

DOI: 10.11830/ISSN.1000-5013.201903012



采用偏最小二乘法的基因-药物 共模块识别

毛玉杰¹, 魏东^{1,2}, 李玉双¹

(1. 燕山大学 理学院, 河北 秦皇岛 066004;
2. 河北数港科技有限公司, 河北 秦皇岛 066004)

摘要: 首先,将药物二维化学结构转化为数值序列,计算药物之间的皮尔逊相关系数,进而构建药物关联网络;然后,在带有基因网络约束的稀疏偏最小二乘算法的基础上,加入药物关联网络信息,提出伴有基因和药物关联网络正则约束的稀疏偏最小二乘(SGDPLS)算法;最后,将 SGDPLS 算法应用于基因-药物共模块识别.结果表明:药物关联网络信息的加入能够有效提高所识别的共模块中基因模块与药物模块的相关性,增加共模块的生物可解释性.

关键词: 偏最小二乘算法; 药物关联网络; 基因模块; 药物模块; 基因-药物共模块

中图分类号: Q 332 **文献标志码:** A **文章编号:** 1000-5013(2020)01-0121-05

Identify Gene-Drug Co-Modules by Partial Least Square Method

MAO Yujie¹, WEI Dong^{1,2}, LI Yushuang¹

(1. School of Science, Yanshan University, Qinhuangdao 066004, China;
2. Hebei Dataport Technology Limited Company, Qinhuangdao 066004, China)

Abstract: First, we transform the two-dimensional chemical structures of drugs into digital sequences, calculate the Pearson correlation coefficients between drugs, and then construct a drug association network. Next, we incorporate the information from drug association network into sparse partial least square algorithm with gene network, and present the sparse partial least square algorithm with gene and drug association networks (SGDPLS) algorithm. Finally, we apply SGDPLS algorithm to identify gene-drug co-modules. The result shows that, the addition of drug association network can improve the correlations between the gene modules and drug modules identified from the common module, and enhance the interpretability of the modules.

Keywords: partial least square algorithm; drug association network; gene module; drug module; gene-drug co-module

近年来,国际上开展了一系列大型的药物筛选实验,其中,规模最大的两个研究项目是癌症细胞系百科全书(CCLE)^[1]和癌症基因组项目(CGP)^[2],它们分别识别与单个药物有关联的基因组分子特征.然而,在癌症的治疗中,往往都是多个药物联合使用,因为一个药物对应的靶标基因不止一个,同时,一个靶标基因所靶向的药物也不止一个^[3].通过探索多个药物与多个基因之间的联系,即模块化分析,可以更直观地发现高维数据之间隐含的生物信息.胡尊胜等^[4]基于复杂网络,研究蛋白质界面网络中的模体和模块,有利于研究蛋白质界面网络的形成机制;Zhang 等^[5]提出稀疏网络正则约束的非负矩阵分

收稿日期: 2019-03-06
通信作者: 李玉双(1980-),女,教授,博士,主要从事生物数学的研究. E-mail: yushuangli@ysu.edu.cn.
基金项目: 国家自然科学基金资助项目(61807029)

解模型,对 microRNA-基因共模块进行识别;Chen 等^[6]提出带有网络正则约束的稀疏偏最小二乘模型(SNPLS),对基因-药物共模块进行识别,通过对各模块的分析,有利于发现不同药物之间相互作用的分子机理.受 SNPLS 模型的启发,本文提出伴有基因和药物关联网络正则约束的稀疏偏最小二乘(SGDPLS)算法,并将其应用于癌症药物敏感性基因组学数据库中.

1 材料和方法

1.1 数据来源和预处理

从癌症药物敏感性基因组学数据库(genomics of drug sensitivity in cancer, GDSC)中下载了细胞系的基因表达数据 $X_1 \in \mathbf{R}^{641 \times 13\,321}$ 和细胞系药物响应数据 $X_2 \in \mathbf{R}^{641 \times 95}$. 从 Pathway Commons 数据库(<http://www.pathwaycommons.org/>)中下载基因关联网络 A , 该网络对应的邻接矩阵为 A . 从 NCBI PubChem 数据库中下载药物的化学结构 SDF 文件,使用 OpenBabelGUI 工具箱读取药物 SDF 文件,并将药物特征转化为分子指纹,通过计算 Jaccard 相关系数^[7],得到药物相似性矩阵 D . 将 D 中小于 0.8 的位置赋值为零,其余位置赋值为 1,得到药物关联网络 B , 该网络对应的邻接矩阵为 B . 为了保证数据都处于同一数量级,且避免离群数据的出现,对模型中的初始数据 X_1 和 X_2 作了 Z-score 处理.

1.2 伴有基因和药物关联网络正则约束的稀疏偏最小二乘算法

在 SNPLS 算法的基础上,增加了药物关联网络信息,构建 SGDPLS 算法,其流程图如图 1 所示. 图 1 中: H 是药物 Docetaxel 对应的化学结构;目标函数(1)中的 L 和 C 是由邻接矩阵 A 和 B 经过拉普拉斯矩阵变换得到的^[8]. SGDPLS 算法对权向量 g 和 d 进行稀疏约束,可以使模型在迭代优化过程中不陷入过拟合状态,增加识别模块的可解释性.

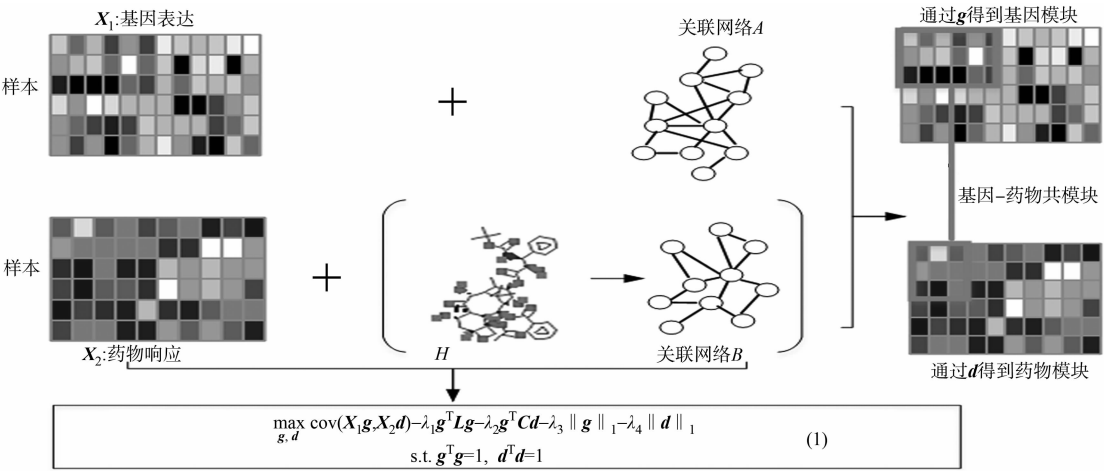


图 1 SGDPLS 算法流程图
Fig. 1 SGDPLS algorithm flow chart

1.3 SGDPLS 模型的求解

图 1 中的目标函数(1)是一个非凸函数,利用传统的优化算法很难得到模型的全局最优解. Chen 等^[6]采用交替坐标下降法,通过交替更新 g 和 d ,求得模型的局部极大值. 文中将该方法用于 SGDPLS 模型的求解,SGDPLS 模型的求解过程有以下 3 个步骤.

步骤 1 以标准的偏最小二乘模型^[9] $\max_{g,d} \text{cov}(X_1 g, X_2 d)$ 的解,作为初始变量 g ,且 $u = X_1 g$.

步骤 2 交替更新 g 和 d .

固定变量 g ,更新变量 d ,则

$$d_q \leftarrow \frac{\text{sign}(z_q)(|z_q| - \lambda_4)_+}{2(\lambda_2 + \omega)}, \quad q = 1, 2, \dots, m, \quad \text{标准化 } d,$$
$$v = X_2 d.$$

上式中: $(\cdot)_+ = \begin{cases} x, & x > 0 \\ 0, & \text{其他} \end{cases}$; $z_q = t_q + 2\lambda_2 \sum_{p=1}^m \frac{b_{p,q} d_p}{\sqrt{f_p f_q}}$, t_q 为向量 $t = \frac{1}{p}(X_2^T X_1 g) = \frac{1}{p}(X_2^T u)$ 的第 q 个元素,

f_p 表示网络 B 对应的度矩阵 \mathbf{F} 的 p 行, $b_{p,q}$ 表示邻接矩阵 \mathbf{B} 的第 p 行, 第 q 列; ω 表示朗格朗日因子.

固定变量 \mathbf{d} , 更新变量 \mathbf{g} , 则

$$g_j \leftarrow \frac{\text{sign}(z_j)(|z_j| - \lambda_3)_+}{2(\lambda_1 + \sigma)}, \quad j = 1, 2, \dots, n, \quad \text{标准化 } \mathbf{g},$$
$$\mathbf{u} = \mathbf{X}_1 \mathbf{g}.$$

上式中: $z_j = t_j + 2\lambda_1 \sum_{i=1}^n \frac{a_{i,j} g_i}{\sqrt{e_i e_j}}$, t_j 表示向量 $\mathbf{t} = \frac{1}{p}(\mathbf{X}_1^T \mathbf{X}_2 \mathbf{d}) = \frac{1}{p}(\mathbf{X}_1^T \mathbf{v})$ 的第 j 个元素, e_i 表示网络 A 对应的度矩阵 \mathbf{E} 的第 i 行, $a_{i,j}$ 表示邻接矩阵 \mathbf{A} 的第 i 行, 第 j 列; σ 表示朗格朗日因子.

步骤 3 重复步骤 2 直到 \mathbf{u} 收敛, 算法终止.

对目标函数(1)做上述迭代运算, 得到权向量 \mathbf{g} 和 \mathbf{d} 识别共模块之前, 先对 \mathbf{g} 和 \mathbf{d} 作 Z-score 标准化处理. 通过设定阈值 T , 筛选共模块的成员. 对于基因(药物)模块, 选取 $g(\mathbf{d})$ 中比阈值 $T_1(T_2)$ 大的位置所对应的基因(药物)作为基因(药物)模块的成员. 对于样本模块, 首先计算 $\mathbf{u} = \mathbf{X}_1 \mathbf{g}$ 和 $\mathbf{v} = \mathbf{X}_2 \mathbf{d}$; 然后, 对 $\mathbf{u} + \mathbf{v}$ 作标准化处理, 取向量中比阈值 T_3 大的位置所对应的样本作为样本模块的成员. 在筛选样本、基因和药物共模块时, 设置的阈值分别为 $T_1 = 3, T_2 = 3, T_3 = 2$. 阈值过大(小), 会使识别出的共模块过小(大). 模块过大, 会导致模块中包含的信息太多, 有用的信息不易被挖掘出来; 模块过小, 则包含的信息过少, 不具有生物可解释性^[4].

2 结果分析

将 SGDPLS 模型用于 GDSC 数据库, 通过 MATLAB 编程^[10] 调参, 识别出 20 个不同的基因-药物共模块. SGDPLS 模型识别共模块分布图, 如图 2 所示. 由图 2 可知: 有 17 个样本模块其样本个数分布在 24~34 范围内(图 2(a)); 有 13 个基因模块其基因个数分布在 91~151 范围内(图 2(b)); 12 个药物模块的药物个数为 2(图 2(c)).

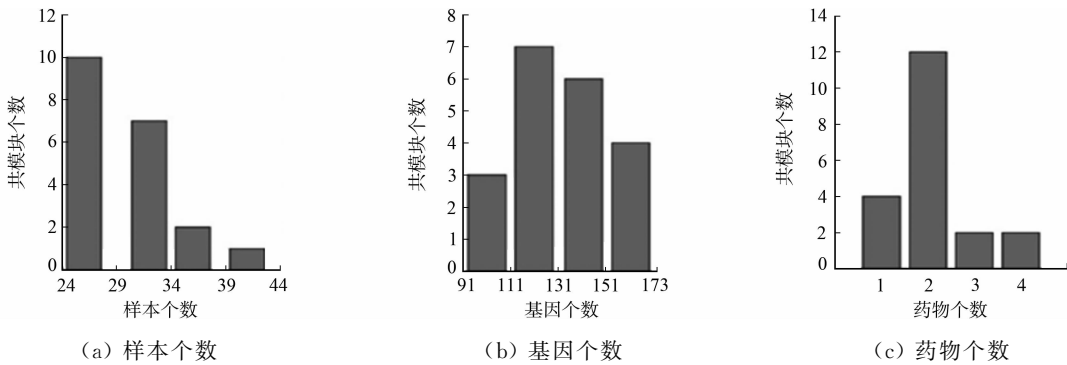


图 2 SGDPLS 模型识别共模块分布图

Fig. 2 Distribution map of SGDPLS model recognition common modules

为了解释文中识别出的共模块具有生物意义, 利用 R 语言中的 clusterProfiler 包对识别出的基因模块进行 GO 生物功能项和 KEGG 通路富集分析. 对于 20 个基因模块, 有 14 个基因模块(70%)至少富集一种 GO 生物过程; 有 12 个基因模块(60%)至少富集一种 KEGG 通路. 例如, 第 2 个共模块, 样本模块包含 28 个样本, 这些样本主要富集的癌症类型是小细胞肺癌和小细胞癌. 第 2 个共模块生物项的富集结果, 如图 3 所示. 图 3 中: p 表示富集分数经过 Benjamini 校正后的值.

基因模块包含 107 个基因, 这些基因主要富集的 GO 生物过程是 DNA 构象变化、核染色体分离、姐妹染色体分离(图 3(a)), 这些过程都发生在细胞增殖过程中. 因为姐妹染色体分离到子细胞的过程是不可逆的, 检测点的缺失会导致染色体不稳定, 所以, 在细胞周期中, 有丝分裂过程最易发生癌变^[11]. 现代医学利用相关药物制止细胞中纺锤体的出现, 从而抑制细胞有丝分裂的进行, 使细胞分裂停留在 G0 阶段, 利用该技术可以有效地遏制癌细胞的恶性增殖和扩散. 另外, 基因模块主要富集的 KEGG 通路为上皮细胞的细菌侵入、小细胞肺癌、细胞周期、焦点粘连、同源重组等细胞生化过程(图 3(b)), 这些生化过程对癌症的产生和转移起着很重要的作用. 药物模块包含 2 个药物, 即 Docetaxel 和 RDEA119.

Riichiroh 等^[12]指出,药物 Docetaxel 已在临床上被证实可以用于小细胞肺癌的治疗. 药物 RDEA119^[13]是一种 MEK 抑制剂,MEK 是一种磷酸化丝裂原活化蛋白激酶(MAPK)的激酶,MEK 酶的靶标 ERK 具有高选择性,且可以驱动细胞增殖,从而实现抑制肿瘤细胞增殖并诱导细胞凋亡. 研究表明,该药物对组织的选择性,可以降低其对中枢神经的毒副作用,且口服性的 MEK 抑制剂 RDEA119 已在临床上被开发应用,有效地推动了药代动力学的发展^[14].

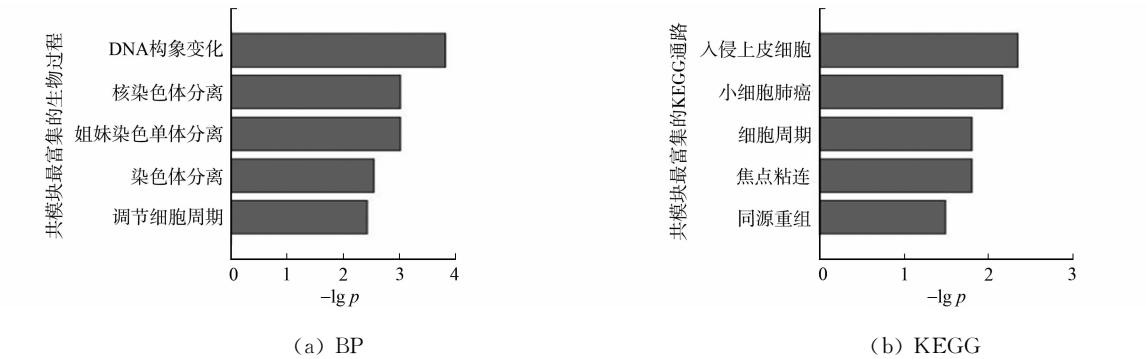


图 3 第 2 个共模块生物项的富集结果

Fig. 3 Enrichment results of the second common module biological item

3 SGDPLS 算法与 SNPLS 算法的比较

为了展示药物关联网络在 SGDPLS 算法中所起的重要作用,将 SNPLS 算法也用于同样的数据集上,分别计算 2 种算法识别出来的 20 个基因-药物共模块之间的皮尔逊相关系数,结果如图 4 所示.

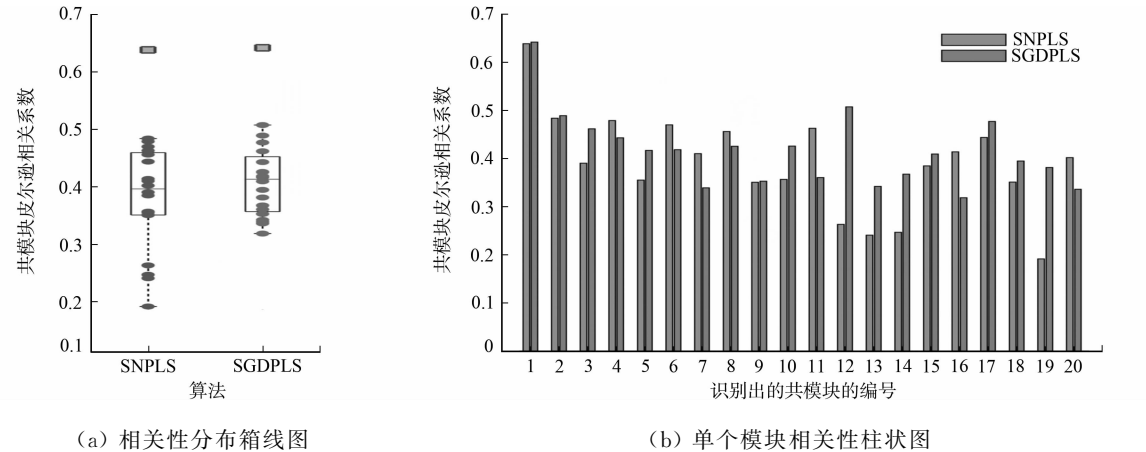


图 4 两种算法共模块皮尔逊相关性比较

Fig. 4 Comparison of co-module Pearson correlation between two algorithms

由图 4 可知:SGDPLS 算法识别的大部分共模块(65%)比 SNPLS 算法具有更好的相关性,说明由药物二维化学结构构造的药物关联网络,在一定程度上,可以提高共模块中基因模块和药物模块数据模式的相似性,使共模块更具生物意义.

4 结束语

提出了 SGDPLS 算法,并深入挖掘多个基因和多个药物之间的对应关系,针对 SGDPLS 算法和 SNPLS 算法,独立识别出的 20 个基因-药物共模块,从共模块之间的皮尔逊相关系数和生物意义角度进行分析. 结果表明:SGDPLS 算法识别的共模块中,有 65%比 SNPLS 算法识别的共模块具有更高的相关性. 由此可见,药物关联网络信息的加入增强了 SGDPLS 算法识别模块的相关性,有助于发现潜在的药物靶标. 同时,SGDPLS 算法有望应用于其他领域中,进行高维数据的降维,如航天遥感数据、金融市场交易数据,用以发现数据的本质结构^[15].

参考文献:

- [1] BARRETINA J, CAPONIGRO G, STRANSKY N, *et al.* The cancer cell line encyclopedia enables predictive modeling of anticancer drug sensitivity[J]. *Nature*, 2012, 483(7391): 603-607. DOI: 10.1038/nature11003.
- [2] GARNETT M J, EDELMAN E J, HEIDORN S J, *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells[J]. *Nature*, 2012, 483(7391): 570-575. DOI: 10.1038/nature11005.
- [3] BOKHARI S U, GOPAL U M, DUCKWORTH W C. Beneficial effects of a glyburide/metformin combination preparation in type 2 diabetes mellitus[J]. *American Journal of the Medical Sciences*, 2003, 325(2): 66-69. DOI: 10.1097/00000441-200302000-00003.
- [4] 胡尊胜, 林锦贤, 吕曦. 蛋白质界面网络中模体和模块的探测[J]. *华侨大学学报(自然科学版)*, 2014, 35(1): 61-66. DOI: 10.11830/issn.1000-5013.2014.01.0061.
- [5] ZHANG Shihua, LI Qingjiao, LIU Juan, *et al.* A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules[J]. *Bioinformatics*, 2011, 27(13): i401-i409. DOI: 10.1093/bioinformatics/btr206.
- [6] CHEN Jinyu, ZHANG Shihua. Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data[J]. *Bioinformatics*, 2016, 32(11): 1724-1732. DOI: 10.1093/bioinformatics/btw059.
- [7] WANG Lin, LI Xiaozhong, ZHANG Louxin, *et al.* Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization[J]. *BMC Cancer*, 2017, 17(1): 513. DOI: 10.1186/s12885-017-3500-5.
- [8] LUXDURG U V. A tutorial on spectral clustering[J]. *Statistics and Computing*, 2007, 17(4): 395-416. DOI: 10.1007/s11222-007-9033-z.
- [9] OLESZKO A, HARTWICH J, WÓJTOWICZ A, *et al.* Comparison of FTIR-ATR and Raman spectroscopy in determination of VLDL triglycerides in blood serum with PLS regression[J]. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2017, 183: 239-246. DOI: 10.1016/j.saa.2017.04.020.
- [10] WANG Bei, CHEN Jinyu, ZHANG Shihua. BMTK: A toolkit for determining modules in biological bipartite networks[J]. *Quantitative Biology*, 2018, 6(2): 186-192. DOI: 10.1007/s40484-018-0132-y.
- [11] JALLEPALLI P V, LENGAUER C. Chromosome segregation and cancer: Cutting through the mystery[J]. *Nature Reviews Cancer*, 2001, 1(2): 109. DOI: 10.1038/35101065.
- [12] RIICHIROH M, YUTAKA N, TOMOHIDE T, *et al.* Phase III study, V-15-32, of gefitinib versus docetaxel in previously treated Japanese patients with non-small-cell lung cancer[J]. *Journal of Clinical Oncology*, 2008, 26(26): 4244-4252. DOI: 10.1200/JCO.2007.15.0185.
- [13] DILLY A K, SONG X, ZEH H J, *et al.* Mitogen-activated protein kinase inhibition reduces mucin 2 production and mucinous tumor growth[J]. *Translational Research*, 2015, 166(4): 344-354. DOI: 10.1016/j.trsl.2015.03.004.
- [14] IVERSON C, LARSON G, LAI C, *et al.* RDEA119/BAY 869766: A potent, selective, allosteric inhibitor of MEK1/2 for the treatment of cancer[J]. *Cancer Research*, 2009, 69(17): 6839-6847. DOI: 10.1158/0008-5472.can-09-0679.
- [15] 华正宇. 基于 SLT 的偏最小二乘分类算法及其优化方法研究[D]. 沈阳: 东北大学, 2013.

(责任编辑: 黄晓楠 英文审校: 黄心中)