ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.9, September 2010, pp.2148−2160
doi: 10.3724/SP.J.1001.2010.03796

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

识别表达不稳定的基因*

杨 昆[1,2+], 李建中[2], 徐德昌[2], 戴国骏[1]

[1](杭州电子科技大学 计算机学院,浙江 杭州 310018)

[2](哈尔滨工业大学 计算科学与技术学院,黑龙江 哈尔滨 150001)

Identifying Unstable Genes in Expression

YANG Kun[1,2+], LI Jian-Zhong[2], XU De-Chang[2], DAI Guo-Jun[1]

[1](School of Computer, Hangzhou Dianzi University, Hangzhou 310018, China)

[2](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: yangkun@hdu.edu.cn

Yang K, Li JZ, Xu DC, Dai GJ. Identifying unstable genes in expression. *Journal of Software*, 2010,21(9):2148−2160. http://www.jos.org.cn/1000-9825/3796.htm

Abstract: An idea of identifying unstable genes by integrative analysis of pair of different data has been proposed. This problem is modelled as a nonlinear integer programming problem. Three approximate methods have been proposed to work out the solution. An index is designed to measure and rank the instability magnitude of unstable gene. The experimental results on two lung cancer datasets from two research groups demonstrate the identified unstable genes have really unstable expression. The identified unstable genes can be used to improve the result of identifying differential expression genes and excluding these genes can effectively enhance the accuracy of microarray data classification. The findings suggest the proposed methods are effective and the unstable genes can provide valuable information for microarray data analysis.

Key words: microarray; gene expression data analysis; integer programming; differential expression gene

摘 要: 提出集成分析来自相同研究问题的不同数据集来识别表达不稳定的基因.把这一问题形式化为一个非线性整数规划问题,三个启发式的算法被提出来求解这一优化问题;进一步地设计了一个统计量来度量基因的不稳定表达程度.提出的方法应用于两个真实数据,实验结果显示:所识别的不稳定基因在两个数据中的表达不一致;利用表达不稳定基因可以提高差异表达基因的筛选结果,而去除表达不稳定基因可以有效地提高微阵列数据分类.实验结果表明,提出的方法是有效的,并且表达不稳定基因可以为微阵列数据分析提供有价值的信息.

关键词: 微阵列;基因表达数据分析;整数规划;差异表达基因

中图法分类号: TP391 **文献标识码**: A

在基因调控、基因功能、新陈代谢、药物反应、疾病监测等的研究中,微阵列芯片(基因芯片)是监测基因表达水平广泛应用的有效技术.基因表达数据已为生物学、医学领域的研究提供了非常有价值的信息,并且人

* Supported by the National Natural Science Foundation of China under Grant No.60903086 (国家自然科学基金); the Zhejiang Provincial Natural Science Foundation of China under Grant No.Y1080973 (浙江省自然科学基金)

Received 2009-02-11; Revised 2009-11-04; Accepted 2009-12-03; Published online 2010-05-31

们期待着基因表达数据能够为生物学医学的发展做出更多的贡献[1−3].对于具体研究中的一个待研究的基因,例如前列腺肿瘤研究中的一个基因,样本个体中生物学和遗传学的变异性会影响这个基因的表达水平,进而导致这个基因在样本中不稳定地表达.尤其重要的是,肿瘤样本中与生俱来的染色体组的不稳定性可以导致某些基因的表达产生更大程度的变化.这种表达不稳定的基因是探索潜在的生物学奥秘和癌症起因的重要线索.在现阶段的实际研究中,由于样本资源的有限性和实验经费的限制,单个研究中的样本数目是极其有限的.在有限样本的约束下,这些表达不稳定的基因可以在同一研究问题的不同数据集间引起差异.于是,集成分析来自相同研究问题的不同数据集是识别表达不稳定基因的一个可行途径.

识别出的不稳定基因可以为生物学医学的研究,例如识别差异表达的基因[4]、识别疾病的鉴别基因[5]、发现基因表达的特定模式[6]、研究基因的功能和相互作用网络[7]等提供有价值的信息.微阵列数据分析中的一个重要目标是识别出在不同类别的两组样本间表达有差异的基因(differentially expressed genes),例如正常组织的样本和疾病组织的样本,或者不同实验条件下的样本,或者不同的外界刺激下的样本[4,8,9].因而,识别的表达不稳定基因可以用来验证、筛选现阶段小样本实验中得到的差异表达基因. Ein-Dor 等人研究发现,要识别可靠的鉴别基因来精确预测癌症样本需要数以千计的实验样本[10].在目前费用昂贵的情况下,由于包含数以千计样本的芯片实验是难以适用的,在具体的操作上也非常耗时,更为重要的是,疾病样本非常稀少,收集大量的样本及其困难.同样,鉴别的不稳定基因可以对鉴别基因的识别提供帮助.我们可以通过排除所识别的表达不稳定基因来提高微阵列数据分类中的预测精度.进一步地,所识别出的不稳定基因可以解释为什么不同数据间存在差异,进而可以为其他类型的微阵列数据集成分析提供帮助.

本文研究通过集成分析来自相同研究问题的不同数据集来识别表达不稳定的基因.我们把这一问题形式化为一个非线性整数(0–1)规划问题,优化目标是最大化所构造的多维目标函数,优化解是一个多维二元向量,其中每个维度对应一个特定的基因.在解向量中,如果一个基因其对应的分量取值为 0,那么这个基因就称为不稳定基因.三个启发式的近似算法被提出来求解这一非线性整数优化问题,进一步地,我们设计了一个统计量来度量和排列所识别的基因的不稳定表达的程度.本文提出的方法应用于不同研究小组产生的两个真实肺癌数据,实验结果显示:所识别的不稳定基因在两个数据中的表达的确不一致,并且可以用来修正差异表达基因的筛选结果,排除所识别的不稳定基因可以有效地提高微阵列数据分类结果.

1 相关的工作

随着微阵列技术在生物学、医学等研究领域中广泛应用,产生了大量的数据.对单个研究中产生的微阵列数据进行独立的分析是一种重要而且有效的数据分析方式,并取得了巨大的成功.然而这种数据分析方式面临一个巨大的障碍,单个研究中的实验是小样本实验,样本的数目有限.集成分析来自不同研究团体的多个实验数据是一个可行的数据分析方式,既可以验证单一数据上的分析结果,又可以增加样本的数目进而提高、扩展单一数据上的分析结果,还有可能获得在单一数据上难以得到的新结果.近年来,有很多方法被提出,在不同尺度上集成分析微阵列数据.

Rhodes 等人提出了一种 *meta* 分析的模型,目的是综合每个基因在单个研究的数据上计算得到的 *p-value* 来得到这个基因针对于多个研究的平均 *p-value*[11].Choi 等人提出了一个关于效应量分析的方法,把基因在两个类别间的差异表达尺度作为效应量,然后利用多个数据上计算的效应量来估计平均的综合效应[12].Hu 等人提出了一个质量测度来估计每个基因在单个数据中的重要程度,并把这一质量测度融合到效应量分析方法中来组合多个数据上的结果[13].Jiang 等人利用他们提出一个数据变换的方法来变换、组合两个肺癌的数据,并且提出基于随机森林和 *fisher* 线性判别的两个基因剔除方法,在组合的数据上识别标记基因[14].在文献[15]提出的方法中,直接组合多个数据来增加样本的数量,然后以每个基因对为考察单元来计算它们的分类能力分数,选择具有最高分类能力分数的基因对的基因为标记基因.Huttenhower 等人给出一个贝叶斯框架从多个数据中预测共表达的基因功能关系[16].Jornsten 等人提出一种基于 *meta* 数据的微阵列数据的缺失值估计方法,他们以样本的绝对值 *Pearson* 相关系数为样本相似性测度从多个数据的数百个样本中选出与目标样本最相似的前 40 个样本,

以这些数据来估计目标样本中的缺失值[17].

本文的研究不同于已有的数据集成研究.首先,研究的目标不一致,我们的目标是通过集成分析不同的数据来鉴别表达不稳定的基因;其次,采用的理论基础和技术不一样,我们把识别不稳定基因的问题抽象为一个非线性整数规划问题,并且给出 3 个求解算法.

2 问题的定义和算法

2.1 问题定义

包含 p 个基因 n 个样本的基因表达数据可以表示成一个 $p\times n$ 的矩阵,其中每个行向量代表一个基因在 n 个样本中的表达水平,每个列向量表示一个样本中 p 个基因的表达水平.令 $A_{p\times n_1}$ 和 $B_{p\times n_2}$ 是针对相同的科学问题而由不同的研究产生的两个微阵列数据,它们共同包含 p 个相同基因,其中矩阵 A 和 B 中的第 i 个行向量对应于第 i 个基因在各自样本中的表达值.由于数据产生方式的差异,例如微阵列平台、实验的操作协议、实验的条件以及数据尺度和分布上的差异,直接比较两个数据是不可靠的,需要对数据进行处理.对于每一个数据,我们计算每个基因在这一数据上的差异表达统计量,用每个基因的差异表达统计量来代替这一原始数据进行后面的集成分析,从而尽可能地消除数据本身带有的特异性.设向量 a 和 b 分别是由数据集 A 和 B 计算得到的两个 p 维的向量,其中每一个维度对应一个基因.接着,采用相关系数为测度函数来衡量两个向量 a 和 b(即分别代表数据集 A 和 B)的一致性.两个向量的相关系数越大表示向量间的一致性越高,也就说明数据集 A 和 B 的一致性程度越高.如果 p 个基因在两个不同的数据集 A 和 B 中的表达稳定,那么 a 和 b 间相关系数就会非常高;如果 p 个基因的部分基因在数据集 A 和 B 中表达不稳定,那么会导致 a 和 b 的相关系数变低.也就是说,从 p 个基因中删除部分表达不稳定的基因会使得新差异表达向量 a 和 b 间的相关系数增大.于是,从 p 个基因中选出至多 m 个表达最不稳定的基因就可以表示成如下的优化问题:

输入:向量 a 和 b,目标函数 F 和维度集合 $P=\{1,2,\ldots,p\}$ 以及参数 m,其中 $m<p$;

输出:维度集合 S 且 $|S|\leqslant m$ (集合 S 对应于不稳定基因的集合);

要求:向量 a 和 b 在维度集合 $P\backslash S$ 上的投影上使得目标函数 F 达到最大值.

优化问题可以形式化为:

$$\begin{cases}\max\ f(x)=\dfrac{\sum_{i=1}^{p}x_i\cdot(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{\sum_{i=1}^{p}x_i\cdot(a_i-\bar{a})^2}\cdot\sqrt{\sum_{i=1}^{p}x_i\cdot(b_i-\bar{b})^2}}\\ \text{s.t.}\quad t=\sum_{i=1}^{p}x_i\geqslant p-m\\ \bar{a}=\dfrac{1}{t}\sum_{i=1}^{p}x_ia_i\\ \bar{b}=\dfrac{1}{t}\sum_{i=1}^{p}x_ib_i\\ x_i\in\{0,1\}\\ 1\leqslant i\leqslant p\end{cases}\tag{1}$$

其中,相关系数函数 $f(x)$是一个有界函数,其上下界为 1 和–1;x 为 p 维二元向量.不失一般性,我们可以假定 $f(x)>0$.进一步地,对目标函数 $f(x)$展开,其可以变成如下形式:

$$f(x)=\frac{\sum_{i=1}^{p}x_ia_ib_i-t\bar{a}\bar{b}}{\sqrt{\sum_{i=1}^{p}(x_ia_i^2)-t\bar{a}^2}\cdot\sqrt{\sum_{i=1}^{p}(x_ib_i^2)-t\bar{b}^2}}\tag{2}$$

优化问题式(1)是一个复杂的非线性整数规划问题,目标函数非凸、非多项式、非可分.线性整数规划是

NP-Complete[18,19],例如背包问题是一个典型的线性整数规划问题[18].求解整数规划问题的多项式时间的精确算法是不存在的,除非 $P=NP$.整数规划问题的常用算法有枚举法、割平面法和分枝定界法[20].枚举法和分枝定界法只适合于求解只含有少数变量的小规模整数规划问题,对大规模的问题其计算时间难以承受,通常不能使用[21].割平面法实质是用解线性规划的方法解整数规划问题,其对复杂非线性问题不适用.非线性整数规划问题比线性整数规划更复杂也更难以求解.文献[22]表明,对一般的非线性规划问题,求解一个满足近似比的解也是 NP-hard 问题,而且获得非线性规划问题的局部最优解是 NP 的.近似方法是求解非线性整数规划的有效方法.近来有很多近似方法被发展出来求解这样的问题[23−25].然而,目前的方法只能处理含有 100~200 变量的中等尺度的非线性整数规划问题,对大尺度问题难以胜任.更重要的是,基因表达数据通常含有数千到几万个基因,微阵列数据的这一超高维数的特点使得式(1)成为一个涉及到数千到几万变量的超大规模尺度的非线性整数规划问题.遗传算法是随机方法,对目标函数要求低,但性能受控制参数影响[21,26];当问题规模较大时计算耗时并且效果有待验证.为此,本文给出 3 个不同复杂程度的基于贪心思想的近似算法来求解本文提出的问题.

2.2 相关的定义和算法

为了叙述方便,首先给出一些约定和定义.令 e_i 为一个 p 维的单位向量,其中第 i 分量为 1,其余的分量为 0 且令 $E=\{\pm e_i \mid 1\leqslant i\leqslant n\}$.

定义 1(可行解集). 令 X_z 为满足式(1)中约束条件的点集,即 $X_z=\{x \mid \sum_{i=1}^{p} x_i \geqslant p-m, x_i\in\{0,1\}\}$,称 X_z 为可行解集.

定义 2(离散邻域). 设 $x\in X_z$,称 $N(x)=\{x\}\cup\{x+e \mid x+e\in X_z, e\in E\}$ 为点 x 的离散邻域,简称为邻域.

定义 3(局部最大点). 设 $x^*\in X_z$,如果任意 $x\in N(x^*)$ 都有 $f(x)\leqslant f(x^*)$,则称 x^* 为 $f(x)$的局部最大点.

对于向量 a 和 b,有一个简单的观察(naive observation),对一个稳定的基因 i,其在数据集 A 中的差异表达量 a_i 和在数据集 B 中的差异表达量 b_i 间会有相对较小的差异;相反,对一个不稳定的基因 i, a_i 和 b_i 间可能具有较大的差异.即 $|a_i-b_i|$ 对向量 a 和 b 间的一致性有着重要的影响.下面给出一个基于 $|a_i-b_i|$ 的近似算法 amno(approximate method based on naive observation)来识别表达不稳定的基因.算法 amno 非常简单方便,但其对计算结果(所识别的不稳定基因集合 S)难以提供必要的保证.

算法. amno.

输入:向量 a 和 b,候选基因集合 $T=\{1,\ldots,p\}$,参数 m.

输出:不稳定基因集合 S.

1. $x^{(0)}=\{1,\ldots,1\}$, $S=\varnothing$, k=1;
2. Compute $d_i=|a_i-b_i|, 1\leqslant i\leqslant p$;
3. Let $\{d_{i_k}\}_k^p=1$ be the descending sequence of d_i;
4. **While** $k\leqslant m$ **do**
5. $S=S\cup\{i_k\}, T=T\setminus\{i_k\}$;
6. $x^{(k)}=x^{(k-1)}-e_{i_k}$, compute $f(x^{(k)})$;
7. k=k+1;
8. **end**

第 2 种算法是一个基于直接搜索(direct search)的算法,称为算法 amds (approximate method based on direct search).在算法 amds 的每次迭代中有两个主要的操作:第 1 个操作是从候选基因集合 T 中搜索 1 个维度 i,把它从 T 取出并放入集合 S 中使得目标函数 $f(x)$的增量大于一个给定的参数 ε,称为取出操作;第 2 个操作是从维度集合 S 中搜索一个维度 j,把它从 S 放回到集合 T 中使得目标函数 $f(x)$的增量大于一个给定的参数 ε,称为放回操作.算法 amds 每次迭代时首先执行取出操作,如果取出操作失败那么执行放回操作;如果两个操作都失败,算法 amds 停止运行.算法 amds 满足下面的定理.

算法. amds.

输入:向量 a 和 b,候选基因集合 $T=\{1,\ldots,p\}$,参数 m.

输出:不稳定基因集合 S.

1. $x^{(0)}=\{1,\ldots,1\}, S=\varnothing, k=1$;
2. **While** $k\geqslant 1$ **do**
3. Find $f(x^{(c)})=\max\limits_{i\in T} f(x^{(k-1)}-e_i)$;
4. **If** $f(x^{(k-1)})+\varepsilon<f(x^{(c)})$ and $x^{(c)}\in X_z$ **then**
5. $S=S\cup\{i\}, T=T\setminus\{i\}, x^{(k)}=x^{(c)}$;
6. **else**
7. Find $f(x^{(c)})=\max\limits_{j\in S} f(x^{(k-1)}+e_j)$;
8. **If** $f(x^{(k-1)})+\varepsilon<f(x^{(c)})$ and $x^{(c)}\in X_z$ **then**
9. $S=S\setminus\{j\}, T=T\cup\{j\}, x^{(k)}=x^{(c)}$;
10. **else**
11. **STOP**.
12. **end**
13. **end**
14. $k=k+1$;
15. **end**

定理 1. 当 $\varepsilon=0$ 时,算法 amds 在有限步内终止于局部最大点,且其迭代次数的上界为 $\left\lfloor \{1-f(x^{(0)})\}/\Delta f \right\rfloor+1$,其中初始点为 $x^{(0)}$ 且 $\Delta f=\min\{|f(x^1)-f(x^2)|\,\big|\,f(x^1)\neq f(x^2), x^1\in X_z, x^2\in X_z\}$.

证明:对于算法 amds 中的点序列 $\{x^{(k)}\}$,由 Δf 的定义可得 $f(x^{(k)})+\Delta f\leqslant f(x^{(k+1)})$.假设算法 amds 不终止,那么一定存在一个正整数 k_0,使得 $f(x^{(k_0)})\geqslant f(x^{(0)})+k_0\cdot\Delta f>1$.然而,相关系数 $f(x)$是有界函数,其最大值为 1,故导致矛盾,因而算法 amds 有限步内终止.

设算法 amds 终止于 $x^{(k_1)}$,那么 $1\geqslant f(x^{(k_1)})\geqslant f(x^{(0)})+k_1\cdot\Delta f$.即有 $k_1\leqslant\{1-f(x^{(0)})\}/\Delta f$.加上可能的第 k_1+1 次不成功的迭代,故迭代次数的上界为 $\left\lfloor \{1-f(x^{(0)})\}/\Delta f \right\rfloor+1$.

证明:$x^{(k_1)}$ 是 $f(x)$的局部最大点.$x^{(k_1)}$ 的邻域 $N(x^{(k_1)})$ 可以表示如下形式:$\{x^{(k_1)}-e_i\in X_z\,|\,i\in T\}\cup\{x^{(k_1)}\}\cup\{x^{(k_1)}+e_j\in X_z\,|\,j\in S\}$.根据算法 *amds*,对于任意的点 $x\in\{x^{(k_1)}-e_i\in X_z\,|\,i\in T\}$,有 $f(x^{(k_1)})-f(x)\geqslant 0$;对于 $\{x^{(k_1)}+e_j\in X_z\,|\,j\in S\}$ 中的任意点 x 同样有 $f(x^{(k_1)})-f(x)\geqslant\varepsilon=0$.根据局部最大点的定义,$x^{(k_1)}$ 是 $f(x)$的局部最大点. □

当算法 amds 计算得到点 $x^{(k)}$ 时,为获得点 $x^{(k+1)}$ 算法 amds 至多只搜索当前点 $x^{(k)}$ 的邻域内的点,对邻域外的点不考察,当 $x^{(k)}$ 为局部最大点时算法停止,显然,对任意的 $i\in T, j\in S$,点 $x=x^{(k)}-e_i+e_j\notin N(x^{(k)})$.对算法 amds 进行改进,在取出操作和放回操作失败之后增加对点 $x=x^{(k)}-e_i+e_j$ 的搜索,也就是在集合 T 和 S 间互换一个维度称为交换操作,得到算法 amdsj(direct search and jump),使之可能从一个局部最大点跳跃到另一个更好的部最大点.令算法 amdsj 的初始值为 $x^{(0)}$,记 $w=f(x^{(0)})\cdot\varepsilon$.

定理 2. 当 $\varepsilon\geqslant 0$ 时,算法 amdsj 在有限步内终止,且其迭代次数的上界为 $\left\lfloor \{1-f(x^{(0)})\}/\max\{w,\Delta f\} \right\rfloor+1$,其中初始点为 $x^{(0)}, w=f(x^{(0)})\cdot\varepsilon$ 且 $\Delta f=\min\{|f(x^1)-f(x^2)|\,\big|\,f(x^1)\neq f(x^2), x^1\in X_z, x^2\in X_z\}$.

证明:对于算法 amdsj 中的点序列 $\{x^{(k)}\}$,由 Δf 的定义可得 $f(x^{(k)})+\Delta f\leqslant f(x^{(k+1)})$,而且有 $f(x^{(k)})+w\leqslant f(x^{(k+1)})$.故有 $f(x^{(k)})+\max\{w,\Delta f\}\leqslant f(x^{(k+1)})$.其他类似定理 1 的证明即可. □

令 $g(x)=\sum_{i=1}^{p}x_ia_ib_i-t\bar{a}\bar{b}$,$u(x)=\sum_{i=1}^{p}x_ia_i^2-t\bar{a}^2$ 且 $v(x)=\sum_{i=1}^{p}x_ib_i^2-t\bar{b}^2$,那么有 $f(x)=\dfrac{g(x)}{\sqrt{u(x)v(x)}}$. $g(x)$ 和 $u(x)$ 以及 $v(x)$ 满足下面的引理.根据引理,在算法中可以方便地计算所需要的 $f(x\pm e_j)$ 和 $f(x+e_j-e_k)$.

算法. amdsj.

输入:向量 a 和 b,候选基因集合 $T=\{1,\ldots,p\}$,参数 m.

输出:不稳定基因集合 S.

1. $x^{(0)}=\{1,\ldots,1\}$, $S=\varnothing$, k=1;
2. **While** $k\geqslant 1$ **do**
3. Find $f(x^{(c)})=\max_{i\in T}f(x^{(k-1)}-e_i)$;
4. **If** $f(x^{(k-1)})+w<f(x^{(c)})$ and $x^{(c)}\in X_z$ **then**
5. $S=S\cup\{i\}$, $T=T\setminus\{i\}$, $x^{(k)}=x^{(c)}$;
6. **else**
7. Find $f(x^{(c)})=\max_{j\in S}f(x^{(k-1)}+e_j)$;
8. **If** $f(x^{(k-1)})+w<f(x^{(c)})$ and $x^{(c)}\in X_z$ **then**
9. $S=S\setminus\{j\}$, $T=T\cup\{j\}$, $x^{(k)}=x^{(c)}$;
10. **else**
11. Find $f(x^{(c)})=\max_{i\in T,j\in S}f(x^{(k-1)}-e_i+e_j)$;
12. **If** $f(x^{(k-1)})+w<f(x^{(c)})$ and $x^{(c)}\in X_z$ **then**
13. $S=S\cup\{i\}\setminus\{j\}$, $T=T\cup\{j\}\setminus\{i\}$, $x^{(k)}=x^{(c)}$;
14. **else**
15. **STOP**.
16. **end**
17. **end**
18. **end**
19. k=k+1;
20. **end**

引理 1. 设 $t=\sum_{i=1}^{p}x_i$, $\bar{a}=\sum_{i=1}^{p}x_ia_i/t$ 并且 $\bar{b}=\sum_{i=1}^{p}x_ib_i/t$.对任意的 $x\in X_z$,如果 $x+e_j\in X_z$,那么如下等式成立: $g(x+e_j)=g(x)+\dfrac{t}{t+1}(a_j-\bar{a})(b_j-\bar{b})$, $u(x+e_j)=u(x)+\dfrac{t}{t+1}(a_j-\bar{a})^2$ 且 $v(x+e_j)=v(x)+\dfrac{t}{t+1}(b_j-\bar{b})^2$.

证明:展开并合并,具体过程略.同理可证引理 2 和引理 3. □

引理 2. 设 $t=\sum_{i=1}^{p}x_i$, $\bar{a}=\sum_{i=1}^{p}x_ia_i/t$ 并且 $\bar{b}=\sum_{i=1}^{p}x_ib_i/t$.对任意的 $x\in X_z$,如果 $x-e_k\in X_z$,那么如下等式成立: $g(x-e_k)=g(x)-\dfrac{t}{t-1}(a_k-\bar{a})(b_k-\bar{b})$, $u(x-e_k)=u(x)-\dfrac{t}{t-1}(a_k-\bar{a})^2$ 且 $v(x-e_k)=v(x)-\dfrac{t}{t-1}(b_k-\bar{b})^2$.

引理 3. 设 $t=\sum_{i=1}^{p}x_i$, $\bar{a}=\sum_{i=1}^{p}x_ia_i/t$ 并且 $\bar{b}=\sum_{i=1}^{p}x_ib_i/t$.对任意的 $x\in X_z$,如果 $x-e_k+e_j\in X_z$,那么如下等式成立: $g(x-e_k+e_j)=g(x)-(a_k-\bar{a})(b_k-\bar{b})+(a_j-\bar{a})(b_j-\bar{b})-(a_j-a_k)(b_j-b_k)/t$,

$u(x-e_k+e_j)=u(x)-(a_k-\bar{a})^2+(a_j-\bar{a})^2-(a_j-a_k)^2/t$ 且 $v(x-e_k+e_j)=v(x)-(b_k-\bar{b})^2+(b_j-\bar{b})^2-(b_j-b_k)^2/t$.

定理 3. 当 $\varepsilon\geqslant 0$ 时,算法 amdsj 的时间复杂性为 $O\left(\left\lfloor\{1-f(x^{(0)})\}/\max\{w,\Delta f\}\right\rfloor\cdot pm\right)$.

证明:算法 amdsj 的一次迭代中,行 3 至多计算 p 个 $f(x-e_i)$,行 7 至多计算 m 个 $f(x+e_j)$,行 11 至多计算 pm 个 $f(x-e_i+e_j)$.根据前面的引理,只要保存部分中间变量,从 $f(x)$ 只需常数次操作就可以得到

$f(x-e_i), f(x+e_j)$ 或 $f(x-e_i+e_j)$.于是行 3,7,11 只需 pm 个常数次操作.行 1,5,9,13 只需 p 个常数次操作;行 2,4,6,8,12,14~20 需常数次操作.由定理 2 可得算法 amdsj 的时间复杂性为 $O\left(\left\lfloor \{1-f(x^{(0)})\}/\max\{w,\Delta f\}\right\rfloor \cdot pm\right)$.

定义 4(不稳定指数). P 为全体基因集合,S 为不稳定基因集合且余集 $T=P\backslash S$.设 $x^*=(x_1,\ldots,x_p)$ 是对应于 S 的解向量,即满足当 $j\in T$ 时 $x_j=1$,当 $j\notin T$ 时 $x_j=0$.那么对于任意的基因 $i\in S$,称 $df_i=f(x^*+e_i)-f(x^*)$ 为基因 i 的关于 S 的不稳定指数,简称不稳定指数.

定义 5(不稳定排序). P 为全体基因集合,S 为不稳定基因集合且 $\{df_i \mid i\in S\}$ 为其不稳定指数集合.那么对于任意的基因 $i\in S$,称其在上升序列 $\{df_{i_k} \mid 1\leqslant k\leqslant |S|\}$ 中的序为基因 i 的关于 S 的不稳定排序,简称不稳定排序.

由于遗传算法对标函数要求低,适范围较广,针对本文的优化问题设计并实现了一个基于遗传算法的方法 amga(approximate method based on genetic algorithm),作为优化问题求解时的参照.因二重结构编码的遗传算法比基本遗传算法性能优越,对取值为 0 或 1 的解进行二重结构编码,具体的遗传操作为基于排序计算适应度,随机遍历抽样选择,顺序交叉(重组)和二进制变异[21].

3 数据集

本文实验的真实数据是来自两个关于肺癌的 Affymetrix microarray 数据,这两个数据是由两个不同的研究团体使用不同型号的 Affymetrix 寡核苷酸微阵列得到的.其中一个数据是由 Affymetrix 公司 HuGeneFL 芯片产生,拥有 7 129 个探针集,称为 Michigan 数据,它包含两个类别共 96 个样本,其中 86 个为肺腺癌 lung adenocarcinoma(AD)样本,10 个正常肺 normal samples 样本(NL)[27].另一个数据由 Affymetrix 公司 HG_U95Av2 芯片产生,拥有 12 600 个探针集,称为 Harvard 数据,包含 5 个类别共 203 个样本[28].为了与 Michigan 数据相对应,我们使用 Harvard 数据中肺腺癌和正常组织两个类别的样本,即 17 个正常样本和 127 个腺癌样本.参考数据产生的原始文献来预处理数据,对 Michigan 数据的密度值以 $\log_{10}\{\max(X+100,0)+100\}$ 进行预处理和变换,对 Harvard 数据以 $\log_{10}\{\max(X,0)+10\}$ 进行预处理和变换.

由于产生数据的微阵列型号不同,用基于序列的探针匹配方法来匹配两种微阵列中使用的探针,得到 6 124 个在两种微阵列中同时出现的公共探针集,以这 6 124 公共探针集在两个数据中的表达值为分析目标.匹配和选择公共探针集的具体细节过程见文献[14]的数据预处理章节.Michigan 以及 Harvard 数据经过预处理和变换之后,6 124 个公共探针中部分探针的表达值在所有样本中完全相等,在后面的实验中去除这些表达值一致的 38 个探针,其中 1 个来自 Michigan 数据,37 个来自 Harvard 数据.最后所得的 6 086 个探针集用于后面的实验,为了描述方便,在后面章节中把每个探针集看成是一个基因.

基因表达数据经常出现样本不平衡数据[9,29].文献[9]研究表明,样本不平衡对差异表达基因的筛选有重要的影响,并且对不同的方法有不同的影响.在两个样本类间的各种样本比例下结合相等和不相等类别方差对 6 个差异表达基因筛选方法进行了性能比较,发现 Regularized t-test (Reg-t)[30],SAM[31]和 Welch t-test (Welch-t)[8]在不同的情况下有各自的优势.本文使用的 Michigan 数据和 Harvard 数据中两个类别间的样本比例分别是 8.6 和 7.47,两个实验数据都是不平衡数据.因此我们选用了 Reg-t,SAM,Welch-t 和常用的 t-test(T)[4]一共 4 种方法来计算差异表达统计量,所得的 4 种统计量的基本信息见表 1.

Table 1 The basic information of four statistics

表 1 4 种同计量的基本信息

Statistic	Harvard data				Michigan data			
	Mean	Variance	Mix	Max	Mean	Variance	Mix	Max
Reg-t	1.03	7.47	−14.38	14.51	0.02	5.20	−22.86	7.58
t-test	1.07	7.46	−13.97	13.95	0.04	5.32	−21.32	8.39
SAM	1.05	7.42	−13.79	13.78	0.02	3.41	−16.79	6.36
Welch-t	1.18	12.37	−19.86	12.37	0.28	12.98	−21.38	12.73
Mean	1.08	8.73	−15.50	13.65	0.09	6.73	−20.59	8.77

4 实验结果

本文的研究问题是识别表达不稳定的基因,并把这一问题形式化为非线性整数规划问题,对此提出 3 种求解算法:算法 amno,算法 amds 和算法 amdsj,其中算法 amno 是最简单,但是得到的结果没有保证,算法 amds 和算法 amdsj 可以获得局部最优解,且算法 amdsj 是对算法 amds 的进一步改进.首先,对提出的 3 种算法进行比较确定算法的特点以及验证前面章节的理论分析结果.然后,把选出的基因应用于分类和差异表达基因的筛选.具体的实验环境是:Pentium4 3.20Ghz 处理器,512 兆内存,Windows XP 系统.在所有的实验中参数 $\varepsilon=0$.

4.1 3种算法的比较

实验目的是测试所提出的 3 种算法,通过目标函数的大小、关于统计量的稳定性和前 k 个不稳定基因的稳定性 3 个指标来比较 3 种算法的求解结果.

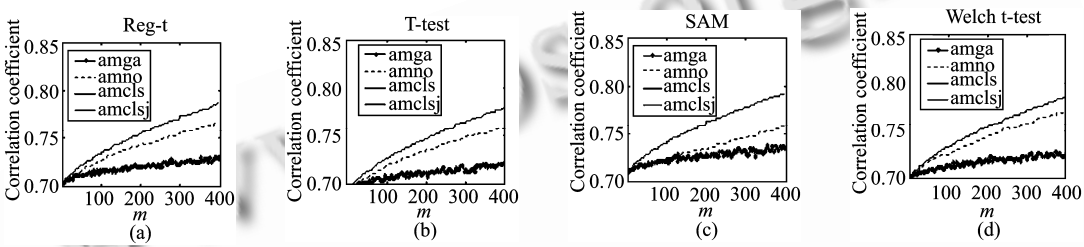


Fig.1 Values of objective function at various m

图 1 不同参数 m 下的目标函数

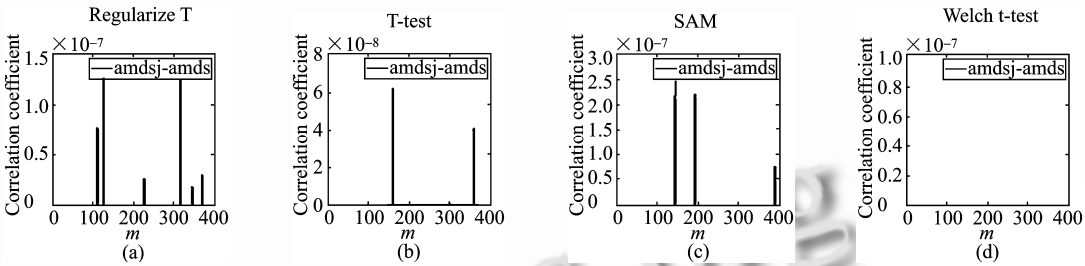


Fig.2 Difference of objective function at various m

图 2 在不同 m 下的值目标函数的差别

4.1.1 目标函数

首先考察 3 种算法中最后的目标函数(即相关系数)在不同的输入参数 m 时变化情况.图 1 描述的是在真实数据上 3 种算法中目标函数随着不同输入参数 m 的变化图(amga 方法的结果作为参照),表明算法 amds 和算法 amdsj 与算法 amno 相比能够得到更大的目标函数(amga 方法的目标函数最小),也就是算法 amds 和算法 amdsj 可以求出更好的解,发现更好的不稳定基因集合.实验的结果与前面的结论是一致的,即算法 amds 和算法 amdsj 可以获得局部最优解,而算法 amno 的解没有保证.amga 方法性能不佳的原因是本文的问题规模太大(染色体长度 6000 多位),遗传算法搜索的解,集占可行解集比例非常小.

图 1 可以区分算法 amds(j)与算法 amno,但是难以区分算法 amds 和算法 amdsj 间的差别.为了区别两者,查看 $f(x)_{amdsj}-f(x)_{amds}$ 在不同输入参数 m 下更高分辨率的图形,结果如图 2 所示.图 2 表明,算法 amds 和算法 amdsj 产生不同结果,虽然不同的结果与 m 有关.尽管两者的目标函数间的差别很小,但是这也说明了算法 amdsj 对比算法 amds 具有优点,其目标函数值更大,能够发现更好的解.

4.1.2 关于统计量的稳定性

本文中需要对基因表达数据进行处理得到差异表达统计量,然后把差异表达统计量作为分析的目标.目前,研究人员提出了多个计算差异表达统计量的方法,如 Reg-t, SAM, Welch t-test 和 t-test 等.把不同的方法应用于同一个数据,可能会得到不同的差异表达统计量,而不同的差异表达统计量进而又可能影响识别表达不稳定的基因.根据文献[9],选定 4 个有代表性的方法来计算不同的差异表达统计量,进而考察所提出的算法关于不同的差异表达统计量的稳定性.为了精确地度量算法关于差异表达统计量的稳定性,引入如下定义的不同差异表达统计量上产生的不稳定基因集合间的重合率: $OR^S=\frac{|S_m^{1*}\cap S_m^{2*}|}{m}$,其中 S_m^{1*} 和 S_m^{2*} 分别表示在给定的 m 下由差异表达统计量 1*和差异表达统计量 2*上计算得到的不稳定基因集合.某个算法产生的 OR^S 曲线位于另一个算法的 OR^S 曲线之上表明这个算法关于差异表达统计量的稳定性越好.4 种统计量两两组合一共 $C_4^2=6$ 种组合的结果见图 3.除了 Reg-t 和 T 组合以外,在其他 5 种组合中,算法 amds 和算法 amdsj 的不稳定基因集合间的重合率比算法 amno 至少高出 20%,优势非常明显.另外,图 3 表明 Welch-t 统计量和其他 3 个统计量间有着更大的差别,因为对于任何一个算法来说,在包含 Welch-t 统计量的组合上(即子图 c,e 和 f)的 OR^S 明显低于其他不含 Welch-t 统计量的组合(即子图 a,b 和 d)上的 OR^S.

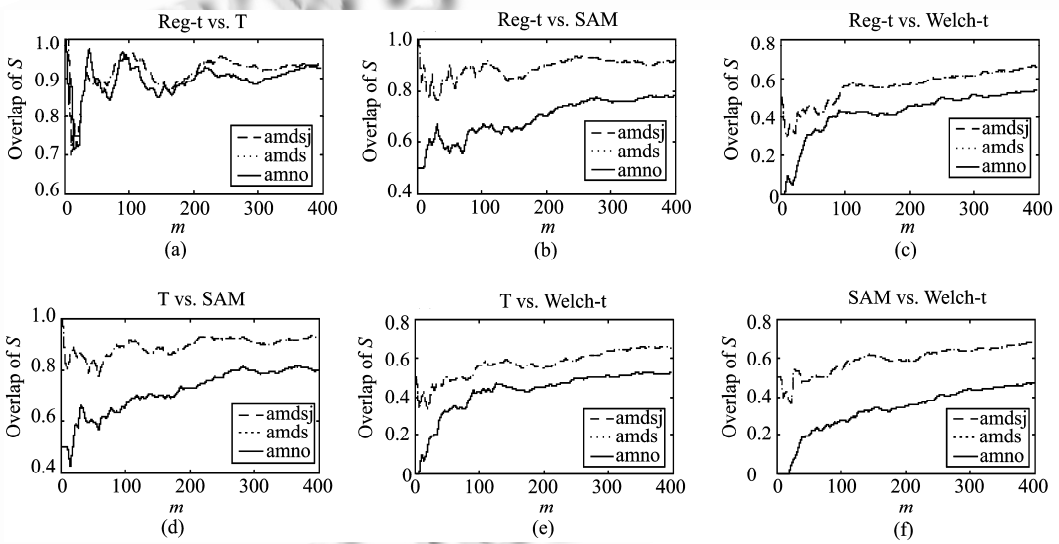


Fig.3 Overlap rates between the sets of unstable genes

图 3 不稳定基因间的重合率

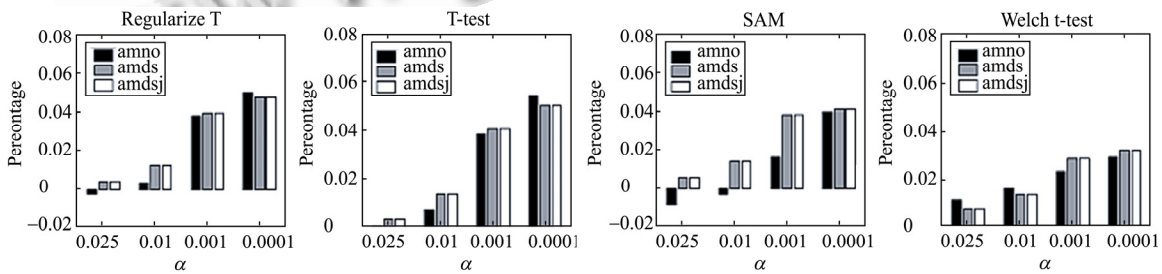


Fig.4 OR^{deg} by the unstable genes S_{100}

图 4 不稳定基因 S_{100} 引起的 OR^{deg} 的变化

4.1.3 前 k 个不稳定基因的一致性

本文中不稳定基因的集合是由输入的参数 m 限定的.对于识别出的不稳定基因集合 S_m 中的每个基因可以根据其不稳定指数进行排序,以确认集合成员间的关系.设 S_{m1} 和 S_{m2} 分别为参数 $m1$ 和 $m2$ ($m1<m2$)下识别的不稳定基因.若直接比较集合 S_{m1} 和 S_{m2},那么这两个集合间至少有 $m2-m1$ 个不同的元素;进一步地,如果 $m1$ 和 $m2$ 相差很大时,那么直接比较是不合适的.然而,我们可以借助基因的不稳定排序来比较 S_{m1} 和 S_{m2}.一个很有意思的问题是:集合 S_{m1} 中的前 k 个基因和集合 S_{m2} 中的前 k 个基因间有多大的不同? S_{50} 和 S_m (m=100,150,200,250,300)的前 50 个不稳定基因间的交见表 2,其中算法 amdsj 的结果与算法 amds 相同,故略去.可以看到,在所有的统计量上,算法 amdsj 与算法 amds 识别的集合 S_{m1} 和 S_{m2} 的前 k 个基因间差别很小.上述结果表明,对于参数 m 下的不稳定基因集合 S_m 中的基因,在后续的分析中可以按照基因的不稳定次序来处理.

Table 2 Intersection of the top-50 ranked elements from S_{50} and S_m

表 2 S_{50} 和 S_m 的前 50 个不稳定基因间的交

Statistic	Algorithm	S_{100}	S_{150}	S_{200}	S_{250}	S_{300}
t-test	amno	32	32	32	32	32
	amds	**49**	**48**	**48**	**47**	**47**
Reg-t	amno	34	34	34	34	34
	amds	**49**	**48**	**48**	**48**	**47**
SAM	amno	27	19	19	18	18
	amds	**49**	**48**	**47**	**47**	**47**
Welch-t	amno	40	37	36	37	37
	amds	**49**	**48**	**46**	**46**	**46**

4.2 应用于发现差异表达基因

识别的不稳定基因可以用来帮助识别差异表达基因.在给定的某个显著性水平下,分别从两个数据上识别的差异表达基因集合间的重现率($OR^{\deg}$)定义为 $OR^{\deg}=\frac{|\deg_\alpha^{1*}\cap\deg_\alpha^{2*}|}{|\deg_\alpha^{1*}\cup\deg_\alpha^{2*}|}$,其中 $\deg_\alpha^{1*}$ 和 $\deg_\alpha^{2*}$ 为显著性水平 α 下分别从两个不同的数据 1*和数据 2*上识别的差异表达基因集合.如图 4 所示,在宽广的显著性水平区域上,删除由算法 amds 和算法 amdsj 所识别的不稳定基因 S_{100} 可以一致地增加不同数据间的差异表达基因的重现率.

4.3 应用于分类

考虑了样本不平衡现象,文献[29]提出了稳定的基因选择方法,用度量函数量化基因的鉴别能力,并由此来排序和选择基因.在分类实验中,我们用文献[29]提出的度量函数来排序基因,从排序的基因中选出前 x 个基因.接着从训练数据中取出与前 x 个基因关联的数据,根据一个特定的分类方法构造一个分类器.然后用所构造的分类器来预测测试数据中的样本类别,预测精度称为分类器精度.因此,不同的基因集合可以根据其上所构造的分类器精度来比较.文献[32]表明支持向量机(SVMs)是最突出的、稳健的微阵列数据分类方法.在下面的实验中,采用支持向量机来构造分类器.

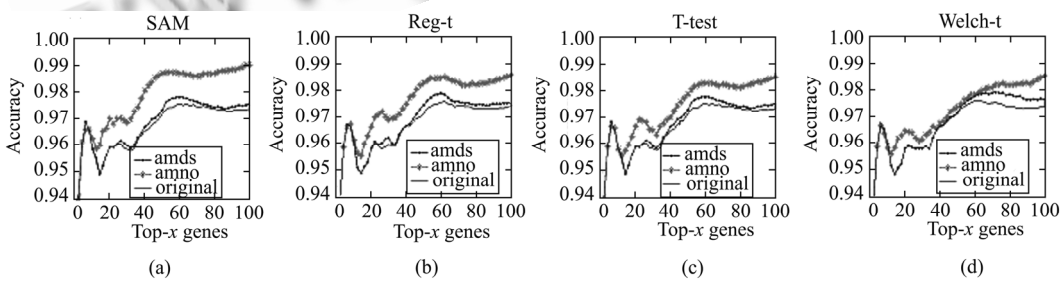


Fig.5 Result of using Harvard data as training data

图 5 Harvard 数据为训练集时的分类结果

在本实验中考虑了一个有意义的问题是:从全部基因中去除所识别的不稳定基因后再进行分类是否有助于分类的结果?我们可以比较去除和没有去除不稳定基因两种情况下的平均分类结果来回答这一问题.下文借助一个修正的 *n*-fold 交叉验证(revised cross-validation,简称 rCV)实验来获得平均分类结果.在此交叉验证实验中,测试数据保持不变,把训练数据随机地分割成近似相等的 *n* 份,然后依次完成 *n* 个流程;在每个流程中按次序把 *n* 份训练数据中的每一份去除,用余下的 *n*–1 份数据来选择基因并构造对应的分类器;然后用所得的分类器预测测试数据中的样本类别,如此 *n* 个流程得到一个平均预测结果为一次交叉验证的结果.多次交叉验证实验可以得到一个稳健的平均结果.图 5 和图 6 所示是 20 次 10-fold 修正交叉验证实验的平均结果,其中 *origianl* 曲线代表没有去除不稳定基因的情况下得到的结果.交叉验证实验表明,删除识别出的不稳定基因可以有效地提高分类精度.额外地,相比于算法 amds,去除由算法 amno 识别的不稳定基因可以更大程度地提高对分类精度;另一方面,从图 6 的 b)、c)子图可以看出去除由算法 amds 识别的不稳定基因可以更稳健地提高对分类精度.

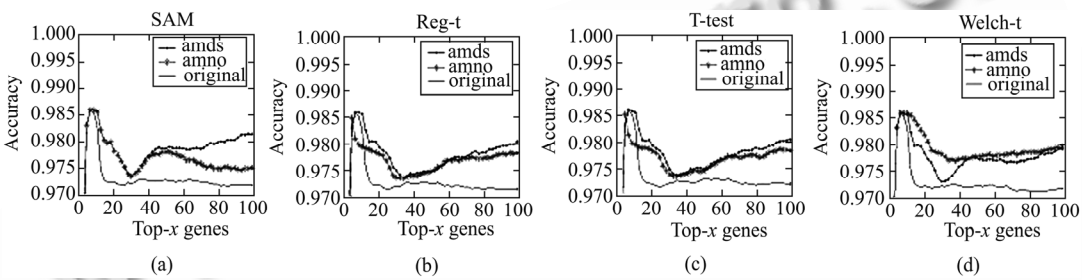


Fig.6 Result of using Michigan data as training data

图 6 Michigan 数据为训练集时的分类结果

部分结果中 amno 算法比 amds 算法更好的原因是:amno 算法依靠单个基因的统计量差值$|a_i–b_i|$(实际上是距离)来计算,而统计量 a_i 和 b_i 可以作为特征选择(基因选择)的指标,其值越大分类能力越强,amno 算法实际上选出的是$|a_i–b_i|$大的基因(不适合作为特征从一个数据分类另一个数据);amds 算法考虑了其他基因并依靠相关系数来计算,注重的是统计量的变化趋势而不是$|a_i–b_i|$;分类应用中 amno 算法的部分结果比 amds 算法好是正常的.然而 amno 算法仅考虑单个基因$|a_i–b_i|$,易受统计量计算方法和数据的影响,相对来说不如 amds 算法稳定.

5 结束语

不稳定基因可以为微阵列数据分析提供有价值的信息.例如,识别的表达不稳定基因可以用来筛选、验证现阶段小样本实验中得到的差异表达基因;另外,我们也可以在分类之前排除所识别的不稳定基因来提高微阵列数据分类中的预测结果;进一步地,所识别出的不稳定基因可以解释为什么不同数据间存在差异,进而可以为其他类型的微阵列数据集成分析提供帮助.

本文提出的 3 种算法中,算法 amno 是最简单且速度最快的一种,算法 amds 居中而算法 amdsj 是最复杂的一种.与算法 amno 相比,算法 amds 和算法 amdsj 可以得到更好的解,因为他们在评价一个基因时同时考虑了其他的基因,而算法 amno 是基于单个基因的信息来评价一个基因,没有考虑其他基因的影响.算法 amdsj 可以看成是算法 amds 的改进版本,它以更多的计算为代价得到一个更优的解,在大部分情况下两者的解相同.可以说,算法 amds 是速度和质量的一个合理折衷.

需要注意的是,用于集成分析的数据可能携带的特异性,如由微阵列平台、实验的协议、探针的匹配等引入的偏斜和噪音,它们会影响最后的结果.因此,差异表达统计量而不是数据本身被用于不稳定基因的识别以减少数据的特异性.在微阵列数据集成分析中,开发新的统计量以代替原始的数据是一个有价值的并且是富有挑战性的研究方向.另外,本文的数学模型隐含:识别的最不稳定基因是相对于其他的候选基因,并且不稳定基因的数目被输入的参数 *m* 所限定.如果没有先验知识的辅助,确定参数 *m* 的值将会是一个困难的问题.

致谢 作者向对本文的工作给予支持和建议的同行表示衷心的感谢.

References:

[1] Schena M, Shalon D Davis RW, Brown PO. Quantitive monitoring of gene expression patterns with a complementary DNA microarray. Science, 1995,270:467−470. [doi: 10.1126/science.270.5235.467]

[2] Golub TR, Slonim DK, Tamayo P, Huark C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Blomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 1999,286:531−537. [doi: 10.1126/science.286.5439.531]

[3] Petricoin EF, Hackett JL, Lesko LJ, Puri RK, Gutman SI, Chumakov K, Woodcock J, Feigal DW, Zoon KC, Sistare FD. Medical applications of microarray technologies: a regulatory science perspective. Nature Genetics, 2002,32(Supp.):474−479. [doi: 10.1038/ng1029]

[4] Cui XQ, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biology, 2003,4:210. [doi: 10.1186/gb-2003-4-4-210]

[5] Song L, Bedo J, Borgwardt LM, Gretton A, Smola A. Gene selection via the BAHSIC family of algorithms. Bioinformatics, 2007, 23:i490−i498. [doi:10.1093/bioinformatics/btm216]

[6] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. of the National Academy of Sciences, 1999,96:6745−6750. [doi: 10.1073/pnas.96.12.6745]

[7] Zhou XJ, Kao MCJ, Huang H, Wong A, Nunez-Iglesias J, Primig M, Aparicio OM, Finch CE, Morgan TE, Wong WH. Functional annotation and network reconstruction through cross-platform integration of microarray data. Nature Biotechnology, 2005,23: 238−243. [doi: 10.1038/nbt1058]

[8] Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics, 2002,18:546−554. [doi: 10.1093/bioinformatics/18.4.546]

[9] Yang K, Li JZ, Gao H. The impact of sample imbalance on identifying differentially expressed genes. BMC Bioinformatics, 2006, 7(Suppl.4):S8. [doi:10.1186/1471-2105-7-S4-S8]

[10] Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc. of the National Academy of Sciences, 2006,103:5923−5928. [doi: 10.1073/pnas.0601231103]

[11] Rhodes DR, Barrete TR, Rubin MA, Ghosh D, Chinnaiyan AM. Meta-analysis of microarrays: inter-study validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Research, 2002,62:4427−4433.

[12] Choi JK, Yu U, Kin S, Yoo OJ. Combining multiple microarray studies and modeling inter-study variation. Bioinformatics, 2003, 19:i84−i90. [doi: 10.1093/bioinformatics/btg1010]

[13] Hu P, Greenwook CMT, Beyene J. Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. BMC Bioinformatics, 2005,6:128. [doi: 10.1186/1471-2105-6-128]

[14] Jiang H, Deng Y, Chen H-S, Tao L, Sha Q, Chen J, Tsai C-J, Zhang S. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics, 2004,5:81. [doi: 10.1186/1471-2105-5-81]

[15] Xu L, Tan AC, Naiman DQ, Geman D, Winslow RL. Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. Bioinformatics, 2005,21:3905−3911. [doi: 10.1093/bioinformatics/bti647]

[16] Huttenhower C, Hibbs M, Myers C, Troyanskaya OG. A scalable method for integration and functional analysis of multiple microarray datasets. Bioinformatics, 2006,22:2890−2897. [doi: 10.1093/bioinformatics/btl492]

[17] Jörnsten R, Ouyang M, Wang H-Y. A meta-data based method for DNA microarray imputation. BMC Bioinformatics, 2007,8:109. [doi: 10.1186/1471-2105-8-109]

[18] Garey MR, Johnson DS. Computers and Intractability: A Guide to the theory of NP-completeness. San Francisco: W.H. Freeman and Company, 1979.

[19] Papadimitriou CH. On the complexity of integer programming. Journal of the ACM, 1981,28:765−768. [doi: 10.1145/322276.322287]

[20] Li F, Hao Y, Sun Y, Wu X. Operations Research Method and Model. Shanghai: Fudan University Press, 2006 (in Chinese).

[21] Wang X, Cao L. Genetic Algorithm - Theory, Application and Software. Xi'an: Xi'an Jiaotong University Press, 2002 (in Chinese).

[22] Bellare M, Rogaway P. The complexity of approximating a nonlinear program. Mathematical Programming, 1995,69:429−441.

[23] Zhu WX. An approximate algorithm for nonlinear integer programming. Applied Mathematics and Computation, 1998,93:183−193. [doi: 10.1016/S0096-3003(97)10083-2]

[24] Gu YH, Wu ZY. A new filled function method for nonlinear integer programming problem. Applied Mathematics and Computation, 2006,173:938−950. [doi: 10.1016/j.amc.2005.04.025]

[25] Ng CK, Li D, Zhang L-S. Discrete global descent method for discrete global optimization and nonlinear integer programming. Journal of Global Optimization, 2007,37:357−379. [doi: 10.1007/s10898-006-9053-9]

[26] Lei Y, Zhang S, Li J, Zhou C. MATLAB Genetic Algorithm Toolbox and Application. Xi'an: Xidian University Press, 2005 (in Chinese).

[27] Beer DG, Kardia SLR, Huang C-C, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JMG, Iannettoni MD, Orringer MB, Hanash S. Gene-Expression profiles predict survival of patients with lung adenocarcinoma. Nature Medicine, 2002,8:816−824.

[28] Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. of the National Academy of Sciences, 2001,98: 13790−13795. [doi: 10.1073/pnas.191502998]

[29] Li JZ, Yang K, Gao H, Luo JZ. Guo Z. Model-Free gene selection method by considering unbalanced samples. Journal of Software, 2006,17(7):1485−1393 (in Chinese with English abstract). http://www.jos.org.cn/1000−9825/17/1485.htm [doi:10.1360/jos171485]

[30] Baldi P, Long AD. A bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. Bioinformatics, 2001,17:509−519. [doi: 10.1093/bioinformatics/17.6.509]

[31] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc. of the National Academy of Sciences, 2001,98:5116−5121. [doi: 10.1073/pnas.091062498]

[32] Li T, Zhang C, Ogihara M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 2004,20:2429−2437. [doi: 10.1093/bioinformatics/bth267]

附中文参考文献:

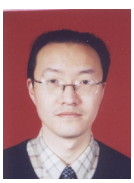
[20] 李枫,郝勇,孙焰,吴晓.运筹学方法与模型.上海:复旦大学出版社,2006.

[21] 王小平,曹立明.遗传算法——理论、应用与软件实现,西安:西安交通大学出版社,2005.

[26] 雷英杰,张善文,李续武,周创明.MATLAB 遗传算法工具箱及应用.西安:西安交通大学出版社,2005.

[29] 李建中,杨昆,高宏,骆吉洲,郭政.考滤样本不平衡的模型无关的基因选择方法.软件学报,2006,17(7):1485−1493. http://www.jos.org.cn/1000−9825/17/1485.htm [doi:10.1360/jos171485]

杨昆(1979－),男,浙江杭州人,博士,讲师,CCF 会员,主要研究领域为计算生物学,数据挖掘.



徐德昌(1964－),男,博士,副教授,主要研究领域为计算生物学,数据挖掘.

李建中(1950－),男,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,传感器网络,数据挖掘,计算生物学.

戴国骏(1965－),男,博士,教授,CFF 高级会员,主要研究领域为汽车电子,传感器网络,计算生物学,生物医学工程及仪器.