第35卷 第3期 北京工业大学学报 Vol.35 No.3
2009年 3月 JOURNAL OF BEIJING UNIVERSITY OF TECHNOLOGY Mar. 2009

基于基因表达谱的白血病分子预测模型研究

王金莲[1,2], 阮晓钢[1], 李晓明[1,3]
(1. 北京工业大学 电子信息与控制工程学院, 北京 100124; 2. 首都医科大学 生物医学工程学院, 北京 100069;
3. 廊坊师范学院, 廊坊 102800)

摘 要: 采用生物信息学方法对肿瘤基因表达数据进行挖掘,以获取和肿瘤不同亚型相关的候选标志基因集合,应用机器学习方法从标志基因集合中提取出甄别肿瘤不同亚型的规则集,进而建立起肿瘤预测模型. 利用Relief、信息增益和分类信息指数从不同角度挖掘蕴含在基因表达谱中的候选特征基因,抽取出候选特征基因公约集合. 以对不同肿瘤组织样本的识别能力为依据,选取分类能力最强的一组基因集合作为特征基因. 利用规则判定树提取出反映这些特征基因相互作用的规则集并以此构建肿瘤预测模型,并将此模型应用于白血病基因表达数据中,建立了白血病分子预测模型. 研究表明,该模型得到的白血病标志基因对肿瘤临床诊断具有一定的参考价值.

关键词: 肿瘤; 基因表达谱; 决策树; 支持向量机
中图分类号: TP 18 **文献标识码**: A **文章编号**: 0254 - 0037(2009)03 - 0301 - 08

随着基因芯片技术的出现,肿瘤分类已进入了分子分类阶段,Golub 等[1]以有权表决法作为分类手段,就急性骨髓性白血病(AML)与急性成淋巴细胞白血病(ALL)的识别问题进行了研究. 肿瘤基因表达谱的主要目的是通过分析特定的基因表达模式与肿瘤临床生物学行为之间的关系,以阐明肿瘤发生发展的分子机制,预测肿瘤临床生物学行为,如转移潜能[2]、药物敏感性[3]、预后判断[4]、肿瘤分子分型[5]、分子诊断标志和药物作用靶点等等;Scherf 等[6]基于基因表达谱,采用聚类算法分析了基因表达模式同118种药物间的联系,并给出某些基因在药物作用后所表现出的敏感性和抗药性;Alizaden 等[7]利用 cDNA 芯片所测定的患有弥漫性大 B 淋巴瘤(DLBCL)的 38 位患者的基因表达谱数据对这些患者化疗后的临床预后进行了研究,发现具有不同亚型的患者在采用同一类型的化疗药物进行治疗时有着明显的临床结果,而且这 2 种亚型的弥漫性大 B 细胞淋巴瘤经治疗后的 5 a 生存率差异较大.

急性白血病(acute leukemia)是造血系统的恶性肿瘤,临床上急性白血病可分为 ALL 和 AML. 利用基因表达谱在分子水平上对肿瘤亚型进行识别成为可能[8].

本文旨在提供一种有效的识别肿瘤不同亚型的分子诊断模型,该模型以分类标志基因为基础,考查这些基因在样本中的表达模式和逻辑规则,从而形成肿瘤预测模型规则集,考虑到模型临床的应用价值和实用性,本模型以每个基因在样本中的表达量作为预测肿瘤亚型的依据,给出基因表达的阈值.

1 实验数据及实验原理

1.1 实验数据

本文采用白血病数据集[9],Golub 等人[1]用 DNA 芯片检测了急性白血病的基因表达谱数据, 数据集共包含 72 个急性白血病样本,其中 47 个为 ALL,25 个为 AML. 每个样本均含 7129 个人类基因在 Affymetrix 芯片上的表达真值.

收稿日期: 2008-01-03.
基金项目: 国家自然科学基金重点资助项目(60234020).
作者简介: 王金莲(1969 -), 女, 陕西千阳人, 博士生, 讲师.

1.2 实验原理和欲解决的关键问题

本文利用 Relief[10]、信息增益(IGI)[11]和分类信息指数[12](CII)算法提取特征基因公约集合. 以SVM作为分类器检验集合的样本分类正确率,以分类正确率为标准对基因集合用交叉验证的方法筛选候选特征基因,选取分类能力最强的一组基因集合作为特征基因. 实验原理如图1所示.

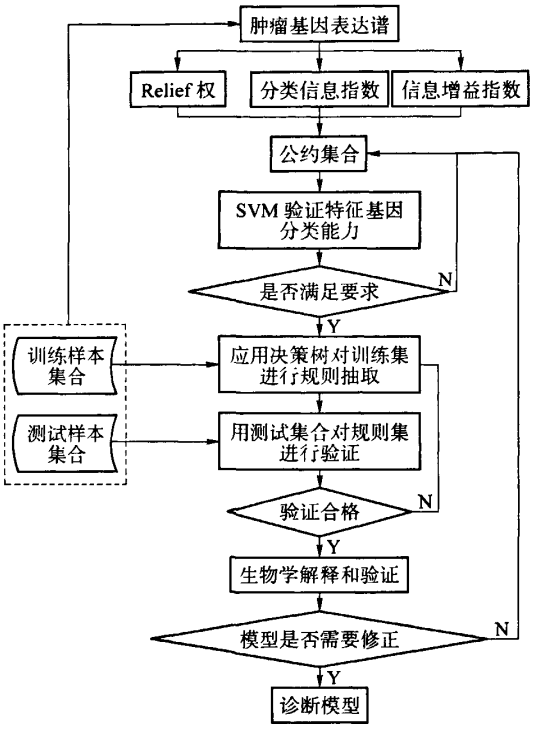

图1 肿瘤预测模型实验原理

Fig.1 The flow chart of the cancer prediction model

特征选择算法可以分为 Filter 和 Wrapper 两大类[13]. Wrapper 类方法直接利用后续学习算法的训练准确率评估特征子集,因此其评估和后续算法性能偏差小,但计算量很大,不适合大数据集;Filter 方法和后续学习算法无关,一般直接利用所有训练数据的统计性能评估特征,如一致度、信息增益以及 Relief 评估等,因此 Filter 方法具有速度快的优点,但评估和后续学习算法的性能有较大偏差.

规则集的抽取大多是利用机器学习方法产生的,常用的分类规则集抽取通常有贝叶斯、决策树、人工神经网络、约略集方法和遗传算法等等. 决策树(decision tree)是一种基于示例的归纳学习算法. 该算法采用自顶向下的递归方式,在决策树的内部节点进行属性值的比较并跟据不同的属性值判断从该结点向下的分支,在决策树的叶结点得到结论[14],本文利用决策树算法对肿瘤样本分类知识进行提取.

2 实验步骤和结果

2.1 公约集合的提取

分别采用 Relief 算法、信息增益和分类信息指数对7 129个肿瘤差异表达基因所包含的分类信息进行度量,分别生成 Relief、信息增益和分类信息指数候选特征基因集合. 这3个集合的交集为候选特征基因的公约集合.

$$\{A,D,E\}=\{A,B,D,E\cdots X\}\cap\{A,C,D,E,F\cdots M\}\cap\{A,D,E\cdots Q\}$$

公约集合　　Relief 集合　　信息增益集合　　分类信息集合

式中,A、B、C、D 表示候选特征基因.

2.1.1 分类信息指数提取候选特征基因

分类信息指数[11]是基于 Golub 的"信噪比"提出来的

$$d(g)=\frac{1}{2}\frac{|\mu_{g+}-\mu_{g-}|}{\sigma_{g+}+\sigma_{g-}}+\frac{1}{2}\ln\left(\frac{\sigma_{g+}^2+\sigma_{g-}^2}{2\sigma_{g+}\sigma_{g-}}\right) \tag{1}$$

式中,u_{g+}、u_{g-} 为基因 g 在肿瘤样本中表达水平的均值;σ_{g+}、σ_{g-} 为基因 g 在正常样本中表达水平的标准差. 基因的分类信息指数由2部分构成:第1部分为 Golub[1]等定义的"信噪比"指标;第2部分为水平分布方差的不同对样本分类的贡献. 依据式(1)在样本集上计算每个基因的分类信息指数,分类信息指数散

点图如图 2 所示.

图 2 中多数基因的分类信息指数小于 0.5，说明这些基因在 ALL 和 AML 中 2 个类别中的表达水平无论是均值还是方差均无明显差异，作为与样本类别不相关的"噪声基因"而存在.

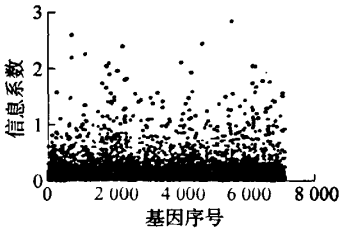

图 2　分类信息指数散点图

Fig.2　The scatter plot of classification information index

2.1.2　信息增益指数提取候选特征基因

从机器学习的角度分析，一个属性的信息增益是指使用了这个属性分割样本而导致的期望熵的降低. 一个属性 A 相对训练样本集合 S 的信息增益[11]

$$G(S,A)=E(S)-\sum_{v\in A}\frac{|S_v|}{|S|}E(S_v) \tag{2}$$

$$E(S)=\sum_{i=1}^{m}-p_i\log_2 P_i$$

式中，$E(S)$为样本集 S 的信息熵，它刻画了样本集的纯度；P_i 为 S 中属于类别 i 的样本比例.

以基因表达谱的连续值作为输入，采用智能阈值选取算法计算每个基因的信息增益指数，具体分为 4 步.

第 1 步　输入基因表达谱的 EXCEL 文件格式，行为基因，列为样本. 最后一行用 0、1 分别标注每个样本的类别；

第 2 步　以每个样本中的基因在样本中的表达值作为表达的阈值；

第 3 步　如果样本中基因表达值残缺，则取该基因在所有样本中的平均值代替；

第 4 步　利用式(2)，计算各个基因的信息增益，并依据基因的信息增益用冒泡排序法，按照从小到大的顺序将所有基因排序后放入有序集合 $\varphi=\{g_1,g_2\cdots g_{7129}\}$.

按照以上算法计算每个基因的信息增益散点图如图 3 所示.

大多数基因的信息增益指数分布在 0.8～0.9，占总基因数的 82.58%，指数分布在 0.7 以下的基因占总基因的 5.35%，只有 5% 的基因其信息增益指数比较小，因此这些基因将被选作更进一步分析的候选基因.

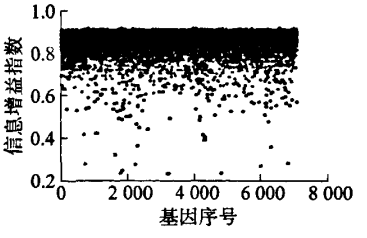

图 3　信息增益散点图

Fig.3　The scatter plot of information gain index

2.1.3　Relief 提取候选特征基因

Relief 算法对属性可分性评价的过程是基于学习样本集中的每个样本(示例)的一个机器学习过程. 对训练集中任一学习样本 S，该算法在训练集中搜索出和该样本最为接近的 $K(K>1)$个同类别样本(称为：H)及 K 个异类样本(称为：M). 对于属性 A_i 而言，样本 S 与异类样本在该属性上的差别越大，而与同类别样本在该属性上的差别越小，则样本在 A_i 属性上的类别可分性就越大，A_i 具有的分类权重也就越大. Relief 算法的具体描述见算法 1.

算法 1：Relief Algorithm(F)　　{F 为待分析的属性集合}

1) Set weights vector W to zeros.

{向量 W 中第 i 个元素对应于 F 中第 i 个属性的分类权重}

2) For $i=1$ to Num do　　{Num 为训练集的样本数}

a) choose ith instance x　　{选择训练集中第 i 个样本}

b) Find its nearest K Hits and nearest K Misses

c) For $j=1$ to card(F)　　{对 F 中的每个属性计算分类权重}

$$W_j=W_j-\frac{\sum_{m=1}^{k}(x_j-H_{mj})^2}{k}+\frac{\sum_{m=1}^{k}(x_j-M_{mj})^2}{k}$$

3) Return W

算法1中采用Mahalanobis平方距离作为Relief算法中对基因子集类别可分性判据,即其分类能力的评价函数

$$J_M(F)=(\boldsymbol{\mu}_{F+}-\boldsymbol{\mu}_{F-})^T\boldsymbol{\Sigma}_F^{-1}(\boldsymbol{\mu}_{F+}-\boldsymbol{\mu}_{F-}) \tag{4}$$

$$\boldsymbol{\Sigma}_F=\frac{\boldsymbol{\Sigma}_{F+}+\boldsymbol{\Sigma}_{F-}}{2} \tag{5}$$

式中,$\boldsymbol{\mu}_{F+}$、$\boldsymbol{\mu}_{F-}$分别为基因子集F中的基因在正常和肿瘤2个类别中表达水平的均值向量,$\boldsymbol{\Sigma}_{F+}$、$\boldsymbol{\Sigma}_{F-}$分别为这些基因在2个类别中表达水平的协方差矩阵.

用Relief算法得到的每个基因的分类权重记为$W=\{W_1, W_2\cdots W_{7\,129}\}$,每个基因的分类权重散点如图4所示.

大多数基因的Relief权重小于100,只有少数基因的权重比较大,说明大多数基因在ALL和AML样本中的差异非常小,不能作为划分这2种样本的特征基因.

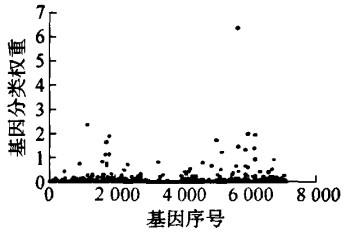

图4 Relief所确定的特征基因分类权重的散点图

Fig.4 The scatter plot of Relief weight

2.1.4 特征基因公约集合确定

用3种方法获得了3个候选特征基因集合分别记为C、R和G. 对C和R集合中的基因按降序对其值排序,对G按升序排序,然后取每个集合前5%的基因进行两两相交得到3个新的子集,分别包含127、257和257个基因,3个集合的交集为93.

2.2 特征基因提取及其样本分类能力检验

若给定样本集的形式$S_T=\{(x_i,y_i)\mid x_i\in R^d, y_i\in\{-1,+1\}, i=1,2\cdots N\}$,则SVM[15]的判别函数为

$$g(x)=\sum_{i\in S}a_iy_ik(x_i,x) \tag{6}$$

式中,S为支持向量的个数;$K(x,x_i)$为核函数;a_i为拉格朗日乘子.

分类器SVM的参数选择问题目前理论上尚未解决,只能通过试验的方法进行选取. 最终选择了径向基核函数

$$K(\boldsymbol{x},\boldsymbol{y})=e^{\frac{\|x-y\|^2}{2\sigma^2}} \tag{7}$$

通过多次实验选取$\sigma=10$,SVM上界控制因子$C=10\,000$.

用SVM分类器对93个公约集合基因分类能力检验的分为4步.

1) 将肿瘤及其对应的正常组织样本均按近似2:1的比例分配在训练集和测试集中,训练集包括27个ALL和11个AML样本,测试集包括20个ALL和14个AML样本,具体分配情况如图5所示.

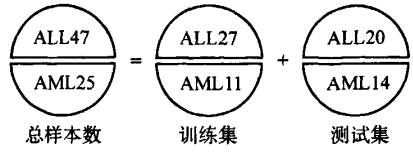

图5 样本划分

Fig.5 Partition of sample in the original dataset

2) 训练集中的样本首先被用作特征选取学习算法的学习样本,学习算法在训练集上对基因表达水平与样本类别间的关系进行学习,从而得到决定样本类别特征的分类特征基因.

3) 训练好的分类器作为区分肿瘤组织样本与正常组织样本的分类模型,对测试集中的样本进行类别预测,以检验求得的分类特征基因是否具有对样本类别良好的判别能力.

4) 为了获得对分类错误率的可靠估计并同已有的研究成果进行比较,本文采用"留一交叉检验法"(Leave-One-Out Cross Validation, LOOCV)在训练集上对样本类型进行识别. 记录所有被错误分类的样本数作为留一法的分类错误数.

利用 SVM 模型对在不同属性集合中的基因作为样本分类特征的情况下，对测试集中样本的分类情况如图 6 所示．纵轴表示进行样本分类测试时被分类器错误分类的样本数，横轴表示不同的属性集合中的基因数．

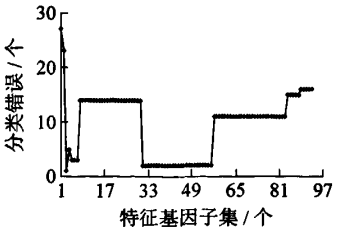

图 6　基因集合样本分类能力曲线

Fig.6　Classification performance of different feature gene sets

对测试集共 34 个测试样本测试的结果如图 6 所示，当基因集合中的基因个数从 56 减少到 31 时错分数为 2 个，这 2 个数时候或者之前错分数增加．因此以这 31 个基因作为特征基因集合．本文确定的 31 个分类特征基因相比 Golub 选取的 50 个基因，仅有 5 个基因重合，它们是：CD33；MYL1；RABAPTIN-5 protein；MB-1 和 TCF3．

3　预测模型规则集抽取和评价

为了确保得到分类能力强的基因也就是对肿瘤发生发展过程具有重大影响的基因，本文采用了信息反馈机制，通过 SVM 方法反复训练淘汰分类效果差的基因，重新优化后选取新的公约集合原理如图 1 所示．预测模型的规则集抽取采用决策树．决策树采用自上而下的递归方法构造，建树分为建树（Tree Building）和剪枝（Tree Pruning）2 个步骤．具体算法见算法．

首先把 72 个样本集合随机的划分成 5 组相同大小的样本子集，每个子集由 9 个 ALL 和 5 个 AML 样本组成且各个子集之间没有重复基因，以其中任意的 4 组子集作为训练集，剩余的一组作为测试样本对该决策树的分类性能进行检验，为了保证每个样本都有一次机会作为测试样本，将 4 个训练集和 1 个测试集的每种组合作为决策树产生和规则集提取的训练和测试集合，因此共有 120 种训练集和测试集的组合方式．然后对每组训练集利用决策树算法进行训练，记录每次测试的错误率，终将产生 120 棵决策树．如此交叉验证后最终得到一个“决策树群”．

算法

1) Set Root node

2) If Examples are normal, return Root(返回单节点树)

3) If Examples are cancer, return Root(返回单节点树)

4) If Attributes is empty, return Root(返回规则集)

5) Else

6) Choose A from Attributes(A 属性集合中分类能力最好的属性)

7) Root classified by A

8) For each attribute (i) in attributes

For each attributes(j) in examples(遍历每个样本中每个基因属性的值)

Set $A = v_{ij}$(分别以样本中的每个基因的表达值作为分类阈值)

令 Example-v_i 为样本中满足 A 属性值为 v_i 的子集

If info gain is under the threshold Then Tree Pruning(如果分类信息增益不满足阈值，就剪掉这个规则分枝，info gain 根据公式(6)得到)

If Examples-vi is empty(如果基因属性集合为空)

Add leaf on the parent node(在父节点下加一个叶子节点名称)

Else

Add new branch under parent node;（在父节点下加一个新的分枝）

9) End

10) Return root(返回预测模型的规则集)

对每棵树的评价指标定义为

$$S=1\Big/\left(e_t+\frac{n_t}{f_t}\right) \tag{8}$$

式中，e_t为决策树在交叉试验中对不同测试集的样本分类错误数均值，该指标反映样本的分类能力；n_t为决策树的节点也即基因个数，用于描述描述决策树的复杂度；f_t为交叉检验过程中该数出现的次数．一棵决策树的样本分类能力越强、复杂度越小且出现的次数越多则这棵树的性能就越好，其 S 值就越大．

利用算法 2 得到 9 棵不同的决策树，表 1 给出了这 9 棵树的样本错分数、所含基因数及其性能指标．

表 1 对决策树群中决策树的评价结果

Table 1 Evaluation of decision trees in "Treeset"

树标号	所含基因名	节点数	错分数	频率	S
t_1	CD33	1	1	8	0.88
t_2	CD33, NPY, KIAA0212	3	0	11	3.666
t_3	KIAA0212	1	0.87	1	0.535
t_4	CD33, KIAA0159, KIAA0212, NPY, MPO	5	0	21	4.5
t_5	CD33, KIAA0159, KIAA0212, NPY, SCYA5, MXS1, MPO	7	0.93	11	0.638
t_6	CD33, NPY	2	0	8	4
t_7	CD33, NPY, KIAA0159, KIAA0212, MXS1	5	0.72	20	1.03
t_8	CD33, NPY, KIAA0159, KIAA0212, MXS1, SCYA5, MPO	7	0.54	2	0.249
t_9	TCF3, NPY, KIAA0159, KIAA0212, MXS1, SCYA5, MPO	7	0.333	1	0.136

由表 1 知，决策树 t_4 具有最大的性能指标，此树共涉及 5 个关键分类特征基因，分别是 CD33, KIAA0159, KIAA0102, NPY, MPO．t_4 在其中一组测试样本中的决策形式如图 7 所示．

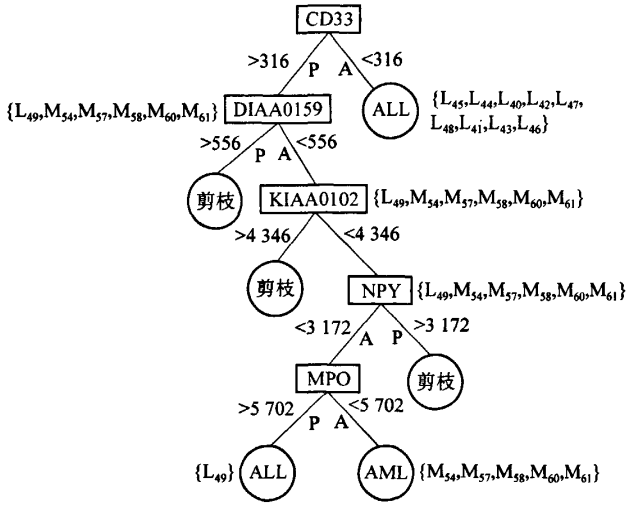

图 7 具有最大预测能力的决策树 42

Fig.7 Decision tree 42 with maximum performance index

图 7 中 P 为基因上调表达，A 为基因下调表达，树枝上的数值是基因区分 2 类样本 ALL 和 ANL 的阈值．$\{L_n, M_n, n=1.23\cdots72\}$为测试样本集合，$L$ 为急性成淋巴细胞白血病样本(ALL)，M 为急性骨髓性白血病样本(AML)，n 为测试样本的个数，斜体为错分的样本，黑体为被划分正确的样本．剪枝表示没有产生分类规则集．

利用 t_4 的5个关键基因在组织样本中的表达模式可形成2组共3条样本分类规则集. 这3条规则集在交叉检验实验中对样本的分类错误率为0，可实现对不同测试集中样本的准确分类. 决策树学习系统以产生式规则对 t_4 进行描述

规则集

IF CD33＜316 THEN 成淋巴细胞白血病(ALL)　　Rule 1

IF CD33＞316 THEN

IF KIAA0159＜556 THEN

IF KIAA0102＜4346 THEN

IF NPY＜3172 THEN

IF MPO＞5702 THEN 成淋巴细胞白血病(ALL)　　Rule 2

IF MPO＜5702 THEN 急性骨髓性白血病(AML)　　Rule 3

本文方法的特征基因为5个，分类正确率100%；Golub方法的特征基因为50个，分类正确率94%. 利用上述规则集中的5个关键基因在组织样本中的表达量准确描述了急性成淋巴细胞白血病和基因骨髓性白血病样本在基因表达模式上的差异.

4 结束语

CD33编码单克隆抗体表面糖蛋白－抗原CD33，是当前用于白血病亚型分型重要的免疫标记[16-17]. MPO(髓过氧化物酶)是一种在髓细胞分化中合成的血红素蛋白，形成一种单链前体，髓过氧化酶随后分裂成一个轻链和一个重链. MPO参与的生物学过程包括抗凋亡，防御应答，过氧化氢分解代谢过程和氧化反应应答.

CD33、MPO是目前已知的白血病免疫分型中常用的免疫标记[16-19]. 抗原CD33和MPO(髓过氧化物酶)则是髓系白血病的2种重要的免疫标记. 从基因表达的角度看，免疫标记对应基因在样本中的表达情况就应当是样本类别信息的直接反映，可称为样本分型的“标记基因”.

参考文献：

[1] GOLUB T R, SLONIM D K, TAMAYO P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286(5439): 531-537.

[2] KANG H C, KIN I J, PARK J H, et al. Identification of genes with different expression in acquired drug-resistant gastric cancer cells using high-density oligonucleotide microarrays[J]. Clinical Cancer Res, 2004, 10(1pt 1): 272-284.

[3] CHEN X, LEUNG S Y, YUEN S T, et al. Variation in gene expression patterns in human gastric cancers[J]. Molecular biology of the cell, 2003, 14(8): 3208-3215.

[4] SAKAKURA C, HAGIWARA A, NAKANISHI M, et al. Different gene expression profiles of gastric cancer cells established from primary tumor and malignant asctites[J]. Br. J. Cancer, 2002, 87(10): 1153-61.

[5] LEUNG S Y, CHEN X, CHU K M, et al. Phospholipase A2 group IIA expression in gastric adenocarcinoma is associated with prolonged survival and less frequent metastasis[J]. PNAS, 2002, 99(27): 16203-16208.

[6] SCHERF U, ROSS D T, WALTHAM M, et al. A gene expression database for the molecular pharmacology of cancer[J]. Nature Genetics, 2000, 24(3): 236-244.

[7] ALIZADEN A A, EISEN M B, DAVIS R E, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling[J]. Nature, 2000, 403(6769): 503-511.

[8] ALON U, BRAKAI N, NOTTERMAN D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays[J]. PNAS, 1999, 96(12): 6745-6750.

[9] INSTITUTE W. Cancer Program Data Sets[DB/OL]. [1999-08-17]. http://www.broad.mit.edu/cgi-bin/cancer/datasets.cg.

[10] KONONENKO I. Estimating attributes: analysis and extensions of Relief. F. Bergadano, L.D.Raedt(eds.). Proceedings of European Conference on Machine Learning[C]. Catania, Springer-Verlag, 1994: 171-182.

[11] Mitchell T M. Machine learning[M]. 曾华军, 张银奎译. 北京: 机械工业出版社, 2003: 15-43.

[12] RUAN Xiao-gang, LI Ying-xin, LI Jian-geng, et al. Study on tumor-specific gene expression patterns[J]. Science of China 2006, 36(1): 89-96.

[13] KOHAVI R, JOHN G. Wrappers for feature subset selection[J]. Artificial Intelligence, 1997, 97: 273-324.

[14] QUINLAN J R. See5.0: Rule Quest Research Data Mining Tools[CP]. [2002-10-20]. http://www.relequest.com.

[15] VAPNIK V N. The nature of statistical learning theory[M]. Berlin, Springer-Verlag, 1994: 22-23.

[16] WANG Ying, XU Shi-rong, LIN Feng-ru, et al. Expressions of Cyclin E2 and Survivin in acute leukemia and their correlation[J]. Journal of Experimental Hematology, 2006, 2: 271-275.

[17] 江雪杰, 王季石, 方琴. 成人急性淋巴细胞白血病中乳腺癌耐药蛋白基因表达及其临床意义[J]. 中国实验血液学杂志, 2008, 1: 31-34.
JIANG Xue-jie, WANG Ji-shi, FANG Qin. Gene expression of breast cancer resistance protein in adult acute lymphocytic leukemia and its clinical significance[J]. Journal of Experimental Hematology, 2008, 1: 31-34. (in Chinese)

[18] 张艳, 江滨, 黄晓军, 等. 多发性骨髓瘤的细胞遗传学研究[J]. 中国实验血液学杂志, 2007, 1: 76-78.
ZHANG Yan, JIANG Bin, HUANG Xiao-jun, et al. Cytogenetic study of multiple myeloma[J]. Journal of Experimental Hematology, 2007, 1: 76-78. (in Chinese)

Study on Leukemia Molecular Prediction Model with Gene Expression Profile

WANG Jin-lian[1,2], RUAN Xiao-gang[1], LI Xiao-ming[3]

(1. College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China;
2. College of Biology Medical Engineering, Capital Medical University, Beijing 100096, China;
3. Lang Fang Normal University, Lang Fang 102800, China)

Abstract: A leukemia molecular prediction model is constructed by using bioinformatics and machine learning methods with gene expression profile. Firstly, three methods including relief, classification information index and information gain index are used to select candidate feature gene set from the leukemia gene expression profile. Secondly, intersection of three candidate feature gene sets is generated, and then the best classification performance of intersection genes which is tested by SVM is selected as feature genes. Thirdly, the classification rule sets are extracted from these feature genes by using decision tree method. Finally, the leukemia molecular prediction model is constructed with these classification rules. The results show that the model is helpful to cancer clinical diagnosis and cancer gene biological experiments. Also, the two key genes (CD33, MPO) are biomarkers of leukemia clinically.

Key words: tumor; gene expression profile; decision tree; support vector machine

(责任编辑 张士瑛)