

## 急性白血病的基因表达谱分析与亚型分类特征的鉴别

李颖新<sup>1</sup> 刘全金<sup>2</sup> 阮晓钢<sup>1\*</sup>

<sup>1</sup>(北京工业大学电子信息与控制工程学院,北京 100022)

<sup>2</sup>(安庆师范学院物理系,安庆 246011)

**摘要:** 本研究基于生物信息学理论,运用模式识别方法和计算技术,对急性白血病的基因表达谱数据进行分析,研究急性白血病的亚型识别与分类信息基因选取问题。首先去除无关基因,然后利用浮动顺序搜索算法搜索特征空间生成候选特征子集,最后以支持向量机作为分类器进行急性白血病的亚型识别,并以误识率为依据鉴别出了 5 个包含完整分类信息的基因。实验结果表明,本研究鉴别出的 5 个信息基因能以 100 % 的正确率准确识别急性白血病亚型。

**关键词:** 急性白血病; 生物信息学; 信息基因; 基因表达谱; 支持向量机

### Analysis of Leukemia Gene Expression Profiles and Subtype Informative Genes Identification

LI Ying-Xin<sup>1</sup> LIU Quan-Jin<sup>2</sup> RUAN Xiao-Gang<sup>1</sup>

<sup>1</sup>(School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100022)

<sup>2</sup>(Department of Physics, Anqing Teacher's College, Anqing 246011)

**Abstract:** In this paper, we analyzed the leukemia gene expression profiles based on the method of bioinformatics, and focused our attention on the leukemia molecular classification and informative genes identification. After having removed the irrelevant genes to the classification task, we employed a suboptimal search method to generate candidate feature subsets for classification, and then each feature subset was applied to support vector machine to classify the samples by "Leave-One-Out Cross Validation" process and independent test. We chose the genes in the feature subset with minimum errors as the informative genes for distinguishing the two classes of samples. The results showed that all the samples were able to be correctly classified with informative genes, and a comparison of the results between this paper and some previous studies was also presented.

**Key words:** leukemia; bioinformatics; informative genes; gene expression profile; support vector machine

中图分类号 TP18 Q617 文献标识码 A 文章编号 0258-8021(2005)02-0240-05

## 引言

随着大规模基因表达谱(Gene expression profile, 或称为基因表达分布图)技术的发展,基因表达数据的分析与建模已经成为生物信息学研究领域中的重要课题。基于基因表达谱,在分子水平上准确地进行肿瘤亚型的识别,对诊断和治疗肿瘤具有重要意义<sup>[1-3]</sup>。在 DNA 芯片上测量的数千个基因中,分析出决定样本类别的一组基因“标签”,即“信息基因”

(informative genes)是正确识别肿瘤类型、给出可靠诊断和简化实验分析的关键所在。

目前人们对肿瘤亚型识别与信息基因选取问题已进行了一定程度上的探索<sup>[4-7]</sup>。1999 年《Science》发表了 Golub 等针对急性白血病亚型识别与信息基因选取问题的研究成果<sup>[4]</sup>。Golub 等以“信噪比”(Signal to noise ratio)指标作为衡量基因对样本分类贡献大小的量度,采用加权投票的方法进行亚型的识别,并从 7 129 个基因中选出了 50 个可能的与亚

收稿日期: 2003-12-22, 修回日期: 2004-08-16。

基金项目: 国家自然科学基金重点资助项目(No. 60234020)。

\*通讯作者

型分类相关的信息基因。Golub 的工作大大缩小了决定急性白血病亚型差异的基因范围,给出了亚型识别的基因依据。此后,Tibshirani 和 Guyon 等进一步推进了对该问题的研究:Tibshirani 等利用收缩质心法选出了 21 个可能的信息基因,并进一步提高了识别正确率<sup>[6]</sup>;Guyon 等则利用支持向量机选出了 8 个可能的信息基因,识别率达到了 100 %<sup>[7]</sup>。Tibshirani 和 Guyon 的研究进一步缩小了决定急性白血病亚型分类的信息基因的范围。然而,仍存在鉴别出更加精简有效的一组信息基因的可能。

本研究基于生物信息学的理论,运用模式识别方法和计算技术,对急性白血病的基因表达谱数据进行分析,研究急性白血病的亚型识别与信息基因选取问题。首先去除无关基因,然后利用浮动顺序搜索算法搜索特征空间生成候选特征子集,最后以支持向量机作为分类工具进行急性白血病亚型的识别,并以误识率为依据鉴别出了 5 个包含完整分类信息的基因作为样本分类信息基因。实验结果表明本研究鉴别出的 5 个信息基因能以 100 % 的正确率准确识别急性白血病亚型。

1 问题描述

Golub 等分析的急性白血病基因表达谱数据集共含有 72 个急性白血病样本,每个样本均含 7 129 个基因的表达数据。其中 47 个样本被诊断为急性成淋巴细胞白血病 (acute lymphoblastic leukemia, ALL), 25 个被诊断为急性骨髓性白血病 (acute myeloid leukemia, AML)。整个数据集被划分为训练集与独立测试集,见图 1。

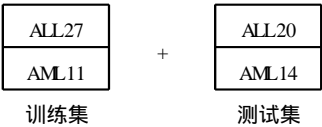


图 1 急性白血病数据集

样本集的数据组织方式如表 1 所示。第一行为样本标号(如:S1,S2,...);第二行是样本所属类别;第一列为基因标号(如:AFFX-BioB-5.at,...,Z78285.at。共 7 129 个)。表中的数值即为基因在样本中的表达水平值。

本研究的目的是:基于 Golub 等人公布的急性白血病的基因表达谱数据,运用计算方法和模式识别技术,研究其中包含的样本分类信息,进行急性白血病的亚型识别,搜索出更加有效的“信息基因组合”,从而给出识别急性白血病亚型的基因依据。

表 1 训练样本集的数据组织形式

	S1	S2	S3	S4	.....	S37	S38
	ALL	ALL	ALL	ALL	.....	AML	AML
AFFX-BioB-5.at	- 241	- 139	- 76	- 135	.....	- 25	- 72
					.....		
Z78285.at	- 37	- 14	- 41	- 91	.....	- 33	0

2 基因表达谱中无关基因的剔除

在基因表达谱中,一些基因的表达水平在所有样本中都非常接近,在 ALL 和 AML 两个类别中的分布无论其均值还是方差均无明显差别,这些基因与样本类别无关,不会对样本类型的判别提供有用信息,反而会增加信息基因搜索的计算复杂度。因此,必须对这些“无关基因”进行剔除。

在衡量基因含有样本分类信息多少的度量问题上,Golub 等人采用了“信噪比”(Signal to noise ratio)指标<sup>[4]</sup>,即:

d = (μ1 - μ2) / (σ1^2 + σ2^2) (1)

其中:d 为基因的信噪比,μ1、μ2 分别为该基因在 ALL 和 AML 中表达水平的均值,σ1、σ2 为其表达水平的标准差。

依据式(1),若某一基因在 ALL、AML 两个类别中的分布均值相同,则其信噪比 d = 0,该基因将被作为无关基因而被剔除。然而,如果该基因在两个类别中分布的方差出现较大差异,比如其在 ALL 中分布方差很小,而在 AML 中分布方差很大。那么从生物学的角度分析,该基因很可能与 ALL 致病机理紧密相关<sup>[8]</sup>。依据这种分布方差的不同仍然可以很好地进行样本类别的判断。

基于上述分析,采用了基因的 Bhattacharyya 距离<sup>[9,10]</sup>来衡量基因中蕴含的分类信息量,即:

B = 1/4 \* ((μ1 - μ2)^2 / (σ1^2 + σ2^2)) + 1/2 \* ln((σ1^2 + σ2^2) / (2 \* σ1 \* σ2)) (2)

其中 B 为基因的 Bhattacharyya 距离。由式(2)知,Bhattacharyya 距离由两部分构成:第一项体现了基因在两个类别中分布均值的差异对样本分类的贡献;第二项体现了分布方差的不同对分类的贡献。依据该距离公式,即使基因在两类不同样本中分布的均值相同,只要分布的方差出现大的差异,仍然可以获得较大的距离值<sup>[10]</sup>。

从模式分类的角度看,基因的 Bhattacharyya 距离越大,利用该基因的信息,样本的可分性就越好。

根据公式(2),计算了每个基因的 Bhattacharyya 距离,并作出了基因的 Bhattacharyya 距离分布的直方图,见图 2。

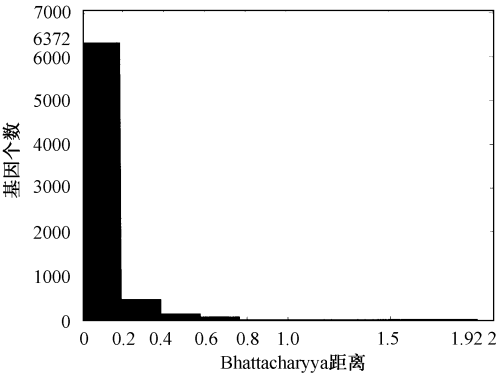


图 2 基因 Bhattacharyya 距离分布的直方图

依据基因所含样本类别信息的多少,将基因分为“信息基因”和“无关基因”两类。设  $S_I$  为信息基因集合,  $S_N$  为无关基因集合,则“信息基因”与“无关基因”可如下定义:

$$g \begin{cases} S_I & B(g) > \\ S_N & B(g) \end{cases} \quad (3)$$

其中  $g$  为基因,  $B(g)$  为基因  $g$  的 Bhattacharyya 距离,  $\theta$  为指定的 Bhattacharyya 距离的阈值。

由图 2 可知,绝大部分基因的 Bhattacharyya 距离小于 0.2。这些基因在两个类别中的分布,无论其均值还是方差均无明显差异,可以作为无关基因剔除。

急性白血病基因表达谱中基因 Bhattacharyya 距离的详细分布情况见表 2。依据表 2 和式(3)对信息基因与无关基因的定义知:在阈值  $\theta = 0.2$  时,  $card(S_I) = 757$ ,即在 7 129 个基因中,有 757 个基因为信息基因;  $card(S_N) = 6\,372$ ,即有 6 372 个基因为无关基因。 $S_I$  中 757 个基因均在不同程度上包含了样本的分类信息,是进一步分析的基础。

表 2 基因 Bhattacharyya 距离发布情况

Bhattacharyya 距离	基因个数	所占百分比
0 ~ 0.2	6372	89.38 %
0.2 ~ 0.4	464	6.51 %
0.4 ~ 0.6	166	2.33 %
0.6 ~ 1.0	99	1.39 %
1.0 ~ 1.92	28	0.39 %

### 3 浮动顺序搜索算法与候选分类特征子集的生成

信息基因集合  $S_I$  中含有的 757 个信息基因,可

以形成  $2^{757} \approx 7.58 \times 10^{227}$  个不同的基因组合,每个基因组合被称为一个特征子集。考虑到计算复杂度与最优解间的平衡关系,本研究采用浮动顺序搜索算法(Floating Sequential Search Method, FSSM)<sup>[11]</sup>对特征子集所构成的空间进行搜索,以得到不同大小的候选分类特征子集。浮动顺序搜索算法需要评价函数以评估特征子集所包含的分类信息量。采用了特征子集的 Bhattacharyya 距离作为评价函数<sup>[9-10]</sup>,即:

$$J(F_i) = \frac{1}{8} (\mu_2 - \mu_1)^T \left( \frac{1 + \sqrt{2}}{2} \right)^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \frac{\left| \frac{1 + \sqrt{2}}{2} \right|}{\sqrt{\frac{1}{2} \left( \frac{1 + \sqrt{2}}{2} \right)^2}}$$

(4)

其中  $J(F_i)$  表示含有  $i$  个基因的特征子集  $F_i$  的 Bhattacharyya 距离。 $\mu_1$ 、 $\mu_2$  为特征子集  $F_i$  中的基因在 ALL 和 AML 中分布的均值向量,  $\Sigma_1$ 、 $\Sigma_2$  为相应的协方差矩阵。

令  $F_{i\_max}$  为含有  $i$  个基因的候选分类特征子集,它是所有由  $i$  个基因构成的基因集合中具有最大评价函数的基因集合。本研究利用浮动顺序搜索算法在特征子集空间进行搜索,寻找具有不同维数的候选分类特征子集  $F_{i\_max}$ 。具体算法如下:

FFSM ( $n$ ,  $\{F_{i\_max} \mid card(F_{i\_max}) = i, i = 2, 3, \dots, n\}$ ) 算法描述:

step1:初始化  $F_{2\_max} = \{g_1, g_2\}$ ,  $g_1, g_2$  为  $S_I$  中 Bhattacharyya 距离最大的两个基因。

step2:如果  $i = n$ ,则退出。

否则 (1)  $S_I = S_I - F_{i\_max}$

(2) 搜索  $g \in S_I$ ,使  $F_{(i+1)\_max} = \{F_{i\_max}, g\}$  的评价值  $J(F_{(i+1)\_max})$  最大。

Step3:令  $F_{i\_max} = F_{i\_max} \cup \arg \max_{F_i \mid card(F_i) = i, F_i \subset F_{(i+1)\_max}} J(F_i)$

step4:如果  $J(F_{i\_max}) > J(F_{(i+1)\_max})$  则  $i = i + 1$ ,并转向 step2。

step5:令  $F_{i\_max} = F_{i\_max}$

如果  $i = 2$ ,则转向 step2。

否则  $i = i - 1$ ,转向 step3。

该算法也称为“加  $l$  减  $r$  法”。整个算法的结构如图 3 所示。

通过 FFSM 算法在特征子集空间中的搜索,产生出了 60 个具有不同维数的候选分类特征子集

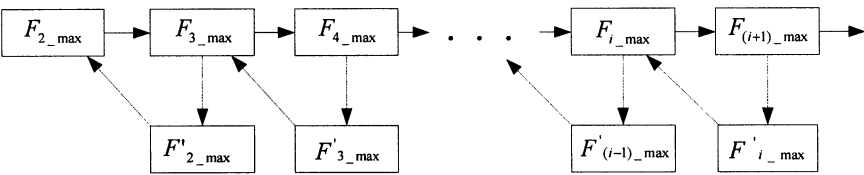


图 3 FFSM 算法示意图

$F_{i\_max} (i = 2, 3, \dots, 61)$ 。

4 基于支持向量机的亚型识别与分类信息基因集合的鉴别

依据生成的 60 个候选分类特征子集  $F_{i\_max} (i = 2, 3, \dots, 61)$ , 以支持向量机 (Support Vector Machine, SVM) 为分类器进行样本的识别, 以考察每个候选分类特征子集的分类能力。能够作为亚型分类特征的基因组合“分类信息基因集合”是具有最佳分类能力和最少基因个数的候选分类特征子集。

支持向量机作为一种基于结构风险最小化原理的机器学习算法<sup>[12]</sup>, 具有比基于经验风险最小化原理的神经网络学习算法更强的理论基础和更好的泛化能力。支持向量机将样本分类问题转化为在约束条件下求解二次规划问题。一个典型的线性支持向量机求解线性可分条件下的样本分类问题, 被转化为如下二次规划问题。

$$\text{Minimize : } J_{\text{svm}}(w) = \frac{1}{2} \|w\|^2 \quad (5)$$
$$\text{Subject to : } y_i (w^T x_i + w_0) \geq 1, i = 1, 2, \dots, N \quad (6)$$

其中  $w$  为权向量,  $x_i$  为第  $i$  个样本的特征向量,  $y_i$  为样本标号。

以线性支持向量机作为分类器, 运用不同的候选分类特征子集  $F_{i\_max}$  对急性白血病的两种不同亚型进行识别。其目的在于计算每一个候选分类特征子集  $F_{i\_max}$  的错误识别样本数  $E_{cv}(F_{i\_max})$  和  $E_{te}(F_{i\_max})$ , 其中:  $E_{cv}(F_{i\_max})$  是在训练样本集上采用“留一交叉检验法”(Leave-One-Out cross validation) 进行样本识别时的累计错误识别样本数;  $E_{te}(F_{i\_max})$  是对独立测试集进行识别时的累计错误识别样本数。

利用支持向量机得到的识别结果如图 4 所示。  
由图 4 知: 基于  $F_{5\_max}$  的识别结果为:  $E_{cv}(F_{5\_max}) = 0, E_{te}(F_{5\_max}) = 0$ 。即利用训练集进行的“留一交叉检验法”和使用独立测试集分别进行的样本识别实验中, 其正确识别率均为 100%。这表明

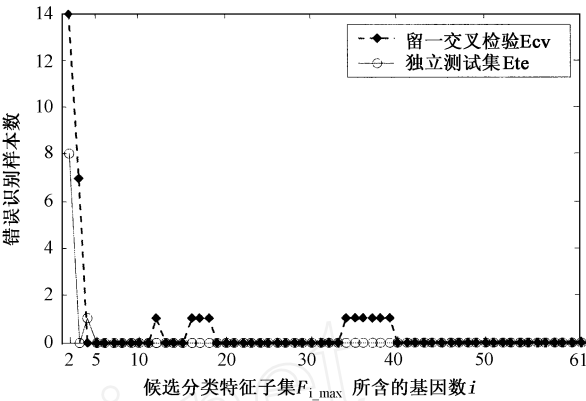


图 4 不同候选分类特征子集  $F_{i\_max}$  下的样本识别情况

候选特征子集  $F_{5\_max}$  包含了完整的样本分类信息。因此,  $F_{5\_max}$  就是所要寻找的“分类信息基因集合”。

分类信息基因集合中的 5 个基因及其描述见表 3。将这些基因同 Golub 选出的 50 个信息基因进行了对比: 其中的 M23197 (CD33) 和 M80254 (CyP3) 出现在了 Golub 等选出的 50 个信息基因中; M19507, M20902, M27891 这三个基因被 Golub 当作了无关基因, 而本研究结果显示这三个基因同 M23197、M80254 作为一个整体时, 包含了完整的样本分类信息, 是决定样本类别的信息基因。

表 3 信息基因及其描述

基因标号	基因描述
M19507	MPO, Myeloperoxidase
M20902	APOC1, Apolipoprotein CI
M23197	CD33, CD33 antigen (differentiation antigen)
M27891	CST3, Cystatin C (amyloid angiopathy and cerebral hemorrhage)
M80254	Peptidyl-prolyl cis-trans isomerase, mitochondrial precursor

针对急性白血病的亚型识别与信息基因选取问题, 将目前一些典型的研究成果同本研究成果进行了对比, 见表 4。从该表可知, 在保持对样本准确识别的基础上, 本研究结果进一步缩小了急性白血病亚型分类信息基因的范围, 为准确区分急性白血病的两种不同亚型提供了一组可供参考的基因。

表 4 不同方法提取的信息基因及其分类性能

	重要性指标	方法类型	信息基因数目	分类方法	正确识别的样本数	
					留一交叉检验法	独立测试集
Golub 等 <sup>[4]</sup>	信噪比	单基因分析	50	加权投票法	36/38	29/34
Tibshirani 等 <sup>[6]</sup>	“质心距离”	改进的单基因分析	21	基于质心距离的近邻法	37/38	32/34
Guyon 等 <sup>[7]</sup>	灵敏度	组合分析	8	支持向量机	38/38	34/34
本研究	Bhattacharyya 距离, 误识率	组合分析	5	支持向量机	38/38	34/34

5 结束语

针对急性白血病的亚型识别与信息基因选取问题,在研究有关文献的基础上,从分析基因组合的角度出发,研究了不同基因组合对样本的分类能力,并鉴别出了 5 个分类信息基因。基于该基因组合的样本识别实验,取得了 100 % 的正确识别率,表明该组基因包含了完整的样本分类信息。本研究结果将急性白血病亚型分类信息基因的范围进一步缩小,从生物信息学的角度,为急性白血病的诊断与研究提供了借鉴和参考。

值得指出的是,鉴别出的 5 个信息基因是一个功能基因组合,其对急性白血病亚型分类的影响是该基因组合作为一个整体共同作用的结果,而非单个基因作用的线性叠加。

参考文献

[ 1 ] Ramaswamy S, Golub TR. DNA microarrays in clinical oncology[J]. Journal of Clinical Oncology, 2002, 20(7) :1932 - 1941.

[ 2 ] Lander ES, Weinberg RA. GENOMICS: journey to the center of biology[J]. Science, 2000, 287(5459) :1777 - 1782.

[ 3 ] Lander ES. Array of hope[J]. Nature Genetics, 1999, 21(supp. 1) :3 - 4.

[ 4 ] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286(5439) :531 - 537.

[ 5 ] Khan J, Wei JS, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks[J]. Nat Med, 2001, 7(6) :673 - 679.

[ 6 ] Tibshirani R, Hastie T, Narasimhan B, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression[J]. PNAS, 2002, 99(10) :6567 - 6572.

[ 7 ] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2000, 46(13) :389 - 422.

[ 8 ] 李泽, 包雷, 黄英武, 等. 基于基因表达谱的肿瘤分型和特征基因的选取[J]. 生物物理学报, 2002, 18(4) :413 - 417.

[ 9 ] Duda OR, Hart PE, Stork GD. Pattern Classification [M]. Second Edition. New York: John Wiley & Sons 2001 :46 - 48.

[ 10 ] Theodoridis S, Koutroumbas K. Patter Recognition [M]. Second Edition. New York: Academic Press, 2003, 177 - 179.

[ 11 ] Padil P, Novovicova J, Kittler J. Floating search method in feature selection[J]. Pattern Recognition Letters, 1994, 15(11) :1119 - 1125.

[ 12 ] Vapnik VN. Statistical Learning Theory [M]. New York: Wiley Interscience, 1998.