# 基于卷积长短时记忆神经网络的蛋白质二级结构预测

郭延哺[1] 李维华[1] 王兵益[2] 金 宸[1]

**摘 要** 鉴于不同类型氨基酸的相互作用对蛋白质结构预测的影响不同,文中融合卷积神经网络和长短时记忆神经网络模型,提出卷积长短时记忆神经网络,并应用到蛋白质 8 类二级结构的预测中. 首先基于氨基酸序列的类别信息和氨基酸结构的进化信息表示蛋白质序列,并采用卷积提取氨基酸残基之间的局部相关特征,然后利用双向长短时记忆神经网络提取蛋白质序列内部残基之间的远程相互作用,最后将提取的蛋白质的局部相关特征和远程相互作用用于蛋白质 8 类二级结构的预测. 实验表明,相比基准方法,文中模型提高 8 类二级结构预测的精度,并具有良好的可扩展性.

**关键词** 生物信息学, 蛋白质二级结构, 卷积神经网络, 长短时记忆神经网络

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

**DOI** 10.16451/j.cnki.issn1003-6059.201806009 **中图法分类号** TP 391

# Protein Secondary Structure Prediction Based on Convolutional Long Short-Time Memory Neural Networks

GUO Yanbu[1], LI Weihua[1], WANG Bingyi[2], JIN Chen[1]

**ABSTRACT** Since the interaction of different types of amino acid has an influence on the prediction of protein structure, convolutional neural networks and long short-term memory neural networks are integrated. A convolutional long short-term memory neural network is proposed to predict 8-class protein secondary structures. Firstly, the protein sequence is represented based on the amino acid sequence class feature and the amino acid structure profile feature. The local correlation characteristics between amino acid residues are extracted by the convolutional operations, and then the long-range interactions between the residues on protein sequences are extracted by the bi-directional long short-term memory network. Finally, the local correlation characteristics and long-range interactions between amino acid residues are employed to predict protein secondary structures. Experimental results show that the proposed model achieves a higher accuracy than the baselines and the framework has good scalability.

**Key Words** Bioinformatics, Protein Secondary Structure, Convolutional Neural Networks, Long Short-Term Memory Neural Networks

收稿日期:2018-01-03;录用日期:2018-03-16
Manuscript received January 3;
accepted March 16, 2018
国家自然科学基金项目(No. 11661081)、教育部科技发展中心"云数融合科教创新"基金(No. 2017B00016)、云南省科技创新人才培养项目、云南省创新团队项目资助
Supported by National Natural Science Foundation of China(No. 11661081), Integration of Cloud Computing and Big Data, Innovation of Science and Education(No. 2017B00016), Training Project of Scientific and Technological Innovation Talents in Yunnan, Project of Innovative Research Team of Yunnan Province
本文责任编委 周水庚
Recommended by Associate Editor ZHOU Shuigeng

1. 云南大学 信息学院 昆明 650500
2. 中国林业科学研究院 资源昆虫研究所 昆明 650224
1. School of Information Science and Engineering, Yunnan University, Kunming 650500
2. The Research Institute of Resource Insects, Chinese Academy of Forestry, Kunming 650224

**Citation**　GUO Y B, LI W H, WANG B Y, JIN C. Protein Secondary Structure Prediction Based on Convolutional Long Short-Time Memory Neural Networks. Pattern Recognition and Artificial Intelligence, 2018, 31(6): 562-568.

蛋白质是组成生物体的重要成分,可以完成免疫、细胞信号传输等功能. 蛋白质结构常分为:一级结构、二级结构、三级结构和四级结构[1-2]. 随着后基因组时代的到来,海量的蛋白质数据已积累完成. 然而传统的实验测定方法由于高昂费用和实验难度[3],已无法满足日益增长的蛋白质和结构数据分析的需求. 因此,蛋白质结构预测的计算方法就成为计算生物学研究的热点问题. 蛋白质三级结构很大程度上取决于蛋白质的一级序列信息[4],直接由蛋白质一级序列预测蛋白质三级结构极其困难. 由于蛋白质一级结构对应的蛋白质二级结构可以有效降低并解决三级结构预测的难度,因此蛋白质二级结构预测被作为三级结构预测的桥梁.

蛋白质二级结构预测问题的研究由来已久,已有很多计算方法用于蛋白质二级结构预测[2,5-6]. 这些方法主要集中在蛋白质序列的特征表示和蛋白质二级结构预测方法两方面. 蛋白质特征表示使用合适的数学模型完整表示蛋白质的序列和结构的信息. 通过蛋白质序列表示进行二级结构预测方法有:基于氨基酸组成及位置的方法[7]和基于氨基酸物理化学特征提取方法[5]. 鉴于使用氨基酸组成表示蛋白质信息较有限,也有基于多特征融合的技术[8-9]表示蛋白质序列,如基于蛋白质的序列结构(Sequence Profiles, SP)的进化信息[10]和通过位置特异性迭代的基本局部比对搜索工具(Position-Specific Iterative Basic Local Alignment Search Tool, PSI-BLAST)[11]产生的位置特定的评分矩阵(Position-Specific Scoring Matrices, PSSM)[12]表示蛋白质序列.

蛋白质二级结构预测通常采用基于统计分析的预测方法[7]和基于机器学习的预测方法[5]. 在传统的结构预测方法中,蛋白质特征提取很大程度上依赖于人工设计. 针对蛋白质序列特征表示的难题,Qian 等[13]使用非线性神经网络模型提取蛋白质特征,并预测蛋白质二级结构. 近期,研究者将循环神经网络(Recurrent Neural Networks, RNN)[14]和卷积神经网络(Convolutional Neural Networks, CNN)[11]用于蛋白质二级结构预测,成功改善蛋白质结构预测精度.

蛋白质二级结构可以分为 3 类[1]或 8 类[15]. 研究者早期主要关注蛋白质 3 类二级结构预测. 针对 8 类二级结构预测方法的研究总结如下. Wang 等[16]基于蛋白质序列的 PSSM 和位置无关的特征,如理化特征,采用条件神经域(Conditional Neural Fields, CNF)构建预测 8 类蛋白质二级结构的预测模型. Zhou 等[11]基于蛋白质序列离散的氨基酸类型特征,利用 PSI-BLAST、PSSM 产生的蛋白质序列列型特征,采用监督学习的卷积生成随机网络(Deep Convolutional Generative Stochastic Networks, DCGSN),进行蛋白质二级结构预测. Sønderby 等[17]基于蛋白质离散的氨基酸类型特征和蛋白质序列列型特征,采用双向长短时记忆(Long Short-Term Memory, LSTM)神经网络构建 8 类二级结构预测模型. Li 等[8]基于密集的蛋白质氨基酸类型特征和蛋白质序列列型特征,采用级联的卷积循环神经网络(Cascaded Convolutional and RNN, Cascaded CRNN)预测 8 类蛋白质二级结构. Busia 等[18]基于规则化的氨基酸类型和 PSSM 特征,采用多卷积核的卷积神经网络(Multi-scale CNN, MCNN)进行 8 类蛋白质二级结构的预测. Wang 等[9]以 PSSM 特征作为深度卷积神经域(Deep Convolutional Neural Fields, Deep CNF)的输入,使用最大化对数似然目标函数进行模型参数学习,获得较好的 8 类蛋白质二级结构预测精度. 虽然神经网络模型在蛋白质二级结构预测中取得较好效果,但仍未充分利用蛋白质序列的复杂局部和远程相互作用,8 类蛋白质二级结构预测仍是一个具有挑战性的研究方向[16].

蛋白质二级结构取决于蛋白质序列氨基酸的局部和远程相互作用[1,11,17]. 由于缺乏完备的专业知识用于设计多种蛋白质的特征提取模式,自动提取蛋白质序列特征[11]对蛋白质结构预测显得尤其重要. 深度神经网络在图像[19]和自然语言处理[20]方面已表现出强大的自动学习数据中复杂结构关系的能力. CNN 和 LSTM 神经网络是广泛应用的神经网络方法[11,14]. CNN 虽然提取局部复杂的特征依赖关系,但不能提取氨基酸之间的远程作用. LSTM 是循环神经网络的拓展,不仅可以处理序列之间的远程依赖的问题,还可以避免梯度消失的问题.

基于上述原因,本文融合 CNN 和 LSTM 神经网络,提出基于卷积长短时记忆(Convolutional LSTM, C-LSTM)神经网络,进行蛋白质 8 类二级结构预测. 首先通过多卷积核提取氨基酸残基间的局部相关特征,然后基于蛋白质氨基酸残基的局部特征,使用双

向 LSTM 神经网络提取氨基酸间的远程作用特征,最后将蛋白质的局部相关特征与远程作用特征融合作为蛋白质的特征表示,实现蛋白质的 8 类二级结构预测.

# 1 蛋白质二级结构预测模型

蛋白质二级结构预测是给定蛋白质序列数据,预测每个氨基酸残基对应的二级结构类型. 每个蛋白质序列的二级结构由氨基酸残基之间的局部和远程相互作用决定. 针对蛋白质序列内部的氨基酸残基的局部相互作用关系,本文采用多卷积核的 CNN,提取蛋白质序列的氨基酸之间复杂的局部作用与位置信息. 对于蛋白质序列氨基酸之间的复杂的远程相互作用,在多卷积核 CNN 提取的局部作用信息的基础上,采用 2 层双向 LSTM 提取蛋白质序列中氨基酸之间的远程作用信息. 本文提出的蛋白质 8 类二级结构预测模型如图 1 所示.

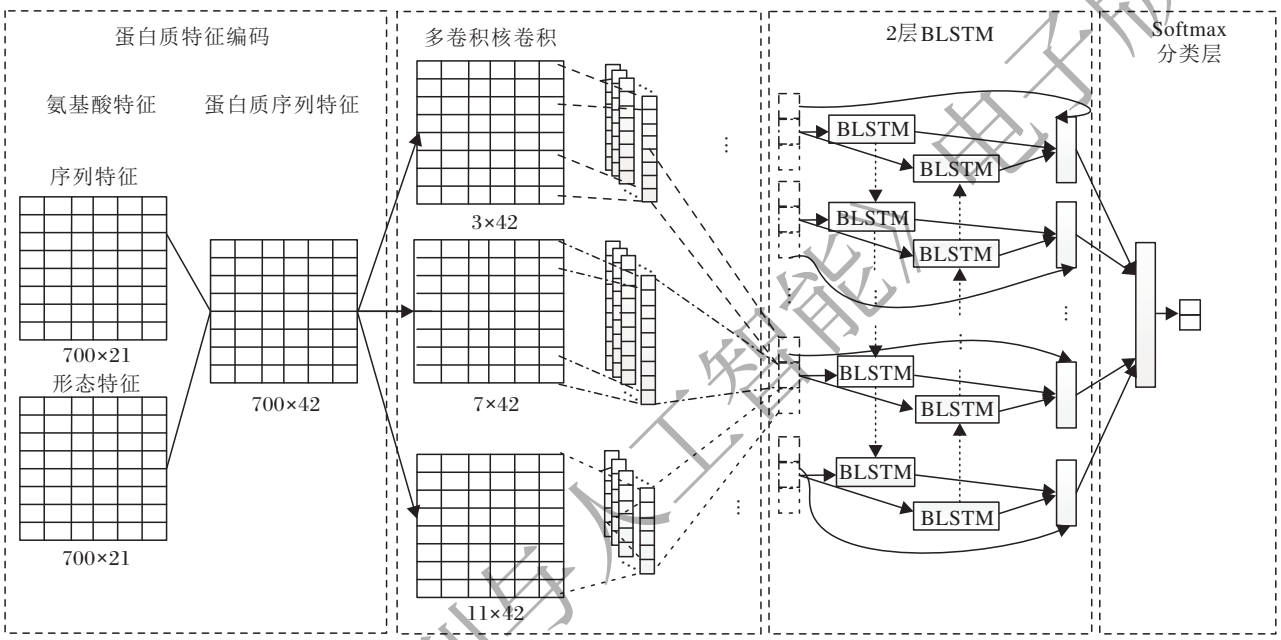


图 1 8 类蛋白质二级结构预测模型

Fig. 1 Prediction model of eight-classes protein secondary structure

## 1.1 蛋白质序列编码

蛋白质二级结构预测中常用的氨基酸编码方式有正交编码、5 位编码、Codon 编码和 Profile 编码. 实验表明使用富含生物进化信息的 Profile 编码方式可以得到较高的预测精度,即充分利用蛋白质序列的生物特征信息可以有效提高蛋白质二级结构预测精度[21]. 蛋白质序列含有 $n$ 个氨基酸,表示为 $\boldsymbol{P}=\{\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_n\}$,其中 $\boldsymbol{R}_i\in\mathbf{R}^m$ 表示氨基酸序列 $\boldsymbol{P}$ 的第 $i$ 个位置的氨基酸的特征向量. 对应的二级结构标签可表示为 $\boldsymbol{S}=\{\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_n\}$,其中 $\boldsymbol{S}_i\in\mathbf{R}^8$ 表示在第 $i$ 个位置的二级结构类型.

本文采用的蛋白质序列特征由蛋白质序列中的氨基酸的类型信息和蛋白质序列的进化特征信息组成,正如文献[11]和文献[22]所示. 每个氨基酸表示成42维向量,前21维是正交编码的氨基酸类型信息,后21维是采用 PSI-BLAST 和 PSSM 谱编码产生的与蛋白质氨基酸的进化相关的信息.

## 1.2 多卷积核卷积层

蛋白质序列的二级结构类型受序列中的氨基酸残基之间的局部作用关系影响,鉴于文献[8]、文献[9]、文献[23]和文献[24]提出的基于 CNN 和 $n$-gram 的蛋白质序列的分布式表示和特征提取方法,本文采用多种卷积核 $f_j(j=1,2,\cdots,n)$ 提取蛋白质序列内部的氨基酸局部近距离的作用关系,按

$$\boldsymbol{l}_i^{f_j}=\boldsymbol{F}*\boldsymbol{R}_{i:i+f_i-1}=Relu(w_{f_j}\cdot R_{i:i+f_j-1}+b_{f_j}),$$

$$\boldsymbol{L}^{f_j}=[\boldsymbol{l}_1^{f_j},\boldsymbol{l}_2^{f_j},\cdots,\boldsymbol{l}_n^{f_j}],$$

得到输出序列 $\boldsymbol{L}^{f_j}$,其中,$\boldsymbol{F}\in\mathbf{R}^{f_j\times m}$、$\boldsymbol{l}^{f_j}\in\mathbf{R}^q$,$\boldsymbol{F}$ 表示卷积窗口函数,$f_j$ 表示对氨基酸序列进行卷积的序列长度,$m$ 表示每个氨基酸的特征维度,$q$ 表示卷积核 $f_j$ 的数目,$w_{f_j}$、$b_{f_j}$ 分别表示长度为 $f_j$ 的卷积核的权重项、偏置项.

本文通过 3 种卷积核($f_1=3$, $f_2=7$, $f_3=11$)卷

积操作得到特征映射(Feature Map),分别为 $\boldsymbol{L}^{f_1}$,$\boldsymbol{L}^{f_2}$,$\boldsymbol{L}^{f_3}$. 然后将提取到的蛋白质序列内部氨基酸之间的局部近距离作用特征连接得到蛋白质序列氨基酸的局部作用特征 $\boldsymbol{L}=[\boldsymbol{L}^{f_1},\boldsymbol{L}^{f_2},\boldsymbol{L}^{f_3}]$.

### 1.3　长短时记忆层

蛋白质的氨基酸残基之间除了局部作用关系,也有长距离相互作用. LSTM 神经网络网络是 RNN 的扩展,针对长期依赖缺失的问题而设计. 与 RNN 神经网络不同,LSTM 的循环单元模块具有不同的结构单元,存在 4 个以特殊方式相互影响的神经元. 在 LSTM 神经网络中,通过门机制对神经单元状态进行更新,门结构的作用是选择性是否让附近氨基酸的信息更新当前位置的特征表示. LSTM 神经元结构如图 2 所示.

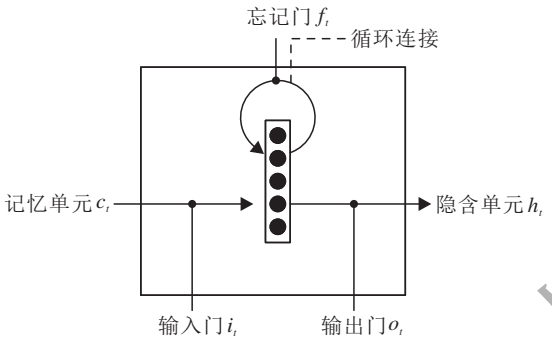


图 2　LSTM 神经元结构图

Fig. 2　LSTM unit structure

LSTM 神经元具有输入门、忘记门和输出门 3 种门结构,保持和更新记忆单元状态如下:

$$f_t=\sigma(w_{xf}l_t+w_{hf}l_{t-1}+b_f),$$

$$i_t=\sigma(w_{xi}l_t+w_{hi}l_{t-1}+b_i),$$

$$c_t=f_tc_{t-1}+i_t\arctan(w_{xc}l_t+w_{hc}l_{t-1}+b_c),$$

$$o_t=\sigma(w_{xo}l_t+w_{ho}l_{t-1}+b_o),$$

$$h_t=o_t\tanh c_t,$$

$$\sigma(z)=\frac{1}{1+\exp(-z)},$$

其中,$i_t$、$f_t$、$o_t$、$c_t$ 表示 $t$ 时刻对应的 3 种门结构和细胞状态,$\sigma(z)$ 为非线性激活函数,$h_t$ 为时刻 $t$ 隐含单元状态.

本文在多卷积核卷积层之后,采用双向 LSTM 神经网络自动提取蛋白质序列的氨基酸残基之间的远程依赖关系. 为了更有效地提取并强化蛋白质序列的氨基酸残基之间复杂的长距离依赖信息,本文采用 2 层双向 LSTM 神经网络. 最后,提取氨基酸残基之间的局部作用信息和远程作用信息,计算并连接融合,得到输出层的输入特征 $\boldsymbol{h}_t$:

$$\overrightarrow{h_t}=LSTM\left(h_t^{z-1},\overrightarrow{h_{t-1}^{z}}\right),$$

$$\overleftarrow{h_t}=LSTM\left(h_t^{z-1},\overleftarrow{h_{t+1}^{z}}\right),$$

$$\boldsymbol{h}_t=\left[\overrightarrow{h_t}\oplus\overleftarrow{h_t},\boldsymbol{L}\right],$$

其中,$\overrightarrow{h_t}$、$\overleftarrow{h_t}$ 分别表示长短时记忆层基于前 $t-1$、后 $n-t$ 个氨基酸残基,在第 $t$ 个位置提取到的特征表示.

### 1.4　输出层

本文提出的卷积长短时记忆神经网络提取的蛋白质序列特征表示为 $\boldsymbol{h}=[h_1,h_2,\cdots,h_t]$,将提取得到的完整的蛋白质序列特征传入 softmax 层,预测蛋白质序列每个氨基酸类别的概率:

$$p_i(y\mid s)=softmax(\boldsymbol{w}^s\boldsymbol{h}+b^s),$$

$$softmax(z)=\frac{e^z}{\sum e^z},$$

其中,$\boldsymbol{w}^s$、$b^s$ 分别为 $softmax$ 层的权重项、偏置项,$p_i$ 为第 $i$ 个氨基酸残基的二级结构类别的预测概率.

### 1.5　模型训练

本文使用随机梯度下降 Adma 算法(Adaptive Moment Estimation)[25] 训练本文模型. 该模型的训练目标是最小化交叉熵损失函数:

$$L(\theta)=-\frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{c}y_{kj}(\log_2(p_{kj}))+\lambda\parallel\theta\parallel^2,$$

其中,$\lambda$ 表示 $l_2$ 范数的正则化项,$y_{kj}$ 表示二级结构类别,$c$ 表示蛋白质二级结构的类别数 $k_j$,$c=8$,$N$ 表示训练集样本数.

参数调节方法:

$$\theta\leftarrow\theta+\alpha\frac{\partial L(\theta)}{\partial\theta},$$

其中,$\alpha$ 为学习率,$\theta$ 为模型所有的参数. 本文采用 Dropout、Early stopping 和正则化策略以避免过拟合程度.

## 2　实验及结果分析

本文实验环境的主要参数如下. 处理器:Intel i7-7700 CPU 3.60 GHz. 图形加速卡:NVIDIA GeForce GTX 1060 6 GB. 内存 8 GB. 操作系统:Ubuntu 16.04 LTS(64bit). 采用深度学习框架 TensorFlow0.12、Keras1.2 构建神经网络.

为了评估本文模型的准确率与鲁棒性,采用 4 个公开的蛋白质数据集:CB6133、CB513、CASP11 和

CASP10. 具体实验数据集信息如表1所示.

CB6133为一个非同源蛋白质结构数据集,共有6 128条蛋白质,其中,5 560条为训练集,256条为验证集,272条为测试集. CB513数据集为一个基准测试数据集,但是与CB6133数据集有冗余的蛋白质序列,与Zhou等[11]一样,本文删除CB6133与CB513之间存在序列一致性大于25%的蛋白质序列,最终得到一个筛选CB6133,共有5 534条蛋白质. 本文与文献[17]一样,当使用筛选CB6133数据集训练模型时,从中随机选择256条作为验证集,其余作为训练集,然后分别采用CB513、CASP10和CASP11数据集测试模型预测的准确率.

**表1 实验数据集的统计信息**

Table 1 Statistical information for experimental datasets

| 名称 | 训练集 | 验证集 | 测试集 | 共计 |
|---|---|---|---|---|
| CB6133 | 5600 | 256 | 272 | 6133 |
| 筛选 CB6133 | 5278 | 256 | 0 | 5534 |
| CB513 | 0 | 0 | 513 | 513 |
| CASP10 | 0 | 0 | 123 | 123 |
| CASP11 | 0 | 0 | 105 | 105 |

在本文模型中,需要设定的参数主要是多卷积核卷积层和长短时记忆层. 针对具体实验,设置各个参数如下.

蛋白质序列数据特征块尺寸大小为700×42.

多卷积核卷积层中,第1种卷积核的个数及尺寸大小为

$$n_1 \times f_1 \times m = 150 \times 3 \times 42.$$

第2种卷积核的个数及尺寸大小为

$$n_2 \times f_2 \times m = 150 \times 7 \times 42.$$

第3种卷积核的个数及尺寸大小为

$$n_3 \times f_3 \times m = 150 \times 11 \times 42.$$

在长短时记忆层中,双向LSTM的层数为2,每层LSTM中隐含单元数为300.

针对蛋白质8类二级结构预测问题,本文设计如下3个实验.

实验1. 为了测试本文提出的基于卷积长短时记忆(C-LSTM)神经网络的蛋白质8类二级结构预测性能,采用基准模型:双向循环神经网络(Bidirectional Recurrent Neural Networks, BRNN)[14]、深度卷积生成随机网络(Deep Convolutional Generative Stochastic Networks, DC-GSN)[11]、CNF[16]、Deep CNF[9]和LSTM[17]. 与基准模型一样,也采用8类二级结构预测准确率作为本文的性能评测指标. 本文模型与基准模型在CB6133、CB513、CASP10和CASP11蛋白质序列数据集上进行性能对比.

由于原文中DC-GSN只给出在CB6133和CB513数据集中的蛋白质8类二级结构预测精度,LSTM也只给出在CB513数据集上的预测精度,因此本文模型的性能未与DC-GSN和LSTM在CASP10、CASP11和CB6133上的预测精度进行对比.

表2给出本文的C-LSTM模型和基准模型在CB6133、CB513、CASP10和CASP11这4个数据集上的蛋白质8类二级结构的预测精度结果. 由表2可看出,本文模型性能在CB513、CASP10数据集上优于其它基准方法,然而在CB6133和CASP11数据集上,本文模型预测精度低于Deep CNF.

**表2 在4个数据集上各模型的预测准确率对比**

Table 2 Prediction performance comparison of different methods on 4 datasets

| 模型 | CB6133 | CB513 | CASP10 | CASP11 |
|---|---|---|---|---|
| BRNN | 66.6 | 63.5 | 64.9 | 65.6 |
| CNF | 69.7 | 64.9 | 64.8 | 65.1 |
| Deep CNF | **75.2** | 68.3 | 71.8 | **72.3** |
| DC-GSN | 72.1 | 66.4 | – | – |
| LSTM | – | 67.4 | – | – |
| 本文模型 | 73.3 | **69.2** | **73.6** | 70.9 |

由表2还可以看出,相比BRNN、CNF、LSTM和DC-GSN,本文模型在4组蛋白质数据集上的蛋白质8类二级结构预测精度明显提高,可见考虑蛋白质序列的氨基酸残基的远程和局部相互作用更符合蛋白质结构预测的实际特点,同时多卷积核的CNN和双向LSTM更能综合并提取蛋白质序列的所有特征,提高8类二级结构预测的性能. 相比Deep CNF,本文模型在CASP10和CB513数据集上8类二级结构的预测准确率更高,然而在CB6133和CASP11数据集上,本文模型的性能略低,主要原因可能是Deep CNF结合条件随机场模型,而条件随机场考虑到相邻氨基酸残基标签的关系,有助于推理全局最优的标记.

实验2. 为了验证本文模型中多卷积窗口对8类蛋白质二级结构预测性能提升的必要性,在2层叠加的双向LSTM神经网络中,设置7组卷积窗口进行对比实验,分别是3,7,11,3与7,3与11,7与11,3、7与11,并测试在7种卷积窗口下,本文的卷积长短时记忆神经网络分别在4个蛋白质序列数据集上的8类二级结构的预测准确率.

表3给出在7种不同卷积核尺寸下,本文模型在

4 个数据集上的预测精度对比结果. 由表可知,在 3 种单独的卷积核(3、7、11) 情况下,卷积核为 3 时,本文模型的预测精度优于卷积尺寸 7 或 11. 相比其它卷积,在 3 种卷积核(3、7、11) 共同作用下,本文模型的预测性能最好,可以更好地提取蛋白质氨基酸残基的局部相互作用.

**表 3　在 4 个数据集上不同卷积窗口的预测性能对比**

Table 3　Performance comparison of multi-scale convolution windows on 4 datasets

| 卷积窗口 | CB6133 | CB513 | CASP10 | CASP11 |
|---|---|---|---|---|
| 3 | 72.8 | 68.8 | 72.7 | 70.4 |
| 7 | 72.4 | 68.3 | 72.2 | 70.3 |
| 11 | 72.3 | 68.5 | 71.8 | 69.5 |
| 3、7 | 72.8 | 69.0 | 73.0 | 70.6 |
| 3、11 | 73.1 | 68.7 | 72.7 | 70.2 |
| 7、11 | 72.7 | 68.5 | 72.0 | 69.6 |
| 3、7、11 | **73.3** | **69.2** | **73.6** | **70.9** |

实验3. 为了验证本文模型中2层双向长短时记忆神经网络对 8 类蛋白质二级结构预测性能提升的必要性. 在实验 2 的卷积窗口的基础上,分别测试 1 层双向 LSTM 和2层双向 LSTM 在4组蛋白质序列数据集上 8 类蛋白质二级结构的预测准确率.

表4给出在不同层的双向 LSTM 下,基于不同卷积核,本文模型在 4 个数据集上的预测精度. 由表可知,在 2 层双向 LSTM 下,本文模型的蛋白质二级结构预测性能最好,因此采用2层双向 LSTM 和3种卷积核(3、7、11) 构造本文的蛋白质结构预测模型.

**表 4　在 4 个数据集上不同层数的双向 LSTM 的预测性能对比**

Table 4　Performance comparison of different layers of bidirectional LSTM on 4 datasets

| 卷积窗口 | 双向 LSTM 层数 | CB6133 | CB513 | CASP10 | CASP11 |
|---|---|---|---|---|---|
| 3 | 1 层 | 67.7 | 64.1 | 65.8 | 64.0 |
| | 2 层 | 72.8 | 68.8 | 72.7 | 70.4 |
| 7 | 1 层 | 70.6 | 66.2 | 69.1 | 67.3 |
| | 2 层 | 72.4 | 68.3 | 72.2 | 70.3 |
| 11 | 1 层 | 70.8 | 66.8 | 69.6 | 68.1 |
| | 2 层 | 72.3 | 68.5 | 71.8 | 69.5 |
| 3、7 | 1 层 | 70.9 | 67.1 | 69.9 | 68.0 |
| | 2 层 | 72.8 | 69.0 | 73.0 | 70.6 |
| 3、11 | 1 层 | 71.4 | 67.7 | 70.1 | 68.7 |
| | 2 层 | 73.1 | 68.7 | 72.7 | 70.2 |
| 7、11 | 1 层 | 71.5 | 67.7 | 70.3 | 68.4 |
| | 2 层 | 72.7 | 68.5 | 72.0 | 69.6 |
| 3、7、11 | 1 层 | 72.2 | 68.2 | 71.7 | 69.6 |
| | 2 层 | **73.3** | **69.2** | **73.6** | **70.9** |

# 3　结 束 语

蛋白质二级结构预测精度的提高对于人们全面了解蛋白质结构和功能极其重要. 针对蛋白质二级结构的预测受氨基酸残基之间的远程和短程相互作用的影响,本文融合卷积神经网络和长短记忆神经网络,提出卷积长短时记忆神经网络,并将其应用于蛋白质 8 类二级结构的预测中. 本文模型可以基于多卷积核的卷积操作提取氨基酸残基之间局部相互作用特征和位置信息,同时基于双向长短期记忆网络提取蛋白质序列内部残基的远程相互作用特征,并将蛋白质氨基酸之间的局部和远程作用特征合并用于二级结构预测. 实验表明,本文模型可以有效提高蛋白质 8 类二级结构的预测精度,具有较好的拓展性. 本文模型忽略蛋白质序列的二级结构类型之间的关联作用,如何结合二级结构类型之间的关联作用,改进蛋白质结构预测性能将是今后改进方向.

## 参　考　文　献


[1] 张海仓,高玉娟,邓明华,等. 蛋白质中残基远程相互作用预测算法研究综述. 计算机研究与发展, 2017, 54(1): 1-19.
(ZHANG H C, GAO Y J, DENG M H, *et al*. A Survey on Algorithms for Protein Contact Prediction. Journal of Computer Research and Development, 2017, 54 (1): 1-19.)

[2] 张燕平,查永亮,赵 姝,等. 基于自相关系数和 PseAAC 的蛋白质结构类预测. 计算机科学与探索, 2014, 8(1): 103-110.
(ZHANG Y P, ZHA Y L, ZHAO S, *et al*. Protein Structure Class Prediction Based on Autocorrelation Coefficient and PseAAC. Journal of Frontiers of Computer Science and Technology, 2014, 8(1): 103-110.)

[3] 李玉岗,张 法,刘志勇. 结合位点进化距离与支持向量机的蛋白质分类方法. 计算机学报, 2008, 31(1): 43-50.
(LI Y G, ZHANG F, LIU Z Y. Combining Position-Specific-Value Method and SVM for Remote Protein Classification. Chinese Journal of Computers, 2008, 31(1): 43-50.)

[4] 韩 跃,冀俊忠,杨翠翠. 基于多标签传播机制的蛋白质相互作用网络功能模块检测. 模式识别与人工智能, 2016, 29(6): 548-557.
(HAN Y, JI J Z, YANG C C. Functional Module Detection Based on Multi-label Propagation Mechanism in Protein-Protein Interaction Networks. Pattern Recognition and Artificial Intelligence, 2016, 29(6): 548-557.)

[5] CHENG J L, TEGGE A N, BALDI P. Machine Learning Methods for Protein Structure Prediction. IEEE Reviews in Biomedical Engineering, 2008, 1: 41-49.

[6] KANNAN D, DIABAT A, ALREFAEI M, *et al*. A Carbon Footprint Based Reverse Logistics Network Design Model. Resources, Conservation and Recycling, 2012, 67: 75-79.

[7] HUA S J, SUN Z R. A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure Support Vector Machine Approach. Journal of Molecular Biology, 2001, 308(2): 397-407.

[8] LI Z, YU Y Z. Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks // Proc of the 25th International Joint Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2016: 2560-2567.

[9] WANG S, PENG J, MA J Z, *et al*. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Scientific Reports, 2016. DOI: 10.1038/srep18962.

[10] ROST B, SANDER C. Prediction of Protein Secondary Structure at Better Than 70% Accuracy. Journal of Molecular Biology, 1993, 232(2): 584-599.

[11] ZHOU J, TROYANSKAYA O G. Deep Supervised and Convolutional Generative Stochastic Network for Protein Secondary Structure Prediction // Proc of the 31st International Conference on Machine Learning. New York, USA: ACM, 2014: 745-753.

[12] JONES D T. Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. Journal of Molecular Biology, 1999, 292(2): 195-202.

[13] QIAN N, SEJNOWSKI T J. Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. Journal of Molecular Biology, 1988, 202(4): 865-884.

[14] POLLASTRI G, PRZYBYLSKI D, ROST B, *et al*. Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles. Proteins, 2002, 47(2): 228-235.

[15] KABSCH W, SANDER C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. Biopolymers, 1983, 22(12): 2577-2637.

[16] WANG Z Y, ZHAO F, PENG J, *et al*. Protein 8-Class Secondary Structure Prediction Using Conditional Neural Fields // Proc of the IEEE International Conference on Bioinformatics and Biomedicine. Washington, USA: IEEE, 2010: 109-114.

[17] SØNDERBY S K, WINTHER O. Protein Secondary Structure Prediction with Long Short Term Memory Networks[C/OL]. [2017-11-25]. https://arxiv.org/pdf/1412.7828v2.pdf.

[18] BUSIA A, COLLINS J, JAITLY N. Protein Secondary Structure Prediction Using Deep Multi-scale Convolutional Neural Networks and Next-Step Conditioning[C/OL]. [2017-11-25]. https://arxiv.org/pdf/1611.01503v1.pdf.

[19] 吕永标,赵建伟,曹飞龙. 基于复合卷积神经网络的图像去噪算法. 模式识别与人工智能, 2017, 30(2): 97-105.
(LÜ Y B, ZHAO J W, CAO F L. Image Denoising Algorithm Based on Composite Convolutional Neural Network. Pattern Recognition and Artificial Intelligence, 2017, 30(2): 97-105.)

[20] BENGIO Y. Deep Learning of Representations: Looking Forward [C/OL]. [2017-11-25]. https://arxiv.org/pdf/1305.0445.pdf.

[21] 阮晓钢,孙海军. 编码方式对蛋白质二级结构预测精度的影响. 北京工业大学学报, 2005, 31(3): 229-235.
(RUAN X G, SUN H J. Research on Encode Influencing Protein Secondary Structure Prediction. Journal of Beijing University of Technology, 2005, 31(3): 229-235.)

[22] ALTSCHUL S F, MADDEN T L, SCHÄFFER A A, *et al*. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Research, 1997, 25(17): 3389-3402.

[23] BUSIA A, JAITLY N. Next-Step Conditioned Deep Convolutional Neural Networks Improve Protein Secondary Structure Prediction [C/OL]. [2017-11-25]. https://arxiv.org/pdf/1702.03865.pdf.

[24] ASGARI E, MOFRAD M R K. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PlOS One, 2015. DOI: 10.1371/journal.pone.0141287.

[25] KINGMA D P, BA J. ADAM: A Method for Stochastic Optimization[C/OL]. [2017-11-25]. https://arxiv.org/pdf/1412.6980.pdf.

## 作者简介

**郭延哺**,硕士研究生,主要研究方向为深度学习、生物信息学. E-mail: guoyb213321@163.com.
(**GUO Yanbu**, master student. His research interests include deep learning and bioinformatics.)

**李维华**(通讯作者),博士,副教授,主要研究方向为数据挖掘、机器学习. E-mail: lywey@163.com.
(**LI Weihua**(Corresponding author), Ph. D., associate professor. Her research interests include data mining and machine learning.)

**王兵益**,博士,副研究员,主要研究方向为植物分子生物学. E-mail: wbykm@aliyun.com.
(**WANG Bingyi**, Ph. D., associate researcher. His research interests include plant molecular biology.)

**金 宸**,硕士研究生,主要研究方向为自然语言处理、机器学习. E-mail: chenjin0721@gmail.com.
(**JIN Chen**, master student. His research interests include natural language processing and machine learning.)