doi: 10. 3969 / j. issn. 1674 - 1242. 2023. 04. 006

基于 DNABERT 编码的启动子预测模型

李伟豪，刘喆，林关宁
（上海交通大学生物医学工程学院，上海 200030）

【摘要】启动子是位于基因上游区域的特定 DNA 序列，通过识别和预测 DNA 序列中的启动子，可以更好地理解基因调控的机制，促进生物学和医学研究的进展。通过实验的方法来预测启动子既昂贵又费时，而通过计算方法进行启动子预测同样存在不足之处，如精度有待提升、序列编码方式所包含的信息量不足等。该文提出了一种新的编码方式，将预训练模型 DNABERT 应用于启动子预测的编码，并测试了使用不同深度学习模型进行预测的效果。实验结果表明，使用经过预训练和微调的 DNABERT 进行编码的 Transformer 模型在启动子预测任务中取得了较好的效果。

【关键词】启动子预测；DNA 编码；深度学习
【中图分类号】Q811.4　　【文献标志码】A　　文章编号：1674-1242（2023）04-0364-07

A Promoter Prediction Model Based on DNABERT Encoding

LI Weihao，LIU Zhe，LIN Guanning
(School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China)

【**Abstract**】Promoters are specific DNA sequences located in the upstream region of genes. By identifying and predicting promoters in DNA sequences, a better understanding of gene regulation mechanisms can be achieved, thereby advancing biological and medical research. Experimental methods for promoter prediction are both expensive and time-consuming. Additionally, computational methods for promoter prediction have limitations such as room for improvement in accuracy and insufficient information content in sequence encoding methods. This paper introduces a novel encoding approach that applies the pre-trained model DNABERT to promoter prediction. Different deep learning models are tested for prediction performance. Experimental results demonstrate that the Transformer model encoded using pre-trained and fine-tuned DNABERT achieves the best performance in promoter prediction tasks.

【**Key words**】Promoter Prediction; DNA Encoding; Deep Learning

0　引言

启动子是基因转录调控的关键区域，在真核生物中，启动子的核心区域被称为核心启动子，包含转录起始位点（Transcription Start Site，TSS）。数十年的体外研究确定了核心启动子的许多功能序列基序，这些基序是蛋白质识别和启动转录的重要结构[1,2]。在这些功能性核心启动子元件中，最著名的是 TATA-box。过去，TATA-box 被认为普遍存在于核

收稿日期：2023-11-21。
基金项目：国家自然科学基金（82150610506）。
作者简介：李伟豪（2001—），男，辽宁省营口市人，硕士研究生，从事生物信息学研究。
通信作者：林关宁，男，教授，博士研究生导师，邮箱（E-mail）：nickgnlin@sjtu.edu.cn。

心启动子中[1]。然而，许多基因的核心启动子中并没有 TATA-box，而是包含启动区域（Initiator Region，INR）或下游启动子元素（Downstream Promoter Element，DPE）等其他元件[3]。最近的研究表明，只有大约 17%的真核核心启动子包含 TATA-box[4]。另外，全基因组结构分析发现，许多核心启动子不具有任何已知的核心启动子元件。这种结构异质性允许核心启动子扩展其功能库，从而充当响应一系列条件的基因和细胞类型特异性转录调节因子[5]。由此可见，启动子具有极高的复杂性和多样性。

生物序列数据蕴含丰富的信息，可用于预测生物大分子的功能或结合位点[6,7]。本文以 DNA 序列为基础，进行了启动子的预测研究。传统的启动子预测方法包括基于实验数据的方法和基于序列特征的方法[8]。基于实验数据的方法使用实验技术（如 5'端 RACE、CAGE 和 ChIP-seq 等）来确定启动子序列的位置和边界。这类方法可以提供准确的启动子信息，但需要大量的实验数据和时间成本。基于序列特征的方法通常使用计算机算法来识别启动子序列中的共识序列，如 TATA-box、INR 和 DPE 等。这类方法通常使用机器学习算法（如神经网络、支持向量机和随机森林等）来训练模型并预测新的启动子序列。目前，许多启动子预测软件被开发出来，包括 iPro-WAEL[9]、PromoterScan[10]、Promoter[11]、Prom-Machine[12]等。然而，这些软件的准确性和特异性仍然有待提高，因为启动子的复杂性和多样性使得设计通用的计算方法来识别启动子非常困难。

基因组中的 DNA 序列可以被视为一种特殊的语言，它包含生命活动的指令和调控信息。然而，DNA 语言与自然语言有很大不同。例如，它只使用 4 种字母（A、C、G、T），并且具有多义性和远距离语义关系等。这些特点使传统的生物信息学方法难以有效地解析 DNA 语言，尤其是在数据稀缺的情况下。为了解决这个问题，Ji 等[13]提出了一种针对 DNA 语言的预训练模型，即 DNABERT 模型。DNABERT 模型在 BERT 模型的基础上进行了一些改进和适应。例如，它将输入单元从字母改为 *k*-mer（长度为 *k* 的连续子序列），并采用一种新颖的掩码策略来预测连续的 *k*-mer 等。DNABERT 模型是一种针对 DNA 语言设计的预训练模型，包含对 DNA 序列语义的学习信息，它给启动子预测带来了新的机遇。

本文将 DNABERT 模型进行预训练和微调，将得到的 DNA 嵌入作为输入编码，通过卷积神经网络（Convolutional Neural Network，CNN）、循环神经网络（Recurrent Neural Network，RNN）、长短期记忆（Long Short-Term Memory，LSTM）网络和 Transformer 等多种深度学习模型进行表示学习，以提高启动子预测的准确性和稳定性。实验结果表明，相比传统的 one-hot 编码，基于 DNABERT 编码的预测模型能够更好地捕获 DNA 序列中的局部特性和全局特征。与现有的其他方法相比，本文提出的方法 DNABERT-Prom 在多个评估指标上都取得了更好的结果。

1　实验方法及材料

1.1　数据集

在预训练过程中，本文使用了下载自 GENCODE 的人类基因组参考序列 GRCh38.p13，版本为 Release 43。GENCODE 是一个国际合作项目，旨在为人类基因组提供高质量的基因注释，包括蛋白质编码基因、非编码 RNA 基因、假基因等。GENCODE 的注释是 ENSEMBL 基因组的默认注释集。GRCh38.p13 是人类基因组参考序列的最新版本，由 Genome Reference Consortium 发布于 2019 年 2 月 28 日。它包含参考染色体、组装补丁和替代序列（单倍型）等序列区域，总长度为 3 088 269 832bp（base pair，碱基对），包含 20 368 个蛋白质编码基因和 21 306 个非编码 RNA 基因。对于预训练使用的人类基因组数据集，本文删除了所有序列间隙和未注释区域（带有 “N” 的序列区域），并将其用 *k*-mer 表示法进行分词。

微调与启动子预测的数据集来自 EPDnew 数据库[14]。EPDnew 是一个真核生物启动子数据库，收录了来自高通量实验（如 CAGE 和 Oligocapping）的转录起始位点的验证数据。在微调和后续的启动子预测中都需要使用启动子数据集。为了避免数据泄露，本文将启动子数据集均分为两部分，分别用于 DNABERT 的微调和启动子预测任务。数据库中共有 3 065 个人类 TATA 启动子序列和 26 533 个非 TATA 启动子序列，范围为 -5 000 ~ +5 000bp，其中+1 是转录起始位点的位置。在数据的处理上，使用从转录起始位点位置的 -249 ~ +50bp 中提取的 300bp 启动子序列作为正类。在负样本的构建上，对于 TATA 启动子序列，在序列中随机选择了 3 065 个 300bp 的基因组

区域，这些区域不在 – 249 ~ +50bp 范围内，但包含 TATA 基序。为了确保这些 TATA 序列尽可能与 TATA 启动子相似，将 TATA 基序定位于相对于实际 TATA 盒（TSS 的上游约 25bp）的相同位置。这样能够强制模型学习不太明显的特征，并仅通过开发对上下文的理解区分这些特征。对于非 TATA 启动子序列，由于它没有单一的区分特征，本文采用了 Oubounyt 等[15]提出的随机替换方法，以这种方式构建的数据集在保持数据生成质量和效率的同时，更具挑战性，使模型更难学习。

1.2　DNABERT-Prom 模型设计与测试

本文开发了 DNABERT-Prom，这是一个深度学习启动子预测工具。DNABERT-Prom 使用经过预训练和微调的 DNABERT 模型对 DNA 序列进行编码，然后将其作为输入，分别通过 CNN、RNN、LSTM 网络和 Transformer 等多种深度学习模型进行预测，使用预测性能最好的网络架构作为最终模型。基于 DNABERT 编码的启动子预测流程如图 1 所示。

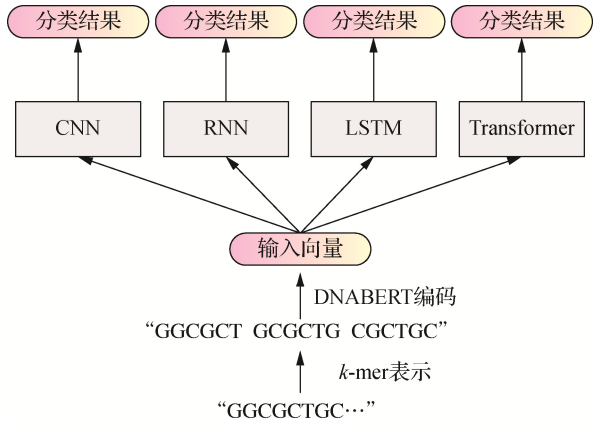

图 1　基于 DNABERT 编码的启动子预测流程

Fig.1　Promoter prediction process based on DNABERT embedding

首先，对 DNABERT 模型进行预训练和微调。在预训练阶段，由于一个标记可以从其周围的 *k*-mer 序列中推断出来，因此 DNABERT 模型屏蔽了连续的 *k*-mer 序列（总共约占输入序列的 15%）。DNABERT 模型使用两种方法从人类基因组中生成训练数据：直接非重叠分割和随机采样，序列长度为 5 ~ 510。与自然语言不同，*k*-mer 表示的独特语法引入了标记掩码问题，因此可以基于其前一个 *k*-mer 和后一个 *k*-mer 轻松地构造掩码的 *k*-mer。以 3-mer 序列{CAT，ATG，TGA，GAC，ACT}为例，“TGA”是“ATG”中的“TG”和“GAC”中的“A”的连接。如果独立地掩码 15% 的 *k*-mer，在大多数情况下，掩码的 *k*-mer 的前一个 *k*-mer 和后一个 *k*-mer 是未掩码的。这将显著简化预训练任务，并防止模型学习 DNA 序列的深层语义关系，因为模型可以轻松地从相邻的标记中推断出掩码标记。因此，本文不是独立地掩码每个 *k*-mer，而是掩码连续的 *k* 个 *k*-mer。对于 DNABERT 模型，将每个掩码标记的最后隐藏状态独立地馈送到分类层，并对整个词汇表执行分类任务。这里的类数等于词汇表中的标记数。在每个步骤中，计算所有掩码 *k*-mer 的交叉熵损失。DNABERT 预训练模型参数如表 1 所示。

表 1　DNABERT 预训练模型参数

Tab.1　Parameters of DNABERT pre-training model

参数名	参数值
max_seq_length	512
masked_lm_prob	0.15
learning_rate	le-4
weight_decay	0.01
batch_size	16
warmup_steps	10 000

在预训练过程中，DNABERT 模型使用了与 BERT 模型相同的架构，由 12 个 Transformer 层组成，每层有 768 个隐藏单元（Hidden Units）和 12 个注意力头（Attention Heads），并将混淆度作为预训练过程中的评估指标。在预训练期间，本文使用混合精度浮点运算在 Nvidia 3090 GPU 上进行训练，耗时约 14 小时。预训练过程中混淆度的变化如图 2 所示。

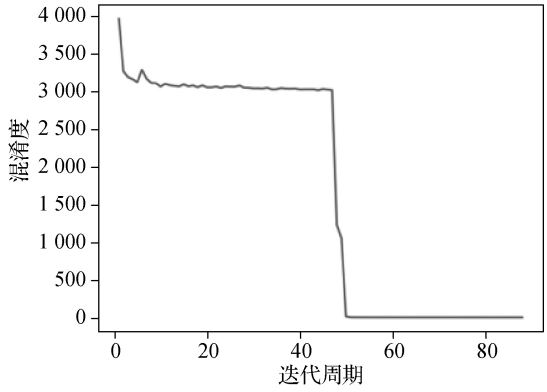

图 2　预训练过程中混淆度的变化

Fig.2　Change of perplexity in the process of pre-training

在微调过程中，为了避免数据泄露（防止用于测试的数据泄露到训练集中），使用 EPDnew 数据库中

一半的 TATA 启动子和非 TATA 启动子对 DNABERT 模型进行微调，另一半用于后续启动子预测。微调同样使用 *k*-mer 表示法对 DNA 序列进行分词，将每个碱基与其后面的碱基连接起来，形成一个长度为 *k* 的子串。在训练过程中，学习率首先线性升高到峰值，然后线性衰减到接近 0。实验中使用 Adam 作为优化器，并在输出层使用 Dropout 方法随机失活神经元。将训练数据分为训练集和开发集以进行超参数调整。对于不同 *k* 的 DNABERT 模型，略微调整了峰值学习率。详细的超参数设置如表 2 所示。在微调任务中，DNABERT 模型中的 *k*=3、*k*=4、*k*=5、*k*=6 实现了非常相似的性能，略有波动，*k*=6 的实验结果最佳，因此后续实验中使用 *k*=6 来实现编码任务。训练完成后，提取模型的最后一层——隐藏层的输出，便得到了 DNABERT- finetune 模型。对于输入模型的每个序列，都能获得一个长度为 768 的向量输出，这个向量输出便是 DNA 序列的编码表示，可以用于后续的启动子预测任务。

表 2　DNABERT 微调的超参数

Tab.2　Super parameters of DNABERT fine-tuning

k-mer	学习率	批次大小	随机失活	预热	最大步数
3	2.00E-04	64	0.1	0.06	2
4	3.00E-04	64	0.1	0.06	3
5	2.00E-04	64	0.1	0.06	3
6	2.00E-04	64	0.1	0.06	3

之后，使用 DNABERT 模型对 DNA 序列进行编码，然后将其作为输入，通过 CNN、RNN、LSTM 网络和 Transformer 等多种深度学习模型执行启动子预测任务。将启动子预测数据集随机打乱后按照 8 : 2 的比例划分为训练数据和测试数据。在训练中使用交叉熵损失函数。

$$L=\frac{1}{n}\sum_{i=1}^{n}\left(-y^{i}\log\hat{y}^{i}-(1-y^{i})\log(1-\hat{y}^{i})\right) \qquad (1)$$

使用 Adam 算法更新网络权重。在实验中，使用 Pytorch 对用于测试的深度学习模型进行了实现。

2　实验结果

2.1　评价指标

由于本实验中启动子预测任务属于平衡的二分类任务，即预测输入的 DNA 序列属于正样本（含启动子序列）还是负样本（不含启动子序列），因此在实验中使用准确率（ACC）、F1 分数和 Matthews 相关系数（MCC）作为评价指标。准确率（ACC）指的是所有预测中预测正确的比例。它的计算公式如下。

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}} \qquad (2)$$

式中，TP（真阳性）表示预测为正例且真实为正例的样本数；TN（真阴性）表示预测为反例且真实为反例的样本数；FN（假阴性）表示预测为反例但真实为正例的样本数；FP（假阳性）表示预测为正例但真实为反例的样本数。

F1 分数是一种用来衡量二分类模型精确度的指标，它兼顾了模型的精确率和召回率。F1 分数可以被看作模型精确率和召回率的一种调和平均，它的最大值是 1，最小值是 0。F1 分数的计算公式如下。

$$\mathrm{F1}=\frac{2\times\mathrm{precision}\times\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}} \qquad (3)$$

式中，precision（精确率）是指预测为正例且真实为正例的样本在所有预测为正例的样本中的比例；recall（召回率）是指预测为正例且真实为正例的样本在所有真实为正例的样本中的比例。

Matthews 相关系数是一种用于评价二分类模型性能的指标。Matthews 相关系数可以被看作观察到的二元分类和预测的二元分类之间的相关系数，或者是问题及其对偶的回归系数的几何平均数。Matthews 相关系数可以由混淆矩阵直接计算出来，公式如下。

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \qquad (4)$$

2.2　不同模型的性能比较

首先对不同模型在基于 DNABERT 编码的启动子预测任务中的性能进行比较。数据集来自最新版本的真核生物启动子数据库 EPDnew 中的人类 TATA 启动子和非 TATA 启动子数据集。将数据集输入经过预训练和微调的 DNABERT-finetune 模型中，得到编码后的 DNA 序列。以此为输入，对不同模型的预测能力进行比较。为了消除随机误差，对每个模型都进行 5 次独立重复的实验，实验结果如图 3 所示。在 4 种模型中，Transformer 取得了最好的效果，在含有 TATA 启动子和非 TATA 启动子的混合数据集上的准确度、F1 分数和 Matthews 相关系数分别达到 0.975、0.977 和 0.967。实验表明，基于 DNABERT 编码的深度学习模型在启动子预测任务中取得了良好的效果，超过了仅

使用 DNABERT 进行启动子预测的结果[13]。在实验使用的几个模型中，Transformer 取得了最好的效果，表明 Transformer 可以利用自注意力机制提升对局部上下文信息的关注度，降低异常点的影响。

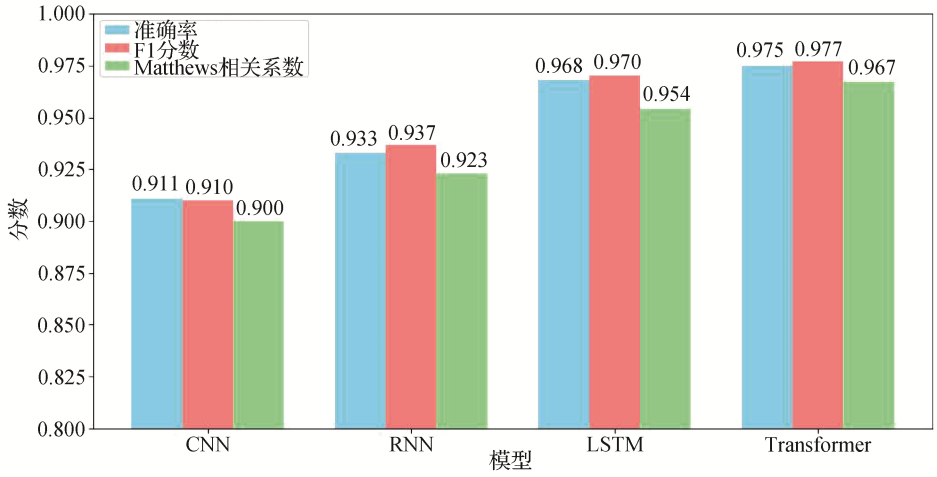

图 3　不同深度学习模型的性能比较

Fig.3　Performance comparison of different Deep Learning models

2.3　预训练与微调对 DNABERT 编码的影响

在 DNABERT 模型的训练中，使用了预训练和微调两种技术来提高 DNABERT 模型的性能和泛化能力。预训练是指在大规模数据上进行无监督学习，让模型学习到 DNA 数据的潜在结构和规律，从而提高模型的泛化能力。微调是指在特定任务上对预训练模型进行有监督学习，通过调整模型参数来适应特定任务的需求。微调可以让模型更好地适应启动子预测任务的数据和特征，提高模型在该任务上的性能。本文分别使用仅在人类基因组上进行预训练的 DNABERT-pretrain 模型和预训练后又在启动子训练数据集上进行微调的 DNABERT-finetune 模型，对启动子预测任务的 DNA 序列进行编码，再将编码得到的结果分别输入 Transformer 模型来得到预测结果。为了验证两种编码方式的有效性，实验中还使用了传统的 one-hot 编码方式进行启动子预测，实验结果如图 4 所示。经过微调的 DNABERT-finetune 编码方式在混合启动子测试集上取得了最好的性能，远超过 one-hot 编码方式和仅经过预训练的 DNABERT-pretrain 编码方式。这说明经过微调的 DNABERT 编码方式有效地学习了对启动子预测任务有利的重要特征，有效提升了在下游问题中的预测表现。

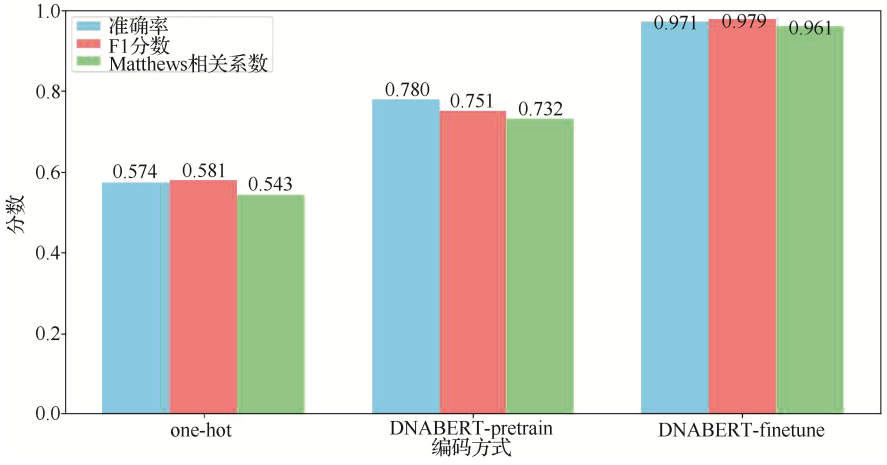

图 4　不同编码方式的性能比较

Fig.4　Performance comparison of different embeddings

2.4　与同类工具的比较

将本文提出的基于 DNABERT 编码的启动子预测工具 DNABERT-Prom 与现有的几种启动子预测工具进行对比，包括 DeePromoter[15]、CNNProm[8] 和 iPro-WAEL[9]，实验结果如图 5 所示。在实验中，DNABERT-Prom 使用了在人类基因组上进行预训练后又在启动子数据集上进行微调的 DNABERT-finetune 模型，对输入 DNA 序列进行编码，并用 Transformer 模型进行预测。对于其他几种模型，使用各自文献中提出的结果最优的参数进行实验。DNABERT-Prom 在启动子预测任务中的准确率和 F1 分数超过 DeePromoter、CNNProm 和 iPro-WAEL。在 Matthews 相关系数上，DNABERT-Prom 取得了与 iPro-WAEL 相近的分数，并超过 DeePromoter 和 CNNProm。

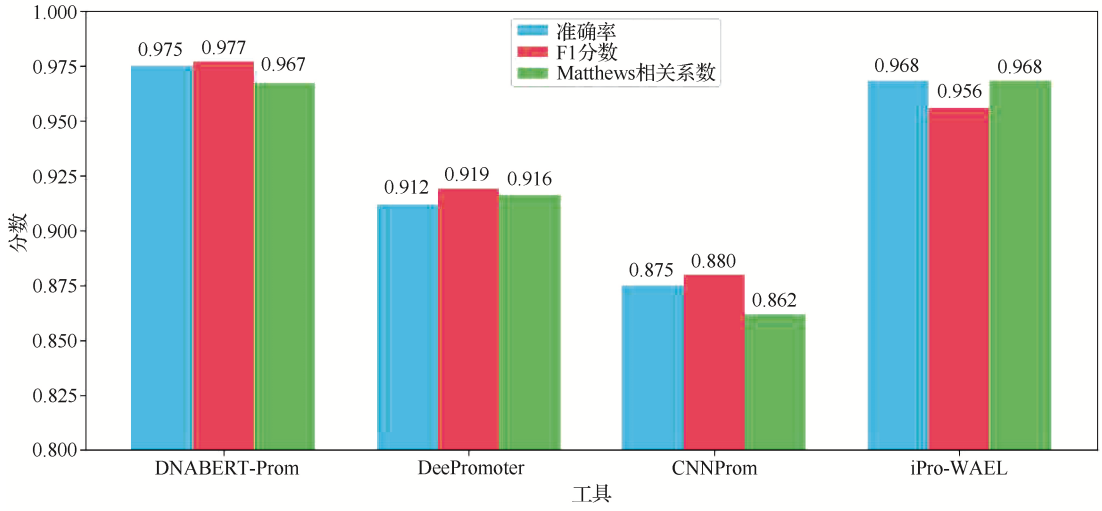

图 5　不同工具在人类启动子预测任务上的性能比较

Fig.5　Performance comparison of different tools in Human Promoter Prediction Task

3　结论

本文主要介绍了使用 DNABERT 对输入数据进行编码，并在多种深度学习模型上进行启动子预测的实验结果。首先，本文使用了 3 种不同的评价指标来评估模型的性能，包括准确率、F1 分数和 Matthews 相关系数。实验结果表明，在含有 TATA 启动子和非 TATA 启动子的预测任务中，使用 DNABERT 编码并在 Transformer 上进行预测的模型在所有指标上都表现出了最好的性能。然后，本文对不同的编码方式进行了比较，包括仅在人类基因组上进行预训练的 DNABERT-pretrain 编码方式和预训练后又在启动子数据集上进行微调的 DNABERT-finetune 编码方式，以及传统的 one-hot 编码方式。在测试集中，使用 DNABERT-finetune 编码方式进行编码的表现远优于其他两种编码方式。最后，本文比较了不同工具的性能差异。DNABERT-Prom 在启动子预测任务中的准确率和 F1 分数超过了 DeePromoter、CNN-Prom 和 iPro-WAEL。在 Matthews 相关系数上，DNABERT-Prom 取得了与 iPro-WAEL 相近的分数，并超过 DeePromoter 和 CNNProm。

本文所提的基于 DNABERT 编码的启动子预测模型 DNABERT-Prom 在预测性能上超过了大多数现有工具，一个可能的解释是 DNABERT 可以将 DNA 序列转换为一个低维的向量，这个向量可以捕捉 DNA 序列中的语义和结构信息，从而方便后续的分析和预测。另外，DNABERT 可以利用大量无标注的 DNA 数据进行预训练，从而学习到 DNA 语言的特征和规律，这些知识可以帮助提高其在下游任务中的预测性能。在本文测试的深度学习模型中，Transformer 取得了最好的效果，原因可能是 Transformer 使用自注意力机制，可以捕捉序列数据中任意位置之间的关系，并且可以并行处理所有输入，这使 Transformer 模型在 DNA 序列数据的处理上取得了很好的效果。

综上所述，本文的研究意义在于提出了一种新的启动子预测方式，将预训练模型 DNABERT 的嵌入输出作为启动子预测的输入编码，并通过实验验证了

DNABERT 语义表示的优越性。同时，本文对比了不同深度学习模型的选择对预测结果的影响。实验结果表明，使用 DNABERT 编码和深度学习模型进行启动子预测的方法具有较高的准确性和稳定性，说明包含 DNA 序列语义的预训练嵌入可以作为下游任务的可靠输入，可以为生物学研究提供有力的支持。同时，本文使用了包括 DNABERT 编码和多种深度学习模型在内的新技术，为深入理解 DNA 序列特征和生物学机制提供了新思路和新方法，具有重要的理论和实践意义。

参考文献

[1] VO N L, WANG Y L, KASSAVETIS G A, *et al*. The punctilious RNA polymerase II core promoter[J]. **Genes Dev**, 2017, 31(13): 1289-1301. DOI: 10.1101/gad.303149.117.

[2] ROY A L, SINGER D S. Core promoters in transcription: old problem, new insights[J]. **Trends Biochem Sci**, 2015, 40(3): 165-171. DOI: 10.1016/j.tibs.2015.01.007.

[3] SANDELIN A, CARNINCI P, LENHARD B, *et al*. Mammalian RNA polymerase II core promoters: insights from genome-wide studies[J]. **Nat Rev Genet**, 2007, 8(6): 424-436. DOI: 10.1038/nrg2026.

[4] YELLA V R, BANSAL M. DNA structural features of eukaryotic TATA-containing and TATA-less promoters[J]. **FEBS Open Bio**, 2017, 7(3): 324-334. DOI: 10.1002/2211-5463.12166.

[5] UMAROV R, KUWAHARA H, 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.

[6] 宋世龙，赵雨桐，王茜，等. 一种基于胶囊网络外膜蛋白拓扑结构预测方法[J]. **生物医学工程学进展**，2022，43（4）：207-218.
SONG Shilong, ZHAO Yutong, WANG Qian, *et al*. A method for predicting the topological structure of outer membrane proteins based on capsule network[J]. **Progress in Biomedical Engineering**, 2022, 43(4): 207-218.

[7] 宗西增，蔡蕊蕊，田若婷，等. 基于多层感知机的 DNA 甲基化年龄预测模型[J]. **生物医学工程学进展**，2023，44（1）：34-41.
ZONG Xizeng, CAI Ruirui, TIAN Ruoting, *et al*. DNA methylation age prediction model based on multilayer perceptron[J]. **Progress in Biomedical Engineering**, 2023, 44(1): 34-41.

[8] 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. DOI: 10.1371/journal.pone.0171410.

[9] ZHANG P, ZHANG H, WU H. iPro-WAEL: a comprehensive and robust framework for identifying promoters in multiple species[J]. **Nucleic Acids Res**, 2022, 50(18): 10278-10289. DOI: 10.1093/nar/gkac824.

[10] PRESTRIDGE D S. Predicting Pol II promoter sequences using transcription factor binding sites[J]. **J Mol Biol**, 1995, 249(5): 923-932. DOI: 10.1006/jmbi.1995.0349.

[11] KNUDSEN S. Promoter 2.0: for the recognition of Pol II promoter sequences[J]. **Bioinformatics**, 1999, 15(5): 356-361. DOI: 10.1093/bioinformatics/15.5.356.

[12] ANWAR F, BAKER S M, JABID T, *et al*. Pol II promoter prediction using characteristic 4-mer motifs: a machine learning approach[J]. **BMC Bioinformatics**, 2008, 9: 414. DOI: 10.1186/1471-2105-9-414.

[13] JI Y, ZHOU Z, LIU H, *et al*. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome[J]. **PloS One**, 2021, 37(15): 2112-2120. DOI: 10.1093/bioinformatics/btab083.

[14] DREOS R, AMBROSINI G, CAVIN P R, *et al*. EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era[J]. **Nucleic Acids Res**, 2013, 41(Database issue): D157-164. DOI: 10.1093/nar/gks1233.

[15] OUBOUNYT M, LOUADI Z, TAYARA H, *et al*. DeePromoter: robust promoter predictor using deep learning[J]. **Front Genet**, 2019, 10: 286. DOI: 10.3389/fgene.2019.00286.