PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2017, 44(11): 1016~1025
www.pibb.ac.cn

乳腺癌发生的特征基因筛选及模式识别 *

温建鑫　王学栋　李晓琴 **　常　宇
(北京工业大学生命科学与生物工程学院，北京 100124)

摘要　本文选取癌症基因组图谱数据库的乳腺癌样本作为数据集，在全基因组的水平上研究乳腺癌病人从正常到发病 I 期基因表达的变化，寻找与乳腺癌发病密切相关的特征基因，建立乳腺癌发生的模式识别分类方法，为乳腺癌预防及早期诊断提供理论支持．研究中，综合利用相关性、t 检验、置信区间等统计学方法，建立乳腺癌发生特征基因筛选方法，获得与乳腺癌发生具有显著性差异的特征基因 336 个．通过机器学习方法建模，得到的分类准确率能达到 98%以上，与之前乳腺癌相关的研究相比，准确率更高．同时采用 KEGG (kyoto encyclopedia of genes and genomes) 通路分析得到与基因显著相关($P < 0.05$)的通路有 8 个，GO(gene ontology) 基因功能富集分析显示与基因显著相关($P < 0.05$)的功能有 18 个．最后对映射在 8 个通路中的一部分基因进行简要功能分析，说明了其在调控水平上的密切关系，表明识别的特征基因在乳腺癌的发生过程中有重要的作用，这对了解乳腺癌发病机理以及乳腺癌的早期诊断非常重要．

关键词　乳腺癌，基因表达，模式识别，肿瘤预测，早期诊断
学科分类号　Q7，Q81　　**DOI**: 10.16476/j.pibb.2017.0221

乳腺癌是全球女性最常见的恶性肿瘤之一，全世界每年新发乳腺癌病例约 138 万，死亡人数达到 46 万，占女性新发癌症病例的 1/3、癌症死亡病例的 17%．美国 2016 年的乳腺癌发病数为 25.5 万例，死亡 4.1 万例，死亡率在女性恶性肿瘤中位居前三[1]．中国的乳腺癌发病率的增长速度是全球最快的国家之一，近十多年来大城市的上升幅度达 20%～30%[2]．尽管乳腺癌的诊断和治疗措施不断提高，但死亡率仍未得到有效控制．因此，对乳腺癌发病模式进行分类识别研究，识别乳腺癌发生的关键基因，对有效预防、早期诊断及提供个性化治疗等方面具有重要意义．

乳腺癌的发生与女性的生理、心理、遗传、环境等息息相关，是多种因素相互影响的共同结果．随着微阵列测序技术和生物信息学的飞速发展，人们能够利用基因芯片检测数千个基因的表达，了解病人特异的表达谱，为系统研究乳腺癌发病的相关基因提供了技术保证，从而成为乳腺癌研究领域的一种常规技术[3]．Sotiriou 等[4]根据基因表达值筛选出了 97 个基因，这些基因能把中分化的乳腺癌组织进一步划分为两个子类，一类具有较高的癌症复发风险，另一类具有较低的癌症复发风险．Feng 等[5]通过基因表达值筛选得到一组 188 个基因组成的基因团，能够较准确地区分乳腺原发癌与癌旁的正常乳腺组织．Ma 等[6]通过分析 99 例乳腺癌患者的 7 650 个基因的表达谱，证实了一组 9 个基因组成的基因群可以较准确区分乳腺癌的组织学分级(准确度 83%，灵敏度 85%)．Xie 等[7]通过辨识度和独立性相结合的特征选择方法，根据基因表达数据对乳腺癌的辨识能力不同，将所有基因按照辨识能力由大到小进行排序，筛选出一组由 10 个基因组成的基因团，能够很好地区分乳腺癌组织和正常

* 国家自然科学基金(11572014)和智能制造领域大科研推进计划(01500054631751)资助项目．
** 通讯联系人．
Tel: 15313254516, E-mail: lxq0811@bjut.edu.cn
收稿日期：2017-08-02，接受日期：2017-09-08

组织(准确率 85.32%).

以上研究都是从基因表达水平对乳腺癌的发生和发展机制进行了相关研究．本文则从全基因组水平上研究基因表达谱信息，从微阵列数据出发，利用生物信息学技术对乳腺癌的癌旁和Ⅰ期进行建模预测，全面寻找对乳腺癌发生相关的特征基因，为进一步揭示乳腺癌发生机制及新靶向药物的研发奠定理论基础.

1 数据与方法

1.1 数据及预处理

本文用到的乳腺癌数据来源于癌症基因组图谱(the cancer genome atlas，TCGA)公共数据库(https://cancergenome.nih.gov/)，下载乳腺癌病人的基因表达谱数据和临床信息数据，得到样本数 1 222 个，基因数 60 483 个．用 python 语言编写程序将病人的表达谱和临床信息进行整合，挑选出癌旁基因表达谱和乳腺癌为Ⅰ期的样本表达谱数据，同时为后续研究方便，删去在所有样本中表达值为空的基因，最终得到样本数 295 例，其中癌旁 113 例，乳腺癌Ⅰ期样本有 182 例，基因数有 57 179 个，具体信息见表 1.

Table 1 Clinical data of breast cancer samples

	Stage Ⅰ	Normal
Number	182	113
Age	69.0±15.9	69.5±13.8

由于基因的表达数据相差较大，为了方便后期建模且训练收敛迅速，需要对表达谱数据进行归一化处理，要求其区间为[0, 1]，取值公式为

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

其中：x 为各样本某一基因的表达值；$x_{\min}$ 为该基因在所有样本表达值的最小值；$x_{\max}$ 为该基因在所有样本表达值的最大值.

1.2 特征基因的筛选

本文涉及分类问题为常见的二分类模型，但是由于特征的维度远远大于样本数，特征之间的关联关系相对复杂、关联关系间依赖性影响等问题，使得学习产生了诸多问题，比如：分析数据、训练模型时间长，数据量大导致“维度灾难”，使得模型过于复杂等等．为了克服这些不利因素的影响，提高特征识别的准确率，本文采用各种统计学的方法，将它们组合起来形成一套特征基因筛选的流程.

a．相关性筛选． 相关性分析是指对两个或多个具备相关性的变量元素进行分析，从而衡量两个变量因素的相关密切程度，用相关系数 r 来表示．常见的相关系数为 Pearson 相关系数，表示两个连续随机变量之间的线性相关程度，但是对于不服从正态分布的等级变量、总体分布类型未知的变量不能再用 Pearson 系数来描述关联性，此时可采用秩相关，也称等级相关，来描述两个变量之间的关联程度与方向．本文分类结果为二分类变量，所以采用 spearman 相关系数来描述基因表达与分类结果的相关性，将相关系数 $r > 0.5$ 的基因保留下来.

b．t 检验筛选．t 检验是用 t 分布理论来推论差异发生的概率，从而比较两个平均数的差异是否显著，t 检验分为单总体检验和双总体检验．由于 t 检验的样本必须满足齐性分布，在对候选基因和分类结果做 t 检验之前必须做齐性检验，对于满足齐性的基因采用双总体 t 检验方法筛选特征基因，保留对分类结果标签具有显著性差异($P < 0.05$)的基因，对于不满足齐性分析的基因采用非参检验中的 Cruskal-Wallis 秩和检验筛选特征基因，保留对对分类结果标签具有显著性差异($P < 0.05$)的基因.

c．置信区间筛选．t 检验筛选虽然能够从众多基因中选择出和分类结果均值有差异的特征基因，但是无法解决基因表达在癌旁和Ⅰ期中分布差异不大，均值分布有差异的问题．置信区间筛选则可以根据基因表达在癌旁和Ⅰ期中分布的差异来筛选特征基因．置信区间筛选的计算公式为：

$$[\mu - i\times\sigma, \mu + i\times\sigma] \tag{2}$$

其中 μ 为基因在癌旁或Ⅰ期中的表达均值，σ 为标准差，为保证区间的有效性，i 取值必须大于 1．分别计算每一个基因在癌旁和Ⅰ期样本的置信区间，比较两区间是否有重合来筛选对分类结果有差异性的基因.

特征筛选的具体流程如图 1 所示，所有的筛选过程均采用 python 语言编写代码实现.

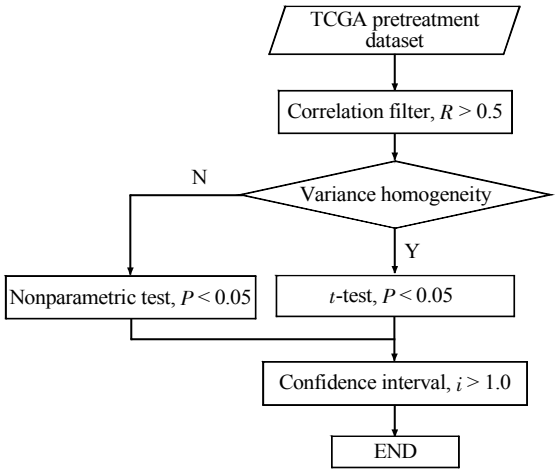

Fig. 1　Flowchart of signature genes identification for breast cancer

1.3　建模预测及模型评价方法

本文的研究内容为模式识别分类问题中的二分类模型，随机选取数据集的 90%作为训练集，剩余 10%作为测试集，采用支持向量机(support vector machine，SVM)、随机森林(random forest，RF)、人工神经网络(artificial neural network，ANN)多种分类模型对该数据集进行建模预测，重复 10 次该建模过程，采用混淆矩阵(confusion matrix，CM)对分类结果进行显示，通过计算总体准确率(accuracy，*ACC*)以及敏感度(sensitivity，*SEN*)、特异度(specificity，*SPE*)、马修斯相关系数(matthews correlation coefficient，*MCC*)指标来评价模型的分类结果．*ACC*，*SEN*，*SPE* 与 *MCC* 的定义如下：

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

$$SEN = \frac{TP}{TP + FN} \tag{4}$$

$$SPE = \frac{TN}{FP + TN} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{6}$$

其中：*TP*(true positive)被判定为正样本，事实上也是正样本的个数；*TN*(true negative)被判定为负样本，事实上也是负样本的个数；*FP*(false positive)被判定为正样本，但事实上是负样本的个数；*FN*(false negative)被判定为负样本，但事实上是正样本的个数．

2　结果与讨论

2.1　特征基因的分类结果

本文研究的目标是识别对乳腺癌的发生有着至关重要的特征基因，为进一步揭示乳腺癌发生和发展的机理奠定基础，为此识别的基因必须具有足够高的癌症分类能力，采用的分类模型必须具有高的分类精度，才能充分地证明基因的有效性．通过本文筛选方法对预处理之后的 57 179 个特征基因进行识别，得到候选基因 336 个．对以上得到的基因分别采用支持向量机(SVM)、随机森林(RF)、人工神经网络(ANN)循环 10 次建模得到的分类结果平均值，如表 2 所示．

Table 2　Performance of classification model for 336 breast cancer signature genes

	SVM	RF	ANN
ACC	0.9833 ± 0.0224	0.9947 ± 0.0100	0.9967 ± 0.0100
SEN	1.0000 ± 0.0000	0.9923 ± 0.0231	0.9889 ± 0.0333
SPE	0.9727 ± 0.0380	1.0000 ± 0.0000	1.0000 ± 0.0000
MCC	0.9668 ± 0.0436	0.9934 ± 0.0199	0.9921 ± 0.0237

从表 2 可以看出，建模的三种分类模型中支持向量机的敏感性达到 1，说明其对于癌旁组织的识别率准确率最高；而随机森林和人工神经网络的特异性达到 1，说明其对于癌症Ⅰ期的样本识别率最高．综合对比三种分类模型可以看出，随机森林的马修斯相关系数达到最高：0.9934，而支持向量机和人工神经网络也都达到 0.96 以上，表明本文筛选的特征基因对乳腺癌癌旁和Ⅰ期的分类具有很高的区分能力，体现出本文筛选方法的可靠性和优越性．

本文的分类结果还与已有的研究成果进行了对比．Xie 等[7]采用辨识度和独立性相结合的特征选择方法 FSDI(feature selection based on discernibility and independence of a feature)对乳腺癌的基因表达数据筛选出多组特征基因的组合，利用这些特征基因的组合对乳腺癌的正常和发病样本进行分类，分类结果表明当 FSDI 方法筛选的特征基因数达到 10 个时，分类准确率最高为 0.8532．同时也采用了经典的特征选择方法比如 Weight[8]、mRMR[9]、ARCO[10]、SVM-RFE[11]和 Relief[12]进行对比，分类结果显示采用 Weight 方法当特征基因达到 6 个时

分类准确率最高，此时为 0.7775，采用 mRMR 方法当特征基因达到 4 个时分类准确率最高，此时为 0.8305，采用 ARCO 方法当特征基因达到 15 个时分类准确率最高，此时为 0.8194，采用 SVM-RFE 方法当特征基因达到 15 个时分类准确率最高，此时为 0.7290，采用 Relief 方法当特征基因达到 3 个时分类准确率最高，此时为 0.6951，具体信息如表 3 所示.

Table 3 Classification of breast cancer by several classical feature selection methods

Genes numbers	FSDI	Weight	mRMR	SVM-RFE	Relief	ARCO
2	0.6922	0.6365	0.6132	0.6188	0.6245	0.7770
3	0.7402	0.5902	0.8108	0.6161	0.6951	0.7854
4	0.7995	0.6848	0.8305	0.6866	0.6640	0.7855
6	0.7967	0.7775	0.7967	0.7234	0.6697	0.7544
7	0.7601	0.7182	0.7910	0.6980	0.6697	0.7544
10	0.8532	0.7200	0.8024	0.6923	0.6837	0.7996
15	0.8237	0.7683	0.7742	0.7290	0.6753	0.8194

上述方法的分类准确率均低于本文的癌旁和癌症Ⅰ期分类模型．考虑到特征基因数量可能对分类结果产生影响，将特征选取流程中的置信区间筛选 i 值改为 1.5，得到基因数 16 个，其中蛋白质编码基因 13 个．重复 **2.1** 节第一段过程得到的分类结果平均值如表 4 所示，从表 4 中可以看出筛选出的 13 个特征编码基因采取人工神经网络建模时得到最高的平均分类准确率 100%，SVM 和 RF 分类模型准确率也分别达到 0.9864 和 0.9899，与之前的 336 个基因建模结果相比，具有更高的准确率，表明本文筛选的特征基因的数目大小不影响乳腺癌的预测准确率.

Table 4 Performance of classification model for 13 breast cancer signature genes while i=1.5

	SVM	RF	ANN
ACC	0.9933 ± 0.0133	0.9950 ± 0.0119	1.0000 ± 0.0000
SEN	1.0000 ± 0.0000	0.9887 ± 0.0272	1.0000 ± 0.0000
SPE	0.9891 ± 0.0219	1.0000 ± 0.0000	1.0000 ± 0.0000
MCC	0.9864 ± 0.0272	0.9899 ± 0.0240	1.0000 ± 0.0000

2.2 特征基因的 KEGG 通路分析和 GO 功能富集分析

本文采用 DAVID 生物信息学资源 6.8/NIAID 功能注释工具对特征基因集进行 KEGG 通路分析，选取 GO 数据库进行功能注释．为研究调控网络上的相互联系，对得到的 336 个特征基因进行综合分析.

KEGG 分析的结果表明，在乳腺癌中的 336 个特征基因与黏着斑、肿瘤通路、ABC 转运蛋白、醚脂类代谢、肌动蛋白细胞骨架调控、轴突导向、钙信号通路等通路显著相关($P < 0.05$)，这些通路按显著性排列中占据最多基因的是肿瘤通路(12 个)，其次是黏着斑通路(9 个)，有文献表明黏着斑通路是指细胞与胞外基质之间的连接方式，它能维持细胞在运动过程中的张力以及细胞生长的信号传递，如果细胞骨架信号传导控制异常会导致细胞外刺激和细胞反应的脱节，这在免疫疾病、发育缺陷和癌症中常见．通路分析表明本文识别的特征基因与肿瘤的发生有重要关系，在癌症的识别上具有可信性．具体信息如表 5 所示.

Table 5 KEGG pathways for the 336 signature genes

Term	Count	*P*-Value	Genes
Focal adhesion	9	0.0081	CAV2, FN1, CAV1, PPP1R12B, PDGFD, MYLK, FIGF, EGFR, PAK7
Pathways in cancer	12	0.0212	EDNRB, FN1, FGF1, GNG11, CXCL12, FOXO1, RUNX1T1, FIGF, CKS2, EGFR, STAT5B, EPAS1
ABC transporters	4	0.0234	ABCA9, ABCA10, ABCA6, ABCA8
Ether lipid metabolism	4	0.0248	PPAP2A, PPAP2B, PAFAH1B3, ENPP2
Regulation of actin cytoskeleton	8	0.0283	FN1, FGF1, GSN, PPP1R12B, PDGFD, MYLK, EGFR, PAK7
Axon guidance	6	0.0326	UNC5B, CXCL12, ABLIM1, DPYSL2, PAK7, SLIT3
Calcium signaling pathway	7	0.0395	EDNRB, ADRB2, RYR3, MYLK, EGFR, GNAL, OXTR
Neuroactive ligand-receptor interaction	9	0.0401	S1PR1, EDNRB, ADRB2, LEPR, THRB, SSTR1, NR3C1, OXTR, ADCYAP1R1

GO 分析发现在乳腺癌中的 336 个特征基因与肌动蛋白结合、电压门控钠通道活性、整合素结合、钙离子结合、蛋白激酶结合、转录激活活性等功能显著相关($P < 0.05$).

2.3 乳腺癌相关特征基因分析

2.3.1 *i* 取 1.5 对应特征基因分析

如 **2.1** 节所述，当置信区间筛选的阈值 *i* 取 1.5 时，得到 16 个基因，它们分别为：JAM2、RP11-875O11.1、HOXA4、TMEM220-AS1、SH3BGRL2、DMD、CACHD1、TMEM220、TSLP、SPRY2、SCN4B、FIGF、ANKRD29、ADAMTS5、CTC-297N7.9 和 MME．已有文献表明，这其中 TSLP、FIGF、MME 3 个基因与乳腺癌的发生和发展密切相关，JAM2、DMD、SPRY2 这 3 个基因参与肿瘤的发生和转移过程，以上 6 个基因都是非常重要的癌症相关基因.

a．TSLP(thymic stromal lymphopoietin)主要由皮肤、肺、胸腺和胃肠道中的上皮细胞产生，是涉及 T 辅助细胞 2 型细胞型炎症免疫应答的主调节因子．TSLP 是一种多功能蛋白，可作为肿瘤抑制因子[13-15]，在乳腺癌原发性肿瘤中的表达与预后相关[16].

b．FIGF(C-fos induced growth factor)是一种蛋白质编码基因，编码的蛋白质是血小板衍生生长因子家族的成员．该基因的 KEGG 通路映射同时出现在黏着斑通路和癌症通路上，与乳腺癌的发生和发展密切相关[17-21]，可作为乳腺癌的新型预后因素.

c．MME(membrane metalloendopeptidase)，别名 CD10，该基因编码一种常见的急性淋巴细胞白血病抗原，是诊断人急性淋巴细胞白血病的一种重要的细胞表面标志物，同时，该基因的表达与乳腺癌的生存率有关，是乳腺癌病人重要的预后标志基因[22-24]，可列为乳腺癌化疗前的标志物[25].

d．JAM2(junctional adhesion molecule 2)，该基因属于连接黏附分子家族．其编码的蛋白质是一种多功能的跨膜蛋白，属于免疫球蛋白超家族，定位于上皮细胞和内皮细胞的紧密连接处，可作为一种黏附配体与多种免疫细胞类型相互作用，调节血管生成、肿瘤转移、细胞增殖等过程[26].

e．DMD(dystrophin)，该基因在基因组范围全长大于 2 Mb，是当今已知的人类最大的基因，其编码的蛋白质复合体是细胞骨架和细胞外基质的重要组成部分．早年有文献指出 DMD 编码的肌营养不良蛋白聚糖复合体在癌症的发生中起着关键作用[27]，后来经过多年的研究 Wang 等[28]提出 DMD 可作为肿瘤抑制基因和可能的抗肿瘤转移因子.

f. SPRY2(sprouty RTK signaling antagonist 2)是一种蛋白编码基因，其编码的蛋白质可作为受体酪氨酸激酶信号传导的抑制剂，其表达或功能的抑制可能促进细胞增殖和血管生成．该基因作为调节细胞致癌转化的肿瘤抑制因子，具有作为治疗性抗癌剂的潜力[29].

2.3.2 通路相关特征基因分析

如 **2.2** 节所述，KEGG 通路分析的 336 个特征基因与黏着斑通路、肿瘤通路具有显著相关性，在这两个通路中，黏着斑通路有 9 个基因，其中 7 个基因至少一次出现在其他通路中，肿瘤通路有 12 个基因，其中除去与黏着斑通路重复的基因，有 3 个基因至少一次出现在其他通路上，这 10 个基因都是非常重要的乳腺癌相关基因．图 2 所示为这 10 个基因在乳腺癌癌旁组织和癌症Ⅰ期表达水平箱线图，从图 2 可以看出这些基因在乳腺癌癌旁组织和癌症Ⅰ期表达量存在显著差异，其中，FN1同时出现在黏着斑通路和肿瘤通路中，在乳腺癌组织中表达上调，其他 9 个基因在乳腺癌组织中表达均下调．本研究还对这 10 个基因采用支持向量机(SVM)、随机森林(RF)、人工神经网络(ANN)循环 10 次建模，得到的平均分类准确率为别为：98.21%、99.67%、99.51%，具有较高的准确率.

a．FN1(fibronectin 1)，该基因编码纤维连接蛋白，该蛋白参与细胞黏附和迁移过程，包括胚胎产生、伤口愈合、凝血、宿主防御和转移等．有文献[30]报道：纤连蛋白水平的增加会导致乳腺上皮细胞的过度增殖和腺泡增大，同时，纤连蛋白分化腺苷会刺激增殖，导致乳腺上皮细胞生长停滞，对维持适当的腺泡形态有不利影响，这些影响表明在乳腺发育和肿瘤形成过程中纤连蛋白表达与上皮细胞生长之间存在联系．测定纤维连接蛋白含量可作为乳腺癌和其他可能的癌症预后的生物标志物[31]．由图 2 可知，FN1 在癌旁组织中表达量低于癌症Ⅰ期表达量，这对于我们更好地理解纤连蛋白动态调节乳腺细胞的发育和肿瘤的发生有一定的指导作用.

b．PPP1R12B(protein phosphatase 1 regulatory subunit 12B)，该基因是一种蛋白质编码基因，位于 1 号染色体 p32.1-p32.2 区域，能调节肌球蛋白磷酸酶活性，增强收缩机器的 Ca^{2+} 敏感性．与该基因相关的 GO 注释包括酶活化剂活性和磷酸酶调节剂活性.

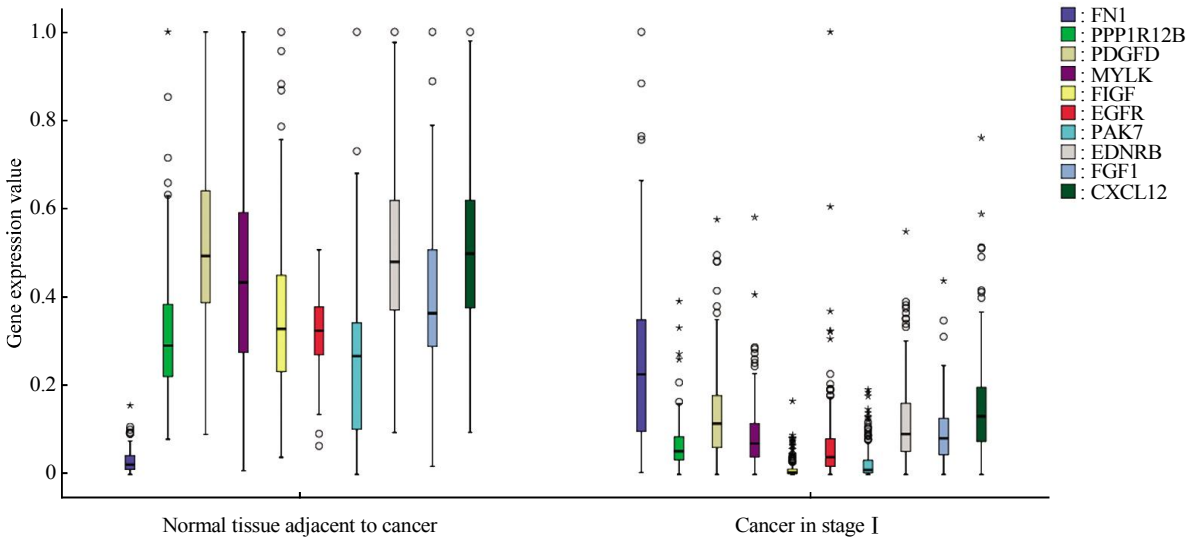

Fig. 2　Box chart of gene expression levels in paracancerous tissue and stage Ⅰ

c．PDGFD、FIGF、FGF1 属于生长因子类基因，其中 PDGFD(platelet derived growth factor D)、FIGF(C-fos induced growth factor)编码的蛋白质是血小板衍生生长因子家族的成员，FGF1(fibroblast growth factor 1)编码的蛋白质是成纤维细胞生长因子(FGF)家族的成员，这些因子在调节胚胎发育、细胞增殖、细胞迁移、存活和趋化性中起重要作用．FIGF 别名 VEGFD (vascular endothelial growth factor D)．Nakamura 等[17]提出 VEGFD 的表达与淋巴结转移相关，可能是乳腺癌的新型预后因素，VEGF-D 可用于治疗乳腺癌，作为手术后侵袭性治疗的决策生物标志物．Akahane 等[18]在研究 VEGFD 对人乳腺癌细胞凋亡的过程中也得出结论：VEGFD 除了具有血管生成和淋巴管生成因子的功能外，还具有作为乳腺癌细胞存活因子的新功能．Teramoto 等[19]通过实时聚合酶链反应(RT-PCR)测量了 109 例患者乳腺癌组织中 VEGF-C 和 VEGF-D 的 mRNA 表达量，结果表明，在乳腺癌中 VEGF-C 和 VEGF-D 参与淋巴结转移前的淋巴管侵袭，淋巴结转移后它们的表达值降低． Wang 等[20]为了研究 VEGFD 作为乳腺癌存活的预后因素，在 2012 年总结搜索了电子数据库 PubMed 和 EMBASE，通过荟萃分析表明，VEGF-C 和 VEGF-D 可以预测乳腺癌患者预后不良，但是需要对 VEGFD 的表达值进行标准化．Zhao 等[21]的研究表明：肿瘤衍生的 VEGF-C/D 诱导肿瘤周围淋巴管生成，从而促进乳腺癌患者淋巴细胞浸润和转移扩散．研究表明，FGF1 与雌激素受体阴性乳腺癌的风险显著相关[32]．Slattery 等[33]研究各种生长因子对于乳腺癌的影响，结果表明 FGF1 和 ERBB2 都显著影响乳腺癌病人整体生存率，这两种基因的遗传变异可能与乳腺癌诊断后的存活有关．

d．MYLK、EGFR、PAK7 属于蛋白激酶类基因，这 3 个基因编码的蛋白质是蛋白激酶超家族的成员，可在各种不同的信号通路中起作用，包括细胞骨架调节、细胞迁移、增殖和细胞存活．其中 MYLK(myosin light chain kinase)编码的蛋白质还能够调节内皮细胞和血管的通透性，是导致成纤维细胞凋亡的关键参与者．有文献表明 MYLK 通过与激活的 ERK1/2 的相互作用来促进乳腺癌细胞的增殖和迁移[34]，MYLK 通过 p38 通路的抗凋亡来实现乳腺癌细胞的高增殖能力[35]．EGFR (epidermal growth factor receptor)编码的蛋白质是一种跨膜糖蛋白，该蛋白质是表皮生长因子家族成员的受体，可与配体结合诱导受体二聚化和酪氨酸自磷酸化从而导致细胞增殖[36]．黏着斑激酶(FAK)和表皮生长因子受体(EGFR)是在人乳腺癌中过度表达和激活的蛋白酪氨酸激酶，Golubovskaya 等[37]通过实验结果表明，FAK 和 EGFR 信号通路的双重抑制可以协同增强乳腺癌细胞凋亡．Ray 等[38]通过对 50 例子宫颈癌和 50 例女性乳腺癌的 EGFR 进行比较发现，在乳腺癌病例中，EGFR 的过度表达与淋巴结转移显著相关，EGFR 可作为乳腺癌的潜在治疗靶点[39]．Brandt 等[40]通过基于人群的病例对照研究，

评估 EGFR 多态性与乳腺癌风险之间的关系得出结论：位于 EGFR 基因的内含子 1 中的 5′序列的长度可能增加家族性乳腺癌的风险，其作用可以通过饮食因素调节．PAK7，别名 PAK5(p21 (RAC1) activated kinase 5)，该基因可与 MARK2 拮抗作用，破坏 F-actin 网络稳定，导致应力纤维和局部黏连的消失，是肿瘤细胞调节中的致癌激酶[41].

e．EDNRB、CXCL12 属于 G 蛋白偶联受体系列基因，该基因编码的蛋白质可作为 G 蛋白偶联受体在许多不同的细胞功能中发挥作用，包括胚胎产生、免疫监视、炎症反应、组织稳态、肿瘤生长和转移．其中 EDNRB(endothelin receptor type B)简称 ETB 受体，主要位于血管内皮细胞中，在血管收缩、血管舒张和细胞增殖中起作用．研究发现，ET-1 在乳腺癌中通过其受体 ET(A)R 和 ET(B)R 以自分泌和旁分泌方式刺激肿瘤细胞生长，ET-1、ET(A)R 和 ET(B)R 的高表达在总体生存率较低的乳腺癌患者中较为常见[42]．研究发现 ET-1、ET(A)R 和 ET(B)R 表达增加和乳腺癌血管分布相关，因此可能参与乳腺癌血管生成的调节，ET 系列的表达模式可能在未来的抗血管生成靶向治疗乳腺癌中具有临床意义[43]．CXCL12(C-X-C motif chemokine ligand 12)，别名 SDF-1，该基因的突变和人类免疫缺陷病毒Ⅰ型感染的抗性有关，外源 SDF-1 可调节细胞的运动、趋化性和黏附，SDF-1 及其受体 CXCR4 可能为预测细胞侵袭提供重要信息，成为人乳腺癌的重要治疗靶点[44]．研究表明，CXCL12 趋化因子能抑制人乳腺癌的生长和转移，在人乳腺癌细胞生长和侵袭中起重要作用，可能是乳腺癌患者的潜在预后标志[45].

2.4 在其他癌症发生中的应用

将乳腺癌特征基因筛选方法应用于其他癌症的分析中，选取 TCGA 数据库中癌旁样本和发病Ⅰ期样本数均大于 40 的癌症，得到肺腺癌、肺鳞状细胞癌、结肠癌、甲状腺癌、肝细胞癌五种癌症，设定置信区间筛选的 i 值为 1.0，得到癌症的特征基因数依次为：586、2162、1442、371、580个，对这些特征基因采用随机森林进行建模，循环 10 次得到的准确率如表 6 所示.

Table 6 Performance of random forest model for five cancers while i=1.0

Cancer (number of signature genes)	*ACC*	*SEN*	*SPE*	*MCC*
Lung adenocarcinoma(586)	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
Lung squamous cell carcinoma(2162)	0.9933 ± 0.0133	0.9875 ± 0.0375	0.9958 ± 0.0125	0.9821 ± 0.0358
Colon adenocarcinoma(1442)	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
Thyroid carcinoma(371)	0.9771 ± 0.0249	0.9600 ± 0.0800	0.9795 ± 0.0225	0.9163 ± 0.0944
Hepatocellular carcinoma(580)	0.9739 ± 0.0288	0.9633 ± 0.0737	0.9768 ± 0.0284	0.9329 ± 0.0748

从表中可以看出采用本文的特征基因筛选方法在多种癌症中准确率均能达到 97%以上，具有较高的准确率，表明本文的特征基因筛选方法普适性较好．将这些特征基因分别做 KEGG 通路分析，结果显示在多种癌症中多次出现的通路有：癌症通路(pathways in cancer)、轴突导向(axon guidance)、细胞黏附分子通路(cell adhesion molecules)．其中轴突导向是形成神经元网络的关键阶段，在神经发育过程中，轴突导向分子引导轴突选择正确生长方向，从而成功到达靶区，近年来关于轴突导向分子的研究进展迅速，许多研究发现其在神经系统、免疫调节、血管生成、肿瘤侵袭转移等方面均起着重要作用[46-47]．细胞黏附分子是介导细胞与细胞、细胞与细胞外基质间相互接触和结合分子的统称，以其受体 - 配体结合的形式发挥作用，使细胞与细胞、细胞与基质，或细胞 - 基质 - 细胞发生黏附，参与细胞的识别、活化和信号转导，细胞的增殖分化，细胞的伸展与移动，是细胞应答、炎性反应、凝血、肿瘤转移以及创伤愈合等一系列重要生理和病理过程的分子基础[48].

3 结 论

利用生物信息学的方法，在全基因组的水平上研究乳腺癌病人从正常到发病Ⅰ期基因表达的变化，从癌症基因组图谱数据库收集乳腺癌数据集，采用统计学中的相关方法组合起来筛选对于分类结果有显著性差异的基因，并对这些基因采用机器学习的方法建模得到准确率均较高，对筛选的 336 个特征基因进行 KEGG 通路分析，得到 10 个重要的基因均和乳腺癌的发生和发展具有重要意义，最后

还将基因筛选方法应用于 TCGA 数据集中其他 5 种癌症的分析，得出的准确率均较高，表明本文方法的普适性较好．筛选得到的特征基因对于揭示乳腺癌发生机理具有重要的意义，也为乳腺癌早期预测及诊断提供一种新的可参考指标．

参 考 文 献

[1] Siegel R L, Miller K D, Jemal A. Cancer Statistics, 2017. Ca A Cancer Journal for Clinicians, 2017, **67**(1): 7−30

[2] 吴 松. 精准医学导论. 广州: 中山大学出版社, 2015: 26−27
Wu S. Introduction to Precision Medicine. Guangzhou: Sun Yat-sen University Press, 2015: 26−27

[3] Bhattacharjee A, Richards W G, Staunton J, *et al*. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA, 2001, **98**(24): 13790−13795

[4] Sotiriou C, Wirapati P, Loi S, *et al*. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. Journal of the National Cancer Institute, 2006, **98**(4): 262−272

[5] Feng Y, Li X, Sun B, *et al*. Evidence for a transcriptional signature of breast cancer. Breast Cancer Research and Treatment, 2010, **122**(1): 65−75

[6] Ma Y, Qian Y, Wei L, *et al*. Population-based molecular prognosis of breast cancer by transcriptional profiling. Clinical Cancer Research, 2007, **13**(7): 2014−2022

[7] Xie J Y, Wang M Z, Zhou Y, *et al*. Coordinating discernibility and independence scores of variables in a 2D space for efficient and accurate feature selection//Proceedings of The ICIC 2016, Lanzhou, China. 2016

[8] Xie J Y, Xie W X. Several feature selection algorithms based on the discernibility of a feature subset and support vector machines. Chinese Journal of Computers, 2014, **37**(8): 1704−1718

[9] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, **27**(8): 1226−1238

[10] Wang R, Tang K. Feature Selection for Maximizing the Area Under the ROC Curve; proceedings of the IEEE 13th International Conference on Data Mining Workshops, USA, F Dec 6-6, 2009, 2009 [C]. IEEE: USA, 2009

[11] Guyon I, Weston J, Barnhill S, *et al*. Gene selection for cancer classification using support vector machines. Machine Learning, 2002, **46**(1): 389−422

[12] Kira K, Rendell L A. The feature selection problem: traditional methods and a new algorithm; proceedings of the Proceedings of the 10th National Conference on Artificial Intelligence, San Jose, California, F July 12-16, 1992, 1992 [C]. AAAI Press: San Jose, California, 1992

[13] Roan F, Bell B D, Stoklasek T A, *et al*. The multiple facets of thymic stromal lymphopoietin (TSLP) during allergic inflammation and beyond. Journal of Leukocyte Biology, 2012, **91**(6): 877−886

[14] Zhang Y, Zhou B. Functions of thymic stromal lymphopoietin in immunity and disease. Immunologic Research, 2012, **52** (3): 211−223

[15] Barooei R, Mahmoudian R A, Abbaszadegan M R, *et al*. Evaluation of thymic stromal lymphopoietin (TSLP) and its correlation with lymphatic metastasis in human gastric cancer. Medical Oncology, 2015, **32**(8): 217

[16] 于海明, 杨俊兰, 李 莹, 等. 乳腺癌原发灶胸腺基质淋巴细胞生成素表达与预后相关. 细胞与分子免疫学杂志, 2015, **31**(2): 239−243
Yu H, Yang J, Li Y, *et al*. Chinese Journal of Cellular and Molecular Immunology, 2015, **31**(2): 239−243

[17] Nakamura Y, Yasuoka H, Tsujimoto M, *et al*. Prognostic significance of vascular endothelial growth factor D in breast carcinoma with long-term follow-up. Clinical Cancer Research: an Official Journal of the American Association for Cancer Research, 2003, **9**(2): 716−721

[18] Akahane M, Akahane T, Matheny S L, *et al*. Vascular endothelial growth factor-D is a survival factor for human breast carcinoma cells. International Journal of Cancer, 2006, **118**(4): 841−849

[19] Teramoto S, Arihiro K, Koseki M, *et al*. Role of vascular endothelial growth factor-C and -D mRNA in breast cancer. Hiroshima Journal of Medical Sciences, 2008, **57**(2): 73−78

[20] Wang J, Guo Y, Wang B, *et al*. Lymphatic microvessel density and vascular endothelial growth factor-C and -D as prognostic factors in breast cancer: a systematic review and meta-analysis of the literature. Molecular Biology Reports, 2012, **39**(12): 11153−11165

[21] Zhao Y C, Ni X J, Li Y, *et al*. Peritumoral lymphangiogenesis induced by vascular endothelial growth factor C and D promotes lymph node metastasis in breast cancer patients. World Journal of Surgical Oncology, 2012, **10**(1): 165

[22] Iwaya K, Ogawa H, Izumi M, *et al*. Stromal expression of CD10 in invasive breast carcinoma: a new predictor of clinical outcome. Virchows Archiv: an International Journal of Pathology, 2002, **440**(6): 589−593

[23] Boberg D R, Batistela M S, Pecharki M, *et al*. Copy number variation in ACHE/EPHB4 (7q22) and in BCHE/MME (3q26) genes in sporadic breast cancer. Chemico-Biological Interactions, 2013, **203**(1): 344−347

[24] Vo T N, Mekata E, Umeda T, *et al*. Prognostic impact of CD10 expression in clinical outcome of invasive breast carcinoma. Breast Cancer, 2015, **22**(2): 117−128

[25] Jana S H, Jha B M, Patel C, *et al*. CD10-a new prognostic stromal marker in breast carcinoma, its utility, limitations and role in breast cancer pathogenesis. Indian Journal of Pathology & Microbiology, 2014, **57**(4): 530−536

[26] Zhao H, Yu H, Martin T A, *et al*. The role of JAM-B in cancer and cancer metastasis (Review). Oncology Reports, 2016, **36**(1): 3−9

[27] Acharyya S, Butchbach M E, Sahenk Z, *et al*. Dystrophin glycoprotein complex dysfunction: a regulatory link between muscular dystrophy and cancer cachexia. Cancer Cell, 2005, **8**(5):

421-432

[28] Wang Y, Marino-Enriquez A, Bennett R R, *et al*. Dystrophin is a tumor suppressor in human cancers with myogenic programs. Nature Genetics, 2014, **46**(6): 601-606

[29] Chitra E, Lin Y W, Davamani F, *et al*. Functional interaction between Env oncogene from Jaagsiekte sheep retrovirus and tumor suppressor Sprouty2. Retrovirology, 2010, **7**(1): 1-18

[30] Williams C M, Engler A J, Slone R D, *et al*. Fibronectin expression modulates mammary epithelial cell proliferation during acinar differentiation. Cancer Research, 2008, **68**(9): 3185-3192

[31] Von Au A, Vasel M, Kraft S, *et al*. Circulating fibronectin controls tumor growth. Neoplasia, 2013, **15**(8): 925-938

[32] Cen Y L, Qi M L, Li H G, *et al*. Associations of polymorphisms in the genes of FGFR2, FGF1, and RBFOX2 with breast cancer risk by estrogen/progesterone receptor status. Molecular Carcinogenesis, 2013, **52**(S1): 52-59

[33] Slattery M L, John E M, Stern M C, *et al*. Associations with growth factor genes (FGF1, FGF2, PDGFB, FGFR2, NRG2, EGF, ERBB2) with breast cancer risk and survival: the Breast Cancer Health Disparities Study. Breast Cancer Res Treat, 2013, **140**(3): 587-601

[34] Zhou X, Liu Y, You J, *et al*. Myosin light-chain kinase contributes to the proliferation and migration of breast cancer cells through cross-talk with activated ERK1/2. Cancer Letters, 2008, **270** (2): 312-327

[35] Cui W J, Liu Y, Zhou X L, *et al*. Myosin light chain kinase is responsible for high proliferative ability of breast cancer cells *via* anti-apoptosis involving p38 pathway. Acta Pharmacologica Sinica, 2010, **31**(6): 725-732

[36] Fernandez Val J F, Losada J, Arregui Murua M A, *et al*. Cell proliferation, nuclear ploidy, and EGFr and HER2/neu tyrosine kinase oncoproteins in infiltrating ductal breast carcinoma. Cancer Genetics and Cytogenetics, 2002, **138**(1): 69-72

[37] Golubovskaya V, Beviglia L, Xu L H, *et al*. Dual inhibition of focal adhesion kinase and epidermal growth factor receptor pathways cooperatively induces death receptor-mediated apoptosis in human breast cancer cells. The Journal of Biological Chemistry, 2002, **277**(41): 38978-38987

[38] Ray A, Naik S L, Sharma B K. Distribution of prognostically unfavourable product of c-erbB-2 oncogene and EGF-R in carcinomas of the breast and uterine cervix. Indian Journal of Physiology and Pharmacology, 2002, **46**(4): 423-433

[39] Tong L, Yang X X, Liu M F, *et al*. Mutational analysis of key EGFR pathway genes in Chinese breast cancer patients. Asian Pacific Journal of Cancer Prevention: APJCP, 2012, **13** (11): 5599-5603

[40] Brandt B, Hermann S, Straif K, *et al*. Modification of breast cancer risk in young women by a polymorphic sequence in the egfr gene. Cancer Research, 2004, **64**(1): 7-12

[41] Wen Y Y, Zheng J N, Pei D S. An oncogenic kinase: putting PAK5 forward. Expert Opinion on Therapeutic Targets, 2014, **18** (7): 807-815

[42] Wulfing P, Diallo R, Kersting C, *et al*. Expression of endothelin-1, endothelin-A, and endothelin-B receptor in human breast cancer and correlation with long-term follow-up. Clinical Cancer Research: an Official Journal of the American Association for Cancer Research, 2003, **9**(11): 4125-4131

[43] Wulfing P, Kersting C, Tio J, *et al*. Endothelin-1-, endothelin-A-, and endothelin-B-receptor expression is correlated with vascular endothelial growth factor expression and angiogenesis in breast cancer. Clinical Cancer Research: an Official Journal of the American Association for Cancer Research, 2004, **10** (7): 2393-2400

[44] Kang H, Mansel R E, Jiang W G. Genetic manipulation of stromal cell-derived factor-1 attests the pivotal role of the autocrine SDF-1-CXCR4 pathway in the aggressiveness of breast cancer cells. International Journal of Oncology, 2005, **26**(5): 1429-1434

[45] Lv Z D, Kong B, Liu X P, *et al*. CXCL12 chemokine expression suppresses human breast cancer growth and metastasis *in vitro* and *in vivo*. International Journal of Clinical and Experimental Pathology, 2014, **7**(10): 6671-6678

[46] Eissler N, Rolny C. The role of immune semaphorins in cancer progression. Experimental Cell Research, 2013, **319** (11): 1635-1643

[47] Rehman M, Tamagnone L. Semaphorins in cancer: biological mechanisms and therapeutic approaches. Seminars in Cell & Developmental Biology, 2013, **24**(3): 179-189

[48] 刘凤茹, 侯振江, 王秀文. 细胞黏附分子. 检验医学与临床, 2007, **4**(8): 748-750
Liu F R, Hou Z J, Wang X W. Laboratory Medicine and Clinic, 2007, **4**(8): 748-750

Signature Genes Identification of The Breast Cancer Occurrence and Pattern Recognition*

WEN Jian-Xin, WANG Xue-Dong, LI Xiao-Qin**, CHANG Yu

(*School of Life Science and Bioengineering, Beijing University of Technology, Beijing* 100124, China)

Abstract To identify signature genes for the pathogenesis of breast cancer, which provides a theoretical support for prevention and early diagnosis of breast cancer. The pattern recognition method was used to analysis the genome-wide gene expression data which was collected from the breast cancer part of TCGA (The Cancer Genome Atlas) database.336 gene expression signature genes were selected by means of a combination of statistical methods such as correlation, t test, confidence interval, etc. The accuracy can be as high as 98% through the machine learning method modeling, which is higher compared with the previous study. The KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis and GO (Gene Ontology) enrichment analysis indicated the significant correlation among eight and eighteen kinds of genes respectively. A functional analysis of the part of the eight pathways showed theirs close relationship at the level of gene regulation which indicted the identified signature genes play an important role in the pathogenesis of breast cancer and is very important for understanding the pathogenesis of breast cancer and the early diagnosis of breast cancer.

Key words breast cancer, gene expression, pattern recognition, tumor prediction, early diagnosis

DOI: 10.16476/j.pibb.2017.0221

*This work was supported by grants from The National Natural Science Foundation of China (11572014) and Major Research Projects in The Field of Intelligent Manufacturing (01500054631751).

**Corresponding author.

Tel: 86-15313254516, E-mail: lxq0811@ bjut.edu.cn

Received: August 2, 2017 Accepted: September 8, 2017