文章编号:1001 - 9081(2005)12 - 2868 - 04

基于马氏距离的缺失值填充算法

杨 涛,骆嘉伟,王 艳,吴君浩
(湖南大学 计算机与通信学院,湖南 长沙 410082)
(taoxtao@ tom. com)

摘 要:提出了一种基于马氏距离的填充算法来估计基因表达数据集中的缺失数据。该算法通过基因之间的马氏距离来选择最近邻居基因,并将已得到的估计值应用到后续的估计过程中,然后采用信息论中熵值的概念计算最近邻居的加权系数,得到缺失数据的填充值。实验结果证明了该算法具有有效性,其性能优于其他基于最近邻居法的缺失值处理算法。

关键词:微阵列;缺失值估计;马氏距离;信息熵
中图分类号: TP311.13 **文献标识码**:A

Missing value estimation for gene expression data based on Mahalanobis distance

YANG Tao, LUO Jia-wei, WANG Yan, WU Jun-hao
(*College of Computer and Communications, Hunan University, Changsha Hunan* 410082, *China*)

Abstract: A imputation method based on Mahalanobis distance was proposed to estimate missing values in the gene expression data. The nearest neighbors were chosen by the Mahalanobis distance between genes, and then the concept of entropy was utilized to obtain estimations of missing values. The imputed values were used for the later imputation. Experiments prove that the method is valid and its performance is higher than the other imputation methods based on k-nearest neighbors for gene expression data.

Key words: microarray; missing value estimation; Mahalanobis distance; entropy

0 引言

DNA 微阵列技术[1]可以同时监测并获得各种环境下的成千上万个基因表达水平值,将基因的活动状态比较完整地展现出来。微阵列实验获得的基因表达数据通常是以大矩阵的形式表现的,矩阵的行表示基因的表达水平值,而矩阵的列表示试验条件。由于 DNA 微阵列试验的各个步骤中存在有许多非理性的因素[2],如:不完全分解、图像损坏、表面有灰尘或划痕、试验错误等,造成获得的基因表达数据中常常包含缺失值。

为了从大量的基因表达数据中提取潜在的生物事实和意义,通常采用分类、聚类等各种数据分析方法,如多元监督分类法中的支持向量机(SVMs)[5]、主成分分析(PCA)[6]和单值分解法(SVD)[7]等,这些方法都要求输入完整的数据矩阵,不能对存在有缺失值的数据进行分析。如果使用填充值代替缺失值进行数据分析,则填充值的准确性会直接影响分析结果。因此,为了保证数据分析和处理的正确性和有效性,确保提供有效的数据,对缺失值进行正确的处理是一个非常重要的预处理过程。

解决生物数据缺失的一个方法就是重复试验[8],但是这个方法的代价太高,非常耗时,在实际运用中不可取。还有一些处理缺失值的简单方法如直接删除或忽略缺失值、使用“0”值填充或者使用样本数据的平均值代替[9],这些方法都存在着很大的不足与缺陷,未考虑数据的属性和数据之间的相关性,没有充分利用数据集所蕴涵的有价值的信息。近年来,在基因表达数据缺失值问题处理上提出了一些更为准确、复杂的方法,如单值分解法[10]、基于最近邻居法的缺失值填充算法 KNN_{impute}[10]、基于贝叶斯公式填充方法[11]以及基于最小平方原则的基因表达缺失值的处理算法[12]。这些算法基于不同模型、从不同角度研究了基因表达数据中缺失值的处理问题。其中,KNN_{impute} 是一种简单、快速的算法,它利用本身具有完全值的相似基因的表达值实现对缺失数据的估计。在 KNN_{impute} 算法的基础上,文献[13]提出了一种新的基于最近邻居的缺失值填充方法——有序最近邻居算法(SKNN),不仅利用了现有的具有完全值的基因,而且考虑了经过缺失值填充后的基因所包含的信息,使得在缺失率较大的情况下,也能获得比较好的性能。虽然 KNN_{impute} 和 SKNN 两个算法对基因表达数据中的缺失值填充有较好的性能和准确度,但是二者都是根据欧氏距离选择相似基因,欧氏距离与各指标的量纲有关,且没有考虑各变量之间的相关性和重要性,可能会造成相似基因的选择并不是最佳选择;而且在计算其估计值时,各邻居基因的加权系数也是根据欧氏距离来确定的,从而影响估计结果的准确度。

针对 KNN_{impute} 和 SKNN 的缺失值填充算法的不足之处,本文提出一种新的基于马氏距离的缺失值处理算法 MKNN。该算法采用了一种新的相似基因度量指标——马氏距离,它不仅考虑了观测变量之间的相关性,而且也考虑到了各个观测指标取值的差异程度,能更好地描述基因之间的相似程度。

收稿日期:2005 - 06 - 03;**修订日期**:2005 - 08 - 22 **基金项目**:湖南省自然科学基金(03JJY3095)

作者简介:杨涛(1977 -),女,四川内江人,硕士研究生,主要研究方向:生物基因数据挖掘; 骆嘉伟(1964 -),女,湖南长沙人,副教授,主要研究方向:生物信息处理、数据挖掘; 王艳(1981 -),女,湖南邵阳人,硕士研究生,主要研究方向:数据挖掘; 吴君浩(1981 -),男,湖南怀化人,硕士研究生,主要研究方向:生物序列数据挖掘.

然后利用信息论中熵值的概念,通过基因之间的马氏距离提供的信息决定各个相似基因的加权系数,其相应位置的加权平均值即为缺失数据的估计值。

1　KNN 与 SKNN 算法

1.1　KNN 算法

文献[10]提出的基于最近邻居的缺失值填充算法 KNN_{impute} 考虑了基因表达数据之间的相关性,因而预测结果较为准确。通过指定最近邻居基因数为 K,根据邻居基因提供的信息,对微阵列数据集中的缺失值进行预测和估计。KNN_{impute} 算法首先根据目标基因(包含有缺失值的基因)与其他具有完全值的基因之间的欧氏距离,在数据集中选择与目标基因的距离最小的 K 个最相近邻居基因,然后对选择出的 K 个最近邻居基因赋予相应的权值,其相应位置值的加权平均值即为目标基因缺失数据的估计值。

KNN 算法流程的伪代码描述:

Input:GeneData[ ][ ]:Gene expression data with missing values,
　　K :the number of nearest neighbors;

Output:EstData[ ][ ]:gene expression data with estimation value;

(1) Initialize data, construct experiment data matrix;

(2) Compute the Euclidian distance $d_i(z_i,z) = \sqrt{(z_i-g)^T(z_i-g)}$, g is the target gene which contains missing values;

(3) Select K closest genes as nearest neighbor genes from data set based on Euclidian distance;

(4) Calculate the weight of the nearest neighbor genes:

$$w_i = \frac{1/d_i}{\sum_{i=1}^{k} 1/d_i};$$

(5) Estimate the missing value: $\tilde{g} = \sum_{i=1}^{k} w_i x_i$, x_i is the corresponding expression value in the nearest genes.

1.2　SKNN 算法

文献[13]提出的有序最近邻居的缺失值填充算法 SKNN 是在 KNN_{impute} 算法基础上发展而来的,二者选择最近邻居基因的度量指标和计算邻居基因加权系数的方法均相同。作为 KNN_{impute} 的改进算法,SKNN 主要在两个方面不同于 KNN 方法:1)根据数据集中的缺失率进行排序,从缺失率最小的基因开始填充;2)SKNN 算法不仅利用数据集中具有完全值的基因,还将经过 SKNN 算法处理后的具有完全值的基因加入到相似基因的选择范围内,因此即使在缺失率较大的情况下,也能获得比较好的性能。在 SKNN 算法中,对包含有缺失数据的基因只需一次相似基因的选择过程,即可对其中的所有缺失值同时进行填充,而不同于 KNN_{impute} 需对每个缺失数据均进行相似基因的选择过程,这样就减少了执行时间。通过上述的改进,SKNN 算法在数据缺失率大的情况下具有较好的性能和实用价值。

2　基于马氏距离的缺失值填充算法 MKNN

在前面分析的 KNN_{impute} 和 SKNN 算法中,两者在选择最近邻居基因时都是以欧氏距离为度量指标,但是欧氏距离具有两个主要的缺点:1)欧氏距离的值与各指标的量纲有关,而各指标计量单位的选择有一定的人为性和随意性,而且任何一个变量计量单位的改变都会使此距离的数值改变,从而使该距离的数值依赖于各变量计量单位的选择;2)欧氏距离的定义没有考虑各个变量之间的相关性和重要性。实际上,欧氏距离是把各个变量都同等看待,将两个样本在各个变量上的离差简单地进行了综合。

为了克服上述欧氏距离的缺点,本文采用马氏距离来度量基因之间的相似程度,马氏距离相对于其他距离如欧氏距离而言具有以下优点[14]:1)马氏距离是欧几里德空间中非均匀分布的归一化距离,不用考虑各特征参数的量纲;2)马氏距离是根据整个空间上的特征分布情况来作为判别依据的,排除了样本之间的相关性影响。因此,它能更好地描述基因之间的相似性,为更高一级的数据分析提供有效的数据。

在信息论[15]中,熵值是系统无序程度或混乱程度的度量,信息被解释为系统无序程度的减少,其表现为系统的某项指标的变异度。熵值越小,不确定性越小,则它所蕴涵的确定性信息就越大;反之,熵值越大,不确定性越大,则它所蕴涵的确定性信息就越小。

在 KNN_{impute} 和 SKNN 算法中,计算缺失数据估计值使用的加权系数是通过欧氏距离的简单计算得到的,并不能准确反映各最近邻居对含有缺失值的目标基因的影响。本文从信息论的观点出发,利用信息熵的概念,计算最近邻居的加权系数,最终得到缺失数据的填充值。

基于马氏距离的缺失值填充算法由以下三个主要部分组成:

2.1　基因数据降维处理

通常来说从试验中得到的数据规模很大,具体表现为基因数目繁多,而试验条件相对较小,一般在 20 ~ 100 范围内。在数量众多的基因中,不是所有基因都对包含有缺失值的目标基因有意义,其中存在有不相关或相关较小的基因,而且当变量较多时,会增加马氏距离的计算复杂度。因此对数据集进行降维处理,减少后续工作的计算量和复杂度。

采用对基因表达数据集进行相关性分析方法,根据基因之间的相似程度对基因进行筛选,计算过程中使用行平均值代替存在的缺失值。基因之间的相似系数越大,它们就越相近,就越能准确描述目标基因的信息。

设目标基因 $g = [g_1, g_2, \cdots, g_m]^T$,则与基因 $z_i = [z_{i1}, z_{i2}, \cdots, z_{im}]^T$ 之间的相似系数计算公式为:

$$\gamma_i = \frac{\sum_{k=1}^{m}(g_k-\bar{g})(z_{ik}-\bar{z_i})}{\sqrt{\left[\sum_{k=1}^{m}(g_k-\bar{g})^2\right]\left[\sum_{k=1}^{m}(z_{ik}-\bar{z_i})^2\right]}},\ i = 1,2,\cdots,n \tag{1}$$

其中:$\bar{g} = \frac{1}{m}\sum_{k=1}^{m} g_k$, $\bar{z_i} = \frac{1}{m}\sum_{k=1}^{m} z_{ik}$, n 为基因总数。

在对数据进行降维处理时,选择与目标基因相似度大的基因作为后续工作的处理数据,并且考虑试验条件的个数(列)。一般说来,当相关系数 r 在[0.5,0.8]之间时,二者为显著相关,r 在[0.8,1]之间时,二者为高度相关,当 $r = 1$ 时,两者的关系为完全相关[16]。为了适应各种数据集的规模大小,预选择相似基因数通过下述公式获得:

$$n' = \begin{cases} n & n < 5m \\ 5m & 5m < n < 10m \\ 10m & n > 10m \end{cases} \tag{2}$$

其中,n 为基因总数,m 为实验条件数,n' 为筛选后得到的相似基因数。

2.2　计算基因之间的马氏距离

马氏距离不仅考虑了观测变量之间的相关性,而且也考虑到了各个观测指标取值的差异程度,从而弥补了欧氏距离

的不足,能更好地描述基因之间的相似性。其计算公式为:

$$d_i = \sqrt{(g - z_i)^T \Sigma^{-1} (g - z_i)}\ ,\ i = 1,2,\cdots,n' \qquad (3)$$

其中 $g = [g_1, g_2, \cdots, g_m]^T$, $z_i = [z_{i1}, z_{i2}, \cdots, z_{im}]^T$, g 为包含有缺失数据的目标基因,z_i 属于矩阵 $\boldsymbol{Z}'$,$\boldsymbol{Z}'$ 是经过降维处理后得到的基因表达数据矩阵,且 $g \neq z_i$。Σ 表示观测变量之间的协方差矩阵。若总体协方差矩阵 Σ 未知,则用样本协方差矩阵代替。从 $\boldsymbol{Z}'$ 中选择具有典型代表性的基因作为代替总体的样本抽样,即选择与 z 相关性较强的基因。当 $n' < 2m$,则选取前 m 个相似系数大的基因作作为代替总体的一个样本;其他情况,则从数据集 $\boldsymbol{Z}'$ 中选取 $2m$ 个与 g 相似系数大的基因作为代替总体的两个样本。总体的协方差 Σ 为:

$$\Sigma = \begin{cases} s & n' < 2m, s \text{ 为样本协方差} \\ \dfrac{1}{2}(s1 + s2) & n' \geqslant 2m; s1, s2 \text{ 为样本协方差} \end{cases} \qquad (4)$$

2.3 估计目标基因的缺失值

根据计算得到的马氏距离,选择距离最短的 K 个基因作为目标基因 g 的最近邻居,然后通过这 K 个邻居基因提供的信息,对目标基因中缺失值进行预测和估计。本文采用对相似基因进行加权平均形式的预测模型,对目标基因中的缺失值进行估计。预测的核心问题就是如何求出加权系数,使得结果具有较高的预测精度。

本文利用信息论中熵值的概念,确定各最近邻居基因在对缺失值估计时的加权系数,其步骤如下:

(1) 将计算得到的 K 个最近邻居基因的马氏距离单位化:

$$p_i = \frac{d_i}{\sum_{i=1}^{k} d_i}\ ,\ i = 1,2,\cdots,k \qquad (5)$$

d_i 为第 i 个邻居与目标基因之间的马氏距离。显然,$\sum_{i=1}^{k} p_i = 1$。

(2) 计算第 i 个邻居基因的熵值:

$$h_i = -mp_i \ln p_i,\ i = 1,2,\cdots,k \qquad (6)$$

其中,m 为大于 0 的常数,且 $m = (\ln k)^{-1}$,ln 为自然对数。

(3) 计算第 i 个邻居基因的变异程度系数:

因为 $0 \leqslant h_i \leqslant 1$,根据熵值的大小与其变异程度相反的原则,所以定义第 i 个相似基因的变异程度系数为:

$$v_i = 1 - h_i,\ i = 1,2,\cdots k \qquad (7)$$

(4) 计算第 i 个邻居基因的加权系数:

$$w_i = \frac{1}{k-1}\left(1 - \frac{v_i}{\sum_{i=1}^{k} v_i}\right)\ ,\ i = 1,2,\cdots,k \qquad (8)$$

某个相似基因的变异程度越小,其包含的确定性信息就越大,则其在预测中对应的加权系数就越大,反之就越小。所有的加权系数满足:$\sum_{i=1}^{k} w_i = 1$。

(5) 计算预测值:

目标基因 g 中的缺失值可由下列公式计算获得:

$$\tilde{g} = \sum_{i=1}^{k} w_i \times x_i,\ i = 1,2,\cdots,k \qquad (9)$$

其中 x_i 为相似基因中与缺失值对应位置的表达水平值,计算得到的 $\tilde{g}$ 值即为目标基因中缺失数据的估计值。

2.4 算法的伪代码描述

Input: GeneData[ ][ ]: Gene expression data with missing values,
K: the number of nearest neighbors;
Output: EstData[ ][ ]: genes expression data with estimation value;
(1) Compute the average values of the genes that contain missing values;
(2) Replace missing values with corresponding average value;
(3) Compute the correlativity of the target gene g and other gene z_i;
(4) Select the similar genes for g on the basis of correlativity, obtain new matrix $\boldsymbol{Z}'$;
(5) Compute the Mahalanobis distance of g and other gene z_i in $\boldsymbol{Z}'$;
(6) Select k closest genes as nearest neighbor genes for g; and compute the weighted value of each nearest gene;
(7) Obtain the estimation of missing values in g;

3 实验结果分析

3.1 实验方法

基因表达数据可从开放的公共基因数据库获取,本文实验所用数据集分别在下述研究中使用:识别酵母中能调节细胞周期的基因的研究[17]、酵母从发酵到氧化过程中新陈代谢变化对应的临时基因表达的探索研究[18]、在酵母中环境变化引起的基因表达变化的研究[19]。前两个数据集是时间序列数据,其中一个包含的噪声较小,称其为时间序列;另一个则有较大的噪声,称为噪声时间序列;最后一个为非时间序列数据集。

从数据库中获取的数据本身可能会包含有缺失值,如果直接作为实验数据,则得到的结果无法进行评价,因此需要将其中包含有缺失数据的行和列删除,人为地获得完整的数据集。在获得的完整的数据集中,根据算法的需要随机删除一定比例的数据产生测试数据,然后再使用各种算法来恢复测试数据中的缺失值,并将估计值与真实值进行比较。

对各种算法的缺失值估计算的性能采用均方根误差(Root Mean Squared error, RMS_{error})来评价:

$$RMS_{error} = \sqrt{\frac{\sum_{i=1}^{N} (R_i - I_i)^2}{N}} \qquad (10)$$

其中,R_i 为真实值,I_i 是估计值,N 为缺失值个数。计算得到的 RMS_{error} 的值越小,其估计值就越准确,反之结果就越差。

本文从各种数据缺失率和不同的最近邻居个数两个方面来测试 MKNN 算法的性能,并与 KNN_{impute} 和 SKNN 算法的实验结果进行比较。

3.2 实验结果

分别使用时间序列、噪声时间序列和非时间序列三个不同类型的数据集,产生不同数据缺失率的测试数据,测试三种算法的性能。对时间序列数据集,使用不同的最近邻居数来获得各种算法的 RMS_{error} 值。

从实验结果中可以得出,算法 MKNN 的性能优于 KNN_{impute} 算法和 SKNN 算法。在 KNN_{impute} 算法中,数据缺失率对算法的影响很大。因为当缺失率较大时,相似基因的选择范围就会变得很小,可能造成选择的基因的相似性并不高,因而当缺失率到一定程度时,性能会急剧下降。而在 SKNN 算法中,不仅利用数据集本身具有完全值的基因,而且还利用了经过算法处理后的具有完全值的基因所蕴含的信息,使其也作为选择相似基因的候选基因。这样,即使在缺失率大的情

况下也能比 KNN_{impute}选择出更为相似的基因,从而提高了性能和准确度。在算法 MKNN 中,摒弃了前两种算法中使用的欧氏距离,采用马氏距离来选择最相似基因,并同 SKNN 一样,充分利用经过 MKNN 算法处理后的具有完全值的基因,而且采用信息论中熵的概念计算最近邻居的加权系数,能更准确地反映各相似基因对目标基因的贡献大小,使得最终得到的缺失数据的填充值更为准确。

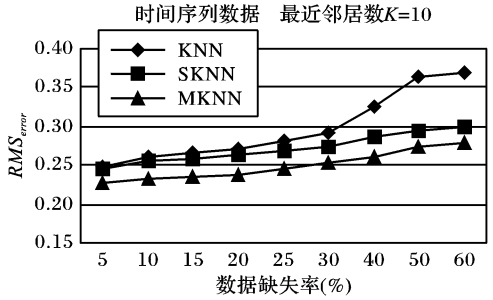

图 1　各算法对时间序列数据处理的 RMS_{error} 值

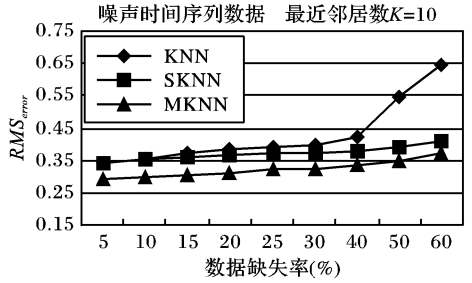

图 2　各算法对噪声时间序列数据处理的 RMS_{error} 值

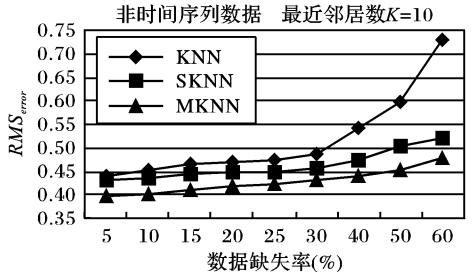

图 3　各算法对非时间序列数据处理的 RMS_{error} 值

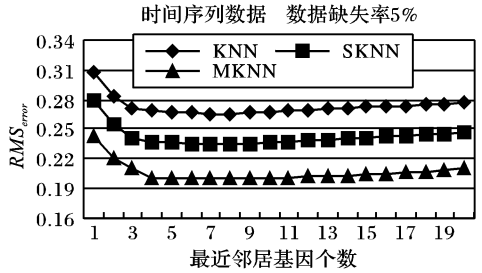

图 4　邻居数对各算法的 RMSerror 值影响变化图

注:图 1 ~ 图 4 中的 KNN 为 KNN_{impute}的简写

4　结语

本文采用马氏距离作为基因之间相似性的度量指标,并利用信息熵的概念,提出了一种对基因表达数据中缺失值的填充算法 MKNN。将 MKNN 算法应用到时间序列、噪声时间序列和非时间序列三种不同类型的数据集中,分析其性能,并与同类算法 KNN_{impute} 和 SKNN 进行比较,实验结果表明 MKNN 算法是一个有效的基因表达数据缺失值的填充算法。

参考文献:

[1] DUDOIT S, YANG YH, CALLOW MJ, *et al.* Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments[J]. Statistica Sinica, 2002, 12(1): 111 – 139.

[2] ARBEITMAN MN, FURLONG EEM, IMAM F, *et al.* Gene expression during the life cycle of Drosophila melanogaster[J]. Science, 2002, 297(5590): 2270 – 2275.

[3] GASCH AP, SPELLMAN PT, KAO CM, *et al.* Genomic expression programs in the response of yeast cells to environmental changes[J]. Molecular Biology of the Cell 2000, 11: 4241 – 4257.

[4] BOHEN SP, TROYANSKAYA OG, ALTER O, *et al.* Variation in gene expression patterns in follicular lymphoma and the response to rituximab[J]. Proc Natl Acad Sci, USA, 2003, 100(4): 1926 – 1930.

[5] BROWN MP, GRUNDY WN, LIN D, *et al.* Knowledge-based analysis of microarray gene expression data by using support vector machines [J]. Proc. Natl Acad. Sci, USA, 2000, 97, 262 – 267.

[6] RAYCHAUDHURI S, STUART JM, ALTMAN R. Principal components analysis to summarize microarray experiments: application to sporulation time series[J]. Pac. Symp. 15Biocomput., 2000, 455 – 466.

[7] ALTER O, BROWN PO, BOTSTEIN D. Singular value decomposition for genome-wide expression data processing and modeling[J]. Proc. Natl Acad. Sci. USA, 2000, 97(18): 10101 – 10106.

[8] BUTTE AJ, YE J, NIEDERFELLNER G, *et al.* Determining significant fold differences in gene expression analysis[J]. Pac. Symp. Biocomput., 2001, 6: 6 – 17.

[9] ALIZADEH AA, EISEN MB, DAVIS RE, *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling [J]. Nature, 2000, 403, 503 – 511.

[10] TROYANSKAYA O, CANTOR M, SHERLOCK G, *et al.* Missing value estimation methods for DNA microarrays[J]. Bioinformatics, 2001, 17: 520 – 525.

[11] SHIGEYUKI OBA, MASA-AKI SATO, ICHIRO TAKEMASA, *et al.* A Bayesian missing value estimation method for gene expression profile data[J]. Bioinformatics, 2003, 19(16) .

[12] KIMY H, GOLUBZ GH, PARKY H. Missing Value Estimation for DNA Microarray Gene Expression Data: Local Least Squares Imputation[J]. Bioinformatics, 2004.

[13] KI-YEOL KIM, BYOUNG-JIN KIM, GWAN-SU YI. Reuse of imputed data in microarray analysis increases imputation efficiency [J]. BMC Bioinformatics 2004, 5: 160.

[14] 何晓群. 多元统计分析[M]. 北京: 中国人民大学出版社, 2004.

[15] 傅祖芸. 信息论——基础理论与应用[M]. 北京, 电子工业出版社, 2001.

[16] 贾俊平. 统计学[M]. 北京: 中国人民大学出版社, 2002.

[17] SPELLMAN PT, SHERLOCK G, ZHANG MQ, *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization[J]. Mol Biol Cell, 1998, 9(12): 3273 – 3297.

[18] DERISI JL, IYER VR, BROWN PO. Exploring the metabolic and genetic control of gene xpression on a genomic scale[J]. Science, 1997, 278, 680 – 686.

[19] GASCH AP, SPELLMAN PT, KAO CM, *et al.* Genomic expression programs in the response of yeast cells to environmental changes [J]. Mol Biol Cell, 2000, 11(12): 4241 – 4257.