小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0019
2026 年 2 月 第 2 期
Vol.47 No.2 2026

利用序列编码及稀疏学习的药物-miRNA 关联预测

任鹏飞,柳 迪,滕志霞

(东北林业大学 计算机与控制工程学院,哈尔滨 150040)
E-mail:tengzhixia@nefu.edu.cn

摘 要: 微小 RNA(miRNA)在许多人类复杂疾病的发生和发展中起着重要的作用,并被广泛认为是未来治疗疾病的有效药物靶点.然而,现有计算模型主要以基于生物学信息的药物(miRNA)间相似度进行学习,缺少其对应的序列特征;此外,目前的 miRNA 与药物的关联数据存在显著的稀疏性和噪声问题,进一步限制了预测模型的性能.针对上述问题,本文提出了一种基于序列特征编码和稀疏学习的药物-miRNA 关联预测模型 SESL.该模型通过深度学习编码器生成序列相似性矩阵,并利用稀疏学习方法对关联矩阵进行优化,最终结合有界核范数正则化实现关联预测.实验结果显示 SESL 在多种情况下优于现有方法,同时在数据更加稀疏的情况下仍保持较高的性能.

关 键 词: 药物;微小 RNA;关联预测;深度自编码器;稀疏学习
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)02-0386-08

Prediction of Drug-miRNA Associations Using Sequence Encoding and Sparse Learning

REN Pengfei, LIU Di, TENG Zhixia
(College of Computer and Control Engineering, Northeast Forestry University, Harbin 150040, China)

Abstract: MicroRNAs (miRNAs) have been widely recognized as potential drug targets for future therapeutic interventions since they play an essential role in the progression of many complex human diseases. However, existing computational models primarily rely on biological information to build drug (miRNA) similarity, while neglecting the sequence features. Additionally, the available association data between miRNAs and drugs is highly sparse and contains substantial noise, which poses significant challenges for accurate prediction. To address these issues, this paper proposes SESL, a model for drug-miRNA association prediction based on sequence feature encoding and sparse learning. The model employs a deep learning encoder to construct sequence similarity matrixs, utilizes sparse learning techniques to optimize the existing association matrix, and integrates bounded nuclear norm regularization to achieve association prediction. The experimental results demonstrate that SESL outperforms existing methods across multiple cases, while maintaining high performance even under conditions of increased data sparsity.

Keywords: drug; miRNA; association prediction; deep auto-encoder; sparse learning

0 引 言

微小 RNA(microRNA,简称 miRNA)是长度约为 22 个核苷酸的单链非编码 RNA,可以靶向调节信使 RNA 的翻译、表达或降解.这种独特的作用方式,使其在许多生命过程中起着重要作用[1].随着研究者的关注和高通量测序技术的出现,越来越多在人类基因表达中具有重要功能的 miRNA 已被确认[2-4],研究表明,许多疾病的发展都与某些特定 miRNA 的异常表达有关.例如,miR-205 和 miR-373 在结肠癌细胞中的表达水平明显高于正常细胞[5].miR-128 的过表达与白血病相关,并且会导致细胞活力降低,对依托泊苷的敏感性增加[6].因此,通过使用药物(包括激活剂和抑制剂)来调控(激活或者抑制)miRNA 的表达被广泛认为是一种治疗相关疾病有效的途径.换而言之,miRNA 是一种潜在的高价值药物靶点.因此,确定药物和 miRNA 之间的关联关系对新药研发具有重要的应用价值.例如,miR-let-7 被认为是一种肿瘤抑制因子,其过表达可抑制 Akt2 的表达并增强其对 5-FU 的敏感性[7].许多具有类似 miR-122 功能的药物激活剂和抑制剂被认为具有治疗丙型肝炎病毒感染的潜力[8].

早期的 miRNA-药物关联预测使用生物学质谱,荧光和报告基因方法[9-11],这些方法既费时又低效.为了解决这个问题,许多研究人员开发了预测这种关联的计算模型并取得了一些成果.

现有的计算方法可以分为两类:基于异构网络推理的预测方法和基于矩阵补全的预测方法.第 1 类方法首先整合药物综合相似度、miRNA 综合相似度,与已知的药物-miRNA 关联,利用这 3 个矩阵构建异构网络,然后应用不同的模型来学习节点特征和拓扑关系,最后进行药物-miRNA 关联预测.例

收稿日期:2025-01-09 收修改稿日期:2025-03-05 基金项目:黑龙江省自然科学基金联合引导项目(LH2024F001)资助;国家自然科学基金面上项目(62271132)资助. 作者简介:**任鹏飞**,男,1999 年生,硕士研究生,研究方向为生物信息学;**柳 迪**,男,1995 年生,博士研究生,研究方向为生物信息学;**滕志霞**(通信作者),女,1982 年生,博士,副教授,CCF 会员,研究方向为生物信息学.

如,SMiR-NBI 模型通过资源在网络中双向传播发现药物-miRNA 未知关联[12];GISMMA 方法通过计算网络中的子图相互作用数量获得药物-miRNA 关联预测得分[13];TLHNSMMA 方法构建了一个药物-miRNA-疾病三层异质网络,并且基于关联有罪原则(Guilt By Association,GBA)推断潜在药物-miRNA 关联关系[14]. 考虑到噪音的存在,SL-GISMMA 方法首先进行了数据增强,然后结合 GISMMA 进行关联预测[15];JMSS-MMA 利用自动编码器和基于概率分布的掩蔽策略的组合来成功抵消噪声数据的影响,从而能够准确预测潜在的药物与 miRNA 关联[16].

第 2 类方法将药物与 miRNA 关联预测作为矩阵元素补全问题,采用矩阵分解等方法. GNMFDMA 方法首先重建关联矩阵,然后基于拉普拉斯图预测关联正则化协同矩阵分解[17]. TSPN 方法通过最小化邻接矩阵的截断 schatten p-范数来构建预测模型,并开发了高效的迭代算法框架来求解[18];RPCAΓNR 构建了一个基于 γ 范数正则化的主成分分析框架,进行求解并获得预测得分[19].

尽管目前的研究方法已经取得了很大进展,但通过深入分析可以发现,已有方法仍然面临两个挑战. 首先,绝大多数方法在数据准备阶段均采用了基于生物学信息的药物和 miRNA 综合相似度,这种生物学相似数据对于基于深度学习和机器学习的计算方法来说缺乏可解释性,也忽略了药物和 miRNA 本身作为序列数据所具备的特征,这会造成潜在嵌入特征利用不充分的问题. 其次,正如以上许多方法都有所提及的,可用的先验 miRNA-药物关联仍旧是十分稀疏的,存在许多未经验证的关联,这对于预测模型而言是噪声,严重影响了模型的预测结果.

自动编码器是一种基于基于深度学习的架构,该架构由编码器-解码器架构的两个联合子模块组成. 编码器将输入转换为特征向量,而解码器将其转换回原始形式. 根据以往的研究可知[20],药物一般表示为 SMILES(Simplified Molecular Input Line Entry System,简称 SMILES)序列数据,并可以通过深度学习编码器-解码器模型,例如循环神经网络(RNN)框架进行建模. 此外,miRNA 序列也可以使用这种模型进行编码[21]. 深度学习药物编码在分子性质预测等方面已经得到有效应用,但在药物-miRNA 关联预测方面却很少使用. Huang 等人首次将图卷积技术与自动编码器结合,开发了预测模型 GCMDR[22]. 其核心在于利用图卷积运算从药物和 miRNA 数据以及药物-miRNA 关联数据中提取特征,从而实现一种端到端的预测方法. 通过构建包含药物、miRNA 及其相互关联的图结构,并应用图卷积运算,模型能够学习到节点的有效表示,进而预测它们之间的关联. 然而,miRNA 和药物的高维属性信息虽然为模型提供了丰富的特征,但也增加了模型的复杂度,导致计算效率低下;同时,该模型没有考虑噪声的影响,缺少降噪处理,影响模型的预测准确性.

基于上述研究的启发,为了解决提出的两个问题,本研究创新性的提出了一种新的预测模型,并将其命名为 SESL(Sequence Encoding and Sparse Learning),该方法的流程如图 1所示. 本方法的核心思想是利用深度学习编码器 LSTMAutoEnc(Long Short-Term Memory Auto-Encoder)[21]获得的基于序列的药物和 miRNA 特征向量构建其各自的相似度矩阵,充分利用被以往的研究忽略的基于序列的相似度信息,并且摆脱对于生物学信息的依赖.

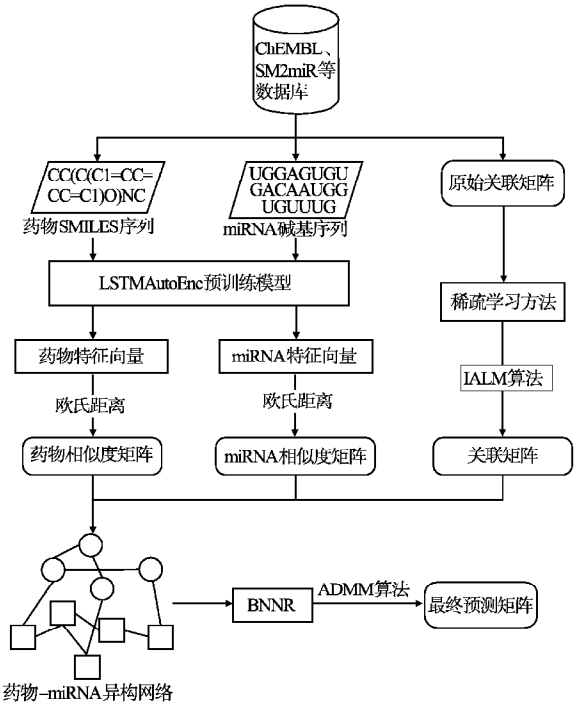

图 1　SESL 模型框架图

Fig. 1　Framework of SESL

本研究的主要贡献有以下几个方面:1)本研究利用了基于序列数据的嵌入特征,提高了预测结果的准确性,同时也摆脱了对于生物学信息的依赖;2)稀疏学习方法的加入有效缓解了先验药物与 miRNA 关联数据的稀疏性,降低了噪声带来的负面影响.

1　问题定义

给定$\{(m_i,s_i),l_i\}$,其中(m_i,s_i)是一个药物-miRNA 关联对,l_i 是他们的关联标签,代表存在关联或不存在关联. 药物 s_i 由 SMILES 序列表示,miRNA m_i 由碱基序列表示. 该问题的主要目标是设计一个方法,以药物-miRNA 关联对作为输入,预测其关联标签.

2　方　法

本节将介绍所提出的 SESL 方法,用于进行基于序列特征的药物-miRNA 关联预测. 首先,本研究通过 LSTMAutoEnc 获取可靠的基于药物 SMILES 序列数据以及 miRNA 序列数据进行编码后的低维特征向量;随后,利用欧式距离对药物(miRNA)进行相似性度量,并构建药物相似度矩阵和 miRNA 相似度矩阵. 同时,本方法利用稀疏学习方法处理已知的药物与 miRNA 关联矩阵,减小噪声数据的负面影响. 最后,将上述药物相似度矩阵、miRNA 相似度矩阵和药物-miRNA 关联矩阵组成异构网络,通过有界核范数正则化方法填充关联矩阵的缺失元素,得到最终的预测结果.

2.1 药物与 miRNA 特征向量获取

本研究采用 LSTMAutoEnc 获取药物与 miRNA 的特征向量. 对于 miRNA,LSTMAutoEnc 训练编码器的数据集来自公开的 RNA Central 数据库(https://rnacentral.org/),其中包含来自各种生物体的最新非编码 RNA 序列. LSTMAutoEnc 提取了 35757 个小调控人类非编码 RNA 序列,包括 3752 个 miRNA 和 32005 个 piRNA(Piwi-interacting RNA). 对于药物,LSTMAutoEnc 使用从 ChEMBL 数据库[23](https://www.ebi.ac.uk/chembl/)检索的大约 70 万个 SMILES 进行训练.

LSTM 编码器基于循环神经网络架构,通过其独特的细胞状态(cell)和门控机制(输入门、遗忘门和输出门),能够有效地保留和更新序列中的关键信息. LSTMAutoEnc 接收药物 SMILES 序列数据和 miRNA 序列数据作为输入. 其中药物 SMILES 序列数据以字符串的形式表示不同的药物分子序列,如图 1所示,由一系列化学原子和化学键组成;而 miRNA 是由腺嘌呤(A)、胞嘧啶(C)、鸟嘌呤(G)和尿嘧啶(U)4 种碱基构成的,长度约为 22 ~ 24 个字节的字符串序列. LSTMAutoEnc 流程图如图 2 所示.

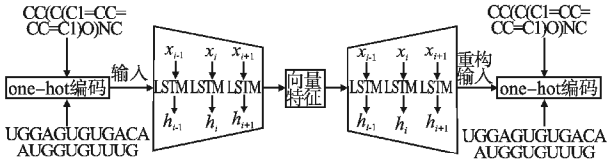

图 2 LSTM 序列自动编码器框架图

Fig. 2 Framework of LSTM sequence auto-encoder

输入的 miRNA 序列和药物 SMILES 被表示为独热(one-hot)编码,由编码器编码成压缩的低维特征向量(64 维),解码器部分从编码中重构输入以验证编码质量. 该网络使用 ReLU 激活函数接受一个分子描述符向量作为一组 6 个全连接层的输入,每个层 256 个单元. 每个单独的全连接层的输出用于设置网络中每个循环层的 cell 状态或 hidden 状态. 网络共有 3 个单向循环层,每层由 256 个 LSTM 神经元组成. 使用 softmax 激活,将最终 LSTM 层的输出馈送到具有 35 个单元的前馈层,并将批处理归一化应用于所有 LSTM 和除最后一个全连接层外的所有输出. 最终获取了基于序列特征编码的 64 维度药物特征向量和 64 维 miRNA 特征向量.

2.2 相似度矩阵构建

对于获得的 64 维药物特征向量和 64 维 miRNA 特征向量,本研究考虑其具有同维度、非负性的特点,决定采用欧氏距离进行相似度计算. 欧式距离是最常见的距离度量方式之一,用于计算多维空间中为向量之间的直线距离. 对于两个 64 维向量 E_i^s 和 $E_i^{s'}$,它们的欧式距离 di 可以通过公式(1)计算:

$$di_{s,s'} = \sqrt{\sum_{i=1}^{64}(E_i^s - E_i^{s'})^2} \tag{1}$$

经过计算药物(miRNA)两两之间的距离,以此来进行相似性度量,本研究最终得到了药物相似度矩阵 S_s 和 miRNA 相似度矩阵 S_m.

2.3 基于稀疏学习方法的关联矩阵优化

在获取了药物与 miRNA 各自的基于序列特征的相似度矩阵后,考虑到药物-miRNA 关联预测问题面临的先验数据稀疏,噪声较多的问题,本研究对原始的药物-miRNA 关联矩阵 A 进行优化.

药物-miRNA 关联矩阵是一个稀疏矩阵,其中绝大多数元素的值为 0,但有些标签是假阴性或假阳性的,可将其视为噪音. 它们的存在对预测有负面影响. 通过前人的研究[24],本研究发现利用稀疏学习方法可以将原本稀疏、噪声较多的矩阵进行降噪,获取具有重要隐藏特征的低秩矩阵,从而有效降低假阴性噪声带来的负面影响. 因此,本文采用稀疏学习方法,具体做法如公式(2):将原始的药物-miRNA 关联矩阵 A 分解为两部分[25]:矩阵 A 与具有重要隐藏特征的低秩矩阵 X 的线性组合;包含大量噪声的稀疏矩阵 P(绝大多数元素值为 0).

$$A = A \times X + P \tag{2}$$

该方程有无限个解,本研究考虑利用矩阵范数进行约束. 最小化矩阵的核范数有利于获得秩较低的矩阵,相应的稀疏范数有利于识别噪声. 因此为了得到一个低秩矩阵 X 和一个稀疏矩阵 P,可以将公式(2)变换为公式(3):

$$\begin{aligned} &\min_{X,P} \|X\|_* + \alpha \|P\|_{2,1} \\ &s.t.\ A = A \times X + P \end{aligned} \tag{3}$$

其中 $\|X\|_* = \sum_i \sigma_i$,$\sigma_i$ 是矩阵 X 的奇异值,该核范数用以约束 X 为低秩矩阵;$\|P\|_{2,1} = \sum_{j=1}^{n}\sqrt{\sum_{j=1}^{n}(P_{ij})^2}$,该稀疏范数用以约束 P 为稀疏矩阵. α 是权重参数,用来平衡这两个范数之间的比重,设置为 0.3.

根据前人的研究[26],如果将公式(3)中的 A 转化为单位矩阵,则模型退化为鲁棒主成分分析(Robust Principal Component Analysis,RPCA). 这是一个具有约束的凸优化问题,因此公式(3)可等价改写为公式(4):

$$\begin{aligned} &\min_{X,P,J} \|J\|_* + \alpha \|P\|_{2,1} \\ &s.t.\ A = A \times X + P, X = J \end{aligned} \tag{4}$$

公式(4)有许多方法可以求解,本文利用非精确增广拉格朗日乘子(Inexact Augmented Lagrange Multipliers,IALM)法来解决该问题,具体如公式(5)所示:

$$\begin{aligned} L = \|J\|_* + \alpha \|P\|_{2,1} + tr(Y_1^T(A - A \times X - P)) + \\ tr(Y_2^T(X - J)) + \frac{\gamma}{2}(\|A - A \times X - P\|_F^2 + \|X - J\|_F^2) \end{aligned} \tag{5}$$

其中 $\gamma > 0$ 为惩罚项,本研究根据 IALM 算法,固定其他变量,并通过更新拉格朗日乘子 Y_1,Y_2,求解了 J、X 和 P 的最小值,并将 X^* 和 P^* 定义为公式(5)的解. X^* 可视为 miRNA 或药物的相似性矩阵,P^* 代表噪声矩阵.

需要注意的是,此处的 X^* 是由算法解得的具有隐藏特征的低秩矩阵,虽然在形式上与 miRNA 或药物相似度矩阵相同,但与 2.2 节中的相似度矩阵没有关系. X^* 的作用是帮助重构新的关联矩阵 A_1. 最终,将稀疏学习后的新关联矩阵表示为公式(6):

$$A_1 = A \times X^* \tag{6}$$

经过以上过程,本方法得到了基于序列特征构建的药物相似度矩阵 S_s,miRNA 相似度矩阵 S_m,以及进行稀疏学习后的药物-miRNA 关联矩阵 A_1. 随后,本研究将药物-miRNA 关联预测问题视为矩阵补全问题,并利用有界核范数正则化(Bounded Nuclear Norm Regularization,BNNR)方法[27]进行解决.

2.4　基于有界核范数正则化的关联预测

BNNR 方法的主要作用是对矩阵 A_1 进行矩阵补全,其优势是在矩阵补全的过程中,通过引入正则化项来平衡近似误差和矩阵的秩特性. 具体来说,它并不是严格地拟合所有已知元素,而是允许存在一定的噪声,增强了模型的泛化能力;同时,通过正则化项来抑制这些噪声的影响,从而高预测的准确性. 该方法流程如下:

首先,利用 S_s、S_m 和 A_1 构建一个异构药物-miRNA 网络,并定义了一个目标矩阵 T 来表示它,初始目标矩阵 T 表示为公式(7):

$$T=\begin{bmatrix} S_s & A_1 \\ A_1^T & S_m \end{bmatrix} \tag{7}$$

其次,基于前人的研究发现[28],矩阵补全问题可转化为目标矩阵秩最小化问题. 而秩最小化问题在计算上是 NP 困难的,因此,可将公式(7)转化为等价的核范数优化问题[29],如公式(8)所示:

$$\min_{T^*} \|T^*\|_* \quad s.t.\ 0\leqslant T_{ij}^*\leqslant 1 \tag{8}$$

其中 $\|T^*\|_*$ 代表最终目标矩阵 T^* 的核范数. 显然,核范数的最小化是一个典型的凸优化问题,求解难度大大降低. 随后,考虑到数据中存在的噪声,加入噪声约束项,得到了 BNNR 的关键公式如公式(9)所示:

$$\min_{T^*} \|T^*\|_* + \frac{\varepsilon}{2}\|R_\Omega(T^*)-R_\Omega(T)\|_F^2, \quad s.t.\ 0\leqslant T_{ij}^*\leqslant 1 \tag{9}$$

Ω 是已知存在关联关系的药物-miRNA 对的坐标集合,R_Ω 是其正交投影算子. ε 用作误差项的权重参数,方程(9)的第 2 项设置了噪声约束并将其放宽为一个正则化项[30,31],它不仅可以最小化数据中噪声的负面影响,由于正则化项本身的特点,该方法也解决了过拟合问题.

该方程由交替方向乘子法(Alternating Direction Method of Multipliers,ADMM)求解[32],经过迭代计算后,最终得到未知元素均被填充完整的关联得分矩阵 A_1^*,即更新后的最终目标矩阵 T^* 对应的部分.

3　实验结果与分析

3.1　数据集与评价指标选取

本文沿用了大多数模型统一使用的开源的公共数据集进行多种实验来评估模型的预测性能. 数据集 1 包含 831 个药物和 541 个 miRNA,以及 664 个药物-miRNA 关联关系,由于绝大多数的药物-miRNA 对不存在关联关系(稀疏度为 0.99),因此将其命名为部分相关数据集 PCD(Partial Correlation Dataset). 所有的药物-miRNA 关联关系都是从 SM2miR[33]数据库获取的,其中药物是从 SM2miR[33]、DrugBank[34] 和 PubChem[35]3 个数据库收集来的,而 miRNA 则来自 SM2miR、HMDD[36]、miR2Disease[37]和 PhenomiR[38]4 个数据库. 数据集 2 是在 PCD 的基础上,去除掉没有任何已知关联关系的 792 个药物和 255 个 miRNA,得到的由 39 个药物和 286 个 miRNA 组成的数据集,本研究将其命名为全相关数据集 FCD(Full Correlation Dataset),稀疏度为 0.94. PCD 与 FCD 的规模和稀疏度都有所不同,可以从不同角度验证模型的性能,其中已确认的药物-miRNA 关联关系为正样本,其余未确认的视作负样本.

在对比实验部分,由于先验关联数据少,该领域研究更加注重在贴合现实场景的 3 种不同实验设置下将 ROC 曲线下面积(the Area Under ROC Curves,AUC)进行对比. 另外,本研究也使用包括 PR 曲线下面积(AUPR)、精确度(Precision)、召回率(Recall)、F1 分数(F1-score)、准确率(Accuracy)和特异度(Specificity)6 种相关性能指标进行综合评价.

3.2　对比实验

本研究与其他的 7 个模型 RPCAΓNR[19]、SNMFSMMA[39]、TSPN[18]、DCMF[40]、EKRRSMMA[41]、GISMMA[13]以及 TLHNSMMA[14] 在两个数据集上进行了多种对比实验. 为了验证 SESL 的预测性能,本研究进行了 3 种不同类型的留一交叉验证以及常规五折交叉验证. 这些实验的结果如表 1 所示.

表 1　在 PCD 和 FCD 数据集上的对比结果

Table 1　Comparison results on PCD and FCD dataset

方法	PCD					FCD				
	全局	固定 miRNA	固定药物	五折	*P* 值	全局	固定 miRNA	固定药物	五折	*P* 值
SESL	0.9808	0.9778	**0.8754**	0.9772 ±0.003	\	**0.9618**	**0.9574**	**0.9507**	**0.9458 ±0.0016**	\
RPCAΓNR	**0.9959**	**0.9937**	0.8628	**0.9958 ±0.0005**	***	0.9534	0.9410	0.9501	0.9449 ±0.0032	*
SNMFSMMA	0.9711	0.9698	0.8329	0.9644 ±0.0034	***	0.8895	0.8884	0.7651	0.8814 ±0.0033	***
TSPN	0.9938	0.9902	0.8162	0.9934 ±0.0004	***	0.8925	0.8915	0.7374	0.8834 ±0.0037	***
DCMF	0.9868	0.9833	0.8377	0.9806 ±0.0029	**	0.8770	0.8836	0.7705	0.8632 ±0.0041	***
EKRRSMMA	0.9799	0.9731	0.8071	0.9767 ±0.0014	**	0.8869	0.8586	0.7706	0.8560 ±0.0027	***
GISMMA	0.9291	0.9505	0.7702	0.9263 ±0.0026	***	0.8209	0.8639	0.6593	0.8088 ±0.0044	***
TLHNSMMA	0.9859	0.9845	0.7645	0.9758 ±0.0029	×	0.8148	0.8245	0.6055	0.8168 ±0.0022	***

注:加粗代表该列的最优值,下划线代表该列的次优值

具体而言,留一交叉验证的实验设置分别是全局验证、固定 miRNA 验证和固定药物验证. 在全局验证中,每个已知的药物-miRNA 关联依次被视为测试样本,而其他 663 对确认的关联被视为训练样本,其他没有已知关联的药物-miRNA 对被认为是候选样本. 固定 miRNA 验证和固定药物验证与全局验证之间的区别在于这两种验证仅选取与所选 miRNA(药物)有关的未知关联对作为候选样本. 在五折交叉实验中,所有已知的药物-miRNA 关联首先被随机分为 5 个大小

基本相等的子集. 每个子集依次作为测试集,其余 4 个子集用于模型训练,其他未经确认的药物- miRNA 对被视为候选样本;然后,将测试集中每个药物-miRNA 关联的得分与候选样本的得分进行比较.

在 3 种实验设置下的留一交叉验证结果表明,在 PCD 上,SESL 在全局验证和固定 miRNA 验证两种情况下的表现略低于 RPCAΓNR,在固定药物验证的情况下,SESL 的 AUC 高于其他所有模型;而在 FCD 上,SESL 在 3 种实验设置下的性能均优于其他所有模型. 在五折交叉验证中 SESL 在 PCD 中排名居中,略低于 RPCAΓNR 和 TSPN;在 FCD 中的表现突出,性能优于其他所有模型.

同时,为了更好地比较 SESL 与其他模型之间的差异性,本研究基于两个数据集的五折交叉验证结果,对 SESL 与其他被测试模型进行假设检验. 本研究对结果进行 t 检验,以评估假设检验的显著性. 当 P 值小于 0.05 时,一般认为被检验模型具有显著性. 为了便于观察本研究用符号 * 代表模型的显著程度, * 表示 P 值小于 0.05, ** 表示 P 值小于 0.01, *** 表示 P 值小于 0.001,而 × 则代表模型对比无显著性. 结果如表 1 所示,通过观察可知,SESL 在 FCD 上具有明显的差异性,表现良好,本研究推测这可能是因为稀疏学习模块在小规模数据集中发挥了更好的作用.

为了从不同角度对 SESL 进行全面衡量,本研究基于两个数据集,采用多种指标,与基准模型 BNNRSMMA[27] 进行对比. BNNRSMMA 使用了生物学信息构建相似度矩阵,同样也采用 BNNR 算法进行预测. 本研究取 10 次五折交叉验证的平均结果作为最终结果,如图 3 所示.

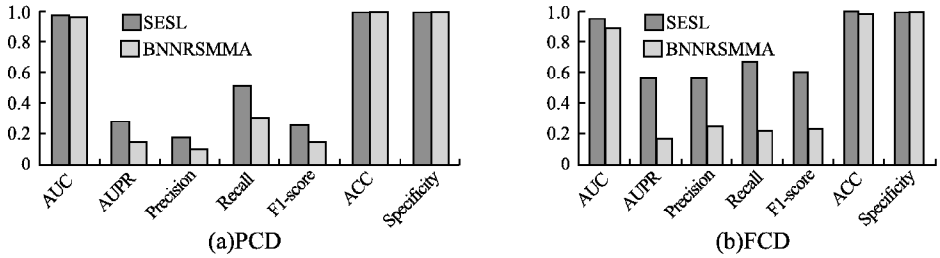

图 3 在 PCD 和 FCD 数据集上的多指标对比结果

Fig. 3 Comparison results of multi-indicators on PCD and FCD dataset

通过分析图 3 可以看出,SESL 在所有指标上均优于基准模型 BNNRSMMA,尤其在 FCD 数据集上,表现更加明显. 本研究推测,这可能是由于本方法中的序列特征与稀疏学习模块既充分利用了序列信息,也对先验的关联矩阵进行了降噪,两部分联合作用,提高了预测性能.

最后,为了证明 SESL 所采用的 LSTM 深度自编码器的有效性,本研究将其与目前流行的大语言模型(Large Language Model)进行对比. 具体而言,本研究所采用的大模型为预训练的 BERT 模型[42] "Bert-base-uncased",它基于谷歌开发 Transforme[43] 架构,采用双向编码器表征(Bidirectional Encoder Representations from Transformers). 与传统的单向语言模型不同,BERT 可以同时从输入文本的前后两个方向学习上下文信息,能够更全面地捕捉文本中的语义依赖关系,从而更精准地把握语义. 本研究使用 Bert-base-uncased 模型基于 FCD 数据集获取药物与 miRNA 的 64 维特征向量,用相同的方式获取相似度矩阵并预测,最后进行五折交叉验证,与基于 LSTM 深度自编码器的模型 SESL 进行多指标对比. 取 10 次五折交叉验证的平均结果作为最终结果,对比结果如表 2 所示.

表 2 LSTM 编码与 BERT 编码的多指标对比结果

Table 2 Comparison results of multi-indicators between LSTM encoding and BERT encoding

编码器	AUC	AUPR	Precision	Recall	F1-score	Acc	Specificity
LSTM	0.9458	**0.5605**	**0.5565**	**0.6657**	**0.5978**	**0.9891**	**0.9934**
BERT	**0.9521**	0.5254	0.5335	0.6575	0.5868	0.9885	0.9926

通过对表 2 多个指标上的对比分析,可以看出:BERT 的 AUC 略高于 LSTM,这表明 BERT 在分类能力上表现较好;而 LSTM 的 AUPR 明显高于 BERT,这表明 LSTM 综合性能良好,识别正样本的能力更强;在 Precision、Recal、F1-score、Acc 和 Specificity 方面,二者表现相近,LSTM 在这些指标上的表现略优于 BERT. 本文推测,这可能是由于药物与 miRNA 的长度相对较短,仅包含几十个字符,从而使得大模型的长距离依赖捕捉能力无法得到充分展现.

此外,与大语言模型相比,LSTM 编码的优势还体现在低资源消耗方面,由于其结构相对简单,通常需要的计算资源和内存较少. 这使得 LSTM 在资源受限的环境下,具有更高的可行性和实用性.

3.3 稀疏学习有效性分析

为了评估稀疏学习模块对于模型预测性能的提升,本研究进行了消融实验. 具体做法如下,本研究分别在带有稀疏学习模块和去除稀疏学习模块的情况下,基于 PCD 和 FCD 两个数据集进行全局验证、固定 miRNA 验证、固定药物验证和五折交叉验证,并将它们各自的 AUC 值进行对比,其中 SE 为去除稀疏学习模块的情况,SESL 为带有稀疏学习模块的情况. 实验结果如表 3 所示.

表 3 稀疏学习模块有效性实验结果

Table 3 Experimental results on the effectiveness of sparse learning modules

数据集	方法	全局	固定 miRNA	固定药物	五折
PCD	SESL	**0.9808**	**0.9777**	**0.8754**	**0.9772**
	SE	0.9590	0.9532	0.7848	0.9733
FCD	SESL	**0.9618**	**0.9574**	**0.9507**	**0.9439**
	SE	0.8709	0.8562	0.7309	0.8362

分析结果可以得知,基于两个数据集,在具有稀疏学习方

法模块的情况下,3 种不同实验设置的留一交叉验证的 AUC 值和五折交叉验证的 AUC 值均高于去除稀疏学习模块情况下的 AUC 值,且这种变化在 FCD 中更加明显,这证明了稀疏学习模块的有效性.

3.4　序列特征有效性分析

由于本研究的核心思路之一是利用深度学习编码器提取的被以往的研究忽略的序列特征向量计算相似度,来代替大多数模型采用的基于生物学信息的相似度,以此来摆脱对生物信息的依赖. 因此,为了证明提出的方法的有效性,本研究在去除稀疏学习模块的情况下,利用 FCD 数据集对两种不同类型的相似度进行全局验证,并对 AUC 和 AUPR 进行对比.

具体来说,本研究进行对比的生物学相似度是被大多数模型广泛采用的药物综合相似度和 miRNA 综合相似度. 药物综合相似度融合了 4 种常见相似性来计算药物综合相似度,避免了单一相似性可能带来的误差,4 个相似性分别为基于功能一致性的相似性 S_S^T[44],基于化学结构的相似性 S_S^C[45],基于副作用的相似性 S_S^S[45] 以及基于适应症表型的相似性 S_S^D[46],计算这 4 种相似性的平均值作为药物综合相似度. 与药物综合相似度的计算方法类似,采用相同的方法来计算 miRNA 之间的两种类型的相似性,分别是基于功能一致性的相似性 S_M^T[44] 和基于适应症表型的相似性 S_M^D[46].

经过以上两个过程,得到了基于生物学信息的药物综合相似度矩阵与 miRNA 综合相似度矩阵. 对比结果如图 4 所示,S-Seq 与 S-Bio 分别代表基于序列的相似度与基于生物学信息的相似度情况.

经过分析可知,在均未使用稀疏学习方法的情况下,基于序列特征的相似度构建的预测模型在 FCD 上的 AUC 和 AUPR 分别为 0.8709 与 0.4035,而同等条件下基于生物学信息的预测模型,其 AUC 和 AUPR 分别为 0.8434 和 0.3882. 这表明了基于序列特征的相似性比生物学相似性更有助于提高预测性能. 值得注意的是,本研究仅凭借药物和 miRNA 的序列信息便完成了预测,充分证明了模型的泛化性能. 本研究推测,这是因为基于序列特征所构建的相似度比基于生物学信息的相似度更具有可解释性,从计算角度来说,更加适合深度学习框架. 同时,该也表明了构建基于序列的特征是一个非常具有潜在研究价值的课题.

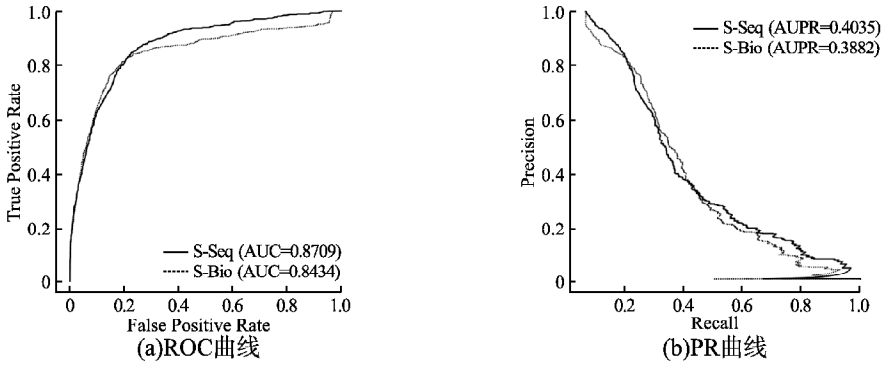

图 4　序列特征与生物学特征对模型性能的影响对比

Fig. 4　Comparison of the impact of sequence features and biological features on model performance

3.5　关联矩阵稀疏度敏感性分析

为了验证本模型对关联矩阵稀疏度的敏感性,本研究从已知的 664 个关联当中随机抽样构建新的关联矩阵. 具体来说,本研究基于 PCD 与 FCD 两个数据集随机选择已知关联的 50% 和 10%,构建新的关联矩阵. 基于新生成的关联矩阵,本研究进行了 10 次 5 折交叉验证,并对 AUC 进行评估,分析结果如图 5 所示.

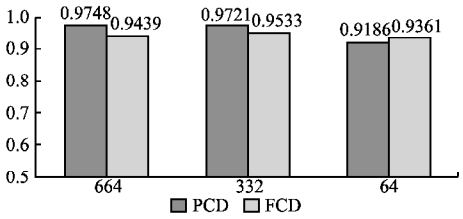

图 5　关联矩阵稀疏度敏感性实验结果

Fig. 5　Experimental results on the sparsity sensitivity of the correlation matrix

结果表明,SESL 算法的性能会随着关联矩阵稀疏度的增加而降低. 这说明已知药物-miRNA 关联的数量是限制预测性能的主要因素. 然而,SESL 在仅有 64 个已知关联的 5 折交叉验证下的 AUC 仍然能达到 0.9186(0.9361),这表明 SESL 在小样本情况下,仍能保持较高的预测准确率,具有可靠的性能.

3.6　案例分析

为了进一步验证 SESL 在现实场景中的应用,本研究进行了两种类型的案例分析. 在第 1 种类型中,根据 SESL 预测的结果,将预测得分降序排列,选出前 10 个不存在已知关联的药物-miRNA 对,之后统计经过文献验证的关联数量. 结果如表 4 所示,在前 10 个预测中有 7 个得到了验证,表 4 中 4 列为相关文献的 PubMed ID,NA 则代表未找到相关证明文献.

在第 2 种类型中,案例分析针对特定的药物,用以评估 SESL 在预测所研究药物的潜在关联 miRNA 的能力. 具体而言,本研究删除了与所研究药物相关的所有已知关联,将其视作一个全新的药物进行预测,之后将预测结果进行降序排列,选取排名前 20 的预测关联对,进行统计,最终观测其中有多少得到了文献验证. 具体来说,本研究选择依诺沙星(Enoxacin,CID:3229)和雌二醇(Estradiol,CID:5757)作为研究药

物,验证结果分别由表5与表6所示.

表4 在PCD数据集中得分前10的预测结果

Table 4 Top 10 predicted results on PCD dataset

预测排名	药物CID	miRNA	PubMed ID
1	CID:36462	hsa-mir-128-1	24846063
2	CID:36462	hsa-mir-128-2	24846063
3	CID:3385	hsa-let-7b	25951903
4	CID:5757	hsa-mir-125b-2	30609807
5	CID:3385	hsa-let-7i	NA
6	CID:3385	hsa-mir-125b-1	24865963
7	CID:3385	hsa-mir-125b-2	24846940
8	CID:3229	hsa-let-7b	NA
9	CID:5757	hsa-mir-125b-1	36575629
10	CID:5757	hsa-mir-126	NA

注:NA表示未找到相关证明文献

表5 在PCD数据集中对于依诺沙星的前20预测结果

Table 5 Top 20 predicted results of Enoxacin on PCD dataset

预测排名	miRNA	PubMed ID	预测排名	miRNA	PubMed ID
1	hsa-mir-124-1	26198104	11	hsa-mir-214	30385810
2	hsa-mir-124-2	26198104	12	hsa-mir-125a	26198104
3	hsa-mir-124-3	26198104	13	hsa-let-7a-1	26198104
4	hsa-mir-520f	NA	14	hsa-mir-34b	29986212
5	hsa-mir-18a	26198104	15	hsa-mir-26b	NA
6	hsa-mir-449b	NA	16	hsa-mir-19a	NA
7	hsa-mir-181b-1	26198104	17	hsa-mir-34c	29986212
8	hsa-mir-663a	NA	18	hsa-let-7i	23220571
9	hsa-mir-181a-2	26198104	19	hsa-mir-34a	33268375
10	hsa-mir-126	NA	20	hsa-mir-221	NA

注:NA表示未找到相关证明文献

表6 在PCD数据集中对于雌二醇的前20预测结果

Table 6 Top 20 predicted results of Estradiol on PCD dataset

预测排名	miRNA	PubMed ID	预测排名	miRNA	PubMed ID
1	hsa-mir-375	27030099	11	hsa-mir-9-3	26198104
2	hsa-mir-29a	22334722	12	hsa-mir-17	23220571
3	hsa-mir-27a	26198104	13	hsa-mir-19a	29416771
4	hsa-mir-22	24715036	14	hsa-let-7e	23220571
5	hsa-mir-124-1	23220571	15	hsa-let-7d	23220571
6	hsa-mir-124-3	23220571	16	hsa-mir-128-2	23220571
7	hsa-mir-630	NA	17	hsa-mir-26b	24735615
8	hsa-mir-150	NA	18	hsa-mir-30c-2	NA
9	hsa-mir-92a-1	NA	19	hsa-mir-200c	23220571
10	hsa-mir-23a	26198104	20	hsa-mir-1469	NA

注:NA表示未找到相关证明文献

在对依诺沙星的案例分析中,预测得分前20的预测结果有13个得到了文献验证,证明其与依诺沙星存在关联关系;而对雌二醇的案例分析中,前20个预测结果则有15个得到了验证.两种不同类型的案例分析均表明,SESL在实际的预测场景中具有可靠性,能够成为预测药物与miRNA关联的有效工具.

4 结 论

本文提出了一种新的药物-miRNA关联预测模型SESL,该模型是基于序列特征编码和稀疏学习来实现潜在药物-miRNA关联预测任务的.相比与大多数模型利用基于生物学信息的相似度进行推断预测,该模型充分利用了序列特征,并以此为基础建立相似度矩阵进行预测,摆脱了对于生物学信息的依赖;同时,稀疏学习模块的加入降低了噪声的负面影响.具体而言,SESL模型利用LSTMAutoEnc编码器预训练后的药物与miRNA特征向量构建相似度矩阵.随后,利用稀疏学习方法对已知的药物-miRNA关联矩阵进行降噪处理.最后,通过BNNR获得最终的预测结果.在两个不同数据集合上的多种实验以及案例分析都证明了SESL具有良好的性能,是预测药物-miRNA关联的可靠工具.

尽管SESL具有良好的性能,但它仍然存在一些局限性.首先,目前已知的药物-miRNA关联仍然很少,这是限制预测性能的主要因素.随着关联信息的增加,模型预测精度将进一步提高.此外,可以考虑利用其他关联关系,如长链非编码RNA或疾病,构建更大的异构网络,进行进一步的研究.

References:

[1] Gorbea C, Mosbruger T, Cazalla D. A viral Sm-class RNA base-pairs with mRNAs and recruits microRNAs to inhibit apoptosis [J]. Nature, 2017, 550(7675): 275-279.

[2] Denzler R, Mcgeary S E, Title A C, et al. Impact of MicroRNA levels, target-site complementarity, and cooperativity on competing endogenous RNA-Regulated gene expression [J]. Molecular Cell, 2016, 64(3): 565-579.

[3] Thomou T, Mori M A, Dreyfuss J M, et al. Adipose-derived circulating miRNAs regulate gene expression in other tissues [J]. Nature, 2017, 542(7642): 450-455.

[4] Tagliafierro L, Glenn O C, Zamora M E, et al. Genetic analysis of α-synuclein 3′untranslated region and its corresponding microRNAs in relation to Parkinson's disease compared to dementia with Lewy bodies [J]. Alzheimer's & Dementia: the Journal of the Alzheimer's Association, 2017, 13(11): 1237-1250.

[5] Eyking A, Reis H, Frank M, et al. MiR-205 and MiR-373 are associated with aggressive human mucinous colorectal cancer [J]. PloS One, 2016, 11(6): e0156871.

[6] Seca H, Lima R T, Almeida G M, et al. Effect of miR-128 in DNA damage of HL-60 acute myeloid leukemia cells [J]. Current Pharmaceutical Biotechnology, 2014, 15(5): 492-502.

[7] Peng J, Mo R, Ma J, et al. let-7b and let-7c are determinants of intrinsic chemoresistance in renal cell carcinoma [J]. World Journal of Surgical Oncology, 2015, 13: 175, doi: 10.1186/S12957-015-0596-4.

[8] Young D D, Connelly C M, Grohmann C, et al. Small molecule modifiers of microRNA miR-122 function for the treatment of hepatitis C virus infection and hepatocellular carcinoma [J]. Journal of the American Chemical Society, 2010, 132(23): 7976-7981.

[9] Seth P P, Miyaji A, Jefferson E A, et al. SAR by MS: discovery of a new class of RNA-binding small molecules for the hepatitis C virus: internal ribosome entry site IIA subdomain [J]. Journal of Medicinal Chemistry, 2005, 48(23): 7099-7102.

[10] Carnevali M, Parsons J, Wyles D L, et al. A modular approach to synthetic RNA binders of the hepatitis C virus internal ribosome entry site [J]. Chembiochem: a European Journal of Chemical Biology, 2010, 11(10): 1364-1367.

[11] Parsons J, Castaldi M P, Dutta S, et al. Conformational inhibition of the hepatitis C virus internal ribosome entry site RNA [J]. Nature Chemical Biology, 2009, 5(11): 823-825.

[12] Li J, Lei K, Wu Z, et al. Network-based identification of microRNAs as potential pharmacogenomic biomarkers for anticancer drugs

[J]. Oncotarget,2016,7(29):45584-45596.
[13] Na-Na, Guan, Ya-Zhou, et al. Prediction of potential small molecule-associated MicroRNAs using graphlet interaction[J]. Frontiers in Pharmacology,2018,9:1152. doi:10.3389/fphar.2018.01152.
[14] Qu J, Chen X, Sun Y Z, et al. Inferring potential small molecule-miRNA association based on triple layer heterogeneous network [J]. Journal of Cheminformatics,2018,10(1):30, doi:10.1186/s13321-018-0284-9.
[15] Jun, Yin, Xing, et al. Prediction of small Molecule-MicroRNA associations by sparse learning and heterogeneous graph inference[J]. Molecular Pharmaceutics,2019,16(7):3157-3166.
[16] Zhou Z, Zhuo L, Fu X, et al. Joint masking and self-supervised strategies for inferring small molecule-miRNA associations [J]. Molecular Therapy-Nucleic Acids,2024,35(1):102103.
[17] Wang M N, Li Y, Lei L L, et al. Combining non-negative matrix factorization with graph Laplacian regularization for predicting drug-miRNA associations based on multi-source information fusion [J]. Frontiers in Pharmacology,2023,14:1132012, doi:10.3389/fphar.2023.1132012.
[18] Wang S, Liu T, Ren C, et al. Predicting potential small molecule-miRNA associations utilizing truncated schatten p-norm[J]. Briefings in Bioinformatics,2023,24(4):1-14.
[19] Wang S, Liu T, Ren C, et al. Identifying potential small molecule-miRNA associations via Robust PCA based on γ-norm regularization[J]. Briefings in Bioinformatics,2023,24(5):1-15.
[20] Kotsias P C, Arús Pous J, Chen H, et al. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks[J]. Nature Machine Intelligence,2020,2(5):254-265.
[21] Abdelbaky I, Tayara H, Chong K T. Identification of miRNA-small molecule associations by continuous feature representation using auto-encoders[J]. Pharmaceutics, 2022, 14(1):3, doi:10.3390/pharmaceutics14010003.
[22] Huang Y A, Hu P, Chan K C C, et al. Graph convolution for predicting associations between miRNA and drug resistance[J]. Bioinformatics,2019,36(3):851-858.
[23] 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:D1100-D1107, doi:10.1093/nar/gkr777.
[24] Vidal R. Subspace clustering[J]. IEEE Signal Processing Magazine,2011,28(2):52-68.
[25] Peng Y, Ganesh A, Wright J, et al. RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2012,34(11):2233-2246.
[26] Chandrasekaran V, Sanghavi S, Parrilo P A, et al. Ranksparsity incoherence for matrix decomposition[J]. SIAM Journal on Optimization: a Publication of the Society for Industrial and Applied Mathematics,2011,21 (2),572-596.
[27] Chen X, Zhou C, Wang C C, et al. Predicting potential small molecule-miRNA associations based on bounded nuclear norm regularization[J]. Briefings in Bioinformatics,2021,22(6):1-14.
[28] Ramlatchan A, Yang M, Liu Q, et al. A survey of matrix completion methods for recommendation systems[J]. Big Data Min Analyt,2018,1(4):308-323.
[29] Candes E, Recht B. Simple bounds for recovering low-complexity models[J]. Math Program,2013,141(1):577-589.
[30] Liu Y J, Sun D, Toh K C. An implementable proximal point algorithmic framework for nuclear norm minimization[J]. Mathematical Programming,2012,133(1-2):399-436.
[31] Chen C, He B, Yuan X. Matrix completion via an alternating direction method[J]. Ima Journal of Numerical Analysis,2012,32(1):227-245.
[32] Cai J F, Candes E J, Shen Z. A singular value thresholding algorithm for matrix completion[J]. SIAM Journal on Optimization: a Publication of the Society for Industrial and Applied Mathematics,2020,20(4):1956-1982.
[33] Liu X, Wang S, Meng F, et al. SM2miR: a database of the experimentally validated small molecules'effects on microRNA expression [J]. Bioinformatics (Oxford, England),2013,29(3):409-411.
[34] 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 Research,2018,46(D1):D1074-D1082.
[35] Kim S, Chen J, Cheng T, et al. PubChem in 2021: new data content and improved web interfaces[J]. Nucleic Acids Research,2021,49(D1):D1388-D1395.
[36] Huang Z, Shi J, Gao Y, et al. HMDD v3.0: a database for experimentally supported human microRNA-disease associations[J]. Nucleic Acids Research,2019,47(D1):D1013-D1017.
[37] Jiang Q, Wang Y, Hao Y, et al. miR2Disease: a manually curated database for microRNA deregulation in human disease[J]. Nucleic Acids Research,2009,37(Database issue):D98-D104.
[38] Ruepp A, Kowarsch A, Schmidl D, et al. PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes [J]. Genome Biology,2010,11(1):R6, doi:10.1186/gb-2010-11-1-r6.
[39] Zhao Y, Chen X, Yin J, et al. SNMFSMMA: using symmetric nonnegative matrix factorization and Kronecker regularized least squares to predict potential small molecule-microRNA association [J]RNA Biol,2020,17(2):281-291.
[40] Wang S H, Wang C C, Huang L, et al. Dual-network collaborative matrix factorization for predicting small molecule-miRNA associations[J]. Briefings in Bioinformatics,2022,23(1):1-12.
[41] Wang C C, Zhu C C, Chen X. Ensemble of kernel ridge regression-based small molecule-miRNA association prediction in human disease[J]. Briefings in Bioinformatics,2022,23(1):1-11.
[42] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2019:4171-4186.
[43] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems,2017:6000-6010.
[44] Lv S, Li Y, Wang Q, et al. A novel method to quantify gene set functional association based on gene ontology [J]. Journal of the Royal Society, Interface,2012,9(70):1063-1072.
[45] Gottlieb A, Stein G Y, Ruppin E, et al. PREDICT: a method for inferring novel drug indications with application to personalized medicine[J]. Molecular Systems Biology,2011,7:496, doi:10.1038/msb.2011.26.
[46] 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.