微阵列数据扰动对错误发现率方法筛选差异表达基因的影响

陈婉婷[1],刘成友[1],吴 静[2],丁 勇[2*]

([1]南京医科大学生物医学工程系,[2]数学与计算机教研室,江苏 南京 210029)

[摘 要] **目的**:探讨微阵列数据的扰动对错误发现率(FDR)方法筛选差异表达基因的影响。**方法**:用计算机模拟仿真的方法,对 1 991 个结肠癌微阵列基因数据给予不同相对误差限的随机扰动,每个扰动进行 1 000 次随机模拟;用 FDR 的 ALSU 方法对无扰动数据与有扰动数据分别筛选差异表达基因,比较两者之间的重复率;分析数据扰动对每次基因排序位次变化的影响。**结果**:差异表达基因的单个平均重复率与总体平均重复率都随数据扰动的增加而下降。差异表达越显著的基因,受扰动误差的影响越小;在扰动误差限≤50%时,数据扰动与差异表达基因总体平均重复率呈线性递减趋势,数据扰动误差限每增加 1%,总体平均重复率约下降 1.85%。扰动误差限越大,基因排序位次的波动越大。**结论**:数据扰动是导致差异表达基因可重复性差的原因,用计算机模拟的方法可定量探讨数据扰动对差异基因筛选的影响。

[关键词] 微阵列;差异表达基因;错误发现率;数据扰动

[中图分类号] R735.5,O212 [文献标志码] A [文章编号] 1007-4368(2014)07-991-05

doi:10.7655/NYDXBNS20140728

Effect of data perturbation in microarray on selecting differentially expressed genes by false discovery rate

Chen Wanting[1], Liu Chengyou[1], Wu Jing[2], Ding Yong[2*]

(*[1]Department of Biomedical Engineering, [2]Department of Mathematics and Computer, NJMU, Nanjing 210029, China*)

[Abstract] **Objective**: To investigate the effect of data perturbation in microarray on selecting differentially expressed genes by false discovery rate (FDR). **Methods**: A total of 1 991 DNA microarray data of colon cancer were afforded random perturbation of different error limits based on a computer simulation. Every perturbation comprised 1 000 random simulations. The differentially expressed genes were selected from data with and without perturbation, respectively, by adaptive linear step-up (ALSU), a method of FDR. The repetition rates between both results were compared. The effect of each gene sort order was analyzed by data perturbation. **Results**: The single average and overall average repetition rates of differentially expressed genes both decreased with increasing data perturbation. The more significant differentially expressed the genes, the less they were affected by perturbation. When the error limit was less than or equal to 50%, the overall average repetition rate of differentially expressed genes decreased with increasing data perturbation linearly. For each 1% increase of perturbation error limit, the overall average repetition rate decreased approximately by 1.85%. The higher the perturbation error limit, the greater the fluctuation the gene sort order had. **Conclusion**: Data perturbation is a reason why differentially expressed genes exhibit low repeatability; the effect of data perturbation on selecting differentially expressed genes can be quantitatively investigated by using computer simulation.

[Key words] microarray; differentially expressed genes; false discovery rate; data perturbation

[Acta Univ Med Nanjing, 2014, 34(07): 991-995, 1002]

DNA 微阵列技术(基因芯片技术),是最近数年发展起来的一种能快速、高效检测 DNA 片段序列、基因表达水平的新技术。基因表达谱是指利用 DNA 微阵列测定组织样本中基因的表达水平,通常用矩阵形式表示。DNA 微阵列技术使“癌变基因”的发现以及对基因组数据库增加功能注释等难题得以解决,大大推进了包括人类基因组在内的各类基因组

[基金项目] 江苏省高校自然科学基金(13KJB310007);南京医科大学科技发展基金重点项目(2013NJMU006)

*通信作者(Corresponding author),E-mail:yding@njmu.edu.cn

研究[1-2]。

在对癌症相关的基因表达谱数据的研究中发现了这样的问题:同一种癌症的不同基因表达谱数据中发现的差异表达基因仅有少部分是重叠的,即可重复性差[3-4],例如文献[5]研究两套与乳腺癌转移相关的数据,结果显示分别有 482 和 462 个差异表达基因,但仅有 64 个基因交叠。可重复性差的原因值得探讨,否则将使高通量技术受到质疑,进一步的应用也必将受到影响。

由于各种原因,基因表达会受到多种干扰因素的影响[6]。例如,表达谱基因芯片的生物学误差和实验系统误差[7],都会使测出的基因表达水平受到影响,这种现象我们称为数据扰动。微阵列数据量是高通量的,每一个基因的小小扰动,组合在一起,效果不容忽视;微阵列数据处理的计算量也是高通量的,这两个高通量的叠加,有可能使微阵列数据的微小扰动对数据处理结果的影响很大。因此表达谱芯片数据的可靠性分析和处理就显得非常重要。

微阵列数据通常样本含量较小,而变量数(基因数)有成千上万个。因此在微阵列数据的基因差异性表达分析中,多重比较问题显得尤为重要[8]。传统的多重比较方法控制的是总体错误率(family-wise error rate,FWER)。对微阵列数据进行多重比较时,如果要将整个检验的 FWER 控制在 0.05,每次检验的水准就必须控制得很低,这样只能发现少数表达差异极大的基因或蛋白等,导致检验效能大大降低。因此不适合微阵列数据的多重比较。近年来,取而代之的是 Benjamini 和 Hochberg[9]做出的开创性的工作,他们提出的错误发现率(false discovery rate,FDR)方法,突破了控制第一类(假阳性)错误的传统方法,将错误识别的比例控制在允许的范围内。基于该方法的重要性和应用的广泛性,对该方法的改进和完善不断出现,如最初的 BH 法(Benjamini and Hbchberg,1995 年)、BL 法(Benjamini and Liu[10],1999 年)、BY 法(Benjamini and Yekutiel[11],2001 年)、ALSU 法[12](adaptive linear step-up,2005 年)。

本研究用结肠癌的微阵列数据,采用计算机模拟仿真数据扰动的方法,探讨基因表达谱数据的扰动对 FDR 方法的影响,找出数据扰动对差异表达基因重复率影响的定量关系。

1 材料和方法

1.1 材料

结肠癌(colon cancer)是常见的恶性肿瘤之一,病因尚未完全清楚,目前认为主要是由环境因素与遗传因素综合作用的结果,以手术切除为主要治疗方法。结肠癌若早发现、早治疗,可以大大提高治愈率和生存期。本文采用 Affymetrix 公司的结肠腺癌基因表达谱实验数据,原实验为采用点有 6 500 个寡聚核甘酸探针组的基因芯片,样本包括 40 例结肠腺癌组织和 22 例正常结肠组织。本文采用 Alon 等[13]筛选出的包含 2 000 个基因的表达谱数据进行分析(http://microarray.princeton.edu/oncology/affy data/index.html)。去除重复和空白的基因后,有 1 991 个不同基因表达数据,经过对数处理(取以 2 为底的对数)。这样得到 1 个 1 991 行、62 列(40 例肿瘤样本和 22 例正常样本)的数据矩阵。

1.2 方法

1.2.1 随机扰动数据的生成

用文献[14]的方法产生随机扰动的数据。其原理为:设近似值 x^* 与准确值 x 的相对误差为 ε,按误差理论,误差服从均值为 0 的正态分布,即 $\varepsilon\sim N(0,\sigma^2)$。由正态分布的 3σ 原则可知,$P(|\varepsilon|\leqslant 3\sigma)=0.997\ 3\approx 1$,设相对误差限为 e,因为误差限是最大的误差,有 $P(|\varepsilon|\leqslant e)=1$,故取 $\sigma=\frac{e}{3}$。再用计算机生成 $\varepsilon\sim N(0,\sigma^2)$随机数,令 $x^*=(1+\varepsilon)x$,则 x^*为满足相对误差限为 e 的随机扰动的数据。

1.2.2 差异表达基因的筛选

在多重比较的 FDR 方法中,ALSU 方法检验效能最高[15],用该方法筛选差异表达基因。

记 m = 1 991,设 m 这个基因表达数据的零假设检验为 $H_{01},H_{02},\cdots,H_{0m}$,其中 H_{0i} 为:第 i 个基因为非差异表达基因;备择假设 H_{1i} 为:第 i 个基因为差异表达基因。首先对每个基因的肿瘤样本和正常样本的数据做方差齐性检验,其中有 94.07%的基因 $P>0.05$,不妨认为数据满足方差齐性要求;然后用成组比较的 t 检验,分别得到这 1 991 个基因对应的 P 值,再将这些 P 值从小到大进行排序,记为$\{P_{(i)}\}$,相应的零假设检验为:$H_{0(1)},H_{0(2)},\cdots,H_{0(m)}$。

使用 FDR 的 ALSU 算法,筛选出差异表达基因,并对$\{P_{(i)}\}$值进行调整,得到相应的调整$\{\widetilde{P}_{(i)}\}$值。步骤如下:

(1)令 $r(\lambda)=\#\{P_{(i)}\leqslant\lambda\}$,即 $P_{(i)}\leqslant\lambda$ 的个数($1\leqslant i\leqslant m$),我们取 $\lambda=0.5$;

(2)估计 $H_{0(i)}$为真的个数 $m_0=[m-r(\lambda)]/(1-\lambda)$;

(3)从 $P_{(m)}$开始,估计$\hat{k}=\max_{1\leqslant k\leqslant m}\{k:P_{(k)}\leqslant\frac{k}{m_0}\alpha\}$,取

检验水平 α=0.05;

(4) 若存在$\hat{k}$, 则拒绝以前的假设:$H_{0\ (1)}$,$H_{0(2)}$,…,$H_{0(\hat{k})}$,即对应的基因为差异表达基因;否则所有的 $H_{0(i)}(1\leqslant i\leqslant m)$均不被拒绝;

(5)调整 $P_{(i)}$值为$\widetilde{P}_{(i)}=\min\limits_{1\leqslant k\leqslant m}\{\min(\frac{m}{k}P_{(k)},1)\}$。

1.2.3　重复率

为了评价数据扰动的影响,提出总体差异表达基因平均重复率指标和单个差异表达基因平均重复率指标。

将原始的 1 991 个肿瘤样本和正常样本的微阵列数据看成无扰动的数据, 对检验水准 α, 用 FDR 的 ALSU 算法筛选出差异表达基因, 该差异表达基因集记为 B_0,其中基因个数为 N_0。将上述微阵列矩阵数据分别给予不同的相对误差扰动,设相对误差限为 e,取 $\sigma = e/3$,再产生正态分布 $N(0,\sigma^2)$的随机数 ε,从而得到原始数据(1+ε)倍的模拟扰动数据。再对有扰动的肿瘤样本和正常样本数据,用 FDR 的 ALSU 算法筛选出差异表达基因,由于数据发生了变化,筛选出的差异表达基因也有所不同,记该差异表达基因集为 B_e,其中基因个数为 N_e。

记 $B_0\cap B_e$ (即无扰动与有误差扰动情况下都被检验为差异表达的基因)的基因个数为 N_{e0},则总体差异表达基因的重复率 $TR_e = 2N_{e0}/(N_0+N_e)$。为了排除随机因素的干扰, 将上述过程随机重复了 1 000 次,得到 1 000 个总体差异表达基因的重复率,再取平均值,得到总体平均差异表达基因重复率 $\overline{TR}_e$。

对确定的相对误差限, 在 1 000 次随机模拟过程中,单个基因被筛选为差异基因的次数的平均值,称为单个差异基因的平均重复率 $\overline{SR}_e$。

2　结　果

2.1　FDR 控制结果

多重比较检验中有 426 个基因的 P 值≤0.05,经 FDR 调整后有 156 个基因的 P 值≤0.05。t 检验的 P 值和 FDR 控制调整后的 P 值见图 1。可以看出,在 0~0.6 的范围内,同一基因的 P 值,FDR 调整后比 t 检验更高; 在> 0.6 的范围内,FDR 调整后比 t 检验更低。将 P 值的范围(0,1)区间 20 等分,观察比较每个区间选中的基因个数(图 2),t 检验基因个数最多的 P 值区间是(0,0.05),并且其个数比其他区间多出几倍。而 FDR 调整后基因个数最的多 P 值区间是(0.7,0.75),且无基因的 P 值> 0.75。因此,FDR 比传统多重假设检验更能控制犯Ⅰ类错误的概率,使得差异性基因的选择更为准确。

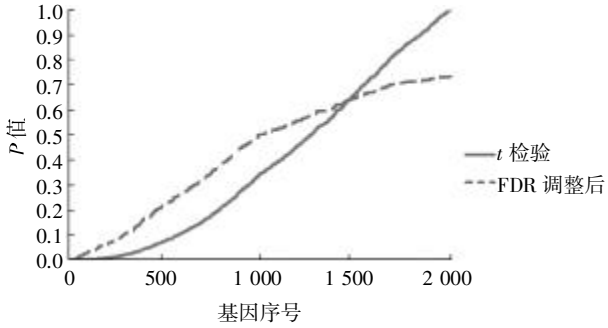


图 1　t 检验和 FDR 调整后的 P 值

Figure 1　The P-value of t-test and after FDR adjusted

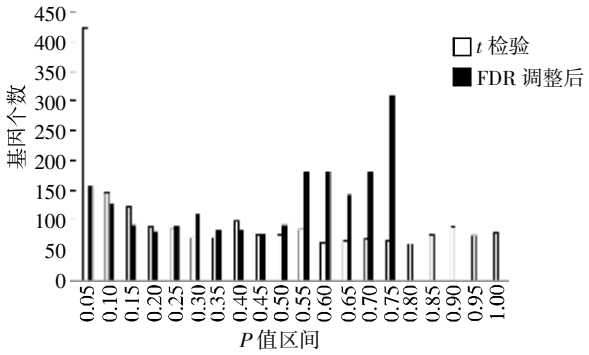


图 2　t 检验与 FDR 调整后 P 值区间的基因个数分布

Figure 2　Gene's number distribution in P-value range of t-test and after FDR adjusted

对 α = 0.01,也得到了类似的结论。

2.2　数据扰动对重复率的影响

共取了 14 个不同的相对误差限:e=5%,10%,…,65,70%,探讨 e 与 $\overline{TR}_e$ 及 $\overline{SR}_e$ 的定量关系。

取 α = 0.05, 对于不同相对误差限 e 的数据扰动,总体差异基因平均重复率 $\overline{TR}_e$ 见表 1,图 3。不难看出, 当误差扰动增大时,$\overline{TR}_e$ 降低。对相对误差限> 50%的情况,误差与 $\overline{TR}_e$ 的线性特征非常明显,考虑到相对误差限为 0 时,$\overline{TR}_e$ 为 1, 故拟合截距为 1 的回归直线,拟合直线为 $\overline{TR}_e = 1-1.849e$,决定系数 $R^2 = 0.993$。可见,微阵列数据扰动对 FDR 控制的结果有直接影响。相对误差增加 1%,$\overline{TR}_e$ 降低约 1.85%。当相对误差限>50%时,$\overline{TR}_e$< 0.1,呈指数下降趋势。

对 α = 0.01,也得到了类似的结论。对相对误差限< 50%的情况, 误差与 $\overline{TR}_e$ 的线性特征也非常明显, 拟合直线为 $\overline{TR}_e = 1-1.847e$, 决定系数 R^2 = 0.993。

将经 FDR 调整后的 P 值< 0.05 的 156 个基因,在相对误差限为 10%、20%和 30%情况下的单个差异基因平均重复率 $\overline{SR}_e$ 汇总于图 4。它们的变化趋

表 1　相对误差限与总体差异基因平均重复率

Table 1　Relative error limits and overall average repetition rates of differential genes

e(%)	$\overline{TRe}$
5	0.931
10	0.850
15	0.752
20	0.647
25	0.540
30	0.431
35	0.328
40	0.239
45	0.158
50	0.102
55	0.065
60	0.045
65	0.029
70	0.020

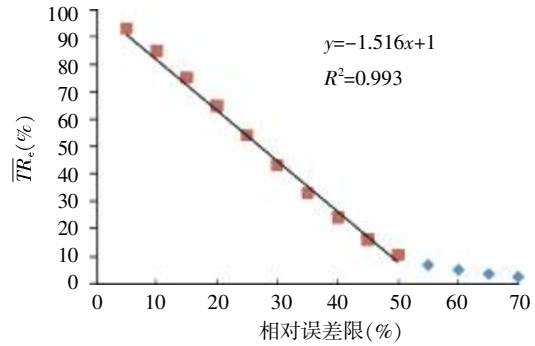


图 3　相对误差限与总体差异基因平均重复率的趋势图

Figure 3　The trend of relative error limits and overall average repetition rates of differential genes

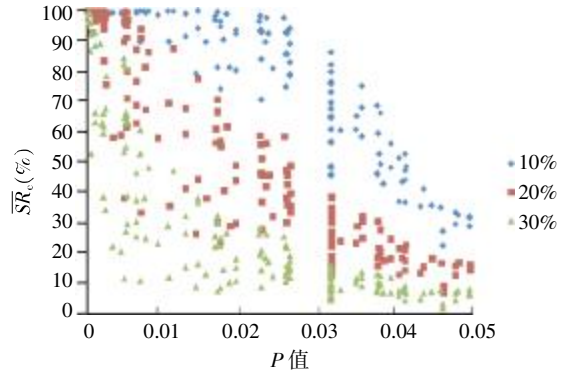


图 4　调整后 P 值与单个差异基因平均重复率的趋势图

Figure 4　The trend of adjusted P-value and single average repetition rates of differential genes

势为:相对误差限越大,$\overline{SR}_e$ 越低;调整后 P 值越大,$\overline{SR}_e$ 就越低;差异表达越显著的基因,受扰动的影响较小,如当相对误差限< 10%,调整后 P < 0.025 时,很多差异表达基因 $\overline{SR}_e$ 为 1,说明它们没有受到误差的影响。

2.3　*数据扰动对基因排序的影响*

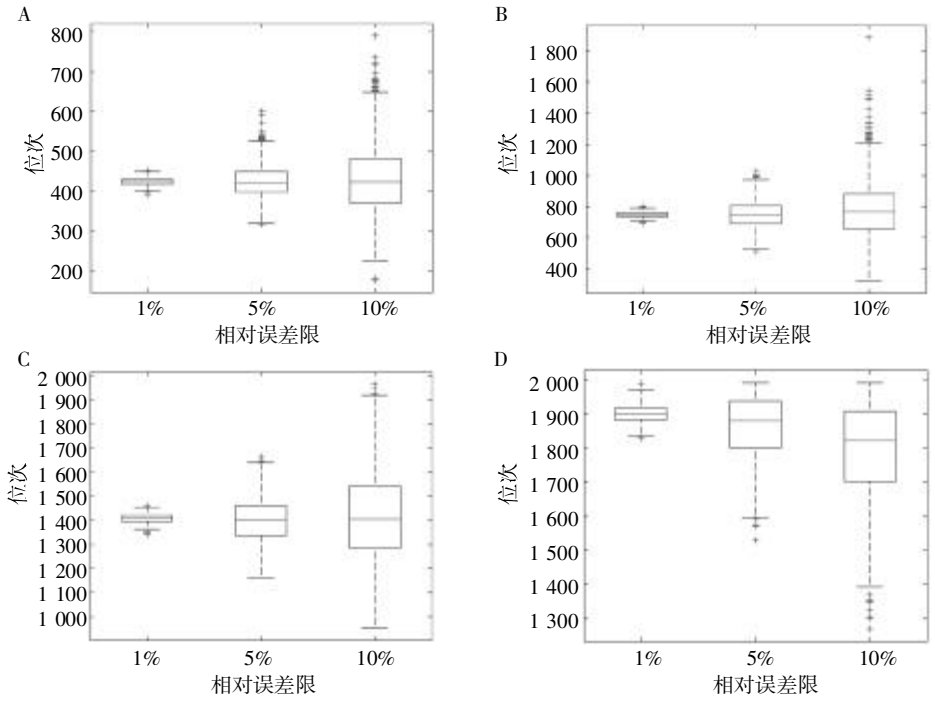
由于基因较多,按分层抽样方法随机选取4 个基因考察数据扰动对基因排序的影响。

将 1 991 个基因按原始数据 t 检验的 P 值从小到大排序编号,平均分为 4 组(前 3 组每组 498 个,P 值分别为 2.884×10^{-8}~0.074 4,0.074 4~0.341 4,0.342 1~0.662 8,第 4 组 497 个基因,P 值为 0.662 8~0.998 8)。每组随机抽样 1 个基因,得到序号为 426、746、1 406、1 901 的基因,分别作为不同大小 P 值的抽样代表。对这 4 个基因分别加入相对误差限 e 为 1%、5%、10%的随机误差扰动重新进行排序,重复 1 000 次,得到 1 000 个排序的位次变化的箱形图如图 5 所示。由图 5 可知,随着相对误差限增大,箱形图的上四分位数与下四分位数之间的差距增大,最大观测值与最小观测值之差也增大,同时离群点增多,说明误差对基因的排序有一定的影响,扰动越大,排序的波动也越大。

3　讨　论

作为一种新的用于多重比较的方法,FDR 的应用领域在不断拓宽,因此该算法的稳定性是一个令人关注的问题。从上面分析可知,原始数据的微小扰动不会引起筛选差异基因较大的差别,1%的相对误差限引起的重复率下降不到 2%;差异表达越显著的基因,其重复率越高,因此,FDR 算法是相对稳定的,是筛选差异表达基因的很好的统计方法。所以,重复率低的主要原因,是数据较大扰动引起的。

表达谱基因芯片的生物学误差和实验测量误差是微阵列数据扰动的主要原因。实验测量误差包括荧光扫描精度、RNA 提取与反转录误差、杂交过程带来的误差、操作过程的误差、克隆准确性等。随着科学技术水平的不断完善,仪器设备质量提高、实验手段、方法的完善,实验测量误差会得到更好的控制。而基因表达的生物学差异,将是使基因表达水平受到影响的更主要因素。当同一种癌症的不同基因表达谱数据中发现的差异表达基因仅有少部分重叠时,要分析数据扰动大小和来源,除了实验测量误差,更要从生物学角度出发,考虑基因表达的时间特异性和空间特异性所导致的差异,因为基因表达是一个非常复杂的生物学过程;另外不同个体或相同个体的不同细胞之间的基因表达模式有时并不一致。通过分析同一种癌症的不同基因表达谱数



A、B、C、D:分别为序号为 426、746、1 406、1 901 的基因。

图 5　分层随机抽样的 4 个基因在不同数据扰动情况下 1 000 个排序位次的箱形图

Figure 5　Box plots of 1 000 sort orders of four genes by stratified random sample in different data perturbations

据的可比性,选择更科学的方法,分析研究基因表达谱数据。

计算机模拟仿真是应用日渐广泛的一种科学方法,应用它便于重复进行试验和控制参数,预测系统的行为效果,通过大量的重复试验,获得其平均意义上的特性指标。因此,该方法为开展生物信息学研究提供了有力的工具。本文用计算机模拟的方法研究了微阵列的扰动对 FDR 方法筛选差异表达基因的影响。事实上,还可以研究模拟微阵列扰动对其他算法(如 SVM[16]等的基因识别算法)的影响,进一步地,研究差异表达基因的重复率对富集分析的影响,从而更深层次、全方位地从生物学通路角度分析数据扰动对重复率低的影响和原因。其他高通量组数据如蛋白质表达组数据、肿瘤癌型微阵列数据分类也存在这种类似的重复率低的问题。因此本文研究的方法也可以应用于评价这些高通量数据的误差扰动影响,因而对疾病标志的发现有理论指导意义。

[参考文献]

[1] Schena M,Shalon D,Davis RW,et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray[J]. Science,1995,270(20):467-470

[2] Schena M,Shalon D,Heller R,et al. Parallel human genome analysis:Microarray based expression monitoring of 1000 genes [J]. Proc Natl Acad Sci USA,1996,93(20):10614-10619

[3] Ein-Dor L,Zuk O,Domany E. Thousands of samples ale needed to generate a robust gene list for predicting outcome in cancer [J]. Proc Natl Acad Sci USA,2006,103(15):5923-5928

[4] Zhang M,Yao C,Guo Z,et al. Apparently low reproducibility of true differential expression discoveries in microarray studies [J]. Bioinformatics,2008,24 (18):2057-2063

[5] 邹金凤,郝春香,洪贵妮,等. 乳腺癌转移相关基因与功能识别的可重复性[J]. 生物信息学,2012,10(1):27-30

[6] Lee ML,Kuo FC,Whitmore GA,et al. Importance of replication in microarray gene expression studies:statistical methods and evidence from repetitive cDNA hybridizations [J]. Proc Natl Acad Sci,2000,97 (18):9834-9839

[7] 罗　瑶,徐　宏,李　瑶,等. 表达谱基因芯片的可靠性验证分析[J]. 遗传学报,2003,30(7):611-618

[8] 荀鹏程,赵　杨,柏建岭,等. 微阵列数据的多重比较[J]. 中国卫生统计,2006,23(1):5-8

[9] Benjamini Y,Hochberg Y. Controlling the false discovery rate:A practical and powerful approach to multiple testing[J]. J R Statist Soc B,1995,57(1):289-300

(下转第 1002 页)

2 结 果

3 例患耳术后 2~3 d 红肿消散,无疼痛,术后 5~6 d 无明显引流液,拔出负压引流管,术后第 10 天拆除缝线,伤口一期愈合,仅病例 1 耳廓外形略增厚,其他 2 例耳廓外形正常,随访 1 年无耳廓畸形发生。

3 讨 论

创面 VAC 是一项促进创面愈合的新技术,国内外相关研究均显示 VAC 除了发挥持续有效的排除积液、积血的作用,而且能扩大毛细血管口径,增大微循环血流量,消除组织水肿,进而增加新生毛细血管密度,促进肉芽组织生长,加速创面修复[2-3]。

对耳廓化脓性软骨膜炎而言,因耳廓外形不规则,以往术后多以棉球或小块纱布打散后按耳廓外形予以加压,力求压力均匀,但实际操作中常因外力不均匀易形成小空腔,成为感染的潜在危险因素,且患者多有疼痛不适感。在临床实践中体会应用 VAC 有以下优点:①引流充分,创面密闭:负压引流不仅将渗液及时引流出,使术腔壁互相持续贴紧,压力均匀,加快创面粘连愈合时间,而且术腔成为密闭系统,有效避免交叉感染的发生;②取材简便,易于操作:引流管用一次性头皮针的软管,其管身柔韧,既利于修剪造孔,又不容易发生塌陷变形。加之管径小,便于创面紧密贴合。操作时剪去针柄后,将管身折叠,用无菌剪修剪已形成尖角的管壁,以形成合适大小的侧孔,一般在末端造侧孔 3~5 个,利于充分引流;③换药简化,便于观察:传统开放换药治疗时间长,每日更换敷料过程繁琐,患者痛苦。而实施负压引流每日换药只需以 75%酒精涂抹创面即可,注意观察耳廓皮肤颜色、引流管位置及是否通畅、引流液的量。整个过程中患者无明显不适感。拔管时间可根据引流量作相应调整,量多时可酌情延长引流管留置时间。

封闭式负压引流治疗耳廓化脓性软骨膜炎操作简单,疗程明显缩短,患者痛苦小,疗效好,值得临床推广应用。

[参考文献]

[1] 田勇泉,韩德民,孙爱华.耳鼻咽喉头颈外科学[M]. 7 版. 北京:人民卫生出版社,2008:315-316

[2] Lable L,Rancan M,Mica L,et al. Vacuum-assisted closure therapy increases local interleukin-8 and vascular endothelial growth factor levels in traumatic wounds [J]. J Trauma,2009,66(3):749-757

[3] 杨 帆,胡 耑,白祥军,等. 负压封闭引流对兔创面氧分压及愈合的影响[J]. 中华急诊医学杂志,2011,20(9):940-944

[收稿日期] 2013-07-10

(上接第 995 页)

[10] Benjamini Y,Liu W. A step-down multiple testing procedure that controls the false discovery rate under independence[J]. J Sta Plan Infer,1999,82(1-2):163-170

[11] Benjamini Y,Yekutieli D. The control of the false discovery rate in multiple testing under dependency [J]. Ann Satitst,2001,29(4):1165-1188

[12] Benjamini Y,Krieger AM,Yekutieli D.Adaptive linear step-up procedures that control the false discovery rate [J]. Biometrika,2006,93(3):491-507

[13] Alon U,Barkai N,Notterman DA,et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays [J]. Biology,1999,96(12):6745-6750

[14] 刘成友,丁 勇. 相对误差直线回归模型两种参数估计方法的比较[J]. 中国卫生统计,2012,29(5):1-3

[15] 张敬华,李金铭,朱 坤. MATLAB 在微阵列数据分析的 FDR 控制中的实现[J]. 福建电脑,2011,8(1):91-92

[16] Debnath R,Kurita T. An evolutionary approach for gene selection and classification of microarray data based on SVM error-bound theories[J]. Biosystems,2010,100(1):39-46

[收稿日期] 2013-10-17