第 42 卷 第 6 期 计 算 机 工 程 2016 年 6 月
Vol. 42 No. 6 Computer Engineering June 2016

·人工智能及识别技术· 文章编号：1000-3428(2016)06-0218-06 文献标志码：A 中图分类号：TP39

一种多信息融合的药物-靶标关联预测算法

彭利红[1]，李泽军[2]，陈 敏[2]，任日丽[1]

（1. 长沙医学院 计算机系，长沙 410219；2. 湖南工学院 计算机与信息科学学院，湖南 衡阳 421002）

摘 要： 在药物结构相似性和靶标序列相似性的基础上，结合药物-靶标相互作用网络信息，考虑分类器和数据集合分布的复杂性，提出一种半监督学习算法预测药物与靶标之间的关联。实验结果表明，该算法的预测性能较 DBSI，KBMF2K 等算法有所提高。对其预测到的药物-靶标相互作用数据进行打分并排序，从中提取前 30% 的数据，其中有部分相互作用可在 KEGG，DrugBank，SuperTarget 和 ChEMBL 数据库中得到验证。

关键词： 多信息融合；半监督学习；药物-靶标相互作用网络；药物相似性；靶标相似性

中文引用格式： 彭利红，李泽军，陈 敏，等. 一种多信息融合的药物-靶标关联预测算法[J]. 计算机工程，2016，42(6)：218-223，229.

英文引用格式： Peng Lihong，Li Zejun，Chen Min，et al. A Drug-target Association Prediction Algorithm with Multi-information Fusion[J]. Computer Engineering，2016，42(6)：218-223，229.

A Drug-target Association Prediction Algorithm with Multi-information Fusion

PENG Lihong[1]，LI Zejun[2]，CHEN Min[2]，REN Rili[1]

（1. Department of Computer，Changsha Medical University，Changsha 410219，China；
2. School of Computer and Information Science，Hunan Institute of Technology，Hengyang，Hunan 421002，China）

【Abstract】 This paper considers complexity of the classifier and the geometric distribution of data，and proposes a semi-supervised learning algorithm to predict the associations between drugs and targets combining drug-target interactions network based on drug structure similarity and target sequence similarities. Experimental results show that，this algorithm has better prediction performance compared with DBSI algorithm，KBMF2K algorithm，etc.. The drugs-target interaction data predicted by the proposed algorithm is scored and sorted，and parts of interactions data can be retired from KEGG，DrugBank，SuperTarget and ChEMBL among the predicted interactions with 30% highest scores.

【Key words】 multi-information fusion；semi-supervised learning；drug-target interactions network；drug similarity；target similarity

DOI： 10.3969/j.issn.1000-3428.2016.06.039

1 概述

预测药物-靶标之间的相互作用关系是现代药物研发中至关重要的一步[1]。传统的实验方法主要依靠药理实验，致使发现的药物疗效不理想、伴有副作用且造成了人力和物资的巨大浪费[2]。计算方法作为对实验方法的有效补充，对节约药物研发成本、提高研发效率、减少药物研发风险具有重要意义[3-5]。

传统的计算方法可分为基于配体的方法[6]、基于受体的方法[7]和文本挖掘方法[8]，然而，基于配体的方法依赖于与靶标关联的配体数量；基于受体的方法不能用于三维结构未知的靶标；而文本挖掘方法则存在文献中药物化合物命名冗余的问题[8]。有研究者假设药物越相似，与相似靶标发生相互作用的可能性越大，从而整合药物的结构相似性、靶标的序列相似性和已知的药物-靶标相互作用等生物信息，基于统计方法来预测药物-靶标相互作用：文献[9]提取来自人类酶、离子通道、GPCRs、细胞核受体的药物-靶标相互作用，把药物的化学结构和靶标

基金项目： 湖南省教育厅优秀青年基金资助项目(14B023)；湖南省教育厅基金资助一般项目(13C1108)。

作者简介： 彭利红(1978－)，女，讲师、博士研究生，主研方向为机器学习、数据挖掘、生物信息学；李泽军(通讯作者)、陈 敏，副教授、博士研究生；任日丽，讲师、硕士。

收稿日期： 2015-04-29 **修回日期：** 2015-07-12 **E-mail：** 304152648@qq.com

的序列信息整合进药理空间,提出一种监督学习方法推断药物与靶标之间的关联;文献[10]利用局部二分图模型来预测未知的相互作用;文献[11]通过分析化学结构、药理结构和药物-靶标相互作用网络的拓扑结构进行预测;文献[12]结合降维、矩阵因子化和二分类器,提出一种新的贝叶斯模型进行预测。然而,上述方法都把未知的药物-靶标相互作用作为负样本,影响了预测的准确度。

半监督学习利用未标记数据对数据进行学习,可提高学习性能,因此,受到越来越多的研究者的青睐[13-14]。为充分利用未知的药物-靶标相互作用数据,提高预测性能,文献[15]基于标记信息和未标记信息提出一种半监督学习方法 NetLapRLS 来推测药物与靶标之间的关联;文献[16]提出 NRWRH 方法,结合药物相似性网络、蛋白质相似性网络和药物-靶标相互作用网络,基于随机游走理论来预测药物-靶标相互作用;文献[17]基于已知药物-靶标相互作用网络提出基于网络的推断方法 NBI。然而,这三种网络推理方法都没有考虑到靶标信息未知的药物。

以上方法存在一些缺陷:缺少已知的药物-靶标相互作用信息,没有经验证的相互作用负样本数据使得研究者很难得到负样本,更不用说从负样本数据中选择数据,另外,很少有方法用于预测靶标信息未知的新药[18]。虽然文献[18]提出半监督学习方法 NetCBP,结合少量的标记数据和大量未标记数据来推测药物与靶标之间的关联,但这种方法严重依赖于由化合物结构相似而得到的药物相似性和由序列相似而得到的靶标相似性。

本文在 NetCBP 方法中药物相似性和靶标相似性的基础上,结合药物-靶标相互作用网络,考虑分类器和数据几何分布的复杂性,提出一种基于半监督学习和多信息融合的药物-靶标相互作用预测(Predicting drug-target interaction based on Semi-supervised Learning and Multi-information Fusion, PreSLMF)算法,用以推测药物与靶标之间的关联。

2　数据与方法

2.1　数据

笔者从 http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/中下载了文献[9]中提供的来自人类酶、离子通道、GPCRs、核受体的 4 类药物-靶标相互作用数据,具体描述如下:

(1)从 KEGG LIGAND 数据库[19]中 DRUG 和 COMPOUND 部分提取药物的化学结构,利用 SIMCOMP 方法[23],把药物作为图形,基于 2 个图形之间共有子结构的大小,使用图形比对方法计算药物之间的结构相似性,从而得到药物相似性矩阵 $\boldsymbol{S}_{sd}$。

(2)从 KEGG GENES 数据库[19]中获得靶标数据,使用 Smith-Waterman 正则化打分函数计算靶标之间的序列相似性:给定 2 个靶标 t_1 和 t_2,令 $\boldsymbol{SW}(t_1,t_2)$ 表示两靶标之间的 Smith-Waterman 打分值[24],则其序列相似性 $\boldsymbol{S}_{sp}(t_1,t_2)=\dfrac{\boldsymbol{SW}(t_1,t_2)}{\sqrt{\boldsymbol{SW}(t_1,t_1)}\cdot\sqrt{\boldsymbol{SW}(t_2,t_2)}}$,从而得到靶标之间的相似性矩阵 $\boldsymbol{S}_{sp}$。

(3)从人类酶、离子通道、GPCRs、核受体中发现 445 个、210 个、223 个和 54 个药物,分别与 664 个、204 个、95 个和 26 个靶标发生相互作用,其相互作用数量分别为 2 926 个,1 476 个,635 个和 90 个[9-12,15-18]。本文令这部分相互作用数据作为标准数据集以评估性能,具体如表 1 所示。

表 1　4 类药物-靶标相互作用数据集

数据集	药物数量	靶标数量	药物与靶标相互作用数据的数量
酶	445	664	2 926
离子通道	210	204	1 476
GPCRs	223	95	635
核受体	54	26	90

2.2　方法

在机器学习领域,半监督学习受到越来越多的关注。本文基于半监督学习[25-26],融合药物相似性、靶标相似性和药物-靶标相互作用数据,预测药物与靶标之间新的关联。

在预测药物与靶标之间的关联时,药物之间的相似性和靶标之间的相似性是非常重要的。给定 n 个药物和 m 个靶标,令矩阵 $\boldsymbol{A}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n]$ 表示原始的药物-靶标相互作用网络,若药物 i 与靶标 j 存在相互作用,则 $a_{ij}=1$,否则 $a_{ij}=0$。

首先,基于余弦相似性和药物-靶标相互作用网络定义靶标之间的关联:

$$\boldsymbol{S}_{np}(i,j)=\frac{\boldsymbol{A}_{i\cdot}\boldsymbol{A}_{j\cdot}^{T}}{\|\boldsymbol{A}_{i\cdot}\|\ \|\boldsymbol{A}_{j\cdot}\|},i,j=1,2,\cdots,m \quad (1)$$

其中,$\boldsymbol{A}_{i\cdot}$ 表示关联矩阵 $\boldsymbol{A}$ 的第 i 行。式(1)描述了由相互作用网络得到的靶标之间的相似性,如果靶标 i 和 j 之间共享了大量药物,但靶标 k 和 l 之间只共享了少量药物甚至没有共享药物,则 $\boldsymbol{S}_{sp}(i,j)>\boldsymbol{S}_{sp}(k,l)$,综合由药物-靶标关系矩阵得到的相似性和靶标相似性得到新的靶标相似性矩阵:

$$\boldsymbol{S}'_{p}=\boldsymbol{S}_{sp}+\alpha\boldsymbol{S}_{np} \quad (2)$$

其中,α 用于平衡由药物-靶标相互作用数据得到的靶标相似性和由序列得到的靶标相似性之间的重要程度,其定义如下:

$$\alpha=\frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\boldsymbol{S}_{\mathrm{sp}}(i,j)}{\sum_{i=1}^{m}\sum_{j=1}^{m}\boldsymbol{S}_{\mathrm{np}}(i,j)}$$

对 $\boldsymbol{S}_{\mathrm{p}'}$进行归一化处理：

$$\boldsymbol{S}_{\mathrm{p}}(i,j)=\frac{\boldsymbol{S}'_{\mathrm{p}}(i,j)}{\sum_{k=1}^{m}\boldsymbol{S}'_{\mathrm{p}}(i,k)} \tag{3}$$

令 $\boldsymbol{A}'=[\boldsymbol{a}'_1,\boldsymbol{a}'_2,\cdots,\boldsymbol{a}'_1]$表示药物与靶标之间的新关联集，其中，$\boldsymbol{A}'=\boldsymbol{S}_{\mathrm{p}}\boldsymbol{A}$，$a'_{ij}$表示经过标签传播后，药物 i 与靶标 j 相互作用的概率；$\boldsymbol{f}_{ij}$表示预测的药物 i 与靶标 j 相互作用的概率，定义带权的损失函数作为目标函数的第一部分：

$$\phi_1(\boldsymbol{f})=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{W}_{ij}(f_{ij}-a'_{ij})^2 \tag{4}$$

其中，矩阵 $\boldsymbol{W}$ 为权重函数，表示药物 i 与靶标 j 发生相互作用的可能性，定义 $\boldsymbol{W}_{ij}$如下：

$$\boldsymbol{W}_{ij}=\begin{cases}1 & a_{ij}=1\\ \boldsymbol{S}_{\mathrm{p}_{j\cdot}}\cdot\boldsymbol{A}_{\cdot i} & a_{ij}=0\end{cases} \tag{5}$$

其中，$\boldsymbol{A}_{\cdot i}$是 $\boldsymbol{A}$ 的第 i 列；$\boldsymbol{S}_{\mathrm{p}_{j\cdot}}$是 $\boldsymbol{S}_{\mathrm{p}}$ 的第 j 行。如果药物 i 的靶标 j 与未知靶标 k 有很强的关联，则药物 i 更有可能与未知靶标 k 发生相互作用。

由于一种药物可能作用于多个靶标，因此药物之间共享的靶标数量也可以用来衡量药物之间的相似性，共享靶标数越多，相似性越大。为此，本文基于已知的药物-靶标相互作用网络中共有的靶标数定义药物之间的相似性：

$$\boldsymbol{S}_{\mathrm{nd}}(i,j)=\frac{\boldsymbol{A}_{\cdot i}^{\mathrm{T}}\boldsymbol{A}_{\cdot j}}{\|\boldsymbol{A}_{\cdot i}\|\|\boldsymbol{A}_{\cdot j}\|},i,j=1,2,\cdots,n \tag{6}$$

综合由相互作用网络得到的相似性和由化学结构得到的相似性，定义药物相似性如下：

$$\boldsymbol{S}'_{\mathrm{d}}=\boldsymbol{S}_{\mathrm{sd}}+\beta\boldsymbol{S}_{\mathrm{nd}} \tag{7}$$

其中，β 用于平衡由药物-靶标相互作用数据得到的药物相似性和由化学结构得到的药物相似性之间重要程度：

$$\beta=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{S}_{\mathrm{sd}}(i,j)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{S}_{\mathrm{nd}}(i,j)}$$

对 $\boldsymbol{S}'_{\mathrm{d}}$进行归一化后得到：

$$\boldsymbol{S}_{\mathrm{d}}(i,j)=\frac{\boldsymbol{S}'_{\mathrm{d}}(i,j)}{\sum_{k=1}^{n}\boldsymbol{S}'_{\mathrm{d}}(i,k)} \tag{8}$$

药物的化学结构越相似，共享靶标数越多，越有可能与相似的靶标发生相互作用，因此，定义平滑函数作为目标函数的第 2 项：

$$\varphi_2(\boldsymbol{f})=\mathrm{tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{F}) \tag{9}$$

其中，$\boldsymbol{F}=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_m]^{\mathrm{T}}$。

$$\boldsymbol{D}_{ii}=\sum_{j=1}^{n}\boldsymbol{S}_d(i,j) \tag{10}$$

$$\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}_{\mathrm{d}}\boldsymbol{D}^{-\frac{1}{2}} \tag{11}$$

其中，$\boldsymbol{I}$ 为 $n\times n$ 单位矩阵；tr( ) 表示矩阵的迹函数。通过最小化平滑函数，能根据与某一药物发生相互作用的靶标预测与此药物关联的新靶标。

目前，已知药物-靶标相互作用数据很少，导致其数据是稀疏的，因此，用以下正则项逼近真实值：

$$\varphi_3(\boldsymbol{f})=\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F} \tag{12}$$

最后定义目标函数为：

$$\begin{aligned}\varphi(\boldsymbol{F})&=\varphi_1(\boldsymbol{f})+\gamma\varphi_2(\boldsymbol{f})+\lambda\varphi_3(\boldsymbol{f})\\&=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{W}_{ij}(f_{ij}-a'_{ij})^2\\&\quad+\gamma\mathrm{tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{F})+\lambda\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}\\&=\frac{1}{2}\|\boldsymbol{W}\circ(\boldsymbol{F}-\boldsymbol{A}')^{\mathrm{T}}(\boldsymbol{F}-\boldsymbol{A}')\|\\&\quad+\gamma\mathrm{tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{F})+\lambda\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}\|\end{aligned} \tag{13}$$

其中，$\circ$表示矩阵的 Hadamard 乘积；参数 γ 和 λ 用于平衡平滑项和稀疏项之间的重要性。第 1 项和第 2 项用于预测训练集和测试集中未知的药物-靶标之间的相互作用，通过最小化式(13)，在保证平滑和稀疏的情况下，尽可能多地预测药物-靶标相互作用。

为求解 $\boldsymbol{F}$，对 $\varphi(\boldsymbol{F})$进行求导：

$$\frac{\partial\varphi(\boldsymbol{F})}{\partial\boldsymbol{F}}=\boldsymbol{W}\circ(\boldsymbol{F}-\boldsymbol{A}')+\gamma\boldsymbol{L}\boldsymbol{F}+\lambda\boldsymbol{I}\boldsymbol{F} \tag{14}$$

把式(14)分解为 m 个问题，令：

$$\boldsymbol{S}_l=\boldsymbol{W}_{l\cdot}\circ\boldsymbol{A}'_{l\cdot},l=1,2,\cdots,m \tag{15}$$

$$\boldsymbol{W}'_{l\cdot}=\mathrm{diag}(\boldsymbol{W}_{l\cdot}) \tag{16}$$

对第 1 个问题，可用下式求解：

$$\boldsymbol{f}_l=(\boldsymbol{W}'_{l\cdot}+\alpha\boldsymbol{L}+\beta\boldsymbol{I})^{-1}\boldsymbol{S}_l \tag{17}$$

从而得到药物-靶标相互作用概率矩阵 $\boldsymbol{DPI}_1=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_m]^{\mathrm{T}}$。

因为药物与靶标之间的关联是相互的，所以为获得更加可靠的结果，以药物相似性矩阵 $\boldsymbol{S}_d$ 为带权系数矩阵，以靶标相似性矩阵 $\boldsymbol{S}_{\mathrm{p}}$ 为基于计算其拉普拉斯矩 $\boldsymbol{L}$，获得类似于式(13)的预测模型，其求解过程类似于式(13)，从而得到另一药物-靶标相互作用概率矩阵 $\boldsymbol{DPI}_2=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n]^{\mathrm{T}}$。

综合 $\boldsymbol{DPI}_1$ 和 $\boldsymbol{DPI}_2$ 得到最终的相互作用矩阵：

$$\boldsymbol{DPI}=\frac{1}{2}(\boldsymbol{DPI}_1^{\mathrm{T}}+\boldsymbol{DPI}_2) \tag{18}$$

算法 1 PreSLMF

输入 药物结构相似性矩阵，靶标序列相似性矩阵，药物-靶标相互作用网络

输出 药物-靶标相互作用概率矩阵

步骤 1 使用式(1)～式(3)、式(5)计算损失函数的权重矩阵 $\boldsymbol{W}$；

步骤 2 使用式(6)～式(8)、式(10)和式(11)计算拉普拉斯矩阵 $\boldsymbol{L}$；

步骤 3 计算 $\boldsymbol{A}'=\boldsymbol{S}_{\mathrm{p}}\boldsymbol{A}$；

步骤 4　把求解问题分解为 m 个问题,对每个问题执行步骤 5 和步骤 6;

步骤 5　利用式(15)和式(16)计算 $\boldsymbol{S}_l$ 和 $\boldsymbol{W}'_l$;

步骤 6　利用式(17)求解 $\boldsymbol{f}_l$ 得 $\boldsymbol{DPI}_1=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_m]^{\mathrm{T}}$;

步骤 7　与步骤 1 ~ 步骤 6 类似,得到 $\boldsymbol{DPI}_2=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n]^{\mathrm{T}}$;

步骤 8　利用式(18)获得最终的药物-靶标相互作用矩阵 $\boldsymbol{DPI}$。

3　实验结果与分析

本文基于 4 个数据集,采用 5 折交叉验证,把标准数据集随机分成大小相等的 5 份,随机选择 1 份作为测试集,其他 4 份作为训练集,实验重复 20 次,取其平均性能作为最终结果,以对 PreSLMF 算法与其他 5 种算法进行分析比较。例如,在 GPCRs 数据集中,有 635 个药物-靶标相互作用数据,在每次交叉验证时,把 508 个已知的药物-靶标相互作用数据作为训练集,其余的 20 677 个药物-靶标对,包括已知的 127 个药物-靶标相互作用数据作为测试集。参数 γ 和 λ 分别取值 0.01 和 0.001。

在药物-靶标相互作用预测中,AUC 值和 AUPR 值是 2 个具有代表性的性能评价标准。AUC 值是 ROC 曲线下的面积,AUPR 是精度-召回曲线下的面积。这 2 个值越大,算法的性能越好。对于药物-靶标相互作用预测来说,由于标准数据集中的数据呈现不平衡性,已知的相互作用数据相对较少,而 AUPR 值能尽可能地减少假阳性数据对预测性能的影响,所以,相对来说,AUPR 值能更合理地评价算法的性能。

3.1　与其他算法的分析比较

本文把 PreSLMF 算法与 DBSI 算法[17]、文献[9]算法、KBMF2K 算法[12]及 NetCBP 算法[18]进行了分析比较,表 2 和表 3 给出了算法 AUC 值和 AUPR 值的比较结果。

表 2　不同算法的 AUC 值比较

数据集	DBSI 算法	文献[9]算法	KBMF2K 算法	NetCBP 算法	PreSLMF 算法
酶	0.807 5	0.821	0.832	0.825 1	0.856 5
离子通道	0.802 9	0.692	0.799	0.803 4	0.830 7
GPCRs	0.802 2	0.811	0.857	0.823 5	0.871 2
核受体	0.757 8	0.814	0.824	0.839 4	0.834 6

表 3　不同算法的 AUPR 值比较

数据集	DBSI 算法	文献[9]算法	KBMF2K 算法	NetCBP 算法	PreSLMF 算法
酶	0.755 6	0.772 3	0.793 1	0.778 4	0.818 7
离子通道	0.749 2	0.648 1	0.761 3	0.752 5	0.794 2
GPCRs	0.747 1	0.760 8	0.819 4	0.780 3	0.837 4
核受体	0.704 2	0.762 9	0.775 8	0.786 9	0.801 3

从表 2 数据可以看出,从 AUC 值来说,与 DBSI 算法、文献[9]算法和 KBMF2K 算法相比,PreSLMF 算法在 4 个数据集上的性能均好于这 3 种方法。与 DBSI 算法相比,PreSLMF 算法在酶、离子通道、GPCRs 及核受体上的性能分别提高了 5.7%,3.3%,7.9% 和 9.2%;与文献[9]算法相比,PreSLMF 算法在酶、离子通道、GPCRs 及核受体上的性能分别提高了 4.1%,16.7%,6.9% 和 2.5%;与 KBMF2K 算法相比,PreSLMF 算法在酶、离子通道、GPCRs 及核受体上的性能分别提高了 2.9%,3.9%,1.6% 和 2.5%;与 NetCBP 算法相比,虽然在个别数据集上性能稍微有点降低,但总体来说,性能较好:PreSLMF 算法在酶、离子通道及 GPCRs 上的性能分别提高了3.7%,3.3% 和 5.5%,在核受体上的性能降低了 0.6%。

从表 3 中数据可以看出,从 AUPR 值来说,与 DBSI 算法、文献[9]算法、KBMF2K 算法和 NetCBP 算法相比,PreSLMF 算法在 4 个数据集上的性能均好于这 4 种方法。与 DBSI 算法相比,PreSLMF 算法在酶、离子通道、GPCRs 及核受体上的性能分别提高了 7.7%,5.7%,10.8% 和 12.1%;与文献[9]算法相比,PreSLMF 算法在酶、离子通道、GPCRs 及核受体上的性能分别提高了 5.7%,18.4%,9.1% 和 4.8%;与 KBMF2K 算法相比,PreSLMF 算法在酶、离子通道、GPCRs 及核受体上的性能分别提高了 3.1%,4.1%,2.1% 和 3.2%;与 NetCBP 算法相比,PreSLMF 方法在酶、离子通道、GPCRs 及核受体上

的性能分别提高了 4.9% ,5.3% ,6.8% 和 1.8% 。

与 NetCBP 算法相比,在核受体数据集中,PreSLMF 算法在 AUC 上的性能略有降低,造成这种现象的原因可能是:核受体数据集中已知的药物-靶标相互作用数据相对较少,NetCBP 算法对该数据集中的数据进行预测时产生了大量的假阳性数据,致使 NetCBP 算法在 AUC 上的性能略高于 PreSLMF 算法。NetCBP 算法利用随机游走理论,分别根据与目标药物和目标靶标的相关性对药物和靶标进行打分排序,基于已标记和未标记的药物-靶标相互作用信息进行标签传播,但该方法没有考虑模型的平滑性和数据几何分布的复杂性,且没有从药物-靶标相互作用网络角度分析药物和靶标之间的相似性,致使得到的数据中假阳率较高,因而在 AUPR 评价标准上,PreSLMF 的性能优于 NetCBP。

3.2 新的药物-靶标相互作用预测

令数据集中每种药物当作一种新的药物,删除所有与其发生相互作用的靶标,然后把此药物作为测试集,其他药物作为训练集,利用 PreSLMF 算法对药物与靶标之间的相互作用进行打分,分值越高发生相互作用的概率越高,并对此分值进行排序,从中提取前 30% 的数据,通过 KEGG, DrugBank, SuperTarget,ChEMBL 数据库检索验证。

酶中 PreSLMF 预测的药物-靶标相互作用中排序前 30% 的部分数据如表 4 所示,在酶中,靶标 hsa:5742 在前列腺素的合成、血小板的活化和聚集过程中起着至关重要的作用,D00097, D00109, D00549, D00827, D02290, D02350, D02709 和 D03716 等 61 种药物对此靶标有影响,本文基于 PreSLMF 方法对酶数据集中对每种药物与 664 个靶标之间的相互作用进行打分,预测 D00418, D03488, D01364 与此靶标有相互作用,且分别在数据库 KEGG, SuperTarget 和 DrugBank 中得到验证。

表 4 酶中 PreSLMF 预测的药物-靶标相互作用中排序前 30% 的部分数据

药物 ID	靶标 ID	数据来源
D00418	hsa:5742	DrugBank
D03488	hsa:5742	DrugBank
D01364	hsa:5742	SuperTarget,DrugBank
D03775	hsa:1636	KEGG
D03781	hsa:1586	KEGG
D00434	hsa:1559	KEGG
D05458	hsa:4128	DrugBank

离子通道中 PreSLMF 预测的药物-靶标相互作用中排序前 30% 的部分数据如表 5 所示,在离子通道中,药物 D00733 作用于靶标 hsa:6331 和 hsa:6338,这 2 个靶标对钠离子通道起着重要作用,影响肾、肺和肠腺等脏器中钠离子的重新吸收,实验预测到此药物还能作用于靶标 hsa:6328, hsa:6334 和 hsa:6336,且此药物与这 3 个靶标之间的相互作用在 KEGG 中能检索到,同时,在 UniProt 数据库中可检索到这 3 个靶标也与钠离子的渗透相关。

表 5 离子通道中 PreSLMF 预测的药物-靶标相互作用中排序前 30% 的部分数据

药物 ID	靶标 ID	数据来源
D00294	hsa:3767	KEGG
D00303	hsa:6326	DrugBank
D00303	hsa:6323	SuperTarget,DrugBank
D00732	hsa:6331	KEGG
D00726	hsa:170572	KEGG
D00726	hsa:200909	KEGG
D00733	hsa:6328	KEGG
D00733	hsa:6334	KEGG
D00733	hsa:6336	KEGG
D00642	hsa:6328	KEGG
D00642	hsa:6334	KEGG
D00617	hsa:2554	SuperTarget,drugBank
D05077	hsa:6328	KEGG

GPCRs 中 PreSLMF 预测的药物-靶标相互作用中排序前 30% 的部分数据如表 6 所示。在 GPCRs 中,药物 D00270 可作用于 hsa:150 和 hsa:151,这两个靶标通过 G 蛋白调解腺苷酸环化酶抑制儿茶酚胺的诱导,实验预测到 D00270 可作用于靶标 hsa:152,此靶标也有类似的作用。另外,D00270 也可作用于靶标 hsa:3356,此靶标影响神经元的感知、认知和情绪等活动,在肠道平滑肌收缩过程中发挥作用,且可能在动脉血管收缩中发挥作用。实验预测到 D00270 可作用于 hsa:3357 和 hsa:3358,并在 KEGG 中检索到了这两条相互作用;同时,hsa:3357 可调节多巴胺和 5-羟色胺的摄取和释放,通过 5-羟色胺和多巴胺的细胞外调节影响神经活动,从而在疼痛的感知方面发挥作用,hsa:3358 结合配体产生一种构象变化,这种变化能通过鸟嘌呤核苷酸结合蛋白(G 蛋白)触发信号,通过 G 蛋白和调节信号通路来刺激家庭神经元的活动。也就是说,hsa:3356、hsa:3357 和 hsa:3358 都可以作为各种药物和受体的精神物质。

表 6　GPCRs 中 PreSLMF 预测的药物-靶标相互作用中排序前 30% 的部分数据

药物 ID	靶标 ID	数据来源
D00283	hsa:3363	KEGG
D00283	hsa:1814	ChEMBL,DrugBank
D00454	hsa:3269	KEGG
D00498	hsa:4986	KEGG,DrugBank
D00528	hsa:134	KEGG
D00604	hsa:148	ChEMBL
D01462	hsa:3350	KEGG
D00270	hsa:152	KEGG
D00270	hsa:3357	KEGG
D00270	hsa:3358	KEGG
D01051	hsa:1813	KEGG
D02357	hsa:3358	ChEMBL
D04375	hsa:151	KEGG
D05113	hsa:4986	DrugBank

核受体中 PreSLMF 预测的药物-靶标相互作用中排序前 30% 的部分数据如表 7 所示。在核受体中,D00348 可作用于 hsa:5914,此靶标是视黄酸受体,调节在不同生物过程中目标基因的表达。实验预测到 D00348 还可作用于 hsa:5915 和 hsa:6258,这 2 个靶标都为视黄酸受体,前者结合异质二聚体,调节不同生物过程中目标基因的表达,主要充当目标基因表达的激活剂,同时与骨骼增长和生长板的功能有关;后者也调节不同生物过程中目标基因的表达。

表 7　核受体中 PreSLMF 预测的药物-靶标相互作用中排序前 30% 的部分数据

药物 ID	靶标 ID	数据来源
D00182	hsa:2099	ChEMBL
D00348	hsa:5915	ChEMBL,SuperTarget
D00348	hsa:6258	ChEMBL
D00554	hsa:2100	KEGG
D00528	hsa:134	KEGG
D00690	hsa:2908	KEGG

4　结束语

本文研究 4 类重要的药物-靶标相互作用网络(酶、离子通道、GPCRs 和核受体),目前,这 4 类药物-靶标网络中仅被识别出极少量的相互作用,还有大量的相互作用数据有待挖掘。因此,本文结合半监督学习算法,融合药物化合物的结构信息、靶标的序列信息及药物与靶标的相互作用信息,提出 PreSLMF 算法。基于文献[9]提供的 4 类标准数据集对药物-靶标相互作用进行预测,结果显示,该算法在性能上优于 DBSI,KBMF2K,NetCBF 等算法,且部分预测到的药物-靶标相互作用数据能在相关数据库中得到验证。今后还将融合更多的信息,如药物的副作用、靶标所对应的基因表达谱等信息来预测药物与靶标之间的相互作用关系。

参考文献

[1] 李　鹏. 基于蛋白质相互作用网络的药物靶标和药物重定位研究[D]. 北京:中国人民解放军军事医学科学院,2011.

[2] 严　鑫,丁　鹏,刘志红,等. 药物分子设计中的大数据问题[J]. 科学通报,2015,60(5/6):558-565.

[3] Mousavian Z, Masoudi-Nejad A. Drug-target Interaction Prediction via Chemogenomic Space: Learning-based Methods[J]. Expert Opinion on Drug Metabolism & Toxicology,2014,10(9):1273-1287.

[4] 赵明珠. 药物-靶标相互作用及药物对组合研究[D]. 上海:上海交通大学,2013.

[5] Zhao Mingzhu, Chang Haoteng, Zhou Qiang, et al. Predicting Protein-ligand Interactions Based on Chemical Preference Features with Its Application to New D-amino Acid Oxidase Inhibitor Discovery[J]. Current Pharmaceutical Design,2014,20(32):5202-5211.

[6] Keiser M J, Roth B L, Armbruster B N, et al. Relating Protein Pharmacology by Ligand Chemistry[J]. Nature Biotechnol,2007,25(2):197-206.

[7] Cheng A C, Coleman R G, Smyth K T, et al. Structure-based Maximal Affinity Model Predicts Small-molecule Druggability[J]. Nature Biotechnology, 2007, 25(1): 71-75.

[8] Zhu Shanfeng, Okuno Y, Tsujimoto G, et al. A Probabilistic Model for Mining Implicit 'Chemical Compound-Gene' Relations from Literature[J]. Bioinformatics,2005,21(S2):245-251.

[9] Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of Drug-target Interaction Networks from the Integration of Chemical and Genomic Spaces[J]. Bioinformatics, 2008,24(13):232-240.

[10] Bleakley K, Yamanishi Y. Supervised Prediction of Drug-target Interactions Using Bipartite Local Models[J]. Bioinformatics,2009,25(18):2397-2403.

[11] Yamanishi Y, Kotera M, Kanehisa M, et al. Drug-target Interaction Prediction from Chemical, Genomic and Pharmacological Data in an Integrated Framework[J]. Bioinformatics,2010,26(12):246-254.

[12] Gönen M. Predicting Drug-target Interactions from Chemical and Genomic Kernels Using Bayesian Matrix Factorization[J]. Bioinformatics, 2012, 28(18): 2304-2310.

[13] 谈　锐,陈秀宏. 半监督的局部保留投影降维方法[J]. 计算机工程,2012,38(6):181-183.

[14] 杜芳华,冀俊忠,吴晨生,等. 基于蚁群聚集信息素的半监督文本分类算法[J]. 计算机工程,2014,40(11):167-171.

[15] Xia Zheng, Wu Lingyun, Zhou Xiaobo, et al. Semi-supervised Drug-protein Interaction Prediction from Heterogeneous Biological Spaces[J]. BMC Systems Biology,2010,4(S2).

(下转第 229 页)

[6] Sun Deyu, Hu Weiling, Wu Wei, et al. Design of the Image-guided Biopsy Marking System for Gastroscopy[J]. Journal of Medical Systems, 2012, 36(5): 2909-2920.
[7] Liu Jiquan, Wang Bin, Hu Weiling, et al. A Non-invasive Navigation System for Retargeting Gastroscopic Lesions[J]. Biomedical Materials and Engineering, 2014, 24(6): 2673-2679.
[8] Wang Bin, Hu Weiling, Liu Jiquan, et al. Gastroscopic Image Graph: Application to Noninvasive Multitarget Tracking Under Gastroscopy[J]. Computational and Mathematical Methods in Medicine, 2014(1): 73-90.
[9] 张雄美,易昭湘,蔡幸福,等. 基于改进 SIFT 的 SAR 图像配准方法[J]. 计算机工程,2015,41(1):223-226.
[10] 刘　斌,孙　斌,余方超,等. 基于不可分小波分解的图像配准方法[J]. 计算机工程,2014,40(10):252-257.
[11] Mikolajczyk K, Schmid C. A Performance Evaluation of Local Descriptors[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(10): 1615-1630.
[12] Oliveira F, Tavares J. Medical Image Registration: A Review[J]. Computer Methods in Biomechanics and Biomedical Engineering, 2014, 17(2): 73-93.
[13] Leitner R, Martin D B, Arnold T, et al. Multi-spectral Video Endoscopy System for the Detection of Cancerous Tissue[J]. Pattern Recognition Letters, 2013, 34(1): 85-93.
[14] Zitova B, Flusser J. Image Registration Methods: A Survey[J]. Image and Vision Computing, 2003, 21(11): 977-1000.
[15] Crum W R, Hartkens T, Hill D. Non-rigid Image Registration: Theory and Practice[J]. British Journal of Radiology, 2014, 27(2): 140-153.
[16] Luong Q T, Faugeras D. The Fundamental Matrix: Theory, Algorithms, and Stability Analysis[J]. International Journal of Computer Vision, 1996, 17(1): 43-75.
[17] Kalal Z, Mikolajczyk K, Matas J. Forward-Backward Error: Automatic Detection of Tracking Failures[C]// Proceedings of the 20th International Conference on Pattern Recognition. Washington D. C., USA: IEEE Press, 2010: 2756-2759.
[18] Selka F, Nicolau S, Vincent A, et al. Evaluation of Endoscopic Image Enhancement for Feature Tracking: A New Validation Framework[M]//Liao Hongen, Linte C A, Masamune K, et al. Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions. Berlin, Germany: Springer-Verlag, 2013: 75-85.
[19] Rosten E, Drummond T. Machine Learning for High-speed Corner Detection[M]//Leonardis A, Bischof H, Pinz A. Computer Vision-ECCV 2006. Berlin, Germany: Springer-Verlag, 2006: 430-443.
[20] Lowe D G. Object Recognition from Local Scale-invariant Features[C]//Proceedings of the 7th IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 1999: 1150-1157.
[21] Bay H, Tuytelaars T, Gool L V. SURF: Speeded Up Robust Features[M]//Leonardis A, Bischof H, Pinz A. Computer Vision-ECCV 2006. Berlin, Germany: Springer-Verlag, 2006: 404-417.
[22] Shi J, Tomasi C. Good Features to Track[C]// Proceedings of CVPR'94. Washington D. C., USA: IEEE Press, 1994: 593-600.
[23] Se S, Lowe D, Little J. Global Localization Using Distinctive Visual Features[C]//Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. Washington D. C., USA: IEEE Press, 2002: 226-231.

编辑　金胡考

(上接第 223 页)

[16] Chen Xin, Liu Mingxi, Yan Guiying. Drug-target Interaction Prediction by Random Walk on the Heterogeneous Network[J]. Molecular BioSystems, 2012, 8(7): 1970-1978.
[17] Cheng Feixiong, Liu Chuang, Jiang Jing, et al. Prediction of Drug-target Interactions and Drug Repositioning via Network-based Inference[J]. PLoS Computational Biology, 2012, 8(5).
[18] Chen Hailin, Zhang Zuping. A Semi-supervised Method for Drug-target Interaction Prediction with Consistency in Nnetworks[J]. PloS ONE, 2013, 8(5).
[19] Kanehisa M, Goto S, Sato Y, et al. KEGG for Integration and Interpretation of Large-scale Molecular Data Sets[J]. Nucleic Acids Research, 2011, 40(Database Issue): 109-114.
[20] Knox C, Law V, Jewison T, et al. DrugBank 3.0: A Comprehensive Resource for 'Omics' Research on Drugs[J]. Nucleic Acids Research, 2011, 39(S1): 1035-1041.
[21] Hecker N, Ahmed J, von Eichborn J, et al. Super Target Goes Quantitative: Update on Drug-target Interactions[J]. Nucleic Acids Research, 2011, 40(Database Issue): 1113-1117.
[22] Gaulton A, Bellis L J, Bento A P, et al. ChEMBL: A Large-scale Bioactivity Database for Drug Discovery[J]. Nucleic Acids Research, 2012, 40(Database Issue): 1100-1107.
[23] Hattori M, Okuno Y, Goto S, et al. Development of a Chemical Structure Comparison Method for Integrated Analysis of Chemical and Genomic Information in the Metabolic Pathways[J]. Journal of the American Chemical Society, 2003, 125(39): 11853-11865.
[24] Smith T F, Waterman M S. Identification of Common Molecular Subsequences[J]. Journal of Molecular Biology, 1981, 147(1): 195-197.
[25] Yu Guoxian, Rangwala H, Domeniconi C, et al. Protein Function Prediction with Incomplete Annotations[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2014, 11(3): 579-591.
[26] Belkin M, Niyogi P, Sindhwani V. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples[J]. Journal of Machine Learning Research, 2006, 7(3): 2399-2434.

编辑　金胡考