科学通报 2020 年 第 65 卷 第 32 期: 3551 ~ 3561
进 展
《中国科学》杂志社
SCIENCE CHINA PRESS

人工智能算法在药物细胞敏感性预测中的应用

李叙潼[1,2], 吴小龙[1,3], 万晓喆[1,2], 钟飞盛[1,2], 崔晨[1,2], 陈颖佳[1,2], 陈立凡[1,2], 陈凯先[1,2], 蒋华良[1,2*], 郑明月[1,2*]

1. 中国科学院上海药物研究所, 上海 201203;
2. 中国科学院大学药学院, 北京 100049;
3. 华东理工大学药学院, 上海 200237
* 联系人, E-mail: hljiang@simm.ac.cn; myzheng@simm.ac.cn

2020-05-09 收稿, 2020-06-15 修回, 2020-06-16 接受, 2020-06-17 网络版发表
国家自然科学基金(81773634)、国家科技重大专项(2018ZX09711002)和中国科学院战略性先导科技专项(XDA12050201)资助

摘要 开发基于癌症患者基因组信息预测有效治疗策略的计算模型是精准医学中的关键挑战. 近年来, 国际多个组织机构公开了针对数百种细胞系的多层次的基因组表征数据. 将这类组学数据与体外肿瘤细胞系的药物细胞敏感性相结合, 研究人员可以剖析癌症治疗药物的分子机制, 并将其转化为精准医学所需的个性化诊疗策略. 基于大数据的人工智能算法在基因组学与药物响应之间建立了新的桥梁, 推进了肿瘤细胞中药物敏感性的预测算法的发展. 本文首先对公开的基因组表征数据集进行了总结, 随后介绍了基因组表征数据和包括机器学习算法、网络算法和多模态神经网络算法在内的人工智能算法在癌细胞的药物敏感性预测中的应用案例. 基于网络的预测方法和多模态深度学习方法有利于实现多组学数据的系统性的整合和应用, 能克服传统的机器学习方法在药物响应预测中的局限性, 是今后药物敏感性研究的发展方向.

关键词 药物敏感性, 机器学习, 网络, 多模态

精准医学的不断发展以及许多大规模药物基因组学数据集公开推动了药物敏感性的预测模型研究. 基于肿瘤的异质性[1], 相同癌症类型的患者可能会对同一种药物表现出不同的响应, 因此开发基于癌症患者基因组信息预测有效治疗策略的计算模型是精准医学中的关键挑战. 个性化医疗指的是针对个别患者量身定制的疗法, 而不是为普通患者设计的“一刀切”的方法. 20世纪发展的个性化医疗主要基于可视化方法, 如X射线或核磁共振成像(magnetic resonance imaging, MRI)及肿瘤组织病理学检查. 21世纪以来, 下一代测序技术(next generation sequencing, NGS)为个性化医疗开辟了许多可能性. 全世界多个组织机构通过其高通量平台产生了针对数百种细胞系的多层次的分子表征数据, 包括DNA结构变异、基因转录水平和DNA甲基化等. 将这类组学数据与药物的肿瘤细胞系敏感性相结合, 研究人员可以剖析癌症治疗药物的分子机制, 并将其转化为精准医疗所需的个性化诊疗策略.

近年来, 基于大数据的人工智能算法的进步彻底改变了生物信息学, 为基因组数据的系统性应用引入了新技术. 它在药物基因组学上的应用为基因组学与药物敏感性之间建立了新的桥梁, 推进了肿瘤细胞中药物敏感性的预测算法的发展[2]. 本文将对公开的基因组学数据集进行总结, 讨论基因组表征数据以及包括机器学习算法、网络算法和多模态神经网络算法在

引用格式: 李叙潼, 吴小龙, 万晓喆, 等. 人工智能算法在药物细胞敏感性预测中的应用. 科学通报, 2020, 65: 3551–3561
Li X T, Wu X L, Wan X Z, et al. The application of artificial intelligence to drug sensitivity prediction (in Chinese). Chin Sci Bull, 2020, 65: 3551–3561, doi: 10.1360/TB-2020-0557

© 2020《中国科学》杂志社

内的人工智能算法在药物癌症细胞敏感性预测中的应用研究.

1 基因组表征数据集

对于机器学习方法, 输入数据的数量和质量很大程度上决定了模型的准确度. 目前, 可以用于药物响应研究的输入数据多为包含了药物剂量响应(dose-response)数据的公开细胞系数据库. 药物敏感性测定通常是在一定的化合物浓度范围内测量与包括细胞死亡、细胞增殖或凋亡等细胞生理事件相关的终点. 一系列浓度的药物-响应曲线计算出的IC_{50}、AUC和E_{max}通常被用作药物敏感性的度量[3]. 药物剂量响应测量是对新疗法和现有疗法进行临床前评估的基础, 尽管单个癌细胞系并不能真实反映临床上的癌细胞组织的复杂性[1], 但如果将其汇总起来剖析, 可以在一定程度上概括出人类癌症的基因多样性, 从而较为容易地建立研究药物响应的预测平台.

20世纪80年代后期, 美国国家癌症研究所(United States National Cancer Institute, NCI)的发展治疗计划(development therapeutics program, DTP)提出了针对9种癌症类型的60个(后为59个)人类肿瘤细胞系抗癌药物筛选项目, 即NCI60[4]. NCI60以每年筛选多达3000个小分子的抗癌活性为目标, 目前已经完成了对超过130000个小分子癌症细胞系敏感性测试. 尽管NCI60所筛选的小分子覆盖了广泛的化学空间, 但细胞系的多样性有其局限性. 2012年3月, *Nature*同期公布了两个大型癌症基因组研究项目, 分别为哈佛大学-麻省理工学院Broad研究院(Broad Institute of MIT and Harvard)与诺华生物医学研究所(Novartis Institutes for BioMedical Research)合作推出的癌细胞系百科全书(cancer cell line encyclopedia, CCLE)[5]和麻省总医院癌症中心(Massachusetts General Hospital Cancer Center)与Sanger研究院(Wellcome Sanger Institute)癌症基因组项目合作公布的生物标志物发现平台, 即癌症药物敏感性的基因组学数据库(genomics of drug sensitivity in cancer, GDSC)[6,7]. CCLE和GDSC分别具有来自15和36种癌症类型约1000种细胞系的遗传分子特征及其对化合物的响应. 随后, Broad研究院发布的癌症治疗反应门户(cancer therapeutics response portal, CTRP v1 & v2)进一步扩展了癌症细胞系敏感性数据集, 包含了481个小分子对的CCLE 860个癌症细胞系体外敏感性数据, 补充了CCLE数据集化合物覆盖较少的不足[8,9]. 2012年, DREAM(Dialogue for Reverse Engineering Assessment and Methods)项目与NCI合作, 发布了NCI-DREAM药物敏感性预测挑战赛, 以53个乳腺癌细胞系的遗传分子特征及针对28个治疗性化合物的响应值作为赛题[10]. 2015年, DREAM与阿斯利康(AstraZeneca)和Sanger研究院合作发起的AstraZeneca-Sanger DREAM挑战药物协同性预测挑战提供了较为独特的资源——一个包含118种药物的910对组合在85种细胞系的响应数据集, 这85种细胞系的分子遗传特征包含在GDSC数据库中[11].

事实上, 癌症本质上具有分子异质性, 即肿瘤由具有独特分子特征的癌细胞组成. 在人群水平上, 肿瘤异质性可分为肿瘤间(intertumoural)异质性和肿瘤内(intratumoural)异质性. 前者指患有相同组织学类型肿瘤患者之间的异质性, 而后者指单个患者肿瘤细胞间的异质性[1]. 因此, 肿瘤细胞系并不能真实地反映临床上癌细胞组织的复杂性, 这限制了基于肿瘤细胞系的药物敏感性研究的临床实用性. 为了描述人群中的肿瘤间异质性, 2006年, NCI和美国国家人类基因组研究所(National Human Genome Research Institute, NHGRI)共同提出癌症基因组图谱(The Cancer Genome Atlas, TCGA)计划. 迄今为止, TCGA已经收集了来自癌症患者的超过20000例原发性癌组织的分子特征, 并匹配了涵盖33种癌症类型的正常样本[12]. 此外, 癌症蛋白质图谱(The Cancer Proteome Atlas, TCPA)作为TCGA项目的一部分, 提供了超过4000个肿瘤样本的蛋白表达数据[13]. 尽管并未提供药物分子的敏感性测试数据, 但TCGA/TCPA常被用作其他数据集的补充或评估预测方法在真实患者上的可靠性[14,15].

除了上述基于未处理细胞的分子表征资源外, Broad研究所开发的CMap(The Connectivity Map)提供了进行遗传性和药理性、扰动的细胞分子特征图谱. 随后, CMap作为美国国立卫生研究院(National Institutes of Health, NIH)LINCS(Library of Integrated Network-based Cellular Signatures)项目的一部分, 基于L1000高通量基因表达谱分析技术, 建立了包含来自约20000种小分子和7000种遗传扰动的约47万个基因表达特征LINCS-L1000数据库[16]. 相比于未处理细胞的分子特征, 基于药理学扰动产生的癌细胞系基因表达特征(signatures)更为直接地与药物敏感性相关, 因此近年发表的文章中, 作为基因表达谱的补充, signatures也逐渐被引入药物响应的预测中以提升预测效果[17,18]. 表1

表 1 可公开获取的基因组数据集

Table 1 The publicly available genomic repositories

数据集	细胞系	化合物数目	主要数据类型[a)]	获取
NCI-60	9种类型, 59个细胞系	>100000	Mut, CNV, GE, Meth, PR	https://discover.nci.nih.gov/cellminer/
TCGA/TCPA	33种类型, 约20000个原发癌细胞	无	Mut, CNV, GE, Meth, PR, Hist	http://cancergenome.nih.gov/ http://tcpaportal.org/tcpa/
CCLE	15种类型, 1457个细胞系	24	Mut, CNV, Meth, GE	http://www.broadinstitute.org/ccle
GDSC	36种类型, 989个细胞系	518	Mut, CNV, Meth, GE	http://www.cancerrxgene.org/
CTRP	860个细胞系	481	Mut, CNV, Meth, GE	http://portals.broadinstitute.org/ctrp/
NCI-DREAM	53个乳腺癌细胞系	28	Mut, CNV, Meth, GE, PR	https://www.synapse.org/#!Synapse:syn2785778/wiki/70252
AstraZeneca-Sanger DREAM	85个癌症细胞系	118个药物的910种组合	Mut, CNV, Meth, GE	https://www.synapse.org/DrugCombinationChallenge
LINCS-L1000	98个癌症细胞系	19811个化合物、18493 shRNAs和3462 个cDNAs(针对5075个基因)、 314 个生物制剂	GE	http://lincsportal.ccs.miami.edu/ https://clue.io/

a) Mut: 基因突变; CNV: 基因拷贝数变异; GE: 基因表达谱; Meth: DNA甲基化; PR: 蛋白表达谱; Hist: 组织病理学图像

列出了上述数据集的内容及其获取方式.

2 基于机器学习的药物响应预测

用于建立药物反应预测模型的机器学习算法主要基于监督学习技术, 线性回归(linear regression)、岭回归(ridge regression)、拉索回归(lasso regression)和弹性网络回归(elastic net regression)等是常用的线性监督学习的算法; 核方法(kernel methods)、基于核的支持向量机(support vector machines, SVM)、决策树(decision trees)、随机森林(random forests, RF)和神经网络(neural networks)等则是非线性监督学习. 对于药物响应预测中常见的机器学习算法术语的解释参见Vougas等人[19]的文章.

2012年, NCI和DREAM共同组织了NCI-DREAM药物敏感性预测挑战赛, 超过65个队伍参与, 提交了44种基于机器学习的方法预测药物响应[10]. 其中, 基于线性监督学习的算法认为基因组学特征的组合线性可以用来推断药物响应, 与拉索回归和岭回归相比, 弹性网络回归表现更好. 实际上, 弹性网络方法可以识别具有高关联性的特征的子集, 同时保持其在预测中的重要性, 这些高关联性的特征可能提示了参与相同生物过程的基因/蛋白质. 因此, 弹性网络方法在赛后几年成为预测药物反应中最流行的线性模型, 在CCLE和GDSC数据集的分析中用于识别药物响应的预测因子, 揭示药物响应的生物标记物[5,6,20].

在非线性模型中, 基于核的BEMKL(Bayesian efficient multiple kernel learning)模型取得了最优表现[10,21]. 基于主办方提供的所有类型的组学数据, 包括体细胞突变、拷贝数变异、DNA甲基化、基因表达谱和蛋白表达谱等, BEMKL联合使用了核回归、多视图学习、多任务学习和贝叶斯推断, 以依据每个视图(异源性的输入)分别得到细胞系的成对相似性作为谱特异性的核, 再通过多核学习对谱特异性核加权求和, 得到组合核. 多任务学习是指所有药物(任务)的同时训练模型. BEMKL首先通过贝叶斯推断细胞系权重来估计每个核药物特异性的中间变量, 然后在所有药物之间共享核权重来估计输出药物响应矩阵. 同时, 通过引入超参数和误差项(偏差)解决由于药物响应数据集中缺失值和噪声的变化率引起的建模偏差, 从而提高药物敏感性预测的能力. 获胜队伍在随后发表的文章中, 将KBMTL(kernelized bayesian multitask learning)与两种单任务基线算法进行比较, 进一步强调引入多任务可以显著提升模型预测表现, 证实了KBMTL在HIV(human immunodeficiency virus)和癌症药物响应预测上的预测能力[22], KBMTL成为后续研究中常见的基准方法之一. 在其他非线性方法中, 基于核的SVM方法也应用于药物反应预测[23~25]. 随机森林方法在比赛中获得第二名[26], 其在后续研究中得到了扩展[24,27~29].

早期的基于机器学习的药物响应预测面临诸多挑战，这些挑战的解决为后续的研究奠定了基础. 比如, 早期的预测多使用药物特异性的模型，然而某个药物的响应可能与其他具有相似结构的药物相关. BEMKL (KBMTL)在竞赛中取胜强调了在药物之间共享信息对提高药物响应预测准确性的重要性，在此基础上发展的其他多任务学习模型的表现均优于单个模型[30~32]. Suphavilai等人[33]认为, 在给定药物的响应预测中, 多任务学习会给所有药物分配相同的重要性，但构建一个优先考虑最相似药物信息的模型可能更具价值，这可以通过推荐系统(recommender system)的协同过滤(collaborative filtering)技术实现. 对此, 他们提出基于推荐系统的药物响应预测方法，通过分析细胞系之间、药物之间的依赖关系, 识别新的药物响应. 我们将在后续小节中继续讨论这类引入药物相似性信息的模型.

药物响应数据的特征-样本比过大是另一个明显的挑战. 相比于包含数千个甚至上万个特征的基因组数据, 药物响应数据的样本量有限, 而在传统的机器学习模型中直接使用高维的组学数据会影响模型预测的精度. 这一问题的一个直接的处理方法是通过统计学手段对基因表达谱进行降维. Papillon-Cavanagh等人[34]使用了4种基于统计学特征选择方法，并结合线性回归模型与弹性网络模型预测药物响应. Parca等人[35]计算了每种基因的表达量与响应的相关性以及每种基因的表达量在样本中的方差，以此对基因表达谱进行降维, 并采用机器学习方法(弹性网络回归、支持向量机)预测药物的IC_{50}. 除了统计学方法，通过自编码器类算法得到的特征低维隐式表征也可以用于基因组数据的降维. Ding等人[36]用基于方差的特征选择方法将基因表达谱、拷贝数变异和突变数据提取并组合成5377位的细胞系特征，发现尽管在特定药物的模型上特征选择方法可以提升弹性网络的预测性能，但对整体的预测能力没有提升，这可能是由于在特征选择过程中丢失了一些有用的信息. 因此, 他们采用提取后的特征建立了无监督堆叠自编码器模型，并使用特定层内的隐变量作为每个细胞系的表征，然后针对每种药物训练弹性网络分类器，提升了采用特征选择方法建模不佳的药物的预测效果. Rampášek等人[18]选择了半监督的方式, 提出了药物反应变分自编码器(drug response variational autoencoder, Dr.VAE). 变分自编码器(variational autoencoder, VAE)是一种基于深度神经网络的生成模型，是基因表达谱的简化表示. 他们将CTRPv2的药物响应数据、基因表达谱和CMap-L1000的化合物扰动基因表达谱进行匹配构造训练数据集，使用半监督的VAE模型, 将基因表达谱嵌入到隐空间, 得到其隐式表征进而预测扰动谱的隐式表征；基于二者的隐式表征, 再使用逻辑回归分类器预测药物响应，从而构建半监督的药物响应和扰动谱的联合模型.

除了上述两个挑战之外，药物响应预测模型开发中的另一个重要问题是所使用的组学数据类型. 一般而言, 在传统机器学习模型中, 使用的组学数据主要包括基因突变、拷贝数变异和基因表达谱等[10,14,25]. 在NCI-DREAM挑战赛和随后的其他研究中，人们发现, 所有细胞系中药物响应水平的大部分差异可以通过全基因组基因表达数据来解释，而其他组学概况只能稍微改善预测[10,25]. 可以说, 药物响应预测模型的开发关键取决于输入数据类型与维数，通过合理地组织输入数据并建立系统性的多组学模型，提升数据的信噪比, 是提升预测性能的关键. Aben等人[37]开发了一种称为TANDEM的两阶段方法，第一阶段使用上游特征(基因突变、拷贝数变异、DNA甲基化和癌症类型)通过弹性网络回归拟合药物响应，第二阶段使用下游特征即基因表达数据来拟合弹性网模型以预测第一阶段的残差. 事实上, 针对异质性高的基因特征数据, 使用传统的机器学习方法预测药物响应存在一定局限性，一般只能通过将模型的流程复杂化来换取效能的提升. 与此不同，基于深度神经网络的表示学习方法可以将高维的输入数据直接嵌入到网络，更适合处理高异质性的基因特征数据. 这种引入多模态概念的深度学习方法可以训练端对端(end-to-end)的预测模型，是药物响应预测算法的发展趋势.

3 基于网络的药物响应预测

近年来，将药物响应转化为网络中的链接预测问题或通过细胞和药物的相似性对网络进行表示学习是药物响应预测的主流方式. 一方面, 将高维度的组学特征直接嵌入到细胞系的相似性网络，或在基因/蛋白质网络进行表示学习可以克服药物响应预测的特征-样本比过大的问题[38]. 另一方面, 将未经处理的组学数据直接作为预测模型的输入会忽略基因组特征之间的多元关联关系，引入基因调控网络(gene regulatory networks, GRN)或蛋白-蛋白相互作用(protein-protein interaction, PPI)网络可以为整合基因组特征提供功能背景, 从而提高药物响应预测模型的性能.

Turki和Wei[39]引入GRN构建出药物响应网络链接预测算法, 提出: (1) 监督链接预测, 使用QBC(query by committee)算法选择具有高质量数据的癌细胞系; (2) 扩展的监督链接预测, 使用CUR矩阵分解选择癌细胞系和重要的基因子集. 链接预测采用支持向量回归(support vector regression)和岭回归. 同样基于链接预测, Stanfield等人[40]将细胞系的基因突变和药物反应与PPI网络整合到一起, 创建了一个由基因、细胞系和药物组成的大型异质网络, 将药物响应预测转换为药物与细胞系在网络中链接的预测问题, 并通过随机重游走(random walk with restart, RWR)计算药物和细胞系的特征向量, 以预测缺失的链接(药物响应). 该文提出使用PPI网络整合细胞系突变数据可以显著提高药物响应预测性能.

基于相似性网络的预测方法也得到广泛的应用, 这类方法可以容易地通过药物相似性网络引入药物的分子结构信息作为特征. Zhang等人[41]提出了双层集成细胞系-药物网络模型, 使用基于基因表达谱的细胞系相似性网络(cell line similarity network, CSN)数据和基于PaDEL[42]计算的1-D和2-D药物结构的药物相似性网络(drug similarity network, DSN)数据来使用线性加权模型预测给定细胞系的药物反应, 对于一对待查询的细胞-药物组合, 其响应是由细胞系相似性网络中与查询细胞系相邻的细胞系的响应, 以及药物相似性网络中与查询药物相邻的药物响应所决定, 其中权重是相似网络中的邻居与查询药物和查询细胞的相似度. 作者使用该模型对MAPK(mitogen-activated protein kinase)和ERK(extracellular regulated protein kinases)信号通路的两个案例研究进行验证, 发现其表现优于作为基线的弹性网络模型. 作者指出, 仅基于DSN的预测比仅基于CSN的预测在多数药物上取得更好表现, 这强调了药物结构信息在预测中的重要性. Le和Pham[38]提出, 在基因表达谱的基础上, 应将遗传变异数据作为细胞系之间的相似性进行嵌入, 由此发展出了GloNetDRP方法. GloNetDRP方法使用细胞系基因表达谱和遗传突变数据共同构建细胞系相似性网络和基于药物分子图的药物相似性网络, 然后通过药物和细胞系之间的已知药物响应将这两个网络连接起来, 形成了药物和细胞系的异质网络. 随后, 作者采用基于随机重游走的全局算法计算未知的药物响应, 发现GloNetDRP的全局算法优于仅考虑局部信息(相似性网络中的邻居)的双层网络方法[41], 且表现超过基于机器学习的代表性方法KBMTL[22]. 此外, 作者认为, 由于以前的研究中直接使用布尔值代表基因突变信息, 这种编码方式可能导致遗传变异数据对药物响应预测的贡献被低估. 在GloNetDRP方法中, 遗传变异信息可以在蛋白-蛋白网络上进行传播, 因此其提供的信息更加丰富, 利用基因变异信息的模型达到了与利用基因表达谱模型相当的预测性能.

除了使用局部的线性加权或全局的网络传播算法直接在网络上计算药物响应, 也可以通过网络表示学习获得新的表征向量作为机器学习模型的输入来预测药物响应. Yang等人[43]提出了整合细胞系遗传异常、药物响应和PPI的NRL2DRP(network representation learning to drug response prediction). NRL2DRP是一种基于邻域相似性的网络表示学习方法, 包含细胞系功能分子谱网络和药物-细胞系响应网络两部分, 前者是通过整合包括突变基因、异常拷贝数片段和启动子甲基化CpG位点的癌症功能事件(cancer functional events, CFEs)[20]并映射到PPI网络而建立的. 模型构建原理是基于具有共享相似功能环境(即相邻网络结构)的细胞系可能具有相似的药物响应这一假设. 为了捕获这种相似性信息, 作者选择基于邻域网络结构的邻近度量进行网络表示学习, 得到的细胞系的表征向量作为特征输入到药物的分类SVM模型中进行药物响应预测, 其表现超过了KBMTL[22]和链接预测算法[40].

除了多组学数据和药物分子结构信息, 基于网络的方法也可以更加方便地引入药物靶标信息. Zhang等人[44]提出HNMDRP(heterogeneous network-based method for drug response prediction)方法, 首先通过计算细胞系基因组图谱、药物化学结构和药物靶标基因之间的皮尔森相关系数来引入相似性度量, 然后构建基于细胞系基因表达谱的细胞系相似性网络、基于化学结构相似性的药物相似性网络、基于PPI的靶标相似性网络, 再通过药物-细胞系响应和药物-靶标相互作用网络将上述相似性网络相连, 最后全部整合到异质网络中. 在该网络上执行基于信息流[45]的算法, 得到所有细胞系-药物对的分数, 即为对药物响应的预测. 作者证实, 已知的药物-靶标相互作用和PPI的引入有助于改善药物响应的预测性能, 而基于1-D和2-D结构的药物分子结构信息的贡献较小. 总体而言, 该模型表现超越了链接预测[40]和相似性网络预测方法[41].

4 基于多模态深度学习的药物响应预测

在使用多组学数据的药物响应预测研究中, 一个

关键的问题是如何集成不同的组学数据. 事实上, 这个问题在人工智能领域普遍存在, 通常采取多模态(multimodal)学习解决. 一般来说, 每一种信息的来源或者形式都可以称为一种模态(modality), 如对视频文件中的图像、文本、语音三种形式的信息进行多模态的处理. 通常认为, 由于不同模态的表现方式和角度不同, 多模态具有冗余性和互补性, 通过合理地处理多模态信息, 使模型达到更好的性能[46,47]. 如图1所示, 多组学分析中, 可以将不同组学的数据视作不同来源的信息, 作为多模态进行融合. 在深度神经网络中, 多模态融合一般采取3种方式: (1) 输入级特征融合(input-level feature fusion), 即早期融合(early fusion), (2) 中

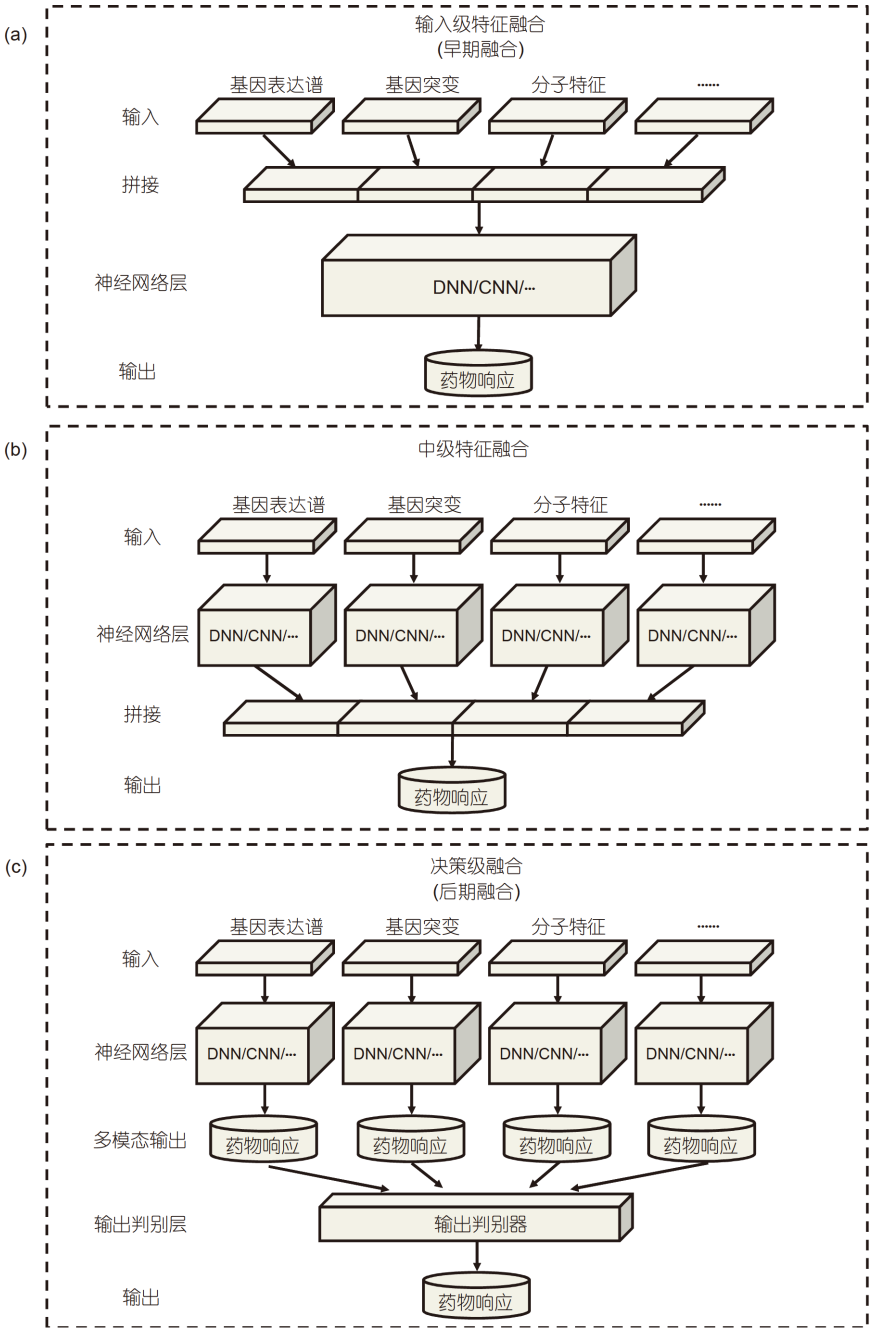


图 1 使用多组学数据预测药物响应的三种多模态特征融合方式. (a) 输入级特征融合(早期融合); (b) 中级特征融合; (c) 决策级融合(后期融合)
Figure 1 Three multimodal feature fusion methods for drug response prediction using multi-omics data. (a) Input-level feature fusion (early fusion); (b) intermediate features fusion; (c) decision-level fusion (late fusion)

级特征融合(intermediate features fusion)和(3) 决策级融合(decision-level fusion), 即后期融合(late fusion). 输入级特征融合使用广泛, 通常是将样本所有类型的特征向量非加权串联, 然后使用一些特征降维或学习方法, 如PCA(principal components analysis)、自动编码器等来创建该样本的集成表示. 中级特征融合采用单模态的网络分别学习每个模态, 将所学习到的权重串联在一起, 再通过全连接层继续训练串联后的特征. 决策级融合通常应用单独的判别器来加权单模态的决策, 即对每个模态单独训练判别器, 其输出再进行加权, 权重可以通过另一个判别器来学习, 整个过程中不执行串联.

通常认为, 早期融合会忽略每种特征类型各自的分布情况, 可能会赋予具有更高维度的模态以更大的权重, 并且进一步增加输入数据的维数. 因此, 使用早期融合方法直接融合多组学数据的研究并不常见, 但可以使用此方法融合基因表达谱和药物的分子特征, 从而对不同细胞系和药物共同建模. Wang等人[48]将细胞系中方差高的前1000基因的基因表达谱与药物分子指纹直接串联输入到神经网络以预测药物响应. 在后续研究中, 他们又对算法进行了改进[49], 先采用堆叠自编码器对基因表达谱降维, 再将降维后的特征与药物的分子指纹进行串联, 输入到深度神经网络中训练以预测药物响应.

近年来, 相比于早期融合, 使用中级特征融合方式的多模态学习在药物响应预测研究中占主导地位. 通过分别学习每种模态的特征后再进行集成, 可以获得多模态数据统一的表征作为分类器或回归器的输入, 这种方式已逐渐成为多组学学习的常见方法. Sharifi-Noghabi等人[15]提出了MOLI(multi-omics late integration), 一种基于深度神经网络的多组学中级特征融合方法. MOLI由3个前馈编码子网络组成, 分别以体细胞突变、拷贝数变异和基因表达谱作为输入, 编码子网络中学习到的特征通过串联拼接, 作为分类子网络的输入用于预测药物响应. 整个网络以端到端的方式进行训练, 在交叉熵损失函数的基础上通过引入三重损失函数(triplet loss)[50]构成组合损失函数来优化模型. 一个MOLI模型并非针对每个药物单独建模, 而是引入迁移学习的思路, 将针对同一通路或靶标的分子作为一组任务, 进行多任务的训练. PDX Encyclopedia外部验证集的结果显示, 靶向药物的迁移学习比单个药物作为输入的方式具有更好预测性能. 分析发现, MOLI预测的EGFR(epidermal growth factor receptor)抑制剂对TCGA数据集中患者的药物反应与EGFR通路中某些基因有统计学意义上的相关性, 证实了体外数据训练的MOLI模型可以在一定程度上反映体内药效, 具有临床预测的潜力. Chiu等人[51]在多模态的基础上引入自编码器对网络进行预训练, 创新性地将TCGA数据应用到训练过程. 他们建立了一种基于癌细胞的突变和表达谱预测药物反应的深度学习模型DeepDR, 该模型由3个深度网络拼接而成, 即突变编码器(mutation encoder, Menc)、表达编码器(expression encoder, Eenc)及预测网络(prediction, P). Menc和Eenc为预训练的自动编码器(autoencoder)的编码器部分, 分别采用TCGA的突变数据和基因表达数据进行预训练. P为整合Menc和Eenc的多任务药物响应预测器. 完整的模型采用CCLE数据集的突变和表达谱作为输入, 以GDSC数据集256个药物的相应数据作为输出进行训练和测试. Menc和Eenc的参数用从TCGA预训练中获得的参数进行初始化, 并在训练过程中更新, P被随机初始化. 预测时只需提供细胞系的突变和表达特征, 该模型即可预测265种药物对于此细胞系的IC_{50}值. 应用该模型预测9059个TCGA样本对265种抗癌药物的反应, 预测结果显示了模型的实用价值. 在泛癌(pan-caner)分析中, 作者检查了11个在所有TCGA样品中变异率高于10%的基因, 发现TTN(titin)可能是对vinorelbine和epothilone等药物敏感的基因标志物. 作者以docetaxel和CX-5461为例, 对与药物耐药性相关的突变和表达谱进行了研究, 证实了docetaxel的耐药机制, 并阐明了新型抗癌药物CX-5461治疗神经胶质瘤(low-grade glioma, LGG)和造血系统恶性肿瘤中的抗癌潜力.

除了融合多组学数据, 中级特征融合方法同样可以将药物的分子结构作为一种模态引入模型. Chang等人[52]提出的CDRscan(cancer drug response profile scan)使用细胞系的基因突变指纹和药物分子指纹作为多模态CNN网络(convolutional neural networks)的输入. 对基因突变, 作者在CGC(cancer gene census)[53,54]的567个基因中选择了28328个位置的序列变异信息进行表征, 对药物则采取3072位的PaDEL分子描符. 为了处理这样的高维度特征, 作者分别使用卷积层来对二者进行降维, 再将降维后的特征进行串联后继续进入卷积网络, 进行端对端的训练. 进一步地, 作者训练了5个不同架构的卷积神经网络, 将预测结果进行集成以提高精度, 最终模型表现超过了基线模型(随机森林和支持向

量机). Manica等人[55]在引入分子结构的基础上进一步增强了模型的可解释性, 建立了基于注意力的多模态卷积编码器模型, 将细胞系基因表达谱、药物分子结构和PPI等信息进行整合来预测药物响应. 对每种药物而言, 作者将与药物靶标相关的权重映射到PPI网络上, 继而进行网络传播, 使用传播后的权重分布确定每种药物的高排名基因, 组合成新的基因集. 细胞系在该集合上的基因表达值即为细胞系特征. 药物分子结构特征则采用SMILES(simplified molecular input line entry specification)进行表征, 作者比较了多种可编码SMILES的神经网络结构, 包括双向递归神经网络(bi-directional recurrent neural network, biRNN)、堆叠卷积编码器(stacked convolutional encoder, SCNN)、自注意力机制(self-attention, SA)、上下文注意力机制(contextual-attention, CA)和多尺度卷积注意力(multiscale convolutional attentive, MCA)编码器等. 细胞系特征和分子结构特征分别进行经过各自的编码层得到编码后的特征, 这些特征串联后进入全连接层, 进行端对端的药物响应预测. 基于多尺度卷积和注意力机制的编码器优于基于摩根分子指纹[56]训练的基线模型和其他SMILES编码器. 此外, 作者以两种作用于白血病细胞系的受体酪氨酸激酶(receptor tyrosine kinase, RTK)抑制剂为案例进行了研究, 通过对注意力权重的分析展示了该模型的可解释性.

基于后期融合的多模态药物响应预测目前尚不多见. Choi等人[57]基于近似的思想构建了RefDNN模型, 他们没有直接将药物的分子结构和基因表达谱输入模型, 而是建立了由多个弹性网络分类器和一个DNN(deep neural networks)分类器组成的RefDNN(reference drug based neural network)模型. M个弹性网络分类器用待查询细胞系的基因表达谱对查询细胞系M种参考药物的响应进行预测, 用以表征基因表达谱; 而待查询药物的分子特征表征为其对M种参考药物的相似性. 得到的两种特征通过哈达玛积(Hadamard product)合并, 输入到DNN中进行端对端的训练来再次预测药物响应. 弹性网络和DNN的权重由这两种网络损失之和的梯度来进行更新. RefDNN利用弹性网络正则化处理高维基因表达数据, 可以鉴定与耐药性相关的基因标记.

5 总结与展望

开发基于癌症患者的基因组信息预测有效治疗策略的计算模型是精准医学中的关键挑战. 因此, 世界多个组织机构公开了大量针对各种癌细胞系和各种抗癌药物的剂量响应数据, 并围绕大规模基因组特征数据开展了许多针对药物细胞敏感性预测的研究. 2012年的NCI-DREAM药物敏感性预测挑战赛可被视为药物敏感性研究的里程碑, 其中涌现出大量机器学习算法的创新性应用, 这也为后续的研究奠定了基础. 然而, 针对高维度、高异质性的基因特征数据, 使用传统的机器学习方法预测药物响应仍存在一定局限, 基于网络的预测方法和多模态深度学习方法有利于实现多组学数据的系统性的整合和应用, 从而获得更好的药物响应预测效果.

事实上, 癌症具有异质性, 即具有不同分子特征的肿瘤亚细胞在患有同一组织类型肿瘤的人群或个体的肿瘤组织之内/之间的分布差异. 癌症的异质性是药物抵抗的重要原因, 这意味着采用肿瘤细胞系分子特征建立的药物敏感性预测方法在临床应用上有很大的局限性. 因此, 相比于预测特定肿瘤细胞系的药物响应, 研究人员更希望通过获得药物响应的预测因子来揭示药物敏感性的生物标记物, 从而指导临床诊断方法并发现新型抗癌药物. 这正符合对人工智能算法的可解释性的需求. 比如, 利用弹性网络模型可以识别具有高关联性特征的子集, 在深度学习模型引入注意力机制可以根据注意力权重来排序重要特征等. 这些人工智能方法为研究复杂生物问题提供了崭新的思路, 但是否真正具有实际应用价值仍不够明确. 总体而言, 可解释性仍是整个人工智能领域目前面临的困境. 随着可解释性人工智能研究的不断发展, 未来的药物敏感性研究会更加聚焦于生物标志物的发现, 这将有助于剖析癌症治疗药物的分子机制, 发现可以作为药物靶标的致癌驱动因素(oncogenic drivers), 并将其转化为精准医疗所需的个性化诊疗策略.

参考文献

1 Dagogo-Jack I, Shaw A T. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol, 2018, 15: 81–94
2 Azuaje F. Computational models for predicting drug responses in cancer research. Brief Bioinform, 2016, 18: 820–829
3 Cortes-Ciriano I, Mervin L, Bender A. Current trends in drug sensitivity prediction. Curr Pharm Des, 2016, 22: 6918–6927

4 Shoemaker R H. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer, 2006, 6: 813–823
5 Barretina J, Caponigro G, Stransky N, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 2012, 483: 603–607
6 Garnett M J, Edelman E J, Heidorn S J, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature, 2012, 483: 570–575
7 Yang W, Soares J, Greninger P, et al. Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res, 2012, 41: D955–D961
8 Basu A, Bodycombe N E, Cheah J H, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell, 2013, 154: 1151–1161
9 Seashore-Ludlow B, Rees M G, Cheah J H, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. Cancer Discov, 2015, 5: 1210–1223
10 Costello J C, Heiser L M, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol, 2014, 32: 1202–1212
11 Menden M P, Wang D, Mason M J, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. Nat Commun, 2019, 10: 2674
12 Chang K, Creighton C J, Davis C, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet, 2013, 45: 1113–1120
13 Li J, Lu Y, Akbani R, et al. TCPA: A resource for cancer functional proteomics data. Nat Methods, 2013, 10: 1046–1047
14 Geeleher P, Zhang Z, Wang F, et al. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. Genome Res, 2017, 27: 1743–1751
15 Sharifi-Noghabi H, Zolotareva O, Collins C C, et al. MOLI: Multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics, 2019, 35: i501–i509
16 Subramanian A, Narayan R, Corsello S M, et al. A next generation connectivity MAP: L1000 platform and the first 1000000 profiles. Cell, 2017, 171: 1437–1452.e17
17 Tan M, Özgül O F, Bardak B, et al. Drug response prediction by ensemble learning and drug-induced gene expression signatures. Genomics, 2019, 111: 1078–1088
18 Rampášek L, Hidru D, Smirnov P, et al. Dr.VAE: Improving drug response prediction via modeling of drug perturbation effects. Bioinformatics, 2019, 35: 3743–3751
19 Vougas K, Sakellaropoulos T, Kotsinas A, et al. Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel *in silico* screening process based on association rule mining. Pharmacol Therapeutics, 2019, 203: 107395
20 Iorio F, Knijnenburg T A, Vis D J, et al. A landscape of pharmacogenomic interactions in cancer. Cell, 2016, 166: 740–754
21 Gonen M. Bayesian efficient multiple kernel learning. ICML, 2012, arXiv:1206.6465
22 Gönen M, Margolin A A. Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning. Bioinformatics, 2014, 30: i556–i563
23 Daemen A, Griffith O L, Heiser L M, et al. Modeling precision treatment of breast cancer. Genome Biol, 2013, 14: R110
24 Hejase H A, Chan C. Improving drug sensitivity prediction using different types of data. CPT Pharmacometrics Syst Pharmacol, 2015, 4: 98–105
25 Jang I S, Neto E C, Guinney J, et al. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Pac Symp Biocomput, 2015, 2014: 63–74
26 Wan Q, Pal R. An ensemble based top performing approach for NCI-DREAM drug sensitivity prediction challenge. PLoS One, 2014, 9: e101183
27 Su R, Liu X, Wei L, et al. Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. Methods, 2019, 166: 91–102
28 Matlock K, De Niz C, Rahman R, et al. Investigation of model stacking for drug sensitivity prediction. BMC BioInf, 2018, 19: 71
29 Haider S, Rahman R, Ghosh S, et al. A copula based approach for design of multivariate random forests for drug sensitivity prediction. PLoS One, 2015, 10: e0144490
30 Le D H, Nguyen-Ngoc D. Multi-task regression learning for prediction of response against a panel of anti-cancer drugs in personalized medicine. In: Proceedings of the 2018 1st International Conference on Multimedia Analysis and Pattern Recognition, Ho Chi Minh City, 2018. IEEE: Piscataway, N.J
31 Tan M. Prediction of anti-cancer drug response by kernelized multi-task learning. Artif Intell Med, 2016, 73: 70–77
32 Yuan H, Paskov I, Paskov H, et al. Multitask learning improves prediction of cancer drug sensitivity. Sci Rep, 2016, 6: 31619
33 Suphavilai C, Bertrand D, Nagarajan N. Predicting cancer drug response using a recommender system. Bioinformatics, 2018, 34: 3907–3914
34 Papillon-Cavanagh S, De Jay N, Hachem N, et al. Comparison and validation of genomic predictors for anticancer drug sensitivity. J Am Med Inform Assoc, 2013, 20: 597–602
35 Parca L, Pepe G, Pietrosanto M, et al. Modeling cancer drug response through drug-specific informative genes. Sci Rep, 2019, 9: 15222

36 Ding M Q, Chen L, Cooper G F, et al. Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. Mol Cancer Res, 2018, 16: 269–278

37 Aben N, Vis D J, Michaut M, et al. TANDEM: A two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. Bioinformatics, 2016, 32: i413–i420

38 Le D H, Pham V H. Drug response prediction by globally capturing drug and cell line information in a heterogeneous network. J Mol Biol, 2018, 430: 2993–3004

39 Turki T, Wei Z. A link prediction approach to cancer drug sensitivity prediction. BMC Syst Biol, 2017, 11: 94

40 Stanfield Z, Coşkun M, Koyutürk M. Drug response prediction as a link prediction problem. Sci Rep, 2017, 7: 40321

41 Zhang N, Wang H, Fang Y, et al. Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. PLoS Comput Biol, 2015, 11: e1004498

42 Yap C W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem, 2011, 32: 1466–1474

43 Yang J, Li A, Li Y, et al. A novel approach for drug response prediction in cancer cell lines via network representation learning. Bioinformatics, 2018, 35: 1527–1535

44 Zhang F, Wang M, Xi J, et al. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. Sci Rep, 2018, 8: 3355

45 Wang W, Yang S, Zhang X, et al. Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics, 2014, 30: 2923–2930

46 Ramachandram D, Taylor G W. Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Process Mag, 2017, 34: 96–108

47 Baltrusaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy. IEEE Trans Pattern Anal Mach Intell, 2019, 41: 423–443

48 Wang Y, Li M, Zheng R, et al. Using deep neural network to predict drug sensitivity of cancer cell lines. In: Proceedings of the Intelligent Computing Theories and Application, Cham, 2018. Springer International Publishing, Cham

49 Li M, Wang Y, Zheng R, et al. DeepDSC: A deep learning method to predict drug sensitivity of cancer cell lines. IEEE/ACM Trans Comput Biol Bioinform, 2019, doi: 10.1109/TCBB.2019.2919581

50 Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7-12 June 2015, 2015. IEEE: Piscataway, N.J

51 Chiu Y C, Chen H I H, Zhang T, et al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Med Genomics, 2019, 12: 18

52 Chang Y, Park H, Yang H J, et al. Cancer drug response profile scan (CDRscan): A deep learning model that predicts drug effectiveness from cancer genomic signature. Sci Rep, 2018, 8: 8857

53 Futreal P A, Coin L, Marshall M, et al. A census of human cancer genes. Nat Rev Cancer, 2004, 4: 177–183

54 Forbes S A, Beare D, Boutselakis H, et al. COSMIC: Somatic cancer genetics at high-resolution. Nucleic Acids Res, 2016, 45: D777–D783

55 Manica M, Oskooei A, Born J, et al. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. Mol Pharm, 2019, 16: 4797–4806

56 Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model, 2010, 50: 742–754

57 Choi J, Park S, Ahn J. RefDNN: A reference drug based neural network for more accurate prediction of anticancer drug resistance. Sci Rep, 2020, 10: 1861

Summary for “人工智能算法在药物细胞敏感性预测中的应用”

The application of artificial intelligence to drug sensitivity prediction

Xutong Li[1,2], Xiaolong Wu[1,3], Xiaozhe Wan[1,2], Feisheng Zhong[1,2], Chen Cui[1,2], Yingjia Chen[1,2], Lifan Chen[1,2], Kaixian Chen[1,2], Hualiang Jiang[1,2*] & Mingyue Zheng[1,2*]

[1] *Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China;*
[2] *School of Pharmacy, University of Chinese Academy of Sciences, Beijing 100049, China;*
[3] *School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China*
* Corresponding authors, E-mail: hljiang@simm.ac.cn; myzheng@simm.ac.cn

The development of computational methods for the prediction of effective therapeutic strategies based on the genomic information of patients is the main challenge of precision medicine. Since the 21st century, next-generation sequencing (NGS) has opened up new possibilities for personalized medicine. Extensive characterization at the molecular level for hundreds of cancer cell lines has been brought to the public eye by many organizations and agencies around the world. For example, the National Cancer Institute 60 Human Cancer Cell Line Screen (NCI-60), Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) have provided large-scale omics data such as genomic, transcriptomic and epigenomic data characterizing cancer cell lines, and The Cancer Genome Atlas (TCGA) has molecularly characterized over 20000 primary cancers of patients. Combined with the drug response data of cancer cell lines, multiomics data could be used to analyse the mechanisms of action of anticancer drugs, which could be incorporated into precision medicine strategies.

Over several decades, artificial intelligence (AI) technologies based on big data have revolutionized bioinformatics. AI has built a bridge between genomics and drug sensitivity by promoting the development of predictive models for the drug response of cancer cell lines. The 2012 NCI-DREAM drug prediction challenge has been particularly influential, as the innovative applications of machine learning that emerged from it have laid the groundwork for future studies. However, classic machine learning models are still challenging in terms of predictability because they limit the systematic integration of high-dimensional multiomics data. Therefore, network-based approaches, including link prediction and network representation, have become mainstream methods for drug response prediction. On the one hand, network-based approaches have not faced the “small n, large p” problem since the multiomics features are either represented in a gene/protein network or embedded in similarity networks between cell lines. On the other hand, the introduction of gene regulatory networks (GRNs) and protein-protein interactions (PPIs) into the predictive model can provide a functional background for the integration of genomic data and thereby improve the predictive performance of drug response. In addition to network-based approaches, multimodal deep learning models can systematically integrate multiomic data by considering them as different modalities. Generally, there are three feature fusion methods in deep neural networks: Input-level feature fusion (early fusion), intermediate feature fusion and decision-level fusion (late fusion). Intermediate feature fusion is predominant in drug response prediction studies, by which features are learned separately for each type of omics data and then integrated into one unified representation to be used as the input for a classifier or a regressor. Moreover, the features of drug structures can be used as a model to improve the performance.

In brief, we summarize the characteristics of publicly accessible genomic databases and discuss the trends of artificial intelligence applications in drug sensitivity prediction for cancer cell lines, including machine learning, networks and multimodal deep neural networks.

drug sensitivity, machine learning, network, multimodal

doi: 10.1360/TB-2020-0557