

Statistical significance analysis of longitudinal gene expression data

Wei Pan

Division of Biostatistics

School of Public Health

University of Minnesota

Email: `weip@biostat.umn.edu`

Http: //www.biostat.umn.edu/~weip

June 20, 2003

Colorado State University

Joint work with Xu Guo, Huilin Qi and Catherine M. Verfaillie.

Outline

- Introduction
- Data
- Method
 1. Test statistic
 2. Use of SAM and MMM
- Simulation
- Discussion

1. Introduction

- Microarray: measure expression levels of thousands of genes
global view
many many applications
data analysis?
- Q: differential gene expression
–most only for independent data; see Pan (2002)
- Use of microarray in time-course studies:
longitudinal data–within-subject correlation
noticed by Perou et al, (2000)

temporal biological processes

- Statistical model-based clustering analysis

(Yeung et al, 2001; Ghosh and Chinnaiyan, 2002; Li et al, 2002 and references therein).

- Not all time-course data induce within-subject correlation

E.g. Lue-Ping Zhao's works

- Goal here: significance analyses to detect differential gene expression with longitudinal array data
- Challenges: features of data
a huge number of genes,

a small number of arrays,
potential within-subject correlation

- Three steps:
 - (1) construct a summary test statistic;
 - (2) specify or estimate its null distribution;
 - (3) determine the significance level
- Approach:
 - (1) test stat: extension of t-stat
 - (2)&(3): SAM (Tusher et al, 2001) or MMM (Pan et al, 2001)
- SAM and MMM: attractive when the sample size is small.

nonparametric

not rank-based, as M-W-W test

using large # of genes

- SAM and MMM: differ in choosing a cut-off point

2. Data

- A rare bone marrow cell was identified: mesodermal progenitor cell (MPC) (Reyes et al, 2001).
- MPC can differentiate at single-cell level into mesenchymal cell types such as osteoblasts, chondroblasts and adipocytes, and also into cells of visceral mesodermal

origin.

- MPC can be an ideal source of cells to generate osteoblasts to treat bone diseases such as osteoporosis or non-healing fractures, and osteogenesis imperfecta (Horwitz et al, 1999).
- Understand the differentiation process of MPC into osteoblasts
gene regulations of specific signaling proteins and transcription factors (Yamaguchi et al, 2000; Ducy et al, 2000)
- Studied gene expression from undifferentiated MPC (at day 0) to osteoblast lineage-

specific differentiation at day 1, day 2 and day 7 by cDNA (Qi et al, PNAS, 2003)

- A key feature: samples taken from the same subject were used to measure gene expression across the seven days.

$J = 3$ subjects

$I = 4132$ genes

- Thus, a longitudinal data set with four different time points was generated.
- Q: identify genes differentially expressed over time

3. Method

- Test stat: Wald-stat based on the robust

or Sandwich or empirical cov estimator;
GEE theory (Liang and Zeger, 1986)

- Example

$$H_0 : \mu_0(i) = \mu_1(i) = \mu_2(i)$$

- $y_{j,1}(i) = y'_{j,1}(i) - y'_{j,0}(i)$

$$y_{j,2}(i) = y'_{j,2}(i) - y'_{j,0}(i)$$

- Hence, $H_0 : \mu_1(i) = \mu_2(i) = 0$

- Note: this $H_0 \neq H'_0: \mu_1(i) = \mu_2(i)$ that
can be tested using the paired t-test or
its analog in SAM

- $\hat{\mu}(i) = (\bar{y}_1(i), \bar{y}_2(i))'$

and its variance is $\hat{V}_W(i) =$

$$\frac{1}{J^2} \sum_{j=1}^J [y_j(i) - \hat{\mu}(i)][y_j(i) - \hat{\mu}(i)]'$$

where $y_j(i) = (y_{j,1}(i), y_{j,2}(i))'$

- With independent data, take $\hat{V}_{W,ind}(i)$ as $\hat{V}_W(i)$ but force non-diagonal elements=0

- Generalized Wald stat is

$$W(i) = (\bar{y}_1(i), \bar{y}_2(i)) \hat{V}_W^{-1}(i) (\bar{y}_1(i), \bar{y}_2(i))'$$

- Gene-specific score (our test-stat) is $w(i) =$

$$(\bar{y}_1(i), \bar{y}_2(i)) [\hat{V}_W(i) + \lambda_w I_{2 \times 2}]^{-1} (\bar{y}_1(i), \bar{y}_2(i))^T,$$

- λ_w selected as in SAM to minimize CV;

stabilize $\hat{V}_W(i)$

$\lambda_w = 0.045$ for our data

- Generate null scores:
 1. Under $H_0 : \mu_1(i) = \mu_2(i) = 0$, we permute $y_1(i)$ and $y_2(i)$ by multiplying a $+1$ or -1 randomly.
 2. With B permutations, we have B longitudinal data sets and obtain B sets of null scores
- SAM: Compare $w(i)$ with

$$w_E(i) = \frac{B}{\sum_{b=1}^B} w_0(i)^{(b)} / B$$

- For most genes, $w(i) \simeq w_E(i)$;
- Deviation of $w(i)$ from $w_E(i)$ gives evidence to reject H_0 for gene i

- How to select a cutoff point?

Use of a threshold Δ to control FDR

$$\text{FDR} = \text{FP} / \text{TP}$$

- SAM result: Table 1
- MMM: similar to SAM, but use w_0 's or $\log(w_0)$'s to estimate the null distr f_0

$$f_0(\log(w_0); \Phi_{g_0}) = \sum_{g=1}^{g_0} \pi_g \Phi(\log(w_0); \mu_g, V_g),$$

where each $\Phi()$ is a normal density.

–use EM to fit for a given g_0 ; EMMIX

(McLachlan et al, 1999)

–use BIC/AIC to select g_0

- For our data, $g_0 = 3$

- For any given FP rate α , solve

$$\alpha = P\{\log(w) > s | f_0\}$$

to obtain cutoff value s

Significant if $w(i) > \exp(s)$.

- LRT more efficient

For simplicity, define RR as two tails of

f_0 (as t-test)

- MMM results: Table 2

- Significant genes

–SAM and MMM give the same top 12 genes

–HSPCB is identified twice

–MYC, UBCH10 and PSMD2 involved in

cell proliferation are down-regulated after the induction of differentiation—reasonable—ODC1 is involved in polyamine biosynthesis—reasonable—The remaining genes: ?

4. Simulation

- Generating simulated data: to mimic the real data and H_0 is always true for each gene
- Consequence of ignoring correlation Rather than using V_W , we use $V_{W,ind}$, to construct test and null stat's

Table 3

5. Discussion

- Accounting for within-subject correlation is necessary
- Zhao et al (2002, Physiol Genomics)
 - same conclusion
 - based on Normal distr
- Our method is general
 - can extend to EB of Efron et al (2001)

Table 1: Identifying temporal gene expression change by SAM for osteoblast differentiation data

Threshold	FP	TP	FDR
$\Delta = 12$	44.5	117	39%
$\Delta = 18$	21.4	67	32%
$\Delta = 35$	2.8	12	23%

Table 2: Identifying temporal gene expression change by MMM for osteoblast differentiation data

Type I error	FP	TP	FDR
$\alpha = 0.001$	4.2	12	35%
$\alpha = 0.005$	22.5	67	34%
$\alpha = 0.01$	43.8	115	38%

Table 3: FDR estimated by SAM and MMM for simulated data with the use of $w(i)$ or $w_{ind}(i)$ (FDR is taken as 100% when FP is greater than TP)

<u>SAM</u>							
$w(i)$				$w_{ind}(i)$			
Threshold	FP	TP	FDR	Threshold	FP	TP	FDR
$\Delta = 2$	9.0	10	90%	$\Delta = 3.0$	3.6	8	45%
$\Delta = 8$	3.8	4	95%	$\Delta = 4.5$	1.6	4	40%
$\Delta = 30$	0.6	1	60%	$\Delta = 6.0$	0.7	2	35%

<u>MMM</u>							
$w(i)$				$w_{ind}(i)$			
Type I error	FP	TP	FDR	Type I error	FP	TP	FDR
$\alpha = 0.001$	3.5	4	88%	$\alpha = 0.001$	4.1	9	46%
$\alpha = 0.005$	20.7	20	100%	$\alpha = 0.005$	18.9	30	63%
$\alpha = 0.01$	44.0	39	100%	$\alpha = 0.01$	42.0	57	74%