

DOI:10.13232/j.cnki.jnju.2021.04.011

一种混合深度神经网络的赖氨酸乙酰化位点预测方法

颜志良, 丰智鹏, 刘 丹, 王会青*

(太原理工大学信息与计算机学院, 太原, 030024)

摘 要: 赖氨酸乙酰化(Lysine acetylation, Kace)普遍存在于人体代谢酶中,与多种代谢疾病密切相关,因此准确识别该位点对于代谢疾病治疗的研究具有重要意义。现有的 Kace 位点预测方法大多采用蛋白质序列层面的信息作为输入,蛋白质结构特性考虑不全面;特征提取时未关注氨基酸残基间顺序相关性,信息丢失严重,降低了预测准确度。提出一种新的 Kace 位点预测深度学习 CL-Kace 模型。CL-Kace 引入蛋白质结构特性,并与蛋白质原始序列、氨基酸理化属性共同构建位点特征空间,采用卷积神经网络(Convolutional Neural Network, CNN)提取特征;引入双向长短期记忆(Bidirectional Long Short-Term Memory, BiLSTM)网络捕获残基间的顺序依赖关系,以提高网络的抽象能力,识别潜在的 Kace 位点。实验结果表明,CL-Kace 模型优于现有的 Kace 位点预测器,能够有效地预测潜在的位点。

关键词: 赖氨酸乙酰化,蛋白质结构特性,卷积神经网络,双向长短期记忆网络,特征学习

中图分类号: TP391

文献标志码: A

A hybrid deep neural network-based method for predicting lysine acetylation sites

Yan Zhiliang, Feng Zhipeng, Liu Dan, Wang Huiqing*

(College of Information and Computer, Taiyuan University of Technology, Taiyuan, 030024, China)

Abstract: Lysine acetylation (Kace) is common in human metabolic enzymes and is closely related to a variety of metabolic diseases. Therefore, accurately identifying this site is of great significance for investigating metabolic disease treatments. Most of existing prediction methods use protein sequence level information as input, and the protein structural properties are not considered comprehensively. During the feature extraction, the sequential correlation between amino acid residues is ignored, and the information loss is serious, which reduces the prediction accuracy. Therefore, we propose a novel Kace site prediction deep learning model, CL-Kace (Kace site Prediction based on Convolutional and Long Short-Term Memory Networks). CL-Kace introduces the protein structural properties and constructs the feature space of the site together with the original protein sequence and the amino acid physicochemical properties. Then, the Convolutional Neural Network (CNN) is used to extract features. We adopt Bidirectional Long Short-Term Memory (BiLSTM) network to capture the sequential dependence between residues to improve the network abstraction ability to identify potential Kace sites. The experimental results show that CL-Kace is superior to the existing Kace site predictors and can effectively predict the potential sites.

Key words: lysine acetylation, protein structural properties, Convolutional Neural Network, Bidirectional Long Short-Term Memory, feature learning

基金项目: 国家自然科学基金(61976150), 山西省重点研发计划(高新技术领域)(201903D121151)

收稿日期: 2021-04-13

* 通讯联系人, E-mail: 1013208257@qq.com

翻译后修饰(Post-Translational Modifications, PTMs)指蛋白质成熟过程中的化学修饰,与蛋白质的结构和功能密切相关,几乎参与细胞所有的生命活动^[1-2]。现今,研究人员已发现 400 多种 PTMs,其中乙酰化(Acetylation)是最重要的 PTMs 之一^[3]。乙酰化修饰通常指蛋白质序列上的赖氨酸(Lysine, K)残基经乙酰基转移酶(Lysine Acetyltransferases, KATs)催化并在 ϵ 氨基上共价结合乙酰基团的过程^[4],其普遍存在于人体的代谢酶中,调节代谢通路及代谢酶的活性,与心血管、癌症、神经退行性等代谢疾病的产生密切相关^[5-8]。研究表明,通过药物调节人体内相关一系列酶的乙酰化修饰和活性,可以控制人体代谢平衡,进而有效地防治代谢疾病^[8-9]。因此,赖氨酸乙酰化(Lysine Acetylation, Kace)位点及其底物的鉴定,对于研究和分析人体乙酰化修饰潜在的分子机制具有重要意义,也为代谢疾病致病过程的研究和药物开发提供了有用的修饰信息^[6,10]。

Kace 作为一种酶促修饰,其肽链序列呈现一定的规律,目前已经开发了许多 Kace 位点预测工具^[1,10-18]。PAIL^[10]基于蛋白质原始序列预测 Kace 位点。N-Ace^[11]为了弥补蛋白质原始序列信息的单一性而增加溶剂可及性和氨基酸理化属性信息,取得了比 PAIL^[10]更高的准确度。PSKacePred^[14]融合氨基酸组成、蛋白质进化相似性和氨基酸理化属性信息,较大提升了 Kace 位点的预测准确度。然而,这些方法大多局限于蛋白质原始序列、氨基酸理化性质等序列层面的信息,很少涉及空间层面的蛋白质结构特性信息。尽管 N-Ace^[11]和 PLMLA^[12]分别考虑了溶剂的可及性和二级结构,但包含的蛋白质结构特性信息并不全面。

随着蛋白质结构特性预测技术的发展^[19],研究人员将更丰富的蛋白质结构特性引入 PTMs 位点预测来描述位点的特征^[20-22]。Success^[21]将二级结构、骨干扭转角和可及表面积三类蛋白质结构特性与蛋白质进化特征结合,在赖氨酸琥珀酰化位点识别上取得了较好的结果。Phoglystruct^[22]则将蛋白质结构特性输入多层感知机中预测赖氨酸磷酸甘油化位点,其结果表明蛋白质结构特性在区分磷酸甘油化和非磷酸甘油化赖氨酸残基上具

有重要作用。因此,蛋白质结构特性包含高度有用的局部和全局结构特征,为 PTMs 的鉴定提供了有力依据。

近年来,大多数 Kace 位点预测方法是基于支持向量机(Support Vector Machine, SVM)^[1,11-16]、朴素贝叶斯(Naive Bayes, NB)^[10]、逻辑回归(Logistic Regression, LR)^[17]等传统机器学习算法设计的。虽然这些方法取得了一定的成果,但依赖人工选择特征,主观性强,难以挖掘潜在信息^[23]。深度学习方法可以自动从原始数据中学习高级表示,是解决上述问题的良好手段,已成功应用于生物信息学领域^[23-27]。MusiteDeep^[23]采用卷积神经网络(Convolutional Neural Network, CNN)提取蛋白质原始序列特征,结合注意力网络进行磷酸化位点预测,取得了较好结果。Long et al^[25]集成 CNN 和长短期记忆(Long Short-Term Memory, LSTM)网络预测了蛋白质羟基化位点,该方法采用 CNN 提取氨基酸的复杂局部特征,通过 LSTM 捕获氨基酸之间的长期依存关系,因而提高了预测器的质量。类似地,Deep-ACLSTM^[26]将非对称 CNN 和双向长短期记忆(Bidirectional Long Short-Term Memory, BiLSTM)网络相结合,有效预测了蛋白质二级结构。因此,将 CNN 与 BiLSTM 组合能够同时关注原始数据的局部信息和长期依赖信息,有效减少信息的丢失,有助于 Kace 预测结果的提高。

在 Kace 位点预测方面,NetAcet^[28]采用神经网络对潜在位点进行了预测,但有限的训练集限制了其性能,无法体现神经网络的优势。近年来,CapsNet^[2]融合五维氨基酸的综合理化特性和一维空缺位置信息作为编码方式,采用 CNN 提取特征,利用胶囊网络预测了 Kace 位点。DeepAcet^[29]基于“F 分数”选择特征,训练深度神经网络(Deep Neural Networks, DNN)预测 Kace 位点。然而上述方法仅考虑了蛋白质序列层面的信息,未考虑蛋白质结构特性信息;且特征提取时仅关注氨基酸残基的局部信息,忽略了不同残基间的顺序依赖关系,信息丢失严重,影响 Kace 位点的预测结果。

基于以上问题,本文首先在考虑蛋白质原始序列和氨基酸理化属性信息的基础上引入二级结

构、骨干扭转角和可及表面积三类蛋白质结构特性信息来更好地描述Kace位点的初始特征空间;然后将CNN与BiLSTM有效组合来同时提取位点的复杂空间特征和氨基酸残基间的长期顺序依赖特性,以提高Kace位点高级表示的质量;最后连接Softmax层构建一种新的Kace位点预测深度学习CL-Kace模型.实验结果表明,CL-Kace模型优于现有的预测方法,有效学习了Kace位点的抽象模式.潜在的Kace位点的预测结果进一步说明,CL-Kace模型是识别未知乙酰化位点的有力工具.

1 赖氨酸乙酰化位点预测模型

Kace位点的预测可以被抽象为二分类问题,即每个潜在的位点可以被分类为Kace或non-Kace^[2].以赖氨酸(K)为中心,本研究提取 $L=2n+1$ 长度的肽链(每侧 n 个残基)作为基序,通过编码方式将基序转化为数值向量作为CL-Kace模型的输入.然后,基于训练数据训练提出的CL-Kace模型,使其学习到Kace位点的深层抽象和基序模式.最后,将训练好的模型用于潜在位点的预测.

1.1 信息编码 本文提出的模型是一种端到端的Kace位点预测深度学习模型,因此训练模型之前需要将蛋白质结构特性、蛋白质原始序列和氨基酸理化属性三类信息向量化.本文通过SPIDER3^[19]的结果(包括八个指数,即二级结构(Secondary structure):helix(H),coil(C),strand(E),骨干扭转角(Backbone torsion angles) $\varphi, \psi, \theta, \tau$,可及表面积(Accessible Surface Area, ASA))对蛋白质结构特性信息进行编码,基序长度为 L 时得到 $L \times 8$ 维的蛋白质结构特性信息的向量表示.对于蛋白质原始序列信息,本文采用one-of-21编码^[2]进行数值化,基序长度为 L 时得到 $L \times 21$ 维的蛋白质原始序列信息的向量表示.另外,本文采用Atchley因子^[30]编码氨基酸理化属性信息,每个氨基酸残基由五个Atchley因子值表示,选择Atchley因子是因为它们统计分析了500种氨基酸理化属性指数,综合反映了氨基酸极性、二级结构、分子体积、密码子多样性和静电荷等信息,基

序长度为 L 时得到 $L \times 5$ 维的氨基酸理化属性信息的向量表示.

1.2 CL-Kace模型结构 本文目的是构建一个可以高效学习Kace位点深层次隐藏特性的深度学习模型.为了充分利用网络中的参数,提取Kace位点基序的深度特征,采用CNN提取位点的蛋白质结构特性、蛋白质原始序列和氨基酸理化属性的高级表示.考虑到基序上不同氨基酸残基间具有顺序相关性,引入BiLSTM关注氨基酸残基间的长期依赖关系,以减少信息丢失.最后,采用Softmax层进行分类来有效预测潜在的Kace位点.CL-Kace模型结构如图1所示.

1.3 CNN提取高级特征 本文引入CNN来进行特征学习,CNN通过共享参数的卷积核进行卷积操作,提取位点的局部空间信息,而不同的卷积核提取不同的局部信息,这些信息组合构成位点的空间特征,比全连接深度网络更高效.本文采用的CNN为一维卷积网络,包含一个卷积层和一个整流线性单元.卷积层通过卷积核矩阵与输入数据矩阵之间的点乘求和,抽取原始数据中的隐藏特征来学习位点基序的空间特性.卷积操作后采用整流线性单元去除负面信息,保留对Kace位点分类有用的信息,该过程通常称为激活.以长度为 L 的位点基序为例,一维CNN的特征提取过程如式(1)至式(3)所示:

$$X' = f_{\text{conv}}(I) \quad (1)$$

$$X'_{ij} = \sum_{k=1}^K \sum_{r=1}^R W_{k,r}^i \cdot I_{k,r}^j + b_i \quad (2)$$

$$X^{(\text{CNN})} = f_{\text{relu}}(X') = \text{relu}(X') = \max(0, X') \quad (3)$$

其中, $I \in \mathbb{R}^{K \times L}$ 为输入层编码向量, $K \in \{8, 21, 5\}$ 为三个分支网络中每种氨基酸的编码长度. X'_{ij} 为第 i 个卷积核滑动到第 j 个窗口的特征表示, $i \in \{1, \dots, N_{\text{filter}}\}$, N_{filter} 为卷积核的数量; $j \in \left(1, \dots, \left\lfloor \frac{L'-R}{S} \right\rfloor + 1\right)$, R 为卷积核的大小, S 为滑动窗口步长,由于本文使用了填充策略,则 $\left\lfloor \frac{L'-R}{S} \right\rfloor + 1 = L$. $W \in \mathbb{R}^{N_{\text{filter}} \times K \times R}$ 为权重矩阵, b 为偏置项. $X^{(\text{CNN})} \in \mathbb{R}^{N_{\text{filter}} \times L}$ 为一维CNN的输出.

在将CNN提取的蛋白质结构特性、蛋白质原

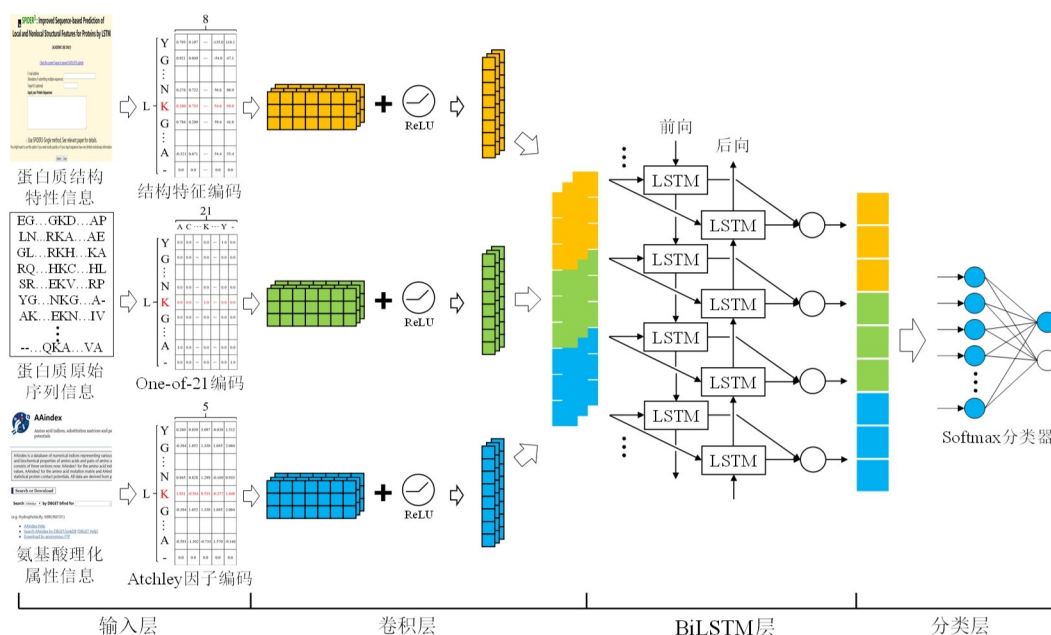


图 1 CL-Kace 的模型框架

Fig. 1 The model framework of CL-Kace

始序列和氨基酸理化属性三类特征信息输入下一层之前,将它们直接串联起来生成单一特征表示,便于 BiLSTM 同时从序列和结构层面提取基序氨基酸残基间的顺序依赖信息。

1.4 BiLSTM 捕获长期依赖信息 考虑到位点基序上不同氨基酸残基之间的顺序相关性,本文引入 BiLSTM 捕获基序上残基间的长期相互依赖信息,以增强网络中的乙酰化信息流,提高网络的判别能力。BiLSTM 由 LSTM 单元组成,单个单元的结构如图 2 所示。

LSTM 单元接收输入数据后,第一步是通过“遗忘门”和上一时刻的输出来决定保留哪些过往

信息;第二步分成两部分,一是通过“输入门”生成新的信息,二是通过 tanh 层添加新的信息,更新当前单元状态;第三步是通过“输出门”和 tanh 层得到 LSTM 单元的输出。LSTM 单元在时间步骤 t 时的计算过程如式(4)至式(8)所示:

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (4)$$

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (5)$$

$$C_t = f_t C_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (6)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (7)$$

$$h_t = o_t \tanh(C_t) \quad (8)$$

其中, f_t, i_t, C_t, o_t 和 h_t 分别代表遗忘门、输入门、单元状态、输出门和隐藏状态; x_t 代表时间步骤 t 时的 LSTM 单元输入; W 和 b 分别代表权重矩阵和偏置项。因此, LSTM 单元通过门控机制调整单元内部信息流,通过“遗忘门”控制历史信息,保证网络学习到残基间的依赖关系。对于 BiLSTM 网络,有来自两个方向的输出,本文采用串联的方式连接它们。

1.5 Softmax 层预测赖氨酸乙酰化位点 基于 BiLSTM 网络的输出 $X^{(LSTM)}$, 本文训练了 Softmax 分类器进行 Kace 位点预测。Softmax 层接收 $X^{(LSTM)}$ 作为输入, 经过加权求和和激活操作后得

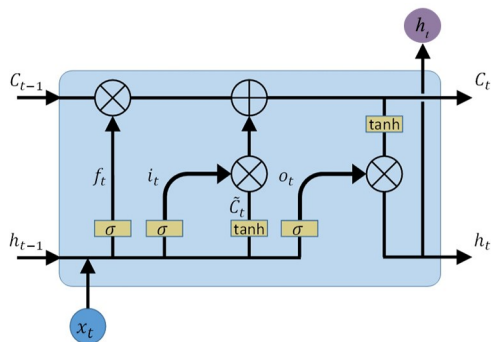


图 2 LSTM 单元结构

Fig. 2 The structure of LSTM unit

到样本的预测类别, Softmax层的前向传播过程如式(9)所示:

$$P(y=i|x) = \frac{e^{w_i^S X^{(LSTM)} + b_i^S}}{\sum_{j=1}^2 e^{w_j^S X^{(LSTM)} + b_j^S}} \quad (9)$$

其中, W_i^S , W_j^S 为权重矩阵, b 为偏置项. $P(y=i|x)$ 表示样本 x 预测为 i 类的概率, 如前所述, 由于 Kace 位点预测可以被抽象为二分类问题, 因此 $i \in \{0, 1\}$, 概率最大所对应的类别即为预测类别.

2 模型训练

本文 CL-Kace 模型采用交叉熵作为代价函数, 以最小化训练误差:

$$L_c = -\frac{1}{N} \sum_{j=1}^N (y^j \ln P(y^j=1|x^j) + (1-y^j) \ln P(y^j=0|x^j)) \quad (10)$$

其中, N 为训练样本总数, y^j 为第 j 个输入基序的真实标签, x^j 为第 j 个输入基序. 为了减轻过拟合的影响, 本文采用 L2 正则化, 因此 CL-Kace 的目标函数定义为:

$$\min_w \left(L_c + \lambda \sum (\|w\|_2)^2 \right) \quad (11)$$

其中, λ 为正则化系数, $\|w\|_2$ 为权重矩阵的 L2 范数, 本文采用 Adam 算法对目标函数进行优化. 本文 CL-Kace 模型的算法流程如下所示.

算法 CL-Kace 模型的算法流程

输入: SPIDER3 编码的蛋白质结构特性、one-of-21 编码的蛋白质原始序列和 Atchley 因子编码的氨基酸理化属性

输出: 位点基序的类别分数

1. 网络参数初始化;
2. 对于每个 epoch:
3. 对于每个 batch:
4. CNN 前向传播, 提取三类信息的复杂空间特征;
5. 沿特征维串联三类特征, 输入 BiLSTM;
6. BiLSTM 前向传递, 学习氨基酸残基间的正向依赖性;
7. BiLSTM 后向传递, 捕获氨基酸残基间的反向依赖性;
8. 串联 BiLSTM 的前向和后向输出, 输入 Softmax 层;

9. Softmax 层前向传播, 预测位点的类别分数;
10. 根据预测分数和真实标签, 计算误差并反向传播, 更新参数;
11. 结束 batch;
12. 结束 epoch.

模型训练过程中还采用早停策略和 dropout 技术进一步防止模型过度拟合. 为了降低模型训练期间数据不平衡的负面影响, 采用类重新加权的方法来增加阳性样本的影响, 迫使模型学习占少数的阳性样本的抽象机制. CL-Kace 模型的超参数配置如表 1 所示. 实验过程中, CL-Kace 模型是基于 Keras 2.1.6 和 TensorFlow 1.13.1 实现的, 硬件环境为 AMD Ryzen 7 4800H CPU 2.9 GHz (Windows 10 系统), 并配备 Nvidia GeForce RTX 2060 (6 GB) 显卡.

表 1 CL-Kace 模型的超参数配置信息

Table 1 Details of the hyperparameters of the proposed model CL-Kace

超参数名称	值
卷积层数	1
卷积核大小	3
卷积核数量	96
卷积核移动步长	1
BiLSTM 隐藏单元数	128
dropout 率	0.6
L2 正则化项系数	0.0001
学习率	0.001
批处理大小	512
最大 epoch 数	2000
早停策略 patience 值	20
阳性:阴性的类权重比	9.0:1.0

3 实验结果与分析

3.1 实验数据与数据预处理 从蛋白质赖氨酸修饰数据库 (Protein Lysine Modifications Database, PLMD)^[31] 收集并下载了 6078 条和 1860 条经实验验证的人类和大肠杆菌赖氨酸乙酰化蛋白质数据. 考虑到 SPIDER3 服务器^[19] 无法处理含有非标准氨基酸的蛋白质序列, 手动删除了这些蛋白质序列 (人类: 28 条, 大肠杆菌: 0 条). 为了避免序列同源性较大而造成模型的偏差, 利用 CD-

HIT^[32]进行序列去冗余,阈值设定为0.3^[14],最终分别保留4681条和1654条乙酰化蛋白质序列.为了方便与其他预测器进行比较,将过滤后的乙酰化蛋白质序列随机选择10%构建独立测试集,剩余乙酰化蛋白质序列作为训练集,数据集的统计信息如表2所示.

表2 本文数据集的统计信息

Table 2 Statistics of the datasets in this paper

物种	数据集类型	蛋白质数目	阳性样本数	阴性样本数
人类	训练集	4212	20354	184097
	独立测试集	469	2540	19374
大肠杆菌	训练集	1488	7242	17162
	独立测试集	166	905	1759

3.2 评估指标 为了合理评估提出的模型,采用五种度量指标进行模型性能的评估,包括灵敏度(Sensitivity, Sn)、特异性(Specificity, Sp)、准确度(Accuracy, ACC)、马氏相关系数(Mathews Correlation Coefficient, MCC)和几何均值(Geometric mean, G -mean),它们的定义如下:

$$Sn = \frac{TP}{TP + FN} \quad (12)$$

$$Sp = \frac{TN}{TN + FP} \quad (13)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FN) \times (TN + FP)}} \quad (15)$$

$$G\text{-mean} = \sqrt{\frac{TP \times TN}{(TP + FN) \times (TN + FP)}} \quad (16)$$

其中, TP , TN , FP 和 FN 分别为真阳性、真阴性、假阳性和假阴性. 当阳性样本和阴性样本不平衡时, MCC 和 G -mean 指标更值得关注,因为它们可以更好地反映模型质量^[22,33-34]. 此外,采用接收器工作特性(Receiver Operating Characteristic, ROC)曲线下面积(Area under ROC, AUC)和精确率召回率(Precision-Recall, PR)曲线下面积(Area under PR, $AUPR$)来衡量模型整体性能, AUC 和 $AUPR$ 越高表明模型整体表现越好.

3.3 CL-Kace 的性能 在进行 Kace 位点预测实

验之前,通过在人类训练集上进行十折交叉验证来确定 CL-Kace 模型中的两个重要参数: CNN 层数、BiLSTM 隐藏单元数. 本实验遵循 Wang et al^[2]的研究,先将 Kace 位点的窗口大小预设为 33, 然后进行参数选择实验. 由于采用网格搜索策略确定参数需要花费大量的时间,为了简化该过程并确保参数的合理性,首先预设 CNN 层数为一层,然后对 BiLSTM 隐藏单元数进行选择,实验结果如图 3a 所示. 由图可见,当 BiLSTM 隐藏单元数为 128 时, CL-Kace 模型取得最好的结果,因此将 BiLSTM 隐藏单元数确定为 128. 接下来,固定 BiLSTM 隐藏单元数为 128, 对 CNN 层数进行选择,实验结果如图 3b 所示. 由图可知,过多的卷积层会造成信息冗余,降低模型性能,因此将 CL-Kace 模型的 CNN 层数设置为 1.

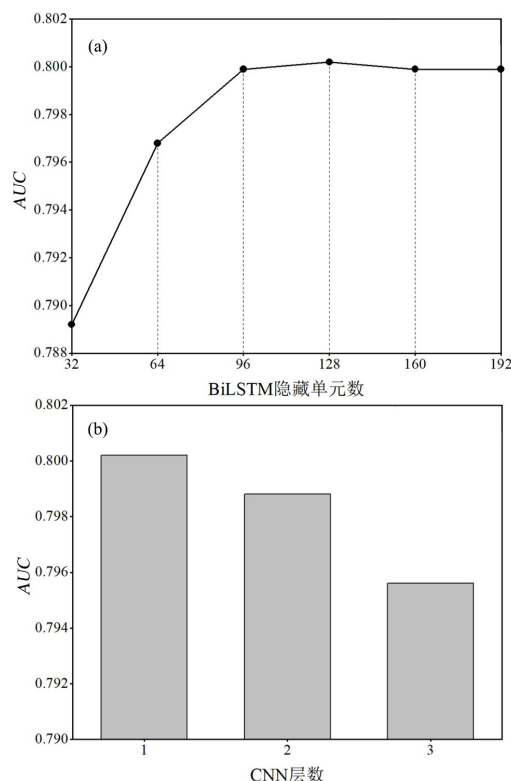


图3 CL-Kace 模型的参数实验结果

Fig. 3 Experimental results of the CL-Kace model with different parameters

蛋白质序列片段的大小直接决定模型学习到的高级表示,因此其值的选取对于 PTMs 位点的预测有重要影响. Wang et al^[2]使用 $L = 33$ 的窗口

进行Kace位点预测;Wu et al^[29]使用 $L=31$ 的窗口;索生宝等^[14]使用 $L=21$ 的窗口. 为了选择CL-Kace模型的最佳Kace位点的窗口大小,保证输入的信息量,尤其要充分利用CNN和BiLSTM自动、高效提取特征的特性,以人类训练集为基准,将窗口大小 L 的初始值设置为13,从13到61,以2递增,共对25个值进行十折交叉验证测试,实验结果取平均值. 如图4所示,随着窗口大小的

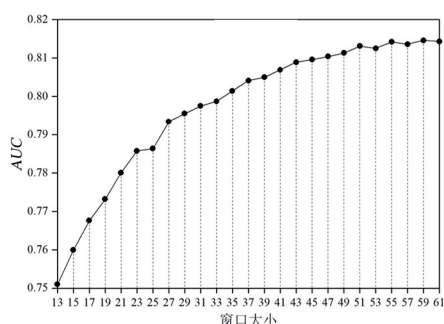


图4 CL-Kace模型的AUC值随Kace位点窗口大小 L 的变化结果

Fig. 4 AUC of the model CL-Kace with different window size L of the Kace site

增加,CL-Kace模型的性能随之提高;当窗口大小 $L>51$ 时AUC的上升趋势变得非常缓慢,且出现了微小波动. 因此,综合考虑计算成本和模型性能,选择CL-Kace的最佳乙酰化位点窗口大小为51,这一结果表明本文的深度学习模型需要较长的序列片段来学习潜在的长距离隐藏特性^[24].

当蛋白质序列片段窗口大小设置为51时,CL-Kace模型在人类训练集上的十折交叉验证结果如图5所示. 由图可知,CL-Kace模型的平均AUC和平均AUPR分别为0.8131和0.3293,它们的标准差分别为0.0028和0.0076. 小的标准差说明该模型是参数健壮的,并未因初始参数的不一致和较大的dropout率(0.6)而出现较大波动,影响模型的训练过程;也表明CL-Kace模型在处理每一折样本时都能较好地学习Kace位点的潜在特性,这也为解释模型的参数健壮性提供了依据.

为了考察模型的可靠性还进行了五次十折交叉验证,结果如表3所示. 由表可见,CL-Kace模型具有相当的稳定性,图5的实验结果并非偶然.

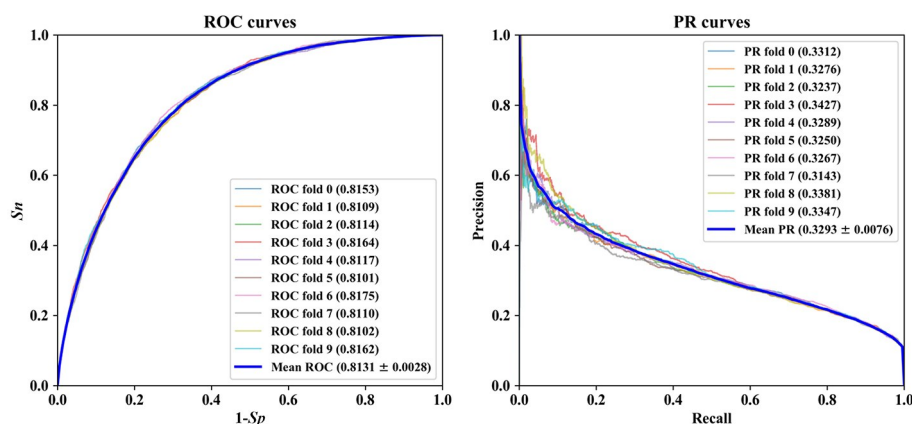


图5 CL-Kace模型在人类训练集上的十折交叉验证结果

Fig. 5 Result of the ten-fold cross-validation of CL-Kace on the training dataset of human

表3 CL-Kace模型在人类训练集上的五次十折交叉验证结果

Table 3 Results of five times ten-fold cross-validation of the model CL-Kace on the training dataset of human

次数	S_n	S_p	ACC	MCC	$G\text{-mean}$	AUC	$AUPR$
1	74.11±2.97	73.36±2.67	73.44±2.12	30.70±0.72	73.68±0.38	81.31±0.28	32.93±0.76
2	70.54±4.17	76.21±3.16	75.64±2.45	31.13±0.73	73.23±0.76	81.38±0.29	33.33±0.88
3	72.49±4.11	74.57±3.44	74.37±2.70	30.85±0.92	73.43±0.56	81.32±0.47	33.15±0.82
4	70.18±4.85	76.04±3.52	75.46±2.70	30.77±0.66	72.93±0.93	81.18±0.42	32.82±0.62
5	74.24±2.90	73.30±2.61	73.45±2.10	30.67±0.70	73.29±0.37	81.25±0.31	32.89±0.73

为了深入探讨 CL-Kace 在区分 Kace 和 non-Kace 位点上的作用,基于人类独立测试集采用 t -SNE^[35],将所有样本的初始编码向量和 CL-Kace 提取的高级抽象特征投影到二维空间中,并将该空间缩放到区间 $[-1, 1]$,如图 6 所示.由图可见, Kace 和 non-Kace 位点基序的初始编码向量在经过 CL-Kace 处理后,生成了区分性良好的高级表示,显示出较好的区分性.这一结果表明,本文提出的 CL-Kace 模型能够从蛋白质结构特性、蛋白质原始序列和氨基酸理化属性三类初级表示中学习它们的深层抽象并进行融合,来增强特征的区分能力,有助于阳性样本和阴性样本的分类.

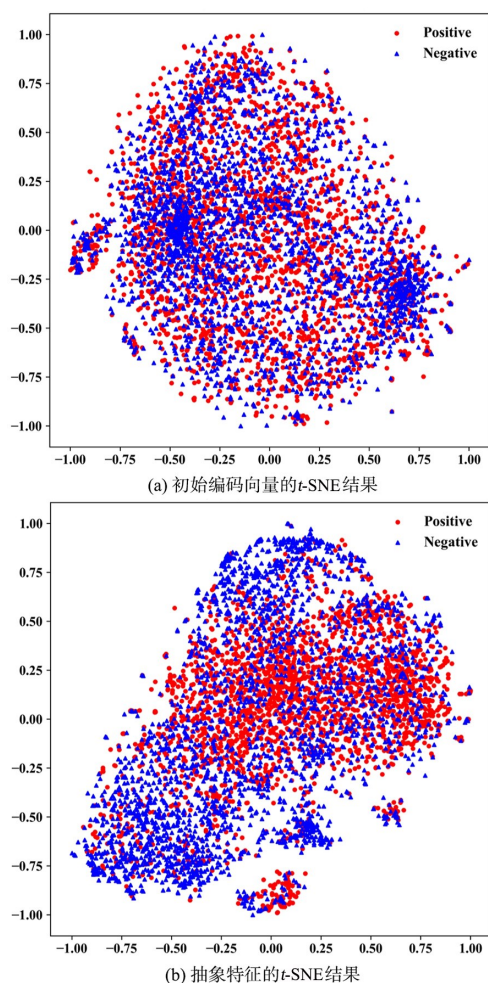


图 6 人类独立测试集样本的初始编码向量和 CL-Kace 提取的高级抽象特征的 t -SNE 可视化结果

Fig. 6 t -SNE visualization of the original coding vectors and abstract features extracted by CL-Kace on the independent test dataset samples of human

3.4 CL-Kace 消融实验 为了验证蛋白质结构特性对 Kace 位点预测的重要作用以及 CL-Kace 模型中各个网络的必要性,在人类训练集上实行十折交叉验证对模型进行消融性实验,主要包括: CNN 的引入, BiLSTM 的引入, 增加蛋白质结构特性作为模型输入. 实验结果如表 4 所示, 表中黑体字代表每个指标的最高值. 基线模型以基序的蛋白质原始序列和氨基酸理化属性信息串联作为输入, 采用 DNN 预测 Kace 位点.

表 4 CL-Kace 消融实验的结果

Table 4 Results of CL-Kace ablation experiments

组件	模型结构(列)			
基线模型	✓	✓	✓	✓
DNN→CNN		✓	✓	✓
+BiLSTM			✓	✓
+蛋白质结构特性				✓
MCC	17.47%	25.90%	27.98%	30.70%
G-mean	58.31%	70.41%	72.31%	73.68%
AUC	70.07%	78.53%	79.72%	81.31%
AUPR	20.03%	28.10%	29.36%	32.93%

由表 4 的第二列和第三列可见, CNN 使基线模型的 MCC, G-mean, AUC 和 AUPR 均有较大提升, 表明 CNN 通过局部感受野提取到基序氨基酸残基的深层隐藏特性, 其权重共享机制一方面减少了网络参数量, 另一方面有助于位点共性特征的获取, 和 DNN 在特征学习方面相比能更充分地利用网络中的乙酰化信息流, 减少信息丢失. 然后引入 BiLSTM, 模型的 MCC, G-mean, AUC 和 AUPR 有进一步提高, 如表 4 的第三列和第四列所示, 表明 BiLSTM 通过记忆单元捕获序列中不同氨基酸残基间的相互依赖关系, 有效关注了位点基序的顺序相关性, 使网络学习到更好的特征表示. 蛋白质结构特性已在相关研究中被证明对 PTMs 位点的预测具有重要作用^[20-22], 因此增加蛋白质结构特性作为模型输入, 如表 4 第四列和第五列所示, 模型的 MCC, G-mean, AUC 和 AUPR 比仅使用蛋白质原始序列和氨基酸理化属性信息分别提高 2.72%, 1.37%, 1.59% 和 3.57%. 这些结果表明, 蛋白质结构特性中的蛋白质局部和全局结构特征是 Kace 位点基序模式

的重要部分,引入蛋白质结构特性使位点的特征空间信息更丰富,且CL-Kace模型可以很好地从该信息中提取到高级表示。

3.5 CL-Kace与其他赖氨酸乙酰化位点预测方法的比较 为了评估CL-Kace的性能,将其与现有的Kace位点预测模型进行比较。由于大多数模型使用不同的训练数据且未提供独立工具,难以进行直接比较,所以本研究选择八种可用且有代表性的模型进行实验,即MusiteDeep^[23],CapsNet^[2],DeepAcet^[29],PSKAcePred^[14],EnsemblePail^[13],GPS-PAIL 2.0^[7],PHOSIDA^[15]和ProAcePred^[16],它们的主要信息如表5所示。八种模型中,PHOSIDA具有人类物种特异性,ProAcePred具有原核生物特异性,因此本文在人类数据集上评估PHOSIDA,在大肠杆菌数据集上评估ProAcePred。

本研究中,MusiteDeep被重新训练且作为深度学习的典型模型;CapsNet是胶囊网络在蛋白质翻译后修饰预测中的成功探索,有重要的指导意义;DeepAcet是DNN在Kace位点预测中的前沿应用;PSKAcePred,EnsemblePail,GPS-PAIL 2.0,PHOSIDA和ProAcePred被看作是传统机器学习方法的代表,前四种提供独立工具,可以进行模型的重新训练,而后四种仅提供了Web服务,所以只在独立测试集上对它们进行评估。上述模型在人类训练集上的十折交叉验证的结果如表6所示(由于EnsemblePail和GPS-PAIL 2.0未提供独立工具用于模型训练,所以无法获取它们的十折交叉验证结果),表中黑体字表示每个指标的最高值。

在对应独立测试集上的结果如图7所示(由于EnsemblePail,GPS-PAIL 2.0和PHOSIDA的

表5 不同对比方法的主要信息

Table 5 The main information of different comparison methods

模型名称	信息编码	分类算法
MusiteDeep	one-of-21 编码蛋白质原始序列	CNN+注意力网络
CapsNet	五维氨基酸综合理化性质和一维空缺位置信息	CNN+胶囊网络
DeepAcet	one-hot 编码、Blosum62、K-间隔氨基酸对组成、信息增益、AAIndex 和位置特异性得分矩阵	DNN
PSKAcePred	氨基酸组成、蛋白质进化相似性和氨基酸理化属性	SVM
EnsemblePail	改进的位置加权矩阵编码序列特征	集成SVM
GPS-PAIL 2.0	BLOSUM62	GPS 2.2 算法
PHOSIDA	蛋白质原始序列信息	SVM
ProAcePred	氨基酸组成、二元氨基酸编码、位置权重氨基酸组成、K空间氨基酸对组成、平均可及表面积、基于分组的权重编码、KNN 进化特征	SVM

预测结果未提供分类概率值,所以无法获取它们的AUC和AUPR)。

从表6可以看到,本文CL-Kace模型除MCC和AUPR外,其他指标均为最高值。如表6第五列和第八列所示,对于MCC和AUPR,DeepAcet和PSKAcePred具有较高结果,这是因为它们需要在基准训练集上实施欠采样构建平衡子集进行模型训练,考虑的阴性样本不全面,且AUPR对于数据集的不平衡很敏感^[34],导致结果存在偏差。从图7可以看出,CL-Kace具有最高的Sn,

表6 不同方法在人类训练集上的十折交叉验证性能

Table 6 The ten-fold cross-validation performance of different methods on the training dataset of human

模型名称	Sn	Sp(%)	ACC(%)	MCC(%)	G-mean(%)	AUC(%)	AUPR(%)
MusiteDeep	73.39±1.47	69.48±1.52	69.87±1.24	26.97±0.54	71.39±0.33	78.46±0.58	28.05±0.96
CapsNet	72.16±1.59	70.93±1.22	71.05±0.96	27.37±0.41	71.53±0.35	78.40±0.39	28.10±0.71
DeepAcet	69.51±1.55	61.40±1.61	65.45±0.64	31.02±1.28	65.31±0.65	70.92±0.81	68.50±0.97
PSKAcePred	71.25±1.06	67.61±0.79	69.43±0.47	38.89±0.95	69.40±0.46	76.46±0.56	75.34±0.65
CL-Kace	74.11±2.97	73.36±2.67	73.44±2.12	30.70±0.72	73.68±0.38	81.31±0.28	32.93±0.76

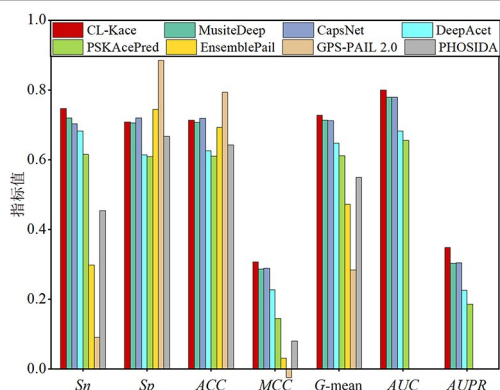


图 7 CL-Kace 与其他方法在人类独立测试集上的结果

Fig. 7 The results of CL-Kace and other predictors on the independent test dataset of human

MCC, G-mean, AUC 和 AUPR, 在独立测试集上表现最优, 和其他预测器相比, Kace 位点预测能力更强. 还可以注意到, PSKAcePred 在独立测试集和训练集上的 MCC, G-mean, AUC 和 AUPR 值相差较大, DeepAcet 亦是如此, 而本文 CL-Kace 模型在独立测试集和训练集上的相应结果基本一致, 表明 CL-Kace 模型通过在模型训练过程中使用类重新加权能有效地处理不平衡数据, 而无需使用随机欠采样构建平衡子集, 一定程度上解决了数据集信息缺失问题. 通过对比上述模型在独立测试集上的结果发现, 采用深度学习的方法 (如 MusiteDeep, CapsNet, DeepAcet 和 CL-Kace) 的泛化性能比传统机器学习方法 (如 PSKAcePred, EnsemblePail, GPS-PAIL 2.0 和 PHOSIDA) 更优异, 表明采用深度学习技术预测 Kace 位点是可行且有效的. 对于 Sp 和 ACC, 本文模型未取得最高值, 这是因为根据公式定义, Sn 与 Sp 有对抗性, CL-Kace 的高 Sn 造成了其相对较低的 Sp , 且 ACC 受不平衡数据的影响^[34]. 从图 7 还可以发现, GPS-PAIL 2.0 的 Sp 和 ACC 过度关注了阴性样本, 使其无法准确地识别真正的 Kace 位点.

基于不同方法在人类训练集上的十折交叉验证结果, 本文对 AUC 指标进行了方差分析, 结果如图 8 所示. 由图可见, 只有 MusiteDeep 与 CapsNet 之间的 P -value 为 0.7839, 其余的 P -value 均小于 0.001, 尤其在将 CL-Kace 模型的 AUC 加入分析后 P -value 更是达到 $2.1E-35$, 表明 CL-

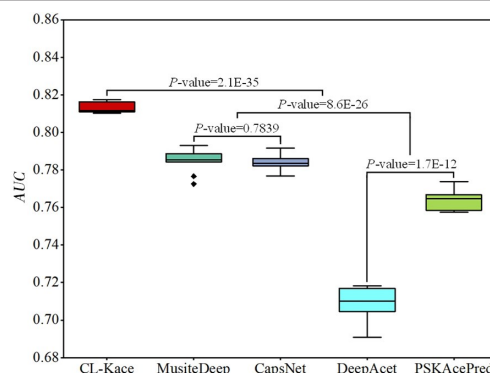


图 8 不同方法在人类训练集上的十折交叉验证 AUC 值的方差分析结果

Fig. 8 The result of an analysis of variance of the AUC values of different methods by the ten-fold cross-validation on the training dataset of human

Kace 模型和其他方法相比, 在预测 Kace 位点上有显著的优势.

由于 ROC 曲线和 PR 曲线可以更直观地比较各个预测器之间的性能, 本文绘制了 CL-Kace 和其他预测器在人类独立测试集上的 ROC 曲线、ROC(01) 曲线 (高特异性下的 ROC 曲线, 不包括提供 Web 服务的模型) 和 PR 曲线. 如图 9 所示, ROC 曲线和 PR 曲线均显示 CL-Kace 模型具有更好的 Kace 位点预测能力; ROC(01) 曲线表明, 即使在高特异性条件下, CL-Kace 也优于其他比较的预测器, 这一点非常重要^[1].

为了进一步验证 CL-Kace 模型的泛化能力, 本文还进行了大肠杆菌数据集的实验, 实验结果如表 7 所示 (由于 EnsemblePail, GPS-PAIL 2.0 和 ProAcePred 的预测结果未提供分类概率值, 因此无法获取它们的 AUC 和 AUPR), 表中黑体字表示每个指标的最高值. 由表可见, 本文 CL-Kace 模型具有较好的泛化性能, 并且适用于不同物种的数据, 这为更多的其他物种的 Kace 位点的预测提供了可用参考.

综上, CL-Kace 模型具有比其他方法更强的归纳能力, 其深度学习的体系结构设计合理、有效, 它通过 CNN 提取到 Kace 位点的蛋白质结构特性、蛋白质原始序列和氨基酸理化属性三类信息的深层次特征, 采用 BiLSTM 捕获不同残基间的长距离顺序相关性, 充分考虑数据的局部特性和全局依赖关系, 构建出更具区分性的高级表示,

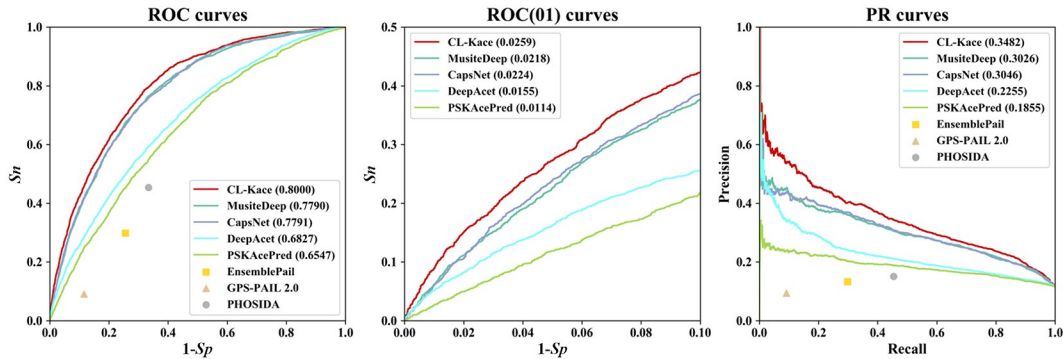


图9 CL-Kace和其他预测器在人类独立测试集上的ROC曲线、ROC(01)曲线和PR曲线

Fig. 9 The ROC,ROC(01) and PR curves of CL-Kace and other predictors on the independent test dataset of human

表7 CL-Kace与其他方法在大肠杆菌独立测试集上的结果

Table 7 The results of CL-Kace and other predictors on the independent test dataset of *Escherichia coli*

模型名称	S_n	S_p	ACC	MCC	G -mean	AUC	$AUPR$
MusiteDeep	75.36%	66.91%	69.78%	40.09%	71.01%	78.15%	61.49%
CapsNet	71.82%	70.15%	70.72%	40.04%	70.98%	78.48%	62.14%
DeepAcet	46.90%	75.67%	64.28%	23.43%	59.57%	63.32%	44.82%
PSKAcePred	62.10%	53.44%	56.38%	14.73%	57.61%	60.87%	41.06%
EnsemblePail	47.29%	54.63%	52.14%	1.83%	50.83%	—	—
GPS-PAIL 2.0	6.08%	90.45%	61.79%	-5.49%	23.45%	—	—
ProAcePred	8.40%	95.28%	65.77%	7.36%	28.29%	—	—
CL-Kace	76.02%	69.19%	71.51%	42.95%	72.52%	79.44%	62.15%

因而能够更准确地分类 Kace 和 non-Kace 位点.

3.6 CL-Kace 预测潜在的赖氨酸乙酰化位点

基于人类独立测试集的结果进行分析,评估 CL-Kace 模型识别未知的 Kace 位点的能力. 表8列出了 CL-Kace 预测为 Kace 的前 20 个候选位点,表

中黑体字表示实际发生乙酰化修饰的位点. 与 Luo et al^[27]类似,在 PLMD 和 Uniprot (<https://www.uniprot.org>) 中手动查验这 20 个候选位点,通过统计查验结果,发现 20 个候选位点中有 13 个是实际发生乙酰化修饰的,占比为 65%. 这些

表8 人类独立测试集上前 20 个候选位点的预测结果

Table 8 The prediction of the top 20 candidate sites from the independent test dataset of human

排名	蛋白质	位置	预测分数	排名	蛋白质	位置	预测分数
1	P53350	9	0.9878	11	Q8IXK0	280	0.9680
2	Q8N8A6	90	0.9840	12	Q13126	40	0.9678
3	Q9NQZ2	12	0.9819	13	O14810	14	0.9672
4	Q92522	143	0.9771	14	Q9P270	459	0.9655
5	Q9BQ61	118	0.9736	15	Q5T5U3	15	0.9651
6	Q9BQ15	203	0.9713	16	Q9UGM6	354	0.9649
7	O43719	190	0.9703	17	Q6ZRV2	740	0.9646
8	Q96TA2	691	0.9693	18	Q8WUM4	19	0.9643
9	P53007	19	0.9692	19	Q99436	72	0.9638
10	Q9H7N4	943	0.9684	20	Q8WUR7	109	0.9635

结果表明,CL-Kace模型充分利用了位点基序的蛋白质结构特性、蛋白质原始序列和氨基酸理化属性三类信息,将CNN和BiLSTM进行合理组合,学习赖氨酸残基周围氨基酸的复杂空间特征和残基间的相互依赖关系,有效预测了潜在的Kace位点,这有助于相关致病过程机制的发现和

4 结 论

本文针对目前赖氨酸乙酰化(Kace)位点预测方法考虑信息不全面、特征学习效率较低的问题,提出CNN和BiLSTM混合的Kace位点预测深度学习CL-Kace模型.该模型引入二级结构、骨干扭转角和可及表面积三类蛋白质结构特性来更好地表征位点特征空间;采用CNN学习位点的局部隐藏特征;通过BiLSTM捕获氨基酸残基间的顺序依赖性,减少信息丢失;最后,基于提取的高级表示预测了潜在的Kace位点.消融性实验研究表明,蛋白质结构特性、CNN和BiLSTM均有助于模型预测能力的提升.对比实验表明,CL-Kace模型的预测性能比现有Kace位点预测器更优或至少相当,有助于潜在Kace位点的识别.潜在Kace位点的预测和分析结果进一步表明:CL-Kace模型可以学习Kace位点的有意义抽象表示,能够发现新的Kace位点,这为代谢疾病药物的开发提供了有用的位点修饰信息.

尽管CL-Kace模型在Kace位点预测中表现出较好的性能,但仍然是一个黑盒,缺乏有意义的生物学过程解释^[36],因此未来的工作将集中于该模型的生物学解释,同时考虑一些有效的结构,如空间注意力^[37]来改进模型.由于已知的Kace位点的数量有限,下一步的工作是在模型训练阶段考虑更多的不平衡数据集的处理方法,以提高模型的预测能力.

参考文献

- [1] Liu Y, Wang M H, Xi J N, et al. PTM-ssMP: A web server for predicting different types of post-translational modification sites using novel site-specific modification profile. *International Journal of Biological Sciences*, 2018, 14(8): 946—956.
- [2] Wang D L, Liang Y C, Xu D. Capsule network for protein post-translational modification site prediction. *Bioinformatics*, 2019, 35(14): 2386—2394.
- [3] Khoury G A, Baliban R C, Floudas C A. Proteome-wide post-translational modification statistics: Frequency analysis and curation of the swiss-prot database. *Scientific Reports*, 2011, 1: 90.
- [4] Nallamilli B R R, Edelmann M J, Zhong X X, et al. Global analysis of lysine acetylation suggests the involvement of protein acetylation in diverse biological processes in rice (*Oryza sativa*). *PLoS One*, 2014, 9(2): e89283.
- [5] 朱志坚, 王兵, 葛玮等. 血清组蛋白去乙酰化酶3对稳定性冠心病患者经皮冠状动脉介入治疗术后主要心血管不良事件的预测价值. *中国医师进修杂志*, 2020, 43(10): 939—943. (Zhu Z J, Wang B, Ge W, et al. Predictive value of serum histone deacetylase 3 on major adverse cardiovascular events in patients with stable coronary artery disease after percutaneous coronary intervention. *Chinese Journal of Postgraduates of Medicine*, 2020, 43(10): 939—943.)
- [6] Shao J L, Xu D, Hu L D, et al. Systematic analysis of human lysine acetylation proteins and accurate prediction of human lysine acetylation through bi-relative adapted binomial score Bayes feature representation. *Molecular BioSystems*, 2012, 8(11): 2964—2973.
- [7] Deng W K, Wang C W, Zhang Y, et al. GPS-PAIL: Prediction of lysine acetyltransferase-specific modification sites from protein sequences. *Scientific Reports*, 2016(6): 39787.
- [8] Butler C A, Veith P D, Nieto M F, et al. Lysine acetylation is a common post-translational modification of key metabolic pathway enzymes of the anaerobe *Porphyromonas gingivalis*. *Journal of Proteomics*, 2015(128): 352—364.
- [9] Zhao S M, Xu W, Jiang W Q, et al. Regulation of cellular metabolism by protein lysine acetylation. *Science*, 2010, 327(5968): 1000—1004.
- [10] Li A, Xue Y, Jin C J, et al. Prediction of N^ε-acetylation on internal lysines implemented in bayesian discriminant method. *Biochemical and Biophysical Research Communications*, 2006, 350(4): 818—824.

- [11] Lee T Y, Hsu J B K, Lin F M, et al. N-Ace: Using solvent accessibility and physicochemical properties to identify protein N - acetylation sites. *Journal of Computational Chemistry*, 2010, 31(15): 2759—2771.
- [12] 施绍萍, 索生宝, 邱建丁. 组合二级结构信息预测赖氨酸甲基化和乙酰化//中国化学会第28届学术年会论文集. 成都: 中国化学会, 2012: 1. (Shi S P, Suo S B, Qiu J D. Incorporating secondary structure for identification of lysine methylation and lysine acetylation//Proceedings of the 28th Annual Conference of the Chinese Chemical Society. Chengdu, China: Chinese Chemical Society, 2012: 1.)
- [13] Xu Y, Wang X B, Ding J, et al. Lysine acetylation sites prediction using an ensemble of support vector machine classifiers. *Journal of Theoretical Biology*, 2010, 264(1): 130—135.
- [14] 索生宝, 孙兴玉, 邱建丁. 结合多特征算法和信息熵预测蛋白质乙酰化位点//第十一届全国计算(机)化学学术会议论文集. 兰州: 中国化学会, 2011: 51. (Suo S B, Sun X Y, Qiu J D. Combining Multi-feature algorithm and information entropy to analyze protein lysine acetylation//Proceedings of the 11th National Conference on Computational Chemistry of the Chinese Chemical Society. Lanzhou, China: Chinese Chemical Society, 2011: 51.)
- [15] Gnad F, Gunawardena J, Mann M. PHOSIDA 2011: The posttranslational modification database. *Nucleic Acids Research*, 2011, 39(S1): D253—D260.
- [16] Chen G D, Cao M, Luo K, et al. ProAcePred: prokaryote lysine acetylation sites prediction based on elastic net feature optimization. *Bioinformatics*, 2018, 34(23): 3999—4006.
- [17] Hou T, Zheng G Y, Zhang P Y, et al. LAceP: Lysine acetylation site prediction using logistic regression classifiers. *PLoS One*, 2014, 9(2): e89575.
- [18] Lu Z K, Cheng Z Y, Zhao Y M, et al. Bioinformatic analysis and post-translational modification crosstalk prediction of lysine acetylation. *PLoS One*, 2011, 6(12): e28228.
- [19] Heffernan R, Yang Y D, Paliwal K, et al. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*, 2017, 33(18): 2842—2849.
- [20] Reddy H M, Sharma A, Dehzangi A, et al. GlyStruct: Glycation prediction using structural properties of amino acid residues. *BMC Bioinformatics*, 2019, 19(S13): 547.
- [21] López Y, Sharma A, Dehzangi A, et al. Success: Evolutionary and structural properties of amino acids prove effective for succinylation site prediction. *BMC Genomics*, 2018, 19(S1): 923.
- [22] Chandra A, Sharma A, Dehzangi A, et al. Phoglystruct: Prediction of phosphoglycerylated lysine residues using structural properties of amino acids. *Scientific Reports*, 2018(8): 17923.
- [23] Wang D L, Zeng S, Xu C H, et al. MusiteDeep: A deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics*, 2017, 33(24): 3909—3916.
- [24] He F, Wang R, Li J G, et al. Large-scale prediction of protein ubiquitination sites using a multimodal deep architecture. *BMC Systems Biology*, 2018, 12(S6): 109.
- [25] Long H X, Liao B, Xu X Y, et al. A hybrid deep learning model for predicting protein hydroxylation sites. *International Journal of Molecular Sciences*, 2018, 19(9): 2817.
- [26] Guo Y B, Li W H, Wang B Y, et al. DeepACLSTM: Deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. *BMC Bioinformatics*, 2019(20): 341.
- [27] Luo F L, Wang M H, Liu Y, et al. DeepPhos: Prediction of protein phosphorylation sites with deep learning. *Bioinformatics*, 2019, 35(16): 2766—2773.
- [28] Kierner L, Bendtsen J D, Blom N. NetAcet: Prediction of N-terminal acetylation sites. *Bioinformatics*, 2005, 21(7): 1269—1270.
- [29] Wu M Q, Yang Y X, Wang H, et al. A deep learning method to more accurately recall known lysine acetylation sites. *BMC Bioinformatics*, 2019, 20(1): 49.
- [30] Atchley W R, Zhao J P, Fernandes A D, et al. Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences of the United States of America*, 2005, 102(18): 6395—6400.

- [31] Xu H D, Zhou J Q, Lin S F, et al. PLMD: An updated data resource of protein lysine modifications. *Journal of Genetics and Genomics*, 2017, 44(5): 243—250.
- [32] Huang Y, Niu B F, Gao Y, et al. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, 2010, 26(5): 680—682.
- [33] Chicco D, Jurman G. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 2020, 21(1): 6.
- [34] He H B, Garcia E A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 2009, 21(9): 1263—1284.
- [35] Van Der Maaten L, Hinton G. Visualizing data using t -SNE. *Journal of Machine Learning Research*, 2008 (9): 2579—2605.
- [36] Ma J Z, Yu M K, Fong S, et al. Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, 2018, 15(4): 290—298.
- [37] Chen L, Zhang H W, Xiao J, et al. SCA - CNN: Spatial and channel - wise attention in convolutional networks for image captioning//*Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA: IEEE, 2017: 5659—5667.

(责任编辑 杨可盛)