

1　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

2　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

3　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

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）：213 ~ 238
Zhang J, Shi X F. Prog Biochem Biophys, 2003, **30** (2): 213 ~ 238

5　Sakurai A, Fujimori S, Kitamura-Abe S, *et al*. On biased distribution in various eukayotes. Gene, 2002, **300** (1 ~ 2): 89 ~ 95

6　Warner J R. The economics of ribosome biosynthesis in yeast. Trends Biochem Sci, 1999, **24** (11): 437 ~ 440

7　Krüger M K, Pedersen S, Hagervall T G, *et al*. The modification of the wobble base of $tRNA^{Glu}$ modulates the translation rate of glutamic acid codons *in vivo*. J Mol Biol, 1998, **284** (3): 621 ~ 631

8　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

9　Carey M, Smale S T. Transcriptional Regulation in Eukaryotes, Concepts, Strategies and Techniques. New York: CSHL Press, 2002, 5 ~ 42

Transcription Rates of Yeast Genes Are Influenced by The Distribution of Introns*

ZHANG Jing[1,2)**], SHI Xiu-Fan[3)], YANG Heng-Fen[1)]

([1)] *The Center of Applied Statistics, Yunnan University, Kunming* 650091, China;

[2)] *Department of Computer Science, Yunnan University of Finance and Economics, Kunming* 650221, China;

[3)] *Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, The Chinese Academy of Sciences, Kunming* 650223, China)

Abstract A comparative analysis on the intron sequence oligonucleotide usages in two sets of yeast genes with higher and lower transcription frequencies, respectively, has shown that the intron sequence structures of the two sets of genes are different. There are more potential binding sites for transcription factors in the introns of the genes with high transcription frequencies. So it is speculated that introns regulate the transcription of genes. But more evidences are needed to favor this speculation. The detailed comparative analyses on the distribution (length and position) of introns and exons in the two sets of gene sequences also show that there is an obvious boundary between the lengths of the two sets of introns. There is no boundary between the lengths of the two sets of exons, although the means of their lengths are of discrepancy. The situation of the gene lengths (length of intron and exon) is similar to exon lengths. As far as the relative position, the introns in two sets of genes all have a bias toward the 5′ ends of genes. But as the actual position is considered, more introns in high transcription genes have a tendency to be located toward the 5′ ends of genes, some even located at 5′-UTR. These results suggest that the gene transcription rates are related to the length of intron, but not to the lengths of exons and genes sequences. The positions of introns may also influence the transcription rates. The transcriptional regulation of introns may be correlative with the transcriptional regulation of the upstream of genes, or be its continuous action.

Key words yeast, transcription frequency of gene, length of intron, position of intron

---

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

** Corresponding author. Tel: 86-871-5036207, E-mail: zhangjing@ynu.edu.cn

Received: May 14, 2003 Accepted: June 28, 2003