中国科学: 生命科学 2023 年 第 53 卷 第 11 期: 1663 ~ 1672
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

论 文

基于核方法的协同药物组合预测方法

张竣[1,2], 袁锐[1], 陈世龙[1,3], 王永翠[1*]

1. 中国科学院西北高原生物研究所, 高原生物适应与进化重点实验室, 西宁 810008;
2. 中国科学院大学生命科学学院, 北京 100049;
3. 中国科学院三江源国家公园研究院, 西宁 810008
* 联系人, E-mail: ycwang@nwipb.cas.cn

收稿日期: 2023-02-28; 接受日期: 2023-06-25; 网络版发表日期: 2023-08-30

摘要 肿瘤因其异质性和复杂的代谢途径, 会对单一药物产生耐药性. 具有协同抑制效应的药物组合策略, 是解决上述问题的有效途径之一. 然而, 筛选有效药物组合往往需要通过一系列药理学、分子生物学实验, 耗时且费用高昂. 生物信息学方法通过对已知协同用药的实验数据进行建模分析, 可以实现有效药物组合的高通量筛选. 本文提出一种基于相似性特征的预测药物组合和细胞系(drug-drug-cell line, DDC)关系的新方法, 用于筛选出特异于细胞系的协同或拮抗的药物组合. 具体地, 首先使用S-kernel和高斯核分别计算药物组合和细胞系基于相似性的特征向量, 然后拼接两向量得到药物组合-细胞系的特征向量, 以此作为机器学习模型的输入特征. 基于药物协同实验的DDC关系作为机器学习的输出. 三种机器学习模型, 包括深度神经网络(deep neural network, DNN)、随机森林(random forest, RF)和支持向量机(support vector machine, SVM), 交叉验证结果表明, 新方法稳定可行, 且深度神经网络和随机森林分类准确率高达89%~91%. 重要的是, 基于新方法的预测模型能够预测包含未知药物分子或细胞系的全新DDC组合. 本文提出的特征计算方法能够使机器学习模型准确预测药物组合同细胞系之间的关系, 为药物组合协同预测提供了一种新方法.

关键词 核方法, 机器学习, 相似性计算, 协同药物组合, 细胞系特异

在2022年中国癌症发生和死亡报告中, 2016年中国新发癌症病例约4064000例, 死亡数约2413500例, 表明癌症仍是威胁中国公共健康的主要因素, 且癌症造成的疾病负担正在上升[1]. 尽管过去20年在致癌机制和药物研发上均有很大进展, 但是耐药性使单一药物在临床上的治疗效果非常有限[2]. 作为肿瘤发展和进化的结果, 肿瘤异质性使不同分子特征的细胞亚群对单一药物的反应差异显著, 耐药细胞亚群残留能够引起癌症复发从而减少生存时间[3]. 此外, 负责药物转运和代谢的功能蛋白的基因突变和抗癌药物的选择压力, 都会导致药物脱敏[4~8]. 近年来关于肿瘤微环境的研究也说明了肿瘤治疗的难度[9]. 近20年来, 大量临床试验表明, 相比单一用药, 协同组合用药可以显著改善癌症患者的生存率[10~18], 是有效的治疗策略之一[19~26].

药物组合疗法是通过同时靶向多条信号通路或多个靶蛋白来避免耐药性的. Mokhtari等人[27]在肺支气

引用格式: 张竣, 袁锐, 陈世龙, 等. 基于核方法的协同药物组合预测方法. 中国科学: 生命科学, 2023, 53: 1663–1672
Zhang J, Yuan R, Chen S L, et al. Kernel-based prediction of a synergistic drug combination (in Chinese). Sci Sin Vitae, 2023, 53: 1663–1672, doi: 10.1360/SSV-2023-0033

管类癌症的体内体外实验中, 证明乙酰唑胺(acetazolamide)和萝卜硫素(sulforaphane)组合能明显抑制肿瘤细胞生长, 且相对于单独用药, 组合治疗的有效剂量明显降低, 减小了大剂量用药的副作用风险. 然而, 药物组合的有效性高度依赖于特定的细胞分子环境并且与剂量和时间有关[28~36], 所以截至目前, 实验验证的有效组合还很少[37]. 在上百种细胞系基础上对上千种小分子进行药物组合筛选需要大量的组合实验, 需要耗费极大的人力物力. 利用计算方法模拟药物-药物相互作用(drug-drug interactions, DDIs)及药物和生物分子之间的相互作用, 用于在特定细胞中预测新的药物组合效果, 得到有效的实验候选集, 可以大大缩小实验范围, 提升药物研发效率.

近年来, 学者们尝试构建多种计算方法准确预测药物组合. Vilar等人[38,39]基于具有相似分子结构的药物具有相似的生物活性这一假设[40,41], 通过计算候选药物与已知DDI中一种药物的相似性, 推断候选药物与DDI中另一种药物是否发生相互作用. 基于结构相似性的数学模型与分子表征(molecular descriptor)和计算相似性的选择有关, 并且对新药的推断依赖于已知DDI. 事实上, Jaaks等人[37]的研究表明, DDI的类型强烈依赖于细胞的分子环境, 文章中所定义的DDI类型建立在药效学(pharmacodynamics)层面. O'Neil等人[42]发表的大型实验数据集加速了机器学习方法的研究, 大量机器学习方法用于结合药物组合和细胞的分子特征来推断药物组合在特异细胞系中的相互作用. 近些年, 计算机算力的大幅提升加速了深度学习在理论和应用上的发展, 如自然语言和图像处理[43,44]. DeepMind公司开发的AlphaFold已经能够预测媲美实验方法的蛋白质三维结构[45]. 在药物组合预测方面, 学者们提出的深度学习模型[46,47]充分表明其在该领域的合理性和强大的学习能力.

基于相似的药物组合在相似的细胞系中应该发挥相近作用这一假设, 本研究提出基于核方法计算机器学习模型的输入特征, 用于预测药物组合同细胞的关系, 进而筛选出细胞特异的协同或拮抗的药物组合. 具体地, 首先引入S-kernel和高斯核分别计算药物组合及细胞系的相似性, 然后将其分别作为表征药物组合和细胞系的特征向量, 并用拼接后的向量作为机器学习模型算法的输入特征, 而将基于NCI-ALMANAC实验数据的DDC关系作为机器学习的输出. 数值实验表明, 基于新特征的预测模型能够学习药物组合在不同细胞系中的表现. 进一步独立测试集的验证实验表明, 模型能够很好地学习药物组合和细胞系之间的关系.

1 材料和方法

1.1 数据准备

本实验机器学习模型训练和验证数据使用NCI-ALMANAC数据集. 数据集包含295688个DDC组合, 每个组合由药物A、药物B和细胞系C组成(两个药物的浓度的组合设计为3×3或5×3, 总共2873515个药物A(浓度*M*)-药物B(浓度*N*)-细胞系C实验测试). 这些DDC组合共包括由105种药物分子得到的5355种药物组合和60株肿瘤细胞系. 药物分子的结构数据文件(structure data file, SDF)取自数据集补充材料. 60株肿瘤细胞系源自NCI-60细胞系库, 这些细胞系的转录组表达值在Cell Model Passports数据库[48]中下载, 共得到57种(SNB-19, NCI/ADR-RES, MDA-MB-435无转录组数据)细胞系的转录组表达值, 囊括白血病、非小细胞肺癌、结直肠癌、中枢神经系统癌症、黑色素瘤、卵巢癌、肾癌、前列腺癌和乳腺癌9种癌症类型[49].

(1) 组合协同分数计算. 本实验是药物组合协同性预测分类任务, 即判定某个药物组合在特定细胞系上的相互作用是协同还是非协同(药效学层面, DDI包括协同、加性和拮抗相互作用)[50]. NCI-ALMANAC实验数据使用生长百分比(observed percent growth, OPG)表示细胞系在特定药物组合下的生长百分比, 通过计算细胞系在药物组合下的期望生长百分比(expected percent growth, EPG)作为不发生相互作用的临界值, 最终得到DDC组合的协同分数[51]: ComboScore=EPG−OPG. 该数据集中ComboScore大于0, 表示协同相互作用; 等于0, 表示加性相互作用; 小于0, 表示拮抗相互作用. 在预实验分析阶段, 使用不同的ComboScore阈值标记药物组合的协同性并训练随机森林模型, 发现使用上四分位数作为更高的阈值使性能表现更好, 计算中把协同DDC组合标签记为1; 加性和拮抗组合标签记为0.

(2) 药物分子表征和转录组特征. 本方法首先需要获取能表征药物分子的物理化学属性值和细胞系的转录组表达值(图1A). 药物分子的分子表征使用RDkit计算, 包括208种物理化学性质, 最终得到含有105×208个

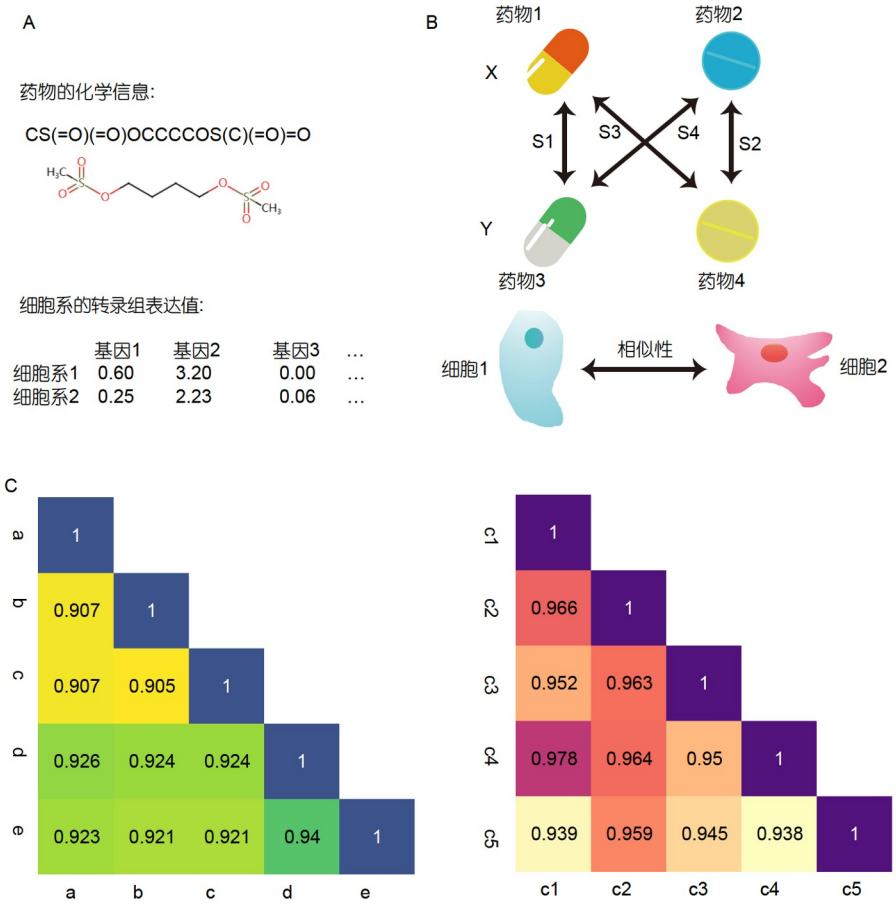

图 1　相似性计算流程. A: 药物分子和细胞系信息收集; B: 药物组合X和Y及细胞系1和2的相似性计算方式; C: 药物组合和细胞系的相似性矩阵

Figure 1　Similarity calculation procedure. A: Collecting information on drugs and cell lines; B: calculating the similarity between drug combination X and Y and cell lines 1 and 2; C: similarity matrix of the drug combinations and cell lines

数值元素的药物分子表征文件. Cell Model Passports数据库中细胞系的转录组数据包含37607个基因, 最终得到含有57×37607个数值元素的细胞系转录组表达值文件. 进一步对每个药物分子的分子表征和细胞系转录组的TPM(transcript per million)表达值做样本水平的L2归一化处理, 并过滤掉药物分子和细胞系的空值特征.

(3) 输入特征计算. 本方法基于相似的药物组合在相似细胞系中的相互作用性质相似这一假设, 通过计算药物组合间和细胞系间的相似性得到DDC组合间的相似性, 并作为机器学习模型的输入特征. 本方法首先分别计算药物组合间的相似性向量以及细胞系间的相似性向量, 然后拼接得到DDC组合的相似性向量. 考虑到药物组合中的相互作用具有对称性质, 即两个药物分子的向量表示顺序不影响药物组合的性质, 本方法采用S-kernel[52,53]计算药物组合间的相似性(图1B). 首先, 药物分子在本实验中被表示为高维向量, 因此使用高斯核函数计算药物组合间四对药物分子的相似性, 如式(4)和(5). 然后得到两个算数平均数, 取最大值, 式(1)~(3). 最后, 相同药物组合的相似性值设为1.

$$S(X,Y)=\max(\mathrm{Sim}_1,\mathrm{Sim}_2), \tag{1}$$

$$\mathrm{Sim}_1=\frac{S1+S2}{2}, \tag{2}$$

$$\mathrm{Sim}_2=\frac{S3+S4}{2}, \tag{3}$$

式(2)和(3)中Si, i=1, 2, 3, 4是计算两个药物之间相似性的高斯核函数(σ=0.1), 即

$$S_1 = e^{-\frac{\|\text{Drug1}-\text{Drug3}\|^2}{2\sigma^2}}, S_2 = e^{-\frac{\|\text{Drug2}-\text{Drug4}\|^2}{2\sigma^2}}, \quad (4)$$

$$S_3 = e^{-\frac{\|\text{Drug1}-\text{Drug4}\|^2}{2\sigma^2}}, S_4 = e^{-\frac{\|\text{Drug2}-\text{Drug3}\|^2}{2\sigma^2}}, \quad (5)$$

$$\text{Sim}_{\text{cell}} = e^{-\frac{\|\text{cell1}-\text{cell2}\|^2}{2\sigma^2}}. \quad (6)$$

而对于组合中细胞系的相似性，本方法使用高斯核函数(σ=0.1)进行计算，如式(6). 如图1C所示，最终得到药物组合和细胞系的相似性对称矩阵.

1.2 机器学习模型构建和评估

(1) 模型构建. 为证实本文提出的输入特征计算方法的可行性和稳健性，本实验构建了传统机器学习方法中的随机森林(random forest, RF)、支持向量机(support vector machine, SVM)模型和深度学习方法中的深度神经网络(deep neural network, DNN)模型三个机器学习模型, 并基于新的特征计算方法进行训练. 随机森林[54]是基于决策树的集成方法，输出结果即决策树的投票结果；支持向量机[55]是一种二分类模型，通过使用核技巧可以作为非线性分类器，本质是寻找最佳分类超平面；深度神经网络[56,57]是包含输入层、隐藏层、输出层的全连接网络，通过反向传播算法和损失函数学习输入和输出间的关系. 本实验的程序基于Python，其中随机森林和支持向量机使用Scikit-Learn机器学习库，深度神经网络使用Keras深度学习平台接口.

(2) 性能评估. 本文采用不同的评估方法对新方法的可行性、稳定性和分类准确性进行评估. 第一, 实验中三种模型的预测值是药物组合属于协同或者非协同的概率，所以使用F_1分数(如式(7)，即精确率(precision)和召回率(recall)在模型准确性评价中权重相同)计算模型最佳得分下的分类概率阈值. 第二, 使用ROC曲线下面积(area under curve, AUC)和PR曲线下面积(area under the precision-recall curve, AUPR)计算并比较三种模型的准确性大小. 为分析本方法的稳定性, 本实验用10个随机种子对三种模型重复训练10次，AUC、AUPR和F_1分数均取平均值和标准差. 式(7)中, Precision=TP/(TP+FP), Recall=TP/(TP+FN). TP表示真阳性, FP表示假阳性, FN表示假阴性.

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (7)$$

训练过程中，传统机器学习算法和深度学习算法采取不同的训练和验证方式，即随机森林和支持向量机算法使用网格搜索和交叉验证调整选取最佳参数, 深度学习算法则使用划分的训练集、验证集训练并验证模型准确性来调整参数，在独立数据集上测试模型的泛化应用能力. 对于深度神经网络，按照8:1:1比例划分训练集、验证集、测试集，三个数据集中的DDC组合可以包括相同的药物-药物组合或者细胞系, 但三个数据集间不包含相同的DDC组合. 此外，本实验使用深度神经网络验证本方法的泛化应用能力，独立的测试数据集来自DrugCombDB数据库[58]中新的DDC组合.

2 结果

2.1 模型超参数选择

对于随机森林模型, 本实验使用网格搜索和5折交叉验证选择最优决策树数量、决策树最大深度、划分内部节点需要的最小样本量、分类特征数量、叶子节点含有最小样本量，即每一个参数组合都使用交叉验证评估模型. 本实验共进行两轮网格搜索以减少参数组合数量和计算量: 第一轮网格搜索设置决策树数量{100, 300, 500, 600}, 最大树深度{8, 16, 32, 64}, 划分内部节点需要的最小样本量{2, 4, 6}; 第二轮在第一轮的最优参数下，网格搜索设置分类特征数量{0.014, 0.05, 0.1}(数值表示占总特征数量的比例)，叶子节点含有最小样本{1, 2, 3, 4}. 经过5折交叉验证，得到最优参数: 决策树数量为100, 最大树深度为32, 划分内部节点需要的最小样本量为6, 分类特征数量0.014, 叶子节点含有最小样本量4.

对于支持向量机模型, 使用网格搜索和5折交叉验证调试最优参数: 正则化参数C、核系数γ. 首先, 进行粗调得到近似的最优参数, C设置为{1, 4, 16, 32, 64, 256}, γ设置为{2^{-10}, 2^{-8}, 2^{-6}, 2^{-4}, 2^{-2}}，较优参数为C=256, γ=0.25; 然后，进行细调得到最优参数, C设置为{250, 256, 262}, γ设置为{0.20, 0.25, 0.30}, 得到最优参数C=256, γ=0.20.

对于深度神经网络，如式(8)~(12)，除输出层激活函数使用sigmoid函数，隐藏层使用ReLU激活函数. 模型输入层到输出层神经元数量依次为{5421, 256, 128, 64, 10, 1}, 损失函数使用二值交叉熵函数, 训练过程使

用随机梯度下降优化器进行迭代优化. 在不同学习率和样本批大小组合的实验中, 学习率设置为10^{-5}, 批大小设置为256. 为避免过拟合, 在训练过程中使用提前终止机制, 如果验证集上的损失连续18个epoch中没有减小就停止训练, 最大训练轮数1000个epoch.

下式中Linear层表示全连接层, BN(batch normalization)表示批归一化:

$$\mathrm{Out}_1=\mathrm{ReLU}(\mathrm{BN}(\mathrm{Linear}(\mathrm{Input}))), \tag{8}$$

$$\mathrm{Out}_2=\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Linear}(\mathrm{Out}_1)\right)\right), \tag{9}$$

$$\mathrm{Out}_3=\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Linear}(\mathrm{Out}_2)\right)\right), \tag{10}$$

$$\mathrm{Out}_4=\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Linear}(\mathrm{Out}_3)\right)\right), \tag{11}$$

$$\mathrm{Out}_5=\mathrm{Sigmoid}\left(\mathrm{Linear}(\mathrm{Out}_4)\right). \tag{12}$$

2.2　模型性能比较

为得到稳定可信的结果, 三个模型均设置10个随机种子重复训练10次, 每次训练计算AUC, AUPR及F_1分数. 另外, 文章与新颖的基于图神经网络的算法GraphSynergy[59]进行了比较. 结果如表1所示, 基于本方法的随机森林模型准确性最高, 用粗体标出, 其次是深度神经网络模型. 此外, 使用插值方法可以计算10次重复ROC曲线的平均值和标准差(图2). 结果表明, 本文提出的输入特征计算方法可重复性高、稳定性好. 更进一步, 实验结果也表明, 本文提出的特征计算方法能较好地表示药物组合和细胞系的特征, 并且基于本方法的机器学习模型能够学习药物组合的协同性和细胞系的依赖关系.

2.3　新药物组合-细胞系预测

最后, 本实验对来自其他数据集的新药物组合-细胞系组合数据作为测试集进行预测. 由于测试数据来自不同实验数据集, 不同的实验设计和协同性计算标准会导致相同的药物组合在相同细胞系上协同性结果不一致[60]. 本文中的独立测试数据集原始数据不能基于ComboScore计算参考标签, 因此基于DrugCombDB数据库中Bliss, Loewe, HSA和ZIP四种标准得到药物组合的协同性标签, 如果超过两种标准的标签为1, 则该药物组合为协同相互作用. 测试数据集中非小细胞肺癌细胞系(A427, NCI-H1650, NCI-H2122, SK-MES-1)和肝癌细胞系(HuH-7)的药物组合实验数据分别来自O'Neil等人的研究和NIH(National Institutes of Health).

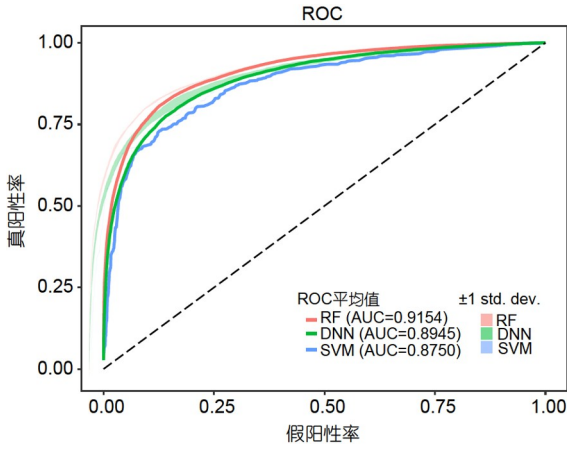

图 2　基于核方法模型的平均ROC曲线

Figure 2　Average ROC curves for the kernel-based model

最终测试集共包括5个细胞系和859个药物-药物组合组成的2428个药物组合-细胞系组合. 预测过程中, 本实验使用训练数据最多的深度神经网络模型、数据预处理和模型参数与训练过程保持一致. 部分预测示例见表2, 其中L表示实验数据标签, P表示预测标签, Odds为深度学习模型输出的概率. 有的药物组合-细胞系的协同性在重复性实验中不一致, 难以确定该组合的标签. 在非小细胞肺癌细胞系和肝癌细胞系中,

表 1　模型的性能比较

Table 1　Comparison of model performance

算法	AUC	AUPR	F_1
DNN	0.8945±0.0056	0.8227±0.0085	0.7416±0.0075
RF	**0.9154±0.0012**	**0.8612±0.0023**	**0.7753±9.5×10^{-4}**
SVM	0.8750±1.0×10^{-5}	0.7871±4.8×10^{-6}	0.7297±0
GraphSynergy	0.8349±0.0011	0.8163±0.0014	0.7287±0.0017

表 2 新药物组合预测

Table 2 Novel drug combination predictions

药物1	药物2	细胞系	ZIP	Bliss	Loewe	HSA	L	P	Odds
Carboplatin	AZD1775	A427	2.62	2.48	3.15	7.29	1	1	0.799
Temozolomide	Geldanamycin	A427	4.24	1.98	1.64	7.51	1	1	0.777
Sunitinib	SN-38	A427	5.10	5.32	2.52	11.23	1	1	0.774
MK-5108	Dasatinib	A427	3.31	2.80	9.43	12.48	1	1	0.756
Etoposide	Sunitinib	A427	5.51	4.27	5.38	12.07	1	1	0.754
Temozolomide	AZD1775	A427	12.86	11.62	14.74	20.01	1	1	0.747
MK-5108	MK-8669	A427	1.11	0.88	4.34	5.79	1	1	0.698
Zolinza	Erlotinib	A427	5.25	4.97	7.28	11.07	1	1	0.694
Sunitinib	Lapatinib	A427	11.88	11.29	13.19	17.58	1	1	0.692
Colchicine	Omacetaxine mepesuccinate	Huh-7	−12.43	−13.72	−36.38	−3.10	0	1	0.760
Clomiphene citrate	Colchicine	Huh-7	−8.99	−2.66	−2.86	−1.12	0	1	0.734
Colchicine	Apilimod	Huh-7	−12.86	−11.81	−13.49	−1.24	0	1	0.680
Amodiaquin dihydrochloride dihydrate	Bepridil	Huh-7	−7.29	−5.38	−3.30	−0.23	0	0	0.048
Sunitinib	Aripiprazole	Huh-7	−6.53	−12.01	−33.66	−9.97	0	0	0.149
Colchicine	Aripiprazole	Huh-7	−8.98	−9.80	−2.70	−1.52	0	0	0.146
Azithromycin	Apilimod	Huh-7	−4.23	−4.47	−71.29	−1.76	0	0	0.030
Clomifene citrate	Mycophenolate mofetil	Huh-7	−4.82	−5.39	−7.99	−1.19	0	0	0.026
Azithromycin	Toremifene citrate	Huh-7	−5.66	−9.81	-6.48	−6.48	0	0	0.024
Colchicine	Favipiravir	Huh-7	−7.77	−8.31	−1.08	−1.08	0	0	0.010

预测为协同的组合占比约10%.

值得注意的是, 肝癌中的药物组合大多为治疗其他非癌症疾病的小分子药物组成, 但近年来有些被用于癌症治疗. 例如秋水仙素和高三尖杉酯碱: 秋水仙素是一种有丝分裂抑制剂, 实验证明在临床可接受的剂量范围内, 秋水仙素对肝癌细胞系有抗癌作用[61]; 高三尖杉酯碱是一种蛋白质合成抑制剂, 有实验证明在肝癌的患者来源类器官中, 低剂量的高三尖杉酯碱是一种很有效的小分子抑制剂[62]. 秋水仙素和高三尖杉酯碱值得进行进一步的体内体外组合实验.

综上, 基于新特征计算方法的DNN模型具有预测药物组合-细胞系协同性标签的能力, 得到具有协同相互作用药物组合或者筛除加性或拮抗相互作用药物组合, 以达到加速药物筛选过程的目的.

3 讨论

药物组合治疗策略相对于单一用药治疗, 除了能在一定程度解决单一药物产生的抗性问题, 协同药物组合还有减少用药剂量, 减轻毒副作用等优点. 新的药物研发周期过长. 综上, 目前药物组合治疗策略是针对复杂疾病的有效治疗策略之一.

20世纪到21世纪初, 对药物组合的计算筛选方法主要是数学建模的方法, 例如Gottlieb等人[63]提出的通过计算组合间的相似性推断未知组合发生DDI的可能性. 2016年, O'Neil等人[42]和NCI-ALMANAC[51]大型药物组合实验数据集发布以来, 利用机器学习方法在细胞分子环境下进行药物组合预测极大提高了筛选准确性和效率, 如深度学习模型DeepSynergy[47]和MatchMaker[46]. 更进一步, 深度学习方法使各种模态数据的融合更加容易, 例如, 细胞系的转录组数据和药物分子的物理化学数据. 药物分子和细胞系各种模态的海量数据让学者将最新的深度学习算法运用到药物组合预测的问题上, 使药物组合筛选获得极大突破.

本文结果表明, 新特征计算方法可以让机器学习算法的准确率达到91%, 重复实验也证明了方法的稳

定性, 并且能合理推测新的DDC组合. 然而, 目前本文和其他机器学习预测方法主要局限于以下几个方面: 首先, 计算相似性作为特征在维度方面的限制性, 尤其是作为机器学习模型的输入, 待预测药物组合数量及细胞系数量需要和训练数据相同. 本实验中通过计算独立数据集和训练数据集的相似性解决这一问题, 对新DDC组合的推测证明了其在实际应用中的合理性. 其次, 目前大型实验数据集都是两个药物分子的组合, 临床上往往会组合两种以上的药物实现更好的效果. 然后, 用于机器学习训练的几乎都是小分子化疗药物, 大分子抗体药物数据量极少且在特征表示方面极具挑战. 另外, 不只是肿瘤细胞内分子环境决定体内药物组合的效应, 肿瘤异质性、肿瘤微环境等多方面因素共同决定了药物组合在体内的效应, 在筛选的过程中如何考虑这些因素仍有待研究. 未来, 基于本方法利用更多模态数据的机器学习模型可能显著提高预测准确性. 综上, 本文提出的特征计算方法为筛选有价值的新药物组合提供了一种新途径.

参考文献

1 Zheng R, Zhang S, Zeng H, et al. Cancer incidence and mortality in China, 2016. J Natl Cancer Center, 2022, 2: 1–9

2 Huang M, Shen A, Ding J, et al. Molecularly targeted cancer therapy: some lessons from the past decade. Trends Pharmacol Sci, 2014, 35: 41–50

3 Lim Z F, Ma P C. Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy. J Hematol Oncol, 2019, 12: 134

4 Zahreddine H, Borden K L B. Mechanisms and insights into drug resistance in cancer. Front Pharmacol, 2013, 4: 28

5 Rupaimoole R, Calin G A, Lopez-Berestein G, et al. miRNA deregulation in cancer cells and the tumor microenvironment. Cancer Discov, 2016, 6: 235–246

6 Hinshaw D C, Shevde L A. The tumor microenvironment innately modulates cancer progression. Cancer Res, 2019, 79: 4557–4566

7 Peltomäki P. Mutations and epimutations in the origin of cancer. Exp Cell Res, 2012, 318: 299–310

8 Vo J N, Wu Y M, Mishler J, et al. The genetic heterogeneity and drug resistance mechanisms of relapsed refractory multiple myeloma. Nat Commun, 2022, 13: 3750

9 Sethi T, Rintoul R C, Moore S M, et al. Extracellular matrix proteins protect small cell lung cancer cells against apoptosis: A mechanism for small cell lung cancer growth and drug resistance *in vivo*. Nat Med, 1999, 5: 662–668

10 Dar T B, Biteghe F A N, Kakar-Bhanot R, et al. Synergistic effects of radiotherapy and targeted immunotherapy in improving tumor treatment efficacy: a review. Clin Transl Oncol, 2022, 24: 2255–2271

11 Grilli R, Oxman A D, Julian J A. Chemotherapy for advanced non-small-cell lung cancer: how much benefit is enough? J Clin Oncol, 1993, 11: 1866–1872

12 Non-small Cell Lung Cancer Collaborative Group. Chemotherapy in non-small cell lung cancer: a meta-analysis using updated data on individual patients from 52 randomised clinical trials. Bmj, 1995, 311: 899–909

13 Cullen M H, Billingham L J, Woodroffe C M, et al. Mitomycin, ifosfamide, and cisplatin in unresectable non-small-cell lung cancer: effects on survival and quality of life. J Clin Oncol, 1999, 17: 3188–3194

14 Sandler A, Gray R, Perry M C, et al. Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. N Engl J Med, 2006, 355: 2542–2550

15 Tanizaki J, Okamoto I, Takezawa K, et al. Combined effect of ALK and MEK inhibitors in EML4-ALK-positive non-small-cell lung cancer cells. Br J Cancer, 2012, 106: 763–767

16 Gandhi L, Rodríguez-Abreu D, Gadgeel S, et al. Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. N Engl J Med, 2018, 378: 2078–2092

17 Paz-Ares L, Luft A, Vicente D, et al. Pembrolizumab plus chemotherapy for squamous non-small-cell lung cancer. N Engl J Med, 2018, 379: 2040–2051

18 Hellmann M D, Paz-Ares L, Bernabe Caro R, et al. Nivolumab plus ipilimumab in advanced non-small-cell lung cancer. N Engl J Med, 2019, 381: 2020–2031

19 Herbst R S, Baas P, Kim D W, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung

cancer (KEYNOTE-010): a randomised controlled trial. Lancet, 2016, 387: 1540–1550
20 Reck M, Rodríguez-Abreu D, Robinson A G, et al. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. N Engl J Med, 2016, 375: 1823–1833
21 Mok T S K, Wu Y L, Kudaba I, et al. Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): a randomised, open-label, controlled, phase 3 trial. Lancet, 2019, 393: 1819–1830
22 Thai A A, Solomon B J, Sequist L V, et al. Lung cancer. Lancet, 2021, 398: 535–554
23 Borghaei H, Paz-Ares L, Horn L, et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. N Engl J Med, 2015, 373: 1627–1639
24 Brahmer J, Reckamp K L, Baas P, et al. Nivolumab versus docetaxel in advanced squamous-cell non-small-cell lung cancer. N Engl J Med, 2015, 373: 123–135
25 Garon E B, Rizvi N A, Hui R, et al. Pembrolizumab for the treatment of non-small-cell lung cancer. N Engl J Med, 2015, 372: 2018–2028
26 Rizvi N A, Hellmann M D, Snyder A, et al. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science, 2015, 348: 124–128
27 Mokhtari R B, Kumar S, Islam S S, et al. Combination of carbonic anhydrase inhibitor, acetazolamide, and sulforaphane, reduces the viability and growth of bronchial carcinoid cell lines. BMC Cancer, 2013, 13: 378
28 Ohsaki Y, Tanno S, Fujita Y, et al. Epidermal growth factor receptor expression correlates with poor prognosis in non-small cell lung cancer patients with p53 overexpression. Oncol Rep, 2000, 7: 603–607
29 Nicholson R I, Gee J M W, Harper M E. EGFR and cancer prognosis. Eur J Cancer, 2001, 37: 9–15
30 Hirsch F R, Varella-Garcia M, Bunn Jr P A, et al. Epidermal growth factor receptor in non-small-cell lung carcinomas: correlation between gene copy number and protein expression and impact on prognosis. J Clin Oncol, 2003, 21: 3798–3807
31 Lynch T J, Bell D W, Sordella R, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med, 2004, 350: 2129–2139
32 Paez J G, Janne P A, Lee J C, et al. *EGFR* mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science, 2004, 304: 1497–1500
33 Kobayashi S, Boggon T J, Dayaram T, et al. *EGFR* mutation and resistance of non-small-cell lung cancer to gefitinib. N Engl J Med, 2005, 352: 786–792
34 Engelman J A, Zejnullahu K, Mitsudomi T, et al. *MET* amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science, 2007, 316: 1039–1043
35 Sharma S V, Bell D W, Settleman J, et al. Epidermal growth factor receptor mutations in lung cancer. Nat Rev Cancer, 2007, 7: 169–181
36 Shepherd F A, Rodrigues Pereira J, Ciuleanu T, et al. Erlotinib in previously treated non-small-cell lung cancer. N Engl J Med, 2005, 353: 123–132
37 Jaaks P, Coker E A, Vis D J, et al. Effective drug combinations in breast, colon and pancreatic cancer cells. Nature, 2022, 603: 166–173
38 Vilar S, Harpaz R, Uriarte E, et al. Drug-drug interaction through molecular structure similarity analysis. J Am Med Inform Assoc, 2012, 19: 1066–1074
39 Vilar S, Uriarte E, Santana L, et al. Similarity-based modeling in large-scale prediction of drug-drug interactions. Nat Protoc, 2014, 9: 2147–2163
40 Matter H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. J Med Chem, 1997, 40: 1219–1229
41 Martin Y C, Kofron J L, Traphagen L M. Do structurally similar molecules have similar biological activity? J Med Chem, 2002, 45: 4350–4358
42 O'Neil J, Benita Y, Feldman I, et al. An unbiased oncology compound screen to identify novel combination strategies. Mol Cancer Ther, 2016, 15: 1155–1162
43 Li H. Deep learning for natural language processing: advantages and challenges. Natl Sci Rev, 2018, 5: 24–26
44 Xu Z, Sun J. Model-driven deep-learning. Natl Sci Rev, 2018, 5: 22–24
45 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583–589
46 Kuru H I, Tastan O, Cicek A E. MatchMaker: a deep learning framework for drug synergy prediction. IEEE ACM Trans Comput Biol Bioinf, 2021, 19: 2334–2344
47 Preuer K, Lewis R P I, Hochreiter S, et al. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. Bioinformatics, 2018, 34:

1538–1546
48 van der Meer D, Barthorpe S, Yang W, et al. Cell Model Passports—a hub for clinical, genetic and functional datasets of preclinical cancer models. Nucleic Acids Res, 2019, 47: D923–D929
49 Shoemaker R H. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer, 2006, 6: 813–823
50 Greco W R, Bravo G, Parsons J C. The search for synergy: a critical review from a response surface perspective. Pharmacol Rev, 1995, 47: 331–385
51 Holbeck S L, Camalier R, Crowell J A, et al. The National Cancer Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. Cancer Res, 2017, 77: 3564–3576
52 Wang Y C, Chen S L, Deng N Y, et al. Computational probing protein-protein interactions targeting small molecules. Bioinformatics, 2016, 32: 226–234
53 Shen J, Zhang J, Luo X, et al. Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci USA, 2007, 104: 4337–4341
54 Breiman L. Random forests. Machine Learn, 2001, 45: 5–32
55 Noble W S. What is a support vector machine? Nat Biotechnol, 2006, 24: 1565–1567
56 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444
57 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313: 504–507
58 Liu H, Zhang W, Zou B, et al. DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. Nucleic Acids Res, 2020, 48: D871–D881
59 Yang J, Xu Z, Wu W K K, et al. Erratum to: GraphSynergy: a network-inspired deep learning model for anticancer drug combination prediction. J Am Med Inf Assoc, 2022, 29: 220
60 Goldoni M, Johansson C. A mathematical approach to study combined effects of toxicants *in vitro*: Evaluation of the Bliss independence criterion and the Loewe additivity model. Toxicol in Vitro, 2007, 21: 759–769
61 Lin Z Y, Wu C C, Chuang Y H, et al. Anti-cancer mechanisms of clinically acceptable colchicine concentrations on hepatocellular carcinoma. Life Sci, 2013, 93: 323–328
62 Li L, Halpert G, Lerner M G, et al. Protein synthesis inhibitor omacetaxine is effective against hepatocellular carcinoma. JCI Insight, 2021, 6: e138197
63 Gottlieb A, Stein G Y, Oron Y, et al. INDI: a computational framework for inferring drug interactions and their associated recommendations. Mol Syst Biol, 2012, 8: 592

Kernel-based prediction of a synergistic drug combination

ZHANG Jun[1,2], YUAN Rui[1], CHEN ShiLong[1,3] & WANG YongCui[1]

1 Key Laboratory of Adaptation and Evolution of Plateau Biota, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining 810008, China;
2 College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China;
3 Institute of Sanjiangyuan National Park, Chinese Academy of Sciences, Xining 810008, China

Complex diseases, such as cancer, often exhibit resistance to single drug therapy because of their heterogeneity and complex metabolic pathways. Combination therapy is an efficient strategy to overcome drug resistance. As experimental screenings consume considerable resources and have low efficacy, the computational method is a good alternative. Thus, this article proposes a new method for computing features of a drug-drug-cell line (DDC) combination based on similarity, where the S-kernel and Gaussian-kernel methods are used to calculate the drug-drug combination similarity and cell line similarity, respectively. The final feature vector for machine learning input was obtained by concatenating these two vectors. The output for machine learning was based on the experimental results of the synergistic drug combination. Cross validation was performed on three machine learning algorithms, including the random forest, support vector machine, and deep neural network models. The results suggested that the novel method had a robust performance with an area under the curve value of 0.89–0.91. Importantly, the model predicted the novel DDC combinations with new drugs or new cell lines based on unique input features. In conclusion, this novel method improved predictive performance and provided a new strategy for predicting synergistic drug combinations.

kernel method, machine learning, similarity computation, synergistic drug combinations, cell type specificity

doi: 10.1360/SSV-2023-0033