基因组学与应用生物学, 2023 年, 第 42 卷, 第 12 期, 第 1344-1352 页

研究论文
Research Article

基于 Transformer 编码器和 Nanopore 数据的 DNA 5-甲基胞嘧啶位点预测

曾佳[1,2]　陈玲玲[1*]

1 亚热带农业生物资源保护与利用国家重点实验室, 广西大学生命科学与技术学院, 南宁, 530004; 2 广西大学计算机与电子信息学院, 南宁, 530004;

*通信作者, llchen@gxu.edu.cn

摘　要　DNA 中的 5-甲基胞嘧啶(5-methylcytosine, 5mC)是通过 DNA 甲基转移酶在胞嘧啶环第 5 个碳原子上共价结合一个甲基产生的, 广泛存在于不同组织中, 在各种生物过程中发挥着重要作用。通过甲基化位点对相应的甲基化修饰进行研究是一种常用手段, 因此, 5mC 位点的准确鉴定对深入理解其生物学功能至关重要。随着人工智能的飞速发展, 深度学习已经成为了生物信息学的重要分析工具, 越来越多的生物学问题通过深度学习得到解决。Transformer 是一种基于注意力机制的深度学习模型, 本研究基于第三代基因测序技术 Nanopore 测序数据进行特征提取, 通过 Transformer 编码器对特征进行编码, 最后输入到双向长短期记忆网络(long short-term memory, LSTM)中以预测 5mC 位点。使用拟南芥(*Arabidopsis thaliana*)和水稻(*Oryza sativa*)对模型进行训练和测试, 结果表明, 本模型能够有效提取 5mC 位点的潜在特征, 从而提高 5mC 位点的预测能力。

关键词　5-甲基胞嘧啶; 深度学习; Nanopore 测序; Transformer 编码器; 双向长短期记忆网络

DNA 5-methylcytosine Site Prediction Based on Transformer Encoder and Nanopore Data

ZENG Jia [1,2]　CHEN Lingling [1*]

1 State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, College of Life Science and Technology, Guangxi University, Nanning, 530004; 2 College of Computer and Electronic Information, Guangxi University, Nanning, 530004

*Corresponding author, llchen@gxu.edu.cn

DOI: 10.13417/j.gab.042.001344

Abstract　5-methylcytosine (5mC) in DNA is produced by covalently binding a methyl group on the fifth carbon atom of the cytosine ring by DNA methyltransferase, and it is widely present in different tissues, and plays an important role in various biological processes. It is a common method to study the corresponding methylation modification through the methylation site. Therefore, the accurate identification of the 5mC site is crucial for a deep understanding of its biological function. With the rapid development of artificial intelligence, deep learning has become an important analytical tool in bioinformatics, and more and more biological problems have been solved through deep learning. Transformer is a deep learning model based on the attention mechanism. In this paper, features are extracted based on the third-generation gene sequencing technology Nanopore sequencing data, and then the features are encoded by the Transformer encoder, and finally input into the bidirectional long short-term memory (LSTM) to predict the 5mC site. We trained and tested the model using Nanopore sequencing data from *Arabidopsis thaliana* and *Oryza sativa*, and the results showed that our model can effectively extract the latent features of 5mC site, thereby improving the predictive power of 5mC site.

Keywords　5-methylcytosine; Deep learning; Nanopore sequencing; Transformer encoder; Bidirectional long short-term memory

DNA 甲基化是高等生物中一种保守的表观遗传　修饰方式, 能够在不改变 DNA 分子一级结构的前提

基金项目: 本研究由国家自然科学基金项目(32270712)资助。

引用格式: 曾佳, 陈玲玲, 2023. 基于 Transformer 编码器和 Nanopore 数据的 DNA 5-甲基胞嘧啶位点预测. 基因组学与应用生物学, 42(12): 1344-1352. [ZENG J, CHEN L L, 2023. DNA 5-methylcytosine site prediction based on transformer encoder and Nanopore data. Genomics and Applied Biology, 42(12): 1344-1352.]

下调节基因组的功能，在基因调控和基因组稳定性中发挥着重要作用（袁超 等，2020）。在真核生物中，5-甲基胞嘧啶（5-methylcytosine，5mC）是 DNA 甲基化的主要存在形式，5mC 可以分为 CpG、CHG、CHH 三种类型（H 代表碱基 A、C、T）。在动物中，大部分的胞嘧啶甲基化发生在 CpG 位点，而在植物中，CpG、CHG、CHH 位点均可发生广泛的甲基化（孙颖 等，2011），三种类型的 5mC 在生物过程的调节中发挥着不同的作用（Zhang et al.，2018；Domb et al.，2020）。

目前，5mC 甲基化检测的主要方法是亚硫酸氢盐测序（Miura et al.，2012），其原理是利用亚硫酸氢盐将基因组中未发生甲基化的 C 碱基转换成 U 碱基，与原来具有甲基化修饰的 C 碱基区分开。然而，亚硫酸氢盐处理的过程会导致大量 DNA 断裂，使高度可变、异质表观的基因组的分析变得复杂（Simpson et al.，2017），并且无法分析重复的基因组区域。第三代 Nanopore 测序技术可以直接对 DNA 测序而无需转换或进行 PCR 扩增，为检测 DNA 中的碱基修饰提供了新的机会（Davis et al.，2013；Xu and Seki，2020）。研究表明，Nanopore 测序中的电信号对 DNA 中的碱基修饰敏感（Laszlo et al.，2013；Schatz，2017），通过将甲基化 5mC 位点的原始电信号与已知序列的相同非甲基化位点的电信号进行对比，可以更高精度地测量特定基因组位置的 DNA 甲基化（Schreiber et al.，2013；Liu et al.，2019a）。除此之外，已有多种基于机器学习的算法使用 Nanopore 测序数据进行 DNA 甲基化检测。Simpson 等（2017）开发了一种基于隐马尔科夫模型（hidden Markov model，HMM）的算法，该方法可以从 Nanopore reads 中检测大肠杆菌（*Escherichia coli*）和人（*Homo sapiens*）CpG 中的 5mC 位点。Rand 等（2017）使用分层 Dirichlet 过程的 HMM 分析了大肠杆菌中三种类型的胞嘧啶（C、5mC、5hmC），以及不同阶段的 N6-甲基腺嘌呤（N6-methyladenosine，6mA），并将该算法集成在 signalAlign 包。Liu 等（2019a）的 DeepMod 和 Ni 等（2021）的 Deepsignal-plant 都用到了具有长短期记忆的双向递归神经网络（bidirectional long short-term memory，Bilstm）检测 DNA 修饰（Hochreiter and Schmidhuber，1997）。其中，DeepMod 使用 Nanopore 测序对人类基因组 HX1 和莱茵衣藻（*Chlamydomonas reinhardtii*）基因组进行测序，然后对三个物种的基因组（大肠杆菌、莱茵衣藻和人）进行评估，对于 5mC 位点，DeepMod 对人工产生的和原始修饰的平均精度都达到了 99%。DeepSignal-plant 可以从植物的原始 Nanopore reads 中准确检测所有类型的 5mC，并且为训练模型开发了一个去噪过程，使得 DeepSignal-plant 能够在所有情况下与亚硫酸氢盐测序实现 5mC 检测的高相关性。

然而，之前的研究并没有充分利用到 Nanaopore 测序数据的序列和电信号特征，并且存在提取特征过程复杂、没有提取到位点的关键信息、模型预测功能不够强大等问题，通过改进这些方面有可能实现更好的预测性能。本研究开发了一种从 Nanopore reads 中预测 5mC 所有类型甲基化位点的深度学习方法 Nanoformer，为了尽可能保留 5mC 位点的原始信息，直接将 Nanopore reads 的原始电信号和 DNA 序列作为输入，通过 Transformer 编码器编码（Vaswani et al.，2017），然后使用 Bilstm 进行解码，最终输入到全连接神经网络中以预测甲基化状态。使用拟南芥（*Arabidopsis thaliana*）和水稻（*Oryza sativa*）基因组数据对 Nanoformer 进行了评估，最终的结果表明其在 DNA 5mC 位点预测中表现良好。

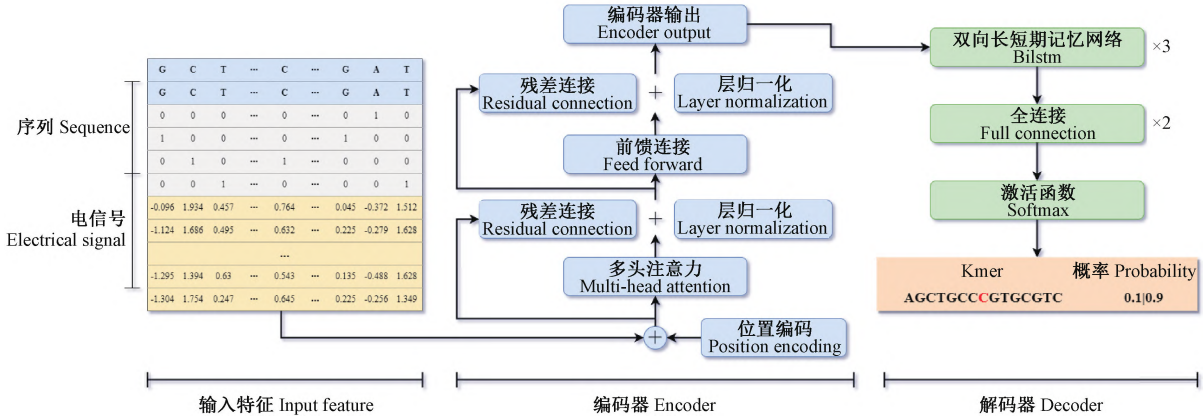
1 结果

1.1 模型架构

Transformer 是 Google 在 2017 年 6 月提出的一种基于自注意力机制的神经网络，在自然语言处理（natural language processing，NLP）领域的多个任务上都取得了非常好的效果。相比于其他模型，它可以高效地并行化运行，比循环神经网络更能有效地处理序列之间的长期依赖关系。鉴于 Transformer 在 NLP 领域的成功以及 DNA 序列和自然语言的相似性，本研究将 Transformer 编码器用于 DNA 5mC 位点预测，Transformer 的注意力机制能够提取到序列中更关键的信息，并减少对其他信息的关注，从而可能获得更好的预测性能。

本研究构建了一个基于 Transformer 编码器和 Bilstm 的深度学习框架，命名为 Nanoformer。如图 1 所示，模型的输入特征是每个 motif 位点的碱基序列和对应电信号结合而成的特征矩阵。输入维度是[batch，k_len，f_nums]，其中，batch 表示每一次输入样本的数量，k_len 表示 kmer 长度，f_nums 表示每个碱基的特征个数，默认为[256，41，12]。因为 Transformer 编码器只考虑了全局信息，不具备位置信

息，所以在将数据输入到模型之前，需要在输入特征　中加入位置编码（position encoding）。



Nanofrmer 分为 3 个部分，Input feature 由序列特征和电信号特征组合而成；Encoder 为 Transformer 编码器，层数为 1；Decoder 由 3 层双向 LSTM 和 2 层全连接神经网络组成。

Nanofrmer can be divided into three parts, and input feature is composed of sequence features and electrical signal features; Encoder is Transformer encoder with 1 layer; Decoder is composed of 3 layers of bidirectional LSTM and 2 layers of fully connected neural network.

图 1　Nanoformer 结构图

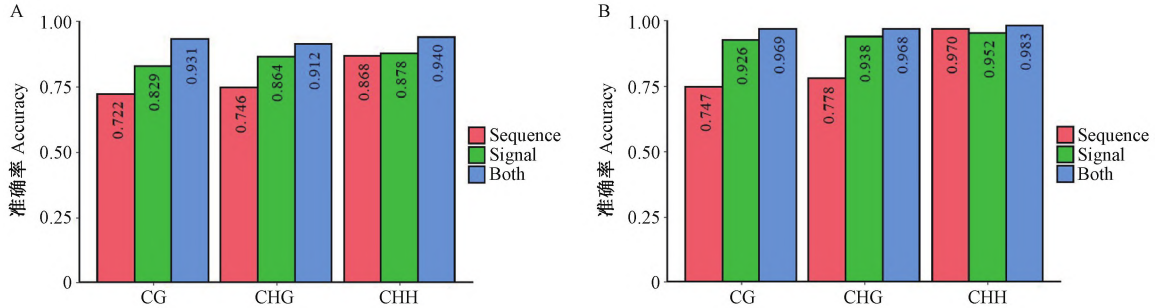
Figure 1　The structure chart of Nanoformer

将进行位置编码后的数据输入到 Transformer 编码层中，Transformer 编码层是由 n 个相同的块结构堆叠而成（本模型 $n=1$），每个块结构又进一步分为两个子层。其中，第一层是一个多头自注意力层（muti-head attention），可以将输入信息映射到多个特征子空间，有助于模型捕捉来自不同子空间的特征关系，提取到不同信息之间更多的潜在关联；第二层是一个全连接前馈网络层（feed forward），由两个具有 Relu 激活函数的全连接层组成。为了防止梯度爆炸和消失，每一个子层之后都进行了残差连接（residual connection）和层归一化（layer normalization）。

经过 Transformer 编码层后，再将编码后的数据输入到 Bilstm 中进行解码。Bilstm 能够提取 DNA 序列两个方向的信息，并根据长短期记忆算法将正向输出与反向输出相结合以获得融合特征。最后经由两层全连接网络和一个 softmax 激活函数输出该位点为 5mC 位点的概率。

1.2　不同编码方式性能比较

为了验证 Nanoformer 编码方式的有效性，本研究将其分别与只用序列作为特征编码和只用电信号作为特征编码的模型性能进行了比较，如图 2 所示，横坐标是 5mC 类型，纵坐标是预测准确率。可以看到，在拟南芥和水稻的数据集上，使用 Sequence+Signal 编码的模型准确率都是最高的，分别达到了 0.940 和 0.983，其次是 Signal 编码，表现最差的是 Sequence 编码。不过在 CHH 类型的 5mC 位点上，Sequence 编码的模型性能要优于 Signal 编码并且接近于 Sequence+Signal 编码。这个结果说明无论是 Sequence 编码还是 Signal 编码，都缺失了各自的一些特征，导致模型预测性能不稳定，而 Signal+Sequence 的编码方式综合了 Sequence 和 Signal 的特征，这种编码方式能够将单一编码方式的特征信息互补，从而显著提高模型预测所有类型 5mC 位点的性能。



Sequence 表示只用序列特征进行编码，Signal 表示只用电信号特征进行编码，Both 表示将两个特征编码组合。

Sequence indicates that only sequence features are used for encoding, Signal indicates that only electrical signal features

are used for encoding, and Both indicates that two feature encodings are combined.

图 2　不同编码方式的准确率比较

（A）拟南芥中的准确率比较；（B）水稻中的准确率比较

Figure 2　Accuracy comparison of different encoding methods

（A）Accuracy comparison in *Arabidopsis thaliana*；（B）Accuracy comparison in *Oryza sativa*

1.3　不同深度学习模型组合性能比较

为了证明 Nanoformer 所用深度学习模型的优势，本研究对不同编码器和解码器的组合性能进行了比较，如图 3 所示，Transformer 和 Bilstm 的三种模型组合在拟南芥和水稻数据集上所有类型 5mC 位点预测的受试者工作特征（receiver operating characteristic，ROC）曲线，可以看到，无论是拟南芥还是水稻，Nanoformer 均具有最高的 ROC 曲线下面积（area under curve，AUC）值，在 CpG 位点，其 AUC 分别达到 0.975 和 0.995。值得注意的是，使用 Bilstm 模型的性能也不低，AUC 值分别达到了 0.951 和 0.988，但是在 Bilstm 的基础上加入了 Transformer 编码器编码后，AUC 均有所提高，这个结果表明，Transformer 编码器能够有效提取到 5mC 位点的潜在特征，从而帮助模型提高最终的预测性能。

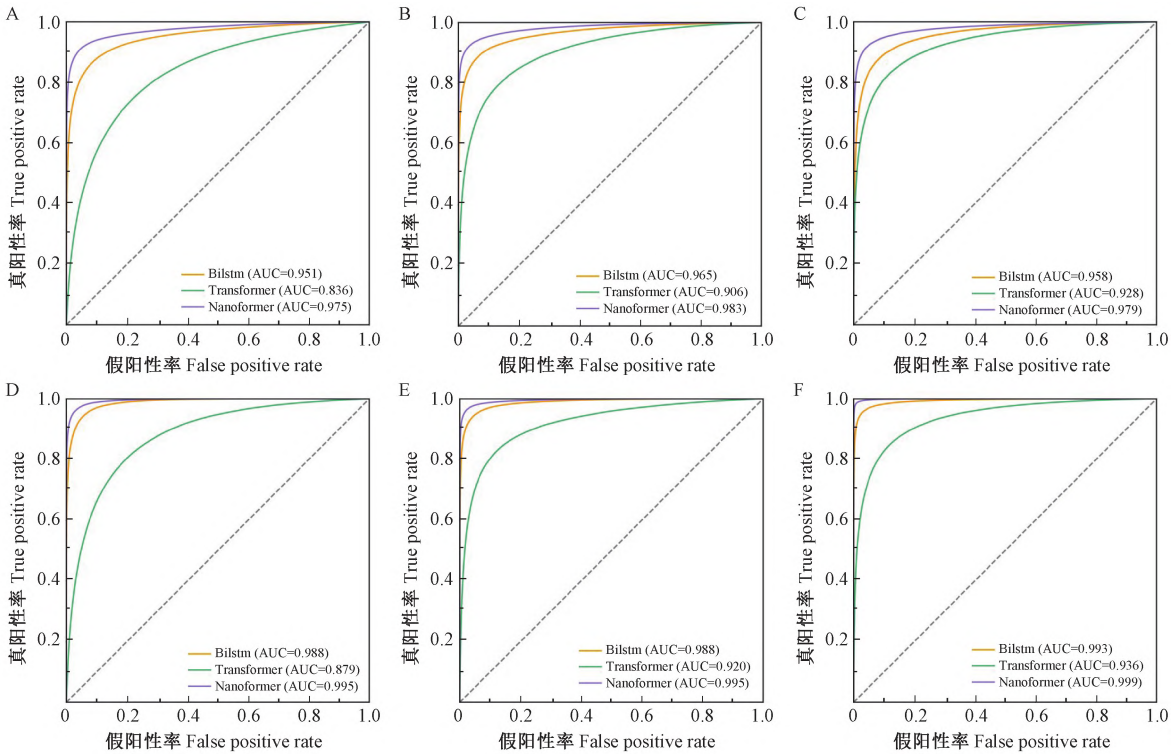


图 3　不同深度学习组合方法 ROC 曲线比较

（A）拟南芥 CpG 位点；（B）拟南芥 CHG 位点；（C）拟南芥 CHH 位点；（D）水稻 CpG 位点；（E）水稻 CHG 位点；（F）水稻 CHH 位点

Figure 3　ROC curve comparison of different deep learning combination methods

（A）CpG sites in *Arabidopsis thaliana*；（B）CHG sites in *A. thaliana*；（C）CHH sites in *A. thaliana*；（D）CpG sites in *Oryza sativa*；（E）CHG sites in *O. sativa*；（F）CHH sites in *O. sativa*

1.4　与现有方法的性能比较

本研究对 Nanoformer 和 Deepsignal-plant 进行了比较。Deepsignal-plant 是目前为止最新发表的，也是表现最好的基于 Nanopore 数据检测基因组 5mC 位点的深度学习工具。为了比较的公平性，使用和训练 Nanoformer 相同的数据重新训练了 Deepsignal-plant，其准确率、精确率、召回率、F_1 值、AUC 等 5 个性能指标对比如表 1 所示，Nanoformer 在拟南芥和水稻的三种类型 5mC 位点预测中性能都要明显优于 Deepsignal-plant 的，其准确率、精准率、召回率、

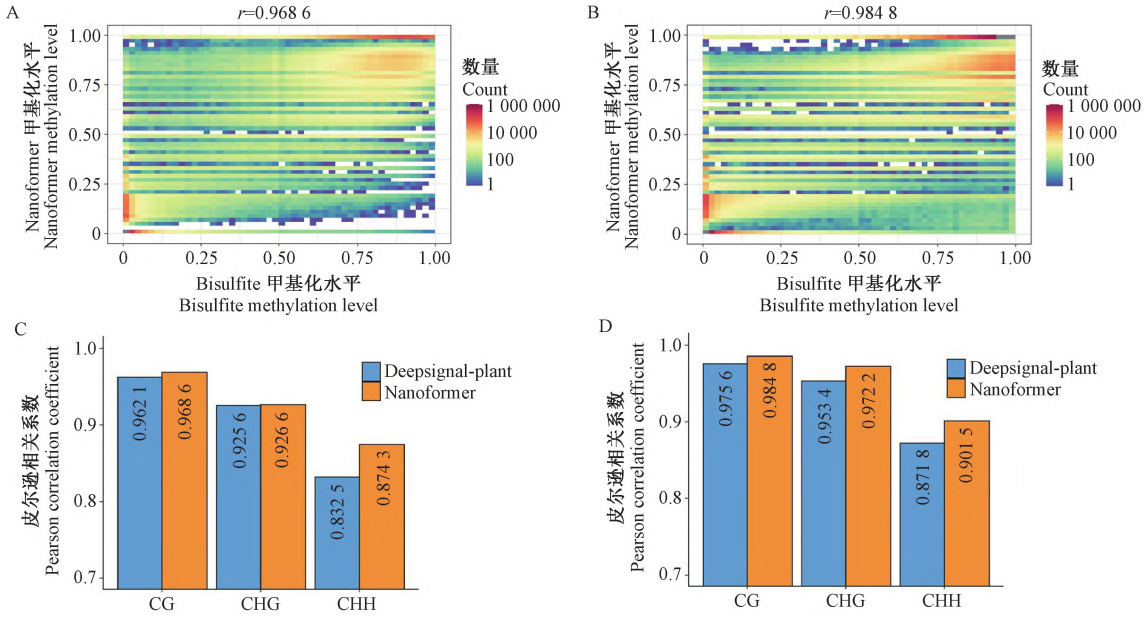
F_1 值、AUC 在拟南芥中分别平均提升了 4.53%、6.07%、3.35%、4.79%、2.78%，在水稻中分别平均提升了 3.20%、4.15%、2.34%、3.26%、1.25%。

表 1　Nanoformer 和 Deepsingal-plant 性能对比

Table 1　Performance comparison between Nanoformer and Deepsingal-plant

物种 Species	基序 Motif	方法 Methods	准确率 Accuracy	精准率 Precision	召回率 Recall	F_1 值 F_1-score	ROC 曲线下面积 AUC
拟南芥 *Arabidopsis thaliana*	CpG	Deepsingal-plant	0.889 9	0.855 2	0.918 9	0.885 9	0.945 4
		Nanoformer	0.930 4	0.911 6	0.947 2	0.929 1	0.973 4
	CHG	Deepsingal-plant	0.898 8	0.857 1	0.935 1	0.894 4	0.957 2
		Nanoformer	0.943 5	0.925 2	0.960 3	0.942 4	0.982 8
	CHH	Deepsingal-plant	0.886 5	0.857 8	0.910 0	0.883 1	0.949 0
		Nanoformer	0.937 1	0.915 3	0.957 0	0.935 7	0.978 9
水稻 *Oryza sativa*	CpG	Deepsingal-plant	0.937 9	0.932 2	0.943 0	0.937 5	0.983 2
		Nanoformer	0.971 6	0.967 1	0.975 9	0.971 5	0.996 0
	CHG	Deepsingal-plant	0.941 0	0.921 3	0.959 0	0.939 8	0.980 5
		Nanoformer	0.969 6	0.964 0	0.974 9	0.969 4	0.994 7
	CHH	Deepsingal-plant	0.954 8	0.941 6	0.967 2	0.954 2	0.988 7
		Nanoformer	0.988 5	0.988 5	0.988 6	0.988 5	0.999 1

为了进一步证明 Nanoformer 的优势，分别使用 9×的拟南芥和 6×的水稻 Nanopore reads 在 Nanoformer 和 Deepsingnal-plant 上进行 5mC 甲基化位点预测，计算和 BS-seq 甲基化水平的皮尔逊相关性，结果如图 4 所示，Nanoformer 在三种 5mC 甲基化位点预测的甲基化水平皮尔逊相关性均要高于 Deepsingnal-plant 的，和表 1 的性能对比结果一致。



r 为皮尔逊相关系数。

r is Pearson correlation coefficient.

图 4　Nanoformer 与 Deepsignal-plant 预测 5mC 位点甲基化水平皮尔逊相关性对比

（A）Nanoformer 预测拟南芥 CpG 5mC 位点；（B）Nanoformer 预测水稻 CpG 5mC 位点；（C）拟南芥甲基化水平比较；（D）水稻甲基化水平比较

Figure 4　Comparison of Pearson correlation between Nanoformer and Deepsignal-plant to predict the methylation level of 5mC sites

（A）Nanoformer prediction of *Arabidopsis thaliana* CpG 5mC sites；（B）Nanoformer prediction of *Oryza sativa* CpG 5mC sites；（C）Comparison of methylation levels in *Arabidopsis thaliana*；（D）Comparison of methylation levels in *Oryza sativa*

1.5　跨物种性能

为了进一步评估 Nanoformer 的跨物种性能，使用拟南芥和水稻数据对模型进行了交叉测试。测试结果如图 5 所示，三种类型 5mC 位点的 AUC 值在使用了不同物种的数据进行测试后均有一定程度的下降，不过下降程度有限。其中，使用水稻样本训练的模型来预测拟南芥 5mC 位点的 AUC 值均在 0.939 以上，使用拟南芥样本训练的模型来预测水稻 5mC 位点的 AUC 值均在 0.976 以上。这个结果说明 Nanoformer 具有一定跨物种检测 5mC 位点的能力。

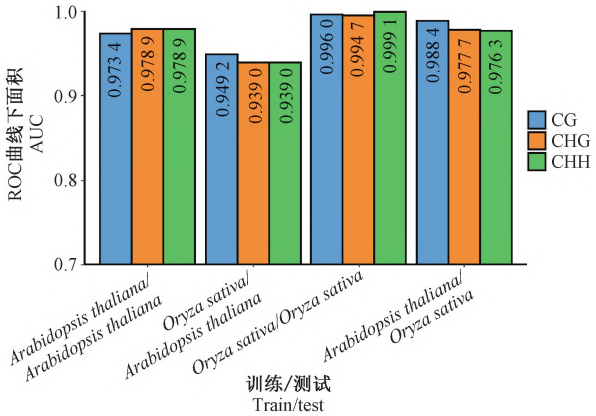


图 5　Nanoformer 跨物种性能测试

Figure 5　Cross-species performance testing of Nanoformer

2　讨论与结论

DNA 5mC 甲基化是一种重要的表观遗传修饰方式，在许多生物的生长和发育过程中扮演着关键的角色。随着机器学习和深度学习等技术的不断发展，近年来已经涌现出了许多基于这些技术的新方法并应用于预测和检测 DNA 甲基化。其中，Nanopore 长读测序技术的出现为直接从信号数据中大规模、低成本地检测 DNA 修饰提供了宝贵的机会（Liu et al., 2019b），这种测序技术可以直接读取单个 DNA 分子的序列，能够区分不同的碱基和 DNA 修饰，包括 5mC 甲基化。

基于这种技术，本研究提出了一种新的深度学习方法 Nanoformer 来进行 DNA 5mC 位点预测。Nanoformer 将 Nanopore 测序数据的原始序列和电信号结合后作为特征输入，这种融合特征相比于单独只用序列或者电信号来说，能够包含更多的 5mC 位点潜在信息，对模型性能有明显提升。此外，Nanoformer 使用 Transformer 编码器进行特征编码，传统的循环神经网络在处理长序列时容易出现梯度消失或爆炸的问题，而 Transformer 编码器通过自注意力（self-attention）机制可以学习到序列中的长期依赖关系，从而提高序列预测的准确性。最后，Nanoformer 使用 Bilstm 进行 5mC 位点的最终分类，Bilstm 具有向前和向后两个方向的处理能力，这意味着它能够利用前后文的信息进行推理和预测，这对于包含 5mC 位点的 DNA 序列来说，使用 Bilstm 非常合适。

本研究使用了拟南芥和水稻的 Nanopore 测序数据对三种类型的 5mC 位点进行了测试，结果表明 Nanoformer 能够准确预测拟南芥和水稻中的 5mC 位点和非 5mC 位点，比现有的基于机器学习的方法取得了更好的性能。此外，对这两个物种的交叉验证表明，Nanoformer 具有跨物种预测的能力。因此，Nanoformer 可以用于全基因组表观遗传学分析，并在未来进行更多物种和修饰类型的预测。

3　材料与方法

3.1　数据集

本文使用了 Ni 等（2021）已发表的拟南芥和水稻的 Nanopore 测序数据和对应的亚硫酸氢盐（Bisulfite）测序数据，该数据可在 NCBI 网站（http://www.ncbi.nlm.nih.gov/）的生物项目 PRJNA764549 和 SRA 数据库的 SRP337810 下载。

Bisulfite 测序是 DNA 甲基化位点检测的金标准，本研究使用 Bismark v0.23.1（Krueger and Andrews, 2011）软件基于 Bisulfite 测序数据，提取基因组中高置信度的甲基化和非甲基化位点。使用 Guppy v4.0.11 对 Nanopore 测序的 fast5 文件进行碱基识别 basecalling，目的是将原始电信号转化为碱基序列，然后使用 Tombo v1.5.1（Stoiber et al., 2016）将原始电信号映射到参考基因组中。

本研究分别从拟南芥和水稻的 Nanopore 测序数据中随机选择了约 9×和 6×的 Nanopore reads。表 2 为从水稻和拟南芥 reads 中提取到的基序位点数量信息。可以发现，在拟南芥和水稻中，高置信度非甲基化胞嘧啶的数量远远高于甲基化胞嘧啶的数量，尤其是在 CHG 和 CHH 类型的基序中。对于每种类型的基序，随机选择出一百万条 kmer 作为模型训练和测试的样本，不足一百万的按最大数量选择。在实验中，使用 5 折交叉验证来评估 Nanoformer 性能，采用 6∶2∶2的比例来划分训练集、验证集和测试集。

表 2 拟南芥和水稻各类型 5mC 位点数量统计

Table 2 Statistics on the number of various types of 5mC sites in *Arabidopsis thaliana* and *Oryza sativa*

物种 Species	基序 Motif	状态 State	基序位点数 Number of motif sites	总位点数 Total number of sites	训练样本数 Number of training samples
拟南芥 *Arabidopsis thaliana*	CpG	甲基化 Methylated	1 166 241	9 334 495	1 000 000
		未甲基化 Unmethylated	2 463 471	20 427 512	1 000 000
	CHG	甲基化 Methylated	124 516	966 529	966 529
		未甲基化 Unmethylated	1 344 336	11 142 868	966 529
	CHH	甲基化 Methylated	46 235	372 315	372 315
		未甲基化 Unmethylated	1 154 896	9 652 391	372 315
水稻 *Oryza sativa*	CpG	甲基化 Methylated	14 672 955	80 528 848	1 000 000
		未甲基化 Unmethylated	6 859 055	39 482 554	1 000 000
	CHG	甲基化 Methylated	4 302 311	23 214 826	1 000 000
		未甲基化 Unmethylated	9 292 069	53 685 174	1 000 000
	CHH	甲基化 Methylated	317 055	1 799 310	1 000 000
		未甲基化 Unmethylated	5 350 947	31 100 690	1 000 000

3.2 特征编码

对于每个基序位点，选取以胞嘧啶为中心、长度为 41 的 kmer 作为一个样本，例如一个 CpG 位点的 kmer(式中为 K)可以用公式(1)表示：

$$K_{CG} = X_1X_2\cdots X_{20}CGX_{23}\cdots X_{40}X_{41}, \quad (1)$$

其中，$X_i \in \{A, G, C, T\}$。每个 kmer 包含两种特征，即序列特征和电信号特征，其中序列特征表示为每个碱基的 one-hot 编码，A=(1, 0, 0, 0)，G=(0, 1, 0, 0)，C=(0, 0, 1, 0)，T=(0, 0, 0, 1)。同时，每个碱基取长度为 12 的电信号(测序产生的每个碱基的电信号个数是不等的，随机取每个碱基对应的 12 个电信号，对于长度小于 12 的以平均值补齐)，并且对电信号进行标准化处理。

3.3 评价指标

本实验使用 6 种常用评价指标来衡量模型性能并与其他方法进行比较，分别是准确率(accuracy)、精准率(precision)、召回率(recall)、F_1 值(精准率和召回率的调和平均)、AUC 以及皮尔逊相关系数(Pearson correlation coefficient)。其中 AUC 是 ROC 曲线与坐标轴围成的面积。ROC 曲线是根据一系列不同的二分类方式，以真阳性率为纵坐标，假阳性率为横坐标绘制的曲线；皮尔逊相关系数用于衡量 Nanoformer 和亚硫酸氢盐测序的甲基化水平的相关程度。各指标计算公式定义如下：

$$A = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}, \quad (2)$$

$$P = \frac{T_P}{T_P + F_P}, \quad (3)$$

$$R = \frac{T_P}{T_P + F_N}, \quad (4)$$

$$F_1 = \frac{2 \times P \times R}{P + R}, \quad (5)$$

其中，A 表示准确率，P 表示精准率，R 表示召回率，T_P、T_N、F_P 和 F_N 分别表示正确预测的 5mC 位点总数、正确预测的非 5mC 位点总数、错误预测的非 5mC 位点总数和错误预测的 5mC 位点总数。

3.4 实验环境及参数设置

本实验的代码运行基于 Python 3.6.13 和 Pytorch 1.4.0 实现；操作系统为 CentOS 7.5，所有模型训练和测试在 NVIDIA Tesla T4 GPU 上进行，内存为 32 G。使用 Adam (Kingma and Ba, 2014) 梯度下降算法和交叉熵来计算损失值。为了防止模型

过拟合，在 Bilstm 层和全连接层插入了 Dropout 层（Srivastava et al., 2014）并且使用早停策略（Wei et al., 2019）。

为了使模型性能达到最佳，需要对 Nanoformer 进行超参数调整。表 3 为超参数调整结果，其中 lr 表示学习率，batch size 表示每一个批次训练的样本数，num_layers_t 表示 transformer 编码器层数，n_head 表示多头注意力的头数，ffdim_t 表示前馈神经网络的维度，num_layers_l 表示 Bilstm 的层数，h_size_l 表示 Bilstm 隐藏层的维度，signal_len 表示每个碱基对应电信号的个数。

表 3 超参数调整结果

Table 3 Results of hyperparameter tuning

参数名称 Prameter name	参数范围 Parameter range	最优参数 Optimal parameters
lr	0.01, 0.001, 0.000 5, 0.000 01	0.000 5
batch size	128, 256, 512, 1 024	256
num_layers_t	1, 2, 3, 4	1
n_head	1, 2, 4, 8	1
ffdim_t	64, 128, 256	128
num_layers_l	1, 2, 3, 4	3
h_size_l	64, 128, 256	128
signal_len	1~16	12

作者贡献

曾佳完成本研究的实验设计、数据分析和论文初稿的写作；陈玲玲是本研究的构思者及负责人。所有作者都阅读并同意最终的文本。

参考文献

孙颖，葛锋，刘迪秋，等，2011. 植物中 DNA 甲基化模式及其相关机制. 植物生理学报，47(8)：745-751.［SUN Y，GE F，LIU D Q，et al.，2011. DNA methylation patterns and its related mechanism in plants. Plant Physiology Journal，47(8)：745-751.］

袁超，张少伟，牛义，等，2020. 植物 DNA 甲基化作用机制的研究进展. 生物工程学报，36(5)：838-848.［YUAN C，ZHANG S W，NIU Y，et al.，2020. Advances in research on the mechanism of DNA methylation in plants. Chinese Journal of Biotechnology，36(5)：838-848.］

DAVIS B M，CHAO M C，WALDOR M K，2013. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. Curr. Opin. Microbiol.，16(2)：192-198.

DOMB K，KATZ A，HARRIS K D，et al.，2020. DNA methylation mutants in *Physcomitrella patens* elucidate individual roles of CG and non-CG methylation in genome regulation. Proc. Natl. Acad. Sci. USA，117(52)：33700-33710.

HOCHREITER S，SCHMIDHUBER J，1997. Long short-term memory. Neural Comput.，9(8)：1735-1780.

KINGMA D P，BA J，2014. Adam：a method for stochastic optimization. arXiv：1412.6980.

KRUEGER F，ANDREWS S R，2011. Bismark：a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics，27(11)：1571-1572.

LASZLO A H，DERRINGTON I M，BRINKERHOFF H，et al.，2013. Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. Proc. Natl. Acad. Sci. U. S. A.，110(47)：18904-18909.

LIU Q，FANG L，YU G L，et al.，2019a. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nat. Commun.，10(1)：2449.

LIU Q，GEORGIEVA D C，EGLI D，et al.，2019b. NanoMod：a computational tool to detect DNA modifications using Nanopore long-read sequencing data. BMC Genom.，20(1)：31-42.

MIURA F，ENOMOTO Y，DAIRIKI R，et al.，2012. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res.，40(17)：e136.

NI P，HUANG N，NIE F，et al.，2021. Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning. Nat. Commun.，12：5976.

RAND A C，JAIN M，EIZENGA J M，et al.，2017. Mapping DNA methylation with high-throughput nanopore sequencing. Nat. Meth.，14(4)：411-413.

SCHATZ M C，2017. Nanopore sequencing meets epigenetics. Nat. Meth.，14(4)：347-348.

SCHREIBER J，WESCOE Z L，ABU-SHUMAYS R，et al.，2013. Error rates for nanopore discrimination among cytosine，methylcytosine，and hydroxymethylcytosine along individual DNA strands. Proc. Natl. Acad. Sci. U. S. A.，110(47)：18910-18915.

SIMPSON J T，WORKMAN R E，ZUZARTE P C，et al.，2017. Detecting DNA cytosine methylation using nanopore sequencing. Nat. Meth.，14(4)：407-410.

SRIVASTAVA N，HINTON G，KRIZHEVSKY A，et al.，2014. Dropout：a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res.，15：1929-1958.

STOIBER M，QUICK J，EGAN R，et al.，2017. *De novo* identification of DNA modifications enabled by genome-guided nanopore signal processing. bioRxiv，DOI：10.1101/094672.

VASWANI A，SHAZEER N，PARMAR N，et al.，2017. Attention is all you need// GUYON I，VON LUXBURG U，BEN-

GIO S, et al., Advances in Neural Information Processing Systems 30 (NIPS 2017). Long Beach: Neural Information Processing Systems Foundation, Inc. (NeurIPS): 5998-6008.

WEI Y T, YANG F, WAINWRIGHT M J, 2019. Early stopping for kernel boosting algorithms: a general analysis with localized complexities. IEEE Trans. Inform. Theory, 65(10): 6685-6703.

XU L, SEKI M, 2020. Recent advances in the detection of base modifications using the Nanopore sequencer. J. Hum. Genet., 65(1): 25-33.

ZHANG H M, LANG Z B, ZHU J K, 2018. Dynamics and function of DNA methylation in plants. Nat. Rev. Mol. Cell Biol., 19(8): 489-506.

(责任副主编 王海峰)
(责任副主编 罗继景)