

基于人工智能的药物-靶标相互作用预测

杨朔, 王洁, 张梦婷, 沈子豪, 李洪林, 李诗良* (华东理工大学新药设计重点实验室, 生物反应器工程国家重点实验室, 华东理工大学药学院, 上海 200237)

摘要: **目的** 建立一个高效的药物-靶标相互作用预测分类模型, 为生物实验提供有力的补充工具。**方法** 研究开发一种基于深度学习的方法来预测药物-靶标相互作用: 通过引入高维分子指纹和蛋白质描述符, 并应用概率矩阵分解算法生成负样本集, 构建一个高效的药物-靶标相互作用预测分类模型。**结果** 与其他已报道的方法相比, 本方法具有可比性或优越性, 预测准确性、特异性、敏感性以及 AUC 值均>90%, 提示该方法在药物靶标预测方面具有良好的应用前景。**结论** 人工智能深度学习模型以及概率矩阵分解算法的结合有助于解决药物-靶标相互作用预测精度低、负样本选择不合理等问题。

关键词: 药物-靶标相互作用; 深度学习; 概率矩阵分解算法; 靶标预测

中图分类号: R966 文献标志码: B 文章编号: 1007-7693(2022)21-2797-07

DOI: 10.13748/j.cnki.issn1007-7693.2022.21.015

引用本文: 杨朔, 王洁, 张梦婷, 等. 基于人工智能的药物-靶标相互作用预测[J]. 中国现代应用药学, 2022, 39(21): 2797-2803.

Drug-target Interaction Prediction with Artificial Intelligence

YANG Shuo, WANG Jie, ZHANG Mengting, SHEN Zihao, LI Honglin, LI Shiliang* (Shanghai Key Laboratory of New Drug Design, State Key Laboratory of Bioreactor Engineering, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China)

ABSTRACT: OBJECTIVE To build an efficient drug-target prediction classification model and provide a useful complementary tool for biological experiments. **METHODS** In this study, a deep learning-based method was developed to predict drug target interaction. By introducing high dimensional molecular fingerprints and protein descriptors, and subsequently applying a probability matrix decomposition algorithm to generate negative samples, a promising drug target interaction classification model was constructed. **RESULTS** The method was comparable or superior to previously developed methods against the test sets, achieving >90% accuracy, specificity, sensitivity, and AUC. This method represented a promising tool for drug target prediction. **CONCLUSION** The combination of artificial intelligence deep learning model and probabilistic matrix factorization algorithm can help to solve the problems of low prediction accuracy of drug-target interaction and unreasonable selection of negative samples.

KEYWORDS: drug-target interaction; deep learning; probabilistic matrix factorization; target prediction

随着计算机性能的快速提高以及计算化学和计算生物学等跨学科领域的发展, 计算机辅助药物设计已逐渐应用于药物发现的各个阶段^[1]。这种方法加快了药物研发过程, 显著降低了研究风险、开发成本和临床试验失败。随着语音识别、计算机视觉和自然语言处理的突破, 人工智能作为一种强大的数据挖掘技术, 已广泛应用于药物设计领域^[2]。众所周知, 科学家将人工智能和大数据结合, 在各个领域都有着重大的突破, 其中相关的工作包括从蛋白质序列预测蛋白质结构^[3]、预测蛋白质-蛋白质和药物-蛋白质相互作用^[4]、药物再利用^[5]、从头设计的生成模型^[6]以及物理化学性质和

ADME/T 值的预测^[7]。

识别药物-靶点之间的相互作用对于药物发现和药物重定位起到非常关键的作用。作为实验方法的补充工具, 多种计算方法可用于预测新的药物-靶点相互作用。传统方法可分为基于配体或基于结构的方法^[8]。首先, 基于结构的方法通常采用分子对接方法来解析靶标的三维结构, 以筛选小分子。但是当靶标的三维结构未知时, 如 G 蛋白偶联受体, 这些方法的应用会受到极大的限制。其次, 基于配体的方法是基于结构方法的有效补充, 该方法采用的原则是相似的药物更有可能靶向于与相似药物相结合的靶标蛋白。这些相似性

作者简介: 杨朔, 男, 硕士 E-mail: syangs@yeah.net

*通信作者: 李诗良, 女, 副教授, 硕导

E-mail: shiliangli@ecust.edu.cn

通常通过药物的化学结构、药物的不良反应及其蛋白质靶序列来量化。但是,这些方法的预测质量在很大程度上取决于如何量化相似性以及如何更多地相似性组合在一起。基于网络的方法也已被用于推断药物-靶标相互作用(drug-target interaction, DTI)^[9]。基于网络推理的概念是根据药物-靶标双边网络拓扑的相似性,并使用评分函数进行分析,以评估药物-靶标之间的关系。然而,基于网络推理不能应用于新开发的药物,因为这些药物在训练集中没有明确的靶标信息。随着许多新的实验仪器和技术的发展,包括高通量筛选实验和下一代测序技术,迫切需要整合多种资源来收集有关药物靶点的相关信息。异构数据为预测新型药物-靶点相互作用提供了广阔的前景,结合异构网络数据可能提高药物靶点预测的准确性,并为药物重定位提供了新的方向。Luo 等^[10]开发了 DTINet,是一种用于药物-靶点结合预测的新型网络集成方法,集成了不同的数据(包括药物、蛋白质、疾病和不良反应),并应用了降维算法。然而,将特征学习与预测任务分离可能不会产生最佳解决方案,因为从无监督学习过程中学习的特征可能并不是最合适的药物或靶标特征的特征。

近年来,深度学习算法在预测药物-靶点相互作用方面显示出比传统机器学习算法具有更强大的表现^[11]。Wen 等^[12]提出了用于预测药物-靶点相互作用的基于深度学习的模型 DeepDTIs。该研究首先通过无监督预训练生成特征,然后将有监督深度学习算法应用于已知药物-靶点相互作用。作者表明,与其他基线模型相比(包括伯努利贝叶斯方法、决策树和随机森林法),DeepDTIs 模型具有最好的性能。然而,量化蛋白质与配体的相互作用仍然是一个挑战。Öztürk 等^[13]提出了一种基于深度学习的 DeepDTA 模型,该模型仅使用靶点和药物的序列信息来预测药物-靶点相互作用结合亲和力和。他们证明,使用靶标和药物的一维表示是预测药物-靶标结合亲和力的有效方法。该模型通过卷积神经网络构建药物的高维表征,并使用 Smith-Waterman 相似度来表示达到最佳一致性指数性能的蛋白质。

本研究开发了一种新型的基于深度学习的小分子药物-靶标相互作用预测方法(deep learning-based method, DLDTIs),其创新点总结如下:

①将推荐算法中常用的概率矩阵分解算法(probabilistic matrix factorization, PMF)用于选择负样本,有效地降低了阴性样本选择中选择潜在结合的药物-靶标对的概率。

②单一的深度神经网络不适用于药物-靶标相互作用预测,本研究验证了集成网络具有优于单一网络的预测能力。

③在正样本数量有限的情况下,验证了负样本数量对模型预测结果的影响。

1 材料

1.1 数据集准备

训练数据集由 Wen 等^[12]提供,其中有关药物和靶点的信息主要来源于 DrugBank 数据库^[14]。数据集包括药物和靶点、药物结构和靶点序列的结合信息,其中无机化合物和分子量较小的分子结构被去除,覆盖 1 412 种药物、1 520 个靶点和 6 262 个已知药物-靶点结合对。数据集也可在 https://github.com/Bjoux2/DeepDTIs_DBN 下载。Wen 等通过随机选择挑选了 6 262 个负样本,而本研究新发展的方法创新点在于使用 PMF 选择 6 262 个负样本,而不是随机选择。最后,整合阳性样本(6 262)和阴性样本(6 262),得到包含 12 524 个样本的训练集。其中,阳性样本表示药物可与靶标相互作用;相反,阴性样本表明药物不能与靶标相互作用。整个数据集分为 3 个子数据集,包括训练集、验证集和测试集,相应比率分别为 0.6(7 514 个样本)、0.2(2 505 个样本)和 0.2(2 505 个样本)。

1.2 化学结构与蛋白质序列表征

本研究采用了 2 种主流的化合物分子指纹,即扩展连接性指纹^[15]和分子访问系统指纹^[16],共包括 6 311 个特征。

对于蛋白质序列表征,从反映蛋白质二级结构的可变长度蛋白质序列中导出固定长度特征,并应用广泛使用的预测蛋白质-蛋白质相互作用的方法计算物理化学属性。共包含 967 个特征值,具体的蛋白质计算的描述符如下。

①氨基酸组成:蛋白质中每种氨基酸的比例。氨基酸组成可以计算多达 20 个特征。其中 N_r 是 r 型氨基酸的数目, N 是序列的长度。

$$f_{r(r)} = \frac{N_r}{N} \quad r=1,2,3,\dots,20。$$

②二肽组成:蛋白质中每种二肽的比例。二肽组成可以计算 400 个固定特征。其中 N 是由 r

和 s 型氨基酸表示的二肽的数量。

$$f_{r(s)} = \frac{N_{(r,s)}}{N} \quad r, s = 1, 2, 3, \dots, 20。$$

③组成、转换和分布：根据每个氨基酸的属性，氨基酸主要可分为三类。这些属性包括疏水性、标准范德华体积以及极性和极化率。因此，组成表示序列中每个编码类的总百分比。转换反映了编码序列中不同编码类的转换比例。每个属性的分布描述符是指定编码类的第一个残基、25%残基、50%残基、75%残基和 100%残基在整个序列中的位置百分比。组成、转换和分布描述符共包括 147 个特征。

$$\text{Composition} = \frac{N_r}{N} \quad r = 1, 2, 3。$$

其中 N_r 是编码序列中 r 的数目， N 是序列的长度。

$$\text{Transition} = \frac{N_{rs} + N_{sr}}{N}。$$

其中 N_{rs} 和 N_{sr} 是序列中分别编码为 rs 和 sr 的二肽的数量， N 是序列的长度。

④准序列顺序描述符：这些描述符清楚地反映了总共 160 个特征的具体物理化学性质的氨基酸分布情况。第 d 秩序列顺序耦合数定义为

$$\tau_d = \sum_{i=1}^{N-d} (d_{i,i+d})^2 \quad d = 1, 2, 3, \dots, \text{maxlag}。$$

公式中 $d_{i,i+d}$ 是位置 i 和 $i+d$ 处 2 个氨基酸之间的距离。 maxlag 是最大 lag ，蛋白质的长度不得 $< \text{maxlag}$ 。在实验中， maxlag 等于 30。对于每种氨基酸类型，I 型准序列顺序描述符可定义为

$$X_r = \frac{f_r}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{\text{maxlag}} \tau_d} \quad r = 1, 2, 3, \dots, 20。$$

其中， w 是 I 型氨基酸的标准化出现率，是加权因子 ($w=0.1$)。II 类准序列顺序描述符定义为

$$X_d = \frac{w \tau_{d-20}}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{\text{maxlag}} \tau_d} \quad r = 21, 22, \dots, 20 + \text{maxlag}。$$

⑤两性性伪氨基酸组成：也称为 2 型伪氨基酸组成，最初由 Chou 等^[17]根据 I 型氨基酸组成开发。本研究总共使用了 240 个特征。

1.3 阴性样本的选取

PMF 模型被常用于推荐系统的应用之中，同时得到了广泛的关注，它属于一种矩阵的分解方法，因为 PMF 从生成概率的方向出发，解释了 X 和 Y 2 个维度的潜在联系。这种方法和奇异值分解法相比

存在一个优势，就是不需要考虑最初的矩阵信息补全的问题，而是将 X 和 Y 维度的信息都表征在一个 P 维度的隐藏特征空间之中，这样极大程度上降低了计算的复杂度。

所以，PMF 是在奇异值分解算法上引入了概率的分布^[18]，假设了隐藏特征空间服从高斯先验分布，还假设了已经收集到的打分数据当中的条件概率也服从高斯先验分布，再利用贝叶斯的推导求得隐藏特征的后验概率。最后，通过梯度下降的算法学习到隐藏特征的向量就可以计算出矩阵中缺失的值，可以很好地实现推荐过程。

PMF 应用到预测药物-靶标相互作用的工作当中，因为可以通过已知的药物-靶标信息去预测缺失的值(图 1)，通过选取有 ≥ 2 个结合信息的靶标和药物构建一个药物-靶标网络，这个网络相当于一个大的矩阵，里面使用 1 或者 0 进行填充，如果药物和靶标有结合信息，则填充 1，反过来填充 0。将这个矩阵放到矩阵分解算法中进行迭代更新训练，最终会得到一个输出矩阵，这个输出矩阵则是将缺失的值进行了填充，最后挑选的策略是选择打分最低的前 6 262 个作为阴性样本集合。

1.4 建模方法

本研究使用了主流的机器学习模型，这些模型主要包括：支持向量机、决策树、k-近邻、随机森林、Adaboost、朴素贝叶斯分类器和神经网络模型^[19-24]。其目的是为了比较不同机器学习模型在靶标-药物相互作用分类预测上的预测能力，模型构建使用了 Python 中 Tensorflow 库和 Scikit-learn 库。本研究建立的药物-靶标相互作用预测模型的工作流程见图 2。首先从文献中获取相关的数据，这些数据的主要来源于 DrugBank 数据库。其中里面包括了药物的 SMILES 结构式、蛋白质序列信息和药物-靶标结合信息(结合信息目前只分析有无结合关系)。然后利用计算分子/蛋白质描述符的计算软件(RdKit 和 Protein Propy)去计算每一个药物的分子指纹和每一个蛋白质描述符。然后，再挑选药物-靶标对的阳性样本和阴性样本。阳性样本集已经由下载的数据提供，药物-靶标对的阴性样本集则是通过采用 PMF 进行获取。将药物-靶标对按照 60%，20%，20%的比例分成训练集、验证集和测试集，再将训练样本特征信息输入到机器学习的模型中进行训练，根据自身的经验以及文献的指导进行模型的调参过程，从而训练得到一

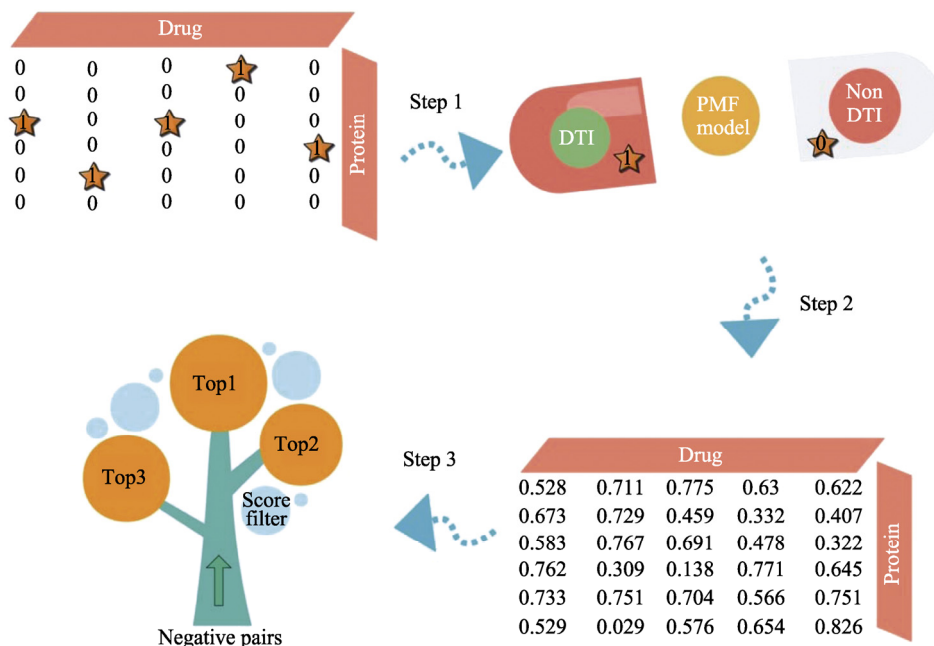


图 1 阴性样本选择过程示意图

Fig. 1 Schematic of the negative sample selection process

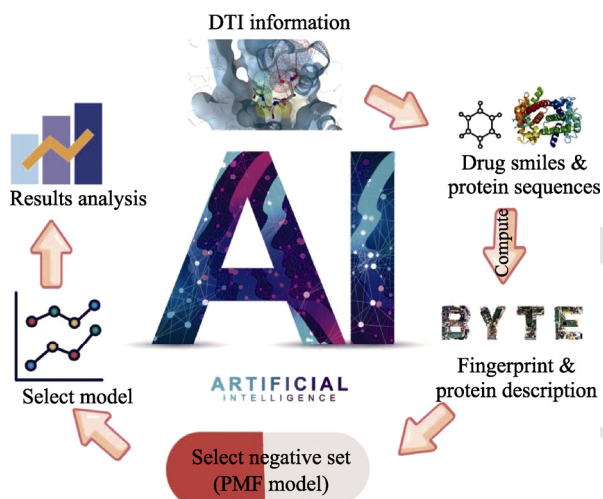


图 2 药物-靶标相互作用预测模型的工作流程

Fig. 2 Workflow of drug-target interaction prediction model

个准确率高泛化的模型。最后建立一个合适的评价标准进行药物-靶标的预测。最终实现的是只需要输入 1 对药物-靶标的信息,模型就可以预测出它们之间是否存在结合的可能性。

1.5 结果评价和分析方法

使用常规的方法来评估预测模型的性能,包括分类准确性、精密度、特异性、敏感性和阴性预测值,分别定义如下:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TN}{TP + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{NPV} = \frac{TN}{TN + FN}$$

其中 TP 、 TN 、 FP 和 FN 分别代表真阳性、真阴性、假阳性和假阴性。

2 结果

2.1 机器学习模型结果分析

训练不同的机器学习模型来预测药物-靶标相互作用,并比较不同训练模型的预测结果,见图 3。虽然朴素贝叶斯和二次判别模型的准确度和灵敏度都很高,但其他评价标准都很差,这表明该分类器存在某些缺陷。在其他模型中,评估标准总体上没有明显的缺陷,表现出相对平均的性能。例如, k -近邻、Adaboost 和支持向量机模型的评估标准都约为 70%。树结构模型的性能最好,包括决策树、随机森林和梯度增强决策树模型。3 种模型的评价标准均 $>80\%$ 。值得注意的是,决策树模型的 $AUC > 90\%$,表明预测 DTI 的能力良好。然而,由于深度学习模型具有从大量数据中提取潜在关键特征的优势,以及其出色的分类或回归能力、强大的特征提取能力和较低的泛化误差。此外,上述传统机器学习模型的性能还需进一步改进,因此,本研究后续采用了深度神经网络模型来预测药物-靶标的分类。

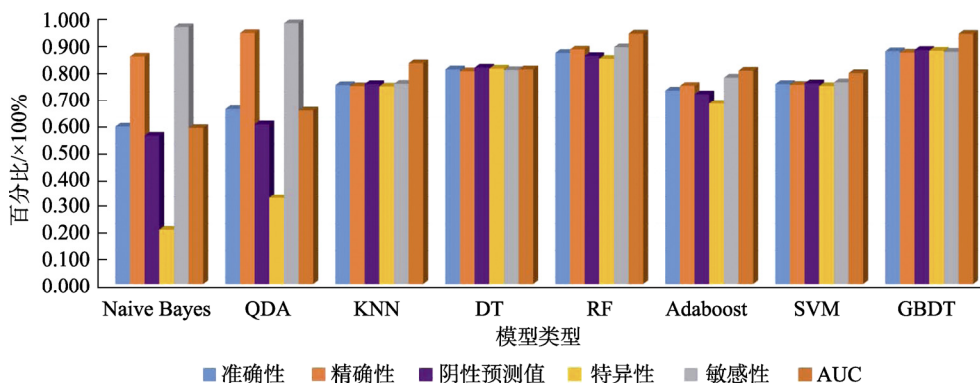


图 3 各种机器学习模型的预测结果

Naive Bayes-朴素贝叶斯分类器; QDA-二次判别分析模型; KNN-k 近邻分类器; DT-决策树分类器; RF-随机森林分类器; Adaboost-Adaboost 分类器; SVM-支持向量机分类器; GBDT-梯度 Boosting 决策树分类器。

Fig. 3 Prediction results for a variety of machine learning models

Naive Bayes-naive bayes classifier; QDA-quadratic discriminant analysis model; KNN -k-nearest neighbor classifier; DT-decision tree classifier; RF-random forest classifier; Adaboost-Adaboost classifier; SVM-support vector machine classifier; GBDT-gradient Boosting decision tree classifier.

2.2 DLDTIs-Con 模型分析和比较

为了验证独立模型优于集成模型,通过比较 2 种不同的网络架构:①2 个独立网络作为药物-靶标对的输入(图 4,即 DLDTIs-Sep);②一个集成网络作为药物-靶标对的输入(图 5,即 DLDTIs-Con)。通过准确度、AUC、敏感性和特异性的评估指标,

将训练模型 DLDTIs-Sep 与先前开发的机器学习模型(表 1)和 DLDTIs-Con 进行比较。DLDTIs-Con 模型的性能优于其他机器学习模型,每个评估标准均>90%。对于 AUC,深度神经网络训练模型达到 96%。值得一提的是,DLDTIs-Con 模型优于 DLDTIs-Sep 模型。这样的结果可能有 2 个原因。

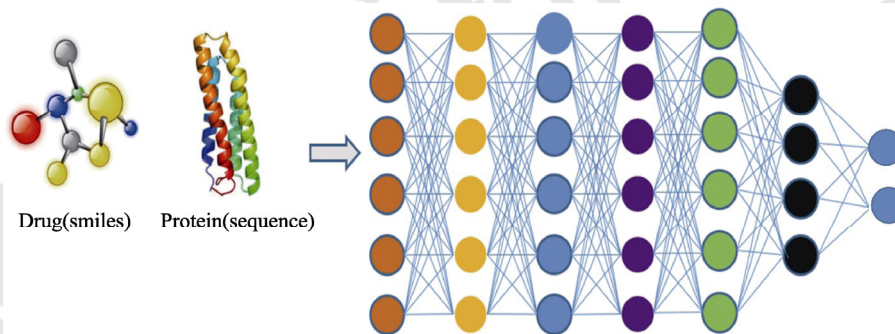


图 4 DLDTIs-Sep 模型架构

Fig. 4 Architecture of DLDTIs-Sep model

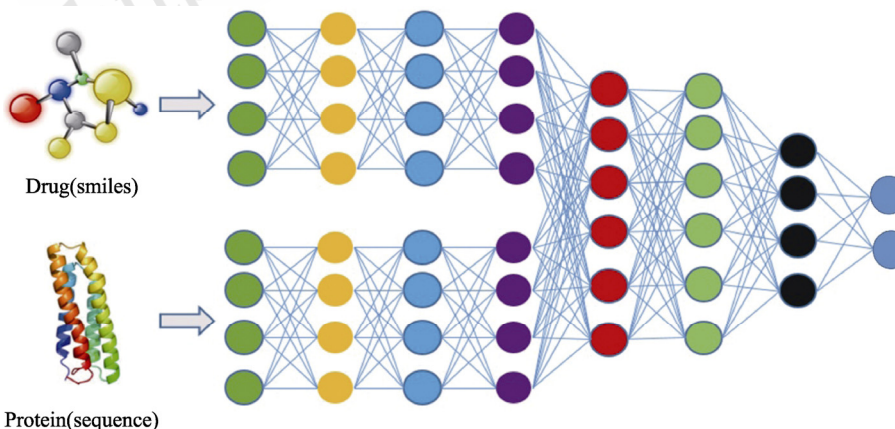


图 5 DLDTIs-Con 模型架构

Fig. 5 Architecture of DLDTIs-Con model

表 1 比较不同的深度学习模型架构和最佳机器学习模型
Tab. 1 Comparison among different deep-learning model architectures and optimal machine-learning models

模型	准确性	敏感性	特异性	AUC
Decision Tree	0.809	0.806	0.811	0.809
Random Forest	0.871	0.892	0.848	0.943
GBDT	0.877	0.875	0.879	0.943
DLDTIs-Con	0.906	0.923	0.890	0.961
DLDTIs-Sep	0.867	0.878	0.857	0.939

首先, 输入维度越高, 越有利于模型的训练和预测。DLDTIs-Con 的输入维度为 7 278, DLDTIs-Sep 的输入维度为 6 311 和 967。其次, DLDTIs-Sep 单独训练蛋白质和药物特征可能会导致某些结合信息的丢失, 这将导致较差的结果。这些结果表明, DLDTIs-Con 可以作为预测 DTI 的最终训练模型。

当今, 许多基于深度学习的模型已经出现, 用于预测药物-靶点相互作用。Wen 等^[12]开发了深度学习无监督模型(深度信念网络), Lee 等^[25]使用蛋白质序列和化合物一维表示训练了基于卷积神经网络的模型。笔者课题组在 Wen 的数据集基础上将 DLDTI 方法与上述 2 个模型的预测性能进行了比较, 见表 2。结果表明, DLDTI 的每个评估指标都高于通过文献方法获得的指标, 证明了所发展模型的优越性。因此, 使用高维特征来表征分子和蛋白质, 以及使用 PMF 方法选择阴性样本有助于提高药物-靶标相互作用预测模型的精度。

表 2 DLDTIs-Con 方法和其他最新研究成果的预测性能比较

Tab. 2 Comparison of the prediction performance between our proposed method and other state-of-the-art works

方法	准确性	敏感性	特异性	AUC
Wen's model	0.858	0.822	0.895	0.915
Lee's model	0.911	0.900	0.923	0.966
DLDTIs-Con	0.952	0.852	0.973	0.969

最后, 为了进一步验证模型有效性, 通过提取 2019 年 Drugbank 数据库中新增加的已验证的药物-靶标相互作用对, 包括了 768 个新增数据。这部分数据可用作模型的阳性数据验证。最后, 模型从 768 个阳性中正确预测出 353 个阳性药物-靶标对。另外, 通过统计发现, 该模型对作用于不同类型钠通道蛋白 α 亚基的药物的预测非常敏感, 见表 3~4。值得一提的是, 帕罗西汀(DB00715)的靶标预测结果基本正确。

表 3 DLDTIs-Con 正确预测基于药物的药物靶标对前五名
Tab. 3 Top 5 predict number of drug-target pairs for drugs based by DLDTIs-Con

排序	DrugBank 数据库 ID 号	DrugBank 数据库药物名称	预测正确的药物靶标对总数
1	DB00715	Paroxetine	15
2	DB01043	Aripiprazole	11
3	DB01238	Memantine	11
4	DB00945	Ranolazine	10
5	DB00243	Acetylsalicylic acid	10

表 4 DLDTIs-Con 正确预测基于蛋白质的药物靶标对的前五名

Tab. 4 Top 5 predict number of drug-target pairs for proteins based by DLDTIs-Con

排序	UniProt 数据库 ID 号	UniProt 数据库中命名	预测正确的药物靶标对总数
1	P35499	Sodium channel protein type 4 Subunit alpha	6
2	Q99250	Sodium channel protein type 2 subunit alpha	6
3	Q9UQD0	Sodium channel protein type 8 subunit alpha	5
4	Q01118	Sodium channel protein type 3 subunit alpha	5
5	Q9NY46	Sodium channel protein type 7 subunit alpha	5

3 结论

为提高现有药物-靶标相互作用预测方法的准确性, 本研究开发了一种基于深度神经网络的新型预测模型。创新性地使用了推荐算法中常用的 PMF 用于负样本的选取。与随机选择方法相比, PMF 选择了得分最低的阴性样本, 这有助于防止使用过度错误分类的潜在阳性样本作为阴性样本来训练模型。通过比较不同类型的模型, DLDTIs-Con 网络结构被证明最适合于药物-靶标相互作用的训练, 其准确度、特异性、敏感性和 AUC 均达到 90%。目前 DLDTIs-Con 可作为预测小分子药物与靶点结合概率以及靶标预测的补充工具。

REFERENCES

- [1] FRYE L, BHAT S, AKINSANYA K, et al. From computer-aided drug discovery to computer-driven drug discovery[J]. Drug Discov Today Technol, 2021(39): 111-117.
- [2] BENDER A, CORTES-CIRIANO I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data[J]. Drug Discov Today, 2021, 26(4): 1040-1052.
- [3] BURLEY S K, BERMAN H M. Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction[J]. Structure, 2021, 29(6): 515-520.
- [4] CHAKRABORTY A, MITRA S, DE D, et al. Determining

- protein-protein interaction using support vector machine: A review[J]. IEEE Access, 2021(9): 12473-12490.
- [5] PUSHPAKOM S, IORIO F, EYERS P A, et al. Drug repurposing: Progress, challenges and recommendations[J]. Nat Rev Drug Discov, 2019, 18(1): 41-58.
- [6] KADURIN A, NIKOLENKO S, KHRABROV K, et al. DruGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico[J]. Mol Pharm, 2017, 14(9): 3098-3104.
- [7] HUANG D Z, BABER J C, BAHMANYAR S S. The challenges of generalizability in artificial intelligence for ADME/Tox endpoint and activity prediction[J]. Expert Opin Drug Discov, 2021, 16(9): 1045-1056.
- [8] SHEN Z H, YAN Y H, YANG S, et al. ProfKin: A comprehensive web server for structure-based kinase profiling [J]. Eur J Med Chem, 2021(225): 113772.
- [9] WU Z R, CHENG F X, LI J, et al. SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug-target interactions and drug repositioning[J]. Brief Bioinform, 2016, 18(2): 333-347.
- [10] LUO Y N, ZHAO X B, ZHOU J T, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information[J]. Nat Commun, 2017, 8(1): 573.
- [11] BALA S A, KANT S O, YAKASAI A G. Deep learning in medical imaging and drug design[J]. J Hum Physiol, 2021, 2(2): 32-37.
- [12] WEN M, ZHANG Z M, NIU S Y, et al. Deep-learning-based drug-target interaction prediction[J]. J Proteome Res, 2017, 16(4): 1401-1409.
- [13] ÖZTÜRK H, ÖZGÜR A, ÖZKIRIMLI E. DeepDTA: deep drug-target binding affinity prediction[J]. Bioinformatics, 2018, 34(17): i821-i829.
- [14] WISHART D S, FEUNANG Y D, GUO A C, et al. DrugBank 5.0: A major update to the DrugBank database for 2018[J]. Nucleic Acids Res, 2018, 46(D1): D1074-D1082.
- [15] ROGERS D, HAHN M. Extended-connectivity fingerprints[J]. J Chem Inf Model, 2010, 50(5): 742-754.
- [16] DURANT J L, LELAND B A, HENRY D R, et al. Reoptimization of MDL keys for use in drug discovery[J]. J Chem Inf Comput Sci, 2002, 42(6): 1273-1280.
- [17] SHEN H B, CHOU K C. PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition[J]. Anal Biochem, 2008, 373(2): 386-388.
- [18] XU S S, ZHUANG H Y, SUN F Z, et al. Recommendation algorithm of probabilistic matrix factorization based on directed trust[J]. Comput Electr Eng, 2021(93): 107206.
- [19] XU R H. Research on drug molecular computing based on support vector machine and neural network[J]. IOP Conf Ser: Earth Environ Sci, 2021, 714(2): 022031.
- [20] HABIB P, ALSAMMAN A, HASSANEIN S, et al. TarDict: A RandomForestClassifier based software predicts drug-target interaction using SMILES[J]. Highlights Bioinform, 2021: bi202101.
- [21] ZHANG Z H. Introduction to machine learning: K-nearest neighbors[J]. Ann Transl Med, 2016, 4(11): 218.
- [22] ZHAO S M, MAO X Y, LIN H H, et al. Machine learning prediction for 50 anti-cancer food molecules from 968 anti-cancer drugs[J]. Int J Intell Sci, 2020, 10(1): 1-8.
- [23] QUINLAN J R. Induction of decision trees[J]. Mach Learn, 1986, 1(1): 81-106.
- [24] KONONENKO I. Semi-naive Bayesian classifier[C]//Mach Learn—EWSL 91, 1991: 206-219.
- [25] LEE I, KEUM J, NAM H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences[J]. PLoS Comput Biol, 2019, 15(6): e1007129.

收稿日期: 2022-06-16
(本文责编: 沈倩)