Hans Journal of Computational Biology 计算生物学, 2020, 10(4), 57-68
Published Online December 2020 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2020.104007

Hans汉斯

基于长短时记忆循环网络和基团特征的蛋白质二级结构预测

韩心怡，刘毅慧

齐鲁工业大学(山东省科学院)计算机科学与技术学院，山东 济南
Email: hanxinyi1234@163.com, yxl@qlu.edu.cn

收稿日期：2020年11月9日；录用日期：2020年11月25日；发布日期：2020年12月2日

摘 要

蛋白质二级结构预测是蛋白质结构研究领域的重要课题，随着机器学习和深度学习的发展，多种多样的预测模型被提出，实验采用双向长短时记忆循环网络模型，取消滑动窗口限制，充分考虑氨基酸长距离相互作用和氨基酸序列前后文之间的相互影响。重新设计了网络的输入特征，在PSSM基础上增加了42基团特征，使用大数据集进行训练，在公共测试集CASP9，CASP10，CASP11和CASP12上Q3准确率分别达到了85.74%，86.83%，84.73%和83.79%。实验结果表明，蛋白质二级结构预测可在新的特征设计，考虑氨基酸长距离相互作用和大数据的使用方向上进一步的研究。

关键词

蛋白质，蛋白质二级结构预测，循环网络，基团，结构预测

Protein Secondary Structure Prediction Based on Long-Short-Term Memory Recurrent Network and Radical Group Features

Xinyi Han, Yihui Liu

College of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan Shandong
Email: hanxinyi1234@163.com, yxl@qlu.edu.cn

Received: Nov. 9th, 2020; accepted: Nov. 25th, 2020; published: Dec. 2nd, 2020

文章引用: 韩心怡, 刘毅慧. 基于长短时记忆循环网络和基团特征的蛋白质二级结构预测[J]. 计算生物学, 2020, 10(4): 57-68. DOI: 10.12677/hjcb.2020.104007

Abstract

Protein secondary structure prediction is an important topic in the field of protein structure research. With the development of machine learning and deep learning, a variety of prediction models have been proposed. The experiment used a bidirectional long-short-term memory recurrent network model, removed the sliding window, and fully considered the long-distance amino acid interaction and the interaction between the context of the amino acid sequence. Redesigned the input features of the network, added 42 radical group features on the basis of PSSM, used large data sets for training, and the accuracy of Q3 on the public test sets CASP9, CASP10, CASP11 and CASP12 reached 85.74%, 86.83%, 84.73% and 83.79% respectively. The experimental results show that protein secondary structure prediction can be further studied in the design of new features, considering the long-range interaction of amino acids and the use of big data.

Keywords

Protein, Protein Secondary Structure Prediction, Recurrent Network, Radical Group, Structure Predict

1. 引言

随着基因组测序工作的完成，蛋白质组的相关工作也随之展开，蛋白质的研究有利于人类对疾病的深入了解和相关药物研发，其中的重点工作便是蛋白质的结构研究。蛋白质结构可划分为一级结构、二级结构、三级结构和四级结构，其中具有生物活性的是三级结构。获取三级结构最直接的方法是使用 X 射线和核磁共振进行观察[1]，但是这种方式效率低，成本高，因此从 1951 年 Pauling 和 Corey 预测了蛋白质多肽骨架的螺旋和片状构象开始，研究者便展开了对蛋白质结构的预测工作。尽管蛋白质结构的信息隐藏在氨基酸序列中，但直接利用氨基酸序列进行三级结构预测难度非常大，因此研究者通常先进行蛋白质二级结构预测，基于该结果做进一步的研究。

早期的蛋白质二级结构预测使用统计学方法和启发式规则[2]，支持向量机[3] [4]，贝叶斯分类算法，马尔可夫模型[5]，前馈神经网络[6] [7]等均被应用在了蛋白质二级结构预测中。近年来，随着深度学习方法在自然语言处理、机器视觉和语音识别等方向取得了巨大的进展，生物信息学领域也开始了对深度学习方法的广泛使用[8]，深度学习通过多层特征变换，可以更好的刻画出数据的隐藏信息，捕捉氨基酸的局部和长距离相互作用。例如 Fang 等提出了 MOUFOLD 方法，使用了由两个嵌套的可进行卷积操作的初始模块、卷积层以及完全联通的致密层组成的 Deep3I 网络(Deep Inception-Inside-Inception Networks, Deep3I)进行蛋白质二级结构预测[9]。WANG 等结合了深度卷积神经网络(Deep Convolutional Neural Networks, DCNN)和条件神经场(Conditional Neural Fields, CNF)，提出名为 Raptor X 的预测方法[10]。SPOT-1D 在 Spider [11]基础结合了双向长短时记忆循环网络(Bidirectional long-short-term memory recurrent network，Bi-LSTM)和残余卷积网络(Residual Convolutional Networks, ResNets)用来识别和传播整个序列中的短期和长期依赖关系[12]。值得注意的是，上述方法中 MOFOLD，Raptor X，Spider3 和 SPOT-1D 对特征输入均有细致的设计，MOFOLD 和 Spider3 网络输入为氨基酸理化性质、PSSM 和隐

马尔可夫特征，Raptor X 网络输入为 PSSM+21 个元素的二进制向量，SPOT-1D 在 Spider3 的特征基础上增加了 SPOT-Contact 预测接触图。另外，大数据集的使用也为模型更好的提取隐藏特征起到了重要的作用。

在本实验中，我们采用双向 LSTM 作为训练模型，为了充分发挥 LSTM 网络对序列数据的处理能力，对训练集取消了滑动窗口限制，以捕捉氨基酸长距离相互作用。同时对训练集特征重新进行了设计，在常用的 PSSM 基础上，增加了蛋白质 42 基团的新编码特征，经过大数据集的训练，网络模型可以充分的提取序列隐藏特征。这种基于大数据和新编码方式的模型预测能力在公共测试集 CASP9，CASP10，CASP11 和 CASP12 中进行了评估实验，Q3 准确率分别达到了 85.74%，86.83%，84.73%和 83.79%。

2. 实验模型

2.1. 模型结构

LSTM 在自然语言处理领域中应用广泛，对序列数据的分类有良好的实验效果，蛋白质二级结构预测是在给定的蛋白质序列下对每一个氨基酸的结构归属做出分类，同样可以看作对序列数据的分类问题，因此本次实验选择了双向 LSTM 作为训练模型。

模型训练示意图如图 1 所示，为了充分发挥 LSTM 对长序列数据的学习优势，实验未设置滑动窗口，而是将训练集拆分成小批量并补零填充序列，使它们具有同一小批量中最长序列的长度。为了防止训练过程中添加过多填充，在训练集输入网络前，根据蛋白质长度对数据进行了排序。在结构预测中，每一个氨基酸的输出状态不仅与它之前的序列信息有关，位置靠后的部分同样会产生影响，而双向 LSTM 可以看作两层标准的 LSTM 网络，分别以开头和结尾作为输入，可以为输出层提供输入序列中每一个氨基酸完整的上下文信息，真正考虑到前后位置的相互作用。

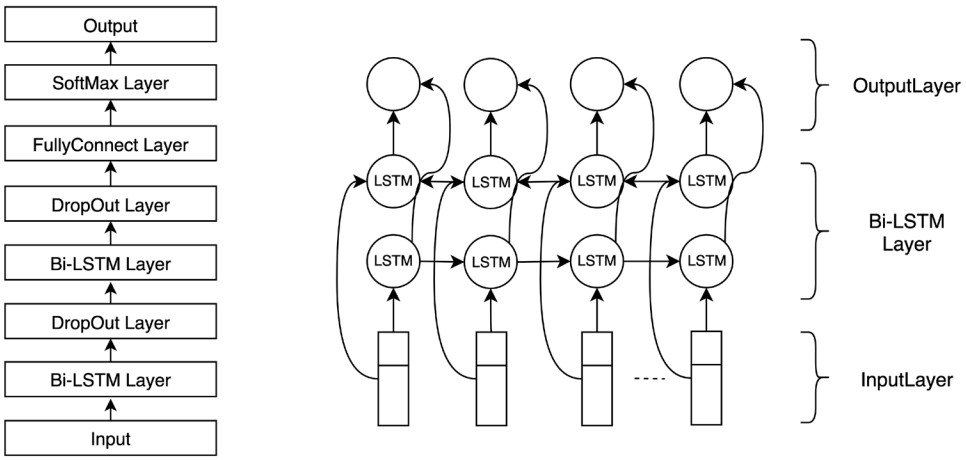

Figure 1. Schematic diagram of network structure and bidirectional LSTM training process
图 1. 网络结构和双向 LSTM 训练示意图

2.2. 模型原理

双向 LSTM 可看作由循环神经网络(Recurrent Neural Network, RNN)逐步演化而来的网络结构，RNN 可较好的针对序列变化的数据进行训练[13]，循环神经网络具有重复的模块单元，展开后的 RNN 单元如图 2 所示。

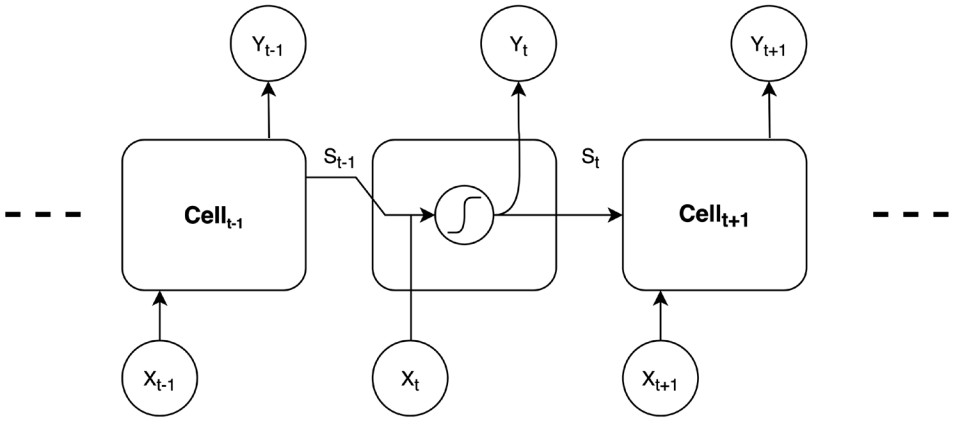

Figure 2. RNN structure diagram
图 2. RNN 网络结构示意图

其中 $t-1$，t，$t+1$ 表示序列的时间位置，X 表示输入的样本，X 经过简单的激活函数得到输出 Y 和记忆数据 S，数学表达式为：

$$y_t = g\left(O * s_t\right) \tag{1}$$

$$s_t = f\left(W * s_{t-1} + U * x_t\right) \tag{2}$$

W 为输入的权重，U 表示 t 时刻下的输入样本权重，O 表示输出样本的权重。

传统的 RNN 面对长序列数据会产生梯度消失和梯度爆炸问题，LSTM 的提出有效的降低了这些情况的风险。在 RNN 的基础上对循环单元做了细致的设计，增设了三种门结构[14]，这种设计方式使 LSTM 单元具备将重要的信息进一步记忆，将不重要信息适当丢弃的能力。标准 LSTM 单元示意图如图 3 所示。

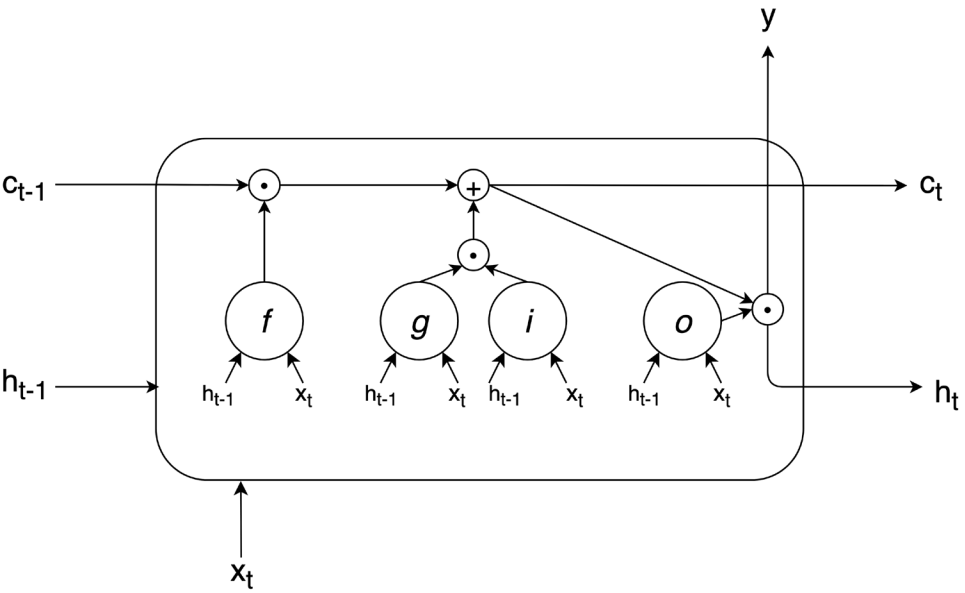

Figure 3. Standard LSTM unit schematic
图 3. 标准 LSTM 单元示意图

单元输入是 t 时刻下的样本 x_t 和 $t-1$ 时刻单元的输出 h_{t-1} (也称为隐藏状态)，其输出可作为下一个单元的输入 h_t 以及本单元的输出 y。C 叫做细胞状态，相比于隐藏状态它的变化很少，通过遗忘门 f，候选

门 g 和输入门 i 进行信息更新，信息在这条线路中传输只有少量的相乘和相加操作，梯度更加稳定，因此 LSTM 对远距离的相互作用有更强大的学习能力。LSTM 单元的数学表达式如下：

$$f_t = \sigma\left(W_f * x_t + R_f * h_{t-1} + b_f\right) \tag{3}$$

$$g_t = \sigma\left(W_g * x_t + R_g * h_{t-1} + b_g\right) \tag{4}$$

$$i_t = \sigma\left(W_i * x_t + R_i * h_{t-1} + b_i\right) \tag{5}$$

$$o_t = \sigma\left(W_o * x_t + R_o * h_{t-1} + b_o\right) \tag{6}$$

其中，W 为输入权值，R 为递归权值，b 为偏置，f 代表遗忘门，g 代表候选门，i 代表输入门，o 代表输出门。首先公式 3 说明隐藏状态 h_{t-1} 和输入 x_t 一同输入至遗忘门，通过激活函数 σ 输出 $f_t \subseteq [0,1]$，0 代表全部遗忘，1 代表全部记忆。公式 4 表示候选门产生候选向量，公式 5 代表输入门来决定更新哪些值，两者的结果进行相乘操作，然后相加操作更新至细胞状态 c_t 如公式 7 所示。最后通过输出门决定最后的输出内容，可作为当前单元的隐藏状态 h_t 或输出 y 如公式 8 所示。

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \tag{7}$$

$$h_t = o_t \otimes g_t\left(c_t\right) \tag{8}$$

最后一层 Bi-LSTM 后是全连接层，全连接层中的每一个神经元都与上一层的神经元连接，将 Bi-LSTM 学习到的高维特征进行整合，最后经过 Softmax 分类器进行分类，计算每一个氨基酸所在的 C、E、H 三种结构的概率，最终完成分类。

3. 实验数据

3.1. 数据集

本次实验的训练集为 CULLPDB 数据集[15]共 15,125 条蛋白质，同源性低于 25%，涵盖类别丰富。我们剔除了同测试集中重复的蛋白质，最终训练集数量为 14,199 条蛋白质。CASP 数据集为全球蛋白质结构预测实验采用的公共测试集，我们选取了其中的 4 组作为测试集，分别为 122 条蛋白质的 CASP9，99 条蛋白质的 CASP10，81 条蛋白质的 CASP11 和 19 条蛋白质的 CASP12。

3.2. 输入特征

实验对网络的输入特征进行了新的设计，一共有 62 维，包含 20 维 PSSM 矩阵和 42 维的基团特征。位置特异性打分矩阵(Position-Specific Scoring Matrix, PSSM)富含生物进化进信息，极大提高了蛋白质二级结构预测的精度，是一种广泛使用的特征信息[16]，本次实验的 PSSM 是通过多序列对比 nr 数据库中的蛋白质，设置 PSI-BLAST 参数阈值为 0.001 和 3 次迭代生成。PSSM 形式为 20*L 矩阵，20 是特征维度，L 代表了不同蛋白质的长度。基团是指蛋白质序列中氢原子和非氢原子之间，或者非氢原子之间形成的官能团，具有稳定的结构特点，根据基团中原子相互连接的共价键进行分类，生成基团信息表[17]，共 42 种基团编号，每个氨基酸都可以用长度 42 的二进制编码唯一表示，氨基酸出现在哪些基团中就将相应位置用 1 表示，其余位置为 0，最终氨基酸的编码结果如表 1 所示。相比于传统的正交编码，基团编码包含了稳定的氨基酸结构信息，对于氨基酸的表示是直观的，用于蛋白质二级结构预测具有更好的效果[18]。因此模型的输入为一条完整的蛋白质序列，形式为 62*L 特征矩阵(L 为序列长度)，模型的输出为该序列对应长度的 C、E、H 序列。

Table 1. Amino acid radical group coding table
表 1. 氨基酸基团编码表

	42 位基团编码
A	010101100000000000000000000000000000000000
R	010101010000000000001000000110101011000000
N	010101010000000010100000000000000000000000
D	010101010000000001100000000000000000000000
C	010101000100000000000000000000000000000000
Q	010101010000000000010000100000000000000000
E	010101010000000000010000010000000000000000
G	100101000000000000000000000000000000000000
H	010101010000000000000000000000000000001000
I	010101000010101000000001000000000000000000
L	010101010000000000010011000000000000000000
K	010101010000000000010000000000010100000000
M	010101010000000000000100100001000000000000
F	010101010000000000000000000000000000000100
P	010101010000000000000000000000000000010000
S	010101001000000000000000000000000000000000
T	010101000010101000000000000000000000000000
W	010101010000000000000000000000000000000010
Y	010101010000000000000000000000000000100001
V	010101000010110000000000000000000000000000

4. 实验结果与分析

针对本次蛋白质二级结构预测，我们对模型中的超参数设置了多组对比试验，分别进行了 Bi-LSTM 层数和隐藏单元的调整，模型学习率的调整，Dropout 值的调整以及正则化系数的调整。实验调参过程体现均以 CASP10 作为测试集。

4.1. 网络层数和隐藏单元

在实验模型中，Bi-LSTM 层的数量会直接影响到网络的复杂程度以及网络的学习能力，不合理的隐藏单元数也会导致模型过拟合，在网络结构上我们希望设计一个复杂程度合理的模型，在适当的训练时间下拥有良好的泛化性。因此我们针对网络层数和隐藏单元数进行了调整实验，U_1 代表第一层 Bi-LSTM 的隐藏单元数，U_2 代表第二层 Bi-LSTM 的隐藏单元数。

Table 2. Experimental results of setting up one layer of Bi-LSTM
表 2. 设置一层 Bi-LSTM 的实验结果

U_1	200	400	600	800	1000	1200
Q_3 (%)	81.39	80.92	81.48	81.27	81.21	81.51

Table 3. Experimental results of setting up two layers of Bi-LSTM
表 3. 设置两层 Bi-LSTM 的实验结果

$U_1 = U_2$	200	400	600	800	1000	1200
Q_3 (%)	81.74	81.79	81.74	81.07	82.38	82.20

从表 2 可以看出，只设置一层 Bi-LSTM 的情况下，网络的学习能力是较为欠缺的，隐藏单元的改变并不会对实验结果有太大的影响，因此需要增加网络层数，以第一层每一时间步的隐藏单元输出作为下一层相应时间步隐藏单元的输入，网络进行了更加深层次的特征学习。实验结果表明，这种操作对结果有一定的提升，但隐藏单元数的确定便十分重要，由于是双层网络，因此两层隐藏单元数之间的选择是相互影响的，如表 3 所示，单纯的使用第一层网络下最好的方案去确定第二层的单元数，得到的结果并不是最优解，因此我们需要找到最优的组合，结果表明，两层隐藏单元为 1000 的组合下，可到的最优组合。考虑到实验的可行性和模型复杂度，最终的网络结构确定为两层 Bi-LSTM，隐藏单元都别为 1000。

4.2. 学习率

网络模型采用了 Adam 优化算法，相比于传统的随机梯度下降方式，Adam 通过计算梯度的一阶矩估计和二阶矩估计为不同的参数设计独立的自适应性学习率，适合解决含大规模数据和参数的优化问题[19]。学习率的初始值会对结果产生较大影响，因此本组实验为确定合适的初始学习率，实验结果如图 4 所示，首先以 10 倍缩量进行粗调节，然后在合适的量级下进行细调节，最终确定为 0.0006，Q3 结果可达 84.64%，训练过程中学习率以每 4 个 epoch 进行 10 倍缩放的方式训练。

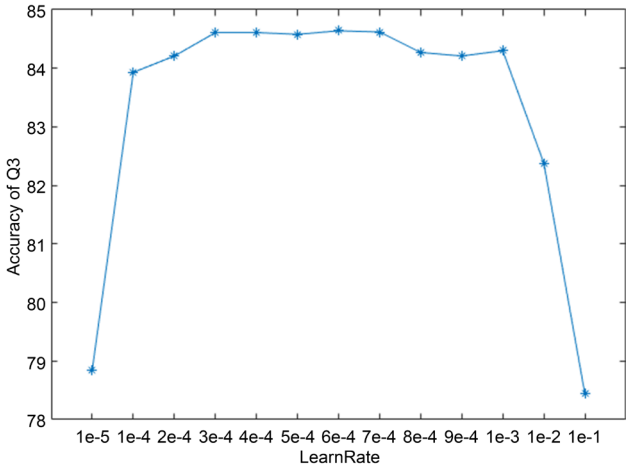

Figure 4. Experimental results at different learning rates
图 4. 不同学习率下的实验结果

4.3. Dropout

由于模型采用的双层 Bi-LSTM 的设计，同时加入了大量的隐藏单元，模型可能会存在过拟合且训练时间过长的问题，因此在每一层后引入了 Dropout，以一定的概率使部分神经元失活，我们将两层 Dropout 设置为相同值进行同步调整，从图 5 中可以看出，在数据集 CASP10 上，两个 Dropout 值同为 0.2 时，结果最优，且明显高于其他值，分析原因可能是因为 Dropout 在 0.2 的状态下失活神经元的概率更符合本模型的训练方式，这样网络模型可以获得更好的泛化性。表 4 中 D_1 代表第一层 Dropout，D_2 代表第二层 Dropout，可以看出，设置两层 Dropout 是更好的设计方案。

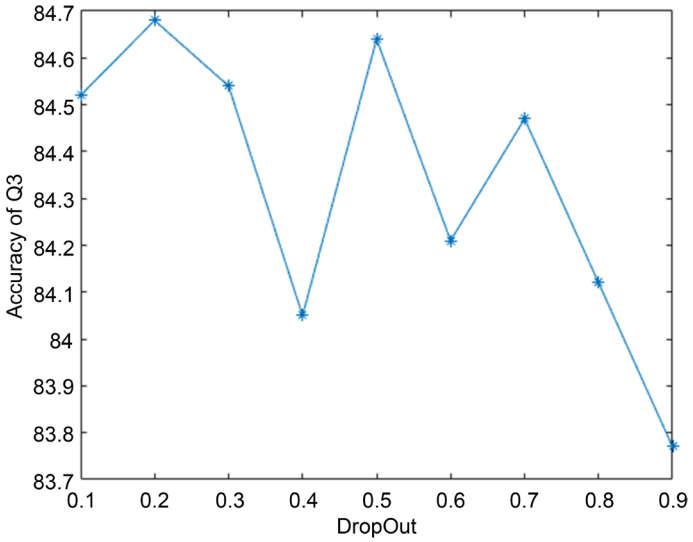

Figure 5. Experimental results of different Dropout values
图 5. 不同 Dropout 值的实验结果

Table 4. The impact of different amounts of Dropout
表 4. 不同数量 Dropout 的影响

Dropout	$D_1 = 0, D_2 = 0$	$D_1 = 0.2, D_2 = 0$	$D_1 = 0, D_2 = 0.2$	$D_1 = 0.2, D_2 = 0.2$
Q3(%)	84.61	84.52	84.57	84.68

4.4. 正则化系数

为防止模型过拟合，实验过程中加入了参数正则化方式，选择 L_2 正则化，在损失函数中增加惩罚项，L_2 正则化的惩罚项是指权值向量 w 中各个元素的平方和然后再求平方根，可表示为 $\lambda\|w\|_2^2$，调节 λ 系数控制惩罚项的大小。实验结果表明，调整 L_2 正则化系数，对实验结果有进一步的提升，在 CASP10 上结果可达 86.83%，实验结果如图 6 所示，L_2 正则化系数最终确定为 0.000005。

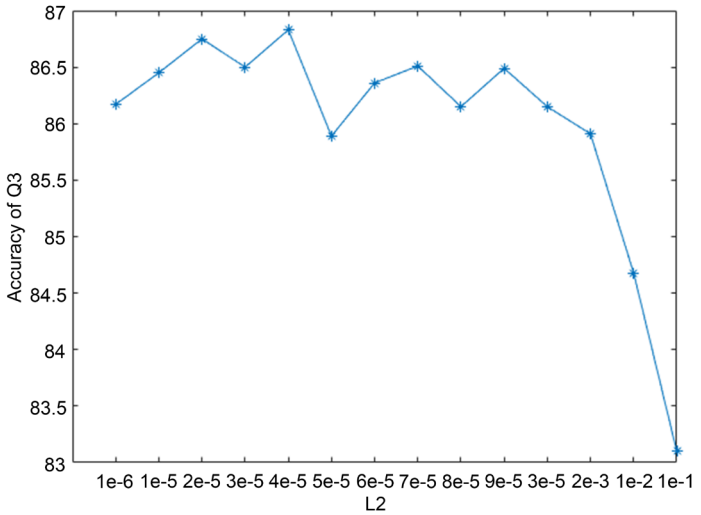

Figure 6. Experimental results of different L2 regularization coefficients
图 6. 不同 L2 正则化系数的实验结果

经过上述的参数调整，网络最终确定的结构如表 5 所示。

Table 5. Finalized network structure parameters
表 5. 最终确定的网络结构参数

参数	参数值
输入维度	62
第一层 Bi-LSTM 隐藏单元	1000
第一层 Dropout	0.2
第二层 Bi-LSTM 隐藏单元	1000
第二层 Dropout	0.2
初始学习率	0.0006
L2 正则化系数	0.000005
全连接层	3
分类器	Softmax
输出结果	C、E、H

4.5. 预测评估

1983 年，Wolfgang Kabsch 提出 DSSP，将二级结构被划分为 α 螺旋(H)、β 折叠(E)、3_{10} 螺旋(G)、β 桥(B)、π 螺旋(I)、转角(S)、β 转角(T)和无规则卷曲(C)8 种结构[20]，实验中将上述 8 态结构以“GHI→H”，“BE→E”和“其他结构→C”的规则转为 3 态结构，进行 3 态二级结构预测。CULLPDB 中的 DSSP 来自于蛋白质的 PDB 文件，评估标准为 Q_3 精度[3] [7] [8] [10] [11]，Q_3 为正确预测的氨基酸数占所有氨基酸的比例，计算公式如下：

$$Q_3 = \frac{Q_C + Q_E + Q_H}{S} * 100\%$$

将测试集输入模型，得出预测数据，同真实 DSSP 进行对比。其中，Q_C 为正确预测的转角数，Q_E 为正确预测的折叠数，Q_H 为正确预测的螺旋数，S 为总的氨基酸数。文献[10]在五折交叉验证下，使用 CULLPDB 数据集获得了 73.2%的八态分类结果。文献[21]使用 TR9993 数据集进行十折交叉验证，Q_3 准确率为 72.42%。文献[22]使用 25PDB 数据集进行三折交叉验证，Q_3 准确率可达 80.18%。表 6 为本实验使用 CULLPDB 做十折交叉验证的结果，为了保证不同长度的蛋白质均匀分布，在划分训练集和测试集之前未将 CULLPDB 根据蛋白质长度排序，数据集均分为 10 份，每次取 9 份作为训练集，1 份作为测试集，十次训练的数据集数量见表 6，十折交叉验证的平均值可达 83.08%，比 CASP 测试集结果低 0.7%~2.7%，结果表明，实验避免了数据集划分不合理而出现过拟合现象。

Table 6. 10-fold cross validation of CULLPDB
表 6. CULLPDB 十折交叉验证

验证次数	1	2	3	4	5	6	7	8	9	10	平均
训练集数量/测试集数量	12780/1419	12779/1420	12779/1420	12779/1420	12779/1420	12779/1420	12779/1420	12779/1420	12779/1420	12779/1420	-
Q_3(%)	83.54	82.36	83.30	83.09	82.68	83.17	83.70	83.19	83.47	82.32	83.08

表 7 列出了在测试集 CASP9，CASP10，CASP11 和 CASP12 [23] [24]上 Q_3 准确率以及转角、折叠和螺旋结构的预测准确率，其中 Q_3 的识别率分别有 85.74%，86.83%，84.73%和 83.79%。值得注意的是，模型对折叠 E 的预测准确值较低，分析后发现实验选取的 CASP 数据有部分蛋白质没有转角结构，相比于预测效果更加稳定的螺旋和转角结构，总体基数较少。为了验证加入 42 基团特征是有效的，我们进行了对比试验，如图 7 所示，在最终确定的网络结构下，分别以 PSSM 和 PSSM+42 基团为输入。单纯使用 PSSM 作为输入特征，在 CASP9-12 上结果分别为 85.17%，85.83%，83.54%和 82.71%。实验结果表明，在四组 CASP 测试集上，加入 42 基团特征对模型分类能力有进一步的提升，Q_3 提升约 1%。

Table 7. Experimental results under the CASP test set
表 7. CASP 测试集下实验结果

实验结果(%) / 测试集	Q_3	Q_C	Q_E	Q_H
CASP9	85.74	84.07	74.65	84.28
CASP10	86.83	85.59	79.32	83.57
CASP11	84.73	83.78	77.25	78.33
CASP12	83.79	78.98	81.93	83.42

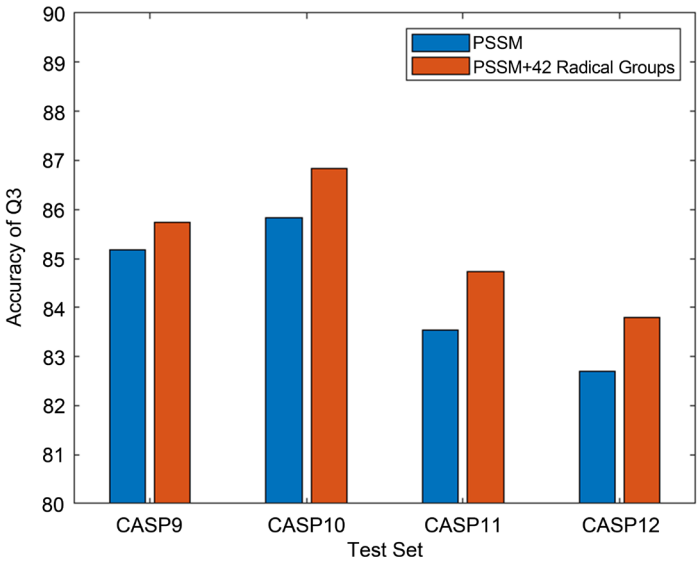

Figure 7. The influence of two characteristics of PSSM and PSSM+42 radical groups on the model
图 7. PSSM 和 PSSM+42 基团两种特征对模型的影响

实验选取了 SPINE-X [25]，SSpro [26]，PSIPred，DeepCNF 和 MUFOLD 五种预测方法，在测试集 CASP10，CASP11 和 CASP12 上进行对比，其中 SPINE-X 采用的多步神经网络，SSpro 模型是双向递归神经网络，PSIpred 采用的两层前馈神经网络，Jpred 使用两层来自 SNNS 神经网络包的人工神经网络，MUFOLD 采用深度神经网络(五种方法结果均摘自文献[2])。对比结果如表 8 所示。相比与上述模型方法，本模型使用双向长短时记忆循环网络和多维特征融合的预测方式有效的提高了 3 态蛋白质二级结构预测结果。

Table 8. Comparison of prediction results
表 8. 预测结果对比

模型	特征输入	CASP10	CASP11	CASP12
SPINE-X	PSSM+物理特性	80.7	79.3	76.9
SSpro	PSSM	78.5	77.6	–
PSIpred	PSSM	81.2	80.7	78.0
Jpred	PSSM+隐马尔可夫模型特征	81.6	80.4	75.1
DeepCNF	PSSM+21 个元素的二进制向量	84.4	84.7	82.1
MUFOLD	氨基酸理化性质+PSSM+隐马尔可夫模型特征	85.98	83.59	80.59
本文模型	PSSM+42 基团特征	86.83	84.73	83.79

5. 结论

蛋白质二级结构预测在蛋白质结构研究领域意义重大，不断有新的模型和方法被提出，本文采用了双向长短时记忆循环网络，在蛋白质序列特征构建上融合了新的 42 基团编码方式，数据包含更多特征，同时使用大数据集进行模型训练，取消了滑动窗口的设计，最大化的捕捉氨基酸之间的长距离相互作用，双向 LSTM 又可以考虑到氨基酸序列前后文的影响，提高了 3 态二级结构的预测结果，相比于多种深度学习方法融合的预测方式，模型更加简洁有效，但取消滑动窗口的训练方式人为增加了训练集的噪声，下一步可进一步改进多维特征构建方式，构建更加有效的特征编码，在网络训练过程方面参数的自动优化将是更加高效的方式，搭建新的网络模型提高预测结果。

基金项目

国家自然科学基金(No. 61375013)，山东省自然科学基金(No. ZR2013FM020)资助。

参考文献

[1] Jiang, Q., Jin, X., Lee, S.J., *et al*. (2017) Protein Secondary Structure Prediction: A Survey of the State of the Art. *Journal of Molecular Graphics & Modelling*, **76**, 379-402. https://doi.org/10.1016/j.jmgm.2017.07.015

[2] Yang, Y., Gao, J., Wang, J., *et al*. (2018) Sixty-Five Years of the Long March in Protein Secondary Structure Prediction: The Final Stretch. *Briefings in Bioinformatics*, **19**, 482-494.

[3] Ma, Y., Liu, Y. and Cheng, J. (2018) Protein Secondary Structure Prediction Based on Data Partition and Semi-Random Subspace Method. *Scientific Reports*, **8**, Article No. 9856. https://doi.org/10.1038/s41598-018-28084-8

[4] 刘斌, 温雪岩. 优化多核 SVM 的蛋白质二级结构预测[J]. 现代电子技术, 2020, 43(8): 139-142.

[5] Lasfar, M. and Bouden, H. (2018) A Method of Data Mining Using Hidden Markov Models (HMMs) for Protein Secondary Structure Prediction. *Procedia Computer Science*, **127**, 42-51. https://doi.org/10.1016/j.procs.2018.01.096

[6] Drozdetskiy, A., Cole, C., Procter, J., *et al*. (2015) JPred4: A Protein Secondary Structure Prediction Server. *Nucleic Acids Research*, **43**, 389-394. https://doi.org/10.1093/nar/gkv332

[7] Jones, D. (1999) Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. *Journal of Molecular Biology*, **292**, 195-202. https://doi.org/10.1006/jmbi.1999.3091

[8] 郭延哺, 李维华, 王兵益, 等. 基于卷积长短时记忆神经网络的蛋白质二级结构预测[J]. 模式识别与人工智能, 2018, 31(6): 562-568.

[9] Fang, C., Shang, Y. and Xu, D. (2018) MUFOLD-SS: New Deep Inception-inside-Inception Networks for Protein Secondary Structure Prediction. *Proteins*: *Structure*, *Function and Bioinformatics*, **86**, 592-598. https://doi.org/10.1002/prot.25487

[10] Wang, S., Peng, J., Ma, J., *et al*. (2016) Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Scientific Reports*, **6**, Article No. 18962. https://doi.org/10.1038/srep18962

[11] Heffernan, R., Yang, Y., Kuldip, P., *et al*. (2017) Capturing Non-Local Interactions by Long Short-Term Memory Bi-directional Recurrent Neural Networks for Improving Prediction of Protein Secondary Structure, Backbone Angles, Contact Numbers, and Solvent Accessibility. *Bioinformatics*, **33**, 3842-3849. https://doi.org/10.1093/bioinformatics/btx218

[12] Hanson, J., Paliwal, K., Litfin, T., *et al*. (2018) Improving Prediction of Protein Secondary Structure, Backbone Angles, Solvent Accessibility and Contact Numbers by Using Predicted Contact Maps and an Ensemble of Recurrent and Residual Convolutional Neural Networks. *Bioinformatics*, **35**, 2403-2410. https://doi.org/10.1093/bioinformatics/bty1006

[13] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[14] Gers, F.A., Schmidhuber, J. and Cummins, F. (2000) Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, **12**, 2451-2471. https://doi.org/10.1162/089976600300015015

[15] Wang, G. and Dunbrack, R. (2005) PISCES: Recent Improvements to a PDB Sequence Culling Server. *Nucleic Acids Research*, **33**, W94-W98. https://doi.org/10.1093/nar/gki402

[16] 泽瓦勒贝 M, 等. 理解生物信息学[M]. 李亦学, 郝沛, 译. 北京: 科学出版社, 2012.

[17] 沈世镒. 蛋白质分析与数学: 生物、医学与医药卫生中的定量化研究.上册[M]. 北京: 科学出版社, 2014.

[18] 张帅燕, 刘毅慧. 基于一种新的基团编码的蛋白质二级结构预测[J]. 智能计算机与应用, 2017, 7(3): 13-16.

[19] Kingma, D. and Ba, J. (2014) Adam: A Method for Stochastic Optimization. Computer Science.

[20] Kabsch, W. and Sander, C. (1983) Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, **22**, 2577-2637. https://doi.org/10.1002/bip.360221211

[21] Heffernan, R., Paliwal, K., Lyons, J., *et al*. (2018) Single-Sequence-Based Prediction of Secondary Structures and Solvent Accessibility by Deep Whole-Sequence Learning. *Computers & Chemistry*, **39**, 2210-2216. https://doi.org/10.1002/jcc.25534

[22] Cheng, J., Liu, Y. and Ma, Y. (2020) Protein Secondary Structure Prediction Based on Integration of CNN and LSTM Model. *Journal of Visual Communication and Image Representation*, **71**, Article ID: 102844. https://doi.org/10.1016/j.jvcir.2020.102844

[23] Moul, T.J., Fidelis, K., Kryshtafovych, A., *et al*. (2011) Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round IX. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **79**, 1-5. https://doi.org/10.1002/prot.23200

[24] Moul, T.J., Fidelis, K., Kryshtafovych, A., *et al*. (2014) Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round X. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **82**, 1-6. https://doi.org/10.1002/prot.24452

[25] Faraggi, E., Zhang, T., Yang, Y., *et al*. (2012) SPINE X: Improving Protein Secondary Structure Prediction by Multistep Learning Coupled with Prediction of Solvent Accessible Surface Area and Backbone Torsion Angles. *Journal of Computational Chemistry*, **33**, 259-267. https://doi.org/10.1002/jcc.21968

[26] Magnan, C.N. and Pierre, B. (2014) SSpro/ACCpro 5: Almost Perfect Prediction of Protein Secondary Structure and Relative Solvent Accessibility Using Profiles, Machine Learning and Structural Similarity. *Bioinformatics*, **30**, 2592-2597. https://doi.org/10.1093/bioinformatics/btu352