Local Pooled Error test for microarray data analysis

Nitin Jain, Michael O'Connell, Jae K. Lee

October 25 2003

Contents

1 Introduction

This document describes local pooled error (LPE) test for identifying significant differentially expressed genes in microarray experiments. Local pooled error test is especially useful when the number of replicates is low (2-3). LPE estimation is based on pooling errors within genes and between replicate arrays for genes in which expression values are similar. This is motivated by the observation that errors between duplicates vary as a function of the average gene expression intensity and by the fact that many gene expression studies are implemented with a limited number of replicated arrays (Chen et al., 1997; Lee, 2002).

Step by step analysis is presented in Section 2 using data from a 6-chip oligonucleotide microarray study of a mouse immune response study.

Details of methodology and application of Local Pooled Error (LPE) test can be found at:

Jain et. al. (2003) *L*ocal pooled error test for identifying differentially expressed genes with a small number of replicated microarrays, Bioinformatics, 19, 1945-1951.

2 Mouse Immune Response Study dataset

Load the library

```
> library(LPE)
```

```
> data(Ley)
```

```
> dim(Ley)
```

```
[1] 12488 7
```

```
> Ley[1:3,]
              ID   c1   c2   c3     t1     t2     t3
1  AFFX-MurIL2_at 16.0 14.1 19.3 2782.7 2861.3 2540.2
2 AFFX-MurIL10_at 22.7  6.9 28.2   18.6   12.7    7.5
3  AFFX-MurIL4_at 33.9 17.1 23.9   24.9   25.2   24.9
```

```
> Ley[,2:7] <- preprocess(Ley[,2:7], data.type = "MAS5")
```

```
> Ley[1:3,]
              ID       c1       c2       c3        t1        t2        t3
1  AFFX-MurIL2_at 4.058556 3.817623 4.282605 11.474255 11.536254 11.340841
2 AFFX-MurIL10_at 4.563176 2.786596 4.829699  4.249216  3.720556  2.937006
3  AFFX-MurIL4_at 5.141769 4.095924 4.591015  4.670059  4.709151  4.668189
```

Mouse immune response study was conducted by Dr. Klaus Ley, Univeristy of Virginia. Three replicates of Affymetrix oligonucleotide chips per condition were used. Based on M vs A sctater plot matrix, IQR normalization was performed, so that interquartile ranges on all chips are set to their widest range. It is performed by multiplying by a scaling factor. Note that this is a simple constant-scale & location normalization step. Finally log based 2 transformation was done. Replcates of Naive condition are named as c1, c2, c3 and those of Actiavted condition are named as t1, t2 and t3 respectively.

Remove the control spots

```
> Ley <- Ley[substring(Ley$ID,1,4) !="AFFX",]

> dim(Ley)

[1] 12422 7

> Ley[1:3,]

           ID        c1        c2        c3       t1        t2        t3
67   92539_at 11.999273 12.151683 12.292905 12.08051 12.180762 11.936893
68 92540_f_at  8.948516  9.003377  8.642889 11.38866 11.429816 11.370188
69   92541_at  6.242440  6.078951  6.101659  5.18579  5.313072  5.937006
```

Calculate the baseline error distribution of Naive contdition, which returns a dataframe of A vs M for selected number of bins (= 1/q), where q = quantile.

```
> var.Naive <- baseOlig.error(Ley[,2:4],q=0.01)

> dim(var.Naive)

[1] 100  2

> var.Naive[1:3,]

             A    var.M
[1,] 0.8360439 1.107993
[2,] 1.4865603 1.069400
[3,] 1.8709628 1.035059
```

Similarly calculate the base-line distribution of Activated condition:

```
> var.Activated <- baseOlig.error(Ley[,5:7], q=0.01)

> dim(var.Activated)

[1] 100   2

> var.Activated[1:3,]

             A     var.M
[1,] 0.2528533 0.9453008
[2,] 0.8687306 0.9474678
[3,] 1.2006186 0.9876654
```

Calculate the lpe variance estimates as described above. The function *lpe* takes the first two arguments as the replicated data, next two arguments as the baseline distribution of the replicates calculated from the *baseOlig.error* function, Gene IDs as probe.set.name. Adjustment for multiple comparison is applied using Bioconductor's multtest package (Dudoit et. al.)

```
> lpe.val <- data.frame(lpe(Ley[,5:7], Ley[,2:4], var.Activated, var.Naive,
      probe.set.name=Ley$ID))

> lpe.val <- round(lpe.val, digits=2)

> dim (lpe.val)

[1] 12422 13

> lpe.val[1:3,]

             x.t1  x.t2  x.t3 median.1 std.dev.1  y.c1  y.c2  y.c3 median.2
92539_at    12.08 12.18 11.94    12.08      0.12 12.00 12.15 12.29    12.15
92540_f_at  11.39 11.43 11.37    11.39      0.14  8.95  9.00  8.64     8.95
92541_at     5.19  5.31  5.94     5.31      0.56  6.24  6.08  6.10     6.10
```

```
           std.dev.2 median.diff pooled.std.dev z.stats
92539_at        0.22       -0.07           0.18   -0.40
92540_f_at      0.23        2.44           0.20   12.50
92541_at        0.51       -0.79           0.55   -1.44
```

Doing FDR correction

```
> fdr.BH <- fdr.adjust(lpe.val, adjp="BH")

> dim(fdr.BH)

[1] 12422    16

> fdr.BH[1, ]

         x.x.t1 x.x.t2 x.x.t3 median.1 std.dev.1 y.y.c1 y.y.c2 y.y.c3 median.2
92539_at  12.08  12.18  11.94    12.08      0.12     12  12.15  12.29    12.15
         std.dev.2 median.diff pooled.std.dev abs.z.stats p.adj.adjp.rawp
92539_at      0.22       -0.07           0.18         0.4       0.6973583
         p.adj.adjp.BH p.adj.index
92539_at      0.812549           2
```

Resampling based FDR adjustment takes a while to run, and returns the critical z-values and corresponding FDR.

```
> fdr.2 <- fdr.adjust(lpe.val, adjp="resamp", iterations=2)

iteration number 1 is in progress
iteration number 1 finished
iteration number 2 is in progress
iteration number 2 finished
Computing FDR...
```

```
> fdr.2

      target.fdr z.critical
 [1,]      0.001  4.2589217
 [2,]      0.010  2.9612657
 [3,]      0.020  2.5032199
 [4,]      0.030  2.2778116
 [5,]      0.040  2.0959562
 [6,]      0.050  1.9955792
 [7,]      0.060  1.8833591
 [8,]      0.070  1.7896138
 [9,]      0.080  1.7184356
[10,]      0.090  1.6488528
[11,]      0.100  1.5894605
[12,]      0.150  1.3653030
[13,]      0.200  1.2058491
[14,]      0.500  0.6876795
```

Note that above table may differ slightly due to generation of 'NULL distribution' by resampling. For each target.fdr, we can note critical z-value, above which all genes are considered significant.

3 Discussion

Using our LPE approach, the sensitivity of detecting subtle expression changes can be dramatically increased and differential gene expression patterns can be identified with both small false-positive and small false-negative error rates. This is because, in contrast to the individual gene's error variance, the local pooled error variance can be estimated very accurately.

Acknowledgments. We wish to acknowledge the following colleagues: P. Aboyoun, J. Betcher, D Clarkson, J. Gibson, A. Hoering, S. Kaluzny, L. Kannapel, D. Kinsey, P. McKinnis, D. Stanford, S. Vega and H. Yan.

Availability of LPE library: UVa School of Medicine

References

References

[1] Chen Y., Dougherty E.R., Bittner M.L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics*, 2:364-374.

[2] Cleveland W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829-836.

[3] Dudoit, S., Yang, Y.H., Callow, M.J., Speed, T.P. (2000),. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Technical Report #578, Department of Statistics, University of California at Berkeley: Berkeley, CA.

[4] Durbin, B., Hardin, J., Hawkins, D.M., and Rocke, D.M. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* 18:105110.

[5] Jain et. al. (2003) Local pooled error test for identifying differentially expressed genes with a small number of replicated microarrays. Bioinformatics, 19, 1945-1951

[6] Lee J.K. (2001). Analysis Issues for Gene Expression Array Data. *Clinical Chemistry* 47:1350-1352.

[7] Lee J.K. and Connell, M.O.(2003). An S-Plus library for the analysis of differential expression. In The Analysis of Gene Expression Data: Methods and Software. Edited by G. Parmigiani, ES Garrett, RA Irizarry ad SL Zegar. Springer, NewYourk.

[8] Mood A.M., Graybill F.A., Boes D.C. (1974). *Introduction to the theory of statistics*, 3rd ed. McGraw-Hill, Inc.: New York.