DOI:10.3969/j.issn.1005-202X.2021.12.020

医学生物信息

基于通路活性的抗癌药物敏感性预测

高冲[1,2],秦玉芳[1,2],陈明[1,2]
1.上海海洋大学信息学院,上海 201306;2.农业农村部渔业信息重点实验室,上海 201306

【摘要】通路数据库是系统分析基因功能,联系基因组信息和功能信息的知识库,通路作为基因功能集合能够提高预测模型的预测能力和解释能力。本研究通过整合KEGG通路数据和CCLE基因表达谱推断通路活性并结合药物数据建立药物敏感性预测模型。在推断通路活性时,并没有把通路简单地作为基因集合,而是选择通路中的高连接度基因,取高连接度基因的平均表达水平作为通路活性值,合并每个通路的活性向量得到通路活性矩阵,然后输入到弹性网进行药物敏感性预测。实验结果表明,对于大多数药物,使用基于通路中高连接度基因的计算分析方法更有利于药物敏感性预测,同时能识别出更多与药物相关基因的通路。

【关键词】癌症;药物敏感性;通路活性;基因表达;个性化医疗
【中图分类号】R318;TP181 【文献标志码】A 【文章编号】1005-202X(2021)12-1569-06

Predicting anticancer drug sensitivity based on pathway activity

GAO Chong[1,2], QIN Yufang[1,2], CHEN Ming[1,2]
1. College of Information Technology, Shanghai Ocean University, Shanghai 201306, China; 2. Key Laboratory of Fisheries Information, Ministry of Agriculture and Rural Affairs of the People's Republic of China, Shanghai 201306, China

Abstract: Pathway database is a knowledge base for analyzing gene functions systematically and linking genomic information and functional information. As a collection of gene functions, pathways can improve the prediction ability and interpretation ability of prediction models. Herein the pathway activity is inferred by the integration of KEGG pathway data and CCLE gene expression profile, and then is combined with drug data to establish a drug sensitivity prediction model. When inferring the activity of the pathway, the pathway is not simply treated as a collection of genes. The genes with high connectivity in the pathway are selected in the study, and the average expression level of the genes with high connectivity is taken as the pathway activity value. The pathway activity matrix which is obtained by combining the activity vector of each pathway is input to the elastic net for drug sensitivity prediction. Experimental results show that for most drugs, the use of computational analysis methods based on the genes with high connectivity in pathways is more conducive to drug sensitivity prediction, and at the same time it can identify more pathways for drug-related genes.

Keywords: cancer; drug sensitivity; pathway activity; gene expression; personalized medicine

前 言

恶性肿瘤(癌症)是严重影响人类健康的疾病之一[1]。虽然传统治疗癌症的方法(放疗、化疗)有明显的治疗效果,但是大量研究表明肿瘤具有异质性[2],患有相同癌症的病人使用相同的治疗方法却有不同的疗效。基于此,个性化医疗应运而生,它关注每一位患者的特异性特征,其中测量患者对药物的反应是一个关键问题[3-4]。

随着高通量基因组学技术的发展,药物基因组学成为测量患者对药物反应的一个重要方法[5]。研究者通常通过基因或蛋白质表达谱等分子图谱来测量细胞对药物的反应,进而建立相应的计算模型预测药物反应[6]。Gillet等[7]发现在细胞系模型和临床具有相关性的前提下,这些计算模型能识别决定药物反应的分子因素,并对患者群体进行相应的个性化药物治疗。许多研究机构开发了诸如癌症细胞系百科全书(Cancer Cell Line Encyclopedia, CCLE)和

【收稿日期】2021-06-25
【基金项目】国家自然科学基金(61702325);国家重点研发计划项目(2018YFD0701003);上海市科技创新计划项目(20dz1203800)
【作者简介】高冲,硕士在读,研究方向:机器学习、生物信息,E-mail: gaochongc@sina.com
【通信作者】秦玉芳,博士,副教授,研究方向:机器学习、生物信息,E-mail: yfqin@shou.edu.cn

肿瘤药物敏感性基因组学(Genomics of Drug Sensitivity in Cancer, GDSC)等包含基因表达数据和拷贝数变异等基因组学数据以及药物反应值在内的大型数据库,这些大型数据集为识别新的药物靶点和药物反应标记物提供了更多的可能性[8];同时,这也为开发药物反应计算模型提供了依据,如Papillon-Cavanagh等[9]利用CCLE和癌症基因组计划(Cancer Genome Project, CGP)数据集建立预测药物反应的线性模型,发现基因组预测因子能够验证对特定药物的反应;Masica等[10]利用CCLE数据集构建多变量组合改变组织(MOCA)模型来识别药物反应的组合生物标志物;Menden等[11]利用GDSC数据集和机器学习算法建立基于细胞系的基因组特征和药物化学特性的药物敏感性预测模型,并通过对实验结果和已有事实的对比验证该模型的有效性。

近年来,许多研究者根据基因水平特征建立抗癌药物敏感性预测模型[6]。如Costello等[12]把基因表达谱或拷贝数变异等基因组学数据用于预测抗癌药物反应,发现基于基因表达数据建立的抗癌药物敏感性预测模型具有很好的预测性能;Geeleher等[13]采用岭回归算法建立抗癌药物反应预测模型,同时使用独立数据集验证了该模型的有效性。这些方法大多基于基因表达数据等基因水平特征,在独立研究中的重复性有限,这对生物学解释提出了挑战[14]。有研究表明考虑基因间相互作用行为比仅仅关注单个基因行为在预测药物反应上具有更好的预测效果[15]。通路数据库是系统分析基因功能,联系基因组信息和功能信息的知识库。通路作为基因功能集合能够提高预测模型的预测能力和解释能力[16]。Wang等[17]把通路数据和基因表达谱应用到药物敏感性预测,研究表明在CCLE数据集的24种药物中,基于通路的模型较基于基因的模型具有更好的预测性能,并且基于通路的模型能识别更多药物相关的基因或通路,具有更好的生物学解释;然而该方法仅仅把通路作为基因集合,没有考虑通路中基因互相作用关系。

针对以上问题,本研究提出一种整合通路网络中高连接度基因和基因表达数据推断通路活性,建立抗癌药物敏感性预测模型,简记为PHG(Pathway Hub Gene)。首先利用通路数据和STRING数据库得到每个通路的基因相互作用网络表,从该网络表中选择高连接度基因;然后分别计算每一个通路的活性向量;最后合并所有通路的活性向量,得到通路活性特征矩阵,以此作为抗癌药物敏感性预测模型的输入。10折交叉验证的实验结果表明,在17-AAG等大多数抗癌药物上,并不是通路中所有基因都对药物敏感性预测有帮助,考虑通路中的关键基因较通路全部基因构建预测模型具有更好的预测效果,同时验证了基于通路的模型较基于基因的模型能给出更好的生物学解释。

1 数据集及预处理

本研究的基因表达和药物IC50值数据来自于CCLE数据库,下载地址为https://portals.broadinstitute.org/ccle/data;同时为了独立检验,也下载了GDSC数据库中的基因表达和药物IC50数据,下载地址为https://www.cancerrxgene.org/。为消除实验技术和实验平台所导致的基因表达量误差,采用以基因为中心的RMA标准化算法对基因表达谱进行标准化处理。经过标准化后,CCLE基因表达谱共有18 900个基因和1 036个细胞系样本,GDSC基因表达谱中共有9 920个基因和697个细胞系样本。

本研究使用IC50值衡量药物敏感性,类似于Wang等[17]的做法,对药物反应IC50值做log变换。由于基因表达数据中的一些细胞系样本在药物反应数据里不存在,所以本研究选取在基因表达数据和药物反应数据中同时存在的细胞系进行分析。例如,对于药物AEW541来说,NCIH2196_LUNG细胞系存在基因表达谱中,但在AEW541药物反应数据中没有该细胞系,所以在做AEW541药物的敏感性预测时需去除该细胞系。

本研究使用的通路数据来自京都基因和基因组数据库(Kyoto Encyclopedia of Genes and Genomes, KEGG)中的通路数据库,在该数据库中下载每个通路的基因集,最终的通路数据集包括389个通路,共有14 097个基因。通路中基因间相互作用关系表可从STRING数据库中获得,下载地址为https://www.string-db.org/cgi/download,STRING数据库包含5 090个物种、24 584 628种蛋白和3 123 056 667个相互作用关系[18],本研究下载的数据来自数据库最新版本(Version 11.0)。

2 模 型

2.1 基于通路高连接基因的通路活性推断

为推断通路活性,本研究不仅考虑通路中每个基因的表达水平,还关注了通路中基因间相互作用关系,基因相互作用关系在预测药物敏感性具有更好的鲁棒性[15]。首先从STRING数据库中得到每个通路(基因集)中基因间互相作用网络表,表中的(G_i, G_j)表示基因G_i和基因G_j在通路中是相互连接的;接着根据通路互相作用网络表计算通路活性向量。

计算每个基因在通路互相作用网络表中的度,

由于通路网络中高连接度的Hub基因对整个通路的功能起着更关键的作用[19]，所以从该网络表中选择高连接度Hub基因来进行分析。将基因的度降序排序，选择排名在前10%的基因作为Hub基因，图1中的$G_{h1}, G_{h2}, \cdots, G_{hk}$为通路$p_1$的Hub基因；计算Hub基因表达值的平均值作为活性值，然后合并每个细胞系样本中的活性值得到该通路的活性向量。活性值计算公式如下：

$$p_{1j} = \frac{1}{h_k}\sum_{i=h_1}^{i=h_k} g_{ij} \quad (1)$$

其中，h_k表示通路P_1中Hub基因的数量；g_{ij}表示基因i在细胞系样本j中表达值；p_{1j}表示细胞系样本j在通路p_1中的活性值。

按照上面的方法，计算所有通路的活性向量，合并得到通路活性矩阵（列为细胞系样本，行为通路）。假定有l个通路，分别记为$P_1, P_2, \cdots, P_l$，按照上述方法计算所有通路的活性向量后得到通路活性矩阵(p_{ij})，其中i为通路，j为细胞系样本。

总的来说，可将基因表达谱和通路中关键基因信息分析整合得到通路活性矩阵，以此来预测癌症药物敏感性。基于通路中高连接度基因模型的流程如图1所示。

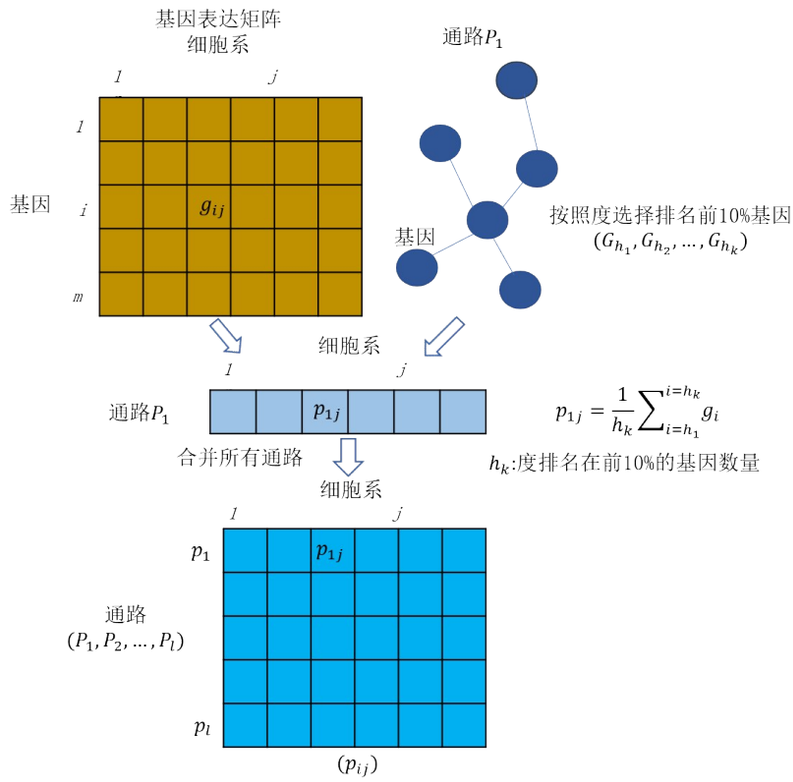

图1 利用通路中高连接度基因表达推断通路活性

Fig.1 Using the expression of genes with high connectivity in the pathway to infer pathway activity

2.2 建立药物反应预测模型

将得到的通路活性特征矩阵作为预测模型的输入，药物敏感性水平作为模型输出，根据均方误差（Mean Square Error, MSE）来调试优化模型的参数，并进行训练与预测。本研究采用机器学习中的弹性网作为预测算法。

弹性网是一种使用L1和L2范数作为先验正则项训练的线性回归模型[20]。这种组合可以学习到类似于Lasso的一个稀疏模型，同时还保留岭回归的正则化属性，既能实现重要特征变量的选择，又能处理强相关性特征数据，具有较好的群组效应，结合了岭回归和Lasso回归的优点。因此，弹性网尤其适用于有多个特征彼此相关的场合。在基于通路/基因的预测模型中，作为特征的通路/基因相互之间实际上都是有联系的。因此，本研究选用弹性网回归算法来构建预测模型，并使用R语言中glmnet包实现弹性网算法。调整和优化模型主要通过网格搜索，在1 000个参数中寻找最优参数，其中α:$[0.1, 1]$设置10个参数，λ:$[\exp^{-5}, \exp^{5}]$设置100个参数，使用10折交叉验证选取最优参数。

3 结果与讨论

3.1 通路之间的重叠性

本研究使用Jaccard指数来评价两个通路之间的

重叠性。通过对通路间重叠性的研究，分析通路是否具有特异性，是否对实验产生较大的误差。Jaccard指数计算公式如下：

$$\text{Jaccard}\left(P_1, P_2\right) = \frac{P_1 \cap P_2}{P_1 \cup P_2} \tag{2}$$

其中，$P_1 \cap P_2$表示同时存在于通路P_1和通路P_2的基因；$P_1 \cup P_2$表示存在于通路P_1或P_2的基因。由式(1)可以发现，当两个通路完全不同时，即两个通路没有相同的基因，则Jaccard指数为0，当两个通路的基因集完全相同时，则Jaccard指数为1。因此，所有通路对的Jaccard指数在0到1变化不等。计算所有通路对的Jaccard指数，结果显示约30%通路对的重叠性小于0.6，大多数通路的Jaccard指数小于0.2，这说明通路之间的重叠性较低，降低了因通路之间的重叠过高而引起的模型误差。

3.2 基于通路活性的药物敏感性预测性能

比较分析文献[17]中的方法(DiffRank)，本研究提出基于通路中所有基因推断通路活性的方法，即PAG(All Gene of Pathway)。为了把基于通路模型和基于基因模型进行对比，还提出基于基因模型的方法AG(All Gene)。

PAG方法和PHG方法的不同在于PHG方法在推断通路活性时使用的是通路中高连接度的关键基因，而PAG方法使用通路中所有基因来计算活性值，进而得到通路活性矩阵，以此作为预测模型的输入。此外，基于基因模型的AG方法是直接使用基因表达矩阵作为药物敏感性预测模型的输入，而不考虑通路信息，基因模型中的细胞系为样本，基因表达值为特征。

本研究使用弹性网算法训练通路活性矩阵，10折交叉验证选择最优参数，并使用最优参数下的MSE作为预测模型性能的评价标准。图2给出了基于CCLE数据集中24种药物在4种模型下进行药物敏感性预测的结果。PHG方法在17-AAG等12种药物上具有最好的预测效果，在AZD6244等6种药物上的预测效果是次好的；PAG方法在Irinotecan等4种药物上具有最好的预测效果，在17-AAG等11种药物上具有次好的效果。通过PHG和PAG对比分析，发现并不是通路中所有基因都会对药物敏感性预测有帮助，只选取通路中连接紧密的基因进行预测可能更具有鲁棒性。AG方法在AZD6244等7种药物上具有最好的效果，在Erlotinib等6种药物上的预测效果是次好的。对比基于通路模型和基于基因模型可以发现基于通路模型有较好的预测性能。总的来说，对于一些药物，使用基于通路中高连接度基因的计算分析方法取得了最好的预测效果，更有利于药物敏感性预测。

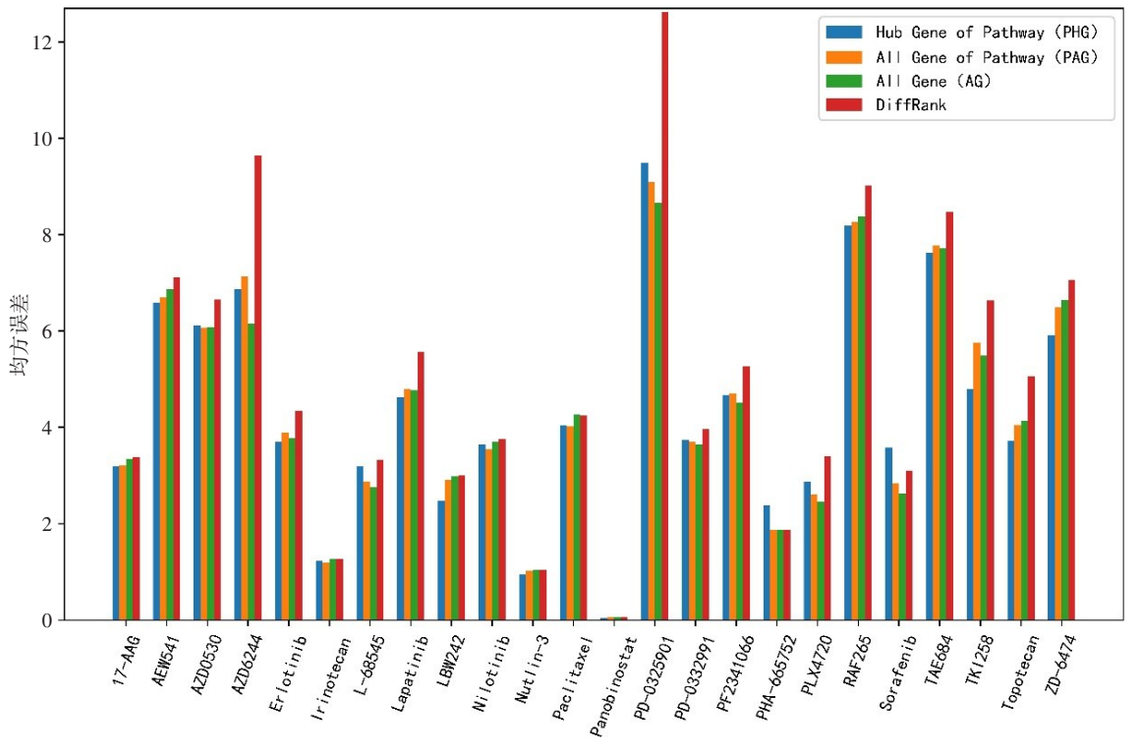

图2 不同模型对CCLE中24种药物的预测性能

Fig.2 Predictive performance of different models for 24 kinds of drugs in CCLE

3.3 模型再现性

本研究中的模型再现性是指在一个数据集上训练数据,在另一个数据集上测试数据,然后再交换数据集重新训练和测试。与CCLE数据集相比,GDSC数据集中基因表达矩阵和通路数据推断通路活性矩阵的特征数量较少。为了实验的有效性,本研究从基于CCLE基因表达谱推断通路活性矩阵中随机抽取和GDSC相同数量的特征,以便训练和预测。

对于给定的抗癌药物,随机选择50次相等数量的特征数据输入到药物敏感性模型,然后计算MSE,把50次MSE的平均值作为验证模型再现性预测性能。共计算了24种药物在CCLE数据集和GDSC数据集上的预测性能,表1中列出了Paclitaxel等4种药物在CCLE数据集和GDSC数据集上的预测性能。在基于药物Paclitaxel敏感性预测的模型再现性中,当以GDSC数据作为训练集,CCLE数据作为测试集时,MSE为5.91;当以CCLE数据为训练集,GDSC数据为测试集时,MSE为5.52,这表明PHG方法在药物Paclitaxel上的药物敏感性预测具有较好的模型再现性。

表1 PHG方法在4种药物的模型再现性

Tab.1 Model reproducibility of PHG method in 4 kinds of drugs

药物	训练集	测试集	MSE
Paclitaxel	GDSC	CCLE	5.91
	CCLE	GDSC	5.52
PD0332991	GDSC	CCLE	5.57
	CCLE	GDSC	175.50
17-AAG	GDSC	CCLE	3.41
	CCLE	GDSC	22.41
PHA665752	GDSC	CCLE	20.82
	CCLE	GDSC	4.52

另外,对于药物17-AAG和PD0332991,以GDSC数据作为训练集训练模型,同时用此模型测试CCLE数据,发现具有较低的MSE,即较好的预测性能,然而当以CCLE数据训练模型,再以GDSC数据测试模型,则有较高的误差,这表明PHG方法在这两种药物上使用基于GDSC基因表达谱作为训练集时会得到较好的模型,具有较好的预测性能。相反,对于药物PHA665752,以CCLE基因表达数据作为训练集构建药物敏感性预测模型则会得到较好的预测性能。

3.4 识别药物相关联基因的通路

本研究把通路数据和基因表达谱整合得到通路活性评分,并以此构建预测模型,进一步识别癌症标记物,从而给出生物学解释。当利用通路中高连接度基因数据和弹性网算法建立预测模型时,弹性网中非零系数对应的特征是预测细胞对药物反应的重要数据[1]。因此,本研究采用了弹性网算法中非零系数统计与抗癌药物相关联基因的通路数量。在24种药物中,19种药物包含靶向基因的通路都能识别出来(表2)。

表2 药物相关基因的通路数量

Tab.2 Number of pathways for drug-related genes

药物	PAG	PHG	药物	PAG	PHG
AZD0530	2	5	AZD6244	2	1
Irinotecan	2	3	Erlotinib	3	2
Lapatinib	7	9	Panobinostat	2	2
Nilotinib	2	3	PD-0332991	1	1
Nutlin-3	3	6	PHA-665752	3	1
Paclitaxel	2	5	PLX4720	0	1
PF2341066	0	2	RAF265	3	3
Sorafenib	2	8	TAE684	5	3
17-AAG	0	1	Topotecan	0	2
AEW541	1	1			

例如,对于药物Lapatinib,使用PHG方法能识别弹性网中非零系数对应的MicroRNAs in cancer、Breast cancer和EGFR tyrosine kinase inhibitor resistance等9个特征通路,其中MicroRNAs in cancer通路包含ABCB1、EGFR和ERBB2等靶向基因,Breast cancer通路包含EGFR和ERBB2等靶向基因。总的来说,基于通路高连接度基因的药物敏感性预测模型能够识别药物相关联基因的通路,具有更好的生物学解释能力。

4 结 论

本研究提出一种基于通路中高连接度基因的抗癌药物敏感性预测方法(PHG);对基因表达谱、通路数据和药物敏感性IC50值进行综合分析,综合考虑不同因素的作用,提取高连接度基因集合,然后计算通路活性矩阵,进而通过机器学习技术进行抗癌药物敏感性预测分析,并把识别的标记与已有研究进行对比分析,验证基因/通路与药物之间的联系。实验表明,基于通路中高连接度基因模型相比其他通路或基因模型有更好的预测性能。通路中并不是所有的基因都对药物敏感性预测起到促进作用,而是一些关键基因更为重要。本研究提出的计算方法为通路活性预测模型的发展提供了参考。

【参考文献】

[1] SUNG H, FERLAY J, SIEGEL R L, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA Cancer J Clin, 2021, 71(3): 209-249.

[2] DINCER A B, CELIK S, HIRANUMA N, et al. DeepProfile: deep learning of cancer molecular profiles for precision medicine[J]. bioRxiv, 2018: 278739.

[3] ASHLEY E A. The precision medicine initiative: a new national effort[J]. JAMA, 2015, 313(21): 2119-2120.

[4] COLLINS F S, VARMUS H. A new initiative on precision medicine[J]. N Engl J Med, 2015, 372(9): 793-795.

[5] CHIN L, ANDERSEN J N, FUTREAL P A. Cancer genomics: from discovery science to personalized medicine[J]. Nat Med, 2011, 17(3): 297-303.

[6] AZUAJE F. Computational models for predicting drug responses in cancer research[J]. Brief Bioinform, 2017, 18(5): 820-829.

[7] GILLET J P, VARMA S, GOTTESMAN M M. The clinical relevance of cancer cell lines[J]. JNCI: J Nat Cancer I, 2013, 105(7): 452-458.

[8] 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.

[9] PAPILLON-CAVANAGH S, DE JAY N, HACHEM N, et al. Comparison and validation of genomic predictors for anticancer drug sensitivity[J]. J Am Med Inform Assoc, 2013, 20(4): 597-602.

[10] MASICA D L, KARCHIN R. Collections of simultaneously altered genes as biomarkers of cancer cell drug response[J]. Cancer Res, 2013, 73(6): 1699-1708.

[11] MENDEN M P, IORIO F, GARNETT M, et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties[J]. PLoS One, 2013, 8(4): e61318.

[12] COSTELLO J C, HEISER L M, GEORGII E, et al. A community effort to assess and improve drug sensitivity prediction algorithms[J]. Nat Biotech, 2014, 32(12): 1202-1212.

[13] GEELEHER P, COX N J, HUANG R S. Clinical drug response can be predicted using baseline gene expression levels and *in vitro* drug sensitivity in cell lines[J]. Genome Biol, 2014, 15(3): R47.

[14] EIN-DOR L, ZUK O, DOMANY E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer[J]. Proc Natl Acad Sci U S A, 2006, 103(15): 5923-5928.

[15] SHI W, JIANG T, NUCIFORO P, et al. Pathway level alterations rather than mutations in single genes predict response to HER2-targeted therapies in the neo-ALTTO trial[J]. Ann Oncol, 2017, 28(1): 128-135.

[16] KHATRI P, SIROTA M, BUTTE A J. Ten years of pathway analysis: current approaches and outstanding challenges[J]. PLoS Comput Biol, 2012, 8(2): e1002375.

[17] WANG X, SUN Z, ZIMMERMANN M T, et al. Predict drug sensitivity of cancer cells with pathway activity inference[J]. BMC Med Genomics, 2019, 12(1): 15.

[18] MERING C V, HUYNEN M, JAEGGI D, et al. STRING: a database of predicted functional associations between proteins[J]. Nucleic Acids Res, 2003, 31(1): 258-261.

[19] CARTER S L, BRECHBÜHLER C M, GRIFFIN M, et al. Gene co-expression network topology provides a framework for molecular characterization of cellular state[J]. Bioinformatics, 2004, 20(14): 2242-2250.

[20] ZOU H, HASTIE T. Regularization and variable selection *via* the elastic net[J]. J Royal Stat Soc B, 2005, 67: 301-320.

（编辑：谭斯允）