基于深度神经网络和局部描述符的大规模蛋白质互作预测方法

桂元苗[1,2] 王儒敬[1,2] 王 雪[1,2,3] 魏圆圆[1*]

[1](中国科学院合肥物质科学研究院智能机械研究所 安徽 合肥 230031)

[2](中国科学技术大学信息技术学院 安徽 合肥 230026)

[3](中国科学院合肥物质科学研究院技术生物与农业工程研究所 安徽 合肥 230031)

摘 要 蛋白质相互作用 PPI(Protein-Protein Interaction)是生物体中众多生命活动过程的重要组成部分,蛋白质互作预测是研究蛋白质互作的重要途径。为了提高蛋白质互作预测性能,构建一个用于预测蛋白质互作的深度神经网络模型 DPPI。采用局部描述符将氨基酸序列编码成具鉴别性的特定维数向量;使用训练集训练 DPPI 模型,并使用测试集对 DPPI 模型进行测试和评价;根据测试和评价的结果调整各参数,优化 DPPI 模型;使用优化后的 DPPI 模型,来对蛋白质互作进行预测。结果表明,DPPI 模型编码简单、代码简洁,实验获得的较高的准确率,可以作为大规模蛋白质互作预测的有益补充。

关键词 深度神经网络 局部描述符 蛋白质互作

中图分类号 TP391 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.04.044

A LARGE-SCALE PREDICTION OF PROTEIN-PROTEIN INTERACTIONS BASED ON DEEP NEURAL NETWORK COMBINED WITH LOCAL DESCRIPTOR

Gui Yuanmiao[1,2] Wang Rujing[1,2] Wang Xue[1,2,3] Wei Yuanyuan[1*]

[1](*Institute of Intelligent Machine, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, Anhui, China*)

[2](*Institute of Information Technology, University of Science and Technology of China, Hefei 230026, Anhui, China*)

[3](*Institute of Technical Biology and Agriculture Engineering, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, Anhui, China*)

Abstract Protein-protein interaction(PPI) plays an important role in many biological processes. PPI prediction is an important way to study protein interaction. In order to improve the prediction performance of protein interaction, a DPPI model was constructed to protein-protein interactions. The local descriptor was used to encode the amino acid sequence into a discriminative specific dimension vector. Then, the training set was used to train DPPI model, and the test set was used for testing and evaluation. We adjusted the parameters according to the results of the test and evaluation to optimize the DPPI model. The optimized DPPI model was used to predict protein interactions. The results find that the DPPI model has simple encoding, simple code, and higher accuracy. It can be used as a useful supplement for large-scale protein interaction prediction.

Keywords Deep neural network Local descriptor Protein-protein interaction(PPI)

0 引 言

蛋白质相互作用 PPI 是生物体中众多生命活动过程的重要组成部分,在许多细胞生物学过程中起着重要的作用。新陈代谢、信号转导、细胞周期调控、新陈代谢、细胞凋亡及免疫应答等一系列生命活动都是通过蛋白质相互作用实现的。蛋白质互作预测是研究蛋白质互作的重要途径。近几年涌现了诸多预测蛋白质相互作用的高通量实验方法,例如:酵母双杂交方

收稿日期:2018-10-19。**桂元苗**,助理研究员,主研领域:机器学习,模式识别。**王儒敬**,研究员。**王雪**,助理研究员。**魏圆圆**,副研究员。

法[1]、质谱蛋白质复合物鉴别[2]、质谱分析[3]以及蛋白质芯片方法[4]等。然而,这些使用化学实验的方法,需要耗费大量的人力、财力和时间,难以应用于大规模的蛋白互作预测。机器学习的出现,使大规模的蛋白互作预测成为可能。到目前为止,已经出现的大量机器学习模型,包括支持向量 SVM[5]、神经网络 NN(Neural Networks)[6]、朴素贝叶斯[7]、K-最近邻[8]等已经被用来预测 PPI。尽管上述 PPI 预测方法很流行,但仍然存在一定的局限性,一般的机器学习模型无法很好地处理蛋白序列噪声输入中的隐性关联[9-11]。深度神经网络的出现,为这类问题提供了强有力的解决方案。

深度神经网络(DNN)是机器学习中最活跃的领域之一,可自动从数据中提取高层抽象信息,用于复杂预测任务,如语音和图像识别[12]、自然语言理解[13]、决策制定[14]以及最近的计算生物学[15-17]。Leung 等[16]利用深度神经网络和 rna-seq 数据,建立了一个预测个体组织和组织间剪接模式差异的模型;Zhou 等[17]使用深度神经网络开发了深度学习框架(DeepSEA http://deepsea.princeton.edu/)。该框架可以从染色质谱分析序列中学习调控序列代码,并且可以改进功能变体的优先级。与基于序列的其他机器学习方法相比,深度神经网络具有几个优点:(1) Bengio 等[18]证明深度神经网络能够减少噪声对原始数据的影响,并学习真正隐藏的高层特征;(2) Alipanahi 等[19]发现,深度神经网络可以使用各种实验数据和评估指标自动学习蛋白质的特定序列基序;(3) Krizhevsky 等[20]人为地将噪声引入基于深度神经网络的方法来减少过度拟合,并且揭示深度神经网络可以增强模型泛化。最近,深度神经网络用于蛋白网络互作也取得了良好的结果[15,21],Tian 等[21]提出了一种称为 DL-CPI(复合蛋白质相互作用预测的深度学习的缩写)的方法。该方法使用深度神经网络通过分层提取来学习复合蛋白对的有用特征,从而在平衡数据集和不平衡数据集上取得了比现有方法更好的预测性能,有效地提高复合蛋白互作的预测性能。Du 等[15]使用深度神经网络基于氨基酸序列来研究蛋白互作预测,并分别获得了 92.50% 的准确率和 90.50% 的召回率。尽管深度神经网络算法在蛋白质互作预测中取得了成功的结果,但基于深度神经网络和局部描述符的蛋白互作预测的研究很少见。

本文首先采用局部描述符将蛋白质序列编码成固定长度的向量,并随机分成训练集和测试集;然后将训练集输入深度神经网络,调整并优化网络结构和学习率、丢弃率等模型参数,训练蛋白互作预测模型 DPPI;最后 DPPI 模型经过测试和验证用于蛋白质互作预测,并将预测的结果和前人的蛋白质互作预测方法进行比较。

1 基础方法

1.1 局部描述符

深度学习等机器学习方法的输入均为某一维数空间中的向量。为使深度学习方法学习并预测蛋白互作关系成为可能,必然要求将长度不统一的蛋白序列编码成某一维数空间中的向量。为了将蛋白质序列编码成维数相同的空间向量,Yang 等[22]首次将局部描述符 LD(Local Descriptor)应用于蛋白互作预测,在酒酿酵母数据集上达到 86.15% 的准确率。LD[23]是一种无需序列比对的方法,其效果在很大程度上取决于潜在的氨基酸分类。首先,依据氨基酸侧链的偶极性和体积将 20 种氨基酸分成 7 组(见表 1),并将蛋白序列中的所有氨基酸替换成对应的分组编码。例如,蛋白序列"MESSKKMDSPGALQTNP"转换成"36335536321124342"。

表 1 基于侧链的偶极子和体积的氨基酸分类

分组	氨基酸种类
1	Ala(A), Gly(G), Val(V)
2	Ile(I), Leu(L), Phe(F), Pro(P)
3	Tyr(Y), Met(M), Thr(T), Ser(S)
4	His(H), Asn(N), Gln(Q), Trp(W)
5	Arg(R), Lys(K)
6	Asp(D), Glu(E)
7	Cys(C)

然后,依据氨基酸官能团在蛋白质初级序列中发生的变化计算 Composition(C)、Transition(T)和 Distribution(D)。其中:Composition 为各组氨基酸在整条蛋白序列中所占的比例;Transition 指一组氨基酸中的氨基酸残基和另外一组氨基酸中的氨基酸残基相邻的频率;Distribution 指在一条蛋白质序列中每组氨基酸的氨基酸残基数目的第一个、25%、50%、75% 和 100% 在整个蛋白质序列中所占位置的比例。所以一个氨基酸片段可以用 63 维的向量表示:7(计算 C 得到的)+21(计算 T 得到的(7×6)/2)+35(计算 D 得到的 7×5)。

为了更好地从蛋白质的氨基酸片段中捕捉蛋白质

相互作用信息,本实验将每条蛋白质序列划分为 10 个局部区域(A－J),见图 1。区域(A－D)是把一条蛋白质序列分成四个相等的区域;区域(E－F)是把一条蛋白质序列分成二个相等的区域;区域 G 表示位于蛋白质序列中间的 50% 氨基酸片段;区域 H 表示整条蛋白质序列的前 75% 的氨基酸片段;区域 I 表示整条蛋白质序列的后 75% 的氨基酸片段;区域 J 表示整条蛋白质序列的中间 75% 的氨基酸片段。一条蛋白质序列的所有局部区域氨基酸片段的编码,串联在一起就形成了一条蛋白质序列的编码,得到 630 维向量。因此,本文构造了一个 1 260 维向量来表示每个蛋白质对,并将其作为 DPPI 模型的输入向量。

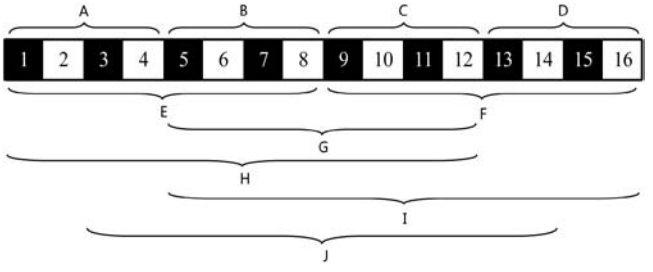


图 1　一条蛋白质划分为 10 个区域的划分方法示意图

1.2　深度神经网络

DNN 是指一组模仿人类大脑设计的,旨在识别模式的算法。DNN 由输入层、一个或多个隐藏层以及输出层三部分组成,如图 2 所示。一般来说,第一层是输入层,最后一层是输出层,而中间的层都是隐藏层。相邻层之间全连接,即第 i 层的任意一个神经元与第 $i+1$ 层的任意一个神经元相连。DNN 类似于一般的人工神经网络,然而,隐藏层的数量和训练过程是不同的。Hinton 等[24]利用预训练方法缓解了局部最优解问题,将隐含层推动到了 7 层后,DNN 才有了真正意义上的“深度”。DNN 在输入层接收数据,在各个节点中将输入数据与权重相结合以非线性方式转换这些数据,通过计算平均梯度并相应地调整权重和激活函数,最后在输出层计算最终输出。

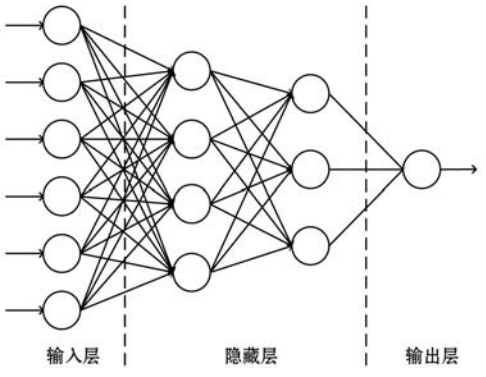


图 2　深度神经网络结构

虽然 DNN 很复杂,但是从局部模型来说,还是和感知机一样,由一个线性关系加上一个非线性激活函数组成,用矩阵法表示,第 l 层的输出为:

$$\boldsymbol{a}^l = \delta(\boldsymbol{Z}^i) = \delta(\boldsymbol{w}^l \boldsymbol{a}^{l-1} + \boldsymbol{b}^l) \tag{1}$$

式中:$l = 1,2,\cdots,N$;a^{l-1} 是第 l 层的输入数据;w^l 是第 $(l-1)$ 层和第 l 层之间的连接权重矩阵;$\boldsymbol{b}^l$ 是第 l 层的偏置,δ 表示第 l 层的激活函数。

目前,在 DNN 中,通常使用 ReLU(Rectified linear unit)作为神经元的激活函数。如式(2)所示,ReLU 具有单侧抑制特性,把所有的负值都变为 0,而正值不变。这种单侧抑制会使神经网络中的神经元具有稀疏激活性,实现稀疏后的模型能够更好地挖掘相关特征和拟合训练数据。

$$\delta(z) = \max(0,z) \tag{2}$$

2　模型构建

2.1　数据集

本实验采用由 Pan 等[30]发布于 http://www.cs-bio.sjtu.edu.cn/bioinf/LR_PPI/Data.htm 的人类蛋白序列对数据集。该数据集包含 36 630 条阳性样本(有互作关系蛋白质序列对)和 36 480 条阴性样本(无互作关系蛋白质序列对)。其中:阳性样本取自人类蛋白质参考数据库(HPRD)2007 版;阴性样本取自瑞士 Swiss-Prot 数据库 57.3 版。

实验侧重 20 种常见氨基酸组成的蛋白质序列,并且蛋白质序列编码方法要求蛋白质序列的长度不易太短。所以,实验过程中除去蛋白质序列长度少于 50 及含有 B、J、O、U、X、Z 的蛋白质序列对,得到 36 591 对阳性样本和 36 324 对阴性样本,分别从阳性样本和阴性样本随机选取 30 000 条蛋白质序列对组成训练集,剩下的 12 915 条蛋白质序列对作为测试集。

2.2　性能评价指标

实验采用准确率(Accuracy)、召回率(Recall)、损失率(Loss)和受试者工作特征曲线下面积 AUC 四个指标来评价模型性能。其中准确率和召回率计算公式如下:

$$A_{\text{ccuracy}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$R_{\text{ecall}} = \frac{TP}{TP + FN} \tag{4}$$

式中:TP、TN、FP 和 FN 分别代表真正、真负、假正和假负。AUC 通过开源代码计算[37]。损失率是用来衡量模型的实际输出与期望输出的距离,损失函数越小,

表示两个概率分布就越接近,模型的拟合性和鲁棒性就越好。损失率通过交叉熵函数计算公式如下:

$$loss(y,y^*) = -\frac{1}{n}\sum_{i=1}^{n}(y_i^*\ln y_i + (1-y_i^*)\ln(1-y_i)) \tag{5}$$

式中:$y=(y_1,y_2,\cdots,y_n)$,为实际输出;$y^*=(y_1^*,y_2^*,\cdots,y_n^*)$为期望输出。

2.3 构建流程

DPPI 模型是基于 Tensorflow 平台在 Python 环境下构建的,其构建流程如图 3 所示。主要包括以下几步:首先,使用 LD 分别对有关序列对和无关序列对的氨基酸序列进行编码,生成有关序列集和无关序列集;其次,使用随机选择的 60 000 条训练集数据对模型进行训练,生成 DPPI 模型;接着,使用剩下的 12 915 条测试集数据对 DPPI 模型进行测试;最后,对 DPPI 模型的预测性能进行评价,并根据评价结果调整参数,优化 DPPI 模型。

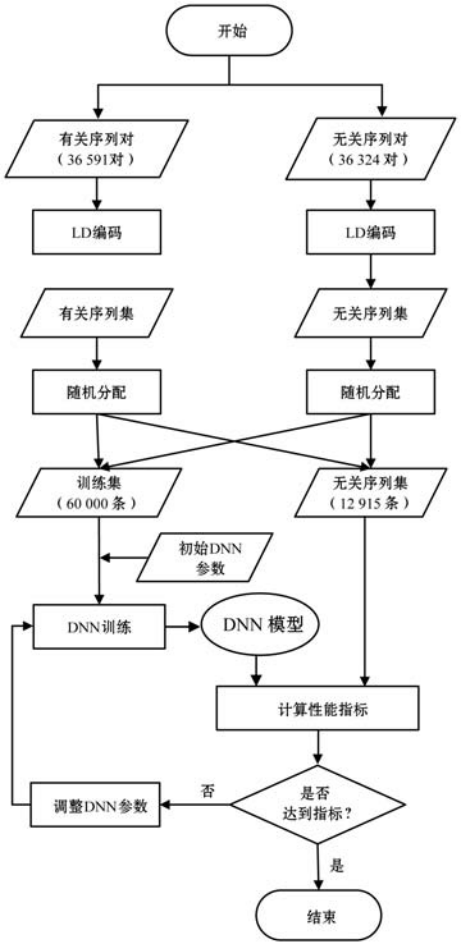


图 3 DPPI 模型构建流程图

2.4 参数调整

参数调整是模型训练过程中很重要的一步,是训练出健壮模型的关键要素之一。实验中,激活函数使用 ReLU、优化器选择 Adam、代价函数使用交叉熵。在优化器方面,目前已经开发了诸如 RMSprop, Adagrad[27] 和 Adam[28] 等优化方法,其中,Adam 集合了 RMSprop 和 Adagrad 两个算法的优点,能够较好地处理噪声样本。交叉熵代价函数是用来衡量深度神经网络的预测值与实际值的一种方式,可以弥补 sigmoid 型函数的导数形式易发生饱和的缺陷,使训练更快收敛。

学习率决定了权值更新的速度,设置得太大易越过最优值,出现振荡现象;太小会使下降速度过慢,长时间无法收敛。因此,学习率直接决定着算法的性能表现。Bengio[29] 认为一般常用的学习率有 0.000 01、0.000 1、0.001、0.01、0.1,推荐使用的是 0.01,同时 Bengio 也指出,学习率的选择要根据数据集的大小、特征提取方法等实际情况来确定。实验中,设定隐含层节点数为 64、激活函数为 ReLU、优化算法为 Adam、批处理大小为 128、迭代次数为 300 000 次,调整学习率的结果如表 2 所示。可以看出,学习率为 0.001 时的准确率最高、平均损失最小。

表 2 DPPI 模型学习率的调整

学习率	准确率/%	平均损失/%
0.01	83.23 ±2.05	37.87 ±3.49
0.001	92.23 ±3.19	21.75 ±5.42
0.000 1	89.55 ±5.51	26.17 ±9.82
0.000 01	78.05 ±4.75	46.39 ±6.61

为了确定模型的宽度,设定激活函数为 ReLU、优化算法为 Adam、学习率为 0.001、批处理大小为 128、迭代次数为 300 000 次,调整模型宽度的结果如表 3 所示。可以看出,宽度为 512 时,模型的准确率最高,同时训练时间比宽度为 256 时增加了近一倍,而准确率、平均损失分别比宽度为 256 时仅仅提高了 0.003 1、0.001 3,考虑到时间复杂度,本模型宽度选择 256。

表 3 DPPI 模型宽度的调整

宽度	准确率/%	平均损失/%	100 步训练时间/s
64	92.23 ±3.19	21.75 ±5.42	0.629 2 ±0.102 4
128	93.90 ±2.58	19.14 ±4.49	0.847 7 ±0.106 7
256	94.52 ±2.48	19.64 ±4.36	1.223 0 ±0.109 8
512	94.83 ±2.06	19.51 ±3.57	2.282 8 ±0.997 8

模型的宽度、激活函数、优化算法、学习率等确定之后,本文通过调整隐含层层数来确定模型的深度,调整深度的结果如表 4 所示。根据表 4,可知网络深度为[256 -128 -64 -32]时的准确率较高、平均损失较低,训练时间较短。

表 4　DPPI 模型深度的调整

深度	准确率/%	平均损失/%	100 步训练时间/s
256 – 128 – 64 – 32	95.74 ± 1.05	22.67 ± 4.18	1.397 ± 0.120 3
256 – 128 – 64	95.46 ± 1.39	24.36 ± 4.55	1.398 7 ± 0.121 5

丢弃率是 DNN 中防止过拟合、提高性能的一个很重要的参数。为了优化 DPPI 模型,本文通过调整丢弃率得到 7 个预测模型,各预测模型的最优结果如表 5 所示。从表 5 可知,丢弃率为 0.025 时准确率最高,达到 96.81%,平均损失为 15.72%;丢弃率为 0.05 时,准确率比丢弃率为 0.025 时的准确率降低了 0.08%,同时平均损失降低了 2.52%;不使用丢弃率时,准确率、AUC、Recall、平均损失分别为 95.84%、97.44%、98.25%、45.38%。虽然不使用丢弃率时准确率和使用丢弃率时准确率差别不大,但不使用丢弃率的平均损失较高,所以不推荐使用。由此,DPPI-2 和 DPPI-3 的性能较好,可以作为 DPPI 的最终预测模型。

表 5　DPPI 模型的最优预测性能

模型	准确率/%	AUC/%	召回率/%	平均损失/%	训练迭代次数	丢弃率/%
DPPI-1	95.84	97.44	98.25	45.38	570 000	N/A
DPPI-2	96.81	98.74	99.06	15.72	600 000	2.5
DPPI-3	96.73	99.00	99.21	13.20	440 000	5.0
DPPI-4	96.41	98.92	99.03	15.26	560 000	7.5
DPPI-5	93.45	97.83	98.17	18.02	380 000	10
DPPI-6	82.61	91.18	92.04	39.92	30 000	30
DPPI-7	81.81	90.84	91.78	44.80	30 000	50

经过超参数调整后,本文构建了一个包含 4 个隐藏层,各隐藏层节点数分别为 256、128、64、32 的 DPPI 模型。经过大量实验和调试,总结了本实验采用的参数,如表 6 所示。

表 6　DPPI 模型参数表

参数名称	范围	参数取值
深度	2,3,4,5,6	4
宽度	128,256,1 024,2 048	256 – 128 – 64 – 32
激活函数	ReLU,tanh,sigmoid,softmax,ELU	ReLU
优化算法	Adadelta,RMSprop,Adam	Adam
学习率	0.01,0.001,0.000 1,0.000 01	0.001
代价函数	cross-entropy	cross-entropy
丢弃率	0.025,0.05,0.075,0.1,0.3,0.5	0.025,0.05

3　实　验

3.1　DPPI 模型实验

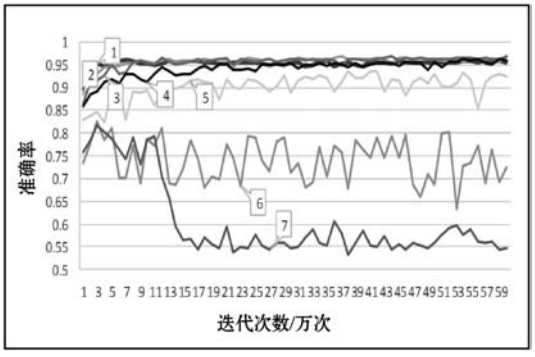
参照表 6 的 DPPI 模型参数,根据使用和不使用丢弃率以及不同的丢弃率值,使用 7 个不同的 DPPI 模型进行计算。每个模型各迭代 60 万次,每 1 万次输出一个测试结果,各输出 60 个实验结果。所有实验结果的统计值见表 7。可以看到,DPPI-2 的准确率最高,DPPI-1 和 DPPI-3 稍微次之。DPPI-3 的 AUC、召回率和平均损失这三个指标比其他 6 组表现更优异。特别地,DPPI-3 的平均损失(15.34%)比 DPPI-2 的平均损失(16.99%)降低了 1.65%。

表 7　DPPI 模型预测性能平均值

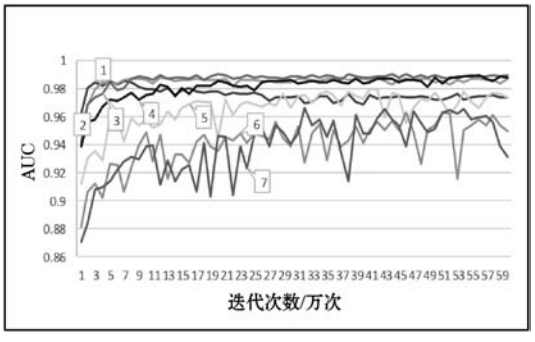
模型	准确率/%	AUC/%	召回率/%	平均损失/%	丢弃率/%
DPPI-1	95.30 ± 0.87	97.57 ± 0.43	98.28 ± 0.30	36.23 ± 9.24	N/A
DPPI-2	95.77 ± 1.30	98.50 ± 0.490	98.87 ± 0.46	16.99 ± 2.34	2.5
DPPI-3	95.60 ± 1.67	98.65 ± 0.71	98.89 ± 0.65	15.34 ± 2.98	5.0
DPPI-4	94.16 ± 1.96	98.13 ± 0.88	98.37 ± 0.78	17.79 ± 3.37	7.5
DPPI-5	90.20 ± 2.74	96.53 ± 1.35	97.01 ± 1.17	24.28 ± 4.73	10
DPPI-6	74.06 ± 4.59	94.07 ± 1.67	94.89 ± 1.51	53.81 ± 11.49	30
DPPI-7	60.58 ± 8.71	93.92 ± 2.14	94.74 ± 1.86	76.62 ± 19.31	50

图 4 为 7 个 DPPI 预测模型不同评价指标的趋势图。其中:1 代表 DPPI-1 预测指标趋势图;2 代表 DPPI-2 预测指标趋势图;3 代表 DPPI-3 预测指标趋势图;4 代表 DPPI-4 预测指标趋势图;5 代表 DPPI-5 预测指标趋势图;6 代表 DPPI-6 预测指标趋势图;7 代表 DPPI-7 预测指标趋势图。从图 4(a)中可以看出,模型 DPPI-6 和 DPPI-7 随着迭代次数的增加,准确率呈下降趋势。DPPI-5 的准确率虽然比 DPPI-6 和 DPPI-3 好,但 DPPI-5 准确率不稳定,振荡幅度稍大。从准确率来看,模型 DPPI-1、DPPI-2、DPPI-3 和 DPPI-4 准确率后期都比较稳定,DPPI-2 的准确率最好。(b)是 7 个 DPPI 模型 AUC 趋势图,可以看出,模型 DPPI-2、DPPI-3 和 DPPI-4 后期结果比较平稳,虽然 DPPI-1 的结果开始较好,但后期较差。(c)是 7 个模型的召回率趋势图,

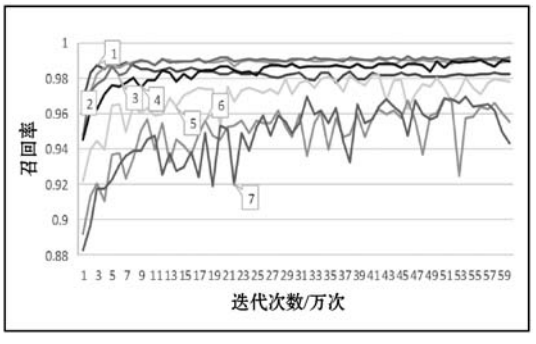
DPPI-3 的召回率性能较好，DPPI-5、DPPI-6 和 DPPI-7 召回率较低，不平稳且振荡幅度较大。(d)是 7 个模型平均损失的趋势图，趋势图显示模型 DPPI-2、DPPI-3 和 DPPI-4 结果较好，但 DPPI-2 振荡幅度比 DPPI-3 稍大。



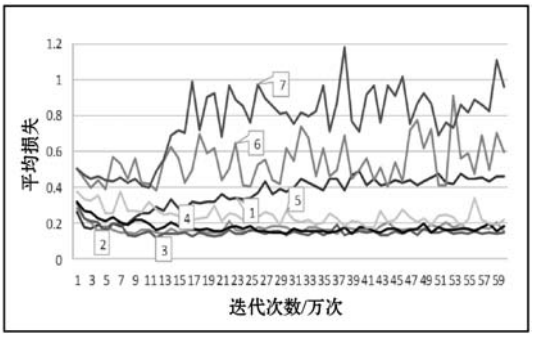
(a)



(b)



(c)



(d)

图 4 DPPI 模型各个指标预测趋势图

综合表 7 和图 4 得到最终的预测模型 DPPI-3，其准确率为 95.6%，平均损失为 15.34%。

3.2 方法比较

近几年，已经有许多研究者对人蛋白互作预测提出了不同的计算方法。这些人蛋白互作预测方法以及 DPPI 的性能比较结果见表 8。可以看出，所列方法获得的精度均在 83.90% 和 97.19% 之间，同时，除了 Sun[34] 使用 SAE + AC 方法获得 97.19% 的准确率以外，DPPI 模型获得了最好的准确率。和 Sun 的结果相比，虽然 DPPI 的准确率不算突出，但是 Sun 使用 SAE + CT 的准确率没有 DPPI 模型的准确率高。从表 9 可见，在酒酿酵母数据集上，DPPI 模型和其他采用 LD 编码的方法比较，也取得最高准确率。

表 8 不同方法的预测性能的比较

参照	方法	准确率	数据集划分方式
Shen 等研究结果[30]	SVM	83.90%	Holdout
You 等研究结果[31]	ELM	84.80%	Holdout
Guo 等研究结果[32]	SVM	90.67%	Holdout
Du 等研究结果[15]	DNN	89.00%	Holdout
Pan 等研究结果[33]	PseAAC + SVM	91.20%	五倍交叉
Sun 等研究结果[34]	SAE + CT	94.52%	十倍交叉
Sun 等研究结果[34]	SAE + AC	97.19%	十倍交叉
DPPI	DNN + LD	95.60%	Holdout

表 9 不同算法采用 LD 编码方式结果比较

参照	方法	数据集	准确率
You 等研究结果[35]	ELM + LD	酒酿酵母	83.90%
Zhou 等研究结果[36]	SVM + LD	酒酿酵母	88.76%
Yang 等研究结果[37]	kNN + LD	酒酿酵母	84.81%
Wang 等研究结果[38]	DNN + LD	酒酿酵母	90.19%

SAE + AC 结果优于 DPPI 的原因可能在于特征提取方法的不同。AC 编码是通过选择物理化学性质，解释了氨基酸与序列中相隔一定数量的氨基酸之间的相互作用，该方法考虑了最长序列 30 bp 的邻近效应[32]。LD 为了更好地从蛋白质的氨基酸片段中捕捉蛋白质相互作用信息，将一条蛋白质序列划分为 10 个局部区域，这样分组，局部信息突出不明显，致使丢失某些关键信息[37]。LD 的这种缺陷在以后的研究中，可以通过增加局部区域的划分等方法来减少特征信息的丢失。

虽然 DPPI 模型准确率和 SAE + AC 方法相比不算突出，但在蛋白互作预测方面也取得了良好的结果，且

DNN 去噪能力优于 SAE,代码也比 SAE 简洁,LD 编码简单、速度快。特别地,由表 10 可知,LD 和 AC 编码 72 915 对人蛋白质的时间可见在相同软硬件计算环境下 LD 编码速度比 AC 编码速度快 3.5 倍以上。通过上面的比较可知,本文的 DPPI 模型与其他方法相比可以显著提高大规模蛋白互作预测性能。

表 10　AC 和 LD 编码时间的比较

编码方式	编码时间/s	数据集
AC	0.028 65	HPRD(36591) + Swiss - Port(36324)
LD	0.007 97	HPRD(36591) + Swiss - Port(36324)

4　结　语

深度学习算法已经涉足许多领域,但是在蛋白互作的研究中还没有被广泛的应用。因此,本文采用深度神经网络 DNN 和 LD 蛋白质序列编码方法相结合的方法构建了蛋白互作预测模型 DPPI。DPPI 模型获得准确率 96.73%、AUC 99.00%、召回率 99.21% 和平均损失 13.2% 的最优性能,以及 95.60% 准确率、98.65% AUC、98.89% 召回率和 15.34% 平均损失的平均性能。和其他研究者提出的人蛋白互作预测方法比较,DPPI 模型的准确率优于 Shen、You、Guo、Du、Pan 等结果,但 DPPI 的结果没有 Sun 采用的 SAE + AC 方法预测结果性能好。LD 的这种缺陷在以后的研究中,可以通过增加局部区域等方法来减少特征信息的丢失。

本文首次采用 DNN 结合 LD 对人蛋白互作数据集构建的,用于蛋白互作预测的模型 DPPI。该模型具有较强的去噪能力、编码简单、代码简洁、计算速度快、运行时间段等优点,可以通过分层抽象学习蛋白质对的有用特征,从数据中自动学习内部分布式特征表示。鉴于以上优点,DPPI 模型可以作为蛋白互作预测的有益补充。

参　考　文　献

[1] Fields S, Song O. A Novel Genetic System to Detect Protein Protein Interactions[J]. Nature, 1989, 340(6230):245-246.

[2] Ho Y, Gruhler A, Herlbut A, et al. Systematic Identification of Protein Complexes in Saccharomyces Cerevisiae by Mass Spectrometry[J]. Nature,2002,415(6868):180-183.

[3] Gavin A C, Bosche M, Krause R, et al. Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes[J]. Nature, 2002, 415(6868): 141-147.

[4] Heng Z, Metin B, Rhonda B, et al. Global Analysis of Protein Activities Using Proteome Chips[J]. Science, 2001, 293(5537):2101-2105.

[5] Chatterjee P, Basu S, Kundu M, et al. PPI_SVM: Prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables[J]. Cellular & Molecular Biology Letters, 2011, 16(2):264-278.

[6] Fariselli P, Pazos F, Valencia A, et al. Prediction of protein—protein interaction sites in heterocomplexes with neural networks[J]. European Journal of Biochemistry, 2002, 269(5): 1356-1361.

[7] Lin X, Chen X W. Heterogeneous Data Integration by Tree Augmented Naive Bayes for Protein-Protein Interactions Prediction[J]. Proteomics, 2013, 13(2): 261-268.

[8] Browne F, Wang H, Zheng H, et al. Supervised Statistical and Machine Learning Approaches to Inferring Pairwise and Module Based Protein Interaction Networks[C]//Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, New York, USA: IEEE, 2007: 1365-1369.

[9] Valente G T, Acencio M L, Martins C, et al. The Development of a Universal in Silico Predictor of Protein-Protein Interactions[J]. PLoS One, 2013, 8(5): e65587.

[10] Chen X W, Liu M. Prediction of Protein-Protein Interactions Using Random Decision Forest Framework[J]. Bioinformatics, 2005, 21(24): 4394-4400.

[11] Saha I, Zubek J, Klingström T, et al. Ensemble Learning Prediction of Protein-Protein Interactions Using Proteins Functional Annotations[J]. Molecular Biosystems, 2014, 10(4): 820-830.

[12] Mohanty S P, Hughes D P, Salathé M. Using Deep Learning for Image-Based Plant Disease Detection[J]. Frontiers in Plant Science, 2016, 7:1419.

[13] Collobert G, Weston J. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning[C]//International Conference on Machine Learning, 2008:160-167.

[14] Silver D, Huang A, Maddison C J, et al. Mastering the Game of Go with Deep Neural Networks and Tree Search [J]. Nature, 2016, 529(7587):484-489.

[15] Du X Q, Sun S W, Hu CL, et al. DeepPPI: Boosting Prediction of Protein-Protein Interactions with Deep Neural Networks[J]. Journal of Chemical Information & Modeling, 2017, 57(6):1499-1510.

[16] Leung M K, Xiong H Y, Lee L J, et al. Deep Learning of the Tissue-Regulated Splicing Code[J]. Bioinformatics, 2014, 30(12): i121-i129.

[17] Zhou J, Troyanskaya O G. Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model[J]. Nature Methods, 2015, 12(10): 931 - 934.

[18] Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 35(8): 1798 - 1828.

[19] Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA—and RNA-binding proteins by deep learning[J]. Nature Biotechnology, 2015, 33(8): 831 - 838.

[20] Krizhevsky A, Sutskever I, Hinton G E. Imagenet Classification with Deep Convolutional Neural Networks[J]. International Conference on Neural Information Processing Systems, 2012, 60(2): 1097 - 1105.

[21] Tian K, Shao M Y, Wang Y, et al. Boosting Compound-Protein Interaction Prediction by Deep Learning[J]. Methods, 2016, 110: 64 - 72.

[22] Yang L, Xia J F, Gui F. Prediction of Protein-Protein Interactions from Protein Sequence Using Local Descriptors[J]. Protein & Peptide Letters, 2010, 17(19):1085 - 1090.

[23] Davies M N, Secker A, Freitas A A, et al. Optimizing Amino Acid Groupings for GPCR Classification[J]. Bioinformatics, 2008, 24(18):1980 - 1986.

[24] Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks[J]. Science, 2006, 313(5786): 504 - 507.

[25] Hinton G E, Osindero S, Teh Y W. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2014, 18(7): 1527 - 1554.

[26] Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks[C]//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS), 2010: 315 - 323.

[27] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A Simple Way to Prevent Neural Networks from Over Fitting. Journal of Machine Learning Research, 2014, 15(1): 1929 - 1958.

[28] Kingma D, Ba J. Adam: A Method for Stochastic Optimization[DB]. arXiv:1412.6980V8, 2014.

[29] Bengio Y. Practical Recommendations for Gradient-Based Training of Deep Architectures[J]. Springer Berlin Heidelberg, 2012, 7700(1 - 3): 437 - 478.

[30] Shen J, Zhang J, Luo X, et al. Predicting protein-protein interactions based only on sequences information[J]. Proceedings of the National Academy of Sciences, 2007, 104(11):4337 - 4341.

[31] You Z H, Li S, Gao X, et al. Large-Scale Protein-Protein Interactions Detection by Integrating Big Biosensing Data with Computational Model[J]. Biomed Research International, 2014, 2014(2):598129.

[32] Guo Y, Li M, Pu X, et al. PRED_PPI: A Server for Predicting Protein-Protein Interactions Based on Sequence Data with Probability Assignment[J]. Bmc Research Notes, 2010, 3(1): 145 - 152.

[33] Pan X Y, Zhang Y N, Shen H B. Large-scale Prediction of Human Protein-Protein Interactions from Amino Acid Sequence Based on Latent Topic Features[J]. Journal of Proteome Research 2010, 9(10):4992 - 5001.

[34] Sun T, Zhou B, Lai H, et al. Sequence-Based Prediction of Protein Protein Interaction Using a Deep-Learning Algorithm[J]. BMC Bioinformatics, 2017, 18(1): 277 - 285.

[35] You Z H, Lei Y K, Zhu L, et al. Prediction of Protein-Protein Interactions from Amino Acid Sequences with Ensemble Extreme Learning Machines and Principal Component Analysis[J]. BMC Bioinformatics, 2013, 14(S8): S10.

[36] Zhou Y Z, Gao Y, Zheng Y Y. Prediction of Protein-Protein Interactions Using Local Description of Amino Acid Sequence[J]. Communications in Computer & Information Science, 2011, 202:254 - 262.

[37] Yang L, Xia J F, Gui J. Prediction of Protein-Protein Interactions from Protein Sequence Using Local Descriptors[J]. Protein & Peptide Letters, 2010, 17(9):1085 - 1090.

[38] Wang J, Zhang L, Jia L Y, et al. Protein-Protein Interactions Prediction Using a Novel Local Conjoint Triad Descriptor of Amino Acid Sequences[J]. International Journal of Molecular Sciences, 2017, 18(11):2373.

(上接第 272 页)

[12] Zhang G, Liu X, Yang Y. Time-series pattern based effective noise generation for privacy protection on cloud[J]. IEEE Transactions on Computers,2015,64(5):1456 - 1469.

[13] Yang J J, Li J Q, Niu Y. A hybrid solution for privacy preserving medical data sharing in the cloud environment[J]. Future Generation Computer Systems, 2015, 43: 74 - 86.

[14] Kung S Y. Discriminant component analysis for privacy protection and visualization of big data[J]. Multimedia Tools and Applications, 2017, 76(3): 3999 - 4034.

[15] 邓劲松, 罗永龙, 俞庆英, 等. 基于非敏感信息分析的轨迹数据隐私保护发布[J]. 计算机应用, 2017, 37(2): 488 - 493.

[16] Chen R, Fung B C M, Mohammed N, et al. Privacy-preserving trajectory data publishing by local suppression[J]. Information Sciences, 2013, 231: 83 - 97.