计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201100117

基于序列特征融合的蛋白质可溶性预测

牛富生　郭延哺　李维华　刘文洋
云南大学信息学院　昆明 650500
(17839164754@163.com)

摘　要　蛋白质可溶性在药物设计的研究中起着重要的作用,传统生物实验测试蛋白质可溶性费时费力,因此基于计算方法对可溶性进行预测成为一个重要的研究方向。针对传统可溶性预测模型不能充分表示蛋白质特征的问题,文中设计了一种基于多种蛋白质序列信息的神经网络模型 PSPNet,并应用到蛋白质可溶性预测中。该模型首先使用氨基酸残基序列嵌入信息和氨基酸序列进化信息表示蛋白质序列;然后采用卷积神经网络提取氨基酸序列嵌入特征的局部关键信息;其次利用双向 LSTM 网络提取蛋白质序列远程依赖特征;最后利用注意力机制将该特征与氨基酸进化信息融合,并将包含了多种序列信息的融合特征用于蛋白质可溶性预测。实验结果表明,相比基准方法,该模型提高了蛋白质可溶性预测的精度,并具有良好的可扩展性。

关键词:蛋白质可溶性;多特征融合;深度学习;注意力机制
中图法分类号　TP391

Protein Solubility Prediction Based on Sequence Feature Fusion

NIU Fu-sheng,GUO Yan-bu,LI Wei-hua and LIU Wen-yang
School of Information Science and Engineering,Yunnan University,Kunming 650500,China

Abstract　Protein solubility plays an important role in the research of drug design. Traditional biological experiments of detecting protein solubility are time-consuming and laborious. Identifying protein solubility based on computational methods has become an important research hot spot in bioinformatics. Aiming at the problem of insufficient representation of protein features by traditional solubility prediction models,this paper designs a neural network model PSPNet based on protein sequence information and applies it to protein solubility prediction. PSPNet uses amino acid residue sequence embedding information and amino acid sequence evolution information to represent protein sequences. Then convolutional neural network is used to extract the local key information of amino acid sequence embedding features. Secondly,bidirectional LSTM network is used to extract the features of remote dependencies of protein sequences. Finally,the attention mechanism is used to fuse this feature and amino acid evolution information,and the fusion feature containing multiple sequence information is used in protein solubility prediction. The experimental results show that PASNet obtains the remarkable performance of protein solubility prediction compared with the benchmark methods and also has a good scalability.

Keywords　Protein solubility,Multi-feature fusion,Deep learning,Attention mechanism

1　引言

蛋白质是一切生命活动中不可或缺的重要物质,蛋白质的结构、性质对功能有着决定性的作用,可溶性是蛋白质的重要特征之一。蛋白质可溶性是指蛋白质溶质溶解于溶剂(通常指的是水)的能力[1]。在人体疾病治疗过程中,蛋白质类药物只有溶解才能发挥重要的作用,因此精确地预测蛋白质可溶性既可以增进人们对疾病发病机制的理解,也可以加速蛋白质类药物的研发和设计进程[2-3]。随着高通量技术的发展,大量的蛋白质序列和结构数据生成[4],对这些蛋白质的可溶

到稿日期:2020-11-16　返修日期:2021-06-29
基金项目:国家自然科学基金项目(32060151);云南省教育厅科学研究基金(2019J0006);云南省创新团队项目(2018HC019);云南大学研究生科研创新基金项目(2020Z73)
This work was supported by the National Natural Science Foundation of China(32060151),Scientific Research Fundation of the Education Department of Yunnan Province,China(2019J0006),Innovative Research Team of Yunnan Province,China(2018HC019) and Yunnan University of Postgraduate Research and Innovation Foundation Project,China(2020Z73).
通信作者:李维华(lywey@163.com)

性进行测试成为迫切需求。传统的生物测定方法主要通过提取出蛋白质实体进行可溶性测试。由于蛋白质的提取过程复杂,这些方法通常费时费力,已无法满足日益增长的蛋白质可溶性分析的需求。因此,使用计算方法对蛋白质可溶性进行预测成为蛋白质功能研究的热点问题。

蛋白质可溶性预测问题的研究由来已久,已有很多的计算方法用于可溶性预测[5-7]。这些计算方法通常关注蛋白质的序列特征表示和可溶性计算方法。蛋白质的序列特征表示就是寻找一种数学模型来尽可能全面地表示出蛋白质序列中隐含的模式和结构信息[8]。已有研究[9]表明,蛋白质的溶解度与序列组成方式和氨基酸的理化特性密切相关。因此,如何充分利用蛋白质的氨基酸组成及其理化性质进行蛋白质特征的表示,已成为蛋白质可溶性预测的重要问题。

蛋白质序列的特征表示通常分为基于氨基酸组成的方法、基于氨基酸理化性质的方法和基于多种特征表示融合的方法。基于氨基酸组成的方法通常统计不同氨基酸或氨基酸组合在蛋白质序列中出现的频率[10]。这种表示方法计算简单且使用方便,但只关注了氨基酸的频率信息,往往忽略了序列中包含的上下文模式和特征。因此,找到一种能够充分表示氨基酸序列特征的方法成为亟待解决的问题。近年来,自然语言处理(Natural Language Processing,NLP)的快速发展,促进了生物序列特征表示的研究。例如,Chen 等[11]采用 skip-gram 模型实现对氨基酸的嵌入表示,在蛋白质相互作用的预测中获得了良好的表现。Guo 等[12]基于 Glove 模型对 DNA 进行嵌入表示,并完成染色质可达性的预测。这些成功的应用表明,使用 NLP 技术对蛋白质序列进行处理,能够获得序列中蕴含的上下文模式和特征,从而更充分地表达出蛋白质序列的特征。

基于氨基酸理化性质的特征表示方法[13]使用的模式和特征信息单一,并且不同氨基酸的理化性质可能相同,所以往往作为其他特征表示方法的补充。基于多种特征表示融合的方法[14-16]一般包含蛋白质的多种信息,如氨基酸组成、蛋白质进化信息和理化性质等。通过多种特征来表示蛋白质的结构信息往往能够在应用中取得更好的效果。例如,基于位置特异性矩阵(Position Specific Scoring Matrices,PSSM)的 RPM_PSSM(PSSM Based Residue Probing Method)[15],它不仅包含了蛋白质进化信息,还包含了理化特征[16]。因此,基于 NLP 技术对蛋白质序列进行特征表示以获取序列中上下文模式和特征,与 RPM_PSSM 向量采用可靠的方式融合成包含多种序列信息的特征,并用于提升蛋白质可溶性预测的性能是可行的。

蛋白质可溶性预测的计算方法通常分为基于传统机器学习的预测方法和基于深度学习的预测方法。基于传统机器学习的预测方法在训练模型时只需要少量数据集,但其却依赖于特征工程。Huang 等[17]使用蛋白质中氨基酸二肽统计频率,采用简单的计分卡模型来预测蛋白质可溶性。Magnan 等[18]通过统计蛋白质序列的 k-mer 和还原氨基酸频率作为特征输入,使用支持向量机[19](Support Vector Machines,SVM)对蛋白质可溶性进行预测。Smialowski 等[20]对蛋白质的氨基酸二肽和三肽的出现频率进行统计并作为输入,使用一个带有径向基函数核的 SVM 和朴素贝叶斯二级分类器进行可溶性预测。Rawi 等[21]使用了氨基酸二肽和三肽的统计频率作为输入,使用梯度增强机器[22]实现了蛋白质可溶性的预测。这些研究在蛋白质可溶性的预测上都获得了一定的成就,但基于传统机器学习的预测方法需要大量的特征选择工作,且无法充分发挥日益增多的海量蛋白质数据的优势。因此,能够自动提取特征并充分发挥海量数据优势的方法成为蛋白质可溶性预测的迫切需求。

近年来,由于可以挖掘海量数据特征和自主特征学习的优势,深度学习方法在 NLP 和计算机视觉等领域获得巨大成功。Shi 等[23]通过卷积神经网络(Convolutional Neural Networks,CNN)来捕捉图像中的局部关键信息,进而实现对汽车车型进行识别。Zeng 等[24]使用双向长短期记忆(Long Short-Term Memory,LSTM)网络挖掘评论中的远程依赖关系,并完成文本的情感分析。深度学习在这些领域取得的成就,为改进蛋白质可溶性的预测性能提供了新的思路。Khurana 等[25]首次将深度学习技术应用在可溶性预测上,并提出了一种蛋白质可溶性预测模型 DeepSol,该模型使用卷积神经网络提取氨基酸序列中的局部关键特征,随后与蛋白质的一些理化性质进行融合,从而改进蛋白质可溶性的预测。然而,DeepSol 模型中蛋白质的特征表示依然使用传统的氨基酸组成方法,因此并不能充分表示序列的上下文信息;其次,该模型仅使用了 CNN。虽然 CNN 在寻找蛋白质序列的局部关键信息中表现良好,但却不能挖掘出序列中前后文之间的远程依赖关系。因此,本文融合了 CNN 和 LSTM 网络来构建模型,以挖掘蛋白质序列的深层特征并对蛋白质可溶性进行预测。

基于上述分析,本文提出了一种新的蛋白质可溶性预测模型(Protein Solubility Prediction Network,PSPNet),该模型基于氨基酸残基的 Word2Vec 嵌入和 RPM_PSSM 向量进行融合并用于可溶性预测。具体地说,PSPNet 首先通过 CNN 提取氨基酸残基间的局部关键特征;然后,使用双向 LSTM 网络提取氨基酸间的远程依赖特征;最后,使用注意力层将蛋白质的局部关键特征、远程依赖特征、包含蛋白质进化信息和理化性质的 RPM_PSSM 向量进行融合,并用于蛋白质可溶性预测。实验表明,PSPNet 在蛋白质可溶性的预测中表现优异。

2 可溶性预测模型

为了预测蛋白质的可溶性,本文基于蛋白质序列的多种特征构建蛋白质可溶性预测模型 PSPNet。PSPNet 模型由特征表示模块、特征学习模块、特征融合模块和输出模块组成,具体如图 1 所示。

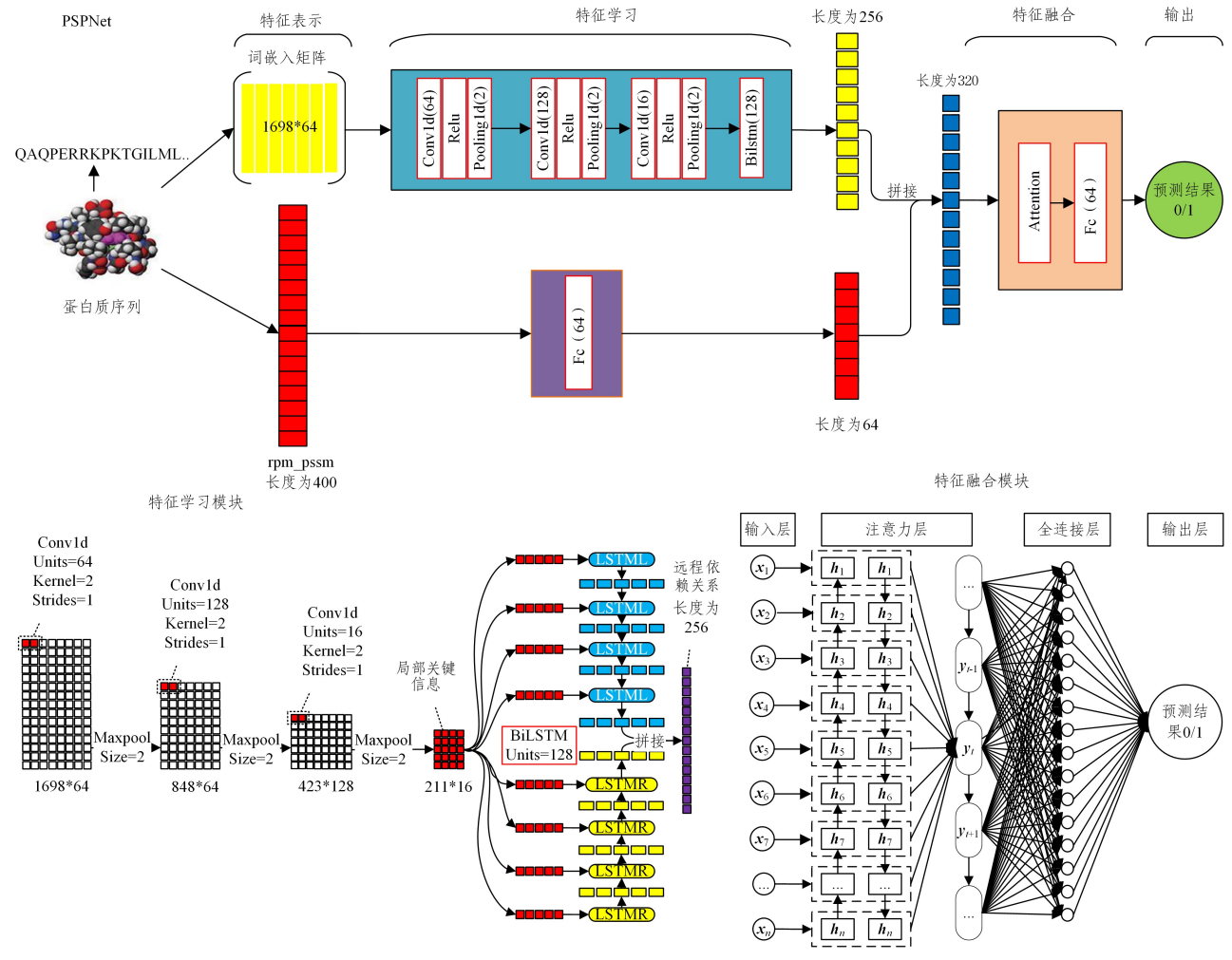

图 1　蛋白质可溶性的深度预测模型

Fig. 1　Deep prediction model of protein solubility

2.1　基本框架

蛋白质序列经过不同的特征表示，得到包含蛋白质上下文模式信息的特征 $\boldsymbol{F}_v$ 和包含蛋白质进化信息以及理化性质的特征 $\boldsymbol{F}_s$。随后，通过卷积神经网络和双向 LSTM 层学习特征 $\boldsymbol{F}_v$ 的局部关键信息和远程依赖关系，同时使用全连接层进行维度压缩。最后，使用一个注意力层和全连接层将二者进行融合并再次压缩维度。特征的提取和融合过程如下：

$$\boldsymbol{F}=g_{fc}(g_{att}([g_{blstm}(g_{cnn}(\boldsymbol{F}_v)),g_{fc}(\boldsymbol{F}_s)])) \tag{1}$$

其中，$\boldsymbol{F}_v$ 和 $\boldsymbol{F}_s$ 分别是基于氨基酸残基的 Word2Vec 嵌入和 RPM_PSSM 向量的蛋白质特征；$\boldsymbol{F}$ 是融合后的特征；g_{fc} 使用全连接神经网络对特征进行维度压缩，在维度压缩的同时最大限度地保存特征信息；g_{cnn} 使用卷积神经网络提取局部关键信息；g_{blstm} 使用双向 LSTM 网络层挖掘远程依赖关系；g_{att} 使用注意力层对两种特征进行融合。

在得到融合特征 $\boldsymbol{F}$ 后，使用一个 sigmoid 层对蛋白质可溶性进行预测。预测层通过一个二元交叉熵损失函数和 Adam 优化器来调整个网络模型的参数。预测过程及损失函数的计算如下：

$$p(y_i)=\mathrm{sigmoid}(\boldsymbol{F}) \tag{2}$$

$$sigmoid=\frac{1}{1+\mathrm{e}^{-x}} \tag{3}$$

$$loss=-\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_i\cdot\log(p(\boldsymbol{y}_i))+(1-\boldsymbol{y}_i)\cdot\log(1-p(\boldsymbol{y}_i)) \tag{4}$$

其中，sigmoid 为激活函数，$p(\boldsymbol{y}_i)$ 是可溶性的预测结果，$loss$ 是样本的损失累加均值。

2.2　特征表示

本文分别基于氨基酸残基的 Word2Vec 嵌入向量和 RPM_PSSM 向量来表示蛋白质序列的特征。

首先，借鉴 NLP 任务中使用 Word2Vec 学习词向量的方法，本文将蛋白质序列看作一个句子，并把其中每个氨基酸残基当作一个单词。将每个蛋白质序列分割成具有 n 个氨基酸残基的子序列，并组成一个语料库。随后通过 Word2Vec 的 Skip-gram 模型[26]得到一个预先训练好的嵌入矩阵。嵌入模块提供一系列氨基酸残基向量表示的索引，将每个氨基酸残基映射到一个稠密向量 $\boldsymbol{v}_i$：

$$\boldsymbol{v}_i=\boldsymbol{M}\cdot\boldsymbol{a}_i \tag{5}$$

其中，$\boldsymbol{M}$ 为氨基酸残基的嵌入矩阵，$\boldsymbol{a}_i$ 表示氨基酸残基组成的词汇表中每个氨基酸残基的索引。使用嵌入矩阵 $\boldsymbol{M}$ 对蛋白质序列进行嵌入表示，最后得到一个蛋白质序列的嵌入特征表示 $\boldsymbol{F}_v=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\boldsymbol{v}_3,\cdots,\boldsymbol{v}_n\}$，作为特征学习模块的输入。

其次，RPM_PSSM 向量是基于 PSSM 的位置特异性矩阵，通过 PSSM 分块特征提取并与氨基酸理化性质进行融合，

得到长度为 200 的向量。最后，每个蛋白质被表示为 $\boldsymbol{F}_s=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3,\cdots,\boldsymbol{s}_n\}$，其中 $\boldsymbol{s}_i\in\boldsymbol{s}^m$ 表示蛋白质序列中第 i 个维度的特征表示。

2.3 特征学习

针对两种不同的蛋白质特征表示，本文采用了不同的学习策略提取特征，针对氨基酸残基的 Word2Vec 嵌入特征，采用 CNN 和双向 LSTM 单元融合的网络结构来挖掘蛋白质序列中隐含的局部关键信息和远程依赖关系，针对 RPM_PSSM 向量，使用了一个全连接层来对特征进行维度压缩。

2.3.1 局部关键信息的提取

蛋白质可溶性与其氨基酸组成有着密切的关联，其中局部的组成信息对预测有很大的影响[9]。因此，挖掘这些局部关键信息对蛋白质可溶性的预测有着很重要的意义。CNN[27]在图像处理领域可以有效地捕捉局部关键特征，因此本文使用 CNN 来提取蛋白质序列中的局部关键信息：

$$\boldsymbol{F}_{conv}=\sum_{i}^{p*q}\boldsymbol{w}_i*\boldsymbol{x}_i,\boldsymbol{x}_i\in\boldsymbol{v}_i,\boldsymbol{v}_i\in\boldsymbol{F}_v \tag{6}$$

$$\boldsymbol{F}_{act}=h(\sum_{i}^{p*q}\boldsymbol{w}_i*\boldsymbol{x}_i+\boldsymbol{b}_i),\boldsymbol{x}_i\in\boldsymbol{v}_i,\boldsymbol{v}_i\in\boldsymbol{F}_{conv} \tag{7}$$

$$\boldsymbol{F}_c=\max(\sum_{i}\sum_{i+1}f(i,j)),f(i,j)\in\boldsymbol{F}_{act} \tag{8}$$

其中，$\boldsymbol{F}_{conv}$，$\boldsymbol{F}_{act}$ 和 $\boldsymbol{F}_c$ 分别是 Word2Vec 嵌入特征经过卷积、激活和最大池化后得到的特征；$h()$ 是激活函数，本文选用线性整流函数[28]（Rectified Linear Unit，Relu）作为激活函数；max()表示最大池化。

初始特征经过网络层后规模被缩小，同时局部关键信息被提取出来。本文在局部关键特征提取模块中设置了 3 个卷积网络层。经过 3 次卷积池化处理，从 Word2Vec 嵌入特征得到包含蛋白质局部关键信息的特征 $\boldsymbol{F}_c$。

2.3.2 远程依赖关系挖掘

除了局部关键信息外，蛋白质中氨基酸残基之间的远程依赖关系也有可能对蛋白质可溶性预测模型产生影响。因此，本文使用双向 LSTM[24]进一步处理获得的特征。双向 LSTM 通过一个正向层和一个反向层（两者均由带有时间顺序的记忆单元组成）捕捉蛋白质序列中氨基酸残基的远程依赖关系。其中，每个记忆单元包含细胞状态、遗忘门、输入门和输出门[29]。细胞状态用来保存 t 时刻的记忆信息，遗忘门用来控制遗忘上一层细胞状态的内容，输入门决定记忆单元的输入信息，输出门决定最后输出的内容。每个单元的信息更新步骤如下：

$$\boldsymbol{f}_t=\gamma((\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t])+\boldsymbol{b}_f) \tag{9}$$

$$\boldsymbol{i}_t=\gamma((\boldsymbol{W}_i\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i)) \tag{10}$$

$$\boldsymbol{C}_t'=\vartheta(\boldsymbol{W}_C\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_c) \tag{11}$$

$$\boldsymbol{C}_t=\boldsymbol{f}_t*\boldsymbol{C}_{t-1}+\boldsymbol{i}_t*\boldsymbol{C}_t' \tag{12}$$

$$\boldsymbol{O}_t=\gamma(\boldsymbol{W}_o\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_o) \tag{13}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t*\vartheta(\boldsymbol{C}_t) \tag{14}$$

其中，$\boldsymbol{f}_t$，$\boldsymbol{i}_t$，$\boldsymbol{O}_t$ 分别表示输入门、遗忘门和输出门的操作状态，γ 代表激活函数 sigmoid，ϑ 代表激活函数 tanh，$\boldsymbol{C}_t'$ 表示输出层的待更新数据，$\boldsymbol{C}_t$ 表示更新后的单元状态，$\boldsymbol{h}_t$ 表示单元的最终输出。

为了同时得到蛋白质序列中氨基酸由前到后和由后向前的远程依赖关系，双向 LSTM 的正向层从 1 时刻到 t 时刻正向计算一遍，得到并保存每个时刻前一个隐含层的输出，反向层与正向相反从 t 时刻到 1 时刻计算并保存输出，最后结合各个时刻正反向结果得到最终输出：

$$\boldsymbol{h}_t=f(\boldsymbol{w}_1\boldsymbol{x}_t+\boldsymbol{w}_2\boldsymbol{h}_{t-1}) \tag{15}$$

$$\boldsymbol{h}_t'=f'(\boldsymbol{w}_3\boldsymbol{x}_t+\boldsymbol{w}_4\boldsymbol{h}_{t+1}') \tag{16}$$

$$\boldsymbol{F}_m=g(\boldsymbol{w}_5\boldsymbol{h}_t+\boldsymbol{w}_6\boldsymbol{h}_t') \tag{17}$$

其中，$\boldsymbol{h}_t$ 表示正向层的输出，包含了蛋白质序列的正向远程依赖关系；$\boldsymbol{h}_t'$ 是反向层的输出，包含了蛋白质序列的反向远程依赖关系；$f()$ 和 $f'()$ 是对正向和反向输出进行计算的函数；$g()$ 是对两个方向的输出组合求和的函数，最终获得长度为 256 的特征向量 $\boldsymbol{F}_m$，该特征包含了蛋白质序列的局部关键信息和远程依赖关系。

2.3.3 特征维度压缩

RPM_PSSM 向量特征的维度会影响模型的预测效果，一方面特征维度过大会导致与特征 $\boldsymbol{F}_m$ 融合时降低 $\boldsymbol{F}_m$ 的重要性；另一方面得到的融合特征维度过大，会使预测模块变得更加复杂。这些都会使整个模型在训练时不易收敛，影响最终预测效果。已有研究表明[30]，全连接网络（Fully Connected Layers，FC）对特征维度的压缩有很好的效果，在实现维度变换的同时能最大限度地保存原有的信息。因此，本文使用一个包含 64 个神经元的全连接层对 RPM_PSSM 向量特征 $\boldsymbol{F}_p$ 进行维度压缩：

$$\boldsymbol{F}_z=Relu(\boldsymbol{w}*\boldsymbol{F}_p+\boldsymbol{b}) \tag{18}$$

其中，$\boldsymbol{w}$ 为该网络层的权重，$\boldsymbol{b}$ 为该层的偏置，Relu 是激活函数。$\boldsymbol{F}_p$ 经过全连接层的压缩后变成长度为 64 的特征向量 $\boldsymbol{F}_z$。$\boldsymbol{F}_z$ 既实现了对特征 $\boldsymbol{F}_p$ 的维度压缩，又最大限度地保存了其包含的进化信息和理化性质。

2.4 特征融合

已有研究表明[31-33]，使用多种特征融合对生物信息的特征表示有更好的效果。本文使用不同的策略对两种蛋白质特征表示进行特征提取，将提取的特征合并得到包含蛋白质序列多种信息的特征 $\boldsymbol{F}_{com}=[\boldsymbol{F}_m,\boldsymbol{F}_z]$。其中 $\boldsymbol{F}_m$ 是长度为 256 的特征向量，包含了蛋白质序列中局部关键信息和远程依赖关系。$\boldsymbol{F}_z$ 是长度为 64 的特征向量，包含了蛋白质的进化信息和理化性质。$\boldsymbol{F}_{com}$ 是长度为 320 的特征向量，由 $\boldsymbol{F}_m$ 和 $\boldsymbol{F}_z$ 进行拼接得到。

经过两个特征拼接而来的特征 $\boldsymbol{F}_{com}$ 虽然在结构上包含了两种不同的蛋白质信息，但是它们彼此独立，在可溶性预测上不能充分发挥出这些特征的作用。为了实现对这些特征的有效融合，本文使用注意力层来对特征 $\boldsymbol{F}_{com}$ 进行融合。注意力机制允许模型动态地关注输入的某些部分，有助于提升模型的表现[34]。简单地说，它会为不同特征赋予不同的权重：

$$\boldsymbol{F}_{att}=\sum_{i=1}^{n}\boldsymbol{a}_i*\boldsymbol{x}_i\boldsymbol{a}_i\in\boldsymbol{a}^*,\boldsymbol{x}_i\in\boldsymbol{F}_{com} \tag{19}$$

$$\boldsymbol{F}=Relu(\boldsymbol{w}*\boldsymbol{F}_{att}+\boldsymbol{b}) \tag{20}$$

其中，$\boldsymbol{a}^*$ 是不同特征的权重空间，$\boldsymbol{F}_{att}$ 是通过注意力层融合后的特征。随后通过一个单元数为 64 的全连接层对该特征实现进一步降维，$\boldsymbol{w}$ 和 $\boldsymbol{b}$ 分别是全连接层的权重和偏置值。经过融合和压缩得到最终用来预测蛋白质可溶性的融合特征 $\boldsymbol{F}$。融合特征 $\boldsymbol{F}$ 是一个长度为 64 的特征向量，其中包含了

蛋白质序列的局部关键信息、远程依赖关系以及进化信息和理化性质。

2.5 输出

对本文提出的蛋白质可溶性预测模型提取出融合特征 $\boldsymbol{F}$，使用 sigmoid 层来实现对可溶性的预测，输出预测结果：

$$p(\boldsymbol{y})=sigmoid(\boldsymbol{w}*\boldsymbol{F}+\boldsymbol{b})=\begin{cases}0, & \text{预测为可溶}\\ 1, & \text{预测为不溶}\end{cases} \tag{21}$$

其中，$P(\boldsymbol{y})$为预测结果，$\boldsymbol{w}$ 和 $\boldsymbol{b}$ 是预测层的权重和偏置，使用 sigmoid 函数将输入映射成 0 或 1 的值。

3 实验及结果分析

3.1 实验环境和数据

本文实验环境中的处理器为 Intel i7-7700 CPU 3.60 GHz，图形加速卡为 NVIDIA GeForce GTX 1060 6 GB，内存为 8GB，操作系统为 Ubuntu 18.04 LTS(64 bit)。本文实验采用深度学习框架 TensorFlow2.3.1 构建神经网络模型。

为了训练本文提出的 PSPNet 模型，本文使用 Rawi 等[21]提供的数据集。该数据集包含了 69 420 个训练集和 2001 个测试集。其中，训练集是由 28972 条可溶的蛋白质序列以及 40448 条不可溶的蛋白质序列组成。在进行特征表示时，去掉长度少于 20 或大于 2000 的序列，最终获得 26018 条可溶的蛋白质序列和 36 380 条不可溶的蛋白质序列，共 62398 条。

为了验证模型预测的性能，本文使用 Chang 等[35]提供的数据集作为测试集，该数据集包含了 1 000 条可溶的蛋白质序列和 1001 条不可溶的蛋白质序列。在实验中，去掉长度少于 20 或大于 2000 的序列，最终获得 997 条可溶的蛋白质序列和 998 条不可溶的蛋白质序列。

3.2 基准模型

为了验证本文提出的蛋白质可溶性预测的深度模型相较于传统模型的优势，本文采用 SCM，SOLpro，PROSO II，PaRSnIP 和 DeepSol 作为基准模型，具体介绍如下：

(1)SCM[17]：该模型使用蛋白质序列的二肽组成作为特征，采用一种计分卡模型来对蛋白质可溶性进行预测。

(2)SOLpro[18]：该模型使用蛋白质序列的 k-mer 组成和还原氨基酸的频率作为特征输入，使用支持向量机技术对蛋白质可溶性进行预测。

(3)PROSO II[20]：该模型使用了氨基酸 k-mer 组成作为模型的输入，采用二级 logistic 回归分类器对蛋白质可溶性进行预测。

(4)PaRSnIP[21]：该模型以蛋白质的二肽和三肽组成作为特征表示，构造了一个渐变增强机器模型对蛋白质可溶性进行预测。

(5)DeepSol[25]：该模型使用氨基酸 k-mer 组成和理化性质融合作为输入，同时首次将深度学习技术应用在蛋白质可溶性预测上。

3.3 评价指标

本文采用准确率(ACC)、马修斯相关系数(Matthews Correlation Coefficient，MCC)、选择性(SEL)和敏感性(SEN)作为评价指标。

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \tag{22}$$

$$MCC=\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{23}$$

$$SEL(s)=\frac{TP}{TP+FN}\ SEL(i)=\frac{TN}{TN+FP} \tag{24}$$

$$SEN(s)=\frac{TP}{TP+FP}\ SEN(i)=\frac{TN}{TN+FN} \tag{25}$$

其中，TP 和 TN 分别指正确预测的可溶和不可溶的蛋白质数量，FP 和 FN 分别指被错误预测的可溶和不可溶的蛋白质数量；ACC 衡量模型对样本的预测能力，MCC 评价二分类模型对样本的均衡能力，选择性和敏感性分别是正反例的查准率和查全率。其中，$SEL(s)$表示可溶蛋白的选择性，$SEL(i)$表示不可溶蛋白的选择性，$SEN(s)$表示可溶蛋白的敏感性，$SEN(i)$表示不可溶蛋白的敏感性。

3.4 实验结果

本文基于 ACC，MCC，SEL 和 SEN 评价指标，将所提模型与基准模型进行对比，结果如表 1 所列。

表 1 PSPNet 同基准模型的对比结果

Table 1 Results of PSPNet compared with benchmark models

模型	ACC	MCC	$SEL(s)$	$SEL(i)$	$SEN(s)$	$SEN(i)$
PSPNet	**0.79**	**0.58**	0.80	**0.78**	**0.77**	0.81
DeepSol	0.77	0.55	**0.84**	0.72	0.65	**0.88**
PaRSnIP	0.74	0.48	0.76	0.72	0.70	0.78
PROSO II	0.64	0.34	0.67	0.68	0.69	0.66
SOLpro	0.60	0.20	0.62	0.58	0.51	0.69
SCM	0.60	0.21	0.65	0.57	0.42	0.77

从整体结果上看，本文模型已经基本超越了当前最好的蛋白质可溶性预测方法 DeepSol，这表明深度学习技术在蛋白质可溶性领域有更好的效果，同时也验证了 PSPNet 使用多特征的融合策略有助于改进可溶性预测的性能。从准确率上看，PSPNet 达到了 0.79，超过了当前最好的 DeepSol 模型 2%，超过了 PaRSnIP 模型 5%。从 MCC 来看，PSPNet 达到了 0.58，超过了当前最好的 DeepSol 模型 3%，超过了 PaRSnIP 模型 10%。ACC 和 MCC 指标的全面领先表明了本文提出的方法对蛋白质可溶性的整体预测结果更加精确且均衡。而在 SEL 和 SEN 两个指标上，无论是在可溶还是不可溶蛋白质的表现上，PSPNet 方法都达到最佳效果或者稍逊于表现最好的方法。以 $SEL(s)$指标为例，本文模型达到了 0.80，仅次于 DeepSol，比 PaRSnIP 方法提升了 4%；在不可溶蛋白上，PSPNet 方法比 DeepSol 和 PaRSnIP 提升了 6%。对这些指标的分析表明，本文基于神经网络对蛋白质多种特征表示和融合，并使用融合特征进行蛋白质可溶性预测的方法能够挖掘出蛋白质序列中的重要信息，从而提升对蛋白质可溶性的预测性能。

3.5 模型分析

为了验证 PSPNet 模型中不同模块的作用，我们设计了 4 个不同的对比实验，以分析不同模块对模型性能的影响。

实验 1 为了验证 Word2Vec 嵌入特征和 RPM_PSSM 向量特征融合对蛋白质可溶性的预测更具优势，本文设计了

分别使用两种特征与使用融合特征的对比实验。

从表 2 中可以看到，在单独使 Word2Vec 嵌入或 RPM_PSSM 向量中的一种特征表示方法时，模型所取得的预测准确率均远远低于融合特征的表现。这说明了本文所使用的特征融合策略对蛋白质可溶性预测有很大的帮助，优于使用单一特征表示的方法。

表 2 使用不同特征时的模型效果

Table 2 Model effect when using different features

特征	*ACC*	*MCC*	*SEL(s)*	*SEL(i)*	*SEN(s)*	*SEN(i)*
融合特征	**0.79**	**0.58**	**0.80**	**0.78**	**0.77**	**0.81**
Word2Vec	0.74	0.49	0.75	0.74	0.74	0.75
RPM_PSSM	0.65	0.31	0.67	0.64	0.61	0.70

实验 2 为了验证 CNN 和双向 LSTM 学习 Word2Vec 嵌入特征的局部信息和远程依赖关系对模型效果的影响，本文设计了分别使用 CNN、双向 LSTM 和两者结合使用的对比实验。

表 3 的结果表明，在对 Word2Vec 嵌入特征应用不同的特征学习策略时，无论是单独使用 CNN 还是双向 LSTM 网络，在各项指标上均低于这两种网络的组合使用。这表明融合 CNN 和双向 LSTM 层在对蛋白质序列的 Word2Vec 嵌入特征表示进行特征学习时，通过捕捉序列中的局部关键信息和远程依赖关系，能够更完整地表示出蛋白质序列的结构信息，从而提升模型的预测性能。

表 3 不同学习策略对模型效果的影响

Table 3 Influence of different learning strategies on model effect

学习策略	*ACC*	*MCC*	*SEL(s)*	*SEL(i)*	*SEN(s)*	*SEN(i)*
CNN	0.73	0.46	0.74	0.72	0.70	0.75
BLSTM	0.74	0.48	0.75	0.72	0.71	0.77
CNN+BLSTM	**0.79**	**0.58**	**0.80**	**0.78**	**0.77**	**0.81**

实验 3 为了验证 RPM_PSSM 向量的特征压缩策略的作用以及选择合适的压缩维度，本文设计了对该特征压缩成不同维度后，分别输入模型进行可溶性预测的对比实验。

表 4 的结果表明，对 RPM_PSSM 向量进行维度压缩对本文模型的预测结果有着积极的影响。在特征的压缩维度变小的过程中，模型的效果在随之提升，直至维度压缩成 64，达到最佳效果。这表明了对 RPM_PSSM 向量进行适量的维度压缩，有助于提升模型的预测性能。

表 4 不同维度压缩对模型的影响

Table 4 Impact of different dimensional compression on model

压缩方式	*ACC*	*MCC*	*SEL(s)*	*SEL(i)*	*SEN(s)*	*SEN(i)*
未压缩	0.69	0.38	0.70	0.68	0.66	0.72
256	0.72	0.44	0.71	0.73	0.73	0.71
128	0.76	0.52	0.74	**0.78**	**0.80**	0.72
64	**0.79**	**0.58**	**0.80**	**0.78**	0.77	**0.81**
32	0.76	0.52	0.74	**0.78**	**0.80**	0.72

实验 4 为了验证使用注意力机制融合多特征的作用，本文对包含了蛋白质序列中局部关键信息、远程依赖关系的特征 $\boldsymbol{F}_m$ 和包含了进化信息、理化性质的 $\boldsymbol{F}_z$，分别使用带注意力机制融合和直接拼接融合的对比实验。

表 5 的结果表明，使用注意力机制对特征进行融合时，模型的预测结果显著提升。这表明本文使用注意力机制能够有效地将不同特征表示进行融合，进而提升模型对蛋白质可溶性预测的性能。

表 5 不同融合策略对模型效果的影响

Table 5 Influence of different fusion strategies on the model effect

融合策略	*ACC*	*MCC*	*SEL(s)*	*SEL(i)*	*SEN(s)*	*SEN(i)*
注意力	**0.79**	**0.58**	0.80	**0.78**	**0.77**	0.81
拼接	0.77	0.54	**0.81**	0.74	0.70	**0.84**

结束语 本文基于 Word2Vec 嵌入和 RPM_PSSM 向量，构建了深度神经网络模型，融合蛋白质的序列信息和结构及理化特征，并进行蛋白质的可溶性预测。实验结果表明，多种特征的融合能够更加充分地表示蛋白质序列的信息，同时也表明深度学习技术可以有效地挖掘出蛋白质中隐含的重要信息，进而提升模型的预测性能。

为了充分利用蛋白质序列中蕴含的信息，使用其他词嵌入模型对蛋白质进行特征表示，丰富蛋白质特征是一个改进的研究方向。另外，融合一些蛋白质的其他结构信息来表示蛋白质序列，或使用一些新的深度模型也许能为模型性能的提升带来机会。

参 考 文 献

[1] ZAYAS J F. Solubility of Proteins[M]. Springer Berlin Heidelberg, 1997.

[2] SMIALOWSKI P, MARTINGALIANO A J, MIKOLAJKA A, et al. Protein solubility[J]. Bioinformatics, 2007, 23(19): 2536-2542.

[3] FROKJAER S, OTZEN D. Protein drug stability: a formulation challenge[J]. Nature Reviews Drug Discovery, 2005, 4(4): 298.

[4] SUN X, LU Z H, XIE J M. Fundamentals of Bioinformatics [M]. Tsinghua University Press, 2005.

[5] WILKINSON D L, HARRISON R G. Predicting the solubility of recombinant proteins in Escherichia coli[J]. Bio/technology, 1991, 9(5): 443-448.

[6] SMIALOWSKI P, MARTIN-GALIANO A J, MIKOLAJKA A, et al. Protein solubility: sequence based prediction and experimental verification [J]. Bioinformatics, 2007, 23 (19): 2536-2542.

[7] AGOSTINI F, VENDRUSCOLO M, TARTAGLIA G G. Sequence-Based Prediction of Protein Solubility[J]. Journal of Molecular Biology, 2012, 421(2): 237-241.

[8] GUO Y B, LI W H, WANG B Y, et al. Protein secondary structure prediction based on convolutional long and short-term memory neural network[J]. Pattern Recognition and Artificial Intelligence, 2018, 31(180): 80-86.

[9] XU G H. Structure and function of protein molecules[J]. Bulletin of Biology, 2010, 45(3): 24-25.

[10] ROY S, MARTINEZ D, PLATERO H, et al. Exploiting Amino Acid Composition for Predicting Protein-Protein Interactions [J]. Plos One, 2009, 4(11): 7813-7826.

[11] CHEN M, JU J T, ZHOU G, et al. Multifaceted protein-protein interaction prediction based on Siamese residual RCNN [J]. Bioinformatics, 2019, 35(14): 305-314.

[12] GUO Y, ZHOU D, NIE R, et al. DeepANF: A deep attentive

neural framework with distributed representation for chromatin accessibility prediction[J]. Neurocomputing,2019,37(9):305-318.

[13] KAWASHIMA S. AAindex:amino acid index database[J]. Nucleic Acids Research,2008,28(1):374.

[14] SHEN H,CHOU K. PseAAC:A flexible web server for generating various kinds of protein pseudo amino acid composition[J]. Analytical Biochemistry,2008,373(2):386-388.

[15] WANG J,YANG B,REVOTE J,et al. POSSUM:a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles[J]. Bioinformatics,2017,33(17):2756-2758.

[16] JEONG J C,LIN X,CHEN X W. On Position-Specific Scoring Matrix for Protein Function Prediction[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics,2011,8(2):308-315.

[17] HUANG H,CHAROENKWAN P,KAO T,et al. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition[J]. BMC Bioinformatics,2012,13(17):1-14.

[18] MAGNAN C N,RANDALL A,BALDI P. SOLpro:accurate sequence-based prediction of protein solubility[J]. Bioinformatics,2009,25(17):2200-2207.

[19] VAPNIK V N. The Nature of Statistical Learning Theory[M]. Springer,1995.

[20] SMIALOWSKI P,DOOSE G,TORKLER P,et al. PROSO II-a new method for protein solubility prediction[J]. The FEBS journal,2012,279(12):2192-2200.

[21] RAWI R,MALL R,KUNJI K,et al. PaRSnIP:sequence-based protein solubility prediction using gradient boosting machine[J]. Bioinformatics,2018,34(7):1092-1098.

[22] FRIEDMAN J H. Greedy Function Approximation:A Gradient Boosting Machine[J]. Annals of Statistics,2001,29(5):1189-1232.

[23] SHI L,WANG Y M,CAO Y J,et al. Car model recognition based on deep convolutional neural networks[J]. Computer Science,2018,45(5):280-284.

[24] ZENG Z,LI L,CHEN J. Bidirectional deep LSTM for sentiment classification[J]. Computer Science,2018,4(5):213-217.

[25] KHURANA S,RAWI R,KUNJI K,et al. DeepSol:a deep learning framework for sequence-based protein solubility prediction[J]. Bioinformatics,2018,34(15):2605-2613.

[26] BOJANOWSKI P,GRAVE E,JOULIN A,et al. Enriching Word Vectors with Subword Information[J]. Transactions of the Association for Computational Linguistics,2017,5:135-146.

[27] ZHOU F Y,JIN L P,DONG J. A review of convolutional neural networks[J]. Chinese Journal of Computers,2017(6):1229-1251.

[28] JIANG A B,WANG W W. ReLU activation function optimization research[J]. Sensors and Microsystems,2018,37(312):56-58.

[29] HOCHREITER S,SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation,1997,9(8):1735-1780.

[30] ZHENG J H. Research on BP Neural Network Method for Image Data Compression[J]. Computer Simulation,2001(2):33-36.

[31] TIAN Q C,ZHANG R S. Overview of Biometric Recognition[J]. Computer Application Research,2009(12):4401-4406.

[32] WANG Y H,DING H W,LI B,et al. Prediction of Protein Subcellular Localization Based on Clustering and Feature Fusion[J]. Computer Science,2021,48(3):206-213.

[33] XIE T Y,ZHOU X G,HU J,et al. Contact Map-based Residue-pair Distances Restrained Protein Structure Prediction Algorithm[J]. Computer Science,2020,47(1):59-65.

[34] LI Y,LI Z X,TENG L,et al. Comment sentiment analysis and sentiment word detection based on attention mechanism[J]. Computer Science,2020,47(1):186-192.

[35] CHANG C C H,SONG J N,TEY B T,et al. Bioinformatics approaches for improved recombinant protein production in Escherichia coli:protein solubility prediction[J]. Briefings in Bioinformatics,2014,15(6):953-962.

NIU Fu-sheng, born in 1993, postgraduate. His main research interests include deep learning and bioinformatics.

LI Wei-hua, born in 1977, Ph.D, associate professor. Her main research interests include data mining and bioinformatics.

(责任编辑:喻藜)