DOI:10.12113/202104017

基于卷积神经网络的细菌转录终止子预测

金 冬,张 萌,贾藏芝*
(大连海事大学 理学院,辽宁 大连 116026)

摘 要:在遗传学中,终止子是位于poly(A)位点下游、长度在数百碱基以内、包含多个回文序列、具有终止转录功能的DNA结构域,其主要作用是使转录终止。在原核生物基因组中有两类转录终止子,即Rho-dependent因子和Rho-independent因子。在本项研究中,提出了一种新的预测模型(TermCNN)来快速准确地识别细菌转录终止子。该模型将具有代表性的6-*mer*特征子集(2 537个特征)和电子—离子相互作用伪电位(EIIP)作为输入向量,利用卷积神经网络(CNN)构建预测模型。五折交叉验证和独立测试的结果表明该模型优于最新的预测模型iTerm-PseKNC。值得注意的是,该模型在跨物种试验中具有明显的优势。它可以高度精确地预测大肠杆菌(*E. coli*)和枯草芽孢杆菌(*B. subtilis*)的转录终止子。

关键词:转录终止子;深度学习;特征选择;卷积神经网络
中图分类号:Q939.1 **文献标志码**:A **文章编号**:1672-5565(2022)03-182-07

Prediction of bacterial transcriptional terminators by using convolutional neural network

JIN Dong, ZHANG Meng, JIA Cangzhi*
(*School of Science*, *Dalian Maritime University*, *Dalian* 116026, *Liaoning*, *China*)

Abstract: In genetics, a transcriptional terminator is a DNA domain located downstream of poly (A) site within a length of hundreds of bases, which contains multiple palindrome sequences and has the function of terminating transcription. Two classes of transcriptional terminators, Rho-dependent and Rho-independent have been found in prokaryotic genomes. In this study, a novel model(Term CNN) was proposed for identifying bacterial transcriptional terminators rapidly and accurately. The model combined representative 6-*mer* sub-set (2 537 features) and electron-ion interaction pseudopotentials (EIIP) of nucleotides as input parameters, and convolutional neural network (CNN) was utilized to train and optimize the model. Extensive 5-fold cross-validation and independent tests showed that the model outperformed the latest prediction model iTerm-PseKNC. It is especially noted that the model achieved obviously superiority on cross-species tests. In summary, the proposed model can predict transcriptional terminators of *Escherichia coli* (*E. coli*) and *Bacillus subtilis* (*B. subtilis*) with high accuray.

Keywords: Transcriptional terminators; Deep learning; Feature selection; Convolutional neural network

在遗传学中,转录终止子通常位于poly(A)位点的下游,提供终止转录的信号。它通过对新合成的转录本RNA提供信号来介导转录终止[1-2]。一般来说,原核生物的转录终止子可分为两类:一类依赖于ρ(Rho)因子才能实现终止作用,记作Rho-dependent;另一类则不依赖ρ因子便能实现终止作用,记为Rho-independent。ρ因子是一种解旋酶,可以破坏mRNA-DNA-RNA聚合酶的转录复合体。通常在细菌和噬菌体中发现Rho-dependent转录终止子[3-4]。Rho-independent的终止位点位于翻译终止密码子的下游,由mRNA上一个无结构的、富含GC碱基对的序列组成[5]。Rho-independent[6-7]因子包含7-20个GC-rich区域,后跟一个短poly-T或T-stretch,在延长转录本上形成自退火发夹结构,转录中的RNA聚合酶遇到发夹

收稿日期:2021-04-27;**修回日期**:2021-07-01.
基金项目:国家自然科学基金(No.62071079).
作者简介:金冬,男,硕士研究生,研究方向:生物信息学.E-mail:jindongstudent@dlmu.edu.cn.
*通信作者:贾藏芝,女,教授,研究方向:生物信息学.E-mail:cangzhijia@dlmu.edu.cn.

结构将会暂停前进。它通过破坏 mRNA-DNA-RNA 聚合酶三元复合物,最终使转录终止。

在传统的实验中,转录终止子是否存在通常是通过测定 mRNA 的长度来确定的,而这种方法往往无法精确的识别终止位点[8]。因此,许多预测转录终止子的方法被开发出来。近年来,Yada 等[9]利用隐马尔可夫模型预测了大肠杆菌基因的转录终止子。Ermolaeva[10]和 Unniraman 等[11]分别使用 TransTerm 算法和 GeSTer 算法预测了转录终止子。2001 年,Lesnik 等[12]提出了一种基于热力学评分系统预测大肠杆菌 K-12 基因组终止子的方法。随着机器学习技术的发展,许多分类任务得到了解决。Feng 等[8]提取伪 *k-tuple* 核苷酸组成特征,并通过二项分布进行特征选择。随后将选择的特征与支持向量机(SVM)相结合,构建了一个名为 iTerm-PseKNC[8]的计算方法来预测转录终止子。最近,Fan 等[13]采用 *k* 核苷酸的位置信息、核苷酸的含量、核苷酸的 47 种物理化学性质作为特征向量,并结合 XGBoost 分类算法构建了预测模型 iterb-PPse,取得了相当不错的效果。值得注意的是,原有的预测方法都是采用传统的机器学习方法作为分类算法。近几年,深度学习种的卷积神经网络框架在生物信息领域得到了广泛应用,并且取得了令人满意的分类性能[14-15]。因此,我们尝试将卷积神经网络应用于细菌终止子的预测。

在本研究中,根据 Feng 等[8]的工作,引入了一种新的转录终止子预测模型,称为 TermCNN。首先从大肠杆菌 DNA 序列中提取 *k-mer* (k= 4,5,6,7)核苷酸组成特征作为 CNN 的输入向量。在五折交叉验证中,挑选出准确率最高的 6-mer 核苷酸组成特征。然后采用最大相关-最大距离(MRMD)、二项分布和 F-score 这三种特征选择方法来寻找 6-mer 特征的最优特征子集,以减少无用信息和节省运行时间。最后将选择出的最优特征子集与电子-离子相互作用伪电位(EIIP)特征相结合,输入到 CNN 进行训练,构建高精度模型。五折交叉验证以及五个独立测试数据集的实验结果一致显示了本文提出的预测模型 TermCNN 的有效性,特别是用于区分不同种类的终止子。

1　材料和方法

1.1　数据收集和预处理

一个客观的基准数据集是建立终止子预测模型的基础。从 RegulonDB[16]中收集大肠杆菌的终止子,去冗余后得到 286 个 Rho-independent 终止子和 19 个 Rho-dependent 终止子[8]。与之前的数据集相比,RegulonDB 新增了 25 个转录终止子。将新发现的 25 个转录终止子视为一个独立的测试集,命名为 E_Ter_25。对于训练数据集,采用了与 Feng[8]相同的数据集,包含 280 个终止子和 560 个非终止子,便于评估和比较不同预测器的性能。对于独立测试,Feng [8]使用了两个终止子独立测试集,分别是 E_Ter_147 和 B_Ter_425。从 Fan 等[13]的工作中选取两个均为负样本构成的独立测试集,样本是分别从大肠杆菌和枯草芽孢杆菌的上游截取的,记为 E_Nonter_159 和 B_Nonter_122。在缩写中,E 表示来自大肠杆菌的序列,B 表示来自枯草芽孢杆菌的序列,数字表示每个数据集中的样本数量(见表 1)。

表 1　不同物种的数据集

Table 1　Datasets of different species

物种	数据集	数量/个
大肠杆菌	正训练样本	280
	负训练样本	560
大肠杆菌	正独立测试集	172
	负独立测试集	159
枯草芽孢杆菌	正独立测试集	425
	负独立测试集	122

1.2　特征提取

特征提取在开发基于机器学习算法的计算模型中起着非常重要的作用。本文从序列中提取了两类特征:一个是 *k-mer*,另一个是 EIIP[17]。

1.2.1　*k-mer* 核苷酸组成

给出一个 DNA 序列 D,它的直观表达式是[18]:

$$D = R_1\ R_2\ R_3\ R_4\ R_5\ R_6\ R_7 \cdots\ R_L \qquad (1)$$

其中 R_i 表示在 DNA 序列中第 i 个位置的核苷酸。

k-mer 核苷酸是将 DNA 序列转化为数字向量的一种简单而常用的方法,这一方法具有重要的生物学意义,在 DNA 调控元件识别中已得到了广泛的应用[19-23]。*k-mer* 可以将任何 DNA 序列表示为 4^k 维的向量如下:

$$R = [\varphi_1\ \varphi_2 \cdots\ \varphi_u \cdots\ \varphi_{4^k}] \qquad (2)$$

其中 $\varphi_u(u = 1,2\cdots,4^k)$ 为沿着序列第 u 个 *k-mer* 的频率。在本工作中,k=1、2、3、4、5、6、7,并与 EIIP 相结合进行测试,寻找最优的特征集。

1.2.2　EIIP

EIIP 作为特征已被广泛的用来预测基因序列[17]。基于 EIIP 的识别方法广泛应用于基因结构识别的关键部分,如 *F56F11.4* 基因的预测[24]、囊性纤维化基因的预测和和增强子的识别[25]等。

四个核苷酸的 EIIP 值分别为,A: 0.126, G:0.0806, C: 0.134, T:0.133。计算每条序列中 A、T、G、C 的平均 *EIIP* 值,构造特征向量为[25]:

$$P = [EIIP_{AAA} \cdot f_{AAA}, EIIP_{AAC} \cdot f_{AAC}, \cdots, EIIP_{TTT} \cdot f_{TTT}] \quad (3)$$

其中 f_{XYZ} 为任意三核苷酸 XYZ 的频率，$EIIP_{XYZ}$ 是三核苷酸 *XYZ* 的 *EIIP* 值之和，$X,Y,Z \in \{A,C,G,T\}$。

1.3 特征选择

特征选择方法可以降低特征向量的维数，为训练分类器找到最优特征子集。近几年，最大相关-最大距离（MRMD）[26]、F-score[27] 和二项分布（BD）[28]，方法在改善预测器性能上具有显著成效，已广泛应用于生物信息学领域。

1.3.1 MRMD

MRMD 利用皮尔逊相关系数计算特征子集与目标类的相关性，并使用欧氏距离函数计算特征子集的冗余度，相关性与距离的和最大的特征被选择到最终的特征子集中。首先定义两个向量的相关系数如下：

$$PCC(\vec{X},\vec{Y}) = \frac{S_{\vec{X}\vec{Y}}}{S_{\vec{X}} S_{\vec{Y}}} \quad (4)$$

其中，

$$S_{\vec{X}\vec{Y}} = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y}) \quad (5)$$

$$S_{\vec{X}} = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(x_k - \bar{x})^2} \quad (6)$$

$$S_{\vec{Y}} = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{y})^2} \quad (7)$$

$$\bar{x} = \frac{1}{N}\sum_{k=1}^{N} x_k \quad (8)$$

$$\bar{y} = \frac{1}{N}\sum_{k=1}^{N} y_k \quad (9)$$

这里 x_k, y_k 是向量 $\vec{X},\vec{Y}$ 的第 k 个元素。

从而，第 i 个特征的最大相关值 MR 定义为：

$$\max MR_i = |PCC(\vec{F}_i, \vec{C}_i)| (1 \leqslant i \leqslant M) \quad (10)$$

其中 $\vec{F}_i$ 是一个 M-D 向量，其元素来自于每一个样本的第 i 个特征；$\vec{C}_i$ 也是一个 M-D 向量，其元素来自于每一个样本的目标标签。

本文中，两个特征的距离采用欧式距离，定义如下：

$$ED(\vec{X}\vec{Y}) = \sqrt{\sum_{k=1}^{N}(x_k - y_k)^2} \quad (11)$$

其中，x_k, y_k 是向量 $\vec{X},\vec{Y}$ 的第 k 个元素。

最大距离就是取所有欧氏距离中的最大值，记为：

$$\max MD_i = ED_i (1 \leqslant i \leqslant M) \quad (12)$$

根据以上结果，第 i 个特征 MRMD 值定义为 $\max(MR_i + MD_i)$，根据此值的大小，对特征进行排序。数值越大，表明此特征与目标标签的相关性越强[26]。

1.3.2 *F-score*

第 j 个特征的 *F-score* 定义为：

$$F\text{-}score(j) = \frac{(\bar{x}_j^{(+)} - \bar{x}_j)^2 + (\bar{x}_j^{(-)} - \bar{x}_j)^2}{\frac{1}{m^+ - 1}\sum_{k=1}^{m^+}(\bar{x}_{k,j}^{(+)} - \bar{x}_j^{(+)})^2 + \frac{1}{m^- - 1}\sum_{k=1}^{m^-}(\bar{x}_{k,j}^{(-)} - \bar{x}_j^{(-)})^2} \quad (13)$$

其中 $\bar{x}_j, \bar{x}_j^{(+)}$ 和 $\bar{x}_j^{(-)}$ 分别表示全部、正、负数据集中第 j 个特征的平均值。m^+ 和 m^- 是正样本和负样本的数目。$\bar{x}_{k,j}^{(+)}$ 和 $\bar{x}_{k,j}^{(-)}$ 分别表示第 k 个正样本和第 k 个负样本的第 j 个特征。*F-score* 值越大，说明特征越有用。

1.3.3 二项分布

Feng[8] 和 Su[29] 采用基于二项分布的技术，通过 SVM 分类器进行五折交叉验证的性能结果对特征进行选择。这里，先验概率 q_i 定义为 *k-mer* 核苷酸的频率，如下所示：

$$q_i = \frac{m_i}{M} \quad (14)$$

其中 $m_i (i = 1, -1)$ 分别表示正、负训练数据集（即终止子和非终止子数据集）中的 *k-mer* 片段总数。M 为全部训练数据集中 *k-mer* 片段的总数。因此，第 j 个 *k-mer* 核苷酸（j=1,2,…, 4^k）在正样本和负样本中的概率可以定义为：

$$p(n_{1,j}) = \sum_{m=n_{1j}}^{N_j} \frac{N_j!}{m!\ (N_j - m)} q_i^m (1 - q_i)^{N_j - m} \quad (15)$$

$$p(n_{-1,j}) = \sum_{m=n_{-1j}}^{N_j} \frac{N_j!}{m!\ (N_j - m)} q_i^m (1 - q_i)^{N_j - m} \quad (16)$$

其中 N_j 表示终止子和非终止子训练数据集中第 j 个 *k-mer* 核苷酸的总数。$n_{1,j}$ 和 $n_{-1,j}$ 分别表示正、负训练数据集中第 j 个 *k-mer* 核苷酸的总数。

最后，根据以下公式计算训练数据集中的第 j 个 *k-mer* 核苷酸的概率：

$$P_j = \min(p(n_{1,j}), p(n_{-1,j})) \quad (17)$$

所有的 *k-mer* 核苷酸可以根据概率的大小进行排序，也就是说，P_j 越小，相应的 *k-mer* 核苷酸对分类效果越有效。

2　卷积神经网络

CNN 已被广泛应用于各种分类任务中,其在图像识别、图像检测、语音识别等方面表现出良好的性能。随着深度学习的深入研究[30],CNN 还用于预测启动子[31]、蛋白质泛素化位点[32]、蛋白质翻译后修饰位点的 capsule 网络[33]、RNA 假尿苷位点[34]。在本研究中,借助 Keras 工具,使用 CNN 模型识别转录终止子。TermCNN 由两个卷积层、两个池化层和连接层组成(见图 1)。转录终止子包含更多的 GC 碱基对,因此使用了一个平均池化层,池化大小为 3×3,这适合于获取序列的 GC 含量。还使用 dropout 来防止模型的过拟合。对于随机梯度下降法,选择了 Adam 优化算法。整个程序在 Python 3.6 中使用,实验环境为:主机 CPU 型号为 AMD Ryzen 74 800 H with Radeon Graphics,主频为 2.90 GHz,物理内存为 16 GB,操作系统为 64 位 Windows10,深度学习框架为 TensorFlow 2.0.0。

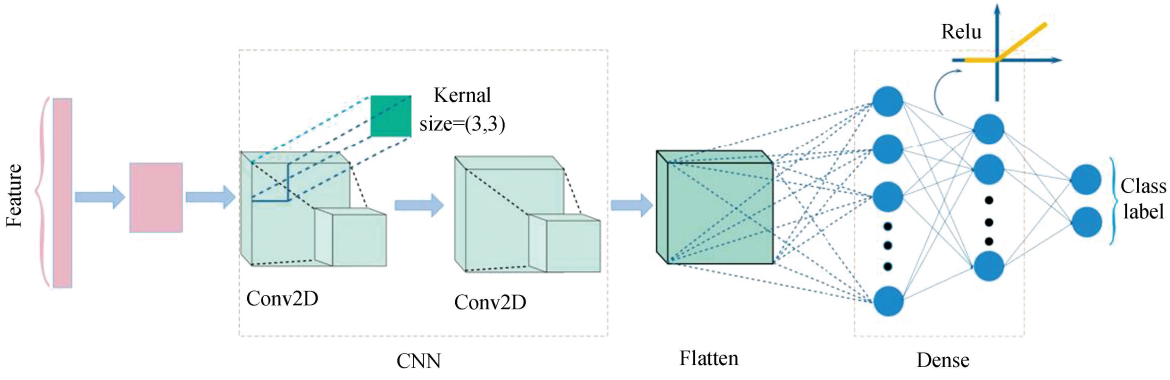

图 1　神经网络模型的架构

Fig.1　Architecture of neural network model

3　模型训练和性能评估

3.1　参数选择

采用贝叶斯对卷积神经网络的神经元个数(a)、批次(b)、dropout(c)、学习率(d)、激活函数(e)以及全连接层数(f)这六种参数进行优化。除上述参数外,所涉及到的其它参数均按照 scikit-learn 库的默认值。其中 a 在[8,64]中取值,b 在集合[8,128]中取值,c 在集合{0.1,0.3,0.5,0.7}中取值,d 在集合[0.000 1,0.01]中取值,e 的选取有两种情况 relu 和 sigmoid,f 在集合[1,10]中取值。根据贝叶斯优化方法对参数组合进行五十次寻优,耗时 1 小时 11 分钟,优化过程以及最佳参数结果(见图 2)。贝叶斯优化器建立了搜索空间的替代模型,并在此维度内进行搜索,而不是在实际搜索空间内进行搜索。优化参数的二维图,最终选取损失值最小(0.122 5)的参数组合。

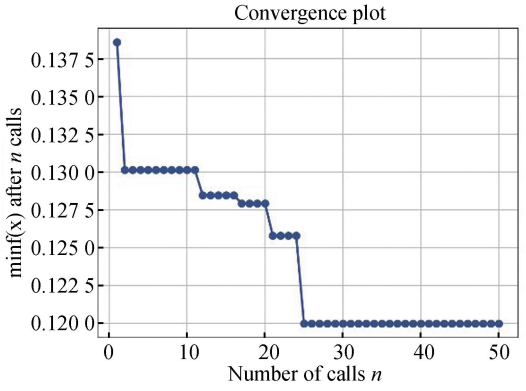

图 2　参数选择的结果

Fig.2　Results of parameters selection

3.2　性能评估

为了评估转录终止预测模型的性能,使用准确性(Acc)、灵敏度(Sn)、特异性(Sp)和马修相关系数(MCC)作为五折交叉验证和独立数据集测试的评估标准。

$$\begin{cases} Sp = \dfrac{TN}{TN + FP} \\ Sn = \dfrac{TP}{FN + TP} \\ Acc = \dfrac{TP + TN}{FN + TP + TN + FP} \\ MCC = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TN + FN)(TN + FP)(TP + FN)(TP + FP)}} \end{cases} \tag{18}$$

其中 TP、TN、FP 和 FN 分别代表真阳性、真阴性、假阳性和假阴性的数量。

4 结果分析

4.1 选择最优的特征子集

为寻找使 CNN 分类器达到最优性能的 *k-mer*(k =4,5,6,7)特征,利用五折交叉验证对每类特征进行测试。如图 3 所示,6-mer 与 CNN 的整合得到了最好的 MCC 和 Acc。6-mer 的 MCC 值为 0.942,比 4-mer高 0.101,比 5-mer 高 0.004,比 7-mer 高 0.035。考虑到 6-mer 的特征维数为 4096,高维度特征可能包含冗余信息,导致过拟合。因此,使用了 MRMD、F-Score 和二项分布这三种常用的特征选择方法来寻找最优特征子集。

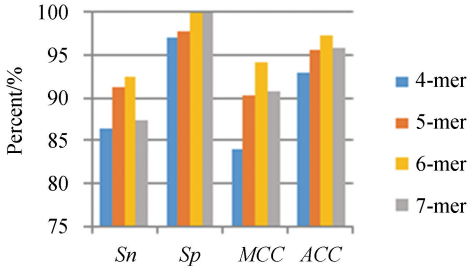

图 3 四种具有不同数量特征的模型性能

Fig.3 Performance of four models with different numbers of features

第 1 步,根据 *F-score* 值、*MRMD* 值和二项分布值对 6-mer 向量中的4 096个元素进行排序;

第 2 步,设 k=30 作为初值。需要指出的是,为特征子集选择的维数是某个数字 k 的平方,特征向量可以转换成一个方阵作为输入。因此,选取排名靠前的 k^2-64 元素与 EIIP 特征结合形成长度为 k 的方0 阵,然后将一维特征向量转换为二维方阵作为 CNN 的输入。

以步长为 5,在特征方阵长度为 k+5,通过五折交叉验证寻找准确率最高的特征子集。最后在精度最高的特征维数周围使用步长为 1 筛选出最优特征子集。并比较最优特征集和无特征选择的特征集的结果。对于 *F-score*、二项分布和 MRMD,k 值分别为 63、64 和 51 时的准确率最高。相对时间成本和准确性,将 MRMD 方法选择的 6-mer 特征向量中前 2 537 个元素与 64 个 EIIP 特征相结合,*Acc* 为 97.62%,*Sn* 为 92.86%,*Sp* 为 100%,*MCC* 为 0.947。这表明所建立的模型 TermCNN 具有良好的识别转录终止子的能力。

4.2 模型对比

为了证明使用深度学习识别转录终止子的优越性,将 CNN 与决策树、多层感知器、逻辑回归、朴素贝叶斯、基于 SVM 的 iTerm-PseKNC、iterb-PPSE 和 CNN +LSTM 进行了比较。结果如表 2 所示。可见,在浅层机器学习中,基于 SVM 的 iTerm-PseKNC 达到了最好的 *Acc*(95.71%),*MCC*(0.888),CNN 实现了较好的性能,达到 97.98%的 Acc 和 0.955 的 MCC,但是基于 XGBoost 的 iterb-PPse 给出了最好的结果,*Acc* 为 99.88%,*MCC* 为0.999。TermCNN 比 iterb-PPse 稍微逊色的原因有两个:1)提取的特征过于单一。本文仅仅考虑的终止子序列的 6-mer 特征,没有考虑位置及核苷酸的物理化学性质;2)数量的样本数量较少,不能够体现 CNN 的优越性。随着越来越多终止子序列的发现,也将会继续优化我们的模型。

表 2 不同分类器在五折交叉验证中识别终止子的比较

Table 2 Comparison of different classifiers for identifying terminators on 5-fold cross-validation

Method	*Sn*/%	*Sp*/%	*Acc*/%	*MCC*
Decision Tree	69.64	90.00	83.21	0.756
Multi-Layer Perceptron	63.57	94.46	84.16	0.708
Logistic Regression	46.07	93.92	77.97	0.384
Naive Bayes	70.71	86.25	81.07	0.752
iTerm-PseKNC	86.07	99.46	95.71	0.888
CNN+LSTM	87.14	98.86	94.93	0.887
iterb-PPse	99.64	100	99.88	0.999
TermCNN	93.93	100	97.98	0.955

4.3 独立测试集表现

为了更好地评价模型的泛化能力,进一步测试了五个独立的数据集 E_Ter_147、B_Ter_425,E_Ter_25,E_Nonter_159 和 B_Nonter_122。对于 E_Ter_147,TermCNN 正确预测了 147 个终止子,iTerm-PseKNC 也正确预测了 147 个终止子。对于 B_Ter_425,TermCNN 型正确预测了 417 个终止子(98.12%),而 iTerm-PseKNC 仅正确预测了 372 个终止子(87.53%)。对于新的独立测试集,TermCNN 正确预测了所有 25 个终止子(100%),而 iTerm-PseKNC 正确预测了 24 个终止子(96%),如图 4 所示。为了多方面检验所建立模型的有效性,从 iterb-PPse 中选取两个负样本数据集 E_Nonter_159 和 B_Nonter_122。对于 E_Nonter_159,TermCNN 预测了 158 个非转录终止子(99.37%)。对于 B_Nonter_122,TermCNN 预测了 121 个非转录终止子(99.18%)。相比于 iterb-PPse, TermCNN 预测对的数目少一个。比较遗憾的是,由于 iTerm-PseKNC 提供的网络服务器不能正常使用,因此无法和它进行比较。

4.4 特征可视化

为了更加直观的可以看到特征的有效性,通过采用 t 分布随机邻居嵌入(t-SNE)进行特征可视化。图 5 中每个点代表一个样本,蓝色点表示转录终止子位点,红色点表示非转录终止子位点。一开始可以清晰的看到只用原始特征表示的两类点很难分开,后经过神经网络层层训练,在全连接层的输出向

量可以比较明显的划分两类。因此,显示 CNN 处理转录终止子数据很有效。

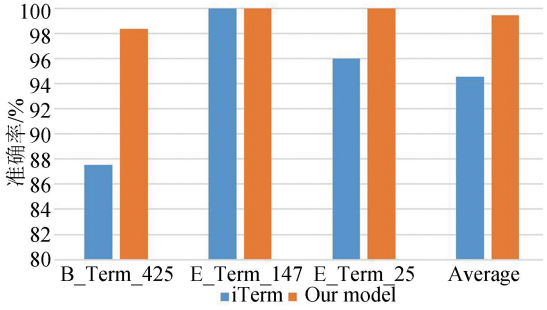

图 4　在独立测试中模型与 iTerm-PseKNC 的准确率比较

Fig.4　Accuracy comparison between iTerm-PseKNC and proposed model on independent datasets

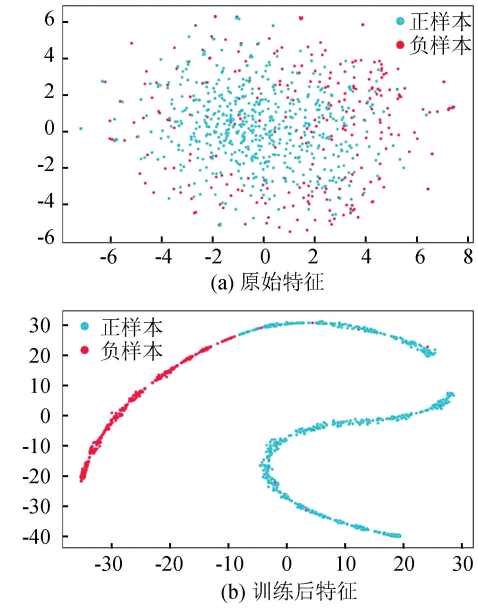

图 5　t-SNE 可视化特征表示

Fig.5　t-SNE visualization for feature representation

5　结　论

1)在这项研究中,提出了一种新的计算模型 TermCNN 可以快速准确地识别转录终止子;

2)将代表性的 6-mer 特征子集和 EIIP 作为输入参数,利用 CNN 对模型进行训练和优化;

3)五折交叉验证和多个独立测试结果证明了模型的竞争力,其性能结果明显优于其他算法和现有计算工具 iTerm-PseKNC,但是在灵敏度方面比 iterb-PPse 稍低。

参考文献(References)

[1]RICHARDSON J P, ROBERTS J W. Transcription termination[J]. Critical Reviews in Biochemistry and Molecular Biology,1993,28(1):1-30. DOI:10.3109/10409239309082571.

[2]HENKIN T M. Control of transcription termination in prokaryotes[J]. Annual Review of Genetics,1996,30(1):35-57. DOI:10.1146/annurev.genet.30.1.35.

[3]RICHARDSON J P. Rho-dependent termination and atpases in transcript termination[J]. Biochimica et Biophysica Acta,2002,1577(2):251-260. DOI:10.1016/S0167-4781(02)00456-6.

[4]RICHARDSON J P. Loading rho to terminate transcription[J]. Cell,2003,114(2):157-159. DOI:10.1016/s0092-8674(03)00554-3.

[5]CIAMPI M S. Rho-dependent terminators and transcription termination[J]. Microbiology,2006,152(9): 2515-2528. DOI:10.1099/mic.0.28982-0.

[6]NUDLER E, GOTTESMAN M E. Transcription termination and anti-termination in *E. coli*[J]. Genes to Cells,2002,7(8):755-768. DOI:10.1046/j.1365-2443.2002.00563.x.

[7]YACHIE N, ARAKAWA K, TOMITA M. On the interplay of gene positioning and the role of Rho-independent terminators in Escherichia coli[J]. FEBS Letters, 2006,580(30): 6909-6914. DOI:10.1016/j.febslet.2006.11.053.

[8]FENG C Q, ZHANG Z Y, ZHU X J, et al. iTerm-PseKNC: A sequence-based tool for predicting bacterial transcriptional terminators[J]. Bioinformatics,2019,35(9):1469-1477. DOI:10.1093/bioinformatics/bty827.

[9]YADA T, NAKAO M, TOTOKI Y, et al. Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models[J]. Bioinformatics, 1999, 15(12):987-93. DOI:10.1093/bioinformatics/15.12.987.

[10]ERMOLAEVA M D, KHALAK H G, WHITE O, et al. Prediction of transcription terminators in bacterial genomes[J]. Journal of Molecular Biology,2000,301(1):27-33. DOI:10.1006/jmbi.2000.3836.

[11]UNNIRAMAN S, PRAKASH R, NAG-ARAJA V. Conserved economics of transcription termination in eubacteria[J]. Nucleic Acids Research,2002,30(3):675-684. DOI:10.1093/nar/30.3.675.

[12]LESNIK E A, SAMPATH R, LEVENE H B, et al. Prediction of Rho-independenttranscriptional terminators in Escherichiacoli[J]. Nucleic Acids Research, 2001, 29(17):3583-3594. DOI:10.1093/nar/29.17.3583.

[13]FAN Y, WANG W, ZHU Q. iterb-PPse: Identification of transcriptional terminators in bacterial by incorporating nucleotide properties into PseKNC[J]. PLoS One,2020,15(5):1-19. DOI:10.1371/journal.pone.0228479.

[14]LIU K,CAO L, DU P, et al. im6A-TS-CNN: Identifying the N(6)-methyladenine site in multiple tissues by using the convolutional neural network - ScienceDirect[J]. Molecular Therapy - Nucleic Acids,2020,21, 1044-1049. DOI:10.1016/j.omtn.2020.07.034.

[15]RAMZAN U, HIROYUKI K, LI Y, et al. Promoter analysis and prediction in the human genome using sequence-

based deep learning models[J]. Bioinformatics, 2019, 35(16):2730-2737.DOI:10.1093/bioinformatics/bty1068.

[16]SOCORRO G C, HELADIA S, ALBERTO S Z, et al. RegulonDB version 9.0: High-level integration ofgene regulation, coexpression, motif clustering and beyond[J]. Nucleic Acids Research, 2016, 44(D1): D133-D143. DOI:10.1093/nar/gkv1156.

[17]NAIR A S, SREENADHAN S P. A coding measure scheme empoying electron-ion interactionpseudo potential (EIIP)[J]. Bioinformation, 2006,1(6):197-202.

[18]CHEN W, LIN H, CHOU K C. Pseudo nucleotide composition or PseKNC: An effective formulation for analyzing genomic sequences[J]. Molecular BioSystems, 2015, 11(10):2620-2634. DOI:10.1039/c5mb00155b.

[19]GHANDI M, LEE D, MOHAMMAD-NOORI M, et al. Enhanced regulatory sequence prediction using gapped k-mer features[J]. PLoS Computational Biology, 2014, 10(7):e1003711. DOI:10.1371/journal.pcbi.1003711.

[20]ZHANG M, LI F, MARQUEZ-LAGO T T, et al. MULTiPly: A novel multi-layer pr-edictor for discovering general and specific types of promoters[J]. Bioinformatics, 2019, 35(17):2957-2965. DOI:10.1093/bioinformatics/btz016.

[21]JIA C, YANG Q, ZOU Q. NucPosPred: Predicting species-specific genomic nucleosome positioning via four different modes of general PseKNC[J]. Journal of Theoretical Biology, 2018, 450:15-21. DOI: 10.1016/j.jtbi.2018.04.025.

[22]YI H C, YOU Z H, ZHOU X, et al. ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation[J]. Molecular Therapy Nucleic Acids, 2019, 17(9):1-9. DOI:10.1016/j.omtn.2019.04.025.

[23]MANAVALAN B, BASITH S, SHIN T H, et al. Meta-4mCpred: A sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation[J]. Molecular Therapy Nucleic Acids, 2019, 16:733-744. DOI:10.1016/j.omtn.2019.04.019.

[24]DEERGHA R D, SWARNY M N S. Analysis ofgenomics and proteomics using DSP techniques[J]. IEEE Transactions on Circuits and Systems I Regular Papers, 2008, 55(1):370-378. DOI:10.1109/TCSI.2007.910541.

[25]HE W, JIA C. EnhancerPred2.0: Predicting enhancers and their strength based on position-specific trinucleotide propensity and electron-ion interaction potentialfeature selection[J]. Molecular Biosystems, 2017, 13(4): 767-774. DOI: 10.1039/C7MB00054E.

[26]ZOU Q, ZENG J, CAO L, et al. A novel features ranking metric with application to scalable visual and bioinformatics data classification[J]. Neurocomputing, 2016, 173(2): 346-354. DOI:10.1016/j.neucom.2014.12.123.

[27]BUI V M, WENG S L, LU C T, et al. SOHSite: Incorporating evolutionary information and physicochemical properties to identify protein S-sulfenylation sites[J]. BMC Genomics, 2016, 17(1):1-9. DOI:10.1186/s12864-015-2299-1.

[28]LAI H Y, CHEN X X, CHEN W, et al. Sequence-based predictive modeling to identify cancerlectins[J]. Oncotarget, 2017, 8(17):28169-28175. DOI:10.18632/oncotarget.15963.

[29]SU Z D, HUANG Y, ZHANG Z Y, et al. iLoc-lncRNA: Predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC[J]. Bioinformatics, 2018, 34(24):4196-4204. DOI:10.1093/bioinformatics/bty508.

[30]谢娟英,刘然.基于深度学习的目标检测算法研究进展[J].陕西师范大学学报(自然科学版), 2019, 47(5):1-9. DOI:10.15983/j.cnki.jsnu.2019.05.151.
XIE Juanying, LIU Ran. Research progress of object detection algorithm based on deep learning[J]. Journal of Shaanxi Normal University (Natural Science Edition), 2019, 47(5):1-9. DOI:10.15983/j.cnki.jsnu.2019.05.151.

[31]UMAROV R, KUWAHARA H, LI Y, et al. Promoter analysis and prediction in the human genome using sequence-baseddeep learning models[J]. Bioinformatics, 2019, 35(16):2730-2737. DOI:10.1093/bioinformatics/bty1068.

[32]FU H, YANG Y, WANG X, et al. DeepUbi: A deep learning framework for prediction of ubiquitination sites in proteins[J]. BMC Bioinformatics, 2019, 20(1):1-10. DOI:10.1186/s12859-019-2677-9.

[33]WANG D, LIANG Y, XU D. Capsule network for protein post-translational modification site prediction[J]. Bioinformatics, 2018, 35(14):2386-2394. DOI:10.1093/bioinformatics/bty977.

[34]TAHIR M, TAYARA H, CHONG K T. iPseU-CNN: Identifying RNA pseudouridine sites using convolutional neural networks[J]. Molecular Therapy Nucleic Acids, 2019, 16: 463-470. DOI:10.1016/j.omtn.2019.03.010.