微生物学报
Acta Microbiologica Sinica
2024, 64(11): 4440–4454
http://journals.im.ac.cn/actamicrocn
DOI: 10.13343/j.cnki.wsxb.20240378

基于 SHAPE-seq 技术解析大肠杆菌典型 5′mRNA 结构特征及其对基因表达的影响

张金鹏[1,2]，任家卫[1,2]，梅子轮[1,2]，张晓梅[3]，徐国强[1,2]，李会[3]，史劲松[3]，许正宏[4]，张晓娟[1,2*]

1 江南大学 生物工程学院，工业生物技术教育部重点实验室，江苏 无锡 214122
2 江南大学，粮食发酵与食品生物制造国家工程研究中心，江苏 无锡 214122
3 江南大学 生命科学与健康工程学院，江苏 无锡 214122
4 四川大学 轻工科学与工程学院，四川 成都 610065

张金鹏, 任家卫, 梅子轮, 张晓梅, 徐国强, 李会, 史劲松, 许正宏, 张晓娟. 基于 SHAPE-seq 技术解析大肠杆菌典型 5′mRNA 结构特征及其对基因表达的影响[J]. 微生物学报, 2024, 64(11): 4440-4454.
ZHANG Jinpeng, REN Jiawei, MEI Zilun, ZHANG Xiaomei, XU Guoqiang, LI Hui, SHI Jinsong, XU Zhenghong, ZHANG Xiaojuan. SHAPE-seq reveals the typical 5′mRNA structure of *Escherichia coli* and its effect on gene expression[J]. Acta Microbiologica Sinica, 2024, 64(11): 4440-4454.

摘　要：【目的】在原核生物基因表达过程中，由于转录本 mRNA 半衰期较短，其抗降解能力和招募核糖体启动翻译反应的能力对基因表达的影响显著。然而 mRNA 在其 5′端存在多个功能区域，可以影响其半衰期的长短进而影响目的基因的表达，其中主要包括：Shine-Dalgarno (SD)序列以及其上下游的核糖体“等候”位点(translational standby site, TSS)、N 端部分编码序列(N-terminal coding sequence, NCS)。各区域由于其自身和跨区域的结构差异会影响基因表达，因此解析各功能区域的构效关系至关重要。【方法】采用引物延伸测序(SHAPE-seq)技术，以大肠杆菌为宿主解析了 7 种结构不同的 5′mRNA 胞内的结构特点，采集了其调控下 mRNA 丰度以及蛋白表达量。【结果】发现非结构化的 NCS 调控下 mRNA 丰度和蛋白量分别提高了 10 倍、19 倍；形成二级结构茎长为 10 nt 的 TSS 有利于提高 mRNA 丰度；SD 序列被包裹形成二级结构时会影响其介导的翻译起始效率(蛋白表达下降 10%)；上述较优的 TSS 和 NCS 的组合调控下，mRNA 丰度和蛋白量显著提高(11 倍、60 倍)。【结论】本研究解析了 5′mRNA 各区域有利于原核生物基因表达的结构特征，初步建立了各功能区域的构效关系，为工业微生物目标基因表达提供了新型调控元件。

资助项目：国家自然科学基金(32171421)
This work was supported by the National Natural Science Foundation of China (32171421).
*Corresponding author. E-mail: zhangxj@jiangnan.edu.cn
Received: 2024-06-20; Accepted: 2024-08-29; Published online: 2024-09-10

关键词： 核糖体等候位点；N 端编码序列；SD 序列；引物延伸测序；构效关系

SHAPE-seq reveals the typical 5′mRNA structure of *Escherichia coli* and its effect on gene expression

ZHANG Jinpeng[1,2], REN Jiawei[1,2], MEI Zilun[1,2], ZHANG Xiaomei[3], XU Guoqiang[1,2], LI Hui[3], SHI Jinsong[3], XU Zhenghong[4], ZHANG Xiaojuan[1,2*]

1 Key Laboratory of Industrial Biotechnology of Ministry of Education, School of Biotechnology, Jiangnan University, Wuxi 214122, Jiangsu, China
2 National Engineering Research Center of Cereal Fermentation and Food Biomanufacturing, Jiangnan University, Wuxi 214122, Jiangsu, China
3 School of Life Sciences and Health Engineering, Jiangnan University, Wuxi 214122, Jiangsu, China
4 College of Biomass Science and Engineering, Sichuan University, Chengdu 610065, Sichuan, China

Abstract: [Objective] Due to the short half-life of mRNA transcript, their anti-degradation ability and ability to recruit ribosomes to initiate translation have significant effects on the gene expression in prokaryotes. However, the multiple functional regions in the 5′ end, of mRNA can affect the half-life and thus the expression of target genes, including the Shine-Dalgarno (SD) sequence and its upstream and downstream translational standby sites (TSSs) and the N-terminal coding sequence (NCS). The own and cross-regional structural differences will affect gene expression. Therefore, it is important to analyze the structure-activity relationship of each functional region. **[Methods]** We employed primer extension sequencing (SHAPE-seq) to analyze the structural characteristics of seven different 5′mRNAs in *Escherichia coli* and collected the mRNA abundance and protein level information under their regulation. **[Results]** Under the regulation of unstructured NCS, the mRNA abundance and protein level were increased by10 and 19 times, respectively. The formation of secondary structure TSS with a stem length of 10 nt increased the mRNA abundance. When the SD sequence was wrapped to form a secondary structure, the efficiency of translation initiation mediated by the SD sequence was affected (protein level downregulation by 10%). The combination of TSS and NCS significantly increased the mRNA abundance and protein level by 11 and 60 folds, respectively. **[Conclusion]** We characterized the structure of each region of 5′mRNA conducive to prokaryotic gene expression and revealed the structure-activity relationship of each functional region, providing a new regulatory element for target gene expression in industrial microorganisms.

Keywords: translational standby site; N-terminal coding sequence; SD sequence; SHAPE-seq; structure-activity relationship

原核生物的基因表达过程由于转录和翻译在空间和时间上是同步发生的，因此目的基因的表达必须在 mRNA 降解前完成[1-2]。然而原核生物体内存在多类型的核酸降解酶[3]，导致 mRNA 降解严重，大多数 mRNA 的半衰期约为 2−4 min[4]，因此在短时间内高效地招募核糖体启动翻译反

应，并且提高 mRNA 抵抗水解的能力，对基因表达至关重要。

据报道 mRNA 5′端至起始密码子后 100−200 nt 的序列对翻译反应和 mRNA 抗降解能力具有重要的影响[5-6]。这一区域可以划分为 3 种：(1) 核糖体等候区(translational standby site, TSS)[7]；(2) 以 Shine-Dalgarno (SD)序列为核心的负责招募核糖体启动翻译反应的区域[8]；(3) 位于起始密码子后的长度约为 50 nt 的 N 端编码序列(N-terminal coding sequence, NCS)[9-10]。其中 TSS 位于 mRNA 上游，是核酸降解酶 E (RNase E)的主要靶向区域，该区域恰当的二级结构会显著提高 mRNA 半衰期，但同时可能对容留核糖体产生负面影响[3,11-12]；SD 序列负责招募核糖体[13-14]，它与核糖体的匹配度以及是否被二级结构包裹，都会直接影响与核糖体结合的速度，进而影响翻译反应效率；而 NCS 由于物理空间上与 TSS 和 SD 区域接近，其序列和是否存在二级结构都会影响上述两个区域的生理功能，影响翻译复合物的延 mRNA 的滑动，进而影响基因表达[5,15]。上述各功能区域的调控效果都会受到其本身结构的影响，但 mRNA 在胞内的结构因其单链化学结构而变得非常复杂[16]，难以精准进行表征其在胞内的结构，这也是 mRNA 水平的基因表达调控机制复杂的重要原因，目前大多研究通过自由能计算来预测 mRNA 二级结构，准确度较低[17-18]。这对解析 mRNA 各功能区域的构效关系造成了严重障碍。

针对这一问题，本研究采用引物延伸的选择性 2′-羟基酰化分析法(selective 2′-hydroxyl acylation analyzed by primer extension sequencing, SHAPE-seq)技术[19]，表征 mRNA 5′端各功能区域在大肠杆菌细胞内的结构。其原理是针对 mRNA 序列中未发生碱基互补配对的区域(单链区)添加化学修饰，这一修饰会造成 mRNA 在进行逆转录时被迫终止，进而形成一个截断的 cDNA 文库。对 mRNA 的逆转录产物(cDNA 文库)进行高通量测序(high-throughput sequencing, HTS)，分析文库多态性、计算 RNA 的反应谱，可以得到各碱基的修饰概率(反应性)，即各位点的未配对的概率[20]。本研究通过在大肠杆菌细胞内解析典型的 mRNA 5′端区域结构特征，并对比通过传统自由能计算预测的 mRNA 二级结构，结合蛋白表达量、转录本丰度的定量分析，建立 mRNA 5′端区域的构效关系，以提高翻译效率和 mRNA 丰度为目标，明确了对基因表达影响显著的各区域的设计参数，为原核生物调控基因表达提供了高性能的元件，也为 mRNA 5′端区域的人工设计提供了重要依据。同时本研究建立的跨区域结构表征-调控表型分析的研究思路也为相关构效关系的研究提供了可借鉴的方法。

1 材料与方法

1.1 材料

1.1.1 菌株和质粒

本研究所用菌株和质粒见表 1。

1.1.2 引物

本研究所用引物名称及序列见表 2，引物均由苏州金唯智生物科技有限公司合成。

1.1.3 主要试剂和仪器

LB 培养基组分均购自国药集团化学试剂有限公司；*Bsa* I 快切酶购自 NEB 公司；cDNA 合成试剂盒购自赛默飞世尔科技公司；1-甲基-7-硝基靛红酸酐(1-methyl-7-nitroisatin anhydride, 1M7)购自 MCE 公司；RNA 提取试剂盒、Power qPCR premix (SYBR Green I)试剂盒、一步法 RT-PCR 扩增试剂盒均购自生工生物工程(上海)股份有限公司；NanoDrop 超微量核酸蛋白测定仪购自 ThermoFisher Scientific 公司。

表 1　本研究所用的菌株质粒

Table 1　Strains and plasmids used in this study

Strains and plasmids	Characteristics	Sources
Strains	*recA1*, *endA1*, *gyrA96*, *thi-1*, *hsdR17*($rk^{-}mk^{+}$), e14^{-} (mcrA^{-}), *supE44*, *relA1*,	Lab stock
Escherichia coli JM109	Δ(*lac-proAB*)/F′[*traD36*, *proAB*$^{+}$, *lacIq*, *lacZ*ΔM15]	
Plasmids		
pDXW-13	It carries a tac promoter and the reporter gene is *egfp*	Lab stock
pDXW-13-*Bsa* I	Two reverse complementary *Bsa* I restriction sites were inserted before the reporter gene *egfp*	This study
6A-NCSSH	The NCS carrying tac promoter is inserted after RBS to form the secondary structure	This study
6A-NCS	Carries tac promoters, an NCS is inserted after an RBS of 6 nt conservatism	This study
S10L4	With tac promoter, *egfp* is preinserted into a secondary structure with a stem length of 10 nt	This study
S10L10	With tac promoter, *egfp* is preinserted into a secondary structure with a loop length of 10 nt	This study
S18L4	With tac promoter, *egfp* is preinserted into a secondary structure with a stem length of 18 nt	This study
S10L10-6-NCS	S10L10 is the skeleton, with NCS inserted before *egfp*	This study
S10L4-6-NCS	S10L4 is the skeleton, with NCS inserted before *egfp*	This study

表 2　本研究所用的引物

Table 2　The primers used in this study

Primers name	Primer sequences (5′→3′)
Bsa I-F	ACAATTCGTAAGAGACCGCTTTCCAGATCTGTAACTTGTGGTCTCGG
Bsa I-R	AAAGCGGTCTCTTACGAATTGTTATCCGCTCACAATTCCACACATTATACG
SH10-F	CGTATTTAACCAGTAGATCATCCAATCTACTGGTGC
SH10-R	GTTTGCACCAGTAGATTGGATGATCTACTGGTTAAA
SH18-F	CGTATTTCCCAACACGGAATACCTACATCCCGGTAGGTATTCCGTGTTGGCT
SH18-R	GTTTAGCCAACACGGAATACCTACCGGGATGTAGGTATTCCGTGTTGGGAAA
L10-F	CGTATTCTAGTAGATCGCCTTCCCCCAAGCGATCTACTCT
L10-R	GTTTAGAGTAGATCGCTTGGGGGAAGGCGATCTACTAGAA
6A-F	AAACCAAGGAGCACACACACATG
6A-R	TCACCATGTGTGTGTGCTCCTTG
N39-F	TGAAAACACAAACCTCAACAAACACCACACACGAATTC
N39-R	TCACGAATTCGTGTGTGGTGTTTGTTGAGGTTTGTGTT
16S RNA-F	CCTACGGGAGGCAGCAG
16S RNA-R	ATTACCGCGGCTGCTGG
QGFP-F	GTGGTGCCCATCCTGGTC
QGFP-R	CTTCATGTGGTCGGGGTAGC
SHAPE-RT	5′-biotin GAACAGCTCCTCGCCCTT
SHAPE-F	GTGACTGGAGTTCAGACGTGTGCTC
1M7	CTTTCCCACACGACGCTCTTCCGATCTRRRYGAACAGCTCCTCGCCCT*T*G*C*T*C*A
DMSO	CTTTCCCTACACGACGCTCTTCCGATCTYYYRGAACAGCTCCTCGCCCT*T*G*C*T*C*A
PET	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
ssDNA	AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC

1.2 结构不同的 5′mRNA 重组质粒构建

为了构建含有结构不同的 5′mRNA 的重组质粒，使用 Golden Gate 无缝组装[21]的方式进行连接，以 pDXW-13 为模板，使用引物 *Bsa* I-F (ACAATTCGTAAGAGACCGCTTTCCAGATCTGTAACTTGTGGTCTCGG)和 *Bsa* I-R (AAAGCGGTCTCTTACGAATTGTTATCCGCTCACAATTCCACACATTATACG)在绿色荧光蛋白(enhanced green fluorescent protein, EGFP)前添加 *Bsa* I 识别位点，所有人工合成的各功能区序列正向引物的黏性末端设计为 CGTA，反向引物的黏性末端设计为 TCAC。将所有功能区的正、反单链引物均稀释至 20 μmol/L，各取 10 μL 混匀置于 PCR 仪中退火形成双链 DNA。PCR 反应条件：95℃ 5 min；每 8 s 下降 0.1 ℃，700 个循环降至 25 ℃。双链 DNA 5′端使用 T4 PNK 酶进行磷酸化并与载体进行连接，将连接产物转入大肠杆菌(JM109)，对转化子进行测序验证。

1.3 基于 SHAPE-seq 解析 mRNA 二级结构

1.3.1 第一链 cDNA 合成

将待测定胞内结构的重组菌株挑取单菌落至 10 mL LB 培养基中培养，以 2%的接种量分别转接到装有 3 mL LB 培养基中，添加浓度为 0.2 mol/L 的 IPTG 进行诱导，然后向培养基中分别添加 22.5 μL 1M7 或 DMSO，于 37 ℃、220 r/min 培养 3 min，然后进行 RNA 提取，将提取的 RNA 溶于 10 μL RNase-free ddH_2O 中，向 RNA 中加入 3 μL 0.5 μmol/L 的 SHAPE-RT 逆转录引物，95 ℃加热 5 min，加入逆转录酶用于合成第一链 cDNA，向 cDNA 中加入 1 μL NaOH (10 mol/L)置于 95 ℃反应 5 min，向每管 cDNA 中加入 5 μL 盐酸(3 mol/L)用以中和 NaOH。向每管 cDNA 中加入 78 μL 预冷的无水乙醇进行洗涤 cDNA，离心后将乙醇吸出，加入 500 μL 70%的预冷乙醇进行重悬后吸出，加入 22.5 μL RNase-free ddH_2O 用于重悬 cDNA。

1.3.2 将 ssDNA 与第一链 cDNA 进行连接

为了明确 cDNA 3′端的序列便于后续进行测序，采用环连接酶将 ssDNA 连接，并在 60 ℃反应 2 h。连接完成后，向 cDNA 中加入 70 μL RNase-free H_2O、10 μL 乙酸钠(3 mol/L)以及 1 μL 糖原(20 mg/mL)用于 cDNA 的可视化，最后加入 300 μL 预冷的 100%乙醇，混合后，进行洗涤，于 12 000 r/min 离心 10 min 后将乙醇吸出，加入 20 μL RNase-free H_2O 和 36 μL AMPure XP Beads 磁珠用于纯化 cDNA，纯化步骤按照说明书进行。纯化完成后，将 cDNA 溶于 20 μL TE 缓冲液中。

1.3.3 构建双链 DNA 文库

以 cDNA 为模板，通过使用引物 1M7/DMSO、SHAPE-F 以及 PET 对 cDNA 进行扩增，其中 1M7/DMSO 引物的作用是用于在合成的 DNA 中添加特定的识别序列，便于后续对序列进行区分，而 PET 的作用是用于将 cDNA 长度增加，便于后续满足测序要求。将获得的双链 DNA 送至武汉希望组生物科技有限公司进行扩增子测序。

1.3.4 5′mRNA 各区域碱基反应性计算

对测序得到的不同结构的 5′mRNA 的序列文件(FASTQ 文件)，首先使用生物信息学中常用的去除引物序列的工具(cutadapt[22]软件)，去除实验过程中添加的引物序列(1M7/DMSO)，指令为：cutadapt-j20-a-A-o-p；cutadapt-j20-g-G-o-p，然后使用序列统计工具(seqkit[23]软件)对每条序列进行读数统计，指令为：seqkit stats。接下来使用 SHAPE-seq 实验流程中的核苷酸反应性计算工具(spats[20]软件)来计算每个核苷酸被修饰的概率，指令为：spats targets.fa RRRY YYYR 1.fastq 2.fastq，其中 targets.fa 为目标 RNA 的序列，

1.fastq 和 2.fastq 为去掉引物序列后的测序数据。对得到的每一个核苷酸的修饰概率 θ，根据公式(1)和公式(2)进行归一化计算每个核苷酸被化学试剂(1M7)修饰的概率 ρ_i。

$$n=\sum_{i=1}^{L}\theta_i \tag{1}$$

$$\rho_i=\frac{\theta_i\cdot L}{n} \tag{2}$$

式中：θ_i 为 spats 输出的每个碱基修饰的反应性，L 为目标 mRNA 去除接头后的序列长度。根据计算得到的 ρ_i 使用 RNA 结构模拟工具(RNAstructure[24])来模拟每种 5′mRNA 在胞内的结构。

1.4 荧光强度分析

取 7 种结构不同的 5′mRNA 的大肠杆菌菌液各 1 h，12 000 r/min 离心 1 min 后，弃上清液，加入 1 mL PBS 洗涤 2 次，取 100 μL 菌液添加至酶标板中以检测其 OD_{600} 值和荧光强度。GFP 检测的激发波长为 488 nm，发射波长为 517 nm。按照公式(3)计算相对荧光强度。

$$\text{相对荧光强度}=\frac{OD_{485,525}-OD_{485,525\,\text{blank}}}{OD_{600}} \tag{3}$$

1.5 qRT-PCR 检测 mRNA 丰度

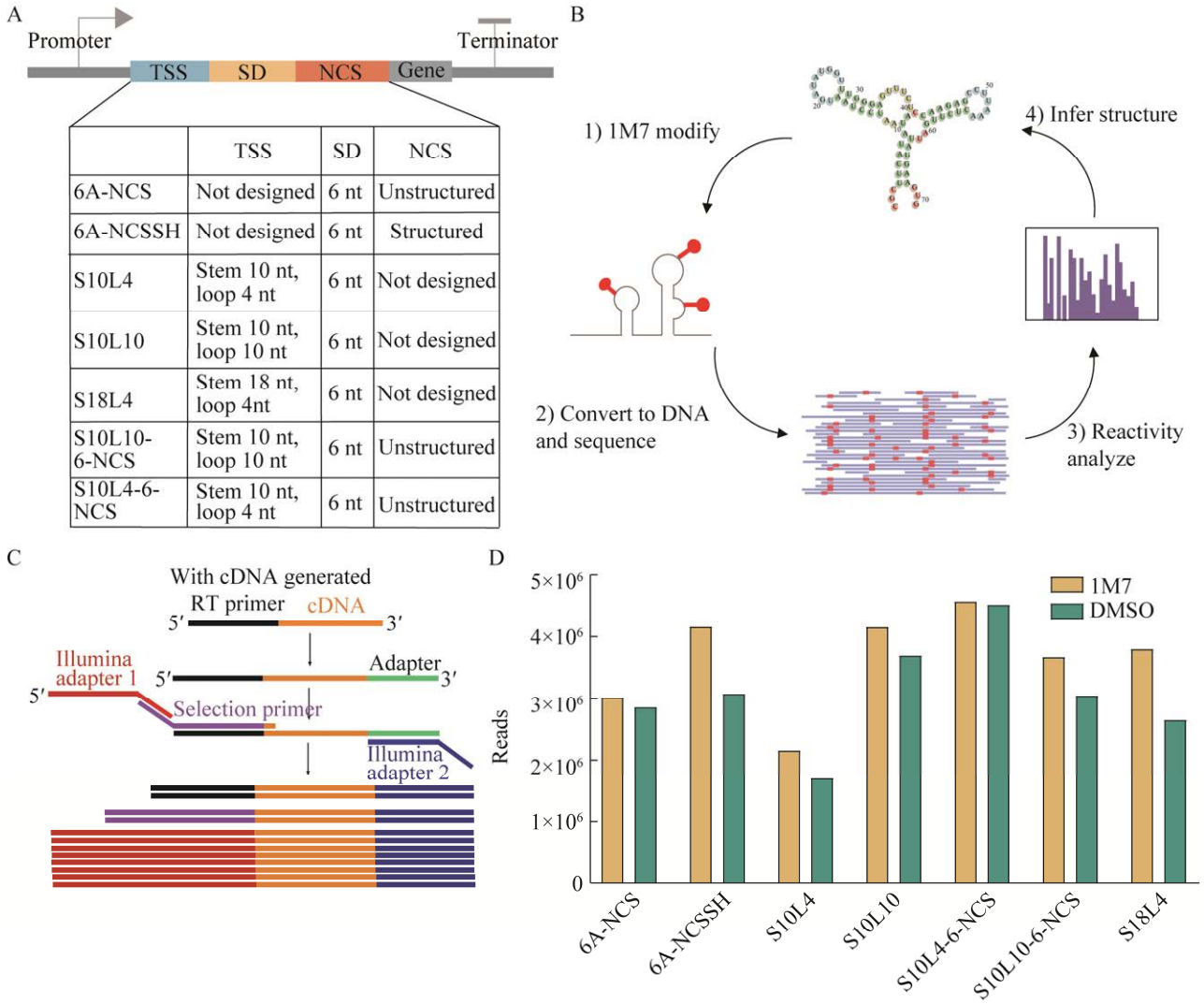
取 7 种结构不同的 5′mRNA 的大肠杆菌菌液各 1 h，然后按照 2%的接种量转接至 2 mL 的 LB 培养基中(添加 1%的抗生素)，于 37 ℃、220 r/min 培养 1 h 后添加 0.5 mmol/L IPTG 进行诱导，继续培养 2 h。取 1 mL 菌液进行 RNA 提取。对提取的 RNA 进行逆转录。将得到的 cDNA 使用引物 QGFP-F/R 以及 16S RNA-F/R 两对引物检测 *egfp* 基因的 mRNA 水平。

2 结果与分析

2.1 包含多功能区 5′mRNA 设计以及 SHAPE-seq 测序数据质量评估

5′mRNA 对基因表达具有显著影响，而这一功能是由其 5′mRNA 结构所决定的。因此，通过解析 mRNA 的构效关系并理性改造其自身结构，可以有效地实现基因表达的人工调控。为探究各区域结构对基因表达的影响，根据已报道的原核 5′mRNA 序列特点，结合 NUPACK[25]设计了 7 种典型的结构不同的 5′mRNA，具体设计以及预测结构如图 1A 所示。本研究利用 SHAPE-seq 重点解析人工设计的包含 TSS、SD 序列和 NCS 区域的 5′mRNA 在胞内的结构。如图 1B 所示，SHAPE-seq 利用 1-甲基-7-硝基靛红酸酐(1-methyl-7-nitroisatin anhydride, 1M7)特异性地与未发生碱基互补配对的碱基结合，导致逆转录时产生一个截断的 cDNA 文库。通过高通量测序读取逆转录得到的 cDNA 文库，统计每个碱基与化学探针的“反应性”，可以获得核苷酸未配对的概率。其中，连续多个高反应性的核苷酸则表示该区域处于单链状态的可能性越大。相反，若连续多个碱基的反应性较低，则该区域的碱基发生碱基互补配对，形成二级结构的概率越高。

在得到 cDNA 文库之后，通过两轮 PCR 的方法在 cDNA 3′端添加接头(图 1C)，同时进行两轮扩增增加 cDNA 文库的长度。对 1M7 处理组的 cDNA 文库进行高通量测序的结果进行初步统计，以同 mRNA 未经 1M7 处理的样品(DMSO 组)为对照，评估测序数据质量(图 1D)。结果发现，各组测序数据中，90%的测序数据都符合 Q 30 的标准，读数最少为 1 696 131 条，并且片段长度存在显著差异，实验组(1M7 组)较对照组(DMSO 组)的片段平均长度短 7−9 nt。整体质量表明文库符合数据质量要求，获得了具有高度多态性的 cDNA 文库，且 1M7 有效插入了 mRNA，截短了 cDNA 片段，满足后续分析各碱基的反应性需求。



	TSS	SD	NCS
6A-NCS	Not designed	6 nt	Unstructured
6A-NCSSH	Not designed	6 nt	Structured
S10L4	Stem 10 nt, loop 4 nt	6 nt	Not designed
S10L10	Stem 10 nt, loop 10 nt	6 nt	Not designed
S18L4	Stem 18 nt, loop 4nt	6 nt	Not designed
S10L10-6-NCS	Stem 10 nt, loop 10 nt	6 nt	Unstructured
S10L4-6-NCS	Stem 10 nt, loop 4 nt	6 nt	Unstructured

图 1　包含 3 个功能区域的 5′mRNA 结构和序列特征以及对这些结构进行 SHAPE-seq 分析流程

Figure 1　Structure and sequence characteristics of 5′mRNA containing three functional regions and SHAPE-seq analysis of these structures. A: Specific structural features of 7 typical 5′mRNA. B: Complete schematic diagram of the SHAPE-seq process. C: Procedure for constructing cDNA library through primer extension. D: Reads of 5′mRNA screened that are suitable for subsequent analysis.

2.2　非结构化的 NCS 有利于目的基因表达

为了研究 N 端编码序列(NCS)结构对目的基因表达的影响，固定了 TSS 序列和 SD 序列，设计了 2 种结构不同的 NCS，2 种 NCS 根据结构预测存在一定差异，具体如下：(1) 6A-NCSSH，NCS 区域形成茎长为 8 nt、环为 6 nt 的二级结构并且其余区域处于单链状态，自由能为−22.90 kcal/mol；(2) 6A-NCS，NCS 区域完全处于单链状态，未发生碱基互补配对，自由能为−13.40 kcal/mol。

按照公式(1)和公式(2)将 SHAPE-seq 的 θ_i 值转换成每一个核苷酸的反应性 ρ_i 值。结果如图 2 所示，6A-NCS 整体的核苷酸反应性明显高于 6A-NCSSH，说明 6A-NCS 较 6A-NCSSH 二

级结构更加简单，链内互补配对的概率较低。

根据计算得到的每个碱基的反应性解析6A-NCSSH 的结构(图 2A)，6A-NCSSH 前 3 个碱基发生了碱基互补配对，后 3 个核苷酸处于单链状态。自第 6 个核苷酸之后，ρ 值开始下降，平均仅为 0.3，存在碱基互补配对现象。8−42 nt 区域内，核苷酸反应性先降后升，该区域形成了一个茎长为 6 nt、环为 21 nt 的二级结构，而 30−37 nt 区域内的 SD 序列位于该二级结构的环上。6A-NCSSH 的 NCS 区域(46−84 nt)的核苷酸反应性普遍较低，平均水平在 0.25 之下，说明该区域内碱基互补配对的概率较高，二级结构较复杂；53 nt 附近 NCS 与 TSS 发生了碱基互补配对，在 58−79 nt 内形成了茎长为 8 nt、环为 6 nt 的二级结构。对这一设计的 SHAPE-seq 分析结果表明该二级结构与预测的结构较一致。

6A-NCS 的 1−42 nt 区域(TSS 以及 SD 序列)内核苷酸反应性大多数都在 2 以上(图 2B)，结

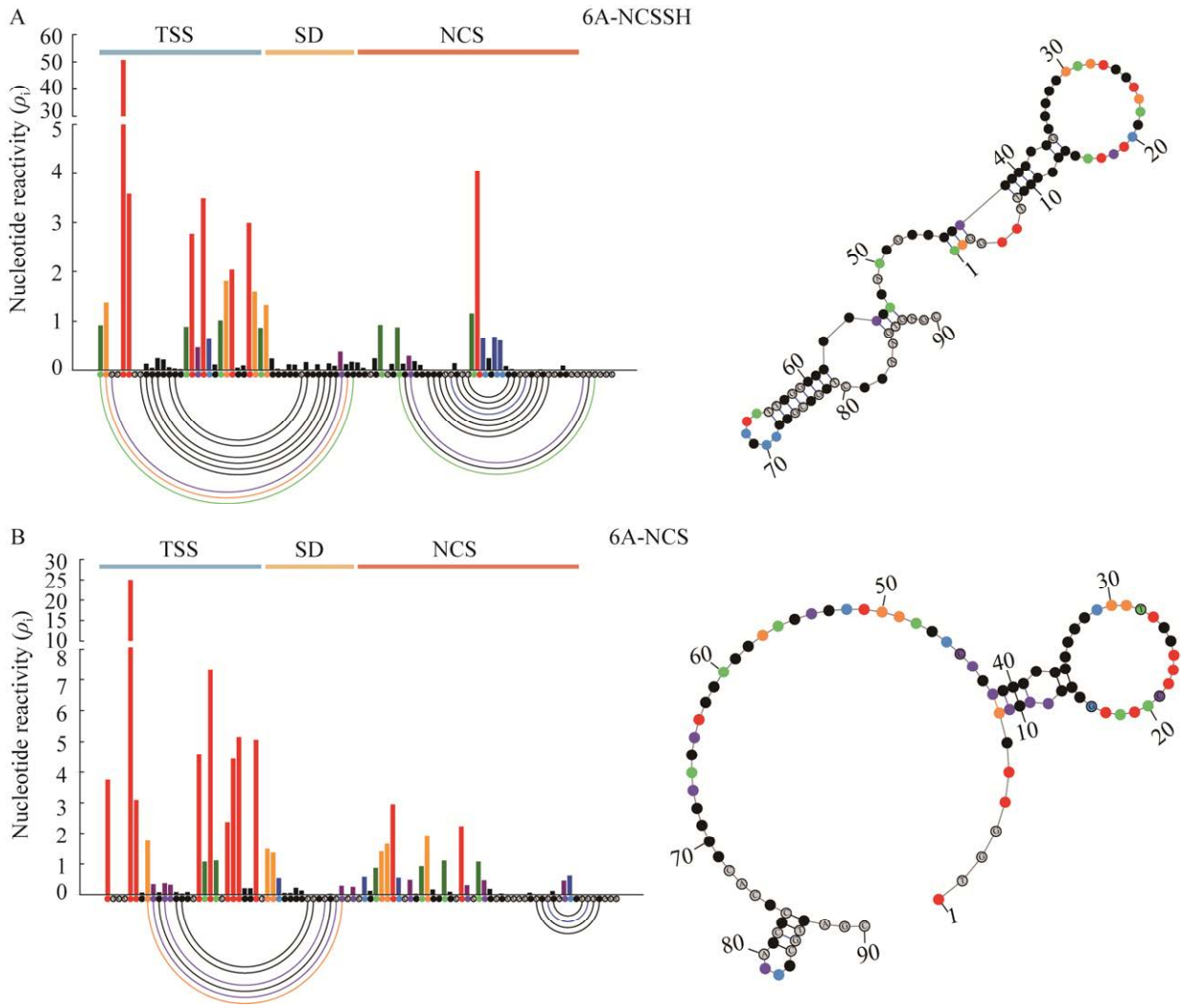


图 2　不同 NCS 序列的各碱基核苷酸反应性分析以及根据各碱基反应性进行的结构模拟分析

Figure 2　Nucleotide reactivity analysis and structure simulation of different NCS sequences based on the results of nucleotide reactivity. A: Nucleotide reactivity and structural simulation of NCS forming secondary structure. B: Nucleotide reactivity and structure simulation of unstructured NCS.

构模拟表明该区域形成了茎长为 8 nt、环为 21 nt 的二级结构，mRNA 起始区域未发生碱基互补配对，SD 序列暴露在二级结构的环上。其 NCS 区域(46−84 nt)的核苷酸反应性都在 0.5 以上，该区域绝大多数碱基未发生碱基互补配对，仅在 76−88 nt 形成了茎长为 4 nt、环为 4 nt 的二级结构。结果表明设计的 NCS 以非结构化(线性化)的形式稳定存在于胞内。

对 6A-NCS 以及 6A-NCSSH 调控的 GFP 的荧光强度以及 mRNA 水平进行定量分析，结果如图 3 所示，较未插入人工设计的 NCS 的组别(WT)相比，结构化的 NCS (6A-NCSSH)和非结构化的 NCS (6A-NCS)调控下，*egfp* 基因的 mRNA 丰度分别提高了 2.8 倍和 10 倍，而荧光强度分别提高了 10 倍和 19 倍。研究结果显示，虽然结构化和非结构化的 NCS 均能显著提高基因表达，但非结构化的 NCS 对 mRNA 丰度和蛋白量都具有更加显著的提高效果。Espah Borujeni 等[5]研究表明，非结构化的 NCS 有利于核糖体在翻译初期的沿 mRNA 链的顺利滑动，而这些覆盖于 mRNA 上的核糖体发挥了保护 mRNA 抵抗水解的作用，因此表现为对 mRNA 丰度的显著提高。结果表明，避免 NCS 区域形成复杂的二级结构，有利于提高目标基因的转录水平以及翻译水平。这与 Allert 等[26]研究结果一致，他们通过对原核生物基因组序列进行分析发现，通常高蛋白产量的基因的 N 端通常具有高 AU 含量以及较高的自由能(较少的二级结构)。

2.3 结构化的 TSS 区域提高 mRNA 丰度

为了研究 TSS 结构对基因表达的影响，设计了 3 种包含相同 SD 序列，但二级结构茎长与环大小不同的 TSS，具体如下，(1) S18L4：茎长为 18 nt、环为 4 nt 的二级结构，自由能为−43.30 kcal/mol；(2) S10L10：茎长为 10 nt、环为 10 nt，自由能为−27.60 kcal/mol；(3) S10L4：茎长为 10 nt、环为 4 nt 的二级结构，自由能为−29.60 kcal/mol。

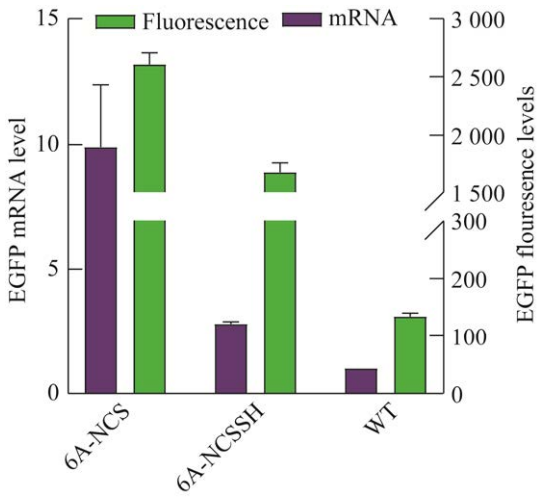


图 3 不同结构的 NCS 调控下 mRNA 丰度和蛋白表达量分析

Figure 3 Analysis of mRNA abundance and protein expression under the regulation of NCS with different structures.

对 3 种 TSS 序列的核苷酸反应性分析发现，其 mRNA 起始区域内核苷酸反应性都较高(图 4)，但是随着 mRNA 的延伸，S10L4 的核苷酸反应性逐渐降低，趋近于 0。虽然 S10L4 自由能不是 3 种设计中最低的，但其链内碱基互补配对概率最高，有多个区域发生了碱基互补配对，二级结构最复杂，结构化程度最高。如图 4C 所示，主要的几个结构化区域为：33−57 nt 区域内形成了茎长为 10 nt、环为 4 nt 的高稳定性的二级结构；在 10−30 nt 形成了茎长为 6 nt、环为 4 nt 的稳定性略差的二级结构；65 nt 附近的 SD 序列被包裹形成二级结构。

S18L4 在 1−30 nt 区域的核苷酸反应性较高(图 4A)，平均在 10，说明该区域大部分碱基都未发生碱基配对，仅有 3 个碱基发生了配对。在 30−75 nt 区域内，存在一个高稳定性局部的二级结构，表现为反应性均低于 0.5，形成了茎长为 18 nt、环为 4 nt 的二级结构，其 SD 序列(80−96 nt)具有较低反应性，形成了较稳定的碱基互补配对。

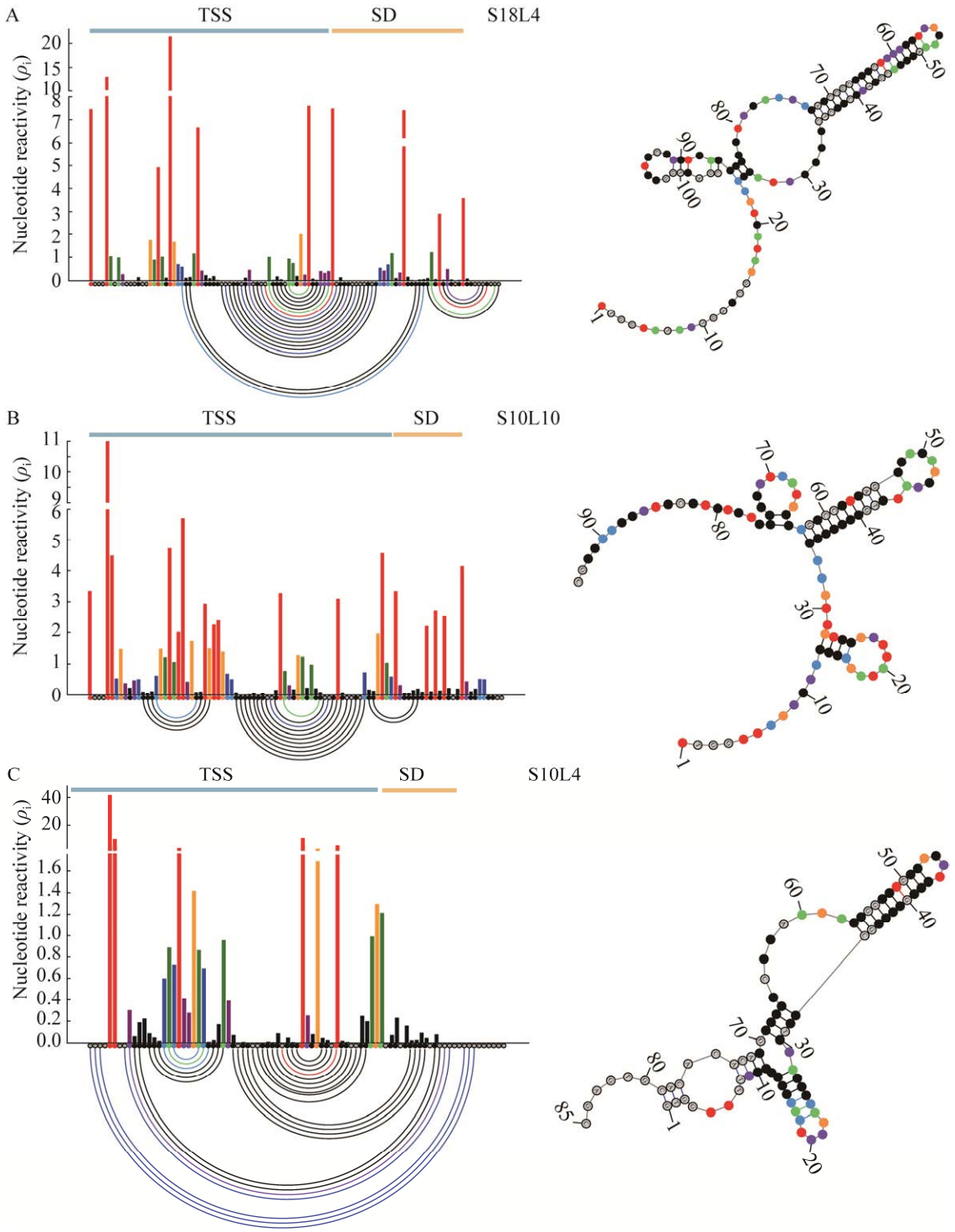


图 4　不同 TSS 序列各碱基的核苷酸反应性分析以及根据各反应性的结构模拟

Figure 4　Nucleotide reactivity analysis and structure simulation of each base of different TSS sequences based on the result of the nucleotide reactivity. A: Nucleotide reactivity and structure simulation of TSS with stem length of 18 nt and ring length of 4 nt. B: Nucleotide reactivity and structure simulation of TSS with stem length of 10 nt and ring length of 10 nt. C: Nucleotide reactivity and structure simulation of TSS with stem length of 10 nt and ring length of 4 nt.

对 S10L10 的核苷酸反应性分析发现(图 4B)，在其 TSS 前 30 nt 区域内反应性较高，平均在 3，在 13−29 nt 形成了茎长 4 nt、环为 8 nt 的二级结构，但由于核苷酸反应性较高，所以该二级结构稳定性较低，形成概率较小。在 34−63 nt 区域内，形成了茎长为 11 nt、环为 7 nt 的二级结构，该二级结构与预测结果并不完全符合，而且二级结构的茎中包含未发生配对的单独碱基。这一设计中 SD 序列所在区域未发生碱基互补配对。

以上二级结构结果表明，mRNA 在细胞内的复杂环境状况下形成的二级结构更加复杂，存在自由能无法预测到的小范围结构。

对 S18L6、S10L10 以及 S10L4 调控下的荧光强度与 mRNA 丰度进行定量分析，结果如图 5 所示，S10L10 与 SD 序列(WT)相比，在其调控下 *egfp* 基因的 mRNA 丰度提高了 1.8 倍，荧光强度提高 2.9 倍。SHAPE-seq 结果表明，设计形成了茎长 11 nt、环 7 nt 的二级结构，这一结构有利于基因表达，但具有茎长为 18 nt 的 S18L6 调控下的 *egfp* 基因的 mRNA 丰度反而不利于提高 *egfp* 基因的 mRNA 丰度，荧光强度虽有提高，但提高幅度不大(46%)。说明进一步增加茎长，不利于基因表达，尤其会使 mRNA 丰度降低。Zhang 等研究报道，当胞内二级结构茎长过长时，会更加容易遭到 RNase III 的靶向降解[6]，当茎长超过 16 nt 时，RNase III 对这一结构的靶向切割效率可以超过 80%[27]。这一机制可以解释 mRNA 丰度的显著降低，另外，需要注意的是，SHAPE-seq 结果表明其 SD 序列发生了碱基互补配对，有可能被二级结构裹挟其中，核糖体需要耗费更大的能量才能与其结合，这也有可能进一步造成了翻译起始效率低的问题[28]。两方面共同造成了这一设计调控下基因表达效果较差。

SHAPE-seq 结果表明，S10L4 具有更加复杂的结构，其调控的 *egfp* 基因的 mRNA 丰度和荧光强度相较于无 TSS 组(WT)提高了 3.7 倍和 2.6 倍。S10L4 调控的 mRNA 丰度比 S10L10 调控的提高了 68%，但荧光强度与 S10L10 相比有一定降低。一方面原因可能是 S10L4 TSS 区域的二级结构更加复杂，可以使 mRNA 更加稳定，mRNA 抵抗水解能力更强；另一方面，由于其 SD 序列与其他区域发生了碱基互补配对，从而导致翻译起始效率大大降低，使蛋白表达水平与 mRNA 丰度提高不成正比[29]。

2.4 优化组合各功能区域的 5′mRNA 促进目的基因表达

为了分析 TSS 和 NCS 之间是否存在二级结构的交互影响，以及是否会对基因表达产生串扰效应，选择了两个较优的 TSS (S10L10 与 S10L4)的序列分别与上述优化得到的非结构化的 NCS 进行连接，解析其构效关系，经预测 S10L4-6-NCS 的自由能为 −28.00 kcal/mol，S10L10-6-NCS 的自由能为−29.90 kcal/mol。经过核苷酸反应性分析发现(图 6)，在 TSS 前 30 nt 区域内，S10L4-6-NCS 的大多数核苷酸反应性都在 1 以下，而 S10L10-6-NCS 的核苷酸反应性平

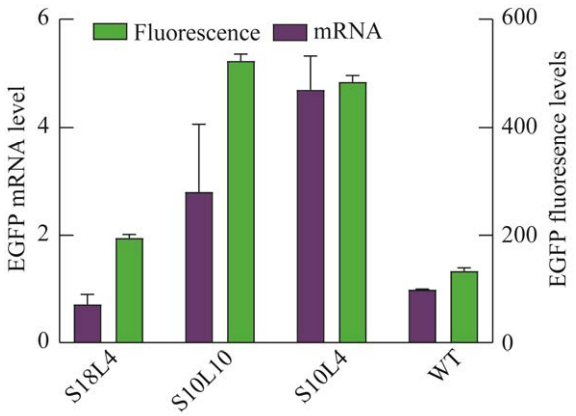


图 5 不同结构的 TSS 调控下 mRNA 丰度和蛋白表达量分析

Figure 5 Analysis of mRNA abundance and protein expression under TSS regulation of different structures.

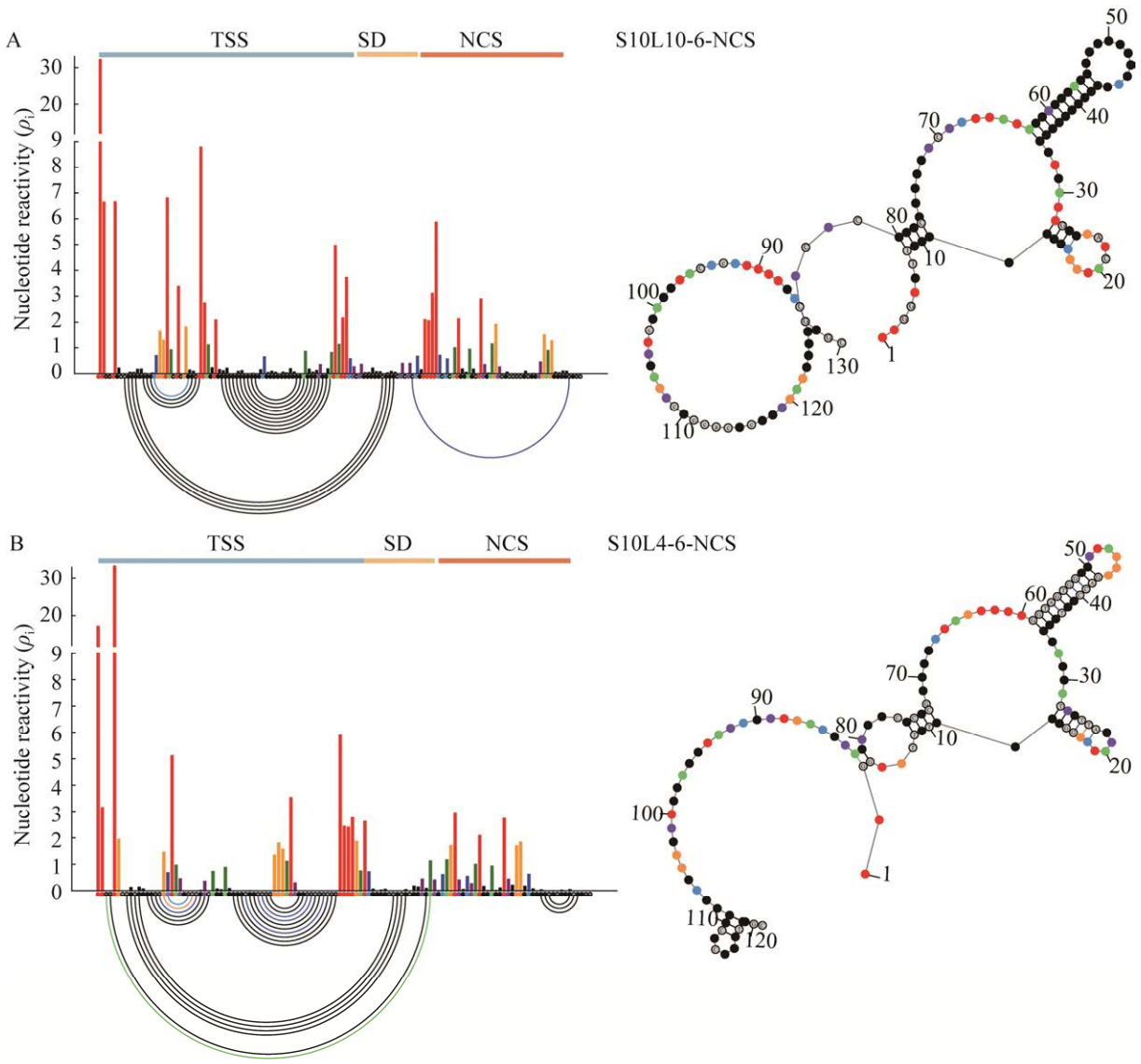


图 6　较优的 TSS 与 NCS 组合的 5′mRNA 序列核苷酸反应性分析以及根据各碱基反应性的结构模拟分析

Figure 6　Nucleotide reactivity analysis and structural simulation of the combination of optimal TSS and NCS based on the nucleotide reactivity. A: Nucleotide reactivity and structural simulation of the combination of a 10 nt stem, a 10 nt loop TSS, and an unstructured NCS. B: Nucleotide reactivity and structural simulation of the combination of a 10 nt stem, a 4 nt loop TSS, and an unstructured NCS.

均介于 1−2，说明在起始区域，S10L4-6-NCS 的二级结构仍然较为复杂，经过结构模拟发现，在 10 nt 附近，S10L4-6-NCS 的 TSS 区域与 SD 序列以及起始密码子发生了碱基互补配对，而 S10L10-6-NCS 的 TSS 仅与 SD 序列末端发生了碱基互补配对。在 10−70 nt 区域内，S10L10-6-NCS 形成了与预测结果一致的茎长为 10 nt、环为 10 nt 的二级结构，并且在 10−30 nt，

形成了茎长为 4 nt、环为 9 nt 的稳定性较差、形成概率较低的二级结构，而 S10L4-6-NCS 在 30-60 nt 区域内形成了茎长为 10 nt、环为 4 nt 的二级结构，同时在 10−30 nt 内形成了稳定性较高、形成概率较高的茎长为 6 nt、环为 4 nt 的二级结构。在 80−130 nt 区域内，两种设计的 NCS 大部分均处于单链状态，但 S10L4-6-NCS 在 110−120 nt 附近，形成了茎长为 4 nt、环为 4 nt 的二级结构。上述结果说明，不同区域之间存在明显的串扰效应，因此非常有必要考虑基因背景的因素，尽量大范围研究 mRNA 水平各元件或功能区的调控机制和效果。

对 S10L10-6-NCS、S10L4-6-NCS 调控下的荧光强度与 mRNA 丰度进行测定，结果如图 7 所示，相比于对照 WT 组，S10L4-6-NCS 和 S10L10-6-NCS 调控下 mRNA 丰度分别提高了 50 倍和 36 倍，而它们调控下荧光强度分别提高了 38 倍和 58 倍。两种组合都对基因表达有显著的促进作用，但 S10L4-6-NCS 对 mRNA 丰度的提高更明显，而 S10L10-6-NCS 对蛋白量提高更明显。结合 SHAPE-seq 结果，认为 S10L4-6-NCS 的 NCS 区域在一定概率上，存在一个局部二级结构，并且其 SD 序列与 TSS 也发生了碱基互补配对。这两个区域的二级结构可能在一定程度上阻碍了核糖体与 mRNA 结合，从而导致翻译延伸速率受到阻滞[5]，但是这些二级结构非常有利于 mRNA 抵抗水解，提高 mRNA 稳定性，从而提高 mRNA 丰度。

综上所述，上述两个较优 TSS 和 NCS 的组合显著提高了基因表达，虽然在提高 mRNA 和蛋白量方面存在具体量比上的差异，但是通过结构分析和表达活性的分析，综合认为具有较优的表达的作用。同时，条件允许的情况下，尽量避免 SD 序列与 NCS 形成二级结构是 5′mRNA 改造中应遵循的重要依据[30]。

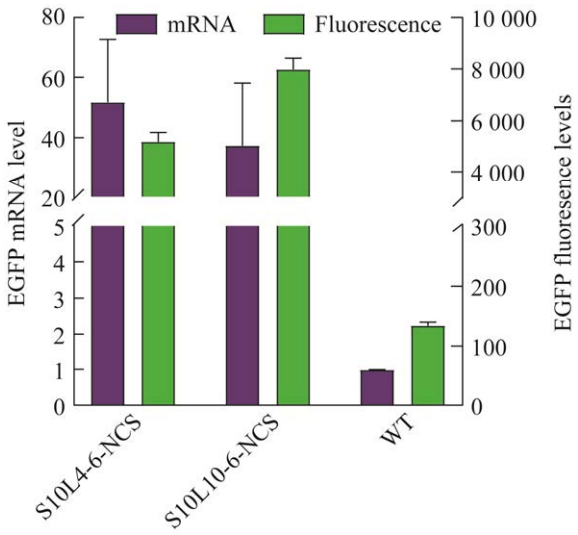


图 7　不同 TSS 和 NCS 组合调控下 mRNA 丰度和蛋白表达量分析

Figure 7　Analysis of mRNA abundance and protein expression under different combinations of TSS and NCS.

3　讨论与结论

原核生物的基因表达受制于 mRNA 的半衰期，而常见的基因表达调控元件不能提高 mRNA 抗水解的能力，目前许多研究发现 mRNA 本身存在多个功能区域可以影响其抗水解的能力以及翻译效率，Zhang 等[6]将 TSS 设计形成二级结构从而开发了一种调节 RNA 降解的 RNA 模块库(dtRNAs)，将其用于增加体内或体外的转录稳定性，而不影响翻译起始效率，因此 TSS 的二级结构可以提高其抗水解能力。Tian 等[30]通过对 96 个合理选择的 NCS 进行系统分析发现，这些 NCS 的调控水平表现出 4 个数量级差异，然后将人工合成的以及天然的 NCS 用于 N-乙酰神经氨酸(NeuAc)生物合成途径中，发现 NeuAc 产量与野生型相比提高了 3.21 倍，而 Espah borujeni 等[5]通过核糖体印迹发现，当 NCS 形成二级结构时会导致其调控的 mRNA 丰度下降，以及核糖体与 mRNA 的结合力降低，因此推测

NCS 形成二级结构不利于目的基因的表达。Sun 等[31]通过使用中等强度且富含 A、U 碱基的 RBS 来调控 3-酮类固醇-9α-羟化酶(KshA、KshB)相关基因的过表达，9α-羟基雄甾-4-烯-3,17-二酮(9-OHAD)的产量得到显著增加。上述研究表明了各功能区可以通过改变自身的特征来改变其调控水平，但多数设计均利用最小自由能进行模拟，不能代表其在胞内的真实结构，并且未探讨多个区域之间是否会形成跨区域的二级结构，因此导致各功能区组合后的调控效果不明确等问题。

本研究采用了 SHAPE-seq 技术，该技术将化学探针与高通量测序技术相结合，通过对典型的 5′mRNA 区域进行结构解析，结合 mRNA 丰度和蛋白表达量的分析，研究了 TSS、SD 序列、NCS 各自以及串扰效应特征对基因表达的影响，三者之间是否会存在结构串扰关系。结果表明具有较高自由能、非结构化的 NCS 核苷酸反应性较高，形成二级结构的概率较小，更加有利于提高目的基因的 mRNA 丰度以及蛋白表达水平，这有可能是当 NCS 形成二级结构时会导致核糖体与 mRNA 的结合能力减弱，不利于核糖体沿 mRNA 的滑动，从而导致其翻译效率降低[5]。改变 TSS 序列可以有效控制该区域形成的结构特征，从而对转录后水平产生 2−3 倍的变化，这一区域形成结构越复杂越有利于提高 mRNA 丰度，然而过长的茎环结构会使 mRNA 更容易降解(mRNA 丰度下降 30%)，可能是由于这种过长的茎会更加容易被核酸降解酶(RNase III)靶向识别[6,32]；另外，SD 序列被二级结构裹挟其中时会影响其介导的翻译起始(蛋白表达下降 10%)。研究中也发现跨区域之间存在形成二级结构的情况，当 SD 序列或者 NCS 被 TSS 包裹形成二级结构时，会导致翻译效率的降低，这是不同区域调控效果存在串扰的重要原因之一。组合二级结构茎长适中的 TSS 以及非结构化的 NCS，其调控下 mRNA 丰度和蛋白量显著提高(11 倍、60 倍)。本研究获得的调控性能较优的 5′mRNA 序列为工业微生物目标基因表达提供了调控元件，研究解析的各区域的构效关系为人工基因回路的构建提供了设计依据。同时本研究采用 SHAPE-seq 技术检测的 mRNA 结构和传统的自由能预测的二级结构结果存在差异，为建立构效关系提供了更精准的信息，是 RNA 二级结构分析的有效方法。然而，需要指出的是，本研究虽然突破了领域内采用基于自由能预测的方法分析二级结构，在一定程度上为解决了 mRNA 结构信息匮乏的问题提供了先例，但是这一方法通量较低，分析工具较少，对实验操作和数据分析要求较高，后续可以开发在测序环节添加样品 barcode，再结合 SHAPE-seq 高通量测序，实现快速获得大规模序列的二级结构解析的目标，更加全面地采集结构信息，建立更加完善的 5′mRNA 构效关系。

参考文献

[1] KANDAVALLI VK, TRAN H, RIBEIRO AS. Effects of σ factor competition are promoter initiation kinetics dependent[J]. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 2016, 1859(10): 1281-1288.

[2] LI TT, LI T, JI WY, WANG QY, ZHANG HQ, CHEN GQ, LOU CB, OUYANG Q. Engineering of core promoter regions enables the construction of constitutive and inducible promoters in *Halomonas* sp.[J]. Biotechnology Journal, 2016, 11(2): 219-227.

[3] CETNAR DP, SALIS HM. Systematic quantification of sequence and structural determinants controlling mRNA stability in bacterial operons[J]. ACS Synthetic Biology, 2021, 10(2): 318-332.

[4] NOUAILLE S, MONDEIL S, FINOUX AL, MOULIS C, GIRBAL L, COCAIGN-BOUSQUET M. The stability of an mRNA is influenced by its concentration: a potential physical mechanism to regulate gene expression[J]. Nucleic Acids Research, 2017, 45(20): 11711-11724.

[5] ESPAH BORUJENI A, CETNAR D, FARASAT I, SMITH A, LUNDGREN N, SALIS HM. Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in

N-terminal coding sequences[J]. Nucleic Acids Research, 2017, 45(9): 5437-5448.
[6] ZHANG Q, MA D, WU FQ, STANDAGE-BEIER K, CHEN XW, WU KY, GREEN AA, WANG X. Predictable control of RNA lifetime using engineered degradation-tuning RNAs[J]. Nature Chemical Biology, 2021, 17(7): 828-836.
[7] de SMIT MH, van DUIN J. Translational standby sites: how ribosomes may deal with the rapid folding kinetics of mRNA[J]. Journal of Molecular Biology, 2003, 331(4): 737-743.
[8] RAMAKRISHNAN V. Ribosome structure and the mechanism of translation[J]. Cell, 2002, 108(4): 557-572.
[9] XU KD, TONG Y, LI Y, TAO J, LI JH, ZHOU JW, LIU S. Rational design of the N-terminal coding sequence for regulating enzyme expression in *Bacillus subtilis*[J]. ACS Synthetic Biology, 2021, 10(2): 265-276.
[10] KUDLA G, MURRAY AW, TOLLERVEY D, PLOTKIN JB. Coding-sequence determinants of gene expression in *Escherichia coli*[J]. Science, 2009, 324(5924): 255-258.
[11] VIEGAS SC, APURA P, MARTÍNEZ-GARCÍA E, de LORENZO V, ARRAIANO CM. Modulating heterologous gene expression with portable mRNA-stabilizing 5′-UTR sequences[J]. ACS Synthetic Biology, 2018, 7(9): 2177-2188.
[12] SHARP JS, BECHHOFER DH. Effect of 5′-proximal elements on decay of a model mRNA in *Bacillus subtilis*[J]. Molecular Microbiology, 2005, 57(2): 484-495.
[13] STEITZ JA. Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA[J]. Nature, 1969, 224(5223): 957-964.
[14] CHEN H, BJERKNES M, KUMAR R, JAY E. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs[J]. Nucleic Acids Research, 1994, 22(23): 4953-4957.
[15] TULLER T, ZUR H. Multiple roles of the coding sequence 5′ end in gene expression regulation[J]. Nucleic Acids Research, 2015, 43(1): 13-28.
[16] DETHOFF EA, CHUGH J, MUSTOE AM, AL-HASHIMI HM. Functional complexity and regulation through RNA dynamics[J]. Nature, 2012, 482(7385): 322-330.
[17] SEETIN MG, MATHEWS DH. RNA structure prediction: an overview of methods[J]. Methods in Molecular Biology, 2012, 905: 99-122.
[18] REUTER JS, MATHEWS DH. RNAstructure: software for RNA secondary structure prediction and analysis[J]. BMC Bioinformatics, 2010, 11: 129.
[19] MORTIMER SA, TRAPNELL C, AVIRAN S, PACHTER L, LUCKS JB. SHAPE-seq: high-throughput RNA structure analysis[J]. Current Protocols in Chemical Biology, 2012, 4(4): 275-297.
[20] WATTERS KE, ABBOTT TR, LUCKS JB. Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq[J]. Nucleic Acids Research, 2016, 44(2): e12.
[21] PENNETTI VJ, LAFAYETTE PR, PARROTT WA. MultiGreen: a multiplexing architecture for GreenGate cloning[J]. PLoS One, 2024, 19(9): e0306008.
[22] MARTIN M. Cutadapt removes adapter sequences from high-throughput sequencing reads[J]. EMBnet Journal, 2011, 17(1): 10.
[23] SHEN W, LE S, LI Y, HU FQ. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation[J]. PLoS One, 2016, 11(10): e0163962.
[24] MUSTOE AM, BUSAN S, RICE GM, HAJDIN CE, PETERSON BK, RUDA VM, KUBICA N, NUTIU R, BARYZA JL, WEEKS KM. Pervasive regulatory functions of mRNA structure revealed by high-resolution SHAPE probing[J]. Cell, 2018, 173(1): 181-195.e18.
[25] ZADEH JN, STEENBERG CD, BOIS JS, WOLFE BR, PIERCE MB, KHAN AR, DIRKS RM, PIERCE NA. NUPACK: Analysis and design of nucleic acid systems[J]. Journal of Computational Chemistry, 2011, 32(1): 170-173.
[26] ALLERT M, COX JC, HELLINGA HW. Multifactorial determinants of protein expression in prokaryotic open reading frames[J]. Journal of Molecular Biology, 2010, 402(5): 905-918.
[27] COURT DL, GAN JH, LIANG YH, SHAW GX, TROPEA JE, COSTANTINO N, WAUGH DS, JI XH. RNase III: Genetics and function; structure and mechanism[J]. Annual Review of Genetics, 2013, 47: 405-431.
[28] 马荣, 尚方正, 潘剑锋, 戎友俊, 王敏, 李金泉, 张燕军. 细胞内 mRNA 翻译影响因素及翻译组学的研究进展[J]. 生物技术通报, 2022, 38(12): 115-126.
MA R, SHANG FZ, PAN JF, RONG YJ, WANG M, LI JQ, ZHANG YJ. Research progress in influencing factors of mRNA translation in cells and translatome[J]. Biotechnology Bulletin, 2022, 38(12): 115-126 (in Chinese).
[29] STERK M, ROMILLY C. Unstructured 5′-tails act through ribosome standby to override inhibitory structure at ribosome binding sites[J]. Nucleic Acids Research, 2018, 46(8): 4188-4199.
[30] TIAN RZ, LIU YF, CHEN JR, LI JH, LIU L, DU GC, CHEN J. Synthetic N-terminal coding sequences for fine-tuning gene expression and metabolic engineering in *Bacillus subtilis*[J]. Metabolic Engineering, 2019, 55: 131-141.
[31] SUN H, YANG JL, SONG H. Engineering mycobacteria artificial promoters and ribosomal binding sites for enhanced sterol production[J]. Biochemical Engineering Journal, 2020, 162: 107739.
[32] LEJARS M, KOBAYASHI A, HAJNSDORF E. RNase III, ribosome biogenesis and beyond[J]. Microorganisms, 2021, 9(12): 2608.