真核生物启动子 TATA-box ·GC-box 和 CAAT-box 的分析

张小辉, 祁艳霞 (河南科技大学动物科技学院, 河南洛阳 471003)

摘要　下载了已经通过试验证实的2 541 条真核生物启动子序列, 运用生物信息学方法分析了 TATA-box、GG-box 和 CAAT-box 的数量及其在启动子中的分布情况。结果表明, 有23 .85 % 真核生物启动子序列中至少有1 个TATA-box , 且 TATA-box 主要分布在转录起始位点前24 ～36 bp 的区域内; 有47 .30 % 真核生物启动子序列中至少有1 个GG-box , 且 GG-box 在转录起始位点前23 ～128 bp 的区域内分布比较集中; 有42 .35 % 真核生物启动子序列中至少有1 个 CAAT-box , 且 CAAT-box 在转录起始位点前51 ～159 bp 的区域内分布比较集中。这说明 TATA-box 在真核生物启动子中的位置比较固定, 对基因转录的正确起始可能起着重要的作用; 而 GG-box 和 CAAT-box 分布区域较广, 数量也明显多于 TATA-box 。

关键词　真核生物; 启动子; TATA-box ; GG-box ; CAAT-box
中图分类号　Q789　　**文献标识码**　A　　**文章编号**　0517 - 6611(2008) 04 - 01380 - 02

Analysis on TATA-box , GG-box and CAAT-box in Eukaryotic Promoters

ZHANG Xiao-hui et al　(College of Animal Science and Technology , Henan University of Science and Technology , Luoyang , Henan 471003)

Abstract　2541 eukaryotic promoter sequences that had been validated by test were downloaded and the number of TATA-box , GG-box and CAAT-box and its distribution conditions in the promoters were analyzed by bioinformatics method . The results showed that 23 .85 % eukaryotic promoter sequences had at least one TATA-box and TATA-box distributed mainly in the area 23 ～128 bp ahead of transcription initiation site . 47 .30 % eukaryotic promoter sequences had at least one GG-box and GG-boxes were concentrated in the area 23 ～128 bp ahead of transcription initiation site . 42 .35 % eukaryotic promoter sequences had at least one CAAT-box and CAAT-boxes were concentrated in the area 51 ～159 bp ahead of transcription initiation site . The results indicated that the position of TATA-box in eukaryotic promoters was relatively fixed and it might play an important role in the correct initiation of gene transcription . But GG-box and CAAT-box were widely distributed and their number was obviously more than that of TATA - box .

Key words　Eukaryote ; Promoter ; TATA-box ; GG-box ; CAAT-box

启动子是位于结构基因5 '端上游的1 段 DNA 序列, 能够指导全酶同模板正确结合, 活化 RNA 聚合酶, 启动基因转录。其中, Ⅱ类启动子主要负责 mRNA 的转录。而 mRNA 是近年来分子生物学研究的热点之一, 对Ⅱ类启动子的研究较为透彻。Ⅱ类启动子主要包括 TATA 框(TATA-box) 、起始子、GGGCGG 框(GG-box) 、CAAT 框(CAAT-box) 等, 其中 TATA 框和起始子又称为核心启动子[1] 。TATA 框一般位于 - 26 ～ - 34 bp 处, 绝大多数在(- 31 ±2) bp 处。已知作用于 TATA 框的转录因子有6 种, 分别为 TFⅡA 、TFⅡB 、TFⅡD 、TFⅡE 、TFⅡF 和 TFⅡH, 从酵母到人类这6 种转录因子的序列都是十分保守的[2] 。TFⅡD 的主要成分是 TBP(TATA-Binding Protein) , 人 TBP 结合 TATA 序列的能力比结合非特异 DNA 的能力约高1 000 倍。TBP 一旦结合 DNA 后, 可以沿着 DNA 单向滑动, 根据滑动速度的不同可判断是否到达目的地[3 - 4] 。GC 框一般位于 TATA 框上游, 是调节蛋白 SP1 的结合之处。与 CAAT 框结合的蛋白称为 CBF。由 CBF-A 、CBF-B 和 CBF-C 组成, CBF-A 和 CBF-C 先组成异源二聚体, 然后吸引 CBF-B 形成三聚体才能结合 CAAT 序列。

随着人类基因组计划的完成, 各种生物学数据库相续建立, 有些是收录核酸、蛋白序列的一级数据库, 更多的是对一级数据库中的序列进行归纳、整理而形成的二级数据库。EBI 就是专门收集通过试验证实的启动子序列的二级数据库。笔者下载了 EBI 数据库中已经通过试验证实的2 541 条真核生物的启动子序列, 利用生物信息学的方法研究了这些启动子中 TATA-box 、GG-box 和 CAAT-box 的分布情况, 为脊椎动物启动子结构研究提供有力依据。

作者简介　张小辉(1978 -) , 男, 河南洛阳人, 讲师, 从事动物遗传育种与繁殖方面的研究。

收稿日期　2007-09-01

1　材料与方法

1 .1　启动子的定义及序列获取　参阅参考文献[5] , 将转录起始位点前599 bp 到转录起始位点后100 bp 的区域(- 599 ～100 bp) 作为基因的启动子序列。启动子序列从 EBI 网站下载(http :// www.epd .isb-sib .ch/ seq_download .html) 。这些序列都是通过试验的方法确定的真核生物启动子序列, 共计2 541 条, 以 fasta 格式保存备用。

1 .2　TATA-box 、GG-box 和 CAAT-box 的定义及模体识别　TATA-box 是转录因子 TBP 识别的 DNA 序列, 其序列格式为 TATAWAW (W 代表 A 或 T) ; GG-box 是转录因子 SP1 所识别的 DNA 序列, 其序列格式为 GGGCGG; CAAT-box 是转录因子 CBF 在 DNA 序列上的结合位点, 其序列格式为 CCAAT 。TATA-box 的模体识别可以通过在启动子序列上是否发现 TATAWAW 特征序列实现。同样, GG-box 和 CAAT-box 的模体识别也可以通过在启动子序列中检测 GGGCGG 和 CCAAT 序列而实现。

2　结果与分析

2 .1　真核生物启动子序列中 TATA-box 、GG-box 和 CAAT-box 的数量　在2 541 个真核生物启动子序列中有498 个基因的启动子序列中含有1 个 TATA-box , 占总启动子数的19 .60 % ; 另有74 个基因的启动子序列中含有2 个 TATA-box , 占总启动子数的2 .91 % ; 只有34 个基因的启动子序列中含有3 个或3 个以上的 TATA-box , 占总启动子数的1 .34 % ; 有1 935 个基因的启动子序列中没有 TATA-box , 占总启动子数的76 .15 % 。

在2 541 个真核生物启动子序列中, 有596 个基因的启动子序列中含有1 个 GG-box , 占总启动子数的23 .46 % ; 有312 个基因的启动子序列中含有2 个 GG-box , 占总启动子数的12 .28 % ; 有294 个基因的启动子序列中含有3 个或3 个以上的 GG-box , 占总启动子数的11 .57 % ; 有1 339 个基因的启动子序列中没有 GG-box , 占总启动子数的52 .70 % 。

在2 541 个真核生物启动子序列中,有736 个基因的启动子序列中含有1 个CAAT-box,占总启动子数的28 .96 %;有223 个基因的启动子序列中含有2 个CAAT-box,占总启动子数的8 .74 %;有117 个基因的启动子序列中含有3 个或3 个以上的CAAT-box,占总启动子数的4 .60 %;有1 465 个基因的启动子序列中没有CAAT-box,占总启动子数的57 .65 %。

2 .2　真核生物启动子序列中TATA-box、GG-box 和CAAT-box 的位置分布　TATA-box、GG-box 和CAAT-box 在真核生物启动子上的分布不是随机的,而是有规律的。TATA-box 主要分布在转录起始位点前30 bp 左右的区域内,在转录起始位点前- 36 ～- 24 bp 的区域内分布着262 个TATA-box,占总数的52 .61 %;而在其他区域,TATA-box 的分布比较均匀。TATA-box 在真核生物启动子上的分布见图1。

GG-box 主要分布在转录起始位点前- 128 ～- 23 bp 的区域内,共有980 个GG-box,占总数的43 .71 %,其中550 bp 左右的区域是GG-box 分布最为密集的区域,而在其他区域,GG-box 的分布比较均匀。GG-box 在真核生物启动子上的分布见图2。

CAAT-box 在真核生物启动子上的分布与GG-box 相似,主要分布在转录起始位点前- 159 ～- 51 bp 的区域内,共有595 个CAAT-box,占总数的37 .92 %,其中510 bp 左右的区域是CAAT-box 分布最为密集的区域;而在其他区域,CAAT-box 的分布比较均匀。CAAT-box 在真核生物启动子上的分布见图3。

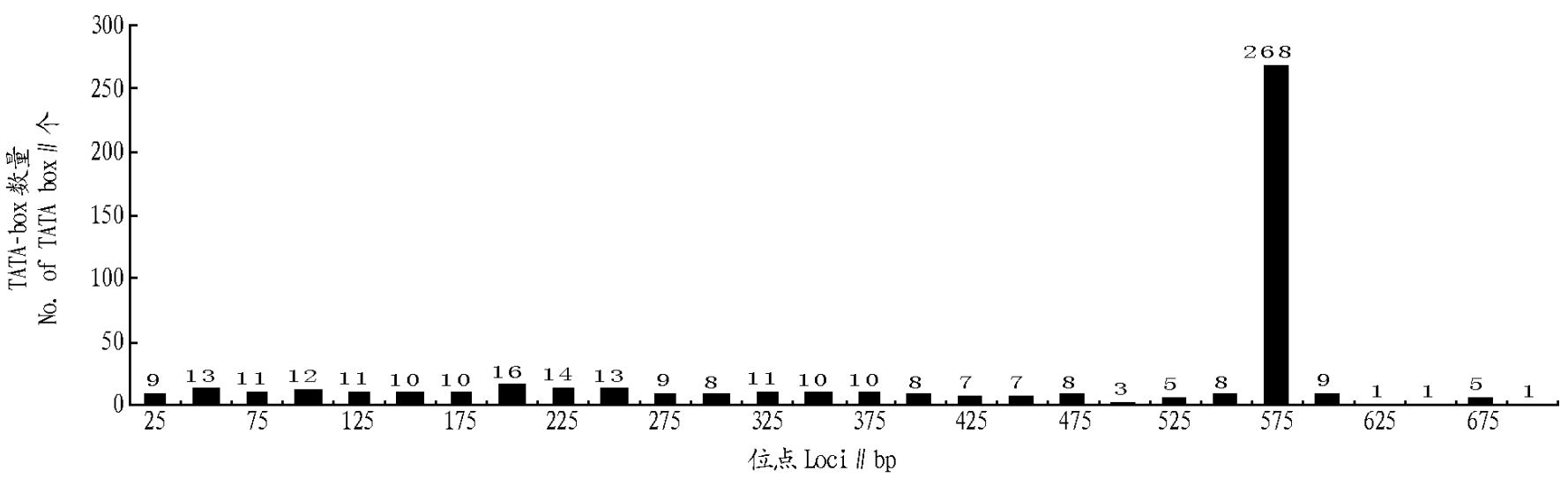

图1　TATA-box 在真核生物启动子中的分布

Fig .1　The distribution of TATA-box in promoters of eukaryotes

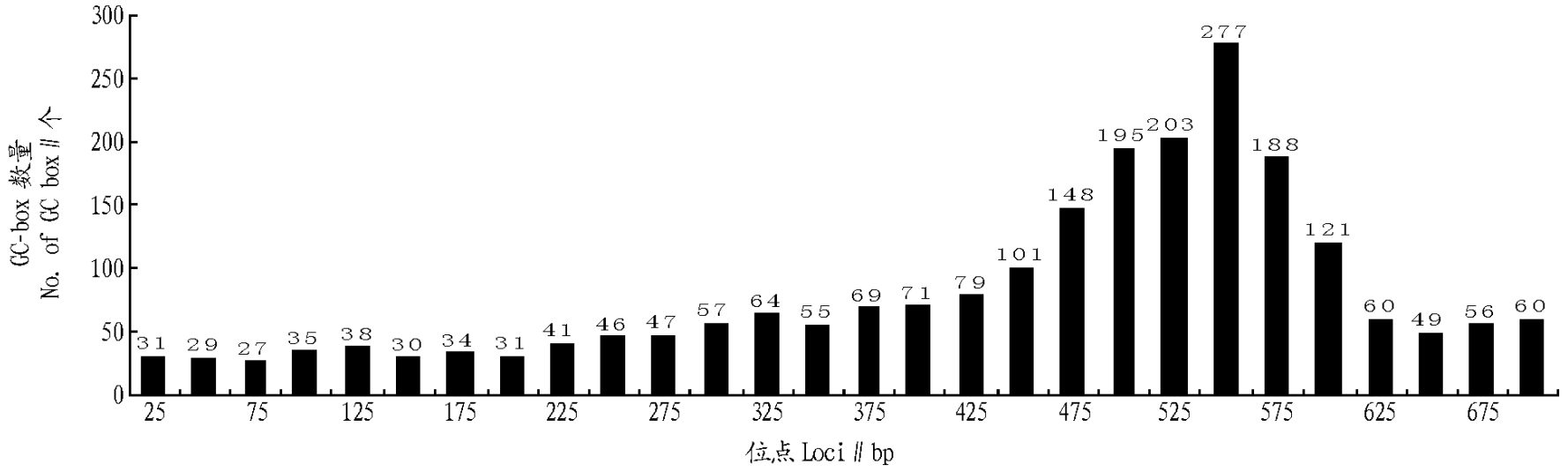

图2　GG-box 在真核生物启动子中的分布

Fig .2　The distribution of GG-box in promoters of eukaryotes

3　讨论

TATA-box 是第一个鉴定出来的真核生物启动子元件。它是通过将果蝇、哺乳动物和一些病毒结构基因的5'侧翼区进行同源性比对发现的。TATA-box 一般位于转录起始位点前25 ～30 bp 的区域内,其一致性序列为TATAWAW。TATA-box 在真核生物启动子内的位置不是一成不变的,如酵母TATA-box 有些就位于起始位点上游40 ～120 bp 的区域内[6]。TATA-box 一旦发生突变将导致转录效率降低,甚至转录被阻止。早期的研究认为,TATA-box 是所有真核生物结构基因所必需的。随着基因组大规模测序计划的完成,越来越多的真核生物基因序列被测定,发现许多基因的5'侧翼区没有TATA-box。对果蝇205 个核心启动子中TATA-box 的分析表明,有43 % 的结构基因的启动子中有TATA-box[7]。另有一项对1 941 个潜在的果蝇启动子分析表明,有33 % 的基因启动子中有TATA-box[8]。对人1 031 个潜在的核心启动子分析表明,有32 % 的结构基因存在TATA-box[9]。笔者通过对2 541 个真核生物启动子中TATA-box 的分析,发现有23 .85 % 的真核生物结构基因的启动子序列中至少含有1 个TATA-box。在不同的研究中对TATA-box 的定义稍有差异。如,对果蝇启动子TATA-box 的研究中,TATA-box 定义为与TATAAA 相差不超过1 个碱基的序列,这样就加大了搜索的范围,导致搜索到的TATA-box 数量增加。笔者将TATA-box 定义得更为严格,也更为客观,是导致搜索到的TATA-box 数量相对较少的重要原因。而TATA-box 主要分布在转录起始位点前- 36 ～- 24 bp 的区域内,与前人的大多数研究相符[10]。

GG-box 和CAAT-box 也是真核生物启动子的重要调控元件,主要分布在转录起始位点及其前面的150 bp 范围内。

(下转第1395 页)

芽,芽的长度总体来说不是最长的。

2.2　重量损失　由表3 可知,处理④的马铃薯重量损失最多,重量损失率达11.11 %,为6 个处理中最高,150 d 后薯块表面明显萎蔫,芽眼丛生,不能食用。对照的马铃薯重量损失为0.45 kg,重量损失率为5.00 %,大部分薯块表面有萎蔫现象。处理①、②、③的马铃薯重量损失和重量损失率均较小,三者差异不是很显著,处理③略高,但薯块外观仍较新鲜,表面光滑,没出芽,水分损失小。处理⑤的马铃薯重量损失和重量损失率与对照基本相当,薯块表面萎蔫,水分损失较大。处理⑥的处理效果也不好,重量损失率仅比对照低1.33 %,薯块外观很不好,已失去使用价值。

表3　贮藏期各处理马铃薯块茎的重量损失

Table 3　Weight loss of potato tuber in each treatment at storage

处理 Treatment	30 d 后重量 Weight after 30 d kg	60 d 后重量 Weight after 60 d kg	150 d 后重量 Weight after 150 d kg	平均重量 Average weight kg	重量损失 Weight loss kg	重量损失率 Loss rate %
①	9.00	8.91	8.91	8.94	0.06	0.67
②	9.00	8.94	8.94	8.96	0.04	0.44
③	9.00	8.88	8.88	8.92	0.08	0.89
④	8.85	8.85	6.15	8.00	1.00	11.11
⑤	8.85	8.70	7.95	8.55	0.45	5.00
⑥	8.88	8.76	8.25	8.67	0.33	3.67
CK	8.88	8.70	7.95	8.55	0.45	5.00

2.3　成本核算　由表4 可知,通过与抑芽效果、重量损失等结果的综合分析,处理②的效价比最高,抑芽剂配合杀菌剂使用,成本低廉,每1 000 kg 仅3 元。处理①、③的抑芽效果也很好,在条件较差的库房可以发挥其优势,成本也不太高。

表4　各处理储存效果与使用成本

Table 4　The storage effect and cost in each treatment

处理 Treatment	成本 Cost /元·kg	储存效果 Storage effect	使用方便程度 Convenient degree for use
①	0.009	好 Good	很容易 Easy
②	0.003	好 Good	较容易 Relatively easy
③	0.018	好 Good	很容易 Easy
④	0.028	差 Bad	较容易 Relatively easy
⑤	0.043	差 Bad	不容易 Uneasy
⑥	0.045	较差 Worse	不容易 Uneasy

3　结论与讨论

在现有的条件下,使用成本低廉的抑芽剂就能解决马铃薯在储存过程中的水分与重量损耗、发芽以至于不能食用等问题。因为CIPC 抑芽剂属于高效、低毒、低残留药剂,安全系数较大,且使用方便,常温下仍可以使用,对湿度的要求也不高。马铃薯撒施该抑芽剂后,随时可以食用,对存放马铃薯的窖、室没有污染,对人体健康也没有影响,试验效果很好,可以考虑进行大规模试验。使用抑芽剂是解决当前马铃薯储存问题最有效的方法,但应注重与实际情况结合。

参考文献

[1] 田丰,张永成,师理,等. 马铃薯不同品系贮藏期品质分析[J]. 中国马铃薯,2006(1):19-23.

[2] 陈彦云,刘成敏,郑学平,等. 马铃薯抑芽剂研制效果试验[J]. 中国马铃薯,2001(5):284-285.

[3] 刘振业. 贵州马铃薯产业现状和发展优势与潜力[J]. 贵州农业科学,2005,33(3):5-8.

[4] 石建宁,候玉霞. 马铃薯贮藏病害化学防腐的研究进展[J]. 宁夏农林科技,1998(1):40-41.

[5] 纳添仓,阮建平,唐小兰,等. 马铃薯贮藏的方式与技术[J]. 青海农林科技,2002(3):34-35.

(上接第1381 页)

GG-box 的数量较多,可能与真核生物基因5'侧翼区GC 含量通常较高有关。

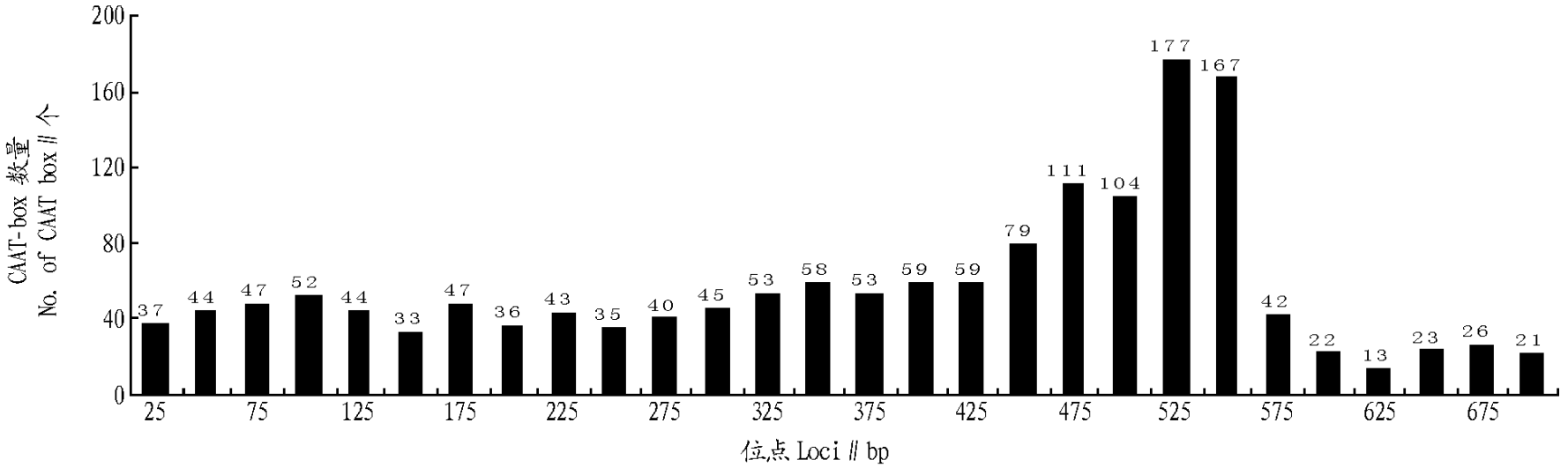

图3　CAAT-box 在真核生物启动子中的分布

Fig.3　The distribution of CAAT-box in promoters of eukaryotes

参考文献

[1] PEDERSEN A G, BALDI P, CHAUVIN Y, et al. The biology of eukaryotic promoter prediction[J]. Compute Chem, 1999, 23(6):91-209.

[2] 童克中. 基因及其表达[M]. 北京:科学出版社, 2001.

[3] MOQTADERI Z. TBP-associated factors are not generally required for transcriptional activation in yeast[J]. Nature, 1996, 383:188-190.

[4] CHEN J. Assembly of recombinant TFIID reveals differential coactivator requirements for distinct transcriptional activators[J]. Cell, 1994, 79:93-105.

[5] 王琦, 李伟鹏, 陈小燕. 基于计算机的启动子识别技术[J]. 医疗卫生装备, 2003(S1):204-206.

[6] STEPHEN T S, JAMES T K. The RNA polymerase II core promoter[J]. Annu Rev Biochem, 2003, 72:449-479.

[7] KUTACH A K, KADONAGA J T. The downstream promoter element DPE appears to be widely used as the TATA box in Drosophila core promoters[J]. Mol Cell Biol, 2000, 20(13):4754-4764.

[8] OHLER U, NIEMANN H, LIAO G C. Modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition[J]. Bioinformatics, 2001, 17:199-206.

[9] YUTAKA S, TATSUHIKO T, JUN S, et al. Identification and characterization of the potential promoter regions of 1031 kinds of human genes[J]. Genome Res, 2001, 11:677-684.

[10] GERALD M R, GUO C L, HEINRICH N. Computational analysis of core promoters in the Drosophila Genome[J]. Genome Biology, 2002, 3(12):101-112.