天津科技大学学报
Journal of Tianjin University of Science & Technology
ISSN 1672-6510,CN 12-1355/N

《天津科技大学学报》网络首发论文

题目：基于自然语言处理的枯草芽孢杆菌启动子强度预测
作者：陈聪葛，郭怡雪，卞亚蕊，刘夫锋，路福平，彭冲
DOI：10.13364/j.issn.1672-6510.20240146
收稿日期：2024-07-16
网络首发日期：2025-01-14
引用格式：陈聪葛，郭怡雪，卞亚蕊，刘夫锋，路福平，彭冲．基于自然语言处理的枯草芽孢杆菌启动子强度预测[J/OL]．天津科技大学学报.
https://doi.org/10.13364/j.issn.1672-6510.20240146

DOI：10.13364/j.issn.1672-6510.20240146

基于自然语言处理的枯草芽孢杆菌启动子强度预测

陈聪葛[1]，郭怡雪[1]，卞亚蕊[1]，刘夫锋[1,2,3]，路福平[1,2,3]，彭　冲[1,2,3]
（1. 天津科技大学生物工程学院，天津 300457；2. 工业发酵微生物教育部重点实验室，天津 300457；
3. 天津市工业微生物重点实验室，天津 300457）

摘　要：作为在转录水平上调节基因表达的关键元件，启动子的强度直接调控基因的表达水平。现有启动子强度预测模型多集中于大肠杆菌，针对其他物种启动子强度预测的模型则相对较少。本研究主要以枯草芽孢杆菌启动子为研究对象，收集多组枯草芽孢杆菌启动子序列强度数据，通过绿色荧光蛋白表达强度作为统一表征指标，对多组启动子强度进行标准化计算，构建枯草芽孢杆菌启动子强度数据集。分别使用 7 种自然语言处理方法，包括 FastText、TextCNN、TextRNN、TextRCNN、TextRNN_Att、DPCNN 和 Transformer，构建启动子强度预测模型。结果表明，Transformer 模型在启动子强度预测任务中取得最好的预测效果，准确率可达 79.49%。本研究自主构建了枯草芽孢杆菌启动子强度数据集，并使用自然语言处理的方法训练得到效果较好的启动子强度预测模型，研究结果可以为枯草芽孢杆菌特定强度启动子的筛选提供依据。

关键词：枯草芽孢杆菌；启动子；启动子强度；自然语言处理
中图分类号：Q812;TP399　　**文献标志码：**A　　**文章编号：**1672-6510 (0000)00-0000-00

Prediction of Promoter Strength in *Bacillus subtilis* Based on Natural Language Processing

CHEN Congge[1], GUO Yixue[1], BIAN Yarui[1], LIU Fufeng[1,2,3], LU Fuping[1,2,3], PENG Chong[1,2,3*]
(1. College of Biotechnology, Tianjin University of Science and Technology, Tianjin 300457, China;
2. Key Laboratory of Industrial Fermentation Microbiology, Ministry of Education, Tianjin 300457, China;
3. Tianjin Key Laboratory of Industrial Microbiology, Tianjin 300457, China)

Abstract: As a key element regulating gene expression at the transcriptional level, promoter strength directly regulates gene expression levels. However, current promoter strength prediction models are mostly focused on *Escherichia coli*, and models for predicting promoter strengths in other species are not yet well developed. This study focuses on *Bacillus subtilis* promoters, amassing a dataset that correlates *Bacillus subtilis* promoter sequences with their respective strengths. By employing green fluorescent protein for normalized characterization of promoter strength, a comprehensive dataset for *Bacillus subtilis* promoter intensities was established. Seven natural language processing methods, including FastText, TextCNN, TextRNN, TextRCNN, TextRNN_ Att, DPCNN, and Transformer were used to train the promoter strength prediction model. The results showed that Transformer model achieved the best prediction effect in the task of promoter strength, with an accuracy of 79.49%. This research pioneers the construction of a *Bacillus subtilis* promoter strength dataset and successfully trains a machine learning model, grounded in natural language processing, that demonstrates remarkable performance in forecasting promoter strength, thereby advancing our ability to computationally decipher regulatory elements in genetic sequences.

Key words: *Bacillus subtilis*; promoter; promoter strength; natural language processing

收稿日期：2024-07-16；**修回日期：**2024-11-07
基金项目：国家自然科学基金项目(32001657)；国家重点研发计划项目(2021YFC2100400)
作者简介：陈聪葛（2000—），女，山西人，硕士研究生；**通信作者：**彭　冲，讲师，cpeng@tust.edu.cn

启动子是 DNA 序列中的一段特定区域，位于转录起始位点（transcription start site，TSS）附近，负责调控基因的转录过程[1-2]。启动子如同基因的激活控制器，决定了基因何时启动以及如何在正确的位置展开转录活动。在这一过程中，RNA 聚合酶扮演着核心角色，它能精准识别并绑定到启动子区域，随后促使 DNA 双螺旋在转录起始位点附近解开，形成一个开放的复合体，从而启动 RNA 分子的合成[3-4]。在原核生物中，RNA 聚合酶通常由核心酶和一个 σ 因子组成。核心酶负责催化 RNA 链从 DNA 模板上以 5′至 3′方向的延长合成，σ 因子负责指导 RNA 聚合酶正确地与 DNA 模板上的启动子区域结合，启动转录过程。不同类型的 σ 因子（如 σ70、σ54 等）可以识别不同类型的启动子，从而控制不同基因的转录[5-6]。原核启动子通常包含高度保守的序列元件，如-10 区（5′-TATAAT-3′）和-35 区（5′-TTGACA-3′）[7]，这些序列直接参与 RNA 聚合酶的识别过程。但是，尽管存在保守序列，启动子序列在不同物种间可能有所不同，甚至同一类型启动子在不同基因间也有变化[8]。作为在转录水平上调节基因表达的关键元件，启动子的活性直接调控基因表达水平。具备高活性的强启动子由于能显著提升转录效率而成为生物技术和生物工程领域内设计高效蛋白质生产体系的优选元件[9]。

随着测序技术的飞速发展以及基因组注释数据的积累，目前已有较多的启动子数据[10-11]，这使构建预测启动子的计算模型成为可能。近年来，各种机器学习算法开始融入到启动子识别的领域中，如 Fisher 线性判别[12]、逻辑回归[13]、支持向量机（SVM）[14-15]、随机森林[16]、决策树[17]、人工神经网络（ANN）[18-19]以及卷积神经网络（CNN）[16,20-21]等。由于启动子的多样性，研究人员也开发了一些启动子二层分类器。Liu 等[22]使用伪 K-tuple 核苷酸组成（PseKNC）构建了一个两层的预测器，第一层用于识别序列是否为启动子，第二层用于识别启动子类型。Xiao 等[23]以 RegulonDB[11]数据库中大肠杆菌的启动子信息构建了 iPSW(2L)-PseKNC，第一层同样是识别 DNA 序列是否为启动子，第二层用于识别启动子的强弱类型[23]。Liang 等[24]在 Xiao 等[23]构建的数据集的基础上，使用 k-mer 核苷酸组成、二进制编码和基于二核苷酸属性矩阵的距离变换进行特征提取，并使用极端随机树（extra-trees）进行特征选择，开发了一个名为 iPromoter-ET 的双层模型，识别启动子强弱的准确率可达 72.59%。Tayara 等[25]也在 Xiao 等[23]构建的数据集的基础上结合卷积神经网络（CNN）和伪二核苷酸组成（PseDNC）开发了一个名为 iPSW(PseDNC-DL)的混合模型，用于识别原核启动子及其强度。与 iPSW(2L)-PseKNC 模型相比，iPSW(PseDNC-DL) 模型在两项任务中都优于 iPSW(2L)-PseKNC 模型。在启动子强度鉴定工作中，iPSW(PseDNC-DL)模型的准确率、敏感性和马修斯相关系数（Mathew’s correlation coefficient，MCC）分别提高了 1.15%、3.58%和 2.27%。

关于启动子强度预测的工作，目前大多集中于大肠杆菌。枯草芽孢杆菌（*Bacillus subtilis*）是一种耐逆性强、非致病性的革兰氏阳性菌，具有安全性高、分泌高效等特点，有利于异源蛋白的高产表达与工业酶制剂的开发，展现出优越的工业应用潜力[26]，但是枯草芽孢杆菌启动子的数据量比较有限。现有的存储枯草芽孢杆菌启动子信息的数据库仅有 2 个，分别是原核生物启动子数据库 PPD[10]和枯草芽孢杆菌转录调控数据库 DBTBS[27]。PPD 数据库仅有 800 余组枯草芽孢杆菌启动子数据，DBTBS 数据库也只有 1400 余组枯草芽孢杆菌启动子信息，尚没有数据库能够提供枯草芽孢杆菌启动子强度的相关数据，有限的数据也限制了枯草芽孢杆菌启动子相关的研究。本研究旨在通过自然语言处理技术预测枯草芽孢杆菌的启动子强度。为此，检索了 9 篇通过荧光强度表征启动子强度的文献，在文献中共收集了 12 组启动子强度信息，在每组中挑选一个启动子序列，以绿色荧光蛋白的表达强度为标准，对这 12 个启动子的强度进行了统一表征。使用线性换算将 12 组启动子在原文献中的强度数据进行标准化计算，获得所有启动子在同一套实验体系下的启动子强度信息，依此建立了枯草芽孢杆菌启动子强度数据集。鉴于 DNA 序列本质上是一种复杂的生物文本信息，其编码的遗传指令与自然语言文本在结构与信息表达上存在相似之处，本研究借鉴自然语言处理（natural language processing, NLP）领域的先进模型，构建了枯草芽孢杆菌启动子强度预测模型。NLP 技术，作为人工智能的一个分支，专长于理解和生成人类语言，其模型通过深度学习方法能有效挖掘文本中的模式与关联，近年来在诸多领域展现出了强大的分析能力[28]。考虑到 DNA 序列分析需同时兼顾局部特征的捕获与长程依赖性的理解，本研究挑选了一系列模型，包括 FastText[29]、TextCNN（Text Convolutional Neural Network）[30]、TextRNN（Text

Recurrent Neural Network）[31]、TextRCNN（Text Recurrent Convolutional Neural Network）[32]、TextRNN_Att（Text Recurrent Neural Network with Attention）[33]和 Transformer[34]。这些模型均在处理序列数据和文本信息方面表现出色。针对 DNA 序列中特定的局部结构特征，本研究还纳入了深度卷积网络模型 DPCNN[35]。本研究分析了不同自然语言处理模型对枯草芽孢杆菌启动子强度的预测效果，旨在为枯草芽孢杆菌特定强度启动子的筛选提供一定参考。

1　材料与方法

1.1　材料

枯草芽孢杆菌（*Bacillus subtilis*）WB600 为荧光强度表达宿主，大肠杆菌（*E. coli*）JM109 为克隆宿主，pLY-3-GFP 为载体质粒。以上材料均来自本实验室。

无缝克隆酶，北京全式金生物技术有限公司，PrimeSTAR Max DNA 聚合酶，TaKaRa 公司；质粒快速提取试剂盒和 DNA 纯化回收试剂盒，Omega 公司；GeneRuler 1 kb DNA Ladder，Thermo Scientific 公司；胰蛋白胨与酵母提取物，Oxoid 公司；卡那霉素（Kan）、氯霉素（Cm），Solarbio 公司；琼脂糖，Sigma 公司；溴化乙锭，Aladdin 公司。

枯草芽孢杆菌培养及大肠杆菌培养均使用 LB 培养基。本研究所用到的引物见表 1，引物由金唯智生物科技有限公司合成。

表 1　引物

Tab. 1　Primers

引物名称	引物序列（5′-3′）
pLY-3-nSP-gfp-F	GGATGATCACATCAAGCAGC
pLY-3-nSP-gfp-R	TCAAATAAGGAGTGTCAAGA
P1-*gfp*-F	GGATTTTTTAAATAAAGCGTTTACAATATATGTATCAAATAAGGAGTGTCAAGAATGG
P1-*gfp*-R	ATTTAAAAAATCCAATGGGTTAAACTTTAATTTTAAACACGGATGATCACATCAAGCAGC
P2-*gfp*-F	GGATTTTTTAAATAAATCGGTTACAATATATGTATCAAATAAGGAGTGTCAAGAATGG
P2-*gfp*-R	ATTTAAAAAATCCAAATATTTAAACTTTAATTTTAAACACGGATGATCACATCAAGCAGC
P3-*gfp*-F	GATAAAAACATTTTCTTTTGATAAACTGAACGGTCAAATAAGGAGTGTCAAGAATGG
P3-*gfp*-R	AATGTTTTTATCACCGAAAAATGGGTGAAAAGTTTCATGCGGATGATCACATCAAGCAGC
P4-*gfp*-F	TGACCTTTATTGACCAAAAATGTATCATGTAACTTCAAATAAGGAGTGTCAAGAATGG
P4-*gfp*-R	TCAATAAAGGTCAAACAAAAAGCTGGCCTGATATGCAAAGGGATGATCACATCAAGCAGC
P5-*gfp*-F	GGATTTTTTACCGCCCGCGTTTACAATATATGTATCAAATAAGGAGTGTCAAGAATGG
P5-*gfp*-R	GGTAAAAAATCCAAATATTTAAACTTTAATTTTAAACACGGATGATCACATCAAGCAGC
P6-*gfp*-F	GACAAAAATGGGCTCGTGTTGGAGAATAAATGTGTCAAATAAGGAGTGTCAAGAATGG
P6-*gfp*-R	GCCCATTTTTGTCAAATAAAATTTAACCGGTATCAACGTTGGATGATCACATCAAGCAGC
P7-*gfp*-F	GACAAGTATTTCCGACACATTCATAATGAAGTTGTCAAATAAGGAGTGTCAAGAATGG
P7-*gfp*-R	GAAATACTTGTCAAGCTTGCCATCTTAACGTTTGCAGGCGGATGATCACATCAAGCAGC
P8-*gfp*-F	GTTGACACTCTTTTGAGAATATGTGATATTATCAGGTCAAATAAGGAGTGTCAAGAATGG
P8-*gfp*-R	CAAAAGAGTGTCAACGTGTATTGACGCAGTAAAGGATAAAGGATGATCACATCAAGCAGC
P9-*gfp*-F	GACATTTTTAAATAAAGCGTTTATAATATATGTATCAAATAAGGAGTGTCAAGAATGG
P9-*gfp*-R	ATTTAAAAATGTCAAATATTTAAACTTTAATTTTAAGCACGGATGATCACATCAAGCAGC
P10-*gfp*-F	CACAGTGAATGAAGACCTGTGCTATATTTAATAGGTCAAATAAGGAGTGTCAAGAATGG
P10-*gfp*-R	CTTCATTCACTGTGAACAAGAGAGATCAGCTGACTTCAACGGATGATCACATCAAGCAGC
P11-*gfp*-F	TAAATTATGATAGAATAAGAAATGTAAAGTATATTCAAATAAGGAGTGTCAAGAATGG
P11-*gfp*-R	CTATCATAATTTAACCACGAGAAGAAATATGAAATGTCGTGGATGATCACATCAAGCAGC
P12-*gfp*-F	GACAATCGTCCTCCAACGTGCTATAATTCTACAATCAAATAAGGAGTGTCAAGAATGG
P12-*gfp*-R	GAGGACGATTGTCAACACTTTTTTTTGATTTTGCTGCCTTGGATGATCACATCAAGCAGC
pLY-3-R(测序)	GTTTGTGATGGCTTCCATGTCG
pLY-3-F(测序)	CGGCACTGAATTTAATTCGGAAGGTC

1.2 方法

1.2.1 数据处理

查阅文献，收集通过实验方法检测获得的枯草芽孢杆菌启动子强度的相关数据，采取 3 种方法对数据进行统一。如果文献的数据中有短序列信息（不足 61 bp）并且标注了 TSS，则使用 BLAST 工具[36]在枯草芽孢杆菌基因组中进行序列比对，以比对到的 TSS 为基点，向上游序列截取共 61 bp 的序列；如果文献中有启动子序列，但没有标注 TSS，则使用 TSSPredator 工具[37]预测转录起始位点，再以预测到的 TSS 为基点向上游截取 61 bp 的序列；如果文献中既没有标注 TSS，提供的启动子序列又小于 61 bp，则既进行序列比对，又进行 TSS 的预测，以预测得到的 TSS 为基准向基因组上游截取 61 bp 的序列。

1.2.2 重组菌株的构建及荧光强度测定

（1）质粒及重组菌株的构建

为了使用荧光强度对每组数据的启动子强度进行表征，选用 pLY-3-GFP 为载体质粒，质粒详情如图 1（a）所示。以 pLY-3-GFP 为模板，以表 1 中列举的片段为引物，采用反向 PCR 方法构建重组质粒。经过琼脂糖凝胶电泳检测质粒是否构建成功，如图 1（b）所示。经过无缝克隆将未连接成环状的线性质粒连接成环状，添加无缝克隆体系并轻轻混匀，将样品放于 50 ℃水浴锅中反应 15 min，反应结束后放置冰上冷却，-20 ℃冰箱保存。使用 DNA 琼脂糖凝胶回收试剂盒进行 DNA 片段的纯化回收，将构建好的质粒转化进入大肠杆菌。

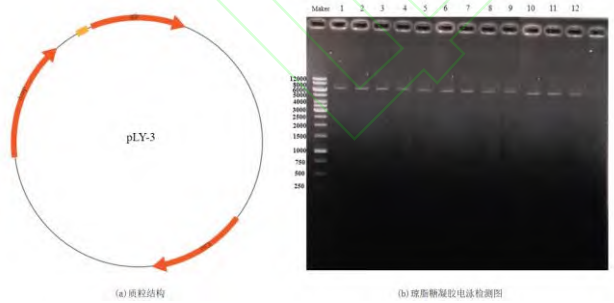

图 1　报告质粒及琼脂糖凝胶电泳检测图

Fig. 1　Reporter plasmid map and agarose gel electrophoresis results

转化后的大肠杆菌接种于 LB 液体培养基在 37 ℃、220 r/min 培养 1 h，5000 r/min 离心 5 min，去除上清液，将剩余菌株与培养基充分混匀，涂布培养基平板，37 ℃倒置培养 12 h。从平板上选取 3 个单菌落进行反向 PCR 验证，以 Ply-3-F、Ply-3-R 为引物，以挑取菌落为模板，通过琼脂糖凝胶电泳进行相对分子质量验证。选取验证正确的菌株进行测序，保存平板，将测序正确的质粒转化进入枯草芽孢杆菌中。

（2）荧光强度的测定

首先将枯草芽孢杆菌接种于 5 mL LB 培养基中，37 ℃、220 r/min 培养至吸光度 A_{600}=0.4~0.6，根据吸光度调整菌体总量，确定不同菌株所需的发酵体积。

将枯草芽孢杆菌接种于 50 mL LB 液体培养基中，进行 3 个平行实验，在 8、12、16、20、24 h 分别测量对应的吸光度和荧光强度，选荧光强度最强的时间进行记录。荧光强度测定方法为：全程避光，取发酵液在 13000 r/min 离心后弃上清液，用 PBS 缓冲液反复清洗 3 次，随后将菌体重悬到 PBS 缓冲液中；吸取 200 μL 混合液加入到黑色不透光的 96 孔板中；酶标仪设定激发波长 480 nm、发射波长 520 nm，扫描完毕后记录数值。

1.2.3 启动子强度标准化计算及数据集的划分

从 12 组数据中分别选择一条强度适中的启动子序列，利用以上方法进行重组菌株的构建及荧光强度的测定，所测强度数据作为启动子强度标准化计算的依据。根据文献中启动子强度与本地表征标准数据的比值重新定义所有启动子序列的强度信息。标准化计算如式（1）所示，其中 x 为启动子标准化计算后强度；F_a 为代表启动子序列本地表征的荧光强度；F_A 为代表启动子序列在原文献中强度；F_X 为需要计算的启动子序列在原文献中的强度。

$$x=\frac{F_a}{F_A}\times F_X \tag{1}$$

将数据集按照荧光强度由强到弱的顺序排列，平均分为 3 组，选取荧光强度高的一组作为正样本（强启动子）；荧光强度低的一组作为负样本（弱启动子），得到正负样本各 193 个。按照 8∶2 的比例划分训练集和测试集，使用训练集构建预测模型，使用测试集数据对模型的预测效果进行验证。

1.2.4 模型选择

由于启动子序列可以看作是一种文本，所以在构建启动子强度预测模型时选择了专为文本处理和自然语言处理任务设计的 FastText[29]、TextCNN（text convolutional neural network）[30]、TextRNN（text recurrent neural network）[31]、TextRCNN（Text recurrent convolutional neural network）[32]、TextRNN_Att（text recurrent neural network with

attention）[33]、DPCNN[35]和 Transformer[34]这 7 个计算模型。所选模型都为深度学习模型，在处理复杂数据时显示出一定的优势。选用 Python 语言编写程序，导入自然语言模型。模型训练所用数据包括长 61 bp 的启动子序列以及序列强度的标签（强启动子标记为 1，弱启动子标记为 0）。本文的计算代码可以通过以下网址访问：https://github.com/ccg56/NLP.

1.2.5　模型评估

为了评估预测模型的质量或比较不同预测模型的性能，需要使用一系列评估指标。本研究采用以下 4 个指标：敏感性（sensitivity，SN，用符号 S_N 表示）、特异性（specificity，SP，用符号 S_P 表示）、准确率（accuracy，ACC，用符号 A 表示）以及马修斯相关系数（Mathew's correlation coefficient，MCC，用符号 M 表示）。在本研究中，SN 是指正确预测为强启动子的比例，SP 为将负样本正确判断的比例，准确率（ACC）是指正确区分强启动子和弱启动子的总体准确率，这是一个相对直观的测量参数。MCC 是一种用于衡量二分类模型性能的指标，综合了敏感性、特异性和召回率。它可以解决数据不平衡和分类器偏好的问题，其值在-1~+1 之间。计算公式为

$$S_N = \frac{N_{TP}}{N_{TP} + N_{FN}} \quad (2)$$

$$S_P = \frac{N_{TN}}{N_{TN} + N_{FP}} \quad (3)$$

$$A = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}} \quad (4)$$

$$M = \frac{N_{TP} \cdot N_{TN} - N_{FP} \cdot N_{FN}}{\sqrt{(N_{TP} + N_{FP}) \cdot (N_{TP} + N_{FN}) \cdot (N_{TN} + N_{FP}) \cdot (N_{TN} + N_{FN})}} \quad (5)$$

其中：N_{TP} 为正确预测为强启动子的数量，N_{TN} 为正确预测为弱启动子的数量，N_{FP} 为将实际为弱启动子错误地预测为强启动子的数量，N_{FN} 为将实际为强启动子错误地预测为弱启动子的数量。

2　结果与讨论

2.1　利用荧光强度对启动子强度数据标准化计算

本研究共收集了来自 9 篇文献中 12 组共计 591 条数据。数据包含启动子序列及其对应的荧光强度数值。Jason 等[38]研究发现，虽然不同研究中的启动子表达强度不能直接通用，但各启动子的相对表达强度比较稳定。基于这一点，从每组数据中各选取 1 条强度适中的启动子，在枯草芽孢杆菌中使用绿色荧光蛋白的荧光强度作为表征指标，对这 12 个启动子序列的强度进行重新表征，从而获得启动子的相对强度关系。根据每条启动子在其来源文献的启动子荧光强度数据，基于相对强度关系进行换算，最后获得所有启动子在同一套实验体系下的启动子强度信息。每组数据在原文献中的荧光强度范围、标准化计算后的荧光强度范围、选取的代表性启动子序列及其在原文献的荧光强度和本地表征荧光强度见表 2。

将标准化计算后的启动子数据集按照启动子强度分为低、中和高 3 份，最低组为负样本，最高组为正样本，最终获得正负样本各 193 个。对正负样本的数据进行多序列比对，并使用 WebLogo[39]工具创建序列标识图，如图 2 所示。从图 2 中可以看出正样本序列，即强启动子的序列具有更好的保守性。强启动子在-10 区的 TATAAT 序列和-35 区的 TTGACA 序列也更明显。

表 2　启动子数据集信息

Tab.2　Promoter data set information

编号	数据量	原文荧光强度范围	代表启动子序列	原文荧光强度	本地表征荧光强度	标准化计算后荧光强度范围	参考文献
1	60	12.6~148.8	GTGTTTAAAATTAAAGTTTAACCCATTGGATT TTTTAAATAAAGCGTTTACAATATATGTA	132.3	4.72	0.45~5.30	[40]
2	18	13449.4~41774.2	GTGTTTAAAATTAAAGTTTAAATATTTGGATT TTTTAAATAGCTGGGTTACAATATATGTA	22096.8	13.89	8.45~26.25	[9]
3	27	7758.62~31854.5	GCATGAAACTTTTCACCCATTTTTCGGTGATA AAAACATTTTCTTTTGATAAACTGAACGG	18327.3	4.97	2.11~8.65	[9]
4	79	0.162262~454.446	CTTTGCATATCAGGCCAGCTTTTTGTTTGACCT TTATTGACCAAAAATGTATCATGTAACT	26.4752	4.81	0.029~82.48	[41]
5	11	2.1~100	GTGTTTAAAATTAAAGTTTAAATATTTGGATT TTTTACCGCCCGCGTTTACAATATATGTA	73.2	4.95	0.14~6.76	[42]
6	10	0.0114~1.25	AACGTTGATACCGGTTAAATTTTATTTGACAA AAATGGGCTCGTGTTGGAGAATAAATGTG	0.303	5.38	0.20~22.18	[43]
7	17	0.00877~2.38	GCCTGCAAACGTTAAGATGGCAAGCTTGACA AGTATTTCCGACACATTCATAATGAAGTTG	0.218	5.34	0.21~58.31	[43]
8	22	0.00574~4.93	TTTATCCTTTACTGCGTCAATACACGTTGACA CTCTTTTGAGAATATGTGATATTATCAGG	0.0871	5.32	0.35~300.88	[43]
9	12	261.236~6922.28	GTGCTTAAAATTAAAGTTTAAATATTTGACAT TTTTAAATAAAGCGTTTATAATATATGTA	2341.97	15.88	1.77~46.95	[44]
10	105	332~20942	GTTGAAGTCAGCTGATCTCTCTTGTTCACAGT GAATGAAGACCTGTGCTATATTTAATAGG	8986	4.37	0.16~10.18	[45]
11	21	0.05~3.91	ACGACATTTCATATTTCTTCTCGTGGTTAAATT ATGATAGAATAAGAAATGTAAAGTATAT	1.7	5.76	0.17~13.25	-
12	209	1.53~94	AAGGCAGGCAAATGCGAAAAAGGTGTTGACA ACAGTGAATGCTTATGGTATAATTAGTGAA	20.48	7.95	0.59~36.49	[46]

注：原文启动子强度因单位不一，这里仅使用其相对强度数值。

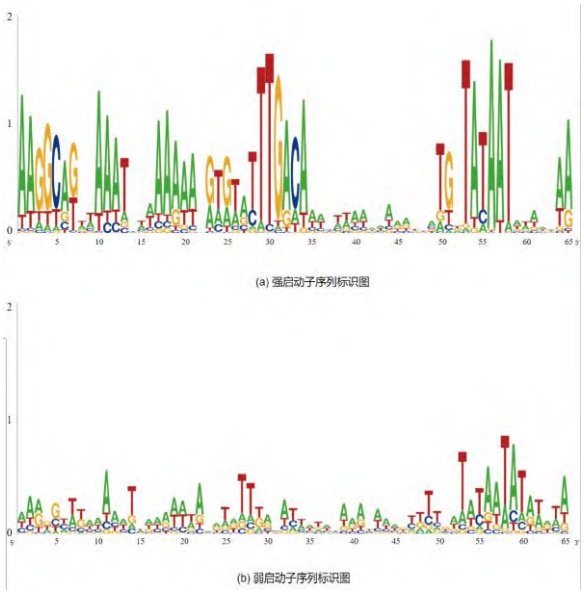

图 2　强启动子、弱启动子序列标识图

Fig. 2　Strong/weak promoter sequence logo

2.2　启动子强度预测模型

以枯草芽孢杆菌启动子强度数据集为基准数据集进行预测。由于选择的 7 个模型 FastText、TextCNN、TextRNN、TextRCNN、TextRNN_Att、DPCNN 和 Transformer 属于深度学习模型，具有强大的特征学习能力，能够自动从原始文本数据中提取和学习最具代表性的特征，因此在进行预测任务时，不需要进行手动特征提取。图 3 为 7 个模型在测试集上预测结果的混淆矩阵。每个混淆矩阵的左上角和右下角的数字，分别代表预测正确的正样本数量和预测正确的负样本数量。从图 3 可以看出 7 个预测模型都可以在一定程度上对不同强度的启动子序列进行区分。

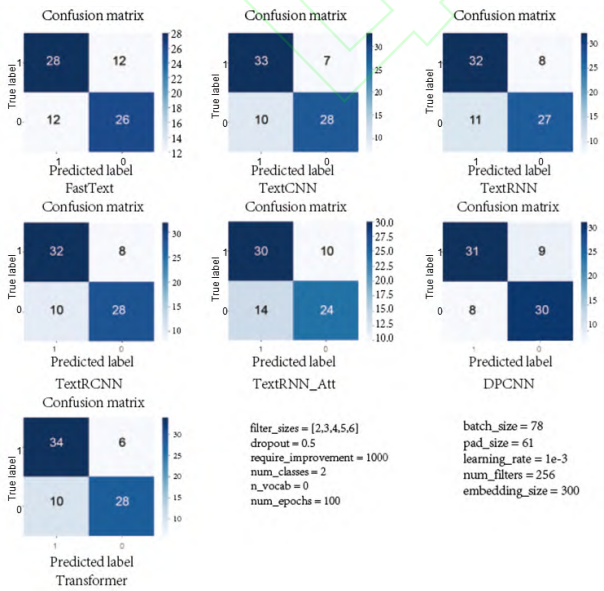

图 3　枯草芽孢杆菌启动子强度预测结果混淆矩阵

Fig. 3　Confusion matrix of prediction results of promoter strength in *Bacillus subtilis*

为了能够更直观地展示每个模型在数据集上的性能，根据混淆矩阵计算出了用于评估模型性能的指标 SN、SP、ACC 和 MCC，如图 4 所示。

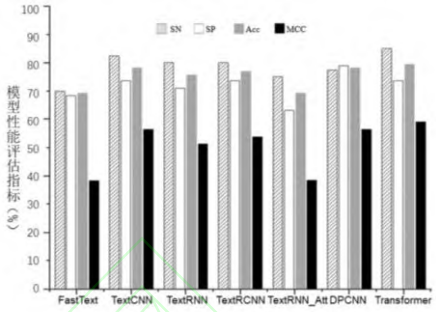

图 4　枯草芽孢杆菌启动子强度预测结果

Fig. 4　Prediction results of promoter strength in *Bacillus subtilis*

由图 4 可以看出，Transformer 模型表现出最佳的性能。该模型对强启动子和弱启动子数据分类效果的敏感性为 85.00%，特异性为 73.68%，准确率可达 79.49%，MCC 值为 0.5915。这表明 Transformer 模型在预测枯草芽孢杆菌启动子强度方面具有很高的精确性和可靠性。这种优势可能源于 Transformer 模型的独特设计，比如其自注意力机制和并行处理能力，使它可以更好地捕捉启动子序列中的关键特征和上下文信息。

TextCNN 和 TextRNN 模型的敏感性和准确率表现良好，但其特异性和马修斯相关系数略逊一筹。这可能意味着它们在识别强启动子方面有一定优势，但在区分弱启动子时可能会出现一些错误。

FastText 模型的所有评估指标均表现不佳，这可能是由于它所采用的基本词汇单元表示方法，即所谓的“词袋模型”（bag of words model），无法有效地捕捉启动子序列中复杂的结构特征和深层次的语义关联。然而，FastText 作为一种基础的预训练模型，可以在与其他模型组合时发挥辅助作用，帮助提取基本的词汇特征。

3　结　论

本研究通过收集 9 篇文献共 12 组枯草芽孢杆菌启动子强度数据，从每组数据集中各选择一条强度适中的启动子序列，通过绿色荧光蛋白对启动子强度进行表征，获得同一套实验体系下的强度信息，建立了枯草芽孢杆菌启动子强度数据集。使用 7 种

自然语言处理方法构建了对应的启动子强度预测模型，其中具有多注意力特性的 Transformer 模型在启动子强度预测任务中取得最好的预测效果，准确率可达 79.49%。本研究建立了枯草芽孢杆菌启动子强度预测模型，并证明了 Transformer 模型在该领域的优越性。这些结果可为进一步优化和开发更有效的启动子强度预测模型提供一定参考。

然而，即使是最优的 Transformer 模型，其准确率仍然有一定的提高空间。启动子强度的决定因素非常复杂，除了序列本身外，还包括染色质状态等多种因素。未来的研究应该考虑整合更多的数据来源，如表观遗传信息和蛋白质互作数据，以期进一步提高预测精度。此外，由于不同物种的基因组结构和表达调控机制有所不同，不同物种间的启动子强度预测模型可能存在差异，因此有必要进一步研究不同物种之间的模型通用性问题，以扩大模型的应用范围。

参考文献：

[1] HENDERSON K L, EVENSEN C E, MOLZAHN C M, et al. RNA polymerase: step-by-step kinetics and mechanism of transcription initiation[J]. Biochemistry, 2019, 58(18): 2339-2352.

[2] KANHERE A, BANSAL M. Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes[J]. Nucleic acids research, 2005, 33(10): 3165-3175.

[3] FEKLISTOV A. RNA polymerase: in search of promoters[J]. Annals of the New York academy of sciences, 2013, 1293(1): 25-32.

[4] HELMANN J D. Where to begin? Sigma factors and the selectivity of transcription initiation in bacteria[J]. Molecular microbiology, 2019, 112(2): 335-347.

[5] BROWNING D F, BUSBY S J. The regulation of bacterial transcription initiation[J]. Nature reviews microbiology, 2004, 2(1): 57-65.

[6] SILVA S, ECHEVERRIGARAY S. Bacterial promoter features description and their application on *E. coli* in silico prediction and recognition approaches[M]//PÉREZ-SÁNCHEZ H. Bioinformatics. Rijeka: InTech, 2012: 241-260.

[7] SZOKE P A, ALLEN T L, DEHASETH P L. Promoter recognition by *Escherichia coli* RNA polymerase: effects of base substitutions in the -10 and -35 regions[J]. Biochemistry, 1987, 26(19): 6188-6194.

[8] BRÁZDA V, BARTAS M, BOWATER R P. Evolution of diverse strategies for promoter regulation[J]. Trends in genetics, 2021, 37(8): 730-744.

[9] HAN L C, CUI W J, SUO F Y, et al. Development of a novel strategy for robust synthetic bacterial promoters based on a stepwise evolution targeting the spacer region of the core promoter in *Bacillus subtilis*[J]. Microbial cell factories, 2019, 18(1): 1-14.

[10] SU W, LIU M L, YANG Y H, et al. PPD: a manually curated database for experimentally verified prokaryotic promoters[J]. Journal of molecular biology, 2021, 433(11): 166860.

[11] SANTOS-ZAVALETA A, SALGADO H, GAMA-CASTRO S, et al. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12[J]. Nucleic acids research, 2019, 47(D1): 212-220.

[12] WANG H Q, BENHAM C J. Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to super helical stress[J]. BMC Bioinformatics, 2006, 7(1): 248.

[13] RAHMAN M S, AKTAR U, JANI M R, et al. iPro70-FMWin: identifying sigma 70 promoters using multiple windowing and minimal features[J]. Molecular genetics and genomics, 2019, 294(1): 69-84.

[14] LIN H, DENG E Z, DING H, et al. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition[J]. Nucleic acids research, 2014, 42(21): 12961-12972.

[15] ZHANG M, LI F Y, MARQUEZ-LAGO T T, et al. MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters[J]. Bioinformatics, 2019, 35(17): 2957-2965.

[16] ZHANG P Y, ZHANG H M, WU H. iPro-WAEL: a comprehensive and robust framework for identifying promoters in multiple species[J]. Nucleic acids research, 2022, 50(18): 10278-10289.

[17] TOWSEY M, HOGAN J M, MATHEWS S, et al. The in silico prediction of promoters in bacterial genomes[J]. Genome informatics, 2007, 19: 178-189.

[18] SHAHMURADOV I A, MOHAMAD RAZALI R, BOUGOUFFA S, et al. bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and *Escherichia*

coli[J]. Bioinformatics, 2017, 33(3): 334-340.

[19] MANN S, LI J Y, CHEN Y P. A pHMM-ANN based discriminative approach to promoter identification in prokaryote genomic contexts[J]. Nucleic acids research, 2007, 35(2): e12.

[20] UMAROV R K, SOLOVYEV V V. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks[J]. PLOS One, 2017, 12(2): e0171410.

[21] LE N Q K, YAPP E K Y, NAGASUNDARAM N, et al. Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous fasttext N-grams[J]. Frontiers in bioengineering and biotechnology, 2019, 7: 305.

[22] LIU B, YANG F, HUANG D S, et al. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC[J]. Bioinformatics, 2018, 34(1): 33-40.

[23] XIAO X, XU Z C, QIU W R, et al. iPSW (2L)-PseKNC: A two-layer predictor for identifying promoters and their strength by hybrid features via pseudo K-tuple nucleotide composition[J]. Genomics, 2019, 111(6): 1785-1793.

[24] LIANG Y Y, ZHANG S L, QIAO H J, et al. iPromoter-ET: Identifying promoters and their strength by extremely randomized trees-based feature selection[J]. Analytical biochemistry, 2021, 630: 114335.

[25] TAYARA H, TAHIR M, CHONG K T. Identification of prokaryotic promoters and their strength by integrating heterogeneous features[J]. Genomics, 2020, 112(2): 1396-1403.

[26] SU Y, LIU C, FANG H, et al. *Bacillus subtilis*: a universal cell factory for industry, agriculture, biomaterials and medicine[J]. Microbial cell factories, 2020, 19: 173.

[27] SIERRO N, MAKITA Y, DE HOON M, et al. DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information[J]. Nucleic acids research, 2008, 36: 93-96.

[28] REDDY D R. Speech recognition by machine: a review[J]. Proceedings of the IEEE, 1976, 64(4): 501-531.

[29] JIN Y, YANG Y. ProtPlat: an efficient pre-training platform for protein classification based on FastText[J]. BMC Bioinformatics, 2022, 23(1): 66.

[30] CHEN N, SU X D, LIU T Y, et al. A benchmark dataset and case study for Chinese medical question intent classification[J]. BMC Medical informatics and decision making, 2020, 20(S3): 125.

[31] LIU P F, QIU X P, HUANG X J. Recurrent neural network for text classification with multi-task learning[C]//ACM. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York: ACM, 2016: 2873-2879.

[32] BANERJEE I, LING Y, CHEN M C, et al. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification[J]. Artificial intelligence in medicine, 2019, 97: 79-88.

[33] POON H K, YAP W S, TEE Y K, et al. Hierarchical gated recurrent neural network with adversarial and virtual adversarial training on text classification[J]. Neural networks, 2019, 119: 299-312.

[34] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//ACM. Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[35] JOHNSON R, ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]//ACL. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 562-570.

[36] ALTSCHUL S F, GISH W, MILLER W, et al. Basic local alignment search tool[J]. Journal of molecular biology, 1990, 215(3): 403-410.

[37] DUGAR G, HERBIG A, FORSTNER K U, et al. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates[J]. PLoS Genetics, 2013, 9(5): e1003495.

[38] COELHO R V, DALL'ALBA G, DE AVILA E SILVA S, et al. Toward algorithms for automation of postgenomic data analyses: *Bacillus subtilis* promoter prediction with artificial neural network[J]. OMICS, 2020, 24(5): 300-309.

[39] CROOKS G E, HON G, CHANDONIA J M, et al. WebLogo: a sequence logo generator[J]. Genome research, 2004, 14(6): 1188-1190.

[40] XU J T, LIU X Q, YU X X, et al. Identification and characterization of sequence signatures in the *Bacillus subtilis* promoter Pylb for tuning promoter strength[J]. Biotechnology letters, 2020, 42(1): 115-124.

[41] SONG Y F, NIKOLOFF J M, FU G, et al. Promoter screening from *Bacillus subtilis* in various conditions hunting for synthetic biology and industrial applications[J].

PLOS ONE, 2016, 11(7): e0158447.

[42] YU X X, XU J T, LIU X Q, et al. Identification of a highly efficient stationary phase promoter in *Bacillus subtilis*[J]. Scientific reports, 2015, 5(1): 1-9.

[43] GUIZIOU S, SAUVEPLANE V, CHANG H J, et al. A part toolbox to tune genetic expression in *Bacillus subtilis*[J]. Nucleic acids research, 2016, 44(15): 7495-7508.

[44] ZHOU C Y, YE B, CHENG S, et al. Promoter engineering enables overproduction of foreign proteins from a single copy expression cassette in *Bacillus subtilis*[J]. Microbial cell factories, 2019, 18(1): 1-11.

[45] YANG S, DU G C, CHEN J, et al. Characterization and application of endogenous phase-dependent promoters in *Bacillus subtilis*[J]. Applied microbiology and biotechnology, 2017, 101(10): 4151-4161.

[46] LIU D Y, MAO Z T, GUO J X, et al. Construction, model-based analysis, and characterization of a promoter library for fine-tuned gene expression in *Bacillus subtilis*[J]. ACS Synthetic biology, 2018, 7(7): 1785-1797.