基因表达谱聚类/分类技术研究及展望

Research on Gene Expression Data Based on Clustering/classification Technology

李 杰[1],唐降龙[1*],王亚东[1],李 霞[1,2*]

LI Jie[1],TANG Xiang-Long[1*],WANG Ya-Dong and LI Xia[1,2*]

1. 哈尔滨工业大学计算机科学与技术学院,哈尔滨 150001

2. 哈尔滨医科大学 生物信息学系,哈尔滨 150086

1. *School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001, *China*

2. *Department of Bioinformatics, Harbin Medical University, Harbin*, 150086, *China*

摘 要 随着人类及多种模式生物全基因组测序基本完成,人类基因组计划的研究进入后基因组时代.后基因组时代研究的焦点已经从测序转向功能研究。聚类/分类技术作为分析基因表达谱和识别基因功能的重要工具之一,近年来获得很大的发展。对目前基因表达谱聚类/分类技术及它们的发展,进行了综述性的研究,分析了它们的优缺点,结合我们的研究,提出了解决问题的思路和方法,为基因表达谱的进一步研究提供了新的途径。

关键词 基因表达谱,分类,聚类,基因调控网络,逆向工程

中图分类号 Q78 **文献标识码** A **文章编号** 1000-3061(2005)04-0667-07

Abstract As the work of sequencing the genome of the human and many model organisms has been partially or fully finished, the "postgenomic era" has begun. Scientists are turning their focus toward identifying gene function from sequencing. Clustering technology, as one of the important tools of analyzing gene expression data and identifying gene function, has been used widely. In this paper we discuss main clustering technology about gene expression data at present, analyze their advantages and disadvantages, present the methods to solve the problems and give new approaches to study gene expression data.

Key words gene expression data, classification, clustering, gene regulatory network, reverse engineering

最近发展起来的基因表达谱芯片技术是分子生物学在实验领域的一项重大突破,为探索生命的本质提供了极大的便利,成为探索生命奥秘的重要工具之一。通过基因表达谱芯片实验可以同时观察成千上万个基因在不同个体、不同组织、不同发育阶段的表达状况,研究它们的功能和相互关系,加深对生命本质的认识。另外也可以根据基因在不同条件下表达的差异性来进行复杂疾病诊断、药物筛选、个性化治疗、基因功能发现、农作物优育和优选、环境检测和防治、食品卫生监督及司法鉴定等,因此基因表达谱的研究具有重要的理论价值和应用意义。近年来通过基因表达谱芯片实验产生了大量的表达谱数据,这类数据的特点是样本数目少、维度高、冗余基因和噪音多。如何对这类"新型"数据进行有效地分析成为研究人员面临的一个全新领域。随着这些海量数据的不断积累,新的数据分析方法、理论和技术也不断涌现。当前主要的分析技术是表达谱的聚类和分类。本文从信息技术的角度对表达谱聚类和分类技术进行研究综述

Received: March 21, 2005; Accepted: April 28, 2005.

This work was supported by the grants from the National Natural Sciences Foundation of China (No. 30170515 and 30370798), the National High-Tech Development Project of China (863 Program) (No. 2003AA2Z2051) and 2002AA2Z2052) and the 211 Project.

* Corresponding author. E-mail: lixia6@yahoo.com

并提出我们解决问题的思路,最后讨论目前面临的挑战和对基因表达谱分析技术发展的展望。

1 基因表达谱聚类和分类分析

数据聚类和分类是重要的数据挖掘方法,表达谱基因聚类可以将那些具有相关功能和共调控关系的基因聚在一起,用于推断调控基元、注释基因功能和确立分子标签,为进一步详细研究基因的功能打下基础。表达谱样本聚类可以帮助发现新的疾病亚型。样本分类可以提高复杂疾病诊断的正确率。比如弥散性大 B 细胞淋巴瘤(DLBCL)分为两种亚型,常规治疗时两种亚型的临床效果差异很大,仅有 10%的患者疗效较好,存活率较高,其余患者的疗效较差,存活率低。因此,鉴别该疾病的亚型具有重要的临床价值。通过对包含 18000 个 DNA 探针的基因芯片数据分析,发现了用于鉴别该疾病亚型的分子标签。研究基因表达谱聚类和分类技术的目的之一是为了发现基因之间的调控关系,从大量的数据中挖掘出具有重要研究和应用价值的基因。由于基因表达谱数据的特殊性,要求新的方法除了具有能够发现数据间的真正关系、分类精度高、方法简单、速度快、鲁棒性强(在分类算法受到随机干扰及其它不确定因素影响时能够保持较高的分类精度)这些特点外,还要求分析结果可视化程度好,可解释性强,具有很好的统计学和生物学意义。这为基因表达谱聚类和分类方法的研究提出了新的挑战。根据映射关系获得的方法及类间相似性的度量方法,可以分为[1] 数据聚类、统计分类、神经网络和结构模式识别 4 种类型,这里就当前基因表达谱聚类和分类的主要算法加以综述。

1.1 数据聚类

数据聚类的目标是用某种相似性度量的方法将数据组织成有意义的和有用的各组数据。数据聚类是一种非监督学习的方法,它不需要利用已知类的信息,聚类的结果完全依赖于数据本身。在基因表达谱研究中,常用的数据聚类方法有分层聚类和 *K*-均值聚类。

1.1.1 分层聚类 :分层聚类(Hierarchical Clustering,HC)是一种传统的聚类方法[2,3],最常用的是 HAC(Hierarchical Agglomerative Clustering)算法,它通过构建一个二叉树图(Dendrogram)来描述基因表达之间的关系。Eisen[2] 等开发了用于基因表达谱分层聚类的软件包,由于该软件简单易用,可视化程度好,可以从网上免费下载,所以很快获得了推广应用。目前该软件包是基因表达谱聚类研究中最常用的工具之一。不过,HAC 算法本身有一些缺点 :解不唯一,无法重新评估 数据的输入顺序对结果可能产生影响 :聚类完成后,数据不能从一个类转移到另一个类 :选择不同的类间距离定义可能影响类的构成并导致不同的注释 :在处理含噪数据时缺乏鲁棒性 ;分层结构是一个二叉树,它有可能不能恰当地描述层与层之间的关系,比如对图 1(1)进行聚类,按照 HAC 算法,最后的分类结果只能是(2)中的一种,不可能是(3) ;当然它更不适合描述那些本身不是分层组织的数据之间的关系。

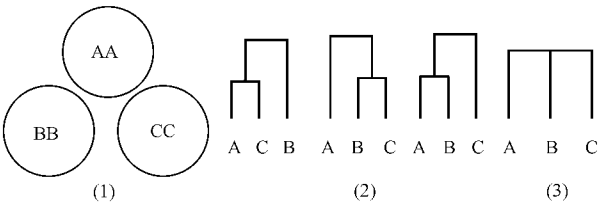

图 1 分层聚类的几种情况

Fig. 1 Several results of hierarchical clustering

为了克服 HAC 的这些缺点,新的分层算法不断被提出。Dopazo[4] 在 SOM(Self-Organizing Map)和 Fritzke[5] 的增长胞结构(Growing Cell Structures)基础上提出了 SOTA(Self-Organizing Tree Algorithm)分层聚类算法。在 SOTA 算法中,采用了神经网络机制,使得 SOTA 具有了鲁棒性。Luo[6] 提出了 HGSOT(Hierarchical Growing Self-Organizing Tree)算法,HGSOT 是在 SOTA 的基础上做了进一步的改进,使得输出的拓扑结构不再固定,每一层的分叉数由 KLD(K-Level Distribution)算法动态查找决定。针对一旦数据被错分到一个类后,就不能再重新评估分配和层间关系不能被很好地描述的缺点,Feng[7] 提出了 DGSOT(Dynamical Growing Self-Organizing Tree)自组织树算法,该算法的优点是 :它采取一种自上而下的方法构建自组织树,在构建树的过程中,在每一层都动态优化分支的个数 :在 DGSOT 中引入了 KLD(K-Level Up Distribution),当数据在上一层没有被分配到合适的类时,在下面各层的分类中仍有机会被调整到其它类中 ;树可以在垂直和水平两个方向上增长。

1.1.2 *K*-均值聚类 :*K*-均值聚类算法是一种叠代算法,它的分类性能依赖于聚类中心的初始值,所以常常在应用 *K*-均值聚类算法以前做一些预处理,以获得好的初始聚类中心。另外为了避免迭代的不收敛,需要预先指定迭代的次数。*K*-均值聚类与分层聚类一样,具有分类结果不确定、鲁棒性不强等缺点,为了克服这些缺陷,*K*-均值聚类法常与其它的一些聚类方法结合起来运用,实现优势互补。比如 :Sugiyam[8] 将 *K*-均值聚类和 SOM 结合起来,对酵母菌基因表达谱数据进行聚类,联合聚类的结果远远好于 *K*-均值单独聚类的结果。

1.2 统计分类

统计分类是模式识别长期发展过程中建立起来的经典方法,它通过概率统计模型得到各类别的特征向量分布,进而进行分类。特征向量的分布是建立在一个类别已知的训练样本集上的,因此,它是一种监督学习的分类方法。

1.2.1 Fisher 线性判别(Fisher 's Linear Discriminant Analysis,FLDA) :FLDA 是 R.A.Fisher 1936 年提出的非参数分类方法,它的主要任务是找到一个投影矩阵 W,通过这个投影矩阵将各类样本数据映射到一个新的空间,在这个新空间中,各类样本能够最大程度地分离开,并且类内离散度越小越好。对于只有两类样本的基因表达谱矩阵来讲,应用 FLDA 可以找到距离最佳投影矢量最近的基因矢量(关键基因矢量),通过关键基因对基因表达谱样本分类,可以获得最大的分类精

度。

FLDA 分类方法,计算简单,可以观测一维情况下的分类情况,能够将 N 维空间样本投影到一维,降低样本空间的维数。对于需要观测二或三维的信息时,该分类方法无能为力,所以为了便于应用,常常需要对传统的 FLDA 算法进行改进。Xiong[9] 采用主成分分析与 Fisher 判别相结合的方法,对结肠癌数据进行了分类研究,结果表明:采用 7 个主成分可以获得精度为 87% 的判别效果。Li[10] 等采用 Fisher 判别与逐步优化过程相结合的办法,对同样的数据进行分类,只需 5 到 6 个基因,预测精度就可以达到 95% 左右,并且,此分类精度已考虑了样品分配方案对预测精度的影响。为了获得一种能够挖掘和解释基因表达谱中的信息,并对癌症进行很好分类的方法,Xinying[11] 提出了 CDFDA(Cross-Weighted Fisher Discriminant Analysis)算法,在 CDFDA 算法的预处理步骤中采用了 FLDA 算法。利用 CDFDA 方法对白血病基因表达谱进行分类处理,取得很好的效果。

1.2.2　最近邻和 *K*-近邻:最近邻法是非参数决策方法的一种,是一种根据样本提供的信息直接决策的方法。在最近邻分类算法中,常用欧氏距离或最小相关性作为距离函数,对于基因表达谱矩阵的两个样本矢量 $x=(x_1\,x_2,\cdots,x_p)$ 和 $x'=(x'_1,x'_2,\cdots,x'_p)$($p$ 是基因个数,x_p 是第 p 个基因的表达值),如果它们之间的距离最近,则它们属于同一类样本。最近邻算法简单,在基因表达谱分类中获得广泛应用。

K-近邻算法是从已知样本中找到未知样本 x 的 K 个近邻,x 的类别由它的 K 个近邻的多数来决定,因此 *K*-近邻算法稳定性比较好,抗噪能力强。*K*-近邻算法难点在于:分类前需要选择 K 值,K 值的选择直接影响分类的结果,对于基因表达谱样本分类来讲,影响就更为显著。

1.2.3　决策树:决策树也是一种常用的分类方法,它的优点是:在学习的过程中不需要使用者了解很多背景知识(这同时也是它的最大缺点),概念简单,计算效率高;作为一种非参数分类方法,使用者不需要输入任何参数;分类的结果意义明确,可解释性强;另外有关决策树的演变算法也很多。正是具有这么多优点,所以在数据挖掘领域有着广泛的应用。

Zhang[12] 等在二叉树的基础上提出了一种递归划分算法,在这种分类算法中,他在熵概念的基础上定义了节点纯度函数 $f(p)=p\log(p)+(1-p)\log(1-p)$($p$ 是一个节点内的的基因处在正常态的概率),以该函数作为内部节点分裂的规则。$f(p)$值越大表示节点纯度越高,越小纯度越低。当 p 值为 0 或 1 时,$f(p)$值最大,表示一个节点上的所有的样本是同一个类型,当 p 值为 0.5 时最小,$f(p)$表示一个节点上的样本是等概率分布的。在构建二叉树的过程中,用节点纯度函数测试所有可能分裂的节点,以选择合适的分叉规则。用该方法对结肠癌数据进行分类,精度达到 92% ~ 94%。

1.2.4　贝叶斯法:贝叶斯分类方法以贝叶斯定理为理论基础,是一种在已知先验概率与条件概率的情况下的模式识别方法。它在文本分类、字母识别、经济预测等领域获得成功的应用。贝叶斯分类器分两种:一种是简单贝叶斯分类器,它假设一个属性对给定类的影响独立于其他属性,即特征独立性假设。当假设成立时,与其他分类算法相比,简单贝叶斯分类器是最精确的。

2000 年 A. Keller[13] 等用贝叶斯分类器分析白血病、大肠杆菌和卵巢瘤 DNA 阵列表达谱数据,分别获得 100% 和 84% 的分类精度。Zhou[14] 等开发了一种非线性贝叶斯分类器,用该分类器分析了遗传乳腺癌、圆的蓝色小细胞瘤、急性白血病,结果表明:该分类器不仅分类精度高,而且能够发现重要的基因。

另一种是贝叶斯网络分类器,它可以考虑属性之间的依赖程度,更能反映类间的真实情况,但计算复杂度比简单贝叶斯分类器高得多。目前已有一些研究人员[15] 建立了用于分析基因表达谱的贝叶斯网络模型,不过还不成熟,有待进一步提高。

与其它的聚类算法,比如决策树、人工神经网络相比,贝叶斯分类法具有如下优点:

第一,贝叶斯分类法有着深厚的统计理论基础;第二,贝叶斯分类法网络能够方便地处理不完整数据;第三,贝叶斯分类法不仅能够分类,而且能够学习变量间的因果关系。这一点对于基因表达谱的研究来讲,非常重要,因为大规模基因表达谱分析不仅仅是聚类的问题,更是一个知识发现的过程。在数据分析中,因果关系有利于对领域知识的理解;在干扰较多时,便于作出精确的预测。

贝叶斯分类法的局限性在于:首先,贝叶斯分类法是建立在数据的分布是某种分布(多数假设为高斯分布)这一假设基础上的,对基因表达谱来讲,它的分布却是未知的,很难确切地说是高斯分布。其次,它假设不同的基因之间是相互独立的,而实际上基因的表达值之间是相互关联的。所以,一般情况下认为:只有在独立性假定成立的时候,贝叶斯方法才能获得精度最优的分类效果;或者在同性相关性较小的情况下,能获得近似最优的分类效果。

1.2.5　隐马尔可夫模型(Hidden Markov Model,HMM):HMM 产生于 20 世纪 60 年代末,已经成功应用于连续语音识别和在线手写体识别。最近,有研究人员将 HMM 应用到基因表达谱数据的分析上。HMM 研究的是一个时间序列的表达谱,如表 1 所示。t 轴表示对同一类细胞样本的测试时间,g 轴表示测试的基因,表内数据表示在不同时间点不同基因的归一化表达值。Ji[16] 等将 HMM 引入基因表达数据聚类分析。他们根据公式 1 将基因表达谱时间序列矩阵转化成一个表达波动序列(表 1)。其中 N 为时间点数,E_i 为基因在每一时间点的表达值,S_i 为序列转换值,a 为容限值(如取 0.05)。这样原来的 N 个时间点的基因表达谱矩阵,转换为由数值 {0,1,2} 构成的 $N-1$ 维波动序列,序列中的值反映了下一时间点表达谱的变化情况,而整个序列则描述了基因表达的波动情况。

$$S_i=\begin{cases}0 & if \parallel E_i-E_{i+1}\parallel\ <a\\ 1 & if \parallel E_{i+1}-E_i\parallel\ \geq a\\ 2 & if \parallel E_{i+1}-E_i\parallel\ \geq a\end{cases}\quad 1\leq i\leq N-1 \qquad (1)$$

表 1 基因表达谱时间序列

Table 1 Time series of gene expression data

	Time series						Fluctuation sequences
	$t1$	$t2$	$t3$	$t4$	$t5$	t	
$g1$ 1	0.46	0.30	0.80	1.51	0.90	…	
$g2$ 2	−0.10	0.49	0.24	0.06	0.46	…	
$g3$ 3	0.15	0.74	0.04	0.10	0.20	…	
$g4$ 4	−0.45	−1.03	−0.79	−0.56	−0.32	…	
5	−0.06	1.06	1.35	1.09	−1.09	…	
…							
g							

The normalized gene expression data

$$\Longrightarrow \begin{pmatrix} 2 & 1 & 1 & 2\cdots \\ 1 & 2 & 2 & 1\cdots \\ 1 & 2 & 1 & 1\cdots \\ 2 & 1 & 1 & 1\cdots \\ 1 & 1 & 2 & 2\cdots \end{pmatrix} \quad S$$

表达波动序列构建的 HMM,其主链由 N 个状态构成,每一状态对应着细胞周期的一个实际状态,每一状态可按一定分布产生一个值(0,1 或 2),代表这一周期细胞状态的调节方向。为了便于研究,他们在波动序列里增加一个 '起始 '状态和一个 '结束 '状态。一个基因表达波动序列可以由一 '随机步 '通过模型产生 :从 '起始 '状态开始,选择跃迁到另一个细胞状态,并依据这一状态的分布产生数值(0,1 或 2),然后再选择下一个跃迁状态,并产生下一个数值,持续这一过程,直到 '结束 '状态。按照这种方法就产生一个表达波动序列。接着应用 Baum-Welch 方法训练模型,并应用前反向算法计算序列的概率。

Schliep[17] 等用 HMM 模型分析了酵母和纤维原细胞基因表达谱时间序列,对于两类数据 :周期(酵母)和非周期(纤维原细胞),都取得同样好的效果。Ji 应用这一 HMM 模型对多个基因表达数据进行聚类,结果表明 :这一方法与 K-均值聚类和 SOM 具有同样的聚类效果。除了能确定正确的类数外,HMM 还能够识别功能基因和调控基因,证实很多基因具有多种功能,并揭示在不同(或同一)过程中不同功能基因间的关系。HMM 抵抗噪声的能力比较强,能够方便地处理具有缺失数据的基因表达谱(在基因表达谱中常出现这种情况),这样在有噪声和某些数据不完整的情况下,利用 HMM 模型同样可以很好地分析和显示数据。HMM 还能够充分利用已知的生物信息(比如某些基因的响应已知),将已知的信息引入模型。

1.3 基于神经网络的融合算法

神经网络是一个由简单的处理单元(神经元)构成的大规模的并行分布式处理器。通过调节神经元之间的权重系数可以实现监督和非监督学习条件下的分类和回归。与统计分类方法相反,神经网络是一个 '模型无关 '的机器。表现出一种非监督学习条件下分类器的性能,具有能够通过调整权值使得输出在特征空间中逼近任意目标的优点。与统计分类相比,存在的不足就是,神经网络的数学解释很复杂,从神经网络本身也得不到任何语义的信息。这对于基因表达谱的研究是不利的,因为从神经网络的输出结果中研究人员很难获得具有生物意义的解释。为了克服神经网络的这一缺点,我们首先用集成决策树的方法选择出具有较强统计意义的特征基因,然后用具有分类精度高、抗噪能力强的支持向量机对结肠癌基因表达谱数据进行分类取得了很好的效果。

神经网络有许多种,这里主要讲自组织映射和支持向量机这两种在基因表达谱分类研究中用的较多的神经网络方法。

1.3.1 自组织映射(Self-Organizing Maps,SOM) :SOM 是一种基于竞争学习的非监督神经网络[18],SOM 可把输入空间的多维数据映射到低维(一维或二维)离散网络上,并能保证相同特征的输入数据映射到低维空间时保持拓扑一致性,其中网格神经元的空间位置表示输入模式包含的内在统计特征。SOM 的这一特性,为多维数据集的可视化提供了一种新的解决方案。

SOM 在基因表达谱研究中获得较多的应用[19,20,21],原因在于 :通过 SOM 可以将高维表达谱映射到二维,从网格上的数据可以清楚地看到数据(基因或样本)的空间聚类情况,这非常有利于理解样本之间的关系 ;另外,它还具有稳健准确和抗噪能力强的优点。Tamayo P 等编制了基于 SOM 的分类软件 GENECLUTER,现已获得广泛的推广使用。

SOM 是一种拓扑保留的神经网络,易产生不均衡分类。若不相关数据过多,感兴趣的数据较少时,分辨率可能会很低。因此,在应用 SOM 对基因表达谱聚类前,需要对数据进行筛选。为了克服 SOM 自身的一些缺陷,一些基于 SOM 的新的算法被提出,比如 :自组织树映射算法(Self-Organizing Tree Maps Algorithm,SOTM)[22],是在 SOM 的基础上引入了分层聚类技术,从而可以有效克服 SOM 的不均衡分类这一缺点。为了使 SOM 的聚类边界明显,Sugiyama[8] 等将 K-均值分类方法引入了 SOM,从而有效解决了这一问题。

1.3.2 支持向量机(Support Vector Machines,SVM) :SVM 是 Vapnik 提出的基于结构风险最小化原理(Structural Risk Minimization Principle,SRM)的统计学习理论,它的主要思想是建立一个超平面作为决策曲面,使得正例和反例之间的隔离边缘被最大化,主要用于模式分类和非线性回归问题。SRM 使 VC(Vapnik Cherovnenkis)维数的上限最小化,这样 SVM 方法比基于经验风险最小化(Empirical Risk Minimization Principle,ERM)的人工神经网方法具有更好的泛化能力。SVM 不仅能够处理高维数据,而且分类精度很高,抗噪能力强,不需要使用者调整和输入大量的参数。它的另外一个优点是它的可度量性,通过训练后支持向量的个数通常比较小,这一点对矩阵维数不断增加的基因表达谱数据来讲非常

重要。正是这些优点使得 SVM 在网上信息的自动分类、手写体相似性识别和蛋白质功能的预测及癌症的分类等很多领域获得了广泛的应用。

SVM 在基因表达谱研究中也获得广泛的应用[23-27]。Furey T[25]、Ben-Dor[26] 和 Ramaswamy[27] 进行癌症分类时都运用了线性可分的 SVM 技术。Chu[28] 将 SVM 方法用于 B-细胞淋巴瘤的分类,结果表明:当采用 62 个分类基因时,精确度达 100%。该方法的基本思路是首先用主成分分析(Principal Components Analysis)方法选择用于分类的基因,然后使用 SVM 进行分类,以区分 B-细胞淋巴瘤的三种亚型:弥漫行大 B 细胞淋巴瘤、滤泡型淋巴瘤、慢性淋巴性白血病。从 62 个样本中选取 31 个训练 SVM,另外 31 个用来测试。Chen[29] 将遗传算法和 SVM 结合起来识别白血病和大肠杆菌癌也取得非常好的效果。

为了达到更好的分类效果,SVM 多与其它方法结合起来应用。比如 Valentini[30] 采用 Bagged Ensembles 与 SVM 相结合的办法,对结肠癌和白血病数据进行了分类研究,结果表明:对白血病而言,即使不进行特征提取,在 16 个基因的情况下,应用该方法进行分类,就可以获得精度超过 80% 的效果。我们[31,32] 先用集成决策树选出 20 个分类意义明确的特征基因,然后用 SVM 对结肠癌基因表达谱数据(40 例结肠腺癌组织,22 例正常结肠组织)分类,结果表明:用 20 个特征基因作为模式的特征所得到的分类器的性能指标(Accuracy、Precision 和 Recall)几乎均高于用 2000 个基因作为模式特征所得到的分类器的性能指标,在降维情况下,分类器的准确性没有下降,反而上升,而且 SVM 的泛化能力获得了提高。

现在 SVM 多用来解决两类问题的分类,面向多类问题的算法还不多,不少研究人员正努力寻求新的算法,以突破这一限制。

2　挑战与展望

基因表达谱芯片作为一种新兴技术,还有许多方面需要完善,基因表达谱数据的分析还处于初始阶段,进行下一步的研究还面临着许多挑战,这些挑战是困难也是后基因时代的主要研究方向,这些挑战有实验技术上的,有数据本身的,还有处理数据的方法上的,这里将基因表达谱研究目前面临的主要的困难和我们解决这些困难的思路进行简单介绍。

2.1　数据本身的问题

表达谱数据的首要问题是数据质量的问题。不同实验室做同样的实验,由于实验材料、实验人员、实验仪器和实验环境的不同,所获得的表达谱数据会有较大差别,这种数据间的不一致性和噪音成为研究表达谱的一大障碍。采用不同表达谱芯片所做实验结果会有较大差异,Ishii M, Evans S J 和 Tony Yuen[34,35,36] 对 cDNA 芯片数据、Affymetric 芯片数据进行比较证实了这点。另外实验中各种噪音的影响,数据不可避免地存在误差,这样对有差异的数据进行分析往往会得到不同的结果,尽管通过预处理可以减少这种差异。但从根本上解决实验数据的不一致性问题需要国际合作,尽快制订出实验的相关标准,这样才能做到数据具有较强的可比性,分析结果具有可信性。

基因表达谱数据的另一个问题是数据的种类和数据量的问题。基因表达谱研究的目的是想通过基因表达谱来认识复杂的生命现象,确定与特定生命现象(如发育、生长、肿瘤发生等)相关的基因,分析不同基因的功能及它们之间的调控关系,推断潜在的调控区域和基因网络。要想达到上述目的则需要多层面(不同的代谢过程,不同的发育阶段,不同的外界、内部刺激等)的高通量实验数据,而目前的表达谱数据大多是单一实验条件下或某几种实验条件下的数据,尽管网上公布的基因表达谱数据以指数的速度增长,但对于所研究的问题来讲,这些数据提供的信息仍然是十分有限,因此需要更多的基因表达谱实验数据。

2.2　基因表达谱数据与其它信息相融合

利用基因表达谱数据分析研究基因功能是非常重要的手段之一,根据生物学中的中心法则可知,基因的调控和表达不是孤立的,它受各种酶和其它因素的影响,将细胞其它方面的各种信息与表达谱结合起来[23,38-40],研究基因的功能会更加有利于对基因功能的全面认识和蛋白质功能的预测,这些信息包括临床医学观测数据,同源序列和顺式序列,DNA、蛋白质序列数据库和公布的各类文库等。如何将这些数据融合起来是我们目前面临的主要困难之一,也是以后研究的主要方向之一。

2.3　集成各种聚类算法方法

对于同一基因表达谱数据用不同的聚类方法进行基因聚类,所得到的结果差别会很大,就是用同一种聚类方法,由于所选初始条件和参数的不同,聚类结果也可能有不小的差别。不同的生物过程和实验中,每个基因的功能和表现是不同的,一个基因可能参与多种生物过程,在不同的过程中扮演不同的角色,因此,在不同的实验数据中,甚至相同数据,应用不同聚类方法,一个基因可能聚到不同的类中,也就是说提供给我们的信息是多方面和多层次的,每种聚类方式只可能挖掘出有关基因信息的某个方面,所以要想从基因表达谱中获得基因的更多信息,需要将各种聚类方法结合起来,去对不同的实验数据聚类,这样才能更加全面地了解基因的功能和基因间的调控关系,才有可能发现新的对健康和疾病治疗有重要意义的基因。具体采用哪些聚类方式,如何将这些聚类方法很好地集成在一起,是我们进一步的研究目标。目前大多数聚类和分类方法是以分类精度作为算法的主要目标,对分类结果的统计意义和分类结果的可解释性考虑的较少,我们提出的集成决策树[32] 方法除了具有较高的分类精度外,分类结果还具有很好的统计意义。我们应用该方法挖掘出的多个基因已被分子生物学实验证实,这些基因在多种人类肿瘤细胞系中总是高表达的[37]。

2.4　建立基因调控网络模型

聚类方法有助于提取共同表达的基因簇,不过要想进一步了解它们之间的调控关系,需要建立调控网络模型,研究人员现已建立了各种调控网络模型,来描述这种关系。这些模型有抽象的 Kauffman 随机布尔网络模型,也有具体的生化互作模型-随机动态模型[41],有连续[42] 的也有离散的[43],还有空间[44] 和非空间模型[45]。这些模型各有优缺点,但都只局限于建立较小规模的简单网络模型,建立较大规模的能够

反映细胞实际情况的基因表达谱网络模型还有很多困难。如何建立大规模的基因调控网络模型将是基因表达谱研究的重要任务之一,也是后基因时代的主要研究方向之一。我们在基因调控网络的构建上也做了一些探索性工作,我们的思路是:首先将基因表达谱通过聚类的方法分为不同的功能模块,然后在每个功能模块的基础上构建一个个小的子网,最后将这些子网连接起来,当然,如何将这些零散的小网连接起来,又让它们具有生物学意义的困难是不言而喻的。建立大规模的网络模型是基因表达谱研究的另一重要任务,是后基因时代的主要研究方向之一。

2.5 逆向工程

逆向工程又叫反向模型,它是从所给的大量基因表达谱数据中推出未知的潜在的网络模型。通过基因表达谱聚类,可以抽取有用的信息,不过聚类技术只能告诉我们哪些基因是共同被调控的,至于谁调控谁,它就无能为力了。逆向工程的目的就是通过建立一个反向模型,来推导出它们之间的因果关系。逆向工程的另一个重要意义是可以指导生物实验的设计,避免生物学实验的盲目性,提高实验的针对性,降低成本,提高效率,而实验反过来又有利于模型的完善和证实。

REFERENCES(参考文献)

[1] Marquesde sa JP. Pattern Recognition Concepts. Methods and Applications. EISBN:3-540-42297-8

[2] Michael B Eisen, Paul T Spellman, Patrick O Brown *et al*. Cluster analysis and display of genome wide expression patterns. *PNAS USA*, 1998, **95**(25): 14 863-14 868

[3] Brazma A, Vilo J. Gene expression data analysis. *FEBS Letters*, 2000, **480**(1): 1724

[4] Joaquin Dopazo, Jose Maria Carazo. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topologyof a phylogenetic tree. *Joural of Molecular Evolution*, 1997, **44**: 226-233

[5] Fritzke B. Growing cell structures-a self-organizing network for unsupervised and supervised learning. *Neural Networks*, 1994, **7**(9): 1441-1460

[6] Luo Feng, Khan L. Ontology construction for information selection. Technical Report, Computer Science Department, University of Texas at Dallas, 2002

[7] Luo Feng, Khan Latifur, Yen I-Ling *et al*. A dynamical growing self-organizing tree (DGSOT) for hierarchical clustering, accepted by Bioinformatics(2004)

[8] Sugiyama A, Kotani M. Analysis of gene expression data by using self-organizing maps and k-means clutering. Neural Networks, 2002, IJCNN '02. Proceedings of the 2002 International Joint Conference, 2002, 2: 1342-1345

[9] Xiong M, Jin L, Li W *et al*. Computational methods for gene expression based tumor classification. *Biotechniques*, 2000, **29**(6): 1264

[10] Li Wuju, Xiong Momiao. Tclass: Tumor classification system based on gene expression profile. *Bioinformatics*, 2002, **18**(2): 325-326

[11] Zhang Xinying, Myers Chad L, Kung SY. Cross-weighted fisher discriminant analysis for visualization of dnamicroarray data. Accepted by ICASSP 2004

[12] Zhang Heping, Yu Chang-Yung, Singer Burton *et al*. Recursive partitioning for tumor classification with gene expression microarray data. *PNAS USA*, 2001, **98**(12): 6730-6735

[13] Andrew D Keller, Michel Schummer, Lee Hood *et al*. Bayesian classification of DNA array expression data. Technical Report, University of Washington, August 2000

[14] Zhou Xiaobo, Wang Xiaodong, Dougherty ER. A Bayesian approach to nonlinear probit gene selection and classification. *Journal of the Franklin Institute*, 2004, **341**(1,2): 137-156

[15] Nir Friedman, Michal Linial, Iftach Nachman *et al*. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 2000, **7**(3,4): 601-620

[16] Ji XL, Li L J, Sun Z R. Mining gene expression data using a novel approach based on hidden Markov models. *FEBS Letters*, 2003, **542**(1-3): 125-131

[17] Schliep A, Schonhuth A, Steinhoff C. Using hidden Markov models to analyze gene expression time course data. *Bioinformatics*, 2003, **19**(Suppl.1): i255-i263

[18] Simon haykin. Neural Networks, A Comprehensive Foundation, 2nd Edition, Pearson Education. Inc, 1999

[19] Yang Yonggao, Chen JX, Kim Woosung. Gene expression clustering and 3D visualization. *Computing in Science & Engineering*, 2003, **5**(5): 37-43

[20] Tamayo P, Slonim D, Mesirov J *et al*. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *PNAS USA*, 1999, **96**(6): 2907-2912

[21] Ressom HH, Wang D, Natarajan P. Adaptive double self-organizing map and its application in gene expression data. *Proceedings of the International Joint Conference on Neural Networks*, 2003, **1**: 39-44

[22] Dopazo J, Carazo JM. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *J Mol Evol*, 1997, **44**(2): 226-233

[23] Brown PS Michael, Grundy Noble William, Lin David *et al*. Knowledge based analysis of microarray gene expression data by using support vector machines. *PNAS, USA*, 2000, **97**(1): 262-267

[24] Fujarewicz K, Kimmel M, Rzeszowska-Wolny J *et al*. Improved classification of gene expression data using support vector machines. *J Med Inf Technol*, 2001, **6**: MI9-MI19

[25] Furey TS, Cristianini N, Duffy N *et al*. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 2000, **16**(10): 906-914

[26] Ben-Dor A, Bruhn L, Friedman N *et al*. Tissue classification with gene expression profiles. *J Comput Biol*, 2000, **7**(3-4): 559-583

[27] Sridhar Ramaswamy, Pablo Tamayo, Ryan Rifkin *et al*. Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS USA*, 2001, **98**(26): 15149-15154

[28] Chu Feng, Wang Lipo. Gene expression data analysis using support vector machines. *Proceedings of the International Joint Conferece on Neural Networks*, 2003, **3**: 2268-2271

[29] Chen Xue-wen. Gene selection for cancer classification using bootstrapped genetic algorithms and support vector machines. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003, pp.504 – 505

[30] Valentini G, Muselli M, Ruffino F. Bagged ensembles of support vector machines for gene expression data analysis. *Proceedings of the International Joint Conference on Neural Networks*, 2003, **3**: 1844 – 1849

[31] Li X(李　霞), Zhang TW(张田文), Li L(李　丽) *et al*. Efficiency of feature gene selection based on decision tree to pattern classification SVM. *Chinese Journal of Biomedical Engineering*(**中国生物医学工程学报**), 2004, **23**(1): 66 – 73

[32] Li X(李　霞), Rao SQ(饶绍奇), Zhang TW(张田文) *et al*. An ensemble method for gene discovery based on DNA microarry data. *Science in China*(Series C)(**中国科学** C **辑**), 2004, **34**(2): 195 – 202

[33] Pierre B, Anthony DL. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 2001, **17**(6): 509 – 519

[34] Ishii M, Hashimoto S, Tsutsumi S *et al*. Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics*, 2000, **68**(2): 136 – 143

[35] Evans SJ, Datson NA, Kabbaj M *et al*. Evaluation of affymetric gene chip sensitivity in rat hippocampal tissue using SAGE analysis. *Eur J Neurosci*, 2002, **16**: 409 – 413

[36] Yuen Tony, Wurmbach Elisa, Pfeffer L Robert *et al*. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res*, 2002, **30**(10): e48

[37] Kowalski J, Denhardt DT. Regulation of the mRNA for monocyte-derived neutrophil-activating peptide indifferentiating HL60 promyelocytes. *Mol Cell Biol*, 1989, **9**: 1946 – 1957

[38] Marcotte EM, Pellegrini M, Thompson MJ *et al*. A combined algorithm for genome-wide prediction of protein function. *Nature*, 1999, **402**: 83 – 86

[39] Drawid A, Gerstein M. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J Mol Biol*, 2000, 301: 1059 – 1075

[40] Califano Andrea, Stolovitzky Gustavo, Tu Yuhai. Analysis of gene expression microarrays for phenotype classification. *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, 2000, **8**: 75 – 85

[41] Arkin Adam, Ross John, McAdams H Harley. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells. *Genetics*, 1998, **149**: 1633 – 1648

[42] Savageau MA. Rules for the evolution of gene circuitry. *Pacific Symposium on Biocomputing*, 1998, **3**: 54 – 65

[43] Liang SD, Fuharman S. Somogyi Rolan. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing*, 1998, **3**: 18 – 29

[44] Mjolsness E, Sharp DH, Reinitz J. A connectionist model of development. *J Theor Biol*, 1991, **152**: 429 – 454

[45] Patrik D'haeseleer, Wen Xiling, Fuhrman Stefanie *et al*. Linear modeling of mRNA expression levels during CNS development and injury. *Pacific Symposium on Biocomputing*, 1999, **4**: 41 – 52