文章编号：1006-3080(2025)04-0552-12 DOI：10.14135/j.cnki.1006-3080.20241104001

基于 MCSP 和 Swin Transformer 的转录因子结合位点预测模型

李 雪，石晋雪，王会青，闫奥煜，王 森
（太原理工大学计算机科学与技术学院(大数据学院)，太原 030600）

摘要：预测转录因子结合位点(Transcription Factor Binding Sites，TFBS)可以帮助识别特定细胞和组织的特异性调控机制，对于理解基因表达调控机制至关重要。现有方法结合DNA的序列和形状信息进行TFBS的预测，生成的形状信息未考虑长侧翼核苷酸的影响，在对序列信息进行特征提取时忽略了不同通道间特征的互补性，模型的预测能力有待提高。本文提出了TFBS预测模型MSSW，考虑长侧翼核苷酸来生成形状信息；利用Swin Transformer提取形状信息中长程依赖和局部关联相结合的特性，将分裂注意力融入多尺度卷积神经网络(Multi-scale Convolution and Split attention，MCSP)来捕获序列中不同通道间特征的互补性，获得跨通道的多尺度序列特征；结合提取的高级序列和形状特征进行TFBS的预测。结果表明，MSSW模型优于现有TFBS预测模型，可有效预测TFBS。

关键词：转录因子结合位点；多尺度卷积；分裂注意力；Swin Transformer；Deep DNAshape

中图分类号：TP391 **文献标志码：**A

转录因子结合位点(Transcription Factor Binding Sites, TFBS)是转录因子结合的特定DNA序列，通常长度在4~30个碱基对(bp)，这些结合位点的作用不仅限于启动转录，还包括调控转录效率以及参与复杂的基因调控网络[1-4]。尽管不同物种的转录因子在序列上可能有所不同，但它们在识别特定DNA序列及其结合方式上具有高度相似性，这种相似性不仅依赖于DNA序列本身，还依赖于DNA的形状信息[5]。Ma等[6]结合DNA序列信息和形状特征，开发了基于核函数的回归和分类框架，能够准确地建模和预测转录因子与DNA(Transcription Factor-DNA, TF-DNA)结合亲和力，实验结果表明，添加形状信息可以显著提高TFBS的预测能力。Zhang等[7]使用DNAshape[8]从一维的核苷酸序列中得到DNA的形状特征，包括次级沟宽度、螺旋桨扭曲、螺旋扭曲和滚动角，整合DNA序列和形状特征，更好地了解蛋白质-DNA的结合偏好。因此，结合序列信息和形状特征可以提高TFBS的预测精度，帮助研究人员更好地理解基因调控网络的复杂性，解释基因表达的调控机制[9-10]。

DNA形状是DNA双螺旋结构在不同区域呈现的三维特征，可以为模型提供DNA的空间结构特征，弥补纯序列信息的局限[11]。Wang等[12]提出的模型CRPTS通过滑动窗口和五聚体查询表从蒙特卡罗模拟[8]获得小沟宽度、滚动、螺旋桨扭转和螺旋扭转4种形状特征，用于TFBS预测。D_SSCA[13]与DeepSTF[14]模型均使用DNAshapeR[15]生成小沟宽度、滚动、螺旋桨扭转、螺旋扭转和小沟静电势5种形状特征，并结合序列信息进行TFBS的预测。然而，上述模型中生成DNA形状信息的方法是基于五

收稿日期：2024-11-04
基金项目：山西省自然科学基金(202203021211121)
作者简介：李 雪(1999—)，女，山东济南人，硕士生，主要研究方向为深度学习与生物信息处理。E-mail: lixue0712@link.tyut.edu.cn
通信联系人：王会青，E-mail: wanghuiqing@tyut.edu.cn
引用本文：李 雪，石晋雪，王会青，等. 基于MCSP和Swin Transformer的转录因子结合位点预测模型[J]. 华东理工大学学报(自然科学版)，2025，51(4)：552-563.
Citation: LI Xue, SHI Jinxue, WANG Huiqing, *et al*. Transcription Factor Binding Site Prediction Model Based on MCSP and Swin Transformer[J]. Journal of East China University of Science and Technology, 2025, 51(4): 552-563.

聚体查询表实现的, 只考虑一阶和二阶邻近核苷酸对中心核苷酸的影响, 忽略了长侧翼核苷酸的影响。Balaceanu 等[16] 指出超过二阶的长侧翼区的核苷酸可以通过改变局部的空间结构, 影响中心核苷酸的滚动角、倾斜角、上升距离等形状参数。DNA 形状变化受到延伸的侧翼区域的影响, 会影响转录因子的结合特异性[17], 这种结合特异性揭示了转录因子与 DNA 相互作用的规律, 帮助模型识别潜在的结合位点。在 TFBS 预测时需要考虑长侧翼核苷酸来生成长侧翼形状信息, 为模型提供更准确的形状信息, 但是五聚体查询表的长度限制了这些方法, 考虑更长查询表时又受到计算复杂度的限制。因此, 本文引入深度学习模型 Deep DNAshape[18] 来生成长侧翼形状信息, 弥补了基于五聚体查询表方式生成形状信息的不足, 为模型提供更加准确的形状信息。

DNA 相邻片段的形状信息相互作用, 形成局部关联性影响转录因子的具体结合方式[19], 同时较远的 DNA 片段之间通过三维空间相互作用影响彼此, 形状信息之间的长程依赖性反映了它们之间的相互作用[19], 局部关联性和长程依赖性的共同作用是揭示转录因子结合规律的重要体现。Zhang 等[7] 使用共享的卷积神经网络(CNN)提取 DNA 形状的局部关联信息和序列信息, 得到序列和形状的共同模式。Wang 等[12] 提出的 CRPTS 使用一种共享的混合 CNN 和循环神经网络(RNN)对 DNA 序列和局部形状信息进行特征提取。上述模型只考虑到形状信息之间的局部关联性, 未考虑形状信息之间存在的长程依赖性。DeepSTF[14] 将 Transformer 用于 TFBS 的预测, 结果表明, 使用 Transformer 来提取形状信息之间的长程依赖性可以有效地提高 TFBS 预测能力。Swin Transformer[20] 可以有效捕获局部关联信息, 同时通过窗口的移动实现全局整合, 获得长程依赖信息。因此, 引入 Swin Transformer 对形状信息进行特征提取, 提供了全局与局部协同的调控网络, 有效提取形状信息中长程依赖和局部关联相结合的特征, 来更全面地捕捉 DNA 的形状特征。

在转录因子识别 DNA 序列的过程中, 通过获得序列特征中跨通道的互补信息, 模型能够更全面地整合序列不同通道中包含的特征, 捕捉序列特征之间的复杂交互关系。Shen 等[21] 提出的 SAResNet, 通过自注意力机制捕获序列中的位置信息, 残差网络使用大小为 1×3 的卷积核提取序列中的结合特征, 并结合迁移学习进行 TFBS 的预测。D_SSCA[13] 将通道注意力用于 TFBS 的预测中, 通道注意力独立地为每个通道分配权重, 将生成的注意力图与高阶特征相乘进行自适应特征细化。DSAC[22] 提出双分支结构用于 TFBS 预测, 自注意力机制与 CNN 交叉使用对序列信息进行特征提取, 有效提取了序列的全局特征。上述注意力机制未能考虑不同通道之间的相互关联性, 导致通道间的协同关系未被充分利用, 从而无法有效捕捉特征之间的互补信息。分裂注意力[23] 对特征进行分组, 对每组通道单独应用注意力机制, 通过组间融合捕捉跨组的协同关系, 增强通道间特征的互补性。因此, 引入分裂注意力并与多尺度卷积相结合, 提出了 MCSP(Multi-scale convolution and split attention)方法, 分裂注意力机制将多尺度卷积捕获的不同长度的基序特征进行分组, 动态地调整不同组中通道的权重, 更好地融合这些多尺度特征, 进一步提高了模型的预测结果。

综上所述, 本文提出了基于 MCSP 和 Swin Transformer、用于 TFBS 预测的模型 MSSW; 使用 Deep DNAshape 来生成长侧翼形状信息, 提供了更全面的形状特征; 对于形状分支使用 Swin Transformer 进行特征提取, 充分提取形状信息之间的局部关联性和长程依赖性; 在序列分支中使用 MCSP 方法捕获不同长度的基序特征, 通过跨通道获得不同通道间特征的互补信息; 最后整合序列信息和形状特征, 进行转录因子结合位点的预测。

1　数据集及预处理

1.1　DNA 序列

本文使用的 165 个染色质免疫沉淀测序(ChIP-seq)数据集与 Zhang 等[13]使用的数据集相同, 便于后续对模型性能的评估, 这 165 个数据集是来自 DNA 元素百科全书(ENCODE)[24]项目生成的 690 个 ChIP-seq 数据集, 这些数据集是由 ChIP-seq 实验得到的, 被广泛应用于 DNA 结合蛋白的结合位点相关研究中。使用到的 165 个 ChIP-seq 数据集包含来自 32 个细胞系的 29 种转录因子, DNA 序列由 101 bp 组成。阳性数据集中的序列通过 ChIP-seq 实验证实与特定的转录因子有相互作用, 被标记为 1, 阴性数据集在保持二核苷酸频率不变的前提下, 通过打乱阳性数据集中碱基顺序来生成, 在数据集中被标记为 0[25]。

1.2　DNA 形状

使用 Deep DNAshape 生成形状信息: 翘曲、螺旋扭曲、小凹槽宽度、开合、螺旋桨扭转、上升、滚动、剪切、移动、滑动、错位、拉伸、倾斜, 为模型提供更细致和全面的 DNA 三维结构描述, Deep DNAshape[18]

通过准确考虑扩展侧翼区域的影响，从根本上改变了当前基于k-mer的DNA形状特征高通量预测，而无需进行广泛的分子模拟或结构生物学实验。通过Deep DNAshape生成的形状信息可以更加深入地了解侧翼区域对序列目标区域中DNA三维结构的影响。

2 TFBS预测模型MSSW

MSSW的总体模型图如图1所示。MSSW模型包括4部分：(1)数据编码阶段生成序列处理分支和形状处理分支的输入特征矩阵；(2)序列处理分支：使用MCSP获得跨通道的多尺度序列特征；(3)形状处理分支：使用Swin Transformer有效地捕捉并提取形状信息的高阶特征；(4)输出层将序列分支和形状分支提取的高级特征进行融合，得到最终的预测结果。

2.1 数据编码

本文使用one-hot对DNA序列进行编码，DNA的序列长度为101 bp，分别使用[1,0,0,0]、[0,1,0,0]、[0,0,1,0]、[0,0,0,1]来代表碱基对A、G、C、T。经过编码，序列被表示为1×4×101的特征矩阵 $\boldsymbol{I}_1$，如下所示：

$$\boldsymbol{I}_1=[\boldsymbol{S}_1,\boldsymbol{S}_2,\boldsymbol{S}_3,\cdots,\boldsymbol{S}_{100},\boldsymbol{S}_{101}] \tag{1}$$

其中，$\boldsymbol{S}_i$ 表示第 i 个核苷酸的one-hot向量。

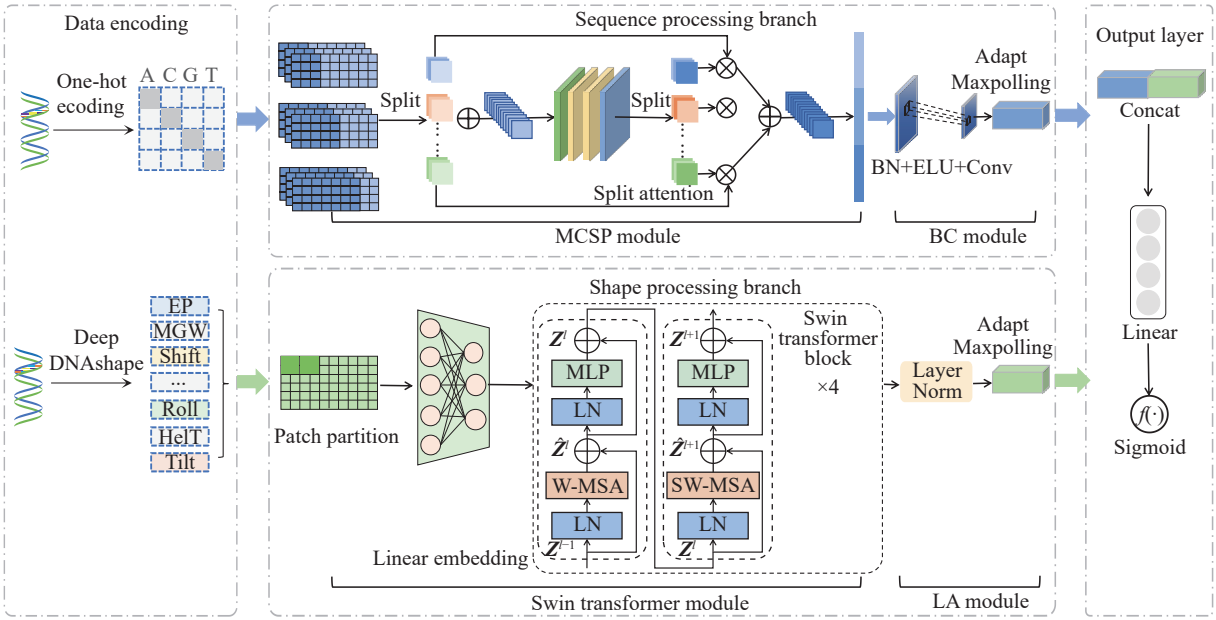


图1 MSSW模型架构

Fig. 1 Model framework of MSSW

为得到长侧翼形状信息，使用Deep DNAshape来生成单个核苷酸的13种形状信息，形状信息包括翘曲、螺旋扭曲、小凹槽宽度、开合、螺旋桨扭转、上升、滚动、剪切、移动、滑动、错位、拉伸、倾斜。给定的DNA序列长度为101 bp，最终形状信息表示为1×13×101的特征矩阵 $\boldsymbol{I}_2$，如下所示：

$$\boldsymbol{I}_2=[\boldsymbol{M}_1,\boldsymbol{M}_2,\boldsymbol{M}_3,\cdots,\boldsymbol{M}_{100},\boldsymbol{M}_{101}] \tag{2}$$

其中，$\boldsymbol{M}_i$ 表示第 i 个核苷酸使用Deep DNAshape得到的模拟向量。

2.2 序列处理分支

2.2.1 MCSP方法 转录因子结合位点是不等长的，传统的卷积层很难捕获到有关不等长的相关特征，使用多尺度卷积可以有效提取此特征[26]，分裂注意力对提取的多尺度特征进行特征处理，对通道进行分组计算注意力，捕捉通道间的交互信息，得到跨通道的多尺度序列特征。MCSP方法使用多尺度卷积来处理序列特征矩阵 $\boldsymbol{I}_1$，卷积核的大小分别为：4×3、4×5、4×7，不同尺度的卷积可以从多个角度有效地提取序列特征，可以更加全面地捕捉和表示不同长度转录因子结合位点的信息。然后，使用ELU激活函数来增加模型的非线性，再对激活函数的输出进行批量归一化(Batch Normalization, BN)处理，使其分布在均值为0，方差为1的范围内，加速模型收敛。以卷积核大小为4×3的卷积为例，计算过程如下所示：

$$O_{4\times3}=\mathrm{BN}(\mathrm{ELU}(\mathrm{Conv}(\boldsymbol{I}_1,\boldsymbol{W}_c,b_c))) \tag{3}$$

其中，$\boldsymbol{W}_c$、b_c 分别表示卷积层的权重矩阵和偏置，$O_{4\times3}$ 表示使用了卷积核大小为 4×3 的卷积，同样的步骤可以得到 $O_{4\times5}$、$O_{4\times7}$。使用最大池化选择每个窗口中的最大值，保留了特征图中的重要特征，同时降低特征图的空间维度，具体操作如下：

$$X_{4\times3}=\mathrm{MaxPool}(O_{4\times3}) \tag{4}$$

其中，以 $O_{4\times3}$ 为输入得到 $X_{4\times3}$，同理可以得到 $X_{4\times5}$、$X_{4\times7}$。

经过多尺度的卷积后，每个尺度的输出都包含了不同的序列特征，为了更好地利用这些多尺度序列特征，每一个尺度的输出都通过 MCSP 中的分裂注意力(Split attention)模块来增强序列特征在不同通道间的交互，获得跨通道的序列特征。沿通道方向拆分为 K 个候选组，每个候选组再拆分为 R 个特征组，输入 X 被分为了 $G=K\times W$ 个组，沿通道维度 $X=\{X_1,X_2,X_3,\cdots,X_G\}$，对每个分组使用映射变换 F_i，每个特征组表示为 $U_i=F_i(X_i),i=\{1,2,3,\cdots,G\}$，将不同特征组中相同通道的分量进行求和,实现特征融合。第 k 个特征组的中间表示为 $\widehat{U}^k$，如下所示：

$$\widehat{U}^k=\sum_{j=R(k-1)+1}^{Rk}U_j \tag{5}$$

其中，$\widehat{U}^k\in R^{H\times W\times C/k},k\in\{1,2,\cdots,K\}$，$H$、$W$、$C$ 分别表示特征图的高度、宽度、总通道数。在空间维度 $S^k\in R^{c/k}$ 上通过全局平均池化得到全局的上下文信息和各通道的统计数据，然后进行两次分组卷积，其中第 c 通道表示为：

$$S_c^k=\mathrm{Conv}_2\left(\mathrm{Conv}_1\left(\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\widehat{U}_c^k(i,j)\right)\right) \tag{6}$$

使用 rsoftmax 来提取不同特征图的注意力分数，将每个特征图通道加权组合，实现不同通道间的特征融合。第 k 个特征图组的分裂注意力信息表示为 $V^k\in R^{H\times W\times C/K}$，其中第 c 通道表示为：

$$V_c^k=\sum_{i=1}^{R}\alpha_i^k(c)\,U_{R(k-1)+i} \tag{7}$$

$$\alpha_i^k(c)=\begin{cases}\dfrac{\exp(\varsigma_i^c(s^k))}{\sum_{j=1}^{R}\exp(\varsigma_j^c(s^k))}, & R>1\\[2ex] \dfrac{1}{1+\exp(-\varsigma_i^c(s^k))}, & R=1\end{cases} \tag{8}$$

其中：$\alpha_i^k(c)$ 表示 rsoftmax 所分配的权重，ς_i^c 表示根据全局上下文 S^k 确定的第 c 通道的第 i 次分裂所对应的权重。3 个尺度的序列特征分别通过分裂注意力块，经 Concat 得到最终的输出 V，如式(9)所示。

$$V=\mathrm{Concat}(V_{(4\times3)},V_{(4\times5)},V_{(4\times7)}) \tag{9}$$

2.2.2　BC模块　BC 模块首先将 MCSP 方法的输出输入到 BN 层和 ELU 激活函数中，进而加速神经网络的训练，然后使用二维卷积来获得更高阶的特征表示，使用自适应最大池化来消除冗余信息，有效地保留了最显著的特征。上述操作如下所示：

$$\boldsymbol{M}=\mathrm{Conv}(\mathrm{ELU}(\mathrm{BN}(V))) \tag{10}$$

$$O_{\mathrm{seq}}=\mathrm{Flatten}(\mathrm{AdaptiveMaxpool}(\boldsymbol{M})) \tag{11}$$

2.3　形状处理分支

2.3.1　Swin Transformer　本文使用 Swin Transformer 对 DNA 的形状信息进行特征提取，在窗口中计算注意力并结合跨窗口交互，可以有效地提取 DNA 形状信息中的局部关联性和长程依赖性相结合的特征，如图 1 中所示。Patch Partition 将输入的特征矩阵分割为特征块，并使用线性层对特征块进行嵌入，这里与 ViT[27] 的嵌入层的作用是相同的，不改变输入特征的大小，只对通道的特征进行转换，形状特征矩阵经过上述处理后的输出为 $\boldsymbol{Z}$。

Swin Transformer 的核心 Swin Transformer Block 如图 2 所示，窗口多头自注意力(Window-based MSA,W-MSA)在局部窗口中计算注意力时，自注意力的计算复杂度与数据的大小成平方正相关，大大降低了运算量。但只使用窗口多头自注意力，每个窗口只关注自己的信息，无法捕获到全局信息，为了捕获跨窗口的长距离依赖关系，使用基于移动窗口多头注意力机制(Shifted Window-based MSA,SW-MSA)，通过移动操作在相邻的窗口计算注意力。由于在 Swin Transformer 中 W-MSA 与 SW-MSA 是成对出现的，这使得 Swin Transformer 可以有效地捕捉输入特征矩阵的局部和全局特征。Swin Transformer Block 的具体计算如下所示：

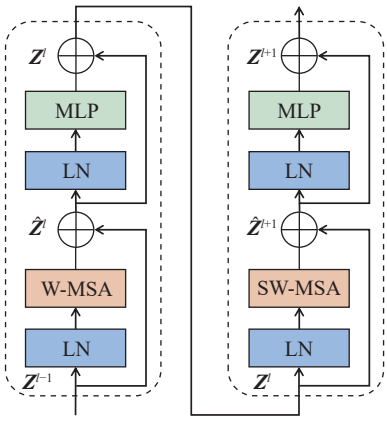


图 2　Swin Transformer Block

Fig. 2　Swin Transformer Block

$$\widehat{\boldsymbol{Z}}^{l}=W-\mathrm{MSA}(\mathrm{LN}(\boldsymbol{Z}^{l-1}))+\boldsymbol{Z}^{l-1} \tag{12}$$

$$\boldsymbol{Z}^{l}=\mathrm{MLP}(\mathrm{LN}(\widehat{\boldsymbol{Z}}^{l}))+\widehat{\boldsymbol{Z}}^{l} \tag{13}$$

$$\widehat{\boldsymbol{Z}}^{l+1}=\mathrm{SW}-\mathrm{MSA}(\mathrm{LN}(\boldsymbol{Z}^{l}))+\boldsymbol{Z}^{l} \tag{14}$$

$$\boldsymbol{Z}^{l+1}=\mathrm{MLP}(\mathrm{LN}(\widehat{\boldsymbol{Z}}^{l+1}))+\widehat{\boldsymbol{Z}}^{l+1} \tag{15}$$

其中，LN 表示层归一化，MLP 为多层感知机，$\boldsymbol{Z}^{l}$ 表示 Swin Transformer 中第 L 层的输出。

2.3.2 LA模块 LA 模块将特征输入经过层归一化，从而增强训练过程的稳定性和模型的收敛性。随后，经过层归一化的特征会传递到自适应平均池化，自适应平均池化将特征图调整到一个固定的尺寸，如式(16)所示：

$$O_{\mathrm{shape}}=\mathrm{Flatten}(\mathrm{AdaptAvgpool}(\mathrm{LN}(\boldsymbol{Z}'))) \tag{16}$$

2.4 输出层

将序列处理分支得到的高级特征与形状处理分支得到的高级特征沿着特征维度进行连接，形成一个综合特征向量，将连接后的特征输入全连接层，将高维特征映射到目标维度，最后使用 Sigmoid 函数来预测转录因子结合位点。上述操作可以表示为：

$$\widehat{y}=\mathrm{Sigmoid}(\mathrm{Linear}(\mathrm{Concat}(O_{\mathrm{seq}},O_{\mathrm{shape}}),\boldsymbol{W}_{\mathrm{f}},b_{\mathrm{f}})) \tag{17}$$

其中，$\boldsymbol{W}_{\mathrm{f}}$、$b_{\mathrm{f}}$ 分别表示线性层的权重矩阵和偏置。

2.5 模型训练

本文模型 MSSW 使用 PyTorch 深度学习框架实现，将训练集按照 8∶2 的比例划分为训练集和验证集，测试集是一个独立的数据集，在训练集上对模型进行训练，使用批量优化对模型进行优化，使用二元交叉熵函数来计算损失，通过反向传播计算梯度，使用 Adam 优化器根据计算得到的梯度调整模型参数，以最小化损失函数。损失函数 Loss 及 Adam 的定义如下所示：

$$\mathrm{Loss}=-\frac{1}{N}\sum_{i=1}^{N}[\widehat{y_i}\times\log(\widehat{y_i})+(1-\widehat{y_i})\times\log(1-\widehat{y_i})] \tag{18}$$

$$\mathrm{Adam}(K,\mathrm{grad}(\mathrm{Loss})) \tag{19}$$

其中，N 为批量大小，$\widehat{y_i}$ 表示 N 个批次中第 i 个序列的标签，K 表示 MSSW 模型中所有的可学习参数，$\mathrm{grad}(\cdot)$ 为梯度函数。每个输入序列的计算过程使用算法 1 表示，见表 1。

3 结果与分析

3.1 模型评估指标

本文对转录因子结合位点进行预测，将此问题转换为二分类问题，使用准确率(ACC)、受试者操作特性曲线下的面积(ROC-AUC)、精确率-召回率曲线下的面积(PR-AUC)指标对分类器的最终预测结果进行评估。

相关公式如下所示：

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{20}$$

$$\mathrm{TPR}=\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{TN}} \tag{21}$$

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{TN}} \tag{22}$$

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{23}$$

表 1 算法 1 的具体步骤

Table 1 Specific steps of algorithm 1

Algorithm 1: MSSW

Input: The DNA sequence containing or not containing TFBSs

Output: The current input sequence's predicted value and the learnt parameters

1. The sequence feature matrix $\boldsymbol{I}_1$ obtained using one-hot encoding is shown in Equation (1), and the shape feature matrix $\boldsymbol{I}_2$ calculated using Deep DNAshape is shown in Equation (2)

2. Initialize all the parameters K of neural network

3.while Epoch < MaxEpoch do:

4.Sequence processing branch:The output V of the MCSP method is computed using Equations (3) to (9), and processed through Equations (10) and (11) to obtain the output of the BC module, which corresponds to the sequence branch output O_{seq}

5.Shape processing branch: The output $\boldsymbol{Z}'$ of the Swin Transformer is obtained using Equations (12) to (15), and the output of the LA module, which represents the shape branch output O_{shape}, is computed using Equation (16)

6.The outputs O_{seq} from the sequence processing branch and O_{shape} from the shape processing branch are concatenated along the feature dimension using Equation (17) to compute the predicted value $\widehat{y}$ for the current input sequence

7.Calculate the Loss using Equation (18)

8.Update the parameter K according to Equation (19)

9.Epoch = Epoch + 1

10.while end

其中, TP 表示真阳性, 正样本被预测为正样本的数量; FP 表示假阳性, 正样本被预测为负样本的数量; FN 表示假阴性, 负样本被预测为正样本的数量; TN 表示真阴性, 负样本被预测为负样本的数量; TPR 与 Recall 都表示正样本被预测为正例的比例, 即真正例率; FPR 表示负样本分错的概率, 即假正例率; Precision 表示预测为真的样本中真正为正样本的比例。

3.2　长侧翼形状信息对于预测结果的提升

为了验证使用长侧翼形状信息在 TFBS 预测中的优势, 本文在 TFBS 预测任务上设计了对比实验, Deep DNAshape 生成的形状信息较 DeepshapeR 考虑了长侧翼核苷酸对中心核苷酸的影响, 在实验中分别以 Deep DNAshape 和 DeepshapeR 生成的形状信息作为 MSSW 模型形状处理分支的输入, 在 165 个 ChIP-seq 数据集上进行实验, 实验结果如图 3 所示。DeepshapeR 是基于五聚体查询表实现的, 生成的形状信息只考虑了一阶、二阶邻近核苷酸的影响, Deep DNAshape 生成的形状信息考虑了长侧翼核苷酸的影响, 具有更加丰富的形状信息。Deep DNAshape 生成的长侧翼形状信息在进行实验时, 在 ACC、ROC-AUC、PR-AUC 3 个指标上明显优于使用 DeepshapeR 生成的形状信息, 这说明生成的形状信息考虑长侧翼核苷酸的影响是有必要的。

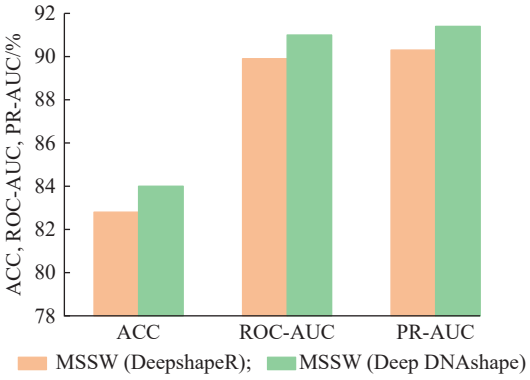


图 3　MSSW 在 165 个数据集上使用 DeepshapeR 与 Deep DNAshape 的性能比较

Fig. 3　Performance comparison of MSSW using DeepshapeR encoding *vs*. Deep DNAshape on 165 datasets

3.3　消融实验

3.3.1　MCSP和Swin Transformer的消融实验　为了更好地理解 MCSP 和 Swin Transformer 的作用, 设计了两个模型: (1)模型 CNNTF: 使用卷积来处理序列分支, 使用 Transformer 处理形状分支; (2)模型 CNNSw: 在模型 CNNTF 的基础上将 Transformer 改为使用 Swin Transformer。

模型 CNNTF 与模型 CNNSw 在序列处理分支上都使用 CNN, 但在形状处理分支中分别使用 Transformer 和 Swin Transformer 来提取形状特征, CNNSw 的预测结果优于 CNNTF, 可以用来验证 Swin Transformer 的有用性。MSSW、CNNTF 和 CNNSw 的实验结果如表 2 所示, 相较于 CNNTF 来说, 在形状处理分支中使用了 Swin Transformer 的 CNNSw 在所有评估指标上均提升了 3.3%。相较于 Transformer 来说, Swin Transformer 的效果要更好。分析原因是: 一方面, Swin Transformer 引入了局部窗口注意力机制, 将计算注意力限制在局部窗口内, 这种机制减少了计算复杂度, 同时捕捉到形状信息的局部关联性; 另一方面, 通过移动窗口技术, Swin Transformer 在不同层次之间滑动和重叠窗口, 使得窗口之间的信息能够有效交互。这种技术进一步提升了模型对形状信息中长程依赖性的捕捉, 弥补了局部注意力的局限性。

表 2　MSSW、CNNTF 和 CNNSW 的实验结果

Table 2　Experimental results of MSSW, CNNTF and CNNSW

Model	ACC/%	ROC-AUC/%	PR-AUC/%
CNNTF(CNN+Transformer)	78.8	86.1	86.5
CNNSw(CNN+Swin)	82.1	89.4	89.8
MSSW(MCSP+Swin)	**84.0**	**91.0**	**91.4**

分析模型 CNNSw 与模型 MSSW, 分别对应不使用 MCSP 模块与使用 MCSP 模块, 来验证 MCSP 的作用。在序列分支中使用了 MCSP 模块的 MSSW 要比不使用 MCSP 模块的 CNNSw 表现好, 如表 2 所示, MSSW 在 ACC 上相对提升 1.9%, 在 ROC-AUC 上相对提升 1.6%, 在 PR-AUC 上相对提升 1.6%。MSSW 及其变体(CNNTF、CNNSw)在 ACC、ROC-AUC、PR-AUC 测试集上的数据分布如图 4 所示, MSSW 及其变体模型 CNNSw 的数据分布都比较稳定, 但相对来说 MSSW 数据的整体分布偏高, 说明大多数数据集在 MSSW 模型上预测的结果更好。使用多尺度卷积和分裂注意力相结合的 MCSP 方法可以实现各自单独应用时无法达到的效果, 多尺度卷积得到不同长度基序特征, 同时通过在提取的多尺度特征上应用分裂注意力, 模型能够捕捉序列特征跨通道的互补信息, 模型可以动态调整和优化特征表示, 提高对复杂模式的识别能力, 更有利于 TFBS 的预测。

MSSW 模型与 CNNTF、CNNSw 相比, 都有了更大的提升, 其中 CNNTF 的预测结果最差。在 165 个 ChIP-seq 数据集的测试集上, 模型在 ACC、ROC-

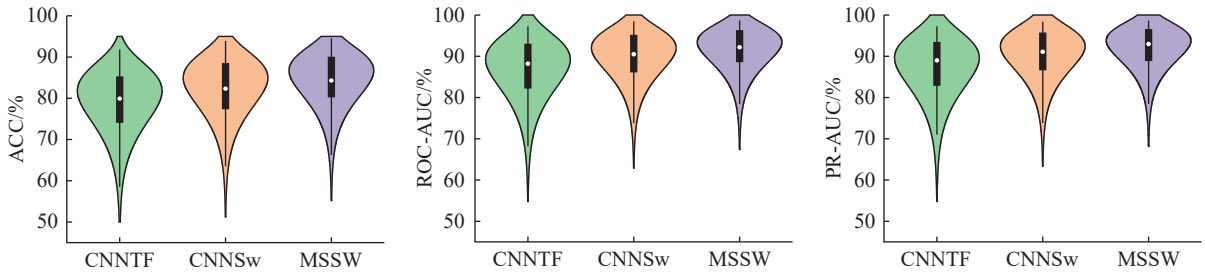


图 4 MSSW 及其变体(CNNTF、CNNSw)在 ACC、ROC-AUC、PR-AUC 上的分布

Fig. 4 Distribution of MSSW and its variants (CNNTF, CNNSw) in ACC, ROC-AUC, and PR-AUC

AUC 和 PR-AUC 的结果可视化如图 4 所示，MSSW 的数据分布最稳定，且在大多数数据集上表现都更好，尤其相较于 CNNTF 模型来说。在序列处理分支中使用 MCSP 模块，并在形状处理分支中使用 Swin Transformer 的模型架构 MSSW，取得了最好的预测结果。

3.3.2 序列处理分支的消融实验 为了验证多尺度卷积的实用性，提出一个新的模型 NSpSw，如图 5(a)所示。该模型为 MSSW 的变体，在序列处理模块中只使用 4×3 这种卷积核大小的卷积。在 165 个 ChIP-seq 数据集的 ACC、ROC-AUC、PR-AUC 的结果如表 3 所示，MSSW 较 NSpSw 在 ACC 上提升 1.4%、在 ROC-AUC 上提升 1.3%、在 PR-AUC 上提升 1.3%。图 6 的散点图示出了 NSpSw 与 MSSW 在 165 个 ChIP-seq 测试子集上 ACC、ROC-AUC、PR-AUC 的比较，其中每个点对应一个测试子集，纵坐标、横坐标分别对应不同模型的 ACC、ROC-AUC、PR-AUC 值。以 ACC 为例，有 142 个数据集在对角线及其上方，23 个数据集位于对角线的下方，说明 MSSW 在 ACC 这个评估指标上有 142 个数据集的结果优于(或等同于)不使用多尺度的模型 NSpSw。多尺度卷积能够更好地适应转录因子结合位点不等长的特性，这种变长特性使得多尺度卷积能够有效捕捉和利用不同长度的序列信息，小尺度卷积核可以捕捉短序列中的关键信息，而大尺度卷积核则能够捕捉较长序列中的关键信息，使得最终提取的特征更加丰富，对于模型预测性能有一定的提升。

本文还比较了使用自注意力机制[28]的模型 MSaSw，如图 5(b)所示，将 MSSW 序列分支中的分裂注意力机制替换为自注意力机制，来验证分裂注意力机制的有效性。MSSW、NSpSw 和 MSaSw 在 165 个 ChIP-seq 数据集的 ACC、ROC-AUC、PR-AUC 的平均结果如表 3 所示，MSSW较 MSaSw 在 ACC、ROC-AUC、PR-AUC 上分别提升了 0.8%、0.7%、0.7%，实验结果的可视化如图 6 所示，多数数据集都位于对角线的上方，说明在多数数据集上 MSSW 模型的结果要优于使用自注意力的模型 MSaSw。分析其原因，在序列长度较长时，自注意力

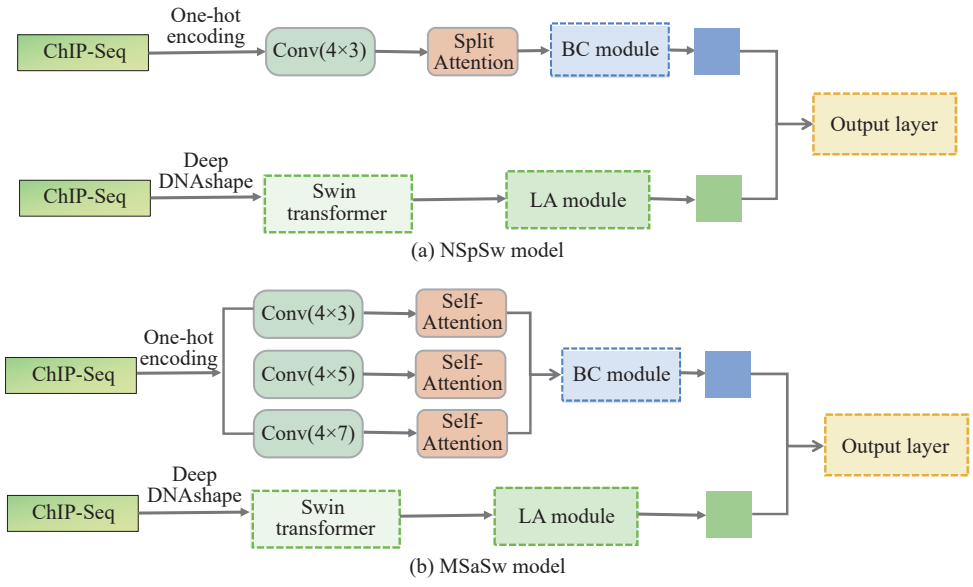


图 5 变体模型

Fig. 5 Variant models

表 3　MSSW、NSpSw 和 MSaSw 实验结果

Table 3　Experimental results of MSSW, NSpSw and MSaSw

Model	ACC/%	ROC-AUC/%	PR-AUC/%
NSpSw(Without multi-scale+Swin)	82.6	89.7	90.1
MSaSw(With Self-attention+Swin)	83.2	90.3	90.7
MSSW(MCSP+Swin)	**84.0**	**91.0**	**91.4**

的计算复杂度要远高于分裂注意力，分裂注意力通过将注意力计算分割到不同的通道，在组内计算注意力获取不同通道之间交互的序列特征，能够更有效地处理复杂数据，自注意力机制在计算每个位置的注意力时考虑了整个序列，这导致在特征提取过程中引入大量冗余信息和噪声。分裂注意力机制可以更好地提取 DNA 序列的局部特征，减少冗余信息和噪声，所以在模型中加入分裂注意力可以有效地提升预测结果。

3.4　跨细胞系来证明模型的泛化性

为了验证所提出模型在不同转录因子(TF)上的广泛适用性，使用不同的细胞系进行训练和测试，选取了 4 种细胞系(分别为 HepG2、Gm12878、K562、HelaS3)，由前人的研究[13]可知这些细胞系中大多数的转录因子是重叠的。首先使用不同的细胞系来构建 MSSW 模型，然后使用训练好的模型在感兴趣的细胞系上预测转录因子结合位点，使用平均 ROC-AUC 作为评估指标，如图 7 所示，跨细胞系预测的平均 ROC-AUC 都在 80% 以上，使用 K562 训练的模型在其他 3 个细胞系测试集上的平均 ROC-AUC 均高于 87%，说明 MSSW 模型可以有效地预测不同细胞系的转录因子结合位点。由图 7 可知，MSSW 模型在传统预测上的结果要优于跨细胞系预测，不同细胞系之间存在生物学上的差异，这使得跨细胞系的预测模型在准确性和可靠性方面受到一定的限制。这些结果也表明，尽管存在这些差异，MSSW 模型依然具有一定的跨细胞系预测能力，这为基于不同细胞系的转录因子结合位点预测提供了有价值的参考。

3.5　对比实验

为了进一步证明本文提出模型的优越性，与在转录因子结合位点预测中最先进的 7 种模型进行了比较。在 165 个 ChIP-seq 数据集的平均 ACC、ROC-AUC、PR-AUC，如表 4 所示，与次优方法(DSAC(2023)：81.6%、89.2%、89.7%)相比，分别提升 2.4%、1.8%、1.7%。实验结果的可视化如图 8 所示，MSSW 模型在 165 个数据集上 ACC、ROC-AUC、PR-AUC 的结果分布更加集中，评估指标的最高点都比其他模型高，并且最低点显著优于其他对比模型，这说明 MSSW 具有更好的泛化性，对于不同类型数据集的预测都有较好的结果。与只包含序列信息的 DeepBind、DanQ、DSAC 相比，性能均有了较大的提升，分析其结果归因于 MSSW 模型结合了 DNA 的序列和形状信息，提供了更加全面的生物学视角，有效增强了模型对转录因子结合位点的识别能力，说明融合多层

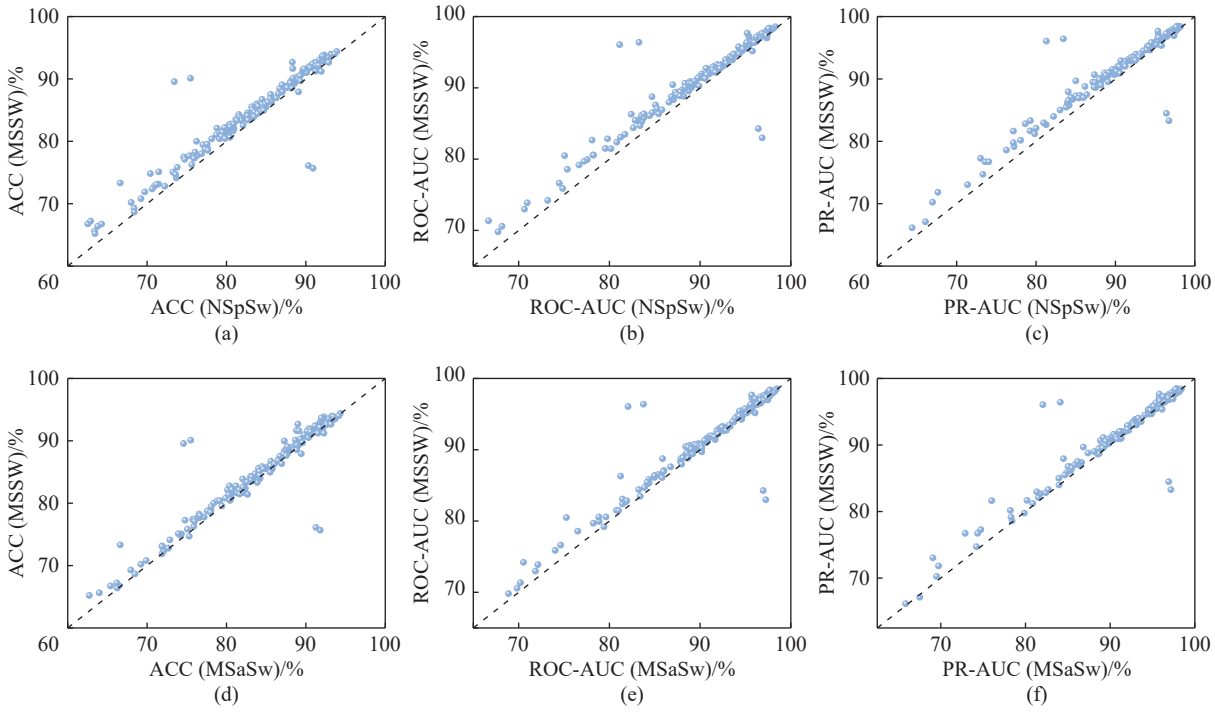


图 6　MSSW 及其变体(NSpSw、MSaSw)在 165 个 ChIP-seq 数据集的测试集上实验结果的比较

Fig. 6　Comparison of experimental results of MSSW and its variants (NSpSw, MSaSw) on the test sets of 165 ChIP-seq datasets

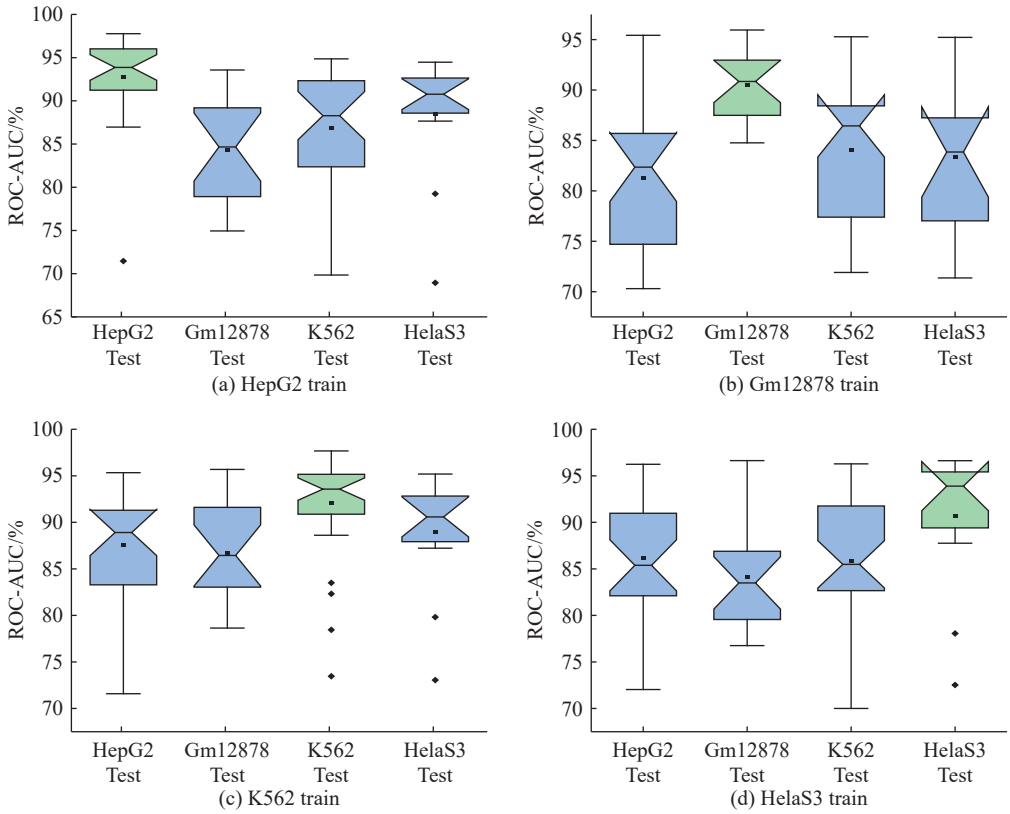


图 7 跨细胞系预测与传统预测之间的 ROC-AUC 比较

Fig. 7 ROC-AUC comparison between cross-cell line prediction and traditional prediction

表 4 MSSW 与 7 种最先进方法在 165 个 ChIP-seq 数据集测试集上的平均 ACC、ROC-AUC、PR-AUC

Table 4 Average ACC, ROC-AUC, PR-AUC of MSSW and seven advanced methods on the test sets of 165 ChIP-seq datasets

Model	ACC/%	ROC-AUC/%	PR-AUC/%
DeepBind(2015)[25]	78.4	85.0	85.5
DanQ(2016)[29]	78.0	84.9	85.5
DLBSS(2019)[7]	79.2	86.6	87.1
CRPTS(2021)[12]	78.9	85.8	86.3
D_SSCA(2022)[13]	78.4	85.3	85.6
DeepSTF(2023)[14]	81.1	88.2	88.8
DSAC(2023)[22]	81.6	89.2	89.7
MSSW	**84.0**	**91.0**	**91.4**

次的信息是有必要的。上述模型中 DLBSS、CRPTS、D-SSCA、DeepSTF 结合了序列和形状信息，使用基于五聚体的方法来生成形状信息；MSSW 使用 Deep DNAshape 来生成形状信息，与基于五聚体的方法相比较，充分考虑了长侧翼核苷酸对形状信息的影响，为模型提供了更全面的形状信息。与使用双分支的自注意力与卷积交替的 DSAC 模型结构相比，MSSW 使用 MCSP 来处理序列分支，自注意力不善于捕获跨通道的序列信息，分裂注意力引入分支结构，在组间计算注意力使得不同通道之间可以进行交互，能更好地捕捉通道间的互补信息，在一定程度上缓解了此问题，同时多尺度卷积较固定大小的卷积可以

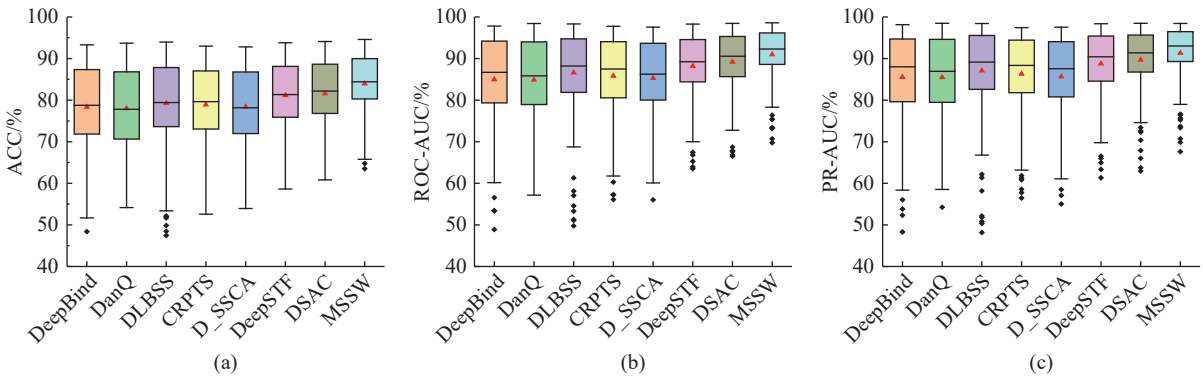


图 8 MSSW 及其对比方法实验结果的可视化

Fig. 8 Visualization of experimental results of MSSW and its comparative methods

捕获到不同长度的序列特征。

为了进一步说明在不同尺度数据集上的表现，以及数据集的大小对于最终预测结果的影响，分别在小型、中型、大型数据集上进行性能比较，不同尺度数据集的划分及其包含数据集的个数如表 5 所示。

表 5　不同尺度数据集的划分及数据集的个数

Table 5　Division of datasets at different scales and the number of datasets

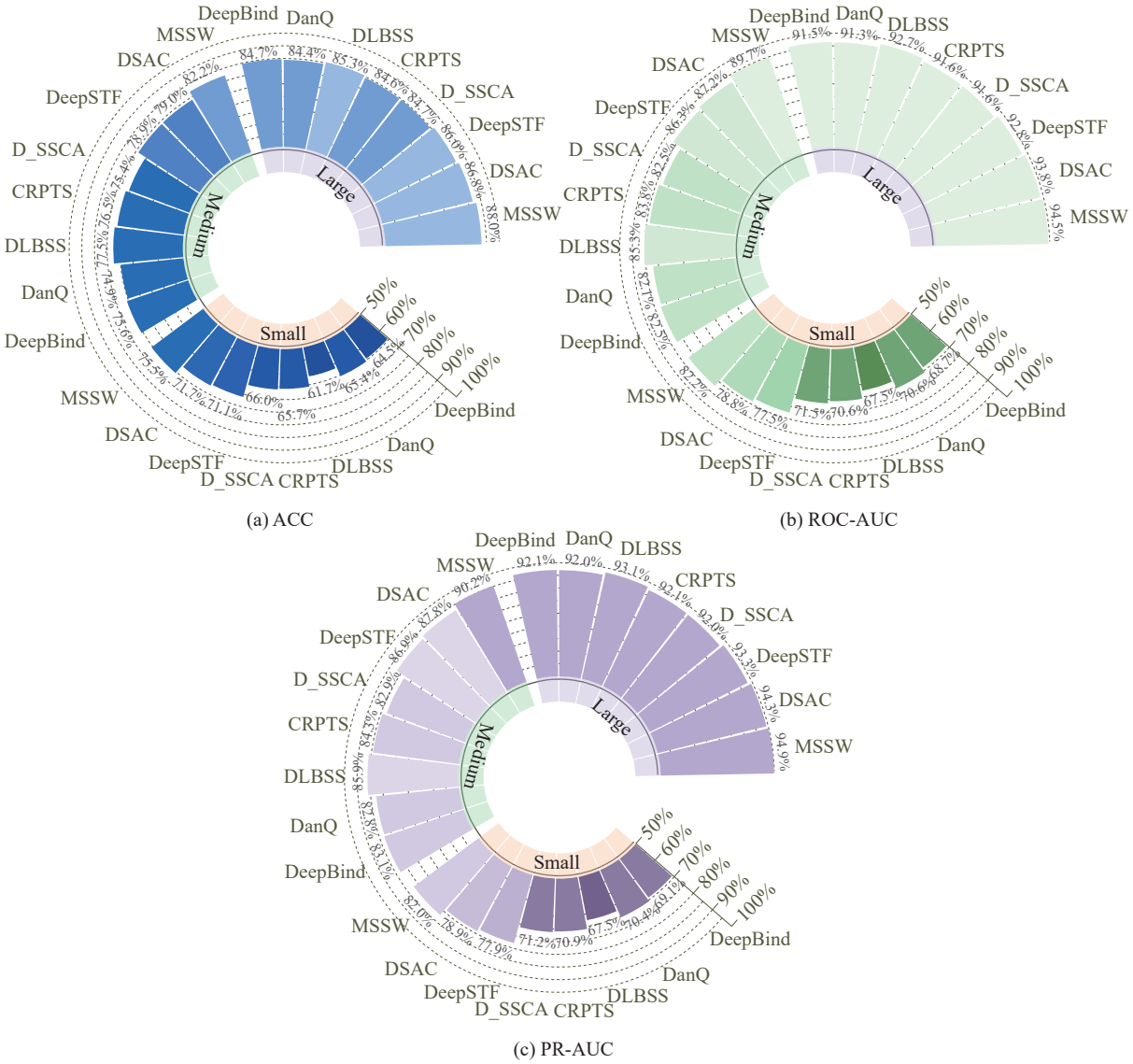
Scale of dataset	Condition of scale	Number of datasets
Small dataset	< 6 000	17
Medium dataset	6 000 ~ 30 000	77
Large dataset	> 300 00	71

在不同尺度数据集上 ACC 的实验结果如表 6 所示，与次优模型 DSAC 相比，MSSW 模型在不同尺度的数据集上 ACC 值分别提升了 3.8%、3.2%、1.2%，在小型数据集上提升效果最为明显，在中型数据集上略低于小型数据集，在大型数据集上也有较

表 6　不同方法在不同尺度数据集上 ACC 性能比较

Table 6　Comparison of ACC performance of different methods on datasets at different scales

Model	ACC/%			
	All	Small	Medium	Large
DeepBind(2015)[25]	78.4	64.5	75.6	84.7
DanQ(2016)[29]	78.0	65.4	74.9	84.4
DLBSS(2019)[7]	79.2	61.7	77.5	85.3
CRPTS(2021)[12]	78.9	65.7	76.5	84.6
D_SSCA(2022)[13]	78.4	66.0	75.4	84.7
DeepSTF(2023)[14]	81.1	71.1	78.9	86.0
DSAC(2023)[22]	81.6	71.7	79.0	86.8
MSSW	**84.0**	**75.5**	**82.2**	**88.0**



(a) ACC

(b) ROC-AUC

(c) PR-AUC

图 9　不同方法在不同尺度数据集上预测结果

Fig. 9　Prediction results of different methods on datasets at different scales

小的提升。MSSW模型不仅在ACC上，在ROC-AUC、PR-AUC上也有显著的提升，实验结果如图9所示，在不同尺度的数据集上MSSW都取得了最好的实验结果，尤其是在中小型数据集上，这说明本文模型架构的设计对于提升中小型数据集有显著优势。

4 结 论

本文提出基于MCSP和Swin Transformer的转录因子结合位点预测模型MSSW，该方法结合了DNA的序列和形状信息。MSSW使用Deep DNAshape来生成长侧翼形状信息，在形状处理分支使用Swin Transformer来提取形状信息中局部关联性和长程依赖性相结合的特征，在序列处理分支使用MCSP得到跨通道的多尺度序列特征，最后将提取到的序列特征和形状特征进行融合，得到最终的预测结果。结果表明，长侧翼形状信息提升了模型对于TFBS预测的性能，并通过跨细胞系实验证明了模型的泛化性。同时，MSSW模型中所使用模块具有一定的优越性，MSSW模型优于现有模型。最后，MSSW在不同尺度的数据集上均有提升，尤其在中小型数据集上提升最为明显。尽管MSSW模型在转录因子结合位点的预测中取得了较好的预测结果，但是此模型只能预测固定长度的DNA序列，在后续的研究中应设计相应的模型使其可以预测任意长度的DNA序列。

参考文献：

[1] GEORGAKOPOULOS-SOARES I, DENG C, AGARWAL V, *et al*. Transcription factor binding site orientation and order are major drivers of gene regulatory activity[J]. Nature Communications, 2023, 14(1): 2333.

[2] LAMBERT S A, JOLMA A, CAMPITELLI L F, *et al*. The human transcription factors[J]. Cell, 2018, 172(4): 650-665.

[3] MITSIS T, EFTHIMIADOU A, BACOPOULOU F, *et al*. Transcription factors and evolution: an integral part of gene expression[J]. World Academy of Sciences Journal, 2020, 2(1): 3-8.

[4] GUALBERTO J M, KÜHN K. DNA-binding proteins in plant mitochondria: Implications for transcription[J]. Mitochondrion, 2014, 19: 323-328.

[5] ZHOU T, SHEN N, YANG L, *et al*. Quantitative modeling of transcription factor binding specificities using DNA shape[J]. Proceedings of the National Academy of Sciences, 2015, 112(15): 4654-4659.

[6] MA W, YANG L, ROHS R, *et al*. DNA sequence+ shape kernel enables alignment-free modeling of transcription factor binding[J]. Bioinformatics, 2017, 33(19): 3003-3010.

[7] ZHANG Q, SHEN Z, HUANG D S. Predicting *in-vitro* transcription factor binding sites using DNA sequence+ shape[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, 18(2): 667-676.

[8] ZHOU T, YANG L, LU Y, *et al*. DNAshape: A method for the high-throughput prediction of DNA structural features on a genomic scale[J]. Nucleic Acids Research, 2013, 41: W56-W62.

[9] TEWHEY R, KOTLIAR D, PARK D S, *et al*. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay[J]. Cell, 2016, 165(6): 1519-1529.

[10] GASPERINI M, TOME J M, SHENDURE J. Towards a comprehensive catalogue of validated and target-linked human enhancers[J]. Nature Reviews Genetics, 2020, 21(5): 292-310.

[11] ROHS R, WEST S M, SOSINSKY A, *et al*. The role of DNA shape in protein-DNA recognition[J]. Nature, 2009, 461(7268): 1248-1253.

[12] WANG S, ZHANG Q, SHEN Z, *et al*. Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture[J]. Molecular Therapy-Nucleic Acids, 2021, 24: 154-163.

[13] ZHANG Y, WANG Z, ZENG Y, *et al*. A novel convolution attention model for predicting transcription factor binding sites by combination of sequence and shape[J]. Briefings in Bioinformatics, 2022, 23(1): bbab525.

[14] DING P, WANG Y, ZHANG X, *et al*. DeepSTF: Predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape[J]. Briefings in Bioinformatics, 2023, 24(4): bbad231.

[15] CHIU T P, COMOGLIO F, ZHOU T, *et al*. DNAshapeR: An R/Bioconductor package for DNA shape prediction and feature encoding[J]. Bioinformatics, 2016, 32(8): 1211-1213.

[16] BALACEANU A, BUITRAGO D, WALTHER J, *et al*. Modulation of the helical properties of DNA: Next-to-nearest neighbour effects and beyond[J]. Nucleic Acids Research, 2019, 47(9): 4418-4430.

[17] GORDÂN R, SHEN N, DROR I, *et al*. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape[J]. Cell Reports, 2013, 3(4): 1093-1104.

[18] LI J, CHIU T P, ROHS R. Deep DNAshape: Predicting DNA shape considering extended flanking regions using a deep learning method[EB/OL]. (2023-10-24) [2024-02-23]. https://doi.org/10.1101/2023.10.22.563383.

[19] GU C, ZHANG J, YANG Y I, *et al*. DNA structural correlation in short and long ranges[J]. The Journal of Physical Chemistry B, 2015, 119(44): 13980-13990.

[20] LIU Z, LIN Y, CAO Y, *et al*. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021: 9992-10002.

[21] SHEN L C, LIU Y, SONG J, *et al*. SAResNet: Self-attention residual network for predicting DNA-protein binding[J]. Briefings in Bioinformatics, 2021, 22(5): bbab101.
[22] YU Y, DING P, GAO H, *et al*. Cooperation of local features and global representations by a dual-branch network for transcription factor binding sites prediction[J]. Briefings in Bioinformatics, 2023, 24(2): bbad036.
[23] ZHANG H, WU C, ZHANG Z, *et al*. Resnest: Split-attention networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New Orleans, LA, USA: IEEE, 2022: 2736-2746.
[24] ENCODE PROJECT CONSORTIUM. An integrated encyclopedia of DNA elements in the human genome[J]. Nature, 2012, 489(7414): 57.
[25] ALIPANAHI B, DELONG A, WEIRAUCH M T, *et al*. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning[J]. Nature Biotechnology, 2015, 33(8): 831-838.
[26] 孙俊静, 顾幸生. 基于注意力机制多尺度卷积神经网络的轴承故障诊断 [J]. 华东理工大学学报 (自然科学版), 2024, 50(2): 247-256.
[27] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2020-10-22) [2021-03-03]. https://doi.org/10.48550/arXiv.2010.11929.
[28] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 122-129.
[29] QUANG D, XIE X. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences[J]. Nucleic Acids Research, 2016, 44(11): e107.

Transcription Factor Binding Site Prediction Model Based on MCSP and Swin Transformer

LI Xue, SHI Jinxue, WANG Huiqing, YAN Aoyu, WANG Sen
(College of Computer Science and Technology(College of Data Science),
Taiyuan University of Technology, Taiyuan 030600, China)

Abstract: Predicting Transcription Factor Binding Sites (TFBS) can help identify specific regulatory mechanisms of cells and tissues, which is crucial for understanding gene expression regulation mechanisms. The existing methods combine DNA sequence and shape information for TFBS prediction, but they typically focus only on neighboring nucleotides to generate shape information, neglecting the influence of longer flanking nucleotides. In the sequence processing branch, these methods neglect the complementarity of features across different channels. Similarly, in the shape processing branch, local correlations and long-range dependencies of shape information are not adequately captured. This lack of deep exploration of both sequence and shape information limits prediction performance. To address these issues, this paper proposes a novel model, MSSW, for predicting transcription factor binding sites. Firstly, Deep DNAshape is used to generate long flanking shape information for the shape branch, considering a more comprehensive set of shape data. Additionally, the Swin Transformer is utilized for feature extraction of the shape information, capturing local correlations through window-based self-attention and obtaining long-range dependency information through window movement. Furthermore, the Multi-scale Convolution and Split attention (MCSP) are employed to extract multi-scale cross-channel features from the sequence. Meanwhile, the sequence and shape features are fused to predict transcription factor binding sites. Finally, MSSW is evaluated on 165 ChIP-seq datasets. The experimental results show that it is superior to existing TFBS prediction models and ablation studies validate the effectiveness of MCSP and the Swin Transformer. Additionally, the model's generalization is verified across different cell lines, providing valuable insights for predicting TFBS in various cellular contexts. The proposed model achieves strong predictive performance across datasets of different scales, particularly excelling with medium and small-sized datasets.

Key words: transcription factor binding sites; multi-scale convolution; split attention; Swin Transformer; Deep DNAshape

（责任编辑：李娟）