응용통계연구 제 **20**권 2호, *2007*년, *pp. 409–422*

# 무 변화 패턴을 갖는 시간경로 유전자발현자료를 제거하기 위한 함수들의 비교*

김경숙[1] 오미라[2] 백장선[3] 손영숙[4]

요 약

시간경로 유전자 발현자료에 대한 본격적인 통계분석을 수행하기에 앞서 의미있는 정보를 제공하지 못할 것으로 여겨지는 유전자들은 선별하여 미리 제거함으로서 자료의 차원을 축소시킬 수 있을 뿐 아니라, 잡음이나 변이가 낮은 자료로 인한 잘못된 판단을 감소시킬 수 있다. 본 논문에서는 관측표본에 대한 백분위수 기준과 붓스트랩 표본에 대한 백분위수 기준 하에서 무 변화 패턴을 갖는 유전자들을 제거시킬 수 있는 기존의 필터링 함수들을 비교하였다. 이스트(yeast) 자료에 적용하여 두 가지 필터링 방식에 대해 가장 유사한 결과를 보인 것은 분산 함수였다.

주요용어: 전처리, 무 변화 패턴 필터링, 시간경로 유전자 발현, 백분위수, 붓스트랩.

## 1. 서론

마이크로어레이 기술의 발전으로 DNA 마이크로어레이 칩을 이용하여 한꺼번에 수천 수만 개 유전자들의 발현 정보를 관측할 수 있게 됨에 따라 생명의약학 분야의 많은 연구가 가능하게 되었고, 이에 따라 마이크로어레이 자료로부터 유용한 유전정보를 밝히기 위한 분석 과정에 있어서도 막대한 계산 양을 필요로 하게 되었다.

일반적으로 마이크로어레이자료의 분석에서는 핵심이 되는 자료 분석을 수행하기 이전에 전처리(preprocessing) 단계로서 정규화 및 이상치 검토 등을 통해 분석결과에 좋지 않은 영향을 미칠 것으로 간주되는 자료라든지, 관측된 유전자의 발현프로파일 내에 결측값 비율이 높은 자료라든지, 또는 유전자 발현프로파일의 변화수준이 너무 미미하여 유전 정보를 밝혀내는데 거의 도움을 주지 못할 것으로 판단되는 자료들은 찾아내어 제거함, 즉,

* 이 논문은 2005년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행되었음 (KRF-2005-204-C00017).

1) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 통계학과, 박사과정 수료
E-mail: ksook620@jnu.ac.kr
2) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 통계학과, 박사과정 수료
E-mail: omr@chonnam.ac.kr
3) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 통계학과, 교수
E-mail: jbaek@chonnam.ac.kr
4) (교신저자)(500-757) 광주광역시 북구 용봉동 300, 전남대학교 통계학과, 교수
E-mail: ysson@chonnam.ac.kr

필터링(filtering) 함으로서 본 분석과정에서 필요한 계산 양을 줄이고 자료분석 결과의 질도 향상시키고자 한다(Chen 등, 1999; Herrero 등, 2003; Kadota 등, 2003).

따라서 전처리 과정은 본 자료 분석 수행 전에 반드시 수행되어져야 할 중요한 단계로 볼 수 있다. Herrero 등(2003)이 개발한 웹기반 툴인 Perl CGI script는 전처리 함수로서 로그변환, 반복수가 다른 자료 제거, 결측치 처리, 무 변화 패턴 필터링(flat pattern filtering) 및 정규화 등을 처리할 수 있는 함수들을 내장하고 있다. 여기서 제공되는 무변화 패턴 필터링 방법은 피크(peak)의 개수, $RMS$(root mean square) 및 표준편차를 이용하는 방법 등이다.

여러 가지 전처리 유형 가운데 본 논문에서는 무 변화 패턴을 갖는 유전자 자료를 제거하는 문제를 논의하려고 한다. 시간경로(time course) 유전자 발현자료는 각 유전자의 발현수준이 시간의 흐름에 따라 관측되므로 시간은 유전자 발현수준에 영향을 주는 중요한 인자이다. 관측된 시간경로 유전자 자료에서 시간의 흐름에 따라 매우 작은 변이를 보이는 유전자들, 즉, 무 변화 패턴을 갖는 유전자들은 유전자 특성을 파악하는데 도움이 되는 정보를 주지 못하며 다른 유전자들과의 관련성도 미미하기 때문에, 추후의 본 자료 분석에 사용되기에 무의미하며 계산 양만 늘릴 뿐 오히려 분석결과에 잡음을 발생시킬 수 있으므로 분석에서 제외시켜야 한다고 하였다(Chen 등, 1999; Lindlöf와 Olsson, 2003).

자료를 필터링 함에 있어서 필터링을 결정하는 기준이 필요하게 되는데 자료의 특성을 고려하여 절대적 기준 또는 상대적 기준을 적용할 수 있다. 자료의 특성에 대한 사전 지식이 있다면 절대적 기준이 상당히 유효하나 그렇지 못한 경우에는 기준 값 결정에 어려움이 따르게 된다. 초기 마이크로어레이 자료 분석에서는 유의미한 유전자를 선택하기 위한 방법으로서 상대적 유전자 발현 비(ratio)가 2.0을 초과하면 선택하는 상대적 2-fold 발현기준을 사용하였으나, 이러한 임계값 선정방법은 변이정도가 서로 다른 자료에 일괄적으로 적용하기에 적합하지 않았다. 이에 반해 관측된 자료 내의 백분위수(percentile: pct)를 기준으로 하는 상대적 기준을 사용하면 적어도 자료 내에서는 상대적인 객관성을 유지할 수 있다. 그러나, 이 방법은 관측표본 자료의 관심 통계량에 대한 순위 분포(ranked distribution)로부터 임의의 백분위수를 기준으로 하여 유전자 제거 여부를 결정하게 되므로 관측되는 자료 자체에 크게 영향을 받게 된다. 이러한 한계점을 해결하는 하나의 대안으로서 최근 생물정보학(bioinformatics)이나 생물통계학(biostatistics) 분야에서는 붓스트랩(bootstrap)과 같은 재표본 추출(resampling) 방법을 이용한 통계적 유의성 검정의 기준을 사용하고 있다. 이러한 붓스트랩 방법 역시 관측 자료로부터 재표집되므로 관측자료와 전혀 무관할 수는 없으나 관측자료에 대한 의존성을 어느 정도는 완화시켜주는 장점이 있다고 본다.

본 논문에서는 무 변화 패턴을 가지는 시간경로 유전자 발현자료를 필터링하는 방법들을 정리하고, 상대적 필터링 기준으로서 관측표본에 대한 백분위수 방식과 붓스트랩 표본에 대한 백분위수 방식을 비교하고자 한다. 2절에서는 지금까지 소개된 무 변화 패턴 필터링 방법을 정리하고, 3절에서는 붓스트랩을 이용한 필터링 방법을 소개한다. 4절에서는 이 스트 시간경로 유전자 발현자료에 적용하여 수치분석한 결과를 검토하며 결론을 맺는다.

## 2. 무 변화 패턴 유전자 필터링 방법

시간경로 유전자 발현자료에 대한 무 변화 패턴 필터링을 위한 방법으로서 먼저, Chen 등(1999)은 절대적 발현값 기준과 상대적 평균 발현기준 및 상대적 2-fold 발현기준을 사용하였다. Herrero 등(2003)은 표준편차 및 $RMS$ 측도를 사용하였고, de Lichtenberg 등(2005)은 각 관측 유전자의 표준편차를 붓스트랩 표본의 표준편차에 대한 경험분포에 근거하여 유의성 검정한 후 해당 유전자의 제거 여부를 결정하였다. Liang 등(2005)은 모수적 방법으로서 선형과 비선형의 Bayesian 모형 및 non-Bayesian 혼합모형을, 준모수적 방법으로서 class dispersion 측도를, 비모수적 방법으로서 Fleury 등(2002)에 의한 파레토(Pareto) 방법을 제시하였다. 또한, $MATLAB$ Bioinformatics toolbox (2003)에 탑재된 방법들과 그 외 $R$ 언어로 구현된 여러 가지 방법들이 있다.

위에서 언급된 필터링 방법들 중에서 확률모형에 근거한 복잡한 추론에 의한 방법은 논외로 하고 비교적 단순한 측도를 사용하는 필터링 방법으로서 $MATLAB$ 및 $R$ 언어로 제공되는 범용화 된 함수들, Chen 등(1999)에 의한 $RA$(relative average) 함수, 그리고 Fleury 등(2002)에 의한 파레토 함수에 대해서만 살펴보기로 한다. Herrero 등(2003)과 de Lichtenberg 등(2005)이 제안한 표준편차와 $RMS$ 측도는 $MATLAB$ 에 탑재된 genevarfilter 함수에서 사용하는 분산 측도와 기본적으로 같은 의미를 가지므로 논의에서 제외하였다.

우선, $P$개의 유전자에 대한 $T$개 관측시점에서 시간경로 유전자 발현프로파일 자료 $\mathbf{X}$가 다음과 같은 행렬형태로 주어져 있다고 하자. 여기서 $x_{it_k}$는 유전자 $i$의 시점 $t_k$에서 관측된 발현 값을 의미한다.

$$\mathbf{X_{P\times T}} = \begin{pmatrix} x_{1t_1} & x_{1t_2} & \cdots & x_{1t_T} \\ x_{2t_1} & x_{2t_2} & \cdots & x_{2t_T} \\ \vdots & \vdots & \vdots & \vdots \\ x_{Pt_1} & x_{Pt_2} & \cdots & x_{Pt_T} \end{pmatrix}.$$

### 2.1. $MATLAB$ 함수

다음의 함수들은 필터링을 위한 기준 값으로서 백분위수와 같은 상대적인 기준뿐만 아니라 임의의 특정 값을 지정하는 절대적인 기준도 적용할 수 있다.

① 'genelowvalfilter' 함수는 관측 유전자 발현 값들의 절대 값들 중 최대값이 기준 절대값 혹은 기준 백분위수 보다 작으면 해당 유전자 자료를 제거한다. 이때 기준 백분위수를 정하는 모집단은 $P \times T$ 개 유전자 발현 값들의 절대 값들이다.

② 'genevarfilter' 함수는 관측 유전자 발현 값들의 분산이 기준 분산 값 혹은 기준 백분위수 보다 작으면 해당 유전자 자료를 제거한다. 이때 기준 백분위수를 정하는 모집단은 $P$개 유전자들의 분산들이다.

③ 'geneentropyfilter' 함수는 관측 유전자 발현 값들의 엔트로피(entropy)가 기준 엔트로피 값 혹은 기준 백분위수 보다 작으면 해당 유전자 자료를 제거한다. 이 때 기준 백분위수를 정하는 모집단은 $P$개 유전자들의 엔트로피 값들이다. 유전자 $i$에 대한 엔트로피

는 관측치열 $\mathbf{x_i} = (x_{it_1}, x_{it_2}, \ldots, x_{it_T})$을 $B$개의 범주로 균등하게 분할하여 각각의 발생확률 $\mathbf{p_i} = (p_{i1}, p_{i2}, \ldots, p_{ij}, \ldots, p_{iB})$를 구한 후 다음의 계산식으로부터 산출된다. *MATLAB* (2003)에서는 $B$의 값으로서 $T/2$보다 크거나 같은 정수들 중에서 최소의 정수값을 사용한다.

$$Ent_i = -\sum_{j=1}^{B} p_{ij} \log_2 p_{ij}.$$

### 2.2. *R* 함수

아래에 소개된 함수들은 각 특정 자료에 대한 필터링을 위해 사용되었던 것으로서 디폴트로 지정된 모든 임계값들은 해당 특정자료에 적합한 값으로 설정되어 있다. 여기에 소개된 함수들을 이용함에 있어서 가장 어려운 점은 각 함수에서 필요로 되는 임계값을 어느 수준으로 정해야 할지에 대한 기준이 마련되어 있지 않다는 점이다. 따라서 관측된 자료의 특성에 적합한 임계값을 정하는 과정이 선행되어야 할 것이다.

① 'preprocess' 함수(http://rss.acs.unt.edu/Rdoc/library/plsgenomics/html/preprocess.html)는 유전자 발현 값이 기준 범위를 벗어나거나, 최소값 대비 최대값의 비(ratio) 혹은 차이(difference)가 임계값 보다 작으면 해당 유전자 자료를 제거한다. 이는 Dudoit 등(2002)에 의한 전처리 방법이 Sophie Lambert-Lacroix 등에 의해 구현된 것으로서 PLSGENOMICS 패키지에 있다.

② 'filtering' 함수(http://pbil.univ-lyon1.fr/library/som/html/filtering.html)는 'preproc-ess' 함수처럼 발현 값이 기준 범위를 벗어나거나, 최소값 대비 최대값의 비 혹은 차이가 임계값 보다 작으면 해당 유전자 자료를 제거한다. 이는 SOM 알고리즘을 적용하기 전에 자료를 필터링하기 위해 구현된 것으로서 SOM 패키지에 있다.

③ 'genefilter' 함수(http://rss.acs.unt.edu/Rdoc/library/genefilter/html/genefilter.html)는 user가 구성한 여러 필터함수들을 순차적으로 적용하고 각각의 필터링 기준에 하나라도 통과하지 못하는 경우가 발생하면 해당 유전자 자료를 제거하는 것으로서, 이는 R. Gentleman 등에 의해 구현된 것으로서 GENEFILTER 패키지에 있다.

### 2.3. 상대적 평균 함수

Chen 등(1999)이 제안한 방법들 가운데 다음과 같은 상대적 평균(relative average: $RA$)함수를 결정기준으로 사용하여 기준 $RA$값 혹은 기준 백분위 수보다 작으면 해당 유전자를 제거한다. 유전자 $i$에 대한 $RA$의 계산식은 다음과 같다.

$$RA_i = \frac{\max_k(x_{it_k}) - \overline{x}_i}{\overline{x}_i}, \quad \text{여기서 } \overline{x}_i = \frac{1}{T}\sum_{k=1}^{T} x_{it_k}.$$

### 2.4. 파레토 함수

Fleury 등(2002)이 제안한 파레토 함수 가운데 유전자의 관측시점 간의 기울기 절대 값들에 대한 평균(average of absolute slopes: $AAS$) 함수 및 기울기 절대값들 중 최대값과 최소값 간의 차이(difference between maximum and minimum absolute slopes: $DAS$) 함수를 결정기준으로 사용하여 기준 값 혹은 기준 백분위 수보다 작으면 해당 유전자를 제거한다. 유전자 $i$에 대한 각각의 계산식은 다음과 같다.

$$AAS_i = \frac{1}{T-1}\sum_{k=1}^{T-1}\left|\frac{x_{it_{k+1}} - x_{it_k}}{t_{k+1} - t_k}\right|,$$

$$DAS_i = \left|\max_k\left(\left|\frac{x_{it_{k+1}} - x_{it_k}}{t_{k+1} - t_k}\right|\right) - \min_k\left(\left|\frac{x_{it_{k+1}} - x_{it_k}}{t_{k+1} - t_k}\right|\right)\right|.$$

## 3. 붓스트랩을 이용한 필터링방법

필터링을 위한 상대적 기준으로 백분위수가 흔히 사용되는데 이는 다소 직관적이고 쉽게 다룰 수 있는 장점 때문일 것이다. 그러나 임의의 특정값을 적용하는 절대적 기준 보다는 더 객관적인 방법으로 고려되지만 주어진 관측표본에 매우 큰 영향을 받는다. 따라서 본 논문에서는 이러한 자료에 대한 의존성을 감소시키며 더 많은 객관성을 확보할 수 있는 방법으로서 붓스트랩을 이용한 방법을 소개하고자 한다.

사용하는 자료의 변이에 따라 달라지는 통계량의 값을 유의확률(p-value)로 바꾸어 해석하면 서로 다른 자료에 대해서도 객관적인 비교가 가능해진다. 그러나 이를 위해서는 귀무가설 하에서 검정 통계량의 분포를 알아야 하지만 흔히 그 분포를 알 수 없는 경우가 대부분이다. 이러한 경우 붓스트랩과 같은 재표본추출법을 통해 관측자료로부터 모 분포를 생성하여 참(true) 분포를 대신하는 방법이 적용되고 있다. 이는 주어진 자료에 의한 영향을 완전히 배제시키진 못하지만 관측표본 자료로부터 랜덤하게 새로운 관측치를 발생시키는 과정을 무수히 많이 반복하여 가상의 모분포를 생성함으로서 단순히 관측된 자료만을 그대로 이용하는 백분위수 방식보다 더 일반적인 결과를 얻을 수 있을 것으로 여겨진다.

de Lichtenberg 등(2005)는 활동성(regulation)을 가지는 의미있는 유전자를 추려내기 위해서 붓스트랩 유의성검정을 제안하였다. 본 논문에서는 de Lichtenberg 등(2005)이 제시한 붓스트랩 유의성검정 방법을 기초로 하였다. 먼저, 관측된 모든 유전자 발현프로파일 가운데 랜덤하게 하나의 유전자에 대한 발현프로파일을 추출하고, 추출된 유전자의 발현 값들을 이용하여 새로운 임의의 발현프로파일을 만드는데, 각 관측시점에 대응하는 발현 값들은 반복을 허용하여 랜덤하게 추출함으로서 하나씩 채워진다. 즉, 원래의 관측된 시점은 무시되고 관측된 발현수준 값만을 이용하여 발생 가능한 새로운 유전자 발현 프로파일이 생성되는 것이다.

이와같은 가상의 유전자 발현 프로파일의 생성과정을 무수히 반복수행하여 최종적인 붓스트랩 표본을 생성하고 붓스트랩 표본에 속한 각 유전자들에 대해 관심 있는 통계량(필

터링 함수)을 계산하여 해당 통계량의 모 분포를 형성한다. 다음으로는 실제 관측된 유전자 발현 프로파일로부터 관측 통계량 값을 계산하고 붓스트랩에 의해 형성된 통계량의 모 분포에 대응시켜 유의확률을 계산한 후 유의수준($\alpha$)과 비교하여 더 큰 유의확률을 갖는 유전자 자료들은 무 변화 패턴을 갖는 것으로 판단하여 제거시킨다.

4절의 자료분석에서 사용된 이스트자료에 분산함수를 사용하여 붓스트랩 방법에 의한 5% 유의성 검정을 하였을때 총 2,425개의 유전자중에서 129(=2425-2296)개의 유전자만이 선택되는 것을 표 4.1로 부터 알 수 있다. 그러나 무 변화 패턴을 가지는 유전자를 제거하는 것은 핵심적인 본 연구에 들어가기 전에 수행하는 전처리 과정에 속하므로 작은 수의 가장 활동적인 유전자를 선택하는 문제가 아니라 본 연구에 기여할 것 같지 않는 가장 비 활동적인 유전자를 선택하여 제거하는 문제이므로 유의한 유전자 선택문제에서 통상 적용하는 5% 혹은 10% 유의수준보다는 90% 혹은 95% 유의수준의 사용이 더 적절하나 이러한 유의수준의 사용은 통계학에서 일반적인 경우가 아니므로 본 논문에서는 붓스트랩 표본에서의 백분위수로 표현하도록 하겠다. 즉, 붓스트랩 $\alpha$% 유의성검정에 의한 유전자 필터링은 붓스트랩 표본에 대한 $(100-\alpha)$pct(percentile) 기준에 의한 필터링을 의미한다.

## 4. 수치 분석

본 연구의 수치분석에서는 2절에서 소개된 여러 가지 필터링 방법들 가운데 객관적이고 직관적으로 쉽게 사용 가능한 것들로서 상대적 기준에 의해 필터링을 수행할 수 있는 $MATLAB$에 내장된 절대값(genelowvalfilter), 분산(genevarfilter) 및 엔트로피(geneentropyfilter) 함수와 파레토 함수($AAS, DAS$) 및 상대적 평균 함수($RA$)에 대해 필터링 방식으로서 관측표본에 대한 백분위수 방식과 붓스트랩 표본에 대한 백분위수 방식에 의한 결과를 비교 검토하였다.

분석에 사용된 자료는 Spellman 등(1998)의 이스트 세포주기(cell-cycle) 자료에서 alpha-factor 부분을 발췌한 것으로서, 6,178개 유전자에 대해 18개 시점에서 관측된 시간경로자료이다. 필터링에 사용된 자료는 결측치를 갖는 자료들을 제외하고 총 2,425개를 사용하였다. 또한, DeRisi 등(1997)의 이스트 자료는 6,400개 유전자에 대해 7개 시점에서 관측된 단기간 시간경로자료로서 최종 6,276개의 결측치 없는 자료를 사용하였다. 2가지 자료에 대한 분석 결과는 서로 거의 유사한 양상을 보였으므로 Spellman의 자료에 대한 결과만을 제시하였다.

그림 4.1은 분산함수를 적용하였을때 관측표본 및 붓스트랩 표본에 대한 5pct 기준으로 필터링된 자료와 남아 있는 자료의 시계열그림이다. 이제 각 필터링 함수에 대해 각 필터링 방식에 따른 차이를 살펴보기 위해 제거되는 유전자 그룹(filtered group)에 대한 분포를 살펴보기로 한다. 각 경우에 해당되는 유전자의 전체 자료에 대한 비율(%) 및 유전자 수($n$), 사용된 필터링 함수값들의 평균(mean) 및 표준편차($SD$)를 표 4.1에 제시하였다. 이 표에서 관측표본 기준은 표본자료에 대한 백분위수에 따른 결과이고, 붓스트랩 기준은 붓스트랩 표본에 대한 백분위수에 따른 결과이다. 붓스트랩의 표본추출 횟수는 de Lichtenberg 등(2005)에서처럼 1,000,000번으로 하였으나 이에 대한 적절한 지침은 아직까지는 없다.

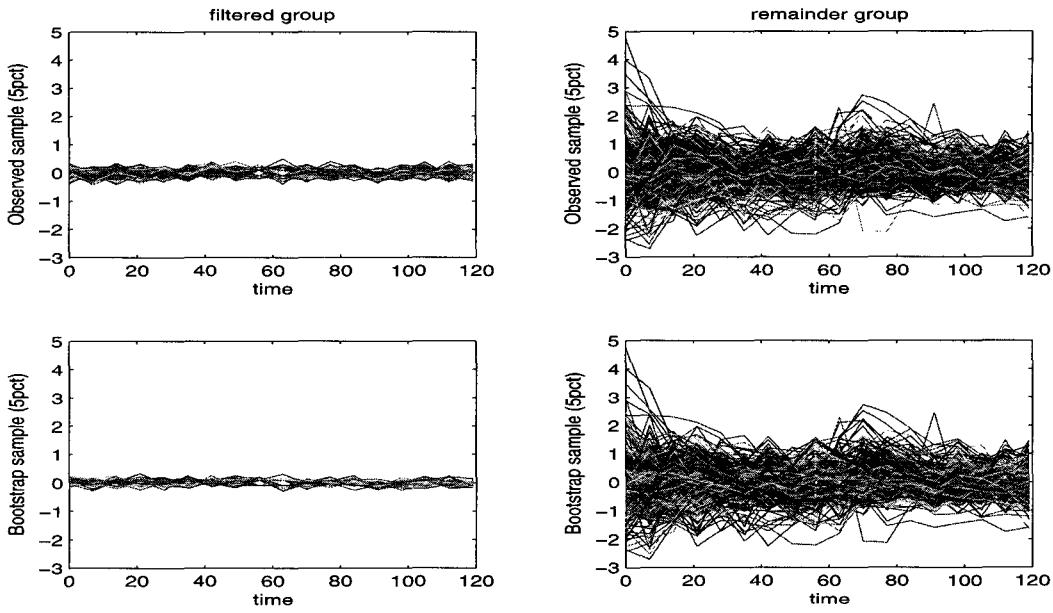


그림 4.1: 분산함수에 의해 제거된 자료와 남아 있는 자료의 시계열그림

절대값 함수를 사용하는 경우 기준 백분위수를 정하는 모집단이 $P \times T$개 유전자 발현값들의 전체가 되므로 이에 상응하는 붓스트랩 표본에 대한 백분위수 기준의 적용은 적절치가 않아 이의 결과는 제시하지 않았다.

한편, 절대값 함수를 사용한 경우에 관측표본에 대한 백분위수 기준에 의한 필터링결과는 5 pct 기준에서 50 pct 기준까지 거의 필터링이 되지 않는 매우 보수적인 경향을 보였다. 이와 같은 이유는 어떤 유전자의 모든 시점에서의 관측값의 절대값을 기준백분위수와 비교하여 기준 백분위수보다도 모두 작아야 그 유전자를 제거하게 되는데 기인한다. 예를 들어 100개의 유전자에 대해 10개 시점에서 관측된 절대값 자료를 고려해 보면 총 관측 자료수는 1,000 개일 것이고, 5 백분위수 이하의 값들은 50개의 가장 작은 값들로 구성되어 있을 것이다. 따라서 어떤 한 유전자의 10개 관측값들 모두가 50개의 가장 작은 값들 중 하나가 되기는 매우 드문 경우로 볼 수 있다.

붓스트랩 표본에서 $RA$ 값들은 관측표본에 비해 0 근처에 매우 밀집하여 분포하는 형태를 갖고 있다. 즉, 붓스트랩 표본에서 $RA$ 값들의 분포는 (minimum, 5pct, 10pct, 50pct, 90pct, 95pct, maximum)$= 1.0e+041*$ $(-2.3347e+017, -5.8, -2.9, 0, 2.8, 5.7, 2.4644e+017)$ 과 같고 관측 표본에서 $RA$ 값들의 분포는 (minimum, 5pct, 10pct, 50pct, 90pct, 95pct, maximum) $= 1.0e+041*$ $(-0.6355e+017, -214.3, -86.5, 0, 78.5, 188.9, 1.8159+017)$ 과 같다. 따라서 5pct, 10pct 기준에서는 붓스트랩 방식이 관측표본의 백분위수 방식에 비해 매우 많은 유전자들을 제거하게 되고 반대로 90pct, 95pct 기준에서는 훨씬 적게 유전자들을 제거하는 현상이 발생하고 있다.

표 4.1: 제거된 유전자들에 대한 필터링 함수값의 기술통계값

| 필터링 함수 | 관측표본 기준 | | | | 백분위수 (pct) | 붓스트랩 표본 기준 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 비율 | $n$ | mean | SD | | 비율 | $n$ | mean | SD |
| 절대값 | 0.0 | 0 | - | - | 5 | - | - | - | - |
| | 0.0 | 0 | - | - | 10 | - | - | - | - |
| | 0.0 | 1 | 0.180 | 0.000 | 50 | - | - | - | - |
| | 41.3 | 1002 | 0.380 | 0.064 | 90 | - | - | - | - |
| | 67.4 | 1636 | 0.447 | 0.102 | 95 | - | - | - | - |
| 분산 | 5.0 | 121 | 0.020 | 0.004 | 5 | 1.2 | 29 | 0.015 | 0.003 |
| | 10.0 | 242 | 0.023 | 0.004 | 10 | 4.6 | 111 | 0.020 | 0.003 |
| | 50.0 | 1,213 | 0.041 | 0.012 | 50 | 43.7 | 1,060 | 0.038 | 0.011 |
| | 90.0 | 2,183 | 0.070 | 0.043 | 90 | 89.1 | 2,160 | 0.068 | 0.040 |
| | 95.0 | 2,304 | 0.081 | 0.065 | 95 | 94.7 | 2,296 | 0.080 | 0.063 |
| 엔트로피 | 5.1 | 124 | 2.186 | 0.480 | 5 | 0.9 | 23 | 1.377 | 0.559 |
| | 10.1 | 244 | 2.467 | 0.450 | 10 | 1.9 | 45 | 1.741 | 0.550 |
| | 51.2 | 1,241 | 3.118 | 0.413 | 50 | 15.1 | 365 | 2.629 | 0.435 |
| | 90.4 | 2,193 | 3.390 | 0.447 | 90 | 59.8 | 1,450 | 3.186 | 0.417 |
| | 97.4 | 2,361 | 3.437 | 0.463 | 95 | 73.4 | 1,779 | 3.280 | 0.425 |
| *AAS* | 5.1 | 123 | 0.021 | 0.002 | 5 | 1.4 | 34 | 0.019 | 0.002 |
| | 10.1 | 244 | 0.023 | 0.002 | 10 | 4.3 | 104 | 0.021 | 0.002 |
| | 50.0 | 1,213 | 0.030 | 0.005 | 50 | 51.8 | 1,255 | 0.030 | 0.005 |
| | 90.0 | 2,183 | 0.037 | 0.019 | 90 | 97.2 | 2,357 | 0.039 | 0.011 |
| | 95.0 | 2,304 | 0.038 | 0.010 | 95 | 99.5 | 2,413 | 0.040 | 0.013 |
| *DAS* | 5.3 | 129 | 0.048 | 0.006 | 5 | 3.1 | 75 | 0.045 | 0.005 |
| | 10.4 | 253 | 0.053 | 0.006 | 10 | 8.4 | 203 | 0.051 | 0.006 |
| | 50.1 | 1,214 | 0.071 | 0.013 | 50 | 56.4 | 1,367 | 0.074 | 0.014 |
| | 90.0 | 2,183 | 0.091 | 0.027 | 90 | 94.6 | 2,295 | 0.095 | 0.032 |
| | 95.1 | 2,306 | 0.096 | 0.033 | 95 | 98.3 | 2,383 | 0.099 | 0.039 |
| *RA* | 5.0 | 121 | -13.1e56 | 12.7e56 | 5 | 49.4 | 1,198 | -1.3e56 | 5.6e56 |
| | 10.0 | 243 | -6.5e56 | 11.1e56 | 10 | 49.9 | 1,211 | -1.3e56 | 5.6e56 |
| | 50.0 | 1213 | -1.3e56 | 5.6e56 | 50 | 49.9 | 1,216 | -1.3e56 | 5.6e56 |
| | 90.0 | 2183 | -0.7e56 | 4.2e56 | 90 | 49.9 | 1,230 | -1.3e56 | 5.6e56 |
| | 95.0 | 2304 | -0.7e56 | 4.1e56 | 95 | 50.9 | 1,251 | -1.3e56 | 5.6e56 |

* 절대값, 분산, 엔트로피 함수: MATLAB 내장된 함수임.
* AAS, DAS 함수: 파레토 함수임.

관측표본의 백분위수 기준에 의한 결과에 대비하여 붓스트랩 표본의 백분위수 기준에 의한 결과를 살펴보면 *RA* 함수를 제외한 필터링 함수들은 전반적으로 낮은 백분위수인 5pct, 10pct 에서는 붓스트랩 방식이 상당히 더 적은 수의 유전자들을 제거하는 보수적인 경향을 보였다. 한편 높은 백분위수인 90pct, 95pct 에서는 분산 함수는 두 가지 방식의 결과가 크게 차이가 나지 않은 반면에 엔트로피 함수는 붓스트랩 방식이 상당히 더 보수적인 경향을 보였고, 파레토 함수는 오히려 관측표본의 백분위수 방식이 더 보수적인 경향을 보였다.

각 필터링 함수별로 관측표본 및 붓스트랩 표본에 대한 5pct 기준으로 필터링된 자료들을 나타낸 그림 4.2를 보면 엔트로피 함수와 *RA* 함수에 의한 필터링 결과에는 시간에 따라 변동성을 가지는 유전자들도 상당수 포함되어 있음을 알 수 있다. 또한 엔트로피 함수 및 *RA* 함수에 비해 상대적으로 비슷한 필터링 결과를 보여주는 분산 함수, *AAS* 함수, 그리고 *DAS* 중에서는 분산 함수가 보다 안정적인 필터링 결과를 주고 있으며 붓스트랩 방법

이 관측표본에 기초한 백분위수 방법보다 더 보수적인 결과를 보여주고 있다. 이러한 평가는 필터링 후의 제거된 자료 그룹과 남아 있는 자료 그룹을 비교하여 그린 상자그림인 그림 4.3으로 부터 다시 확인 할 수 있다. 그림 4.3에서 필터링함수 및 백분위수의 대상이 되는 표본에 따라 필터링의 결과가 많이 달라짐을 볼 수 있다. 필터링 된 그룹의 상자그림을 살펴보면 엔트로피 함수는 관측표본이나 붓스트랩 표본 모두에서 0을 중심으로 매우 넓게 퍼져있는 반면, 분산함수 및 $AAS$함수, $DAS$함수는 0을 중심으로 더 많이 밀집된 형태를 보였으며 이러한 경향은 관측표본 보다 붓스트랩 표본에서, 특히 5pct 기준에서 더 강하게 나타났다. 또한 이 세 가지 함수들 중에서도 분산함수는 관측표본이나 붓스트랩 표본 모두에서 전반적으로 더 밀집한 형태를 보였다. 따라서 사용되는 각 함수 혹은 각 표본별로 남겨진 유전자 그룹과 제거되는 유전자 그룹에 대해 상자그림(box plot) 등을 그려 그룹 간의 분포 특성을 살펴보는 것이 무 변화 패턴을 가지는 시간경로 유전자 자료를 필터링 할 때 많은 도움을 줄 것이다.

다음으로는 여러 필터링 함수 및 필터링 방식을 동시에 적용하여 공통으로 남겨지는 유전자 수를 검토하고 그 특성을 파악하기 위해 다차원적으로 접근한 결과를 표 4.2에 제시하였다. 필터링 함수는 절대값 함수를 제외한 5가지 함수 가운데 1가지, 2가지 또는 3가지를

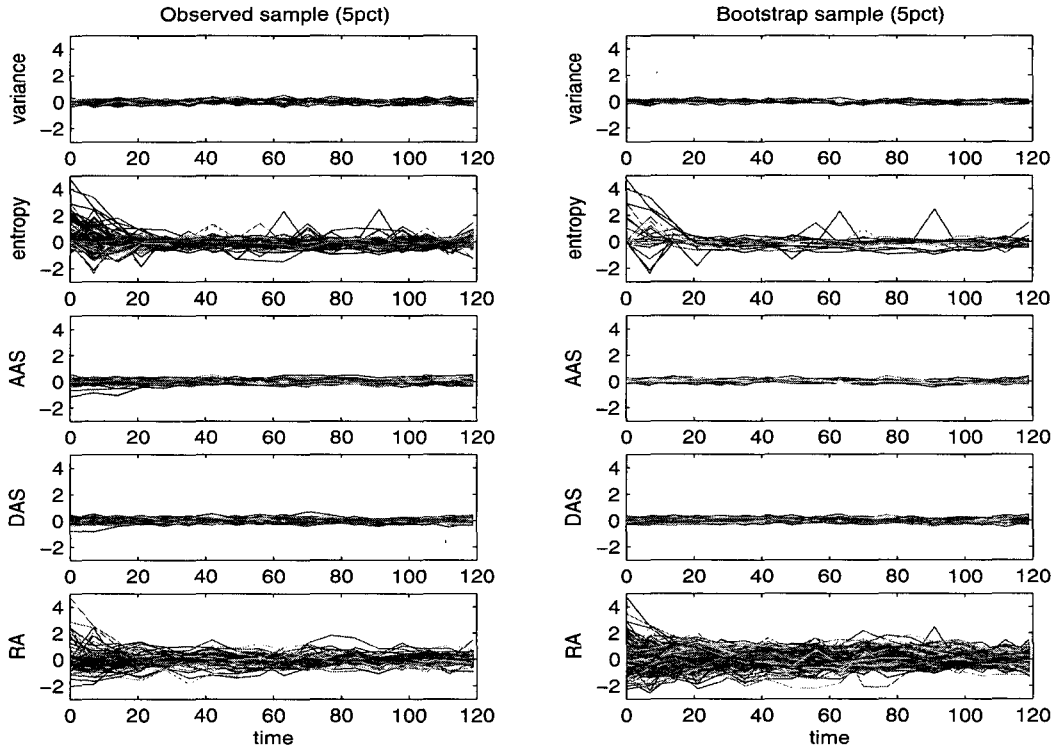


그림 4.2: 여러가지 필터링 방법에 따라 제거된 자료들의 시계열그림

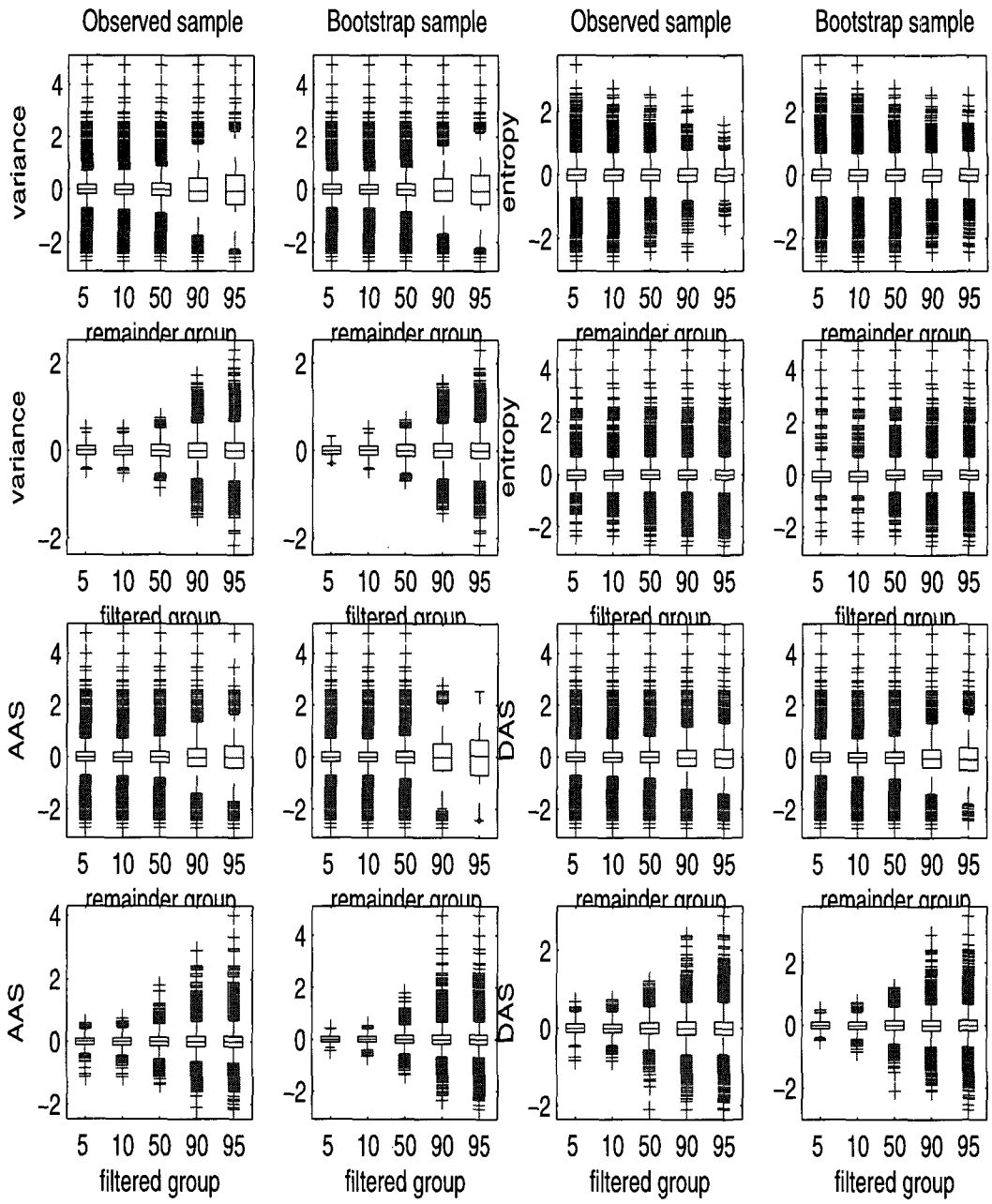


그림 4.3: 여러가지 필터링 방법에 따라 제거된 자료와 남아 있는 자료들의 백분위수별 상자그림

표 4.2: 여러 가지 필터링 방법을 적용하여 공통으로 남겨지는 유전자 수들의 비교

| 필터링 함수 | 5pct | | | 10pct | | | 90pct | | | 95pct | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *O* | *B* | *O&B* | *O* | *B* | *O&B* | *O* | *B* | *O&B* | *O* | *B* | *O&B* |
| 1 | 2,304 | 2,396 | 2,304 | 2,182 | 2,314 | 2,183 | 242 | 265 | 242 | 121 | 129 | 121 |
| 2 | 2,301 | 2,402 | 2,301 | 2,181 | 2,380 | 2,181 | 232 | 975 | 232 | 64 | 646 | 64 |
| 3 | 2,302 | 2,391 | 2,302 | 2,181 | 2,321 | 2,181 | 242 | 68 | 68 | 121 | 12 | 12 |
| 4 | 2,296 | 2,350 | 2,296 | 2,172 | 2,222 | 2,172 | 242 | 130 | 130 | 119 | 42 | 42 |
| 5 | 2,304 | 1,227 | 1,227 | 2,182 | 1,214 | 1,214 | 242 | 1,195 | 242 | 121 | 1,174 | 121 |
| 1, 2 | 2,183 | 2,373 | 2,183 | 1,954 | 2,269 | 1,954 | 28 | 101 | 28 | 5 | 41 | 5 |
| 1, 3 | 2,238 | 2,374 | 2,238 | 2,072 | 2,259 | 2,072 | 147 | 60 | 59 | 69 | 11 | 11 |
| 1, 4 | 2,229 | 2,336 | 2,229 | 2,054 | 2,179 | 2,054 | 139 | 98 | 93 | 55 | 31 | 31 |
| 1, 5 | 2,189 | 1,208 | 1,156 | 1,952 | 1,152 | 1,078 | 49 | 113 | 49 | 16 | 47 | 16 |
| 2, 3 | 2,183 | 2,368 | 2,183 | 1,957 | 2,276 | 1,957 | 30 | 35 | 9 | 3 | 6 | 0 |
| 2, 4 | 2,175 | 2,327 | 2,175 | 1,938 | 2,178 | 1,938 | 10 | 31 | 5 | 0 | 5 | 0 |
| 2, 5 | 2,190 | 1,216 | 1,176 | 1,976 | 1,194 | 1,103 | 30 | 485 | 30 | 5 | 317 | 5 |
| 3, 4 | 2,232 | 2,337 | 2,232 | 2,059 | 2,184 | 2,059 | 139 | 42 | 42 | 54 | 7 | 7 |
| 3, 5 | 2,188 | 1,207 | 1,157 | 1,950 | 1,154 | 1,078 | 45 | 25 | 17 | 17 | 3 | 0 |
| 4, 5 | 2,184 | 1,191 | 1,163 | 1,941 | 1,115 | 1,091 | 39 | 48 | 20 | 16 | 15 | 8 |
| 1, 2, 3 | 2,120 | 2,351 | 2,120 | 1,852 | 2,214 | 1,852 | 20 | 30 | 7 | 3 | 6 | 0 |
| 1, 2, 4 | 2,110 | 2,313 | 2,110 | 1,829 | 2,135 | 1,829 | 10 | 30 | 5 | 0 | 5 | 0 |
| 1, 2, 5 | 2,078 | 1,197 | 1,107 | 1,761 | 1,132 | 976 | 6 | 48 | 6 | 1 | 19 | 1 |
| 1, 3, 4 | 2,187 | 2,325 | 2,187 | 1,992 | 2,152 | 1,992 | 108 | 41 | 41 | 43 | 7 | 7 |
| 1, 3, 5 | 2,128 | 1,195 | 1,121 | 1,850 | 1,122 | 1,017 | 29 | 19 | 14 | 10 | 3 | 1 |
| 1, 4, 5 | 2,118 | 1,181 | 1,120 | 1,828 | 1,087 | 1,016 | 26 | 35 | 16 | 9 | 10 | 5 |
| 2, 3, 4 | 2,114 | 2,314 | 2,114 | 1,838 | 2,140 | 1,838 | 9 | 19 | 4 | 0 | 3 | 0 |
| 2, 3, 5 | 2,079 | 1,196 | 1,110 | 1,764 | 1,134 | 978 | 7 | 15 | 3 | 1 | 2 | 0 |
| 2, 4, 5 | 2,073 | 1,180 | 1,115 | 1,745 | 1,096 | 985 | 3 | 11 | 2 | 0 | 3 | 0 |
| 3, 4, 5 | 2,123 | 1,181 | 1,120 | 1,834 | 1,087 | 1,018 | 26 | 11 | 6 | 8 | 3 | 1 |

* pct: 백분위수
* 필터링 함수: 1. 분산, 2. 엔트로피, 3. AAS , 4. DAS, 5. RA 함수.
* *O*: 관측표본 기준, *B*: 붓스트랩 표본 기준, *O&B*: 관측표본 및 붓스트랩 표본 기준 모두 적용.

동시에 적용하고, 이 때 필터링 방식은 관측표본의 백분위수 방식($O$), 붓스트랩 표본의 백분위수 방식($B$) 및 두 가지 혼합한 방식($O\&B$) 등이 적용되었다. 관측표본의 백분위수 방식에 의해서는 분산, $AAS$, $DAS$ 함수가 그 외 함수들에 비해 상대적으로 서로 유사한 결과를 보였다. 또한 관측표본 및 붓스트랩 표본에서의 결과를 비교할 때 분산 함수가 다른 함수들에 비해 상대적으로 서로 유사한 결과를 갖는 것으로 나타났다.

## 5. 결론

시간경로 유전자 발현자료에 대한 본격적인 자료분석을 수행하기에 앞서 전처리 과정을 통해 유전 정보를 밝히는데 거의 도움을 주지 못하거나 자료분석 결과에 오히려 좋지 않은 영향을 줄 것으로 판단되는 자료들을 미리 찾아내어 제거함으로써 방대한 양의 마이크로어레이 자료를 분석하는 과정에서 부딪히는 막대한 계산 양을 다소 줄일 수 있다.

본 논문에서는 여러 전처리 방법 중에서 기존의 무 변화 패턴 필터링 방법들을 살펴보고, 그 중에서 직관적으로 쉽게 사용가능한 몇 가지 필터링 함수를 중심으로 관측표본에 대한 백분위수 기준 방식과 붓스트랩 표본에 대한 백분위수 기준 방식의 결과를 비교 검토하

였다.

관측표본의 백분위수 방식은 관측된 자료 자체 내에서만 상대적 비교를 하므로 관측 자료에 매우 의존적인 단점이 있다. 한편 이를 보완할 수 있는 하나의 대안으로서 붓스트랩 방식은 상대적으로 계산 양은 많으나 관측자료를 기반으로 발생가능한 값들로 구성된 가상의 유전자 발현프로파일을 무수히 반복생성하여 확률기반 하에 필터링을 수행 하므로 주어지는 자료에 대한 의존성을 약화시킬 수 있다.

본 연구에서 고려된 여러 필터링 함수들 가운데 분산 함수를 적용한 결과가 그 외 다른 필터링 함수의 결과에 비해 관측표본에서의 백분위수 방식과 붓스트랩 표본에서의 백분위수 방식에 따른 결과가 상대적으로 더 유사한 것으로 나타났다. 따라서 붓스트랩 방식과 같은 다소 복잡한 절차를 거치지 않고 간편하게 관측표본의 백분위수 기준에 의해 유전자 필터링을 하고자 하는 경우에는 분산 함수가 가장 안정적이라고 볼 수 있다. 한편, 무 변화 패턴을 가지는 유전자를 제거하는 것은 전처리 과정에 속하므로 유전자 제거에 신중을 기해야 한다. 이때 제거된 유전자는 핵심적인 본 연구에 참여 기회를 잃게 된다. 이러한 측면에서 보자면 보수적이고 보다 안전한 경향을 보여준 붓스트랩 표본에 기초한 백분위수 방법을 적용하고 이때 필터링 함수로는 분산함수를 사용할 것을 제안한다.

## 참고문헌


Chen, Y., Bittner, M. L. and Dougherty, E. R. (1999). Issues associated with microarray data analysis and integration, *Nature Genetics*, **22**, 213–216.

DeRisi, JL., Iyer VR. and Brown PO. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, **278**, 680–686.

Dudoit, S., Fridlyand, J. and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, **97**, 77–87.

Fleury, G.A., Hero, O., Yoshida, S., Carter, T., Barlow, C. and Swaroop, A. (2002). Pareto analysis for gene filtering in microarray experiments, *In Proceedings of the European Signal Processing Conference(EuSIPCO), Toulo- use, France.*

Herrero, J., Díaz-Uriate, R. and Dopazo, J. (2003). Gene expression data preprocessing, *Bioinformatics*, **19**, 655–656.

Kadota K., Tominaga D., Akiyama Y. and Takahashi K. (2003). Detection outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification, *Chem-Bio Informatics Journal*, **3**, 30–45.

de Lichtenberg U., Jensen, L. J., Fausboll, A., Jensen, T. S., Bork, P. and Brunak, S. (2005). Comparison of computational methods for the identification of cell cycle-regulated genes, *Bioinformatics*, **21**, 1164–171.

Lindlöf, A. and Olsson, B. (2003). Genetic network inference: the effects of preprocessing, *BioSystems*, **72**, 229–239.

Liang, Y., Tayo, B., Cai, X. and Kelemen, A. (2005). Differential and trajectory methods for time course gene expression data, *Bioinformatics*, **21**, 3009–3016.

Spellman P., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O, Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-

regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Molecular Biology of the Cell*, **9**, 3273–3297.
The Math Works, Inc. (2003). *MATLAB/Bioinformatics toolbox*, Version 1.0, Natick, MA.

[ 2006년 10월 접수, 2007년 3월 채택 ]

# Comparison of Functions for Filtering Time Course Gene Expression Data with Flat Patterns*

Kyungsook Kim[1] Mira Oh[2] Jangsun Baek[3] Young Sook Son[4]

ABSTRACT

Filtering genes that do not appear to contribute to regulation prior to the statistical analysis of time course gene expression data can reduce the dimensions of data and the possibility of misinterpretation due to noise or lack of variation. In this paper, we compare six different functions for filtering genes with flat pattern under the percentile criterion on an observed sample and that on a bootstrap sample. The result of applying to the yeast cell cycle data shows that the variance function is most similar in both samples.

*Keywords:* Preprocessing, flat pattern filtering, time course gene expression, percentile, bootstrap.

* This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD)(KRF-2005-204-C00017).

1) Doctoral Course, Department of Statistics, Chonnam National University, 300, Yongbong-dong, Buk-gu, Gwangju 500-757, Korea
E-mail: ksook620@jnu.ac.kr

2) Doctoral Course, Department of Statistics, Chonnam National University, 300, Yongbong-dong, Buk-gu, Gwangju 500-757, Korea
E-mail: omr@chonnam.ac.kr

3) Professor, Department of Statistics, Chonnam National University, 300, Yongbong-dong, Buk-gu, Gwangju 500-757, Korea
E-mail: jbaek@chonnam.ac.kr

4) (Corresponding author) Professor, Department of Statistics, Chonnam National University, 300, Yongbong-dong, Buk-gu, Gwangju 500-757, Korea
E-mail: ysson@chonnam.ac.kr