基于GIS的基因表达图谱模型的建立与应用

赵龙山[1] 于小玲[1] 胡国杰[2] 吴发启[1]* 魏晓妹[2]

([1]西北农林科技大学资源环境学院, 杨凌 712100; [2]西北农林科技大学水利与建筑工程学院, 杨凌 712100)

摘要 本文在合理假设的基础上, 根据2010年全国研究生数学建模竞赛A题提供的数据及相关信息, 在GIS的支持下构建了基因表达图谱模型(简称*GEPM*), 并对其进行空间分析, 从而达到对肿瘤识别信息基因提取的目的。结果表明, 在参与分析的1 991个基因中, 有7个基因可以作为肿瘤识别的信息基因; 通过GIS技术构建*GEPM*对于肿瘤的识别与诊断是可行的。因此, 通过本文的研究为基因的识别和研究提供了新的方法。

关键词 栅格数据; *GEPM*; 半变异函数; GIS; 肿瘤; 信息基因

利用DNA芯片技术测定的基因表达谱数据, 建立分类模型, 找出决定待测样本类别的基因“标签(或称之为信息基因)”是当前生物信息学研究领域中的重要课题[1~4], 也是正确识别肿瘤类别、给出可靠诊断的关键所在[5,6]。

Golub等[7]以“信噪比”为衡量基因对样本分类贡献大小的指标, 采用加权投票的方法进行急性白血病亚型的识别, 从7 129个样本基因中选出了50个可能的信息基因, 大大缩小了决定急性白血病亚型的基因范围。因此, 信息基因的选择就是对大规模基因表达数据降维的过程, 逐步缩小基因搜索范围, 以便能更加准确地找出基因“标签”。

本文要解决的问题是采用数学方法建立模型, 从海量的基因表达数据中提取出对肿瘤诊断、治疗等有价值的信息基因。近年来, GIS以其强大的空间数据综合分析能力, 逐渐被人们应用到各个方面[8]。GIS数据库中含有大量的空间数据和非空间数据, 比一般的数据具有更加丰富和复杂的语义信息, 隐藏着丰富的知识, 它主要依据用户要求提出假设, 并以空间可视化的方式表现数据的内容。因此, 本文我们在采用基于GIS的方法构建基因表达图谱模型的基础上, 将基因提取转化成一个空间分析的问题, 从而找出基因“标签”。为了达到这一目的, 首先, 我们假设(1)基因表达数据能够真实反映基因表达水平与肿瘤的关系; (2)所有基因表达数据之间相互独立; (3)同一肿瘤的基因表达水平在不同样本间是相近的(或接近相同)。

基于以上3个假设, 我们得出合理推断, 正常样本基因表达的空间变化相对平稳, 而肿瘤样本则相反, 因此, 我们可以以正常样本基因表达数据为基准构建包含肿瘤基因信息的基因表达图谱模型, 并通过分析基因表达图谱模型来找出信息基因。

1 材料与方法

1.1 研究数据

研究数据来自2010年全国研究生数学建模竞赛A题数据(http://ngmcm.sysu.edu.cn/), 共2 000个基因, 62个样本, 包括22个正常样本和40个癌症样本。对数据初步分析后, 去除重复基因和不完整基因表达数据, 最终确定1 991个基因表达数据作为本文的研究数据。

这里定义正常样本为*N*, 癌症样本为*C*, 可用如下公式表示:

$$N=\begin{pmatrix} n_{1,1} & n_{1,2} & \mathrm{L} & n_{1,22} \\ n_{2,1} & n_{2,2} & & \mathrm{M} \\ \mathrm{M} & & \mathrm{O} & \\ n_{1991,1} & \mathrm{L} & & n_{1991,22} \end{pmatrix}$$, 且其每一列定义为

$$N_i=\begin{pmatrix} n_{1,i} \\ n_{2,i} \\ \mathrm{M} \\ n_{1991,i} \end{pmatrix}, i=1, 2, \mathrm{L}, 22 \quad (1)$$

收稿日期: 2010-10-06 接受日期: 2010-12-20
国家自然科学基金(No.40871133)资助项目
*通讯作者。E-mail: wufaqi@263.net, zls7759989@163.com

$$C=\begin{pmatrix} c_{1,1} & c_{1,2} & \mathrm{L} & c_{1,40} \\ c_{2,1} & c_{2,2} & & \mathrm{M} \\ \mathrm{M} & & \mathrm{O} & \\ c_{1991,1} & \mathrm{L} & & c_{1991,40} \end{pmatrix}$$，且其每一列定义为

$$C_i=\begin{pmatrix} c_{1,i} \\ c_{2,i} \\ \mathrm{M} \\ c_{1991,i} \end{pmatrix}, i=1,2,\mathrm{L},40 \quad (2)$$

并定义矩阵T如下:

$$T=\begin{pmatrix} t_{1,1} & t_{1,2} & \mathrm{L} & t_{1,22} \\ t_{2,1} & t_{2,2} & & \mathrm{M} \\ \mathrm{M} & & \mathrm{O} & \\ t_{1991,1} & \mathrm{L} & & t_{1991,22} \end{pmatrix}, \quad (3)$$

其中, $t_{i,j}=\sum_{m=1}^{40}\left(n_{i,j}-c_m\right),(i=1,2,\mathrm{L},1991;\ j=1,2,\mathrm{L},22)$。

式中, T为基因表达差异; $t_{i,j}$为正常基因与癌症基因表达的差异值, 其值等于每个癌症基因与22个癌症基因表达数据的差值的和。

1.2 研究方法

1.2.1 基因表达图谱模型　基因表达谱(gene expression profile)是利用DNA微阵列探测的功能型组织的生物学表象, 根据DNA微阵列探测的原理与方法[9~11], 基因表达谱具有空间连续变化的特征, 这一特征和地理现象类似, 可用某一变量对其描述, 就像对地形空间变化描述的DEM。本文将基因表达图谱模型定义为: 基于DNA芯片测定的基因表达数据在二维空间上的连续变化过程模拟, 其数学表达为:

$$GEPM=\{G_i=\xi(P_j)|P_j(x_j,y_j,g_j)\in D, j=1,2,\cdots,\mathrm{m}\} \quad (4)$$

式中, D为研究区; P_j为采样点或某一样本的第j号基因, 该基因的空间坐标为(x_j, y_j), 基因表达数据为g_j; ξ为构成GEM的数据结构, 可以是呈规则分布的格网或不规则分布的格网, 本文中ξ定义为正方形格网, 故格网的空间位置(x_j, y_j)隐含在格网的行列号i, j中, 此时的GEM相当于一个$n\times m$的基因表达矩阵:

$$GEPM=T=\begin{pmatrix} t_{1,1} & t_{1,2} & \mathrm{L} & t_{1,22} \\ t_{2,1} & t_{2,2} & & \mathrm{M} \\ \mathrm{M} & & \mathrm{O} & \\ t_{1991,1} & \mathrm{L} & & t_{1991,22} \end{pmatrix} \quad (5)$$

由此可知, $GEPM$为描述基因表达谱空间变化差异的模型, 其值与由正常基因和肿瘤基因的表达水平共同决定。

图1列出了利用半变异函数对正常样本(N)的结果。从图中可以看出, 在以0.01为步长的情况下, 样本数据在空间相对距离小于0.06的情况下, 具有较强的空间相关性, 即N中相邻6个基因表达数据在空间上是相关的(图1)。

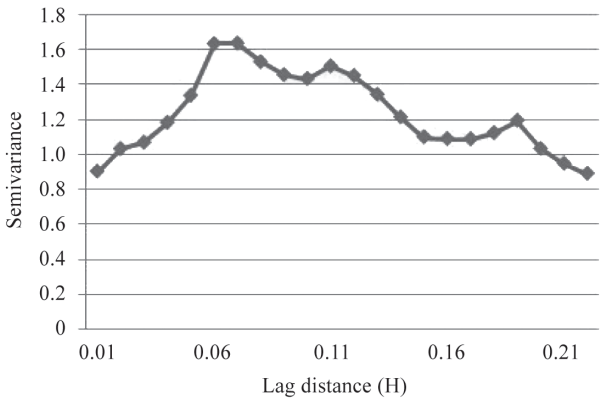

图1　样本N的空间相关性

Fig.1　Spatial correlation of the sample N

1.2.2 去噪　通过邻域分析法来消除基因表达图谱模型中变化过大的格网值, 这些值可能是由于数据获取过程中产生的一些误差, 对分析结果产生影响, 其计算公式为[12]:

$$g(i,j)=\frac{1}{m^2}\sum_{m=1}^{M}\sum_{n=1}^{M}f(m,n) \quad (6)$$

式中, $g(i, j)$为去噪后的基因表达图谱模型栅格值; 为了避免中心值过高对平均值的影响, 在运算时取相邻8个值进行计算, 其运算方法见图, 其模板如下:

$$f(m,n)=\begin{array}{|c|c|c|}\hline 1&1&1\\\hline 1&0&1\\\hline 1&1&1\\\hline\end{array}$$

模板的运算如图2所示, 为了保持数据大小不变, 运算时在数据的上、下、左、右各加一行或一列。

且定义,

$$g(i,j)^*=\begin{cases} g(i,j), \text{当}\left|f(i,j)-g(i,j)\right|>T \\ f(i,j), \text{当}\left|f(i,j)-g(i,j)\right|\leqslant T \end{cases} \quad (7)$$

式中, T为阈值, 可通过基因表达图谱模型频率分布图来确定; $g(i, j)^*$为最优栅格值。

7	7	5	2	4	…
7	7	5	2	4	…
4	4	5	6	3	
5	5	9	4		
3	3	7	5		
8	8				
⋮	⋮				

图2　模板的运算方法

Fig.2　Algorithm for template

此时, 基因表达图谱模型的矩阵表示为:

$$GEPM_{n\times m}=\begin{pmatrix} g_{11}^{*} & \mathrm{L} & g_{1m}^{*} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ g_{n1}^{*} & \mathrm{L} & g_{nm}^{*} \end{pmatrix} \tag{8}$$

1.2.3 梯度变换　　如果某一基因为突变基因, 则在该位置处数据差异较大, 在图像上可用一组线性目标来表示, 因此, 对特征基因的提取就转换成一个线性目标的提取的问题。

设基因表达谱模型V_i(i=1, 2, …, *v*)格网点(*x*, *y*)处的值为*f*(*x*, *y*), 可定义[12]:

$$|gradf(x,y)|=\sqrt{t_1^2+t_2^2}\ ,\ t_1=\frac{\partial f(x,y)}{\partial x},\ t_2=\frac{\partial f(x,y)}{\partial y} \tag{9}$$

对于基因表达图谱模型而言, 式(9)中的连续导数形式可以用求差来近似表示。即

$$\begin{cases} t_1=f(x,y)-f(x+1,y) \\ t_2=f(x,y)-f(x,y+1) \end{cases} \tag{10}$$

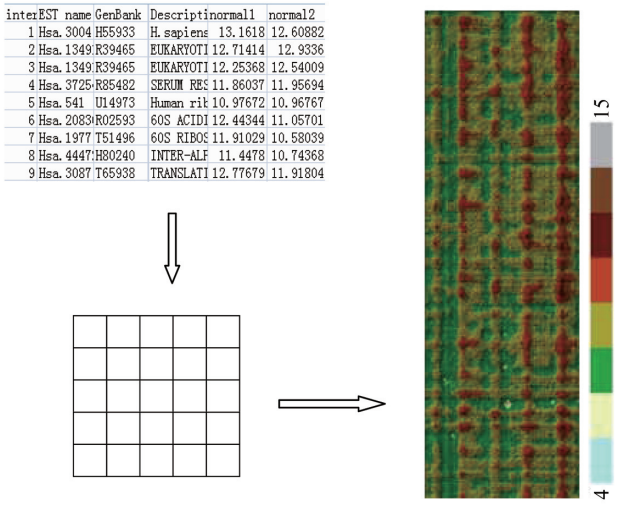

图3　*GEPM*的空间表示方法

Fig.3　Space representation method of *GEPM*

则

$$|gradf(x,y)|\cong|f(x,y)-f(x+1,y)|+|f(x,y)-f(x,y+1)| \tag{11}$$

利用该方法对基因表达图谱模型进行计算, 可以达到突出信息基因, 有效缩小搜索范围的目的。

1.2.4 水平搜索　　水平搜索的目的是提取信息基因, 其计算公式为[12]:

$$g(i,j)=\frac{\sum_{m=1}^{M}\sum_{n=1}^{N}f(m,n)\varphi(m,n)}{\sum_{m=1}^{M}\sum_{n=1}^{N}\varphi(m,n)} \tag{12}$$

其中, $g(i, j)$表示信息基因, $\varphi(m, n)$, $f(m, n)$为数据分析的窗口, 其原理可用如下模板表示:

$f(m,n)$，$\varphi(m,n)=$

-1	-1	-1
0	0	0
1	1	1

2　结果

基于GIS建立的*GEPM*见图3, 图中表示了从原数据到*GEPM*的数据结构, 基于规则格网的*GEPM*在4到15之间变化, 均值为9.5, 标准偏差为6.32。图4为对原数据去噪和梯度变换的效果图, 从图中可以看出, 去噪对数据的影响比较明显, 去噪后的数据整体变化较为平滑。图5显示了利用*GEPM*提取信息基因的结果, 通过对其的分析, 我们提取了标号为R38636、X51416等的7个基因为信息基因, 结果见表1。

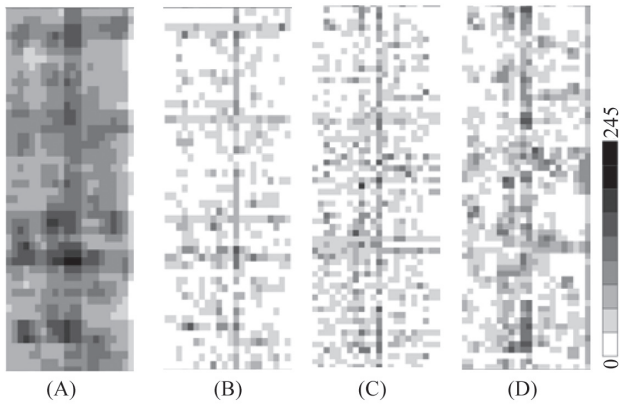

图4　去噪与梯度变换效果图

A: 去噪前; B: 去噪后; C: 梯度变换前; D: 梯度变换后。

Fig.4　Compared figure of denoising and gradient change

A: before denoising; B: denoising; C: before gradient change; D: gradient change.

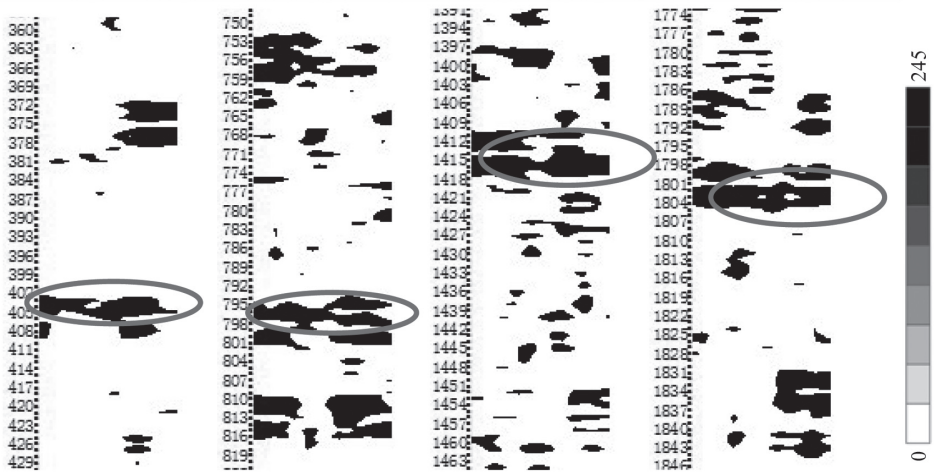

图5 信息基因分布图

Fig.5 Distribution figure of informative genes

表1 *GEPM*提取的信息基因

Table 1 Informative genes extracted from *GEPM*

基因标号 GenBank Acc No.	基因描述 Description of the genes
R38636	Urokinase plasminogen activator surface receptor, GPI-anchored (Human)
X51416	Human mRNA for steroid hormone receptor hERR1
X12548	Human mRNA for lysosomal acid phosphatase (EC 3.1.3.2)
M28882	Human MUC18 glycoprotein mRNA, complete cds.
L20859	Human leukemia virus receptor 1 (GLVR1) mRNA, complete cds.
T40674	30S RIBOSOMAL PROTEIN S18 (Cyanophora paradoxa)
U18299	Human damage-specific DNA binding protein DDBa p127 subunit (DDB1) mRNA, complete cds.

3 讨论

本文主要研究了识别肿瘤的信息基因提取方法。通过对已有研究成果的分析, 在合理假设的基础上, 构建了基因表达图谱模型(*GEPM*), 从而将信息基因的提取问题转换为一个空间分析的问题, 再利用GIS的空间分析方法, 建立了信息基因提取的模型, 最终从1 991个样本基因中提取了7个作为肿瘤识别的信息基因, 为肿瘤的临床诊断与研究提供了新的方法。

参考文献(References)

1 Theodoridis S, Koutroumbas K. Pattern Recognition, 2nd Edition. New York: Academic Press, 2003, 177-9.

2 Ramaswamy S, Golub TR. DNA microarrays in clinical oncology. J Clin Oncol 2002; 20(7): 1932-41.

3 Lander ES, Weinberg RA. Genomics: journey to the center of biology. Science 2000; 287(5459): 1777-82.

4 李颖新, 阮晓钢。基于基因表达谱的肿瘤亚型识别与分类特征基因选取研究。电子学报 2005; 33(4): 651-55.

5 李 泽, 包 雷, 黄英武, 孙之荣。基于基因表达谱的肿瘤分型和特征基因选取。生物物理学报 2002; 18(4): 413-7.

6 张 娅, 饶妮妮, 王 敏, 徐尚蕾。一种基于基因表达谱的结肠癌特征提取方法。航天医学与医学工程 2008; 21(4): 356-60.

7 Golub TR, Slonim DK, Tamayo P. Molecular classification of cancer: class discovery and prediction by gene expression monitoring. Science 1999; 286(5439): 531-7.

8 黄杏元, 马劲松, 汤 勤。地理信息系统概论。北京: 高等教育出版社, 2001, 18-9.

9 Lishutz D, Morris D, Chee M, Hubbell E, Kozal MJ, Shah N, *et al*. Using oligonucleotide probe arrarys to access genetic diversity. Bio Techniques 1995; 19(3): 442-7.

10 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al*. Gene ontology: tool for the unification of biology. Nature Genet 2000; 25: 25-9.

11 Beissbarth T, Speed TP. Gostat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 2004; 20(9):1464-5.

12 汤国安, 张友顺, 刘咏梅。遥感数字图像处理。北京: 科学出版社, 2004, 25-36.

Research on the Gene Expression Profile Model Based on GIS

Long-Shan Zhao[1], Xiao-Ling Yu[1], Guo-Jie Hu[2], Fa-Qi Wu[1]*, Xiao-Mei Wei[2]

([1]*College of Resources and Environment*, *Northwest A & F University*, *Yangling 712100*, *China*; [2]*College of Water Resources and Architectural Engineering*, *Northwest A & F University*, *Yangling 712100*, *China*)

Abstract In this paper, in support of GIS technology, we established gene expression profile models (in short of *GEPM*) and analyzed its spatial information based on the reasonable assumption, according to the dates and related information provided by the question A of 2010 National Graduate Mathematical Modeling Contest, so as to achieve extraction of informative genes which can identify tumor. The results showed that among 1 991 genes analyzed, there were 7 genes can be regarded as informative genes which could identify the tumor; It was feasible to identify and diagnose the tumor through GIS technology to build *GEPM*. Therefore, this study provided a new approach to identify and study genes.

Key words Grid data; gene expression profile model; semivariogram; geography information system; tumor; informative genes

Received: October 6, 2010 Accepted: December 20, 2010

This work was supported by the National Natural Science Foundation of China (No.40871133)

*Corresponding author. E-mail: wufaqi@263.net, zls7759989@163.com