专题: 生物分子模拟中的机器学习

RNA 扭转角预测的深度学习方法*

欧秀娟 肖奕†

(华中科技大学物理学院, 武汉 430074)

(2023 年 6 月 29 日收到; 2023 年 8 月 2 日收到修改稿)

RNA 分子三级结构模建是分子生物物理学研究的基本问题之一, 对理解 RNA 的功能和设计新的结构有重要意义. RNA 三级结构主要由主链和侧链上的 7 个扭转角确定, 准确预测这些扭转角是 RNA 分子三级结构模建的基础. 目前只有个别采用深度学习模型预测 RNA 分子扭转角的方法, 要用于模建 RNA 分子的三级结构其预测精度还有待进一步提高. 本文提出了一种预测 RNA 分子扭转角的深度学习方法 1dRNA, 采用了考虑相邻核苷酸的卷积模型 (DRCNN) 和考虑全链核苷酸的超长短期记忆模型 (DHLSTM) 两种不同的深度学习模型. 结果显示, 与现有方法相比, 这两种模型都能提高 RNA 分子大部分扭转角的预测精度, DRCNN 预测精度提高在 5% 到 28% 之间, DHLSTM 预测精度提高在 6% 到 15% 之间. 结果还显示, α 和 γ 角是最难预测的, 环区扭转角比螺旋区的扭转角难预测, 模型对预测序列长度的变化不敏感, 模型预测角度与 decoys 的角度偏差可用于模型质量评估.

关键词: RNA 结构, 扭转角预测, 深度学习

PACS: 87.14.gn, 87.15.A–, 87.15.bg **DOI:** 10.7498/aps.72.20231069

1 引 言

RNA 分子三级结构模建是分子生物物理学研究的基本问题之一, 对理解 RNA 的功能和设计新的结构有重要意义[1–3]. RNA 分子三级结构模建是给出 RNA 分子的核苷酸序列构建其三级结构[4–10]. RNA 三级结构可以分为主链结构和侧链结构, 主链结构由螺旋区和环区构成, 由 6 个扭转角 (α, β, γ, δ, ε, ζ) 确定, 侧链方向由扭转角 χ 确定 (图 1). RNA 分子主链和侧链结构还涉及共价键键长和键角, 但这些键长和键角会相对平衡位置进行微振动, 在生理温度这些参数的变化关于平衡位置对称, 影响将相互抵消[11]. 因此, 扭转角被认为是 RNA 分子三级结构的决定因素, 预测这些扭转角可以帮助模建 RNA 分子的三级结构.

扭转角预测在蛋白质分子三级结构模建中已经有深入的研究. 与 RNA 分子不同, 蛋白质分子三级结构主要由主链上的 2 个扭转角 ψ 和 ϕ 确定. 从 2007 年以来, 人们提出了不同的神经网络模型预测扭转角 ψ 和 ϕ. 2007 年, Real-SPINE1.0 使用一层全连接神经网络预测蛋白质主链 ψ 角, 角度的平均绝对误差 (mean absolute error, MAE) 为 54°[12]; 2008 年, Real-SPINE2.0 使用同样神经网络和输入特征, 角度标签 [0°, 180°] 不变, [–180°, 0°] 加上 360°做一个平移, 同时预测蛋白质主链 ψ 和 ϕ 角, 角度的 MAE 分别为 38°和 25°[13]; 2009 年, Real-SPINE2.0 使用两层全连接网络, ψ 和 ϕ 角预测精度进一步改进, MAE 分别为 36°和 22°[14]; 2009 年和 2012 年, SPINE XI 和 SPINE X 使用多步神经网络, ψ 角预测的 MAE 分别为 33°[15] 和 35°[16]; 2015 年 SPIDER2 使用深度学习 3 层全连接神经网络预测角度的正弦和余弦值, ψ 角预测的 MAE 降低到 30°[17]; 2017 年, SPIDER3 使用 4 层双向

* 国家自然科学基金 (批准号: 32071247) 资助的课题.

† 通信作者. E-mail: yxiao@hust.edu.cn

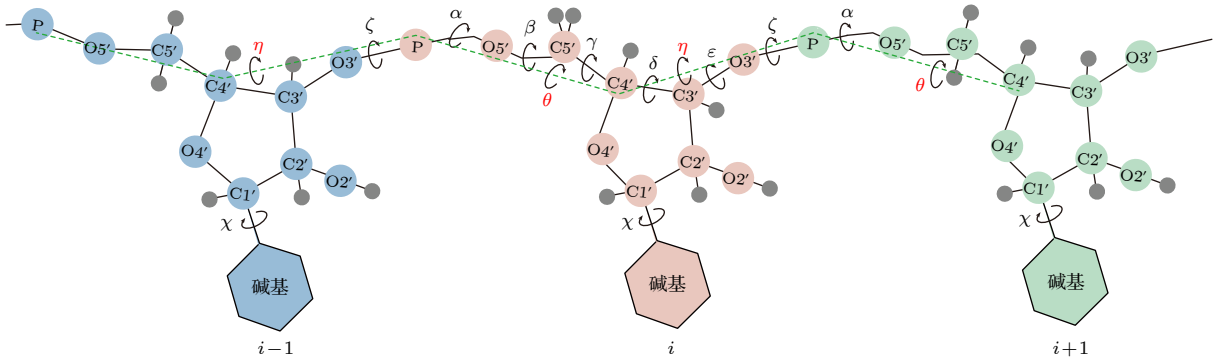

图 1 RNA 分子主链和侧链 7 个扭转角和 2 个伪角的示意图

Fig. 1. Diagram of RNA seven backbone torsion and two pseudo-torsion angles.

LSTM 模型使 ψ 角预测的 MAE 进一步下降为 27°[18]; 2019 年, SPOT-1D 使用 10 层以上的 LSTM (long short-term memory) 残差网络预测角度的正弦和余弦值, ψ 角预测的 MAE 为 23°[19]; 2020 年, 使用 3 层全连接网络, 滑动窗口特征, ψ 角预测的 MAE 仅为 18°[20]. 对于 RNA 分子, 2021 年, SPOT-RNA-1D 首次使用 1 层普通卷积和 2 层膨胀卷积预测 RNA 的 7 个扭转角和 2 个自定义伪角 (η, θ) (图 1) 的正弦和余弦值, α, β, γ, δ, ε, ζ, χ, η, θ 的平均绝对误差分别为 43.94°, 21.94°, 32.98°, 14.61°, 20.69°, 33.27°, 19.59°, 30.25°和 32.91°[21]. 可以看到, 相对于蛋白质分子, RNA 分子扭转角预测的精度还有待提高.

本文提出了一种基于时序网络深度学习模型预测 RNA 分子扭转角的方法 1dRNA, 分别使用深度残差卷积模型 (deep residual CNN, DRCNN) 和深度超长短期记忆模型 (deep HyperLSTM, DHLSTM) 预测 RNA 分子的 7 个扭转角和 2 个伪角, 以此分析抓取相邻核苷酸特征的卷积网络和抓取全局核苷酸特征的循环网络, 哪种网络更合适扭转角预测问题, 并将两个模型的结果和抓取间隔核苷酸特征的 SPOT-RNA-1D 比较. DRCNN模型基于只能看到相邻核苷酸特征的一维卷积, 卷积过程不改变序列长; DHLSTM 模型基于能看到全部核苷酸的特征、并能改变常规长短期记忆 (LSTM) 网络权重共享范式的超 LSTM 网络. 结果表明, 本文采用的两个深度学习模型都可以进一步提高 RNA 分子扭转角的预测精度, 不同模型在不同角度上各有优势, δ, ζ, χ, η 和 θ 角的预测更适合卷积网络, β 和 ε 角的预测更适合循环网络, 而在 α 和 γ 角中, 抓取间隔核苷酸的膨胀网络更好.

2 深度学习模型

2.1 深度学习模型

DRCNN 模型架构如图 2 所示, 由一个一维卷积层 [22] 开始, 输入通道为 4, 输出通道为 512 (卷积输出通道超参数 512 比 256 效果好和 1024 效果类似), 训练批次为 8 (本文模型在一张 11G 显存 GTX 1080 Ti 显卡上能容下的最大样本数), 卷积核为 15 (卷积核超参数 15 比 7 和 30 效果好), 填充方式为“same”, 其他为默认值. 初始卷积层之后, 是 4 个残差块的依次叠加 (残差块的数目 1 到 6 测试显示 4 个残差块效果最好), 每个残差块 [23] 依次包含: 一维批归一化层 BatchNorm1d[24] (特征维度为 512, 添加在卷积网络中, 有助于模型训练的稳定, 效果比 LayerNorm 样本归一化要好), ReLU 激活函数 [25] (对本文模型激活函数 ReLU 比 tanh 和 Leaky ReLu 效果好), 一维卷积层 (输入通道维度为 512, 输出通道维度为 512, 卷积窗口一次能看到的核苷酸数目为 15, 填充方式为“same”, 其他为默认值), 再一维批归一化层, ReLU 激活函数和一维卷积层, 最后将此层卷积的输出和残差块的输入相加, 相加的结果再输入下一个残差块中, 重复 4 次. 数据流出残差块后, 经过一个 ReLU 激活函数 (激活函数放在残差块外训练效果更好), 一维批归一化层 (特征维度为 512), dropout 层 (和全连接层连用, 减少网络的过拟合, 采样概率 0.4, 比 0.2 和 0.5 效果好), 全连接层 (输入维度 512, 输出维度 18), tanh 激活函数 (输出区间在 [–1, 1], 和预测角度的正弦和余弦值区间一致) 得到输出.

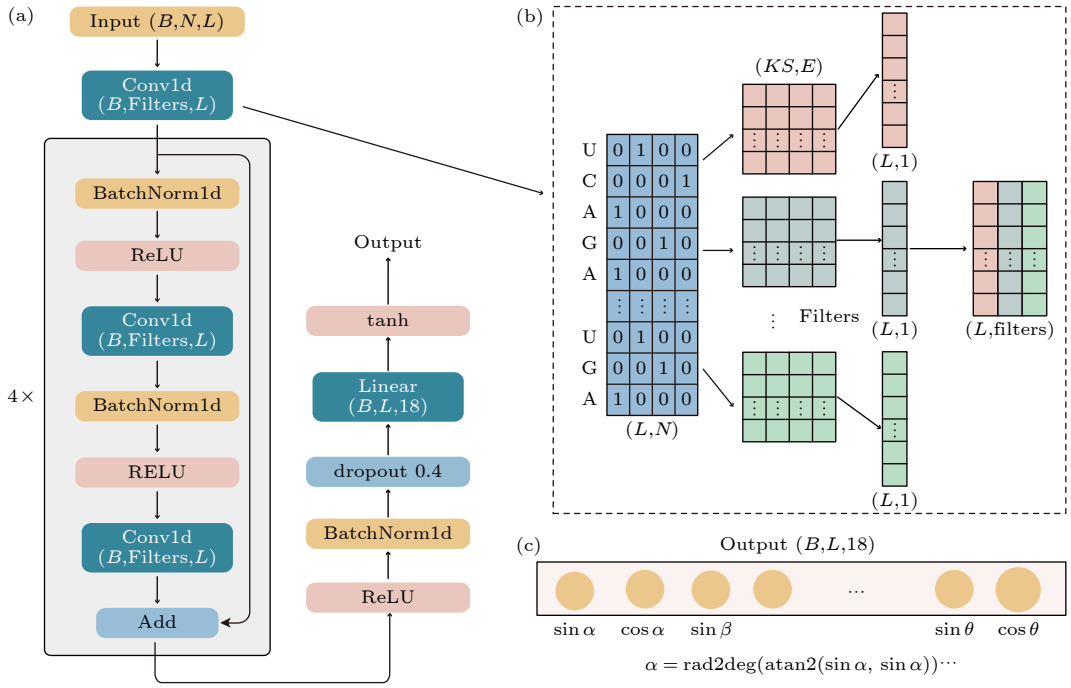

图 2 DRCNN (a) 模型架构; (b) 模型中一维卷积层的原理; (c) 输出层. B, L, N, KS 和 Filters 分别为训练中更新一次模型参数选择的序列数目、序列的长度、输入特征维度、卷积核的小大 (卷积窗口一次能看到的相邻核苷酸数目)、卷积核的数目 (卷积层的输出维度)

Fig. 2. DRCNN: (a) Network architecture; (b) Conv1d layer; (c) output layer. B, L, N, KS and Filters are batch size, sequence length, the size of the input, the size of the filter (the filter can see the number of nucleotides at one time), the number of filters.

DHLSTM 模型结构如图 3 所示, 里面的 HyperLSTM 层原理来自于文献 [26], 输入数据的维度是 (512, 8, 4), 模型更新一次参数选取的样本批次数目为 8, 描述一个核苷酸的初始特征向量维度为 4; 然后经过一个 HyperLSTM 层 (这里的超参数, 外部大 LSTM 层 [27] 的输出维度 Hidden 取 64、内部小 LSTM 层的输出维度和改变 LSTM 层权重的 Hypercell 单元里线性投影的维度 Hyper 都取 16; Hidden 超参数 64 比 16, 32 和 128 效果好, Hyper 超参数 16 比 32 和 64 效果好), 具体来说, 第 t 个核苷酸特征向量和两类隐藏态进入 HyperLSTM cell 单元, 得到第 $t+1$ 个核苷酸新的特征向量和两类隐藏态, 这里每个核苷酸使用不同的 HyperLSTMcell 权重参数, 依次算完所有核苷酸, 得到描述一个批次每个核苷酸新特征数据维度 (512, 8, 64); 接着经过另一个 HyperLSTM 层 (这里三层 HyperLSTMcell 单元的超参数 Hidden 都取 64, Hyper 都取 16), 具体来说, 上一层输出的第 t 个核苷酸特征向量和两类隐藏态 (维度 (8, 64)) 依次进入三个 HyperLSTMcell 单元, 得到第 $t+1$ 个核苷酸新的特征向量 (维度 (8, 64)) 和两类隐藏态输出 (维度分别为 (8, 64), (8, 16)), 依次算完所有核苷酸, 得到描述一个批次每个核苷酸的新特征数据维度 (512, 8, 64); 最后将第二层 HyperLSTM 的输出和第一层的 HyperLSTM 输出相加, 作为一个残差块; 数据流出残差块后, 进入全连接层 (输入维度 512, 输出维度 18), tanh 激活函数得到输出.

DHLSTM 和 DRCNN 训练都使用 MSE 损失函数和 RMSprop 优化器 [28] 训练 (优化器学习率取 0.001、正则化系数取 0.0001, 此优化器比 Adam 和 AdamW 优化器效果好, 学习率 0.01 比 0.1, 0.001, 0.0001 和 0.00001 效果好, 正则化系数经过尝试取学习率的百分之一 0.0001 比较好); 同时预测 9 个角和单独预测一个角, 预测结果基本一致, 故 DHLSTM 和 DRCNN 都同时预测 9 个角; DHLSTM 模型在训练过程中, 训练损失随着 epoch 的增大一直下降, 验证损失在第 85 个 epoch 后开始逐步上升, 如图 4(a) 所示, 故取第 85 个 epoch 的模型为最终模型; DRCNN 模型在训练过程中, 训练损失随着 epoch 的增大一直下降, 验证损失在第 109 个 epoch 后开始逐步上升, 如图 4(b) 所示, 故取第 109 个 epoch 的模型为最终模型. DHLSTM 和 DRCNN 的实现都使用 Facebook 的 PyTorch 深度学习框架 [29].

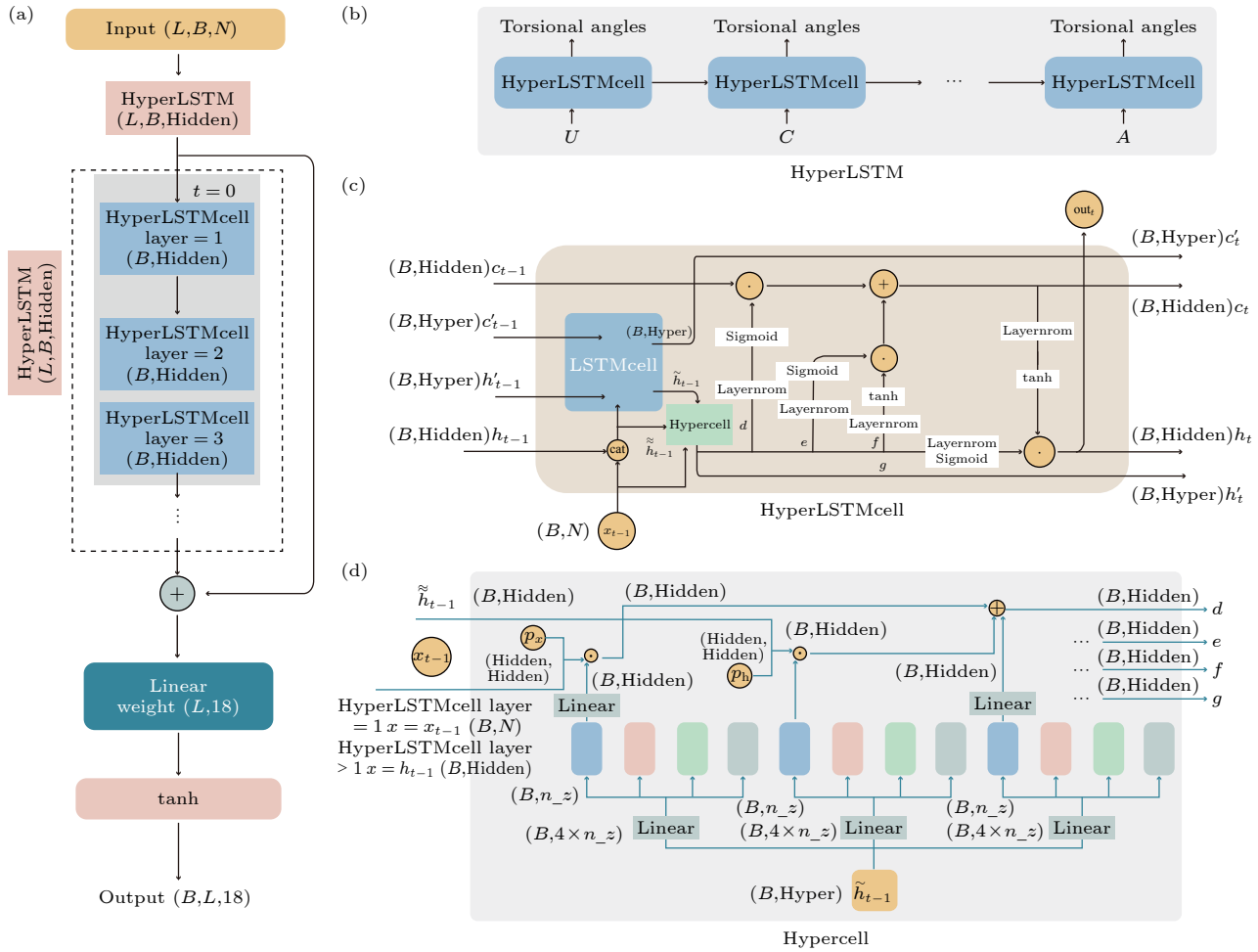

图 3 DHLSTM (a) 模型架构; (b) HyperLSTM 层; (c) 对每个核苷酸的处理单元 HyperLSTMcell, 其中 h_t, c_t 和 h_{t-1}, c_{t-1} 分别是外部更大的 LSTM 在 t 和 $t-1$ 时刻的隐藏态; h'_t, c'_t 和 h'_{t-1}, c'_{t-1} 分别是更小的 LSTM 在 t 和 $t-1$ 时刻的隐藏态; (d) Hypercell 单元. L, B, N, Hidden, Hyper 和 n_z 分别为序列的长度、训练中更新一次模型参数选择的序列数目、输入特征维度、大 LSTM 层的输出维度、内部 LSTM 层的输出维度和改变大 LSTM 层权重的 Hypercell 单元里线性投影的维度, P_x 和 P_h 为动态可训练参数, 绑定在内部超网络里, 作用在输入态 x_{t-1} 和隐藏态, 初始值为全零张量

Fig. 3. DHLSTM: (a) Network architecture; (b) HyperLSTM layer; (c) HyperLSTMcell; h_t, c_t and h_{t-1}, c_{t-1} are the states of the larger outer LSTM at time t and $t-1$, respectively; h'_t, c'_t and h'_{t-1}, c'_{t-1} are the states of the smaller LSTM at time t and $t-1$. (d) Hypercell. L, B, N, Hidden are sequence length, batch size, the size of the input, the size of the LSTM, and Hyper is the size of the smaller LSTM that alters the weights of the larger outer LSTM, n_z is the size of the feature vectors used to alter the larger LSTM weights, P_x and P_h are dynamically trainable parameters, bound in the internal hypernetwork, acting on the input state x_{t-1} and the hidden state, and the initial value is an all-zero tensor.

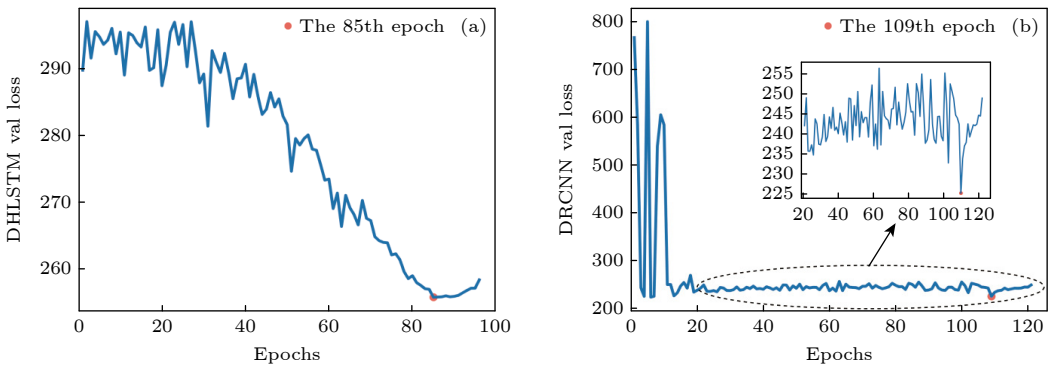

图 4 (a) DHLSTM 模型和 (b) DRCNN 模型验证损失 (MAE) 随 epoch 的变化

Fig. 4. Validation loss curve with the epoch by (a) DHLSTM and (b) DRCNN.

2.2 数据集

为了比较, 采用了 SPOT-RNA-1D 使用的训练集、验证集和测试集 (https://github.com/jaswindersingh2/SPOT-RNA-1D/tree/main/datasets)[21]. 训练集含有 286 个结构, 从 PDB 结构数据库[30] 目前可以下载到 284 个结构 (6N5R_A, 6N5L_A 下架), 本文训练集为这 284 个结构; 验证集含有 30 个结构, 都可从 PDB 下载; 测试集有 3 个分别含有 63, 30 和 54 个结构, 从 PDB 数据库分别下载到 62 (5Y85_B 内含脱氧核苷酸下架)、30 和 54 个结构.

SPOT-RNA-1D 数据集来自于 2020 年 10 月 3 日 PDB 数据库中所有 X 衍射分辨率小于 3.5 Å的 RNA 结构; 用 CD-HIT-EST[31] 软件对所有这些结构的序列设置相似度 0.8 进行聚类, 多簇类中的代表序列构成训练集; 然后将训练集和单簇类利用 BLAST-N[32] 软件设置截断值为 10 处理, 训练集与单簇类有命中的序列被删除, 单簇类中有命中的序列也被删除; 经过这些处理, 训练集剩下的序列作为最终训练集, 单簇类剩下的序列随机分为验证集、测试集 I 和测试 II; 另外, 对 2021 年 4 月 5 日 PDB 数据库中所有 NMR 结构, 使用相同方法, 去除和训练集、验证集、测试集 I 和测试 II 的冗余, 作为测试集 III. 数据集的长度和二级结构分布信息如表 1 所列.

2.3 输入和输出

模型的输入为核苷酸序列特征, 大小为 $L\times4$ 的 one-hot 编码, 四个核苷酸 (A, U, G 和 C) 分别用 (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0) 和 (0, 0, 0, 1) 表示, L 为序列长度, 序列长度最长为 512, 长度不够的补 0. 数据集中最长序列为 414, 常规做法是将所有序列用 0 补齐到最长序列长度. 在预测时, 模型预测的目标序列长度应不大于最长序列长度. 这里取 512 是借鉴很多蛋白质模型中取值 512, 又观察到所有序列长度补齐到 414 和 512 的预测结果类似, 故为了模型能预测更长的序列, 取值 512. 在训练中测试过将所有序列补 0 区域采用 mask 机制, 补 0 区域值虽然被计算但不参与下层值的计算, 模型性能改善不明显. 输出具体如图 2(c) 所示, 有 18 个节点用于预测 9 个角的正弦和余弦值, 然后利用 atan2 函数将角度的正弦和余弦值转化为角度的弧度值, 再利用 rad2deg 函数将角度的弧度值转化为角度值. 这种变换在蛋白质扭转角预测里也常用.

2.4 评 估

使用 MAE 评估整体性能, 具体如 (1) 式, 预测角度值和实验确定的角度值的绝对差, 360°和这个绝对差的差值, 取两者的小值:

$$\mathrm{MAE}=\sum_i \min\Big(\left|\mathrm{torsion}_{\mathrm{pred}}-\mathrm{torsion}_{\mathrm{true}}\right|, \left(360-\left|\mathrm{torsion}_{\mathrm{pred}}-\mathrm{torsion}_{\mathrm{true}}\right|\right)\Big). \tag{1}$$

3 计算结果和讨论

本文两个深度学习模型使用上面的训练集、验证集和 3 个独立的测试集进行训练、验证和测试. 为了了解模型每个角度在每个测试集的总体表现, 表 2 列出了 DRCNN, DHLSTM 和 SPOT-RNA-1D[21] 在验证集和 3 个测试集上整体的性能评估. 在含有 62 个 RNA 的测试集 I 上, DRCNN 预测的 β, δ, ζ, χ, η 和 θ 角的 MAE 比 SPOT-RNA-1D 分别减小了 5%, 28%, 17%, 16%, 24% 和 20%, α, γ 和 ε 角的 MAE 比 SPOT-RNA-1D 分别增大

表 1 训练集、验证集和 3 个测试集的长度和二级结构信息 (百分数是数据集不同配对类型的核苷酸数目占比)

Table 1. Length and secondary-structure information of training, validation and test sets. The number mentioned along with the base pairing type is the percentage of total nucleotides in the region.

数据集	序列长度区间数目						二级结构		
	20—50	50—100	100—200	200—300	300—400	400—512	括号	假结	不配对
训练集	50	179	46	1	7	1	55.10%	5.63%	39.36%
验证集	20	10	0	0	0	0	52.19%	9.8%	38.01%
测试集I	11	41	10	0	0	0	57.58%	2.81%	39.61%
测试集II	8	16	6	0	0	0	58.42%	5.25%	36.33%
测试集III	40	13	1	0	0	0	65.02%	2.67%	32.31%

表 2 DHLSTM, DRCNN 和 SPOT-RNA-1D 在验证集和 3 个测试集上的 MAE

Table 2. Performance comparison in terms of MAE on validation sets and three test sets by three models.

	数据集	7个标准扭转角							伪角	
		α/(°)	β/(°)	γ/(°)	δ/(°)	ε/(°)	ζ/(°)	χ/(°)	η/(°)	θ/(°)
DHLSTM	验证集	47.91	20.22	37.18	16.57	18.23	35.02	19.85	28.09	32.85
	测试集I	48.20	20.66	37.13	13.08	18.82	30.27	17.33	25.74	29.22
	测试集II	47.95	19.89	35.30	15.19	17.87	30.99	17.67	27.20	31.49
	测试集III	45.45	22.30	40.80	13.51	21.43	30.69	16.96	23.87	29.84
DRCNN	验证集	44.67	19.96	35.31	13.86	22.20	31.62	19.49	24.77	30.22
	测试集I	44.84	20.74	36.27	10.51	21.48	27.53	16.39	23.12	26.34
	测试集II	43.41	19.55	35.45	12.19	22.71	28.13	17.16	24.28	28.12
	测试集III	27.14	15.81	25.20	9.73	14.51	17.98	11.58	13.67	17.77
SPOT-RNA-1D [21]	验证集	45.18	20.58	33.88	17.99	20.72	37.50	23.01	33.55	37.02
	测试集I	43.94	21.94	32.98	14.61	20.69	33.27	19.59	30.25	32.91
	测试集II	39.50	18.92	29.47	16.01	17.46	28.91	18.20	28.14	30.25
	测试集III	37.89	21.04	34.68	13.83	22.32	27.87	17.01	25.31	27.22

了 2%, 10% 和 4%; DHLSTM 预测的 β, δ, ε, ζ, χ, η 和 θ 角的 MAE 比 SPOT-RNA-1D 分别减小了 6%, 10%, 9%, 9%, 12%, 15% 和 11%, α 和 γ 角的 MAE 比 SPOT-RNA-1D 分别增大了 10% 和 13%, 这表明在 δ, ζ, χ, η 和 θ 角这些角中, 每层考虑相邻核苷酸特征的 DRCNN 比每层考虑全部核苷酸特征的 DHLSTM 要好, 在 β 和 ε 角中, 每层考虑全部核苷酸特征的 DHLSTM 比每层考虑相邻核苷酸特征的 DRCNN 要好, 在 α 和 γ 角中, 每层考虑间隔核苷酸的 SPOT-RNA-1D 比 DRCNN和 DHLSTM 都要好. MAE 值越大预测难度越大, 在 DRCNN 中角度预测难度 δ, χ, ε, β, η, θ, ζ, γ 和 α 依次递增, 在 DHLSTM 中角度预测难度 δ, χ, β, ε, η, θ, ζ, γ 和 α 依次递增, 在 SPOT-RNA-1D 中角度预测难度 δ, χ, ε, β, η, θ, γ, ζ 和 α 依次递增, 可以看到 δ, χ, η, θ 和 α 角在 3 个模型里预测难度的排序一致, 考虑相邻核苷酸的 DRCNN 和考虑间隔核苷酸的 SPOT-RNA-1D 都表明 ε 比 β 容易预测, 而对于 DHLSTM, ε 比 β 难预测, DRCNN 和 DHLSTM 都表明 ζ 比 γ 容易预测, 而对于 SPOT-RNA-1D, ζ 比 γ 难预测. 这 3 种方法都认为 α 是最难预测的, 表明 3 个模型在角度预测难度方面有一定相似性, 也各有特点. 在测试集 II 和测试集 III 观察到类似的性能趋势, 表明模型对不同类型的测试集具有鲁棒性.

为了了解模型在单个序列上的表现, 图 5 给出了 DRCNN, DHLSTM 和 SPOT-RNA-1D 在 3 个测试集上单个 RNA 分子扭转角预测的 MAE 分布图, 其中 SPOT-RNA-1D 绘制每个盒子需要五类值 (最大值、最小值、中位数、上下四分位数和异常值), 由论文图形数据获取工具 WebPlotDigitizer[33] 得到. 每个模型在 3 个数据集 9 个角度的 27 个 MAE 最小值上, DRCNN 占 18 次, DHLSTM 占 3 次, SPOT-RNA-1D 占 6 次, 而在 27 个 MAE 最大值上, DRCNN 占 4 次, DHLSTM 占 8 次, SPOT-RNA-1D 占 15 次, 表明考虑相邻核苷酸特征的卷积模型 DRCNN 最有可能预测到最小的 MAE 值, DHLSTM 次之, SPOT-RNA-1D 很难预测相比比较小的 MAE 值. 箱子越窄意味着每次预测 MAE 变化更小, 模型预测更稳定, 每个模型在 3 个测试集 9 个角度的 27 个箱子中, DRCNN 出现 9 次, DHLSTM 出现 15 次, SPOT-RNA-1D 出现 3 次, 表明预测最稳定的模型是考虑全部核苷酸特征的 DHLSTM, 且性能中规中矩, 其次是 DRCNN, 对样本反应比较敏感的是 SPOT-RNA-1D. 在 27 个盒子相对较小的中位数上, DRCNN 占 18 次, DHLSTM 占 2 次, SPOT-RNA-1D 占 7 次, 表明 DRCNN 预测的一半数目链的总 MAE 比其他两个模型值要低. 在异常值方面, 3 个测试集 9 个角度上, DRCNN, DHLSTM 和 SPOT-RNA-1D 出现的异常值的数目分别为 24, 21 和 38, 且 DRCNN 和 DHLSTM 出现的异常值本身是比较小, 同样表明 DHLSTM 预测比较稳定. 以上说明, 考虑相邻核苷酸特征的 DRCNN 模型性能整体更强大, 考虑全部核苷酸特征的 DHLSTM 模型预测更稳定.

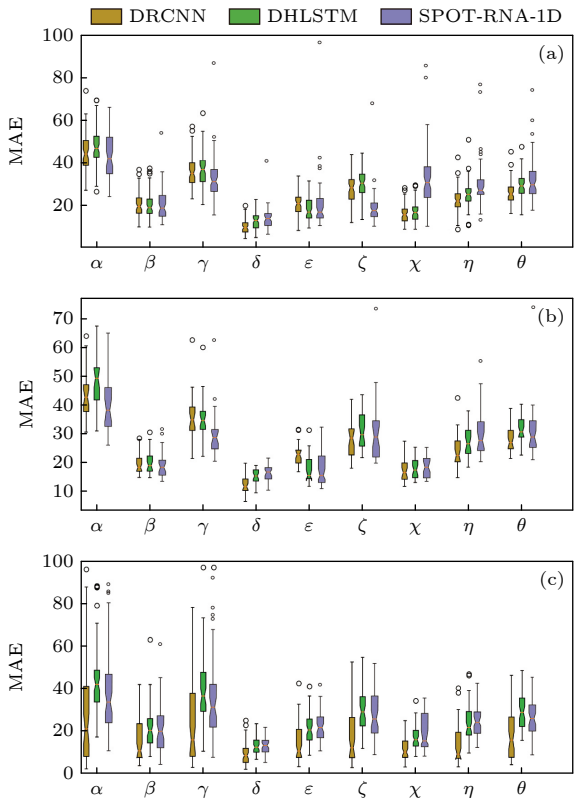

图 5 DRCNN(黄色)、DHLSTM(绿色) 和 SPOT-RNA-1D (紫色) 在测试集 I (a)、测试集 II (b) 和测试集 III (c) 上单个 RNA 链的 MAE 分布图. 每个盒子显示出一组数据的最大值、最小值、中位数、上下四分位数和异常值

Fig. 5. Distribution of MAE for individual RNA chains on test set I (a), test set II (b) and test set III (c) by DRCNN predictor (yellow), by DHLSTM (in green) and SPOT-RNA-1D (in purple). Each box shows the minimum, the maximum, the sample median, the first and third quartiles and outlier.

另外绘制了角度的实验值分布, 如图 6 橙色虚线所示, 可以看出每个角度的实验值的分布是比较陡峭的, 大部分角度都集中在跨度在 40°左右的角度空间, 有少部分角度值分布在跨度在 360°的角度空间中, 最容易预测的 δ 角跨度也是最窄的, 最难预测的 α 角分布有 3 个峰, 跨度是最广的. 为了了解本文模型在预测分布上的能力, 绘制了 DRCNN 和 DHLSTM 在测试集 I 的预测分布如图 6 黄色和绿色虚线所示, DRCNN 预测所有的角度分布都比 DHLSTM 好; 在测试集 II 和测试集 III 上, DRCNN 在 β 和 γ 角上预测的分布比 DHLSTM 要好, 两个模型在预测其他 7 个角的分布类似.

二级结构对 RNA 建模起着重要角色, 根据 DSSR 软件 [34] 输出的 RNA 二级结构, 可将 RNA 二级结构分为三种类型, 括号 (['(', ')']), 假结 (['[', ']', '{', '}', '<', '>', 'A', 'a']), 环区 ['.']. 比较了 DRCNN 和 DHLSTM 在测试集 III 中对 3 种二级结构类型的整体预测性能 (表 3), 可以看出, 对 DRCNN 和 DHLSTM 来说括号类型的核苷酸的扭转角最容易预测的, 处于环区的核苷酸的扭转角是最难预测; 还可以观察到, DRCNN 预测 3 种类型的 MAE 误差都比相应的 DHLSTM 预测的要低; 在其他两个测试集观察到同样结果, 因此, 扭转角预测的误差主要来自于环区和假结区域, 在预测括号、假结和环区区域的扭转角上 DRCNN 都比 DHLSTM 好.

表 1 统计了训练集、验证集和 3 个测试集的序列长度分布. 由表 1 可以看出, 在训练集和验证集中各个长度分布并不均匀, 长度在 50 到 100 区间的有 179 个结构, 在 100 到 200 区间的只有 46 个. 为了了解这种差异是否会导致 DRCNN 和 DHLSTM 对长 RNA 扭转角预测性能较差, 图 7 绘制了两个模型在 9 个角度上的表现与序列长度的关系. 观察 DHLSTM 和 DRCNN 的预测结果, 9 个角的 MAE 值在数目少的长度区间 [78, 94], [155, 171] 和 [171, 186] 并不大; 还观察到 DRCNN 在短长度区间 [1, 47] 结果比 DHLSTM 结果好; 因此, 虽然训练集和验证集对不同长度的 RNA 数目分布不均匀, 但并没有造成 DRCNN 和 DHLSTM 在预测上的长度偏好.

表 3 DHLSTM 和 DRCNN 在测试集 III 不同配对类型中扭转角预测的 MAE

Table 3. Performance according to mean absolute error by DHLSTM and DRCNN for nucleotides in different pairing type on test set III.

	配对类型	七个标准扭转角							伪角	
		α/(°)	β/(°)	γ/(°)	δ/(°)	ε/(°)	ζ/(°)	χ/(°)	η/(°)	θ/(°)
DHLSTM	括号	34.08	16.48	30.21	9.76	17.98	21.38	11.23	18.03	21.91
	假结	34.20	14.98	27.06	6.80	14.25	20.29	10.98	27.41	18.02
	环区	66.77	32.60	60.72	21.05	27.54	47.85	28.52	35.41	46.16
DRCNN	括号	19.43	11.40	18.54	6.65	11.84	12.0	8.30	10.90	12.94
	假结	20.42	14.25	16.75	6.73	12.86	13.54	10.25	16.14	13.52
	环区	40.84	23.26	37.44	15.59	19.07	29.07	18.44	19.25	27.08

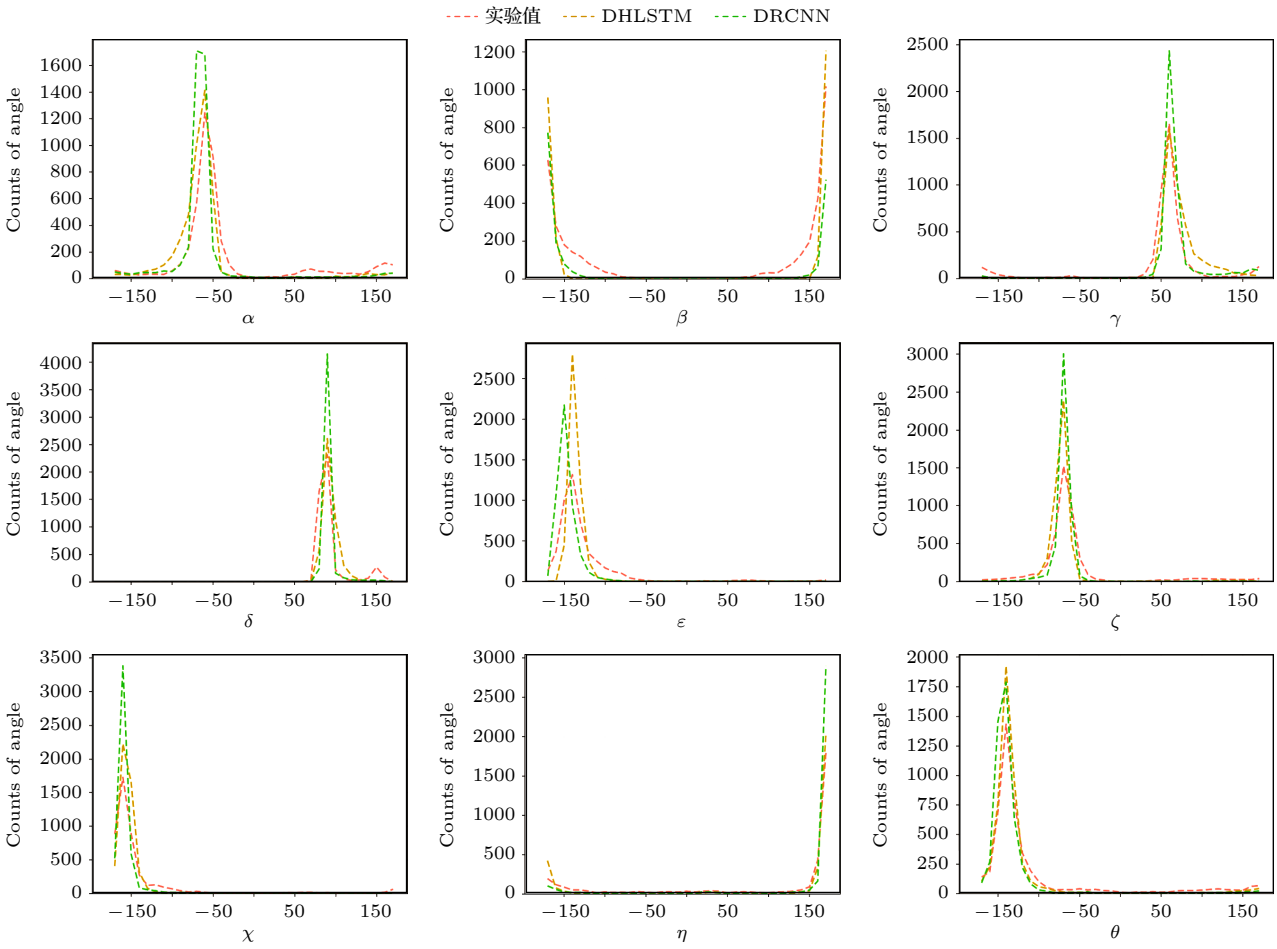

图 6 测试集 I 扭转角的实验值 (橙色)、DHLSTM 预测值 (黄色) 和 DRCNN 预测值 (绿色) 分布图

Fig. 6. Distribution plots of native (in orange), DHLSTM predicted (in yellow), and DRCNN predicted (in green) nine torsion angles on test set I.

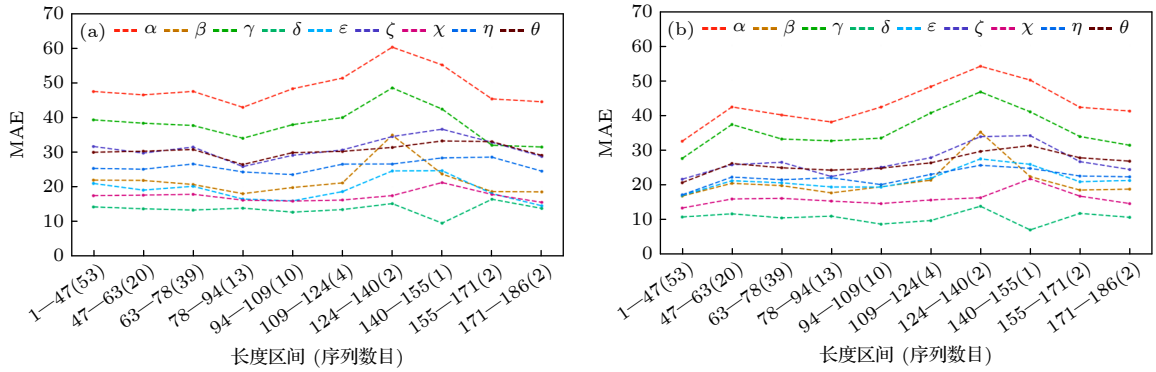

图 7 (a) DHLSTM 和 (b) DRCNN 分别在 3 个测试集 (147个 RNA) 的 9 个扭转角的 MAE 与 RNA 序列长度的函数

Fig. 7. On 147 RNAs in the three test sets, the MAE is measured as a function of the length for the nine torsion angles by (a) DHLSTM and (b) DRCNN.

和 SPOT-RNA-1D 方法一样, 为了了解扭转角之间的相关性, 在测试集 I 上绘制了如图 8 所示的扭转角相关矩阵. 一般情况下, 相邻扭转角之间高度相关, 而较远扭转角相关性较小, 但是矩阵显示, 对于 DRCNN 和 DHLSTM, α 和 γ 角有很强的相关性, 两者也是模型预测难度最大的两个角, ζ 和 θ 有最强的相关性, 两者预测难度排名也是相邻的. 在其他两个测试集的结果相同.

观察一条链中预测的每个角度, 预测的大部分扭转角比一些近天然态或者类天然态结构的扭转角更接近天然态结构扭转角的值. 和 SPOT-RNA-1D 方法一样, 也测试了 DRCNN 和 DHLSTM 这

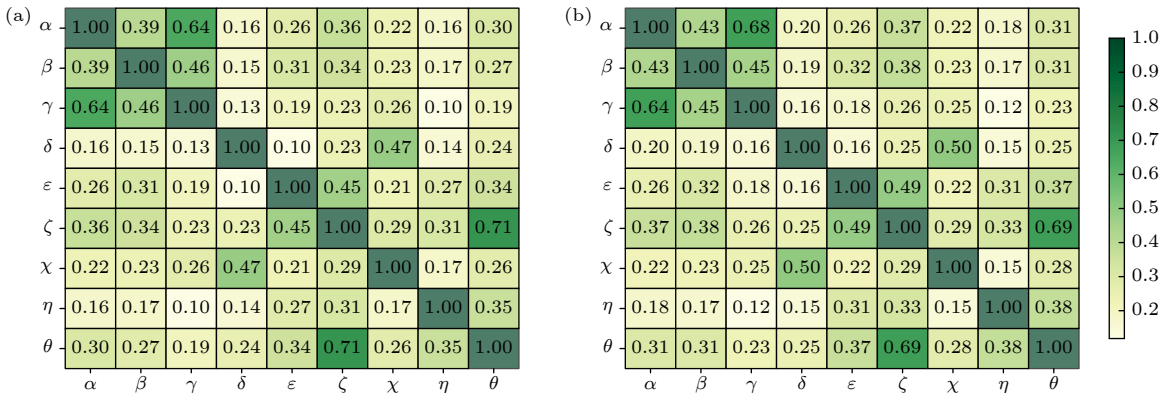

图 8 (a) DHLSTM 和 (b) DRCNN 分别在测试集 I 上扭转角的 MAE 的相关系数 (CCs), 值越大表示两个角度越相关

Fig. 8. Correlation coefficient (CCs) for MAE of between the nine torsion angles of test set I by (a) DHLSTM and (b) DRCNN, the larger the CC value, the more correlated between the two torsions.

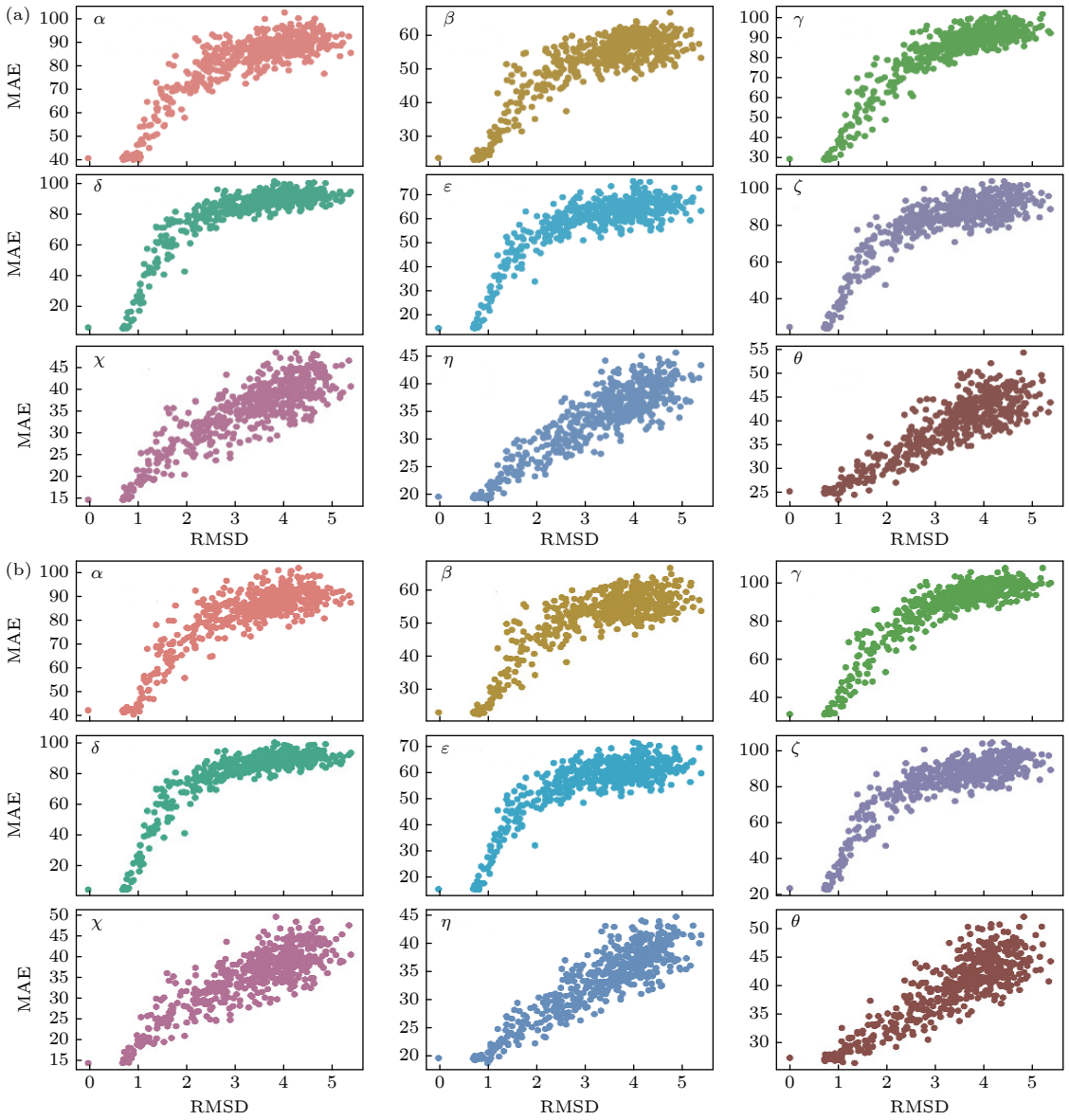

图 9 (a) DRCNN 和 (b) DHLSTM 分别在 RNA 1Y69(链 9) 上预测角度与 decoys 结构角度之间的 MAE 与 RMSD 的关系

Fig. 9. On RNA 1Y69 (chain 9), the MAE is measured as a function of RMSD for the nine torsion angles by (a) DRCNN and (b) DHLSTM.

两种深度学习模型预测的角度和不同 RMSD 结构的角度之间的差异是否可以用于结构的质量评估. 为此, 使用 3dRNA[3] 测试集 85 个 RNA 和它们的 decoys 进行了测试. 图 9 绘制了 DRCNN 和 DHLSTM 在其中一个 RNA(PDB ID 号 1Y69, 链 9) 在预测角度与诱饵模型结构角度之间的 MAE 和结构精度的函数关系, MEA 随 RMSD 持续增加. 在 85 个数据集中的其余 84 个 RNA 中也观察到类似的趋势, 这表明与模型预测角度的偏差或结合其他参量可用于模型质量评估.

4 结 论

本文提出了一种预测 RNA 分子扭转角的深度学习方法 1dRNA, 采用了 DRCNN 和 DHLSTM 两个基于时序网络的模型去预测 RNA 的 7 个扭转角 (α, β, γ, δ, ε, ζ 和 χ) 和 2 个伪角 (η 和 θ), 并和现有方法 SPOT-RNA-1D 进行了比较. 结果表明不同网络在不同角度上各有优势, 当序列长度不超过 50 时, 在预测 9 个角时, 考虑相邻核苷酸特征的 DRCNN 比考虑全部核苷酸特征的 DHLSTM 和考虑间隔核苷酸特征的 SPOT-RNA-1D 都好; 当序列长度超过 50, 在 δ, ζ, χ, η 和 θ 角这些角中, DRCNN 预测的结果整体上比 DHLSTM 和 SPOT-RNA-1D 要好, 在 β 和 ε 角中, DHLSTM 预测的结果整体上比 DRCNN 和 SPOT-RNA-1D 要好, 在 α 和 γ 角中, SPOT-RNA-1D 预测的结果整体上比 DHLSTM 和 DRCNN 要好; 3 个模型在 9 个角度的预测难度上类似, 角度的实验值和预测值分布可以看出角度预测的难度主要在于角度分布的复杂程度, 分布越复杂越难预测, DRCNN 和 SPOT-RNA-1D 预测出来的角度分布比 DHLSTM 丰富; 序列环区的角度分布比配对区域复杂, 角度预测难度也比配对区域大很多; 每个模型在链长度集中在非长链区的训练集和验证集上训练, 但在预测时对长链预测效果也不错; 在模型预测稳定性上, 考虑全链核苷酸的 DHLSTM 比考虑相邻核苷酸的 DRCNN 和考虑间隔核苷酸的 SPOT-RNA-1D 要稳定很多, 异常值少; 模型的各个结果在 3 个测试集上表现类似, 表明模型性能对不同数据集稳定. 从结果来看, 面对比较短序列, 9 个角度都用考虑相邻核苷酸特征的卷积网络更好, 当序列长时, 在预测 δ, ζ, χ, η 和 θ 角用考虑相邻核苷酸特征的卷积网络更好, 预测 β 和 ε 用考虑全链核苷酸特征的超循环网络更好, 预测 α 和 γ 角用考虑间隔核苷酸特征的膨胀卷积网络更好. 在数据集方面, 尝试过加入新发表的 RNA 结构增大数据集训练, 精度能提高但不明显; 可以设计其他类型的网络, 尝试使用单纯的全连接网络和 Transformer[35] 网络训练, 角度预测整体 MAE 比 DRCNN 和 DHLSTM 更好, 但预测的角度分布很差, 很难预测出角度分布峰值之外的区域; 尝试过在 DRCNN 和 DHLSTM 这个两个模型上改进, 精度能提高但不明显; 在加入新特征方面, 加入二级结构特征, 能提高精度但也不明显. 在改进角度预测方面, 从结果可以看出角度分布决定了预测难度, 在预测前如何预先处理这种分布, 和如何把这种分布加入损失函数, 应该可以很大提高预测精度; 另外直接预测角度实值难度大, 可以考虑将跨度 360°的角度分布分成 36 个 bin 去预测.

参考文献

[1] Jiao K, Hao Y Y, Wang F, et al. 2021 *Biophys. Rep.* **7** 21
[2] Sun S, Chen X Z, Chen J, et al. 2021 *Biophys. Rep.* **7** 8
[3] You Y L, Tang Z M, Lin H, Shi J L 2021 *Biophys. Rep.* **7** 159
[4] Zhang Y, Wang J, Xiao Y 2022 *J. Mol. Biol.* **434** 167452
[5] Zhang Y, Wang J, Xiao Y 2020 *Comput. Struct. Biotechnol. J.* **18** 2416
[6] Wang J, Wang J, Huang Y Z, Xiao Y 2019 *Int. J. Mol. Sci.* **20** 4116
[7] Wang J, Xiao Y 2017 *Curr. Protoc. Bioinf.* **57** 5.9.1
[8] Wang J, Zhao Y J, Zhu C Y, Xiao Y 2015 *Nucleic Acids Res.* **43** e63
[9] Zhao Y J, Huang Y Y, Gong Z, et al. 2012 *Sci. Rep.* **2** 734
[10] Wang J, Mao K K, Zhao Y J, Zeng C, Xiang J J, Zhang Y, Xiao Y 2017 *Nucleic Acids Res.* **45** 6299
[11] Olson W K 1982 *Topics in Nucleic Acid Structures* (Part 2) (London: Macmillan Press) pp1–79
[12] Dor O, Zhou Y Q 2007 *Proteins* **68** 76
[13] Xue B, Dor O, Faraggi E, Zhou Y Q 2008 *Proteins* **72** 427
[14] Faraggi E, Xue B, Zhou Y Q 2009 *Proteins* **74** 847
[15] Faraggi E, Yang Y D, Zhang S H, Zhou Y Q 2009 *Structure* **17** 1515
[16] Faraggi E, Zhang T, Yang Y D, Kurgan L, Zhou Y Q 2012 *J. Comput. Chem.* **33** 259
[17] Heffernan R, Paliwal K, Lyons J, et al. 2015 *Sci. Rep.* **5** 11476
[18] Heffernan R, Yang Y D, Paliwal K, Zhou Y Q 2017 *Bioinformatics* **33** 2842
[19] Hanson J, Paliwal K, Litfin T, Yang Y D, Zhou Y Q, Valencia A 2019 *Bioinformatics* **35** 2403
[20] Mataeimoghadam F, Newton M A H, Dehzangi A, Karim A, Jayaram B, Ranganathan S, Sattar A 2020 *Sci. Rep.* **10** 19430
[21] Singh J, Paliwal K, Singh J, Zhou Y Q 2021 *J. Chem. Inf. Model.* **61** 2610
[22] Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M, Inman D J 2021 *Mech. Sys. Signal Proc.* **151** 107398

[23] He K M, Zhang X Y, Ren S Q, Sun J 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Las Vegas, NV, USA, June 27–30, 2016 p770
[24] Nam H, Kim H E 2018 arXiv: 1805.07925v3 [cs.CV]
[25] Clevert D A, Unterthiner T, Hochreiter S 2015 arXiv: 1511.07289v5 [cs.LG]
[26] Jayasiri V, Wijerathne N 2020 https://nn.labml.ai/ [2023-04-02]
[27] Hochreiter S, Schmidhuber J 1997 *Neural Comput.* **9** 1735
[28] Tieleman T, Hinton G 2012 *Lecture 6.5-RMSProp: Divide the Gradient by a Running Average of its Recent Magnitude* (COURSERA: Neural Networks for Machine Learning)
[29] Paszke A, Gross S, Massa F, et al. 2019 *33rd Conference on Neural Information Processing Systems* Vancouver, Canada, December 8, 2019 pp8026-8037
[30] Burley S K, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow G V, et al 2021 *Nucleic Acids Res.* **49** D437
[31] Fu L M, Niu B F, Zhu Z W, Wu S T, Li W Z 2012 *Bioinformatics* **28** 3150
[32] Altschul S F, Gish W, Miller W, Myers E W, Lipman D J 1990 *J. Mol. Biol.* **215** 403
[33] Rohatgi A 2022 Software available at https://automeris.io/WebPlotDigitizer Version 4.6[software]
[34] Lu X J, Bussemaker H J, Olson W K 2015 *Nucleic Acids Res.* **43** e142
[35] Vaswani A , Shazeer N, Parmar N, et al. 2017 arXiv: 1706.03762v7 [cs.CL]

SPECIAL TOPIC—Machine learning in biomolecular simulations

Deep learning methods of predicting RNA torsion angle*

Ou Xiu-Juan Xiao Yi†

(*School of Physics, Huazhong University of Science and Technology, Wuhan* 430074, *China*)

(Received 29 June 2023; revised manuscript received 2 August 2023)

Abstract

Modeling of RNA tertiary structure is one of the basic problems in molecular biophysics, and it is very important in understanding the biological function of RNA and designing new structures. RNA tertiary structure is mainly determined by seven torsions of main-chain and side-chain backbone, the accurate prediction of these torsion angles is the basis of modeling RNA tertiary structure. At present, there are only a few methods of using deep learning to predict RNA torsion angles, and the prediction accuracy needs further improving if it is used to model RNA tertiary structure. In this study, we also develop a deep learning method, 1dRNA, to predict RNA backbone torsions and pseudotorsion angles, including two different deep learning models, the convolution model (DRCNN) that considers the features of adjacent nucleotides and the Hyper-long-short-term memory model (DHLSTM) that considers the features of all the nucleotides. We then empirically show that DRCNN and DHLSTM outperform existing state-of-the-art methods under the same datasets, the prediction accuracy of DRCNN model is improved by 5% to 28% for β, δ, ζ, χ, η, and θ angle, and the prediction accuracy of DHLSTM model is improved by 6% to 15% for β, δ, ζ, χ, η, θ angle. The DRCNN model predicts better results than the DHLSTM model and the existing models in the δ, ζ, χ, η, θ angle, and the DHLSTM model predicts better results than the DRCNN model and the existing model in the β and ε angles, and the existing models predicted better results than the DRCNN model and DHLSTM model in the α and γ angles. The DRCNN model and the existing models predict a richer distribution of angles than the DHLSTM model. In terms of model stability, the DHLSTM model is much more stable than the DRCNN model and the existing models, with fewer outliers. The results also show that the α angle and γ angle are the most difficult to predict, the angles of the ring region is more difficult to predict than the angles of the helix region, the model is also not sensitive to the change of the target sequence length, and the deviation of the model prediction angle from the decoys can also be used to evaluate the RNA tertiary structures quality.

Keywords: RNA structure, torsional angle prediction, deep learning

PACS: 87.14.gn, 87.15.A–, 87.15.bg **DOI:** 10.7498/aps.72.20231069

* Project supported by the National Natural Science Foundation of China (Grant No. 32071247).

† Corresponding author. E-mail: yxiao@hust.edu.cn

物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

RNA扭转角预测的深度学习方法

欧秀娟　肖奕

Deep learning methods of predicting RNA torsion angle

Ou Xiu-Juan　Xiao Yi

引用信息 Citation: Acta Physica Sinica, 72, 248703 (2023)　DOI: 10.7498/aps.72.20231069
在线阅读 View online: https://doi.org/10.7498/aps.72.20231069
当期内容 View table of contents: http://wulixb.iphy.ac.cn
