The Sliding Window Gene-Shaving Algorithm for Microarray Data Analysis

이혜선 (포항공과대학교 산업공학과 책임연구원)
(054-279-8222, 016-807-3122, hyelee@postech.ac.kr)

Abstract :

Gene-shaving(Hastie et al, 2000) is a very useful method to identify a meaningful group of genes when the variation of expression is large. By shaving off the low-correlated genes with the leading principal component, the primary genes with the coherent expression pattern can be identified. Gene-shaving method works well if expression levels are varied enough, but it may not catch the meaningful cluster in low expression level or different expression time even with coherent patterns. The sliding window gene-shaving method which is to apply gene-shaving in each sliding window after hierarchical clustering is to compensate losing a meaningful set of genes whose variation is not large but distinct. The performance to identify expression patterns is compared for the simulated profile data by the different variance and expression level.

약력

1985 서울대학교 소비자아동학과 (학사)
1987 서울대학교 소비자경제학 (석사)
1993 Cornell University 통계학 (석사)
1996 ~ 현재 포항공과대학교 산업공학과 책임연구원
1997 ~ 1998 경일대학교 산업공학과 겸임교수 (통계학)
1994 ~ 1996 미국 국립조사 연구소, 통계프로그래머
1994 ~ 1995 시카고대학교 의과대학 Statistician 겸직(Public Health funded by NIH)
1993 ~ 1994 시카고대학교 경제연구소 연구원

The Sliding Window Gene-Shaving Algorithm for Microarray Data Analysis

이 혜 선

포항공과대학교 산업공학과 책임연구원

e-mail: stat@postech.ac.kr

최대우 (한국외국어대학교 통계학과 부교수)

전치혁(포항공과대학교 산업공학과 교수)

1. Motivation for Analyzing Microarray data

(1) To find molecular pathway how some genes function to a certain disease (cancer)

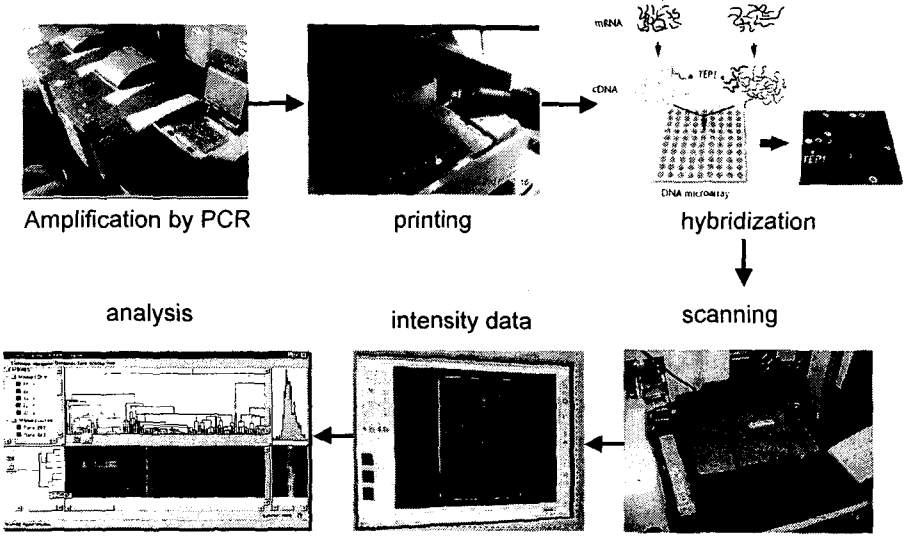
(2) to find new drug discovery by a clinical material suppressing a certain disease

⇒ What can statistician do for that motivation ?

Ans. : Identify the coherent expression pattern from DNA chip microarray data

=> provide valuable results to biologist and let them interpret and develop clinical outcome.

2. Microarray data : Process



3. Gene Expression Data

Slide (experiment)

Genes	slide1	slide2	slide3	slide4
1	0.26	0.30	0.80	0.90
2	-0.10	0.46	0.24	0.06
3	0.15	0.74	0.04	0.10
4	-0.45	-1.03	-0.79	-0.56
5	0.06	1.06	1.35	-1.09

Gene expression level of gene *4* on slide *3*

$=\text{Log}_2$(Red intensity / Green intensity)

Gene Shaving

By Hastie, Tibshirani,... (2000)

Gene shaving

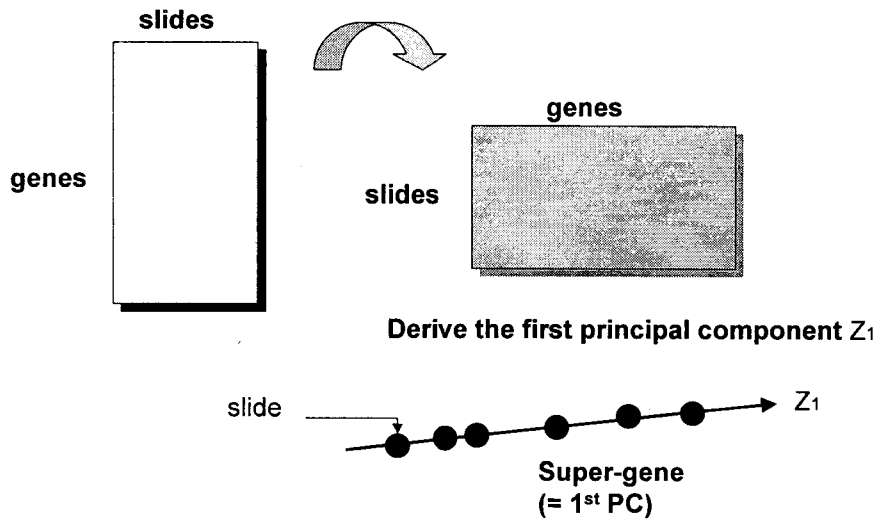
The Basic Idea is

shaving off the low-correlated genes

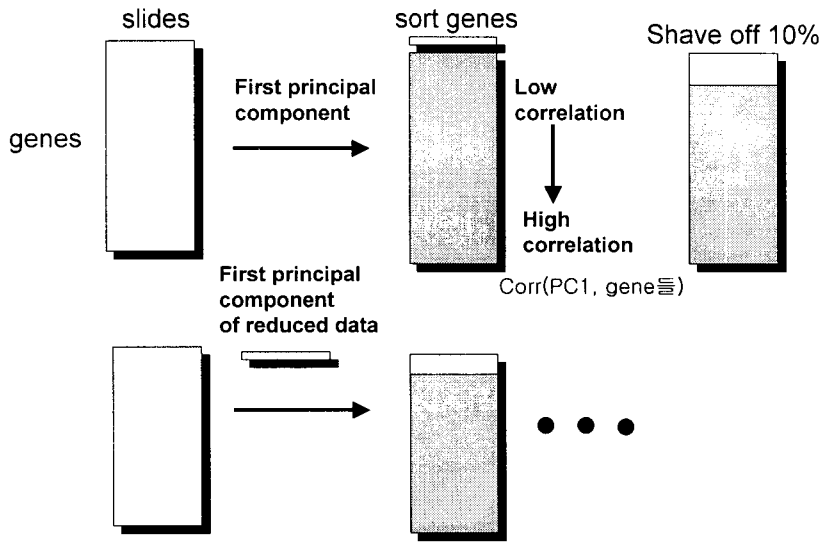
shaving off some noise

=> Better find the meaningful group of genes.

4. Gene Shaving Method



Gene Shaving - schematic process



Simulated Data #1-formula

- Generated data with N=1000, p=60, z_{ij} ~ N(0,1)

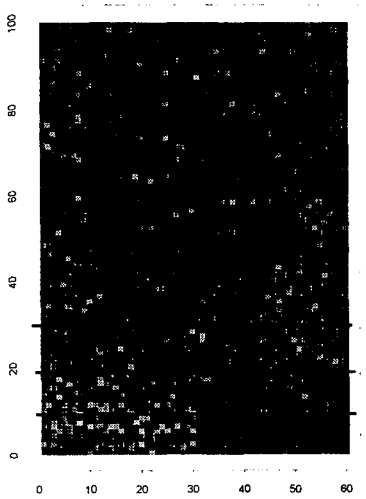
$$x_{ij} = z_{ij} + b_{ij}$$

where for $1 \le i \le 10$ $\quad b_{ij} = \begin{cases} -1 & \text{if } j \le 30 \\ +1 & \text{if } j > 30 \end{cases}$

where for $11 \le i \le 20$ $\quad b_{ij} = \begin{cases} -0.5 & \text{if } j \le 30 \\ +0.5 & \text{if } j > 30 \end{cases}$

where for $21 \le i \le 30$ $\quad b_{ij} = \begin{cases} -0.2 & \text{if } j \le 30 \\ +0.2 & \text{if } j > 30 \end{cases}$

Simulated Data #1-image



- image for the first 100 genes among 1,000 genes

$21 \le i \le 30$
$11 \le i \le 20$
$1 \le i \le 10$

Clusters from gene shaving- Data #1

- **For simulated data #1**
  - **Cluster-1: 3, 6, 8, 7, 9, 2, 1, 4, 10**
  - **Cluster-2: 817, 464**
- **Only find one Cluster which has large variation (shifted –1, 1)**
- **Gene Shaving does not catch clusters with small variation**

* Gene-Shaving : Problem & Suggestion

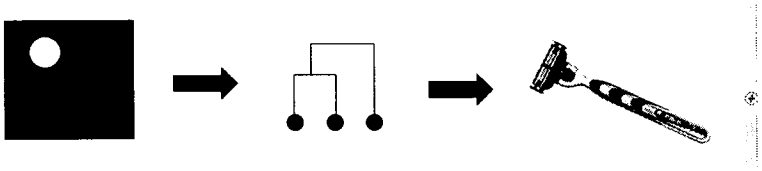
- **In different expression time :**

- There are genes expressed rapidly in the beginning and the ones expressed in slow pace.

- In this case, gene-shaving may not find the cluster with genes responded in the beginning.

⇒ **try to find the coherent patterns in a rough cluster**

☒ Gene-Shaving after Hierachical clustering !!



Array data Hierarchical clustering Gene shaving

5. Gene shaving-after Hierachical Clustering(HC)

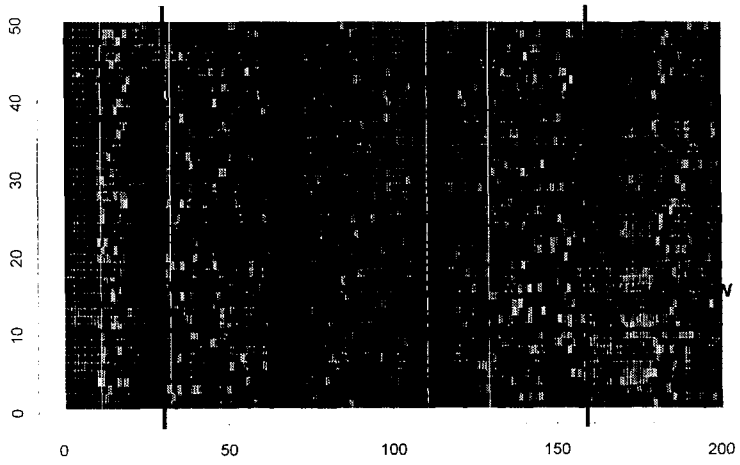
Simulated Data - #2

- **The following 6 patterns are generated:**
  - **Pattern 1:** $\sin(x_i)+\varepsilon_i$
  - **Pattern 2:** $\cos(x_i)+\varepsilon_i$
  - **Pattern 3:** $0.03x_i-2+\varepsilon_i$
  - **Pattern 4:** $-0.08x_i+2+\varepsilon_i$
  - **Pattern 5:** $\sin(x_i)+0.04x_i-3+\varepsilon_i$
  - **Pattern 6:** $\cos(x_i)-0.05x_i+2+\varepsilon_i$

  with $x_i = i,\ (i = 1,\square\ ,50),\quad \varepsilon_i \sim U[-1/2,1/2]$

- **Each cluster with specific pattern has 10 observations (genes).**
- **Rest of 140 genes are from** $N(0,1)$.

Hierarchical clustering - Image

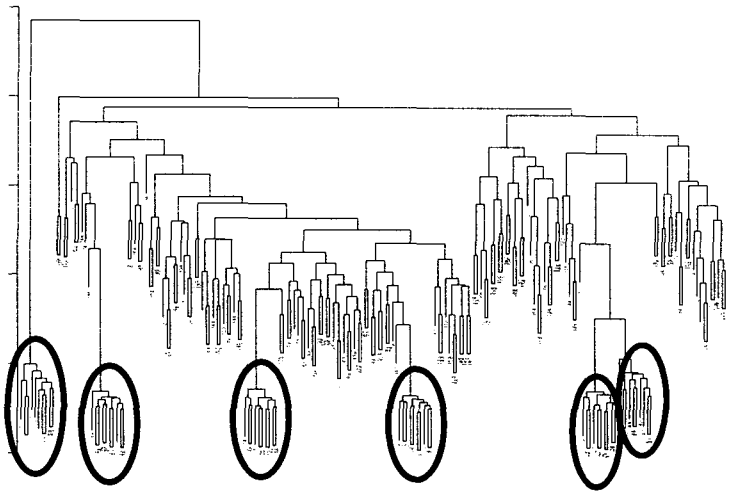


5. Gene shaving-after HC

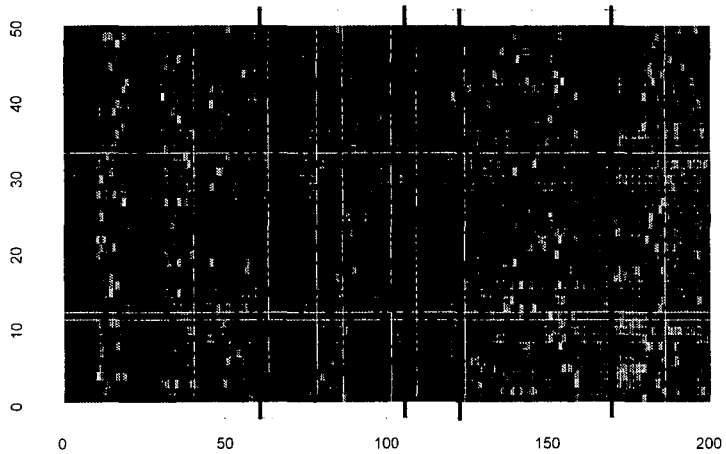
- **Apply Gene-Shaving in each of 3 window from HC.**
  - **1st window : first 30 genes**
  - **2nd window : 129 genes**
  - **3rd window : 41 genes**

- **From 1st window**
  - **Cluster 1 : "34" "40" "31" "36" "39" "33" "35" "32" "37" "38"**
  - **Cluster 2 : "52" "58" "59" "54" "60" "56" "51" "53" "55" "57"**
- **From 2nd window**
  - **Cluster 3: "160" "8" "7" "3" "9" "2" "10" "180" "1" "6" "158" "4"**
- **From 3rd window**
  - **Cluster 4 : "43" "50" "49" "44" "48" "41" "47" "42" "45"**

5. Hierarchical clustering



Another clusters from H.C. (additional 2 partitions)



Finding cluster from additional 2 windows

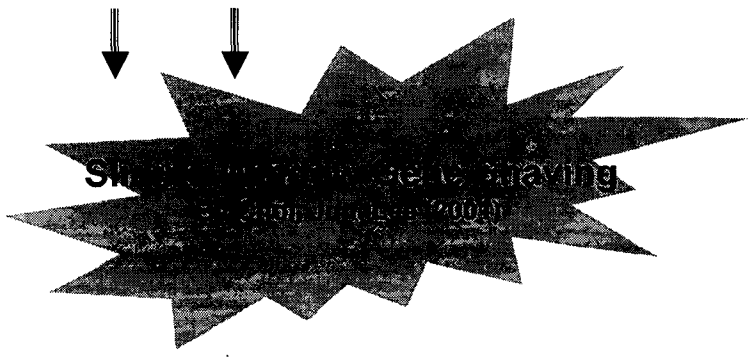
- **From 5th window**
  - **Nothing!**
- **From 6th window**
  - **Cluster5 - "12" "16" "14" "18" "13" "19" "15" "20"**
- Not to find the cluster with "linearly increased" pattern

(reason : there are very small variations in that rough cluster)

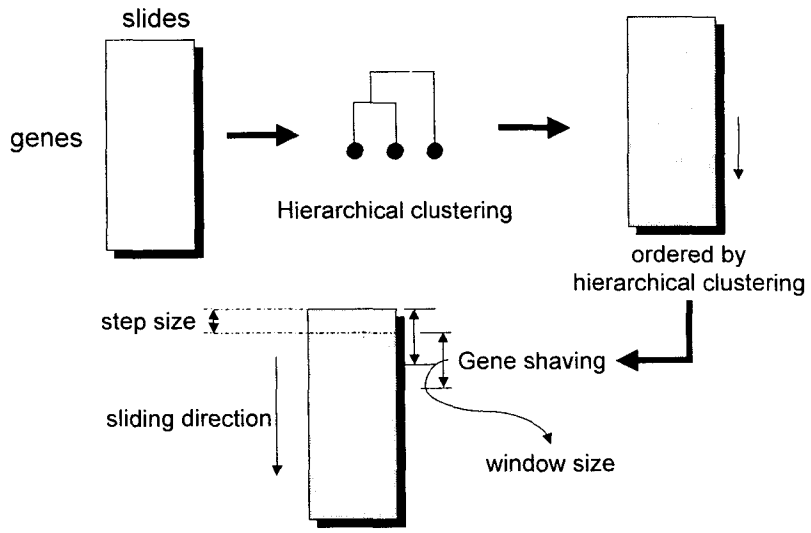
=> mutually exclusive partitioning in H.C may lose a meaningful cluster !

* Question for Gene Shaving-after-H.C.

- With exclusive partitioning, all meaningful clusters can not be found !
- Then how to partition from Hierarchical clustering results?
- => looking carefully Image plot ? (not quite good answer)



6. Sliding Window Gene Shaving



6. Sliding Window Gene Shaving -Algorithm

- STEP 0. Input the following parameters:
  - Step size: the size for which window is shifting
  - Window size: the number of genes applying gene shaving
  - Minimum gap statistics: a criterion for discarding clusters
  - Some numbers for gene shaving (alpha, B, the no. of clusters)
- STEP 1. Order the data by the similarity derived from hierarchical clustering.
- STEP 2. Do gene shaving to the genes as much as window size.
- STEP 3. Discard the cluster from gene shaving if gap statistics for the cluster is below than minimum gap statistics.
- STEP 4. Select more significant cluster (in some sense); If two clusters contain the same gene, select one cluster with the higher gap statistics.

6. Sliding Window Gene Shaving -Algorithm

- STEP 5. Repeat STEP 2~STEP 4 sliding a window as much as step size.
- STEP 6. Combine the results from gene shaving applying to each window;
  - Link the clusters if there is any intersection.

6. Sliding Window Gene Shaving - Result

- **Parameter settings**
  - **Window size=75, Step size=10, Min. gap stat.=45**
- **Identified clusters**
  - **Pattern 4: "31", "32", "33", "34", "35", "36", "37", "38", "39", "40"**
  - **Pattern 6: "51", "52", "54", "56", "58", "59", "60"**
  - **Pattern 2: "11", "12", "13", "14", "15", "16", "17", "18", "19", "20"**
  - **Pattern 1: "1", "2", "3", "4", "6", "7", "8", "9", "10", "180"**
  - **Pattern 5: "41", "42", "43", "44", "45", "46", "47", "48", "49", "50"**
- **Pattern 3 is not found, since the variation for pattern 3 is so small.**

- **Comparison - Results of Gene-Shaving only**
  - **Cluster1: "34" "40" "31" "36" "33" "39" "35" "32" "37" "38"**
  - **Cluster2: "48" "43" "49"**
  - **Cluster3: "52" "12" "59" "58" "60" "56" "54"**

Discussion

- Sliding Window Gene Shaving provides the better result than original gene shaving.
- We expect that Sliding Window Gene Shaving work well for exploring the clusters with different expression time.
- But, if a cluster "A" has small variation and there is another cluster with bigger variation and almost the same pattern, "A" cannot be found easily.
  - For example, due to Pattern-5 and noise, the algorithm cannot catch Pattern-3.
- To overcome the latter point, we may adjust some parameters such as window size and step size.