微生物学报
Acta Microbiologica Sinica
2019, 59(11): 2218–2228
http://journals.im.ac.cn/actamicrocn
DOI: 10.13343/j.cnki.wsxb.20190030

Research Article 研究报告

瘤胃微生物元基因组来源的新的组成型启动子获取

王丽[1*]，赵云[2]，杨茜[1]，戴欣[1]，朱雅新[1]，董志扬[1*]

[1] 中国科学院微生物研究所，微生物资源前期开发国家重点实验室，北京 100101
[2] 中国科学院生物物理研究所，蛋白质与多肽药物所重点实验室，北京 100101

摘要：【目的】自极端环境来源的微生物的基因组中筛选新型的可用于合成生物学底盘细胞设计的启动子元件。【方法】本研究以含有绿色荧光蛋白结构基因和核糖体结合位点的探针型质粒 pUC18-GFP 为载体，通过构建瘤胃微生物元基因组质粒文库，从文库中快速高效筛选具有启动子功能的 DNA 片段。并且通过基于神经网络的启动子预测分析，获得可能的启动子区域。以绿色荧光蛋白和施氏假单胞菌 *Pseudomonas stutzeri* 来源的麦芽四糖淀粉酶作为报告基因验证所获得的新启动子片段的功能。【结果】我们从约 3750 个转化子中筛选到 22 条具有组成型启动子功能的 DNA 片段。这些片段与 NCBI 数据库中已报道的基因序列同源性较低，启动效率高低不等。我们通过启动子预测和亚克隆的方法获得两条全新的启动子片段 *RFa1p2* (76 bp)和 *RFb4p* (547 bp)。此新的组成型启动子可以在不添加任何诱导剂的情况下启动异源蛋白在大肠杆菌基因工程菌中高效表达。

关键词：元基因组文库，组成型启动子，瘤胃微生物，绿色荧光蛋白，麦芽四糖淀粉酶

启动子[1]是一段能够被RNA聚合酶识别并起始转录的 DNA 序列，在转录水平上的基因表达调控主要取决于启动子，启动子本身的序列结构特征以及与蛋白调控因子的相互作用决定了基因的转录效率以及时空特异性表达。优化启动子是构建表达载体、调控微生物发酵的重要途径[2–3]。目前启动子的优化主要有两种方法，一是设计寡核苷酸探针随机引物PCR扩增[2,4]或者突变PCR[5]构建人工合成的启动子文库(SPL，synthetic promoter library)[6]，一是从各种不同来源微生物基因组中随机筛选启动子。两者均能筛选到许多不同强度范围的启动子，但后者更有利于获得一些新的未知的遗传信息。虽然原核细胞与真核细胞以及古菌基因转录水平的调控有很大差别，RNA 聚合酶、启动子功能保守区域均有所不同，但是许多不同来源的基因片段均可被 *E. coli* RNA 聚合酶识别发挥启动子功能，如嗜盐古菌

基金项目：国家自然科学基金(31300007，30770053，31240050)
[*]**通信作者**。董志扬，Tel/Fax：+86-10-64807337，E-mail：dongzy@im.ac.cn；王丽，Tel：+86-10-64807331，E-mail：wangli07@im.ac.cn
收稿日期：2019-01-22；**修回日期：**2019-04-18；**网络出版日期：**2019-04-29

Halobacterium halobium[7]、λ噬菌体、枯草芽孢杆菌、*Lactococcus lactis*，甚至四膜虫、果蝇等真核生物[8–11]。而且在 *E. coli* 中发挥启动子功能的外源片段转入 *Lactococcus lactis*、*Saccharomyces cerevisiae* 等大部分仍能发挥启动子作用。这就决定了我们可以利用大肠杆菌表达系统直接从环境微生物总 DNA 中筛选启动子功能片段。

瘤胃中的微生物基因资源非常丰富，数量高达每毫升瘤胃液 10^{14} 个微生物，分属 3000 种以上的基因组类型，据统计每毫升瘤胃液中大约含古菌和细菌 10^9–10^{11} 个，真菌 10^5 个，原生动物 10^5–10^6 个，噬菌体 10^7–10^9 个，有近 1000 种细菌、真菌和原生动物[12–13]。因而通过构建元基因组文库的方法来保存和筛选瘤胃微生物的调控序列和功能基因十分必要。而且据报道环境中来源的酶基因有潜在酶活力的超过 40%可以在 *E. coli* 中克隆表达[14]。而调控序列本身能在 *E. coli* 中发挥作用的所占比例可能会更高。

本研究提取牛瘤胃胃液混合微生物总 DNA，以启动子探针型质粒 pUC18GFP 为载体，*E. coli* Top10、DH5α 为宿主菌，构建瘤胃胃液元基因组质粒文库，从中筛选启动子片段，并进一步对筛选到的片段进行序列分析和启动子功能区域确定。筛选到 22 个新的组成型启动子片段，对其中的中等强度启动子 *RFa1* (1202 bp)功能区域缩小到了 76 bp (*RFa1p*)。对强启动子 *RFb4* (~2.8 kb)功能区域缩小到 547 bp (*RFb4p*)。

1 材料和方法

1.1 启动子探针型载体 pUC18-GFP 构建

以含有增强型绿色荧光蛋白 mutGFP2 (S65A, V68L, S72A) (吸收峰红移至 480–510 nm，发射峰变为 507–511 nm)(由中国科学院生物物理研究所系统生物学研究中心杭海英教授馈赠)结构基因的质粒为模板，以下列序列为引物 GFP-F/GFP-R (表 1)进行 PCR 扩增，得到 mutGFP2 结构基因片段(717 bp)以及三联体终止密码子 TAATTAATTAA 和核糖体结合位点 AAGGAG，连接到 pUC18 质粒载体的多克隆位点 *Kpn* Ⅰ、*Eco*R Ⅰ之间。三联体终止密码子的设计是为了使得 mutGFP2 结构基因前的编码基因到此终止翻译[15–16]。

表 1. 本文所用的部分 PCR 扩增引物

Table 1. List of PCR primers used in this study

Primer names	Sequences (5′→3′)
GFP-F	GCATCCGGTACCTAATTAATTAAGAAGGAGATATACAATGAGTAAA
GFP-R	GCGAATTCTTATTTGTATAGTTCATCC
GFP2-F	GCATCCGGTACCATGAGTAAAGGAGAAGAAC
RFb4-F	GCGTTGGTCGGCGGCGATAGAG
pUC18G-R	GGCACCCCAGGCTTTACACTTTATG
GFP-r-sp	CGCGAAAGTAGTGACAAGTG
Psmta-F	CCGGAATTCATGAGCCAGACCCTACGTG
Psmta-R	CCCAAGCTTTCAGAACGAGCC
Psmta-F’	CGGGGTACCTAATTAATTAAGAAGGAGATATACATATGAGCCAGACCCTACGTG
Psmta-R’	CCGGAATTCTCAGAACGAGCCGCTGGTGCTC

F stands for forward primer, and R stands for reverse primer. Dsmta represent glucan 1,4-alpha-maltohexaosidase from *Pseudomonas stutzeri*; the underscore characters represent restriction enzyme cleavage sites include *Pst* Ⅰ CTGCAG, *Kpn* Ⅰ GGTACC, *Hin*d Ⅲ AAGCTT and *Eco*R Ⅰ GAATTC.

1.2 瘤胃胃液混合微生物元基因组文库构建

瘤胃胃液样品采集：利木赞牛(来源于北京金维福仁清真食品有限公司肉牛)，宰杀后取瘤胃，瘤胃胃液用灭菌的 4 层纱布过滤，4 °C 离心沉淀菌体，用 1×TE 悬浮洗涤后–70 °C 保存。瘤胃混合微生物总 DNA 提取参考杨瑞红论文[17–19]中的方法。将上述提取出来的混合微生物总 DNA 采用 Wizard SV gel and PCR clean-up system (Promega)胶回收试剂盒纯化(图 2-A)。将纯化后的瘤胃 DNA 用不同的核酸内切限制酶酶切鉴定：*Kpn* Ⅰ(10 U/μL) 0.75 μL 37 °C 孵育 2 h； *Hin*d Ⅲ (15 U/μL) 0.5 μL 37 °C孵育 2 h；*Pst* Ⅰ(15 U/μL) 0.5 μL 37 °C孵育 2 h；*Bam*H Ⅰ(15 U/μL) 0.5 μL 30 °C 孵育 2 h。*Pst* Ⅰ酶切效果最好，*Bam*H Ⅰ次之。纯化后的瘤胃胃液微生物总 DNA 用 *Pst* Ⅰ(15 U/μL) 0.5 μL 于 37 °C 孵育 2 h，琼脂糖凝胶电泳后，利用 QIAquick Gel Extraction Kit (QIAGEN)凝胶回收试剂盒回收 250 bp–8 kb DNA 片段；pUC18-GFP 载体用 *pst* Ⅰ酶切后用碱性磷酸酶 CIAP (alkaline phosphatase (Calf intestine) TaKaRa)去磷酸化。纯化后的瘤胃 DNA *Pst* Ⅰ酶切片段(约 30 ng/3 μL) 7 μL，*Pst* Ⅰ酶切、去磷酸化回收的载体 pUC18-GFP (约 10 ng/μL) 1 μL，T4 DNA Ligase (TaKaRa)连接 4 °C 过夜；化学转化或者电转化 *E. coli* Top10、DH5α 感受态细胞。

1.3 组成型启动子筛选、鉴定

1.3.1 组成型启动子筛选：固体平板上直接筛选。转化后的复活产物涂 LB 固体平板(含氨苄 100 μg/mL)，37 °C 培养 16–20 h 后，显绿色的克隆可以肉眼观察到直接从平板上挑出，固体平板上的绿色克隆经液体培养后的菌液进一步用荧光显微镜观察确证(蓝光激发)，不同克隆荧光强度不同。转化后的复活产物加氨苄(100 μg/mL)摇菌 3–4 h 后可以利用流氏细胞仪快速高效分选，以未连接外源片段的探针型质粒载体的转化子为阴性对照，荧光强度低于阴性对照的部分废弃，荧光强度高于阴性对照的细胞保留，这些细胞为组成型表达绿色荧光蛋白的阳性克隆[21]。

1.3.2 荧光强度测定：首先将菌液浓度校正一致。将 37 °C 培养约 16 h 的不同克隆的菌液稀释 10 倍，取 1 滴菌液滴在血球计数板上，显微镜下计数，取不同克隆一定量的菌液使细胞数目都在 1.44×10^9，离心沉淀菌体后用 100 μL PBS 悬浮。加在 96 孔板上，设置荧光分光光度计 Fluostar Optima (BD)激发光(excitation) 485 nm，发射光(emission) 520 nm，阈值(gain)450，以含 pUC19-GFP 质粒的 *E. coli* BL21 (DE3)菌液为阳性对照，含 pUC18-GFP 质粒的 *E. coli* DH5α 菌液为阴性对照，测量荧光强度。

1.3.3 启动子片段鉴定：将组成型表达 GFP 克隆提质粒、*Pst* Ⅰ酶切鉴定插入片段长度，用 *Sau*3A Ⅰ消化，鉴定插入片段的限制性内切酶片段长度多态性。在 GFP 结构基因内部(距起始密码子 ATG175bp 处)设计反向测序引物 GFP-r-sp，对部分组成型表达绿色荧光蛋白的克隆外源插入片段近 GFP 端测序。对荧光强度较强的克隆 RFa1、RFc1、RFd1 等进行了插入片段全序列测定。

1.3.4 部分启动子启动麦芽四糖淀粉酶基因表达鉴定：将 *Pseudomonas stutzeri* strain 537.1 来源的麦芽四糖淀粉酶(1,4-α-glucan maltotetraohydrolase EC 3.2.1.60)基因(1.6 kb)利用 Psmta-F/Psmta-R 引物将该结构基因连接在 pET28a 表达载体上多克隆位点 *Eco*R Ⅰ和 *Hin*d Ⅲ之间，利用 Psmta-F/Psmta-R 引物将该结构基因通过 PCR 扩增并连接到 pGEM-T 克隆载体上或者 pUC18 质粒载体的多克隆位点 *Kpn* Ⅰ和 *Eco*R Ⅰ之间。将启动子序列 *RFb4-truncation* (~2000 bp) (RFb4 外源插入片段

的 *kpn* Ⅰ酶切片段)、*RFb4* (~2800 bp)、里氏木霉 QM9414 来源的 DNA 片段 *Qmam1* 分别连接到 pUC18-*Psmta* 载体的 *Pst* Ⅰ和 *Kpn* Ⅰ之间，并转化大肠杆菌 DH5α 得到 b4mta-1 (*RFb4-truncation*)、b4mta-3 (*RFb4-truncation*) 、b4mta-4 (*RFb4*) 、b4mta-5 (*RFb4*)和 Qmam 五个转化子；将麦芽四糖淀粉酶结构基因与 pGEM-T (Promega)连接，结构基因编码序列位于 SP6 启动子下游得到 pGEM-T-m (DH5α)转化子；将麦芽四糖淀粉酶结构基因与 pET28a (Novagen) 连接，结构基因编码序列位于 T7 启动子及乳糖操纵子阻遏蛋白结合位点下游；得到 pET28a-m (BL21(DE3))转化子。分别将重组单克隆活菌液点在含有 1%可溶性淀粉和 100 μg/mL 氨苄的 LB 固体平板上，37°C 培养 16 h 看是否显示透明圈(图 3-A)。将 b4-1、b4-3、b4-4、b4-5、Qmam、DH5α、pGEM-T-m 单克隆接种在含有 100 μg/mL 氨苄的液体 LB 培养基中，37 °C、200 r/min 培养 16 h；pET28a-m 单克隆接种在含有 50 μg/mL 卡那霉素的液体 LB 培养基中，37 °C、200 r/min 培养 12 h 后，加入终浓度 0.1 mol/L IPTG，30 °C、200 r/min 继续培养 4 h，将上述发酵液菌体浓度统一校正到 6×10^8 cell/mL，取上述破碎后的菌液 8 μL 点在含有 1%可溶性淀粉的琼脂固体平板上，50 ℃ 处理 10 min，碘液染色(图 3-B)。将 b4-1、b4-3、b4-4、b4-5、Qmam、DH5α 单克隆接种在含有 100 μg/mL 氨苄的液体 LB 培养基中，pGEM-T-m、pET28a-m [BL21(DE3)]接种在含有 100 μg/mL 氨苄或者 50 μg/mL 卡那霉素以及终浓度 0.1 mol/L IPTG 的液体 LB 培养基中，37 °C、180 r/min 培养 16 h，将上述发酵液菌体浓度统一校正到 6×10^8 cell/mL，取上述破碎后的菌液 4 mL，测定粗酶液淀粉酶活力。测活反应体系(5mL)：底物为终浓度 1% (*W/V*)可溶性淀粉；缓冲液为终浓度 0.05 mol/L 磷酸钠；pH 6.6，温度 50 °C，时间 10 min；酶液为 4 mL 发酵菌液超声波破壁后的粗酶液。麦芽四糖淀粉酶活力单位(U)定义：在上述反应条件下，每分钟释放 1 μmol 的还原糖(以葡萄糖作标准曲线)所需酶量，定义为一个酶活力单位。

1.4 启动子功能区域确定

将测得的序列输入 NCBI Blastn 比对，并且用基于神经网络的启动子预测方法 1999 NNPP version 2.2 进行原核、真核启动子预测[20–22]。利用预测软件预测的最有可能的启动子区域，在其附近设计 PCR 引物(RFa1p-F/RFa1p-R、RFa1p-F1//RFa1p-R、RFa1p-F2//RFa1p-R、RFa1p-F3//RFa1p-R)，得到 *RFa1p* (224 bp)、*RFa1p1* (137 bp)、*RFa1p2* (76 bp)、*RFa1p3* (31 bp)序列片段。利用预测软件预测的最有可能的启动子区域，在其附近设计 PCR 引物(RFb4p-F/pUC18G-R、RFb4p-F/RFb4p-R1、RFb4p-F/RFb4p-R2)，得到 *RFb4p* (547 bp)、*RFb4p1* (395 bp)、*RFb4p2* (134 bp)的片段扩增下来并重新连到探针型载体 pUC18GFP 或者 pUC18GFP2 (引物 GFP2-F/GFP-R)的 *pst* Ⅰ、*kpn* Ⅰ位置上，鉴定重组克隆是否显绿色。并将缩小的功能区域进一步 Blastn 比对验证是否为新序列。

2 结果和分析

2.1 瘤胃胃液元基因组质粒文库组成型启动子分析

瘤胃胃液元基因组质粒文库外源插入片段长度 750–8000 bp，平均插入片段长度约 2000 bp，在固体平板上获得了约 3750 个转化子，库容约 7.5 Mb。从中筛选到 27 株组成型表达 GFP 的阳性克隆，*Sau*3A Ⅰ酶切鉴定表明 RFf2、RFf3、RFf4、RFf6、RFf7、RFf8 有完全相同的限制性内切酶图谱，为包含同一插入片段的克隆。经鉴定共获得

22 个各不相同的组成型启动子片段。外源插入片段大小 1.2–4.9 kb 不等，2 kb 左右的居多，启动子强度也有差异，较强的启动 GFP 表达所达到的荧光强度是弱的 5–10 倍(图 1)。启动子强度和外源插入片段的长度大小无线性关系。对这些已测定的序列在 NCBI BLASTn 2.2.15 序列相似性比对表明，均只有局部 17–26 bp、比对分值在 40–50 的相似性，为新序列。用 1999 NNPP version 2.2 (Neural Network Promoter Prediction)预测启动子位置表明，这些序列片段在许多位置都含有与原核或真核启动子保守序列区域(–10 区、–35 区、TATA 框等)相一致的序列，分值为 0.80–1.00。有的插入片段在预测的启动子区域上游、下游或内部的位置上有 cAMP-CAP 结合位点特征序列 TGTGA。将部分启动子 RFb4、RFd1 等构建到麦芽四糖淀粉酶基因编码序列的上游在 *E. coli* 中重组表达，表明均能启动该酶的表达。

2.2 组成型启动子 RFa1 分析及启动子功能区域确定

从瘤胃混合微生物质粒文库中筛选到的组成型启动子 *RFa1*，外源插入片段长度 1202 bp，Blastx 比对表明其插入片段 325–1125 bp 所翻译的蛋白序列与拜氏梭菌 *Clostridium beijerincki* NCIMB 8052 卤代烷脱卤酶超家族水解酶亚家族(HAD-superfamily hydrolase subfamily IIB) Cof 蛋白有 41%的相似性。

RFa1 外源插入片段序列经 NNPP version 2.2 预测，在插入片段 250–295 bp 位置预测有原核启动子，分值为 0.98，在与其重叠的位置 261–306 bp 处也有一原核启动子，分值为 0.98 (表 2)。考虑在这段位置上含有类似于 *galP1*、*galP2* 的双启动子。在此段区域后面利用 ORF 预测软件预测有结构基因开放阅读框，起始密码子 ATG 在 322 bp 处，终止密码子 TGA 在 1135 bp 处。在此预测的最有可能的启动子区域 261–306 bp 位置的转录起始位点上游 188 bp 处下游 36 bp (预测的 ORF 起始密码子下游 15 bp 处)设计 PCR 引物 RFa1P-F/RFa1P-R 扩增 224 bp 的基因片段，将其克隆到启动子探针型载体 pUC18-GFP 的 *Pst* Ⅰ、*Kpn* Ⅰ之间，截取的这一段启动子 224 bp 片段在有核糖体结合位点存

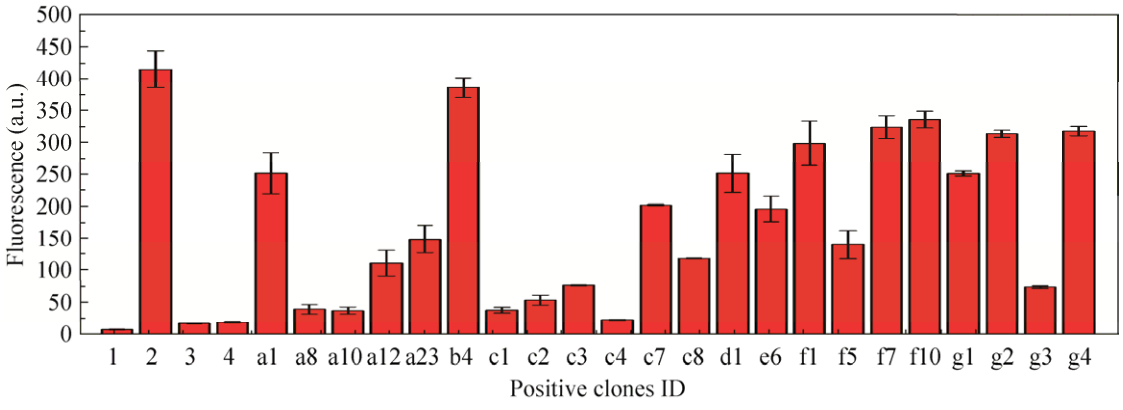

图 1. 22 个显绿色荧光克隆菌液荧光值

Figure 1. Fluorescence intensity of GFP positive clones. 1: pUC18-GFP (DH5α); 2: pUC19-GFP [BL21(DE3)]; 3: pUC19-GFP (DH5α); 4: pET28a-GFP (BL21 (DE3)). a1, a8, a10, a12, a23, b4, d1 are Top10 transformants with respective plasmids; c1, c2, c3, c4, c7, c8,. e6, f1, f5, f7, f10, g1, g2, g3, g4 are DH5αt transformants with respective plasmids.

表 2. *RFa1* 序列启动子预测

Table 2. Promoter sequences prediction of *RFa1*

Start	End	Score	Protein Sequence (5′→3′)
50	95	0.97	GTGTTTTTAGCTTGATTATTTGCTTCTTCTAAGATTTTCT**G**AGCTTCTTT
99	144	0.91	TCTTGGCATTCTTACCGATTACAGAGTAATACAGTAAAAT**G**ACAACAGCT
168	213	0.86	AAATGGAAAGTACATCAAAATTCATAATCTATACCTCCAT**T**TATTGCTAC
250	295	0.98	TATATTGGACCGATTTTCACAATAAGAAACAGAATCGGTGT**A**CTTTCAGA
261	306	0.98	GATTTTCACAATAAGAAACAGAATCGGTGTACTTTCAGAG**T**TCTTCGTGT
309	354	0.86	GTATGATTTAGATATGTATAAAATGATCGTAACAGATCTC**G**ATGAGACTC
470	515	0.97	TTTTGAAAGAAATCGGATTATACGACAAAGAAAATACCTA**C**TCCATTTCA
512	557	0.86	CCATTTCACTCAATGGCGCTATCATCACTGAAAATAAAGG**A**AACAGGATC
744	789	0.94	ATTCCTGAAAAACGACAGGATCATGAAGATATTATTCGTC**A**ATACGGATA
886	931	0.91	CCGGGTGTCAATAAAGGCGATGGCTTACATAAACTGTGTG**A**AATACTTGA
1190	1235	0.81	TCTTTGGATTGGCCTGCAGGTCGACTCTAGAGGATCCCCG**G**GTACCTAAT

Bold character is the possible transcription initiation site for the predicted promoters.

在时能够启动 GFP 的转录和翻译(克隆 *RFa1p*-pUC18GFP)，当此段直接与不含核糖体结合位点的 GFP 编码基因相连，则不能启动 GFP 翻译(克隆 *RFa1p*-pUC18GFP2)。说明该 224 bp 启动子区域内部并没有核糖体结合位点，需要额外添加核糖体结合位点才能启动下游基因的翻译。*RFa1p* (224 bp)非但保留了 *RFa1* (1202 bp)片段的启动子活性，启动子强度还略有增高。*RFa1*-pUC18GFP (DH5α)荧光强度为 258.5；*RFa1p*-pUC18GFP (DH5α)荧光强度为 297.5。蛋白电泳的结果表明，*RFa1p* 启动子片段启动 GFP 蛋白的转录、翻译得到的蛋白量与 pUC19 质粒载体上乳糖操纵子 *LacZ* 的启动子 IPTG 诱导下，启动活性相当。

对于截短的 *RFa1p* (224 bp)启动子序列，在距第一个预测的转录位点 GTG**T**AC 上游 90 bp，第三个预测的转录起始位点 CTC**G**AT 上游 88 bp，第三个转录起始位点 CTC**G**AT 上游 43 bp 处设计 PCR 引物，分别截取长度为 137、76、31 bp 的启动子片段 *RFa1p1*、*RFa1p2* 和 *RFa1p3*，连接在质粒型探针载体 pUC18GFP 的酶切位点 *Pst*Ⅰ、*Kpn*Ⅰ之间(图 2)，实验结果表明，*RFa1p1*、*RFa1p2* 片断均能够启动 GFP 表达，*RFa1p3* 片段不能启动 GFP 表达。启动子功能区域成功缩小到 76 bp。

2.3 组成型启动子 RFb4 分析及启动子功能区域确定

将 *RFb4* 外源基因片段(长度约为 2800 bp)连接到以施氏假单胞菌麦芽四糖淀粉酶为报告基因的探针型质粒载体 pUC18-*Psmta* 上，并转化大肠杆菌 DH5α 得到 b4m (DH5α)-4 和 b4m (DH5α)-5 两个克隆。将 *RFb4* 外源基因片段 *Kpn*Ⅰ酶切片段(长度约为 2000 bp)连接到以施氏假单胞菌麦芽四糖淀粉酶为报告基因的探针型质粒载体 pUC18-Psmta 上，并转化大肠杆菌 DH5α 得到 b4m (DH5α)-1 和 b4m (DH5α)-3 两个克隆。完整的 *RFb4* 以及 *RFb4-truncation* 两段序列均可以启动麦芽四糖淀粉酶的转录、翻译。*RFb4* 启动子强度显著高于 pGEM-T 载体(Promega)上的 SP6 启动子以及乳糖操纵子 *lacZ* 基因启动子强度(图 3、图 4)。*RFb4-truncation* 启动子强度与乳糖操纵子 *lacZ* 基因启动子强度相当(图 4)。

RFb4 外源插入片段长度约为 2800 bp，在距离绿色荧光蛋白编码基因前 *Pst*Ⅰ酶切位点 547 bp 的位置设计 PCR 引物(RFb4p-F/pUC18F-R)，得到截短的 547 bp 的启动子序列片段(图 5)可以启动绿

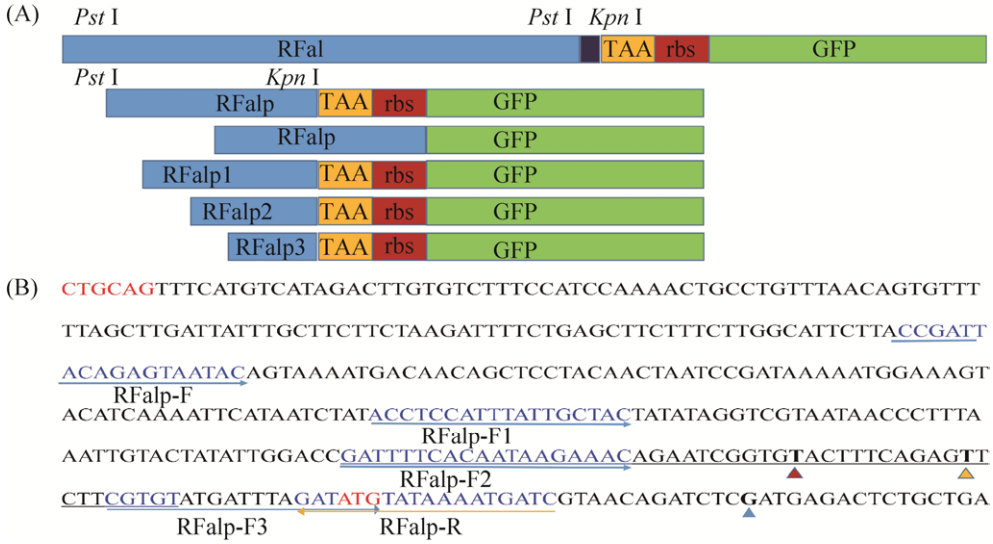

图 2. ***RFa1*** **序列截短设计**

Figure 2. Design of *RFa1* promoter sequence truncation. *RFa1* (1202 bp), *RFa1p* (224 bp), *RFa1GFP2*, *RFa1p1* (137 bp), *RFa1p2* (76 bp), *RFa1p3* (31 bp) sequence segments. TAA: TAATTAATTAA Triplet stop codon. Rbs: Ribosome binding site. GFP: Green fluorescent protein (mutGFP2). B: Design of RFa1 Promoter sequence truncation. *RFa1p* (224 bp) amplification primer RFa1p-F/RFa1p-R; *RFa1p1* (137 bp) amplification primer RFa1p-F1/RFa1p-R, *RFa1p2* (76 bp) amplification primer RFa1p-F2/RFa1p-R, *RFa1p3*(31 bp) amplification primerR Fa1p-F3/RFa1p-R; The underlined sequence is the predicted promoter region. Bold character is the possible transcription initiation site for the predicted promoters.

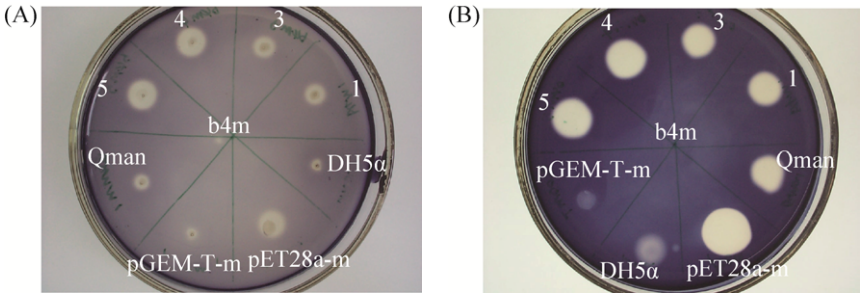

图 3. ***RFb4*** **启动子启动麦芽四糖淀粉酶基因异源表达**

Figure 3. 1,4-α-glucan maltotetraohydrolase gene expression in *E. coli* induced by different kinds of promoters. b4m-1, b4m-3: *RFb4-truncation*-pUC18- *Psmta* (DH5α); b4m-4, b4m-5: *RFb4*-pUC18-*Psmta* (DH5α); Qmam: *Qmam1*-pUC18-*Psmta* (DH5α), DNA segments obtained from *Trichoderma reesei* strain QM9414 which can initiated GFP transcription in *E. coli* (obtained by our laboratory); pGEM-T-m: *Psmta*-pGEM-T (DH5α) (*Psmta* coding gene were inserted into pGEM-T vector just after SP6 promoter sequence); pET28a-m: pET28a-*Psmta* [BL21(DE3)]. A: Different kinds of single clones were inoculated on LB solid plate with 100 μg/mL Amp and 0.5% soluble starch. They were cultured at 37 °C for 16 hours and stained with iodine. B: The lysed bacterial solutions (8 μL) were placed on agarose solid plate with 1% soluble starch. They were incubated at 50 ° C for 10 minutes and then were stained with iodine. pET28a-*Psmta* [BL21(DE3)] transformant were cultured with 0.1 mol/L IPTG for 4 h at 30 °C after cultured for 12 h at 37 °C.

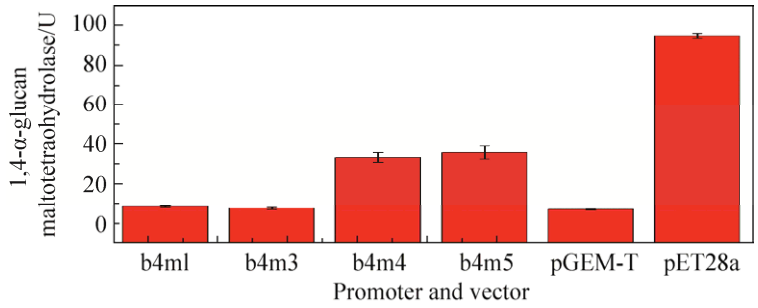

图 4. ***RFb4*** **启动子启动麦芽四糖淀粉酶基因异源表达的转化子的粗酶液活性**

Figure 4. 1,4-α-glucan maltotetraohydrolase activity of different promoter system. The b4m-1, b4m-3: *RFb4-truncation*-pUC18-*Psmta* (DH5α); b4m-4, b4m-5: *RFb4*-pUC18-*Psmta* (DH5α), were inoculated in liquid LB medium at 37 °C, pGEM-T-*Psmta* (DH5α) pET28a-*Psmta* [BL21(DE3)] were inoculated in liquid LB medium with 0.1 mol/L IPTG and 100 μg/mL Amp [pGEM-T-*Psmta* (DH5α)] or 50 μg/mL Kan [pET28a-*Psmta* (BL21(DE3))]. After cultivated at 37 °C 200 r/min for 16 hours, 4 mL of the bacterial suspension with 6×10^8 cell/mL was extracted and then the cells were broken using ultrasonication, and the activity of 1,4-α-glucan maltotetraohydrolase was measured.

GGTACCGTACGAGATGTGGAAGATCCTTCGCAAGGCGAACCGCTGTTAAAGCACTCTGAAATCGACT

TATGCCGATGGTCGCGGCAGAAGGTGAACCGACGCGATTGGCTTGCGGGGAAAGGTCGGATCGTCA

AGAACAGTCTCTACCTTGGTGTGAAGGAAGATGCACTCTTGCGGACGCGCGTTGGTCGGCGGCGATA

RFb4-F

GAGGCCCAAGGCCCGTCTGGGGAGGTCGTGCCGTCTGAGGGTTTCCCGAAGACGATCATATATGACC

GTTGGCCGGCCGTTTCCGTCAAGATGGTACGCAAGGATGTTTGGCTCGTCGATTTTGGCGTCAACTTC

RFb4p-R2

GCCGGAACGTACAGGGCGACATTGCGCGGCGTTCCGAGAGGTGTGACGGTGATGTTTCGTGCGGGG

GAGCGCGTGAACGACGATGGCACGGTGAACGTCAAGACGGCGGTGGCGGGACAGATAAAGAATCC

GGCAAGAGGACCTCTTTTCGATCTGGCGGAACAGCGCGCCGAATGGGTGTCTGGCGGAGACCCCGT

GGCAACGTTCGAGCCGCGCTTCACGTTCCATGCGTTCCGCTATCTTCAGGTGGAGGGGCTTAAGGAT

RFb4p-R1

GATCTGTCGCCTGGGGATTTTGAGGCACTGGCATGGTCGGCCGATGTTCGGGACGGCGCGCATTTCG

AGTGTTCCAATCCGAAGATCAACCTTCTGCACGAGGTTTGCCGCCGCACTTTCCGCGCAAACCTGCA

GGTCGACTCTAGAGGATCCCCGGGTACCTAATTAATTAAG**AAGGAG**ATATACATATGAGTAAAGGA

图 5. ***RFb4*** **序列截短设计**

Figure 5. Design of *RFb4* promoter sequence truncation. Design of RFb4 promoter sequence truncation: *RFb4p* (547 bp) amplification primer RFb4-F/pUC18G-R; *RFb4p1* (395 bp) amplification primer RFb4-F/RFb4p-R1, *RFb4p2* (134 bp) amplification primer RFb4-F/RFb4p-R2. The underlined sequences are the restriction endonuclease cleavage sites. The red characters are the predicted eukaryotic promoter region. Bold characters indicates the possible transcription initiation site for the predicted.

色荧光蛋白的表达。通过 NNPP version 2.2 预测，在距离 *Pst* Ⅰ插入位点 480–430 bp 位置预测有真核启动子，分值为 0.93，在该位置预测的转录起始位点上游 27 bp 至下游 288 bp 处设计 PCR 引物 RFb4p-F/RFb4p-R2，扩增得到 134 bp 的 DNA 片段 (*RFb4p2*) (图 5)，这段序列与 pUC18GFP 探针型质粒载体 *Pst* Ⅰ和 *Eco*R Ⅰ酶切片段连接，不能启动 GFP 表达。在该位置预测的转录起始位点上游 107 bp 至

下游 288 bp 处设计 PCR 引物 RFb4p-F/RFb4p-R1，扩增得到 395 bp 的 DNA 片段(*RFb4p1*)，这段区域与不含核糖体结合位点的 GFP 结构基因连接(*RFb4p1*-pUC18GFP2)，不能启动 GFP 表达。

3 讨论

目前已报到的原核和真核生物的启动子数据库有超过 40 多个[23]。PromEC 数据库包含大肠杆菌 472 个启动子(−75:+25 bp)和转录起始位点(TSS)[24]，RegulonDB 数据库包含 8597 个大肠杆菌 *E. coli* K-12 的启动子序列[25]，EcoCyc 数据库包含 3841 个大肠杆菌 *E. coli* K-12 的启动子序列[26]。从瘤胃胃液混合微生物总 DNA 所构建的元基因组质粒文库，筛选到的组成型启动子片段基本上都是新序列，Blastn 比对只有局部 20 bp 左右的相似性，Blastx 比对蛋白相似性也基本上在 40%左右，充分说明元基因组文库能够获得许多新的遗传信息。这些启动子强度各有不同，涵盖了一定的范围，可不同程度调节目的基因的表达量。后续构建载体可依不同的需要进行选择。报告基因由绿色荧光蛋白换作麦芽四糖淀粉酶或木聚糖酶时启动子仍能发挥作用，假阳性低。利用现有的各种启动子预测软件分析验证启动子功能区域与传统的利用放射性同位素的 DNA 足迹法和 5′-RACE 验证 RNA 聚合酶结合序列以及转录起始位点的方法相辅相成。前者可避免使用同位素，方法简单，但如果需要准确确定转录起始位点的话需要进一步作 5′-RACE 验证。

筛选到的 *RFa1p2*(76 bp)的启动子片段 Blastn 比对表明，其类似–35 区的部分 GTATGATTTAGATATGTAT 在许多真核生物的染色体上都有分布，如斑马鱼、小鼠 1 号、17 号、9 号染色体、人 1 号染色体，水稻 4 号染色体，相似性在 96%–100%。这些区域有的位于基因间重复序列区，有的位于基因的内含子序列区，有的距重复序列 AT 富含区 200–300 bp，有的位于已注释结构基因上游 3 kb 左右，有的还没有被注释。但这些区域能够被大肠杆菌的 RNA 聚合酶识别，一方面在其附近很可能存在结构基因或者调控序列，为这些全基因组测序的真核生物后续基因组注释提供了一定的信息，另一方面可以作为构建原核、真核生物穿梭载体的元件。筛选到的 *RFb4p* (547 bp)的启动子片段 Blastn 比对表明，局部区域与链霉菌属、慢生根瘤菌属、假诺卡氏菌属的基因组 DNA 片段以及真核生物醉蝶花属、猎豹、绵羊、山羊、空齿鹿的 mRNA 序列片段有 82%–93%的相似性。

对于用于微生物发酵调控的各种表达载体的构建，往往趋向于选择诱导型启动子，通过加入诱导物或者温度调控等开启外源基因的表达，这样可以当菌体稳定生长获得一定的菌体量后再实行信号调控更有利于外源基因的表达。但是往往诱导型强启动子如 pET 系统(T7 启动子和 LacZ 操作子序列共同调控)。一方面需要加 IPTG，对于大规模发酵成本较高，对于食品药品等发酵后续还需要去除诱导物；一方面大量表达容易形成包涵体，而且不利于一些对宿主菌有毒害的毒性蛋白的稳定表达。对于其他以代谢底物为诱导物的诱导型启动子，往往随着发酵过程底物的消耗产物的积累出现反馈抑制，不利于发酵的稳定进行。所以强度适中的组成型启动子鉴于可以在菌体的生长期稳定表达目的蛋白，不需要额外添加诱导物或温度控制增加成本，有一定的研究利用价值。而且已经缩小到 76 bp (*RFa1p2*)、547 bp (*RFb4p*)的启动子片段可方便地直接构建到相应载体上，调控外源基因表达。此外，构建的文库也可进一

步添加相应底物筛选诱导型启动子，而且由于以绿色荧光蛋白为选择标记，可以方便地利用流氏细胞仪进行高通量筛选。

参 考 文 献

[1] Gilbert W. Starting and stopping sequences for the RNA polymerase//Losick R, Chamberlin M. RNA Polymerase. Cold Spring Harbor, NY: Cold Spring Harbor Lab, 1976: 193–205.

[2] McClure WR. Mechanism and control of transcription initiation in prokaryotes. *Annual Review of Biochemistry*, 1985, 54: 171–204.

[3] Jensen PR, Hammer K. The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters. *Applied and Environmental Microbiology*, 1998, 64(1): 82–87.

[4] Hammer K, Mijakovic I, Jensen PR. Synthetic promoter libraries-tuning of gene expression. *Trends in Biotechnology*, 2006, 24(2): 53–55.

[5] Jensen PR, Hammer K. Artificial promoters for metabolic optimization. *Biotechnology and Bioengineering*, 1998, 58(2/3): 191–195.

[6] Gilman J, Love J. Synthetic promoter design for new microbial chassis. *Biochemical Society Transactions*, 2016, 44(3): 731–737.

[7] Kagiya G, Ogawa R, Hatashita M, Takagi K, Kodaki T, Hiroishi S, Yamamoto K. Generation of a strong promoter for *Escherichia coli* from eukaryotic genome DNA. *Journal of Biotechnology*, 2005, 115(3): 239–248.

[8] Neve RL, West RW, Rodriguez RL. Eukaryotic DNA fragments which act as promoters for a plasmid gene. *Nature*, 1979, 277(5694): 324–325.

[9] West RW Jr, Neve RL, Rodriguez RL. Construction and characterization of *E. coli* promoter-probe plasmid vectors I. Cloning of promoter-containing DNA fragments. *Gene*, 1979, 7(3/4): 271–288.

[10] West RW Jr, Rodriguez RL. Construction and characterization of *E. coli* promoterprobe plasmid vectors II. RNA polymerase binding studies on antibiotic-resistance promoters. *Gene*, 1980, 9(3/4): 175–193.

[11] West RW Jr, Rodriguez RL. Construction and characterization of *E. coli* promoter-probe plasmid vectors III. pBR322 derivatives with deletions in the tetracycline resistance promoter region. *Gene*, 1982, 20(2): 291–304.

[12] Tajima K, Aminov RI, Nagamine T, Ogata K, Nakamura M, Matsui H, Benno Y. Rumen bacterial diversity as determined by sequence analysis of 16S rDNA libraries. *FEMS Microbiology Ecology*, 1999, 29(2): 159–169.

[13] Koike S, Yoshitani S, Kobayashi Y, Tanaka K. Phylogenetic analysis of fiber-associated rumen bacterial community and PCR detection of uncultured bacteria. *FEMS Microbiology Letters*, 2003, 229(1): 23–30.

[14] Gabor EM, Alkema WBL, Janssen DB. Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environmental Microbiology*, 2004, 6(9): 879–886.

[15] Lu CH, Bentley WE, Rao G. A high-throughput approach to promoter study using green fluorescent protein. *Biotechnology Progress*, 2004, 20(6): 1634–1640.

[16] Cormack BP, Valdivia RH, Falkow S. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene*, 1996, 173(1): 33–38.

[17] Yang RH, Wang JQ, Luo SP, Dong ZY. Extraction and purification of DNA from environmental rumen samples. *Journal of Xinjiang Agricultural University*, 2005, 28(2): 39–42. (in Chinese)
杨瑞红, 王加启, 罗淑萍, 董志扬. 奶牛瘤胃胃液微生物总DNA的提取和纯化. 新疆农业大学学报, 2005, 28(2): 39–42.

[18] Krause DO, Smith WJ, McSweeney CS. Extraction of microbial DNA from rumen contents containing plant tannins. *Biotechniques*, 2001, 31(2): 294–298.

[19] Sharma R, John SJ, Damgaard M, McAllister TA. Extraction of PCR-quality plant and microbial DNA from total rumen contents. *Biotechniques*, 2003, 34(1): 92–94, 96–97.

[20] Uchiyama T, Abe T, Ikemura T, Watanabe K. Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nature Biotechnology*, 2005, 23(1): 88–93.

[21] Hampshire JB, Waibel AH. A novel objective function for improved phoneme recognition using time-delay neural networks. *IEEE Transactions on Neural Networks*, 1990, 1(2): 216–228.

[22] Reese MG, Eeckman FH. Novel neural network algorithms for improved eukaryotic promoter site recognition//Proceedings of the Seventh International Genome Sequencing and Analysis Conference. Hilton Head Island, South Carolina: Hyatt Regency, 1995.

[23] Majewska M, Wysokińska H, Kuźma Ł, Szymczyk P. Eukaryotic and prokaryotic promoter databases as valuable tools in exploring the regulation of gene transcription: a comprehensive overview. *Gene*, 2018, 644: 38–48.

[24] Hershberg R, Bejerano G, Santos-Zavaleta A, Margalit H. PromEC: an updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites. *Nucleic Acids Research*, 2001, 29(1): 277.

[25] Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, Alquicira-Hernández K, Martínez-Flores I, Pannier L, Castro-Mondragón JA, Medina-Rivera A, Solano-Lira H, Bonavides-Martínez C, Pérez-Rueda E, Alquicira-Hernández S, Porrón-Sotelo L, López-Fuentes A, Hernández-Koutoucheva A, Del Moral-Chávez V, Rinaldi F, Collado-Vides J. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, 2016, 44(D1): D133-D143.

[26] Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, Fulcher C, Gama-Castro S, Kothari A, Krummenacker M, Latendresse M, Muñiz-Rascado L, Ong Q, Paley S, Peralta-Gil M, Subhraveti P, Velázquez-Ramírez DA, Weaver D, Collado-Vides J, Paulsen I, Karp PD. The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Research*, 2017, 45(D1): D543–D550.

New constitutive promoters screened from metagenomic library of rumen microbes

Li Wang[1*], Yun Zhao[2], Qian Yang[1], Xin Dai[1], Yaxin Zhu[1], Zhiyang Dong[1*]

[1] State Key Laboratory of Microbial Resources (SKLMR), Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

[2] Laboratory of Protein and Peptide Drugs (LPPD), Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

Abstract: [Objective] Screening of novel promoter elements from the genome of microorganisms of extreme environmental origin for the design of synthetic biological chassis cells. **[Methods]** We used a promoter-probe plasmid pUC18-GFP containing a green fluorescent protein structural gene and a ribosome bind site to construct a rumen metagenomic library. This method allows us to obtain the DNA fragments with constitutive promoter function rapidly and efficiently from this library. We obtained possible promoter regions through the neural network-based promoter prediction analysis. Then, we verified the function of the promoter initiation by using GFP and maltotetraose amylase from *Pseudomonas stutzeri* as the reporter. **[Results]** We obtained twenty-two DNA fragments functioning as constitutive promoters from about 3750 transformants. These fragments share very low sequence identities with the reported gene sequences in the NCBI database, and present different starting efficiencies. In addition, we obtained two new promoter fragments *RFa1p2* (76 bp) and *RFb4p* (547 bp) by promoter prediction and sub-cloning. These new constitutive promoters are able to express heterologous proteins efficiently in the absence of any inductor in the genetically engineered *E. coli* cells.

Keywords: metagenome, constitutive promoter, rumen microbes, green fluorescent protein, maltotetraohydrolase

(本文责编：张晓丽)

Supported by the National Natural Science Foundation of China (31300007, 30770053, 31240050)

*Corresponding authors. Zhiyang Dong, Tel/Fax: +86-10-64807337, E-mail: dongzy@im.ac.cn; Li Wang, Tel: +86-10-64807331, E-mail: wangli07@im.ac.cn

Received: 22 January 2019; Revised: 18 April 2019; Published online: 29 April 2019