计算机科学
COMPUTER SCIENCE

基于Transformer架构的RNA二级结构预测方法

喻定, 李章维

引用本文

喻定, 李章维. 基于Transformer架构的RNA二级结构预测方法[J]. 计算机科学, 2026, 53(3): 375-382.

YU Ding, LI Zhangwei. Prediction Method of RNA Secondary Structure Based on Transformer Architecture [J]. Computer Science, 2026, 53(3): 375-382.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于深度学习的GIFT-128与ASCON算法神经差分区分器研究

Deep Learning-based Neural Differential Distinguishers for GIFT-128 and ASCON

计算机科学, 2026, 53(3): 453-458. https://doi.org/10.11896/jsjkx.250600176

基于Transformer的域自适应物联网流量入侵检测方法

Transformer-based Domain Adaptation Method for IoT Traffic Intrusion Detection

计算机科学, 2026, 53(3): 443-452. https://doi.org/10.11896/jsjkx.241200167

基于对比学习的双通道源代码漏洞检测模型

Dual-channel Source Code Vulnerability Detection Model Based on Contrastive Learning

计算机科学, 2026, 53(3): 424-432. https://doi.org/10.11896/jsjkx.250200124

基于多任务学习的眼科视频特征融合与多维画像

Multi-task Learning-based Ophthalmic Video Feature Fusion and Multi-dimensional Profiling

计算机科学, 2026, 53(3): 383-391. https://doi.org/10.11896/jsjkx.260200058

基于少量目标数据和深度学习的行人重识别方法

Pedestrian Re-identification Methods Based on Limited Target Data and Deep Learning

计算机科学, 2026, 53(3): 287-294. https://doi.org/10.11896/jsjkx.260100073

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250100005

基于 Transformer 架构的 RNA 二级结构预测方法

喻　定　　李章维
浙江工业大学信息工程学院　杭州 310023
(2112003059@zjut.edu.cn)

摘　要　RNA 二级结构预测是生物信息学中的核心问题,近年来,深度学习技术的发展为该领域带来了显著进步。然而,现有方法在预测精度和对外部先验模型的依赖性方面仍存在不足,这些限制可能对模型的鲁棒性和泛化能力造成影响。针对上述问题,提出了一种基于 Transformer 架构的 RNA 二级结构预测模型。该模型设计了两条特征编码通路,通过线性嵌入和独热编码生成序列特征,并利用交叉注意力机制高效融合两种特征表示。在特征提取阶段,模型采用改进的 Swin-Transformer 与 U-Net 相结合的架构(Swin-UNet),实现深层次特征提取,并最终生成 RNA 二级结构配对概率矩阵。实验结果表明,该模型在多个标准数据集上的 F1 得分领先了其他模型 3%以上,且无须依赖外部模型的先验信息。研究结果为 RNA 结构预测提供了新的解决方案,同时展现了 Transformer 架构在生物序列分析中的广阔前景。

关键词:RNA 二级结构预测;深度学习;Swin-Transformer;交叉注意力;U-Net
中图分类号　TP389

Prediction Method of RNA Secondary Structure Based on Transformer Architecture

YU Ding and LI Zhangwei
College of Information Engineering,Zhejiang University of Technology,Hangzhou 310023,China

Abstract　RNA secondary structure prediction is a core problem in bioinformatics,and recent advancements in deep learning have significantly propelled progress in this field. However,existing methods still face limitations in prediction accuracy and reliance on external prior models,which may compromise the robustness and generalization capabilities of these models. To address these issues,this paper proposes a Transformer-based model for RNA secondary structure prediction. The model designs dual feature encoding pathways,generating sequence features through linear embedding and one-hot encoding,and efficiently fuses these two feature representations using a cross-attention mechanism. During the feature extraction phase,the model employs an improved architecture combining Swin-Transformer and U-net(Swin-Unet) to achieve deep-level feature extraction,ultimately producing a pairing probability matrix for RNA secondary structures. Experimental results show that the proposed model achieves over 3% higher F1-scores than other models on multiple benchmark datasets without relying on prior information from external models. This study provides a novel solution for RNA structure prediction and highlights the promising potential of Transformer architectures in biological sequence analysis.

Keywords　Prediction of RNA secondary structure,Deep learning,Swin-Transformer,Cross-attention,U-Net

1　引言

核糖核酸(RNA)作为生命活动中的关键分子,不仅参与从 DNA 到蛋白质的遗传信息传递,还在细胞调控、基因表达等多个生物学过程中发挥重要作用[1]。RNA 分子由腺苷酸(AMP)、胞苷酸(CMP)、鸟苷酸(GMP)和尿苷酸(UMP)4 种核苷酸通过核糖和磷酸基团形成的磷酸二酯键连接而成,其中 RNA 二级结构是由碱基对(主要是 A-U 和 C-G)通过氢键配对,形成茎环、假结等局部结构单元。RNA 的功能与其结构密切相关[2],准确预测 RNA 的二级结构对帮助研究人员理解其生物学功能[3]、开发 RNA 靶向药物[4]等具有重要的价值。然而,传统的实验方法(如 X 射线晶体衍射[5]和冷冻电镜[6]等)虽然可靠,但往往耗时长,成本高,且需要专业的设备和人员支持,这促使研究者转向研究计算机辅助的结构预测方法。

在 RNA 结构预测领域,早期研究主要基于最小自由能(MFE)原理,试图找到在符合热力学原理的基础上最合理与最稳定的结构。例如,Sloma 等[7]基于动态规划方法提出了热力学模拟算法,Ding 等[8]则从统计力学角度探索了中短链 RNA 结构预测问题。一些研究人员也尝试以 MFE 为约束

到稿日期:2025-01-02　返修日期:2025-04-03
基金项目:国家自然科学基金(61573317)
This work was supported by the National Natural Science Foundation of China(61573317).
通信作者:李章维(lzw@zjut.edu.cn)

条件，使用群体优化算法来预测 RNA 的结构。例如，Shapiro 等[9]与 Chen 等[10]相继提出了基于遗传算法(GA)的 RNA 预测方法，Geis 等[11]则提出使用粒子群优化(PSO)来实现 RNA 二级结构的预测。这些传统方法虽然取得了一定成果，但在预测精度和计算效率上仍存在瓶颈，特别是群体优化算法往往面临较高的计算复杂度。随着机器学习技术的发展和 RNA 结构数据的积累，基于机器学习的预测方法在过去一段时间内成为研究热点。例如，Do 等[12]提出了一种名为 Contrafold 的方法，该方法结合了条件对数线性模型(CLLM)与下文无关文法(SCFG)，实现了基于概率的 RNA 二级结构建模；Zakov 等[13]提出的 Contextfold 模型，通过大幅增加机器学习参数量，进一步提高了 RNA 结构的预测正确率。但是，这些基于机器学习的算法依然存在过于依赖特征工程，对长序列 RNA 的复杂结构预测失效等缺点。

在深度学习方法蓬勃发展的背景下，RNA 结构预测领域也出现了多项突破性的进展，一些研究者尝试将神经网络模型与优化算法结合。例如，Zhang 等[14]提出的 CDPfold 模型，使用卷积神经网络(CNN)预测 RNA 序列中各个碱基的状态分布，并采用动态规划(DP)寻找最大概率的二级结构；Quan 等[15]提出的 DpacoRNA 模型，以 Bi-LSTM 作为特征提取器，并使用并行蚁群优化方法(ACO)来预测 RNA 结构。而随着计算机性能的提升，当前研究人员都希望通过训练一些具有百万甚至千万参数的超大型神经网络模型来进一步提高 RNA 结构预测的精度。例如，Chen 等[16]借鉴了 AlphaFold在蛋白质结构预测中的成功经验，提出了将卷积神网络与注意力机制(Attention)结合的 E2Efold 模型，通过多尺度特征提取来预测 RNA 结构；Singh 等[17]提出了一种基于迁移学习的 SPOT-RNA 模型，该模型在泛化性能上表现尚可，但在计算复杂度上仍存在很大的优化空间；Fu 等[18]在 2022 年提出的 UFold，其通过独热编码将 RNA 序列转换为二维矩阵，并使用 U-Net 来提取特征，取得了当时最好的预测性能；随后，Yang 等[19]基于 UFold 框架，引入图神经网络(GNN)，设计了 GCNfold 模型，通过捕获核苷酸之间的拓扑关系进一步提升了 RNA 二级结构预测的精度。然而，这两种方法都依赖外部先验模型来辅助预测。其中，UFold 使用先验预测来缓解 RNA 序列转换为矩阵时产生的数据稀疏问题；而 GCNfold 则需要先验信息来初始化 RNA 拓扑图结构。由于这些先验模型输出的 RNA 二级结构预测往往存在较大的不确定性，其低保真度特性会传播并影响最终的预测精度。

针对上述问题，本文提出了一种基于 Transformer 架构的新方法。该方法通过交叉注意力机制融合序列嵌入和矩阵表示两种编码方式，并使用基于 Swin-Transformer 的 U-Net 进行深层特征提取。在 RNAStralign 与 ArchiveII 数据集上的实验与模型对比结果证明，本文模型无须依赖外部先验信息就能实现更高精度的 RNA 二级结构预测。

2 RNA 二级结构预测模型

本文模型使用基于 Swin-Transformer 架构的 U-Net 来对 RNA 的二级结构进行预测。模型总体框架如图 1 所示。

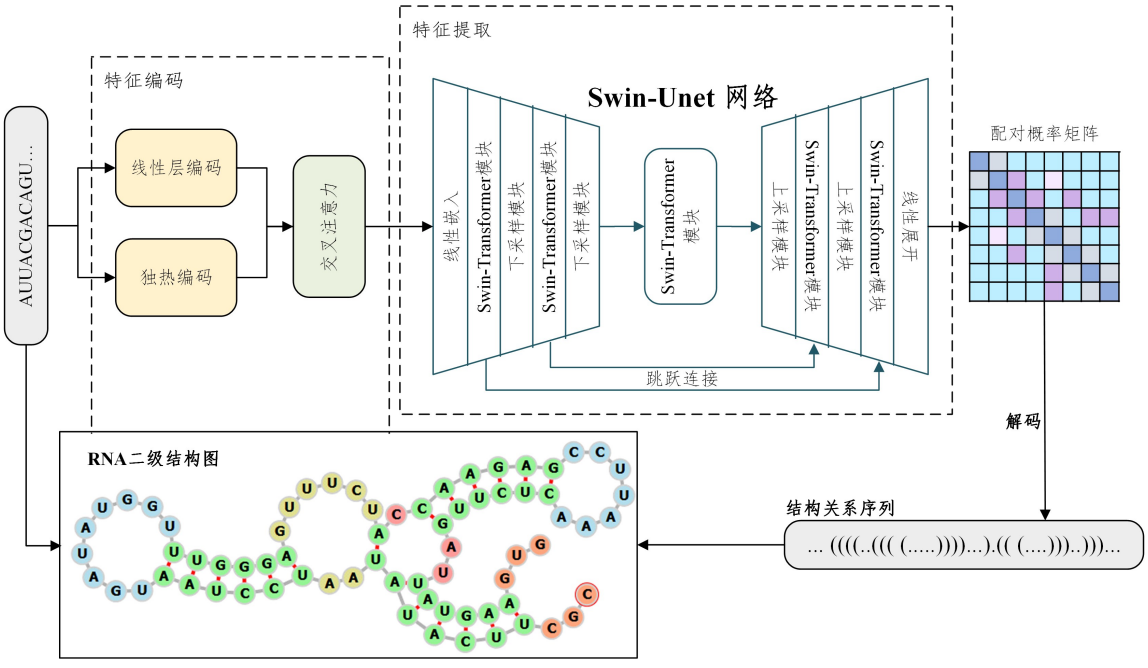


图 1 模型的整体框架图

Fig. 1 Overall framework diagram of the proposed model

2.1 RNA 序列特征编码

考虑到深度学习模型在图像领域的出色表现[20-21]，将输入的 RNA 序列转换为与图像类似的多通道二维矩阵后再进行特征提取。输入长度为 L 的 RNA 序列 $S=\{s_1, s_2, \cdots, s_L\}$，其由 A，U，G，C 这 4 种不同的碱基组成，本文采用了线性层嵌入编码与独热编码两种策略，具体过程如图 2(a)与图 2(b)所示。

线性层嵌入编码策略首先将 RNA 序列中的字母(Char)映射为可被机器理解的整数(Int)。然后，使用神经网络的线性层对整数序列进行两次嵌入，包括一次行嵌入与一次列嵌入，以从单维度的 RNA 序列生成二维的潜在

特征。该过程可表示为：

$$\boldsymbol{E}_{\text{row}}=Embedding_{\text{row}}(S)$$
$$\boldsymbol{E}_{\text{col}}=Embedding_{\text{col}}(S) \tag{1}$$

其中，$\boldsymbol{E}_{\text{row}}\in\mathbb{R}^{L\times d}$ 和 $\boldsymbol{E}_{\text{col}}\in\mathbb{R}^{L\times d}$ 分别为行嵌入与列嵌入后的特征，d 为嵌入层的维度数。行嵌入 $\boldsymbol{E}_{\text{row}}$ 捕捉序列中每个位置的局部上下文信息，而列嵌入 $\boldsymbol{E}_{\text{col}}$ 则通过转置操作捕捉序列全局的上下文信息。接着，将这两个矩阵叠加形成一个二维潜空间：

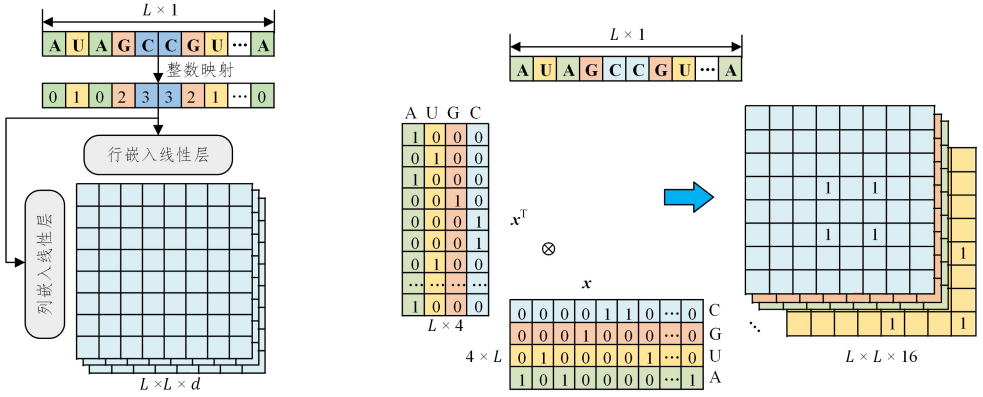
$$\boldsymbol{F}_1=\boldsymbol{E}_{\text{row}}\oplus\boldsymbol{E}_{\text{col}}^{\text{T}} \tag{2}$$

其中，$\boldsymbol{F}_1\in\mathbb{R}^{L\times L\times 2d}$ 为线性层嵌入编码策略最终得到的特征，其大小为 $L\times L$，通道数为 $2d$；$\oplus$ 为矩阵的广播加法运算（Broadcasting），具体法则为 $\boldsymbol{F}_1[i,j]=\boldsymbol{E}_{\text{row}}[i]+\boldsymbol{E}_{\text{col}}^{\text{T}}[j]$，$i,j\in[0,L]$。通过横纵信息的交互，$\boldsymbol{F}_1$ 中融合了 RNA 序列的局部细节特征与全局结构特征，既提升了模型对长距离碱基配对关系的建模能力，又保留了对局部序列细节的关注。

独热编码策略首先将长度为 L 的 RNA 序列 S 编码为 $L\times 4$ 的二进制矩阵 $\boldsymbol{X}\in\{0,1\}^{L\times 4}$，其中的 4 列代表 4 种不同的碱基，当该位置为某一碱基时，对应列标记为 1，其余 3 个位置标记为 0；然后通过与自身之间的克罗内克积（用⊗符号表示），将 $\boldsymbol{X}$ 转换为二维特征：

$$\boldsymbol{F}_2=\boldsymbol{X}\otimes\boldsymbol{X} \tag{3}$$

其中，$\boldsymbol{F}_2\in\mathbb{R}^{L\times L\times 16}$ 中每个通道指定 16 个可能的碱基配对规则之一，$\boldsymbol{F}_2[i,j,k]$ 表示碱基 $\boldsymbol{X}[i]$ 和 $\boldsymbol{X}[j]$ 是否按照第 k 条碱基配对规则配对，如 $k=2$ 的矩阵就记录了所有 A-C 出现配对的位置。该方法通过克罗内克积生成的二维特征，捕捉局部碱基配对规则和序列对齐信息，$\boldsymbol{F}_2$ 中记录着 RNA 序列高精度的局部特征。



(a)线性层嵌入编码策略示意图　　(b) 独热编码策略示意图

图 2　特征编码策略示意图

Fig. 2　Schematic diagram of feature encoding strategies

2.2　基于交叉注意力机制的特征融合

在通过上述的 RNA 序列特征编码后，得到了分别来自线性层嵌入编码的特征 $\boldsymbol{F}_1$ 与独热编码的特征 $\boldsymbol{F}_2$。为了整合这两种不同编码方式下的特征信息，本文采用交叉注意力机制进行特征融合，该过程具体如图 3 所示。

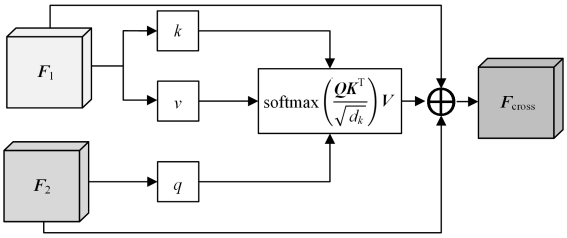


图 3　交叉注意力模块示意图

Fig. 3　Schematic diagram of cross-attention module

交叉注意力与自注意力[22]的计算方法相同，但该机制的核心思想是让两个模态的特征相互学习和补充。与直接将特征拼接或加权平均相比，计算两个模态间的注意力权重的方法能够更精细地选择和对齐相关信息，提升融合特征的表征能力。

$$\boldsymbol{F}_{\text{cross}}=Cross\text{-}Attention(\boldsymbol{F}_1,\boldsymbol{F}_2)=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{4}$$

其中，键矩阵 $\boldsymbol{K}=\boldsymbol{F}_2\boldsymbol{W}_K$ 与值矩阵 $\boldsymbol{V}=\boldsymbol{F}_2\boldsymbol{W}_V$ 由线性层嵌入编码特征 $\boldsymbol{F}_1$ 主导，而查询矩阵 $\boldsymbol{Q}=\boldsymbol{F}_1\boldsymbol{W}_Q$ 则受独热编码特征 $\boldsymbol{F}_2$ 影响，$\sqrt{d}$ 为缩放因子，最终得到交叉融合之后的特征 $\boldsymbol{F}_{\text{cross}}$。

2.3　融合 Swin-Transformer 改进的 U-Net

受 UFold 将 U-Net 架构[23]成功应用于 RNA 二级结构预测任务的启发，本文设计了一种融合 Swin-Transformer 的改进型 U-Net 模型（Swin-UNet），具体如图 4 所示。该模型保留了 U-Net 的经典架构，包含编码器、解码器、瓶颈层和跳跃连接 4 个关键组件，改进点在于将传统 CNN 替换为在当前的计算机视觉领域表现出色的 Swin-Transformer[24]，以进一步增强模型的特征提取能力。

在 Swin-UNet 中，编码器利用 Swin-Transformer 模块实现对 RNA 序列特征的多尺度层次化提取，解码器采用对称结构进行特征重建，瓶颈层通过独立的 Swin-Transformer 模块执行特征压缩与增强，同时引入跳跃连接机制实现高低层级之间的特征融合，以防止出现梯度消失。

Swin-Transformer 模块是 Swin-Unet 网络的基本单元，由层归一化（LN）、基于多头窗口的自注意（WSA）单元、基于多头滑动窗口的自注意（SWSA）单元以及具有 GELU 激活函数[25]的全连接（MLP）层组成。该模块可以表示如下：

$$\hat{z}^{l}=WSA(LN(z^{l-1}))+z^{l-1} \tag{5}$$
$$z^{l}=MLP(LN(\hat{z}^{l}))+\hat{z}^{l} \tag{6}$$

$$\hat{z}^{l+1}=SWSA(LN(z^{l}))+z^{l} \tag{7}$$

$$z^{l+1}=MLP(LN(\hat{z}^{l+1}))+\hat{z}^{l+1} \tag{8}$$

其中，$\hat{z}^{l}$代表 WSA 的输出，$\hat{z}^{l+1}$代表 SWSA 的输出，z^{l}代表第一个模块中 MLP 的输出，z^{l+1}代表第二个模块中 MLP 的输出。

这样的架构虽然被证实能够有效提高模型对图像特征提取的能力，但伴随着一个显著的问题，即 SWSA 将导致出现长度不统一的窗口。为保证高效率的计算，进行循环移位，这样批处理窗口将保持与常规窗口划分相同的数量，从而可以采用掩码(Mask)方法添加相对位置偏置 $\boldsymbol{B}\in\mathbb{R}^{M\times M}$，以将自注意力计算限制在每个子窗口内。因此，改进后的注意力计算式为：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V},\boldsymbol{B})=\mathrm{Softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d}}+\boldsymbol{B}\right)\boldsymbol{V} \tag{9}$$

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbb{R}^{M\times d}$的含义与式(4)中的相同。需要额外注意的是，对于相对位置偏置矩阵 $\boldsymbol{B}$ 中的任意元素$b_{i,j}$，若 i 与 j 位于同一子窗口，则$b_{i,j}=0$；若 i 与 j 位于不同子窗口，则 $b_{i,j}\to-\infty$。在经历 Softmax 函数之后，不同子窗口之间的注意力权重将趋于 0，而相同子窗口的注意力权重将以较大的占比得以保留。

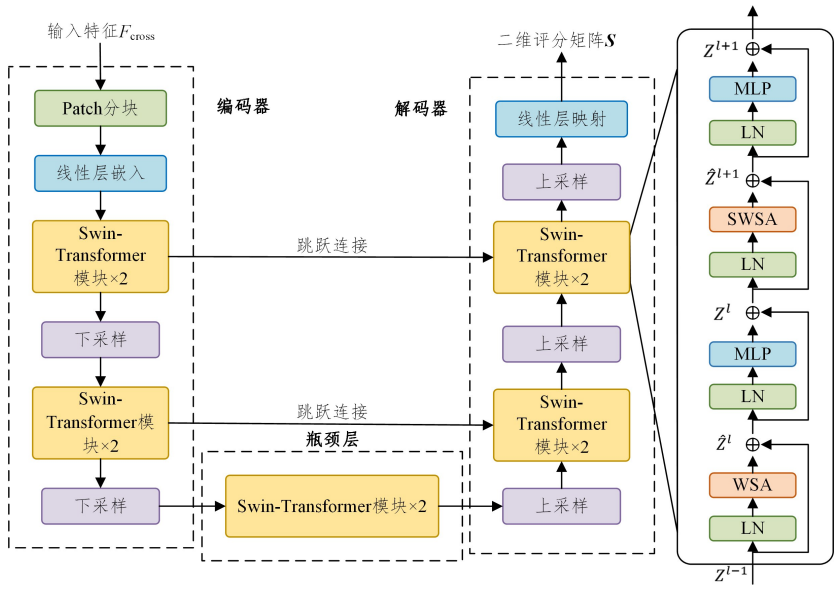


图 4 Swin-Transformer 改进的 U-Net

Fig. 4 U-Net improved by Swin-Transformer

2.4 RNA 二级结构解码

在通过 Swin-UNet 模型完成特征提取后，将输出一个 $L\times L$大小的二维评分矩阵 $\boldsymbol{S}$，其中 L 为输入的 RNA 序列长度，矩阵中的每个元素$S_{i,j}$表示第 i 个核苷酸与第 j 个核苷酸形成碱基对的可能性得分。然而，考虑到这个原始评分矩阵可能包含与 RNA 分子结构物理特性相违背的预测结果，本文设计了一个基于多重约束优化的解码模块，以期望最终预测的 RNA 二级结构在符合生物学规则的前提下，能最大程度地与模型的预测结果一致。

首先，根据 RNA 分子的物理特性定义了 4 个关键的硬约束条件。

(1)碱基配对的对称性约束：要求如果核苷酸 i 与 j 配对，则 j 也必须与 i 配对。

(2)碱基互补配对约束：只允许 A-U 和 C-G 这两种规范配对以及可能存在的 G-U 非规范配对方式。

(3)结构稳定性约束：规定任意发生配对的核苷酸之间的最小序列距离不得小于 4。

(4)单一配对约束：规定每个核苷酸最多只能与一个其他核苷酸形成配对。

为了将这些约束整合到解码过程中，构建一个约束矩阵 $\boldsymbol{M}(x)$，用于编码碱基互补配对约束和结构稳定性约束。矩阵 $\boldsymbol{M}(x)$中的元素定义如下：

$$\boldsymbol{M}(x)_{ij}=\begin{cases}1, & \text{if } x_i\text{与}x_j\text{配对 and } |i-j|\geqslant 4\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

同时，通过以下变换操作确保评分矩阵满足对称性约束：

$$\boldsymbol{T}(\boldsymbol{S})=\frac{1}{2}(\boldsymbol{S}+\boldsymbol{S}^{\mathrm{T}})\circ\boldsymbol{M}(x) \tag{11}$$

其中，$\circ$表示哈达玛(Hadamard)积运算，$\boldsymbol{T}(\boldsymbol{S})$即为满足对称性和碱基配对约束的修正评分矩阵。然而，这个修正后的矩阵仍可能违反单一配对约束。为了解决这个问题，本文将其形式化为一个带约束的优化问题，即：

$$\begin{cases}\hat{\boldsymbol{S}}^{*}=\underset{\hat{\boldsymbol{S}}\in\mathbb{R}^{L\times L}}{\mathrm{argmax}}\langle\hat{\boldsymbol{S}},\boldsymbol{T}(\boldsymbol{S})\rangle-\rho|\hat{\boldsymbol{S}}|_1\\ \text{s.t. } \hat{\boldsymbol{S}}1\leqslant 1\end{cases} \tag{12}$$

其中，目标函数中的$\langle\boldsymbol{S},\boldsymbol{T}(\boldsymbol{S})\rangle$用来度量网络输出的预测矩阵$\hat{\boldsymbol{S}}$与修正评分矩阵 $\boldsymbol{T}(\boldsymbol{S})$的一致性；$|\hat{\boldsymbol{S}}|_1$为$\hat{\boldsymbol{S}}$的 L1 正则化项，用于控制预测结构的稀疏度；$\rho$ 为权重参数；约束条件$\hat{\boldsymbol{S}}1\leqslant 1$ 确保二维矩阵的每一行或列的值之和的最大值为 1。

通过求解这个优化问题，得到的$\hat{\boldsymbol{S}}^{*}$是一个满足所有约束条件的概率矩阵。最后，通过设定一个经验阈值 θ，将$\hat{\boldsymbol{S}}^{*}$转换为二值矩阵，即最终的 RNA 二级结构预测结果$\boldsymbol{S}^{*}$中每个位置元素$s^{*}_{i,j}$为：

$$s_{i,j}^{*}=\begin{cases}1, & \text{if } s_{i,j}^{*}>\theta\\ 0, & \text{otherwise}\end{cases} \tag{13}$$

2.5　损失函数设计

RNA 二级结构预测任务的本质是判断序列中任意两个核苷酸位点是否形成碱基配对的二分类问题。考虑到 RNA 序列中实际形成配对的核苷酸对数量远少于非配对核苷酸对，存在显著的类别不平衡，本文采用加权二元交叉熵损失函数来优化模型，即：

$$\mathcal{L}=-\frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L}\left[w_p s_{i,j}\log\left(\hat{s}_{i,j}\right)+w_n\left(1-s_{i,j}\right)\log\left(1-\hat{s}_{i,j}\right)\right] \tag{14}$$

其中，L 为 RNA 序列的长度，$s_{i,j}$ 表示位置 i 和 j 的核苷酸对的真实标签（1 表示配对，0 表示不配对），$\hat{s}_{i,j}$ 为模型预测的配对概率，w_p 和 w_n 分别为正样本（代表碱基发生配对）和负样本（代表碱基没有发生配对）的权重系数。

$$w_p=\frac{N_n}{N_p+N_n} \tag{15}$$

$$w_n=\frac{N_p}{N_p+N_n} \tag{16}$$

其中，N_p 和 N_n 分别表示数据集中配对和非配对样本的数量。

3　实验

本文实验在 PC 端进行，CPU 为 13th Gen Intel(R) Core(TM) i5-13600K，GPU 为 Nvidia GeForce GTX 3090，操作系统为 Linux 的 ubuntu 22.04，编程语言为 Python 3.9，选择 PyTorch 2.1.0 深度学习框架，搭配 Torchvision 0.16.0 视觉工具库，使用 CUDA 11.8 作为 GPU 加速运算平台。

3.1　数据集的构建与划分

研究使用 RNAStralign[26] 与 ArchiveII[27] 两个 RNA 二级结构预测标准数据集。其中，RNAStralign 数据集包含 16S rRNA、5S rRNA、信号识别颗粒 RNA 以及转运-信使 RNA 等 8 种重要类型，总计 30451 条 RNA 序列，序列长度分布在 31 至 1851 个核苷酸之间。ArchiveII 数据集则涵盖 10 种 RNA 类型，包含 23S rRNA 和Ⅱ组内含子（grp2）等 RNA 结构，共计 3975 条序列，序列长度范围为 28 至 2968 个核苷酸。

本文在实验环节采用了分层随机抽样的方式对 RNAStralign 数据集进行划分，按照 8∶2 的比例将数据集划分为训练集和测试集，并确保各个子集中 RNA 类型的占比基本一致，而整个 ArchiveII 数据集则作为非同源测试集。

3.2　评价指标

本文采用多个互补的评价指标来评估 RNA 二级结构预测模型的性能。首先定义真正例（TP）表示该核苷酸对在真实结构和预测结构中均形成配对，真负例（TN）表示在两个结构中均不配对，假正例（FP）表示预测为配对但实际不配对，假负例（FN）表示预测为不配对但实际配对。

准确率反映预测配对中正确的比例：

$$Precision=\frac{TP}{TP+FP} \tag{17}$$

召回率则衡量真实配对被正确预测的比例：

$$Recall=\frac{TP}{TP+FN} \tag{18}$$

考虑到准确率和召回率往往存在权衡关系，采用 F1 得分作为平衡两者的综合指标：

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{19}$$

此外，由于 RNA 二级结构预测中存在非配对核苷酸对远多于配对核苷酸对的类别不平衡问题，因此额外引入马修斯相关系数（MCC）作为更严格的评价标准。

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{20}$$

MCC 的取值范围为[−1，1]，其特点是只有当模型在所有混淆矩阵元素（TP，TN，FP，FN）上都表现良好时才能获得较高分数，因此能更全面地反映模型的预测性能。

3.3　模型对比

将本文模型与多种传统及深度学习模型在 RNA 二级结构预测任务中的性能评估结果进行了对比，对比模型包括基于最小自由能经典理论与概率上下文无关文法的模型 Contrafold[12]，结合上下文特征和最小自由能方法的混合模型 Contextfold[13]，最早被行业广泛使用的 RNA 二级结构预测软件 Mfold[28]，2020 年提出的端到端一体化的深度学习模型 E2Efold[16]，2022 年提出的基于深度学习但依赖于 CDPFold 提供的先验推理结果的模型 UFold[18]，以及 2024 年提出的在编码器中融合多头自注意力机制的模型 Wfold[29]。

3.3.1　RNAStralign 数据集上的实验结果

在 RNAStralign 数据集上的模型对比结果如表 1 所列。可以看到，基于最小自由能优化的传统方法的准确率与召回率都普遍偏低，而基于深度学习的方法在 RNA 二级结构预测任务中明显更具优势。对比当前较为先进的 UFold 模型，本文模型在没有依赖外部模型提供先验推理的情况下，准确率提高了 4.2%，召回率提高了 4.8%，F1 得分提高了 4.6%，MCC 提高了 4.6%，即在所有评价标准上均超过了所有以前的方法。

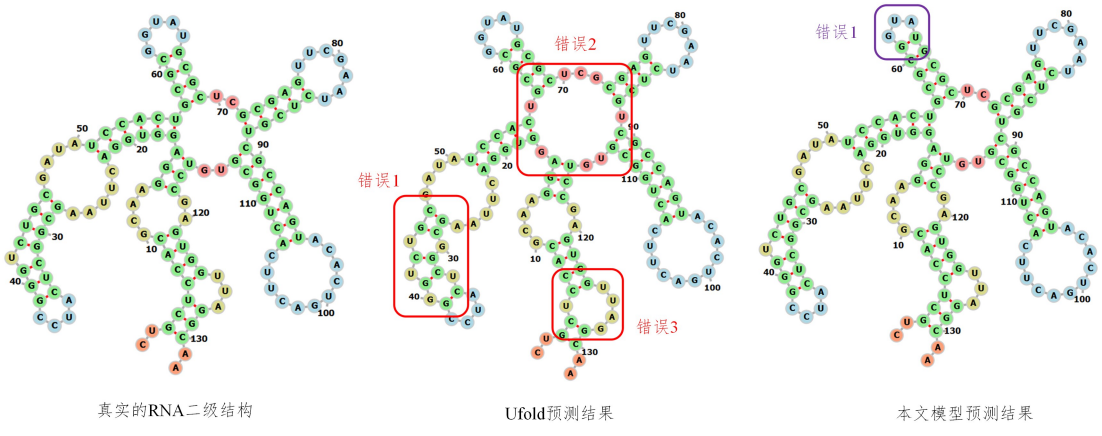
表 1　RNAStralign 数据集上各模型性能的对比

Table 1　Comparison of the performance of various models on the RNAStralign dataset

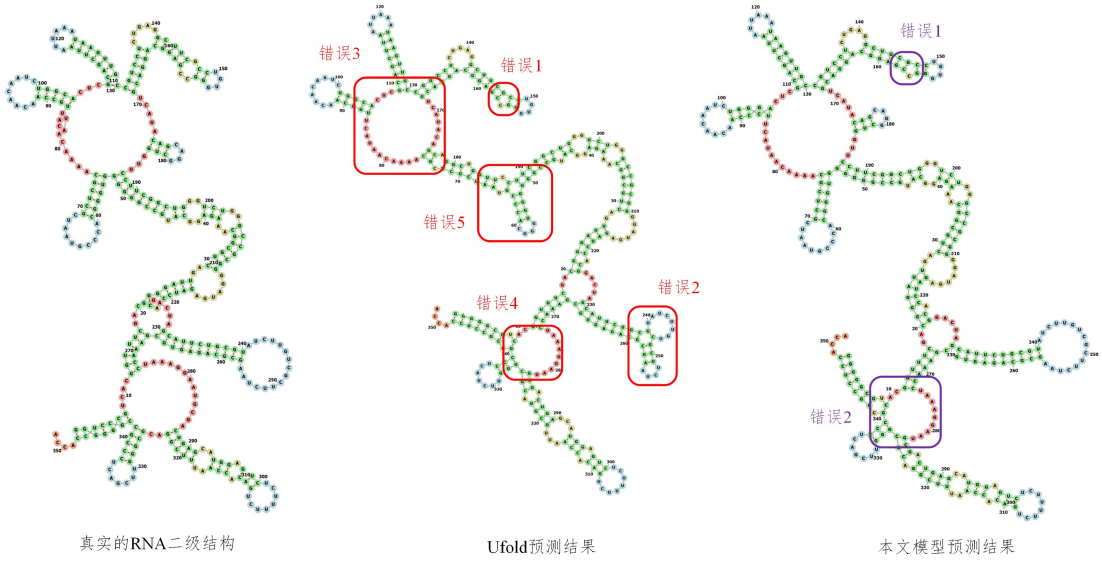
模型	准确率	召回率	F1 得分	MCC	参数
Contrafold	0.664	0.565	0.611	0.476	—
Contextfold	0.718	0.627	0.670	0.537	—
Mfold	0.676	0.560	0.613	0.498	—
E2Efold	0.832	0.803	0.817	0.752	1.865×10^{7}
WFold	0.872	0.796	0.832	0.788	1.978×10^{7}
UFold	0.861	0.836	0.848	0.786	$\mathbf{8.630\times10^{6}}$
Ours	**0.903**	**0.884**	**0.894**	**0.832**	1.257×10^{7}

此外，本文模型中由于融合了计算复杂度比较高的 Transformer 架构，因此参数量高于基准模型 UFold，但低于 E2Efold 和 WFold 的参数量。在使用 NVIDIA 3090 GPU 进行本地部署后，本文模型的平均推理时间花销为 0.17 s，能够满足实际应用需求。

为了直观地展示本文模型在 RNA 二级结构预测任务上的性能，利用 ViennaRNA 软件[30] 对预测结果进行了二级结构可视化，具体如图 5 所示。



(a)短链 RNA 预测对比



(b)长链 RNA 预测对比

图 5 本文模型与 Ufold 模型在短链 RNA 与长链 RNA 上的预测结果可视化对比(电子版为彩图)

Fig. 5 Visual comparison of the prediction results between the proposed model and the Ufold model on sRNA and LRNA

本文选取了一个长度为 132 的短链核糖体 RNA 序列样本与一个长度为 353 的长链转移信使 RNA 序列样本,来展示本文模型的预测结果、UFold 模型的预测结果以及真实结构的比较,并用红色框与紫色框强调出现识别错误的区域。如图 5(a)所示的短链 RNA 预测可以看到,在错误 1 红框中,UFold 模型由于多次未能识别到 U-G 配对的情况,茎环结构出现了严重失真;在错误 2 红框中则可以看到,UFold 模型对 RNA 环状结构的长度预测偏大,这些预测错误的情况在其他样本里也时有发生,而本文所提出的这套基于 Transformer 的 RNA 二级结构预测模型的输出则更接近原本的真实结构,仅出现如紫色框所标注的一处配对错误。如图 5(b)所示的长链 RNA 预测可以看到,UFold 模型预测出现错误的情况显著增多,包括未能识别出环状结构以及环状结构大小判断错误等;而本文模型的预测结果更贴近真实长链 RNA 结构,相比于 UFold 模型,其识别错误率更低,尤其在环状结构、茎环结构及非典型碱基配对的预测上表现出更强的能力。

3.3.2 ArchiveII 数据集上的实验结果

为了严格评估模型的泛化能力,并验证其在实际 RNA 结构预测场景中的应用潜力,本文采用了跨数据集测试策略,将 RNAStralign 数据集上训练的模型直接应用于独立的 ArchiveII 数据集,且为确保评估的公平性,预先剔除了两个数据集间的重叠序列。这种评估方式也有助于验证模型是否存在过拟合现象。各模型的性能测试结果如表 2 所列。

表 2 ArchiveII 数据集上各模型性能的对比

Table 2 Comparison of the performance of various models on the ArchiveII dataset

模型	准确率	召回率	*F*1 得分	*MCC*
Contrafold	0.681	0.604	0.640	0.499
Contextfold	0.693	0.616	0.652	0.511
Mfold	0.718	0.643	0.677	0.532
E2Efold	0.710	0.652	0.681	0.560
UFold	0.818	0.794	0.806	0.705
Wfold	0.859	0.821	0.829	0.767
Ours	**0.882**	**0.859**	**0.870**	**0.793**

实验结果表明,基于最小自由能优化的传统方法在两个数据集上表现相对稳定,这主要归因于其优化策略依赖于固定的热力学参数,而非数据驱动的学习过程。相比之下,深度学习方法在跨数据集测试中普遍出现性能下降;然而,本文模型还是展现出了优异的鲁棒性,相较于 UFold 模型的准确率

上从 0.861 大幅降至 0.818(降幅 5.0%)，本文模型的准确率仅从 0.903 小幅下降至 0.882(降幅 2.3%)。此外，WFold 模型虽然在切换陌生的 RNA 家族后 MCC 得分只从 0.788 下降到 0.767(降幅 2.66%)，略优于本文模型，但评估的整体得分都落后于本文模型。这一结果不仅展示了本文模型在 RNA 二级结构预测任务中的优秀性能，还凸显出了出色的泛化能力。

此外，图 6 绘制了 4 种深度学习模型的 F1 得分分布散点图，每一个 RNA 序列样本都在图上用一个点表示。可以看到，本文模型在 F1 得分分布上呈现集中趋势，且主要聚集在 1.0 分值附近。这种分布特征有力地证实了模型在处理多样化 RNA 序列时具有稳定且高效的预测能力，体现出优异的算法鲁棒性和预测一致性。

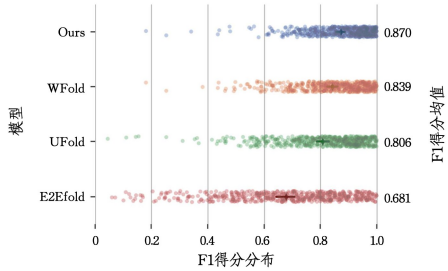


图 6　不同模型在 ArchiveII 数据集上的 F1 得分散点分布图

Fig. 6　Scatter plot of F1 scores for different models on the ArchiveII dataset

3.4　消融实验

为了系统评估模型各组件的有效性和贡献度，本节开展了详细的消融实验。以 UFold 作为基准模型，在其基础上实现了 3 个关键的改进点，包括：(1)引入新的序列线性层编码方法；(2)采用交叉注意力机制进行特征融合；(3)将基于 CNN 的 U-Net 替换为基于 Transformer 的 Swin-Unet。分别去除这 3 点改进，并重新进行模型训练与测试，在 RNAStralign 数据集上的表现结果如表 3 所列。

表 3　模型消融实验结果

Table 3　Results of the model ablation experiments

消融方法	准确率	召回率	*F1* 得分	*MCC*
UFold	0.861	0.836	0.848	0.786
去除(1)	0.891	0.873	0.882	0.821
去除(2)	0.897	0.877	0.887	0.825
去除(3)	0.871	0.853	0.862	0.800
无去除	**0.903**	**0.884**	**0.894**	**0.832**

可以看到，移除任一改进点都会导致模型性能下降，从而证明了各个改进模块的有效性。其中，序列线性层编码方法能够捕捉序列中的全局上下文信息和潜在的结构模式，特别是在映射长链 RNA 碱基之间的关系中发挥了重要作用，使整体模型的准确率提高了 0.012。交叉注意力机制实现了线性层嵌入编码和独热编码特征的自适应融合，提升了特征的表征能力，使整体模型准确率提高了 0.006。而将 U-Net 替换为 Swin-Unet 是本文最重要的改进，引入的分层 Transformer 结构能够处理 RNA 序列中的长程依赖和局部信息，对模型性能的贡献程度最大，使整体模型准确率提高了 0.032。

结束语　本文提出了一种基于 Transformer 的深度学习模型用于 RNA 二级结构预测。与现有方法不同，模型设计了双通路特征编码策略，同时利用线性嵌入编码和独热编码二维矩阵来捕获 RNA 序列的多维特征表示，并通过交叉注意力机制实现了两种编码的融合。模型还创新性地将 Swin-Transformer 引入改进的 U-Net 架构中，增强了模型对 RNA 序列长程依赖关系的建模能力。实验结果表明，本文模型在预测性能上优于现有方法，尤其在泛化能力方面表现突出。本文不仅为 RNA 二级结构预测提供了新的解决方案，也展示了 Transformer 架构在生物序列分析任务中的应用潜力，为该领域的深度学习方法的发展提供了新的研究思路。

参 考 文 献

[1] SUN J, ZHANG J, FANG X Y. Research progress on RNA high-order structure determination[J]. Chemistry of Life, 2024, 44(9): 1638-1649.

[2] ZENG C W, ZHAO Y J. Advances in RNA-protein structure prediction[J]. Science China Physics, Mechanics & Astronomy, 2023, 53(9): 222-232.

[3] CHEN Z R, HUANG J H, LI B, et al. Applications of Computational Biology inRNA Research[J]. Science China: Life Sciences, 2024, 54(4): 668-693.

[4] DONG Y Y, PENG Q, WANG M, et al. Research progress on the replication mechanisms of important human-infecting RNA viruses and polymerase-directed drug development[J]. Biomedical Transformation, 2024, 5(1): 2-11.

[5] LING X Y, LIU R. The impact of X-ray diffraction crystallography on DNA double helix[J]. Emerging Science and Technology, 2023, 2(1): 9-18.

[6] PAN Z L, JIA X Y, SU Z M. Recent advances in RNA cryo-EM structure determination[J]. Science China: Life Sciences, 2024, 54(8): 1424-1438.

[7] SLOMA M F, ZUKER M, MATHEWS D H. Predictive methods using RNA sequences[M]// RNA Structure and Folding. New York: Humana Press, 2014: 27-43.

[8] DING Y, LAWRENCE C E. A statistical sampling algorithm for RNA secondary structure prediction[J]. Nucleic Acids Research, 2003, 31(24): 7280-7301.

[9] SHAPIRO B A, NAVETTA J. A massively parallel genetic algorithm for RNA secondary structure prediction[J]. The Journal of Supercomputing, 1994, 8: 195-207.

[10] CHEN J H, LE S Y, MAIZEL J V. Prediction of common secondary structures of RNAs: a genetic algorithm approach[J]. Nucleic Acids Research, 2000, 28(4): 991-999.

[11] GEIS M, MIDDENDORF M. Particle swarm optimization for finding RNA secondary structures[J]. International Journal of Intelligent Computing and Cybernetics, 2011, 4(2): 160-186.

[12] DO C B, WOODS D A, BATZOGLOU S. CONTRAfold: RNA secondary structure prediction without physics-based models[J]. Bioinformatics, 2006, 22(14): e90-e98.

[13] ZAKOV S, GOLDBERG Y, ELHADAD M, et al. Rich parameterization improves RNA structure prediction[J]. Journal of Computational Biology, 2011, 18(11): 1525-1542.

[14] ZHANG H, ZHANG C, LI Z, et al. A new method of RNA secondary structure prediction based on convolutional neural network and dynamic programming[J]. Frontiers in Genetics, 2019, 10: 467.

[15] QUAN L, CAI L, CHEN Y, et al. Developing parallel ant colonies filtered by deep learned constrains for predicting RNA secondary structure with pseudo-knots[J]. Neurocomputing, 2020, 384: 104-114.

[16] CHEN X, LI Y, UMAROV R, et al. RNA Secondary Structure Prediction By Learning Unrolled Algorithms[C]// Proceedings of the 2020 International Conference on Learning Representations(ICLR). 2020: 1-19.

[17] SINGH J, HANSON J, PALIWAL K, et al. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning[J]. Nature Communications, 2019, 10(1): 5407.

[18] FU L, CAO Y, WU J, et al. UFold: fast and accurate RNA secondary structure prediction with deep learning[J]. Nucleic Acids Research, 2022, 50(3): e14.

[19] YANG E, ZHANG H, ZANG Z, et al. GCNfold: A novel lightweight model with valid extractors for RNA secondary structure prediction[J]. Computers in Biology and Medicine, 2023, 164: 107246.

[20] KANG L, SU Z J. Principle and prospect of image data enhancement technology[J]. Information Technology, 2024(9): 176-185.

[21] LYU J, WANG Z Y. A Survey of Retinal Vessel Segmentation Algorithms Based on Deep Learning[J]. Journal of Chongqing Normal University(Natural Science), 2024, 41(4): 110-125.

[22] PAN X, GE C, LU R, et al. On the integration of self-attention and convolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 815-825.

[23] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]// Medical Image Computing and Computer-assisted Intervention-MICCAI 2015: 18th International Conference. Springer, 2015: 234-241.

[24] 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. 2021: 10012-10022.

[25] HENDRYCKS D, GIMPEL K. Gaussian error linear units (gelus)[J]. arXiv: 1606.08415, 2016.

[26] TAN Z, FU Y, SHARMA G, et al. TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs[J]. Nucleic Acids Research, 2017, 45(20): 11570-11581.

[27] SLOMA M F, MATHEWS D H. Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures[J]. RNA, 2016, 22(12): 1808-1818.

[28] ZUKER M. Mfold web server for nucleic acid folding and hybridization prediction[J]. Nucleic Acids Research, 2003, 31(13): 3406-3415.

[29] YUAN Y, YANG E, ZHANG R. Wfold: A new method for predicting RNA secondary structure with deep learning[J]. Computers in Biology and Medicine, 2024, 182: 109207.

[30] LORENZ R, BERNHART S H, HÖNER ZU SIEDERDISSEN C, et al. ViennaRNA Package 2.0[J]. Algorithms for Molecular Biology, 2011, 6: 1-14.

YU Ding, born in 1999, postgraduate. His main research interest is intelligent information processing.

LI Zhangwei, born in 1967, Ph.D, associate professor. His main research interest is intelligent information processing.

（责任编辑：柯颖）