科学通报 2010 年 第 55 卷 第 26 期：2602 ~ 2610
论 文 www.scichina.com csb.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

基于基因表达谱与系统发育树的肿瘤细胞多药耐药性表型预测

刘鑫奕[①*], 李作峰[②*], 文静然[①], 蔡青青[①], 徐烨[③], 张晓艳[①†]

① 同济大学生命科学与技术学院生物信息学系, 上海 200092;
② 上海生物信息技术研究中心, 上海 200235;
③ 复旦大学附属肿瘤医院大肠外科, 上海 200032
* 同等贡献
† 联系人, E-mail: xyzhang@tongji.edu.cn

2010-01-29 收稿, 2010-04-13 接受
国家高技术研究发展计划(2007AA02Z332, 2008AA02Z126, 2009AA02Z308)和上海市重大项目(07DZ19505)资助

摘要 应用基因芯片技术预测抗肿瘤药物多药耐药性(multiple drug resistance, MDR)表型时, 探针表达值归一化策略和特征基因集的选取往往是导致实验间结果不一致性的重要原因. 从基因芯片数据出发, 如何建立一个统计学上稳定的预测模型, 已成为 MDR 表型预测建模研究中迫切需要解决的问题. 本研究以多药耐药肿瘤细胞系的基因表达数据为研究对象, 将探针表达定性为有无表达(1/0)两种状态, 再将其归类到由蛋白质结构域组序(protein domain organization, PDO)定义的基因集中. 在此基础上, 通过引入系统发育学中的基因含量方法(gene content), 在 PDO 基因集水平上建立了系统发育学模型(细胞树), 并用于 MDR 表型的预测. 结果显示, 肿瘤细胞系的分类主要受细胞病理分型和 MDR 表型(紫杉醇和长春碱)的影响. 系统发育学模型在预测样本的 MDR 表型方面优于特征基因模型. 尽管本文方法的应用受到样本混杂度的限制, 但其对于血液系统肿瘤或纯度较高的细胞系仍具有潜在的应用价值.

关键词
基因芯片
系统发育树
肿瘤
多药耐药性
样本聚类分析
细胞树

肿瘤细胞的多药耐药性(multiple drug resistance, MDR)是导致化疗方案失败的重要原因. 进入后基因组时代, 对常见肿瘤的研究积累了大量基因芯片(microarray)和 MDR 实验数据, 使抗癌药物在基因组水平的耐药性预测成为可能[1~5]. 目前, 由单个训练集挑选用于耐药预测的特征基因是一种有监督机器学习方法, 已成为应用最广泛的预测方案. 与单基因分析相比, 基因集富集分析(gene set enrichment analysis, GSEA)更具有生物学意义[6]. 然而, 这些方法依赖于探针表达值归一化策略和特征基因集的选取, 限制了基因芯片与临床应用的接轨[7,8].

样本聚类分析是基因芯片分析领域重要的无监督学习方法, 被广泛用于寻找肿瘤亚型和病人亚群, 以辅助发现可用于预测新样本的特征基因[9]. 此外, 一些与系统发育树构建的相关方法被借鉴到基因芯片分析中来, 例如用自展法(bootstrap)评估层次聚类树拓扑结构的稳定性[10,11].

本研究提出了一种基于系统发育学基因含量方法(gene content)的无监督基因芯片聚类算法, 利用以蛋白质结构域组序(protein domain organization, PDO)定义的基因集处理基因芯片数据, 通过邻接法(neighbor-joining, NJ)和自展法构建细胞树. 方法被成功用于肿瘤建系细胞样本聚类和 MDR 表型预测.

1 材料与方法

(ⅰ) 基因芯片数据. 从 GEO 数据库[12]下载

英文版见: Liu X Y, Li Z F, Wen J R, et al. Prediction of multiple drug resistance phenotype in cancer cell lines using gene expression profiles and phylogenetic trees. Chinese Sci Bull, 2010, 55, doi: 10.1007/s11434-010-4131-2

Affymetrix 芯片的 CEL 文件. 表 1 列出了本文所使用的 4 套 GEO 数据集[13~16], 均为人类单通道 Affymetrix HG-U133A 平台的基因表达数据.

(ⅱ) 利用蛋白质结构域组序建立基因集. 蛋白质结构域组序(PDO)定义为一个蛋白质上二级结构域的组成及排序[17]. 利用 PDO 建立基因集的方法如下: 首先, 根据平台注释中的 GenBank[18]登录号和基因 ID, 从 NCBI FTP 服务器获取对应的蛋白质 GI 号和氨基酸序列; 其次, 通过 GI 号从 Pfam FTP 服务器获取蛋白质的二级结构域并根据排序得到 PDO, 对于尚未被 Pfam 预测过的蛋白质(约 50%), 将其氨基酸序列作为本地 Pfam 程序的输入进行预测[19]; 再次, 若一个探针对应多个 GI, 则选择出现频率最高且结构域数量最少的 PDO; 最后, 具有相同 PDO 的探针将合并为一个基因集.

(ⅲ) 建立样本表达概型. 采用 MAS 5.0 算法中的检测 Call 值判断基因的表达水平, 即存在/边际存在/不存在(P/M/A)[20]. Call 值为“P”记为“1”, 表示基因表达; Call 值为“M”或“A”记为“0”, 表示基因未表达, 这样就建立了有关基因表达的(0, 1)矩阵. 引申到 PDO 的表达概型, 若一个 PDO 单元中有至少一个探针的 Call 值为“P”, 则认为 PDO 表达, 记为“1”, 反之记为“0”, 最终得到 PDO 的(0, 1)矩阵.

(ⅳ) 自展法评估与样本聚类树构建. PDO 的(0, 1)矩阵可以看作离散形态学数据, 利用 PHYLIP 软件包[21]中的 seqboot 对其进行重复随机抽样, 生成 1000 个数据集. 根据 Fukami-Kobayashi 等人[17]提出的 PDO 集距离计算公式, 将数据集计算成 1000 个距离矩阵, 再通过 PHYLIP 提供的邻接法 neighbor 和一致性算法 consense 程序构建具有自展值的样本聚类树. 在本文中, 利用已建立的人类细胞系所构建的样本聚类树被称为细胞树.

(ⅴ) MDR 表型预测模型. 将 GSE11812 的表达数据作为建立 MDR 表型预测模型的训练集. 为了进行方法比较, 我们同时用 GSE11812 建立了基于特征基因的参照模型.

(a) 系统发育树模型. (1) 利用训练集的 PDO (0, 1)矩阵对样本进行聚类, 生成细胞树. 在不考虑细胞表型的情况下, 根据树型拓扑对样本进行分组. (2) 对每组样本的 MDR 表型进行评价. 策略如下: 统计每组样本对每种药物的耐药率并计算平均值, 然后对每两组数据进行配对方差检验. 此外, 统计每个样本的耐药数, 并计算组内均值. (3) 对于每组样本, 如果平均耐药率超过 50%, 并且统计结果表明其具有显著的 MDR 趋势, 那么这一组被定义为耐药组(R), 否则定义为敏感组(S).

(b) 参照模型. (1) 根据样本对某种药物的耐药表型, 将训练集分为敏感组(S)和耐药组(R), 排除中间样本(M). (2) 使用 ArrayTools[22]中的 MAS 5.0 算法计算探针表达值并进行归一化, 根据分组信息, 采用 SAM[23]工具挑选差异表达基因集. (3) 利用 WEKA 软件[24]中整合的 Logistic 回归算法对特征变量构建模型, 取阈值为 1×10^{-5}. (4) 对阿霉素、紫杉醇和长春碱重复上述工作, 分别构建 3 个独立的模型.

(ⅵ) NCI-60 细胞系的 MDR 表型预测. 利用第(ⅴ)节中的两个模型预测 NCI-60 细胞系(GSE5720)的 MDR 表型.

(a) 系统发育树模型. (1) 得到 NCI-60 的 PDO 矩阵. (2) 为了尽可能保证模型的树型拓扑结构不受测试样本影响, 每次用一个测试样本与训练集的 30 个样本构树. (3) 若测试样本与训练集中的 R 组聚在一起, 认为其 MDR 表型为耐药, 若与 S 组聚在一起, 则 MDR 表型为敏感.

(b) 参照模型. (1) 使用 ArrayTools 中的 MAS 5.0 算法计算 NCI-60 样本的探针表达值并进行归一化. (2) 将特征基因的表达值作为模型的输入, 预测对相应药物的耐药性(阿霉素/紫杉醇/长春碱).

为了比较两种模型的预测结果, 我们根据预测结果将 NCI-60 分为耐药组和敏感组. 从“开发治疗项目”(DTP)的网站[25]获取 NCI-60 对多种药物的 50%生长抑制浓度指标(GI50). 对于两组样本对每种药物的 GI50 值, 一方面用 Wilcoxon 检验计算其排序

表 1 4 套 GEO 基因芯片数据集及其用途

GEO 系列 ID	用途	描述
GSE11812	训练集	30 个癌症细胞系对 11 种抗癌药物的耐药性分析[13]
GSE5720	测试集	美国国立癌症研究所(NCI)提供的 60 种癌症细胞系(NCI-60)[14]
GSE1133	测试集	选取 HL-60 细胞系[15]
GSE11466	测试集	选取 MDA-MB-435 细胞系[16]

是否具有显著差异, 另一方面用 t 检验分析其均值是否具有显著差异. 两种检验结果均可以展示耐药组 GI50 的整体分布是显著高于(P<0.05)或是低于敏感组(P>0.95).

(ⅶ) 获取差异表达 PDO 基因集. 将训练集的 PDO 矩阵作为 SAM 样本的输入. 阈值选取标准为: 保证用筛选出的 PDO 能构建出对以上 3 种药物耐药性递增的邻接树. 使用富集分析工具 DAVID[26]提取差异表达 PDO 基因集所承担的主要生物学功能, 进而分析与 MDR 表型的关系.

2 结果与讨论

2.1 PDO 基因集的建立

Affymetrix 的 HG-U133A 平台包含 22283 个探针, 有 20518 个探针可以获得蛋白质 GI 号, 其中 18371 个能映射到 PDO 信息, 未能映射到 PDO 的探针均为缺少结构域信息的基因. 探针的利用率为 82.4%. 最终建立了 5459 个无重复 PDO 基因集.

本研究之所以使用 Affymetrix HG-U133A 平台的基因表达数据, 是因为该平台所特有的寡核苷酸探针设计和质量控制算法为样本表达概型的建立创造了条件. MAS 5.0[20]是最能胜任此项工作的算法.

为了将探针表达与生物学功能相联系, 我们在分析过程中将 PDO 作为一种序列信息进行整合. 尽管基因芯片的平台注释提供了获取序列信息的入口, 但是将其引入研究的报道并不多[27,28]. 最近, PDO 被成功用于微生物的系统发育树构建[17]. 受到这一研究的启发, 我们假设 PDO 可以作为一个独立的细胞功能单元, 在基因芯片样本聚类中比较细胞间 PDO 集合的差异性, 那么所构建的细胞树将反映出基因型与表型间的潜在关系.

2.2 训练集细胞树的构建

为了将已有的系统发育学方法应用于基因芯片分析, 我们在 PDO 基因集分析中借鉴了基因含量方法[29]. 基因芯片样本表达概型的涵义为: 对于每个基因应将详细的表达值“粗粒化”为是否表达或是否被检测到, 而对于每个样本应尽可能考虑基因的整体表达情况. 概型化的优点在于规避统计学上不稳定的特征基因集合提取和表达值归一化问题, 能够保证样本聚类结果和新样本表型预测的可靠性.

我们对 GSE11812 实验中的 30 个癌症细胞样本构建了样本树, 如图 1 所示. 与 Gyorffy 等人[13]在原文中构建的层次聚类树进行比较(见图 S1), 本文的方法主要有以下优势:

首先, 系统发育树具有自展值. 自展树允许对树型拓扑的可靠性进行直观评价, 而且可以将自展值较高的分支作为分组分析的依据. 根据图 1 的拓扑结构可将样本分为 4 组.

其次, 耐药谱较广的样本在图 1 树中有聚集趋势. 通过对比每组样本的细胞类型和耐药信息可以发现: 一方面, 样本树受细胞病理分型的影响, 如卵巢癌样本聚集在第 3 组, 而黑色素瘤样本则聚集在第 4 组; 另一方面, 耐药谱较广的样本有聚集在第 1 组的趋势. 尽管原文的层次聚类树也能展现细胞类型的规律, 但是高耐药样本则较为分散. 我们对图 1 树的 4 组样本的耐药信息进行了统计分析, 结果如表 2 所示. 第 1 组的平均样本耐药数约为 6, 为 4 组中最高. 此外, 第 1 组与其他组的样本耐药率具有统计学差异, 且耐药率的组内平均值最高. 因此, 我们推测 MDR 表型是影响样本树的重要因素, 而本文的方法能够将这种因素以系统发育树的方式揭示出来.

最后, 系统发育树中对阿霉素、紫杉醇和长春碱 3 种药物具有耐药性的样本都集中在第 1 组, 而 Gyorffy 等人[13]构建的层次聚类树则没有明显规律. 在 MDR 研究领域, 挖掘肿瘤对不同药物耐药的潜在联系对于临床诊断和个性化医疗方案的设计具有重要意义. Efferth 等人[30]曾在 2008 年发表文章, 认为 MDR 表型的预测可以通过检测阿霉素的耐药性而得到可靠的结果. 为了进一步研究分组结果与每种药物耐药性之间的关系, 我们将系统发育树作为预测模型, 选取 NCI-60 癌症细胞系进行检验.

2.3 NCI-60 细胞系的耐药表型预测

预测结果如表 3 所示. 系统发育学模型预测结果包含组别、分枝自展值以及广谱耐药性; 参照模型由于采用了 3 种药物的耐药性分组, 包含对 3 种药物的耐药结果. “1”和“0”分别代表耐药与敏感. 我们从两方面对结果进行比较.

(ⅰ) 细胞病理分型. 因为样本树的树型拓扑主要受细胞类型的影响, 所以系统发育树模型能够对测试样本的细胞类型进行预测, 而参照模型则无法做到. 例如, 肾癌、中枢神经肿瘤分布在第 1 组; 肠癌和白血病在第 2 组; 卵巢癌和黑色素瘤分别在第 3

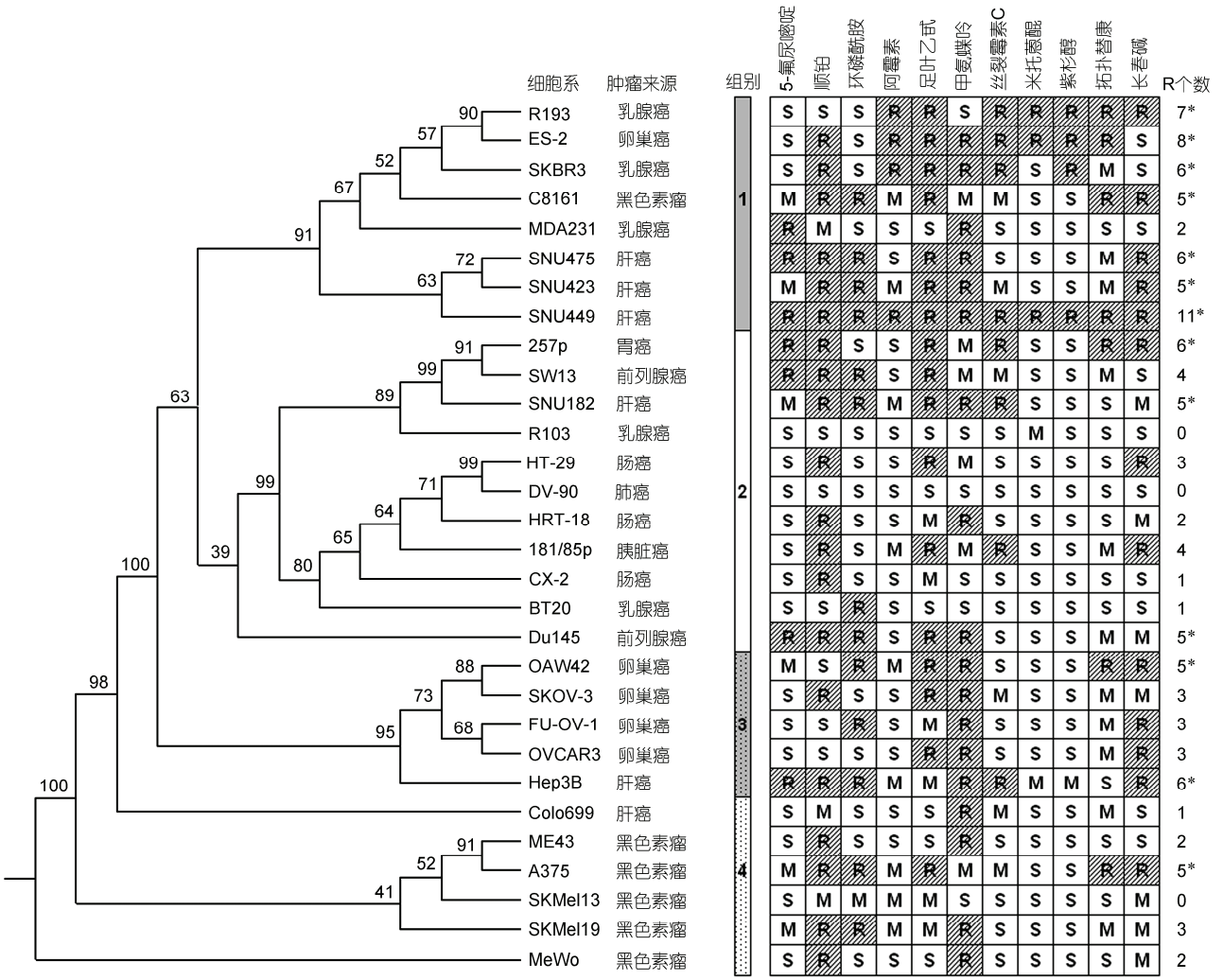

图 1　对 GSE11812 中 30 个肿瘤细胞系构建的细胞树

根据拓扑结构将样本分为 4 组. 树枝上注释了样本耐药信息, R/M/S 分别代表耐药、居中和敏感, 其中 R 的背景用阴影标记. 图中统计了样本耐药数, 大于 4 的值用*标记

表 2　对图 1 中样本耐药信息的统计分析 a)

组别	分组信息统计		样本耐药率(%)												耐药率的方差检验 P 值		
	样本数	样本平均耐药数	5-氟尿嘧啶	顺铂	环磷酰胺	阿霉素	足叶乙甙	甲氨蝶呤	丝裂霉素C	米托蒽醌	紫杉醇	拓扑替康	长春碱	平均值	2 组	3 组	4 组
1	8	6	38*	75*	50	50*	88*	75	50*	63	35*	50*	50*	57*	0.02082	0.03985	0.000531
2	11	3	27	73	36	0	55	27	27	27	0	9	0	36	–	0.9144	0.2644
3	5	4	20	40	60*	0	60	100*	20	80*	0	20	0	36		–	0.2393
4	6	2	0	67	33	0	17	67	0	17	0	17	0	20			–

a) 统计每组样本对每种药物的耐药率以及组内平均值, 并对每两组进行配对方差检验. *示 4 组中的耐药率最大值

组和第 4 组, 与训练集中的样本一致. 系统发育树模型将 MDA-MB-435 预测在第 4 组(主要由黑色素瘤样本组成). 此细胞系最初被认为是乳腺癌, 近几年经相关文献更正为黑色素瘤[31,32]. MDA-N 是由 MDA-MB-435 培养得到的细胞系[33], PDO 法同样将其预测为黑色素瘤. 此外, 来自不同数据集的相同细胞系能

表 3 NCI-60 癌症细胞系 MDR 预测结果[a)]

细胞系名	细胞类型	系统发育学模型			参照模型*		
		组别	Bootstrap	耐药性	阿霉素	紫杉醇	长春碱
ACHN	肾癌	1	83	1	0	0	0
TK-10	肾癌	1	63	1	0	1	0
786-0	肾癌	1	29	1	0	0	0
RXF 393	肾癌	1	41	1	1	0	0
SN12C	肾癌	1	36	1	0	1	0
A498	肾癌	1	62	1	0	0	0
UO-31	肾癌	1	40	1	1	1	0
CAKI-1	肾癌	1	44	1	1	1	0
SF-268	中枢神经系统癌	1	22	1	1	0	0
SF-295	中枢神经系统癌	1	77	1	1	1	0
SF-539	中枢神经系统癌	1	42	1	0	1	0
SNB-19	中枢神经系统癌	1	51	1	1	0	0
SNB-75	中枢神经系统癌	1	80	1	1	1	0
U251	中枢神经系统癌	1	59	1	0	0	0
HS 578T	乳腺癌	1	50	1	1	1	0
MDA-MB-231	乳腺癌	1	94	1	0	0	0
BT-549	乳腺癌	1	41	1	1	1	1
MCF7	乳腺癌	2	62	0	0	0	0
T-47D	乳腺癌	3	50	0	0	0	0
DU-145	前列腺癌	1	100	1	0	0	0
PC-3	前列腺癌	1	69	1	1	1	0
HOP-62	非小细胞肺癌	1	57	1	1	0	0
HOP-92	非小细胞肺癌	1	74	1	1	1	0
NCI-H226	非小细胞肺癌	1	31	1	1	0	0
NCI-H460	非小细胞肺癌	2	72	0	0	1	0
A549	非小细胞肺癌	2	62	0	0	1	0
NCI-H322M	非小细胞肺癌	2	50	0	0	0	0
EKVX	非小细胞肺癌	2	40	0	1	1	0
NCI-H522	非小细胞肺癌	2	79	0	0	0	0
HT29	肠癌	2	62	0	0	0	0
HCT-116	肠癌	2	87	0	0	0	0
COLO 205	肠癌	2	63	0	0	0	0
SW-620	肠癌	2	100	0	0	0	0
HCC-2998	肠癌	2	30	0	0	0	0
HCT-15	肠癌	2	46	0	0	0	0
KM12	肠癌	2	35	0	0	0	0
CCRF-CEM	白血病	2	100	0	0	1	0
MOLT-4	白血病	2	100	0	0	1	0
HL-60(TB)	白血病	2	100	0	0	0	0
RPMI-8226	白血病	2	90	0	0	0	0
K-562	白血病	2	76	0	0	0	0
SR	白血病	2	52	0	0	0	0
OVCAR-5	卵巢癌	1	75	1	0	1	0
NCI/ADR-RES	卵巢癌	1	48	1	0	1	0
OVCAR-8	卵巢癌	2	48	0	0	0	0
OVCAR-3	卵巢癌	3	100	0	0	0	0
IGROV1(1)	卵巢癌	3	45	0	0	0	1
IGROV1(2)	卵巢癌	3	50	0	0	0	0
SK-OV-3	卵巢癌	3	100	0	0	1	0
OVCAR-4	卵巢癌	3	71	0	0	0	0
UACC-257	黑色素瘤	1	38	0	0	0	0
LOX IMVI	黑色素瘤	1	54	1	1	0	1

续表 3

细胞系名	细胞类型	系统发育学模型			基本模型*		
		组别	Bootstrap	耐药性	阿霉素	紫杉醇	长春碱
SK-MEL-2	黑色素瘤	4	32	0	0	0	0
UACC-62	黑色素瘤	4	95	0	0	0	0
M14	黑色素瘤	4	82	0	0	0	0
MDA-MB-435	黑色素瘤	4	55	0	0	0	0
MDA-N	黑色素瘤	4	100	0	0	1	0
MALME-3M	黑色素瘤	4	31	0	0	0	0
SK-MEL-28	黑色素瘤	4	100	0	0	0	0
SK-MEL-5	黑色素瘤	4	66	0	0	0	0

a) 预测结果中的 1 和 0 分别代表耐药和敏感. *, 3 种药物的分组阈值均为 1×10^{-5}

很好地聚在一起, 说明系统发育学方法对表型预测具有稳定性和可重复性.

(ⅱ) MDR 表型. Wilcoxon 检验和 *t* 检验的结果如表 4 所示. 在 3 个参照模型中, 仅有阿霉素的模型能够预测 5-氟尿嘧啶和甲氨蝶呤的耐药性. *t* 检验的结果显示, 由长春碱的特征基因构建的参照模型能够预测环磷酰胺的耐药性. 有趣的是, 参照模型对于建模所用药物的耐药性预测性均不佳. 我们推测其原因是基于特征基因的建模依赖探针表达值归一化和特征变量集筛选策略.

系统发育树模型能够很好地区分出测试集中对 5-氟尿嘧啶和紫杉醇有耐药趋势的样本, 体现在 Wilcoxon 和 *t* 检验都具有显著性差异(P<0.05). 而且, 从 Wilcoxon 检验来看, 系统发育树模型能够较好地预测长春碱的耐药性. 此外, 结合训练集样本树与药物的关系, 我们推测, 系统发育树预测得到的广谱耐药趋势可能与细胞系对紫杉醇和长春碱的耐药性有紧密联系.

因此, 本文方法的预测结果显著优于参照模型. 我们认为这与系统发育学概念和方法的引入密不可分, 主要包括以下几个方面: (1) 基因表达概型借鉴了基因含量思想, 规避了特征基因集合提取和表达值归一化步骤, 在跨实验分析中表现稳定. (2) PDO 考虑了蛋白质二级结构域的组成与排序, 因此被假设为具有序列信息的细胞功能单元, 那么用 PDO 矩阵进行聚类分析则相当于比较不同细胞的整体功能; 在某些情况下, 不同类型的细胞要适应类似的“环境压力”, 如抗癌药物, 因此有可能表达出相同的 PDO 子集, 这种子集有可能并不影响整体表型, 但也有几率出现类似进化上的“趋同”现象, 例如本文中的广谱耐药组, 其 PDO 子集与紫杉醇和长春碱有关的微管蛋白等生物分子相关. (3) 本研究使用了经典的进化距离法构建细胞树, 由于此类构树算法建立在进化假说之上, 其思想可以借鉴到基因芯片领域来计算组织细胞的时空表达差异, 并且有利于解释聚类结果的生物学意义.

总而言之, 细胞类型并不是影响树型拓扑的唯一因素. MDR 也会影响到肿瘤细胞的整体表型. 我们发现, 被本文方法聚在第 1 组中的训练集和测试集样本, 主要包含肾癌、肝癌和中枢神经癌等, 大多为临床化疗预后较差的癌症种类或者分化程度较低的恶性腺癌. 一些具有组织特异性的肿瘤, 如卵巢癌和黑色素瘤等, 当广谱耐药性增强时, 其表型可能与天然高耐药肿瘤趋同. NCI-60 中的 NCI/ADR-RES (别名 OVCAR-8/ADR)是一个很好的例子, 该细胞系最初被认为是 MCF7 经阿霉素诱导产生的乳腺癌细胞系, 近几年的研究发现其是由卵巢癌 OVCAR-8 诱导而来[34], 预测结果中 OVCAR-8 位于聚类树拓扑结构的第 2 组, 而 NCI/ADR-RES 则有向第 1 组靠近的趋势(表 3), 说明存在由 MDR 表型引起的趋同现象.

2.4 差异表达 PDO 基因集

提取差异表达 PDO 的分组策略为: 将训练集细胞树 R 组中对阿霉素、紫杉醇或者长春碱耐药的 4 个样本作为耐药组, 非广谱耐药组中对阿霉素、紫杉醇和长春碱均敏感的 14 个样本作为敏感组.

对训练集进行差异表达 PDO 分析, 最终得到 45 个 PDO, 其中有 8 个上调 37 个下调. 回溯得到 65 个基因, 包括 9 个上调, 56 个下调. 表 5 展示了部分结果, 包含 SAM 分数较高或者与 DAVID 注释高度相关的 16 个基因. 全部结果请见表 S1.

对基因集进行富集分析, 大部分基因(61/65)与分子功能类的本体较为相关, 包括锌离子结合、碳酸

表 4 NCI-60 细胞系 GI50 值分布的 Wilcoxon 检验与 *t* 检验 *P* 值

药物	Wilcoxon 检验				*t* 检验 a)			
	系统发育学模型	参照模型			系统发育学模型	参照模型		
		阿霉素	紫杉醇	长春碱		阿霉素	紫杉醇	长春碱
5-氟尿嘧啶	0.0148*	0.0063*	0.3994	0.2669	0.0158*	0.0059*	0.1833	0.5482
顺铂	0.7995	0.8006	0.9250	0.6928	0.6598	0.7755	0.5738	0.8395
环磷酰胺	0.3570	0.7238	0.3705	0.5770	0.1067	0.2942	0.2697	0.0103*
阿霉素	0.4761	0.4603	0.4642	0.8130	0.2030	0.4501	0.1959	0.6353
足叶乙甙	0.9542	0.4572	0.9831	0.6278	0.9319	0.6883	0.8544	0.5186
甲氨蝶呤	0.0669	0.0028*	0.6627	0.3283	0.0852	0.0373*	0.1875	0.4508
丝裂霉素 C	0.6006	0.5397	0.8906	0.3391	0.5895	0.3963	0.4722	0.3154
米托蒽醌	0.9753	0.7994	0.9860	0.7090	0.9816	0.9167	0.7402	0.4231
紫杉醇	0.0011*	0.0791	0.1686	0.4063	0.0077*	0.0659	0.1887	0.7887
拓扑替康	0.9635	0.9116	0.9933	0.9014	0.8476	0.9405	0.7031	0.7730
长春碱	0.0326*	0.4210	0.4434	0.2669	0.0874	0.4028	0.0567	0.8555

a) *示 $P<0.05$

表 5 与耐药性相关的差异表达 PDO 及对应基因(部分)

PDO	探针 ID	基因 ID	基因名	调控	描述
PF00039, PF00039, PF00039	210495_x_at	2335	*FN1*	上调	纤维连接蛋白 1
PF00354	204684_at	4884	*NPTX1*	上调	神经元正五聚蛋白 1
PF01186	204298_s_at	4015	*LOX*	上调	赖氨酰氧化酶
PF02244, PF00246	205832_at	51200	*CPA4*	上调	羧肽 A4
PF04360	201858_s_at	5552	*SRGN*	上调	丝甘蛋白聚糖
PF01039	209623_at	64087	*MCCC2*	下调	甲基丁烯酰辅酶 A 羧化酶 2 (β)
PF02807, PF00217	202712_s_at	1159	*CKMT1B*	下调	线粒体酸激酶 1B
PF00194	203963_at	771	*CA12*	下调	碳酸酐酶Ⅻ
PF09507	212836_at	10714	*POLD3*	下调	聚合酶(DNA 引导), δ3, 辅亚基
PF00634, PF00634, ……, PF09104	208368_s_at	675	*BRCA2*	下调	乳腺癌易感基因 2
PF07992, PF00070	205512_s_at	9131	*AIFM1*	下调	细胞凋亡诱导因子, 线粒体相关, 1
PF08311	209642_at	699	*BUB1*	下调	苯并咪唑出芽抑制解除同源物 1（酵母）
PF00096, PF00096, ……, PF00096	206182_at	7693	*ZNF134*	下调	锌指蛋白 134
PF05622	219976_at	51361	*HOOK1*	下调	鞭毛钩同源物 1 (果蝇)
PF05622	221078_s_at	55704	*CCDC88A*	下调	卷曲螺旋域含 88A
PF08758, PF00028, ……, PF01049	201130_s_at	999	*CDH1*	下调	E-钙黏连素 1 (上皮)

酐酶活性和微管蛋白结合等. 这些研究领域的确都与癌症机制和肿瘤多药耐药性紧密相关[35~37].

碳酸酐酶是一种与细胞呼吸作用紧密相关的锌酶, 在肿瘤细胞中的表达量会下调. 在系统发育树模型的耐药组中, 碳酸酐酶家族所对应的 PDO(PF00-194)表达下调.

微管蛋白是细胞骨架的重要组成部分, 也是紫杉醇与长春碱的作用靶点, 结合后会抑制肿瘤细胞的有丝分裂. *BRCA2*, *HOOK1*, *CCDC88A*, *CKMT1B* 和 *AIFM1* 等基因的注释均与微管蛋白相关, 在耐药中发生了下调, 我们认为这是揭示紫杉醇和长春碱与 MDR 表型关系的切入点. 其中, 抑癌基因 BRCA2 承担着 DNA 损伤修复的重要功能, 其突变不但会增加乳腺、宫颈和前列腺等器官的癌变几率, 而且会引起癌细胞对抗癌药物的耐药性(如铂类药物)[38,39]. BRCA2 所对应的 PDO 差异表达, 对样本聚类具有很大贡献. 因此, 由差异表达 PDO 分析所筛选的基因集的确与肿瘤耐药表型联系紧密.

3 结论

本研究假设每一个 PDO 都可以作为一个独立的细胞功能单元. 基于这个假设, 我们将系统发育学中的基因含量方法引入到基因芯片分析中来, 利用自展法等系统发育学算法在 PDO 基因集水平构建细胞树. 对 MDR 基因表达数据的预测结果说明, 本文的方法对于肿瘤分类学研究具有一定意义[40].

本文方法的局限性在于, 对于样本的混杂度[41]较敏感, 聚类过程中会导致树型拓扑结构不稳定, 出现较低的自展值. 这是由于混杂样本可能表达多种细胞的 PDO, 因此无法确定与哪一组织细胞更接近. 因此, 本方法更加适合分析血液系统肿瘤或纯度较高的细胞系.

综上所述, 基于系统发育学方法的样本聚类方法对肿瘤多药耐药基因芯片的研究有着一定的指导意义, 并且为基因芯片与临床分析的结合设计了一套整体的解决方案. 由于本方法同时具有聚类与预测的能力, 使其有可能应用于基因芯片数据的临床实践, 特别是肿瘤的预测及诊断方面的相关应用.

参考文献

1 Kang H C, Kim I J, Park J H, et al. Identification of genes with differential expression in acquired drug-resistant gastric cancer cells using high-density oligonucleotide microarrays. Clin Cancer Res, 2004, 10: 272—284

2 Mallory J C, Crudden G, Oliva A, et al. A novel group of genes regulates susceptibility to antineoplastic drugs in highly tumorigenic breast cancer cells. Mol Pharmacol, 2005, 68: 1747—1756

3 Tai I T, Dai M, Owen D A, et al. Genome-wide expression analysis of therapy-resistant tumors reveals SPARC as a novel target for cancer therapy. J Clin Invest, 2005, 115: 1492—1502

4 Lee C H, Macgregor P F. Using microarrays to predict resistance to chemotherapy in cancer patients. Pharmacogenomics, 2004, 5: 611—625

5 Yabuki N, Sakata K, Yamasaki T, et al. Gene amplification and expression in lung cancer cells with acquired paclitaxel resistance. Cancer Genet Cytogenet, 2007, 173: 1—9

6 Subramanian A, Tamayo P, Mootha V K, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA, 2005, 102: 15545—15550

7 Pollack J R. A perspective on DNA microarrays in pathology research and practice. Am J Pathol, 2007, 171: 375—385

8 Verhaak R G, Staal F J, Valk P J, et al. The effect of oligonucleotide microarray data pre-processing on the analysis of patient-cohort studies. BMC Bioinformatics, 2006, 7: 105

9 Redestig H, Repsilber D, Sohler F, et al. Integrating functional knowledge during sample clustering for microarray data using unsupervised decision trees. Biom J, 2007, 49: 214—229

10 Kerr M K, Churchill G A. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci USA, 2001, 98: 8961—8965

11 Suzuki R, Shimodaira H. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. Bioinformatics, 2006, 22: 1540—1542

12 Barrett T, Troup D B, Wilhite S E, et al. NCBI GEO: Archive for high-throughput functional genomic data. Nucleic Acids Res, 2009, 37: D885—890

13 Gyorffy B, Surowiak P, Kiesslich O, et al. Gene expression profiling of 30 cancer cell lines predicts resistance towards 11 anticancer drugs at clinically achieved concentrations. Int J Cancer, 2006, 118: 1699—1712

14 Shankavaram U T, Reinhold W C, Nishizuka S, et al. Transcript and protein expression profiles of the NCI-60 cancer cell panel: An integromic microarray study. Mol Cancer Ther, 2007, 6: 820—832

15 Su A I, Wiltshire T, Batalov S, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA, 2004, 101: 6062—6067

16 Chen M, Sinha M, Luxon B A, et al. Integrin alpha6beta4 controls the expression of genes associated with cell motility, invasion, and metastasis, including S100A4/metastasin. J Biol Chem, 2009, 284: 1484—1494

17 Fukami-Kobayashi K, Minezaki Y, Tateno Y, et al. A tree of life based on protein domain organizations. Mol Biol Evol, 2007, 24: 1181—1189

18 Benson D A, Karsch-Mizrachi I, Lipman D J, et al. GenBank. Nucleic Acids Res, 2009, 37: D26—D31

19 Finn R D, Tate J, Mistry J, et al. The Pfam protein families database. Nucleic Acids Res, 2008, 36: D281—D288

20 Pepper S D, Saunders E K, Edwards L E, et al. The utility of MAS5 expression summary and detection call algorithms. BMC Bioinformatics, 2007, 8: 273

21 Retief J D. Phylogenetic analysis using PHYLIP. Methods Mol Biol, 2000, 132: 243—258

22 Zhao Y, Simon R. BRB-arraytools data archive for human cancer gene expression: A unique and efficient data sharing resource. Cancer Inform, 2008, 6: 9—15

23 Tusher V G, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci

USA, 2001, 98: 5116—5121
24 Hall M, Frank E, Holmes G, et al. The WEKA data mining software: An update. SIGKDD Explorations, 2009, 11: 10—18
25 Shoemaker R H. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer, 2006, 6: 813—823
26 Huang da W, Sherman B T, Lempicki R A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc, 2009, 4: 44—57
27 Carter S L, Eklund A C, Mecham B H, et al. Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. BMC Bioinformatics, 2005, 6: 107
28 Li Z, Liu Q, Song M, et al. Detecting correlation between sequence and expression divergences in a comparative analysis of human serpin genes. Biosystems, 2005, 82: 226—234
29 Snel B, Huynen M A, Dutilh B E. Genome trees and the nature of genome evolution. Annu Rev Microbiol, 2005, 59: 191—209
30 Efferth T, Konkimalla V B, Wang Y F, et al. Prediction of broad spectrum resistance of tumors towards anticancer drugs. Clin Cancer Res, 2008, 14: 2405—2412
31 Rae J M, Creighton C J, Meck J M, et al. MDA-MB-435 cells are derived from M14 melanoma cells—A loss for breast cancer, but a boon for melanoma research. Breast Cancer Res Treat, 2007, 104: 13—19
32 Ross D T, Scherf U, Eisen M B, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet, 2000, 24: 227—235
33 Lorenzi P L, Reinhold W C, Varma S, et al. DNA fingerprinting of the NCI-60 cell line panel. Mol Cancer Ther, 2009, 8: 713—724
34 Liscovitch M, Ravid D. A case study in misidentification of cancer cell lines: MCF-7/AdrR cells (re-designated NCI/ADR-RES) are derived from OVCAR-8 human ovarian carcinoma cells. Cancer Lett, 2007, 245: 350—352
35 Wakasugi T, Izumi H, Uchiumi T, et al. ZNF143 interacts with p73 and is involved in cisplatin resistance through the transcriptional regulation of DNA repair genes. Oncogene, 2007, 26: 5194—5203
36 Hunakova L, Bodo J, Chovancova J, et al. Expression of new prognostic markers, peripheral-type benzodiazepine receptor and carbonic anhydrase IX, in human breast and ovarian carcinoma cell lines. Neoplasma, 2007, 54: 541—548
37 Canta A, Chiorazzi A, Cavaletti G. Tubulin: A target for antineoplastic drugs into the cancer cells but also in the peripheral nervous system. Curr Med Chem, 2009, 16: 1315—1324
38 Wang W, Figg W D. Secondary BRCA1 and BRCA2 alterations and acquired chemoresistance. Cancer Biol Ther, 2008, 7: 1004—1005
39 Sakai W, Swisher E M, Karlan B Y, et al. Secondary mutations as a mechanism of cisplatin resistance in BRCA2-mutated cancers. Nature, 2008, 451: 1116—1120
40 Weigelt B, Reis-Filho J S. Histological and molecular types of breast cancer: Is there a unifying taxonomy? Nat Rev Clin Oncol, 2009, 6: 718—730
41 Demichelis F, Magni P, Piergiorgi P, et al. A hierarchical Naive Bayes Model for handling sample heterogeneity in classification problems: An application to tissue microarrays. BMC Bioinformatics, 2006, 7: 514

补充材料

图 S1 由 Gyorffy 等人构建的层次聚类树
表 S1 差异表达 PDO 分析结果

本文的以上补充材料见网络版 csb.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.