实验项目 4：启动子的分析和预测【应用】

一、分析平台：

1.1. **硬件平台**：（硬件配置）：win10：CPU 1.70GHz　2.40GHz

1.2. **系统平台**：（操作系统及其版本号）WIN10 专业版

1.3. **软件平台**:（软件系统及其版本号，若是在线分析平台，还需要提供 URL 地址）

Promoter：
http://www.cbs.dtu.dk/services/Promoter/
Tssw：
http://linux1.softberry.com/berry.phtml?topic=tssw&group=programs&subgroup=promoter
Tssg：
http://linux1.softberry.com/berry.phtml?topic=tssg&group=programs&subgroup=promoter
FPROM：
http://www.softberry.com/cgi-bin/programs/promoter/fprom.pl
primerblast：
https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi
cister
http://zlab.bu.edu/~mfrith/cister.shtml
p-match
http://gene-regulation.com/cgi-bin/pub/programs/pmatch/bin/p-match.cgi
NEBcutter
http://nc2.neb.com/NEBcutter2/
绘图使用：PicPick

1.4. **数据资源**：NCBI-GENEdatabase → “TOMM40”

NEB 内切酶产品具体 Buffer 活性表：
https://wenku.baidu.com/view/1accd41ba76e58fafab0033c.html
限制性内切酶保护碱基表：
https://wenku.baidu.com/view/1ef43e4acfc789eb162dc850.html

二、实验步骤：

2.1、获取启动子区域序列

（1） 任选一个人类已知基因；

NCBI-GENE 数据库中选择“TOMM40”:TOMM40 translocase of outer mitochondrial membrane 40 [*Homo sapiens* (human)]

（2） 利用 UCSC genome browser 查看该基因上游 5kb 范围内有无其他基因，确定该基因的上游 promoter 的大致区间（不超过 5kb）；

（3） 在 Genbank 的 Gene 数据库中搜索该基因，查看该基因在基因组中的定位和基因结构；

（4） 点击该图示右上方的“Genbank 链接”，打开该基因的序列信息，查看该基因的 Feature 区域信息；

（5） 在该页面右侧的基因组序列位置信息框中，重新输入数据，获取该基因的启动子序列（包含 exon1 和 intron1 区域），注意基因编码方向；

2.2、启动子的分析和预测

（1） 使用 5 种包含狭义启动子分析和预测功能的工具，对这段启动子序列进行计算分析（cister 生成的图在 2.3 中展示）；

（2） 使用 3 种包含转录因子分析和预测功能的工具，对这段启动子序列进行计算分析，如果结果太多，自行修改阈值，保持结果在 10 个左右即可。

2.3、启动子区域结构模型图的绘制： 使用 PicPick 软件根据上计算结果绘制启动子区 域的结构模型图，标注每个结果的定位和分值。

2.4、启动子 PCR 引物设计：

（1） 根据第 2.3 步绘制的启动子结构模型图，给出设计启动子 PCR 扩增引物区间界定的建议；

（2） 使用 PrimerBlast 设计扩增不同 Promoter 区间的特异性引物，得到引 物设计结果，并把引物设计结果（位置）绘制到启动子区域的结构模型 图上；

（3） 查询 pGL4.17 的载体数据，获得酶切信息；

（4） 使用 NEBcutter 分析该启动子序列，保存没有酶切位点“0 cutters” 的核酸内切酶数据；

（5） 选择合适的核酸内切酶；

（6）根据 pGL4.17 的载体酶切数据，把其识别序列 链接到相应的引物末端，根据 pGL4.17 序列信息，添加保护碱基至相应的引物末端。

三、实验结果：

3.1、获取启动子区域：

3.1.1、基因详细信息：

TOMM40 translocase of outer mitochondrial membrane 40 [Homo sapiens (human)]

Transcript (Including UTRs)

Position: hg38 chr19:44,891,220-44,903,689 **Size:** 12,470 **Total Exon Count:** 10 **Strand:** +

Coding Region

Position: hg38 chr19:44,891,416-44,903,169 **Size:** 11,754 **Coding Exon Count:** 9

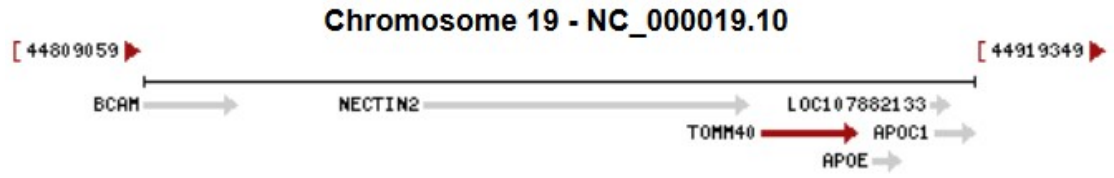
##是正链编码。

Location: 19q13.32

Exon count: 10

表格 1.TOMM40 genomic context

Annotation release	Status	Assembly	Chr	Location
109	current	GRCh38.p12 (GCF_000001405.38)	19	NC_000019.10 (44891220..44903689)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	19	NC_000019.9 (45394477..45406946)



图表 1.TOMM40 基因组定位

3.1.2、Feature 区域信息部分截取：

source 1..12470
/organism="Homo sapiens"
/mol_type="genomic DNA"
/db_xref="taxon:9606"
/chromosome="19"

gene 1..12470
/gene="TOMM40"
/gene_synonym="C19orf1; D19S1177E; PER-EC1; PEREC1; TOM40"
/note="translocase of outer mitochondrial membrane 40; Derived by automated computational analysis using gene prediction method: BestRefSeq,Gnomon."
/db_xref="GeneID:10452"
/db_xref="HGNC:HGNC:18001"
/db_xref="MIM:608061"

mRNA
join(1..470,1174..1241,1618..1710,2561..2662,2742..2847,
9511..9633,9809..9885,9989..10091,11811..12470)
/gene="TOMM40"
/gene_synonym="C19orf1; D19S1177E; PER-EC1; PEREC1;
TOM40"
/product="translocase of outer mitochondrial membrane
40,
transcript variant 1"
/note="Derived by automated computational analysis
using
gene prediction method: BestRefSeq."
/transcript_id="NM_001128917.1"
/db_xref="GeneID:10452"
/db_xref="HGNC:HGNC:18001"
/db_xref="MIM:608061"

3.1.3、获取该基因的启动子序列：

原始序列 TOMM40 在 19 染色体上位点起始位置:from: 44891220, to: 44903689，因为是正链编码的，所以向前 2000 个位置开始查询 form:44889220, to:44903689，查看结果：

gene <1..9
/gene="NECTIN2"
mRNA <1..9

gene 2001..14470
/gene="TOMM40"
mRNA
join(2001..2470,3174..3241,3618..3710,4561..4662,

4742..4847,11511..11633,11809..11885,11989..12091,
13811..14470)
/gene="TOMM40"

这是该区域下的部分 feature 结果，可以看出，前九个位置是 TOMM40 上游区域的基因(需去除)，为了更有效的预测出 TOMM40 上游启动子区域所以保留第一个外显子(from:2001, to:2470)和第一个内含子(from:2471, to:3173)区域，所以本实验中截取的 TOMM40 的启动子区域在 19 号染色体位置从 44889230 到 22892393，共 3163 个位点。以下为截取的序列：

cttgcctttt ctgtgggatg tgaaggggat ggggtgagac atgaagggga atgaggaagg
agtctaaggg cttgagagaa tcagtgaact ggggtcaaga gagttaaaga caaaaacaaa
gggaccaaag atacaaaata agaattgaga gagtaatgag caaaaatgta tcaatcaaaa
acaaaatttt ttttttttt gaggcggagt ctcgctctgt cgcccaggct ggagtgcagt
ggcatgatct cggctcactg caagctccac cttctgagtt aacgccattc ctctgcctca

gcctcccgag tagctgggac tacaggcgcc caccaccacg cccggcttat tttttgtatt
tttagtagag acggggtttc accgcgttag ccaagatggt ctcgatctcc tgacctcgtg
atccacccgt ctcggcctcc caaagtgctg ggattacagg cgtgagccac tgcgcccagc
cctaaaaaca gaatgttgag taatggaatg gggaggggtc cagggcatgc tgtttaatga
acattcataa ttgcaacagc agtatgatga atgggacaaa gataaaaata gcaattggcc
agatgcagtg gctcattcct gtaattccag cacttcggga aactgaggcg ggaagctcga
ctgagtccag gagttaagag accagcctgg gaaacacagt gagaccccgc ctctacgaaa
gttagccggg cgtggtggca cgcacctgta gttcagctac tctggaggct gaggtaggag
gatcgattga ggccagaagt tcaaggctgc agtaagctat gatggcgcca ctgcactcca
gcctgagtga cagagtccaa ccttgtcttt aaaaaaataa aaattaaaaa agcagtaacg
aaagtataag aggctcaagt ttaatgaata aacggcaagt aaaagtagac taaaggcaaa
atcatgaata atgttgtaat aggagcccaa ggcatgacgg ataataaacc tgaatgaatg
ggccagaggc agagtgatga ataatggaaa aatagggacc gggggcaagg gaggtaacgg
acaggagttc atggatccaa aacatgatgg ctaattagag aagggccgga agaacgacga
aagtgagcag gcaggctgga gctaaaagta gtctgggcaa tgaagctcaa atgaatgggg
cagaggcatg atgggtaatg ggagaggaat gaatgggcca aagatagaag ccgcagtgac
gcgaggaaca agaggcatga tgggtaacga aaggggtggg tctaaggcaa ctgtttcgct
aagaggtgat gagggcattg tgggtaacga ggagcagcgc agggttcgca gaacagatta
gaattctccc gaggcactct gggaagggcc agcacttccg gttttaggtc ggctactccg
aaccagaggt ggggtggggg cccggctgcc gcggtgcctg gtgggacgcg aggcctgacc
ttgctgccta gccgcctctg ccgcgcaacc caccttt acc tgtccttcga ccctggaacg
ttagccaatg agagtaccaa gctgatacgc caccaaggtc gaccccgtac acgctggagc
caatcaaaat gctgcaaggg tcaaagccga ccaattatca cagcaacctc gccgcggggc
ggaatcaaaa gagggcctcc tccaggagag aggcggggcg atgcctcagc gggcgtggca
aaatgctcag cacagaccaa tggcgggtta gcaccggaac ccgcggcgac gcgagccaat
aggcgcaggc gctgcgagcc aatgggaagg gtgggagggg cgccgtggct accctgcgag
tgagaaccaa tacaaaagga catttcaggg aaagtgggcg ggactttatg cacaagtcca
atgggaagac cgagtcttga cgctggtggg cgggcctcag ggcacactaa accaatgggc
taggtggggc ggggcgacgg tggtggcggc ggcggcagcg ggttcggttg cgcgtggcgc
acggggtggg agcggagccc aggccgggag caggcgccgc cgccagtgag aaccggggcc
ggagccgggt gcggatttgc tggggctgag tcgggggcgc gcgggccctg acctctgccc
tctgacctct cccctagcag gcgaccatgg ggaacgtgtt ggctgccagc tcgccgcccg
cagggccgcc accgccgcct gcgccggccc tcgtggggct gccgccacct ccgccctcgc
cgccgggctt cacgctgccg ccgctgggag gcagcctggg cgccggcacc agtacgagtc
gaagttcgga acggaccccc ggggctgcaa ccgccagcgc ctcaggggcc gccgaggatg
gggcctgcgg ctgcctgccc aacccgggca cattcgagga gtgccaccgg aagtgcaagg
gtgaggggcg aggggccccc gctgggctgc gatggcctgg atctcggggg aaggggggagg
acactgggga ctctgggatt tggcgcgcac cattggaatt atttaacagc actaggaggt
gatgttggga tcgaatggtg gaacgttgga cttggggctt agaatgatgg aatcaaatgc
tggaaacggg atggaatgtc atagcagtag agaaaagcct ttagggacct gaggagcccc
gggatcagcc aagccagact tctcttgtga tcgggaaggc aactgaggcc caaggtcacg
gtgtcagcaa ggtgtcagcg aggttccttg ggtatgggac ccaaagcctc cggatcccag
cctggagcaa ttagagtagt agtagtggtg gagatttatg gagttctgtt ctggtgttca
ttatacgtta actcattaga tccttgggac aattctgtgt ggtgagggtc ccatctttca

```
gatgatgagt ttgacctaaa gttgctcagc ttggtggcag tgagatttga gcaagcaaag
gccctggccc tgtctaacta ggctgtactg cctctttaca ggtggaatcc tttgtgagat
gttctgctgt gggtctctgg agagagctgg gggtggtagg gaaggaagag atgagagttg
        gtgtggggtt ggagtggagt gtgacagcgt ttctcttctc caga
```

3.2、启动子的分析和预测：

3.2.1、Promoter、TSSG、TSSW 和 NNPP 四种启动子分析结果：

（1）Promoter2.0

```
Sequence, 3164 nucleotides
  Position  Score  Likelihood
       300  0.574  Marginal prediction
      1200  0.686  Marginal prediction
      2000  0.708  Marginal prediction
      2700  0.534  Marginal prediction
```

（2）TSSG

```
> test sequence
 Length of sequence-      3164
 Threshold for LDF-  4.00
     2 promoter(s)  were predicted
 Pos.:   1995 LDF- 14.92
 Pos.:   1694 LDF-  8.26
```

（3）TSSW

```
> test sequence
 Length of sequence-      3164
 Thresholds for TATA+ promoters -  0.45, for TATA-/enhancers
-  3.70
     4 promoter/enhancer(s) are predicted
Enhancer Pos:   2051 LDF- 16.48  ##增强子不需要
Enhancer Pos:   1693 LDF-  5.37  ##增强子不需要
Promoter Pos:   2457 LDF-  4.69
Promoter Pos:   1905 LDF-  1.14  TATA box at   1870    17.87
```

（4）FPROM

```
Sequence    1 of    1,  Name:  test sequence
Length of sequence:      3164
     2 promoter/enhancer(s) are predicted
Promoter Pos:      1900 LDF:   +3.472 TATA box at       1871
+4.285 TACAAAAG Enchancer at:      1999 Score:  +12.013
Promoter Pos:       615 LDF:   -0.231 TATA box at        581
+4.156 GATAAAAA
```

3.2.2、TSSG、TSSW 和 CISTER 预测转录因子的分析结果：

（1）TSSG 预测的两个启动子位置附近的转录因子：

Pos.: 1995 LDF- 14.92

Pos.: 1694 LDF- 8.26

其对应的转录因子结合位点：

Transcription factor binding sites:

for promoter at position - 1995

表格 2.TSSW 预测启动子 1995 位点附近的转录因子

1865	(+)	S00098	AACCAAT
1970	(+)	S00098	AACCAAT
1875	(+)	S01153	AARKGA
1709	(+)	S00922	AGAGG
1794	(+)	S00696	AGCCAAT
1817	(+)	S00696	AGCCAAT
1881	(+)	S00089	CANYYY
1881	(+)	S01616	CATTW
1757	(+)	S00633	CCAAT
1796	(+)	S00633	CCAAT

for promoter at position - 1694

表格 3.TSSG 预测启动子 1694 位点附近的转录因子

1567	(+)	S01090	AATGA
1445	(+)	S00922	AGAGG
1690	(+)	S00922	AGAGG
1563	(+)	S00696	AGCCAAT
1618	(+)	S00696	AGCCAAT
1610	(+)	S00395	CACGCW
1395	(+)	S00089	CANYYY
1413	(+)	S00089	CANYYY
1526	(+)	S00089	CANYYY
1531	(+)	S00089	CANYYY
1664	(+)	S00089	CANYYY

（2）TSSW 预测的两个启动子位置附近的转录因子：

Promoter Pos: 2457 LDF- 4.69

Promoter Pos: 1905 LDF- 1.14 TATA box at 1870 17.87

for promoter at position - 2457

表格 4.TSSW 预测启动子 2457 位点附近的转录因子

2318	(+)	HS$A4_01	GGGCGCgGG
2214	(+)	CHICK$ACRA	CCGCCC
2271	(+)	CHICK$ACRA	CCGCCC
2172	(+)	CHICK$AAC_	ccaaatatGGCGACggccgggg
2252	(+)	MAIZE$ADH1	CGTGG
2357	(+)	MAIZE$ADH1	CCCCGG
2168	(+)	Y$ADH2_01	TCTCC
2403	(+)	RAT$ANTEN_	ccacagttgggatttCCCAACctgaccag
2266	(+)	RAT$A12COL	CACCTCC
2184	(+)	Y$CYC1_09	ctcatttggcgagcGTTGGt

for promoter at position - 1905

表格 5.TSSW 预测启动子 1905 位点附近的转录因子

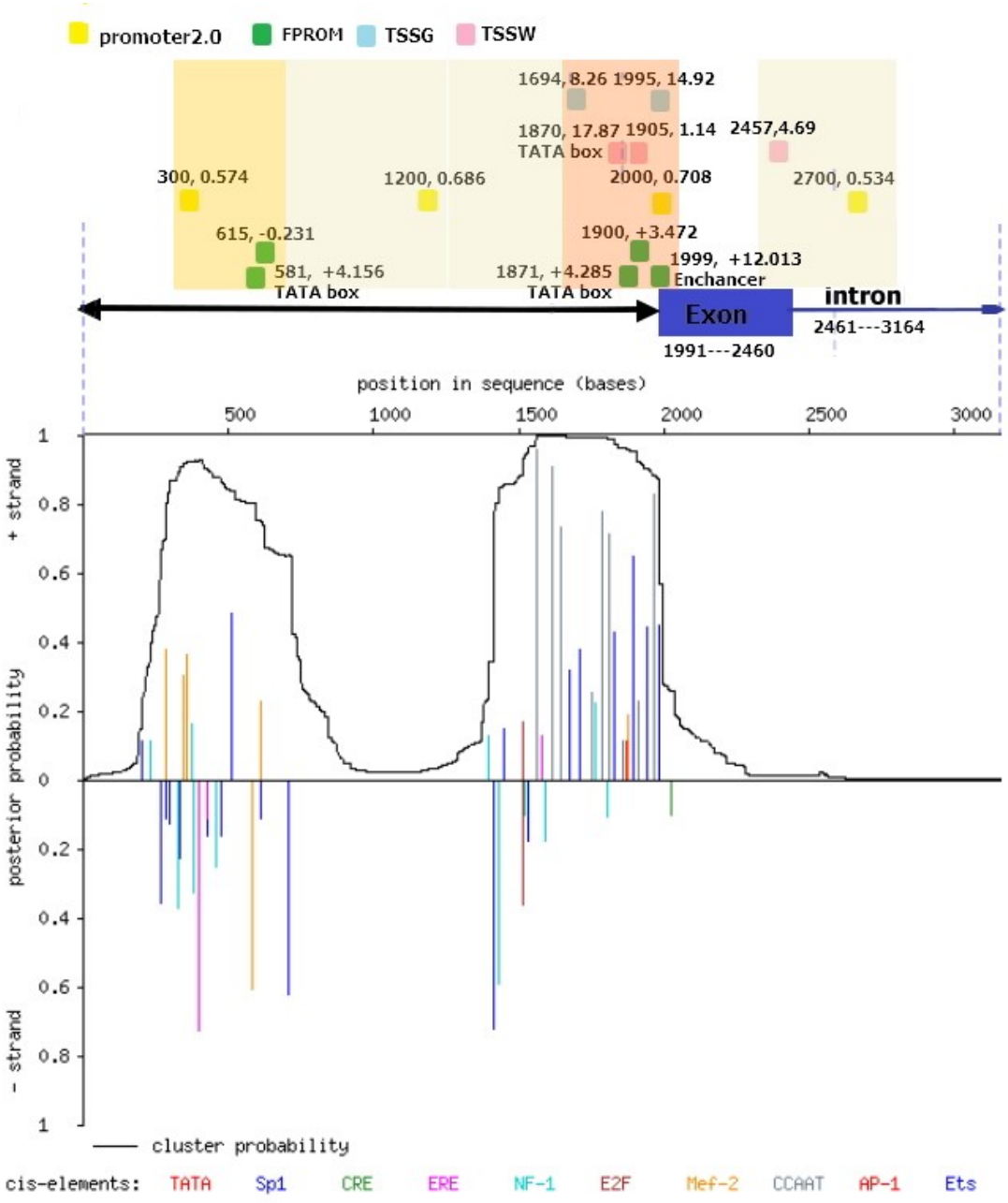
1777	(+)	CHICK$AAC_	ccaaatatGGCGACggccgggg
1620	(+)	HS$BAC_03	CCAAT
1651	(+)	HS$BAC_03	CCAAT
1757	(+)	HS$BAC_03	CCAAT
1796	(+)	HS$BAC_03	CCAAT
1819	(+)	HS$BAC_03	CCAAT
1867	(+)	HS$BAC_03	CCAAT
1734	(+)	MAIZE$ADH1	CGTGG
1844	(+)	MAIZE$ADH1	CGTGG

（3）CISTER 预测的转录因子：

表格 6.cister 预测的转录因子

type	position	strand	sequence	probability
CCAAT	1560 to 1575	+	gttagccaatgagagt	0.96
CCAAT	1615 to 1630	+	tggagccaatcaaaat	0.9
CCAAT	1967 to 1982	+	ctaaaccaatgggcta	0.82
CCAAT	1791 to 1806	+	gcgagccaataggcgc	0.77
CCAAT	1646 to 1661	+	gccgaccaattatcac	0.72
CCAAT	1814 to 1829	+	gcgagccaatgggaag	0.7
Sp1	1893 to 1905	+	agtgggcgggact	0.63
Sp1	508 to 520	+	atggggaggggtc	0.49
Sp1	1984 to 1996	+	gtggggcggggcg	0.43
Sp1	1945 to 1957	+	ggtgggcgggcct	0.42
Sp1	1830 to 1842	+	ggtgggaggggcg	0.41
Tef	284 to 295	+	gccattcctctg	0.39

3.3、启动子区域结构模型图的绘制：

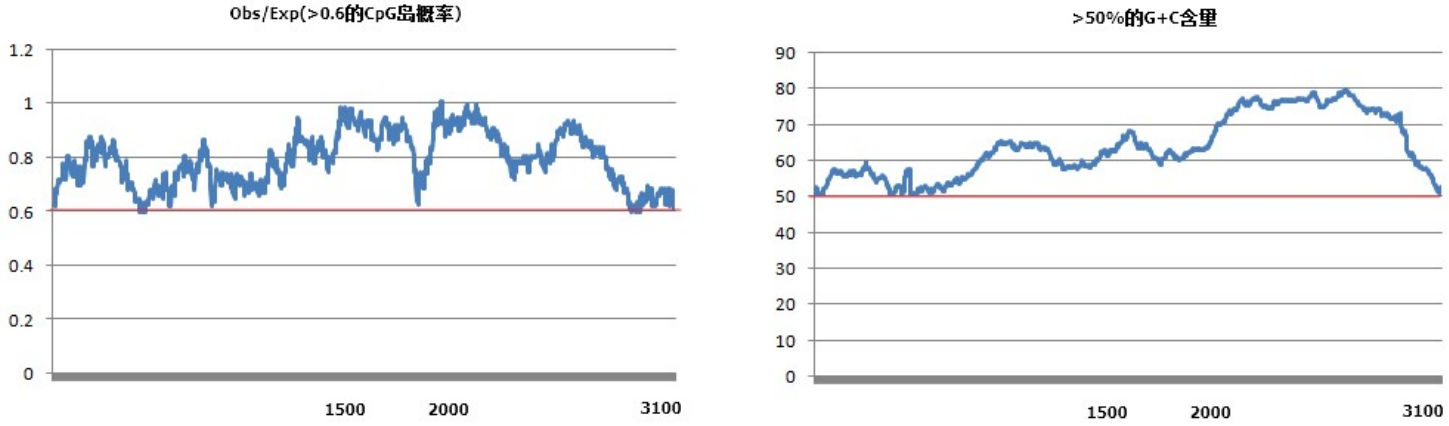


图表 2.cister、promoter2.0、FPROM、TSSG、TSSW 结果汇总

3.4、启动子 PCR 引物设计：

（1） 根据第 3.3 步绘制的启动子结构模型图图表 2，启动子 PCR 扩增引物建议设计在外显子上游 300-350bp 的位置到外显子的开始位置间的区段内，即四种软件都有预测到的区域，本实验的 1650-2000bp 区间；

进一步验证，通过查看序列是否存在 CpG 岛，而对启动子预测的准确性做出辅助性的推测=》将全序列导入 CpG Prediction 软件，每 200bp 进行 CpG 岛的预测打分（Obs/Exp>0.6 并且 %GC ＞ 50 时更好），CpG 岛是预测启动子并提高预测准确性的重要序列；

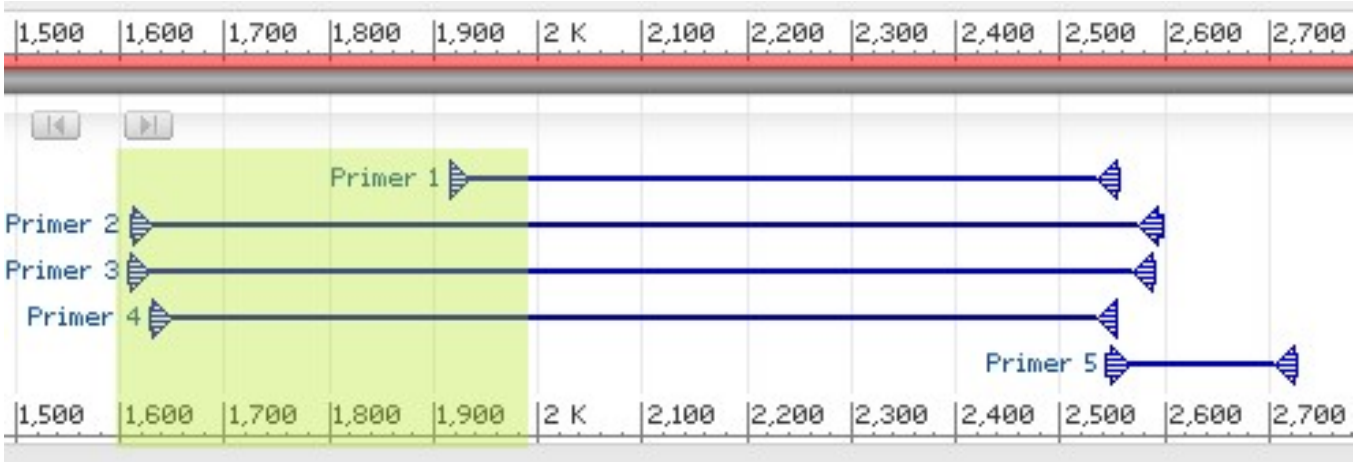


图表 3.CpG 岛检验

并且很多真核生物的核心启动子为 TATAbox，看 cister 的预测结果，TATAbox 区域也在 1600-2000bp 的区域内。

（2） PrimerBlast 设计扩增 Promoter 结果：

1).将全序列导入，没有设置参数时的结果，可以看出 Forward primer 区域集中在 1600-2000，与图表 2 中各软件预测的区域相似；

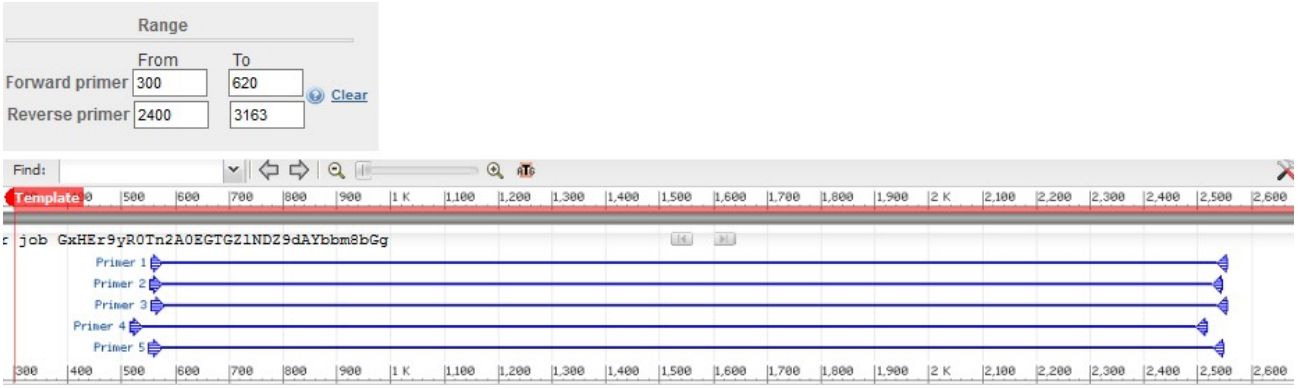


图表 4.Primer Blast 结果

2). 设置正向引物和反向引物的区域：

图表 5.正向引物和反向引物区间设置（1）

图表 6.正向引物和反向引物区间设置（2）

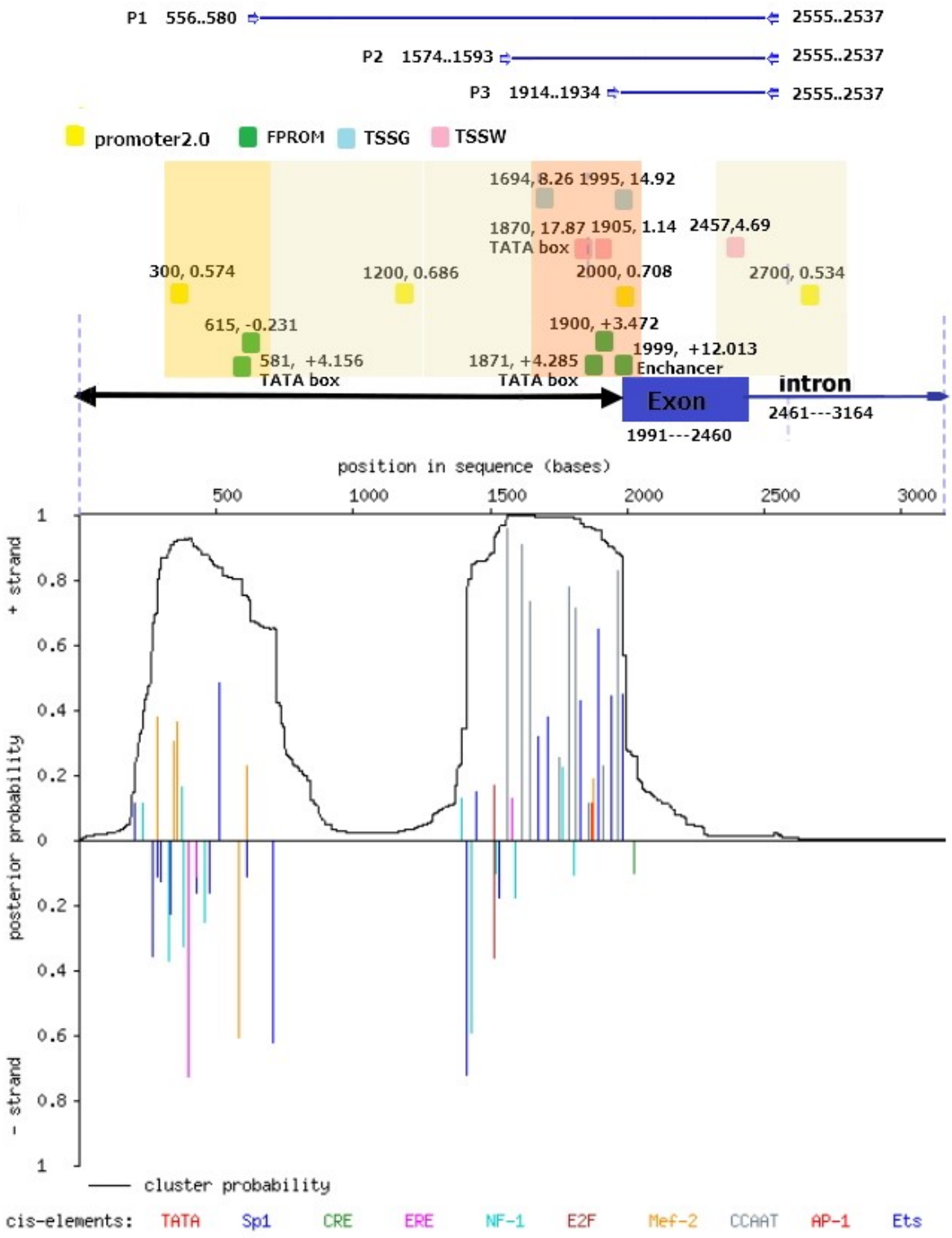


图表 7.正向引物和反向引物区间设置（3）

3). 根据不同区段的结果选取三个反向引物（下游引物）位置相同的引物：

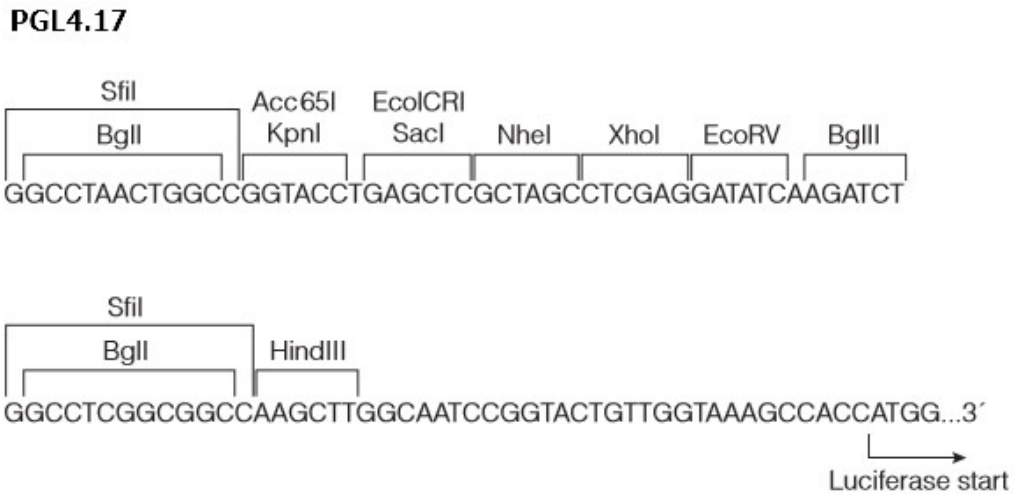
表格 7.三种不同区段的 primer 信息

P1									
	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	ACAGCAGTATGATGAATGGGACAAA	Plus	25	556	580	60.81	40	4	0
Reverse primer	CAATGGTGCGCGCCAAATC	Minus	19	2555	2537	61.16	57.89	8	2
Product length:2000									
P2									
	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	GTACCAAGCTGATACGCCAC	Plus	20	1574	1593	58.71	55	6	0
Reverse primer	CAATGGTGCGCGCCAAATC	Minus	19	2555	2537	61.16	57.89	8	2
Product length:982									
P3									
	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	AAGTCCAATGGGAAGACCGAG	Plus	21	1914	1934	59.72	52.38	6	0
Reverse primer	CAATGGTGCGCGCCAAATC	Minus	19	2555	2537	61.16	57.89	8	2
Product length:642									



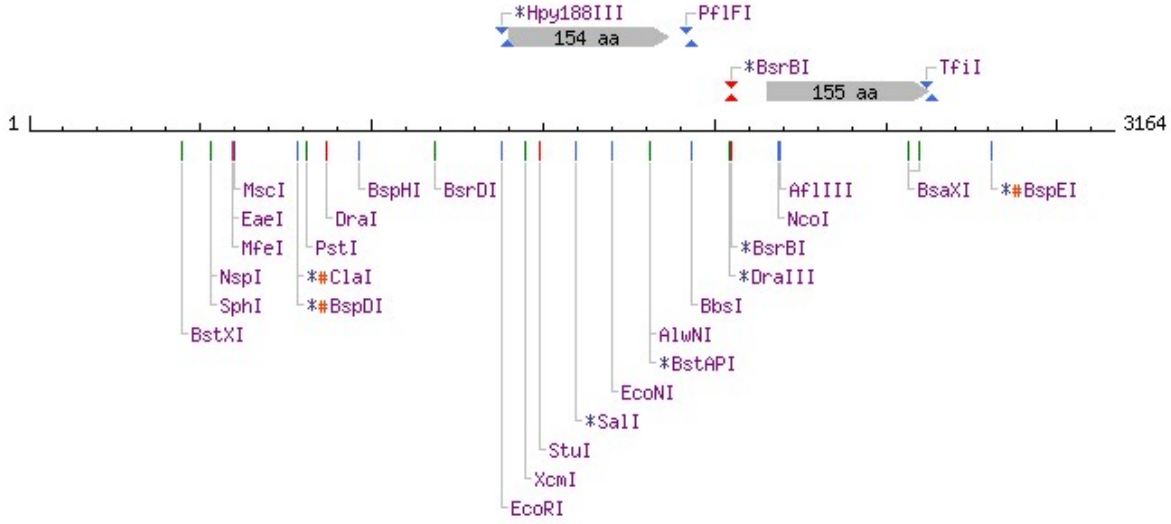
图表 8.添加特异性引物到启动子区域的结构模型图

（3） 查询 pGL4.17 的载体数据，酶切信息



图表 9.pGL4.17 酶切数据

（4） 使用 NEBcutter 分析该启动子序列



图表 10.NEBcutter 分析结果

=>没有酶切位点的核酸内切酶数据：

Enzymes that don't cut

AatII (G_ACGT^C)

Acc65I (G^GTAC_C)

AcuI (CTGAAGNNNNNNNNNNNNNN_NN^)

AfeI (AGC|GCT)

AflII (C^TTAA_G)

AgeI (A^CCGG_T)

AhdI (GACNN_N^NNGTC)

AleI (CACNN|NNGTG)

ApaLI (G^TGCA_C)

AscI (GG^CGCG_CC)

AseI (AT^TA_AT)

AsiSI (GCG_ATˆCGC)
AvrII (CˆCTAG_G)
BaeI (_NNNNNˆNNNNNNNNNNACNNNNGTAYCNNNNNNN_NNNNNˆ)
BciVI (GTATCCNNNNN_Nˆ)
BclI (TˆGATC_A)
BfuAI (ACCTGCNNNNˆNNNN_)
BglII (AˆGATC_T)
BmgBI (CAC|GTC)
BmtI (G_CTAGˆC)
BsaAI (YAC|GTR)
BsaBI (GATNN|NNATC)
BsiEI (CG_RYˆCG)
BsiHKAI (G_WGCWˆC)
BsiWI (CˆGTAC_G)
BsmI (GAATG_CNˆ)
BspMI (ACCTGCNNNNˆNNNN_)
BspQI (GCTCTTCNˆNNN_)
BsrGI (TˆGTAC_A)
BstBI (TTˆCG_AA)
BstEII (GˆGTNAC_C)
BstZ17I (GTA|TAC)
CspCI (_NNˆNNNNNNNNNNCAANNNNNGTGGNNNNNNNNNN_NNˆ)
DrdI (GACNN_NNˆNNGTC)
EagI (CˆGGCC_G)
Eco53kI (GAG|CTC)
EcoRV (GAT|ATC)
FseI (GG_CCGGˆCC)
FspI (TGC|GCA)
HindIII (AˆAGCT_T)
KpnI (G_GTACˆC)
MluI (AˆCGCG_T)
MslI (CAYNN|NNRTG)
NdeI (CAˆTA_TG)
NheI (GˆCTAG_C)
NotI (GCˆGGCC_GC)
NruI (TCG|CGA)
NsiI (A_TGCAˆT)
PacI (TTA_ATˆTAA)
PaeR7I (CˆTCGA_G)
PciI (AˆCATG_T)
PflMI (CCAN_NNNˆNTGG)
PmeI (GTTT|AAAC)
PmlI (CAC|GTG)
PshAI (GACNN|NNGTC)

PsiI (TTA|TAA)
PspXI (VC^TCGA_GB)
PvuI (CG_AT^CG)
PvuII (CAG|CTG)
RsrII (CG^GWC_CG)
SacI (G_AGCT^C)
SapI (GCTCTTCN^NNN_)
SbfI (CC_TGCA^GG)
ScaI (AGT|ACT)
SexAI (A^CCWGG_T)
SfiI (GGCCN_NNN^NGGCC)
SgrAI (CR^CCGG_YG)
SnaBI (TAC|GTA)
SpeI (A^CTAG_T)
SrfI (GCCC|GGGC)
SspI (AAT|ATT)
SwaI (ATTT|AAAT)
XbaI (T^CTAG_A)
XhoI (C^TCGA_G)
XmnI (GAANN|NNTTC)
ZraI (GAC|GTC)

（5） 选择合适的核酸内切酶

上面信息中黄标的为与 Pg14.17 载体酶数据序列和名称都对应的，绿色标的为与 Pg14.17 载体酶序列相对应但是名称不相同(本实验中不进行讨论），最终整合黄色标的数据，查询载体酶的信息并进行选择。

表格 8.选取核酸内切酶依据

Enzyme	Protection base	NEBuffer	Temp./℃	Heat Inaction
XhoI (C^TCGA_G)	1	4+BSA	37	65
SfiI (GGCCN_NNN^NGGCC)	1	4+BSA	50	NO
KpnI (G_GTAC^C)	1or2	1+BSA	37	NO
SacI (G_AGCT^C)	1	1+BSA	37	65
NheI (G^CTAG_C)	1or2	2+BSA	37	65
Acc65I (G^GTAC_C)	2	3+BSA	37	65
EcoRV (GAT|ATC)	1	3+BSA	37	80
HindIII (A^AGCT_T)	2or3	2	37	65
BglII (A^GATC_T)	2	3	37	NO

最终选择 KpnI（5′... GGTACC...3′） 和 SacI（5′... GAGCTC...3′）内切酶分别作为正向引物和反向引物的核酸内切酶。

（6） 根据选定的合适的内切酶，添加识别序列链和保护碱基至相应的引物末端

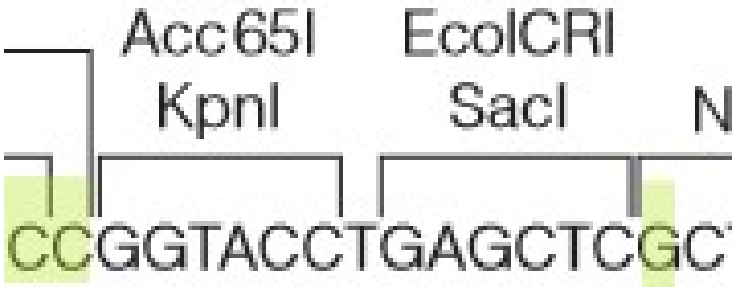
表格 9.原始预测的引物与添加内切酶识别序列以及保护碱基的引物

	primer	original sequence	add enzyme, add protection base(s)
p1	Forward primer	ACAGCAGTATGATGAATGGGACAAA	CCGGTACCACAGCAGTATGATGAATGGGACAAA
	Reverse primer	CAATGGTGCGCGCCAAATC	CGAGCTCCAATGGTGCGCGCCAAATC
p2	Forward primer	GTACCAAGCTGATACGCCAC	CCGGTACCGTACCAAGCTGATACGCCAC
	Reverse primer	CAATGGTGCGCGCCAAATC	CGAGCTCCAATGGTGCGCGCCAAATC
p3	Forward primer	AAGTCCAATGGGAAGACCGAG	CCGGTACCAAGTCCAATGGGAAGACCGAG
	Reverse primer	CAATGGTGCGCGCCAAATC	CGAGCTCCAATGGTGCGCGCCAAATC

【说明：】表格 9 中红色字显示的为特异性引物，绿色的为保护碱基。

【注：】设计正向引物时，该引物序列应该具有待扩增 DNA 序列接近 5’段处的序列。反向引物设计，该引物序列应该具有的是和待扩增 DNA 序列接近 3’端处的序列的反向互补序列。

==》 根据表格 8 可以知道这两个核酸内切酶可以都只要一对保护碱基， KpnI 为正向引物添加识别序列，所以保护碱基选择前面一对即可；但 SacI 为反向引物添加识别序列，所以需要选择下游的一个碱基（没有成对碱基），并且将该碱基的互补碱基添加到反向引物上。



图表 11.保护碱基添加说明

四、讨论与结论：

4.1、启动子是参与特定基因转录及其调控的 DNA 序列。包含核心启动子区域和调控区域。核心启动子区域产生基础水平的转录，调控区域能够对不同的环境条件作出应答，对基因的表达水平做出相应的调节。 区域：启动子的范围非常大，可以包含转录起始位点上游 2000bp，有些特定基因的转录区内部也存在着转录因子的结合位点，因此也属于启动子范围。

所以 TSSW 预测的其中一个启动子位点 2457 和 promoter 预测的一个启动子位点 2700，不一定是完全预测错的，但是经过多个软件预测，这两个位点附近都没有其他软件预测到，所以本实验中不推荐研究这两个位于转录区域的位点。

4.2、启动子预测软件大体分为三类，第一类是启发式的方法，它利用模型描述几种转录因子结合部位定向及其侧翼结构特点，它具有挺高的特异性，但未提供通用的启动子预测方法；第二类是根据启动子与转录因子结合的特性，从转录因子结合部位的密度推测出启动子区域，这方法存在较高的假阳性；另一类是根据启动子区自身的特征来进行测定，这种方法的准确性比较高。同时，还可以结合是否存在 CpG 岛，而对启动子预测的准确性做出辅助性的推测。

PROMOTER 2.0，用神经网络方法确定 TATA 盒、CCAAT 盒、加帽位点(cap site)和 GC 盒(GCbox) 的位置和距离，识别含 TATA 盒的启动子。

注：选择软件进行 promoter 区域的预测时，需注意软件是根据什么物种进行预测以及其原理：1).bprom：预测是原核生物的-10box 和-35box 启动子区域

http://www.softberry.com/cgi-bin/programs/gfindb/bprom.pl

2).nnpp：是根据伯克利果蝇基因组计划进行 promoter 预测

http://www.fruitfly.org/seq_tools/promoter.html

Promoter predictions for seq0 :

Start	End	Score
895	945	0.83

Promoter Sequence:aagcagtaacgaaagtataagaggctcaagtttaatgaataaacggcaag

1037	1087	0.88

Promoter Sequence:gagtgatgaataatggaaaaatagggaccgggggcaagggaggtaacgga

1675	1725	0.97

Promoter Sequence:cgccgcggggcggaatcaaaagagggcctcctccaggagagaggcggggc

1730	1780	0.82

ctcagcgggcgtggcaaaatgctcagcacagaccaatggcgggttagcac

3). PROMOTER 2.0，用神经网络方法确定 TATA 盒、CCAAT 盒、加帽位点(cap site) 和 GC 盒(GCbox) 的位置和距离，识别含 TATA 盒的启动子。

4.3、可以用 PCR Primer Stats 评估潜在的 PCR 引物，PCR Primer Stats: https://sites.ualberta.ca/~stothard/javascript/pcr_primer_stats.html

4.4、在真核生物中，一个基因附近存在多个启动子是正常的。不同软件依据不同原理预测的启动子准确度不一定，而且真核生物启动子的种类比原核生物的要多，其保守序列和结合转录因子的结合也不同的，所以在多种软件都预测到的启动子区域进行实验验证比较好。

4.4、在添加引物和保护碱基是一定要注意正向引物和反向引物的保护碱基添加方式，具体已在 3.4 的第（6）步中详述。

附录：

表格 10.简并碱基

简并碱基	正常碱基
R	A/G
Y	C/T
M	A/C
K	G/T
S	G/C
W	A/T
H	A/T/C
B	G/T/C
V	G/A/C
D	G/A/T
N	A/T/C/G