

识别关键蛋白质的混合深度学习模型

刘文洋 郭延哺 李维华

云南大学信息学院 昆明 650500

(wyl20180901@163.com)



摘要 关键蛋白质是有机体生存不可或缺的蛋白质。关键蛋白质的识别有助于理解细胞生命的最低要求、发现致病基因和药物靶点,对疾病的诊治和药物设计等有重要意义。现有方法表明整合蛋白质互作网络和序列的相关特征可以提高对关键蛋白质的识别精度和鲁棒性。文中整合了基因表达谱、蛋白质互作网络和亚细胞位置信息,设计了一种混合神经网络模型 IEPHDL。该模型首次使用双向门控循环单元对基因表达谱进行特征学习,使用由多个全连接层组成的深度神经网络对 3 种数据特征进行深度再学习,充分发挥双向门控循环单元网络、全连接网络、Node2vec 在特征学习和表示方面的优势,实现对关键蛋白质的有效识别。实验表明,IEPHDL 对关键蛋白质识别的准确率为 88.7%,精确率为 86.2%,AUC 为 85.2%,其准确率比当前最优的中心性方法、机器学习方法、深度学习方法依次高出 13%,8.9%,3.8%,其他指标也均高于这三者。最后,通过实验分析,证实了双向门控循环单元网络依赖自身强大的特征学习能力,在关键蛋白质识别中起着关键作用。

关键词 关键蛋白质;蛋白质互作网络;深度学习;Node2vec;双向门控循环单元网络

中图法分类号 TP391

Identifying Essential Proteins by Hybrid Deep Learning Model

LIU Wen-yang, GUO Yan-bu and LI Wei-hua

School of Information Science and Engineering, Yunnan University, Kunming 650500, China

Abstract Essential proteins are those proteins that are essential to the viability of the organism. The identification of essential proteins helps to understand the minimum requirements of cell life, discover disease-causing genes and drug targets, and is of great significance for the diagnosis and treatment of diseases and drug design. Existing methods show that integrating protein interaction networks and the relevant features of sequences can improve the accuracy and robustness of essential proteins identification. In this paper, gene expression profiles, protein interaction networks and subcellular location information are integrated, and a hybrid neural network model IEPHDL is designed. The IEPHDL model uses bidirectional gated recurrent unit to perform feature learning on gene expression profiles for the first time, and uses a deep neural network composed of multiple fully connected layers to perform deep relearning of three data features, to give full play to the advantages of bidirectional gated recurrent unit network, fully connected network and Node2vec in feature learning and representation, to achieve effective identification of essential proteins. Experiment results show that, IEPHDL has an accuracy of 88.7% for essential protein identification, an precision of 86.2%, and an AUC of 85.2%. The accuracy is 13%, 8.9%, 3.8% higher than the current optimal centrality method, machine learning method, and deep learning method in turn, and other indicators are also higher than the three methods. Finally, through experimental analysis, it is confirmed that the bidirectional gated recurrent unit network relies on its strong feature learning ability and plays a key role in essential protein identification.

Keywords Essential proteins, Protein interaction network, Deep learning, Node2vec, Bidirectional gated recurrent unit network

1 引言

关键蛋白质是有机体生存不可或缺的蛋白质,关键蛋白质的移除或缺失使蛋白质复合物功能模块丧失,甚至导致生

物无法存活^[1]。关键蛋白质的识别有助于理解细胞的生长调控机制、发现疾病的致病基因和潜在药物靶点,同时对药物设计和疾病诊治等具有重要的理论意义和应用价值。

在生物实验室中,主要采用基因剔除、基因抑制、转座子

收稿日期:2020-07-13 返修日期:2020-09-19 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目: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(2019152)

This work was supported by the Scientific Research Foundation 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(2019152).

通信作者:李维华(lywey@163.com)

突变等方法来识别关键蛋白质,但这些实验方法费时又费力。随着高通量数据的累积,基于计算方法的关键蛋白质识别成为可能。因此,鲁棒的计算方法不仅能为关键蛋白质识别实验提供指导,也成为关键蛋白质识别的有力补充^[2]。

基于计算方法的关键蛋白质识别就是利用现有数据确定影响蛋白质重要程度的关键特征,并基于这些特征评估蛋白质生物功能的重要程度。拓扑特征是从生物网络如蛋白质互作(Protein-Protein Interaction, PPI)网络^[3]中获取的最重要的特征。基于节点中心性的度量方法是最典型的网络拓扑特征获取方法,主要包括度中心性(Degree Centrality, DC)^[4]、接近中心性(Closeness Centrality, CC)^[5]、特征向量中心性(Eigenvector Centrality, EC)、介数中心性(Betweenness Centrality, BC)^[6]等。基于拓扑特征的关键蛋白质识别方法对网络结构比较敏感,因此假阳性噪声和数据缺失会降低预测的性能^[7]。为了弥补基于拓扑特征识别方法的不足,整合多维度生物特征成为关键蛋白质识别的主要研究思路。除了网络拓扑特征,关键蛋白质识别中涉及的生物特征主要包括序列特征和功能特征。序列特征主要通过基因组或转录组学等数据来获取,如 Tang 等^[8]和 Li 等^[9]从基因表达谱中学习特征并将其与网络拓扑特征相融合,进而识别关键蛋白质。功能特征主要涉及亚细胞位置、分子功能等。例如, Zeng 等^[10]通过融合 PPI 网络、亚细胞位置与基因表达谱来识别关键蛋白质。虽然整合多维度生物特征^[11]可以提高关键蛋白质识别的鲁棒性,但是预测性能仍有待提高。

深度学习(Deep Learning, DL)依托深度神经网络的建模能力,不仅可以自动从原始数据中获得多重水平的特征,还可以对特征之间的非线性关系进行建模。深度学习自提出以来在图像处理、自然语言理解等任务中取得了突破性进展,在生

物信息领域也得到了广泛应用^[12]。深度学习不仅可以为序列数据的学习提供良好的支撑,也可以从网络模型中捕捉拓扑特征,将网络节点映射为低维稠密向量,并进一步直接支持分类、聚类、关联推理等应用。Zeng 等^[10]基于卷积神经网络(Convolutional Neural Network, CNN)^[13]、长短期记忆网络(Long Short-Term Memory, LSTM)与 Node2vec 技术,探索并验证了深度学习框架在关键蛋白质识别中的优势。

门控循环单元(Gated Recurrent Unit, GRU)^[14]是简化的 LSTM 模型。相比 LSTM 模型,GRU 具有训练参数少、张量运算少、训练速度快的优点,尤其是双向 GRU(Bidirectional Gated Recurrent Unit, BiGRU)可以从两个不同的方向捕获序列特征,在很多问题的建模和预测中均表现出更好的性能。例如, Guo 等^[15]结合多尺度卷积和注意力网络,验证了 BiGRU 网络在染色质开放性预测中的优越性。鉴于 BiGRU 在特征学习中的优势,本文首次基于 BiGRU 和 Node2vec,设计了一个识别关键蛋白质的混合神经网络模型(Identifying Essential Proteins by Hybrid Deep Learning Model, IEPHDL)。该模型利用 BiGRU 从两个不同的方向自动捕获基因表达谱中影响蛋白质重要性的关键特征,利用 Node2vec 技术将 PPI 网络的拓扑特征表示为低维向量,融合从 PPI 网络、基因表达谱和亚细胞信息中学习得到的特征,来实现对关键蛋白质的识别。

2 关键蛋白质预测模型

2.1 模型的基本框架

为了识别关键蛋白质,本文整合基因表达谱、蛋白质互作网络以及亚细胞定位信息,设计出一个关键蛋白质预测模型 IEPHDL。该模型由 3 个模块组成:输入模块、特征提取模块和输出模块。其基本框架如图 1 所示。

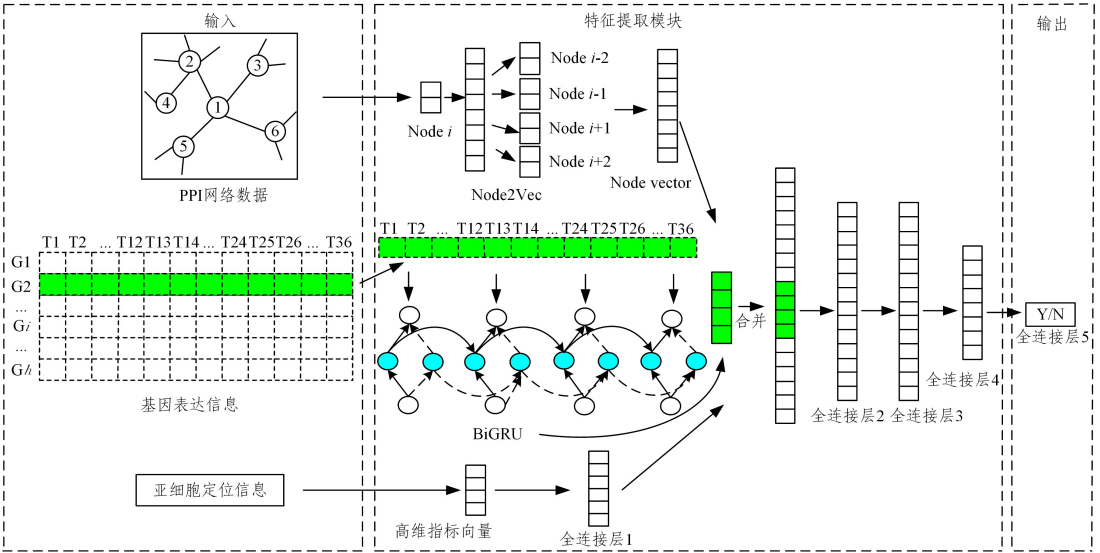


图 1 关键蛋白质识别模型

Fig. 1 Essential protein recognition model

输入模块包括 PPI 网络、基因表达谱和亚细胞信息 3 种数据。在特征提取模块,通过网络表征方法 Node2vec 对 PPI 网络节点进行表示学习,获取网络节点的拓扑特征;基因表达信息通过 BiGRU 网络获取特征;亚细胞定位

信息首先通过高维指标向量对节点进行表示,然后通过全连接层 FC₁ 学习获取特征。将蛋白质的以上 3 种特征合并作为节点的最终表示,依次输入全连接网络层 FC₂, FC₃ 和 FC₄ 学习并获取节点最终特征。最后,在输出模块通过

全连接层 FC_5 预测其是否为关键蛋白质。

2.2 特征提取模块

IEPHDL 模型的特征提取模块的主要任务是利用基因表达数据、PPI 网络和亚细胞位置信息获得评估蛋白质重要程度的关键特征, 这个任务包括 4 个部分: PPI 网络的节点映射、基因表达谱的特征提取、亚细胞位置信息的特征表示和全连接层特征获取。

2.2.1 PPI 网络的特征提取

关键蛋白质与蛋白质在 PPI 网络中的拓扑特性密切相关。因此, 有效利用 PPI 网络中蛋白质的拓扑特征对关键蛋白质识别具有重要意义。Node2vec 技术可以将网络中的节点映射为低维向量, 这样不仅可以有效表示网络中的拓扑特征, 也可以进一步与相关特征进行融合, 能明显提高识别关键蛋白质的效率。因此, IEPHDL 模型利用 Node2vec 技术捕捉 PPI 网络中的拓扑特征并获得低维向量表示。

IEPHDL 模型首先根据蛋白质互作数据, 建立节点为蛋白质的 PPI 网络 $N \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$; 其次, 学习映射函数 $\Phi: \mathcal{V} \rightarrow \mathbb{R}^{|\mathcal{V}| \times d}$, 将 $\mathcal{V}$ 映射到低维空间 $\mathbf{X} \rightarrow \mathbb{R}^{|\mathcal{V}| \times d}$, 获取低维特征向量 $\mathbf{X}_v \in \mathbb{R}^d$, d 是特征表示中向量的维度。最终获得蛋白质 p 的拓扑特征向量 $\mathbf{f}_N(p) = \mathbf{X}_v = [v_1, v_2, \dots, v_m]$, 其中 m 是节点特征的维度。

2.2.2 基因表达谱的特征提取

基因表达谱反映蛋白质功能发生的时间与空间等信息, 因此充分有效地从基因表达数据中捕获影响蛋白质识别重要程度的关键特征是模型的主要任务之一。

GRU 是 LSTM 的简化模型, 具有更新门和重置门 2 种门结构, 如图 2 所示。相比 LSTM, GRU 具有训练参数少、张量运算较少、训练速度快的优点, 并在很多应用中表现出良好的性能。GRU 利用重置门 r_t 控制与当前输入结合的先前信息的多少, 利用更新门 u_t 决定对先前记忆保留的程度。在第 t 步, 当前状态由先前学习的特征和即将学习的特征共同决定, 即当前状态向量 $\mathbf{h}_t$ 是通过在先前激活状态 $\mathbf{h}_{t-1}$ 和候选激活状态 $\tilde{\mathbf{h}}_t$ 之间的线性插值计算。

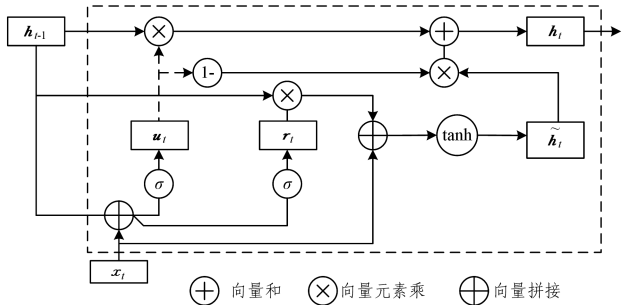


图 2 GRU 循环单元结构

Fig. 2 GRU cycle unit structure

h_t = (1 - u_t)h_{t-1} + u_t \tilde{h}_t

u_t = \sigma(W^u x_t + U^u h_{t-1})

其中, u_t 代表更新门, σ 是 sigmoid 激活函数。候选状态 $\tilde{\mathbf{h}}_t$ 的计算如下:

\tilde{h}_t = \delta(W^h x_t + U^h (r_t \odot h_{t-1}))

r_t = \sigma(W^r x_t + U^r h_{t-1})

其中, r_t 代表重置门, δ 代表 $\tanh$ 激活函数, $\odot$ 表示按元素相乘。

BiGRU 可以从两个不同的方向捕捉序列的关键信息, 在很多问题的建模和预测中表现出比 GRU 更好的性能。因此, IEPHDL 模型采用一个由前向后的 GRU 层和一个由后向前的 GRU 层分别提取基因表达谱中的关键特征。

基因表达序列为 $\mathbf{X} = [x_1, x_2, \dots, x_T]$, 其中 $x_i \in \mathbb{R}^k$, 在 t 时刻 $\mathbf{h}_t$ 的计算如下:

h_t = BiGRU(x_t, \overrightarrow{h_{t-1}}, \overleftarrow{h_{t-1}})

其中, $\overrightarrow{h_{t-1}}$ 和 $\overleftarrow{h_{t-1}}$ 分别是 $t-1$ 时刻 BiGRU 自前向后提取的上下文特征和自后向前学习的上下文特征。具体地说, $\mathbf{h}_t$ 由一个前向上下文 $\overrightarrow{h}_t$ 和一个后向上下文 $\overleftarrow{h}_t$ 组成, 即 $\mathbf{h}_t$ 按照如下公式进行计算:

h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t]

\overrightarrow{h}_t = GRU(x_t, \overrightarrow{h_{t-1}})

\overleftarrow{h}_t = GRU(x_t, \overleftarrow{h_{t-1}})

最终, IEPHDL 模型从基因表达序列中获得特征 $\mathbf{f}_x(p) = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n]$, 其中 n 是特征的维度。

2.2.3 亚细胞位置信息的特征表示

与非关键蛋白质相比, 关键蛋白质在某些亚细胞位置中更频繁地出现, 并且其进化更为保守^[16]。亚细胞位置通常分为 11 类: 1) 细胞骨架; 2) 胞质溶胶; 3) 内质网; 4) 内体; 5) 细胞外; 6) 高尔基体; 7) 线粒体; 8) 细胞核; 9) 过氧化物酶; 10) 血浆; 11) 液泡。每类亚细胞中关键和非关键蛋白质的数量如表 1 所列。

表 1 每类亚细胞中关键和非关键蛋白质的数量

Table 1 Number of essential and non-essential proteins in each type of subcellular localization

亚细胞定位	关键蛋白质数	非关键蛋白质数
细胞骨架	95	133
胞质溶胶	138	289
内质网	137	292
内体	22	109
细胞外	1	70
高尔基体	61	184
线粒体	173	753
细胞核	809	1 407
过氧化物酶	4	61
血浆	53	354
液泡	19	238

为了嵌入亚细胞信息, 同时与基因表达谱序列特征和 PPI 拓扑特征进行融合, IEPHDL 模型采用一个 11 维的 multi-hot 向量描述蛋白质的亚细胞信息。向量中每一维与一类亚细胞信息一一对应, 如果蛋白质在某个亚细胞中执行其功能, 则将其对应维度设置为 1, 否则设置为 0。然后将蛋白质 p 的亚细胞位置信息的表示输入全连接层 FC_1 提取特征, 最终获得 p 的亚细胞特征 $\mathbf{f}_s(p) = [s_1, s_2, \dots, s_q]$, 其中 q 是特征的维度。

2.2.4 全连接网络

IEPHDL 模型根据蛋白质 p 的 PPI 网络 N 、基因表达谱 X 以及亚细胞信息 S , 将影响关键蛋白质判定的主要特征映射为统一的形式 $\mathbf{h}_1$:

$$\mathbf{h}_1 = [\mathbf{f}_N(\mathbf{p}), \mathbf{f}_X(\mathbf{p}), \mathbf{f}_S(\mathbf{p})] \quad (9)$$

其中, $\mathbf{f}_N(\mathbf{p})$ 是从 PPI 网络 N 中捕捉的拓扑特征, $\mathbf{f}_X(\mathbf{p})$ 是 BiGRU 从基因表达谱 X 中捕捉的序列特征, $\mathbf{f}_S(\mathbf{p})$ 是从亚细胞定位信息 S 中获得的特征, $\mathbf{h}_1$ 是这 3 个特征的连接。

将 $\mathbf{h}_1$ 依次输入具有不同权重项 $\mathbf{W}$ 和偏置项 b 的 3 个全连接层 FC_2, FC_3 和 FC_4 中进行特征提取, 获取关键蛋白质的最终特征 $\mathbf{h}_2$:

$$\mathbf{h}_2 = \text{FC}_4(\text{FC}_3(\text{FC}_2(\mathbf{h}_1))) \quad (10)$$

最后, IEPHDL 模型使用权重项和偏置项分别为 $\mathbf{W}'$ 和 b' 的全连接层 FC_5 确定每个蛋白质的类别 y :

$$y = \text{FC}_5(\mathbf{h}_2) = (\mathbf{W}' \times \mathbf{h}_2 + b') \quad (11)$$

其中, $y=1$ 表示 p 是关键蛋白质, $y=0$ 表示 p 是非关键蛋白质。

2.3 模型训练

模型训练的目的在于根据训练样本获得模型的参数集合。定义损失函数:

$$l(\bar{y}_i, y_i) = y_i \bar{y}_i + (1 - y_i \log(1 - \bar{y}_i)) \quad (12)$$

其中, y_i 是真实标签, $\bar{y}_i$ 是模型预测的标签。模型训练的目标是最小化目标函数 ζ :

$$\zeta = - \sum_{i=1}^n l(\bar{y}_i, y_i) \quad (13)$$

其中, n 为训练集的样本数。

训练过程使用 Dropout^[17] 来实现随机让网络中某些隐藏层节点不工作, 进而阻止某些特征仅在其他特定特征下才有效果的情况。随后, 使用随机梯度下降 Adam^[18] 算法 (Adaptive Moment Estimation) 对模型进行训练, 参数更新规则如下:

$$\theta \leftarrow \theta + \alpha \frac{\partial L(\theta)}{\partial \theta} \quad (14)$$

其中 α 为学习率, θ 为模型的所有参数。

3 实验与结果分析

3.1 实验环境和参数

本文实验环境为: Windows10 (64 bit) 操作系统; TensorFlow 深度学习框架。

特征提取模块中 BiGRU 包括两个不同方向的 GRU 层, 每个 GRU 具有 16 个隐藏神经元, 整个 BiGRU 层共具有 32 个隐藏神经元, 并且使用 dropout (取值为 0.05) 避免过拟合。全连接层 $\text{FC}_1, \text{FC}_2, \text{FC}_3$ 和 FC_4 的激活函数均为线性整流函数 (Rectified Linear Unit, ReLU)。全连接层 FC_5 的激活函数为 sigmoid 函数, 用于最后的预测。具体参数如表 2 所列。

表 2 IEPHDL 的参数设置

Table 2 Parameter settings for IEPHDL

参数	取值
迭代次数	50
批大小	64
优化器	Adam
学习率	0.001
Dropout	0.05
数据维度 $m:n:q$	64:32:11
BiGRU 单元个数	32
$\text{FC}_1 - \text{FC}_5$ 单元个数	512:512:312:128:1

3.2 性能评估指标

本文采用准确率 (A_{cc})、精确率 (P_{re})、召回率 (R_{ec})、F1 分数 ($F_{measure}$) 和 AUC 作为评价指标, 使用十折交叉验证评估模型。

$$A_{cc} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (15)$$

$$P_{re} = \frac{TP}{(TP + FP)} \quad (16)$$

$$R_{ec} = \frac{TP}{(TP + FN)} \quad (17)$$

$$F_{measure} = \frac{(1 + \beta^2) \cdot P_{re} \cdot R_{ec}}{\beta^2 \cdot P_{re} + R_{ec}} \quad (18)$$

其中, TP 和 TN 分别代表正确分类的关键蛋白质和非关键蛋白质的样本数, 而 FN 和 FP 分别代表错误分类的关键蛋白质和非关键蛋白质的样本数, β 是精确率和召回率之间调整权重的参数。AUC 定义为接收器工作特性 (ROC) 曲线下的面积。精确率和 AUC 是模型最重要的评估指标。通常, AUC 值越大, 分类器的性能越好, 当 AUC 值为 1 时, 模型的性能最佳。

3.3 数据集

采用酿酒酵母的关键蛋白质数据集^[10]评估 IEPHDL 模型和基准模型。该数据集融合了关键蛋白质数据集、基因表达数据集、PPI 网络数据集和亚细胞位置数据集。数据集中涉及 1199 个关键蛋白质。

PPI 网络数据从 BioGRID^[19] 数据库中下载。除去重复的蛋白质相互作用和蛋白质自身相互作用, 该数据集由 5616 个蛋白质和 52833 种相互作用组成。

基因表达数据 GSE3431 从 GEO (Gene Expression Omnibus)^[20] 中抽取, 该数据集共包含 6777 个蛋白质和 36 个时间点, 且具有 3 个连续的代谢周期, 每个周期中有 12 个时间点。

亚细胞位置数据从 COMPARTMENTS^[21] 数据库下载, 它集成了 MGD^[22], SGD^[23], UniProtKB^[24], WormBase^[25] 和 FlyBase^[26] 数据库, 最终获取 3923 个具有亚细胞位置信息的蛋白质。

3.4 实验设计与结果分析

为了评估 IEPHDL 模型的性能, 本文设计了 5 个实验。

实验 1 为了验证 IEPHDL 模型在捕捉 PPI 网络拓扑特征方面的能力, 采用传统基于拓扑特征的中心性方法作为基准方法, 包括 DC、BC、CC、EC、网络中心性 (Network Centrality, NC)^[27] 方法和局部连通中心性 (Local Average Connectivity, LAC)^[28] 方法。

实验 2 为了评价本文 IEPHDL 模型在特征自动学习上的优势, 采用传统机器学习^[10] 的方法作为基准方法。

实验 3 为了验证 IEPHDL 模型优于现有的深度学习模型, 采用 Zeng 模型^[10] 为基准模型。

实验 4 为了验证 IEPHDL 模型的鲁棒性, 对交叉验证进行分析。

实验 5 为了验证本文所使用的 BiGRU 在捕获序列特征方面的优势, 采用 3 种同类神经网络作为基准方法。

实验 1 首先按照每种基于拓扑的中心性方法计算蛋白质得分, 并根据得分将它们降序排列, 然后选择按这些方法排名

最高的 1185 个蛋白质作为待预测的关键蛋白质,其余蛋白质被认为是非关键蛋白质。最后,将其与关键蛋白质和非关键蛋白质的真实标记进行比较,获得了一个混淆矩阵,用以计算每个评价指标。实验结果如图 3 所示。

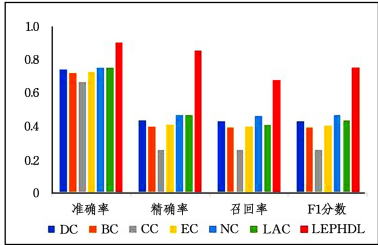


图 3 IEPHDL 和基于拓扑的方法的性能比较

Fig. 3 Performance comparison between IEPHDL and topology-based methods

从图 3 中可以明显看出,通过 IEPHDL 获得的所有评估指标均高于其中任何一种方法的性能指标。原因可能是基于拓扑特征的中心性方法采用的评分函数依赖于大量的先验知识,而 IEPHDL 模型可以自动从 PPI 网络中捕捉关键特征并利用这些特征完成关键蛋白质的识别。

实验 2 采用决策树 (Decision Tree, DT)、随机森林 (Random Forest, RF)、支持向量机 (Support Vector Machine, SVM) 和 Adaboost 作为基准方法,分析 IEPHDL 模型在特征学习上的性能。实验 2 的所有方法都将未通过神经网络层的 3 种数据获取的表示向量进行连接,并将连接的长向量作为输入。实验结果如表 3 所列。可以看出,IEPHDL 模型在所有评估指标上都优于其他方法。这表明深度学习不仅可以自动获取关键蛋白质的特征,也有利于捕捉决定关键蛋白质的主要特征。

表 3 IEPHDL 和传统机器学习方法的性能比较

Table 3 Performance comparison between IEPHDL and traditional machine learning methods

模型	准确率	精确率	召回率	F1 分数	AUC
DT	0.74	0.41	0.32	0.36	0.64
RF	0.80	0.67	0.19	0.29	0.73
SVM	0.79	0.62	0.20	0.29	0.75
Adaboost	0.81	0.63	0.23	0.34	0.76
IEPHDL	0.89	0.86	0.68	0.76	0.85

实验 3 以现有基于深度学习的 Zeng 模型^[10]作为基准模型,验证 IEPHDL 模型的效果。实验将 Zeng 模型的参数赋予 IEPHDL 模型,以确保两个模型具有相同的参数。在相同的环境下,实验结果如表 4 所列。从表 4 可知,IEPHDL 的准确率、精确率、召回率、F1 分数和 AUC 均优于 Zeng 模型。该结果表明,IEPHDL 在关键蛋白质识别研究中具有很好的效果。

表 4 IEPHDL 和 Zeng 模型的性能比较

Table 4 Performance comparison between IEPHDL and Zeng model

模型	准确率	精确率	召回率	F1 分数	AUC
Zeng	0.848	0.721	0.427	0.536	0.831
IEPHDL	0.854	0.793	0.507	0.642	0.835

实验 4 以交叉验证为基准方法,验证 IEPHDL 模型的鲁

棒性。实验对 Zeng 模型和 IEPHDL 模型分别进行变量控制,以交叉验证作为每个模型的唯一变量,对每个模型分别使用五折和十折交叉验证进行评估,实验结果如表 5 所列。从表 5 中可以看到,无论是 Zeng 模型还是 IEPHDL 模型,十折交叉验证均优于其自身的五折交叉验证。这表明,十折交叉验证在一定程度上提高了 IEPHDL 的泛化能力,是提升关键蛋白质预测模型鲁棒性的主要原因。

表 5 交叉验证分析

Table 5 Cross-validation analysis

模型	准确率	精确率	召回率	F1 分数	AUC
Zeng 五折	0.848	0.721	0.427	0.536	0.831
Zeng 十折	0.849	0.746	0.436	0.537	0.833
IEPHDL 五折	0.862	0.820	0.639	0.719	0.842
IEPHDL 十折	0.887	0.862	0.683	0.758	0.852

实验 5 对 IEPHDL 模型中基因表达信息特征处理方法 BiGRU 进行研究,以同类神经网络 LSTM,GRU 和 BiLSTM 为基准方法,与 BiGRU 进行比较。实验在相同参数下(见表 2)控制变量,其中 LSTM 和 GRU 的单元数均为 32,BiLSTM 由 2 个单元数分别为 16 的 LSTM 组成,BiGRU 由 2 个单元数分别为 16 的 GRU 组成,以确保 4 种神经网络类型为唯一变量,表 6 列出了该实验结果。可以看出,在关键蛋白质的基因表达信息特征学习中,IEPHDL 模型所使用的 BiGRU 比前三者表现得更出色,其捕获序列特征的能力明显优于前三者。

表 6 IEPHDL 和同类神经网络性能比较

Table 6 Performance comparison between IEPHDL and similar neural networks

模型	准确率	精确率	召回率	F1 分数	AUC
LSTM	0.832	0.695	0.365	0.503	0.694
GRU	0.834	0.703	0.354	0.517	0.712
BiLSTM	0.856	0.751	0.437	0.543	0.839
IEPHDL	0.887	0.862	0.683	0.758	0.852

结束语 关键蛋白质的识别不仅是了解生物体生存的最低要求,而且对发现人类疾病基因和新的药物靶点也具有重要意义。本文设计了一个关键蛋白质识别模型,该模型自动从基因表达数据、PPI 网络和亚细胞位置信息中学习影响蛋白质重要程度的关键特征,进一步基于特征融合实现对关键蛋白质的预测。本文模型充分利用 BiGRU 从基因表达谱中捕捉序列特征,并与表示 PPI 网络节点拓扑特征的低维稠密特征向量和亚细胞位置特征向量融合。与基于拓扑中心性方法、传统机器学习方法及现有的深度学习模型进行实验比较,结果表明,本文模型有效地提高了关键蛋白质预测的性能。从现有的工作来看,如何设计更有效的深度学习模型以提高模型的鲁棒性是关键蛋白质识别未来研究的方向。在今后的工作中,我们将考虑使用注意力机制进一步增强模型提取特征的能力,并尝试将本文模型应用于其他物种。

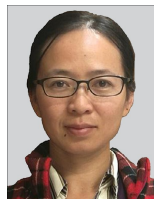
参 考 文 献

[1] LI M,ZHENG R,ZHANG H,et al. Effective identification of essential proteins based on priori knowledge, network topology and gene expressions [J]. Methods,2014,67(3):325-333.

- [2] ZHANG X, ACENCIO M L, LEMKE N. Predicting essential genes and proteins based on machine learning and network topological features: a comprehensive review [J]. *Frontiers in Physiology*, 2016, 7: 75.
- [3] HONG H Y, LIU W. Efficient prediction algorithm of essential proteins based on PPI network [J]. *Computer Science*, 2016, 43(S2): 16-20, 25.
- [4] JEONG H, MASON S P, BARABASI A, et al. Lethality and centrality in protein networks [J]. *Nature*, 2001, 411(6833): 41-42.
- [5] WUCHTY S, STADLER P F. Centers of complex networks [J]. *Journal of Theoretical Biology*, 2003, 223(1): 45-53.
- [6] JOY M P, BROCK A, INGBER D E, et al. High-Betweenness Proteins in the Yeast Protein Interaction Network [J]. *BioMed Research International*, 2005, 2005(2): 96-103.
- [7] HU J, ZHU H W, MAO Y M. Identification of essential proteins based on time-weighted PPI network [J]. *Computer Engineering and Applications*, 2019, 942(23): 155-167.
- [8] TANG X, WANG J, ZHONG J, et al. Predicting essential proteins based on weighted degree centrality [J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2014, 11(2): 407-418.
- [9] LI M, ZHANG H, WANG J, et al. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data [J]. *BMC Systems Biology*, 2012, 6(1): 15-30.
- [10] ZENG M, LI M, FEI Z, et al. A deep learning framework for identifying essential proteins by integrating multiple types of biological information [J]. *IEEE/ACM Trans Comput Biol Bioinform*, 2019, 1(18): 296-305.
- [11] HONG H Y, LIU W. Research on essential Protein Recognition Method Based on Improved PSO Algorithm [J]. *Computer Science*, 2017, 44(10): 38-44.
- [12] FAKOOR R, LADHAK F, NAZI A, et al. Using deep learning to enhance cancer diagnosis and classification [C]// *Proceedings of the International Conference on Machine Learning*. New York, USA: ACM, 2013: 1-6.
- [13] ZENG M, LI M, WU F X, et al. DeepEP: a deep learning framework for identifying essential proteins [J]. *BMC Bioinformatics*, 2019, 20(S16): 506.
- [14] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation [J]. *arXiv*: 1406. 1078, 2014.
- [15] GUO Y, ZHOU D, NIE R, et al. DeepANF: A deep attentive neural framework with distributed representation for chromatin accessibility prediction [J]. *Neurocomputing*, 2020, 379: 305-318.
- [16] LI G, LI M, WANG J, et al. Predicting essential proteins based on subcellular localization, orthology and PPI networks [J]. *BMC Bioinformatics*, 2016, 17(8): 279.
- [17] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet Classification with Deep Convolutional Neural Networks [C]// *Proceedings of the International Conference on Neural Information Processing Systems*. 2012: 1106-1114.
- [18] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization [J]. *arXiv*: 1412. 6980, 2014.
- [19] STARK C, BREITKREUTZ B, REGULY T, et al. BioGRID: a general repository for interaction datasets [J]. *Nucleic Acids Research*, 2006, 34(90001): 535-539.
- [20] TU B P, KUDLICKI A, ROWICKA M, et al. Logic of the Yeast Metabolic Cycle: Temporal Compartmentalization of Cellular Processes [J]. *Science*, 2005, 310(5751): 1152-1158.
- [21] BINDER J X, PLETSCHER-FRANKILD S, TSAFOU K, et al. COMPARTMENTS: unification and visualization of protein subcellular localization evidence [J]. *Database*, 2014, 2014(2): bau012.
- [22] EPPIG J T, BLAKE J A, BULT C J, et al. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse [J]. *Nucleic Acids Research*, 2012, 40: 881-886.
- [23] CHERRY J M, ADLER C, BALL C A, et al. SGD: Saccharomyces Genome Database [J]. *Nucleic Acids Research*, 1998, 26(1): 73-79.
- [24] MAGRANE M. UniProt Knowledgebase: a hub of integrated protein data [J]. *Database*, Volume 2011, 2011(3): bar009.
- [25] HARRIS T W, ANTOSHECHKIN I, BIERI T, et al. WormBase: a comprehensive resource for nematode research [J]. *Nucleic Acids Research*, 2010, 38(2): 463-467.
- [26] MCQUILTON P, PIERRE S E S, THURMOND J. FlyBase 101: the basics of navigating FlyBase [J]. *Nucleic Acids Research*, 2012, 40: 706-714.
- [27] WANG J, LI M, WANG H, et al. Identification of Essential Proteins Based on Edge Clustering Coefficient [J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2012, 9(4): 1070-1080.
- [28] LI M, WANG J, CHEN X, et al. A local average connectivity-based method for identifying essential proteins from the network level [J]. *Computational Biology and Chemistry*, 2011, 35(3): 143-150.



LIU Wen-yang, 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.