基因聚类分析中数据预处理方式和相似度的选择*

杨春梅[1] 万柏坤[1**] 高晓峰[2]

1. 天津大学生物医学工程与科学仪器系，天津 300072；
2. Motorola (China) Electronics Ltd. 天津 300457

摘要 聚类分析是基因表达数据分析研究的主要技术之一. 选择恰当的数据预处理方式和相似性度量，是获得高质量聚类结果的前提. 采用具有外部标准的基因表达数据集，分别以Pearson相关系数和Euclidean距离为相似度，以校正的Rand指数比较了使用分层聚类、K-均值聚类和SOMs聚类算法对经过行标准化、列标准化和对数化处理后数据的基因聚类质量. 结果表明：K-均值聚类和SOMs聚类质量显著优于分层聚类，SOMs聚类质量稍好于K-均值聚类. 而且，分层聚类偏好于Pearson相关系数准则和行标准化处理，而应用K-均值聚类和SOMs算法时，则最好是对数据进行对数化处理、并以Euclidean距离为相似性度量准则. 上述研究结论将为基因表达聚类分析的实施提供有价值的参考依据.

关键词 基因表达 聚类分析 数据预处理 相似度 Rand指数

随着人类基因组计划的实施和人基因组工作草图的完成，生命科学已进入了产生大量基因表达数据、生命信息量爆炸性增长的时代. 呈指数增长的生命信息也为生命、数学、物理、化学和信息等科学领域提供了巨大的研究平台，科学家的主要工作就是从这些数据信息中去探索生命的奥秘. 其中，DNA芯片技术因其能对大量的基因表达谱进行同步、快速测量，同时提供成千上万条基因的表达水平，而被广泛应用于生命科学的各个领域，产生了海量的基因表达数据[1,2]. 如何分析和处理这些数据，从中提取有用的生物学或医学信息，已成为后基因组时代研究的瓶颈[3,4].

聚类分析技术是目前基因表达分析研究的主要计算技术之一[4,5]. 它能将功能相关的基因按表达谱的相似程度归纳成共同表达类别，有助于对基因功能、基因调控、细胞过程及细胞亚型等进行综合研究. 有多种聚类算法已被成功地用于基因表达分析，如分层聚类(hierarchical clustering)，K-均值聚类(K-means clustering)，主成分分析(principal component analysis，PCA)及自组织映射(self-organizing maps，SOMs)等[6,7]. 但由于不同聚类算法，甚至同一聚类算法使用不同参数，一般都将产生不同的类别(clusters). 故生物学家面对大量基因表达数据集的第一个棘手问题即是如何选择合适的聚类算法. 然而，目前尚未有达成共识的统一性指导方针[8].

基因表达聚类分析的第一步是对芯片试验所产生的基因表达矩阵进行预处理，以确保基因表达水平的可比性[3]. 目前，数据预处理方法可分为对原始数据作标准化处理和取表达比率的对数值两大类. 其中标准化包括对行(基因表达矢量)和列(样品表达矢量)进行标准化处理两种情况. 此外，常规的基因表达聚类分析算法基于个体间的相似性度量(简称相似度)来衡量两个表达谱的

2005-07-21 收稿，2005-09-23 收修改稿

* 天津市重点建设学科基金资助（批准号：2001-31）

** 通讯作者，E-mail：bkwan@tju.edu.cn

相似程度，如Pearson相关系数、Euclidean距离、Jackknife相关、信息熵等[9]. 其中应用较成功的是Pearson相关系数和Euclidean距离[10,11]. 由于不同的聚类算法可能偏好不同的数据预处理方式和相似度，因此，在基因表达聚类分析时，选择合适的数据预处理方法和相似度至关重要，是获得正确聚类结果的前提[12]，也是生物学家面对大量基因表达数据集的第二个棘手难题.

本文针对上述两个难题，采用具有外部“金标准”的数据集，比较了几种常用聚类算法在不同数据预处理方式和相似性度量准则下的分析质量，以期为基因表达聚类分析的实施提供有价值的参考依据.

1 材料与方法

1.1 基因表达数据

本文采用下述具有外部标准的基因表达数据集：

(1) 酵母孢子化数据集(Spor)

Chu等[13]应用含97%已知或推测酵母(*Saccharomyces cerevisiae*)基因的DNA微阵列研究了酵母双倍体细胞孢子化过程中基因表达情况，在减数分裂早期、中期和末期的7个时间点测量了每一条基因mRNA转录水平变化，并对比研究了生长期细胞Ndt80异常表达和缺失导致的基因表达变化，共得到10个实验样品的表达数据. 实验中，他们同时通过显微镜对基因表达情况进行了观测，从形态学信息中获得了7个显著不同的功能表达模式，为数百条未知基因的潜在功能提供了线索. 本文选取其中分别属于6个不重叠功能表达模式的161条基因表达谱组成数据集(记为Spor数据集，161×10表达矩阵)，并以这6个功能类作为外部标准类(external standard classes).

(2) 酵母半乳糖数据集(Gal)

为探测酵母半乳糖利用GAL通路中基因表达调控信息与基因和蛋白质间的相互作用，Ideker等[14]将野生型(wt)和分别删除9条GAL基因之一的9个突变株(gal1△，gal5△，gal7△，gal10△，gal3△，gal4△，gal6△，gal80△，gal12△)在有或无2%半乳糖(+gal，−gal)的介质中培养，然后采用含酵母全基因组的DNA微阵列检测了这20个试验样品中GAL通路受扰动时的mRNA表达水平，并使用最大似然估计法鉴别出997条差异表达基因. Yeung等[15]从Gene Ontology(GO)数据库列表中搜索出其中205条基因，分属于生物合成(蛋白质代谢与修饰)、能量通路(碳水化合物合成与分解)、核酸代谢及细胞转运等4个功能类别. 本文选取这205条基因的表达矢量构成Gal数据集，组成205×20基因表达矩阵，并将这4个功能类作为外部标准类.

(3) 鼠中枢神经系统发育数据集(CNS)

为探索哺乳动物中枢神经系统发育中潜在的复杂自组织过程、研究基因家族间可能的功能关系，Wen等[16]基于生物学先验知识精选出对鼠颈部脊髓发育重要的4个基因家族(Neuro-Glial Markers，Neurotransmitter Receptors，Peptide Signaling，Diverse)中112条基因，采用逆转录－PCR(RT-PCR)技术检测了这112条基因在鼠颈部脊髓发育过程中9个时间点的表达水平. 本文将其组成的112×9表达矩阵作为CNS数据集，并以这4个功能类作为外部标准类.

(4) 酵母细胞周期数据(Cellcycle)

Cho等[17]采用含芽殖酵母(*S. cerevisiae*)全基因组6220个基因的cDNA微阵列监控了酵母基因组在两个细胞周期(17个时间点)中表达水平的起伏. 在有效表达(不含无法注释的负值和缺失值)基因中，他们发现416条基因表达水平表现出与细胞周期显著相关的周期性，分别在细胞周期的G1早期、G1晚期、S期、G2期和M期等5个时相达到峰值. Yeung等[12]从其中挑选出384条仅在单个时相达到峰值的基因，将同一个时相达到峰值的基因归入一个标准类，获得了5个不相交的标准类，分别对应于细胞周期的5个时相；这384条基因表达谱即构成Cellcycle的一个数据子集—Cellcycle_384数据集，5个标准类被视为外部金标准. 此外，Tavazioe等[18]通过搜索蛋白质序列数据库MIPS发现其中237条基因可分别归入4个功能类别：DNA合成与复制、中心体组织、氮及硫代谢、核糖体蛋白. 本文将这237条基因表达谱构成Cellcycle的另一个数据子集—Cellcycle_237，其4个功能类作为外部标准类.

1.2 聚类算法

本文采用分层聚类、K-均值聚类和自组织映射等算法进行基因表达数据分析. 各种算法简述如下:

(1) 分层聚类(hierarchical clustering)

分层聚类是一种经典的聚类方法[19],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. 该算法开始时将每一基因表达矢量当成一类,根据给定的相似度准则,反复地将最相近的两类合并为一类(类数相应减1),直到期望的类数 K. 因而,当反复合并过程终止时有 K 个子树,对应于K类. 分层聚类算法的输入包括基因表达矩阵(或相似性矩阵)及预定的类数 K. 采用不同的类间相似性定义方式,形成不同的分层聚类算法,常见有:单连接(single-linkage)、均连接(average-linkage)、全连接(complete-linkage)和质心连接(centroid-linkage)分层聚类,具体定义参见文献[7].

(2) K-均值聚类(K-means clustering)

K-means 聚类是一种很受欢迎的实时聚类算法[6],简便且能处理大量数据. 其目标是:在最小化误差函数基础上,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. 开始先指定 K 个质心,根据一定的相似度准则,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.

质心初始化是K-均值聚类算法的关键之一. 一种初始化方法是随机选择 K 个基因表达矢量作为初始质心,另一种方法是使用其他聚类方法(比如,均连接分层聚类)得到的类中心作为初始质心[15]. 为了说明质心初始化方式对聚类结果的影响,我们以 Cellcycle_237 数据集为例,分别计算了标准类数下(这里为4类)100次随机初始化(random initialization)和均连接初始化(hierarchical average-linkage initialization) K-均值聚类结果的 Rand 指数,结果如图1所示. 其中,蓝色散点为随机初始化聚类结果的 Rand 指数,蓝色虚线为100次的平均水平,红色实线为均连接初始化聚类结果的 Rand 指数. 从图1可以看到:随机初始化时聚类结果具有随机性,质量很不稳定;均连接初始化则有效地排除了随机初始化过程中引入的随机性因素,算法是确定性的,得到稳定的聚类结果. 而且,采用均连接分层聚类结果进行质心初始化能够利用数据中的类信息,可以保证较好的聚类质量. 因此,本文以均连接分层聚类结果作为K-均值聚类的初始质心.

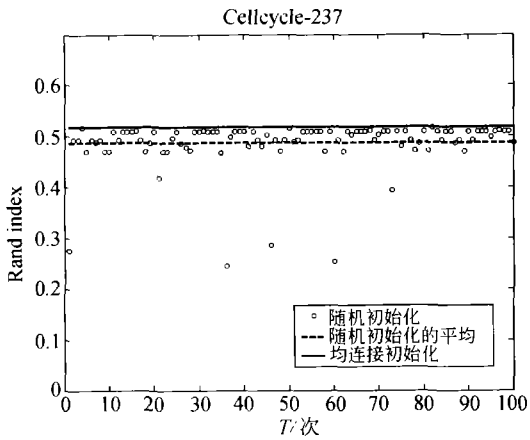

图1 质心随机初始化和均连接初始化K-均值聚类结果的质量比较

(3) 自组织映射(SOMs)

自组织映射是在K-means聚类算法基础上发展起来的一种非监督的神经网络学习方法[20],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. 给网格节点(神经元)赋予一定权重,来表示类的质心. 该方法需计算一种相似度来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SOMs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.

由于SOMs算法中节点的初始权重是随机产生的,使得算法结果有一定的不确定性[21]. 为此,本文对SOMs算法运行100次,选取其中最好的结果[22].

1.3 聚类结果的评价

聚类结果评价又称为聚类确认(cluster validation),是指以客观定量的方式对不同聚类过程所得的结果进

行质量和可靠性评价. 若将聚类结果与外部“金标准”进行一致性比较，则可望获得对聚类质量的独立、无偏的评价.

在基因表达聚类分析中，Rand 指数(Rand index)被广泛用来评价聚类结果与外部标准的一致性[12,22]. 设 U 与 V 为一个数据集的两种独立划分，若 a 为 U 和 V 中都属于同一类的个体对数，b 为 U 中属于同一类而 V 中不属同一类的个体对数，c 为 V 中属于同一类而 U 中不属于同一类的个体对数，d 为 U 和 V 中都不属于同一类的个体对数，则 Rand 指数定义为 $(a+d)/(a+b+c+d)$. 即便是在两种划分类数不同的情况下，Rand 指数同样能有效地比较其间的吻合程度. Rand 指数位于 0 与 1 之间，其数值越大，两种划分的一致程度越高. 当两种划分完全一致时，Rand 指数为 1. 但是，两个随机划分的 Rand 指数的期望值并不为 0. 为此，对 Rand 指数进行归一化校正为(指数值－期望值)/(最大指数值－期望值). 校正的 Rand 指数的期望值为 0，最大值为 1，取值范围更宽，可能取负值. 本文采用校正的 Rand 指数来衡量聚类结果的优劣.

为了研究不同预处理方式和相似度准则对聚类结果质量的影响，我们分别以 Pearson 相关系数和 Euclidean 距离作为相似度准则，采用单连接、全连接、均连接和质心连接 4 种形式的分层聚类算法以及 K-均值聚类和 SOMs 聚类算法，将经行标准化(normalize by line，NL)、列标准化(normalize by column，NC)和以 2 为底对数化(logarithmize，lb)预处理的 Spor，Gal，CNS，Cellcycle _ 384 和 Cellcycle _ 237 数据集聚为标准类数，计算结果类(clusters)与标准类(classes)间改进的 Rand 指数值，最后由指标的大小确定聚类结果的质量.

2 结果与讨论

2.1 分层聚类对相似度和预处理方法的选择

表 1 给出 5 个数据集经过行标准化(NL)、列标准化(NC)和对数化(lb)预处理后，分别以 Pearson 相关系数和 Euclidean 距离为相似性度量准则，进行单连接分层聚类结果的 Rand 指数值. 从表 1 可以看到：以 Pearson 相关系数作为相似性度量准则时，Spor，CNS 和两个 Cellcycle 数据集皆在行标准化预处理后得到最好的聚类结果，而 Gal 数据集则在对数化预处理后得到最好的聚类结果. 由于 Gal 数据集对数化后聚类结果的 Rand 指数值占据绝对优势，使得平均 Rand 指数值也表现出对数化的优势. 为了消除单个 Rand 指数值对均值的不均衡影响，对同一行的 Rand 指数值进行权重投票：按从小到大顺序对其排序，最小的权重为 1，依次增加，最大的权重值为 3. 权重投票结果如表 1 括号中数值所示. 平均权重值排序结果表明：以 Pearson 相关系数作相似性度量，则对数据进行标准化预处理是最可取的. 由于将数据行标准化处理后，基因表达矢量间的 Euclidean 距离和 Pearson 相关系数是等价的[7]. 因此，在 Euclidean 准则下，行标准化处理的结果与 Pearson 相关系数下相同，而经另外两种数据预处理后的聚类结果则差得多.

表 1 不同相似度和预处理方法下单连接分层聚类结果的 Rand 指数值

数据集	Pearson 相关系数			Euclidean 距离		
	NL	NC	lb	NL	NC	lb
Spor	0.0191(3)	0.0094(2)	0.0080(1)	0.0191	0.0114	0.0124
Gal	0.7451(2)	−0.0068(1)	0.8664(3)	0.7451	0.0175	0.0175
CNS	0.0063(3)	−0.0118(1)	0.0005(2)	0.0063	0.0081	0.0081
Cellcycle_384	0.0077(3)	−0.0023(1)	0.0044(2)	0.0077	−0.0015	−0.0020
Cellcycle_237	0.0181(3)	−0.0044(1)	0.0166(2)	0.0181	−0.0151	0.0166
平均	0.1593(2.8)	−0.0032(1.2)	0.1792(2)	0.1593	0.0041	0.0105

括号中的数表示权重投票值.

表 2—4 分别列出了全连接、均连接和质心连接分层聚类结果的 Rand 指数. 同样可以看到：以相关系数度量相似性时，这几种形式的分层聚类同样偏好于对数据作行标准化预处理. 而且，总体数

值表明，Pearson 相关系数准则下的聚类结果显著地优于 Euclidean 距离准则. 由此可见：采用分层聚类算法进行基因聚类时，应该对数据作行标准化处理，并以 Pearson 相关系数为相似性度量准则.

此外，比较上述 4 种连接形式的分层聚类结果可以看到：单连接分层聚类结果显著差于其他 3 种分层聚类结果. 这与文献[12]的结果相吻合，其原因是由于单连接分层聚类通过两个最相似的个体来确定两类间的相似性，可能潜在地导致类连锁，使得类取几何拉长的形状，而同一类中两个体的相似性可能很低. 其他连接形式的分层聚类算法都能避免类连锁，故可以得到优于单连接聚类质量的结果.

表 2 不同相似度和预处理方法下全连接分层聚类结果的 Rand 指数值

数据集	Pearson 相关系数			Euclidean 距离		
	NL	NC	lb	NL	NC	lb
Spor	0.4522	0.3494	0.3958	0.4522	0.0397	0.3377
Gal	0.7447	0.6431	0.7004	0.7447	0.6666	0.7004
CNS	0.1506	0.0759	0.1568	0.1506	0.0087	0.0841
Cellcycle_384	0.4828	0.3844	0.4435	0.4828	0.0173	0.0191
Cellcycle_237	0.5527	0.0760	0.5091	0.5527	−0.0451	0.3753
平均	0.4766	0.3058	0.4411	0.4766	0.1374	0.3033

表 3 不同相似度和预处理方法下均连接分层聚类结果的 Rand 指数值

数据集	Pearson 相关系数			Euclidean 距离		
	NL	NC	lb	NL	NC	lb
Spor	0.3910	0.3632	0.3977	0.3910	0.0203	0.1733
Gal	0.8563	0.9122	0.8563	0.8563	0.1578	0.8692
CNS	0.1058	0.0894	0.1053	0.1058	0.0107	0.0061
Cellcycle_384	0.4606	0.4587	0.4236	0.4606	−0.0015	−0.0051
Cellcycle_237	0.5387	0.0774	0.5077	0.5387	0.0738	0.2551
平均	0.4705	0.3802	0.4581	0.4705	0.0522	0.2597

表 4 不同相似度和预处理方法下质心连接分层聚类结果的 Rand 指数值

数据集	Pearson 相关系数			Euclidean 距离		
	NL	NC	lb	NL	NC	lb
Spor	0.3910	0.3878	0.4154	0.3910	0.0203	0.0511
Gal	0.8664	0.9210	0.8664	0.8664	0.1588	0.8614
CNS	0.1058	0.1019	0.1053	0.1058	0.0107	0.0022
Cellcycle_384	0.4819	0.4536	0.4222	0.4819	−0.0015	0.0064
Cellcycle_237	0.3384	0.2162	0.3173	0.3384	0.0818	0.2214
平均	0.4367	0.4161	0.4253	0.4367	0.0540	0.2285

2.2 K-均值聚类对相似度和预处理方法的选择

表 5 是采用 K-均值聚类的结果，从表 5 中 Rand 指数可以看到：在 Pearson 相关系数准则下，行标准化和对数化都具有一定的优势. 进行权重投票后(括号中数值)，平均权重表明：对数化略优于行标准化. 而在 Euclidean 距离准则下，则所有数据都一致性地支持对数化. 由此可见：采用 K-均值聚类算法进行基因聚类分析时，最好是对数据进行对数化预处理，并以 Euclidean 距离为相似性度量准则.

表5 不同相似度和预处理方法下K-均值聚类结果的Rand指数值

数据集	Pearson相关系数			Euclidean距离		
	NL	NC	lb	NL	NC	lb
Spor	0.3866(2)	0.3387(1)	0.4032(3)	0.3866	0.0641	0.4087
Gal	0.9436(2)	0.7241(1)	0.9436(2)	0.9436	0.8603	0.9545
CNS	0.1312(1)	0.1617(2)	0.1671(2)	0.1310	0.0044	0.1622
Cellcycle_384	0.4663(2)	0.4310(1)	0.4726(3)	0.4663	0.0048	0.4756
Cellcycle_237	0.5131(3)	0.0094(1)	0.3689(2)	0.5131	−0.0331	0.5268
平均	0.4882(2)	0.3330(1.5)	0.4711(2.4)	0.4881	0.1801	0.5056

括号中的数表示权重投票值

2.3 SOMs聚类对相似度和预处理方法的选择

表6列出了SOMs聚类分析的结果，可以看到：在Pearson相关系数准则下，行标准化预处理下的效果比对数化好；而在Euclidean距离准则下，对数化预处理则均比行标准化预处理好. 而且，从总体数值上看：Euclidean距离准则下的聚类结果显著地优于Pearson相关系数，正如Kohonen最早指出的“SOMs最好以Euclidean距离来度量输入矢量和节点神经元间的相似性”[20].

并且，从分层聚类、K-均值聚类和SOMs聚类的总体结果可以看到：K-均值聚类和SOMs聚类结果显著优于分层聚类，SOMs聚类结果稍好于K-均值聚类.

表6 不同相似度和预处理方法下SOMs聚类结果的Rand指数值

数据集	Pearson相关系数			Euclidean距离		
	NL	NC	lb	NL	NC	lb
Spor	0.3626(2)	0.3457(1)	0.4045(3)	0.3601	0.3758	0.4674
Gal	0.8665(3)	0.7231(1)	0.8605(2)	0.8573	0.9393	0.9596
CNS	0.1258(2)	0.1782(3)	0.0986(1)	0.1359	0.1747	0.1664
Cellcycle_384	0.4688(3)	0.4228(1)	0.4607(2)	0.4695	0.4581	0.4900
Cellcycle_237	0.5046(3)	0.0252(1)	0.2904(2)	0.4963	−0.0596	0.5222
平均	0.4657(2.6)	0.3390(1.4)	0.4229 (2)	0.4638	0.3777	0.5211

括号中的数表示权重投票值

3 结论

综上所述，在生命信息量呈爆炸性增长的时代，加快基因表达数据分析速度、提高信息检测质量是探索生命奥秘的基础. 聚类分析是基因表达数据分析研究的主要技术之一. 选择恰当的数据预处理方式和相似性度量，是获得高质量、高速度聚类结果的前提. 本文采用具有外部标准的基因表达数据集，分别以Pearson相关系数和Euclidean距离为相似度，以校正的Rand指数对分层聚类、K-均值聚类和SOMs聚类结果的质量进行了比较，可以得到如下结论：K-均值聚类和SOMs聚类质量显著优于分层聚类，SOMs聚类质量稍好于K-均值聚类. 此外，分层聚类偏好于Pearson相关系数准则和行标准化预处理，而应用K-均值聚类和SOMs算法时，则最好对数据进行对数化预处理、并以Euclidean距离为相似性度量准则. 当然，若能应用基因表达数据内在信息进行SOMs节点权重的初始化，而非随机性选择，必将显著改善SOMs算法的准确性，进一步提高聚类结果的质量. 上述研究结论将为基因表达聚类分析的实施提供有价值的参考依据.

参考文献

1 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

2 Neelam D, Ruben B, Dennis O, et al. Gene expression microarrays: A 21st century tool for directed vaccine design. Vaccine, 2002, 20 (1): 22--30

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

4 Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: A survey. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(11): 1370 1386

5 Amir B, Friedman N, Yakhini Z. Class discovery in gene expression data. Recomb, 2001: 31—33

6 Sherlock G. Analysis of large-scale gene expression data. Curr Opin Immunol, 2000, 12(2): 201--205

7 Yang C M, Wan B K, Gao X F. Actuality and development of the clustering technologies for gene expression. Prog Biochem Biophys (in Chinese), 2003, 30(6): 974 979

8 Yeung K Y. Model-based clustering and validation techniques for gene expression data. http: //u. washington. edu/kayee/report, 2004 06 28

9 Kaski S. Learning metrics for exploratory data analysis. Neural Networks for Signal Processing, 2001, 11: 53—62

10 Jiang D, Pei J, Zhang A. DHC: A density-based hierarchical clustering method for time-series gene expression data. In: Proceedings of 3rd Annual IEEE International Symposium on Bioinformatics and Bioengineering (BIBE 2003). Los Alamitos: IEEE Comput Soc, 2003, 393--400

11 Tang C, Zhang L, Zhang A, et al. Interrelated two-way clustering: An unsupervised approach for gene expression data analysis. In: Proceedings of 2nd Annual IEEE International Symposium on Bioinformatics and Bioengineering (BIBE 2001). Los Alamitos: IEEE Comput Soc, 2001, 41—48

12 Yeung K Y, Haynor D R, Ruzzo W L. Validating clustering for gene expression data. Bioinformatics, 2001, 17(4): 309—318

13 Chu S, DeRisi J, Eisen M, et al. The transcriptional program of sporulation in budding yeast. Science, 1998, 282(5398): 699—705

14 Ideker T, Thorsson V, Ranish J A, et al. Integrated genomic and proteomic analyses of a systemically perturbed metabolic network. Science, 2001, 292(5518): 929—934

15 Yeung K Y, Medvedovic M, Bumgarner R E. Clustering gene expression data with repeated measurements. Genome Biology, 2003, 4(5): R34

16 Wen X L, Fuhrman S, Michaels G S, et al. Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci USA, 1998, 95(1): 334—339

17 Cho R J, Campbell M J, Winzeler E A, et al. A genome-wide transcriptional analysis? of the mitotic cell cycle. Molecular Cell, 1998, 2 (1): 65—73

18 Tavazoie S, Huges J D, Campbell M J, et al. Systematic determination of genetic network architecture. Nature Genetics, 1999, 22(3): 281- 285

19 David R G, Michael S, Jacques V H. Interactive visualization and exploration of relationships between biological objects. TIBTECH, 2000, 18(12): 487—494

20 Kohonen T. The self-organizing map. Proc IEEE, 1990, 78(9): 1464--1480

21 Petri T, Mikko K, Wonga G, et al. Analysis of gene expression data using self-organizing maps. FEBS Letters, 1999, 451(2): 142—146

22 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