·论　　著·

小波-神经网络方法在基因表达数据分析中的应用研究*

张　玲[1,2]，伍亚舟[1△]，陈　军[1]，易　东[1]
（第三军医大学：1. 卫生统计学教研室；2. 健康教育与医学人文教研室，重庆 400038）

摘　要：目的　针对基因表达数据，探索新的有效特征提取和分类方法。**方法**　采用小波多分辨率分析（MRA）方法提取基因表达的特征和前馈式神经网络（BP 神经网络）方法进行特征分类。**结果**　基因表达具有明显的多尺度特征，最大平均分类率为 94.72%。**结论**　采用多尺度理论对基因表达数据进行分析是一种新的有效的生物信息学方法，值得进一步探索与研究。

关键词：基因表达数据；多分辨率分析；前馈式神经网络
doi：10.3969/j.issn.1671-8348.2010.17.003
中图分类号：Q786；R195.1　　**文献标识码：**A　　**文章编号：**1671-8348(2010)17-2246-02

Applied research of gene expression data based wavelet-neural network method*

ZHANG Ling[1,2]，*WU Ya-zhou*[1△]，*CHEN Jun*[1]，*et al*.
（1. *Department of Health Statistics*；2. *Department of Health Education and Medical Humanities*，*Third Military Medical University*，*Chongqing* 400038，*China*）

Abstract：Objective　To search a new and effective method for feature extraction and classification based on gene expression data. **Methods**　The features of gene expression were extracted by the wavelet multi-resolution analysis，and the features were classified by the BP neural network methods. **Results**　There was multi-scale feature for gene expression；the maximum average classification rate was 94.72%. **Conclusion**　Multi-scale theory analysis of gene expression is a new and effective bioinformatics method，which is worth further exploration and research.

Key words：gene expression data；multi-resolution analysis；BP neural network

基因表达芯片具备大规模、高通量的特点，可以获得样品中大量基因序列和表达信息（数据），根据基因表达数据进行肿瘤诊断是当今生物信息学领域中的一个重要研究方向。利用这些基因表达数据（如癌症数据），可以建立有效的分类模型，实现对肿瘤样本与正常组织的正确分类；也可以找出决定样本类别的一组特征信息，加快疾病的诊断和对应药物研究。

目前主要采用聚类分析[1-2]和遗传算法[3]等方法对基因表达数据进行分类。但是，在基因表达谱数据分析过程中，由于微阵列表达数据具有样本少、维数高（基因数量巨大）、非线性等特点，使得有意义的基因表达信息被大量的噪声所淹没，且基因表达信号具有非常复杂的特性，利用各种统计方法对差异基因进行识别会产生大量的假阳性结果，建立分类模型则由于其中含有大量对分类不起作用的基因使其效能降低，其主要瓶颈集中在有效特征的选取以及对属于不同种类的样本进行正确诊断方面，而特征提取的质量和分类方法的优劣将会直接影响分类的效果。本文从信号处理的角度出发，利用多尺度理论[4-5]对白血病实验样本的基因表达数据进行处理和分析，具体采用小波多分辨率分析（multi-resolution analysis，MRA）方法[6]进行不同层次的特征提取，随后利用前馈式神经网络（Back-propagation neural network，BP 神经网络）方法进行识别分类，以正确区分不同的样本。

1　实验数据

本研究的数据集来自于 Golub 等[7]人进行的白血病实验样本，总共 72 个样本，其中 47 个为急性淋巴细胞白血病（ALL）样本，25 个为急性髓性白血病（AML）样本，每个样本包含有 7 129 个基因。该实验结果表明，对于属于不同种类的样本中的基因表达数据，其表达强度不一致（即存在差异），通过数据处理方法提取对分类有影响的特征，以达到区分这两类样本疾病的目的。

2　方　　法

基于 DNA 微阵列的芯片可以在同一时间点上或同一样本下提供大规模的基因表达数据，从信号的角度来看，基因表达数据也可以被视为一个信号集[8]。利用多尺度理论中的 MRA 方法进行去噪和提取特征，随后利用 BP 神经网络方法来识别这些特征，以便正确区分 ALL 和 AML 样本。

2.1　小波多分辨率提取基因表达特征　小波分析是近年来发展起来的一种新的时频分析方法，它能以不同的时间和频率分辨率分析信号，使得它具有多分辨率分析的特点，即在低频部分具有较高的频率分辨率和较低的时间分辨率，在高频部分具有较高的时间分辨率和较低的频率分辨率。正是这种特性，使得小波变换具有对信号的自适应性，而且小波分析并非是对单个点或单个频率进行处理的过程，因而具有很强的抑制噪声的能力。小波多尺度理论可以参考文献资料[4-6]。

小波变换系数（或部分系数）能反映信号在时域及频域的局部信息，各个小波系数实际上是信号时间-尺度（时频）特征的一种表现；且它们比较完备地描述了信号的主要特征，是特征表示的基础，这些系数可以重构出信号（表达），因此可以考虑将小波系数作为表达（信号）的特征。

本研究利用小波变换和多分辨率分析方法，分别选择 3 种小波函数 bio3.3、db5 和 sym4，并且在分解层数为 6、7、8 和 9

*　基金项目：国家自然科学基金资助项目（30901242）；第三军医大学科研创新基金资助项目（2007XG20）。　△　通讯作者，E-mail：asiawu5@sina.com。

情况下进行对基因表达数据处理和提取表达的特征，提取特征记为：$F_{ALL}=\{F_1,F_2,\cdots\cdots F_{47}\}$，$F_{AML}\{F_1,F_2,\cdots\cdots F_2\ 25\}$，从上述提取的特征中，采用随机抽取的方法分别构建训练集和测试集的特征向量：$F_{训练}=\{F_{ALL},F_{AML}\}$，$F_{测试}=\{F_{ALL},F_{AML}\}$。由于白血病数据集的原始基因表达数据大小不一，范围相差大，导致运算复杂，训练时间长，处理结果不佳，所以在训练分类前，对提取的特征首先进行标准化，将所有的数据转换到一个范围内，便于数据的处理。标准化函数采用 MATLAB7.0 软件自带的内部函数 premnmx( )完成。

2.2 BP 神经网络方法分类特征　对于提取后的特征分类，目前有很多种方法，而神经网络以其强大的非线性映射能力，在模式识别领域得到了广泛的应用；本研究属于典型的二分类问题，这里采取 BP 神经网络进行识别分类。具体识别分类时，采用 newff()函数创建一个前向 BP 网络，输入层神经元个数随着分解层数的改变而变化，隐含层传递函数为 tansig，输出层传递函数采用 logsig，训练函数为 traingscg，学习函数为 learngdm，输出值范围为(0,1)；以 0.5 为临界值，小于 0.5 判别为 ALL 样本，大于或等于 0.5 判别为 AML 样本；当平均误差率小于 0.0001，训练停止。上述提取特征和识别分类的具体算法程序均采用 Matlab7.0 软件编写、调试和运行处理。

3 结　果

小波函数选择 sym4 时的 MRA 的图示结果，见图 1(其他小波函数的结果略)。每幅图片的上面部分是原始基因的表达分布，中间部分是经过去噪后的表达，下面部分是提取的特征系数。从图上发现，随着分解层数的增加，提取的特征系数逐渐减少，每种情况下的特征数目相对于原始的表达数目减少了很多，而且这些表达特征系数主要反映了原始基因的表达变化情况，因此可以被用来进行特征的有效分类。

图 1　基因表达数据的多分辨率分析结果(小波函数 sym4)

采用 BP 神经网络对提取的特征进行分类，样本划分法将白血病数据集随机化平均分为两大类，其中一半为训练集样本 36 个(ALL25 个、AML11 个)，剩余的另一半为测试集样本 36 个(ALL22 个、AML14 个)，使用样本错判的个数作为判别效果的评价标准。为了检验分类效果的稳定性，每种情况均进行 10 次训练和测试，分类结果见表 1。

由表 1 结果发现，3 种小波函数分别在 4 种分解层数情况下，分类效果均比较理想，平均都达到 84%以上。从它们相互比较的结果分析发现，当选择不同的小波函数时，得到的结果有所不同，小波函数 sym4 的分类效果最好，平均达到 91.18%，小波函数 bio3.3 的分类效果次之，平均达到 89.79%，而小波函数 db5 的分类结果稍差；另一方面，从提取特征数目的多少来看，在分解层数为 8(特征数目为 229，小波函数 sym4)时得到的结果最好，平均达到 94.72%。

表 1　不同小波函数和分解层数下 ALL 和 AML 样本的 BP 分类结果(%)

分解层数	bio3.3		db5		sym4	
	特征数	分类结果	特征数	分类结果	特征数	分类结果
6	897	89.72	899	89.17	897	92.50
7	452	91.67	454	89.45	452	92.22
8	229	93.05	231	90.83	229	94.72
9	118	84.72	120	84.17	118	85.28
平均		89.79		88.40		91.18

表中的数据是 10 次平均后的结果。

4 讨　论

目前，针对基因表达数据研究的方法虽然很多，但能够对样本完全分类正确的并不太理想，其中一个重要的原因就是被识别分类的特征不是很明显。本研究针对 ALL 和 AML 样本数据集，采用多尺度特征提取研究方法，不仅能很好地起到降低维数作用，而且还能有效地提取表达的特征，在多次训练和测试运行的基础上，BP 神经网络方法的分类效果比较理想，说明该特征提取和分类算法效率高、运算速度快、耗时短。但小波函数的选择，对于分类的效果也有一定的影响。当分解层数越小时，即提取的特征越多时，包含对分类有促进作用的信息的机会也越多，但同时在该特征信息中也增加了那些影响分类的无效信息，因此选择合适的特征数，对于样本识别分类效果的准确性、稳定性、收效性都有着较大的影响。从上述结果的比较发现，当提取的特征包含了充分有效基因信息的同时，也避免分类无效基因信息的干扰，从而达到最优的分类效果；并且还保证了算法运输的耗时最短，即训练和测试时很快达到收效效果。

另外，识别分类方法对分类正确率也会有较大的影响。BP 神经网络方法克服了传统分类方法(如主成分分析等)的一些不足，解决了基因芯片存在样本少、维数高、非线性等问题，使得分类效果更加理想，最大的平均分类率达到 94.72%。同时，网络参数的选择和设计优化效果对于分类正确率也有着较大的影响，比如各层神经元个数的确定(特别是隐含层神经元个数的确定)，传递函数的选择，训练函数的选择和学习函数的选择等；而且从实际数据的处理分析中发现，该方法在分类效果的收敛性方面还需要改善。因此，在下一步的研究中将进一步探讨基因表达特征的分类方法。

总之，本研究将小波分析融入基因表达数据处理是一种重要的思想，其本质是通过对基因表达数据功能的重排列，利用多尺度或多分辨率算法对数据作适当的变换和分解，去除对分类造成干扰的噪声，降低数据的不确定性和复杂性，提取基因或样本在不同尺度上或不同层次方面的分类特征，改善特征识别的正确分类率，提高应用数值分类技术寻找复杂致病基因的效果，以便建立相应的数据分析技术平台，从而为生物信息学实验提供重要的信息，进一步丰富生物信息学的内容。本文所提供的方法能够把属于不同种类的疾病进行正确区分，对于疾病的诊断以及确定正确的治疗方案具有重大意义，更为重要的是本研究中利用小波多尺度技术研究基因表达数据也是生物信息学方法研究上一次有益的尝试，值得进(下转第 2250 页)

C_3 水平低于复苏前($P<0.05$),与同组复苏后即刻比较,B 组复苏后 6 h C_3、C_4 水平分别下降 45.24%,49.97%,差异有统计学意义($P<0.01$),C、D 组 C_3、C_4 水平与复苏前比较差异无统计学意义,提示随着复苏后缺血再灌注发展补体被激活而大量消耗,与张慧等[6]报道心脏体外循环手术患者 C_3、C_4 在转流中 30 min 和转流停止直至术后 4 h 出现大量消耗相吻合。C、D 组 C_3、C_4 水平降低较小,可能与血必净能减少缺血再灌注损伤,拮抗内毒素诱导单核/巨噬产生的内源性炎性介质失控性释放的作用有关,而与血必净的剂量无关。

近年有研究认为,心肺复苏术后患者 SIRS 发生率高达 90.6%,是影响心肺复苏最终死亡的主要危险因素[7]。TNF-α 是触发缺血再灌注损伤时炎症瀑布效应的关键因子,其表达高度依赖于 IL-12 或 IL-23,如果其表达被抑制,则再灌注时 TNF-α 合成及缺血再灌注损伤程度均明显减轻[8-9]。Th1/Th2 细胞平衡受诸多因素的调节,Th1 主要分泌 IL-2、IFN-γ、TNF-β,主要介导细胞免疫反应。Th2 主要分泌 IL-4、IL-5 和 IL-13,主要介导体液免疫反应。IL-4 在 Th 细胞分化过程中起到基本调节作用,可使 Th0 B 细胞向 Th2 细胞分化,IL-12 使 $CD4^+$ 和 Th2 细胞向 Th1 细胞分化[10];作者研究发现,与 A 组及复苏后即刻比较,B、C、D 组 IL-12、TNF-α 水平复苏后均高于 A 组和复苏后即刻($P<0.01$),与 B 组比较,C、D 组复苏后 IL-12、TNF-α 水平低于 B 组($P<0.01$),提示 TNF-α、IL-12 参与了心肺复苏后再灌注损伤的发生、发展过程,本实验中未检测到血清 IL-4 浓度,可能与 Thl/Th2 细胞失衡及细胞因子之间交互抑制有关,由此推测心肺复苏后机体存在免疫失衡,主要以 Th1 细胞因子表达异常为主。血必净可抑制血清 IL-12、TNF-α 水平,调节过高或过低的免疫反应,保护和修复应激状态下受损的脏器。

通过本实验可推测,心肺复苏后存在补体水平的消耗,炎性因子 TNF-α、IL-12 过度表达及 Th1/Th2 细胞失衡,主要以 Th1 细胞因子表达异常为主。血必净可抑制补体水平的过度活化,抑制血清 IL-12、TNF-α 水平,保护和修复应激状态下受损的脏器。

参考文献:

[1] Eisenberg MS, Mengert TJ. Cardiac Resuscitation[J]. N Engl J Med, 2001, 344(17): 1304.

[2] Hendrickx HH, Rao GR, Safar P, et al. Asphyxia cardiac arrest and resuscitation in rats: Ⅰ Short term recovery [J]. Resuscitation, 1984, 12(2): 97.

[3] Lazar HL, Bokesch PM, van Lenta F, et al. Soluble human complement receptor 1 limits ischemic damage in cardiac surgery patients at high risk requiring cardiopulmonary bypass[J]. Circulation, 2004, 110(11 Suppl 1): Ⅱ 274.

[4] 马虹,吴素华,董吁钢,等.急性心肌梗死患者补体激活与心肌缺血损伤的关系[J].中华老年心脑血管病杂志,2002,2(6):372.

[5] 刚丽,孙晓义,蒋志宏,等.血必净注射液治疗急性呼吸窘迫综合征的疗效观察[J].疑难病杂志,2007,6(6):359.

[6] 张慧,苗庄,刘刚,等.心脏体外循环手术患者检测 C_3、C_4、CRP 结果分析[J].吉林医学,2007,28(9):1132.

[7] 朱英,黄淮,颜景华,等.心肺复苏后多器官功能障碍综合征危险因素和预后分析[J].中国医师杂志,2006,8(8):1048.

[8] Lentsch AB, Yoshidome H, Kato A, et al. Requirement for interleukin-12 in the pathogenesis of warm hepatic ischemia/reperfusion injury in mice[J]. Hepatology, 1999, 30(6): 1448.

[9] Husted TL, Blanchard J, Schuster R, et al. Potential role for IL-23 in hepatic ischemia/reperfusion injury[J]. Inflamm Res, 2006, 55: 177.

[10] 王兵,张畔.多器官功能障碍综合征中急性虚证发病与辅助 T 淋巴细胞 1/2 平衡之间的关系及治疗对策[J].中国中西医结合急救杂志,2005,12(1):58.

(收稿日期:2010-03-05 修回日期:2010-04-10)

(上接第 2247 页)

一步探索研究。

参考文献:

[1] Eisen MB, Spellman PT, Brown PO, et al. Cluster analysis and display of genome-wide expression patterns[J]. Proc Natl Acad Sci USA, 1998, 95(25): 14863.

[2] Seal S, Komarina S, Aluru S. An optimal hierarchical clustering algorithm for gene expression data[J]. Inform Proc lett, 2005, 93(3): 143.

[3] 孟范静,刘毅慧,王洪国,等.遗传优化算法在基因数据分类中的应用[J].生物信息学,2008,6(20):119.

[4] Chen XF, He ZJ, Xiang JW, et al. A dynamic multi-scale lifting computation method using Daubechies wavelet[J]. J Comput Appl Math, 2006, 188(2): 228.

[5] 罗万春,陈军,伍亚舟,等.基于小波多尺度的人类胚胎期大脑皮层基因表达分析[J].重庆医学,2009,38(12):1462.

[6] 胡昌华,张军波,夏军,等.基于 MATLAB 的系统分析与设计——小波分析[M].西安电子科技大学出版社,1999.

[7] 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: 531.

[8] 闫晓光,游顶云,李康.基因表达数据与小波变换分析的思想与方法[C]//2007 年中国卫生统计学大会. 2007 年中国卫生统计学术大会论文集,西安,2007.

(收稿日期:2009-11-27 修回日期:2010-02-07)