응용통계연구 (2009)
22(3), 585–594

마이크로어레이 자료의 사전 처리 순서에 따른 검색의 일치도 분석

김상철[1] · 이재휘[2] · 김병수[3]

[1]연세대학교 응용통계학과, [2]연세대학교 응용통계학과, [3]연세대학교 응용통계학과

(2008년 9월 접수, 2008년 11월 채택)

요 약

마이크로어레이 실험의 실험자들은 원 측정치인 영상을 조사하여 통계적 분석이 가능한 자료의 형태로 변환하는데 이러한 과정을 흔히 사전 처리라고 부른다. 마이크로어레이의 사전 처리는 불량 영상의 제거(filtering), 결측치의 대치와 표준화로 세분되어질 수 있다. 표준화 방법과 결측치 대치 방법 각각에 대하여서는 많은 연구가 보고되었으나, 사전 처리를 구성하는 원소들간의 적정한 순서에 대하여서는 연구가 미흡하다. 표준화 방법과 결측치 대치 방법 중 어느 것이 먼저 실시되어야 하는지에 대하여서 아직 알려진 바가 없다. 본 연구는 사전 처리 순서에 대한 탐색적 시도로서 대장암과 위암을 대상으로 실시한 두 조의 cDNA 마이크로어레이 실험 자료를 이용하여 사전 처리를 구성하는 원소들간의 다양한 순서에 따라 검색된 특이 발현 유전자 군이 어떻게 변화하는지를 분석하고 있다. 즉, 결측치 대치와 표준화의 여러가지 방법들의 조합에 따라 검색된 특이 발현 유전자 군이 얼마나 일치적인가를 확인하고자 한다. 결측치 대치 방법으로는 K 최근접 이웃 방법과 베이지안 주성분 분석을 고려하였고, 표준화 방법으로는 전체 표준화, 블럭별 국소(within-print tip group) 평활 표준화 그리고 분산 안정화를 유도하는 표준화 방법을 적용하였다. 따라서 사전 처리를 구성하는 두개 원소가 각각 2개 수준과 3개 수준을 가지고 있고, 두개 원소의 순열에 따른 모든 가능한 사전 처리 개수 수는 12개가 된다. 본 연구에서는 12개 사전 처리 방법 각각에 따라 정상 조직과 암 조직간 특이적으로 발현하는 유전자 군을 검색하였고, 사전 처리 순서를 바꾸었을때 유전자 군이 얼마나 일치적으로 유지되는지를 파악하고 있다. 표준화 방법으로 분산 안정화 표준화를 사용할 경우는 사전 처리 순서에 따라 특이 발현 유전자 군이 다소 민감하게 변하는 것을 보이고있다.

주요용어: 결측치 대치, 마이크로어레이, 사전 처리 단계, 일치도, 표준화, t_3통계량.

1. 서론

수 만개의 유전자 발현을 동시에 관찰 가능한 마이크로어레이 실험은 유전체를 이용한 생물학 분야에서 광범위하게 이용되는 방법이다. 이 기술은 인간의 암 발생 및 진전에 따른 유전자의 변이에 대한 연구, 생존 분석을 통한 암의 분자 예후 지표 개발 그리고 특정 처리와 시간에 따른 유전자의 변동에 대한 연구에 활발히 응용된다. 마이크로어레이 실험은 여러 단계의 실험 과정을 거치므로 각 단계마다 시스템 변동에 의한 편향이 발생할 수 있다. 편향의 발생 원인으로는 열이나 빛 등에 대한 두 염료[4]의 민감도 차

본 연구는 2007년도 연세대학교 상경대학 기초학문분야 연구 지원을 받아 수행된 연구임.
[3]교신저자: (120-749) 서울특별시 서대문구 성산로 262, 연세대학교 응용통계학과, 교수.
E-mail: bskim@yonsei.ac.kr

[4]cDNA 마이크로어레이에서는 흔히 적색 염료나 녹색 염료가 사용되며, 각각 처리군과 대조군의 표본에 표지된다. 대조군과 처리군을 같은 표본으로 구성하여 cDNA 마이크로어레이 실험을 하면 일반적으로 녹색염료의 보합 강도(G)가 적색염료의 보합 강도(R) 보다 다소 크게 나타난다.

이, 염료 혼합의 효율성 차이, 스캐닝 과정의 차이, 프린트 팁 간의 차이, 슬라이드의 공간적 효과와 슬라이드 간 이질성 등이 있다. 이상의 편향을 제거하는 방법으로 여러 가지의 표준화 방법이 제안되었다 (Yang 등, 2002; Huber 등, 2002). Yang 등 (2002)은 염료강도의 로그비를 $M = \log_2(R/G)$으로 정의하고, $A = 1/2\log_2(RG)$를 정의하여 M은 y축으로 A를 x축으로 하는 M-A 그림을 이용하여 몇 가지 표준화를 제안하였다. 전체 자료의 로그비의 평균이나 중앙값을 0으로 표준화하는 전체 표준화, 자료의 비선형 관계를 lowess 함수를 이용하여 표준화하는 발현 강도 의존 표준화, 슬라이드의 공간적 영향과 프린트 팁 마다 발생하는 비 선형 관계를 표준화하는 블럭별 국소 평활 표준화와 블럭별 국소 평활 표준화 후 각 프린트 팁 별 분산의 크기를 일치 시키는 척도 표준화이다. Huber 등 (2002)은 Affymetrix 올리고 칩의 유전자 발현 강도나 cDNA 마이크로어레이의 각 염료 별 형광강도 또는 형강강도의 로그비를 이용하여 분산 안정화를 유도하는 표준화 방법을 제시하였다. 다양한 통계적 방법을 적용하기 위해서는 실험 자료를 $n \times p$ (n: 어레이의 개수, p: 유전자의 개수)의 완전한 자료로 생성해야한다. 그러나, 이미지 분석시 불량 스팟의 제거(filtering)에서 제거된 유전자는 결측치로 존재하며, 소수의 결측치가 있는 유전자를 제거하면 많은 정보의 손실을 가져온다. 이 경우 전체 실험에서 유전자의 결측치 비율이 0~40% 이하의 결측된 자료를 대치하여 분석에 이용하는 결측치 대치 방법이 제안되었다. 결측치 대치 방법으로는 Troyanskaya 등 (2001)이 제시한 행 평균 방법, K 최근접 이웃 방법, 주성분 분석 방법과 Oba 등 (2003)의 베이지안 주성분 분석 방법 등이 있다. 대부분의 마이크로어레이는 실험의 질적 상태와 동질성을 확인하기 위해서 동일 유전자를 여러 번 점적하여 사용한다. Kim 등 (2005)은 중복된 유전자를 처리하기 위하여 중복 유전자의 평균값으로 대표하여 사용하였다. 위에서 언급한 불량 스팟의 제거, 표준화, 결측치 대치와 중복 유전자 처리 단계 등을 사전 처리 단계라 한다 (Kim 등, 2005). 이를 통하여 얻어진 $n \times p$ 행렬 자료를 이용하여 암 환자의 정상 조직과 암에서 특이적으로 발현하는 유전자를 찾기 위한 방법으로 Ge 등 (2003)과 Westfall과 Young (1993)이 제시한 비모수적 t검정과 족별 오류율 방법이 있다. 짝을 이룬 설계 실험에서는 종종 시료의 부족 등으로 결측된 자료가 발생된다. 이 경우 짝을 이룬 자료와 독립 자료가 혼합된 자료 형태를 가지게 되며, 이 자료를 결합하여 특이 발현 유전자를 검색할 수 있는 방법으로서 Kim 등 (2005)은 t_3통계량을 제안한 바 있다. 본 연구는 Kim 등 (2005)이 발표한 대장암 마이크로어레이 자료와 Kim 등 (2006)의 위암 마이크로어레이 자료에 대해 3가지 표준화 방법(분산 안정화 표준화, 전체 표준화, 블럭별 국소 평활 표준화)과 2가지 결측치 대치 방법(K 최근접 이웃 방법, 베이지안 주성분 분석 방법)의 순서 조합으로 12가지 사전 처리 방법을 구성하였다. 그리고 각 사전 처리 과정에 따라 t_3통계량 방법과 Benjamini와 Hochberg의 다중 검정 절차로 특이 발현 유전자를 검색하고, 그 일치도를 측정하였다.

2. 사전 처리 방법, t_3통계량과 일치도

마이크로어레이 분석을 위한 사전 처리 과정은 불량 스팟의 제거, 표준화, 결측치 대치, 중복 유전자 처리 등의 여러 단계로 구성된다. 본 논문에서 사용한 표준화 방법은 Yang 등 (2002)의 전체 표준화 및 블럭별 국소 평활 표준화 방법과 쌍곡선 아크사인 변환을 사용하는 Huber 등 (2002)의 분산 안정화를 유도하는 표준화 방법이다. 이미지 분석 단계와 제거 단계에서 결측된 유전자 중에서 결측치가 20% 이하인 유전자는 Troyanskaya 등 (2001)의 K 최근접 이웃 방법과 Oba 등 (2003)의 베이지안 주성분 분석 방법을 적용하여 결측치를 대치하였다. 위의 방법을 조합하여 사전 처리를 수행하여 얻은 자료를 대상으로 t_3통계량과 Benjamini와 Hochberg의 다중 검정 절차로 정상 조직과 암 간의 특이 발현 유전자를 검색하였다. 두 특이 발현 유전자 군 G_1과 G_2의 일치도를 $C(G_1 \cap G_2)/C(G_1 \cup G_2)$로 정의한다. 단 $C(A)$는 유한 집합 A의 원소의 개수를 나타낸다. 상기 일치도를 사용하여 다양한 사전 처리 순서에 따라 검색되는 특이 발현 유전자 군간 일치도를 평가하였다.

2.1. 표준화

마이크로어레이 실험에서 발생하는 편향를 보정하기 위해서 제시된 다양한 표준화 방법들 중 본 논문에서는 세 가지 방법을 사용한다. 두 염료의 형광강도의 로그 비를 이용하는 Yang 등 (2002)의 전체 표준화 및 블럭별 국소 평활 표준화 방법과 Huber 등 (2002)의 분산 안정화를 유도하는 표준화 방법이다.

2.1.1. 전체 표준화 전체 표준화는 적색 염료(R)의 강도와 녹색 염료(G)의 강도가 상수에 의해 연관되어있다고 가정하고($R = k \cdot G$), 로그 비의 분포의 중심을 아래 식 (2.1)과 같이 상수의 가감에 의해 0에 맞춘다.

$$\log_2\left(\frac{R}{G}\right) => \log_2\left(\frac{R}{G}\right) - c = \log_2 \frac{R}{k \cdot G}, \tag{2.1}$$

식 (2.1)에서 위치변수 $c = \log_2 k$는 유전자 전체의 로그 비의 평균이나 중앙값이다. 이러한 전체 표준화는 여전히 많이 쓰이는 표준화 기법이지만 염료 간 민감도 차이의 비선형 관계나, 프린트 팁 별 또는 슬라이드 상의 공간적 요인으로 인한 편향를 보정 할 수 없다는 단점이 있다.

2.1.2. 블럭별 국소 평활 표준화 마이크로어레이 실험은 마이크로어레이 슬라이드 제작 중에 발생할 수 있는 슬라이드 표면의 공간적 편향이나 프린트 팁의 물리적 영향에 의한 불일치를 가지고 있다. M-A 그림과 lowess(locally weighted scatter plot smoothing) 적합값에서 공간적 편향과 프린트 팁의 불일치를 확인 할 수 있다. 여기서 한 프린트 팁이 하나의 블럭을 유도한다. M-A 그림은 염료강도의 로그비를 $M = \log_2(R/G)$으로 정의하고, $A = (1/2)\log_2(RG)$를 정의하여 M은 y축으로 A를 x축으로 그린 그림이다. Lowess는 국소 가중 산점도 평활 방법으로 x축의 각 자료의 이웃하는 일정 비율의 자료를 이용하여 y값들의 중심 경향을 파악하는 방법이다. t개의 불럭이 있을 때 블럭별 국소 평활 표준화는 다음 식 (2.2)와 같다.

$$\log_2\left(\frac{R}{G}\right) => \log_2\left(\frac{R}{G}\right) - c_i(A) = \log_2 \frac{R}{k_i(A) \cdot G}, \tag{2.2}$$

$i = 1, \ldots, t$이다. 식 (2.2)에서 $c_i(A)$는 i번째 프린트 팁 군에서의 M-A 그림의 lowess 적합 값이다. 즉, 이 방법은 A-의존적 표준화 방법으로 각 프린트 팁 군 마다 lowess 적합을 한 후 표준화를 시키는 방법이다.

2.1.3. 분산 안정화 표준화 마이크로어레이 실험에서 각 유전자의 발현의 변동은 mRNA의 양, 보합효과, 표지 효과의 혼합된 결과이다. 이와 같이 여러 가지 효과의 교락 효과로 인하여 마이크로어레이 실험에서 대조군과 처리군 유전자의 발현을 직접적으로 비교할 수가 없다. 일반적으로 반복된 마이크로어레이 실험의 유전자들의 발현 강도의 분산은 평균과 더불어 증가한다. Huber 등 (2002)은 분산이 평균에 따라 변하는 이분산을 등분산으로 바꾸어 줄 수 있는 변환을 모색하였다. 전체 강도 범위내에서 등분산을 유도할 수 있는 분산 안정화 방법으로 다음 식 (2.3)과 같은 쌍곡선 아크 사인을 사용하였다.

$$h(x) = \operatorname{arsinh}(a + bx), \tag{2.3}$$

a, b는 슬라이드에 의존하는 매개변수를 나타낸다. 식 (2.3)의 최대 가능도 추정량을 구함으로써 마이크로어레이 실험들 간의 분산-평균의 관계의 안정화를 모색하였다.

2.2. 결측치 대치

마이크로어레이 실험은 다양한 실험 과정을 거치고, 각 과정마다 실험의 오류나 편향를 포함하게 된다. 특히 이미지 분석 단계와 불량 스팟 제거 단계는 이러한 오류나 편향을 포함한 유전자를 제거하여 양질의 자료를 만드는 과정이다. 이러한 단계를 수행하여 얻어진 $n \times p$ 행렬 자료에는 측정된 값과 결측된 값이 존재한다. 전체 실험에서 결측된 값이 포함된 유전자를 모두 분석에서 제외한다면 많은 정보의 손실을 가져온다. 이를 해결하기 위해서 결측치 비율이 일정 수준을 넘지 않는 유전자에 한해 결측치를 대치하는 방법이 결측치 대치 과정이다. 본 논문에서는 Troyanskaya 등 (2001)의 K 최근접 이웃 방법과 Oba 등 (2003)의 베이지안 주성분 분석 방법을 사용한다.

2.2.1. K 최근접 이웃 방법 Troyanskaya 등 (2001)이 제안한 방법으로 결측치를 가지는 유전자의 나머지 발현값의 프로파일과 결측치를 가지지 않는 유전자들 중에 발현값의 프로파일이 유사한 유전자들을 이용하여 결측치를 대치하는 방법이다. 결측치를 가진 유전자와 결측치를 가지지 않는 가장 가까운 K개의 유전자를 이용하여, 유사성을 가중치로 한 가중 평균으로 결측치를 보정한다. 여기서 유사성의 기준으로는 유클리드 거리, 상관계수 등이 사용된다.

2.2.2. 베이지안 주성분 분석 방법 Oba 등 (2003)이 제시한 방법으로 베이즈 추론의 계층 모형과 반복 알고리즘을 사용하여 잠재 변수를 추정하고, 결측치를 대치하는 방법이다. 베이지안 주성분 분석 대치법은 주성분 회귀분석, 베이지안 추정, 기대-극대화(Expectation-Maximization)의 순서로 구성된다. 이 과정들로부터 계층 모형에서 가정한 사전 분포와 기대-극대화 유사 반복 알고리즘을 이용, 모수의 사후 분포 및 결측치의 분포를 추정하여 그 기대값으로 결측치를 대치한다.

2.3. t_3통계량 방법

처리군과 대조군의 두 집단의 차이를 검색하기 위한 실험에서는 짝을 이룬 자료나 두 집단이 독립적으로 관측된 자료를 사용하여 t검정을 수행한다. 만약, 짝을 이룬 자료를 이용한 실험을 수행할 경우 실험 과정에서 발생되는 오류나 시료의 부족으로 결측치가 발생할 수 있다. 표 3.1은 cDNA 마이크로어레이 실험에서 간접 설계 사용시 나타날 수 있는 부분적으로 짝을 이룬 자료 구조를 나타낸다. X, Y는 각각 공통 준거대 정상 조직, 공통 준거대 암 보합시 나타나는 유전자 발현 강도이고, U, V는 각각 X, Y의 독립적 복제 관찰치이다. Kim 등 (2005)은 표 3.1과 같은 부분적으로 짝을 이룬 자료에서 특이 발현 유전자를 검색할 수 있는 t_3통계량을 다음 식 (2.4)와 같이 제안하였다.

$$t_3 = \frac{n_1 \overline{D} + n_H \left(\overline{V} - \overline{U} \right)}{\sqrt{n_1 S_D^2 + n_H^2 \left(\frac{1}{n_2} S_U^2 + \frac{1}{n_3} S_V^2 \right)}}, \tag{2.4}$$

단, $\overline{D} = 1/n_1 \sum_{j=1}^{n_1} D_j$, $D_j \equiv X_j - Y_j$, $\overline{U} = 1/n_2 \sum_{k=1}^{n_2} U_k$, $\overline{V} = 1/n_3 \sum_{l=1}^{n_3} V_l$이고, S_D^2, S_U^2, S_V^2은 각각 D, U, V 표본 분산이다. n_1, n_2, n_3는 각각 표 3.1에서 보듯이 짝을 이룬 자료의 크기, Y가 결측된 자료의 크기, X가 결측된 자료의 크기를 나타내며, n_H는 n_2와 n_3의 조화 평균이다. 식 (2.4)의 t_3-통계량은 귀무가설하에서 근사적으로 $N(0,1)$을 따른다.

표 3.1. 대장암 자료와 위암 자료에서 정상 조직과 암 조직 모두 실험한 짝을 이룬 자료와 둘 중 하나가 결측된 자료의 예수

보합				
공통준거대 정상 조직	공통준거대 암	例數	대장암 例數	위암 例數
X	Y	n_1	36	85
U	결측	n_2	19	13
결측	V	n_3	32	11

2.4. 일치도

두 집합의 A, B의 일치도로써 다음 식 (2.5)의 M_{AB}를 사용한다.

$$M_{AB} = \frac{C(A \cap B)}{C(A \cup B)} \tag{2.5}$$

단, $C(E)$는 유한집합 E의 원소의 갯수를 나타낸다.

3. 자료 설명 및 사전 처리 방법

본 논문에서 사용한 마이크로어레이 자료는 연세대학교 의과대학 암전이 연구 센터에서 실험한 자료로써 11개의 암 세포 주를 이용한 간접 설계로 실험되었으며, 한 환자의 암과 암 주변의 정상이라 판단되는 조직을 채취하여 짝을 이룬 설계로 마이크로어레이 실험을 실시하였다. 슬라이드에는 32개의 프린트 팁을 이용한 총 17,664개의 스팟이 점적되어 있다. 표 3.1은 대장암과 위암 각각을 대상으로 실시한 마이크로어레이 실험에서 예수를 나타낸다. 대장암의 경우 36예에 대하여 완전하게 짝을 구성하지만, 19예에 대하여서는 X 자료만, 또 다른 32예는 Y자료만 관측되었다. 이 자료를 이용하여 그림 3.1의 사전 처리 과정 및 자료 분석을 수행한다.

그림 3.1은 마이크로어레이 자료를 이용하여 특이 발현 유전자를 검색하는 분석 단계를 나타낸 그림이다. 사전 처리 분석 과정은 총 12개의 방법으로 구성되었고, 그림 3.1에서는 크게 6개의 방법과 결측치 대치에 따른 각각 2개의 방법(K: K 최근접 이웃 방법, B: 베이지안 주성분 분석 방법)으로 세분화하였다. 방법1, 방법4는 분산화 안정 표준화 방법을 이용한 절차이다. 방법1은 분산 안정화 방법으로 표준화를 수행한 후, 실험과정에서 발생된 오류가 있는 불량 스팟을 제거하고, 전체 실험에서 결측치 비율이 20% 이상인 유전자를 선별하여 결측치 대치(K 최근접 이웃 방법, 베이지안 주성분 분석)를 수행하는 과정이다. 방법4는 방법1과 동일한 사전 처리 방법을 사용하지만 사전 처리 순서가 불량 스팟 제거, 비결측 비율 결정, 결측치 대치, 분산 안정화 표준화 순으로 구성된다. 방법2, 방법5는 전체 표준화 방법을 이용한 절차이다. 방법2은 실험 과정에서 발생된 오류가 있는 스팟을 제거하고, 전체 표준화 과정을 수행 후, 전체 실험에서 결측치 비율이 20% 이상인 유전자를 선별하고, 결측치 대치(K 최근접 이웃 방법, 베이지안 주성분 분석)를 수행하는 과정이다. 방법5는 방법2와 동일한 사전 처리 방법을 사용하지만 사전 처리 순서가 비결측 비율 결정, 결측치 대치, 전체 표준화 순으로 구성된다. 방법3, 방법6은 블럭별 국소 평활 표준화 방법을 이용한 절차이다. 방법3은 실험 중 발생된 오류를 포함한 스팟을 제거하고, 블럭별 국소 평활 표준화 과정을 수행 후, 전체 실험에서 결측치 비율이 20% 이상인 유전자를 선별하고, 결측치 대치(K 최근접 이웃 방법, 베이지안 주성분 분석)를 수행하는 과정이다. 방법6은 방법3과 동일한 사전 처리 방법을 사용하지만 사전 처리 순서가 비결측 비율 결정, 결측치 대치, 블럭별 국소 평활 표준화 순으로 구성된다. 위의 사전 처리 과정 후 중복 점적된 유전자는 평균으로 대치한다. 최종적으로 특이 발현 유전자 검색을 위해 t_3통계량과 Benjamini와 Hochberg의 다중 검정 절차를 사용한다.

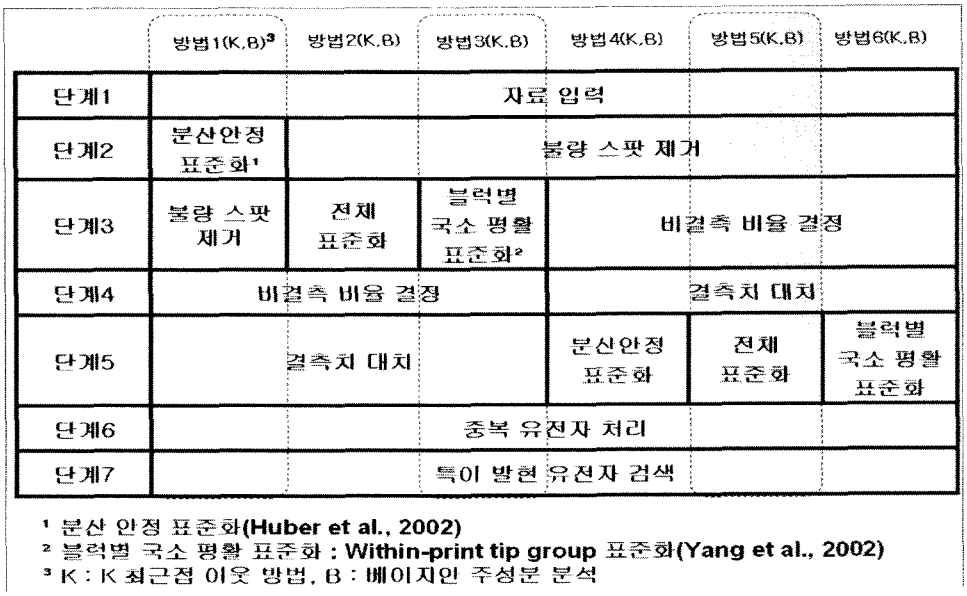

그림 **3.1.** 특이 발현 유전자 검색을 위한 7단계의 마이크로어레이 분석 과정과 불량 스팟 제거, 표준화, 비결측 비율 결정, 결측치 대치의 사전 처리 순서를 조합한 12가지의 사전 처리 방법

4. 결과

짝을 이룬 자료와 결측된 자료가 병합된 암과 정상 조직을 구별할 수 있는 특이 발현 유전자 선별을 위해 t_3통계량의 p값을 이용하여 Benjamini와 Hochberg의 다중 검정 절차를 사용하여 FDR(False Discovery Rate)을 얻는다. 표 4.1–4.4은 FDR($\leq$ 0.00001, $\leq$ 0.0001, $\leq$ 0.001)에 따른 특이 유전자 검색 결과이다. 표준화 방법과 결측치 보정 방법에 따른 결과와 상기 결정한 유의 수준에서 유의하게 검색된 유전자의 개수의 일치도를 나타낸 것이 표 4.1–4.4이다. 그림 3.1의 6개 사전 처리 방법에서 l번째 방법과 l'번째 방법 각각에서 검색된 특이 발현 유전자 군간 일치도를 $M_{ll'}$으로 표기한다($l, l' = 1, \ldots, 6, l \neq l'$). 표 4.1–4.2는 대장암 환자의 36명의 짝을 이룬 자료와 독립 자료(암 32명, 정상 19명)를 K 최근접 이웃 방법 방법과 베이지안 주성분 분석으로 결측치를 대치 후 특이 발현 유전자의 개수를 검색한 결과와 일치도 결과이다. 전체 표준화 방법(방법2, 방법5)의 사전 처리 순서에 따른 특이 발현 유전자의 일치도가 평균 0.9891로 가장 높게 측정되었다. 분산 안정화 표준화 방법(방법1, 방법4)의 경우는 평균 0.8622의 일치도가 측정되어 세 가지 표준화 방법 중 가장 낮게 측정되었다. 결측치 보정 방법에 차이에 따른 일치도는 거의 차이를 보이지 않는다.

표 4.3–4.4는 위암 환자의 85명의 정상 조직과 암 조직의 짝을 이룬 자료와 독립 자료(암 11명, 정상 13명)를 K 최근접 이웃 방법과 베이지안 주성분 분석 방법으로 결측치를 대치 후 특이 발현 유전자의 개수를 검색한 결과와 사전 처리간 일치도 결과이다. 전체 표준화 방법(방법2, 방법5)이 사전 처리 순서에 따른 특이 발현 유전자의 일치도가 평균 0.9942로 가장 높게 측정되었다. 분산 안정화 표준화 방법(방법1, 방법4)의 경우는 평균 0.8874의 일치도가 세 가지 표준화 방법 중 가장 낮게 측정되었다. 결측치 보정 방법에 따른 일치도는 동일한 FDR에 따라 유사한 결과를 보인다. 표 4.1–4.4를 종합할 때 전체 표준화 방법(방법2, 방법5)이 사전 처리 방법에 따라 특이 발현 유전자 군의 일치도가 안정적인 방

표 4.1. 대장암 환자 자료(K 최근접 이웃 방법으로 결측치 대치)에서 Benjamin와 Hochberg (1995)의 FDR을 사용하여 검색한 특이 발현 유전자 수와 표준화 방법간의 일치도

FDR	분산 안정화 표준화			전체 표준화			블럭별 국소 평활		
	방법1	방법4	일치도(M_{14})	방법2	방법5	일치도(M_{25})	방법3	방법6	일치도(M_{36})
≤ 0.00001	4419	4232	0.8552	4326	4356	0.9885	4464	4441	0.9745
≤ 0.0001	5056	4881	0.8616	4988	5007	0.9898	5112	5094	0.9771
≤ 0.001	5914	5722	0.8680	5805	5821	0.9901	5944	5901	0.9735

표 4.2. 대장암 환자 자료(베이지안 주성분 분석 방법으로 결측치 대치)에서 Benjamin and Hochberg의 FDR을 사용하여 검색한 특이 발현 유전자 수와 표준화 방법간의 일치도

FDR	분산 안정화 표준화			전체 표준화			블럭별 국소 평활		
	방법1	방법4	일치도(M_{14})	방법2	방법5	일치도(M_{25})	방법3	방법6	일치도(M_{36})
≤ 0.00001	4435	4265	0.8570	4368	4384	0.9859	4508	4490	0.9758
≤ 0.0001	5086	4921	0.8638	5028	5047	0.9903	5143	5153	0.9754
≤ 0.001	5947	5764	0.8677	5857	5873	0.9898	5980	5971	0.9750

표 4.3. 위암 환자 자료(K 최근접 이웃 방법 방법으로 결측치 대치)에서 Benjamin and Hochberg의 FDR을 사용하여 검색한 특이 발현 유전자 수와 표준화 방법간의 일치도

FDR	분산 안정화 표준화			전체 표준화			블럭별 국소 평활		
	방법1	방법4	일치도(M_{14})	방법2	방법5	일치도(M_{25})	방법3	방법6	일치도(M_{36})
≤ 0.00001	5902	6170	0.8845	5830	5867	0.9883	6143	6126	0.9849
≤ 0.0001	6660	6931	0.8879	6631	6639	0.9933	6966	6933	0.9830
≤ 0.001	7658	7884	0.8946	7613	7624	0.9956	7951	7934	0.9849

표 4.4. 위암 환자 자료(베이지안 주성분 분석 방법으로 결측치 대치)에서 Benjamin and Hochberg의 FDR을 사용하여 검색한 특이 발현 유전자 수와 표준화 방법간의 일치도

FDR	분산 안정화 표준화			전체 표준화			블럭별 국소 평활		
	방법1	방법4	일치도(M_{14})	방법2	방법5	일치도(M_{25})	방법3	방법6	일치도(M_{36})
≤ 0.00001	6194	6449	0.8856	6141	6152	0.9952	6438	6409	0.9837
≤ 0.0001	6953	7203	0.8859	6891	6897	0.9962	7237	7222	0.9831
≤ 0.001	7938	8243	0.8856	7896	7897	0.9965	8196	8180	0.9840

표 4.5. 대장암, 위암 환자 자료에서 K 최근접 이웃 방법 방법, 베이지안 주성분 분석 방법으로 결측치 대치 자료에서 t_3 통계량을 이용한 특이 발현 유전자 상위 50개의 일치도

자료 및 결측치 대치 방법	일치도(M_{14})	일치도(M_{25})	일치도(M_{36})
대장암(K 최근접 이웃 방법)	0.62	1	0.98
대장암(베이지안 주성분 분석 방법)	0.60	1	1.00
위암 (K 최근접 이웃 방법)	0.72	1	0.98
위암 (베이지안 주성분 분석 방법)	0.68	1	0.98

법이고, 분산 안정화 표준화 방법(방법1, 방법4)이 가장 민감한 방법이라 할 수 있다. 결측치 보정 방법 중 베이지안 주성분 분석 방법이 K 최근접 이웃 방법 보다 특이 발현 유전자를 많이 검색하지만 일치도는 유사한 결과를 보인다.

정상 조직과 암 조직을 이용한 마이크로어레이 실험에서 생물학자들은 상위 몇 개의 유전자들에 많은 관심을 가지고 있다. 표 4.5는 각 자료 별, 결측치 대치 방법 별 t_3통계량을 이용한 특이 발현 유전자의 상

위 50개의 일치도 계산한 결과이다. 그림 3.1의 6개 사전 처리 방법에서 l번째 방법과 l'번째 방법 각각에서 검색된 특이 발현 유전자 군간 일치도를 $M_{ll'}$으로 표기한다. 방법2와 방법5의 경우 모든 자료 및 결측치 대치 방법에서 일치도가 1로 같고, 방법1과 방법4의 경우 평균 0.655의 일치도를 보인다. 표 4.5의 결과는 표 4.1–4.4와 같이 일치도의 민감도가 같은 결과를 보임을 확인하였다.

5. 결론 및 토의

표준화와 결측치 대치 각각에 대한 연구는 활발히 진행되었지만, 두 절차를 연결하여 최종 분석에 사용될 자료를 만드는 사전 처리 순서에 대한 연구는 미흡하다. 본 논문은 사전 처리 순서에 대한 정상 조직과 암의 특이 발현 유전자의 일치도 측정을 위해 대장암 환자와 위암 환자의 마이크로어레이 실험 자료를 이용하여 세 가지 표준화 방법, 두 가지 결측치 보정 방법과 그 순서의 조합으로 12 가지의 사전 처리 과정으로 구성하였다. 각 사전 처리 단계를 수행하여 얻어진 자료를 바탕으로 t_3통계량을 이용하여 정상 조직과 암에서 특이적으로 발현하는 유전자의 개수를 Benjamini와 Hochberg의 다중 검정의 동일한 FDR 하에서 검색하고, 일치도로 식 (2.5)를 사용하였다. 표준화 방법에 따른 방법 별 일치도가 약 10% 차이를 보이고, 동일한 표준화 방법과 동일한 결측치 대치 방법의 순서에 따른 일치도 차이는 미비함을 확인하였다. 이는 사전 처리를 구성하는 방법에 따라 특이 발현 유전자가 다소 영향을 받는다는 사실을 보여준다. Wit과 McClure (2004, 71쪽)은 사전 처리 순서를 공간(보합) 보정, 배경 보정, 염료 효과 보정, 실험내 블럭간 척도 보정, 조건간(실험간) 척도 보정의 과정을 제안하였다. Wit과 McClure의 사전 처리 과정은 마이크로어레이 내부의 블럭별 국소 지역에서 마이크로어레이 간의 넓은 지역으로 확장하는 방법이 통계적 분석시 안정적이라 보고하고있다. 본 논문의 결과는 전체 표준화를 이용한 사전 처리 방법이 Wit과 McClure가 제시한 마이크로어레이 내부의 국소 지역에서 넓은 지역으로 확장하는 절차와 일치하는 결과를 나타내고 있다. 그러나, Huber 등 (2002)이 제시한 분산 안정화 표준화를 이용한 사전 처리 방법의 경우 마이크로어레이 간의 넓은 지역을 보정하는 절차로 일치도 측면에서도 다소 불안정하다. 결측치 보정 방법 간의 사전 처리 순서에 따른 일치도는 유사한 결과를 보인다. 상위 50개의 유전자를 각 방법별로 선별하여 일치도를 조사한 경우 표 4.5의 결과를 얻었고, 표 4.1–4.4의 결과와 유사한 결과를 보인다. 본 논문에서 사용한 방법 외 현재까지 제시된 다양한 표준화 (Bolstad 등, 2003; Smyth 등, 2003; Workman 등, 2002)및 결측치 대치 방법 (Bø 등, 2004; Kim 등, 2004; Ouyang 등, 2004; Sehgal 등, 2005)을 고려할 때, 사전 처리 단계의 순서 및 방법에 있어서 다양한 조합을 구성할 수 있음은 명백하다. 향후 이러한 여러 가지 방법들 중 어떤 방법을 선택하고, 어떤 순서로 수행하는 것이 가장 정확하고 의미 있는 결과를 가져올 것인가에 대한 연구가 진행되어야 할 것이다.

참고문헌

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society Series B,* **57**, 289–300.

Bolstad, B. M., Irizarry, R. A., Astrand, M. and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics*, **19**, 185–193.

Bø, T. H., Dysvik, B. and Jonassen, I. (2004). LSimpute: Accurate estimation of missing value in microarray data with least squares methods, *Nucleic Acide Research,* **32**, e34.

Ge, Y., Dudoit, S. and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis, *Test*, **12**, 1–77.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics,* **18**, S96–S104.

Kim, B. S., Benner, A. and Kim, S. C. (2006). Development of a molecular prognostic indicator of gastric cancer using the penalized Cox regression, <한국통계학회 2006년 춘계학술발표회 논문집>, 41.

Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y. and Chung, H. C. (2005). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer, *Bioinformatics*, **21**, 517–528.

Kim, H., Golub, G. H. and Park, H. (2005). Missing value estimation for DNA microarray gene expression data: Local least squares imputation, *Bioinformatics,* **21**, 187–198.

Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K. I. and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data, *Bioinformatics,* **19**, 2088–2096.

Ouyang, M., Welsh, W. J. and Georgopoulos, P. (2004). Gaussian mixture clustering and imputation of microarray data, *Bioinformatics,* **20**, 917–923.

Sehgal, M. S., Gondal, I. and Dooley, L. S. (2005). Collateral missing value imputation: A new robust missing value estimation algorithm for microarray data, *Bioinformatics,* **21**, 2417–2423.

Smyth, G. K. and Speed, T. (2003). Normalization of cDNA microarray data, *Methods,* **31**, 265–273.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics,* **17**, 520–525.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*, John Wiley & Sons, New York, 116–117.

Wit, E. and McClure, J. (2004). *Statistics for Microarrays: Design, Analysis and Inference*, Wiley, New York, 71.

Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielser, H. B., Saxild, H. H., Nielsen, C., Brunak, S. and Knudsen, S. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments, *Genome Biology,* **3**, research0048.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research,* **30**, e15.

A Concordance Study of the Preprocessing Orders in Microarray Data

Sang-Cheol Kim[1] · Jae-hwi Lee[2] · Byung Soo Kim[3]

[1]Department of Applied Statistics, Yonsei University;
[2]Department of Applied Statistics, Yonsei University;
[3]Department of Applied Statistics, Yonsei University

(Received September 2008; accepted November 2008)

Abstract

Researchers of microarray experiment transpose processed images of raw data to possible data of statistical analysis: it is preprocessing. Preprocessing of microarray has image filtering, imputation and normalization. There have been studied about several different methods of normalization and imputation, but there was not further study on the order of the procedures. We have no further study about which things put first on our procedure between normalization and imputation. This study is about the identification of differentially expressed genes(DEG) on the order of the preprocessing steps using two-dye cDNA microarray in colon cancer and gastric cancer. That is, we check for compare which combination of imputation and normalization steps can detect the DEG. We used imputation methods(K-nearly neighbor, Baysian principle comparison analysis) and normalization methods(global, within-print tip group, variance stabilization). Therefore, preprocessing steps have 12 methods. We identified concordance measure of DEG using the datasets to which the 12 different preprocessing orders were applied. When we applied preprocessing using variance stabilization of normalization method, there was a little variance in a sensitive way for detecting DEG.

Keywords: Concordance measure, imputation, microarray, normalization, preprocessing, *t3* statistic.

This research was supported by the Yonsei University College of Business and Economics Research Fund of 2007.

[3]Corresponding author: Professor, Department of Applied Statistics, Yonsei University, 262 Seongsanno, Seodaemoon-Gu, Seoul 120-749, Korea. E-mail: bskim@yonsei.ac.kr