Journal of the Korean Data & Information Science Society
2012, **23**(1), 1–11

한국데이터정보과학회지

마이크로어레이 자료에서 생존과 유의한 관련이 있는 유전자집단 검색†

이선호[1] · 이광현[2]

[12]세종대학교 응용통계전공

접수 2011년 10월 15일, 수정 2011년 11월 25일, 게재확정 2011년 12월 6일

요 약

환자의 생존시간과 함께 유전자 마이크로어레이 자료가 주어진 경우 생존에 유의한 영향을 미치는 대사경로를 찾는 방법을 연구하였다. 기존의 방법인 유전자 집합 농축도 분석, 글로벌 검정과 왈드 형태 검정을 비교 분석하였고, 치환을 통하여 p값을 구하는 단점을 개선한 수정된 왈드 형태 검정을 제안하였다. 모의실험과 실제자료 분석을 이용하여 새로운 방법의 적용 가능성을 보였다.

주요용어: 마이크로어레이, 생존분석, 왈드 통계량, 유전자집단.

1. 서론

동시에 수 만개 유전자를 관찰할 수 있는 마이크로어레이 기술의 개발은 제한적이었던 유전자 발현 정보 연구의 한계를 허물었고 특정 질병의 유전학적 특성과 기전 연구를 활성화시켰다. 광범위해진 유전자 발현 정보를 분석하기 위해 단순한 발현비 비교법에서부터 복잡한 전산학적 기술이 요구되는 통계 방법에 이르기까지 유전자 발현 변화를 평가하는 다양한 데이터 분석법이 개발되고 있다.

초기의 마이크로어레이 분석은 환자들의 임상 결과가 이진자료 (종양/정상, 전이 여부 등)인 표현형 (phenotype)에 따라 특이발현 양상을 보이는 유전자를 개별적으로 찾아내는 단일유전자 위주의 분석으로서 Tusher 등 (2001)의 SAM (Significance Analysis of Microarrays)과 Tibshirani 등 (2002)의 PAM (Prediction Analysis for Microarrays)이 대표적이다. 하지만 단일유전자 분석은 결과 해석이 어려울 수 있고, 동일 질병에 관한 서로 다른 자료의 분석 결과들 사이에 일치성이 낮으며 유전자수가 많아서 생기는 다중 검정의 한계 등이 문제점으로 제기되었다 (Subramanian 등, 2005).

단일유전자 분석의 단점을 보완하기 위해 염색체 위치가 같거나 동일한 대사기능을 수행하는 유전자집단을 대상으로 질병의 발생에 유의한 역할을 하는 집단을 찾아내는 유전자집단 분석법 (gene set analysis)이 대두하게 되었다. Pavlidis 등 (2002)의 기능이 같은 유전자들에 점수를 주는 세 가지 방법을 시작으로 유전자 집합 농축도 분석 (Gene Set Enrichment Analysis; GSEA) (Mootha 등, 2003; Subramanian 등, 2005), 모수적 방법을 이용한 유전자 집합 농축도 분석 (Parametric Analysis of Gene set Enrichment; PAGE) (Kim과 Volsky, 2005)과 마이크로어레이 자료의 유전자집단 유의성 분석 (Significance Analysis of Microarrays - Gene Set; SAM - GS) (Dinu 등, 2007) 등이 대표적이다.

---

† 이 논문은 2009년도 세종대학교 교내연구비 지원에 의한 논문임.
[1] 교신저자: (143-747) 서울시 광진구 군자동 98, 세종대학교 수학통계학부, 교수.
E-mail: leesh@sejong.ac.kr
[2] (143-747) 서울시 광진구 군자동 98, 세종대학교 대학원 응용통계전공, 석사.

환자의 임상결과가 이진표현형에 국한되었던 분석법도 연속형과 생존자료 형태의 표현형을 대상으로 확장하여 발전되었으며 대표적인 방법으로는 GSEA, 글로벌 검정 (Global test; GT) (Geoman 등, 2004, 2005)과 왈드 형태 검정 (Wald-type test; WT) (Adewale 등, 2008)이 있다. 그러나 실제로 환자의 표현형이 생존시간인 경우는 축적된 자료가 많지 않으며 중도절단의 문제가 있어 활발히 다루어지지 않았다.

본 논문에서는 종양환자들을 장기간 추적한 생존시간과 그들의 마이크로어레이 자료로부터 생존에 유의한 영향을 끼치는 유전자집단을 검색하는 기존의 방법들을 비교 분석하고 이들의 단점을 보완하여 더 정확하고 빠른 방법을 제안하려 한다.

2절에서는 기존의 분석 방법을 소개하고 모의실험을 통해 방법들 간의 통계적 유의성을 비교해본다. 3절에서는 비모수적 방법의 단점을 개선할 수 있는 새로운 방법을 제안하며, 4절에서 실제 난소암 119예의 마이크로어레이 자료와 환자들의 생존시간을 추적한 자료로부터 204개 유전자집단의 생존과 관련한 유의성을 검정하였다.

2. 기존의 분석방법과 검정력 비교를 위한 모의실험

n개 표본의 유전자 수가 g인 마이크로어레이 발현자료와 생존자료를 통하여 생존에 유의한 영향을 미치는 유전자집단을 찾아보려 한다. 즉, i번째 환자 ($i = 1, 2, \cdots, n$)로부터 유전자자료 $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \cdots, x_{ig})$와 반응변수 $Y_i = (T_i, c_i)$를 얻고 m개 유전자로 구성된 유전자집단 S가 생존과 관련이 있는지 검정하려는 것이다. 여기서 T_i는 생존시간을, c_i는 중도절삭여부를 나타내는 변수로서 중도절삭이 일어난 경우 0의 값을, 그렇지 않은 경우 1의 값을 갖는다.

임상결과가 절삭이 포함된 생존자료 형태인 경우에도 유전자집단 분석이 가능한 GSEA, GT와 WT에 대해 좀 더 자세히 알아보고 모의실험을 통하여 기존 방법들 간의 검정력 차이를 비교해 보도록 하겠다.

2.1. 여러 가지 분석방법

유전자 집합 농축도 분석 (GSEA)

Mootha 등 (2003)은 특이발현 유전자들의 관심 유전자집단 포함 여부를 분석에 반영하는 방법을 개발하였고 Subramanian 등 (2005)은 이를 보완 발전시켜 GSEA-P라는 소프트웨어를 공개하였다. 또한 Barry 등 (2005)은 GSEA를 확장하여 3개 이상, 연속적 또는 생존자료 형태의 임상결과에도 적용가능하도록 하였다.

m개 유전자로 구성된 관심 유전자집단 S에 특이발현 유전자가 얼마나, 어떤 순위로 포함되었는지를 나타내는 통계량 MES (maximum enrichment score)를 사용한 GSEA는 유전자집단 분석의 선두주자였고 아래와 같은 탄탄한 알고리즘으로 인해 많이 사용되었다.

i) 각 유전자의 발현값과 환자 표현형 사이의 연관성을 보여주는 대푯값 r (상관계수, t-통계량, Cox 모형의 회귀계수 등)을 구하고 크기순으로 정렬한다.

ii) 정렬된 순서로 각 유전자의 R_j값을 구한다 ($j = 1, \cdots, g$)

$$R_j = \begin{cases} \dfrac{r_j^d}{\sum_{k\in S} r_k^d} & \text{, 유전자가 S에 속할 때} \\ -\dfrac{1}{g-m} & \text{, 유전자가 S에 속하지 않을 때} \end{cases}$$

단, 가중치 $d \in [0, 1]$

iii) 유전자집단 S의 점수는 $MES_S = \max_{1\le k\le g} \sum_{j=1}^{k} R_j$이다.

환자 표현형이 생존자료인 경우, 시점 t에서 환자 $(i = 1, \cdots, n)$의 j번째 유전자 $(j = 1, \cdots, g)$와 관련한 위험함수는 단변량 Cox 비례위험모형 (Cox, 1972)을 사용하여 $h_i(t|x_{ij}) = h_o(t)\exp(\beta_j x_{ij})$와 같이 표현되며 GSEA에서 j번째 유전자 발현 정보와 생존시간 사이의 연관성을 보여주는 대푯값으로 표준화한 β_j추정량의 절대값 $|\widehat{\beta}_j/s.e.(\widehat{\beta}_j)|$을 사용한다.

가중치가 $d = 0$인 경우는 본래 Mootha 등 (2003)이 제안한 콜모고로프-스미르노프 (Kolmogorov-Smirnov) 통계량에 해당되며 본 연구에서는 가중치가 각각 d =0, 0.5, 1.0인 세 경우의 GSEA0, GSEA5, GSEA10을 비교하였다.

왈드 형태 검정 (WT)

Adewale 등 (2008)은 회귀 모형 구조를 이용한 유전자집단 분석 방법으로 왈드 통계량을 사용한 방법을 제시하였다. 이 방법은 회귀모형에 적합한 모든 형태의 표현형에 적용이 가능하며 다른 알려진 요인의 정보를 사용할 수 있다는 장점이 있다.

크기가 m인 관심 유전자집단에 속한 j번째 유전자 $(j = 1, \cdots, m)$와 임상결과 사이의 연관성을 r_j라 할 때 다음과 같이 왈드 통계량에 기초한 통계량 W를 구한다.

$$W = \sum_{j=1}^{m} \left(\frac{r_j}{s.e.(r_j)} \right)^2$$

환자 표현형이 생존자료일 때 연관성 r은 GSEA와 마찬가지로 단변량 Cox 비례위험모형에서 구한 추정회귀계수 $\widehat{\beta}$를 사용하였고 결국 W는 관심 유전자집단에 속한 각 유전자들의 왈드 통계량의 제곱합을 구한 형태이다. 왈드 통계량은 대표본에서 얻은 최우추정량의 함수로서 표준정규분포를 따르는 것으로 알려져 있지만 Adewale 등 (2008)은 표본치환 (sample permutation)을 사용하여 분석하였다.

글로벌 검정 (GT)

GT는 대사경로 (pathway)에 속한 각 유전자들의 계수에 해당하는 모수들의 임의효과 모형에서 유도한 스코아 검정이다. 처음에는 일반화선형모형에 대한 GT가 제안되어 환자들의 임상 결과가 이진표현형일 때 관심 유전자집단이 결과와 관련 있는지 검정할 수 있었고 (Goeman 등, 2004), 이 아이디어가 발전하여 임상 결과가 생존시간인 경우 Cox 비례위험 모형으로부터 마팅게일 잔차 (martingale residual)를 이용하여 각 대사경로가 환자들의 생존율과 관계있는지 검정할 수 있는 방법이 유도되었다.

Goeman 등 (2005)은 관심 유전자집단 S에 속한 m개 유전자의 Cox 비례위험모형의 회귀계수가 $H_0 : \beta_1 = \cdots = \beta_m = 0$을 만족하는지 검정하는 대신 모든 유전자의 회귀계수는 평균 0, 공통분산 σ^2을 따른다는 가정아래 $H_0 : \sigma^2 = 0$의 가설검정으로 대체하여 스코아검정을 유도한 것이다. $\boldsymbol{X}_{n\times m}$은 n명 환자들의 유전자 발현자료 중 S에 속한 m개 유전자들만의 발현자료, $\widehat{u}_i = \widehat{H}(t_i)$는 환자 i의 관찰시간 t_i까지의 추정 위험함수라 할 때, $\widehat{\boldsymbol{u}} = (\widehat{u}_1, \cdots, \widehat{u}_n)^T$, $\widehat{\boldsymbol{U}} = diag(\widehat{\boldsymbol{u}})$, $\boldsymbol{R} = \boldsymbol{X}_{n\times m}^T \boldsymbol{X}_{n\times m}$, $\boldsymbol{c} = (c_1, \cdots, c_n)^T$, $T = (\boldsymbol{c} - \widehat{\boldsymbol{u}})^T \boldsymbol{R}(\boldsymbol{c} - \widehat{\boldsymbol{u}}) - trace(\boldsymbol{R}\widehat{\boldsymbol{U}})$라 하자. 이때 GT의 통계량 Q는 다음과 같다.

$$Q = \frac{T - \widehat{E}(T)}{\widehat{var}(T)}$$

여기서 T의 기댓값과 분산의 추정량을 유도하는 것은 기술적으로 간단하지는 않지만 (Goemane 등, 2005, 3.5절) R package의 'globaltest'를 사용하면 Q의 값을 쉽게 구할 수 있다.

각 환자의 표현형을 임의로 재분배하는 치환과정을 반복하여 각 통계량의 분포를 구한다. 생성된 분포로부터 MES, W와 Q의 p값을 각각 구한다.

2.2. 모의실험을 통한 분석방법의 비교

모의실험을 통하여 생존율에 유의한 영향을 미치는 유전자집단을 찾아내는 분석방법들인 GSEA0, GSEA5, GSEA10, GT와 WT의 크기와 검정력을 비교해 보았다. 실험방법은 Dinu 등 (2007)이 GSEA와 SAM-GS를 비교하였던 방법과 동일하게 하였으며 Loi 등 (2007)이 분석한 원발성 유방암 환자들의 유전자 발현값, 전이 여부와 전이까지 관찰시간 자료를 미국 국립생물정보센터에서 다운 (www.ncbi.nlm.nih.gov/ geo/query/acc.cgi?acc=GSE6532) 받아 사용하였다.

자료는 284명의 환자, 12133개의 유전자로 구성되어 있고 각 유전자에 대해 단변량 Cox 모형의 회귀계수를 이용하여 왈드 통계량 $\widehat{\beta}/s.e.(\widehat{\beta})$과 p값을 구하였다 (그림 2.1).

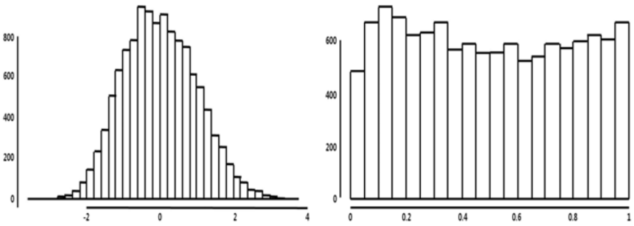


그림 2.1 12133개 유전자의 왈드 통계량과 p값의 분포

GSEA, GT와 WT의 크기와 검정력 비교를 위하여 이 유전자들로부터 다음과 같이 서로 다른 구성의 크기가 m인 유전자집단 S를 만들었다.

[A] p값<0.05를 만족하는 유전자가 하나도 섞여 있지 않다.
[B-1] p값<0.05를 만족하는 유전자가 $m * r$개 포함.
[B-2] 양의 회귀계수를 갖고 p값<0.05를 만족하는 유전자가 $m * r$개 포함.
[C-1] p값<0.001를 만족하는 유전자가 $m * r$개 포함.
[C-2] 양의 회귀계수를 갖고 p값<0.001를 만족하는 유전자가 $m * r$개 포함.

주어진 S에 대하여 GSEA0, GSEA5, GSEA10, WT와 GT의 통계량 값을 구하고 각 환자의 표현형을 임의로 재분배하는 치환 과정을 반복하여 p값을 계산하였다. 이 때 S의 크기 m은 20과 50, 유의한 유전자 비율 r은 0.1, 0.2로 하였고 1000번의 치환을 수행하였다.

[A]를 만족하는 유전자집단을 100개 생성하여 얻은 p값들 중 0.05 보다 작은 비율이 검정의 크기이고, [B]나 [C]의 유의한 유전자가 포함된 집단을 100개 생성하여 얻은 p값들 중 0.05 보다 큰 비율이 검정력이다.

먼저 [A]의 유의한 유전자를 포함하지 않은 유전자집단에 대한 결과를 보면 WT와 GT는 100개의 유전자집단 중 유의한 군이 하나도 없다고 판단하였으며, GSEA는 가중치가 커질수록 유의하지 않다고 판단하는 경우가 많아져 GSEA10은 100개의 유전자집단 중 생존과 유의한 관련이 있다고 판단된 집단은 하나도 없었다. 또한 생존과 관련있는 유전자로 p값이 0.05 보다 작은 유전자를 10% 포함한 [B]는 모든 방법의 검정력이 미미하였으나 포함율이 20%가 되면서 WT의 검정력이 제일 높게 나타났고 GSEA는 안 좋은 결과를 보였다. [B]에 비해 생존과의 관련성이 좀 더 높은 유전자들 (p값<0.001)을 포함시킨 [C]에서도 WT의 검정력이 GT를 앞섰고 GSEA는 상대적으로 약했다. 특히 GSEA0는 사용할 필요가 없다는 결론이 나올 정도였다. [B-1]과 [B-2]를, 그리고 [C-1]과 [C-2]를 비교해 볼 때 회귀계수의 부호가 모두 양수인 경우가 부호의 구분을 두지 않은 경우에 비해 약간의 검정력 우세를 보였다.

전체적으로는 유의한 유전자의 유의성과 포함비율이 높을수록, 유전자군의 크기가 클수록, 그리고 생존에 영향을 미치는 방향이 같을 때 검정력이 높았다. 또한 WT가 생존과 관련있는 유전자집단을 가장 잘 구분하는 것으로 나타났으며 GT, GSEA10, GSEA5, GSEA0의 순이었다.

표 2.1 모의실험을 통한 5가지 분석 방법의 크기와 검정력

r	m= 20					m= 50				
	GSEA0	GSEA5	GSEA10	WT	GT	GSEA0	GSEA5	GSEA10	WT	GT
[A] 유의한 유전자가 하나도 없는 유전자집단										
0.0	0.05	0.01	0.00	0.00	0.00	0.03	0.01	0.00	0.00	0.00
[B-1] 유의한 유전자 (p값<0.05) $m*r$개 포함										
0.1	0.08	0.05	0.04	0.03	0.10	0.06	0.05	0.04	0.03	0.12
0.2	0.10	0.17	0.23	0.48	0.33	0.04	0.55	0.63	0.76	0.55
[B-2] 양의 회귀계수를 갖는 유의한 유전자 (p값<0.05) $m*r$개 포함										
0.1	0.08	0.04	0.04	0.07	0.08	0.07	0.05	0.04	0.07	0.10
0.2	0.11	0.17	0.23	0.54	0.36	0.04	0.56	0.64	0.84	0.67
[C-1] 매우 유의한 유전자 (p값<0.001) $m*r$개 포함										
0.1	0.08	0.07	0.09	0.85	0.37	0.08	0.07	0.12	1.00	0.49
0.2	0.10	0.28	0.83	1.00	0.86	0.04	1.00	1.00	1.00	0.98
[C-2] 양의 회귀계수를 갖는 매우 유의한 유전자 (p값<0.001) $m*r$개 포함										
0.1	0.09	0.08	0.10	0.87	0.33	0.06	0.08	0.14	1.00	0.89
0.2	0.10	0.27	0.82	1.00	0.84	0.04	1.00	1.00	1.00	0.98

3. 새로운 방법의 제안

2절에서 기존의 분석법을 비교한 결과, GSEA에 비해 WT나 GT의 검정력이 더 좋은 것을 알 수 있었다. 하지만, 세 방법 모두 비모수적 분석방법으로 치환을 사용하여 p값을 구하기 때문에 방대한 양의 마이크로어레이 자료를 분석하는데 많은 시간이 걸린다는 단점이 있다. 본 연구에서는 빠르고 정확한 분석이 가능한 모수적 분석법을 제안하고자 한다.

기존의 방법 중 통계적 유의성이 뛰어났고 이론적으로는 카이제곱분포를 따르지만 마이크로어레이 자료의 특성상 활용되지 못했던 WT를 보정하여 카이제곱분포를 만족하는 새로운 통계량을 제안하려 한다.

3.1. 새로운 통계량 제안

GSEA나 WT는 각 유전자의 왈드 통계량을 분석에 사용하고 있다. 이론적으로 왈드 통계량은 최우추정량의 특성상 표준정규분포를 따르지만 실제 마이크로어레이 분석에서는 만족하지 못하는 경우가 많아 (Mansmann 등, 2005; Liu 등, 2007) Adewale 등 (2008)은 유전자집단에 속한 각 유전자들의 왈드 통계량의 제곱합을 의미하는 WT의 분포로 카이제곱분포 대신 치환을 통해 생성된 분포를 사용하였다. 본 논문에서는 이진표현형 자료에 대하여 이선호 등 (2009)이 제안했던 수정된 t^2 (modified t square; MTS) 방법을 응용하여 모수적 가정을 만족할 수 있는 통계량을 제안하려 한다.

MTS는 각 유전자의 t-통계량과 그의 정규화 점수를 혼합하여 정규분포에 근접한 통계량으로 변환하는 방법을 사용하였지만, 본 논문에서는 t-통계량 대신 왈드 통계량과 그의 정규화 점수를 적용하여 다음과 같은 변환을 시도하였다.

i) 단변량 Cox 모형을 적용하여 각 유전자의 왈드 통계량 $w_i = \widehat{\beta}_i / se(\widehat{\beta}_i)$을 구한다. $(i = 1, \cdots, n)$

ii) w_i의 순위를 이용하여 정규점수 $NS_i = \Phi^{-1}(w_i - 0.375)/(n + 0.25)$를 구한다. 이때, $\Phi^{-1}(\cdot)$는 Blom (1985)의 방식을 이용한 역누적 표준정규분포를 의미한다.

iii) 각 유전자의 대푯값으로 $w_i^M = (w_i + NS_i)/2$를 사용한다.

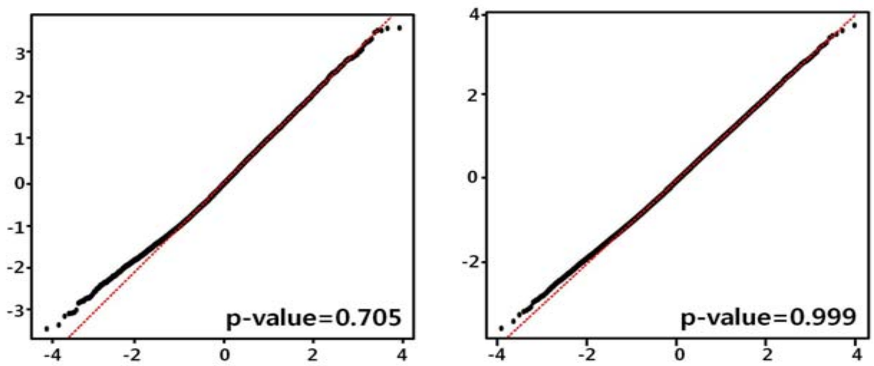
이렇게 구한 수정된 왈드 통계량 w^M은 각 유전자들의 왈드 통계량 순위는 그대로 유지하면서 그들의 분포를 정규분포에 가깝도록 하였고, 유전자집단 S에 속한 m개 유전자의 w^M의 제곱합인 수정된 왈드 형태 통계량 (modified Wald-typed test; MWT)는 카이제곱분포의 가법성에 의해 자유도가 m인 카이제곱분포를 따르게 되어 S의 유의성에 대한 모수적 검정이 가능하게 된다.

$$MWT = \sum_{i \in S} w_i^M \quad \sim \chi^2(m)$$

3.2. 모의실험을 통한 MWT 의 검증

새로운 통계량 MWT가 치환을 이용한 WT와 같은 효과를 갖는지 알아보기 위해 2절에서 실행한 모의실험과 같은 조건으로 비교해 보았다.

먼저, 모의실험에서 사용한 유방암자료의 12133개 유전자들의 왈드 통계량 w과 정규점수를 이용하여 수정한 통계량 w^M가 정규분포에 가까운지 보기 위하여 분위수대조도 (Quantile-Quantile plot)를 그렸으며 (그림 3.1) Kormogorov-Smirnov 검정을 통해 두 통계량이 모두 정규분포를 따른다는 것을 확인하였다.



(a) 왈드 통계량, w (b) 정규점수로 수정 후, w^M

그림 3.1 유방암 자료에서 각 유전자에 대한 정규성 검정을 위한 분위수대조도

2절에서 실시한 모의실험과 동일한 조건에서 MWT의 크기와 검정력을 측정한 후, [A], [B]와 [C]의 경우에 대하여 각 100번의 반복 중 p값=0.05를 기준으로 채택과 기각에 대한 WT와 MWT의 결과 일치율을 조사하였다 (표 3.1). 유의한 유전자를 포함하지 않은 [A]의 결과를 보면 WT나 MWT 모두 유의한 유전자집단이 없다고 판단함으로써 두 방법이 동일한 결과를 보였다. [B]의 p값<0.05의 유의한 유전자를 10∼20% 포함한 유전자집단의 경우에는 최하 83%, 평균 93%의 결과 일치율을 보였으나, [C]의 좀 더 유의성이 큰 유전자들을 포함한 경우에는 두 방법의 결과 일치율이 거의 100%에 가까움을 볼 수 있다.

표 **3.1** WT와 MWT의 크기와 검정력 비교

r	m= 20 WT	MWT	일치율	m= 50 WT	WMT	일치율
[A] 유의한 유전자가 하나도 없는 유전자집단						
	0.00	0.00	100%	0.00	0.00	100%
[B-1] 유의한 유전자 (p값<0.05) $m*r$개 포함						
0.1	0.03	0.03	100%	0.03	0.04	99%
0.2	0.48	0.43	87%	0.76	0.82	94%
[B-2] 양의 회귀계수를 갖는 유의한 유전자 (p값<0.05) $m*r$개 포함						
0.1	0.07	0.04	95%	0.07	0.05	98%
0.2	0.54	0.39	83%	0.84	0.88	92%
[C-1] 매우 유의한 유전자 (p값<0.001) $m*r$개 포함						
0.1	0.85	0.82	91%	1.00	1.00	100%
0.2	1.00	1.00	100%	1.00	1.00	100&
[C-2] 양의 회귀계수를 갖는 매우 유의한 유전자 (p값<0.001) $m*r$개 포함						
0.1	0.87	0.79	90%	1.00	1.00	100%
0.2	1.00	1.00	100%	1.00	1.00	100%

4. 자료 분석 및 결과

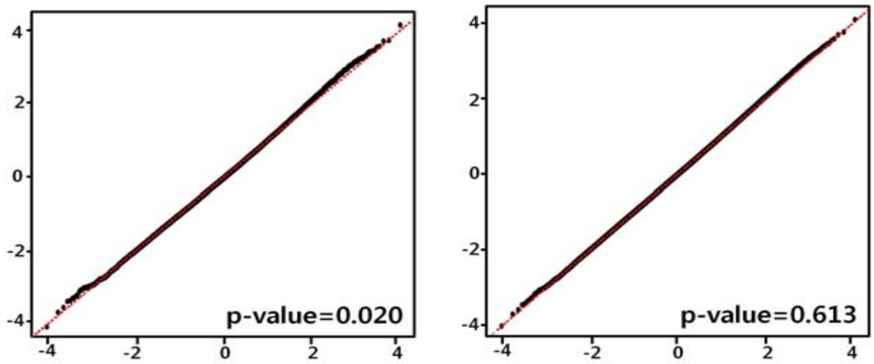
실제 자료를 분석하여 앞에서 논한 GSEA, GT, WT와 MWT 방법들이 유의한 집단을 찾아내는데 어떤 차이를 보이는지 비교하여 보았다.

미국 듀크대학교 병원에서 원발성 난소암이란 진단으로 난소 적출술을 받은 119명 환자들의 22115개 유전자 정보를 담은 마이크로어레이 자료와 그들의 생존시간을 추적 관찰한 생존분석 자료를 다운 (data.cgt.duke.edu/platinum.php)받아 사용하였다 (Dressman 등, 2007; Lee 등, 2011). 또한 유전자들의 기능과 대사경로의 정보가 구축되어 있는 KEGG (Kyoto Encyclopedia of Genes and Genomes; www.genome.jp/kegg)로 부터 204개의 대사경로를 찾아냈다.

먼저 각 유전자들의 왈드 통계량 w와 정규점수로 보정하여 수정한 통계량 w^M가 정규분포를 따르는지 분위수대조도 (그림 4.1)를 그려 본 결과, 보정이전에는 정규분포를 벗어나던 분포 (p값=0.020)가 보정이후 정규분포를 따르는 것 (p값=0.613)을 확인하였다.

모두 204개나 되는 대사경로를 검정하기 때문에 발생하는 다중비교의 문제는 유의하다고 판명되었지만 실제로는 유의하지 않은 대사경로의 비율 (falsely discovered rate; FDR) (Benjamini 와 Yekutieli, 2001)을 이용하여 보정하였고 각 대사경로의 FDR을 나타내는 q값이 0.05 보다 작은지 여부를 각 유전자집단의 유의성 판별 기준으로 삼았다. GSEA 방법으로는 q값<0.05를 만족하는 생존과 관련있는 대사경로를 하나도 찾지 못하였지만 GT와 WT는 모두 3개, MWT는 13개가 유의하였다. 표 4.1은 최소한 한번은 q값<0.05를 만족한 14개 대사경로의 GSEA10, GT, WT와 MWT의 q값이고 q값<0.05를 만족하는 경우는 진하게 표시하였다. 세 방법에서 모두 유의하다는 공통 결과를 얻어낸 대사경로는 Pentose phosphate pathway와 Histidine metabolism 2개 뿐이었다.

MWT는 GT와 WT에서 유의하다는 판단을 내리지 않은 10개의 대사경로를 추가로 찾아냈는데 이들이 실제로 생존과 관련이 있는지 점검해 보기 위하여 대사경로에 속한 유전자들만 사용하여 다변량 Cox 모형을 적합하고 여기서 구한 각 환자의 위험점수 (risk score)를 기준으로 고위험군과 저위험군으로 나누어 로그순위 검정을 하였다. 그 결과 10개 중 4개 대사경로는 p값<0.01을, 1개는 p값<0.05, 1개는 p값<0.1을 만족하였다. 나머지 4개의 대사경로는 Purine metabolism, Cell cycle, Type II di-



(a) 왈드 통계량, w (b) 정규점수로 수정 후, w^M

그림 4.1 난소암 자료의 분위수대조도

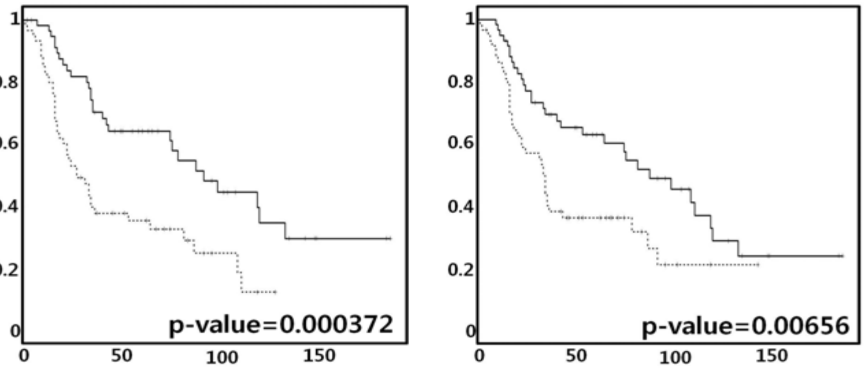
abetes mellitus, Parkinson's disease인데 Purine metabolism의 경우 세 개의 방법에서 모두 유의하다고 판별한 2개의 대사경로와 밀접한 관련이 있음을 KEGG 홈페이지를 통해 확인할 수 있다. 또한 Cell cycle은 Colorectal cancer와 관련된 대사경로이며 당뇨 (Type II diabetes mellitus)와 파킨슨병 (Parkinson's disease)은 다양한 합병증과 암을 유발하는 질병으로 이미 잘 알려져 있다.

그림 4.2는 세 방법에서 모두 유의하다는 결론을 얻은 Pentose phosphate pathway와 MWT에서만 유의하다는 결론을 얻은 Mismatch repair에 관련하는 대사경로에서 환자를 두 군으로 나누어 Kaplan-Meier 생존곡선을 그린 것이고 두 개의 대사경로가 모두 생존과 유의한 연관이 있음을 볼 수 있다 (p값은 각각 0.000372와 0.00656). 전체 204개 유전자집단에 대한 WT와 MWT의 유의성 판별 결과를 비교해 보면 94.6%의 높은 일치도를 보였다.

표 4.1 난소암 자료 분석 결과 (q값<0.05)

대사경로 이름	유전자수	GSEA10	GT	WT	MWT
Pentose phosphate pathway**	39	0.6120	**0.0000**	**0.0000**	**0.0016**
Histidine metabolism*	54	0.6375	**0.0000**	**0.0000**	**0.0007**
One carbon pool by folate	28	0.7013	0.1077	**0.0000**	**0.0095**
Aminophosphonate metabolism*	21	0.7013	**0.0000**	0.1360	0.1129
DNA replication**	52	0.6375	0.1077	0.0765	**0.0000**
Leukocyte transendothelial migration	197	0.8113	0.0742	0.1658	**0.0086**
Mismatch repair**	35	0.8113	0.1077	0.1311	**0.0095**
Tryptophan metabolism*	86	0.7013	0.0612	0.0850	**0.0169**
Purine metabolism	201	0.7013	0.0742	0.1360	**0.0169**
Cell cycle	209	0.8465	0.2451	0.1785	**0.0169**
Nucleotide excision repair**	56	0.8113	0.1286	0.0850	**0.0182**
Type II diabetes mellitus	74	0.8113	0.1166	0.1360	**0.0320**
Colorectal cancer**	165	0.8113	0.0729	0.1360	**0.0391**
Parkinson's disease	34	0.8113	0.3383	0.1714	**0.0499**

* 로그순위검정 결과 p값<0.05 ** 로그순위검정 결과 p값<0.01
q값<0.05 인 경우는 q값을 진하게 표시



Pentose phosphate pathway   Mismatch repair

그림 **4.2** 두 군으로 나누었을 때의 Kaplan-Meier 생존곡선과 로그순위검정 결과

5. 결론

환자 표현형이 생존시간인 경우 마이크로어레이 자료에서 얻은 유전자집단의 유의성 분석은 모두 비모수적인 방법이므로 각 군의 p값을 구하는데 많은 시간이 걸린다. 우리는 각 유전자에 약간의 보정을 해 주는 방법으로 기존의 유전자집단 유의성 분석 방법 중 효과가 좋다고 판단된 왈드 형태 검정법을 대신할 수 있는 모수적 방법을 제안하였다.

실질적으로 동일한 대사경로에 속하는 유전자들은 서로 비슷한 기능을 수행하므로 유전자간 상관관계가 존재할 수 있지만, 유전자집단의 유의성을 다룰 때 상관관계는 부수적인 문제이기 때문에 대부분의 연구에서는 유전자간 유사독립성을 전제로 하였고 본 논문에서 제시한 MWT도 유전자들이 서로 독립이라는 가정 아래에서 유도되었다. 그러나 치환 과정에서는 유전자간의 관계를 유지하기 위하여 표본간 치환을 실시하였다.

앞에서 실시한 난소암 분석에서 204개 유전자집단의 p값을 구하기 위하여 1000번의 치환을 수행하였는데 각 실행마다 모든 유전자를 대상으로 단변량 Cox 모형으로 회귀계수를 구하는 작업이 동반되어 일반 컴퓨터 1대로 일주일 이상의 시간이 걸렸다. 이와 같은 이유로 검정력을 알아보기 위한 모의실험에서 유전자집단은 100개 밖에 생성하지 못한 아쉬움이 있다.

모의실험과 실제 자료 분석을 통하여 WT와 MWT는 최소한 80% 이상 결과가 일치함을 확인하였다. 확실한 생물학적 검증 없이 방법의 효과를 단정 지을 수는 없겠지만 비모수적 방법을 위한 작업시간을 고려할 때 모수적 방법인 MWT는 충분히 경쟁력있는 방법이라 생각한다.

참고문헌

이선호, 이승규, 이광현 (2009). 마이크로어레이 자료분석에서 모수적 방법을 이용한 유전자군의 유의성 검정. <한국통계학회논문집>, **16**, 397-407.

Adewale, A. J., Dinu, I., Potter, J. D. and Yasui, Y. (2008). Pathway analysis of microarray data via regression. *Journal of Computational Biology*, **15**, 269-277.

Barry, W. T., Nobel, A. B. and Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: A structured permutation approach. *Bioinformatics*, **21**, 1943-1949.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165-1188.

Blom, G. (1958). *Statistical estimates and transformed beta-variables*, John Wiley & Sons, New York.

Cox, D. R. (1972). Regression models and life-tables. *Journal of Royal Statistical Society*, **34**, 187-220.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S. Halloran, P. and Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, **8**, 242.

Dressman, H. K., Berchuck, A., Chan, G., Zhai, J., Bild, A., Sayer, R., Cragun, J., Clarke, J., Whitaker, R. S., Li, L. et al. (2007). An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *Journal of Clinical Oncology*, **25**, 517-525.

Goeman, J. J., Van de Geer, S. A., De Kort, F. and Van Houwelingen, H. C. (2004). A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics*, **20**, 93-99.

Goeman, J. J., Oosting, J., Cleton-Jansen, A. M., Anninga, J. K. and Van Houwelingen, H. C. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics*, **21**, 1950-1957.

Kim, S. Y. and Volsky, D. J. (2005). PAGE: Parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 14.

Lee, S. Y., Kim, J. H. and Lee. S. (2011). A comparative study on gene-set analysis methods for assessing differential expression associated with the survival phenotype. *BMC Bioinformatics*, **12,** 377.

Liu, Q., Dinu, I., Adewale, A. J., Potter, J. D. and Yasui, Y. (2007). Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, **8**, 431.

Loi, S., Haibe-Kains, B., Desmedt, C., Wirapati, P., Lallemand, F., Tutt, A. M., Gillet, C., Ellis, P., Ryder, K., Reid, J. F. et al. (2008). Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics*, **22**, 239.

Mansmann, U. and Meister, R. (2005). Testing differential gene expression in functional groups, Goeman's global test versus an ANCOVA approach. *Methods on Information in Medicine*, **44**, 449-53.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E. et al. (2003). PGC-1alpha-responsive genes involved in oxidative phospho- rylation are coordinately downregulated in human diabetes. *Nature Genetics*, **34**, 267-273.

Pavlidis, P., Lewis, D. and Nobel, W. S. (2002). Exploring gene expression data with class scores. *Pacific Symposium on Biocomputing.* 474-485.

Subramanian, A., Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B.L., Gillette, M. A., Paulovich, A., Pomeroy, S.L., Golub, T. R., Lander, E. S., Mesirov, J. P. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**, 15545-15550.

Journal of the Korean Data &
Information Science Society
2012, **23**(1), 1–11

한국데이터정보과학회지

Detecting survival related gene sets in microarray analysis†

Sunho Lee[1] · Kwanghyun Lee[2]

[12]Department of Applied Statistics, Sejong University

Received 15 October 2011, revised 25 November 2011, accepted 6 December 2011

Abstract

When the microarray experiment developed, main interest was limited to detect differentially expressed genes associated with a phenotype of interest. However, as human diseases are thought to occur through the interactions of multiple genes within a same functional category, the unit of analysis of the microarray experiment expanded to the set of genes. For the phenotype of censored survival time, Gene Set Enrichment Analysis(GSEA), Global test and Wald type test are widely used. In this paper, we modified the Wald type test by adopting normal score transformation of gene expression values and developed a parametric test which requires much less computation than others. The proposed method is compared with other methods using a real data set of ovarian cancer and a simulation data set.

Keywords: Censored survival data, gene set analysis, microarray, Wald type statistics.

---

† This work was supported by the faculty research fund of Sejong University in 2009.
1 Corresponding author: Professor, Department of Applied Statistics, Sejong University, Seoul 143-747, Korea. E-mail: leesh@sejong.ac.kr
2 Master, Department of Applied Statistics, Sejong University, Seoul 143-747, Korea.