新技术讲座

# 基因表达聚类分析技术的现状与发展*

杨春梅[1)]** 万柏坤[1)] 高晓峰[2)]

（[1)]天津大学精密仪器与光电子工程学院，天津 300072；[2)]摩托罗拉（中国）电子有限公司，天津 300457）

摘要 随着多个生物基因组测序的完成、DNA 芯片技术的广泛应用，基因表达数据分析已成为后基因组时代的研究热点. 聚类分析能将功能相关的基因按表达谱的相似程度归纳成类，有助于对未知功能的基因进行研究，是目前基因表达分析研究的主要计算技术之一. 已有多种聚类分析算法用于基因表达数据分析，各种算法因其着眼点、原理等方面的差异，而各有其优缺点. 如何对各种聚类算法的有效性进行分析、并开发新型的、适合于基因表达数据分析的方法已是当务之急.

关键词 DNA 芯片，基因表达，聚类分析，非监督聚类，监督聚类，基于模型聚类

学科分类号 Q78

DNA 芯片技术能对大量的基因表达谱进行同步、快速测量，同时提供数千条基因的表达水平[1,2]. 例如，一个只有 5 个样品和 2 个复制的小规模实验就可以得到近 100 000 个数据点[3]. 如何对这些数据进行注释，从中提取有用的生物学信息，已成为后基因组时代的研究热点.

就目前所知，基因表达数据至少可以从以下三个逐渐复杂的层次上进行分析[4]：第一，分析单条基因的表达水平，此时人们着眼于一定实验条件下每条基因的表达是否与对照不同；第二，考虑基因组合，将基因分成不同的类别以研究基因的共同功能、相互作用以及协同调控等；第三，尝试推断潜在的调控区域及基因网络，以从机理上解释观察到的图谱. 目前对基因表达数据的研究主要集中于第二层，而第三层则是更深一步的研究目标.

已有多种数据挖掘和信息处理技术被应用于基因表达数据分析，主要包括：聚类分析、多元统计、模式识别及神经网络等几大类. 功能相关的基因通常共同表达，检测具有相似表达谱的基因群是研究基因功能的一种有效方法. 因此，基因表达数据分析的一类重要方法就是聚类分析. 聚类分析是指将一组个体按其相互间的相似程度归入几个子类，根本思想是确定类群，使同一类内的各个体间差异最小，而不同类间的差距最大. 聚类分析的首要目标是将表达谱相似的基因归纳成类，然后聚焦于那些可能参与某些生物过程的基因群，对这些类进行生物学注释，同时获得新的生物学知识[3].

基因表达聚类分析一般分三个步骤：数据标准化、数据筛选和模式识别. 为有效地比较表达水平，首先必须将数据标准化，然后剔除表达水平低于给定阈值的基因，以减少数据量或降低维数，最后，寻找数据中的模式，为表达谱赋予一定的生物学功能. 在基因表达矩阵中，不同的基因有不同的强度值范围，单个的强度值并没有太大的意义，而相对值则更能说明问题. 在数千条基因中，并非所有的基因都对类别的划分做出同样的贡献，实际上有些基因可能没有贡献. 因此，需要排除那些对实验条件几乎不起反应的基因. 不管使用哪种方法，进行什么分析，排除实验过程中表达水平不变的基因都是明智的. 选择富含信息的基因是降低数据的复杂性、提高信噪比的第一步（Viann C. Gene Expression. http:/www. cbi. pku. edu. cn，2000）.

聚类算法有多种不同的分类方式：依其先验知识的有无，可分为非监督聚类和监督聚类；依其是否假定内在的概率框架，又可分为基于判断的聚类和基于模型的聚类. 非监督聚类算法不需要基因实际功能类别方面的先验知识，通常使用相似性或距离测量之类的准则来划分类别；监督聚类方法则事先指定经学习训练后哪些数据应归为一类. 基于判断的聚类采用直观推理步骤，而基于模型聚类则假定数据的内在概率框架.

* 天津市重点学科基金资助（2000-31）.

** 通讯联系人.

Tel：022-27401410，E-mail：yangcm2000@ eyou. com

收稿日期：2003-04-03，接受日期：2003-07-31

## 1 非监督聚类

非监督聚类使用递归的分割方法将基因或样品划分为统计上有意义的类，几乎不采用先验知识[5]. 对于先验知识难以获得或者不完全的情况下，这种方法就显得特别重要，而大多数基因表达实验正是如此[6]. 分层聚类、K-means 聚类及自组织聚类等非监督聚类方法已广泛应用于各种基因阵列数据的分析.

### 1.1 学习准则

非监督聚类需要一个准则来衡量两个表达谱的相似程度，选择一种合适的比较准则是至关重要的[7]. 基因表达数据分析中普遍使用的准则为欧氏距离和 Pearson 相关系数[8].

欧氏距离测量空间中两个点的绝对距离，故同时考虑了矢量的方向和幅度. 若直接使用原始数据进行计算，则表达谱幅度相似的基因将被认为是相似的. 但生物学上更倾向于寻找表达水平不同而表达谱形状相似的基因，故使用欧氏距离前需对数据作适当的转换，如重新进行标度或作归一化处理. Pearson 相关系数从本质上说是测量两个表达矢量所指方向的相似性，处理时将其视为单位矢量，因而对幅度的变化不敏感. 但若两个不很相似的基因表达谱在某一个突出的峰或谷特别相关的话，Pearson 相关准则可能给出假阳性. 相关系数的一个有趣的性质是它可用来检测负相关的基因.

### 1.2 分层聚类

分层聚类（hierarchical clustering）是一种经典的聚类方法[9,10]，它用二元树形状的系统树图（dendrogram）来描述数据间的关系，其中最相似的谱形成嵌套子集中的一层. 分层聚类开始时将每一矢量当成一类，反复地将最近的两类合并为一类. 在一定水平上终止系统树，可以将数据分成类.

分层聚类的性能随待分类基因数的平方下降，因其需要计算每一对基因表达谱间的距离，若分析大量的基因，计算工作量将非常大. 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[11]. 系统树图也只是原始数据的一维排列，其中每一项有两个邻项，多种关系无法在单一维数上得到表现. 而且，分层聚类缺乏稳健性，解也可能不是唯一的，数据的输入次序对结果也可能产生影响. 若待比较的关系是对称的，则会产生类质同像现象（isomorphism），即，系统树图的树枝间可以互换，产生一个节点次序不同的等价系统树图，如图 1 所示. 另外，这种分类方法也易于出错，可能将那些属于不同类别但很接近的节点误分. 当然，系统树图只限于描述分层组织的数据，并不适合于显示非分层聚类的结果，更不能用于分析本身不是分层组织的数据.

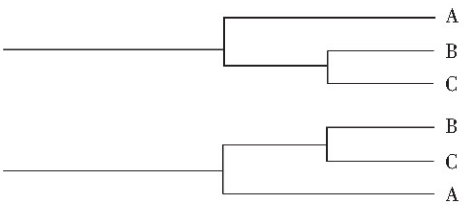


**Fig. 1 Dendrograms are isomorphic**

图 1 同构的系统树图

### 1.3 K-means 聚类

K-means 聚类是一种很受欢迎的实时聚类算法[12]，简便且能处理大量数据，其目标是在最小化误差函数的基础上将数据划分为预定的类数 $K$. 在运行算法前必须先指定类数 $K$ 和迭代次数或收敛条件，开始随机指定 $K$ 个质心，根据一定的距离准则（如欧氏距离），将每一个表达谱分配到最接近或“相似”的质心，形成类，然后以每一类的平均矢量作为这一类的质心，重新分配，反复迭代直到类收敛（类的质心不变）或达到最大的迭代次数.

K-means 聚类算法的关键问题是如何初始化质心和怎样更新质心. 由于有多种初始化 $K$ 类的可能，故难于选择最优化的结果. 另外，对有些实验，无法确定预期的类数. 而且，也没有很好的方法来选择算法应运行的确切迭代次数.

### 1.4 自组织映射

自组织映射（self-organizing maps，SOMs）是在 K-means 聚类算法基础上发展起来的一种非监督的神经学习方法[13]，算法目标是找到能描述输入数据集的原型矢量，同时将高维输入空间连续映射到网格上. 这个网格由一定数目的神经元组成，一般为六角型或方形排列的二维网络，从而易于显示. 给网格节点（神经元）赋予一定权重，来表示类的质心. 如同 K-means 聚类，计算一种距离确定各输入矢量的匹配节点，并由输入矢量调整匹配节点及其邻域的权重，这与 K-means 聚类不同. 经反复学习，模拟矢量以有序的方式描述数据的概率

分布. 训练好的 SOM 网格节点上已分配好相应的基因表达谱，节点的权重矢量代表相应类内表达谱的平均，且相邻节点表示相似的类，类差别越大，其节点相距越远.

SOMs 聚类克服了 K-means 聚类的一些缺点[14]：它应用类间的全局关系，能提供大数据集内相似性关系的综合看法，便于研究数据变量值的分布及发现类结构. 而且，它具有更稳健更准确的特点，对噪声稳定，一般不依赖于数据分布的形状，同时能高性能地显示大量数据. 但 SOMs 是一种拓扑保留的神经网络，产生不均衡分类[15]. 若不相关数据（如不变的“平坦”谱）或某种特定形式的谱过多，在 SOMs 的输出中，这种数据将占据大部分类，从而感兴趣的数据只能处于少数类，分辨率就可能很低. 因此，在应用 SOMs 聚类前，一般要对数据进行筛选，或使用自组织树映射算法（self-organizing tree maps algorithm，SOTA）. SOTA 是一种树形布局的 SOM[16]，它将分层聚类和 SOM 的优势相结合，运用于基因表达谱时能避免这些方法中的一些问题.

### 1.5 超顺磁性聚类

超顺磁性聚类（super-paramagnetic clustering，SPC）是一种基于模拟非均匀铁磁物质的物理特性的聚类方法，将数据聚类问题视为检验不均匀 Potts 模型的平衡特性[17]. 给每一数据点分配一个 Potts 自旋子，相邻数据点间引入强度随距离下降的相互作用函数. 非均匀 Potts 模型系统随温度变化表现出三相：在低温下，所有自旋子呈现完全有序的排列，系统为铁磁相；随着温度的升高，小区域自旋子形成磁化“颗粒”，附属于同一“颗粒”者相互间产生强耦合，而无关者间相互作用很弱，不同“颗粒”的排列呈无序状态，为超顺磁相；在高温下，系统不表现任何有序性，为顺磁相. 在超顺磁相的转换温度下，磁化率表现出显著的峰值. 原则上，超顺磁相可以有一系列的转换点. 随着温度的升高，系统可以首先分裂为两类，其中每一类又可以分裂为更多的子类，依此类推. 这样，数据就分层组织为类.

SPC 算法的优点是对噪声及初始化不敏感，因为类由系统的综合性质产生. 由磁化率的峰值很容易鉴别主要的分界，从而能清楚显示类的构成和分界，并且在每一个分辨率上能自动确定类数.

### 1.6 双向聚类

基因表达数据为矩阵形式，其行代表基因，列代表样品如组织或实验状态等，可以分别以行或列作为矢量在样本空间或基因空间进行分析，双向聚类（two-way clustering）方法则同时在两个方向上对其进行聚类分析.

#### 1.6.1 耦合双向聚类

在一个典型的微阵列试验中，同时测量数千条基因的表达水平. 而一个感兴趣的生物过程可能只包含基因中的一小部分，且过程也可能只发生于少数样品中. 耦合双向聚类（coupled two-way clustering，CTWC）的主要思想是鉴别基因和样品的子集[18]，以减少应用的特征和聚类的数据点，降低其他样品和基因引入的噪声. 即：分别寻找特征（基因或样品）和个体（样品或基因）的一个小子集 $F_i$ 和 $O_j$，组成（$F_i$，$O_j$）对. 当 $O_j$ 中个体只用 $F_i$ 中特征描述时，聚类能产生稳定、显著的划分.

CTWC 通过迭代的聚类过程来寻找这样的子集，其中只考虑和测试上一次迭代中确认的稳定类的基因（或样品），从而大大提高了计算的可行性. 开始，将所有基因（$G^0$）和样品（$S^0$）作为特征和个体运行标准双向聚类算法. 用 $G_i^1$ 和 $S_j^1$ 表示这一步找到的稳定基因和样品类，（$G_i^1$，$S_j^1$）定义了表达数据的一个子阵. 对这一子阵进行双向聚类，产生的稳定基因（或样品）类表示为 $G_i^2$（或 $S_j^2$）. 反复进行，直到不再有符合条件（如稳定性、临界尺寸等）的新稳定类出现终止.

#### 1.6.2 相关双向聚类

基因表达矩阵通常有数千行，列数则不到一百，故若直接应用传统的聚类方法对样品聚类，很难得到满意的结果[19]. 为有效地聚类样品，首先应对样品空间进行降维处理. 相关双向聚类（interrelated two-way clustering）即是针对这一问题提出的非监督聚类算法.

相关双向聚类算法的目的是在发现重要基因模式的同时，对样品分类[3]. 即：先找到与实验条件高度相关的基因子集，再利用这些基因将样品分类. 相关双向聚类也是通过迭代过程实现：首先，采用常规的聚类算法（如 K-means，SOMs 等）将所有基因聚为 $K$ 类（$G_i$，$i=1\sim k$）；其次，在每一类 $G_i$ 内，对样品聚类，如最典型的情况聚为二类，表示为 $S_{i,a}$，$S_{i,b}$；然后，将前面两步的聚类结果进行组合，寻找异类组（heterogeneous groups），对每一异类组内的基因进行排序，去除一部分基因；最后，将剩余的基因组合起来，进行下一次循环，

直到满足终止条件结束.

双向聚类方法的优点是：通过在基因维和样品维反复聚类，可以动态地使用基因和样品间的相互关系. 反复聚类过程中，基因维的下降可以改善类的准确性，而这反过来又有助于进一步降基因维.

非监督方法寻找数据间的统计依赖性，它们常被用作揭示大数据集中隐藏信息的工具，但结果却往往受限于模型的选择和数据集中变量的选择或特征的提取. 在一个更宽广的数据分析或模型结构下，模型的选择指定了要寻找哪种依赖性，数据变量的选择、预处理与转换（特征提取）限定了数据的感兴趣或重要方面.

## 2 监督聚类

监督方法假定全部或部分表达谱具有额外信息，如基因的功能类别，或样品的疾病/正常属性，有了这些信息以后，典型的工作是建立一个分类器来预报表达谱的类别. 监督学习分为两个阶段：分类模型的训练和测试. 数据集分为两个互斥的子集，其中一个用来训练模型，另一个用来测试模型. 在基因表达数据分析中最常用的监督方法是支持矢量机（support vector machines，SVM）.

SVM 是一种基于结构风险最小化原理的统计学习方法[20]，是当前应用最广的分类技术之一. 它将输入矢量非线性地映射到高维特征空间，在新的空间中给出最优分类超平面将原始数据分开. SVM 是最大边界分类器，它选择特征空间中能区分开正负样本的边界（margin）最大的超平面，使超平面与训练点的最小距离最大化，从而避免过拟合. 距最优超平面最近的点为“支持矢量”，是训练样本的一个小子集，支持矢量的线性组合可以表示最大边界超平面.

从基因维对实验样品进行分类时，由于数据特征数远大于训练样本数，常规分类方法极易导致过拟合. SVM 则能有效地避免这个问题，而且具有很好的推广能力. 即使有缺失的数据点，SVM 也可以通过使用软边界寻找到最优分界超平面来区分数据（Lu G. Gene classification with support vector machines. http://www.cse.ucsc.edu/~pyang/pyang/biofinal.html）.

虽然分层聚类、SOMs 及 SVM 都借助距离函数来比较基因表达测量，SVM 允许使用多种形式的函数，尤其是它可以使用高维特征空间中的距离函数，从而可能考虑基因表达测量间的相关性. 另外，由于 SVM 采用先验知识区分基因，即使相关基因按距离函数相隔很远，也能较准确判断类别.

## 3 基于模型的聚类方法

基于判断的算法对聚类分析中遇到的一些实际问题，如“选择哪种聚类方法?”，无法给出系统的指导. 基于模型的聚类方法（model-based clustering）假定聚类个体是由一种内在的概率框架产生，是基于判断聚类算法的主要替代[21]. 在这一概率框架下，聚类方法的选择和类数的确定等问题转化为模型选择问题，因而比判断型算法具有更大的优势，也因此成为近来新型聚类方法研究的焦点. 目前，应用到基因表达聚类分析中的概率模型主要有混合模型和隐马尔可夫模型（hidden Markov model，HMM）.

### 3.1 基于混合模型聚类（**mixture model-based clustering**）

这类算法假定数据中蕴含的每一组（成分）由一种内在的概率分布混合产生[22]. 例如，高斯混合模型中[23]，独立多元观测 $y_1$，$y_2$，…，$y_n$ 组成的数据 $Y$ 中每一成分 $k$ 由参数为 $u_k$（均值矢量）和 $\sum_k$（协方差矩阵）的多元正态分布模拟：

$$f_k(y_i \mid u_k, \sum_k) = \frac{\exp\left\{-\frac{1}{2}(y_i - u_k)^T \sum_k^{-1}(y_i - u_k)\right\}}{\sqrt{\det(2\pi \sum_k)}} \quad (1)$$

算法的目标是：由数据预测参数 $u_k$ 和 $\sum_k$，并确定相应于这些参数预测的类.

为了便于参数预测，对 $\sum_k$ 进行特征值分解：$\sum_k = \lambda_k D_k A_k D_k^T$. 其中，$D_k$ 为特征向量组成的正交矩阵，确定成分的取向，$A_k$ 为对角矩阵，元素与 $\sum_k$ 的特征值成正比，确定成分的形状，$\lambda_k$ 为标量，确定成分的体积. 约束或改变部分参数，可以得到这一框架下的一组模型，如等体积球模型（EI：$\sum_k = \lambda I$，$I$ 为单位矩阵）、不等体积球模型（VI：$\sum_k = \lambda_k I$）等，以适应数据特征的变化.

每一种协方差矩阵与类数组合情况相应于一个不同的概率模型，由贝叶斯信息准则（BIC）估计每一种模型下数据被观测到的概率，计算结果的 BIC 得分，最后选择 BIC 得分最大的模型和类数.

Yeung 等[24]应用不同约束的高斯混合模型，对多个具有外部分析准则的基因表达数据和人工合成数据进行聚类分析，聚类结果的可靠性和稳定性

都得到显著的改善.

**3.2 基于隐马尔可夫模型聚类（HMM-based clustering）**

最近，Ji 等[25]新颖地将 HMM 引入基因表达数据聚类分析. 他们假定每一基因表达谱由一条马尔可夫链以一定的概率产生，基因表达谱经标准化处理后按（2）式转化为一个表达波动序列. 其中 $N$ 为时间点数，$E$ 为基因在每一时间点的表达水平，$S$ 为序列转换值，$a$ 为容限值（如取 0.05）. 这样，cDNA 微阵列实验得到的 $N$ 个时间点的基因表达谱，被转换为由字符集 {0，1，2} 构成的 $N$-1 点的波动序列，序列中每个字符反映了下一时间点表达水平的变化情况，而整个序列则描述了基因表达的波动情况.

$$S_i = \begin{cases} 0 & \text{if} \quad |E_i - E_{i+1}| < a \\ 1 & \text{if} \quad E_{i+1} - E_i \geqslant a \\ 2 & \text{if} \quad E_j - E_{i+1} \geqslant a \end{cases} \quad 1 \leqslant i \leqslant N-1 \tag{2}$$

对所有的表达波动序列构建一 HMM，其主链由 $N$ 个状态构成，每一状态相当于一个实际细胞状态. 每一状态可按一定分布产生一个字符（0，1 或 2），代表细胞状态的调节方向. 为便于模拟，他们虚拟了一个“BEGIN”状态和一个“END”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“BEGIN”状态开始，选择跃迁到一个细胞状态，并依据这一状态的分布产生字符（0，1 或 2），然后再选择跃迁到另一状态，并产生下一个字符，持续这一过程，直到“END”状态. 跟随这个过程就产生一个表达波动序列. 采用 Baum-Welch 方法训练模型，并应用前向 - 后向算法计算序列的概率. 最后，依据概率确定序列的类属.

Ji 等[25]应用这一 HMM 模型对多个基因表达数据进行聚类，结果表明，这一方法不仅能确定正确的类数，而且能揭示不同功能组间的关系，并可进一步用于寻找多功能基因、区别功能基因和调控基因. 但是，由于这一方法基于马尔可夫过程，即观测只取决于先前和当前状态，目前只限于分析时间序列数据.

## 4 总结与展望

综上所述，基因表达数据分析是一个成长的跨学科领域，它与数理学、医学、计算机科学等许多领域交叉. 随着多个生物基因组测序的完成、DNA 芯片技术的广泛应用，基因表达信息的日益积累，数据分析将变得越来越重要. 基因表达数据分析已成为后基因组时代的研究热点. 对该领域研究工作的一个主要挑战是：开发通过靶细胞或组织的表达谱能预测表型反应的方法，以期建立细胞的转录状态与其对刺激反应倾向之间的关系，鉴别出能够影响反应的基因（如受体、抑制路径等）.

基因表达数据分析亦是一个全新的领域，目前尚无分析这些数据的最佳方法，通常是统计和分类方法相结合以更好地理解结果. 由于聚类方法并不足以对不同的分类结果进行系统的比较，开发能胜任这一特定工作的方法具有极其重要的意义，这些工具需要执行特定的方法来鉴别和刻画两个或多个子系统间的相似性或差异，并提供统计显著性评价. 由于分析基因的数量极大，系统间相似性可能是偶然出现的，能评价这种相似显著性和有效性的算法十分重要，并急待开发. 因此，如何对各种聚类算法的有效性进行分析、并开发新型的、适合于基因表达数据分析的方法已是当务之急.

## 参 考 文 献


1 Mark S，Dari S，Renu H，*et al*. Parallel human genome analysis：Microarray-based expression monitoring of 1 000 genes. Proc Natl Acad Sci USA，1996，**93**（20）：10 614 ~10 619

2 Iyer V R，Eisen M B，Ross D T，*et al*. The transcriptional program in the response of human fibroblasts to serum. Science，1999，**283**（5398）：83 ~87

3 Chun T，Li Z，Zhang A，*et al*. Interrelated Two-way Clustering：An Unsupervised Approach for Gene Expression Data Analysis. In：IEEE Comput Soc eds. Proceedings 2nd Annual IEEE International Symposium on Bioinformatics and Bioengineering（BIBE 2001），Los Alamitos：IEEE Comput Soc，2001. 41 ~48

4 Pierre B，Anthony D L. 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

5 Amir B，Friedman N，Yakhini Z. Class discovery in gene expression data. In：Lengauer T，eds. Proceedings of the Fifth Annual International Conference on Computational Molecular Biology（RECOMB 2001）. New York：ACM，2001，31 ~38

6 Amir B，Ron S，Zohar Y. Clustering gene expression patterns. Journal of Computational Biology，1999，**6**（3/4）：281 ~297

7 Kaski S. Learning metrics for exploratory data analysis. Neural Networks for Signal Processing，2001，**XI**：53 ~62

8 Joaquin D，Edward Z，Ilaria D，*et al*. Methods and approaches in the analysis of gene expression data. Journal of Immunological Methods，2001，**250**：93 ~112

9 Michael B，Eisen M B，Paul T，*et al*. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA，1998，**95**（25）：14 863 ~14 868

10 Brazma A，Vilo J. Gene expression data analysis. FEBS Letters，2000，**480**（1）：17 ~24

11 David R G, Michael S, Jacques V H. Interactive visualization and exploration of relationships between biological objects. TIBTECH, 2000, **18** (12): 487 ~494
12 Tavazoie S, Hughes D, Campbell M J, *et al*. Systematic determination of genetic network architecture. Nature Genet, 1999, **22** (3): 281 ~285
13 Kohonen T. The self-organizing map. Proc IEEE, 1990, **78** (9): 1464 ~1480
14 Tamayo P, Solni D, Mesirov J, *et al*. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA, 1999, **96** (6): 2907 ~2912
15 Fritzke B. Growing cell structures——a self-organizing network for unsupervised and supervised learning. Neural Networks, 1994, **7** (9): 1441 ~1460
16 Dopazo J, Carazo J M. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. J Mol Evol, 1997, **44** (2): 226 ~233
17 Marcelo B, Shai W, Eytan D. Data clustering using a model granular magnet. Neur Comput, 1997, **9** (8): 1805 ~1842
18 Gad G, Erel L, Eytan D. Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci USA, 2000, **97** (22): 12079 ~12084
19 Hartigan J, Wong M. Algorithm AS136: a k-means clustering algorithms. Applied Statistics, 1979, **28** (1): 100 ~108
20 Berand H, Alessandro V, Tomaso P. Learning and vision machines. Proc IEEE, 2002, **90** (7): 1164 ~1177
21 Fraley C, Raftery A E. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc, 2002, **97** (458): 611 ~631
22 McLachlan G J, Bean R W, Peel D. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics, 2002, **18** (3): 413 ~422
23 Yeung K Y, Fraley C, Murua A, *et al*. Model-based clustering and data transformations for gene expression data. Bioinformatics, 2001, **17** (10): 977 ~987
24 Yeung K Y, Medvedovic M, Bumgarner R E. Clustering gene expression data with repeated measurements, Genome Biology, 2003, **4** (5): R34
25 Ji X L, 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

# Actuality and Development of The Clustering Technologies for Gene Expression*

YANG Chun-Mei[1)**], WAN Bai-Kun[1)], GAO Xiao-Feng[2)]

([1)] *College of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin* 300072, China;

[2)] *Motorola (China) Electronics Ltd. Tianjin* 300457, China)

**Abstract** With many genomes completed and extensive applications of DNA chips, analysis of the gene expression data has become a hotspot in the postgenomic age. Clustering is the art to group genes with related functions according to the similarities in their expression profiles. A number of clustering algorithms have been developed for gene expression data analysis. For their respective focuses and principles, every method has its own advantages and disadvantages, which are reviewed. How to evaluate the capabilities of these algorithms, and to develop new methods more suitable for gene expression analysis, should be urgent.

**Key words** DNA chip, gene expression, clustering analysis, unsupervised clustering, supervised clustering, model-based clustering

* This work was supported by a grant from The Tianjin Key Subject Fund (2000-31).

** Corresponding author. Tel: 86-22-27401410, E-mail: yangcm2000@eypu.com

Received: April 3, 2003 Accepted: July 31, 2003